EISTI
Département de mathématiques

Statistiques descriptives bivariées

Objectifs

- Observer simultanément des individus d'une population sur deux caractères
- Mesurer un lien éventuel entre deux caractères en utilisant un résumé chiffré qui traduit l'importance de ce lien
- Qualifier ce lien :
  - en cherchant une relation numérique approchée entre deux caractères quantitatifs
  - en cherchant des correspondances entre les modalités de deux caractères qualitatifs

2 types de variables $\Rightarrow$ 3 types de croisements :
- **qualitatif $\times$qualitatif**
- **qualitatif $\times$quantitatif**
- **quantitatif $\times$quantitatif**

Croisement qualitatif × qualitatif

Croisement Qualitatif ×Qualitatif

Département de mathématiques

Distribution conjointe

- Les seuls calculs possibles sur des variables qualitatives sont des

  effectifs et/ou des fréquences

- Chercher un lien entre deux variables qualitatives X et Y reviendra à étudier l'ensembles des effectifs des sous-populations définies par les couples de modalités (x_i,y_j) prises respectivement par X et Y.

Tableau de contingence

X \ Y	y_1		y_j		y_l
x_1	$n_{1,1}$				$n_{1,1}$
x_i	$n_{i,1}$		$n_{i,j}$		$n_{i,l}$
x_k	$n_{k,1}$		$n_{k,j}$		$n_{k,l}$

$n_{i,j}$ est le nombre d'individus ω tels que $X(\omega) = x_i$ et $Y(\omega)= y_j$

Exemple : Etude du lien entre la couleur des yeux et la couleur des cheveux sur un échantillon de 100 personnes

Yeux \ Cheveux	bruns	chatains	roux	blonds
bleus	11	10	1	8
verts	5	8	1	4
marrons	16	22	2	12

Croisement Qualitatif ×Qualitatif

Distribution marginale

	y_1		y_j		y_l	
x_1	$n_{1,1}$				$n_{1,l}$	$\mathbf{n_{1,.}}$
x_i	$n_{i,1}$		$n_{i,j}$		$n_{i,l}$	$\mathbf{n_{i,.}}$
x_k	$n_{k,1}$		$n_{k,j}$		$n_{k,l}$	$\mathbf{n_{k,.}}$
	$\mathbf{n_{.,1}}$		$\mathbf{n_{.,j}}$		$\mathbf{n_{.,l}}$	$\mathbf{n}$

Effectifs marginaux

pour X: $n_{i,.} = \sum_{j=1}^{l} n_{i,j}$ pour Y: $n_{.,j} = \sum_{i=1}^{k} n_{i,j}$

Effectif total

$$n = \sum_{j=1}^{l} n_{.,j} = \sum_{i=1}^{k} n_{i,.} = \sum_{i=1}^{k}\sum_{j=1}^{l} n_{i,j}$$

<u>Exemple</u> : Etude du lien entre la couleur des yeux et la couleur des cheveux

Yeux \ Cheveux	bruns	chatains	roux	blonds	
bleus	11	10	1	8	**30**
verts	5	8	1	4	**18**
marrons	16	22	2	12	**52**
	32	**40**	**4**	**24**	**100**

Comparaison des effectifs non pertinente

Croisement Qualitatif ×Qualitatif

Distribution marginale

Des effectifs ne sont pas directement comparables tandis que des fréquences sont toujours comparables

	y_1		y_j		y_l	
x_1	$f_{1,1}$				$f_{1,l}$	$\mathbf{f_{1,.}}$
x_i	$f_{i,1}$		$f_{i,j}$		$f_{i,l}$	$\mathbf{f_{i,.}}$
x_k	$f_{k,1}$		$f_{k,j}$		$f_{k,l}$	$\mathbf{f_{k,.}}$
	$\mathbf{f_{.,1}}$		$\mathbf{f_{.,j}}$		$\mathbf{f_{.,l}}$	**1**

$f_{i,j}$ est la *proportion* d'individus ω tels que $X(\omega) = x_i$ et $Y(\omega) = y_j$

Fréquences marginales

pour X : $f_{i,.} = \sum_{j=1}^{l} f_{i,j}$ pour Y: $f_{.,j} = \sum_{i=1}^{k} f_{i,j}$

$$1 = \sum_{j=1}^{l} f_{.,j} = \sum_{i=1}^{k} f_{i,.} = \sum_{i=1}^{k}\sum_{j=1}^{l} f_{i,j}$$

Exemple : Etude du lien entre la couleur des yeux et la couleur des cheveux

Yeux \ Cheveux	bruns	chatains	roux	blonds	
bleus	0,11	0,1	0,01	0,08	**0,3**
verts	0,05	0,08	0,01	0,04	**0,18**
marrons	0,16	0,22	0,02	0,12	**0,52**
	0,32	**0,4**	**0,04**	**0,24**	**1**

Croisement Qualitatif ×Qualitatif

Profils ligne et colonne

L'analyse croisée consiste à chercher des correspondances entre des modalités de X et des modalités de Y.

Profils lignes	y_1		y_j		y_l	
x_1	$f_{1/1}$				$f_{l/1}$	$f_{1,.}$
x_i	$f_{1/i}$		$f_{j/i}$		$f_{l/i}$	$f_{i,.}$
x_k	$f_{1/k}$		$f_{j/k}$		$f_{l/k}$	$f_{k,.}$
	$f_{.,1}$		$f_{.,j}$		$f_{.,l}$	

Profil ligne : répartition en fréquence de la variable Y dans une sous-population définie par une modalité de la variable X

$$f_{j/i} = \frac{n_{i,j}}{n_{i,.}}$$

comparable avec $f_{.j}$

Profil colonne : répartition en fréquence de la variable X dans une sous-population définie par une modalité de Y

$$f_{i/j} = \frac{n_{i,j}}{n_{.,j}}$$

comparable avec $f_{i.}$

La ligne des fréquences marginales de Y est appelée ***profil moyen***.

Croisement Qualitatif ×Qualitatif

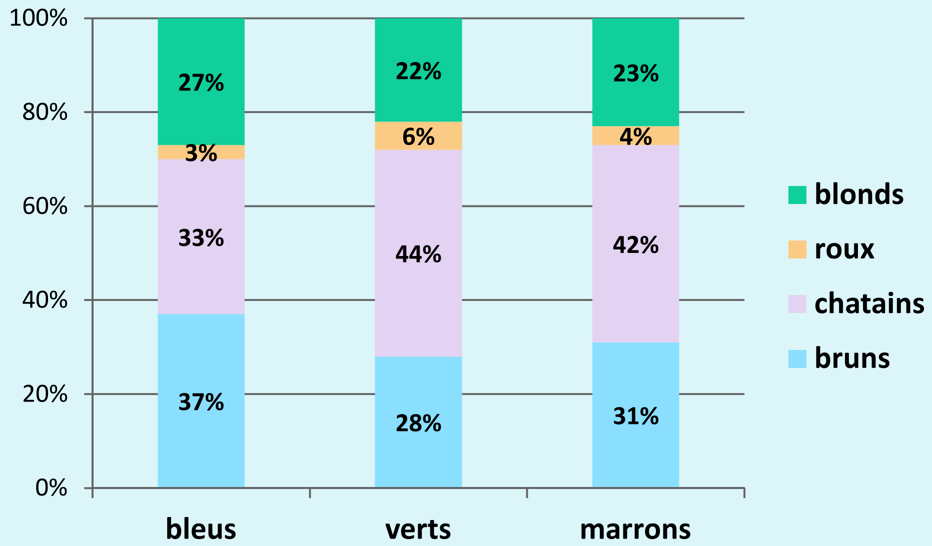
Profils ligne et colonne

Département de mathématiques

- **Exemple précédent**

Profils lignes

	bruns	chatains	roux	blonds	
bleus	0,37	0,33	0,03	0,27	1,00
verts	0,28	0,44	0,06	0,22	1,00
marrons	0,31	0,42	0,04	0,23	1,00
Freq. marginales	0,32	0,4	0,04	0,24	



- *28% des personnes ayant les yeux verts ont les cheveux bruns*
- *32% des personnes ont les cheveux bruns*

➢ Pour les profils lignes, on note que la répartition des couleurs de cheveux est la même quelle que soit la couleur des yeux et est la même que celle de la population totale..

Il ne semble pas y avoir de lien entre les modalités de ces deux caractères

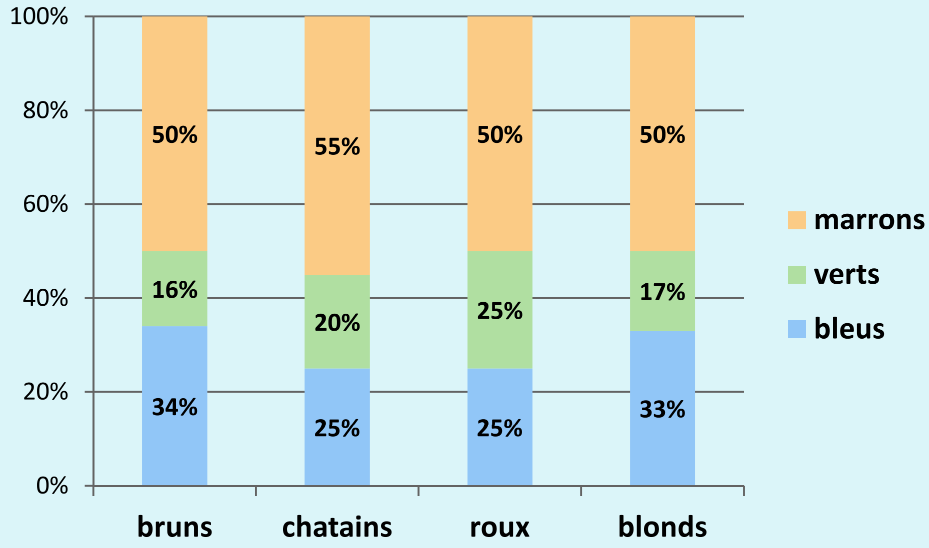
Croisement Qualitatif ×Qualitatif

Profils ligne et colonne

- **Exemple précédent**

Profils colonnes

	bruns	chatains	roux	blonds		Freq. marginales
bleus	0,34	0,25	0,25	0,33		0,3
verts	0,16	0,20	0,25	0,17		0,18
marrons	0,50	0,55	0,50	0,50		0,52
	1,00	1,00	1,00	1,00		



- *16% des bruns ont les yeux verts*
- *18% des personnes ont les yeux verts*

➢ Pour les profils colonnes, on note que la répartition des couleurs des yeux est la même quelle que soit la couleur des cheveux et est la même que celle de la population totale.

Il ne semble pas y avoir de lien entre les modalités de ces deux caractères

Croisement Qualitatif ×Qualitatif

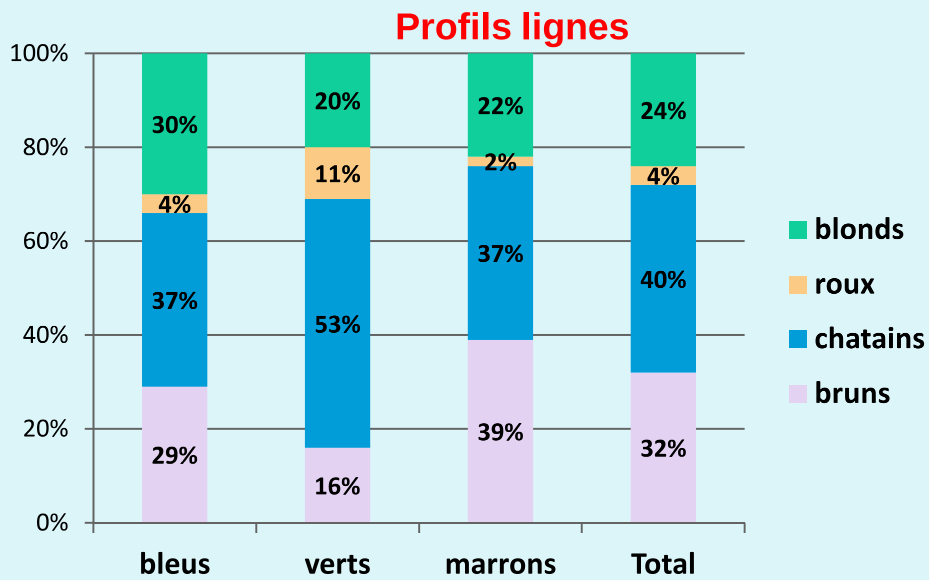
Profils ligne et colonne

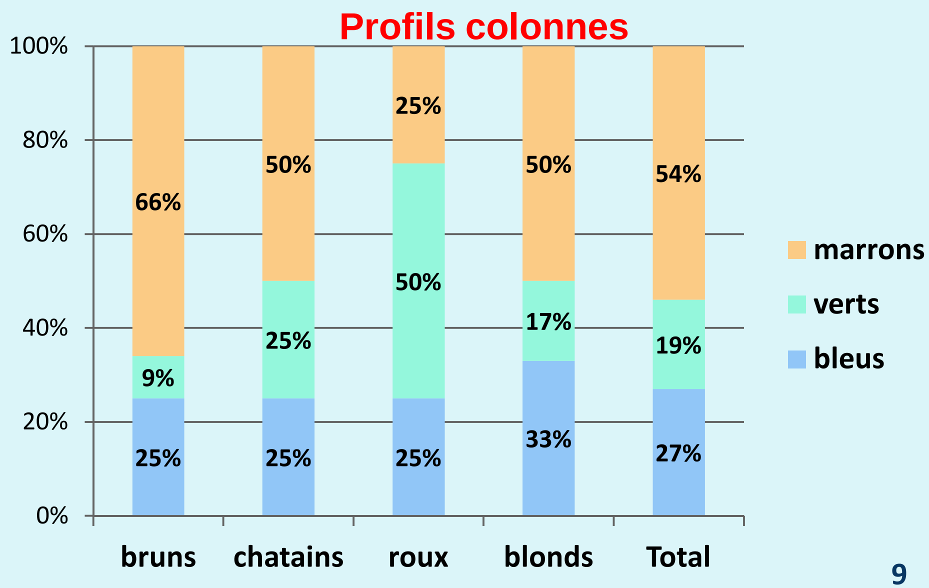
- **Exemple précédent modifié**

	bruns	chatains	roux	blonds	
bleus	8	10	1	8	**27**
verts	3	10	2	4	**19**
marrons	21	20	1	12	**54**
	32	**40**	**4**	**24**	**100**

Dans cet exemple, la répartition de la couleur des cheveux suivant la couleur des yeux n'est pas la même que celle de la population totale.

Il semble qu'il y ait un lien entre les modalités de ces deux variables





Croisement Qualitatif ×Qualitatif

Indépendance

X et Y ne sont pas liés

$\Leftrightarrow$ les profils lignes sont égaux $\Leftrightarrow$ les profils colonnes sont égaux

$$\Leftrightarrow \boxed{f_{i,j} = f_{i,.} \times f_{.,j}} \quad \forall i \in \{1,\dots,k\}, \forall j \in \{1,\dots,l\}$$

Comparaison des tableaux

Tableau de contingence théorique si X et Y sont indépendants

	y_1		y_j		y_l	
x_1	$f_{1,.}\times f_{.,1}$		$f_{1,.}\times f_{.,j}$		$f_{1,.}\times f_{.,l}$	$f_{1,.}$
x_i	$f_{i,.}\times f_{.,1}$		$f_{i,.}\times f_{.,j}$		$f_{i,.}\times f_{.,l}$	$f_{i,.}$
x_k	$f_{k,.}\times f_{.,1}$		$f_{k,.}\times f_{.,j}$		$f_{k,.}\times f_{.,l}$	$f_{k,.}$
	$f_{.,1}$		$f_{.,j}$		$f_{.,l}$	1

Tableau de contingence observé

	y_1		y_j		y_l	
x_1	$f_{1,1}$		$f_{1,j}$		$f_{1,l}$	$f_{1,.}$
x_i	$f_{i,1}$		$f_{i,j}$		$f_{i,l}$	$f_{i,.}$
x_k	$f_{k,1}$		$f_{k,j}$		$f_{k,l}$	$f_{k,.}$
	$f_{.,1}$		$f_{.,j}$		$f_{.,l}$	1

EISTI

Département de mathématiques

Croisement Qualitatif ×Qualitatif

Indépendance

	bruns	chatains	roux	blonds	
bleus	11	10	1	8	**30**
verts	5	8	1	4	**18**
marrons	16	22	2	12	**52**
	32	**40**	**4**	**24**	**100**

Tableau des effectifs observés

On divise par la taille de l'échantillon

	bruns	chatains	roux	blonds	
bleus	0,11	0,1	0,01	0,08	**0,3**
verts	0,05	0,08	0,01	0,04	**0,18**
marrons	0,16	0,22	0,02	0,12	**0,52**
	0,32	**0,4**	**0,04**	**0,24**	**1**

Tableau des fréquences observées

Comment comparer les deux tableaux?

$f_{i,.} \times f_{.,j}$

On multiplie par la taille de l'échantillon

	bruns	chatains	roux	blonds	
bleus	9,6	12	1,2	7,2	**30**
verts	5,76	7,2	0,72	4,32	**18**
marrons	16,64	20,8	2,08	12,48	**52**
	32	**40**	**4**	**24**	**100**

Tableau des effectifs théoriques

	bruns	chatains	roux	blonds	
bleus	**0,3×0,32**	0,12	0,01	0,07	**0,3**
verts	0,06	0,07	0,01	0,04	**0,18**
marrons	0,17	0,21	0,02	0,12	**0,52**
	0,32	**0,4**	**0,04**	**0,24**	**1**

Tableau des fréquences théoriques

Croisement Qualitatif ×Qualitatif

Indépendance et chi-deux

Comment mesurer le lien de dépendance entre les X et Y ? Comment mesurer la « distance » entre les deux tableaux? Mr Pearson a créée la ***distance du χ^2*** :

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(n_{i,j} - t_{i,j}\right)^2}{t_{i,j}}$$

où $t_{i,j} = n \times f_{i,.} \times f_{.,j}$ est l'effectif théorique de la case (i,j).

1. La distance du χ^2 est d'autant plus grande que X et Y sont liées entre eux.
2. La distance du χ^2 accorde plus d'importance aux différences entre les effectifs observés et effectifs théoriques sur les petits effectifs théoriques. S'écarter de 2% par rapport à 75% est moins significatif que de s'écarter de 2% par rapport à 5% .
3. La distance du χ^2 respecte le principe d'équivalence distributionnelle.
   - Si deux colonnes ont des effectifs proportionnels alors la fusion des modalités correspondantes du caractère X ne change pas la distance du χ^2 entre X et Y .
   - Si deux lignes ont des effectifs proportionnels alors la fusion des modalités correspondantes du caractère X ne change pas la distance du χ^2 entre X et Y .
4. Malheureusement la distance du χ^2 dépend aussi :
   - du nombre de modalités de X et Y
   - du nombre d'individus.
5. On ne peut donc comparer deux distance du χ^2 que sur deux tableaux strictement équivalents en modalités et en nombre d'individus.

Croisement Qualitatif ×Qualitatif

Indépendance et coefficients normalisés

- *Coefficient de contingence* : $$CC = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

  CC varie entre 0 et presque 1. Plus il est proche de 0 plus X et Y sont indépendants et plus il est proche de 1 plus X et Y sont liés. Par contre il dépend de k et s. On ne peut donc comparer que des tableaux de mêmes dimensions.

- *Coefficient de Van Cramer :* $$V = \sqrt{\frac{\chi^2}{n \times [\min(k,s) - 1]}}$$

  Même interprétation que le coefficient précédent avec l'avantage de ne plus dépendre de k et s. C'est le coefficient normalisé le plus utilisé.

- Il existe d'autres coefficients comme le coefficient phi de Pearson ou le PEM Le PEM (Pourcentage de l'Écart Maximum).
- Mais il faut retenir :
  1. que ces coefficients ne varie proportionnellement avec l'importance du lien
  2. que plus ils sont proches de 0 plus X et Y sont indépendants et plus ils sont proches de 1 plus X et Y sont liés.
  3. qu'il faut comparer l'évolution dans le temps de ces coefficients sur des tableaux équivalents