Cours 1 : Variables qualitatives (discrètes)

Cours 1 : Variables qualitatives (discrètes)
Plan du cours

Cours 1 : Variables qualitatives (discrètes)
Plan du cours

Données du Titanic

- 2201 voyageurs à bord du Titanic
- Variable `Class` : classe dans laquelle un passager voyageait
- 4 classes :

$$\{1^{re}, 2^{e}, 3^{e}, \text{équipage}\} = \{\text{'1st'}, \text{'2nd'}, \text{'3rd'}, \text{'Crew'}\}$$

Données :

Passager	Class
1	3rd
2	Crew
3	3rd
4	3rd
5	Crew
6	Crew
⋮	⋮

Quelques définitions et notations

- $n =$ nombre d'observations ($n = 2021$ passagers)
- $i =$ indice de l'observation : $i \in \{1, \dots n\}$ (le passager)
- $x =$ variable d'intérêt (la classe)
- $x_i =$ valeur de la variable x pour l'observation i (la classe du i-ème passager)
- $\mathcal{X} =$ ensemble des modalités (ici, les classes) :

$$x_i \in \mathcal{X}.$$

- $k =$ nombre de modalités ($k = 4$) :

$$|\mathcal{X}| = k.$$

Variable indicatrice

Définition 1 (Variable indicatrice)

On note u_{ia} variable indicatrice de l'appartenance de l'individu i à la modalité a (pour $1 \leq i \leq n$ et $1 \leq a \leq k$) :

$$u_{ia} = \begin{cases} 1 & \text{si } x_i = a, \\ 0 & \text{sinon.} \end{cases}$$

Par construction, on a pour tout $1 \leq i \leq n$,

$$\sum_{a=1}^{k} u_{ia} = 1.$$

Variable indicatrice

Données du Titanic : variable `Class`

i	Class	1st	2nd	3rd	Crew
1	3rd	0	0	1	0
2	Crew	0	0	0	1
3	3rd	0	0	1	0
4	3rd	0	0	1	0
5	Crew	0	0	0	1
6	Crew	0	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮

Notation vectorielle

On peut associer à la modalité a de la variable le vecteur $\mathbf{u}_a$ de dimension n :

$$\mathbf{u}_a = \begin{bmatrix} u_{1a} \\ u_{2a} \\ \vdots \\ u_{ia} \\ \vdots \\ u_{na} \end{bmatrix}.$$

On note $\mathbf{u}_a^\intercal$ sa transposée :

$$\mathbf{u}_a^\intercal = \begin{bmatrix} u_{1a} & u_{2a} & \cdots & u_{ia} & \cdots & u_{na} \end{bmatrix}.$$

Effectifs

Définition 2 (Effectif)

Pour $1 \leq a \leq k$, on note n_a l'effectif de la modalité a :

$$n_a = |\{i : x_i = a\}| = \sum_{i=1}^{n} u_{ia}.$$

Proposition 1

On a :

$$\sum_{a=1}^{k} n_a = n$$

Démonstration de la proposition 1 :

$$\sum_{a=1}^{k} n_a = \sum_{a=1}^{k} \left(\sum_{i=1}^{n} u_{ia} \right) = \sum_{i=1}^{n} \underbrace{\left(\sum_{a=1}^{k} u_{ia} \right)}_{=1} = \sum_{i=1}^{n} 1 = n.$$

□

Effectifs

Données du Titanic : variable `Class`

1st	2nd	3rd	Crew
325	285	706	885

Fréquences

Définition 3 (Fréquence)

Pour $1 \leq a \leq k$, on note f_a la fréquence de la modalité a :

$$f_a = n_a/n.$$

Proposition 2

On a :

$$\sum_{a=1}^{k} f_a = 1$$

Fréquences

Fréquences

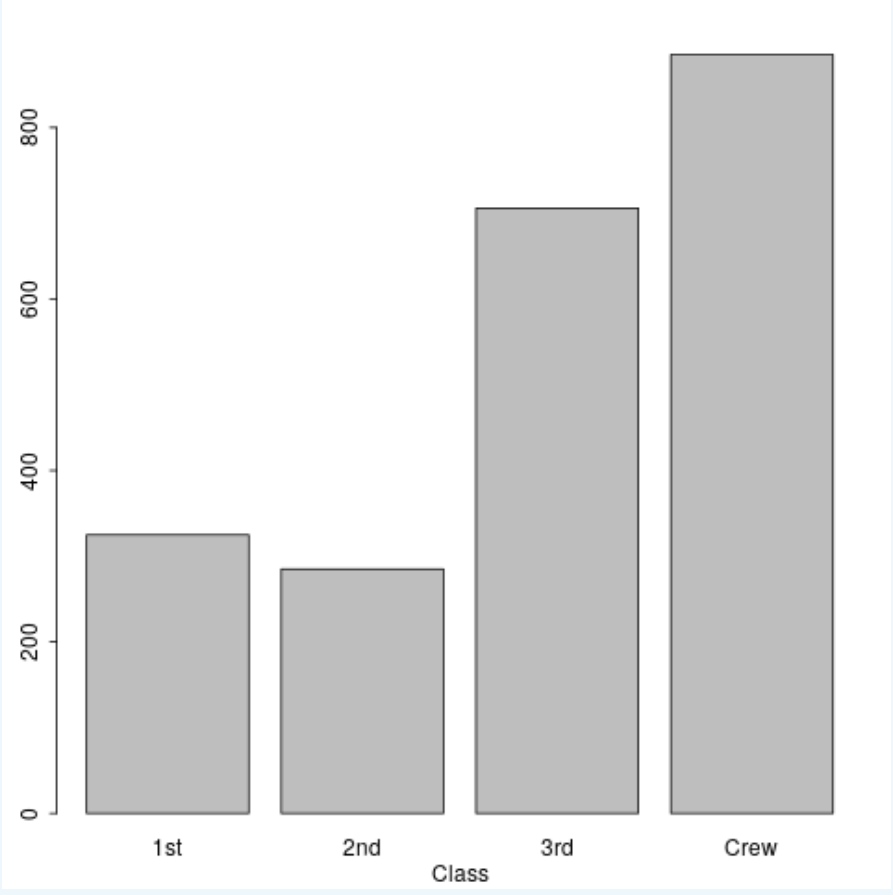
Données du Titanic : variable `Class`

1st	2nd	3rd	Crew
0.15	0.13	0.32	0.40

Histogramme

- k barres
- n_a = hauteur de la barre a (histogramme des effectifs)
- f_a = hauteur de la barre a (histogramme des fréquences)

Données du Titanic : variable `Class`



Deuxième variable qualitative

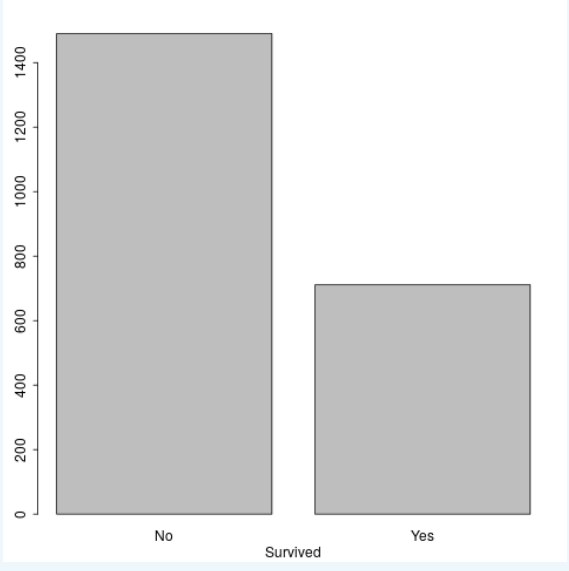
On considère une deuxième variable y ayant ℓ modalités.

Données du Titanic : variable `Survived`

- y = survie du passager au naufrage
- Modalités $\mathcal{Y} = \{\text{'No'}, \text{'Yes'}\}$
- Effectifs :

No	Yes
1490	711

- Histogramme



Variable indicatrice

Comme pour la variable x, on définit pour la variable y la variable indicatrice v_{ib}, pour $1 \leq i \leq n$ et $1 \leq b \leq \ell$:

$$v_{ib} = \begin{cases} 1 & \text{si } y_i = b, \\ 0 & \text{sinon.} \end{cases}$$

On associe également le vecteur $\mathbf{v}_b$ à la modalité b de la variable y.

Données du Titanic : variables `Class` et `Survived`

	Class				Survived	
i	1st	2nd	3rd	Crew	No	Yes
1	0	0	1	0	1	0
2	0	0	0	1	1	0
3	0	0	1	0	1	0
4	0	0	1	0	1	0
5	0	0	0	1	1	0
6	0	0	0	1	1	0

Notation vectorielle

On associe le vecteur $\mathbf{v}_b$ à la modalité b de la seconde variable qualitative :

$$\mathbf{v}_b = \begin{bmatrix} v_{1b} \\ v_{2b} \\ \vdots \\ v_{ib} \\ \vdots \\ v_{nb} \end{bmatrix}.$$

Données du Titanic : variables `Class` et `Survived`

	Class				Survived	
i	1st	2nd	3rd	Crew	No	Yes
1	0	0	1	0	1	0
2	0	0	0	1	1	0
3	0	0	1	0	1	0
4	0	0	1	0	1	0
5	0	0	0	1	1	0
6	0	0	0	1	1	0

Produit scalaire

Définition 4 (Produit scalaire)

Le produit scalaire entre deux vecteurs $\mathbf{u}$ et $\mathbf{v}$ *de même dimension* n

$$\mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_i \\ \vdots \\ u_n \end{bmatrix} \qquad \text{et} \qquad \mathbf{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ v_n \end{bmatrix}$$

est défini comme le produit terme à terme de leur coordonnées respectives :

$$< \mathbf{u}, \mathbf{v} > = \mathbf{u}^\intercal \, \mathbf{v} = \begin{bmatrix} u_1 & \cdots & u_i & \cdots & u_n \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ v_n \end{bmatrix}$$

$$= u_1 v_1 + u_2 v_2 + \cdots + u_n v_n = \sum_{i=1}^{n} u_i v_i .$$

Table de contingence / effectif croisé

Définition 5 (Table de contingence, effectif croisé)

La table de contingence est le tableau $k \times \ell$ dont le terme général n_{ab} est l'effectif croisé (ou joint) de la modalité a de la variable x et de la modalité b de la variable y :

$$n_{ab} = |\{i : x_i = a \text{ et } y_i = b\}| = \sum_{i=1}^{n} u_{ia} v_{ib}.$$

Proposition 3

$$n_{ab} = < \mathbf{u}_a, \mathbf{v}_b > .$$

Proposition 4

$$\sum_{a=1}^{k} \sum_{b=1}^{\ell} n_{ab} = n.$$

Table de contingence / effectif croisé

Table de contingence / effectif croisé

Données du Titanic : variables `Class` et `Survived`

n_{ab}	No	Yes
1st	122	203
2nd	167	118
3rd	528	178
Crew	673	212

Fréquence jointe

Définition 6

La fréquence jointe des modalités a et b désigne la proportion qu'elle représente par rapport à l'ensemble de la population

$$f_{ab} = n_{ab}/n.$$

Proposition 5

$$\sum_{a=1}^{k}\sum_{b=1}^{\ell} f_{ab} = 1.$$

Données du Titanic : variables `Class` et `Survived`

f_{ab}	No	Yes
1st	0.06	0.09
2nd	0.08	0.05
3rd	0.24	0.08
Crew	0.31	0.10

Fréquence jointe

Mesure d'association

Mesurer l'association entre deux variables qualitative x et y n'est pas immédiat car

- les variables x et y n'ont pas nécessairement le même nombre de modalités k et ℓ et
- il n'existe le plus souvent pas d'appariement prédéfini entre les modalités de l'une et de l'autre.

Définition 7 (Indice de Rand)

L'indice de Rand est la proportion de paires concordantes entre les variables x et y :

$$RI(x, y) = \sum_{a=1}^{k} \sum_{b=1}^{\ell} \binom{n_{ab}}{2} \Big/ \binom{n}{2} .$$

Données du Titanic : association entre les variables `Class` et `Survived`

n_{ab}	No	Yes
1st	122	203
2nd	167	118
3rd	528	178
Crew	673	212

$$RI(x,y) = \frac{\binom{122}{2} + \binom{203}{2} + \binom{167}{2} + \binom{118}{2} + \binom{528}{2} + \binom{178}{2} + \binom{673}{2} + \binom{212}{2}}{\binom{2201}{2}} \simeq 0.187.$$

Remarques.

1. Interprétation pas évidente ($RI(x,y) \in [0,1]$ ...).
2. Indice sensible aux effectifs des modalités.
3. Plusieurs variantes de cet indice visant à corriger des effets indésirables (ex : indice de Rand ajusté).

Effectif marginal

Définition 8 (Effectif marginal)

On note n_{a+} (resp. n_{+b}) l'effectif marginal de la modalité a de la variable x (resp. b de la variable y).

$$n_{a+} = \sum_{b=1}^{\ell} n_{ab}, \qquad n_{+b} = \sum_{a=1}^{k} n_{ab}.$$

Données du Titanic : variables `Class` et `Survived`

n_{ab}	No	Yes	n_{a+}
1st	122	203	325
2nd	167	118	285
3rd	528	178	706
Crew	673	212	885
n_{+b}	1490	711	$n = 2201$

Fréquence marginale

Définition 9 (Fréquence marginale)

Fréquences marginales

$$f_{a+} = n_{a+}/n, \qquad f_{+b} = n_{+b}/n.$$

Proposition 6

$$f_{a+} = \sum_{b=1}^{\ell} f_{ab}.$$

Données du Titanic : variables `Class` et `Survived`

f_{ab}	No	Yes	f_{a+}
1st	0.06	0.09	0.15
2nd	0.08	0.05	0.13
3rd	0.24	0.08	0.32
Crew	0.31	0.10	0.40
f_{+b}	0.68	0.32	1.00

Fréquence marginale

Fréquence relative

Définition 10 (Fréquences relatives)

$$f_{a|b} = n_{ab}/n_{+b}, \qquad f_{b|a} = n_{ab}/n_{a+}.$$

Données du Titanic : variables `Class` et `Survived`

$f_{a\|b}$	No	Yes
1st	0.08	0.29
2nd	0.11	0.17
3rd	0.35	0.25
Crew	0.45	0.30
	1.00	1.00

$f_{b\|a}$	No	Yes	
1st	0.38	0.62	1.00
2nd	0.59	0.41	1.00
3rd	0.75	0.25	1.00
Crew	0.76	0.24	1.00

Généralisation

On peut bien sûr vouloir analyser simultanément plus de deux variables.

Données du Titanic : variables x =`Class`, y =`Survived` et z =`Sex`

(3^{e} classe et équipage seulement : $n = 1591$)

	Class	Survived	Sex
1	3rd	No	Male
2	Crew	No	Male
3	3rd	No	Male
4	3rd	No	Male
5	Crew	No	Male
6	Crew	No	Male

On peut généraliser toutes les notations et quantités :

effectifs :

$$n_{abc} = |\{i : x_i = a, y_i = b, z_i = c\}|, \qquad n_{++c} = \sum_{a=1}^{k} \sum_{b=1}^{\ell} n_{abc}$$

fréquences relatives :

$$f_{ab|c} = n_{abc}/n_{++c}, \qquad f_{a|bc} = n_{abc}/n_{+bc}.$$

Comparaison de fréquence relatives (1/2)

Données du Titanic : taux de survie en fonction de la classe parmi les hommes

$a \in \mathcal{X} = \{\text{3rd}, \text{Crew}\}$, $b \in \mathcal{Y} = \{\text{No}, \text{Yes}\}$, $c = \text{Male}$

n_{abc}	No	Yes
3rd	422	88
Crew	670	192

$\Rightarrow$

$f_{b\|ac}$	No	Yes
3rd	0.83	0.17
Crew	0.78	0.22

Le taux de survie est plus élevé parmi les hommes d'équipage (22%) que parmi les passagers de 3^{e} classe (17%) .

Données du Titanic : taux de survie en fonction de la classe parmi les femmes

$a \in \mathcal{X} = \{\text{3rd}, \text{Crew}\}$, $b \in \mathcal{Y} = \{\text{No}, \text{Yes}\}$, $c = \text{Female}$

n_{abc}	No	Yes
3rd	106	90
Crew	3	20

$\Rightarrow$

$f_{b\|ac}$	No	Yes
3rd	0.54	0.46
Crew	0.13	0.87

Le taux de survie est plus élevé parmi les femmes d'équipage (87%) que parmi les passagères de 3^{e} classe (46%).

Comparaison de fréquence relatives (2/2)

Données du Titanic : taux de survie en fonction de la classe

$a \in \mathcal{X} = \{\text{3rd}, \text{Crew}\}$, $b \in \mathcal{Y} = \{\text{No}, \text{Yes}\}$

n_{ab}	No	Yes
3rd	528	178
Crew	673	212

$\Rightarrow$

$f_{b\|a}$	No	Yes
3rd	0.75	0.25
Crew	0.76	0.24

Le taux de survie est (légèrement) supérieur parmi les passagers (femmes ou hommes : 25%) de 3^{e} classe que parmi les membres d'équipage (24%).

(Alors qu'on observe l'écart inverse parmi les hommes *et* parmi les femmes)

‘Paradoxe’ de Simpson

Le ’paradoxe’ de Simpson n’est pas un paradoxe :

$$\left.\begin{array}{rcl} f_{\texttt{Survived|3rd,Male}} & > & f_{\texttt{Survived|Crew,Male}} \\ f_{\texttt{Survived|3rd,Female}} & > & f_{\texttt{Survived|Crew,Female}} \end{array}\right\} \not\Rightarrow \quad f_{\texttt{Survived|3rd}} > f_{\texttt{Survived|Crew}}$$

mais une simple conséquence de la définition des fréquences conditionnelles :

$$\left.\begin{array}{rcl} \dfrac{n_{\texttt{3rd,Male,Survived}}}{n_{\texttt{3rd,Male}}} & > & \dfrac{n_{\texttt{Crew,Male,Survived}}}{n_{\texttt{Crew,Male}}} \\ \dfrac{n_{\texttt{3rd,Female,Survived}}}{n_{\texttt{3rd,Female}}} & > & \dfrac{n_{\texttt{Crew,Female,Survived}}}{n_{\texttt{Crew,Female}}} \end{array}\right\} \not\Rightarrow \quad \frac{n_{\texttt{3rd,Survived}}}{n_{\texttt{3rd}}} > \frac{n_{\texttt{Crew,Survived}}}{n_{\texttt{Crew}}}$$

Effet de structure

Dans le cas du Titanic, ce résultat (un peu surprenant) est notamment dû aux proportions très différentes d'hommes et de femmes dans les deux classes.

Données du Titanic : taux de survie en fonction de la classe

$a \in \mathcal{X} = \{\text{3rd}, \text{Crew}\}$, $c \in \mathcal{Z} = \{\text{Male}, \text{Female}\}$

n_{ac}	Male	Female
3rd	510	196
Crew	862	23

$\Rightarrow$

$f_{c\|a}$	Male	Female
3rd	0.72	0.28
Crew	0.97	0.03

Données de recensement

Données sur le naufrage du Titanic : *tous* les passagers.

- Pas d'aléa d'échantillonnage (tous les passagers sont pris en compte).
- Pas de valeur de représentativité (il n'y a eu qu'un Titanic)

Exemple. La proportion de morts parmi les membres d'équipages (76%) est supérieures à la proportion de morts parmi les passagers de 3^{e} classe (75%).

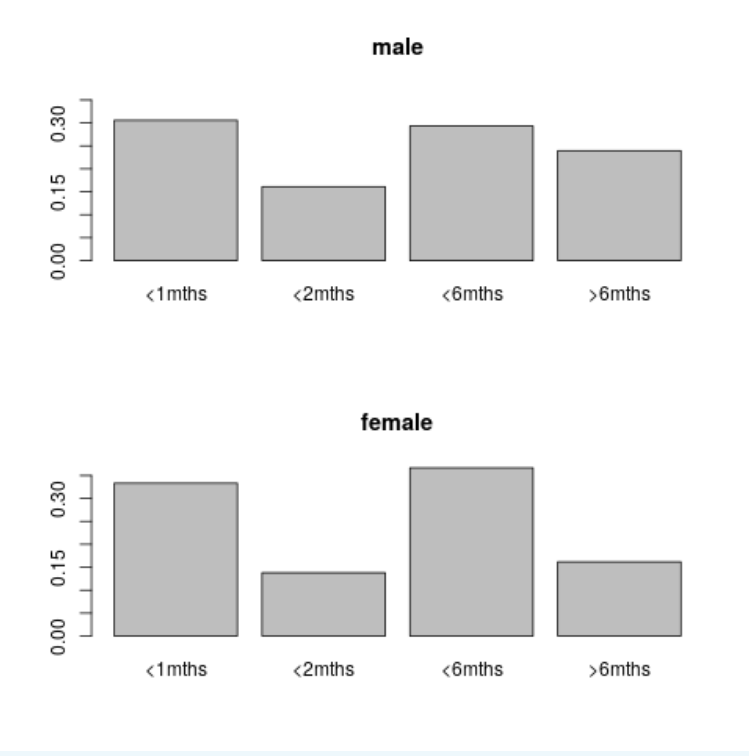
(De peu, certes, mais supérieure.)

Données issues d'un échantillonnage

Données de chômage

$n = 452$ personnes ayant connues une période de chômage (données USA, 1993)

	sex	reason	time
1	male	reentr	<1mths
2	male	lose	<2mths
3	male	lose	<1mths
4	male	reentr	<1mths
5	female	reentr	<1mths
6	female	reentr	<1mths



Inférence.

- Conclusions à tirer pas seulement sur ces $n = 452$ personnes.
- Vocation à être généralisable à une population plus large (ex : la population au chômage aux USA en 1993).