Statistique et Informatique — UFR STAPS
Licence 3 — Université de Paris X - Nanterre

TD n°6 : χ^2

1 χ^2 à la main

On s'intéresse à deux variables qualitatives. On veut estimer s'il y à un lien entre deux variables. Pour cela, on va calculer le χ^2 (prononciation Khi-deux).

1.1 Contexte

Les psychologues sociaux se proposent d'étudier l'impact du toucher sur la consommation des ménagères. A l'entrée d'un supermarché, ils proposent à des ménagères de goûter des pizzas. A certaines, ils proposent simplement de goûter ; à d'autres ils proposent de goûter après avoir établi un léger contact physique (une demi seconde via l'avant bras). Ils observent ensuite si les ménagères achètent ou n'achètent pas la pizza.

1.2 Effectifs observés

Le premier tableau est celui des effectifs observés, il correspond aux résultats observés pendant enquête. Pour cela, on utilise la fonction `table()` (la même que pour les effectifs en univarié) en lui donnant le nom des DEUX variables à croiser (en univarié, on ne lui donnait qu'un seul nom). Dans notre cas, nous cherchons à savoir s'il y a un lien entre [CONTACT] et [ACHAT].

1. Chargez le fichier "ContactAchatPizza.csv" en mémoire, stockez le dans `donnees`.
2. Dressez le tableau croisé des variables [CONTACT] et [ACHAT] et stockez le dans la variable `Tobs`. Pour cela, tapez `Tobs <- table(donnees$contact,donnees$achat)`.

1.3 Effectifs attendus

Le deuxième tableau est celui des effectifs attendus. 58,8% des personnes ont été touché et 63 personnes ont acheté. S'il y avait indépendance entre les variables (et que le contact n'influence pas l'achat), alors il y aurait 58,8% de 63 personnes touchées qui n'auraient pas acheté (c'est à dire $0,588 \times 63 = 31,164$) ; de même, il y aurait 41,1% de 63 personnes touchés qui auraient acheté, c'est à dire $0,411 \times 63 = 25,9$. Symétriquement, $0,588 \times 148 = 87.0$ des touchés n'achèteraient pas et $0,411 \times 148 = 61,0$ des non touchés n'achèteraient pas. Ces quatre nombres représentent les effectifs attendus.

3. On a besoin de calculer les totaux de chaque ligne. Pour cela, on doit utiliser l'instruction `apply`. On peut ensuite stocker le résultat dans la variable `totalLigne` : `totalLigne <- apply(Tobs,1,sum)`.
4. On a besoin de calculer les totaux de chaque colonne. Pour cela, on doit utiliser l'instruction `apply`. On peut ensuite stocker le résultat dans la variable `totalColonne` : `totalColonne <- apply(Tobs,2,sum)`.
5. Enfin, on a besoin de calculer les pourcentages de chaque ligne. Pour cela, on doit calculez le nombre total d'individu (cela se fait avec `sum(Tobs)`). Puis on divise les totaux de chaque ligne par le nombre total d'individu : `frequenceLigne <- totalLigne / sum(Tobs)`.
6. Enfin, pour calculer le tableau des effectifs attendu, on utilise la commande `outer` : `outer(frequenceLigne,totalCol`
   Stockez ce tableau dans la variable `Tatt`.

1.4 Tableau des écarts

Le troisième tableau est celui des écarts entre les effectifs observés et les effectifs attendus. Il s'agit simplement de soustraire case à case le deuxième tableau au premier.

7. R permet de faire directement les soustractions de tableau. Pour soustraire un tableau à un autre, il suffit de taper `Tobs-Tatt`. Stockez le résultat dans la variable `Tecart`

1.5 Écarts au carré pondérés

Le quatrième et dernier tableau est celui des écarts au carré pondérés. Si nous sommions les écarts, nous obtiendrions une somme nulle. Nous pourrions sommer la valeur absolue des écarts, mais la fonction "valeur absolue" est une fonction pas très sympathique (non dérivable). On va donc élever les écarts au carré.

Ensuite, "5 personnes en plus si on en attend 300", ça n'est pas la même chose que "5 personnes en plus si on en attend 3". On va donc diminuer l'importance de l'écart en fonction de l'effectif attendu. Plus l'effectif attendu est grand, moins l'écart devra avoir d'importance. Pour obtenir cela, on va simplement diviser l'écart au carré par l'effectif attendu.

8. R permet de passer au carré et de faire les divisions directement avec des matrices. `Tecart^2/Tatt` calcule les écarts au carré pondéré. Stockez le résultat dans `Tfinal`

1.6 χ^2 et degré de liberté

Le χ^2 est simplement la somme de toutes les cellules du tableau des écarts au carré pondérés `Tfinal`. Le degré de liberté est (le nombre de lignes moins un) multiplié par (le nombre de colonne moins un)

9. Calculez le χ^2. Pour cela, utilisez la fonction `sum(Tfinal)` et stockez le résultat dans `khid`.
10. Calculez le DDL du tableau, stockez le résultat dans `ddl`.

Nous pouvons maintenant déterminer s'il y a oui ou non un lien entre les deux variables. Si le χ^2 est grand, il y a un lien. S'il est petit, il n'y a pas de lien. Plus précisément, R permet de trouver directement le petit p associé à un χ^2. Pour cela, il a besoin du χ^2 et de son DDL.

11. La fonction `pchisq(khid,ddl,lower.tail=FALSE)` donne le petit p.
12. Conclusion ? Y a-t-il un lien entre les variables ?

2 χ^2, automatiquement

Il est également possible de calculer le χ^2 automatiquement.

13. Pour cela, il faut utiliser la fonction `chisq.test()` sur le tableau croisé. Tapez `chisq.test(Tobs)`. Combien vaut le χ^2 ?
14. **R** donne le χ^2 mais aussi beaucoup d'autres choses : dans tout ce qu'il donne, quelque part se trouve le DDL et le petit p. Combien valent-ils ?
15. Trouvez-vous la même chose que dans le calcul à la main ?

Si le résultat est différent, c'est parce que **R** applique automatiquement la "correction de Yates". Pour la supprimer, il faut ajouter `correct=FALSE`

16. Calculez le χ^2 sans la correction de Yates. Pour cela, utilisez `chisq.test(Tobs,correct=FALSE)`.
17. Trouvez le χ^2, le petit p et le DDL.
18. Trouvez-vous la même chose qu'à la main ?

3 Exemple réel

19. Chargez en mémoire le fichier "miniESPAD99.csv"
20. Grâce à un `summary()`, examinez rapidement les données. Quelles sont les variables nominales ?
21. Choisissez deux variables nominales (un couple de variables qui, selon vous, sont peut-être liées), et calculez le χ^2 correspondant.
22. Les variables sont-elles liés au risque 5 %
23. Recommencez-avec deux autres couples de variables (vous aurez donc au final 3 χ^2). Éventuellement, une même variable pourra être dans plusieurs couples. Quelles sont les variables liées ?