

TD n°4

Lien entre deux variables

1 Variables numériques

Il est possible de mesurer le lien entre deux variables numériques, mais c'est assez compliqué : on garde ça pour l'année prochaine. Par contre, la représentation graphique de deux variables numériques peut déjà donner une bonne indication quant à l'existence d'un lien.

1. Ouvrez le fichier *Reponses2009.xls*
2. Sélectionnez simultanément les colonnes [1-AGE] et [14E-GRAVITEANTISECHE]. Pour cela, sélectionnez une des deux colonnes, pressez la touche Ctrl puis, en maintenant la touche Ctrl appuyée, sélectionnez la deuxième colonne.
3. Cliquez sur l'assistant graphique (juste à droite de la touche ↓
A).
4. Dans la colonne de gauche, sélectionnez *Nuages de points*. À droite, conservez le choix par défaut (les points non reliés entre eux).
5. Cliquez sur *Terminez*
6. Selon votre étude, l'âge est-elle liée à la l'estimation de la gravité du copiage ?

2 Variables qualitatives, le χ^2

On s'intéresse à deux variables qualitatives. On veut estimer la force d'un lien entre deux variables. Pour cela, on calcule un indice appelé le χ^2 (prononcation Khi -deux). Pour calculer le χ^2 , on doit dresser une suite de 4 tableaux.

2.1 Effectifs observés

Le premier tableau est celui des effectifs observés, il correspond aux résultats observés après l'enquête.

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

Eff. Observés	Achete	N'achète pas
Contact	46	78
Non contact	17	73

8. 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.
9. A droite du tableau, dans les cases E3 :E5, calculez le total des lignes.
10. Changez la couleur de fond des cases ou vous venez de calculez un total.
11. Dans les cases C6 :D6, calculez le pourcentage de personnes qui ont acheté et le pourcentage de personnes qui n'ont pas acheté.

2.2 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 touchés. Donc $124 \times 29\%$ des gens touchés achèterait. De même, $90 \times 29\%$ des gens non touchés achèteraient. Symétriquement, $124 \times 71\%$ des touchés n'achèteraient pas et $90 \times 71\%$ des non touchés n'achèteraient pas. Ces quatre nombres représentent les effectifs attendus.

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

Eff. Attendus	Achete	N'achète pas
Contact		
Non contact		

13. Dans la case C10, calculez les effectifs attendus de personne qui ont été touché et qui ont acheté. N'écrivez pas directement 124*0,19 mais utilisez les numeros de cellule : =E3*C6
14. Complétez le tableau.
15. Ajoutez une ligne total et une colonne total. Présentent-elles un intérêt ?

2.3 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.

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

Ecarts	Achete	N'achète pas
Contact		
Non contact		

17. 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 numeros de cellule : =C3-C10
18. Complétez le tableau.
19. Ajoutez une ligne total et une colonne total. Présentent-elles un intérêt ?

2.4 Aïl aïl aïl...

Une erreure s'est glissée dans l'énoncé, il ne fallait pas lire 73 mais 70...

20. Modifiez la case D4 et remplacez le 73 par un 70.

Si vous avez utilisé systématiquement des numéros de cellule pour faire les calculs, tout est automatiquement modifié. Sinon, c'est la galère, il faut tout refaire...

2.5 Écarts au carré pondérés

Le quatrième et dernier tableau est celui des écarts au carré pondérés. Comme nous venons de le voir, si nous sommions 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 sympatique (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.

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

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

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

2.6 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)

25. Calculez le χ^2 (ou s'il est déjà calculé, changez la couleur de sa cellule).
26. 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. Quand est-ce qu'un χ^2 est grand ? Suspens... Hélas, pour le savoir, il faudra passer en L3... (enfin, si vous êtes sage, je vous donnerais quand même un indice en fin de TD)
Pour l'instant, sachez simplement qu'un χ^2 avec un DDL de 1 est grand s'il est supérieur à 4.

3 Lien entre [2-SEXE] et [5-REDOUBLEMENT]

Concernant l'étude que nous sommes en train de mener, nous souhaitons voir s'il y a un lien entre les variables [2-SEXE] et [5-REDOUBLEMENT]. Pour cela, nous allons dresser le tableau des effectifs observés puis calculez son χ^2 .

3.1 Tableau croisé dynamique

Dresser le tableau observé n'est qu'une histoire de comptage : il faut compter pour chaque couple de modalité possible. Heureusement, Excel sait faire ça automatiquement, grâce aux *tableaux croisés dynamiques*.

27. Donnez un nom à l'ensemble de toutes les données (par exemple **BDD**).
28. Allez dans *Données* → *Rapport de données croisé dynamique*
29. Une fenêtre s'ouvre. Les choix proposés sont les bons, cliquez sur *Suivant*.
30. Excel vous demande où se trouve vos données. Tapez le nom que vous avez choisi (**BDD**) puis cliquez sur *Suivant*.
31. Choisissez *Nouvelle feuille* puis cliquez sur *Terminez*.
32. Excel crée une nouvelle feuille. En haut à gauche, il prépare un tableau. Au centre, il ouvre une fenêtre appelée *Liste des champs de tableau croisé*. Cette fenêtre contient toutes les variables de la base de donnée. Dans cette fenêtre, sélectionnez la variable [2-SEXE] et faites la glisser sur le tableau dans la case *Déposer les champs de colonne ici*
33. Toujours dans cette fenêtre *Liste des champs de tableau croisé*, sélectionnez la variable [5-REDOUBLEMENT] et faites la glisser sur le tableau dans la case *Déposer les champs de ligne ici*
34. Enfin, sélectionnez la variable [2-SEXE] (une deuxième fois) et faites la glisser sur le tableau dans la case *Déposer Données ici*

C'est prêt ! Vous venez de dresser le premier tableau, celui des effectifs observés. Cependant, ce tableau n'est pas facile à utiliser, en particulier il est compliqué de faire référence aux cases qu'il contient. Pour pouvoir l'utiliser, on va le *Recopier par valeur*

35. Sélectionnez tout le tableau
36. Dans *Edition*, choisissez *Copier*
37. Deux ou trois lignes sous le tableau, faites un clic droit et choisissez *Collage spécial*, puis *Valeur*

Excel vient de copier les valeurs du tableau dans les cellules que vous avez choisi. Ce nouveau tableau est plus facilement utilisable. Vous allez maintenant pouvoir calculer son χ^2 .

3.2 Calcul du χ^2

38. Dressez le tableau des effectifs attendus
39. Dressez le tableau des écarts
40. Dressez le tableau des écarts au carré pondérés
41. Calculez le χ^2 .
42. Calculez le DDL du tableau.

3.3 Conclusion

Reste à déterminer ce qu'est un χ^2 petit ou grand. Nous ne rentrerons pas ici dans les détails de la théorie mais nous donnerons simplement une recette de cuisine : Excel possède une fonction qui, étant donné un χ^2 et son DDL, nous dit si les variables sont liées ou non. Si le χ^2 est dans la case EE et que son DDL est dans la case C'est la fonction `=0.05/loi.khideux(KHI2 ;DDL)`. Si cette fonction (appelée *fonction test*) est plus grande ou égale à 1, alors il y a un lien entre les variables. Sinon, il n'y a pas de lien.

43. On suppose que le χ^2 est dans la case EE et que son DDL est dans la case FF (si ça n'est pas le cas, adaptez la formule qui suit. Dans la case TT, tapez `=0.05/loi.khideux(EE ;FF)`
44. Résultat ? Y a-t-il un lien entre les variables ?
45. Choisissez deux autres variables qualitatives pour lesquelles vous pensez qu'il existe un lien et calculez leur χ^2 .