TD de statistique : tests du Chi 2

Jean-Baptiste Lamy

6 octobre 2008

1 Test du Chi 2

C'est l'équivalent de la comparaison de moyenne, mais pour les variables qualitatives.

1.1 Cas 1 : comparer les répartitions observées dans des échantillons différents (chi 2 d'indépendance)

Dans ce cas, on distingue en général deux échantillons (ou sous-échantillons) différents (par exemple un groupe "témoin" et un groupe "test") et l'on souhaite déterminer si la répartition des valeurs d'une variable qualitative (par exemple malade - pas malade) est la même dans les deux groupes.

1.1.1 À partir d'un tableau de contingence

	temoin	test	...
1	x1	y1	...
2	x2	y2	...
...	...	...	

Le tableau de contingence doit être entré dans R comme une matrice ; on exécute ensuite le test du Chi 2 sur cette matrice :

```
> matrice = matrix(c(temoin1,temoin2,..., test1,test2,...), ncol=2)
> chisq.test(matrice)
```

1.1.2 À partir d'un tableau de données

Il est possible d'effectuer directement le test du Chi 2 à partir des variables d'un tableau de données. La variable1_qualitative est la variable qualitative étudiée, et la variable2_qualitative est une variable qui permet de distinguer les différents échantillons ou groupe.

```
> chisq.test(variable1_qualitative, variable2_qualitative)
```

1.2 Cas 2 : comparer les répartitions observées sur un même échantillon

Dans ce cas, il n'y a qu'un seul échantillon, et on souhaite comparer plusieurs répartitions observées sur ce même échantillon (par exemple avant et après intervention), les valeurs étant appariées 2 à 2. On utilise alors le test du Chi 2 apparié de McNemar :

```
> matrice = matrix(c(avant1,avant2,..., apres1,apres2,...), ncol=2)
> mcnemar.test(matrice)
```

Ou à partir d'un tableau de données :

```
> mcnemar.test(variable1_qualitative, variable2_qualitative)
```

1.3 Cas 3 : comparer une répartition observée dans un échantillon à une répartition théorique (chi 2 d'ajustement / de conformité)

Pour comparer les proportions observées à des proportions théoriques, on utilise le test du Chi 2 (on est obligé de passer par un tableau de contingence) :

	x
1	x1
2	x2
...	...

```
> matrice = matrix(c(x1,x2,...), ncol=1)
> chisq.test(matrice, p=c(x1_theorique, x2_theorique,...)) # x1_theorique + x2_theorique + ... = 1
```

2 Exercice 1

Une étude a été réalisé sur 100 patients d'un service hospitalier afin de vérifier la relation entre le tabac et les problèmes pulmonaires. Pour cela, nous avons demandé à chaque personne son âge, son sexe, sa situation (célibataire, mariée,...), sa consommation de tabac (nombre de cigarettes par jour), la présence de tabagisme passif, et la présence de problème pulmonaire (cancer du poumon, BPCO,...) chez cette personne.

1. Le tableau de données est enregistré dans le fichier tabac.csv. Charger ce fichier.

   Réponse :

   ```
   > tableau = read.table("tabac.csv", sep=",", header=TRUE)
   > attach(tableau)
   ```

2. Ajouter une colonne "fumeur" de type booléenne.

   Réponse :

   ```
   > tableau["fumeur"] = tabac > 0
   > attach(tableau)
   ```

3. Les fumeurs ont-ils significativement plus de problèmes pulmonaires ?

   Réponse : Les variables fumeur et probleme_pulmonaire sont qualitatives, on utilisera donc un test du Chi 2.

   Nous avons ici 2 groupes de personnes (groupe fumeur et groupe non fumeur), et nous voulons comparer la répartion entre personne ayant des problème pulmonaire et personne sans problème pulmonaire au sein de ses deux groupes. Nous sommes donc dans le cas n°1 (comparer les répartitions observées dans des échantillons différents).

   Nous avons ici un tableau de données, nous allons donc prendre la formule pour les tableaux de données (et pas celle pour les tableaux de contingence).

   ```
   > chisq.test(probleme_pulmonaire, fumeur)
           Pearson's Chi-squared test with Yates' continuity correction
   data :  probleme_pulmonaire and fumeur
   X-squared = 26.8279, df = 1, p-value = 2.224e-07
   ```

   $p < 0.05$, donc il y a une différence significative dans la répartition des problèmes pulmonaires au sein des groupes fumeurs et non fumeurs. En revanche, le test statistique ne donne pas le SENS de cette différence. Pour cela il faut soit calculer les moyennes, soit faire un graphique. Ici, la seconde option est la plus facile :

   ```
   > plot(factor(fumeur), factor(probleme_pulmonaire))
   ```

   Sur le graphique, la barre de gauche correspond aux non-fumeurs et celle de droite aux fumeurs. La partie gris clair correspond aux personnes ayant des problèmes pulmonaires. Nous observons qu'il y a donc plus de problèmes pulmonaires chez les fumeurs que chez les non-fumeurs, et cette différence est significative (d'après le test du Chi 2).

4. Les fumeurs sont-ils significativement plus souvent des personnes célibataires ?

   Réponse : Cette question est du même type que la précédente.

   ```
   > chisq.test(situation=="celibataire", fumeur)
           Pearson's Chi-squared test with Yates' continuity correction
   data :  situation == "celibataire" and fumeur
   X-squared = 4.8867, df = 1, p-value = 0.02706
   > plot(factor(fumeur), factor(situation))
   ```

   On observe qu'il y a plus de célibataires chez les fumeurs, et cette différence est significative ($p < 0.05$). En revanche, le test statistique ne permet jamais de déduire un lien de cause à effet : on ne peut pas en déduire que le tabac rend célibataire !

5. Le pourcentage de fumeur en France est de 31,8% (en 2006, chez les 15-75 ans). La répartition observée ici est-il compatible avec ce pourcentage ? Pourquoi ?

   Réponse : La question porte sur une variable qualitative (fumeur ou non-fumeur), et nous souhaitons comparer cette variable à la proportion théorique de 31,8% de fumeurs. Nous allons donc faire un test du Chi 2, cas n°3 (comparer une répartition observée dans un échantillon à une répartition théorique). Dans ce cas, il n'est pas possible de travailler directement sur le tableau de données et il faut construire le tableau de contingence.

   La fonction summary permet de compter le nombre d'individus pour chacune des valeurs d'une variable.

   ```
   > summary(fumeur)
      Mode   FALSE    TRUE
   logical      51      49
   ```

   Il y a donc 49 fumeurs et 51 non fumeurs, soit le tableau de contingence suivant (tableau à une colonne car nous sommes dans le cas de comparaison à une répartition théorique) :

	échantillon
fumeur	49
non-fumeur	51

```
> matrice = matrix(c(49,51), ncol=1)
> chisq.test(matrice, p=c(0.318, 1-0.318))
        Chi-squared test for given probabilities
data :  matrice
X-squared = 13.641, df = 1, p-value = 0.0002213
```

Il y a une différence significative. Ici, nous avons 49 / (49 + 51) = 49% de fumeurs dans l'échantillon, ce qui est significativement supérieur à la moyenne national de 31,8%. Cela peut s'expliquer car les fumeurs ont plus de problèmes de santé et sont donc plus souvent à l'hôpital.

6. Le tabagisme passif augmente-t-il la probabilité d'avoir des problèmes pulmonaires ?

Réponse :

```
> chisq.test(probleme_pulmonaire, tabagisme_passif)
        Pearson's Chi-squared test with Yates' continuity correction
data :  probleme_pulmonaire and tabagisme_passif
X-squared = 6.9193, df = 1, p-value = 0.008527
```

7. Y a-t-il un lien entre le fait d'être célibataire et le tabagisme passif ?

Réponse :

```
> chisq.test(tabagisme_passif, situation=="celibataire")
        Pearson's Chi-squared test with Yates' continuity correction
data :  tabagisme_passif and situation == "celibataire"
X-squared = 2.3481, df = 1, p-value = 0.1254
```

Pas célibataire => plus de chance de souffrir du tabagisme passif.

8. En France, 60% des fumeurs souffrent de problèmes pulmonaires. Est-ce que cela correspond à ce que nous observons ici ?

Réponse :

```
> matrice = matrix(c(19,31), ncol=1)
> chisq.test(matrice, p=c(.4, .6))
        Chi-squared test for given probabilities
data :  matrice
X-squared = 0.0833, df = 1, p-value = 0.7728
```

9. 75% des problèmes pulmonaires sont liés au tabac. Observons-nous cela dans notre échantillon ?

Réponse : "75% des problèmes pulmonaires sont liés au tabac" signifie que, parmi les personnes ayant des problèmes pulmonaires, 75% d'entre elles sont des fumeurs. Nous allons donc comparer la proportion de fumeurs parmi les personnes de notre échantillon qui ont des problèmes pulmonaires, par rapport au chiffre théorique de 75%.

Pour cela, nous allons d'abord extraire du tableau les personnes ayant des problèmes pulmonaires (Groupe Problème Pulmonaire ou GPP).

```
> gpp = tableau[tableau["probleme_pulmonaire"] == TRUE,]
> attach(gpp)
```

Maintenant nous procédons comme à la question précédente ; comme le tableau "gpp" est attaché, c'est sur lui que portent les calculs.

```
> summary(fumeur)
   Mode   FALSE    TRUE
logical       5      30
> matrice = matrix(c(30,5), ncol=1)
> chisq.test(matrice, p=c(.75, .25))
        Chi-squared test for given probabilities
data :  matrice
X-squared = 2.1429, df = 1, p-value = 0.1432
```

La différence n'est pas significative. Notre échantillon est donc conforme.

10. Dans ce service hospitalier, l'âge moyen est de 40 ans. Est-ce que notre échantillon est conforme à ce chiffre ?

Réponse : Nous nous intéressons ici à l'âge, qui est une variable numérique. Nous allons donc faire une comparaison de moyenne (cf TD4) pour le comparer à la moyenne théorique de 40 ans (cas n°3, comparer une moyenne observée dans un échantillon à une moyenne théorique).

```
> t.test(age, mu=40)
        One Sample t-test
data :  age
t = 1.0505, df = 99, p-value = 0.2961
alternative hypothesis : true mean is not equal to 40
95 percent confidence interval :
 38.77333 43.98667
sample estimates :
mean of x
    41.38
```

p > 0.05 => pas de différence entre la moyenne d'âge théorique et la moyenne observée.

3 Exercice 2

Un laboratoire pharmaceutique a mis au point un nouveau médicament contre la migraine. Pour évaluer l'efficacité de ce traitement, une étude a été réalisée sur 10000 patients. Chaque patient a reçu successivement trois traitements dans un ordre aléatoire : un placebo (lactose), le nouveau médicament, et le médicament "gold standard" de référence (le sumatriptan). Chaque traitement a été testé sur une période de un mois, suivi d'une période d'un mois sans traitement. Pour chacun des traitements, les patients ont indiqués s'ils ont ressenti une amélioration de leur état migraineux, et s'ils ont constaté des effets indésirables.

1. Le tableau de données est enregistré dans le fichier migraine.csv. Charger ce fichier.

   Réponse :

   ```
   > tableau = read.table("migraine.csv", sep=",", header=TRUE)
   > attach(tableau)
   ```

2. Quel est le pourcentage de patient ayant ressenti une amélioration de leur état avec le médicament ? Donner un interval de confiance à 95% de ce pourcentage.

   Réponse :

   ```
   > n=10000
   > moyenne=mean(amelioration_avec_medicament)
   > moyenne
   [1] 0.5685
   > variance = moyenne * (1 - moyenne)
   > moyenne - 1.96 * sqrt(variance / n) ; moyenne + 1.96 * sqrt(variance / n)
   [1] 0.5587924
   [1] 0.5782076
   ```

3. Le nouveau médicament est-il significativement plus efficace que le placébo ?

   Réponse : Nous voulons comparer la proportion de patients ayant eu une amélioration de leur état migraineux, pour le placébo et pour le nouveau médicament. Nous allons donc regarder les variables amelioration_avec_placebo et amelioration_avec_medicament. Il s'agit de variables qualitatives, nous allons donc faire un test du Chi 2. Il s'agit ici de valeurs appariées (chaque patient a eu les trois traitements : nous n'avons PAS un groupe test et un groupe témoin), nous sommes donc dans le cas n°2 (comparer les répartitions observées sur un même échantillon).

   ```
   > mcnemar.test(amelioration_avec_medicament, amelioration_avec_placebo)
          McNemar's Chi-squared test with continuity correction
   data :  amelioration_avec_medicament and amelioration_avec_placebo
   McNemar's chi-squared = 3078.941, df = 1, p-value < 2.2e-16
   > summary(tableau)
   amelioration_avec_placebo  amelioration_avec_medicament
   FALSE :8049                 FALSE :4315
   TRUE :1951                 TRUE :5685
   ```

   Nous constatons qu'il y a plus d'amélioration avec le nouveau médicament qu'avec le placébo, et cette différence est significative ($p < 0.05$).

4. Le nouveau médicament cause-il significativement plus d'effets indésirables que le placébo ?

   Réponse :

   ```
   > mcnemar.test(ei_avec_medicament, ei_avec_placebo)
          McNemar's Chi-squared test with continuity correction
   data :  ei_avec_medicament and ei_avec_placebo
   McNemar's chi-squared = 292.1902, df = 1, p-value < 2.2e-16
   > summary(t)
   ei_avec_placebo ei_avec_medicament
   FALSE :9138      FALSE :8376
   TRUE :862       TRUE :1624
   ```

5. Le nouveau médicament est-il significativement plus efficace que le "gold standard" ?

   Réponse :

   ```
   > mcnemar.test(amelioration_avec_medicament, amelioration_avec_gold)
         McNemar's Chi-squared test with continuity correction
   data :  amelioration_avec_medicament and amelioration_avec_precedent
   McNemar's chi-squared = 1.8716, df = 1, p-value = 0.1713
   ```

6. Le nouveau médicament cause-il significativement moins d'effet indésirable que le "gold standard" ? Que concluez-vous ?

   Réponse :

   ```
   > mcnemar.test(ei_avec_medicament, ei_avec_gold)
          McNemar's Chi-squared test with continuity correction
   data :  ei_avec_medicament and ei_avec_precedent
   McNemar's chi-squared = 7.7204, df = 1, p-value = 0.00546
   > summary(t)
   ei_avec_precedent ei_avec_medicament
   FALSE :8229        FALSE :8376
   TRUE :1771        TRUE :1624
   ```

Le nouveau médicament a un niveau d'efficacité équivalent au "gold standard", en revanche il cause un peu moins d'effets indésirables. On peut cependant s'interroger sur la puissance de l'étude, sans doute excessive (trop de patients !)... et donc sur la pertinence de ce résultat qui met en évidence une différence certes réelle mais très faible...

4 À savoir pour l'examen

Lire un tableau de données depuis un fichier et l'attacher (TD 1)
Ajouter une colonne à un tableau de données (TD 1)
Extraire certaines lignes d'un tableau de données (TD 1)
Tracer des graphiques (TD 2)
Calculer un interval de confiance (TD 3)
- variable qualitative
- variable numérique

Comparer les moyennes de variables numériques (TD 4)
- 2 moyennes issues de deux échantillons
- 2 moyennes issues du même échantillon
- 1 moyenne issue d'un échantillon à une moyenne théorique

Effectuer une régression linéaire (TD 4)
Comparer les répartitions de variables qualitatives (TD 5)
- 2 répartitions issues de deux échantillons
- 2 répartitions issues du même échantillon
- 1 répartition issue d'un échantillon à une répartition théorique