X et Y, deux variables indépendantes ??

On peut se demander si le nombre de crayons cassés par une personne dans une journée est dépendant ou non de l'état de son stress.

Même si l'on peut penser qu'il existe bien une forte dépendance entre ces deux aspects, on doit mener une étude statistique sérieuse pour conclure de manière scientifique. Au passage, nous apprendrons peut-être quelques notions en lien avec les probabilités !*!*

1) Situation exemple

Voici en avant-première dans ce tableau le résultat d'une enquête (réalisée avant hier) au sein d'une école d'architecture où les crayons représentent un outil de travail fondamental.

X est le nombre de crayons cassés par une personne au cours de la journée et Y l'état de son stress du jour.

X \ Y	Stress faible	Stress moyen	Stress fort	
0	16	12	4	**32**
1	10	11	9	30
2	7	**9**	14	30
3	2	7	10	19
	35	39	37	**111**

Comment lire un tel tableau ? Voici quelques indices :

9 personnes moyennement stressées ont cassé 2 crayons ; **35** personnes avait un stress faible parmi les N = **111** personnes sondées ; **32** personnes n'ont cassé aucun crayon.

Avant d'essayer de répondre à la question : « Le nombre X et l'état Y sont-elles des variables indépendantes ? », il faut d'abord rappeler quelques notions liées aux probabilités.

2) Rappels sur les probabilités (ça peut servir…)

* On peut ici qualifier X de « variable aléatoire », car si l'on rencontre une personne par hasard dans cette école, son nombre de crayons cassés sera un entier entre 0 et 3 (d'après l'étude précédente), mais on ne pourrait a priori rien savoir de plus avant de l'interroger. De même, Y est aussi une variable aléatoire.

* On peut estimer la probabilité pour qu'une personne choisie au hasard soit faiblement stressée selon la formule de Laplace : $p(Y = \text{faible}) = \frac{nombre\ de\ personnes\ de\ faible\ stress}{nombre\ total\ de\ personnes}$

Ainsi : $p(Y = \text{faible}) = \frac{35}{111}$ (bien sûr, ceci au sein de cette école…)

* On peut aussi calculer la probabilité de rencontrer une personne ayant cassé 3 crayons **et** dont l'état de stress est fort selon le même principe, mais, le nombre considéré en haut de la fraction doit tenir compte de deux critères. On le trouve donc à l'intersection de la ligne X = 3 et de la colonne Y = fort. Ainsi : $p(X = 3 \cap Y = \text{fort}) = \frac{10}{111}$

Remarque : Vous l'aurez compris, le symbole mathématique $\cap$ signifie « et ».

* Enfin, il faut dire un mot sur la notion d'indépendance de deux variables. A est indépendante de B lorsque la réalisation de A n'influence pas du tout la réalisation de B. Exemple :
A = nombre de chips présents dans un paquet de 135 g acheté chaque samedi.
B = couleur des chaussettes du prof de maths portées au moment de l'achat.

Une propriété mathématique permet de signifier et/ou de vérifier l'indépendance entre deux variables A et B. On a : $p(A = i \cap B = j) = p(A = i)\ p(B = j)$ pour tout couple (i ; j) avec i et j les différents cas possibles pour ces variables.

3) Revenons à notre affaire de crayons…

Pour mieux comprendre la suite, nous allons supposer un instant que nos variables X et Y sont effectivement des variables aléatoires indépendantes.

Comment cette propriété se traduirait-elle dans les données du tableau ?

On va examiner le cas du couple (X = 2 ; Y = moyen*)* pour lequel on observe 9 personnes dans le tableau. On compte 39 personnes de stress moyen (total au bas de la colonne) et 30 personnes ayant cassé 2 crayons (total à droite de la ligne).

Si X et Y étaient indépendant*e*s, la case à l'intersection de (X = 2) et (Y = moyen) devrait comporter la valeur : $\frac{30 \times 39}{111} = 10{,}54.$ (démo possible en utilisant la formule ci-dessus)

Aparté : comme vous êtes très sympa ce matin, je vous fais la mini démo…
D'après le tableau : $\mathrm{p}(\mathrm{X} = 2) = \frac{30}{111}$ et de même, $\mathrm{p}(\mathrm{Y} = \text{moyen}) = \frac{39}{111}$
Ainsi, la probabilité de l'événement $(\mathrm{X} = 2 \cap \mathrm{Y} = \text{moyen})$ si l'on suppose ces deux variables indépendantes sera : $\mathrm{p}(\mathrm{X} = 2 \cap \mathrm{Y} = \text{moyen}) = \mathrm{p}(\mathrm{X} = 2)\ \mathrm{p}(\mathrm{Y} = \text{moyen}) = \frac{30}{111}\ \frac{39}{111} = \frac{30 \times 39}{111^2}$
Enfin, le nombre de personnes dans ce cas sera : $N \times \mathrm{p}(\mathrm{X} = 2 \cap \mathrm{Y} = \text{moyen}) = 111 \times \frac{30 \times 39}{111^2} = \frac{30 \times 39}{111}$

Bien entendu, l'école étudiée ne peut comporter 10,54 personnes dans cette case !

On comprend ici que cet aspect théorique donne une indication de la valeur que l'on devrait avoir lorsque X et Y sont indépendantes. Comme on travaille dans le domaine des probabilités, on doit comprendre que si l'on observait de nombreuses écoles du même type comportant toujours les même totaux 39 et 30 déjà cités, on devrait observer des variations aléatoires proches de la valeur 10,54 dans la case étudiée (parfois 11, parfois 10, parfois 9, …).

4) Etudions la situation avec soin…

Notons o_{ij} la valeur observée dans le tableau à l'intersection de la ligne n° i et de la colonne n° j. On utilisera ces valeurs par la suite.

On note ensuite t_{ij} les valeurs théoriques que prendraient les o_{ij} si les variables X et Y étaient indépendantes. On a vu comment calculer l'une de ces valeurs.

En général, on a : $t_{ij} = \frac{S_i \times S_j}{N}$ avec S_i la somme des valeurs observées sur la ligne i, S_j la somme des valeurs observées dans la colonne j et N le nombre total de personnes sondées.

En utilisant cette formule, les valeurs théoriques de la première ligne devraient donc être dans l'ordre : $\frac{32 \times 35}{111} = 10{,}09$; $\frac{32 \times 39}{111} = 11{,}24$; $\frac{32 \times 37}{111} = 10{,}67$.

Vérifiez que la deuxième ligne serait : 9,46 ; 10,54 ; 10. La troisième ligne sera identique.

Enfin, pour la quatrième ligne, on trouve : 5,99 ; 6,68 ; 6,33.

Pour la suite, lorsque les variables X et Y sont indépendantes, on utilise l'hypothèse que chaque variable $(o_{ij} - t_{ij})$ (autrement dit, l'écart entre la valeur observée et la valeur théorique dans chaque case) suit une loi normale de moyenne nulle. Selon l'école étudiée, ayant les mêmes totaux sur chaque ligne et dans chaque colonne, on pourrait observer de petites variations au sein d'une même ligne par exemple.

Le problème de la mesure d'un écart entre les valeurs observées et les valeurs théoriques n'est pas si simple. En effet, un écart positif dans une case pourrait être compensé par un écart négatif dans la case d'à côté ! Au final, on pourrait croire qu'il n'y a pas d'écart alors que chaque case donne un écart non nul !

Pour éviter ce souci, on mesure les écarts puis, on les élève au carré. Ainsi, les écarts au carré ne pourront pas se compenser puisqu'ils seront tous positifs.

Mais à ce moment, on fait intervenir une nouvelle variable aléatoire liée à la somme des écarts au carré. Le problème (en fait assez complexe) se pose selon la méthode suivante…

5) Mise en place de la résolution (enfin !!)

On calcule les valeurs t_{ij} dans chaque case, puis on calcule les écarts $(o_{ij} - t_{ij})$ et enfin, on calcule la valeur : $Q^2 = \sum \frac{(o_{ij} - t_{ij})^2}{t_{ij}}$ (avec nos valeurs, vérifiez que $Q^2 = 15{,}1$)

On démontre que cette valeur suit une loi particulière appelée loi du Khi-2 : χ^2 (en grec).

Cette loi dépend d'un paramètre appelé « degré de liberté ». Il faut donc déterminer le nombre de degrés de liberté dans notre cas pour ensuite résoudre notre problème. Nous ferons cela dans quelques instants… On admet pour le moment que ce nombre est égal à 6 pour notre exercice.

Donc, on sait que la variable aléatoire notée Q^2 suit la loi du χ^2 à 6 degrés de liberté.

Pour que nos deux variables soient bien indépendantes, on comprend que la valeur de Q^2 ne doit pas être trop grande, car les écarts au global ne doivent pas être si grands que cela entre les valeurs observées et les valeurs théoriques.

On va enfin comparer notre valeur de Q^2, notée souvent χ^2_{obs}, à la valeur maximale attendue dans le cas où nos variables X et Y seraient indépendantes. Si notre Q^2 dépasse cette valeur, on pourra conclure par la négative : X et Y ne sont pas finalement indépendantes.

Pour cela, on va étudier l'extrait ci-contre du tableau classique lié à la loi du χ^2 :

ddl = ν	Valeurs de χ^2	
1	3,84	6,63
2	5,99	9,21
3	7,81	11,34
4	9,49	13,28
5	11,07	15,09
6	12,59	16,81
7	14,07	18,48
…	…	…
α	0,05	0,01

Comment interpréter ces valeurs ?

Voilà une bonne question !

Comme on travaille avec 6 degrés de liberté (ddl en abrégé ou parfois ν), seule la ligne ddl = 6 nous est utile.

Ensuite, on doit comparer la valeur de notre Q^2 valant 15,1 avec les valeurs de cette ligne 6 (dans un tableau complet, il y a une dizaine de colonnes pour les valeurs de χ^2, mais pour faire simple, je n'en propose ici que deux).

Mais laquelle regarder : 12,59 ou bien 16,81 ? Et d'ailleurs pour faire quoi après ??

Il faut dire ici un petit mot sur la notion de test statistique et de risque associé.

Il faut comprendre que la réponse à notre question « X et Y sont-elles des variables indépendantes ? » sera soit Oui, soit Non. Ce côté tranché n'est pas vraiment en accord avec la notion de probabilité. En réalité, on répond plutôt avec une phrase du genre suivant : « Il y a 95 % de chances pour que ces variables soient indépendantes ». Le côté probabiliste de la réponse apparaît dans le côté suivant : il reste 5 % de chances que l'on se trompe en donnant cette conclusion. C'est un risque à courir, on dit aussi « au risque de 5% ».

Mais bon, être sûr à 95 %, c'est déjà pas mal !! C'est ça, un test statistique.

Donc, il reste à comprendre le rôle de la ligne avec le alpha (α). Si l'on me demande une réponse « au risque de 5% », ou aussi « avec 95 % de confiance », je dois regarder la valeur dans la colonne $\alpha = 0{,}05$.

Pour conclure sur notre cas de crayons cassés, on lit la valeur 12,59 car on va travailler au risque de 5 % (cas généralement proposé).

On conclut alors ainsi : notre valeur de Q^2 valait 15,1. Elle dépasse la valeur 12,59 du tableau qui représente le maximum admissible pour la variable Q^2 lorsque ddl = 6.

On peut donc conclure que X et Y ne sont pas des variables indépendantes au risque de 5%.

Voilà, nous avons résolu cette question…

6) Et la question : « Monsieur, c'est quoi cette affaire de degré de liberté ? »

Avant de nous quitter, il reste à éclaircir certains points concernant les degrés de liberté !

Tournez la page S.V.P. !!!

Rappelons d'abord pour notre tableau de départ que la somme au bout de chaque ligne (ou de chaque colonne) est une valeur fixée par notre sondage. Par ailleurs, on peut concevoir une certaine variabilité des valeurs au sein d'une même ligne en fonction des échantillons étudiés. On étudie finalement des variabilités dans chaque case, mais en supposant le total sur la ligne restant constant (de même pour la colonne).

Ainsi, lorsqu'une ligne (par exemple notre première ligne) comporte trois valeurs (notre cas !), elles ne sont pas réellement indépendantes les unes des autres, libres de tout mouvement… Non, car leur somme doit rester celle imposée au bout de la ligne (pour nous **32**). Cette somme impose donc 1 condition qui restreint les libertés des coefficients de cette ligne. Au final, le degré de liberté à considérer pour cette ligne vaut 3 – 1 = 2.

De même, on va prendre 4 – 1 = 3 pour les degrés de liberté liés aux colonnes, dans lesquelles on peut trouver 4 valeurs à chaque fois.

Enfin, le nombre des degrés de liberté pour notre tableau entier vaut le produit $2 \times 3 = 6$.

En général, pour un tableau de n lignes et m colonnes, on obtient $(n - 1)(m - 1)$ degrés de liberté. Simple finalement !!

7) Un petit bonus pour la route ?

Posons-nous la question suivante, à la suite d'une étude très sérieuse réalisée à Cesi (une école d'ingénieurs proche de Rouen) :

« Le fait d'avoir réussi un devoir peut-il être lié au temps de révision ? ».

Question délicate…

Voyons les faits grâce à un sondage sur quelques étudiants et étudiantes.

Notons X la note obtenue (A = très bien ; B = bien ; C = insuffisant ; D = mauvais) et Y le temps passé aux révisions (Top+ = entre 39 et 30 heures ; Top = entre 29 et 20 heures ; Bof = entre 19 et 10 heures ; Bof– = entre 9 et 0 heures). (durées typiques par semaine……)

Voici le résultat du sondage :

	Top+	Top	Bof	Bof–
A	35	27	9	5
B	8	49	21	11
C	6	15	27	13
D	5	8	31	43

A vous de suivre une démarche similaire permettant de répondre à cette énigmatique question !!!

Remarque : vous trouverez facilement des tables complètes de la loi du χ^2 sur le Web…