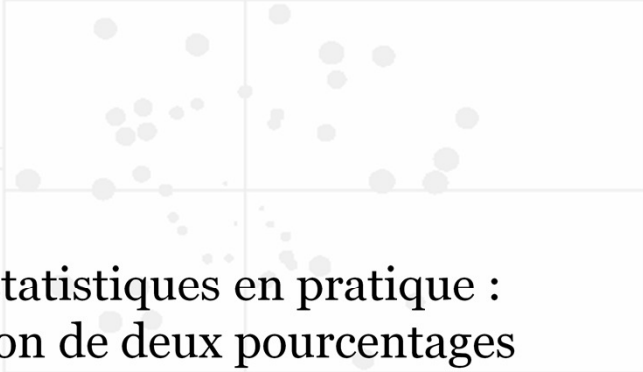


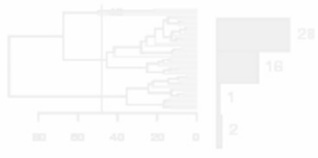
Chapitre 10

Introduction à la statistique avec R





Les tests statistiques en pratique : comparaison de deux pourcentages







1

Pr. Bruno Falissard



[0:01] Il s'agit maintenant de se lancer dans la pratique des tests statistiques. Et nous allons commencer par le cas de la comparaison de deux pourcentages.

Comparaison de deux pourcentages : le test du Chi-2

Introduction à la statistique avec R > Comparaison de 2 pourcentages







- Comparer deux pourcentages : le test du « Chi-2 »
- Attention aux faibles effectifs et aux petits pourcentages → test exact de Fisher



2

Pr. Bruno Falissard



[0:10] Le test de comparaison de deux pourcentages est bien connu parce qu'il porte un nom curieux qui se perd dans l'histoire des statistiques. Il s'agit du test du χ^2 . Avant d'utiliser un test statistique il faut toujours avoir en tête ses conditions de validité. En ce qui concerne le test du χ^2 , il fonctionne si l'effectif sur lequel vous travaillez n'est pas trop petit (c'est-à-dire plus que quelques dizaines) et si les pourcentages ne sont pas trop proches de 0 ni de 100%. Alors ces conditions peuvent sembler un peu vagues. Heureusement, R vérifie automatiquement ces conditions de validité et vous signale s'il y a une difficulté potentielle, auquel cas il existe un test de substitution que l'on appelle le test exact de Fisher.

```
> smp.c$ed.b <- ifelse(smp.c$ed>2,1,0)
> str(smp.c)
'data.frame': 799 obs. of 10 variables:
 $ age      : int  31 49 50 47 23 34 24 52 42 45 ...
 $ prof     : Factor w/ 8 levels "agriculteur",...: 3 NA 7 6 8 6 3 ...
 $ dep.cons : int  0 0 0 0 1 0 1 0 1 0 ...
 $ scz.cons : int  0 0 0 0 0 0 0 0 0 0 ...
 $ grav.cons: int  1 2 2 1 2 1 5 1 5 5 ...
 $ n.enfant : int  2 7 2 0 1 3 5 2 1 2 ...
 $ rs       : int  2 2 2 2 2 1 3 2 3 2 ...
 $ ed       : int  1 2 3 2 2 2 3 2 3 2 ...
 $ dr       : int  1 1 2 2 2 1 2 2 1 2 ...
 $ ed.b     : num 0 0 1 0 0 0 1 0 1 0 ...
```

[0:52] Dans un cours précédent, nous avons créé la variable `ed.b` qui correspond à l'existence d'un haut niveau d'évitement du danger chez les détenus.

Nous allons essayer de tester dans ce cours si la prévalence de la dépression est plus élevée chez les détenus qui ont un haut niveau d'évitement du danger que chez les détenus qui ont un bas niveau d'évitement du danger.

Comparaison de deux pourcentages : le test du Chi-2

```

> table(smp.c$ed.b, smp.c$dep.cons, deparse.level=2, useNA="always")
      smp.c$dep.cons
smp.c$ed.b      0      1 <NA>
0             335    135      0
1              96    126      0
<NA>          51     56      0

```

4

Pr. Bruno Falissard

[1:18] Commençons par quelques statistiques descriptives, notamment dans le but de croiser nos deux variables binaires d'intérêt : existence d'un haut niveau d'évitement du danger, existence d'un diagnostic de dépression. Pour cela nous allons utiliser la fonction `table()`, avec, c'est préférable, l'option `deparse.level=2`, dans le but de renseigner le nom de nos variables et l'instruction `useNA="always"` dans le but de déterminer le nombre de détenus ayant des données manquantes, soit pour la variable évitement du danger, soit pour la variable dépression. On obtient les résultats suivants, avec par exemple, 126 détenus ayant à la fois un haut niveau d'évitement du danger et un diagnostic de dépression. Ces effectifs sont intéressants mais comme on compare des pourcentages, on aimerait avoir des pourcentages plutôt que des effectifs. On peut obtenir cela à partir de la fonction `prop.table()`.

Comparaison de deux pourcentages : le test du Chi-2

```

> table(smp.c$ed.b, smp.c$dep.cons, deparse.level=2, useNA="always")
      smp.c$dep.cons
smp.c$ed.b      0      1 <NA>
      0      335  135      0
      1       96  126      0
      <NA>   51   56      0

> tab <- table(smp.c$ed.b, smp.c$dep.cons, deparse.level=2)
> prop.table(tab, 1)
      smp.c$dep.cons
smp.c$ed.b      0      1
      0 0.7127660 0.2872340
      1 0.4324324 0.5675676

> prop.table(tab, 2)
      smp.c$dep.cons
smp.c$ed.b      0      1
      0 0.7772622 0.5172414
      1 0.2227378 0.4827586

```

4'

Pr. Bruno Falissard

[2:16] Dans un premier temps, nous stockons les résultats issus de la fonction `table()` dans un objet que l'on appelle `tab`. Au passage, on remarque que nous enlevons ici l'option `useNA="always"` dans le but d'avoir des pourcentages de déprimés parmi les détenus ayant un haut niveau ou un bas niveau d'évitement du danger. On oublie l'option `useNA`.

Utilisons donc la fonction `prop.table()`. Nous l'appliquons à l'objet `tab` que nous venons de calculer. Le nombre 1 signifie que nous souhaitons estimer le pourcentage de dépression selon que les détenus ont ou n'ont pas un haut niveau d'évitement du danger. Si nous avions utilisé le nombre 2 à la place du nombre 1, nous aurions le pourcentage contraire, c'est-à-dire le pourcentage de détenus ayant un haut niveau d'évitement du danger selon que ces détenus sont ou ne sont pas déprimés. Nous allons faire ce calcul dans quelques secondes mais d'abord regardons les résultats. Nous avons 28,7% de dépression chez les détenus ayant un bas niveau d'évitement du danger, alors que ce nombre est quasiment le double (56,7%) chez les détenus ayant un haut niveau d'évitement du danger. Nous passons donc de 28% de pourcentage de dépression à 56% de pourcentage de dépression. Certes, il est toujours utile de calculer un p pour objectiver le fait que le hasard puisse expliquer à lui tout seul ou pas expliquer une différence au moins aussi importante. Mais a priori, le p dans une telle situation sera très petit.

Alors, avant de faire ce calcul, comme promis, estimons le pourcentage d'évitement du danger selon que les détenus sont ou ne sont pas déprimés. Nous avons ainsi 48% de détenus ayant un haut niveau d'évitement du danger chez les détenus déprimés alors que ce pourcentage tombe à 22% chez les détenus n'étant pas déprimés.

Comparaison de deux pourcentages : le test du Chi-2





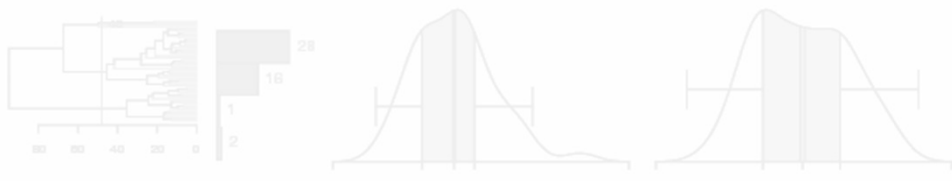
```

> chisq.test(smp.c$ed.b, smp.c$dep.cons, correct=FALSE)

Pearson's Chi-squared test

data:  smp.c$ed.b and smp.c$dep.cons
X-squared = 50.4416, df = 1, p-value = 1.228e-12

```





5

Pr. Bruno Falissard 

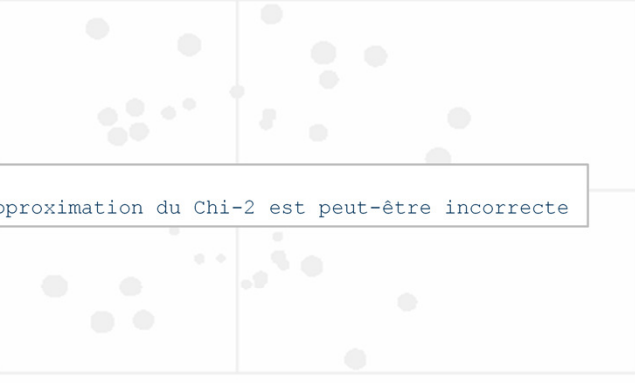
[4:11] Il est grand temps maintenant de faire notre test du chi-2. Pour cela nous allons utiliser la fonction R `chisq.test()`. Nous l'appliquons d'une part à la variable existence oui ou non d'un haut niveau d'évitement du danger (`ed.b`) puis à la variable existence oui ou non d'une dépression (`dep.cons`). Il faut surtout ne pas oublier l'instruction `correct=FALSE`, sinon R propose un test du chi-2 avec correction de continuité, qui est un test plus robuste mais nettement moins puissant. Le résultat est ici, avec un p égal à 10^{-12} . Le p est comme prévu très largement inférieur à 5%. On peut donc affirmer avec un haut niveau de certitude, que le hasard à lui tout seul ne pourrait pas expliquer une telle différence de prévalence de dépression.

Comparaison de deux pourcentages : le test de Fisher

Introduction à la statistique avec R > Comparaison de 2 pourcentages



Message d'avis :
In `chisq.test(x)` : l'approximation du Chi-2 est peut-être incorrecte







6

Pr. Bruno Falissard


[5:04] Nous sommes dans une situation où la taille de l'échantillon est substantielle (plusieurs centaines de sujets) et où les pourcentages comparés ne sont ni proches de 0 ni proches de 100%. Les conditions de validité du chi-2 sont donc parfaitement respectées. Si tel n'avait pas été le cas, R nous aurait prévenus, il nous aurait prévenus à l'aide du message ci-contre.

```
> fisher.test(smp.c$ed.b, smp.c$dep.cons)
```

Fisher's Exact Test for Count Data

data: smp.c\$ed.b and smp.c\$dep.cons

p-value = 2.033e-12

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

2.303664 4.603460

sample estimates:

odds ratio

3.250819

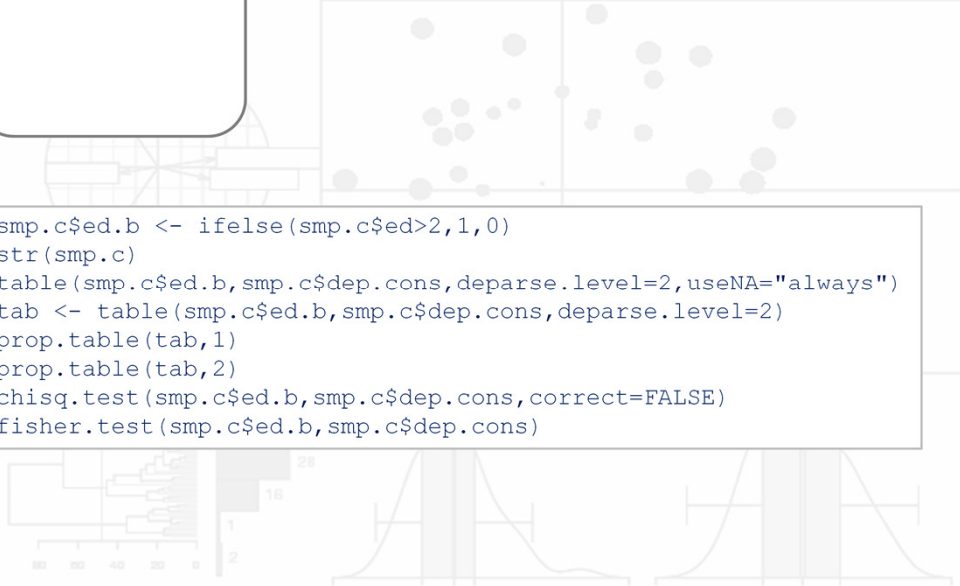
[5:26] Dans une telle situation où on ne peut pas utiliser le test du chi-2, il existe une alternative qui correspond au test dit exact de Fisher. La fonction R correspondante est la fonction `fisher.test()` avec une syntaxe très proche de la syntaxe du test du chi-2. Nous obtenons ici un p qui vaut $2 \cdot 10^{-12}$, voisin du p que nous avons pour le test du chi-2 et ici aussi très largement significatif.

Divers

Introduction à la statistique avec R > Les tests statistiques en pratique



```
smp.c$ed.b <- ifelse(smp.c$ed>2,1,0)
str(smp.c)
table(smp.c$ed.b, smp.c$dep.cons, deparse.level=2, useNA="always")
tab <- table(smp.c$ed.b, smp.c$dep.cons, deparse.level=2)
prop.table(tab, 1)
prop.table(tab, 2)
chisq.test(smp.c$ed.b, smp.c$dep.cons, correct=FALSE)
fisher.test(smp.c$ed.b, smp.c$dep.cons)
```





8

Pr. Bruno Falissard



[5:55] Et voici, pour conclure, un récapitulatif des différentes instructions R que nous avons utilisées pendant ce cours.