Année universitaire 2012-2013

Université de Toulouse Le Mirail

LICENCE 1 DE PSYCHOLOGIE

PY0106X - Statistique Descriptive

Exercices de Statistique

Frédéric Ferraty

Table des matières

Chapitre 1

De l'enquête aux données

1.1 Introduction

Il s'agit d'une enquête intitulée "Attachement au quartier" menée en 2010 par des étudiants inscrits en deuxième année de Psychologie. L'objectif est d'étudier, parmi une population plutôt jeune (de 14 ans à 35 ans), l'attachement ou le "désattachement" que les habitants vouent à leur lieu de vie. 189 personnes ont participé à cette enquête ; elles ont été soumises à un questionnaire permettant de mesurer plus d'une vingtaine de caractéristiques. Les données recueillies serviront de fil conducteur pour illustrer les différentes notions abordées dans ce cours de statistique descriptive.

1.2 Présentation du questionnaire

1- Age : …………..

2- Sexe : ☐ Homme ☐ Femme

3- Quel est le pays d'origine de votre père : né en France ☐ né à l'étranger ☐

Quel est le pays d'origine de votre mère : née en France ☐ née à l'étranger ☐

Quel est votre pays de naissance: né en France ☐ né à l'étranger ☐

4- Quel est votre lieu d'habitation ?
Nom de votre ville : ………………………………………………….
Habitez vous : en centre ville ☐ en banlieue ☐
dans un village ☐ dans une cité ☐
Autres : ☐ …………………………………

5- Votre type de logement actuel :
☐ chambre universitaire
☐ T1
☐ T2
☐ T3 et +

6-Votre mode de logement actuel :
☐ en cité Universitaire
☐ en HLM
☐ dans une résidence
☐ dans une maison
☐ autres :……………………

7- Depuis combien de temps vivez-vous dans ce logement ? ……………………………

8. Vous êtes : Locataire ☐ propriétaire ☐

9- Dans ce logement vous vivez :
☐ seul
☐ en couple
☐ en famille (combien êtes vous ?......................................)
☐ chez vos parents (combien êtes vous ?......................................)
☐ en colocation (combien êtes vous ?......................................)

10- Dans votre quartier :
- Avez-vous de la famille qui habite dans le quartier ?
Oui ☐ Non ☐
Si oui combien ? ……………………………………………..

- Avez-vous des amis qui habitent dans le quartier ?
Oui ☐ Non ☐
Si oui combien ? ……………………………………………..

11 Fréquentez vous des voisins dans votre quartier ?
Oui ☐ Non ☐ Si oui combien ? ……………………………………………..
Quelle est la nature de vos relations avec vos voisins ? cochez la réponse de votre choix.

Très mauvaise	mauvaise	bonne	Très bonne

12- Dans votre quartier y a-t'il des commerces?
un marché ☐ une boucherie ☐
une boulangerie ☐ un supermarché ☐
une pharmacie ☐ une librairie ☐ autres ☐……………………………

13- Dans votre quartier y a-t'il des lieux culturels ?
théâtre ☐ cinéma ☐ associations ☐ bibliothèque ☐
musée ☐ autres ☐…………………………………

14- Dans votre quartier y a-t'il des lieux de sortie ?
fast-food ☐ restaurant ☐ boite de nuit ☐ bar ☐ autres ☐………………

15- Dans votre quartier y a-t'il des lieux de formation ?
école☐ collège☐ lycée☐ université ☐ formation pour adulte ☐ autres ☐………………

16- <u>Dans votre quartier,</u> dans quels lieux vous rendez-vous le plus souvent ?
(codez de 1 très souvent à 4 rarement)
_______ les commerces
_______ les lieux culturels
_______ les lieux de sortie
_______ les lieux de formation

17- Cochez la case qui correspond à votre réponse :

Que représente votre quartier pour vous ?	**Tout à fait en désaccord**				**Tout à fait d'accord**
1 Pour y vivre, c'est le quartier idéal.	**1**	**2**	**3**	**4**	**5**
2 Ce quartier fait partie de moi-même.	**1**	**2**	**3**	**4**	**5**
3 Je suis très attaché(e) à certains endroits de ce quartier.	**1**	**2**	**3**	**4**	**5**
4 Il me serait très difficile de quitter définitivement ce quartier.	**1**	**2**	**3**	**4**	**5**
5 Je pourrais facilement quitter ce quartier.	**1**	**2**	**3**	**4**	**5**
6 Je n'aimerais pas à avoir à quitter ce quartier pour un autre.	**1**	**2**	**3**	**4**	**5**

1.3 Codage des réponses

Une étape importante dans le dépouillement des données consiste à les saisir informatiquement afin de pouvoir les traiter. Cette phase exige que l'on soit capable de codifier les réponses données par les individus interrogés. Ci-dessous, vous trouverez quelques exemples de codages utilisés pour les dix premières questions :

variables	**Codage variable** (8 caractères maxi)	**modalités**	**code**
1-Age	age	Rentrer l'âge tel quel	Mettre l'âge
2-Sexe	sexe	Homme Femme	1 2
3-Pays d'origine du père	ppere	Né en France Né hors de France	1 2
3bis- Pays d'origine de la mère	pmere	Née en France Née hors de France	1 2
3bisbis-Votre pays de naissance	pnaiss	Né en France Né hors de France	1 2
4- Nom de votre ville	nomville		Ecrire le nom de la ville
4bis Lieu d'habitation	lieuhab	- En centre ville - En banlieue - Dans un village - En cité - Autres	1 2 3 4 5
5-Type de logement	typlog	- chambre universitaire - T1 - T2 - T3 et +	1 2 3 4
6-Mode de logement actuel :	modlog	- en cité U - en HLM - dans une résidence - dans une maison - autres	1 2 3 4 5
7-Temps vécu dans ce logement	durelog		Ecrire le temps 1an 6 mois s'écrit 1,5

1.4 Données brutes

L'emploi des codages nous permet de construire un tableau contenant les données brutes suivant :

age	sexe	ppere	pmere	pnaiss	nomville	lieuhab	typlog	modlog	durelog	···
16	2	1	2	1	toulouse	1	4	4	16	···
33	2	1	1	1	toulouse	2	4	3	2	···
19	2	1	1	1	toulouse	1	2	3	1	···
34	1	1	1	1	toulouse	2	4	3	4	···
19	1	1	2	1	toulouse	1	3	3	1	···
20	2	1	1	1	toulouse	1	2	3	4	···
20	1	1	1	1	colomiers	2	4	4	13	···
21	1	2	2	2	toulouse	5	0	3	0.17	···
20	2	1	1	1	labarthe/lèze	2	4	4	0.5	···
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Chaque ligne correspond à un individu ; chaque colonne correspond à une caractéristique mesurée sur tous les individus. Par exemple, le premier individu composant ce tableau a 16 ans ; c'est une étudiante (sexe=2) dont le père est né en France (ppere=1) et la mère à l'étranger (pmere=2). Cette étudiante est née en France (pnaiss=1), habite le centre ville (lieuhab=1) de Toulouse (nomville=toulouse) dans une maison (modlog=4) de type "T3 ou +" (typlog=4) depuis 16 ans (durelog=16).

Remarque 1.1 *Le codage peut parfois être la source de confusions. En effet, le sexe est codé 1 pour un homme et 2 pour une femme. Il est clair qu'il s'agit là d'un codage arbitraire qui ne représente en aucun cas une quantité. Il en est de même pour "ppere", "pmere", "pnaiss", "lieuhab", "typlog" et "modlog". Réaliser des opérations arithmétiques (addition, multiplication,...) sur ces codes n'aurait aucun sens. En revanche, les valeurs présentes dans la colonne "age" ou "durelog" représentent des quantités avec lesquelles on peut effectuer des opérations arithmétiques comme par exemple une moyenne.*

Le tableau des données brutes possède systématiquement autant de lignes que d'individus (soit ici 189) et autant de colonnes que de caractéristiques mesurées. Le nombre d'individus participant à cette enquête ainsi que le nombre de caractéristiques mesurées étant importants, nous avons donné uniquement les premières lignes et colonnes, les points de suspension symbolisant les données restantes.

1.5 Statistique descriptive : définition

On appelle **statistique descriptive** l'ensemble des méthodes permettant d'organiser, de présenter, de décrire et de synthétiser l'information recueillie. Ces techniques s'appuient sur des outils graphiques et analytiques. L'objectif de ce cours est de familiariser le lecteur avec différents outils statistiques, à la fois simples et pertinents.

Chapitre 2

Exercices relatifs à la partie I : Statistique descriptive univariée

2.1 Vocabulaire de base et mise en forme des données brutes

Exercice 1 *Préciser l'ensemble des modalités possibles ainsi que la nature des variables suivantes (lorsqu'il s'agit de variables directement extraites de l'enquête, le numéro permet de les localiser dans le questionnaire) :*
"Pays de naissance" (n°3) ; "Temps vécu dans votre logement actuel" ; "Mode de logement" (n°6) ; "Nombre de personnes vivant avec vous" ; "Niveau de difficulté à quitter définitivement ce quartier ?" (n°17 item 4).

Exercice 2 *Pour la variable "Type de logement", on a extrait les observations suivantes :*

$$4\ ,2\ ,4\ ,4\ ,2\ ,3\ ,4\ ,1\ ,2\ ,4\ ,3\ ,3\ ,4\ ,4\ ,4\ ,2\ ,1\ ,3\ ,2\ ,4.$$

où "1" = "chambre universitaire", "2" = "T1", "3" = "T2" et "4" = "T3 et +". Quelle est le type de cette variable ? Dresser le tableau donnant les effectifs, fréquences et pourcentages ; quelle est la taille de l'échantillon ?

Exercice 3 *Pour la variable "Age" (exprimé en années), on a extrait les observations suivantes :*

$$21, 20, 19, 18, 20, 20, 19, 22, 18, 19, 20, 21, 16, 20, 25, 19, 22, 23, 20, 19, 20, 24, 16, 21, 18.$$

Que signifie par exemple l'observation 18 ? Quelle est la nature de la variable "Age" ? Dresser le tableau des effectifs, fréquences et pourcentages en regroupant les modalités selon les classes suivantes : $[16; 19[$, $[19; 20[$, $[20; 21[$, $[21; 25]$.

Exercice 4 *Voici le tableau des pourcentages obtenu pour la variable "Mode de logement" :*

x_i	%
"Cité U"	*4.8*
"HLM"	*16.4*
"Résidence"	*38.6*
"Maison"	*28.6*
"Autre"	*11.6*
TOTAL	*100*

Sachant que la taille de l'échantillon $N = 189$*, retrouver les effectifs pour chaque modalité.*

2.2 Représentation des variables qualitatives

Exercice 5 *On s'intéresse à la variable* X *= "Nature du lieu d'habitation" pour laquelle on a observé les effectifs suivants :*

x_i	*Centre ville*	*Banlieue*	*Village*	*Cité*	*Autre*
n_i	*87*	*30*	*32*	*30*	*10*

Quel est le type de cette variable ? Quel est son mode ? Représenter le diagramme en barres des fréquences ainsi que le diagramme unicolonne des fréquences.

Exercice 6 *On s'intéresse à la variable* X *= "Locataire/Propriétaire" pour laquelle on a observé les effectifs suivants :*

x_i	*Locataire*	*Propriétaire*
n_i	*134*	*55*

.

Représenter le diagramme en secteurs.

2.3 Médiane et quartiles

Exercice 7 *On considère la variable "attaché à certains endroits" dont les modalités (qui s'échelonnent de '1'='tout à fait en désaccord' à '5'='tout à fait d'accord')* *mesurent le niveau d'adéquation avec l'affirmation "Je suis très attaché(e) à certains endroits de ce quartier". On a les deux tableaux d'effectifs suivants :*

Hommes

x_i	1	2	3	4	5
n_i	14	27	17	19	15

Femmes

x_i	1	2	3	4	5
n_i	21	23	18	14	21

1. *Quel est le type de cette variable ?*
2. *Représenter la distribution de cette variable pour les hommes d'une part, puis pour les femmes d'autre part.*
3. *Déterminer la médiane pour chacun de ces deux tableaux ; y-a-t-il une différence pour ces deux groupes (Femmes/Hommes) du point de vue de la médiane ?*

Exercice 8 *On considère la variable "Temps vécu dans le logement" pour laquelle on a obtenu le tableau d'effectifs suivants :*

x_i	[0 ;1[	[1 ;2[	[2 ;3[	[3 ;5[	[5 ;11[	[11 ;16[	[16 ;21[	[21 ;26]
n_i	35	36	32	25	20	18	16	7

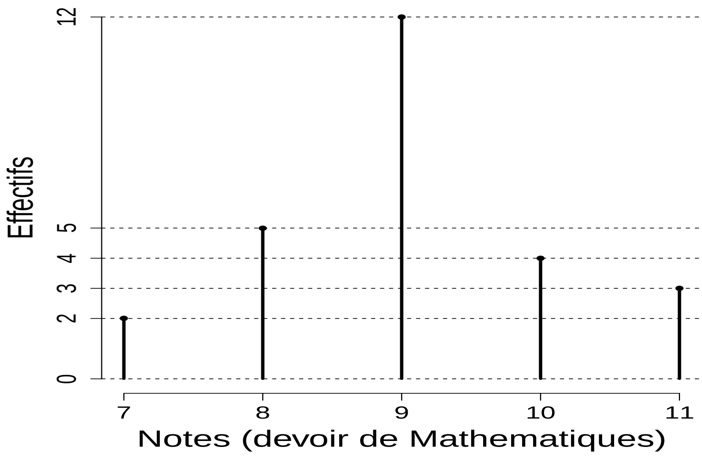
1. *Quel est le type de cette variable ?*
2. *Déterminer la médiane ainsi que les 1er et 3ème quartiles ; interpréter ces différents indices de position.*
3. *A cause d'une erreur de saisie, la borne supérieure 26 a été remplacée par 66 ; cela a-t-il un impact sur la détermination de la médiane ?*

2.4 Variable quantitative : mode, moyenne, variance et représentation

Exercice 9 *Les notes (variable X) obtenues par une classe d'élèves de 5ème lors d'un devoir de Français fournissent le tableau suivant :*

x_i	n_i	$n_i \times x_i$	$n_i \times (x_i)^2$	N_i *(eff. cum.)*
4	*2*			
5	*3*			
6	*5*			
7	*3*			
8	*2*			
9	*2*			
10	*4*			
11	*4*			
12	*3*			
14	*2*			
TOTAL			**2432**	

1) Préciser la variable étudiée ainsi que son type.
2) Compléter le tableau ci-dessus.
3) Réaliser le diagramme en bâtons représentant la distribution de X.
4) Calculer la moyenne et la variance de X.
5) Déterminer la valeur de la modalité qui permet de séparer l'échantillon en 2 sous-échantillons de même taille.
6) La figure ci-après représente la distribution des notes obtenues par la même classe lors d'un devoir de Mathématiques (variable Y) :



6.a) À partir des représentations des distributions de X et Y, sans faire de cal-

cul, quelle est d'après vous la variable de plus petite variance ? Pour laquelle des 2 variables la moyenne est-elle la plus représentative ?

6.b) Déduisez du graphique la valeur de la médiane

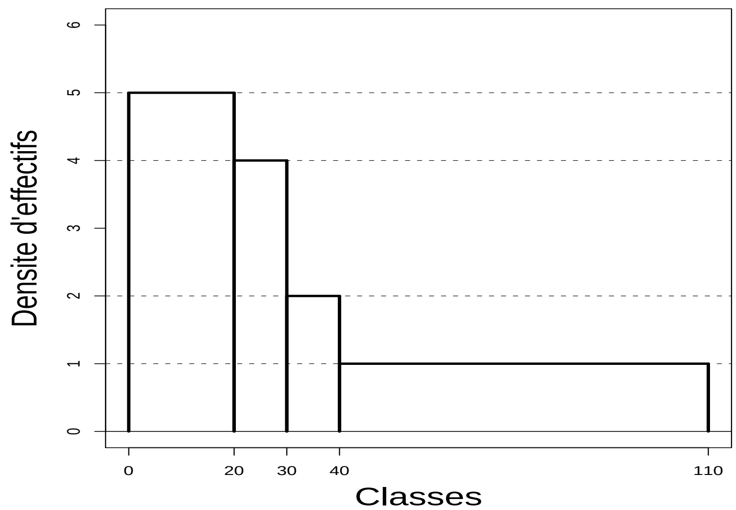
6.c) À partir de la figure représentant la distribution de Y, vérifier par le calcul que la moyenne de $Y \simeq 9.04$

Exercice 10 *On considère la variable X "Temps vécu dans le logement" pour laquelle on a obtenu le tableau d'effectifs suivants :*

x_i	[0 ;1[	[1 ;2[	[2 ;3[	[3 ;5[	[5 ;11[	[11 ;16[	[16 ;21[	[21 ;26]
n_i	35	36	32	25	20	18	16	7

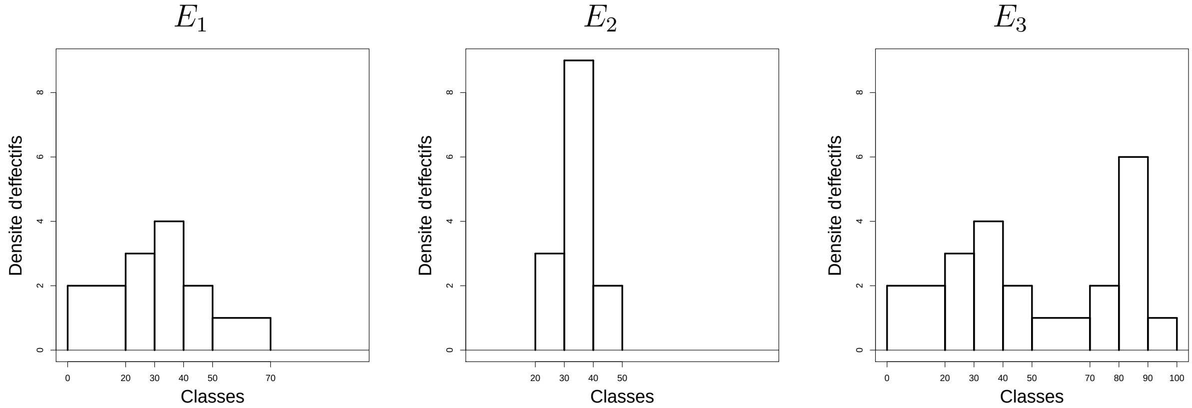
1. *Représenter l'histogramme de X ; peut-on dire que la moyenne sera un résumé pertinent pour cette distribution ?*
2. *Déterminer le(s) mode(s).*
3. *Calculer la moyenne $\overline{x}$.*
4. *Calculer la médiane de X et comparer avec la moyenne.*
5. *Déterminer $Var(X)$ puis σ_X.*

Exercice 11 *On considère la variable X "Temps de trajet domicile/travail" pour laquelle on a obtenu l'histogramme suivant :*



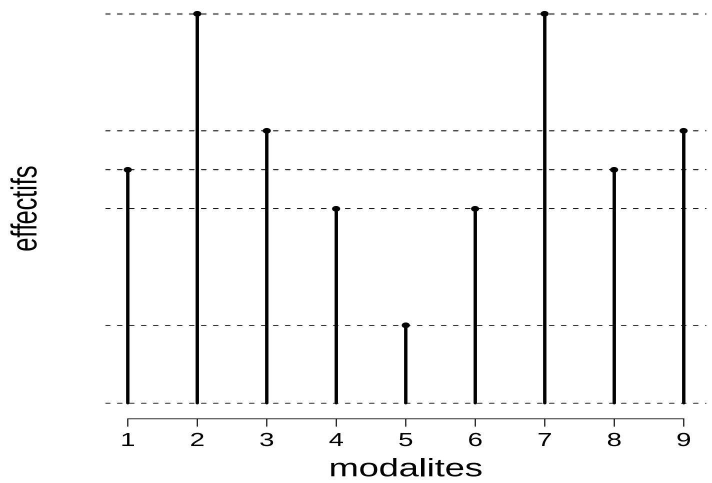
1. *Que pouvez-vous dire sur l'allure générale de la distribution ainsi représentée ?*
2. *A votre avis et sans faire de calculs, la moyenne est-t-elle proche de la médiane ?*
3. *A partir de cet histogramme, dresser le tableau des effectifs.*
4. *Calculer la moyenne puis déterminer la médiane.*

Exercice 12 *On dispose de 3 échantillons (E_1, E_2 et E_3) pour lesquels on a observé une même variable. On a obtenu pour chaque échantillon les 3 histogrammes suivants :*



Soient σ_1^2 (resp. σ_2^2 et σ_3^2) la variance correspondant à E_1 (resp. E_2 et E_3). Sans faire de calculs, pouvez-vous ranger par ordre croissant (de la plus petite à la plus grande) ces 3 quantités σ_1^2, σ_2^2 et σ_3^2 en expliquant concisément pourquoi ?

Exercice 13 *Soit X le nombre de fautes réalisé lors d'une dictée. A partir d'un échantillon d'élève, on obtient le diagramme en bâtons suivant pour lequel on a omis de préciser les effectifs :*



1. *Commenter ce diagamme en bâtons puis déterminer la médiane.*
2. *Sachant que les effectifs manquants sur le graphique sont 2, 5, 6, 7 et 10, calculer la moyenne puis l'écart-type.*

Chapitre 3

Exercices relatifs à la partie II : Statistique descriptive bivariée

3.1 Distribution conjointe, marginale, conditionnelle

Exercice 14 *Le tableau ci-après fournit les effectifs conjoints de la distribution conjointe des deux variables X = "Pays de naissance de la mère" et Y = "Pays de naissance du père" :*

Y (père) \ X (mère)	né en France	né à l'étranger
né en France	129	17
né à l'étranger	13	30

1. *Préciser la nature des variables étudiées.*
2. *Représenter la distribution conjointe de (X, Y).*
3. *Compléter le tableau des effectifs conjoints en donnant les lois marginales (i.e. marges de X et Y) ; représenter la distribution marginale de Y.*
4. *Représenter la distribution de X conditionnellement à Y. Que peut-on dire à partir de ce graphique concernant la relation entre X et Y ?*

Exercice 15 *Le tableau ci-après fournit les effectifs conjoints de la distribution conjointe des deux variables X = "Locataire/Propriétaire" et Y = "attachement à certains endroits" mesurant le niveau d'adéquation (de "niv. 1"="tout à fait en désaccord" à "niv. 5"="tout à fait en accord") avec l'affirmation* "Je suis très attaché(e) à certains endroits de ce quartier" *:*

X \ Y	*niv. 1*	*niv. 2*	*niv. 3*	*niv. 4*	*niv. 5*
x_1 = "locataire"	22	45	25	20	22
x_2 = "propriétaire	13	5	10	13	14

1. *Préciser la nature des variables étudiées.*
2. *Représenter la distribution conjointe de (X, Y).*
3. *Compléter le tableau des effectifs conjoints en donnant les lois marginales ; représenter la distribution marginale de Y.*
4. *Représenter la distribution de X conditionnellement à Y. Que peut-on dire à partir de ce graphique concernant la relation entre X et Y ?*

Exercice 16 *On a interrogé une partie des élèves d'un collège pour connaître la distance regroupée selon trois catégories (courte, moyenne et longue) qu'ils doivent parcourir pour se rendre à l'établissement scolaire (i.e. distance domicile/collège). On s'intéresse de plus à la variable Y = "niveau scolaire". L'objectif est d'étudier l'éventuel impact de la distance domicile/collège sur les résultats scolaires. On obtient ainsi le tableau suivant :*

X \ Y	*faible*	*moyen*	*élevé*	**Marge de *X***
courte	23		79	**127**
moyenne		85		**223**
longue	102		27	
Marge de *Y*		**131**	**161**	**N=500**

1. *Préciser les variables étudiées ainsi que leur type. Quelle est la population étudiée ? Quelle est la taille de l'échantillon ?*
2. *Compléter le tableau ci-contre.*
3. *Représenter la loi marginale de X.*
4. *Déterminer la distribution de Y conditionnellement à X. Représenter graphiquement cette distribution. Que pouvez-vous dire concernant le lien entre X et Y ?*

Exercice 17 *On dispose de deux variables X et Y pour lesquelles on a obtenu le graphique représenté à la figure 3.1 (l'axe vertical est exprimé en %).*

1. *S'agit-il de la distribution : conjointe de (X,Y) ? marginale de X ? marginale de Y ? de X conditionnellement à Y ? de Y conditionnellement à X ?*
2. *Construire le tableau correspondant à ce graphique.*
3. *Sachant que la distribution marginale de X est donnée par $n_{1\bullet} = 50$ et $n_{2\bullet} = 60$, dresser la table de contingence correspondante en explicitant les marges (i.e. lois marginales).*

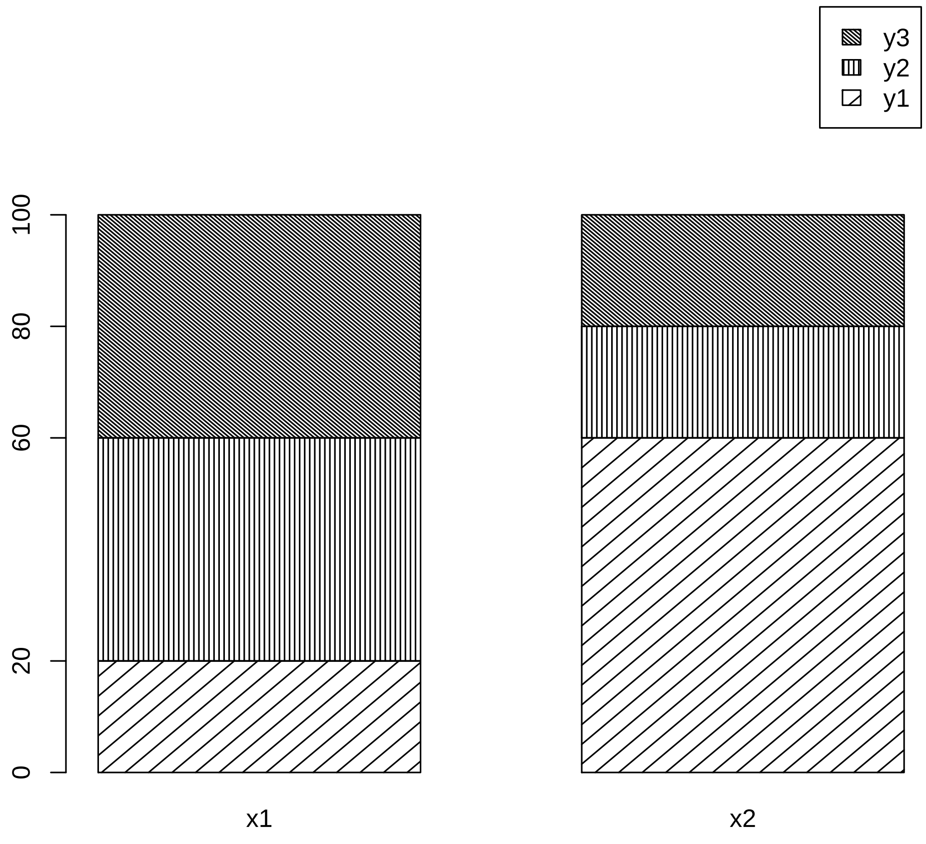


Figure 3.1 –

3.2 Khi-deux, V de Cramér et coefficient phi

Exercice 18 *Le tableau ci-après fournit les effectifs conjoints de la distribution conjointe des deux variables X = "Pays de naissance de la mère" et Y = "Pays de naissance du père" :*

X (mère) \ Y (père)	né en France	né à l'étranger
née en France	129	17
née à l'étranger	13	30

1. *Donner le tableau des effectifs théoriques.*
2. *Donner le tableau fournissant les contributions ; en déduire la valeur du χ^2.*
3. *Calculer le coefficient φ ; que pouvez-vous dire sur l'intensité du lien entre ces deux variables ?*

Exercice 19 *Le tableau ci-après fournit les effectifs conjoints de la distribution conjointe des deux variables X = "Locataire/Propriétaire" et Y = "attachement à certains endroits" mesurant le niveau d'adéquation (de "niv. 1"="tout à fait en désaccord" à "niv. 5"="tout à fait en accord") avec l'affirmation* Je suis très attaché(e) à certains endroits de ce quartier *:*

X \ Y	niv. 1	niv. 2	niv. 3	niv. 4	niv. 5
x_1 = "locataire"	22	45	25	20	22
x_2 = "propriétaire	13	5	10	13	14

1. *Donner le tableau des effectifs théoriques.*
2. *Donner le tableau fournissant les contributions ; en déduire la valeur du χ^2.*
3. *Calculer le V de Cramer ; que pouvez-vous dire sur l'intensité du lien entre le fait d'être propriétaire ou non et le fait d'être attaché à certains endroits.*

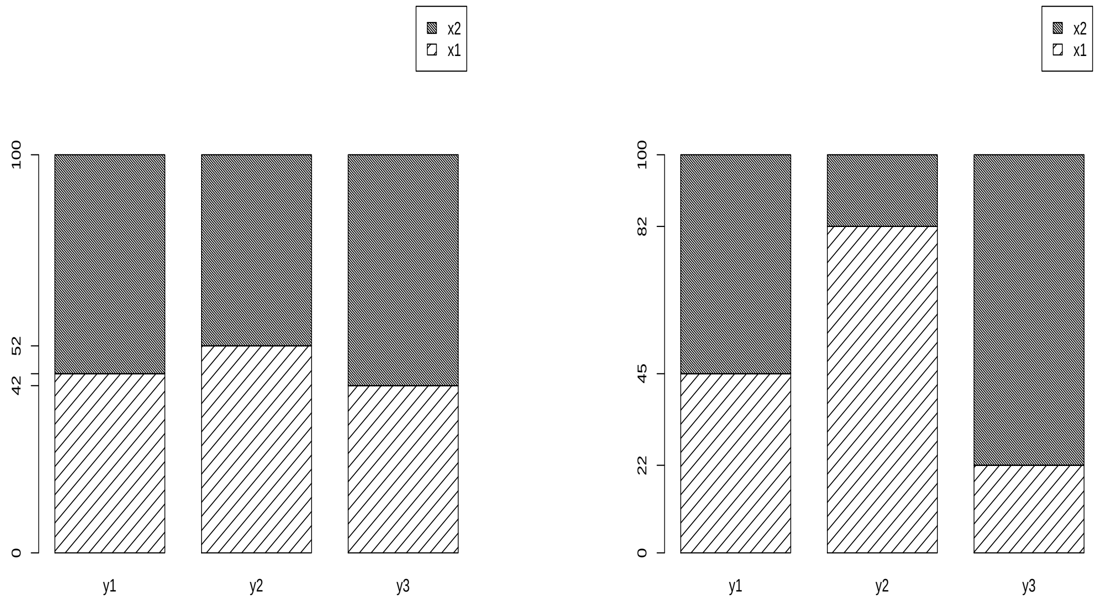
Exercice 20 *On a interrogé une partie des élèves d'un collège pour connaître la distance regroupée selon trois catégories (courte, moyenne, longue) qu'ils doivent parcourir pour se rendre à l'établissement scolaire (i.e. distance domicile/collège). On s'intéresse de plus à la variable Y="niveau scolaire". L'objectif est d'étudier l'éventuel impact de la distance domicile/collège sur les résultats scolaires. On obtient ainsi le tableau suivant :*

X \ Y	*faible*	*moyen*	*élevé*
courte	*23*	*25*	*79*
moyenne	*83*	*85*	*55*
longue	*102*	*21*	*27*

1. *Donner le tableau des effectifs théoriques.*
2. *Donner le tableau fournissant les contributions ; en déduire la valeur du* χ^2.
3. *Calculer le V de Cramer ; que pouvez-vous dire sur l'intensité du lien entre la distance séparant l'élève du collège et le niveau scolaire.*

Exercice 21 *On considère une variable X (resp. Y) ayant 2 (resp. 3) modalités* x_1 *et* x_2 *(resp.* y_1, y_2 *et* y_3*). On a obtenu pour 2 échantillons les 2 distributions représentées à la figure 3.2.*

1. *De quelle distribution s'agit-il ?*
2. *A partir de ces graphiques, lequel de ces 2 échantillons possède le V de Cramér le plus élevé ?*
3. *Proposer un graphique représentant le même type de distribution mais correspondant à un V de Cramér nul.*



(a) Échantillon 1

(b) Échantillon 2

Figure 3.2 –

Chapitre 4

Exercices récapitulatifs

Exercice 22 *30 élèves en CM2 ont fait 6 dictées. On s'intéresse au nombre total de fautes (variable X) cumulées lors de ces 6 dictées et on obtient le tableau ci-dessous :*

x_i	n_i	$n_i \times x_i$	$n_i \times (x_i)^2$	N_i *(eff. cum.)*
6	*1*			
12	*2*			
15	*6*			
18	*3*			
21	*1*			
23	*2*			
26	*3*			
30	*7*			
32	*4*			
34	*1*			
TOTAL			**17725**	

1. *Préciser la variable étudiée ainsi que son type. Quelle est la population ? Préciser la taille de l'échantillon.*
2. *Compléter le tableau ci-dessus.*
3. *Réaliser le diagramme en bâtons représentant la distribution de X. Que pouvez-vous déduire de ce graphique concernant la répartition du nombre de fautes ?*

4. *Calculer la moyenne et la variance de X.*

5. *Déterminer la médiane ; que représente cette valeur ?*

Exercice 23 . *Sur les 6 dictées réalisées par les élèves de CM2 (voir exercice 1), 3 se sont déroulées dans un environnement bruyant et 3 autres dans un environnement silencieux. Le tableau ci-après donne le nombre de fautes cumulées (regroupées en trois catégories) selon le type d'environnement :*

X \ Y	*bruyant*	*silencieux*	**Marge de X**
moins de 25	*37*		**241**
de 25 à 30		*86*	**288**
plus de 30		*21*	
Marge de Y			**N=691**

1. *Préciser les variables X et Y étudiées dans cette étude ainsi que leur type puis compléter le tableau ci-dessus.*

2. *Déterminer la distribution de X conditionnellement à Y. Représenter graphiquement cette distribution. Que pouvez-vous dire concernant le lien entre X et Y pour cet échantillon ?*

3. *Donner le tableau contenant les effectifs théoriques.*

4. *Calculer le χ^2 puis en déduire le coefficient φ. Que peut-on dire du lien entre X et Y ?*

Exercice 24 . *Une grande entreprise nommée a mené une enquête interne afin d'étudier, selon différents secteurs d'activités (variable X), le niveau de stress ressenti par ses employés (variable Y). Les données ont été regroupées dans la table de contingence ci-dessous :*

Y / X	*Faible*	*Moyen*	*Important*	*Extrême*	*Marge de X*
Commercial	*2*	*4*	*18*	*13*	
Production	*15*	*11*	*5*	*1*	
Marge de Y					

1. *Compléter le tableau précédent.*

2. *Préciser les variables étudiées ainsi que leur type. Quelle est la population étudiée ? Quelle est la taille de l'échantillon ?*

3. *Calculer la médiane de la loi marginale Y.*

4. *Déterminer la distribution de Y conditionnellement X. Représenter graphiquement cette distribution. À partir du graphique, peut-on en déduire qu'il existe un lien entre X et Y ?*

5. *Calculer le coefficient du χ^2 puis le coefficient φ. Que peut-on dire de la liaison entre X et Y ?*

Exercice 25 . *Une compagnie d'assurance souhaite étudier la distribution de l'âge de ses assurés (variable X dont l'unité est la décennie). Dans ce but, on donne le tableau suivant :*

y_i	n_i	c_i	$n_i\, c_i$	$n_i\, (c_i)^2$	d_i : *densité d'effectifs*	N_i
$[1.8; 2.5[$	*11*					
$[2.5; 3[$	*25*					
$[3; 3.5[$	*32*					
$[3.5; 4[$	*23*					
$[4; 5[$	*56*					
$[5; 6[$	*31*					
$[6; 7[$	*22*					
TOTAL						

1. *Compléter le tableau.*
2. *Représenter l'histogramme.*
3. *Déterminer le(s) mode(s).*
4. *Calculer la moyenne, variance et écart-type.*
5. *Déterminer la médiane ainsi que les 1er et 3ème quartiles.*

Chapitre 5

Annales

☐ **Régime Contrôle Continu**	☐ **Régime Examen Terminal**
Nom :	**Prénom :** ....
Nom d'épouse :	**Numéro d'étudiant :** ....
	Signature : ....

UE6 PY0106X - Partie Statistique - 1ère session/Mai 2012

Sujet à compléter puis à insérer dans la ou les copie(s)

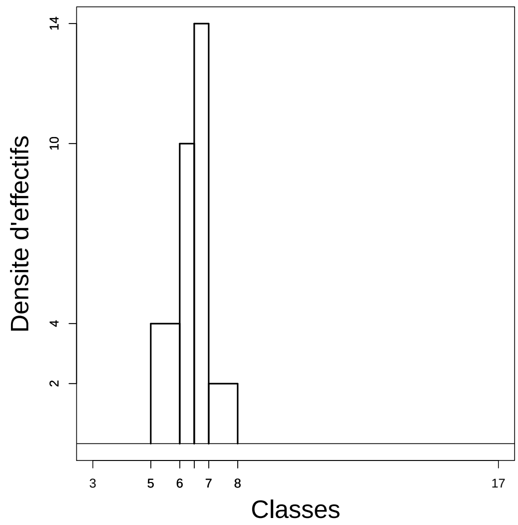
Exercice 1

On souhaite évaluer la capacité de mémorisation auprès d'élèves scolarisés en CM2. Dans ce but, on mesure sur un 1er échantillon d'élèves (groupe 1), le temps (variable X exprimée en minutes) nécessaire pour mémoriser un petit texte. Le tableau suivant fournit les résultats observés (où c_i désigne génériquement les centres des classes et a_i leur amplitude) :

x_i	n_i	c_i	$n_i \times c_i$	$n_i \times (c_i)^2$	N_i (eff. cum.)	$d_i = n_i/a_i$
[3 ;4[	3					
[4 ;5[	7					
[5 ;6[	5					
[6 ;11[	5					
[11 ;13[	6					
[13 ;17[	2					
TOTAL				$\boldsymbol{S_2 = 2005}$		

1) Quelle est la population étudiée ? Quelle est la taille de l'échantillon ? Préciser la variable étudiée ainsi que son type (en détaillant l'ensemble des valeurs possibles).
2) Compléter le tableau ci-dessus.

3) Représenter la distribution de X ; en déduire le(s) mode(s).
4) Calculer la moyenne, la variance et l'écart-type de X.
5) Déterminer la médiane ; que représente cet indice ? Expliquez brièvement pourquoi il y a une différence non négligeable entre la moyenne et la médiane.



6) La Figure ci-contre représente la distribution de cette variable obtenue à partir d'un 2ème échantillon d'élèves (groupe 2). À partir des histogrammes obtenus pour chacun de ces 2 groupes et sans faire de calculs, répondez aux questions suivantes en les justifiant :

6.a) Quel groupe d'élève paraît le plus homogène ? Que peut-on en déduire du point de vue de la variance ? Pour quel groupe d'élèves la moyenne est-elle la plus représentative ?

6.b) Pour quel groupe d'élèves l'écart entre la moyenne et la médiane est-il le plus grand ?

Exercice 2

0n s'intéresse aux 2 variables X="assiduité" et Y="niveau des résultats" que l'on a observées sur un échantillon d'étudiants inscrits en 1ère année. On a obtenu le tableau suivant :

X \ Y	faible	moyen	bon	**Marge de *X***
Assidu	14	42	97	
Non assidu	103	51	26	
Marge de *Y*				

1. Préciser les variables étudiées ainsi que leur type. Quelle est la population étudiée ? Quelle est la taille de l'échantillon ?
2. Compléter le tableau ci-dessus.
3. Déterminer la distribution de Y conditionnellement à X. Représenter graphiquement cette distribution. Que pouvez-vous dire concernant le lien entre X et Y pour cet échantillon ?

4. Compléter les 2 tableaux ci-dessous en détaillant vos calculs sur votre copie.

Tableau des effectifs théoriques

X \ Y	faible	moyen	bon
Assidu			
Non assidu			

Tableau des contributions

X \ Y	faible	moyen	bon
Assidu			
Non assidu			

5. Calculer le χ^2 d'indépendance.
6. En déduire le V de Cramér ϕ_c. Que peut-on dire de l'intensité du lien entre X (assiduité) et Y (niveau des résultats) pour cet échantillon ?
7. Question bonus (hors barême). Complétez en justifiant votre réponse le tableau ci-dessous par 2 effectifs conjoints de sorte qu'ils correspondent à la situation où l'on aurait le V de Cramér $\phi_c = 0$:

X \ Y	faible	moyen	bon
Assidu	14	42	84
Non assidu	103		

☐ **Régime Contrôle Continu**	☐ **Régime Examen Terminal**
Nom :	**Prénom :** ....
Nom d'épouse :	**Numéro d'étudiant :** ....
	Signature : ....

UE6 PY0106X - Partie Statistique - 2ème Session/Juin 2012

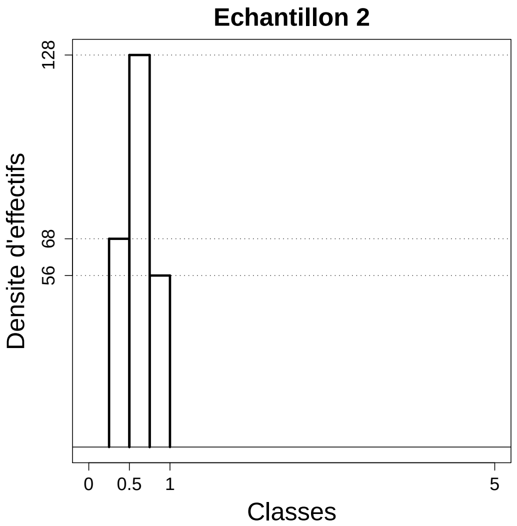
Sujet à compléter puis à insérer dans la ou les copie(s)

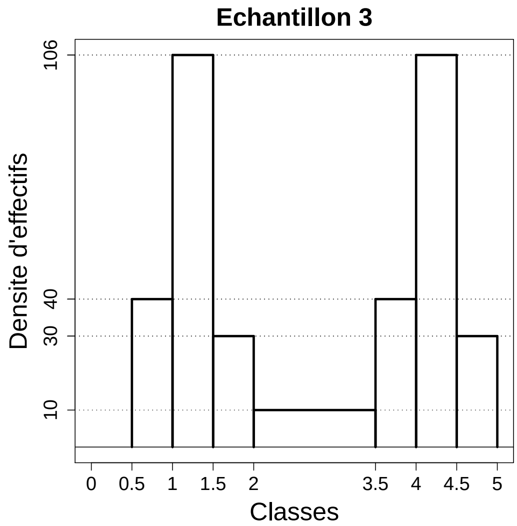
Exercice 1

On s'intéresse au temps (variable X exprimée en heure) passé quotidiennement par les élèves d'un collège sur les différents réseaux sociaux existant sur internet. Dans ce but, on a interrogé un 1er groupe d'élèves (échantillon 1) de ce collège. Le tableau suivant fournit les résultats observés (où c_i désigne génériquement les centres des classes et a_i leur amplitude) :

x_i	n_i	c_i	$n_i \times c_i$	$n_i \times (c_i)^2$	N_i (eff. cum.)	$d_i = n_i/a_i$
[0 ;0.25[	10					
[0.25 ;0.5[	13					
[0.5 ;0.75[	29					
[0.75 ;1[	9					
[1 ;2[	32					
[2 ;5]	7					
TOTAL				$\boldsymbol{S_2 = 177.95}$		

1) Quelle est la population étudiée ? Quelle est la taille de l'échantillon ? Préciser la variable étudiée ainsi que son type (en détaillant l'ensemble des valeurs possibles).
2) Compléter le tableau ci-dessus.
3) Représenter la distribution de X ; en déduire le(s) mode(s).
4) Calculer la moyenne, la variance et l'écart-type de X.
5) Déterminer la médiane ; que représente cet indice ?
6) Les 2 figures ci-dessous représentent la distribution de cette variable obtenue à partir de deux autres échantillons d'élèves :





Sans faire de calculs, répondez aux questions suivantes en les justifiant :

6.a) À partir des trois histogrammes dont vous disposez, quel échantillon d'élèves possède la plus petite variance ? Pour quel échantillon d'élèves la moyenne est-elle la plus représentative ?

6.b) Quelle est la médiane pour l'échantillon 3 ?

Exercice 2

On dispose d'un échantillon d'élèves en cours d'étude dans le secondaire pour lesquels on a observé X="cycle d'étude". On leur a demandé d'évaluer leur volume de travail personnel (faible, moyen ou important), cette dernière variable étant notée Y. On a obtenu le tableau suivant :

X \ Y	faible	moyen	important	**Marge de X**
Collège			16	144
Lycée	47		46	
Marge de Y	143			290

1) Préciser les variables étudiées ainsi que leur type. Quelle est la population étudiée ? Quelle est la taille de l'échantillon ?

2) Compléter le tableau ci-dessus.

3) Déterminer la distribution de X conditionnellement à Y. Représenter graphiquement cette distribution. Que pouvez-vous dire concernant le lien entre X et Y pour cet échantillon ?
4) Compléter les 2 tableaux ci-dessous en détaillant (pour chaque tableau) vos deux premiers calculs sur votre copie.

Tableau des effectifs théoriques

X \ Y	faible	moyen	important
Collège			
Lycée			

Tableau des contributions

X \ Y	faible	moyen	important
Collège			
Lycée			

5) Afin de répondre à la question “Que peut-on dire de l’intensité du lien entre le cycle d’étude et le volume de travail personnel sur cet échantillon ?” et en utilisant ce qui précède, calculer un indice adéquat puis conclure.

Chapitre 6

Synoptique et formulaire

Le prochain tableau est un synoptique proposant une vision synthétique du programme couvert par ce module. A la suite de ce synoptique, vous trouverez un formulaire/aide-mémoire qui sera distribué avec le sujet d'examen. Ce formulaire ne peut en aucun cas refléter de façon exhaustive le contenu de ce module ; il s'agit simplement d'un aide-mémoire qui vient en complément d'une étude approfondie de ce cours.

SYNOPTIQUE DES SITUATIONS ABORDÉES DANS LA PARTIE STATISTIQUE DU PY0106X

Situations	**Exemples**	**Organisation des données**	**Représentations graphiques**	**Ce qu'on peut déterminer**
1 variable qualitative nominale	Zone d'habitat (Centre ville, Banlieue, Village, Autre)	Tableau d'effectifs	Diagramme en barres ou en secteurs ou unicolonne	Mode
1 variable qualitative ordinale	Niveau d'adéquation avec une affirmation (1 = "tout à fait en désaccord"),…, 5 = "tout à fait d'accord"	Tableau d'effectifs	Diagramme en barres ou en secteurs ou unicolonne	Mode(s), médiane, quartiles
1 variable quantitative discrète	Nombre de personnes vivant dans le même logement dont l'ensemble des modalités possibles est $\{1, 2, 3, \ldots\}$	Tableau d'effectifs	Diagramme en bâtons	Mode(s), médiane, quartiles, moyenne, variance, écart-type
1 variable quantitative continue	Temps vécu dans le logement (valeurs possibles = tous les nombres positifs)	Tableau d'effectifs dont les modalités sont données sous forme de classes	Histogramme (on représente les densités d'effectifs)	Mode(s), médiane, quartiles, moyenne, variance, écart-type
Couple de variables (X, Y)	X = "Locataire/Propriétaire" et Y = "Zone d'habitat"	Tableau des effectifs conjoints (appelé aussi table de contingence)	Distribution conjointe, Distributions marginales, Distributions conditionnelles	En plus des indices usuels liés à l'étude d'une seule variable, on peut calculer les mesures d'association : khi-deux, V de Cramér (ou coefficient phi si on a seulement 2 modalités pour X et Y)

Aide mémoire - Formulaire

Étude d'une variable

Variable qualitative

Représentations : diagramme en barres ou diagramme unicolonne ou diagramme en secteurs.

Variable quantitative discrète

$\boldsymbol{x_i}$ (modalités)	$\boldsymbol{n_i}$ (effectifs)	$\boldsymbol{f_i}$ (fréquences)	$\boldsymbol{n_i \times x_i}$	$\boldsymbol{n_i \times (x_i)^2}$	$\boldsymbol{N_i}$ (eff. cum.)
x_1	n_1	$f_1 = n_1/N$	$n_1 \times x_1$	$n_1 \times (x_1)^2$	$N_1 = n_1$
x_2	n_2	$f_2 = n_2/N$	$n_2 \times x_2$	$n_2 \times (x_2)^2$	$N_2 = N_1 + n_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_K	n_K	$f_K = n_K/N$	$n_K \times x_K$	$n_K \times (x_K)^2$	$N_K = N_{K-1} + n_K$
TOTAL	$\boldsymbol{N}$	1	$\boldsymbol{S_1}$	$\boldsymbol{S_2}$	

Pourcentage = fréquence$\times$100.

$$\boxed{\overline{X} = \frac{1}{\boldsymbol{N}}\Big\{(n_1 \times x_1) + (n_2 \times x_2) + \cdots + (n_K \times x_K)\Big\}} = \frac{\boldsymbol{S_1}}{\boldsymbol{N}}$$

$$\boxed{Var(X) = \frac{1}{\boldsymbol{N}}\Big\{\big(n_1 \times (x_1)^2\big) + \big(n_2 \times (x_2)^2\big) + \cdots + \big(n_K \times (x_K)^2\big)\Big\} - (\overline{X})^2} = \frac{\boldsymbol{S_2}}{\boldsymbol{N}} - (\overline{X})^2$$

La variance est aussi notée S_X^2 ou encore σ_X^2.

Écart-type : $\boxed{\sigma_X = \sqrt{Var(X)}}$. σ_X est aussi parfois noté S_X.

Représentation : diagramme en bâtons.

Variable quantitative continue

Les modalités se présentent sous la forme de classes de valeurs. Pour calculer moyenne, variance et écart-type, il suffit de remplacer les modalités $x_1, x_2, \ldots, x_K$ par les centres des classes $c_1, c_2, \ldots, c_K$.
Représentation : histogramme (on représente les bornes des classes sur l'axe horizontal et sur l'axe vertical les densités d'effectifs $d_1 = n_1/a_1, d_2 = n_2/a_2, \ldots, d_K = n_K/a_K$ où $a_1, a_2, \ldots, a_K$ sont les amplitudes des classes).

Médiane et quartiles

Médiane et quartiles sont définis pour des variables $\left\{\begin{array}{c}\text{qualitatives ordinales}\\ \text{ou}\\ \text{quantitatives}\end{array}\right.$; on suppose que les modalités sont ordonnées dans l'ordre croissant.

Médiane : modalité séparant l'échantillon de taille N en 2 sous-échantillons de même taille $N/2$.

Détermination de la médiane :

1) on sélectionne l'effectif cumulé N_* immédiatement supérieur à $N/2$,

2) la modalité correspondant à N_* est la médiane (dans le cas d'une variable quantitative continue, la classe correspondant à N_* est applée classe médiane et la médiane est le centre de la

classe médiane).
Quartiles : modalités notées Q_1, Q_2 et Q_3 séparant l'échantillon de taille N en 4 sous-échantillons de même taille $N/4$.
Détermination de Q_1 : même procédé que pour la médiane en remplaçant $N/2$ par $N/4$.
Détermination de Q_2 : par définition, Q_2 = médiane.
Détermination de Q_3 : même procédé que pour la médiane en remplaçant $N/2$ par $(3 \times N)/4$.

Mode(s)

Variable qualitative nominale : mode = modalité de plus grand effectif.
Variable qualitative ordinale ou quantitative : mode(s) = modalité(s) correspondant au(x) pic(s) observé(s) sur le graphique représentant la distribution (lorsqu'on dispose de classes, on parle de classes modales ; modes = centres des classes modales).

Étude d'un couple de variables (X, Y)

Distribution conjointe, marginale, conditionnelle

X \ Y	y_1	y_2	$\cdots$	y_C	**Marge de X**
x_1	$n_{1,1}$	$n_{1,2}$	$\cdots$	$n_{1,C}$	$\boldsymbol{n_{1\bullet}}$
x_2	$n_{2,1}$	$n_{2,2}$	$\cdots$	$n_{2,C}$	$\boldsymbol{n_{2\bullet}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_L	$n_{L,1}$	$n_{L,2}$	$\cdots$	$n_{L,C}$	$\boldsymbol{n_{L\bullet}}$
Marge de Y	$\boldsymbol{n_{\bullet 1}}$	$\boldsymbol{n_{\bullet 2}}$	$\cdots$	$\boldsymbol{n_{\bullet C}}$	$\boldsymbol{N}$

Distribution conjointe de (X, Y) = ensemble des informations contenues dans le tableau fournissant les $L \times C$ effectifs conjoints $n_{1,1}, n_{1,2}, \ldots, n_{L,C}$.

Distribution marginale de X = $(x_1, n_{1\bullet}), (x_2, n_{2\bullet}), \ldots, (x_L, n_{L\bullet})$ où $n_{1\bullet}, n_{2\bullet}, \ldots, n_{L\bullet}$ sont les L effectifs marginaux de X (somme des effectifs conjoints ligne par ligne).

Distribution marginale de Y = $(y_1, n_{\bullet 1}), (y_2, n_{\bullet 2}), \ldots, (y_C, n_{\bullet C})$ où $n_{\bullet 1}, n_{\bullet 2}, \ldots, n_{\bullet C}$ sont les C effectifs marginaux de Y (somme des effectifs conjoints colonne par colonne).

Distribution de X conditionnellement à Y

$X \backslash Y$	y_1	y_2	$\cdots$	y_C
x_1	$\frac{n_{1,1}}{\boldsymbol{n_{\bullet 1}}}$	$\frac{n_{1,2}}{\boldsymbol{n_{\bullet 2}}}$	$\cdots$	$\frac{n_{1,C}}{\boldsymbol{n_{\bullet C}}}$
x_2	$\frac{n_{2,1}}{\boldsymbol{n_{\bullet 1}}}$	$\frac{n_{2,2}}{\boldsymbol{n_{\bullet 2}}}$	$\cdots$	$\frac{n_{2,C}}{\boldsymbol{n_{\bullet C}}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_L	$\frac{n_{L,1}}{\boldsymbol{n_{\bullet 1}}}$	$\frac{n_{L,2}}{\boldsymbol{n_{\bullet 2}}}$	$\cdots$	$\frac{n_{L,C}}{\boldsymbol{n_{\bullet C}}}$
TOTAL	1	1	$\cdots$	1

Distribution de Y conditionnellement à X

$X \backslash Y$	y_1	y_2	$\cdots$	y_C	**TOTAL**
x_1	$\frac{n_{1,1}}{\boldsymbol{n_{1\bullet}}}$	$\frac{n_{1,2}}{\boldsymbol{n_{1\bullet}}}$	$\cdots$	$\frac{n_{1,C}}{\boldsymbol{n_{1\bullet}}}$	1
x_2	$\frac{n_{2,1}}{\boldsymbol{n_{2\bullet}}}$	$\frac{n_{2,2}}{\boldsymbol{n_{2\bullet}}}$	$\cdots$	$\frac{n_{2,C}}{\boldsymbol{n_{2\bullet}}}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_L	$\frac{n_{L,1}}{\boldsymbol{n_{L\bullet}}}$	$\frac{n_{L,2}}{\boldsymbol{n_{L\bullet}}}$	$\cdots$	$\frac{n_{L,C}}{\boldsymbol{n_{L\bullet}}}$	1

Il suffit de multiplier toutes ces quantités par 100 pour obtenir des pourcentages.

Mesures d'association : khi-deux, V de Cramér, coefficient phi

Tableau des effectifs théoriques $(T_{1,1}, T_{1,2}, \ldots, T_{L,C})$

$X \backslash Y$	y_1	y_2	$\cdots$	y_C	**Marge de $\boldsymbol{X}$**
x_1	$T_{1,1} = \frac{n_{1\bullet} \times n_{\bullet 1}}{N}$	$T_{1,2} = \frac{n_{1\bullet} \times n_{\bullet 2}}{N}$	$\cdots$	$T_{1,C} = \frac{n_{1\bullet} \times n_{\bullet C}}{N}$	$\boldsymbol{n_{1\bullet}}$
x_2	$T_{2,1} = \frac{n_{2\bullet} \times n_{\bullet 1}}{N}$	$T_{2,2} = \frac{n_{2\bullet} \times n_{\bullet 2}}{N}$	$\cdots$	$T_{2,C} = \frac{n_{2\bullet} \times n_{\bullet C}}{N}$	$\boldsymbol{n_{2\bullet}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_L	$T_{L,1} = \frac{n_{L\bullet} \times n_{\bullet 1}}{N}$	$T_{L,2} = \frac{n_{L\bullet} \times n_{\bullet 2}}{N}$	$\cdots$	$T_{L,C} = \frac{n_{L\bullet} \times n_{\bullet C}}{N}$	$\boldsymbol{n_{L\bullet}}$
Marge de $\boldsymbol{Y}$	$\boldsymbol{n_{\bullet 1}}$	$\boldsymbol{n_{\bullet 2}}$	$\cdots$	$\boldsymbol{n_{\bullet C}}$	$\boldsymbol{N}$

Tableau des contributions $(cont_{1,1}, cont_{1,2}, \ldots, cont_{L,C})$

$X \backslash Y$	y_1	y_2	$\cdots$	y_C
x_1	$cont_{1,1} = \dfrac{(n_{1,1} - T_{1,1})^2}{T_{1,1}}$	$cont_{1,2} = \dfrac{(n_{1,2} - T_{1,2})^2}{T_{1,2}}$	$\cdots$	$cont_{1,C} = \dfrac{(n_{1,C} - T_{1,C})^2}{T_{1,C}}$
x_2	$cont_{2,1} = \dfrac{(n_{2,1} - T_{2,1})^2}{T_{2,1}}$	$cont_{2,2} = \dfrac{(n_{2,2} - T_{2,2})^2}{T_{2,2}}$	$\cdots$	$cont_{2,C} = \dfrac{(n_{2,C} - T_{2,C})^2}{T_{2,C}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_L	$cont_{L,1} = \dfrac{(n_{L,1} - T_{L,1})^2}{T_{L,1}}$	$cont_{L,2} = \dfrac{(n_{L,2} - T_{L,2})^2}{T_{L,2}}$	$\cdots$	$cont_{L,C} = \dfrac{(n_{L,C} - T_{L,C})^2}{T_{L,C}}$

Khi-deux d'indépendance : χ^2 = somme de toutes les contributions $= cont_{1,1} + cont_{1,2} + \cdots + cont_{L,C}$

V de Cramér : $\phi_c = \sqrt{\dfrac{\chi^2}{N \times \left\{ \min(L,C) - 1 \right\}}}$.

On a $0 \leq \phi_c \leq 1$: $\begin{cases} 0 \leq \phi_c < 0.3 & \longrightarrow \text{ lien d'intensité faible,} \\ 0.3 \leq \phi_c < 0.5 & \longrightarrow \text{ lien d'intensité moyenne,} \\ 0.5 \leq \phi_c < 1 & \longrightarrow \text{ lien d'intensité forte.} \end{cases}$

Cas particulier quand $L = C = 2$: Coefficient phi $= \sqrt{\dfrac{\chi^2}{N}}$.