

Introduction

Quand nous regardons les fréquences des variables catégorielles, il est possible d'avoir plus d'une variable et nous cherchons donc à voir si une variable dépend d'une autre. Par exemple, est-ce que le fait de préférer les films d'actions ou les films romantiques dépend du fait d'être un garçon ou une fille ; est-ce que le fait de préférer rock ou le rap dépend de la génération à laquelle on appartient ; est-ce que le fait d'aller à l'université dépend du fait d'avoir au moins un parent qui y a été ; etc.

On calcule encore le khi-carré pour voir si nos résultats sont significatifs, mais dans une table de contingence, il existe une formule pour calculer vos fréquences attendues. Pour ce faire, vous avez besoin de vos totaux marginaux. Il s'agit du nombre d'observations de la ligne et la colonne où se situe la fréquence que vous cherchez.

Illustration:

	Colonne 1	Colonne 2	Total
Ligne 1			total mar. L1
Ligne 2			total mar. L2
Total	total mar. C1	total mar. C2	TOTAL

Les totaux marginaux sont donc le total des lignes et le total des colonnes. Le fait est qu'au moins un "groupe" de totaux marginaux (c.-à-d., soit ceux des colonnes soit ceux des lignes) ne seront pas les mêmes. En d'autres mots, le total marginal de la ligne 1 ne sera pas le égal au total marginal de la ligne 2 et/ou le total marginal de la colonne 1 ne sera pas égal au total marginal de la colonne 2.

Exemple de table de contingence:

<u>Type de film préféré selon le genre - O (A)</u>			
	Action	Romantique	Total
Garçon	43 (30.5)	7 (19.5)	50
Fille	18 (30.5)	32 (19.5)	50
Total	61	39	100

Formule

Fréquences attendues

$$A_{ij} = \frac{L_i C_j}{N}$$

Diagram illustrating the formula for expected frequency A_{ij} :

- L_i : Total marginal de la Ligne i
- C_j : Total marginal de la Colonne j
- N : Effectif total de l'échantillon
- A_{ij} : Fréquence attendue à la cellule située à la ligne i et la colonne j

Khi-carré

$$\chi^2 = \sum \frac{(O - A)^2}{A}$$

Diagram illustrating the chi-square formula:

- χ^2 : Symbole du khi-carré
- \sum : Symbole de la somme
- O : Fréquence observée
- A : Fréquence attendue

Mise en situation

Donald T. est chercheur et il se demande qui est le fabricant de perruques préféré des ses concitoyens de la République de Santa-Banana entre les deux plus grands du monde: Al O. Pesci et Pierre-Hugues Wiggs. Il se demande également si le fabricant de perruques préféré est le même pour les chauves et les chevelus. Il a donc posé la question à deux groupes: les chevelus et les chauves.

H: Le fabricant de perruques préféré dépend de la présence de cheveux.

H₀: Le fabricant de perruques préféré ne dépend pas de la présence de cheveux

Voici ses résultats :

	Pesci	Wiggs	Total
Chevelus	33	39	72
Chauves	28	44	72
Total	61	83	144

Calculs

Étape 1: calculer les fréquences attendues de votre table de contingence

	Pesci	Wiggs	Total
Chevelus	$A_{1,1}$	$A_{1,2}$	72
Chauves	$A_{2,1}$	$A_{2,2}$	72
Total	61	83	144

$$A = \frac{L * C}{N}$$

$$A_{1,1} = \frac{L1 * C1}{N} = \frac{72 * 61}{144} = \frac{4392}{144} = 30.5$$

$$A_{1,2} = \frac{L1 * C2}{N} = \frac{72 * 83}{144} = \frac{5976}{144} = 41.5$$

$$A_{2,1} = \frac{L2 * C1}{N} = \frac{72 * 61}{144} = \frac{4392}{144} = 30.5$$

$$A_{2,2} = \frac{L2 * C2}{N} = \frac{72 * 83}{144} = \frac{5976}{144} = 41.5$$

Étape 2: Calculer le khi-carré

	Pesci	Wiggs	Total
Chevelus	33 (30.5)	39 (41.5)	72
Chauves	28 (30.5)	44 (41.5)	72
Total	61	83	144

$$\begin{aligned} \chi^2 &= \sum \frac{(O - A)^2}{A} \\ &= \frac{(33-30.5)^2}{30.5} + \frac{(39-41.5)^2}{41.5} + \frac{(28-30.5)^2}{30.5} + \frac{(44-41.5)^2}{41.5} \\ &= \frac{6.25}{30.5} + \frac{6.25}{41.5} + \frac{6.25}{30.5} + \frac{6.25}{41.5} \\ &= 0.20 + 0.15 + 0.20 + 0.15 \\ &= 0.70 \end{aligned}$$

Étape 3: Le degré de liberté

Quand nous avons deux variables, pour calculer notre degré de liberté, il faut multiplier le nombre de ligne - 1 par le nombre de colonne - 1

$$dl = (\text{nombre de ligne} - 1) * (\text{le nombre de colonne} - 1)$$

$$dl = (2-1)*(2-1)$$

$$df = 1*1$$

$$df = 1$$

Étape 4: Distribution khi-carré

L'étape est la même qu'un khi-carré standard. La colonne de notre seuil de rejet et la ligne de notre degré de liberté nous indiquent notre valeur critique.

TABLE C.4 Critical Values of Chi Square (χ^2)

df	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	1.32330	2.70554	3.84146	5.02389	6.63490	7.87944	10.828
2	2.77259	4.60517	5.99147	7.37776	9.21034	10.5966	13.816
3	4.10835	6.25139	7.81473	9.34840	11.3449	12.8381	16.266
4	5.38527	7.77944	9.48773	11.1433	13.2767	14.8602	18.467
5	6.62568	9.23635	11.0705	12.8325	15.0863	16.7496	20.515
6	7.84080	10.6446	12.5916	14.4494	16.8119	18.5476	22.458
7	9.03715	12.0170	14.0671	16.0128	18.4753	20.2777	24.322
8	10.2188	13.3616	15.5073	17.5346	20.0902	21.9550	26.125
9	11.3887	14.6837	16.9190	19.0228	21.6660	23.5893	27.877
10	12.5489	15.9871	18.3070	20.4831	23.2093	25.1882	29.588
11	13.7007	17.2750	19.6751	21.9200	24.7250	26.7569	31.264

Dans ce cas, notre valeur critique est de 3.84. Cela signifie que si votre χ^2 est **supérieur** à 3.84, vous pouvez rejeter votre hypothèse nulle. Si vous rejetez l'hypothèse nulle, cela signifie que la guimauve est plus populaire que le chocolat. Notre χ^2 est de 0.70, on ne peut pas rejeter l'hypothèse nulle.

Signification des symboles

X : Valeur

N : Nombre d'observations total

n : Nombre d'observations d'un groupe

- n_1 : nombre d'observations du groupe 1 ;
- n_2 : nombre d'observations du groupe 2 ;
- etc.

K : Nombre de groupes

Σ : Somme

\bar{X} : Moyenne d'un échantillon

μ : Moyenne d'une population

s : Écart-type d'un échantillon

σ : Écart-type d'une population

s^2 : Variance d'un échantillon

σ^2 : Variance d'une population

χ^2 : Khi-carré

A : Fréquence attendue

O : Fréquence observé

L : Ligne

C : Colonne

t : Statistique t (ou score t dans le cas d'une corrélation/régression linéaire)

r : Coefficient de Pearson

Z : Score Z

b : Coefficient de régression

a : Ordonnée à l'origine

\hat{Y} : Valeur de la VD qu'on veut prédire l'aide la VI

F : Statistique F

Ouvrages de référence

- Field, A. (2017). *Discovering Statistics Using IBM SPSS Statistics: North American Edition* (5th ed.). Sage Edge
- Howell, D.C. (2008). *Méthodes statistiques en sciences humaines* (M. Rogier, V. Yzerbyt, & Y. Bestgen, Trans.). (6th ed.). De boeck. (Original work published 2008)
- Tabachnick, B. G., & Fidell, L. S. (2021). *Using Multivariate Statistics* (7th ed.). Pearson