Chapitre III. Observation d'un couple de variables

1) Distribution conjointe et tableau de contingence

On observe simultanément 2 variables X et Y sur un échantillon de n individus d'une population donnée. A chaque individu de l'échantillon est donc associé un couple de réponses à X et Y.

On notera (x_i, y_i) la réponse à (X, Y) pour l'individu numéro i de l'échantillon. On notera aussi K et K' les nombres de modalités (ou de classes dans le cas d'une variable quantitative continue) de X et de Y.

Pour des variables quantitatives discrètes ou des variables qualitatives, les ensemble des modalités pourront alors s'écrire $\mathcal{M}_X = \{m_1, \ldots, m_K\}$ et $\mathcal{M}_Y = \{m'_1, \ldots, m'_{K'}\}$.

Comme dans le cas d'une seule variable, les données peuvent être présentées sous la forme d'un tableau d'effectifs où, pour chaque couple de modalités, on a compté le nombre d'individus ayant pour réponse ce couple de modalités.

Ce tableau est appelée **tableau de contingence** ou **distribution conjointe en effectifs de** (X, Y).

$X \setminus Y$	$\mathbf{m'_1}$	$\mathbf{m'_2}$	$\dots$	$\mathbf{m'_j}$	$\dots$	$\mathbf{m'_{K'}}$
$\mathbf{m_1}$	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	$n_{1K'}$
$\mathbf{m_2}$	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	$n_{2K'}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{m_i}$	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	$n_{iK'}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{m_K}$	n_{K1}	n_{K2}	$\dots$	n_{Kj}	$\dots$	$n_{KK'}$

On a donc un tableau à K lignes (Nbre de modalités de X) et K' colonnes (Nbre de modalités de Y) avec les effectifs pour les $K \times K'$ couples de modalités (m_i, m'_j), $(1 \leq i \leq K, 1 \leq j \leq K')$. Par exemple, à l'intersection de la $i$ème ligne et de la $j$ème colonne, l'effectif n_{ij} représente le nombre d'individus de l'échantillon ayant à la fois les modalités (réponses) m_i pour X et m'_j pour Y.

La somme des $K \times K'$ effectifs n_{ij} $(1 \leq i \leq K, 1 \leq j \leq K')$ est égale à n, ce qui se traduit par la formule suivante:

$$\sum_{i=1}^{K} \sum_{j=1}^{K'} n_{ij} = n$$

On peut également remplacer les effectifs par les fréquences. Pour ceci, il suffit de diviser chaque effectif par n,

$$f_{ij} = \frac{n_{ij}}{n}$$

Le tableau obtenu représentera alors la **distribution (conjointe) en fréquences de X et Y**.
La somme des fréquences est égale à 1 (ou 100% s'il s'agit de pourcentages), c'est-à-dire,

$$\sum_{i=1}^{K}\sum_{j=1}^{K'} f_{ij} = 1.$$

Note: Le tableau de contingence est beaucoup plus lisible que la liste des données brutes mais résulte en une perte d'information. En effet, à partir du tableau de contingence, on ne peut pas reconstituer la liste des données brutes (alors que le contraire est possible), en particulier on ne peut pas connaître le couple de réponses à (X, Y) pour un individu donné.

2) Distributions marginales

A partir de la distribution (conjointe) de X et Y, on peut en déduire la **distribution marginale de X** (appelée aussi distribution de X) et la **distribution marginale de Y** (ou distribution de Y). Le mot "marginal" vient du fait qu'on les présente souvent en "marge" du tableau de contingence, en parallèle à la liste de modalités.

Effectifs marginaux: **l'effectif marginal de la modalité m_i** de X correspond au nombre d'individus dont la réponse à X est m_i. On le note $n_{i\bullet}$ et on l'obtient en faisant la somme des K' effectifs sur la $i$ème ligne, $n_{i1}, n_{i2}, \ldots, n_{iK'}$, ce qui se traduit par la formule:

$$n_{i\bullet} = \sum_{j=1}^{K'} n_{ij}$$

Note: Le "point" en deuxième position signifie donc que l'on somme sur le deuxième indice j (i est fixé).

De même, on peut calculer **l'effectif marginal de la modalité** m'_j de Y en faisant la somme des K effectifs sur la $j$ème colonne:

$$n_{\bullet j} = \sum_{i=1}^{K} n_{ij}$$

La **fréquence marginale de la modalité** m_i est notée $f_{i\bullet}$ et est égale à l'effectif marginal $n_{i\bullet}$ (somme des effectifs de la $i$ème ligne) divisé par la taille de l'échantillon n.
De même la **fréquence marginale de la modalité** m'_j est notée $f_{\bullet j}$ et est égale à l'effectif marginal $n_{\bullet j}$ (somme des effectifs de la $j$ème colonne) divisé par la taille de l'échantillon n.

Le tableau ci-dessous est le tableau de contingence avec les marges (en effectifs).

$X \setminus Y$	$\mathbf{m'_1}$	$\mathbf{m'_2}$	$\dots$	$\mathbf{m'_j}$	$\dots$	$\mathbf{m'_{K'}}$	Marge X
$\mathbf{m_1}$	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	$n_{1K'}$	$n_{1\bullet}$
$\mathbf{m_2}$	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	$n_{2K'}$	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{m_i}$	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	$n_{iK'}$	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{m_K}$	n_{K1}	n_{K2}	$\dots$	n_{Kj}	$\dots$	$n_{KK'}$	$n_{K\bullet}$
Marge Y	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet K'}$	n

Notons que la somme des effectifs marginaux de X est égale à n, $\sum_{i=1}^{K} n_{i\bullet} = n$. Idem pour les effectifs marginaux de Y, $\sum_{j=1}^{K'} n_{\bullet j} = n$.

Ce tableau fournit les 3 distributions (en effectifs): la distribution (marginale) de X (dernière colonne) avec ses modalités (1ère colonne), la distribution (marginale) de Y (dernière ligne) avec ses modalités (1ère ligne) et la distribution conjointe de X et Y (à l'intérieur du tableau) avec ses couples de modalités (m_i, m'_j), $i = 1, \ldots, K$, $j = 1, \ldots, K'$.

Exemple 1: Notes d'étudiants et filières d'étude

$X \setminus Y$	$[0, 6[$	$[6, 10[$	$[10, 14[$	$[14, 20]$	Total X
Filière A	26	6	4	1	**37**
Filière B	12	9	3	1	**25**
Filière C	1	4	5	6	**16**
Filière D	10	8	3	1	**22**
Total Y	**49**	**27**	**15**	**9**	**n=100**

Sur les 100 étudiants de l'échantillon, il y a donc, par exemple, $n_{\bullet 2} = 27$ étudiants (toutes filières confondues) qui ont une note comprise entre 6 et 10 et il y a en $n_{3\bullet} = 16$ qui sont issus de la filière C (quelque soit leur note).

Exemple 5: Couleurs des yeux et des cheveux

$X \setminus Y$	Brun	Noir	Roux	Blond	Total X
Marron	113	72	7	39	**231**
Gris-vert	38	41	10	27	**116**
Bleu	20	17	8	53	**98**
Total Y	**171**	**130**	**25**	**119**	**n = 445**

Sur les 445 élèves de l'échantillon, il y a donc, par exemple, $n_{\bullet 1} = 171$ élèves qui ont les cheveux bruns (quelque soit la couleur de leur yeux) et il y a en $n_{2\bullet} = 116$ qui ont les yeux gris-vert (quelque soit la couleur de leur cheveux).

3) Distributions conditionnelles

A partir d'un couple de variables donné (X, Y), on peut définir les distributions conditionnelles de X et les distributions conditionnelles de Y.

Etant donné une modalité m'_j de Y, on définit la **distribution conditionnelle de X sachant $Y = m'_j$** (appelée aussi distribution de X conditionnée par une modalité m'_j de Y ou plus simplement **distribution conditionnelle de** $X_{Y=m'_j}$) comme la distribution de X restreinte au sous-échantillon (c'est-à dire une partie, un sous-groupe de l'échantillon) des seuls individus dont la réponse (la modalité) à Y est m'_j.

De même, étant donné une modalité m_i de X, on définit la **distribution conditionnelle de Y sachant $X = m_i$** (ou **distribution conditionnelle de** $Y_{X=m_i}$) comme la distribution de Y restreinte au sous-échantillon des seuls individus dont la réponse (la modalité) à X est m_i.

Par la suite par souci de simplicité dans les notations, $X_{m'_j}$ (resp. Y_{m_i}) pourra désigner $X_{Y=m'_j}$ (resp. $Y_{X=m_i}$), pour tout $j = 1, \ldots, K'$ (resp. pour tout $i = 1, \ldots, K$).

La distribution conditionnelle de $X_{m'_j}$ (en effectifs et en fréquences) est donnée dans le tableau ci-dessous:

mod. X	$\mathbf{m_1}$	$\mathbf{m_2}$	$\ldots$	$\mathbf{m_i}$	$\ldots$	$\mathbf{m_K}$	Tot.
Eff.	n_{1j}	n_{2j}	$\ldots$	n_{ij}	$\ldots$	n_{Kj}	$n_{\bullet j}$
Fréq.	$n_{1j}/n_{\bullet j}$	$n_{2j}/n_{\bullet j}$	$\ldots$	$n_{ij}/n_{\bullet j}$	$\ldots$	$n_{Kj}/n_{\bullet j}$	1

Pour obtenir les fréquences de la distribution conditionnelle de $X_{m'_j}$, on a divisé chaque effectif par le total des effectifs de la $j$ème colonne, à savoir $n_{\bullet j}$. On notera $f_{i|j} = n_{ij}/n_{\bullet j}$, la fréquence conditionnelle de $X_{m'_j}$ associée à la modalité m_i

Attention! La distribution conditionnelle en fréquences de $X_{m'_j}$ est différente de la $j$ème colonne du tableau de contingence en fréquences, vu qu'ici on a divisé par $n_{\bullet j}$ et non par n.

De même, la distribution conditionnelle de Y_{m_i} (en effectifs et en fréquences) est donnée dans le tableau ci-dessous:

mod. Y	$\mathbf{m'_1}$	$\mathbf{m'_2}$	$\ldots$	$\mathbf{m'_j}$	$\ldots$	$\mathbf{m'_{K'}}$	Tot.
Eff.	n_{i1}	n_{i2}	$\ldots$	n_{ij}	$\ldots$	$n_{iK'}$	$n_{i\bullet}$
Fréq.	$n_{i1}/n_{i\bullet}$	$n_{i2}/n_{i\bullet}$	$\ldots$	$n_{ij}/n_{i\bullet}$	$\ldots$	$n_{iK'}/n_{i\bullet}$	1

Pour obtenir les fréquences de la distribution conditionnelle de Y_{m_i}, on a divisé chaque effectif par le total des effectifs de la $i$ème ligne, à savoir $n_{i\bullet}$. On notera $f_{j|i} = n_{ij}/n_{i\bullet}$, la fréquence conditionnelle de Y_{m_i} associée à la modalité m'_j

Il y a en tout K' distributions conditionnelles de X (autant que de modalités de Y), à savoir $X_{m'_1}, X_{m'_2}, \ldots, X_{m'_{K'}}$ représentées dans le tableau ci-dessous (distributions en fréquences).

mod. X	$\mathbf{X_{m'_1}}$	$\mathbf{X_{m'_2}}$	$\ldots$	$\mathbf{X_{m'_j}}$	$\ldots$	$\mathbf{X_{m'_{K'}}}$
$\mathbf{m_1}$	$n_{11}/n_{\bullet 1}$	$n_{12}/n_{\bullet 2}$	$\ldots$	$n_{1j}/n_{\bullet j}$	$\ldots$	$n_{1K'}/n_{\bullet K'}$
$\mathbf{m_2}$	$n_{21}/n_{\bullet 1}$	$n_{22}/n_{\bullet 2}$	$\ldots$	$n_{2j}/n_{\bullet j}$	$\ldots$	$n_{2K'}/n_{\bullet K'}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{m_i}$	$n_{i1}/n_{\bullet 1}$	$n_{i2}/n_{\bullet 2}$	$\ldots$	$n_{ij}/n_{\bullet j}$	$\ldots$	$n_{iK'}/n_{\bullet K'}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{m_K}$	$n_{K1}/n_{\bullet 1}$	$n_{K2}/n_{\bullet 2}$	$\ldots$	$n_{Kj}/n_{\bullet j}$	$\ldots$	$n_{KK'}/n_{\bullet K'}$
Total	1	1	$\ldots$	1	$\ldots$	1

Ce tableau donne donc les distributions conditionnelles de X en fréquences (les K' distributions se lisent en colonnes). Pour avoir les distributions en effectifs, il suffit de prendre les mêmes effectifs que dans le tableau de contingence (en effectifs).
Notons que les modalités de $X_{m'_j}$, pour tout $j = 1, \ldots, K'$, sont les mêmes que celles de X, à savoir $m_1, \ldots, m_K$.

Il y a K distributions conditionnelles de Y (autant que de modalités de X), $Y_{m_1}, Y_{m_2}, \ldots, Y_{m_K}$ représentées dans le tableau ci-dessous.
Rappelons que Y_{m_i} désigne $Y_{X=m_i}$ pour $i = 1, \ldots, K$.

mod. Y	$\mathbf{m'_1}$	$\mathbf{m'_2}$	...	$\mathbf{m'_j}$	...	$\mathbf{m'_{K'}}$	Total
$\mathbf{Y_{m_1}}$	$n_{11}/n_{1\bullet}$	$n_{12}/n_{1\bullet}$	...	$n_{1j}/n_{1\bullet}$	...	$n_{1K'}/n_{1\bullet}$	1
$\mathbf{Y_{m_2}}$	$n_{21}/n_{2\bullet}$	$n_{22}/n_{2\bullet}$	...	$n_{2j}/n_{2\bullet}$	...	$n_{2K'}/n_{2\bullet}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	1
$\mathbf{Y_{m_i}}$	$n_{i1}/n_{i\bullet}$	$n_{i2}/n_{i\bullet}$	...	$n_{ij}/n_{i\bullet}$	...	$n_{iK'}/n_{i\bullet}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	1
$\mathbf{Y_{m_K}}$	$n_{K1}/n_{K\bullet}$	$n_{K2}/n_{K\bullet}$	...	$n_{Kj}/n_{K\bullet}$	...	$n_{KK'}/n_{K\bullet}$	1

Ce tableau donne donc les distributions conditionnelles de Y en fréquences (les K distributions se lisent en lignes). Pour avoir les distributions en effectifs, il suffit de prendre les mêmes effectifs que dans le tableau de contingence (en effectifs).
Notons que les modalités de Y_{m_i}, pour tout $i = 1, \ldots, K$, sont les mêmes que celles de Y, à savoir $m'_1, \ldots, m'_{K'}$.

Nous donnons à présent les distributions conditionnelles de X et celles de Y(en effectifs et en fréquences) pour **Exemple 1** et **Exemple 5** (les deux exemples qui font intervenir 2 variables).

Exemple 1: Notes d'étudiants et filières d'étude

Les 4 distributions conditionnelles de X sont données dans le tableau ci-dessous (les fréquences sont entre parenthèses en caractères gras):

mod. X	$X_{[0,6[}$	$X_{[6,10[}$	$X_{[10,14[}$	$X_{[14,20]}$
A	26 **(0.531)**	6 **(0.222)**	4 **(0.267)**	1 **(0.111)**
B	12 **(0.245)**	9 **(0.333)**	3 **(0.2)**	1 **(0.111)**
C	1 **(0.02)**	4 **(0.148)**	5 **(0.333)**	6 **(0.667)**
D	10 **(0.204)**	8 **(0.296)**	3 **(0.2)**	1 **(0.111)**
Total	49 **(1)**	27 **(1)**	15 **(1)**	9 **(1)**

Par exemple, la fréquence conditionnelle $f_{1|1}$ a été obtenue en faisant $26/49 = 0.531$. De même, la fréquence conditionnelle $f_{3|4}$ a été obtenue en faisant $6/9 = 0.667$.

Interprétation: la fréquence $f_{1|1} = 0.531$ veut dire que parmi les étudiants ayant eu une note comprise entre 0 et 6, 53.1% sont issus de la filière A. La fréquence $f_{3|4} = 0.667$ signifie que parmi les étudiants ayant eu une note comprise entre 14 et 20, 66.7% sont issus de la filière C.

Les 4 distributions conditionnelles de Y sont données dans le tableau ci-dessous (les fréquences sont entre parenthèses en caractères gras):

mod. Y	$[0, 6[$	$[6, 10[$	$[10, 14[$	$[14, 20]$	Total
Y_A	26 **(0.703)**	6 **(0.162)**	4 **(0.108)**	1 **(0.027)**	37 **(1)**
Y_B	12 **(0.48)**	9 **(0.36)**	3 **(0.12)**	1 **(0.04)**	25 **(1)**
Y_C	1 **(0.063)**	4 **(0.25)**	5 **(0.313)**	6 **(0.375)**	16 **(1)**
Y_D	10 **(0.455)**	8 **(0.364)**	3 **(0.136)**	1 **(0.045)**	22 **(1)**

Par exemple, la fréquence conditionnelle $f_{3|2}$ a été obtenue en faisant $3/25 = 0.12$.

Interprétation: la fréquence $f_{3|2} = 0.12$ signifie que parmi les étudiants issus de la filière B, 12% ont obtenu une note comprise entre 10 et 14.

Exemple 5: Couleurs des yeux et des cheveux

Les 4 distributions conditionnelles de X sont données dans le tableau ci-dessous (les fréquences sont entre parenthèses en caractères gras):

mod. X	X_{Brun}	X_{Noir}	X_{Roux}	X_{Blond}
Marron	113 **(0.661)**	72 **(0.554)**	7 **(0.28)**	39 **(0.328)**
Gris-vert	38 **(0.222)**	41 **(0.315)**	10 **(0.4)**	27 **(0.227)**
Bleu	20 **(0.117)**	17 **(0.131)**	8 **(0.32)**	53 **(0.445)**
Total	171 **(1)**	130 **(1)**	25 **(1)**	119 **(1)**

Interprétation: par exemple, la fréquence conditionnelle $f_{2|4} = 0.227$ signifie que parmi les élèves blonds, 22.7% ont les yeux gris-vert.

Les 3 distributions conditionnelles de Y sont données dans le tableau ci-dessous (les fréquences sont entre parenthèses en caractères gras):

mod. Y	Brun	Noir	Roux	Blond	Total
Y_{Marron}	113 (**0.489**)	72 (**0.312**)	7 (**0.030**)	39 (**0.169**)	231 (**1**)
Y_{Gris}	38 (**0.328**)	41 (**0.353**)	10 (**0.086**)	27 (**0.233**)	116 (**1**)
Y_{Bleu}	20 (**0.204**)	17 (**0.173**)	8 (**0.082**)	53 (**0.541**)	98 (**1**)

Interprétation: par exemple, la fréquence conditionnelle $f_{2|1} = 0.312$ signifie que parmi les élèves ayant les yeux marrons, 31.2% ont les cheveux noirs.

4) Représentations graphiques

Représentation simultanée des conditionnelles de X ou de Y : cas des variables qualitatives et des variables discrètes

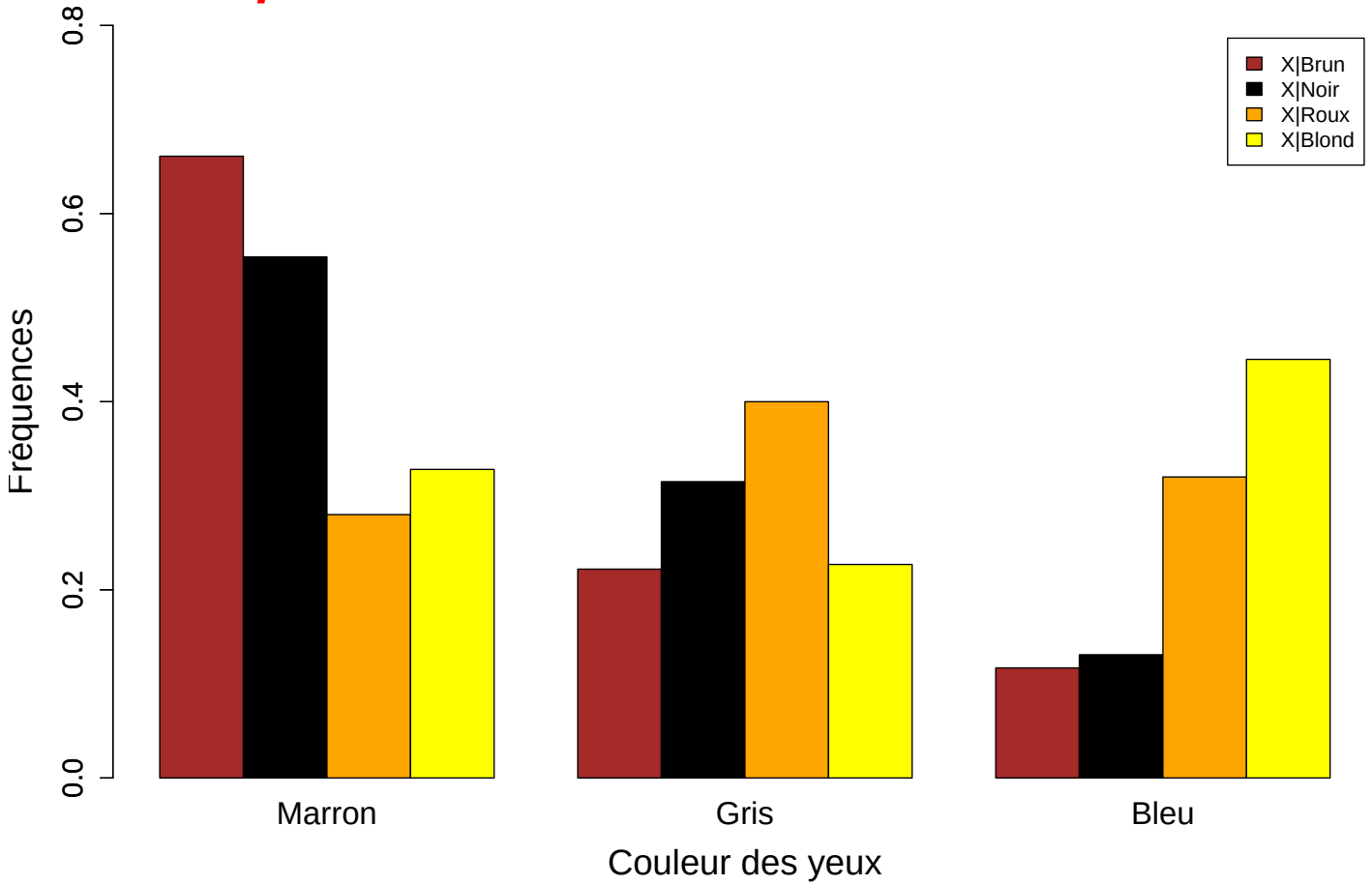
Si la variable X (resp. Y) n'est pas continue, on peut représenter simultanément (sur un même graphique) les diagrammes en barres de ses distributions conditionnelles.

On place les modalités de X (resp. Y) sur l'axe horizontal, dans l'ordre si la variable est ordonnée (qualitative ordinale ou quantitative discrète) et en respectant l'échelle si elle est quantitative discrète.

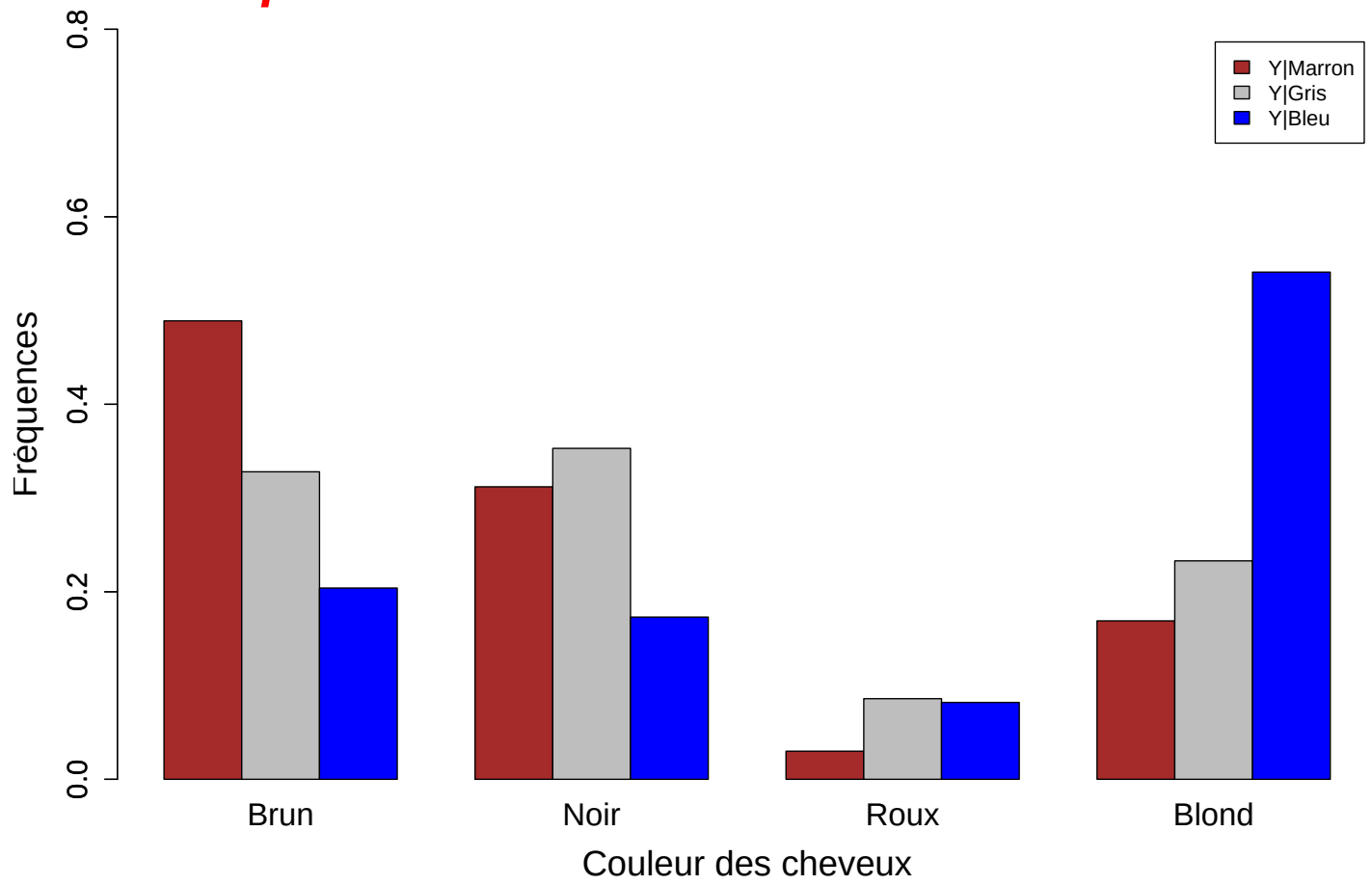
Au niveau de chaque modalité de X (resp. Y), on trace K' (resp. K) barres (ou bâtons) côte à côte (1 par distribution conditionnelle), chacune de longueur égale à la la fréquence correspondante.

Exemple 5: Couleurs des yeux et des cheveux

Graphe des distributions conditionnelles de X



Graphe des distributions conditionnelles de Y



Représentation simultanée des conditionnelles de X ou de Y : cas des variables continues

Dans le cas où X (resp. Y) est continue, on peut représenter les fonctions de répartition des distributions conditionnelles de X (resp. de Y) sur le même graphe.

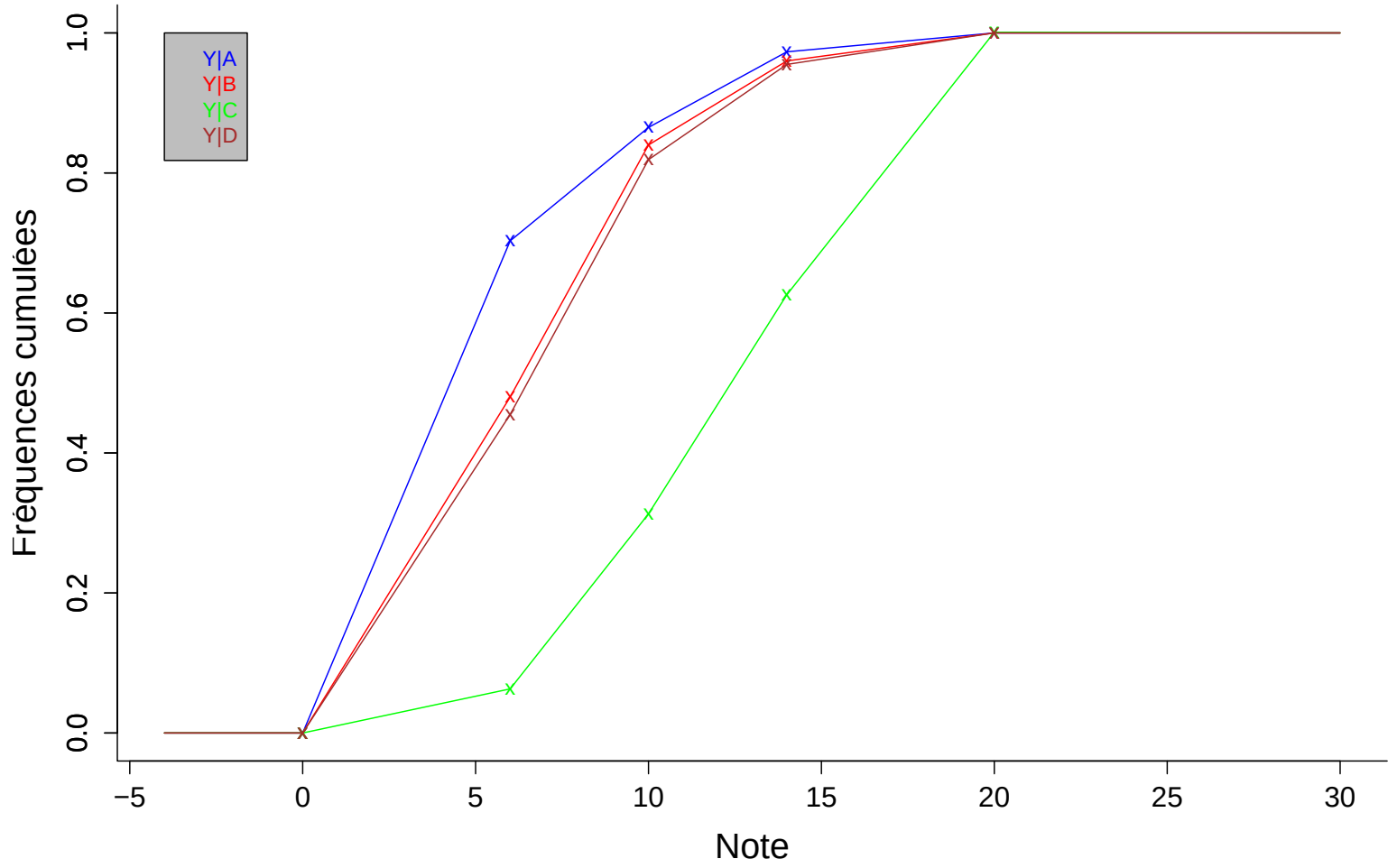
Les distributions conditionnelles de X (resp. Y) peuvent être classées par ordre de grandeur globale de la variable X (resp. Y) en comparant leur fonction de répartition (voir Chapitre IV).

Exemple 1: Notes d'étudiants et filières d'étude

Les fréquences (par exemple f_{Y_A} pour Y_A) et fréquences cumulées (par exemple F_{Y_A} pour Y_A) sont données pour les 4 distributions conditionnelles de Y dans le tableau ci-dessous:

Note	$[0,6[$	$[6,10[$	$[10,14[$	$[14,20]$	Total
Bornes	0	6	10	14	20	
f_{Y_A}	0.703	0.162	0.108	0.027	1
$\mathbf{F_{Y_A}}$	**0**	**0.703**	**0.865**	**0.973**	**1**	
f_{Y_B}	0.48	0.36	0.12	0.04	1
$\mathbf{F_{Y_B}}$	**0**	**0.48**	**0.84**	**0.96**	**1**	
f_{Y_C}	0.063	0.25	0.313	0.375	1
$\mathbf{F_{Y_C}}$	**0**	**0.063**	**0.313**	**0.626**	**1**	
f_{Y_D}	0.455	0.364	0.136	0.045	1
$\mathbf{F_{Y_D}}$	**0**	**0.455**	**0.819**	**0.955**	**1**	

Graphe des fonctions de répartition conditionnelles



Chapitre IV. Indices statistiques

Dans ce chapitre du cours, le cadre considéré sera celui de variables quantitatives. On pourra avoir une seule variable quantitative X mesurée sur un échantillon, ou un couple de variables (X, Y) où au moins l'une des deux variables est quantitative.

Les indices statistiques permettent de "décrire" les données en synthétisant l'information fournie par la distribution des observations. Ils se calculent uniquement sur des variables quantitatives. On considérera deux types d'indices:

- **Indices de localisation**: donnent une information sur un endroit spécifique de la distribution. Exemples d'indices de localisation: min, max, mode, moyenne, médiane, quartiles, etc...
- **Indices de dispersion**: servent à mesurer la dispersion d'une distribution, c'est-à-dire "la variation" des valeurs d'une variable. Exemples d'indices de dispersion: étendue, écart-type, variance, ...

1) Premiers indices de localisation et de dispersion

Nous présentons dans un premier temps quelques indices de localisation et de dispersion simples dont le calcul est direct:

- Pour les indices de localisation:
  - Le **minimum (min)** = la plus petit valeur observée dans l'échantillon
  - Le **maximum (max)** = la plus grande valeur observée dans l'échantillon
  - Le **mode** d'une distribution: pour le calcul du mode, on distingue 2 cas:

1. Cas d'une variable discrète: c'est la modalité (ou valeur) la plus observée (donc avec le plus grand effectif ou la plus grande fréquence). Sur un diagramme en bâtons, c'est la valeur où le "bâton est le plus haut".
2. Cas d'une variable continue: on introduit la classe modale qui est la classe de plus forte densité (la classe avec le plus "haut" rectangle dans l'histogramme). Le mode est alors défini comme le centre de la classe modale.

- Pour les indices de dispersion:

  L'**étendue**: l'étendue de la distribution est définie par:
  étendue = max - min.

2) Quantiles

Les quantiles sont des indices de localisation basés sur les rangs. Pour les calculer, on range les individus de l'échantillon par ordre croissant de la variable.

La **médiane** Me est la valeur observable qui partage en 2 parties d'effectif égal l'échantillon des individus rangés par ordre croissant de la variable, ou encore c'est la valeur observable telle qu'il y ait autant d'observations supérieures à cette valeur que d'observations inférieures.

Calcul de la médiane dans le cas d'une variable discrète ou de données brutes

Lorsque le nombre d'observations est impair $n = 2k + 1$ alors Me est la $(k + 1)^e$ observation de la série d'observations rangées par ordre croissant.
Lorsque le nombre d'observations est pair $n = 2k$ alors Me est le milieu observable de la k^e et $(k + 1)^e$ observation de la série d'observations

rangées par ordre croissant. Si le milieu n'est pas observable, on prend la valeur observable (inférieure) la plus proche.

Exemples:

1) Considérons la série de $n = 6$ observations issues d'une variable discrète (seuls les entiers sont observables): $1, 1, 2, 3, 5, 5$. La médiane est $Me = 2$ car c'est la valeur observable la plus proche de 2.5 qui est le milieu de 2 et 3.

2) Considérons maintenant la série de $n = 8$ observations issues d'une variable continue (tous les réels sont observables):
$4.35, 7.45, 9.5, 11.5, 14, 15.5, 17, 18.25$. La médiane est $Me = 12.75$ car c'est le milieu (observable) de 11.5 et 14.

3) **Exemple 3: Habitants d'une résidence de Montpellier**
Les observations rangées par ordre croissant sont: $0, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4$.
Comme $n = 11$, la médiane est donc la 6ème observation, c'est-à-dire $Me = 2$.

La médiane Me doit vérifier les 2 propriétés P_1 et P_2:

- Au moins 50% des individus ont une valeur inférieure ou égale à Me:

$$P_1 : \text{fréq}(\text{observations} \leq Me) \geq \frac{1}{2}$$

- Au moins 50% des individus ont une valeur supérieure ou égale à Me:

$$P_2 : \text{fréq}(\text{observations} \geq Me) \geq \frac{1}{2}$$

Remarque: pour une variable continue, on peut obtenir une valeur de la médiane à partir du graphe de la fonction de répartition (graphe des fréquences cumulées). En effet la médiane est la valeur Me vérifiant $F(Me) = 0.5$ où F est la fonction de répartition.

Généralisation à un quantile quelconque q_α d'ordre α

Un quantile d'ordre α ($0 \leq \alpha \leq 1$) est la valeur observable notée q_α telle que la proportion des individus de l'échantillon, ayant une valeur de X inférieure ou égale à q_α, est au moins égale à α et la proportion des individus, ayant une valeur de X supérieure ou égale à q_α, est au moins égale à $1 - \alpha$. Ceci est donc équivalent aux 2 propriétés P_1 et P_2 vérifiées par le quantile q_α:

- La proportion des individus qui ont une valeur inférieure ou égale à q_α est au moins égale à α:

$$P_1 : \text{fréq}(\text{observations} \leq q_\alpha) \geq \alpha$$

- La proportion des individus qui ont une valeur supérieure ou égale à q_α est au moins égale à $1 - \alpha$:

$$P_2 : \text{fréq}(\text{observations} \geq q_\alpha) \geq 1 - \alpha$$

Remarques:

- Pour calculer un quantile dans le cas d'une variable discrète (ou pour des données brutes issues d'une variable discrète ou continue), on utilise les 2 propriétés P_1 et P_2.
- Pour calculer un quantile dans le cas d'une variable continue où les observations sont réparties dans des classes, on utilisera des méthodes d'interpolation linéaire (voir page 60). Une valeur du quantile peut aussi être obtenue en utilisant le graphe de la fonction de répartition (q_α vérifie $F(q_\alpha) = \alpha$ où F est la fonction de répartition).

Le quantile d'ordre $\alpha = 0.5$, $q_{0.5}$ est la médiane. Les 3 quantiles d'ordre respectif $1/4$, $1/2$ et $3/4$ partagent en 4 effectifs égaux l'échantillon des individus classés par ordre croissant de la variable. On les appelle quartiles et on les note:

$$Q_1 := \text{1er quartile} = q_{0.25},$$

$$Q_2 := \text{2ème quartile} = \text{Me} = q_{0.5},$$

$$Q_3 := \text{3ème quartile} = q_{0.75}.$$

Parmi les autres quantiles couramment utilisés, on peut citer les déciles $D_1, D_2, ..., D_9$, qui partagent en dix parties d'effectif égal l'échantillon rangé par ordre croissant de la variable. On peut donc les réécrire $D_i = q_{i/10}$.

De même, les centiles $c_1, c_2, ..., c_{99}$, où $c_i = q_{i/100}$, partagent en 100 parties d'effectif égal l'échantillon rangé par ordre croissant de la variable.

3) Indices de localisation centrale et indices de dispersion: moyenne, médiane et variance

La moyenne et la médiane sont des indices de localisation centrale. Un indice de localisation centrale est une valeur qui est la plus proche possible de toutes les observations à la fois. C'est une sorte de "résumé" des observations. Pour définir ces 2 indices, nous allons introduire une notion de distance (en particulier pour pouvoir quantifier la "proximité" entre une valeur et des observations).

Soit une variable X observée sur un échantillon de n individus. Nous disposons donc des observations $x_1, x_2, \cdots, x_n$.

On appelle dispersion des observations de la variable X autour d'une valeur v la somme des distances de chaque observation à v:

$$\text{Disp}(v) = \sum_{i=1}^{n} d(x_i, v),$$

où pour la distance $d(x_i, v)$ entre x_i et v, on considérera soit la distance au sens des valeurs absolues $d(x_i, v) = |x_i - v|$, soit la distance au sens des carrés $d(x_i, v) = (x_i - v)^2$.

Cas d'une **variable discrète** où les observations sont regroupées par modalité (aussi appelée valeur):

Variable X	v_1	v_2	$\cdots$	v_K	Total
Effectifs n_k	n_1	n_2	$\cdots$	n_K	n

Le regroupement des observations par valeur (modalité) permet de réécrire la dispersion comme ceci:

$$\text{Disp}(v) = \sum_{k=1}^{K} n_k d(v_k, v).$$

Pour les deux distances on aura donc:

$$\text{Disp}_a(v) = \sum_{k=1}^{K} n_k |v_k - v| \qquad \text{et} \qquad \text{Disp}_c(v) = \sum_{k=1}^{K} n_k (v_k - v)^2.$$

Un indice de localisation centrale sera définie comme la valeur v qui rend la dispersion $\text{Disp}(v) = \sum_{i=1}^{n} d(x_i, v)$ minimale. Il dépend par conséquent de la distance choisie.

Si c'est la distance au sens des valeurs absolues $d(x_i, v) = |x_i - v|$, alors le minimum de la dispersion est obtenu pour $v = Me$, la médiane des observations.

Si c'est la distance au sens des carrés $d(x_i, v) = (x_i - v)^2$, alors le minimum de la dispersion est obtenu pour la moyenne des observations $v = \bar{x}$ où

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

La dispersion (au sens des carrés) des observations autour de la moyenne $\bar{x}$ est :

$$\text{Disp}_c(\bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

La dispersion (au sens des carrés) moyenne des observations autour de la moyenne s'appelle la **variance**:

$$\text{Var}(x) = \frac{1}{n}\text{Disp}_c(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

On a également:

$$\text{Var}(x) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\bar{x})^2.$$

La racine carrée de la variance est l'**écart-type**: $\sigma(x) = \sqrt{\text{Var}(x)}$.
La variance et l'écart-type sont des indices de dispersion. L'écart-type est un indice de même unité que les données.

La dispersion (au sens des valeurs absolues) des observations autour de la médiane Me est :

$$\text{Disp}_a(Me) = \sum_{i=1}^{n} |x_i - Me|,$$

aussi appelée écart absolu à la médiane.

L'écart absolu moyen à la médiane est donc:

$$\frac{1}{n}\text{Disp}_a(Me) = \frac{1}{n}\sum_{i=1}^{n} |x_i - Me|.$$

Calcul de $\bar{x}$ et Var(x) dans le cas d'une variable discrète où les observations sont regroupées par modalité (aussi appelée valeur):

Variable X	v_1	v_2	$\cdots$	v_K	Total
Effectifs n_k	n_1	n_2	$\cdots$	n_K	n

Le regroupement des observations par valeur permet d'écrire la moyenne comme ceci:

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{K} n_k v_k,$$

et la variance:

$$\text{Var}(x) = \frac{1}{n} \sum_{k=1}^{K} n_k (v_k - \bar{x})^2$$

ou

$$\text{Var}(x) = \frac{1}{n} \sum_{k=1}^{K} n_k v_k^2 - (\bar{x})^2.$$

Calcul de $\bar{x}$ et Var(x) dans le cas d'une variable continue où les observations sont réparties dans K classes (intervalles):

Variable X	$[b_0; b_1[$	$[b_1; b_2[$	$\cdots$	$[b_{K-1}; b_K[$	Total
Effectifs n_k	n_1	n_2	$\cdots$	n_K	n

Vu la perte d'information (voir chapitre I), la moyenne calculée ne peut être qu'une approximation. Son expression est la suivante:

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{K} n_k c_k$$

où c_k est le centre de la $k$ème classe $[b_{k-1}; b_k[$, c'est-à-dire $c_k = \frac{b_{k-1}+b_k}{2}$.
De même pour la variance, on a:

$$\text{Var}(x) = \frac{1}{n} \sum_{k=1}^{K} n_k (c_k - \bar{x})^2 \quad \text{ou} \quad \text{Var}(x) = \frac{1}{n} \sum_{k=1}^{K} n_k c_k^2 - (\bar{x})^2.$$

Remarques:

- La moyenne n'est presque jamais une valeur entière. Dans le cas d'une variable discrète elle ne sera que très rarement une valeur observable.
- Par contre, par définition, la médiane est une valeur observable. Donc dans le cas discret, c'est en général une valeur entière.
- Généralement on a l'ordre suivant :

  $$\text{Mode} \leq \text{Médiane} \leq \text{Moyenne}$$

  ou inversement.

  L'égalité de ces 3 indices traduit la symétrie de la distribution.

4) Moyennes et variances dans le cas d'un couple de variables (X, Y)

4.1) Moyennes conditionnelles et moyenne globale

Considérons un couple de variables (X, Y) observé sur un échantillon de taille n où K et K' sont les nombres de modalités (ou éventuellement d'intervalles pour une variable continue) de X et de Y. On suppose que les observations sont réparties dans un tableau de contingence (voir page 2) et on rappelle que n_{ij} est l'effectif associé au couple de modalités (m_i, m'_j). Supposons qu'au moins une des 2 variables est quantitative, par exemple Y. On sait (page 15) qu'il existe K distributions conditionnelles de Y (autant que de modalités de X) et 1 distribution (marginale) de Y. Vu que Y est quantitative, on peut donc calculer $K + 1$ moyennes pour les $K + 1$ distributions (les K distributions conditionnelles de Y_{m_i}, $(i = 1, ..., K)$, et la distribution de Y).

Le calcul des ces moyennes est identique au cas d'une seule variable.

Pour la distribution conditionnelle de Y_{m_i}, la moyenne est mesurée uniquement sur les individus du sous-échantillon $\{X = m_i\}$ (c'est-à-dire les individus de l'échantillon dont la réponse à X est m_i). **La moyenne conditionnelle de Y_{m_i}**, que l'on notera $\bar{y}_i$, est aussi appelée **moyenne du $i$ème groupe**. Reprenant les notations habituelles, on a alors:

$$\bar{y}_i = \frac{1}{n_{i\bullet}} \sum_{j=1}^{K'} n_{ij} m'_j$$

où les n_{ij}, $(j = 1, ..., K')$, sont les effectifs du sous-échantillon $\{X = m_i\}$ (effectifs de la distribution conditionnelle de Y_{m_i}) et donc $n_{i\bullet} = \sum_{j=1}^{K'} n_{ij}$ est le nombre d'individus du sous-échantillon $\{X = m_i\}$. Enfin, m'_j est la $j$ème modalité de Y (c'est une valeur numérique).

Pour la distribution de Y, la moyenne est mesurée sur tous les individus de l'échantillon. **La moyenne de Y** (notée simplement $\bar{y}$) est aussi appelée **moyenne globale de Y**. La formule est:

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{K'} n_{\bullet j} m'_j$$

où les $n_{\bullet j}$, $(j = 1, ..., K')$, sont les effectifs de la distribution (marginale) de Y (ou effectifs marginaux de Y).

Remarque: Notons que les formules ci-dessus sont valables que Y soit discrète ou continue. En effet dans le cas où Y est continue, il suffit de remplacer dans les formules, la modalité m'_j (aussi appelée valeur v_j dans le cas d'une variable discrète) par le centre c_j de la $j$ème classe.

4.2) Décomposition de la moyenne

Il est possible de calculer la **moyenne globale de Y** (moyenne de Y mesurée sur tout l'échantillon) à partir des moyennes conditionnelles. Une erreur fréquente consiste à calculer la moyenne globale en faisant la moyenne arithmétique des moyennes conditionnelles. En réalité, la moyenne globale de Y est la moyenne des moyennes des sous-échantillons (ou moyennes conditionnelles) pondérée par les effectifs associés (donnés par la distribution de X). La formule est la suivante:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{K} n_{i\bullet} \bar{y}_i$$

où les $\bar{y}_i$, $(i = 1, ..., K)$, sont les K moyennes conditionnelles et $n_{i\bullet}$, $(i = 1, ..., K)$, les effectifs associés aux K sous-échantillons (ce sont les effectifs marginaux de X).

Cette formule est appelée formule de **décomposition de la moyenne.**

4.3) Variances conditionnelles et variance globale

On peut également calculer les variances des distributions conditionnelles et de la distribution marginale. La **variance conditionnelle de Y_{m_i}** (ou **variance du $i$ème groupe**) est:

$$\text{Var}(y)_i = \frac{1}{n_{i\bullet}} \sum_{j=1}^{K'} n_{ij}(m'_j - \bar{y}_i)^2 = \frac{1}{n_{i\bullet}} \sum_{j=1}^{K'} n_{ij}(m'_j)^2 - (\bar{y}_i)^2.$$

La **variance de Y** (ou **variance globale de Y**) est:

$$\text{Var}(y) = \frac{1}{n} \sum_{j=1}^{K'} n_{\bullet j}(m'_j - \bar{y})^2 = \frac{1}{n} \sum_{j=1}^{K'} n_{\bullet j}(m'_j)^2 - (\bar{y})^2.$$

Remarque: Notons que les formules ci-dessus sont valables que Y soit discrète ou continue. En effet dans le cas où Y est continue, il suffit de remplacer dans les formules, la modalité m'_j par c_j (centre de classe).

Exemple 1: Notes d'étudiants et filières d'étude

Rappelons que c_j est le centre de la $j$ème classe. Par exemple, le centre de la 2ème classe est $c_2 = (6 + 10)/2 = 8$.

Centres c_j	3	8	12	17	
Classes Y	$[0, 6[$	$[6, 10[$	$[10, 14[$	$[14, 20]$	Total X
Y_A	26	6	4	1	**37**
Y_B	12	9	3	1	**25**
Y_C	1	4	5	6	**16**
Y_D	10	8	3	1	**22**
Y	**49**	**27**	**15**	**9**	n=100

Calcul des moyennes conditionnelles $\bar{y}_i$ et de la moyenne globale $\bar{y}$:

$$\bar{y}_1 = \frac{26 \times 3 + 6 \times 8 + 4 \times 12 + 1 \times 17}{37} = 5.162$$

$$\bar{y}_2 = \frac{12 \times 3 + 9 \times 8 + 3 \times 12 + 1 \times 17}{25} = 6.44$$

$$\bar{y}_3 = \frac{1 \times 3 + 4 \times 8 + 5 \times 12 + 6 \times 17}{16} = 12.313$$

$$\bar{y}_4 = \frac{10 \times 3 + 8 \times 8 + 3 \times 12 + 1 \times 17}{22} = 6.682$$

$$\bar{y} = \frac{49 \times 3 + 27 \times 8 + 15 \times 12 + 9 \times 17}{100} = 6.96$$

Calcul des variances conditionnelles et de la variance globale:

$$\sigma^2(y)_1 = \frac{26\times(3-5.162)^2+6\times(8-5.162)^2+4\times(12-5.162)^2+1\times(17-5.162)^2}{37}$$

$$= 13.433$$

$$\sigma^2(y)_2 = \frac{12\times(3-6.44)^2+9\times(8-6.44)^2+3\times(12-6.44)^2+1\times(17-6.44)^2}{25}$$

$$= 14.726$$

$$\sigma^2(y)_3 = \frac{(3-12.313)^2+4\times(8-12.313)^2+5\times(12-12.313)^2+6\times(17-12.313)^2}{16}$$

$$= 18.34$$

$$\sigma^2(y)_4 = \frac{10\times(3-6.682)^2+8\times(8-6.682)^2+3\times(12-6.682)^2+(17-6.682)^2}{22}$$

$$= 15.49$$

$$\sigma^2(y) = \frac{49\times(3-6.96)^2+27\times(8-6.96)^2+15\times(12-6.96)^2+9\times(17-6.96)^2}{100}$$

$$= 20.858$$

Remarques: Dans tout ce paragraphe 4, on a supposé que Y était quantitative (X pouvant être soit quantitative soit qualitative) et on a mesuré les moyennes et les variances de Y dans chacun des sous-échantillons (groupes) induits par les modalités de X (les moyennes et variances conditionnelles de Y). Il est bien sûr possible de faire les mêmes calculs pour les distributions conditionnelles de X, à condition que X soit quantitative. On mesure alors les moyennes (et variances) de X dans chacun des sous-échantillons induits par les modalités de Y (ce sont les moyennes et les variances conditionnelles de X).

Le calcul des autres indices (mode, quartiles, médiane,...) est similaire au calcul des moyennes et des variances, c'est-à-dire qu'on se restreint aux sous-échantillons pour les distributions conditionnelles et on considère l'échantillon en entier pour la distribution marginale.

5) Comparaison des distributions conditionnelles

La comparaison des distributions conditionnelles Y_{m_i}, $(i = 1, \cdots, K)$, lorsque Y est quantitative, permet en général de classer les sous-échantillons (les groupes) correspondants par ordre croissant de la variable Y.

Lorsque Y est discrète, on peut comparer les distributions conditionnelles par leur représentation simultanée en bâtons sur un même graphique; si Y est continue, la représentation simultanée des histogrammes est difficile à lire, et il est préférable de les comparer par d'autres moyens, par exemple en utilisant les fonctions de répartition. Dans les deux cas, on peut en général classer (partiellement ou totalement) les sous-échantillons en comparant leur moyenne (moyenne conditionnelle) ou leur boxplot (résumé graphique de la distribution).

5.1) Classement par moyenne

Comme la **moyenne** est un indice de localisation centrale, un classement des moyennes conditionnelles conduit naturellement à classer les sous-échantillons (les groupes).

On rappelle que les moyennes conditionnelles de la variable Y (ou moyennes des groupes) sont les moyennes des distributions conditionnelles de Y, c'est-à-dire les moyennes de Y mesurées sur les sous-échantillons $\{X = m_i\}$, $i = 1, ..., K$.

Exemple 1: Notes d'étudiants et filières d'étude

Dans l'exemple 1 (page 52), on a le classement suivant des moyennes conditionnelles:

$$\bar{y}_3 > \bar{y}_4 > \bar{y}_2 > \bar{y}_1.$$

Ce qui peut s'interpréter comme ceci: la note obtenue au concours par les étudiants de la filière C est globalement supérieure à celle obtenue par les étudiants de la filière D, qui eux mêmes ont une note globalement supérieure à celle obtenue par les étudiants de la filière B.

5.2) Classement par fonction de répartition (cas continu)

Définition: La **fonction de répartition** F d'une variable Y est une fonction mathématique qui prend en entrée n'importe quel nombre réel r, et retourne comme valeur notée $F(r)$ la proportion des individus de l'échantillon dont la valeur de Y est inférieure ou égale à r; $F(r)$ est donc un nombre compris entre 0 et 1 (ou entre 0 et 100 en pourcentage).

Exemple 1: Notes d'étudiants et filières d'étude

On reprend l'exemple 1 (où la variable Y = "Note" est continue) pour classer les distributions conditionnelles par fonction de répartition.
D'après le graphe page 27, on a le classement suivant pour les fonctions de répartition (conditionnelles): $F_{Y_C} < F_{Y_D} < F_{Y_B} < F_{Y_A}$.

Par conséquent, la distribution conditionnelle de Y_C est globalement supérieure à la distribution conditionnelle de Y_D, qui elle même est globalement supérieure à la distribution conditionnelle de Y_B (attention l'ordre est inversé).

Ceci peut s'interpréter de la manière suivante: la note obtenue au concours par les étudiants de la filière C est globalement supérieure à celle obtenue par les étudiants de la filière D, qui eux mêmes ont une note globalement supérieure à celle obtenue par les étudiants de la filière B.

5.3) Classement par boxplot

Définition: Le **boxplot** (ou boîte à pattes) d'une distribution est un résumé graphique de localisation et de dispersion de la distribution: on trace un axe gradué limité par le minimum et le maximum des observations. Ensuite, on dessine une boîte dont les 2 extrêmités se situent au niveau de Q_1 (1er quartile) et de Q_3 (3ème quartile), et on trace un trait coupant la boîte au niveau de la médiane Me.

Le boxplot étant un bon résumé graphique d'une distribution, la comparaison des distributions conditionnelles par boxplot peut être instructive: plus un boxplot est situé à droite sur l'axe gradué, plus sa distribution est grande. Donc comme pour les fonctions de répartition, on peut comparer les boxplots des distributions conditionnelles, et classer (partiellement ou totalement) les sous-échantillons par ordre de grandeur globale de la variable Y.

Remarque: l'intervalle $[Q_1; Q_3]$, appelé **intervalle interquartile**, est un indice de dispersion. Cet intervalle comprend 50% des observations et est centré autour de la médiane (la quantité $Q_3 - Q_1$ est appelée l'écart interquartile).

Calcul des quantiles pour une variable continue

Pour une variable quantitative continue où les observations sont réparties dans des classes (intervalles), on utilise des méthodes d'interpolation linéaire pour trouver la valeur de la médiane ou de n'importe quel quantile.

Pour un quantile quelconque d'ordre α, q_α, la méthode est la suivante:

1. Localiser la classe où se trouve le quantile q_α: q_α appartient à la classe $[b_{i-1}; b_i[$ où les bornes b_{i-1} et b_i vérifient: $F(b_{i-1}) < \alpha < F(b_i)$.

2. Le calcul de q_α utilise alors la formule suivante:

$$q_\alpha = b_{i-1} + (b_i - b_{i-1}) \times \frac{\alpha - F(b_{i-1})}{F(b_i) - F(b_{i-1})}.$$

   Comme $b_i - b_{i-1} = a_i$ et $F(b_i) - F(b_{i-1}) = F_i - F_{i-1} = f_i$, on peut réécrire

$$q_\alpha = b_{i-1} + a_i \times \frac{\alpha - F(b_{i-1})}{f_i}.$$

Pour calculer le 1er quartile, la médiane ou le 3ème quartile, il suffit donc de remplacer α par 0.25, 0.5 ou 0.75 dans la méthode ci-dessus.

Exemple 1: Notes d'étudiants et filières d'étude

Nous traçons les boxplots pour comparer les 4 sous-échantillons dans l'exemple 1. Pour cela, nous devons calculer les 3 quartiles (Q_1, Me et Q_3) dans chacun des 4 sous-échantillons (A, B, C et D).

Nous rappelons dans le tableau ci-dessous les fréquences cumulées (et les fréquences) pour les 4 distributions conditionnelles:

Note	$[0,6[$	$[6,10[$	$[10,14[$	$[14,20]$	Total
Bornes	0	6	10	14	20	
f_A	0.703	0.162	0.108	0.027	1
$\mathbf{F_A}$	**0**	**0.703**	**0.865**	**0.973**	**1**	
f_B	0.48	0.36	0.12	0.04	1
$\mathbf{F_B}$	**0**	**0.48**	**0.84**	**0.96**	**1**	
f_C	0.063	0.25	0.313	0.375	1
$\mathbf{F_C}$	**0**	**0.063**	**0.313**	**0.626**	**1**	
f_D	0.455	0.364	0.136	0.045	1
$\mathbf{F_D}$	**0**	**0.455**	**0.819**	**0.955**	**1**	

Pour le 1er sous-échantillon (groupe) A (étudiants de la filière A) on a:
$Q_1 \in [0, 6[$ car $F_A(0) = 0 < 0.25 < 0.703 = F_A(6)$ donc
$Q_1 = 0 + (6 - 0) \times \frac{0.25-0}{0.703-0} = 2.13$
$Me \in [0, 6[$ car $F_A(0) = 0 < 0.5 < 0.703 = F_A(6)$ donc
$Me = 0 + (6 - 0) \times \frac{0.5-0}{0.703-0} = 4.27$
$Q_3 \in [6, 10[$ car $F_A(6) = 0.703 < 0.75 < 0.865 = F_A(10)$ donc
$Q_3 = 6 + (10 - 6) \times \frac{0.75-0.703}{0.865-0.703} = 7.16$

Pour le 2ème sous-échantillon (groupe) B (étudiants de la filière B), on trouve:
$Q_1 = 3.125$, $Me = 6.22$ et $Q_3 = 9$

Pour le 3ème sous-échantillon (groupe) C (étudiants de la filière C), on trouve:
$Q_1 = 8.99$, $Me = 12.39$ et $Q_3 = 15.989$

Pour le 4ème sous-échantillon (groupe) D (étudiants de la filière D), on trouve:
$Q_1 = 3.3$, $Me = 6.49$ et $Q_3 = 9.24$

Les tracés des boxplots (1 par distribution conditionnelle) sont présentés à la page suivante. Ils confirment le classement des distributions conditionnelles et donc des sous-échantillons obtenu par les deux méthodes précédentes (5.1 et 5.2).

Boxplots des 4 distributions conditionnelles