Tableaux de contingence et tests Chi–2

Chapitre I

Anouar El Ghouch

LSBA, Université catholique de Louvain, Belgium

Introduction

Définitions et Rappels

Une variable catégorielle ou discrète est une variable aléatoire (va) qui ne peut prendre qu'un nombre dénombrable (fini ou infini) de valeurs possibles (quantitatives ou qualitatives).

Type de variables catégorielles:

» Numérique discrète: Le résultat du lancer d'un dé $(1, \ldots, 6)$; Nombre d'années d'études après le bac; Nombre de personnes frappées par une maladie.

» Nominale: Sexe; Religion; Profession; Nationalité. L'ordre des modalités n'a pas de sens, et l'analyse statistique ne doit pas en dépendre.

» Ordinale: État de santé du patient (mauvais, passable, bon, excellent); Niveau de satisfaction ($1 =$ très insatisfait, $\ldots, 5 =$ très satisfait). Les modalités peuvent être ordonnées bien que la distance entre elles ne soit pas toujours quantifiable.

» Intervalle: Revenu annuel ($\leqslant 10000, \ldots, > 250000$); Âge ($\leqslant 18, \ldots, > 80$). Ce sont des variables de nature continue qui ont été catégorisées (regroupement des valeurs en catégories).

Quelques remarques

- Le type des variables (Nominale, Ordinale, ...) influence le choix de la méthode d'analyse à appliquer.
- Le plus souvent, l'intérêt d'une étude statistique réside dans la détection et l'analyse des interactions possibles entre deux ou plusieurs variables observées.
- En fonction du contexte et de la nature de l'étude, ces interactions sont traitées soit de façon symétrique, soit de façon non-symétrique. Dans le premier cas, les variables sont traitées de façon identique alors que dans le deuxième cas les variables sont réparties en deux groupes: variable(s) à expliquer et variable(s) explicative(s) (prédicteur(s)).
- Par exemple, on peut étudier (de façon symétrique) l'association entre la couleur des yeux et la couleur de cheveux. Ou étudier comment le niveau d'éducation de la mère affecte les performances scolaires de son enfant (et pas l'inverse !).

Distributions discrètes

Les distributions de probabilité les plus courantes pour les données catégorielles sont

- Distribution Binomiale
- Distribution Multinomiale
- Distribution de Poisson

Ces distributions sont brièvement passées en revue dans les slides suivantes.

Distribution Binomiale

- On considère une expérience "Bernoulli" avec deux résultats possibles étiquetés

  échec (E) et succès (S)

- Cette expérience est répétée n fois, de manière *indépendante* et *identique*.

- Soit Y_i, $i = 1, \ldots, n$, la va correspondant au résultat de la ième répétition/expérience. $Y_1, \ldots, Y_n \in \{0 \equiv E, 1 \equiv S\}$.

- La probabilité de succès $p = P(Y_i = 1)$ est la même $\forall i = 1, \ldots, n$.

- Soit $Y = \sum_{i=1}^{n} Y_i$ = nombre total de succès obtenus parmi n. Aux conditions décrites ci-dessus,

$$Y \sim Bin(n, p), \quad \text{càd}$$

$$P(Y = y) = C_n^y p^y (1-p)^{n-y}, \; y = 0, 1, \ldots, n,$$

où $C_n^y = \frac{n!}{y!(n-y)!}$.

À titre d'exemple, considérons le cas de $n = 100$ individus interrogés et classés selon leurs habitudes tabagiques: fumeur (Fum) ou non-fumeur (Nfu).

Fum (1)	Nfu (0)
40	60

$Y_i = 1(0)$ si l'un individu i, $i = 1, \ldots, 100$, est fumeur (non-fumeur)

Y_i, $i = 1, \ldots, 100$, est notre échantillon aléatoire.

$Y = \sum_{i=1}^{n} Y_i =$ nbr. total de fum. parmi n.

$Y \sim \text{Bin}(100, p)$, avec $p = P(\text{un individu tiré au hasard fait partie des fumeurs})$

La valeur qu'on a observée $y = 40$ est une simple réalisation de Y.

La Binomiale dans R: les fonctions `dbinom()`, `pbinom()`, `rbinom()`

Soit $Y \sim \text{Bin}(n, p)$. Voici comment calculer des probabilités sur Y à l'aide de R.

Prob.	Commande R
$P(Y = y)$	`dbinom(y, size = n, prob = p)`
$P(Y \leqslant y)$	`pbinom(y, size = n, prob = p)`
$P(Y > y)$	`pbinom(y, size = n, prob = p, low = FALSE)`

Par exemple, voici $P(Y = 5)$ pour $Y \sim \text{Bin}(15, 0.3)$,

```
dbinom(5, size = 15, prob = 0.3)
```

```
[1] 0.20613
```

On peut aussi simuler des résiliations d'une Binomiale à l'aide de la fonction `rbinom()`. Par exemple, voici comment générer 10 réalisations qui proviennent d'une $\text{Bin}(15, 0.3)$.

```
rbinom(10, size = 15, prob = 0.3)
```

```
 [1] 7 5 2 5 6 5 6 1 2 9
```

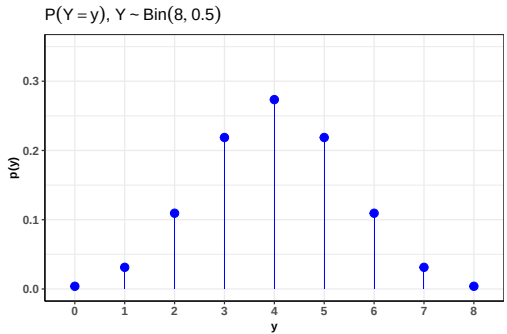
Quelques propriétés.

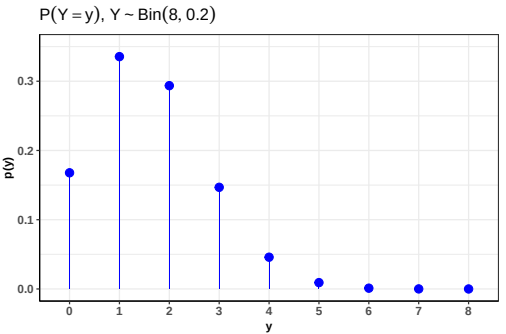
- $E(Y) = np$ et $V(Y) = np(1-p)$
- La somme de variables binomiales indépendantes ($\perp\!\!\!\perp$) *avec la même probabilité de succès* est une variable binomiale:

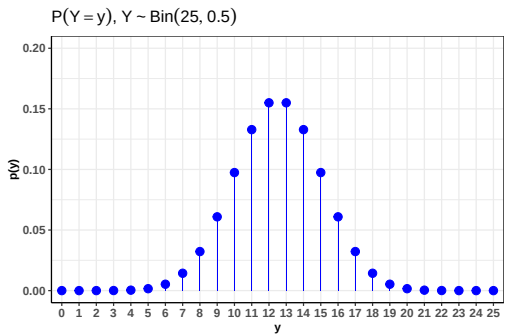
$$\sum_k Bin(n_k, p) = Bin(\sum_k n_k, p)$$

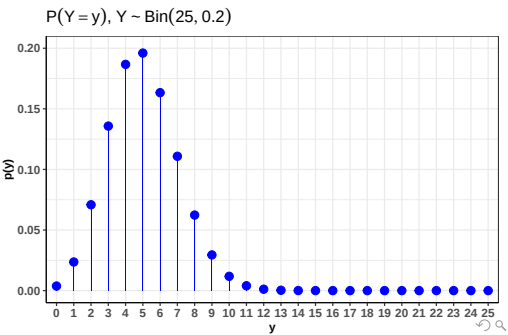
- Pour un n "suffisamment" grand (np et $np(1-p) \geqslant 5$),

$$Bin(n, p) \approx N(np, np(1-p))$$


P(Y = y), Y ~ Bin(8, 0.5)
p(y)
y


P(Y = y), Y ~ Bin(8, 0.2)
p(y)
y


P(Y = y), Y ~ Bin(25, 0.5)
p(y)
y


P(Y = y), Y ~ Bin(25, 0.2)
p(y)
y

Distribution Multinomiale

La loi multinomiale est une généralisation de la binomiale au cas où le nombre de résultats possible dépasse 2 :

- Une expérience avec K ($K \geqslant 2$) résultats/catégories possibles $(C_1, C_2, \ldots, C_K)$
- n répétitions *indépendantes et identiques*.
- Le résultat de la ième répétition/expérience est la va $Y_i \in \{1 \equiv C_1, \ldots, K \equiv C_K\}$.
- Les probabilités $p_k = P(Y_i = k)$ sont les même $\forall i = 1, \ldots, n$ et $\sum_k p_k = 1$
- $N_k = \sum_{i=1}^{n} I(Y_i = k) =$ le nombre de fois où le résultat C_k est observé au cours des n répétitions. Aux conditions décrites ci-dessus,

$$(N_1, \ldots, N_K) \sim Mul_{(K)}(n, (p_1, \ldots, p_K)), \text{ càd}$$

$$P(N_1 = n_1, \ldots, N_K = n_K) = \frac{n!}{n_1! \ldots n_K!} p_1^{n_1} \ldots p_K^{n_K},$$

$n_1, \ldots, n_K \in \{0, \ldots, n\}$ et $\sum_k n_k = n$.

À titre d'exemple, considérons le cas de 100 individus interrogés et classés selon leurs habitudes tabagiques: Accro au tabac (Acc), fumeur-occasionnel (Occ), non-fumeur (Nfu).

Acc (1)	Occ (2)	Nfu (3)
30	10	60

$Y_i = 1, 2$, ou 3 si iéme individu est, respectivement, accro au tabac, un fumeur occasionnel, ou non-fumeur, $i = 1, \ldots, 100$.

$N_k =$ nbr. d'individus observés dans la catégorie "k" $(1 \equiv Acc, 2 \equiv Occ, 3 \equiv Nfu)$.
$N_k \sim Bin(100, p_k)$, $p_k = P(\text{ind. tiré au hasard} \in \text{cat. "k"}) = P(Y_i = k)$, $k = 1, 2, 3$.

$(N_1, N_2, N_3) \sim Mul_{(3)}(100, (p_1, p_2, p_3))$

$(n_1, n_2, n_3) = (30, 10, 60)$ est la valeur observé (une réalisation) de (N_1, N_2, N_3).

Quelques propriétés

$(N_1, \ldots, N_K) \sim Mul(n, (p_1, \ldots, p_K)) \Rightarrow N_k \sim Bin(n, p_k)$ et

$$Cov(N_k, N_{k'}) = -np_kp_{k'}$$

Une multinomiale $Mul_{(2)}(n, (p_1, 1-p_1))$ est équivalente à une binomiale $Bin(n, p_1)$. Autrement dit, dire que $Y \sim Bin(n, p_1) \Leftrightarrow (Y, n-Y) \sim Mul_{(2)}(n, (p_1, 1-p_1))$.

La multinomiale dans R: Soit $(N_1, N_2, N_3) \sim Mul(10, (0.25, 0.25, 0.5))$. Pour calculer, par exemple, $P(N_1 = 3, N_2 = 2, N_3 = 5)$, taper

```
dmultinom(c(3, 2, 5), prob = c(0.25, 0.25, 0.5))
```

```
[1] 0.076904
```

Pour générer, par exemple, 5 réalisations qui proviennent de cette distribution, taper

```
rmultinom(5, size = 10, prob = c(0.25, 0.25, 0.5))
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    2    3    4    1    1
[2,]    2    2    3    4    2
[3,]    6    5    3    5    7
```

Distribution de Poisson

Dans beaucoup de situation, les données ne proviennent pas d'un nombre déterminé d'essais.

Exemple: Nombre d'accidents de voiture survenus, dans une région, durant un certain laps de temps. Ce qui est fixé dans une telle expérience, c'est un intervalle dans lequel nous comptons les occurrences d'événements.

L'intervalle est généralement un intervalle du temps mais peut être un espace physique (Nombre de bactérie par microlitre, par exemple).

La distribution la plus simple convenant aux configurations de ce type est la Poisson.

$Y =$ nombre d'occurrences d'un événement dans un intervalle donné $\sim \text{Pois}(\mu)$, si

$$P(Y = y) = \frac{\mu^y}{y!} e^{-\mu}, \quad y = 0, 1, 2, \ldots,$$

où le paramètre $\mu > 0$ est le nombre d'occurrences attendu dans l'intervalle spécifié, càd $\mu = E(Y)$

Quelques propriétés

- $V(Y) = E(Y) = \mu$
- Pour un μ "suffisamment" grand, $\dfrac{Y - \mu}{\sqrt{\mu}} \underset{a}{\sim} N(0,1)$
- Si $N_k \sim Pois(\mu_k)$, $k = 1, \ldots, K$, sont $\perp\!\!\!\perp$ alors
  1. $\sum_k N_k \sim Pois(\mu)$, avec $\mu = \sum_k \mu_k$
  2. $(N_1, \ldots, N_K) | \sum_k N_k = n \sim Mul(n, \mathbf{p})$, avec $\mathbf{p} = (\mu_1/\mu, \ldots, \mu_K/\mu)$ Voir App.

  $\rightarrow$ L'imposition d'un total fixe sur le nombre d'événements observés dans plusieurs populations de Poisson conduit à la distribution multinomiale.

Le tableau suivant donne le nombre d'incidents observés au cours d'un weekend dans trois régions différentes (A, B, C):

Rég. A	Rég. B	Rég. C
30	20	50

Soit N_1, N_2, N_3 les variables aléatoires qui représentent le nombre d'incidents qui peuvent survenir dans A, B, et C, respectivement.

On peut supposer que ces trois variables sont indépendantes et que $N_k \sim \text{Pois}(\mu_k)$ où μ_k , $k = 1, 2, 3$, est le nombre moyen d'incidents dans A, B, ou C.

Si, par exemple, on restreint notre analyse aux 100 incidents les plus graves, alors l'hypothèse Poissonnienne revient à supposer que $(N_1, N_2, N_3) \sim \text{Mul}(100, (p_1, p_2, p_3))$ où $p_k = \mu_k/(\mu_1 + \mu_2 + \mu_3)$, $k = 1, 2, 3$, est la probabilité qu'un incident survienne dans A, B, ou C.

Ainsi, la modélisation par Multinomiale, peut être vue comme un cas particulier de la modalisation par Poisson. Pour cette raison, et parce que l'hypothèse Poissonnienne facilite certains traitements mathématiques, c'est cette dernière qui est le plus souvent utilisée dans la pratique.

La Poisson dans R: les fonctions `dpois()`, `ppois()`, `rpois()`

Soit $Y \sim \text{Pois}(1.8)$. Voici quelques opérations que l'on peut effectuer dans R sur cette variable.

```
# P(Y = 4)
dpois(4, lambda = 1.8)
```

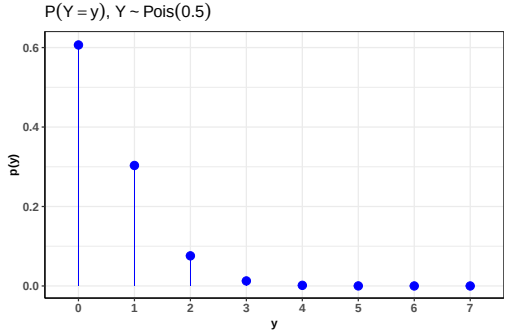
```
[1] 0.072302
```

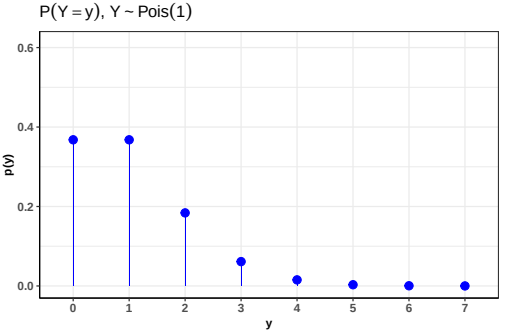
```
# P(Y > 2)
ppois(2, lambda = 1.8, low = FALSE)
```

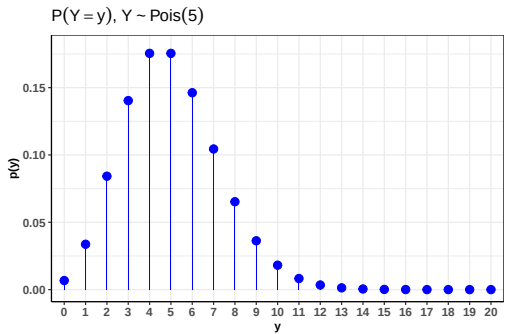
```
[1] 0.26938
```

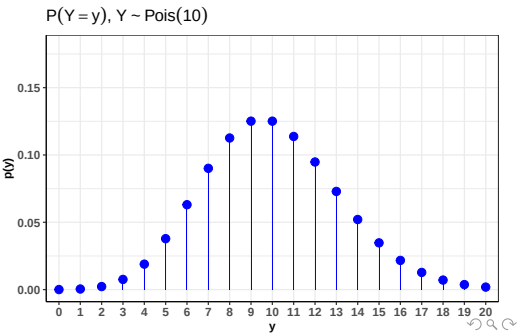
```
# Générer 10 observations
rpois(10, lambda = 1.8)
```

```
 [1] 0 2 2 0 3 1 2 4 3 0
```


P(Y = y), Y ~ Pois(0.5)
p(y)
y


P(Y = y), Y ~ Pois(1)
p(y)
y


P(Y = y), Y ~ Pois(5)
p(y)
y


P(Y = y), Y ~ Pois(10)
p(y)
y

Binomiale ou Poisson ?

Ajuster une distribution aux données

Exemple 1

Voici les fréquences (Freq) de nombre d'enfants de sexe masculin ($nMales$) dans 6115 familles. Chacune de ces familles est composée de 12 enfants.[1]

nMales	0	1	2	3	4	5	6	7	8	9	10	11	12
Freq	3	24	104	286	670	1033	1343	1112	829	478	181	45	7

Quel modèle (distribution) est approprié pour la variable $nMales$?

(a) Binomiale ou (b) Poisson

[1] Étude réalisée en Saxe entre 1876 et 1885; Geissler, A. (*1889*)

Exemple 1

Voici les fréquences (Freq) de nombre d'enfants de sexe masculin ($nMales$) dans 6115 familles. Chacune de ces familles est composée de 12 enfants.[1]

nMales	0	1	2	3	4	5	6	7	8	9	10	11	12
Freq	3	24	104	286	670	1033	1343	1112	829	478	181	45	7

Quel modèle (distribution) est approprié pour la variable $nMales$?

(a) Binomiale ou (b) Poisson

$nMales$ ne peut prendre que les valeurs de 0 à 12 $\rightarrow$ la Binomiale semble plus plausible que la Poisson.

[1] Étude réalisée en Saxe entre 1876 et 1885; Geissler, A. (*1889*)

Question: le modèle Binomiale s'ajuste-t-il correctement aux données; càd peut-on vraiment dire que $nMales \sim Bin(12, p)$, où p est la probabilité (inconnue), pour une famille tirée au hasard, d'avoir un garçon ?

Question: le modèle Binomiale s'ajuste-t-il correctement aux données; càd peut-on vraiment dire que $nMales \sim Bin(12, p)$, où p est la probabilité (inconnue), pour une famille tirée au hasard, d'avoir un garçon ?

Pour répondre à cette question, on peut commencer par

» Estimer p par

$$\hat{p} = \frac{\text{nombre total de garçons}}{\text{nombre total d'enfants}} = \frac{3 \times 0 + 24 \times 1 + \ldots 7 \times 12}{6115 \times 12} = 0.51922$$

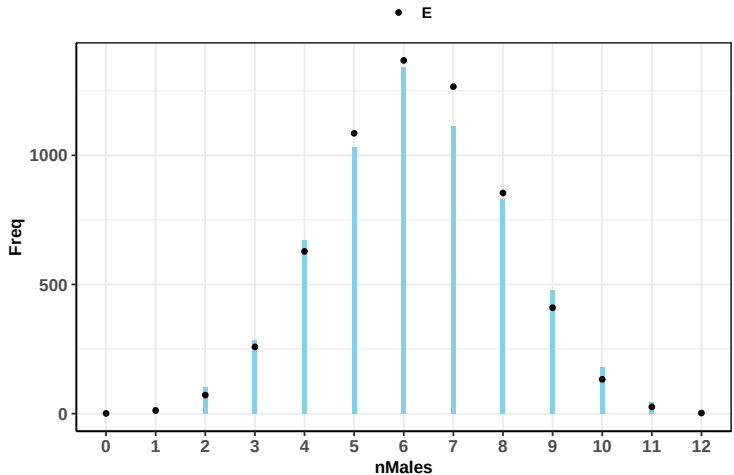
» Puis comparer les probabilités théoriques $P(Bin(12, 0.52) = k)$, $k = 0, \ldots, 12$, aux proportions observées $Freq_k/6115$. Ce qui revient à comparer les effectifs observés $Freq_k$ aux effectifs attendus (E_k) sous l'hypothèse de la Binomial. Ces derniers donnés par

$$E_k = 6115 \times P(Bin(12, \hat{p}) = k),\ k = 0, \ldots, 12$$

```
6115 * dbinom(0:12, size = 12, prob = 0.51922)
```

Les résultats de ces calculs sont résumés dans le tableau et le graphique suivant.

nMales	Freq	E
0	3	0.93
1	24	12.09
2	104	71.80
3	286	258.48
4	670	628.06
5	1033	1085.21
6	1343	1367.28
7	1112	1265.63
8	829	854.25
9	478	410.01
10	181	132.84
11	45	26.08
12	7	2.35



Que peut-on conclure ?

Exemple 2

Voici les fréquences ($Freq$) de nombre de naissances ($nBaby$) par heure dans un hôpital. Ces données concernent une période de 24 heures.

nBaby	0	1	2	3	4
Freq	3	8	6	4	3

Quel modèle (distribution) est approprié pour $nBaby$?

(a) Binomiale ou (b) Poisson

Exemple 2

Voici les fréquences ($Freq$) de nombre de naissances ($nBaby$) par heure dans un hôpital. Ces données concernent une période de 24 heures.

nBaby	0	1	2	3	4
Freq	3	8	6	4	3

Quel modèle (distribution) est approprié pour $nBaby$?

(a) Binomiale ou (b) Poisson

$nBaby$ peut prendre les valeurs $0, 1, 2, ... \rightarrow$ la distribution de Poisson peut être envisageable.

Question : le modèle Poisson s'ajuste-t-il correctement aux données : $\texttt{nBaby} \sim \text{Pois}(\lambda)$, pour un certain $\lambda > 0$ (inconnu) ?

Question : le modèle Poisson s'ajuste-t-il correctement aux données : $nBaby \sim Pois(\lambda)$, pour un certain $\lambda > 0$ (inconnu) ?

» Estimons λ:

$$\hat{\lambda} = \frac{\text{total des naissances}}{\text{temps total}} = \frac{3 \times 0 + 8 \times 1 + 6 \times 2 + 4 \times 3 + 3 \times 4}{3 + 8 + 6 + 4 + 3} = 1.833$$

» Calculons les effectifs attendus (E) sous l'hypothèse de Poisson :

$$E_k = 24 \times P(Pois(\hat{\lambda}) = k),\ k = 0, \ldots, 4$$

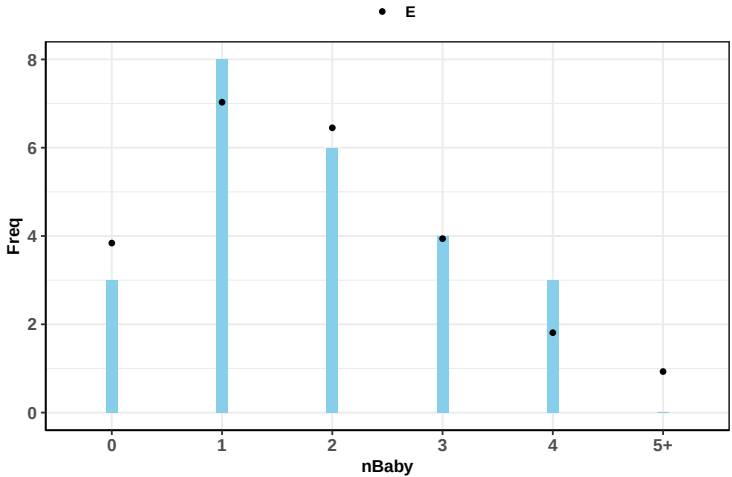
Puisque une Poisson $\in [0, \infty)$, il est logique d'inclure dans nos calculs la catégorie "5+", càd celle qui correspond à $nBaby > 4$.

$$E_{5+} = 24 \times P(Pois(\hat{\lambda}) > 4) \approx 0.93$$

```
24 * dpois(0:4, lambda = 1.8333)
24 * ppois(4, lambda = 1.8333, low = FALSE)
```

Comparons les effectifs observés (Freq) aux effectifs attendus.

nBaby	Freq	E
0	3	3.84
1	8	7.03
2	6	6.45
3	4	3.94
4	3	1.81
5+	0	0.93



Que peut-on conclure ?

Maximum de vraisemblance

La méthode de maximum de vraisemblance est la méthode la plus utilisée pour l'estimation comme pour l'inférence en statistique. Nous donnons ici un bref rappel des points les plus importants concernant cette méthode. Pour plus de détails, il convient de consulter le document "Likelihood method" extrait du livre Analysis of categorical data with R [2].

[2] Bilder, Christopher R., and Thomas M. (CRC Press, 2014)

Estimateur du maximum de vraisemblance

Soit Y_i, $i = 1, \dots, n$ un échantillon iid qui provient d'une distribution discrète $p_\theta(y) = P_\theta(Y_i = y)$, $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$, $d \geqslant 1$, est un (vecteur) paramètre inconnu. La fonction de vraisemblance est

$$L_n(\theta) = \prod_{i=1}^{n} p_\theta(y_i)$$

L'estimateur de maximum de vraisemblance (EMV) est

$$\hat{\theta}_n = \arg\max_\theta L_n(\theta) = \arg\max_\theta l_n(\theta),$$

ou $l_n(\theta) = \log L_n(\theta)$ est la log-vraisemblance.

Dans un cas régulier, trouver $\hat{\theta}$ revient à résoudre l'équation suivante

$$S_n(\theta) = 0,$$

où $S_n(\theta) = \left(\frac{\partial l_n}{\partial \theta_j}(\theta) \right)_{j=1,\dots,d} \in \mathbb{R}^d$. Cette quantité est appelée le Score.

Les principaux propriétés du EMV sont

- Si $\hat{\theta}$ est le EMV de θ alors $g(\hat{\theta})$ est le EMV de $g(\theta)$.
- $\hat{\theta}_n$ est un estimateur consistent: $\hat{\theta}_n \xrightarrow{p} \theta$.
- $\hat{\theta}_n \sim_a N_d(\theta, I_n^{-1}(\theta))$,

où $I_n(\theta) = -E[H_n(\theta)] = -E\left[\frac{\partial^2 l_n}{\partial\theta_j\theta_k}(\theta)\right]_{(j,k)}$ est la matrice d'information de Fisher.

Exemple: Poisson

Soit Y_i, $i = 1, \ldots, n$, un échantillon iid d'une $\text{Pois}(\mu)$. Il est facile de vérifier que

$$l_n(\mu) = \sum_i (Y_i \ln(\mu) - \mu - \log(Y_i!)) \qquad \text{(Log-Vraisemblance)}$$

$$S_n(\mu) = \frac{\sum_i Y_i}{\mu} - n \Rightarrow \hat{\mu} = \frac{\sum_i Y_i}{n} \qquad \text{(Score et EMV)}$$

$$I_n(\mu) = -E(-\sum_i Y_i/\mu^2) = n/\mu \qquad \text{(Information Fisher)}$$

Notez que de $n\hat{\mu} \sim \text{Pois}(n\mu)$. Cette écriture représente la distribution exact de $\hat{\mu}$. La théorie (asymptotique) de vraisemblance, nous dit que $\hat{\mu} \underset{a}{\sim} N(\mu, \mu/n)$.

Exemple: Multinomiale

Soit $(N_1, N_1, N_3) \sim \text{Mul}_{(3)}(n, (p_1, p_2, p_3))$. La log-vraisemblance est donnée par

$$l_n(\mathbf{p}) = \log\left(\frac{n!}{N_1!N_2!N_3!} p_1^{N_1} p_2^{N_2} p_3^{N_3}\right) = N_1 \log(p_1) + N_2 \log(p_2) + N_3 \log(p_3) + \text{const}$$

Puisque $p_3 = 1 - p_1 - p_2$, on peut considérer que les *paramètres effectifs* de cette fonction sont $\mathbf{p} = (p_1, p_2)$.

Score: $S_n(\mathbf{p}) = (N_1/p_1 - N_3/p_3, N_2/p_2 - N_3/p_3)^t \Rightarrow$ EMV: $\hat{p}_k = \frac{N_k}{n}$, $k = 1, 2, 3$.

Information Fisher: $I_n(\mathbf{p}) = n \begin{bmatrix} 1/p_1 + 1/p_3 & 1/p_3 \\ 1/p_3 & 1/p_2 + 1/p_3 \end{bmatrix}$

Et on a que

$$\hat{\mathbf{p}} \underset{a}{\sim} \mathbf{N}_2(\mathbf{p}, I_n^{-1}(\mathbf{p}))$$

où $I_n^{-1}(\mathbf{p}) = n^{-1} \begin{bmatrix} p_1(1-p_1) & -p_1 p_2 \\ -p_1 p_2 & p_2(1-p_2) \end{bmatrix}$.

Les tests de Wald, Score et LR

La vraisemblance permet de construire des tests et des intervalles de confiances qui sont asymptotiquement valides et qui ont des propriétés optimales.

Pour tester,

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta \neq \theta_0$$

on utilisera une des statistiques suivantes:

- Statistique de Wald: $Z_n^2 = (\hat{\theta} - \theta_0)^t I_n(\hat{\theta})(\hat{\theta} - \theta_0)$
- Statistique du Score: $X_n^2 = S^t(\theta_0) I_n^{-1}(\theta_0) S(\theta_0)$
- Statistique du rapport de vraisemblance (LR): $G_n^2 = -2(l_n(\theta_0) - l_n(\hat{\theta}))$

Ces trois tests sont asymptotiquement équivalents. Ainsi, les remarques suivantes s'appliquent aux trois statistiques.

» Une grande valeur de Z_n^2 indique que les données sont plus plausibles sous l'hypothèse alternative que sous l'hypothèse nulle. Par conséquent, H_0 sera rejetée pour les grandes valeurs de Z_n^2.

» Sous certaines conditions, *si* H_0 *est vraie*, alors Z_n^2 suit asymptotiquement la loi de χ_d^2. Symboliquement, on écrira $Z_n^2 \underset{H_0,a}{\sim} \chi_d^2$.

» Soit z^2 la valeur observée de Z_n^2. Lorsque n est suffisamment grand, la p−valeur du test de Wald, qui est donnée par $P(Z_n^2 > z^2|H_0)$, peut être approximée par $P(\chi_d^2 \geqslant z^2)$. On appelle cette dernière la p−valeur asymptotique.

» À grande taille d'échantillon, on rejettera H_0 si

$$p - \text{valeur} < \alpha \text{ ou } z^2 > \chi_{d,1-\alpha}^2,$$

où $\chi_{d,\alpha}^2$ désigne le quantile d'ordre α d'une χ_d^2.

Considérant un échantillon provenant d'une distribution multinational $(n, \mathbf{p})$, où $\mathbf{p} := (p_1, \ldots, p_K)$ est inconnu.

Catégorie	1	2	...	K
Fréquence	n_1	n_2	...	n_K
Probabilité	p_1	p_2	...	p_K

Nous cherchons ici à savoir s'il existe une conformité entre les probabilités théoriques inconnues $(p_1, \ldots, p_k)$ et des probabilités spécifiques $\mathbf{p}^0 := (p_1^0, \ldots, p_k^0)$. En d'autres termes, nous voulons tester

$$H_0 : p_k = p_k^0, \forall k = 1, \ldots, K \quad \text{vs} \quad H_1 : p_k \neq p_k^0 \text{ pour au moins une valeur de } k$$

Ainsi, pour $K = 3$, nous pourrions vouloir tester

$$H_0 : p_1 = p_2 = 1/4, p_3 = 1/2 \quad \text{vs} \quad H_1 : p_1 \neq 1/4, \text{ ou } p_2 \neq 1/4, \text{ ou } p_3 \neq 1/2$$

Nous pouvons effectuer ce test de différentes manières, en utilisant les statistiques de Wlad, Score ou LR. Pour cette dernière, il est facile de vérifier que

$$G_n^2 = -2(l_n(\mathbf{p}_0) - l_n(\hat{\mathbf{p}})) = 2\sum_{k=1}^{K} N_k \log\left(\frac{\hat{p}_k}{p_k^0}\right) = 2\sum_{k=1}^{K} N_k \log\left(\frac{N_k}{np_k^0}\right)$$

Ce que nous pouvons symboliquement écrire comme

$$G_n^2 = 2\sum \mathbf{O} \log\left(\frac{\mathbf{O}}{\mathbf{E}}\right), \text{ ou}$$

- $\mathbf{O}$ = fréquences observées (Observed).
- $\mathbf{E} = E(\mathbf{O}|H_0)$ = fréquences attendues sous l'hypothèse nulle (Expected).

De la même manière, mais avec un peu plus de calculs, il est possible de montrer que les statistiques de Wald et du Score peuvent s'écrire comme suit

$$Z_n^2 = \sum \frac{(\mathbf{O}-\mathbf{E})^2}{\mathbf{O}}, \text{ et } X_n^2 = \sum \frac{(\mathbf{O}-\mathbf{E})^2}{\mathbf{E}}$$

Dans les trois cas/formules, nous comparons simplement les effectifs observés et ceux attendus sous H_0. Une "trop" grande valeur pour l'une de ces statistiques traduit un écart important entre ces effectifs, ce qui devrait conduire au rejet de H_0.

Si H_0 est vrai et que la taille de l'échantillon $n \to \infty$, alors ces trois statistiques convergent en distribution vers une χ^2_{K-1}.

À grande taille d'échantillon, on rejettera H_0 si $P(\chi^2_{K-1} \geqslant g^2) < \alpha$, où g^2 désigne la valeur observée de G^2_n. La même règle s'applique aux tests de Wald et du Score en remplaçant G^2_n par Z^2_n et X^2_n, respectivement.

Dans le contexte des données discrètes, la statistique du Score telle que formulée ci-dessus est plus connue sous le nom de la statistique Khi-deux de Pearson proposée, à l'origine, par Karl Pearson en 1900.

Exemple

$K = 3, (n_1, n_2, n_3) = (58, 59, 127), n = 244$

$$H_0 : p_1 = 0.25 \text{ et } p_2 = 0.25 \text{ vs } H_1 : p_1 \neq 0.25 \text{ ou } p_2 \neq 0.25$$

	$k = 1$	$k = 2$	$k = 3$
Observed $(O) = N_k = n\hat{p}_k$	58	59	127
Expected $(E) = np_{0k}$	61	61	122

```
O <- c(58, 59, 127)
E <- 244 * c(0.25, 0.25, 0.5)
z2 <- sum((O - E)^2/O)
x2 <- sum((O - E)^2/E)
g2 <- 2 * sum(O * log(O/E))
```

$\Rightarrow z^2 \approx x^2 \approx g^2 \approx 0.42.$

$\Rightarrow p-\text{valeur} = P(\chi^2_2 \geqslant 0.42) = 0.81$

Test khi-deux de Pearson avec R

```
chisq.test(x = c(58, 59, 127), p = c(0.25, 0.25, 0.5),
           correct = FALSE)

 Chi-squared test for given probabilities

data:  c(58, 59, 127)
X-squared = 0.418, df = 2, p-value = 0.81
```

Pour autant que $np_k^0 \geqslant 5$, $\forall k = 1, \ldots, K$, la distribution χ^2_{K-1} fournit une très bonne approximation pour les vraies distributions des statistiques G_n^2, Z_n^2 et X_n^2. Dans le cas contraire, l'approximation peut être inadéquate et `chisq.test()` est programmé pour afficher un Warning.

La statistique de Pearson (X_n^2) *converge plus rapidement* vers un χ^2_{K-1} que les deux autres. C'est pourquoi elle est plus largement utilisée dans la pratique.

Sans `correct = FALSE` , R réalise le test de Pearson en appliquant une correction dite de continuité ou correction de Yates. En pratique, il est coutume, mais pas nécessaire, d'appliquer une telle correction.

Les trois statistiques (Wald, Pearson/Score, LR) peuvent être appliquées pas uniquement pour faire un test de conformité sur les probabilités d'une multinomiale (cas traité ici), mais aussi comme test d'ajustement, test d'homogénéité ou test d'indépendance, ... qu'on découvrira progressivement.

Les formules qui définissent ces statistiques restent fondamentalement inchangées. Ce qui change c'est la loi limite sous H_0. Plus précisément, c'est toujours un χ^2 mais avec un degré de liberté qui varie d'un contexte à l'autre.

Tests par simulations (Monte Carlo testing)

Plutôt que d'utiliser la distribution asymptotique, qui suppose un échantillon de grande taille, il est possible de calculer la p−valeur à l'aide de simulations. Pour ce faire, il suffit de générer un grand nombre, disons $B = 10000$, de réalisations d'un $\mathrm{Mul}(n, \mathbf{p}^0)$. Pour chaque réalisation $\mathbf{n}_b := (n_{b1}, \ldots, n_{bK})$, $b = 1, \ldots, B$, on calcule la statistique de Pearson, que l'on note par X_b^2, et on estime ensuite la p−valeur par $B^{-1} \sum_1^B I(X_b^2 \geqslant x^2)$.

Pour notre exemple, pour effectuer ces calculs dans R, il suffit de taper

```
{rmultinom(10000, size = 244, prob = c(0.25, 0.25, 0.5)) |>
    apply(2, FUN = \(O) sum((O - E)^2 / E)) >= 0.418} |> mean()
```

ou, plus simple, d'appeler la fonction `chisq.test()`, en spécifiant l'argument `simulate.p.value = TRUE`.

```
chisq.test(x = c(58, 59, 127), p = c(0.25, 0.25, 0.5), sim = TRUE, B = 10000)

X-squared = 0.418, df = NA, p-value = 0.81
```

Test χ^2 d'ajustement

On applique souvent le test χ^2 pour vérifier si l'échantillon dont on dispose provient d'une certaine loi de probabilité théorique. L'objectif dans ce cas est de réaliser le test suivant

- H_0 : les données observées sont conformes à une certaine loi théorique.
- H_1 : les données observées ne sont pas conformes à la distribution théorique.

On parle alors d'un test d'ajustement ou tests d'adéquation (Goodness-of-Fit Test). Il s'agit en vrai d'une application directe du test χ^2 de conformité.

Cette démarche est décrite plus en profondeur dans le slide suivant.

Soit Y une variable discrète à K modalités. Tester si $Y \sim \mathcal{D}$, pour une distribution $\mathcal{D}$ donnée, est équivalent à tester

$$H_0 : p_k = p_k^0, \forall k = 1, \ldots, K,$$

où $p_k = P(Y = k)$ et $p_k^0 = P(\mathcal{D} = k)$, avec $\sum_k p_k = \sum_k p_k^0 = 1$.

Sur base d'un échantillon $Y_1, \ldots, Y_n$, la statistique de Pearson est donnée par

$$X_n^2 = \sum_{k=1}^{K} \frac{(N_k - np_k^0)^2}{np_k^0} \equiv \sum \frac{(O - E)^2}{E} \underset{H_0, a}{\sim} \chi^2_{K-1},$$

où $N_k = \sum_{i=1}^{n} I(Y_i = k)$ est la fréquence de la modalité k.

Le même résultat/formulation est aussi valable pour les deux autres statistiques (Wald, LR).

Comme illustration, reprenant des 6115 familles Saxe; voir Slide 17.

nMales	0	1	2	3	4	5	6	7	8	9	10	11	12
Freq	3	24	104	286	670	1033	1343	1112	829	478	181	45	7

On se donne comme objectif de tester

$$H_0 : nMales \sim \text{Bin}(12, 0.5) \text{ vs } H_1 : nMales \not\sim \text{Bin}(12, 0.5)$$

$$H_0 \Leftrightarrow p_k = p_k^0, \quad \text{pour } k = 0, \ldots, 12,$$

avec $p_k = P(nMales = k)$ et $p_k^0 = P(\text{Bin}(12, 0.5) = k)$.

```
Freq <- c(3, 24, 104, 286, 670, 1033, 1343, 1112, 829, 478, 181, 45, 7)
p0 <- dbinom(0:12, 12, 0.5)
qsres <- chisq.test(Freq, p = p0, cor = FALSE) |>
  print()
```

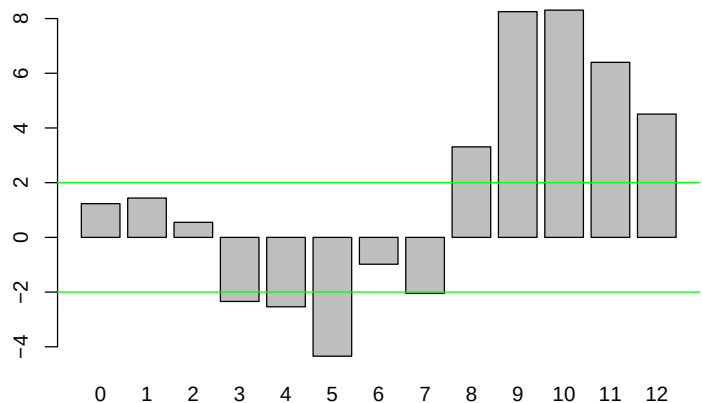
```
Warning in chisq.test(Freq, p = p0, cor = FALSE): Chi-squared approximation may be incorrect
```

```
X-squared = 249, df = 12, p-value <2e-16
```

$\rightarrow$ la distribution des données ne semble pas se conformer à une $\text{Bin}(12, 0.5)$.

Le graphique suivant montre les résidus standardisés $r_k := (O_k - E_k)/\sqrt{E_k}, k = 0, \ldots, 12$.
$\rightarrow$ Plus r_k est grande, plus la contribution de la k—ième case à la valeur du khi-deux est importante.

```
barplot(qsres$residuals, names = 0:12)
abline(h = c(-2, 2), col = "green")
```



Tester avec des fréquences attendues estimées

Lors de l'application de test χ^2 de conformité, il arrive bien souvent que les fréquences attendues $\mathbf{E} = \mathrm{E}(\mathbf{O}|\mathrm{H}_0)$ soient (en partie) inconnues et doivent en l'occurrence être estimées à partir des données.

Par exemple, avec les données sur les enfants de Saxon, on peut être intéressé à tester l'hypothèse que les fréquences observées proviennent de la famille Binomiale avec $p \in (0, 1)$. Plus précisément, on souhaite tester

$$\begin{aligned} \mathrm{H}_0 &: nMales \sim \mathrm{Bin}(12, p), \text{ pour un certain } p \in (0,1) \text{ vs} \\ \mathrm{H}_1 &: \text{ il n'y a pas de } p \text{ tel que } nMales \sim \mathrm{Bin}(12, p). \end{aligned}$$

Puisque p est inconnu, nous ne pouvons pas appliquer le test de Pearson (ni aucun autre test vu précédemment), pour la simple raison que nous ne pouvons pas calculer les fréquences attendues sous H_0.

Pour couvrir de telles situations, les tests classiques (Wald, Score/Pearson, LR) ont été généralisés comme expliqué ci-après.

Pour la statistique LR, la théorie générale peut être synthétisée comme suit.
Si $\mathbf{N} \sim \mathrm{Mul}_{(K)}(n, \mathbf{p}^0(\theta))$, ou $\theta \in \mathbb{R}^l$, $l \leqslant K-2$. Et si $\hat{\theta}$ est l'EMV de θ, alors

$$G_n^2 = 2\sum_{k=1}^{K} N_k \log \frac{N_k}{np_k^0(\hat{\theta})} \equiv 2\sum O \log \frac{O}{\hat{E}} \sim \chi^2_{K-1-l}.$$

Le degré de liberté (dl) qui figure dans cette formule peut être exprimé, de façon générale, comme dl = le nombre de paramètres à estimer dans le modèle sans tenir compte de l'hypothèse nulle (ici $K-1$) moins le nombre de paramètres à estimer sous H_0 (ici l)

De même, pour les statistiques de Wald et de Pearson/Score, nous avons

$$Z_n^2 = \sum \frac{\left(O-\hat{E}\right)^2}{O} \sim \chi^2_{dl}, \text{ et } X_n^2 = \sum \frac{\left(O-\hat{E}\right)^2}{\hat{E}} \sim \chi^2_{dl}.$$

Ces résultats s'appliquent à un large éventail de situations, comme l'illustrent les exemples suivants et comme nous le verrons plus loin.

Exemple 1

Reprenant de nouveau l'exemple des 6115 familles Saxe; voir Slide 17. Voici comme tester $H_0 : nMales \sim \text{Bin}(12, p)$, pour un certain $p \in (0, 1)$, dans R.

```
that <- 0.51922
p0hat <- dbinom(0:12, 12, that)
Ehat <- 6115 * p0hat

# Test LR
{2 * sum(Freq * log(Freq / Ehat))} |> print() |> pchisq(df = 11, low = FALSE)

[1] 97.007
[1] 6.9782e-16

# Test Pearson
sum((Freq - Ehat)^2/Ehat) |> print() |> pchisq(df = 11, low = FALSE)

[1] 110.5
[1] 1.4531e-18
```

Exemple 2

Reprenant l'exemple des naissances à l'hôpital; voir Slide 20.

nBaby	0	1	2	3	4	5+
Freq	3	8	6	4	3	0

$$H_0 : \text{nBaby} \sim \text{Pois} \quad \text{vs} \quad H_1 : \text{nBaby} \not\sim \text{Pois}$$

```
Freq <- c(3, 8, 6, 4, 3, 0)
that <- 1.833
p0hat <- dpois(0:4, lambda = that) |> c(ppois(4, lambda = that, low = FALSE))
Ehat <- 24 * p0hat

# Test Pearson
sum((Freq - Ehat)^2/Ehat) |> print() |> pchisq(df = 4, low = FALSE)

[1] 2.0699
[1] 0.72291
```

Exemple 3

Soit $(N_1, N_2, N_3) \sim Mul_{(3)}(n, (p_1, p_2, p_3))$. Nous souhaitons tester

$$H_0 : p_1 = \pi \quad \text{vs} \quad H_1 : p_1 \neq \pi,$$

où π est une probabilité donnée (par exemple 0.25 !). Sous H_0, nous avons que

$$(p_1, p_2, p_3) = (\pi, p_2, 1 - \pi - p_2)$$

et que la log-vraisemblance (voir Slide 27) est donnée par

$$l_n = N_1 \log(\pi) + N_2 \log(p_2) + N_3 \log(1 - \pi - p_2) + \text{ const}$$

Nous en déduisons que, sous H_0, l'EMV de p_2 est $\hat{p}_2^0 = (1 - \pi)\dfrac{N_2}{N_2 + N_3}$.

Pour réaliser le test, on peut utilise la statistique $G_n^2 = 2 \sum \mathbf{O} \log\left(\dfrac{\mathbf{O}}{\hat{\mathbf{E}}}\right) \underset{a, H_0}{\sim} \chi_1^2$,

avec $\mathbf{O} = (N_1, N_2, N_3)$ et $\hat{\mathbf{E}} = n \times \hat{\mathbf{p}}_0 = \left(n\pi, n\hat{p}_2^0, n(1 - \pi - \hat{p}_2^0)\right)$.

Exemple 4

n veaux sont répartis en 3 catégories : (A) "pas de pneumonie", (B) "pneumonie, sans infection secondaire" [3] ou (C) "pneumonie suivie d'une infection secondaire". Les données récoltées $\mathbf{N} = (N_1, N_2, N_3)$ sont considérées comme un échantillon multinomial avec les probabilités $\mathbf{p} = (p_1, p_2, p_3)$. Nous souhaitons tester si la probabilité, pour un veau, de contracter une pneumonie est égale à la probabilité conditionnelle de contracter une infection à la suite d'une pneumonie :

$$H_0 : \; p_2 + p_3 = \frac{p_3}{p_2 + p_3}.$$

Soit $\pi = p_2 + p_3$. Sous H_0, il est facile de vérifier que $(p_1, p_2, p_3) = \left(1 - \pi, \pi(1-\pi), \pi^2\right)$. et que l'EMV de π est $\hat{\pi} = \dfrac{N_2 + 2N_3}{2n - N_1}$.

Pour effectuer le test, on peut utilise la statistique G_n^2, exactement comme dans l'exemple précédent, sauf qu'ici

$$\hat{E} = n \times \left(1 - \hat{\pi}, \hat{\pi}(1-\hat{\pi}), \hat{\pi}^2\right).$$

[3] Infection secondaire = infection qui se rajoute, en raison de l'état de faiblesse du malade, à une première infection.

Tableaux de contingence à deux variables

Définitions et outils descriptifs

On considère une certaine population dans laquelle on s'intéresse à deux variables:

» X à I valeurs possibles $(1, 2, \ldots, I)$ qui correspondent à I catégories différentes.

» Y à J valeurs possibles $(1, 2, \ldots, J)$ qui correspondent à J catégories différentes.

On possède un échantillon de n individus:

$$(X_l, Y_l), l = 1, \ldots, n.$$

À partir de cet échantillon on calcule

$$\begin{aligned} N_{ij} = & \text{ Nombre d'individus pour lesquelles } X = i \text{ et } Y = j \text{ (simultanément)} \\ & = \sum_{l=1}^{n} I(X_l = i, Y_l = j) \end{aligned}$$

En analysant les N_{ij}, on aimerait comprendre et étudier l'association (lien), éventuelle, entre X et Y.

Les (x_l, y_l), $l = 1, \ldots, n$, constituent les données brutes alors que les (i, j, n_{ij}), $(i, j) = (1, 1), \ldots, (I, J)$, représentent les données groupées. Celles-ci peuvent être représentées, de manière équivalente, sous forme d'un dataframe ou sous forme d'un tableau à double entrée (tableau croisé):

X	Y	N
1	1	n_{11}
⋮	⋮	⋮
1	J	n_{1J}
⋮	⋮	⋮
I	1	n_{I1}
⋮	⋮	⋮
I	J	n_{IJ}

Format "classique" (data-frame)

X\Y	1	2	...	J
1	n_{11}	n_{12}	...	n_{1J}
2	n_{21}	n_{22}	...	n_{2J}
⋮	⋮	⋮	⋮	⋮
I	n_{I1}	n_{I2}	...	n_{IJ}

Format "tableau croisé" (contingency table)

Le vecteur des $\{n_{ij}\}_{i,j}$ peut être vue comme une réalisation du vecteur aléatoire $\mathbf{N} = (N_{11}, \ldots, N_{IJ})$.

Probabilités d'intérêt

Les probabilités en lien avec un tableau de contingence à deux variables sont les suivantes, pour $i = 1, \ldots, I$ et $j = 1, \ldots, J$:

Probabilités conjointes : $P(X, Y)$	p_{ij}	$P(X = i, Y = j)$
Probabilités marginales : $P(X)$ et $P(Y)$	$p_{i.}$	$P(X = i) = \sum_j p_{ij}$
	$p_{.j}$	$P(Y = j) = \sum_i p_{ij}$
Probabilités conditionnelles : $P(X\|Y)$ et $P(Y\|X)$	$p_{i\|j}$	$P(X = i\|Y = j) = p_{ij}/p_{.j}$
	$p_{j\|i}$	$P(Y = j\|X = i) = p_{ij}/p_{i.}$

Exemple : Tabagisme

Les données utilisées pour cet exemple se trouvent dans la dataframe `Whickham` du package `mosaicData`, qui donne une petite partie d'une enquête menée à Whickham au Royaume-Uni au début des années 1970. On a demandé aux participantes (uniquement des femmes) leurs âges et si elles fumaient. Un suivi vingt ans plus tard a révélé si la participante était encore en vie.

Whickham contiennent 1314 observations et les variables suivantes :

- outcome : statut de survie après 20 ans: un facteur avec deux niveaux (Alive, Dead)
- smoker : statut de fumeur au départ: un facteur avec deux niveaux (No, Yes)
- age: âge (en années) au départ

Voici un aperçu des données brutes.

```
Whickham <- mosaicData::Whickham
head(Whickham)
```

```
  outcome smoker age
1   Alive    Yes  23
2   Alive    Yes  18
3    Dead    Yes  71
4   Alive     No  67
5   Alive     No  64
6   Alive    Yes  38
```

```
str(Whickham)
```

```
'data.frame': 1314 obs. of  3 variables:
 $ outcome: Factor w/ 2 levels "Alive","Dead": 1 1 2 1 1 1 1 2 1 1
 $ smoker : Factor w/ 2 levels "No","Yes": 2 2 2 1 1 2 2 1 1 1 ...
 $ age    : int  23 18 71 67 64 38 45 76 28 27 ...
```

```
# Réordonner les niveaux de smoker (facultatif)
Whickham <- Whickham |> transform(smoker = factor(smoker, levels = c("Yes", "No")))
```

Les slides suivantes montrent quelques fonctions R utiles pour manipuler et afficher des données de ce type.

```r
# base R function table(), xtabs(), and addmargins()
```

```r
Whickham["smoker"] |> table()
```

```
smoker
Yes  No
582 732
```

```r
#or# xtabs(~ smoker, data = Whickham)
```

```r
tb <- xtabs(~ smoker + outcome,
  data = Whickham) |> print()
```

```
      outcome
smoker Alive Dead
   Yes   443  139
   No    502  230
```

```r
#or# Whickham[c("smoker","outcome")] |> table()
```

```r
gdf <- as.data.frame(tb) |> print()
```

```
  smoker outcome Freq
1    Yes   Alive  443
2     No   Alive  502
3    Yes    Dead  139
4     No    Dead  230
```

```r
xtabs(Freq ~ smoker + outcome, data = gdf) |>
  addmargins()
```

```
      outcome
smoker Alive Dead  Sum
   Yes   443  139  582
   No    502  230  732
   Sum   945  369 1314
```

```r
# base R function proportions()
```

```r
# simple proportions
xtabs(~ smoker, data = Whickham) |>
  proportions()
```

```
smoker
    Yes      No
0.44292 0.55708
```

```r
# joint and marginal proportions
proportions(tb) |> addmargins()
```

```
      outcome
smoker   Alive    Dead     Sum
   Yes 0.33714 0.10578 0.44292
   No  0.38204 0.17504 0.55708
   Sum 0.71918 0.28082 1.00000
```

```r
# conditional proportions given "smoker"
proportions(tb, "smoker") |> addmargins(2)
```

```
      outcome
smoker   Alive    Dead     Sum
   Yes 0.76117 0.23883 1.00000
   No  0.68579 0.31421 1.00000
```

```r
# conditional proportions given "outcome"
proportions(tb, "outcome") |> addmargins(1)
```

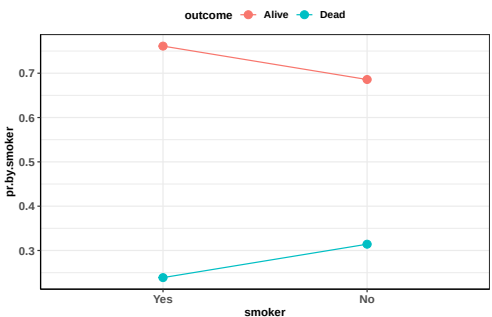
```
      outcome
smoker   Alive    Dead
   Yes 0.46878 0.37669
   No  0.53122 0.62331
   Sum 1.00000 1.00000
```

```
# counts, proportions and plots using using tidyverse packages
```

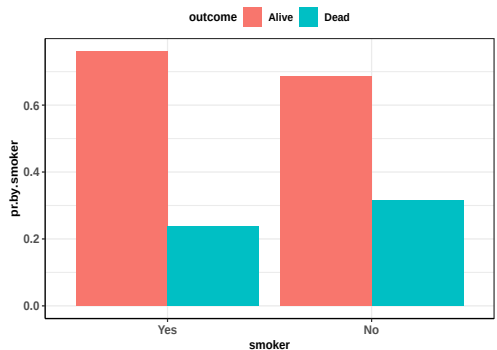
```
gdfp <- Whickham |> count(smoker, outcome) |> mutate(pr = proportions(n)) |>
  mutate(pr.by.smoker = proportions(n), .by = smoker) |> print()

gdfp |> ggplot(aes(x = smoker, y = pr.by.smoker, color = outcome)) + geom_point(size = 3) +
  geom_line(aes(group = outcome))
```

```
  smoker outcome   n      pr pr.by.smoker
1    Yes   Alive 443 0.33714      0.76117
2    Yes    Dead 139 0.10578      0.23883
3     No   Alive 502 0.38204      0.68579
4     No    Dead 230 0.17504      0.31421
```



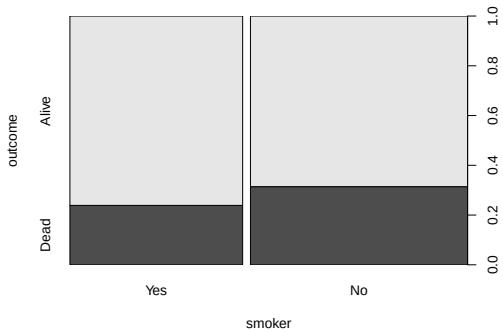

```
gdfp |> ggplot(aes(x = smoker,
  y = pr.by.smoker, fill = outcome)) + geom_col(position = "dodge")
```



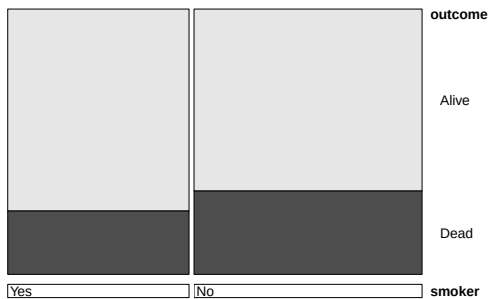
Mosaic plot

```
# base R function spinplot() and the vcd's package function doubledecker()
```

```
spineplot(tb)
#or# plot(outcome ~ smoker, data = tb, weights = Freq)
```




```
vcd::doubledecker(tb)
#or# doubledecker(outcome ~ smoker, data = tb)
```



Cette analyse semble indiquer un effet "protecteur" du tabagisme: seulement 69% des non-fumeuses ont survécu au bout de 20 ans contre 76% des fumeuses!

Modélisation de tables de contingence

Avant d'être analysées, les données d'un tableau de contingence doivent être récoltées. Pour ce faire, plusieurs schémas ou modes d'échantillonnage sont possibles, dont les plus connus sont les suivants

- Échantillonnage Multinomial Simple
- Échantillonnage Poisson
- Échantillonnage Multinomial Multiple

Échantillonnage multinomial simple

On dispose d'un seul échantillon de taille n fixée avant la collecte. Après la collecte, les individus sont classés dans les $I \times J$ catégories formées par le croisement de deux variables catégorielles X et Y.

Exemple: On interroge 1000 individus, prises au hasard, et on note leurs sexes ($X: 1 \equiv F, 2 \equiv M$) et leurs orientations politiques ($Y: 1 \equiv$ gauche, $2 \equiv$ centre, $3 \equiv$ droite).

Soit $N_{ij} = \sum_{l=1}^{n} I(X_l = i, Y_l = j)$. Dans un tel cas, il est raisonnable de supposer que

$$\mathbf{N} = (N_{11}, \ldots, N_{IJ}) \sim Mul(n, \mathbf{p}),$$

avec $\mathbf{p} = (p_{11}, \ldots, p_{IJ})$, $p_{ij} = P(X = i, Y = j)$. La vraisemblance de ce modèle est

$$L(\mathbf{p}) = \frac{n!}{\prod_{i,j} n_{ij}!} \prod_{i,j} p_{ij}^{n_{ij}} \Rightarrow \hat{p}_{ij} = N_{ij}/n$$

$$\Rightarrow \hat{p}_{i.} = N_{i.}/n, \ \hat{p}_{.j} = N_{.j}/n$$

$$\Rightarrow \hat{p}_{j|i} = N_{ij}/N_{i.}, \ \hat{p}_{i|j} = N_{ij}/N_{.j}$$

On échantillonne les individus dans un intervalle de temps et/ou d'espace bien déterminé, sans fixer préalablement la taille de l'échantillon.

Exemple: Pendant une journée, on interroge les individus qu'on croise à la sortie d'une gare. *La taille d'échantillon est ici aléatoire* et n'est connue qu'après la collecte.

Dans un tel cas, il est raisonnable de supposer que les N_{ij} sont des variables indépendantes et que

$$N_{ij} \sim \text{Pois}(\mu_{ij}),\ \forall i, j.$$

Soit $\boldsymbol{\mu} = (\mu_{11}, \ldots, \mu_{IJ})$. La vraisemblance de ce modèle est

$$L(\boldsymbol{\mu}) = \prod_{ij} \frac{\mu_{ij}^{n_{ij}}}{n_{ij}!} \exp(-\mu_{ij}) \Rightarrow \hat{\mu}_{ij} = N_{ij}$$

$$\Rightarrow \hat{p}_{ij} = \frac{\hat{\mu}_{ij}}{\sum_{ij} \hat{\mu}_{ij}} = \frac{N_{ij}}{n}, \text{ avec } n = \sum_{ij} N_{ij}.$$

Voir App.

Malgré que les formules sont différentes, on obtient les mêmes estimations que dans le cas d'un échantillonnage multinomial.

Avant la collecte, la population est stratifiée selon les modalités de l'une des deux variables, disons X. Dans chacune des sous-populations ($X = 1, \ldots, X = I$), on récolte un nombre $n_{i\cdot}$ prédéfini d'observations.

Contrairement aux deux autres modes d'échantillonnages (multinomial multiple et poisson), où l'on dispose de deux variables aléatoires, cette manière de récolter les données implique que la variable de stratification, ici X, ne peut être considérée comme aléatoire.

Exemple: On interroge 500 femmes et 500 hommes pour identifier leurs orientations politiques. Dans une telle situation, chaque ligne (ou chaque colonne) du tableau de contingence représente un échantillon de la sous-population définie par la variable de stratification (ici, sous-population des femmes et sous-population des hommes).

Dans un tel cas, il est raisonnable de supposer que pour chaque sous-population $i = 1, \ldots, I$,

$$\mathbf{N}_i = (N_{i1}, \ldots, N_{iJ}) \sim \text{Mul}(n_{i.}, \mathbf{p}_i),$$

avec $\mathbf{p}_i = (p_{1|i}, \ldots, p_{J|i})$, où $p_{j|i} = P(Y = j|X = i)$.

En supposant, en plus, que les $\mathbf{N}_i$ sont indépendantes, la vraisemblance de ce modèle est

$$L = \prod_i \frac{n_{i.}!}{\prod_j n_{ij}!} \prod_j p_{j|i}^{n_{ij}} \Rightarrow \hat{p}_{j|i} = N_{ij}/n_{i.}$$

Encore une fois, on obtient les mêmes estimations, pour $\hat{p}_{j|i}$, MAIS, dans ce cas, à cause de la stratification, les probabilités $p_{i|j}, p_{ij}$ et $p_{i.}$ ne sont pas estimables.

Tableaux de contingence à deux variables

Test d'indépendance et test d'homogénéité

Test d'indépendance

L'objectif est de voir s'il y un lien entre deux variables catégorielles X et Y sans préciser une causalité (X et Y jouent un rôle symétrique). Cette formulation est souvent utilisée lorsque un échantillonnage simple (multinomiale ou poisson) a été effectué.

Exemple: Est-ce qu'il existe un lien entre la couleur des yeux et la couleur des cheveux ?

En termes statistiques, on souhaite tester si $P(X, Y) = P(X)P(Y)$, càd tester

$$\begin{aligned} H_0 &: \ P(X = i, Y = j) = P(X = i)P(Y = j) \ \forall (i, j) \text{ vs} \\ H_1 &: \ \exists (i, j) \ P(X = i, Y = j) \neq P(X = i)P(Y = j). \end{aligned}$$

Test d'homogénéité

Nous désirons *comparer plusieurs groupes* (définies par les modalités d'une variable catégorielle X) et voir si elles sont homogènes au regard d'un facteur donné (Y). Càd voir si la distribution de Y est *la même* dans tous les groupes.

Exemple: Les hommes et les femmes votent-ils de la même façon ? Si nous résumons le vote par le facteur Y : $1 \equiv$ gauche, $2 \equiv$ centre, $3 \equiv$ droite, alors cette question peut être traduite en H_0 : $P(Y = j|Homme) = P(Y = j|Femme)$, $\forall j$.

Ici X (ci-dessus, *Sexe*) et Y (ci-dessus, *Vote*) jouent un rôle asymétrique: X est la variable qui est censée avoir une influence sur Y et non l'inverse. On peut parler de variable explicative (X) et variable à expliquer (Y). Cette formulation est souvent utilisée lorsque un échantillonnage multinomiale multiple a été effectué.

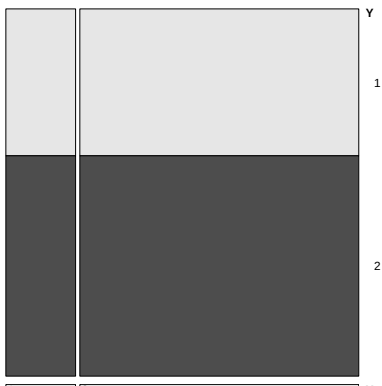
En termes statistiques, on souhaite tester si $X \mapsto P(Y|X)$ reste constante, ou, de façon équivalente, tester

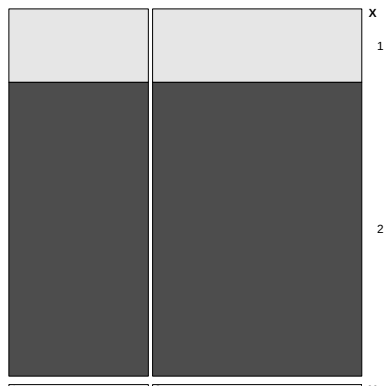
$$H_0' : \ P(Y = j|X = 1) = \cdots = P(Y = j|X = I) \ \forall j.$$

Ce que l'on peut aussi écrire comme suit H_0' : $P(Y = j|X = i) = P(Y = j) \ \forall i, j$.

Il est facile de vérifier que, de point de vue mathématique, H_0 et H_0' sont équivalentes. Ces hypothèses expriment au fond la même chose, mais la formulation et le vocabulaire utilisés diffèrent.

Graphiquement H_0, comme H_0', se traduit par un bar/mosaic plot plat:





Nous pouvons tester ces hypothèses (H_0 ou H_0') en utilisant l'une des statistiques de test définies précédemment: Pearson/Score, LR ou Wald.

Statistique de Pearson	Statistique de LR
$X_n^2 = \sum_{i,j=1}^{I,J} \frac{(N_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$	$G_n^2 = 2 \sum_{i,j=1}^{I,J} N_{ij} \log \frac{N_{ij}}{\hat{\mu}_{ij}}$

avec $\hat{\mu}_{ij}$ = les fréquences espérées sous H_0, càd $E(N_{ij}|H_0)$

Rappelons que, sous H_0, ces statistiques suivent asymptotiquement une loi du khi-deux dont le degrés de liberté est la différence entre le nombre de paramètres à estimer dans le modèle, sans tenir compte de l'hypothèse nulle, et le nombre de paramètres à estimer sous H_0.

Pour le test d'indépendance :

$$\mu_{ij} = np_{ij}$$

$$\underset{H_0}{=} np_{i.}p_{.j} \Rightarrow \hat{\mu}_{ij} = n\hat{p}_{i.}\hat{p}_{.j} = \frac{N_{i.}N_{.j}}{n}$$

$$dl = (IJ - 1) - ((I - 1) + (J - 1))$$
$$= (I - 1)(J - 1)$$

Pour le test d'homogénéité :

$$\mu_{ij} = n_{i.}p_{j|i}$$

$$\underset{H_0'}{=} n_{i.}p_{.j} \Rightarrow \hat{\mu}_{ij} = n_{i.}\hat{p}_{.j} = \frac{n_{i.}N_{.j}}{n}$$

$$dl = (J - 1)I - (J - 1)$$
$$= (I - 1)(J - 1)$$

On constate que les deux procédures aboutissent aux mêmes formules. En conséquence, en pratique, on peut effectuer (calculer) l'un ou l'autre test sans faire attention à l'énonciation utilisée.

MAIS il faut comme même veiller à avoir une cohérence entre, d'une part, la nature des données et du problème posé et, d'autre part, la formulation des hypothèses et les interprétations/conclusions qui s'ensuivent.

Test d'indépendance/homogénéité dans R

Test de Pearson

```
chtb <- chisq.test(tb, correct = FALSE) |>
  print()

 Pearson's Chi-squared test

data:  tb
X-squared = 9.12, df = 1, p-value = 0.0025

#or# summary(tb)
```

Test de LR

```
(2 * sum(tb * log(tb/chtb$expected))) |>
  print() |> pchisq(1, low = FALSE)

[1] 9.2003
[1] 0.0024198
```

Test de Pearson et de LR

```
vcd::assocstats(tb)

                    X^2 df  P(> X^2)
Likelihood Ratio 9.2003  1 0.0024198
Pearson          9.1209  1 0.0025271
```

La cote et le rapport des cotes

Considérant le tableau suivant qui donne la distribution conditionnelle de $Y|X$.

	Y		
$X =$ Groupe	$1 =$ Succès	$2 =$ Échec	Sum
$1 = G1$	p_1	$1 - p_1$	1
$2 = G2$	p_2	$1 - p_2$	1

où $p_i = p_{1|i} = P(Y = 1|X = i) = P(S|Gi)$.

Pour comparer p_1 et p_2, on peut considérer la différence $p_1 - p_2$ ou le rapport p_1/p_2, parfois appelé *risque relatif*. En général, ce dernier est plus approprié/informatif, comme illustré dans le tableau suivant

p_1	p_2	$p_1 - p_2$	p_1/p_2
0.01	0.001	0.009	10
0.41	0.401	0.009	1.02

Une autre façon de comparer deux probabilités, qui est aussi un instrument bien pratique pour mesurer l'association entre deux variables dichotomiques, est le rapport des cotes, que nous examinerons ci-après.

La cote (odds en anglais) d'un évènement S (pensez à "Succès", par exemple) de probabilité $p = P(S)$ est le rapport

$$o(S) = \frac{p}{1-p}.$$

On peut passer d'une cote à une probabilité (et vice versa) :

$$p = \frac{o}{o+1}$$

Le tableau suivant nous donne quelques exemples de passage de l'une à l'autre quantité :

p	o	p	o
0	0	0.5	1
0.001	0.001001	0.67	2
0.01	0.0101	0.9	9
0.05	0.053	0.95	19
0.1	0.11	0.99	99
0.2	0.25	0.999	999
0.33	0.5	1	∞

Tout comme une probabilité, la cote d'un évènement est une mesure de sa vraisemblance. Autrement dit, la cote d'un évènement est (une façon de mesurer) la chance de le voir se réaliser.

Si un évènement S a été observée un nombre s de fois sur un total de t réalisations, alors:

» la probabilité $P(S)$ exprime le ratio du nombre de résultats favorables (à S) sur le nombre de résultats total. $\rightarrow s/t$.

$P(S) = 0.5 \rightarrow$ sur un total de 100 expériences, on s'attend à observer 50 fois S.

» la cote $o(S)$ exprime le ratio du nombre de résultats favorables sur le nombre de résultats défavorables. $\rightarrow s/(t-s) = s/e$, où e est le nombre d'échecs, càd nombre de fois où l'on n'a pas observé S.

$o(S) = 0.1 \rightarrow$ on s'attend à observer 10 fois moins de succès que d'échecs
$o(S) = 1 \rightarrow$ on s'attend à constater autant de succès que d'échecs.
$o(S) = 10 \rightarrow$ on s'attend à constater 10 fois plus de succès que d'échecs.

Exemple

Les résultats d'une étude sur l'efficacité d'un traitement contre l'insuffisance cardiaque sont résumés dans le tableau croisé (tableau 2×2) suivant

	Y = Insuffisance		
X = Groupe	1 = Non	2 = Oui	Total
1 = Traité	10898	139	11037
2 = Placebo	10795	239	11034

Soit $p_1 = P(Y = 1|X = 1)$ et $o_1 = o(Y = 1|X = 1)$ la probabilité et la cote de *succès (= pas d'insuffisance cardiaque)* dans le groupe traité. De même, nous définissons $p_2 = P(Y = 1|X = 2)$ et $o_2 = o(Y = 1|X = 2)$ pour le groupe placebo.

Le tableau suivant donne, pour chaque groupe, l'estimation de ces quantités.

Group	$\hat{p}$	$\hat{o}$
(1) Traité	$10898/11037 = 0.987$	$10898/139 = 78.4$
(2) Placebo	$10795/11034 = 0.978$	$10795/239 = 45.2$

$\rightarrow$ dans le groupe traité, nous estimons qu'il y a 78 fois plus de succès que d'échecs, contre seulement 45 dans le groupe placebo.

La quantité

$$\text{or} = \frac{o_1}{o_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)},$$

qu'on appelle le rapport des cotes, est une mesure qui permet simplement de comparer la cote (d'un événement) dans deux groupes/populations (ici traité et placebo).

Dans notre exemple,

$$\widehat{\text{or}} = 78.4/45.2 \approx 1.7$$

$\rightarrow$ on estime qu'un individu du groupe traité à 1.7 fois plus de *chance (= cote)* de succès q'un individu du groupe placebo. Autrement dit, un individu du groupe traité à 70% plus de chance de succès qu'un individu du groupe placebo.

Quelques caractéristiques de l'odds ratio

- Connaître or ne permet pas de savoir p_1/p_2 et encore moins de calculer p_1 et p_2.
- $\text{or} = 1 \Leftrightarrow p_1 = p_2 \Leftrightarrow X \perp\!\!\!\perp Y$.
- Plus or s'éloigne de 1, plus la différence entre p_1 et p_2 est importante, révélant un lien fort entre X et Y.
- En termes de X et Y, nous pouvons exprimer or comme suit

$$\begin{aligned}\text{or} &= \frac{P(Y=1|X=1)P(Y=2|X=1)}{P(Y=1|X=2)P(Y=2|X=2)} = \frac{P(X=1,Y=1)P(X=2,Y=2)}{P(X=1,Y=2)P(X=2,Y=1)} \\ &= \frac{P(X=1|Y=1)P(X=2|Y=1)}{P(X=1|Y=2)P(X=2|Y=2)}.\end{aligned}$$

X|Y	succès	échec	
Groupe 1	s_1	e_1	n_1
Groupe 2	s_2	e_2	n_2

Puisque l'EMV de o_1 est s_1/e_1, et que celui de o_2 est s_2/e_2, l'EMV de or est

$$\hat{or} = \frac{s_1 e_2}{e_1 s_2}$$

Pour une taille de l'échantillon grande, on peut montrer que

$$\ln(\hat{or}) \sim_a N\left(\ln(or), \sigma^2\right), \text{ avec}$$

$$\sigma^2 = \frac{1}{n_1 p_1 (1-p_1)} + \frac{1}{n_2 p_2 (1-p_2)}.$$

Un estimateur consistant de cette variance est donné par

$$\hat{\sigma}^2 = 1/s_1 + 1/e_1 + 1/s_2 + 1/e_2$$

Ce résultat nous permet d'effectuer des tests et de construire un intervalle de confiance pour $\ln(\text{or})$. Par exemple, un intervalle à 95% est donné par

$$\ln(\hat{\text{or}}) \pm 1.96\hat{\sigma}$$

En utilisant la fonction exponentielle, on obtient un intervalle de confiance pour or:

$$[\exp(\ln(\hat{\text{or}}) - 1.96\hat{\sigma}), \exp(\ln(\hat{\text{or}}) - 1.96\hat{\sigma})]$$

Dans notre exemple

$$\text{ic}_{95\%}(\text{or}) = [1.41, 2.14].$$

Au risque de 5%, on peut donc conclure que $\text{or} \neq 1$. Et nous pouvons affirmer que le traitement réduit de manière significative le risque d'insuffisance cardiaque.

```
dt <- read.table(text = "
      X   Y     N
Placebo Oui   239
Placebo Non 10795
Traite  Oui   139
Traite  Non 10898
", header = TRUE)
```

```
dt <- dt |> transform(X = factor(X, levels = c("Traite", "Placebo")),
                      Y = factor(Y, levels = c("Non", "Oui")))

xtabs(N ~ X + Y, data = dt) |> vcd::loddsratio(log = FALSE) |>
  confint()

                       2.5 % 97.5 %
Traite:Placebo/Non:Oui 1.406 2.1431
```

Tableaux de contingence à trois variables

Notations et structure

Un tableau de contingence à trois variables (ou à trois niveaux) est une classification croisée d'observations par les niveaux de trois variables catégorielles (facteurs). Plus généralement, les tableaux de contingence à S niveaux classent les observations par les niveaux de S variables catégorielles.

Soit X, Y et Z nos trois variables d'intérêts: X à I valeurs possibles $(1, 2, \ldots, I)$, qui correspondent à I catégories différentes; Y à J valeurs possibles $(1, 2, \ldots, J)$; et Z à K valeurs possibles $(1, 2, \ldots, K)$.

Nous disposons d'un échantillon de n individus: $(X_l, Y_l, Z_l), l = 1, \ldots, n$.

À partir de cet échantillon on calcule

$$\begin{aligned} N_{ijk} = & \text{ Nombre d'individus pour lesquelles } X = i,\ Y = j \text{ et } Z = k \text{ (simultanément)} \\ & = \sum_{l=1}^{n} I(X_l = i, Y_l = j, Z_l = k) \end{aligned}$$

En analysant les N_{ijk}, on aimerait comprendre et étudier l'association (lien), éventuelle, entre X, Y et Z.

Voici à quoi ressemble un tableau à trois variables/niveaux:

	$Z=1$				$Z=2$				
$X \backslash Y$	1	2	...	J	1	2	...	J	...
1	n_{111}	n_{121}	...	n_{1J1}	n_{112}	n_{122}	...	n_{1J2}	...
2	n_{211}	n_{221}	...	n_{2J1}	n_{212}	n_{222}	...	n_{2J2}	...
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	...
I	n_{I11}	n_{I21}	...	n_{IJ1}	n_{I12}	n_{I22}	...	n_{IJ2}	...

Dans cet exemple, X occupe le niveau 1 (ligne), Y occupe le niveau 2 (colonne) et Z occupe le niveau 3. Dans un tel cas, on parle d'un tableau $I \times J \times K$.

On peut aussi représenter ces données sous forme d'un dataframe classique de colonnes $(X, Y, Z, Freq)$, où Freq est le vecteur des fréquences observées n_{ijk}. Les lignes de ce dataframe sont données par (i, j, k, n_{ijk}).

Dans un tableau $I \times J \times K$, il est possible de considérer différents types de probabilités/distributions, notamment,

» la distribution conjoint $P(X, Y, Z)$:

$$p_{ijk} = P(X = i, Y = j, Z = k)$$

» la distribution conditionnel $P(X, Y|Z)$:

$$p_{ij|k} = P(X = i, Y = j|Z = k)$$

» la distribution marginal $P(X, Y)$:

$$p_{ij.} = P(X = i, Y = j) = \sum_{k=1}^{K} p_{ijk}$$

Nous pouvons également définir différents types de fréquences marginales, comme par exemple

$$N_{ij.} = \sum_{k=1}^{K} N_{ijk} \quad \text{et} \quad N_{..k} = \sum_{i=1}^{I} \sum_{j=1}^{J} N_{ijk},$$

où le "$\cdot$" représente la somme sur l'un des indices.

Tableaux de contingence à trois variables

Analyse conditionnelle versus analyse marginale

Lorsque l'on cherche à étudier la relation entre deux variables X et Y alors que l'on possède aussi des observations pour d'autres variables, on fait la distinction entre l'association entre X et Y conditionnelle à la valeur des autres variables, et l'association marginale entre X et Y, soit l'association ne tenant pas compte des autres variables.

Pour, par exemple, étudier la distribution conditionnelle $(X, Y)|Z$, il faut analyser les tableaux dits partiels (ou conditionnels) où l'on croise X et Y pour chacune des modalités de Z. Voici un exemple d'un tableau partiel où Z est fixée à k (une valeur donnée)

	$Z = k$			
$X \backslash Y$	1	2	...	J
1	n_{11k}	n_{12k}	...	n_{1Jk}
2	n_{21k}	n_{22k}	...	n_{2Jk}
⋮	⋮	⋮	⋮	⋮
I	n_{I1k}	n_{I2k}	...	n_{IJk}

Nous pouvons analyser ce tableau $I \times J$ en utilisant les techniques que nous avons appris dans le chapitre précédent. En se faisant, nous limitons nos objectifs à l'étude de l'assocaition entre X et Y sachant que $Z = k$.

Tableau marginal

Dans le cas où l'on souhaite, par exemple, étudier l'association entre X et Y, sans tenir compte de Z, alors il faut analyser le tableau marginal suivant, que l'on obtient en sommant sur Z.

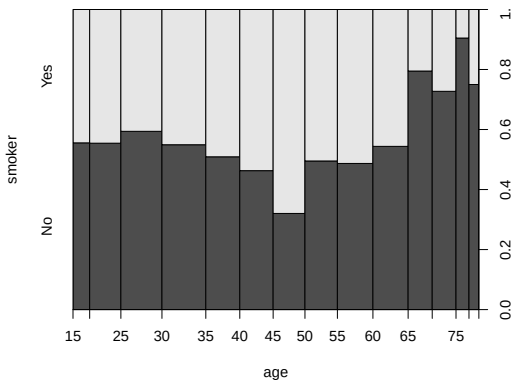
X\Y	1	2	...	J
1	$n_{11.}$	$n_{12.}$	...	$n_{1J.}$
2	$n_{21.}$	$n_{22.}$	...	$n_{2J.}$
⋮	⋮	⋮	⋮	⋮
I	$n_{I1.}$	$n_{I2.}$	...	$n_{IJ.}$

À la différence d'un tableau partiel dans lequel la variable Z est contrôlée, i.e. sa valeur est maintenue fixe, ce tableau ignore complètement Z.

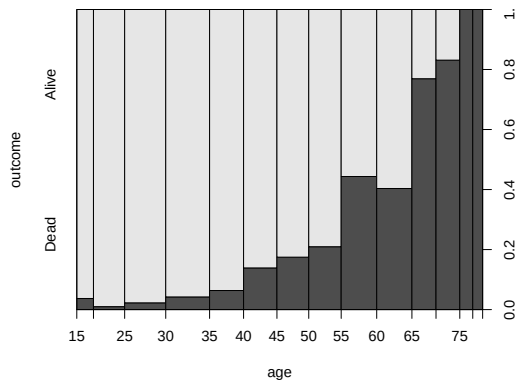
L'analyse des tableaux partiels, d'une part, et marginal, d'autre part, peut donner lieu à des conclusions diamétralement opposées. Nous allons illustrer cela avec les données Tabagisme (voir Slide 49).

Pour rappel, lorsque on a analysé le tableau (marginal) smoker × outcome, nous avons noté "un effet bénéfique du tabac". À présent nous allons reprendre l'analyse en intégrant l'âge. Voici, pour commencer, deux graphiques qui aident à percevoir le lien entre l'âge d'une part, et les variables outcome et smoker d'autre part.

```
plot(smoker ~ age, data = Whickham)
```




```
plot(outcome ~ age, data = Whickham)
```



Remarque

Parfois, nous avons besoin de catégoriser une variable numérique. Pour effectuer cette opération, il convient d'utiliser la fonction `cut()`.

```
cut(0:10, breaks = c(-Inf, 1, 4, 6, Inf), right = FALSE)
```

```
 [1] [-Inf,1) [1,4)    [1,4)    [1,4)    [4,6)    [4,6)    [6, Inf) [6, Inf)
 [9] [6, Inf) [6, Inf) [6, Inf)
Levels: [-Inf,1) [1,4) [4,6) [6, Inf)
```

La nouvelle variable obtenue est un facteur avec les niveaux `[-Inf,1)`, `[1,4)`, `[4,6)`, `[6, Inf)`. Par défaut les intervalles sont fermés à droite et ouverts à gauche; `right = FALSE` sert à préciser le contraire. Nous pouvons contrôler les libellés des niveaux du facteur crée par `cut()` avec l'argument `labels`.

```
cut(0:10, breaks = c(-Inf, 1, 4, 6, Inf), labels = c("moins que 1", "entre 1 et 4",
  "entre 4 et 6", "plus que 6"), right = FALSE)
```

```
 [1] moins que 1  entre 1 et 4 entre 1 et 4 entre 1 et 4 entre 4 et 6
 [6] entre 4 et 6 plus que 6   plus que 6   plus que 6   plus que 6
[11] plus que 6
Levels: moins que 1 entre 1 et 4 entre 4 et 6 plus que 6
```

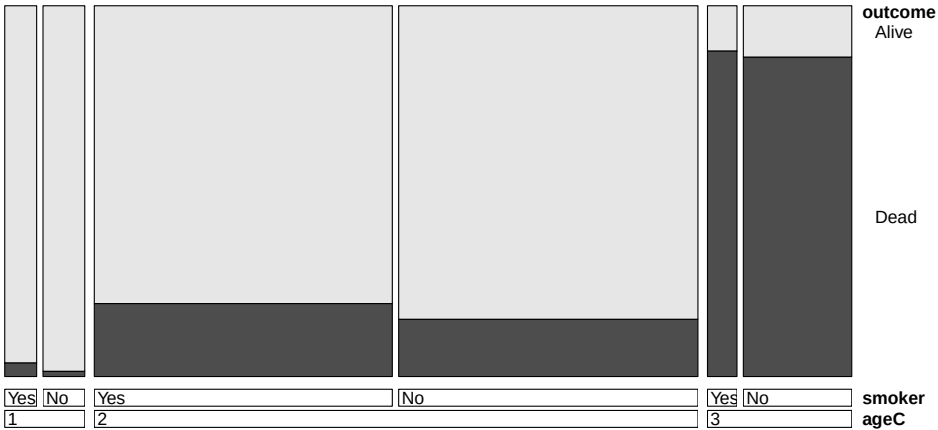
Nous pouvons intégrer l'âge dans notre analyse en le considérant comme une variable continue, ou le transformer préalablement en une variable catégorielle. C'est ce que nous ferons ici en créant la variable `ageC` : "1" $= age \leqslant 24$, "2" $= 24 < age \leqslant 65$, "3" $= age > 65$.

```
Whickham <- Whickham |> transform(ageC = cut(age, breaks = c(-Inf, 24, 65, Inf),
                                             labels = c("1", "2", "3")))
tb <- xtabs(~ smoker + outcome + ageC, data = Whickham) |> print()
ptb <- tb |> proportions(c("smoker", "ageC")) |> print()
```

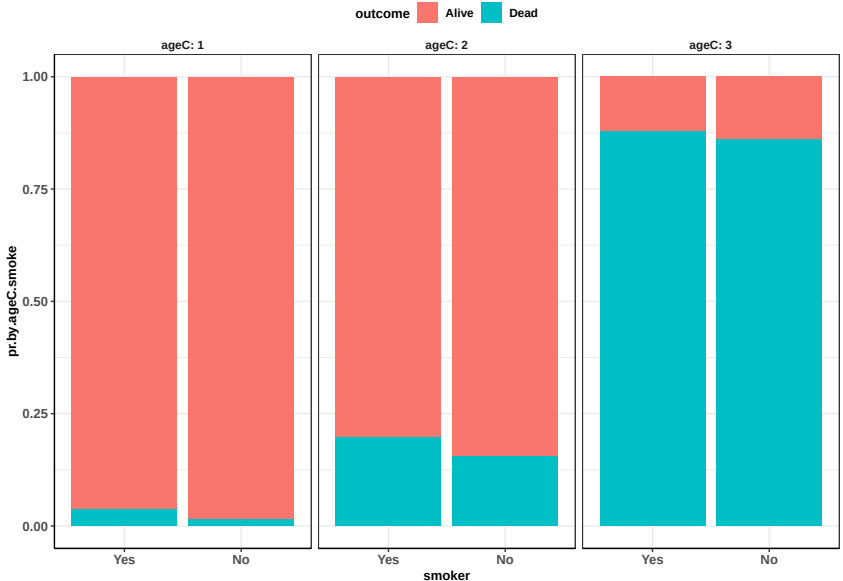
ageC	smoker | outcome	Alive	Dead
1	Yes	53	2
	No	71	1
2	Yes	384	94
	No	406	74
3	Yes	6	43
	No	25	155

ageC	smoker | outcome	Alive	Dead
1	Yes	0.964	0.036
	No	0.986	0.014
2	Yes	0.803	0.197
	No	0.846	0.154
3	Yes	0.122	0.878
	No	0.139	0.861

```
vcd::doubledecker(outcome ~ ageC + smoker, data = tb)
```



```
ptb |> as.data.frame() |> ggplot(aes(x = smoker, y = Freq, fill = outcome)) +
  geom_col() + facet_wrap(~ageC, labeller = label_both) + labs(y = "pr.by.ageC.smoke")
```



Contrairement à ce qu'on a constaté auparavant, on peut voir ici que, dans chaque tranche d'âge, la survie chez les fumeuses est inférieure à celle des non-fumeuses. Comment expliquer alors que la tendance s'inverse lorsque l'on combine tous les groupes d'âge ?

Pour comprendre ce qui se passe, il faut examiner la relation entre l'âge et les autres deux variables (outcome et smoker). Une lecture attentive du "mosaicplot" ci-dessus révèle que, contrairement aux autres catégories d'âges, peu de femmes âgées (> 65 ans au début de l'étude) fument, mais beaucoup d'entre elles décèdent, vraisemblablement de causes "naturelles", au bout de 20 ans. Ce qui a induit l'effet "protecteur" du tabagisme observé lorsque l'âge a été ignoré.

Cet exemple illustre ce qui est connu sous le nom du paradoxe de Simpson. Ce dernier est rencontré lorsqu'une analyse de tableaux partiels montre une certaine tendance, mais cette tendance s'inverse lorsque la même analyse est réalisée sur le tableau marginal (que l'on obtient en sommant sur les tableaux partiels).

Rapports de cotes conditionnels

Les rapports de cotes conditionnels sont des rapports de cotes entre deux variables pour des niveaux fixes d'une troisième variable. Ces rapports nous permettent d'étudier l'association conditionnelle de deux variables, étant donné la troisième.

Le tableau suivant illustre cela dans le cas de notre exemple sur le tabagisme.

$Z =$ ageC	$X \backslash Y$	$1 =$ Alive	$2 =$ Dead	$\hat{p}(Alive\|X,Z)$	$\hat{o}(Alive\|X,Z)$	
1	$1 =$ Yes	53	2	0.964	26.5	$0.37 = \hat{or}^{XY\|Z=1}$
	$2 =$ No	71	1	0.986	71	
2	Yes	384	94	0.803	4.085	$0.74 = \hat{or}^{XY\|Z=2}$
	NO	406	74	0.846	5.486	
3	Yes	6	43	0.122	0.1395	$0.87 = \hat{or}^{XY\|Z=3}$
	NO	25	155	0.139	0.1612	
Tab. marginal	Yes	443	139	0.761	3.187	$1.46 = \hat{or}^{XY}$
	No	502	230	0.686	2.183	

```
vcd::loddsratio(tb, log = FALSE)

 odds ratios for smoker and outcome by ageC

      1       2       3
0.37324 0.74458 0.86512
```

$\widehat{or}^{XY|Z=k}$ est un estimateur du rapport de cotes conditionnel

$$or^{XY|Z=k} = \frac{p_{1k}/(1-p_{1k})}{p_{2k}/(1-p_{2k})},$$

où $p_{ik} = P(Y = 1|X = i, Z = k)$, $i = 1, 2$. Alors que $\widehat{or}^{XY}$ est un estimateur du rapport de cotes marginal

$$or^{XY} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)},$$

avec $p_i = P(Y = 1|X = i)$, $i = 1, 2$.

Dans notre exemple, les rapports conditionnels sont tous inférieurs à 1 (dans les trois tranches d'âges, les proportions de survie pour X = Yes sont plus petites que pour X = No); alors que le rapport marginal est supérieur à 1 (tout âge confondu, la proportion de survie pour X = Yes est plus grande que pour X = No). C'est une autre manifestation du paradoxe de Simpson: *la direction de l'association entre X et Y est inversée lorsque l'on tient compte de Z.*

Tableaux de contingence à trois variables

Types d'associations

Contrairement au cas d'un tableau de contingence classique (à double entrée) où seule l'indépendance entre la variable ligne et la variable colonne est d'intérêt, dans un tableau à trois niveaux, il existe une multitude d'associations/modèles possibles. Les voici du plus au moins complexe.

» Modèle saturé, dénoté par (XYZ): absence de toute forme d'indépendance → toutes les variables sont liées les unes aux autres.

» Association homogène, dénoté par (XY, XZ, YZ): les rapports de cotes conditionnels ne dépendent pas de la valeur de la troisième variable.

» Indépendance conditionnelle: 2 variables sont indépendantes étant donné la 3e → $(XY, XZ) : Y \perp\!\!\!\perp Z|X$, $(XY, YZ) : X \perp\!\!\!\perp Z|Y$, et $(XZ, YZ) : X \perp\!\!\!\perp Y|Z$.

» Indépendance partielle: 2 variables sont conjointement indépendantes de la 3e → $(XY, Z) : Z \perp\!\!\!\perp (X, Y)$, $(YZ, X) : X \perp\!\!\!\perp (Y, Z)$, et $(XZ, Y) : Y \perp\!\!\!\perp (X, Z)$.

» Indépendance mutuelle ou totale, dénoté par (X, Y, Z): absence de toute forme de dépendance → toutes les variables sont indépendantes les unes des autres.

Avant d'examiner plus en détail ces différents types d'associations, notez que

1. Le concept d'indépendance conditionnelle est très important et constitue la base de nombreux autres associations.

2. Comme nous allons le voir plus tard, chaque type d'association correspond à un modèle log-linéaire particulier. Les notations (en bleu) utilisées pour dénoter ces différents types d'associations font référence aux termes d'interactions qui figurent dans ces modèles. Par exemple, l'écriture (XYZ) signifie la présence d'une interaction triple entre les variables X, Y, Z; et l'écriture (XY, XZ, YZ) signifie la présence de toutes les interactions doubles.

3. Ces différents types d'associations (modèles) sont reliés entre elles par une structure hiérarchique qui fait qu'on peut passer du modèle le plus complexe (saturé) au modèle le plus simple (indépendance mutuelle) en imposant des conditions/hypothèses de plus en plus strictes.

Il y a une association homogène entre X et Y conditionnellement à Z lorsque $or^{XY|Z=k}$ ne change pas en fonction de k, i.e.,

$$or^{XY|Z=1} = or^{XY|Z=2} = \ldots = or^{XY|Z=K}$$

Une association homogène est une propriété symétrique : association homogène X-Y|Z $\Leftrightarrow$ association homogène X-Z|Y $\Leftrightarrow$ association homogène Y-Z|X. Par conséquence, on parle d'une association homogène entre X et Y et Z.

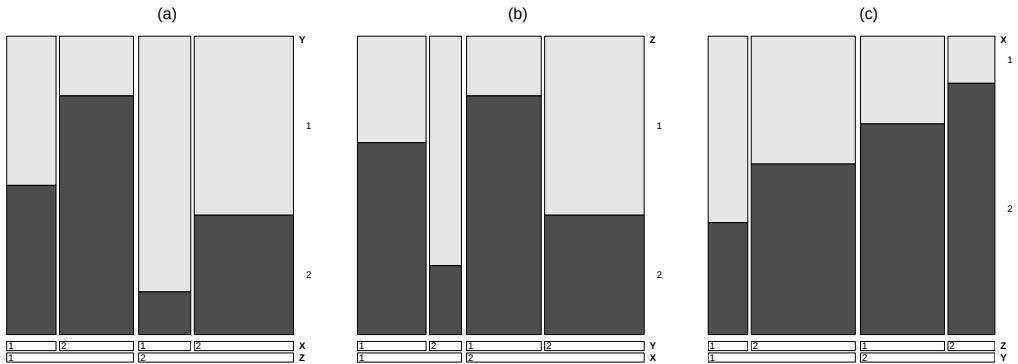
Lorsqu'on a une association homogène alors il n'est pas nécessaire de présenter/détailler les résultats de toutes les analyses conditionnelles (ce qui peut être fastidieux) puisque, dans un tel cas, ces analyses aboutiront aux mêmes constatations/conclusions.

Exemple fictif

Z	X	Y = Outcome				
Clinic	Treatment	1 = Success	2 = Failure	$\hat{p}$(Succ.)	$\hat{o}$	$\hat{or}^{XY}$
1	1 = A	10	10	0.5	1	4
	2 = B	6	24	0.2	0.25	
2	A	18	3	0.857	6	4
	B	24	16	0.6	1.5	
	A	28	13	0.683	2.15	2.87
	B	30	40	0.428	0.75	

```
dt <- data.frame(Z = rep(c(1, 2), c(4, 4)), X = rep(c(rep(1, 2), rep(2, 2)), 2),
  Y = rep(c(1, 2), 4), Freq = c(10, 10, 6, 24, 18, 3, 24, 16))
tb <- xtabs(Freq ~ X + Y + Z, data = dt)
```

```r
doubledecker(Y ~ Z + X, tb, main="(a)")
doubledecker(Z ~ X + Y, tb, main="(b)")
doubledecker(X ~ Y + Z, tb, main="(c)")
```



Il n'est pas facile de déceler graphiquement une association homogène. Néanmoins, lorsqu'on compare les blocs "Z=1" et "Z=2" dans (a), on constate une certaine similitude dans leurs structures/tendances. La même remarque s'applique aux deux autres graphiques.

Indépendance conditionnelle

X et Y sont indépendantes étant donné que $Z = k$, $X \perp\!\!\!\perp Y|Z = k$, si

$$P(X, Y|Z = k) = P(X|Z = k)P(Y|Z = k)$$

Si cette dernière égalité est vraie quelque soit $k = 1, \ldots, K$, alors on dit que X et Y sont indépendantes étant donné Z, $X \perp\!\!\!\perp Y|Z$. Cela est équivalent à dire que

$$or^{XY|Z} = 1, \forall Z.$$

Interprétation. X et Y peuvent sembler liés si Z n'est pas prise en compte, mais ce lien apparent disparaitrait si Z est contrôlée (prise en compte).

Exemples

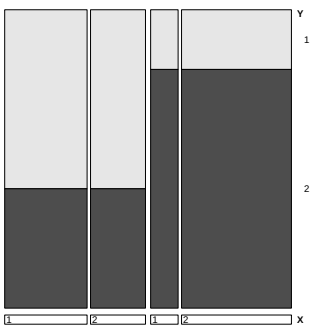
Souffrir d'une pathologie $\perp\!\!\!\perp$ Âge | L'état d'un organe
Excès de vitesse $\perp\!\!\!\perp$ Niveau socioéconomique | Puissance du moteur

Quelques propriétés

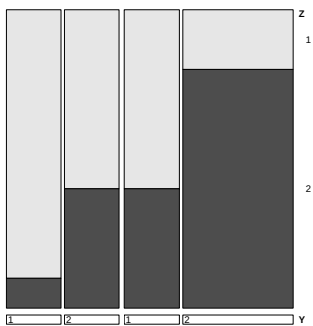
» Ind. conditionnelle $\Rightarrow$ Ass. homogène
» Ass. homogène et $or^{XY|Z=1} = 1 \Rightarrow X \perp\!\!\!\perp Y|Z$
» $X \perp\!\!\!\perp Y|Z \not\Rightarrow X \perp\!\!\!\perp Y$

Z	X	Y = Outcome				
Clinic	Treatment	1 =Success	2 =Failure	$\hat{p}$(Succ.)	$\hat{o}$	$\hat{or}^{XY}$
1	1 =A	18	12	0.6	1.5	1
	2 =B	12	8	0.6	1.5	
2	A	2	8	0.2	0.25	1
	B	8	32	0.2	0.25	
	A	20	20	0.5	1	2
	B	20	40	0.33	0.5	

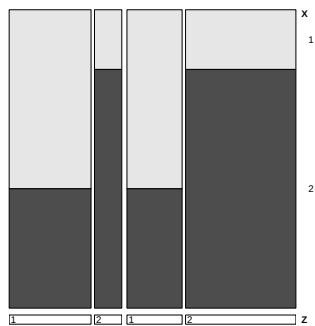
Y ind X | Z



Z not−ind Y | X



X not−ind Z | Y



Indépendance partielle

X et Z sont conjointement indépendantes de Y, $(X, Z) \perp\!\!\!\perp Y$, si

$$P(X, Y, Z) = P(X, Z)P(Y)$$

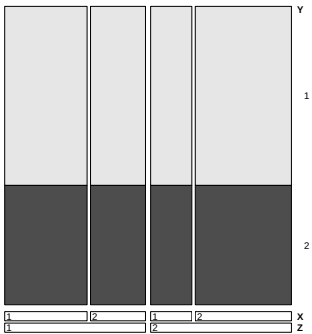
Interprétation. On peut complètement dissocier Y de (X, Z) et effectuer des analyses marginales (Y et X-Z) sans risque de tirer des conclusions erronées.

Quelques propriétés

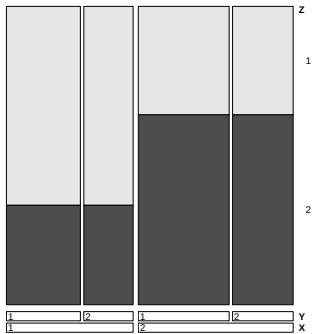
$$\begin{aligned}
(X, Z) \perp\!\!\!\perp Y &\Leftrightarrow Y \perp\!\!\!\perp X|Z \text{ et } Y \perp\!\!\!\perp Z|X \\
&\Leftrightarrow Y \perp\!\!\!\perp X|Z \text{ et } Y \perp\!\!\!\perp Z \quad \text{▸ Voir App} \\
&\Leftrightarrow Y \perp\!\!\!\perp Z|X \text{ et } Y \perp\!\!\!\perp X \\
&\Rightarrow Y \perp\!\!\!\perp Z \text{ et } Y \perp\!\!\!\perp X
\end{aligned}$$

Z	X	Y = Outcome				
Clinic	Treatment	1 =Success	2 =Failure	$\hat{p}$(Succ.)	$\hat{o}$	$\hat{or}^{XY}$
1	1 =A	18	12	0.6	1.5	1
	2 =B	12	8	0.6	1.5	
2	A	9	6	0.6	1.5	1
	B	21	14	0.6	1.5	
	A	27	18	0.6	1.5	1
	B	33	22	0.6	1.5	

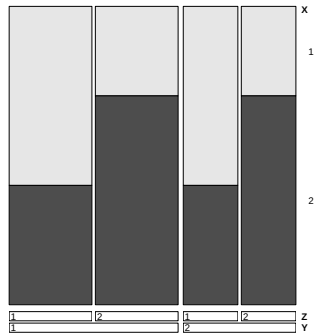
Y ind X | Z



Z ind Y | X



X not−ind Z | Y



Au moins un des mosaicplots est plat.

Indépendance mutuelle ou totale

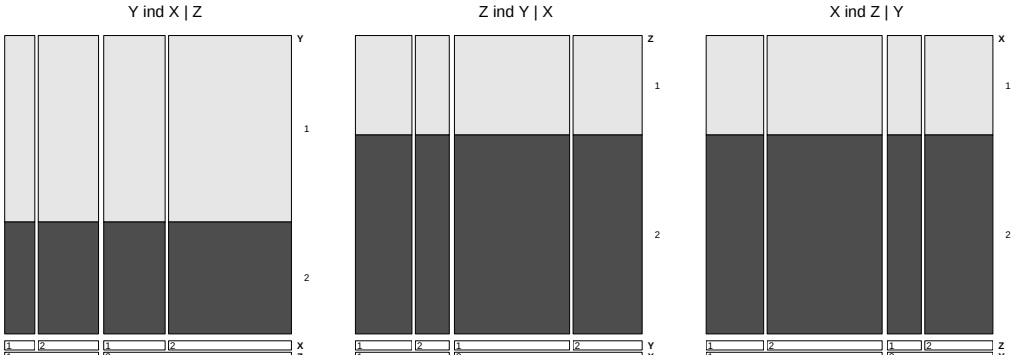
X, Y et Z sont mutuellement indépendantes, $X \perp\!\!\!\perp Y \perp\!\!\!\perp Z$, si

$$P(X, Y, Z) = P(X)P(Y)P(Z)$$

Interprétation. Les trois variables n'ont aucun lien entre elles et peuvent donc être dissociées et analysées chacune séparément.

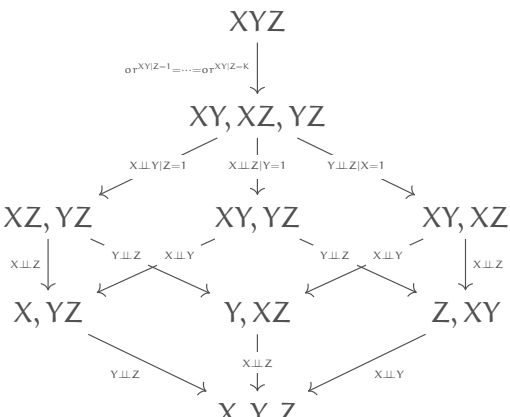
Quelques propriétés

$$\begin{aligned} X \perp\!\!\!\perp Y \perp\!\!\!\perp Z &\Leftrightarrow (X, Z) \perp\!\!\!\perp Y, (Y, Z) \perp\!\!\!\perp X \text{ et } (X, Y) \perp\!\!\!\perp Z \\ &\Leftrightarrow (X, Z) \perp\!\!\!\perp Y \text{ et } X \perp\!\!\!\perp Z \\ &\Leftrightarrow X \perp\!\!\!\perp Y|Z, Z \perp\!\!\!\perp Y|X \text{ et } X \perp\!\!\!\perp Z|Y \\ &\Rightarrow X \perp\!\!\!\perp Y, X \perp\!\!\!\perp Z \text{ et } Y \perp\!\!\!\perp Z \end{aligned}$$



Tous les mosaicplots sont plats.

Comme signalé auparavant, ces différents modèles (formes d'associations) sont liés par une structure hiérarchique. Cette structure est schématisée dans la figure suivante (du plus complexe au plus simple).



Saturé

Asso. homogène

Ind. conditionnelle

Ind. partielle

Ind. mutuelle

Le plus souvent en pratique, une telle structure est utilisée pour choisir une forme d'association qui soit le plus simple possible et qui décrit correctement les données.

Pour cela, le principe est le suivant: on commence par le modèle saturé pour ensuite "descendre" progressivement dans la hiérarchie jusqu'en "bas". À chaque étape on compare le dernier modèle non rejeté au modèle simple qui le suit immédiatement dans la hiérarchie des modèles possibles. Si ce dernier est rejeté alors on stoppe la progression et on adopte le dernier modèle non rejeté. La comparaison s'effectue, en général, à l' aide de l'une des statistiques usuelles (Wald, Pearson ou LR).

Nous en apprendrons davantage sur ce sujet dans le cadre des régressions Log-linéaire et Logit que nous traiterons dans les prochains chapitres.

Ces modèles facilitent beaucoup cette tâche et ils donnent un sens plus clair aux associations. Un autre avantage de ces modèles c'est qu'ils (1) peuvent être utilisés pour modéliser des tableaux de contingences non seulement à deux ou trois variables, mais de toutes tailles; et (2) qu'ils permettent d'incorporer dans l'analyse à la fois des variables discrètes et continues.

De la Poisson à la Multinomiale (Slide 13)

Soit $N_k \sim \text{Pois}(\mu_k)$, $k = 1, 2, 3$ trois va indépendantes et soit $\mu = \mu_1 + \mu_2 + \mu_3$.

$$
\begin{aligned}
&P(N_1 = n_1, N_2 = n_2, N_3 = n_3 | N_1 + N_2 + N_3 = n) \\
&= \frac{P(N_1 = n_1, N_2 = n_2, N_3 = n_3, N_1 + N_2 + N_3 = n)}{P(N_1 + N_2 + N_3 = n)} \\
&= \frac{P(N_1 = n_1, N_2 = n_2, N_3 = n_3)}{P(N_1 + N_2 + N_3 = n)} I(n = n_1 + n_2 + n_3) \\
&= \frac{(\mu_1^{n_1}/n_1!)e^{-\mu_1}(\mu_2^{n_2}/n_2!)e^{-\mu_2}(\mu_3^{n_3}/n_3!)e^{-\mu_3}}{(\mu^n/n!)e^{-\mu}} I(n = n_1 + n_2 + n_3) \\
&= \frac{n!}{n_1! n_2! n_3!} \frac{\mu_1^{n_1}\mu_2^{n_2}\mu_3^{n_3}}{\mu^n} I(n = n_1 + n_2 + n_3) \\
&= \frac{n!}{n_1! n_2! n_3!} (\mu_1/\mu)^{n_1}(\mu_2/\mu)^{n_2}(\mu_3/\mu)^{n_3} I(n = n_1 + n_2 + n_3).
\end{aligned}
$$

$$\Rightarrow (N_1 = n_1, N_2 = n_2, N_3 = n_3) | (N_1 + N_2 + N_3 = n) \sim \text{Mul}(n, \mu_1/\mu, \mu_2/\mu, \mu_3/\mu).$$

On a que $N_{ij} = \sum_{l=1}^{n} I(X_l = i, Y_l = j)$ et $n = \sum_{i,j} N_{ij}$. Ces deux quantités sont des va. Par définition,

$$\begin{aligned}
\mu_{ij} &= E\left[N_{ij}\right] = E\left[\sum_{l=1}^{n} I(X_l = i, Y_l = j)\right] \\
&= E\left[E\left(\sum_{l=1}^{n} I(X_l = i, Y_l = j)\middle| n\right)\right] \\
&= E\left[\sum_{l=1}^{n} p_{ij}\right] = E\left[np_{ij}\right] \\
&= p_{ij}E(n) = p_{ij}E(\sum_{i,j} N_{ij}) = p_{ij}\sum_{i,j}\mu_{ij}.
\end{aligned}$$

$(X, Z) \perp\!\!\!\perp Y \Leftrightarrow Y \perp\!\!\!\perp X|Z$ ET $Y \perp\!\!\!\perp Z$ (SLIDE 96)

Notez que

$$
\begin{aligned}
(X, Z) \perp\!\!\!\perp Y &\Leftrightarrow p(x, y, z) = p(x, z)p(y) && (1)\\
Y \perp\!\!\!\perp X|Z &\Leftrightarrow p(x, y|z) = p(x|z)p(y|z) && (2)\\
Y \perp\!\!\!\perp Z &\Leftrightarrow p(y, z) = p(y)p(z) && (3)
\end{aligned}
$$

- (1) $\Rightarrow$ (2) et (3):

$$(1) \Rightarrow \sum_x p(x, y, z) = \sum_x p(x, z)p(y) \Rightarrow p(y, z) = p(y)p(z) : (3), \text{ et}$$

$$(1) \Rightarrow p(x, y, z)/p(z) = p(x, z)p(y)/p(z) \Rightarrow p(x, y|z) = p(x|z)p(y) \underset{(3)}{=} p(x|z)p(y|z) : (2)$$

- (2) et (3) $\Rightarrow$ (1) :

$$(2) \Rightarrow p(x, y, z) = p(x, z)p(y|z) \underset{(3)}{=} p(x, z)p(y) : (1)$$