5.1 DISTRIBUTION CONJOINTE ET TEST D'INDÉPENDANCE

cours 27

Il arrive souvent qu'on veuille savoir s'il y a un lien entre deux ou plusieurs variables statistiques sur une même population.

Par exemple, si on fait une enquête on pourrait s'intéresser à l'âge, le sexe, le salaire, l'état civil, le niveau de scolarité, la langue maternelle, etc.

Ensuite on pourrait s'intéresser à voir s'il y a un lien entre le salaire et le sexe par exemple.

Pour des raisons de simplicité, nous n'allons que regarder deux variables statistiques à la fois.

Considérons une population ou un échantillon.

Intéressons-nous à deux variables statistiques

X Y

pouvant être des variables statistiques qualitatives ou quantitatives.

Ayant comme modalités

$$X = \{x_1, x_2, \ldots, x_k\}$$

$$Y = \{y_1, y_2, \ldots, y_p\}$$

Si les variables sont continues, on les regroupera par classes pour n'avoir qu'un nombre fini de modalités.

Par exemple on pourrait prendre comme population les enseignants d'un cégep.

X : L'âge $\qquad\qquad$ Y : L'état civil

Dans ce cas on pourrait avoir comme modalités

$$X = \{[20, 30[,\ \ [30, 40[,\ \ [40, 50[,\ \ [50, 60[\}$$

$$Y = \{\text{Marié, Célibataire, Divorcé, Veuf, Autre}\}$$

Avec les données recueillies pour ces variables, on les présente dans un tableau qu'on nomme **tableau de contingence**

	Marié	Célibataire	Veuf	Divorcé	Autre	Totaux
$[20, 30[$	4	41	0	0	0	57
$[30, 40[$	22	15	0	17	0	92
$[40, 50[$	30	12	4	25	2	74
$[50, 60[$	22	10	12	16	4	27
Totaux	78	108	16	58	6	250

Tableau de contingence

	y_1	y_2	$\dots$	y_j	$\dots$	y_p	Totaux
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1p}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2p}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ip}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kp}	$n_{k\bullet}$
Totaux	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet p}$	N ou n

le nombre d'individus ayant la modalité y_i pour la variable statistique Y et la modalité x_i pour la variable statistique X

Effectifs partielles de x_i et y_i

Tableau de contingence

	y_1	y_2	$\dots$	y_j	$\dots$	y_p	Totaux
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1p}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2p}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ip}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kp}	$n_{k\bullet}$
Totaux	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet p}$	N ou n

$$n_{\bullet 2} = \sum_{t=1}^{k} n_{t2}$$

Effectifs marginaux de Y

$$n_{2\bullet} = \sum_{t=1}^{p} n_{2t}$$

Effectifs marginaux de X

Tableau de contingence

	y_1	y_2	$\dots$	y_j	$\dots$	y_p	Totaux
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1p}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2p}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ip}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kp}	$n_{k\bullet}$
Totaux	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet p}$	N ou n

Distribution de Y

Distribution de X

Tableau de contingence

	y_1	y_2	$\dots$	y_j	$\dots$	y_p	Totaux
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1p}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2p}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ip}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kp}	$n_{k\bullet}$
Totaux	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet p}$	N ou n

Fréquence relative partielle $f_{ij} = \dfrac{n_{ij}}{N}$ $f_{ij} = \dfrac{n_{ij}}{n}$

Tableau de contingence

	y_1	y_2	$\dots$	y_j	$\dots$	y_p	Totaux
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1p}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2p}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ip}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kp}	$n_{k\bullet}$
Totaux	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet p}$	N ou n

Fréquences relatives conditionnelles

$$f_{x_i|y_j} = \frac{n_{ij}}{n_{\bullet j}}$$

$$f_{y_j|x_i} = \frac{n_{ij}}{n_{i\bullet}}$$

Tableau de contingence

	y_1	y_2	$\dots$	y_j	$\dots$	y_p	Totaux
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1p}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2p}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ip}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kp}	$n_{k\bullet}$
Totaux	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet p}$	N ou n

Fréquences relatives marginales

$$f_{\bullet j} = \frac{n_{\bullet j}}{N} \qquad f_{\bullet j} = \frac{n_{\bullet j}}{n}$$

$$f_{i\bullet} = \frac{n_{i\bullet}}{N} \qquad f_{i\bullet} = \frac{n_{i\bullet}}{n}$$

Faites les exercices suivants

#5.1

Test d'indépendance

	y_1	y_2	$\dots$	y_j	$\dots$	y_p	Totaux
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1p}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2p}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ip}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kp}	$n_{k\bullet}$
Totaux	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet p}$	N ou n

Y est indépendante de X si pour tous j

$$f_{y_j|x_1} = f_{y_j|x_2} = \cdots = f_{y_j|x_k} = f_{\bullet j}$$

	y_1	y_2	$\dots$	y_j	$\dots$	y_p	Totaux
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1p}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2p}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ip}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kp}	$n_{k\bullet}$
Totaux	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet p}$	N ou n

et X est indépendante de Y si pour tous i

$$f_{x_i|y_1} = f_{x_i|y_2} = \cdots = f_{x_i|y_p} = f_{i\bullet}$$

Naturellement, si Y est indépendante de X alors X est indépendante Y

Supposons que les deux variables statistiques sont indépendantes.

alors pour tous i et pour tous j

$$\frac{n_{ij}}{n_{i\bullet}} = \frac{n_{\bullet j}}{N}$$

$$n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{N} = T_{ij}$$

On nomme ces effectifs les **effectifs théoriques** puisque ce sont les effectifs qu'on devrait avoir si les variables sont indépendantes

C’est donc en comparant les effectifs partiels aux effectifs théoriques qu’on peut mesurer à quels points nos variables dépendent l’une de l’autre

Dans les faits, on travaille avec des échantillons plutôt qu’avec la population au complet.

Donc les écarts entre les effectifs partiels observés et les effectifs théoriques peuvent être imputés au fait qu’on est dans un échantillon.

On va donc faire un test d’hypothèse sur l’indépendance des variables statistiques.

	y_1	y_2	$\dots$	y_j	$\dots$	y_p	Totaux
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1p}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2p}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ip}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kp}	$n_{k\bullet}$
Totaux	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet p}$	n

Effectif observé

$O_{ij} = n_{ij}$

Effectif théorique

$T_{ij} = \dfrac{n_{i\bullet} n_{\bullet j}}{n}$

Si les variables sont indépendantes $T_{ij} = O_{ij}$

Par contre si les $T_{ij} \neq O_{ij}$

c'est peut-être dû au hasard d'échantillonnage

Pour vérifier si l'écart est dû au hasard d'échantillonnage ou au fait que les variables sont interdépendantes, on fait un test d'indépendance en utilisant comme hypothèse nulle

H_0 : X et Y sont indépendantes

Comment obtenir la zone d'acceptation?

On calcule une statistique qu'on nomme le χ^2

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} \frac{(O_{ij} - T_{ij})^2}{T_{ij}}$$

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} \frac{(O_{ij} - T_{ij})^2}{T_{ij}}$$

C'est bien beau calculer cette grosse somme, mais on doit savoir quoi faire avec le résultat.

Naturellement plus le χ^2 est grand plus on doit s'attendre à rejeter l'hypothèse nulle.

Par contre on doit tenir en compte la taille du tableau de contingence, car plus il est gros plus on additionne des écarts et donc plus grand sera le χ^2.

On va donc regardé dans une table de χ^2 avec degré de liberté d

	y_1	y_2	$\dots$	y_j	$\dots$	y_p	Totaux
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1p}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2p}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ip}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kp}	$n_{k\bullet}$
Totaux	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet p}$	n

On va donc regardé dans une table de χ^2avec degré de liberté d

$$d = (k-1)(p-1)$$

Donc on peut faire un test d'indépendance entre deux variables statistique.

Lorsqu'on rejette l'hypothèse, ça veut dire que les variables sont interdépendantes.

Mais à quel point?

Il existe plusieurs façons de mesurer l'interdépendance en fonction du contexte. On nomme ces mesures les coefficients d'association.

Le plus courant et le seul que nous verrons est le **coefficient de contingence**

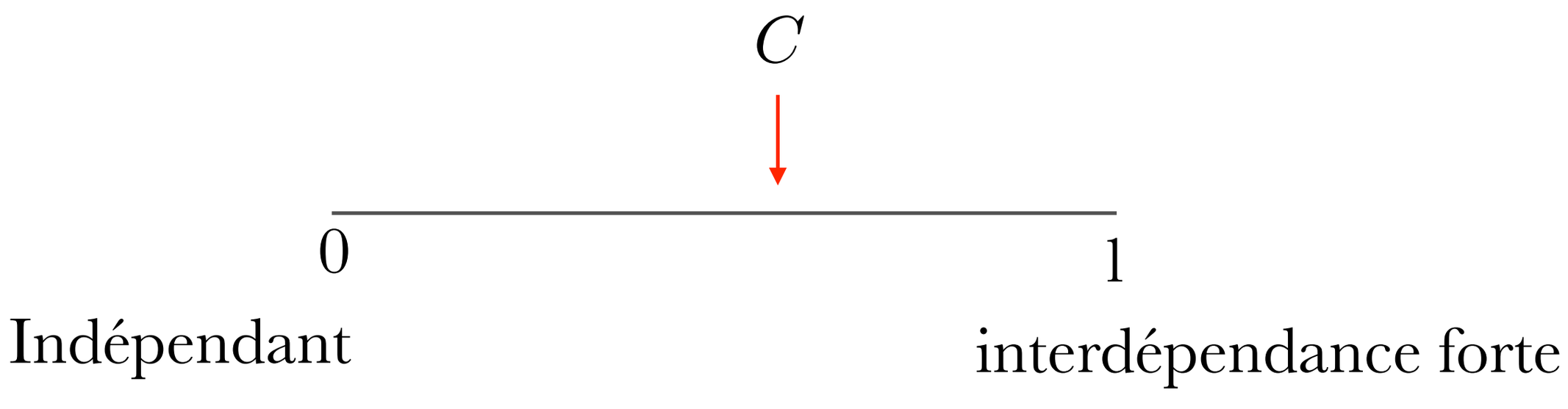
$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Le plus courant et le seul que nous verrons est le **coefficient de contingence**

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

On peut aisément vérifier que

$$0 \leq C < 1$$



Faites les exercices suivants

#5.2 à 5.7

Test d'ajustement

On peut aussi utiliser le test du χ^2

pour vérifier si une distribution donnée suit une certaine loi.

Mais pour ça, commençons par préciser comment on trouve le degré de liberté à utiliser dans la loi du χ^2

Le degré de liberté d'une loi du χ^2 est le nombre de termes de la somme

$$\sum_{i=1}^{k} \frac{(O_i - T_i)^2}{T_i}$$

moins le nombre de paramètres provenant des données observées qu'on doit utiliser dans le calcul des effectifs théoriques

Exemple Supposons qu'on ait une liste de 250 chiffres qu'on dit aléatoire. Si ces chiffres sont vraiment aléatoires, il est naturel de supposer que la fréquence de ces chiffres soit distribuée uniformément.

$$H_0 : X \sim U \qquad H_1 : X \nsim U \qquad \alpha = 0,05$$

Il y a 10 chiffres donc il y aura 10 termes dans le calcul du χ^2

Pour calculer la fréquence théorique, on n'a besoin que d'un nombre soit 250

Donc le degré de liberté est 9.

Donc si $\chi^2 > 16,92$ on rejettera H_0

Exemple

Supposons qu'on ait une liste de 250 chiffres qu'on dit aléatoire. Si ces chiffres sont vraiment aléatoires, il est naturel de supposer que la fréquence de ces chiffres soit distribuée uniformément.

Donc si $\chi^2 > 16,92$

	O_i	T_i	$O_i - T_i$	$(O_i - T_i)^2$
0	28	25	3	9
1	29	25	4	16
2	25	25	0	0
3	26	25	1	1
4	21	25	-4	16
5	34	25	9	81
6	20	25	-5	25
7	17	25	-8	64
8	24	25	-1	1
9	26	25	1	1

Exemple

Supposons qu'on ait une liste de 250 chiffres qu'on dit aléatoire. Si ces chiffres sont vraiment aléatoires, il est naturel de supposer que la fréquence de ces chiffres soit distribuée uniformément.

Donc si $\chi^2 > 16,92$

$(O_i - T_i)^2$
9
16
0
1
16
81
25
64
1
1

Exemple

Supposons qu'on ait une liste de 250 chiffres qu'on dit aléatoire. Si ces chiffres sont vraiment aléatoires, il est naturel de supposer que la fréquence de ces chiffres soit distribuée uniformément.

Donc si $\chi^2 > 16,92$

$$(O_i - T_i)^2$$

9	16	0	1	16	81	25	64	1	1

$$\chi^2 = \frac{9}{25} + \frac{16}{25} + \frac{0}{25} + \frac{1}{25} + \frac{16}{25} + \frac{81}{25} + \frac{25}{25} + \frac{64}{25} + \frac{1}{25} + \frac{1}{25}$$

$$= \frac{214}{25} = 8,56 < 16,92$$

Donc on accepte H_0

Faites les exercices suivants

#5.8 à 5.11

Devoir: # 5.1 à 5.11