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

TD n°6 : χ^2

1 Sous Excel

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.

1. Sur une feuille d'Excel, dressez le tableau observé suivant (dans les cases `B2 :D4`) :

Eff. Observés	Achète	N'achète pas
Contact	46	78
Non contact	17	70

2. Sous le tableau dans les cases `C5` ET `D5`, calculez le total des colonnes. Pour calculer la somme des cases `C3 :C4`, utilisez la fonction `somme(C3 :C4)`. Pareil pour la somme de la colonne `D`.
3. A droite du tableau, dans les cases `E3 :E5`, calculez le total des lignes.
4. Changez la couleur de fond des cases où vous venez de calculez un total.
5. Dans les cases `C6 :D6`, calculez le pourcentage de personnes qui ont acheté et le pourcentage de personnes qui n'ont pas acheté.

1.3 Effectifs attendus

Le deuxième tableau est celui des effectifs attendus. On constate qu'en moyenne, 29% des gens achètent. S'il y a indépendance entre les variables (et que le contact n'influence pas l'achat), alors 29% des gens touchés vont acheter et 29% des gens non touchés vont acheter. Il y a 124 personnes qui ont été touchées. Donc $124 \times 29\%$ des gens touchés achèteraient. De même, $87 \times 29\%$ des gens non touchés achèteraient. Symétriquement, $124 \times 71\%$ des touchés n'achèteraient pas et $87 \times 71\%$ des non touchés n'achèteraient pas. Ces quatre nombres représentent les effectifs attendus.

6. Dressez le tableau suivant (dans les cases `B9 :D11`) :

Eff. Attendus	Achète	N'achète pas
Contact		
Non contact		

7. Dans la case `C10`, calculez les effectifs attendus de personne qui ont été touché et qui ont acheté. N'écrivez pas directement `124*0,29` mais utilisez les numéros de cellule : `=E3*C6`
8. Complétez le tableau.
9. Ajoutez une ligne total et une colonne total. Présentent-elles un intérêt ?

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.

10. Sous Excel, dressez le tableau suivant (dans les cases `B15 :D17`) :

Ecarts	Achète	N'achète pas
Contact		
Non contact		

11. Dans `C16`, calculez l'écart entre les effectifs observés des "acheteurs contactés" et leurs effectifs attendus. Là encore, n'utilisez pas de nombre mais des numéros de cellule : `=C3-C10`
12. Complétez le tableau.
13. Ajoutez une ligne total et une colonne total. Présentent-elles un intérêt ?

1.5 Écarts au carré pondérés

Le quatrième et dernier tableau est celui des écarts au carré pondérés. Comme nous venons de le voir, si nous sommons les écarts, nous obtenons 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.

14. Sous Excel, dressez le tableau suivant (dans les cases `B21 :D23`) :

Ecarts au carré pondérés	Achète	N'achète pas
Contact		
Non contact		

15. Dans `C22`, calculez l'écart au carré pondéré des "acheteurs contactés" : `=C16^2/C10`
16. Complétez le tableau.
17. Ajoutez une ligne total et une colonne total.

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. Le degré de liberté est (le nombre de ligne SANS compter la ligne "total" moins un) multiplié par (le nombre de colonne SANS compter la colonne "total" moins un)

18. Calculez le χ^2 (ou s'il est déjà calculé, changez la couleur de sa cellule).
19. Calculez le DDL du tableau.

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, Excel permet de trouver directement le petit p associé à un χ^2. Pour cela, il a besoin du χ^2 et de son DDL.

20. Utilisez la fonction `loi.khideux(`χ^2`;DDL)` pour connaître le petit p du χ^2 que vous avez trouvé.
21. Conclusion ?

2 Sous R

Pour calculer le χ^2 de deux variables, il faut construire un tableau croisé. 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]`.

22. Chargez le fichier "ContactAchatPizza.csv" en mémoire, stockez le dans `donnees`.
23. Dressez le tableau croisé des variables `[CONTACT]` et `[ACHAT]` et stockez le dans la variable `tableau`. Pour cela, tapez `tableau <- table(donnees$contact,donnees$achat)`.
24. On peut ensuite calculer le χ^2 du tableau en utilisant la fonction `chisq.test()` sur le tableau croisé. Tapez `chisq.test(tableau)`. Combien vaut le χ^2 ?
25. **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 ?
26. Trouvez-vous la même chose que sous Excel ?

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

27. Calculez le χ^2 sans la correction de Yates. Pour cela, utilisez `chisq.test(tableau,correct=FALSE)`.
28. Trouvez le χ^2, le petit p et le DDL.
29. Trouvez-vous la même chose que sous Excel ?

3 Exemple réel

30. Chargez en mémoire le fichier "miniESPAD99.csv"
31. Grâce à un `summary()`, examinez rapidement les données. Quelles sont les variables nominales ?
32. Choisissez trois couples de variables nominales (des couples qui vous semblent pertinents et calculez les χ^2 correspondant. Une variable peut être dans plusieurs couples.
33. Quels couples de variables sont liés au risque 5 %