Tests paramétriques
Comparaison de deux pourcentages

Le test statistique le plus couramment employé pour comparer deux pourcentages est le test du Khi2 (on peut aussi écrire Chi2)

1. Pourcentages et table de contingence

Le test du Khi2 n'utilise pas directement les pourcentages que l'on souhaite comparer. La statistique du test du Khi2 repose sur une somme de différences entre des effectifs observés et des effectifs théoriques, et de ce fait, il se base sur une table de contingence.

Exemple

Prenons le jeu de données « Melanoma » du package MASS, si l'on souhaite comparer le pourcentages de femmes ayant un ulcer (37.3%) au pourcentage d'hommes ayant cette affection (54.4%), les données nécessaires à la réalisation du test du Khi2 ne seront pas directement ces pourcentages, mais la table de contingence suivante :

	Absence d'un ulcer	Presence d'un ulcer
Female	79	47
Male	36	43

2. Echantillons indépendants et échantillons appariés.

L'élément le plus important à définir avant de faire un test du Khi2 est celui de la nature des échantillons sur lesquels ont été estimés les 2 pourcentages à comparer. Est ce qu'il s'agit d'échantillons indépendants ou d'échantillons appariés ?

2. Echantillons indépendants et appariés.

L'élément le plus important à définir avant de faire un test du Khi2 est celui de la nature des échantillons sur lesquels ont été estimés les 2 pourcentages à comparer. Est ce qu'il s'agit d'échantillons indépendants ou d'échantillons appariés ?

Des échantillons sont indépendants si les sujets sur lesquels ont été estimés les deux pourcentages sont différents.

Par exemple donner un médicament actif à un groupe de personnes et donner un placebo inactif à un autre groupe, puis comparer la tension artérielle entre les groupes. Ces échantillons seraient probablement indépendants parce que les mesures sont prises chez des personnes différentes. Quoi que vous sachiez sur la distribution des valeurs dans le premier échantillon, cela ne vous apprend rien sur la distribution des valeurs du second.

2. Echantillons indépendants et appariés.

Des échantillons sont appariés si les sujets sur lesquels ont été estimés les deux pourcentages sont les mêmes.

Par exemple échantillonner la tension de quelques personnes avant et après la prise des médicaments. Ces échantillons seront dépendants parce qu'ils concernent les mêmes individus. Les personnes ayant la tension la plus élevée dans le premier échantillon garderont également la tension la plus élevée dans le second échantillon.

3. Les hypothèses

Quel que soit la variante du test du Khi2 employée, les hypothèses du test sont toujours les suivantes:

- En bilatéral :
- $H_0 : p1 = p2$
  $H_1 : p1 \neq p2$
- En unilatéral :
- $H_0 : p1 = p2$
- $H_1 : p1 > p2$

ou

- $H_1 : p1 < p2$

4. Principe du test du Khi2

La table de contingence contient les effectifs observés des croisements des 2 modalités des 2 variables étudiées (sexe et ulcer dans l'exemple du jeu de données Melanoma).On note généralement ces effectifs "O" pour "observés.

La statistique du test du Khi2 nécessite de calculer un effectif théorique pour chacun des effectifs observés. Ces effectifs théoriques sont calculés sous l'hypothèse nulle, c'est à dire en faisant l'hypothèse que les pourcentages sont égaux. Ils sont généralement notés E pour "expected".

Les effectifs théoriques de chaque combinaison (des modalités des variables catégorielles) sont obtenus en multipliant, dans la table de contingence, l'effectif total de la ligne (nb_l) par l'effectif total de la colonne (nb_c) et en divisant par l'effectif total de la table (nb_T) :

$$E = \frac{(nb_l * nb_c)}{nb_T}$$

Le principe du calcul est simple. Prenons l'exemple de la table de contingence concernant la présence ou l'absence d'un ulcère chez les hommes et les femmes, en modifiant les effectifs de la table de contingence comme ceci :

```
##                      Femmes Hommes
## Présence d'un ulcer     10     40
## absence d'un ulcer      90    160
```

Sous l'hypothèse nulle, les pourcentages d'ulcère chez les hommes et les femmes sont égaux. Cette hypothèse nulle se traduit alors en termes d'effectif. Il s'agit de répartir le nombre total d'ulcère (ici 50) en fonction de la fréquence d'hommes (200/300) et de femmes (100/300) dans l'échantillon étudié.

Au final, les effectifs théoriques de la table de contingence précédente sont :

```
##                          Femmes     Hommes
## Présence d'un ulcer     16.66667   33.33333
## absence d'un ulcer      83.33333   166.66667
```

5. Statistiques et conditions d'application

- *Lorsque les échantillons sont indépendants:*

Dans ce cas, la statistique du test du Khi2 est :

$$\chi_1^2 = \sum_{i=1}^{4} \left(\frac{O_i - E_i}{E_i} \right)^2$$

Le test du Khi2 peut être employé si tous les effectifs théoriques sont >5.
Si au moins un effectif théorique est <5 alors, le test du Khi2 avec correction de Yates, ou bien le test exact de Fisher doivent être employés.
La correction de Yates consiste à soustraire la quantité 0.5 à chaque différence Oi-Ei.

- *Lorsque les effectifs sont appariés*

Dans ce cas, c'est le test de Mac Nemar qui doit être employé. Sa statistique est :

$$\chi_1^2 = \frac{(b-c)^2}{b+c}$$

En utilisant la notation suivante des paires concordantes et discordantes :

```
##                Traitement B
## Traitement A Soulagé Pas soulagé
## Soulagé          "a"       "c"
## Pas soulagé      "b"       "d"
```

Le test de Mac Nemar nécessite que le nombre de paires discordantes soient >10. Dans le cas contraire une correction, dite de de continuité doit être appliquée.

6. Synthèse des tests à employer

- Les échantillons sont indépendants :
- si tous les Ei ≥ 5 : test du Khi2 avec la fonction chisq.test(TC, correct=FALSE)
- si au moins un Ei <5 : test du Khi2 avec correction de continuité (correction de Yates) , avec la fonction chisq.test(TC, correct=TRUE) ou test exact de Fisher avec la fonction fisher.test(TC).

- Les échantillons sont appariés:
- si le nombre de paires discordantes ≥ 10 : test de Mac Nemar avec la fonction mcnemar.test(TC, correct=FALSE)
- sinon : test Mac Nemar avec correction de continuité avec la fonction mcnemar.test(TC, correct=TRUE)

Application

Etudions l'effet du Tabac sur la survenue de cancer sur notre échantillon de 32 sujets dans le jeu de données « Tabac »

15 non fumeurs P_{nf} = ? %

17 fumeurs P_f = ? %

Pourcentage de cancer

tab<-table(Tabac,K)

	NK	K
NF	12	3
F	5	12

prop.table(tab,1)

	NK	K
NF	0,8	0,2
F	0,294	0,706

1. Hypothèses:

H0: $P_{NF} = P_F$ le pourcentage de cancer est identique chez les fumeurs et les non fumeurs

H1: $P_{NF} \neq P_F$ le pourcentage de cancer est différent chez les fumeurs et les non fumeurs

2. Prédictions:

Sous H0 on doit observer P=15/32=47% cancers

	NK	K	
NF	12 7,97	3 7,03	15
F	5 9,03	12 7,97	17
	17	15	32

1. Hypothèses
2. Prédictions

Sous H0 et si les conditions d'application sont respectées

	NK	K	
NF	12 **7,97**	3 **7,03**	15
F	5 **9,03**	12 **7,97**	17
	17	15	32

1. Hypothèses
2. Prédictions

Sous H0 et si les conditions d'application sont respectées

	NK	K	
NF	12 **7,97**	3 **7,03**	15
F	5 **9,03**	12 **7,97**	17
	17	15	32

Conditions

- **E_{ij}>5**
- Indépendance des individus

Confrontation: observation ↔ théorie sous H0

chisq.test(Tabac, K, correct=FALSE)

Confrontation: observation ↔ théorie sous H0

chisq.test(Tabac, K, correct=FALSE)

```
 Pearson's Chi-squared test

data:  Tabac and K
X-squared = 8.1893, df = 1, p-value = 0.004214
```

Confrontation: observation ↔ théorie sous H0

chisq.test(Tabac, K, correct=FALSE)

```
 Pearson's Chi-squared test

data:  Tabac and K
X-squared = 8.1893, df = 1, p-value = 0.004214
```

Test du Chi2 de Pearson

Confrontation: observation ↔ théorie sous H0

chisq.test(Tabac, K, correct=FALSE)

```
 Pearson's Chi-squared test

data:  Tabac and K
X-squared = 8.1893, df = 1, p-value = 0.004214
```

Données

Confrontation: observation ↔ théorie sous H0

chisq.test(Tabac, K, correct=FALSE)

```
 Pearson's Chi-squared test

data:  Tabac and K
X-squared = 8.1893, df = 1, p-value = 0.004214
```

χ^2_0 **sous H0**

Confrontation: observation ↔ théorie sous H0

chisq.test(Tabac, K, correct=FALSE)

```
 Pearson's Chi-squared test

data:  Tabac and K
X-squared = 8.1893, df = 1, p-value = 0.004214
```

Petit « p »

Confrontation: observation ↔ théorie sous H0

chisq.test(Tabac, K, correct=FALSE)

```
 Pearson's Chi-squared test

data:  Tabac and K
X-squared = 8.1893, df = 1, p-value = 0.004214
```

- p<0,05
- Test significatif
- Rejet de H0 au risque α
- Il y a une différence entre les 2 pourcentages
- Dans le sens « les fumeurs développent plus souvent de cancer »

Conditions d'application:

Où trouver les Eij?

E<-chisq.test(Tabac, K, correct=FALSE)

attributes(E)

```
$names
[1] "statistic" "parameter" "p.value"   "method"    "data.name" "observed"
[7] "expected"  "residuals"

$class
[1] "htest"
```

C$expected

Tabac	K = 0	K = 1
0	7.96875	7.03125
1	9.03125	7.96875

Conditions d'application:

remarque

par défaut

chisq.test(Tabac, K, correct=TRUE)

Pearson's Chi-squared test with Yates' continuity correction

data: Tabac and K

X-squared = 6.2839, df = 1, p-value = 0.01218

Pour $3<C_{ij}<5$

Exercice

- fichier TABAC.csv
- Y a-t-il une différence entre le pourcentage de cancer cher les hommes et chez les femmes ?

Références

- Jean Bouyer: *Méthodes statistiques, Médecine-Biologie,* éditions INSERM
- Pr Jean Gaudart ,Cours statistique univariée, Faculté de Médecine de Marseille
- Christophe Chesneau. Introduction aux tests statistiques avec R
- Claire Della Vedova, Comparaison de deux pourcentages avec le logiciel R