

Chapitre 7 – Tableaux de fréquences et tests du khi-deux

STT-1920 Méthodes statistiques

Département de mathématiques et de statistique



UNIVERSITÉ
LAVAL

Plan

Test d'adéquation pour une variable discrète

Test d'homogénéité de I populations

Test d'indépendance de deux variables discrètes

Tableaux 2×2 et comparaison de deux proportions

Synthèse : quel test choisir ?

Test d'adéquation pour une variable discrète

Test d'adéquation : de quoi s'agit-il ?

- **L'objet** : une variable qualitative ou quantitative discrète ayant un nombre fini de valeurs possibles.

Test d'adéquation : de quoi s'agit-il ?

- **L'objet** : une variable qualitative ou quantitative discrète ayant un nombre fini de valeurs possibles.
- **La question** : le **modèle théorique** de probabilités est-il adéquat pour décrire la distribution de cette variable dans la population qui nous intéresse ?

Test d'adéquation : de quoi s'agit-il ?

- **L'objet** : une variable qualitative ou quantitative discrète ayant un nombre fini de valeurs possibles.
- **La question** : le **modèle théorique** de probabilités est-il adéquat pour décrire la distribution de cette variable dans la population qui nous intéresse ?
- **Le but** : comparer un ensemble d'observations avec les **fréquences attendues** sous le modèle théorique pour déterminer si ce modèle s'ajuste bien aux données.

Exemple 1 : les lois de Mendel

On croise des individus de génotype AAbb avec des individus de génotype aaBB. Tous les hybrides seront de la forme AaBb.

- A et B sont des allèles **dominants** (qui s'expriment avec 1 ou 2 copies);
- a et b sont des allèles **récessifs** (qui s'expriment avec 2 copies seulement).

Si on croise des hybrides, il y a 16 génotypes possibles, mais seulement 4 expressions possibles des gènes (**phénotypes**) : AB, Ab, aB et ab.

Exemple 1 : les lois de Mendel

Selon les lois de Mendel, qui supposent que les deux gènes ne sont pas liés entre eux, les quatre phénotypes devraient se retrouver dans la progéniture du croisement dans les proportions suivantes :

Phénotype	AB	Ab	aB	ab
Proportion théorique	9/16	3/16	3/16	1/16

Que penser de l'hypothèse mendélienne lorsqu'on la confronte aux observations ?

Exemple 1 : hypothèses

On veut donc tester les hypothèses suivantes :

$$H_0 : p_1 = \frac{9}{16}, p_2 = \frac{3}{16}, p_3 = \frac{3}{16}, p_4 = \frac{1}{16}$$

H_1 : au moins une des égalités n'est pas vérifiée.

Il faudra récolter des données dans une population (i.e. effectuer plusieurs croisements et noter le phénotype), et comparer les fréquences observées et théoriques pour décider si on rejette H_0 ou non au seuil α .

Exemple 1 : hypothèses

On veut donc tester les hypothèses suivantes :

$$H_0 : p_1 = \frac{9}{16}, p_2 = \frac{3}{16}, p_3 = \frac{3}{16}, p_4 = \frac{1}{16}$$

H_1 : au moins une des égalités n'est pas vérifiée.

Il faudra récolter des données dans une population (i.e. effectuer plusieurs croisements et noter le phénotype), et comparer les fréquences observées et théoriques pour décider si on rejette H_0 ou non au seuil α .

Hypothèse nulle, cas général

$$H_0 : (p_1, p_2, \dots, p_J) = (p_1^*, p_2^*, \dots, p_J^*)$$

Tableau des fréquences observées

Le tableau de fréquences donne le nombre de cas observés ($O_1, O_2, \dots, O_J$) dans l'échantillon pour chaque valeur de la variable ($v_1, v_2, \dots, v_J$).

Valeur	v_1	v_2	$\dots$	v_J	Total
Fréquence observée O_j	O_1	O_2	$\dots$	O_J	n

Fréquences espérées sous H_0

Les fréquences observées seront comparées aux fréquences attendues sous H_0 , donc avec le modèle théorique $(E_1, E_2, \dots, E_J)$.

Valeur	v_1	v_2	$\dots$	v_J	Total
Fréq. observée O_j	O_1	O_2	$\dots$	O_J	n
Fréq. espérée E_j	E_1	E_2	$\dots$	E_J	n

Ces fréquences espérées se calculent comme suit :

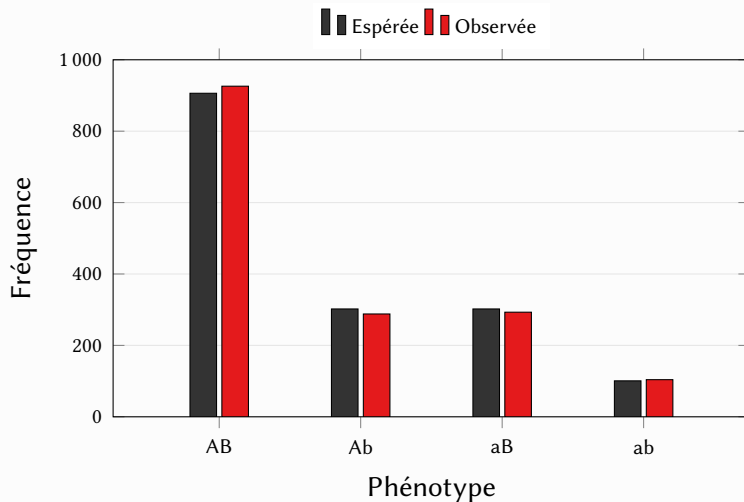
$$E_j =$$

Exemple 1 : les données

Voici les résultats d'une expérience effectuée sur 1611 vaches. Quelles sont les fréquences espérées sous H_0 , i.e. sous l'hypothèse de Mendel?

Phénotype	AB	Ab	aB	ab	Total
Fréq. obs. (O_j)	926	288	293	104	1611
Fréq. esp. (E_j)					

Exemple 1 : fréquences espérées et observées

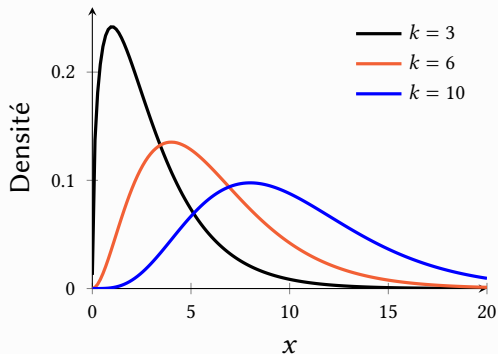


« Distance » entre le modèle et les observations

Pour déterminer à quel point les données et le modèle sont en accord, il faut comparer les O_j et les E_j . Si l'écart est trop grand, on rejettera H_0 et on acceptera H_1 .

Comment mesure-t-on cet écart ?

Rappel : la loi du khi-deux



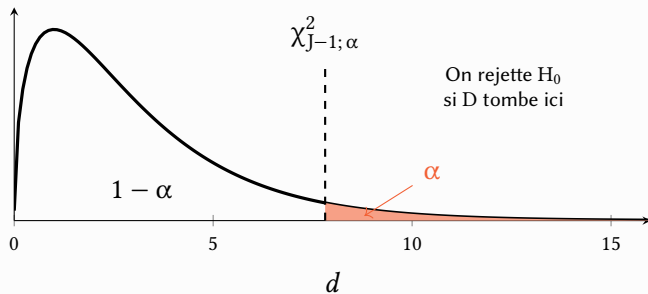
Déjà rencontrée au chapitre 4
(comparaison de variances), la loi χ_k^2 à k
degrés de liberté :

- ne prend que des valeurs positives ;
- est asymétrique (longue queue à droite) ;
- a pour moyenne k : plus k est grand, plus la courbe se déplace vers la droite.

Règle de décision

- Si H_0 est vraie,
- On rejettera donc H_0 si...

Région de rejet : toujours à droite



- Un écart entre O_j et E_j , dans un sens ou dans l'autre, fait augmenter D (les écarts sont mis au carré).
- Le test est donc unilatéral à droite, même si H_1 est « bilatérale ».
- Valeur-p = $P(\chi_{J-1}^2 > d_{\text{obs}})$: l'aire à droite de la distance observée.

Table de la loi du khi-deux

k	γ (aire à droite)				
	0,10	0,05	0,025	0,01	0,005
1	2,71	3,84	5,02	6,63	7,88
2	4,61	5,99	7,38	9,21	10,60
3	6,25	7,81	9,35	11,34	12,84
4	7,78	9,49	11,14	13,28	14,86
5	9,24	11,07	12,83	15,09	16,75
6	10,64	12,59	14,45	16,81	18,55
7	12,02	14,07	16,01	18,48	20,28
8	13,36	15,51	17,53	20,09	21,95
9	14,68	16,92	19,02	21,67	23,59
10	15,99	18,31	20,48	23,21	25,19

Extrait de la table complète (jusqu'à $k = 100$) fournie avec le cours.

La table donne la valeur $\chi^2_{k;\gamma}$ ayant une aire γ à sa droite sous la loi χ^2_k .

Exemple : avec $k = 3$ degrés de liberté et $\gamma = 0,05$:

$$\chi^2_{3;0,05} = 7,81.$$

En R :

`qchisq(0.95, df = 3)`
donne 7,81.

QUIZ!

On calcule $D = \sum_j (O_j - E_j)^2 / E_j$ pour tester un modèle génétique.

- Que signifie une **grande** valeur de D ?

QUIZ!

On calcule $D = \sum_j (O_j - E_j)^2 / E_j$ pour tester un modèle génétique.

- Que signifie une **grande** valeur de D ?
- Que signifie $D = 0$? Est-ce une bonne nouvelle pour H_0 ?

QUIZ!

On calcule $D = \sum_j (O_j - E_j)^2 / E_j$ pour tester un modèle génétique.

- Que signifie une **grande** valeur de D ?
- Que signifie $D = 0$? Est-ce une bonne nouvelle pour H_0 ?
- D peut-il être négatif ?

QUIZ!

On calcule $D = \sum_j (O_j - E_j)^2 / E_j$ pour tester un modèle génétique.

- Que signifie une **grande** valeur de D ?
- Que signifie $D = 0$? Est-ce une bonne nouvelle pour H_0 ?
- D peut-il être négatif?
- Si on double la taille de l'échantillon et que les proportions observées restent les mêmes, que devient D ?

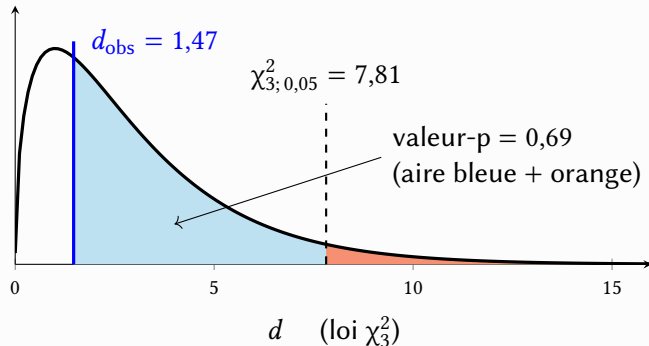
Exemple 1

Au seuil de 5 %, rejette-t-on l'hypothèse de Mendel ?

Phénotype	AB	Ab	aB	ab	Total
Fréq. obs. (O_j)	926	288	293	104	1611
Fréq. esp. (E_j)					
Distance					

Exemple 1 : conclusion

Exemple 1 : la décision en image



La distance observée est bien en deçà de la valeur critique : les écarts entre observations et proportions mendéliennes sont du même ordre que les fluctuations d'échantillonnage attendues.

Conditions d'application et remarques

- La loi χ^2 n'est qu'une **approximation** de la loi de D, valable si les **fréquences espérées** sont assez grandes : **$E_j \geq 5$ pour tout j** .
(C'est une condition sur les E_j , pas sur les O_j !)

Conditions d'application et remarques

- La loi χ^2 n'est qu'une **approximation** de la loi de D, valable si les **fréquences espérées** sont assez grandes : **$E_j \geq 5$ pour tout j** .
(C'est une condition sur les E_j , pas sur les O_j !)
- Si certaines $E_j < 5$, on **regroupe des classes voisines** (p. ex. « 4 et plus ») jusqu'à ce que la condition soit satisfaite, puis on refait le test avec le nouveau nombre de classes J .

Conditions d'application et remarques

- La loi χ^2 n'est qu'une **approximation** de la loi de D, valable si les **fréquences espérées** sont assez grandes : **$E_j \geq 5$ pour tout j** .
(C'est une condition sur les E_j , pas sur les O_j !)
- Si certaines $E_j < 5$, on **regroupe des classes voisines** (p. ex. « 4 et plus ») jusqu'à ce que la condition soit satisfaite, puis on refait le test avec le nouveau nombre de classes J .
- Pour un tableau 2×2 à petits effectifs, on utilise plutôt le **test exact de Fisher** (voir la section sur les tableaux 2×2).

Conditions d'application et remarques

- La loi χ^2 n'est qu'une **approximation** de la loi de D, valable si les **fréquences espérées** sont assez grandes : **$E_j \geq 5$ pour tout j** .
(C'est une condition sur les E_j , pas sur les O_j !)
- Si certaines $E_j < 5$, on **regroupe des classes voisines** (p. ex. « 4 et plus ») jusqu'à ce que la condition soit satisfaite, puis on refait le test avec le nouveau nombre de classes J.
- Pour un tableau 2×2 à petits effectifs, on utilise plutôt le **test exact de Fisher** (voir la section sur les tableaux 2×2).
- Si le modèle théorique est une loi paramétrique (normale, Poisson, etc.) dont les m paramètres sont estimés à partir des données, la loi du χ^2 perd un **degré de liberté supplémentaire** pour chaque paramètre estimé : $D \approx \chi^2_{J-1-m}$.

QUIZ!

On compte les pucerons sur 200 feuilles de soya et on veut vérifier si le nombre de pucerons par feuille suit une loi de **Poisson**. On estime $\hat{\lambda} = \bar{x} = 1,2$, ce qui donne les fréquences espérées suivantes :

Pucerons	0	1	2	3	4	5 et +	Total
E_j	60,24	72,29	43,37	17,35	5,20	1,55	200

- Les conditions d'application sont-elles respectées? Sinon, que faire?

QUIZ!

On compte les pucerons sur 200 feuilles de soya et on veut vérifier si le nombre de pucerons par feuille suit une loi de **Poisson**. On estime $\hat{\lambda} = \bar{x} = 1,2$, ce qui donne les fréquences espérées suivantes :

Pucerons	0	1	2	3	4	5 et +	Total
E_j	60,24	72,29	43,37	17,35	5,20	1,55	200

- Les conditions d'application sont-elles respectées? Sinon, que faire?
- Combien de degrés de liberté aura la loi du χ^2 utilisée?

QUIZ!

On compte les pucerons sur 200 feuilles de soya et on veut vérifier si le nombre de pucerons par feuille suit une loi de **Poisson**. On estime $\hat{\lambda} = \bar{x} = 1,2$, ce qui donne les fréquences espérées suivantes :

Pucerons	0	1	2	3	4	5 et +	Total
E_j	60,24	72,29	43,37	17,35	5,20	1,55	200

- Les conditions d'application sont-elles respectées? Sinon, que faire?
- Combien de degrés de liberté aura la loi du χ^2 utilisée?
- Et si la valeur $\lambda = 1,2$ avait été fixée à l'avance (sans utiliser les données)?

Test d'homogénéité de I populations

Test d'homogénéité

On s'intéresse à la distribution d'une variable discrète dans I populations. On veut savoir si cette distribution est la même pour toutes les populations ou non, i.e. si ces populations sont **homogènes** ou non.

- Distribution de la variable dans la population 1 : $(p_{11}, p_{12}, \dots, p_{1J})$
- Distribution de la variable dans la population 2 : $(p_{21}, p_{22}, \dots, p_{2J})$
- $\vdots$
- Distribution de la variable dans la population I : $(p_{I1}, p_{I2}, \dots, p_{IJ})$

Les hypothèses du test

On voudra tester les hypothèses suivantes :

H_0 : les I distributions sont toutes les mêmes

H_1 : au moins une de ces distributions diffère des autres

- On tire un échantillon de taille n_1 dans la population 1;
- On tire un échantillon de taille n_2 dans la population 2;
- $\vdots$
- On tire un échantillon de taille n_I dans la population I .

Les hypothèses du test

On voudra tester les hypothèses suivantes :

H_0 : les I distributions sont toutes les mêmes

H_1 : au moins une de ces distributions diffère des autres

- On tire un échantillon de taille n_1 dans la population 1 ;
- On tire un échantillon de taille n_2 dans la population 2 ;
- $\vdots$
- On tire un échantillon de taille n_I dans la population I .

Les tailles $n_1, \dots, n_I$ sont **fixées** par le plan d'échantillonnage.

Tableau de fréquences

Population	Valeurs de la variable				Total
	v_1	v_2	$\cdots$	v_J	
1	O_{11}	O_{12}	$\cdots$	O_{1J}	n_1
2	O_{21}	O_{22}	$\cdots$	O_{2J}	n_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
I	O_{I1}	O_{I2}	$\cdots$	O_{IJ}	n_I
Total	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdots$	$O_{\cdot J}$	n

Exemple 2 : vivre seul après le décès du conjoint

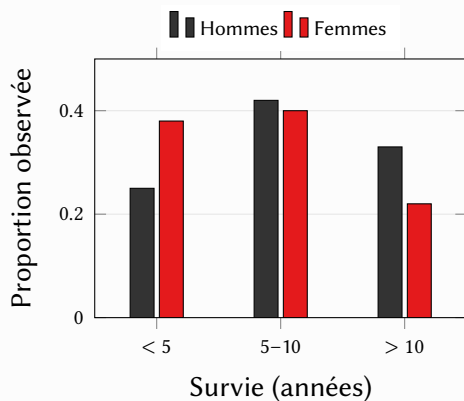
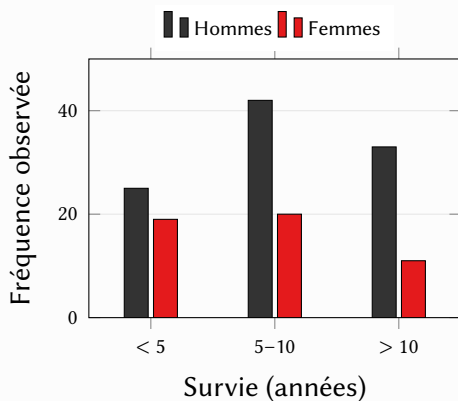
On veut comparer la capacité à vivre seul(e) des hommes et des femmes après le décès de leur conjoint(e). On obtient des données sur 150 personnes.

Population	Années vécues après le décès			Total
	< 5 ans	5 à 10 ans	> 10 ans	
Hommes (veufs)	25	42	33	100
Femmes (veuves)	19	20	11	50
Total	44	62	44	150

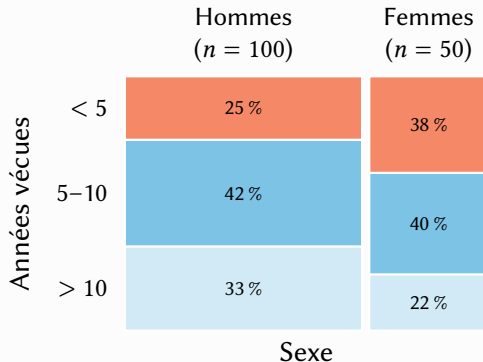
Source : Walpole et Myers, *Probability and Statistics for Engineers and Scientists*.

Exemple 2 : quel graphique ?

Lequel des deux graphiques ci-dessous illustre le mieux les distributions que nous souhaitons comparer ?



Exemple 2 : le diagramme en mosaïque



En R : `mosaicplot(table)`.

- La **largeur** de chaque colonne est proportionnelle à la taille de l'échantillon ($n_1 = 100$, $n_2 = 50$).
- La **hauteur** de chaque rectangle donne la proportion à l'intérieur de la population.
- Sous H_0 (homogénéité), les lignes horizontales seraient à la **même hauteur** dans chaque colonne.

Si H_0 était vraie...

Si H_0 est vraie, toutes les distributions théoriques sont homogènes, i.e. égales à une distribution commune (inconnue) $p_1, p_2, \dots, p_J$.

On pourra estimer ces proportions comme suit :

$$\hat{p}_1 =$$

$$\hat{p}_2 =$$

$$\vdots$$

$$\hat{p}_J =$$

Calcul des fréquences espérées

Si H_0 est vraie, on s'attend à ce que :

$$\begin{array}{lcl} \frac{O_{11}}{n_1} \approx \frac{O_{\cdot 1}}{n} & \Longleftrightarrow & O_{11} \approx \\ \frac{O_{21}}{n_2} \approx \frac{O_{\cdot 1}}{n} & \Longleftrightarrow & O_{21} \approx \\ & \vdots & \\ \frac{O_{ij}}{n_i} \approx \frac{O_{\cdot j}}{n} & \Longleftrightarrow & O_{ij} \approx \end{array}$$

Règle de décision

Si H_0 est vraie,

$$D = \sum_{i=1}^I \sum_{j=1}^J \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim$$

On rejettera donc H_0 si la « distance » entre le modèle d'homogénéité et les observations est trop grande, i.e. si...

QUIZ!

Dans l'exemple 2, on connaît les totaux des lignes et des colonnes. Combien de cases peut-on remplir « librement » avant que toutes les autres soient déterminées ?

Population	< 5 ans	5 à 10 ans	> 10 ans	Total
Hommes	?	?	?	100
Femmes	?	?	?	50
Total	44	62	44	150

QUIZ!

Dans l'exemple 2, on connaît les totaux des lignes et des colonnes. Combien de cases peut-on remplir « librement » avant que toutes les autres soient déterminées ?

Population	< 5 ans	5 à 10 ans	> 10 ans	Total
Hommes	?	?	?	100
Femmes	?	?	?	50
Total	44	62	44	150

- Quel lien avec les degrés de liberté $(I - 1)(J - 1)$?

QUIZ!

Dans l'exemple 2, on connaît les totaux des lignes et des colonnes. Combien de cases peut-on remplir « librement » avant que toutes les autres soient déterminées ?

Population	< 5 ans	5 à 10 ans	> 10 ans	Total
Hommes	?	?	?	100
Femmes	?	?	?	50
Total	44	62	44	150

- Quel lien avec les degrés de liberté $(I - 1)(J - 1)$?
- Combien de degrés de liberté aurait un tableau comparant 4 populations sur 5 catégories ?

Exemple 2

Au seuil de 1 %, peut-on dire que la distribution du nombre d'années de vie est la même chez les hommes et chez les femmes ?

Population	Années vécues après le décès			Total
	< 5 ans	5 à 10 ans	> 10 ans	
Hommes (veufs)	25	42	33	100
Femmes (veuves)	19	20	11	50
Total	44	62	44	150

Exemple 2 : calcul de la distance

$$D = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \dots$$

Exemple 2 : conclusion

Test d'indépendance de deux variables discrètes

Test d'indépendance

On est ici en présence de **deux variables**, toutes les deux qualitatives ou quantitatives discrètes, mesurées sur les unités d'un **même échantillon**. On voudra tester les hypothèses

H_0 : les deux variables sont indépendantes

H_1 : les deux variables sont dépendantes.

Indépendance

Deux variables sont indépendantes si le fait de connaître la valeur de l'une des deux variables ne nous apprend rien sur la valeur de l'autre.

Exemple 3 : génétique et alcoolisme

Une étude est menée en Virginie auprès de 1030 paires de femmes jumelles pour étudier un lien possible entre la génétique et l'alcoolisme. Une entrevue a permis de classer chaque femme comme ayant une consommation abusive d'alcool ou non. Voici les résultats.

Jumelles	Nb de jumelles avec problème d'alcool			Total
	0	1	2	
Identiques	443	102	45	590
Fraternelles	301	113	26	440
Total	744	215	71	1030

K.S. Kendler et al., A population-based twin study of alcoholism in women, *Journal of the American Medical Association*, 268 (1992).

Exemple 3 : sous l'indépendance

Y a-t-il une relation significative entre le type de jumelles et la présence de problème d'alcool ? Si les variables étaient indépendantes :

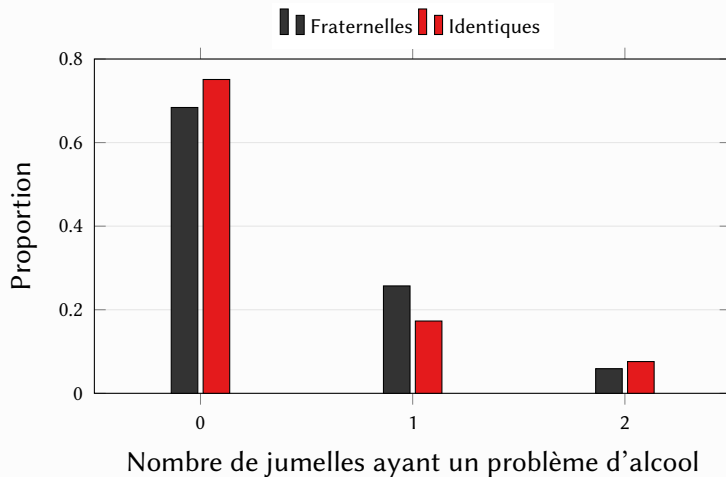
- on devrait observer la même proportion de paires sans problème d'alcool parmi les jumelles identiques que parmi les jumelles fraternelles (i.e. $744/1030 \approx 0,72$);

Exemple 3 : sous l'indépendance

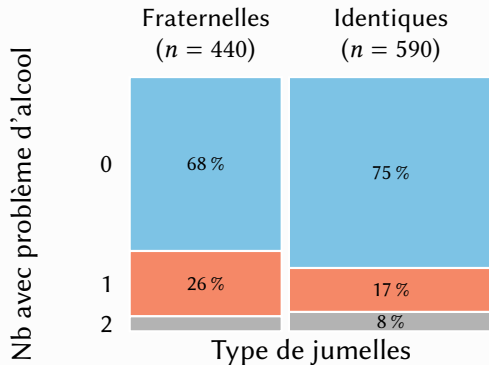
Y a-t-il une relation significative entre le type de jumelles et la présence de problème d'alcool ? Si les variables étaient indépendantes :

- on devrait observer la même proportion de paires sans problème d'alcool parmi les jumelles identiques que parmi les jumelles fraternelles (i.e. $744/1030 \approx 0,72$);
- on devrait observer la même proportion de jumelles identiques parmi les paires sans problème d'alcool que parmi celles avec problème d'alcool (i.e. $590/1030 \approx 0,57$).

Exemple 3 : quelle est votre impression ?

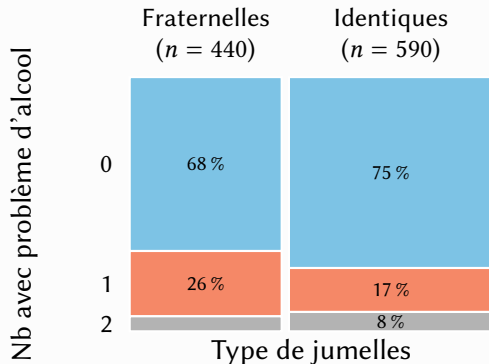


Exemple 3 : le diagramme en mosaïque



- Colonnes : type de jumelles (largeurs $\propto$ 440 et 590 paires).
- Rectangles : nombre de jumelles ayant un problème d'alcool (0, 1 ou 2).
- Si les variables étaient **indépendantes**, les découpages horizontaux seraient les mêmes dans les deux colonnes.

Exemple 3 : le diagramme en mosaïque



- Colonnes : type de jumelles (largeurs $\propto$ 440 et 590 paires).
- Rectangles : nombre de jumelles ayant un problème d'alcool (0, 1 ou 2).
- Si les variables étaient **indépendantes**, les découpages horizontaux seraient les mêmes dans les deux colonnes.

Ici, la bande « 1 » (paires **discordantes**) semble plus épaisse chez les fraternelles.

Tableau de fréquences

Variable 1	Valeurs de la variable 2				Total
	v_1	v_2	$\dots$	v_J	
u_1	O_{11}	O_{12}	$\dots$	O_{1J}	$O_{1\cdot}$
u_2	O_{21}	O_{22}	$\dots$	O_{2J}	$O_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
u_I	O_{I1}	O_{I2}	$\dots$	O_{IJ}	$O_{I\cdot}$
Total	$O_{\cdot 1}$	$O_{\cdot 2}$	$\dots$	$O_{\cdot J}$	n

Tableau de fréquences

Variable 1	Valeurs de la variable 2				Total
	v_1	v_2	$\dots$	v_J	
u_1	O_{11}	O_{12}	$\dots$	O_{1J}	$O_{1\cdot}$
u_2	O_{21}	O_{22}	$\dots$	O_{2J}	$O_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
u_I	O_{I1}	O_{I2}	$\dots$	O_{IJ}	$O_{I\cdot}$
Total	$O_{\cdot 1}$	$O_{\cdot 2}$	$\dots$	$O_{\cdot J}$	n

Ici, seule la taille totale n est fixée à l'avance.

Calcul des fréquences espérées

Si H_0 est vraie, on s'attend à ce que :

$$\frac{O_{11}}{O_{1\bullet}} \approx \frac{O_{\bullet 1}}{n} \iff O_{11} \approx$$

$$\frac{O_{21}}{O_{2\bullet}} \approx \frac{O_{\bullet 1}}{n} \iff O_{21} \approx$$

$\vdots$

$$\frac{O_{ij}}{O_{i\bullet}} \approx \frac{O_{\bullet j}}{n} \iff O_{ij} \approx$$

Règle de décision

Si H_0 est vraie,

$$D = \sum_{i=1}^I \sum_{j=1}^J \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim$$

On rejettera donc H_0 si la « distance » entre le modèle d'indépendance et les observations est trop grande, i.e. si...

Exemple 3

Peut-on dire au seuil de 5 % que la relation entre le type de jumelles et la présence de problème d'alcool est significative ?

Jumelles	Problème d'alcool			Total
	0	1	2	
Identiques	443 426,2 0,66	102 123,16 3,63	45	590
Fraternelles	301 317,8 0,89	113 91,84 4,87	26	440
Total	744	215	71	1030

Dans chaque case : fréquence observée O_{ij} , fréquence espérée E_{ij} et contribution $(O_{ij} - E_{ij})^2 / E_{ij}$.

Exemple 3 : conclusion

Interpréter un rejet : où sont les écarts ?

Rejeter H_0 dit *qu'il y a* un écart au modèle, pas *où*. On examine alors chaque case :

Contributions et résidus

Contribution de la case (i, j) à D : $\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ **Résidu** : $r_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$

Le **signe** du résidu indique s'il y a plus (+) ou moins (−) de cas qu'attendu sous H_0 ; un $|r_{ij}|$ supérieur à 2 environ est notable.

Résidus r_{ij} (ex. 3)	0	1	2
Identiques	+0,82	−1,91	+0,68
Fraternelles	−0,94	+2,21	−0,79

Interpréter un rejet : où sont les écarts ?

Rejeter H_0 dit *qu'il y a* un écart au modèle, pas *où*. On examine alors chaque case :

Contributions et résidus

Contribution de la case (i, j) à D : $\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ **Résidu** : $r_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$

Le **signe** du résidu indique s'il y a plus (+) ou moins (−) de cas qu'attendu sous H_0 ; un $|r_{ij}|$ supérieur à 2 environ est notable.

Résidus r_{ij} (ex. 3)	0	1	2
Identiques	+0,82	−1,91	+0,68
Fraternelles	−0,94	+2,21	−0,79

La colonne « 1 » fournit $3,63 + 4,87 = 8,50$ des 11,14 unités de D (76 %) : les jumelles fraternelles sont plus souvent **discordantes** (une seule a un problème) que prévu, les identiques moins.

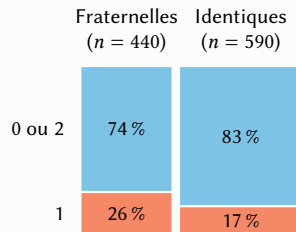
Question

Si le but de l'étude est d'étudier l'existence d'un lien entre la **génétique** et l'alcoolisme, est-ce que le test d'indépendance sur ces données répondait à la question ?

Exemple 3 revisité

Jumelles	Jumelles ayant un problème d'alcool		Total
	0 ou 2 (concordantes)	1 (discordantes)	
Identiques	488	102	590
Fraternelles	327	113	440
Total	815	215	1030

Au seuil de 5 %, la proportion de paires concordantes est-elle la même chez les jumelles identiques et fraternelles ?



En R

Test d'adéquation (exemple 1) :

```
> chisq.test(c(926, 288, 293, 104), p = c(9, 3, 3, 1)/16)
X-squared = 1.4687, df = 3, p-value = 0.6895
```

Test d'homogénéité ou d'indépendance (exemple 3) :

```
> jumelles <- matrix(c(443, 102, 45, 301, 113, 26), nrow = 2, byrow = TRUE)
> test <- chisq.test(jumelles)
> test
X-squared = 11.141, df = 2, p-value = 0.003808
> test$expected      # fréquences espérées
> test$residuals     # résidus (O - E)/sqrt(E)
```

Dans R Commander : menu Statistiques → Tableaux de contingence → Saisir et analyser un tableau à double entrée.

Tableaux 2×2 et comparaison de deux proportions

Rappel : au chapitre 4...

Un tableau d'homogénéité avec $I = 2$ populations et $J = 2$ catégories (« succès » / « échec ») revient à tester $H_0 : p_1 = p_2$, ce qu'on a déjà fait au chapitre 4 avec le test Z de comparaison de deux proportions.

Tableau 2×2 : khi-deux et test Z

Pour un tableau 2×2 , la distance du khi-deux est exactement le carré de la statistique Z du chapitre 4 :

$$D = Z^2 \quad \text{et} \quad \chi^2_{1;\alpha} = z^2_{\alpha/2} \quad (\text{p. ex. } 3,84 = 1,96^2).$$

Les deux tests bilatéraux donnent donc la même valeur-p.

Le test Z permet en plus un test **unilatéral** et un IC pour $p_1 - p_2$; le khi-deux se généralise à plus de 2 groupes ou de 2 catégories.

Exemple 4 : tabagisme de la mère et poids à la naissance

On suit 400 grossesses : 150 mères ayant fumé pendant la grossesse et 250 non-fumeuses. On note si le bébé présente un **faible poids** à la naissance (< 2500 g).

	Faible poids	Poids normal	Total
Fumeuses	30	120	150
Non-fumeuses	25	225	250
Total	55	345	400

Au seuil de 5 %, la proportion de bébés de faible poids diffère-t-elle selon le tabagisme de la mère ? Répondez

- (a) avec le test Z du chapitre 4;
- (b) avec le test du khi-deux.

Données fictives, mais d'un ordre de grandeur réaliste.

Exemple 4 : les deux approches

(a) Test Z. $\hat{p}_1 = 30/150 = 0,20$, $\hat{p}_2 = 25/250 = 0,10$, $\hat{P} = 55/400 = 0,1375$.

$$z_{\text{obs}} = \frac{0,20 - 0,10}{\sqrt{0,1375 \times 0,8625 \left(\frac{1}{150} + \frac{1}{250} \right)}} = \frac{0,10}{0,0356} = 2,81 > z_{0,025} = 1,96.$$

Exemple 4 : les deux approches

(a) **Test Z.** $\hat{p}_1 = 30/150 = 0,20$, $\hat{p}_2 = 25/250 = 0,10$, $\hat{P} = 55/400 = 0,1375$.

$$z_{\text{obs}} = \frac{0,20 - 0,10}{\sqrt{0,1375 \times 0,8625 \left(\frac{1}{150} + \frac{1}{250} \right)}} = \frac{0,10}{0,0356} = 2,81 > z_{0,025} = 1,96.$$

(b) **Khi-deux.** Fréquences espérées $E_{ij} = O_{i.}O_{.j}/n$ (toutes ≥ 5) :

	Faible poids	Poids normal
Fumeuses	30 (20,625)	120 (129,375)
Non-fumeuses	25 (34,375)	225 (215,625)

$$d_{\text{obs}} = 4,26 + 0,68 + 2,56 + 0,41 = 7,91 > \chi_{1;0,05}^2 = 3,84.$$

Exemple 4 : les deux approches

(a) Test Z. $\hat{p}_1 = 30/150 = 0,20$, $\hat{p}_2 = 25/250 = 0,10$, $\hat{P} = 55/400 = 0,1375$.

$$z_{\text{obs}} = \frac{0,20 - 0,10}{\sqrt{0,1375 \times 0,8625 \left(\frac{1}{150} + \frac{1}{250} \right)}} = \frac{0,10}{0,0356} = 2,81 > z_{0,025} = 1,96.$$

(b) Khi-deux. Fréquences espérées $E_{ij} = O_{i.}O_{.j}/n$ (toutes ≥ 5) :

	Faible poids	Poids normal
Fumeuses	30 (20,625)	120 (129,375)
Non-fumeuses	25 (34,375)	225 (215,625)

$$d_{\text{obs}} = 4,26 + 0,68 + 2,56 + 0,41 = 7,91 > \chi_{1;0,05}^2 = 3,84.$$

Même conclusion, même valeur-p (0,005) : $d_{\text{obs}} = 7,91 = 2,81^2 = z_{\text{obs}}^2$. On rejette H_0 : la proportion de faible poids est plus élevée chez les fumeuses (20 % contre 10 %).

QUIZ!

Dans l'exemple 4 :

- Peut-on conclure que le tabagisme de la mère **cause** le faible poids à la naissance?

QUIZ!

Dans l'exemple 4 :

- Peut-on conclure que le tabagisme de la mère **cause** le faible poids à la naissance?
- On voulait en fait montrer que la proportion est **plus élevée** chez les fumeuses. Quel test utiliser? Conclusion?

QUIZ!

Dans l'exemple 4 :

- Peut-on conclure que le tabagisme de la mère **cause** le faible poids à la naissance?
- On voulait en fait montrer que la proportion est **plus élevée** chez les fumeuses. Quel test utiliser? Conclusion?
- Si on avait plutôt choisi 400 naissances au hasard (sans fixer le nombre de fumeuses), quel test ferait-on? La valeur de D changerait-elle?

Petits effectifs : le test exact de Fisher

- Si une fréquence espérée est < 5 dans un tableau 2×2 (maladie rare, petit nombre d'animaux...), l'approximation χ^2 n'est plus fiable.
- Le **test exact de Fisher** calcule la valeur-p exactement, à partir de toutes les tables ayant les mêmes totaux (calcul fait par le logiciel).
- Par défaut, `chisq.test()` applique aux tableaux 2×2 une **correction de continuité** (Yates) qui diminue un peu D; pour retrouver $D = Z^2$, on ajoute `correct = FALSE`.

```
> bebe <- matrix(c(30, 120, 25, 225), nrow = 2, byrow = TRUE)
> chisq.test(bebe, correct = FALSE)
X-squared = 7.9051, df = 1, p-value = 0.004929
> prop.test(c(30, 25), c(150, 250), correct = FALSE)    # chapitre 4
X-squared = 7.9051, df = 1, p-value = 0.004929
> chisq.test(bebe)          # avec correction de Yates
X-squared = 7.0844, df = 1, p-value = 0.007776
> fisher.test(bebe)         # test exact
p-value = 0.006662
```

Synthèse : quel test choisir ?

Trois tests, une même distance

	Adéquation	Homogénéité	Indépendance
Variables	1	1	2
Échantillons	1 (taille n)	I (tailles n_i fixées)	1 (seul n fixé)
H_0	$p_j = p_j^*$	mêmes distributions dans les I populations	variables indépendantes
E	$E_j = np_j^*$	$E_{ij} = n_i O_{.j} / n$	$E_{ij} = O_{i.} O_{.j} / n$
Degrés de liberté	$J - 1$ ($-m$ si m paramètres estimés)	$(I - 1)(J - 1)$	$(I - 1)(J - 1)$

- Dans tous les cas : $D = \sum (O - E)^2 / E$, rejet si $D > \chi^2_{dl; \alpha}$, condition $E \geq 5$ partout.
- Homogénéité et indépendance : **mêmes calculs**; seuls le plan d'échantillonnage et la formulation de la conclusion changent.

QUIZ!

Quel test du khi-deux utiliser ? Combien de degrés de liberté ?

- On capture 60 truites dans chacun de 3 lacs et on note si elles sont infectées par un parasite (oui/non).

QUIZ!

Quel test du khi-deux utiliser? Combien de degrés de liberté?

- On capture 60 truites dans chacun de 3 lacs et on note si elles sont infectées par un parasite (oui/non).
- On croise des plants de pois hétérozygotes et on obtient 705 fleurs pourpres et 224 blanches. Le rapport mendélien 3 :1 est-il respecté?

QUIZ!

Quel test du khi-deux utiliser ? Combien de degrés de liberté ?

- On capture 60 truites dans chacun de 3 lacs et on note si elles sont infectées par un parasite (oui/non).
- On croise des plants de pois hétérozygotes et on obtient 705 fleurs pourpres et 224 blanches. Le rapport mendélien 3 :1 est-il respecté ?
- On choisit 500 patients au hasard dans une clinique et on note leur groupe sanguin (O, A, B, AB) et la présence d'un ulcère gastrique (oui/non).

QUIZ!

Quel test du khi-deux utiliser ? Combien de degrés de liberté ?

- On capture 60 truites dans chacun de 3 lacs et on note si elles sont infectées par un parasite (oui/non).
- On croise des plants de pois hétérozygotes et on obtient 705 fleurs pourpres et 224 blanches. Le rapport mendélien 3 :1 est-il respecté ?
- On choisit 500 patients au hasard dans une clinique et on note leur groupe sanguin (O, A, B, AB) et la présence d'un ulcère gastrique (oui/non).
- Dans un essai clinique, 120 patients reçoivent un vaccin et 120 un placebo ; on note s'ils contractent la grippe durant l'hiver.

Réponses aux exemples (1)

- $E_j = np_j^*$. **Exemple 1** : $E = (906,19; 302,06; 302,06; 100,69)$.
- Distance : $D = \sum_j (O_j - E_j)^2 / E_j$. Si H_0 est vraie, $D \approx \chi_{J-1}^2$; on rejette H_0 si $D > \chi_{J-1;\alpha}^2$.
- **Exemple 1** : contributions 0,43; 0,65; 0,27; 0,11, donc $D = 1,47 < \chi_{3;0,05}^2 = 7,81$ (valeur- $p = 0,69$) : on ne rejette pas l'hypothèse de Mendel.
- **Quiz (distance)** : un grand D signifie que les observations s'éloignent plus du modèle que ne l'expliquent les fluctuations d'échantillonnage (preuve contre H_0). $D = 0$: $O_j = E_j$ partout, accord parfait; on ne rejette pas H_0 (mais un accord « trop parfait », répété d'une expérience à l'autre, est suspect : Fisher a jugé en 1936 les données de Mendel trop proches de la théorie). $D \geq 0$ toujours (somme de carrés divisés par des $E_j > 0$). Si n double avec les mêmes proportions, D double : un même écart relatif devient plus significatif.
- **Quiz (pucerons)** : non, $E_6 = 1,55 < 5$. On regroupe « 4 et plus » ($E = 6,75$), d'où $J = 5$ classes et $5 - 1 - 1 = 3$ degrés de liberté (λ estimé). Si λ était fixé à l'avance : $5 - 1 = 4$ degrés de liberté.
- **Exemple 2** : le graphique des *proportions* (les tailles d'échantillon diffèrent). $\hat{p}_j = O_{\cdot j}/n$ et $E_{ij} = n_i O_{\cdot j}/n$. Sous H_0 , $D \approx \chi_{(I-1)(J-1)}^2$. E : hommes 29,33; 41,33; 29,33; femmes 14,67; 20,67; 14,67. $D = 3,33 < \chi_{2;0,01}^2 = 9,21$ (valeur- $p = 0,19$) : on ne rejette pas l'homogénéité.
- **Quiz (degrés de liberté)** : 2 cases (p. ex. deux cases de la ligne « Hommes »); toutes les autres se déduisent des totaux. C'est exactement $(I - 1)(J - 1) = (2 - 1)(3 - 1) = 2$. Avec 4 populations et 5 catégories : $3 \times 4 = 12$.

Réponses aux exemples (2)

- **Exemple 3** : $E_{ij} = O_{i.}O_{.j}/n$; $E_{13} = 40,67$, $E_{23} = 30,33$ (contributions 0,46 et 0,62). $D = 11,14 > \chi^2_{2; 0,05} = 5,99$ (valeur-p = 0,004) : relation significative.
- **Question** : pas directement ; le lien génétique se manifeste par la *concordance* des jumelles (0 ou 2) plus fréquente chez les identiques.
- **Revisité** : $D = 10,75 > \chi^2_{1; 0,05} = 3,84$ (valeur-p = 0,001) : les jumelles identiques sont plus souvent concordantes (83 % contre 74 %). (Avec la correction de Yates, `chisq.test()` donne 10,25.)
- **Exemple 4** : $z_{\text{obs}} = 2,81$, $d_{\text{obs}} = 7,91 = z_{\text{obs}}^2$, valeur-p = 0,005 : on rejette H_0 . Test exact de Fisher : valeur-p = 0,007.
- **Quiz (exemple 4)** : non : il s'agit d'une étude *observationnelle*; les fumeuses peuvent différer des non-fumeuses (âge, alimentation, revenu, suivi prénatal...). Le test montre une **association**, pas une causalité. Test unilatéral ($H_1 : p_1 > p_2$) : test Z du chapitre 4, $z_{\text{obs}} = 2,81 > z_{0,05} = 1,645$, valeur-p = 0,0025. Avec 400 naissances au hasard : test d'indépendance; les calculs, donc D, sont identiques, seule la formulation de la conclusion change.
- **Quiz (quel test ?)** : truites : homogénéité (3 lacs $\times$ 2 catégories), $(3 - 1)(2 - 1) = 2$ dl. Pois : adéquation, $2 - 1 = 1$ dl ($E = 696,75$ et $232,25$, $D = 0,39$, valeur-p = 0,53 : rapport 3 : 1 compatible avec les données). Groupe sanguin et ulcère : indépendance, $(4 - 1)(2 - 1) = 3$ dl. Vaccin : homogénéité 2×2 , c'est-à-dire une comparaison de deux proportions (chapitre 4), 1 dl.