

Statistique descriptive bivariable

Couple de variables

Marie-Luce Taupin
marie-luce.taupin@genopole.cnrs.fr

2015-2016

Objectifs

- **Statistique univariée** : analyse descriptive séparée de chaque variable d'un tableau *individus* \times *variables*.
- **Statistique bivariée** : analyse descriptive des variables deux à deux :
 - ▶ étude d'un couple de variables statistiques
 - ▶ étude de la liaison entre deux variables quantitatives, qualitatives, quantitative/qualitative
 - ▶ **étape indispensable de toute analyse de jeux de données : croisement systématique des variables 2 à 2.**
- Statistique descriptive multivariée : Analyse des données.

Données brutes et données groupées

Étude de deux variables X et Y sur une **même population** de taille n :

- x_k et y_k : valeurs prises par X et Y pour un même individu k , $1 \leq k \leq n$.
- **Données brutes** $(x_k, y_k)_{k=1, \dots, n}$: les n couples d'observations

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

Exemple

Extrait des données brutes :

Individu	Sexe X	Salaire horaire Y
1	F	13.25
2	F	12.50
3	H	14.00
4	F	13.00
5	H	7.00
6	F	29.80
...		
599	H	14.50

- Le salaire horaire dépend-il du sexe des individus ?

Exemple

Extrait des données brutes :

Employé	Catégorie de personnel	Age	Région
1	A	58	NE
2	B	42	W
3	A	35	S
4	B	26	NE
5	B	22	W
6	C	32	NW
7	A	42	NE
...
597	C	41	S
598	C	33	NW
599	C	29	S

- La répartition des âges est-elle différente selon la catégorie de personnel (et dans quelle mesure) ?
- La catégorie des employés est-elle liée à la région (et de quelle manière) ?

Extrait des données

```
> head(Donnees)
```

	AGE	SEXE	REGION	STAT_MARI	SAL_HOR	SYNDICAT	CATEGORIE	NIV_ETUDES	NB_PERS	NB_ENF	REV_FOYER
1	58	F	NE	C	13.25	non	5	43	2	0	11
2	40	M	W	M	12.50	non	7	38	2	0	7
3	29	M	S	C	14.00	non	5	42	2	0	15
4	59	M	NE	D	10.60	oui	3	39	4	1	7
5	51	M	W	M	13.00	non	3	35	8	1	15
6	19	M	NW	C	7.00	non	3	39	6	0	16

```
> tail(Donnees)
```

	AGE	SEXE	REGION	STAT_MARI	SAL_HOR	SYNDICAT	CATEGORIE	NIV_ETUDES	NB_PERS	NB_ENF	REV_FOYER
594	63	M	NE	M	10.5	non	4	40	2	0	13
595	51	F	S	M	29.8	non	2	42	2	0	14
596	29	F	NE	C	27.0	oui	1	43	2	0	15
597	57	F	NW	D	21.0	non	4	40	1	0	14
598	29	F	W	M	13.0	oui	5	39	6	4	11
599	47	M	S	C	14.5	non	4	39	1	0	12

Description des données

```
> dim(Donnees)
[1] 599 11

> attach(Donnees)

> names(Donnees)

[1] "AGE" "SEXE" "REGION" "STAT_MARI" "SAL_HOR" "SYNDICAT" "CATEGORIE" "NIV_ETUDES"
"NB_PERS" "NB_ENF" "REV_FOYER"

> str(Donnees)

'data.frame': 599 obs. of 11 variables:
 $ AGE      : int  58 40 29 59 51 19 64 23 47 66 ...
 $ SEXE     : Factor w/ 2 levels "F","M": 1 2 2 2 2 1 1 2 1 ...
 $ REGION   : Factor w/ 4 levels "NE","NW","S",...: 1 4 3 1 4 2 3 1 2 3 ...
 $ STAT_MARI : Factor w/ 5 levels "C","D","M","S",...: 1 3 1 2 3 1 3 1 3 2 ...
 $ SAL_HOR  : num  13.2 12.5 14 10.6 13 ...
 $ SYNDICAT : Factor w/ 2 levels "non","oui": 1 1 1 2 1 1 1 1 2 1 ...
 $ CATEGORIE : int  5 7 5 3 3 3 9 1 8 5 ...
 $ NIV_ETUDES: int  43 38 42 39 35 39 40 43 40 40 ...
 $ NB_PERS   : int  2 2 2 4 8 6 3 2 3 1 ...
 $ NB_ENF    : int  0 0 0 1 1 0 0 0 0 0 ...
 $ REV_FOYER : int  11 7 15 7 15 16 13 11 12 8 ...
```

Description des données

```
## Modification du type des variables
```

```
Donnees$CATEGORIE=as.factor(Donnees$CATEGORIE)
Donnees$NIV_ETUDES=as.factor(Donnees$NIV_ETUDES)
Donnees$REV_FOYER=as.factor(Donnees$REV_FOYER)
```

```
> str(Donnees)
```

```
'data.frame': 599 obs. of 11 variables:
 $ AGE      : int  58 40 29 59 51 19 64 23 47 66 ...
 $ SEXE     : Factor w/ 2 levels "F","M": 1 2 2 2 2 1 1 2 1 ...
 $ REGION   : Factor w/ 4 levels "NE","NW","S",..: 1 4 3 1 4 2 3 1 2 3 ...
 $ STAT_MARI : Factor w/ 4 levels "C","D","M","V": 1 3 1 2 3 1 3 1 3 2 ...
 $ SAL_HOR  : num  13.2 12.5 14 10.6 13 ...
 $ SYNDICAT : Factor w/ 2 levels "non","oui": 1 1 1 2 1 1 1 1 2 1 ...
 $ CATEGORIE : Factor w/ 10 levels "1","2","3","4",...: 5 7 5 3 3 3 9 1 8 5 ...
 $ NIV_ETUDES: Factor w/ 15 levels "32","33","34",...: 12 7 11 8 4 8 9 12 9 9 ...
 $ NB_PERS  : int   2 2 2 4 8 6 3 2 3 1 ...
 $ NB_ENF   : int   0 0 0 1 1 0 0 0 0 0 ...
 $ REV_FOYER : Factor w/ 16 levels "1","2","3","4",...: 11 7 15 7 15 16 13 11 12 8 ...
```


Résumé des données

```
> summary(Donnees)
```

AGE	SEXE	REGION	STAT_MARI	SAL_HOR	SYNDICAT	CATEGORIE	NIV_ETUDES
Min. :16.00	F:297	NE:129	C:193	Min. : 2.0	non:496	2 :133	39 :187
1st Qu.:29.00	M:302	NW:122	D: 75	1st Qu.:10.5	oui:103	3 :125	40 :148
Median :42.00		S :200	M:325	Median :15.0		5 : 94	43 :114
Mean :41.85		W :148	V: 6	Mean :17.9		4 : 48	42 : 45
3rd Qu.:53.50				3rd Qu.:22.0		1 : 46	44 : 29
Max. :80.00				Max. :99.0		9 : 39	41 : 22
						(Other):114	(Other): 54

NB_PERS	NB_ENF	REV_FOYER
Min. : 1.00	Min. :0.0000	14 : 89
1st Qu.: 2.00	1st Qu.:0.0000	15 : 77
Median : 3.00	Median :0.0000	13 : 71
Mean : 3.11	Mean :0.5326	12 : 70
3rd Qu.: 4.00	3rd Qu.:1.0000	11 : 61
Max. :13.00	Max. :6.0000	16 : 48
		(Other):183

X et/ou Y qualitatives ou quantitatives discrètes

- $x_1, x_2, \dots, x_i, \dots, x_p$: les p modalités de X (p observations distinctes de X)
- $y_1, y_2, \dots, y_j, \dots, y_q$: les q modalités de Y (q observations distinctes de Y)

X et/ou Y quantitatives continues

- Valeurs de X regroupées en p classes

$$[e_0^X, e_1^X[, \dots, [e_{i-1}^X, e_i^X[, \dots, [e_{p-1}^X, e_p^X[$$

de centres $x_1, \dots, x_i, \dots, x_p$

- Valeurs de Y en q classes

$$[e_0^Y, e_1^Y[, \dots, [e_{j-1}^Y, e_j^Y[, \dots, [e_{q-1}^Y, e_q^Y[$$

de centres $y_1, \dots, y_j, \dots, y_q$

- Confusion parfois entre la classe $[e_{i-1}^X, e_i^X[$ et son centre x_i

Données groupées

- n_{ij} : nombre d'individus pour lesquels à la fois X prend la valeur x_i **et** Y la valeur y_j

$$n_{ij} = \#\{k = 1, \dots, n \mid x_k = x_i \text{ et } y_k = y_j\}$$

- Si X est continue, $x_k = x_i$ signifie $x_k \in [e_{i-1}^X, e_i^X[$ de centre x_i
- **Données groupées** : $(x_i, y_j, n_{ij})_{i=1, \dots, p, j=1, \dots, q}$

Tableaux statistiques et distribution d'une série bvariée

Distribution jointe - Tableau de contingence

- **Distribution jointe en effectifs** de X et de Y :

$$\{(x_i, y_j, n_{ij}) ; 1 \leq i \leq p, 1 \leq j \leq q\}$$

- Pour $i = 1, \dots, p$ et $j = 1, \dots, q$

- ▶ n_{ij} : nombre d'individus possédant la modalité x_i de X **et** la modalité y_j de Y .
- ▶ $n_{i\bullet} = \sum_{j=1}^q n_{ij}$: nombre d'individus possédant la modalité x_i (\in classe de centre x_i) de X
- ▶ $n_{\bullet j} = \sum_{i=1}^p n_{ij}$: nombre d'individus possédant la modalité y_j de Y
- ▶ $n = \sum_{i=1}^p \sum_{j=1}^q n_{ij} = \sum_{i=1}^p n_{i\bullet} = \sum_{j=1}^q n_{\bullet j}$; nombre total d'individus de la population.

Tableau de contingence en effectifs (p lignes, q colonnes)

X	Y	y_1	y_2	\dots	y_j	\dots	y_q	Total
x_1		n_{11}	n_{12}	\dots	n_{1j}	\dots	n_{1q}	$n_{1\bullet}$
x_2		n_{21}	n_{22}	\dots	n_{2j}	\dots	n_{2q}	$n_{2\bullet}$
\vdots		\vdots	\vdots		\vdots		\vdots	\vdots
x_i		n_{i1}	n_{i2}	\dots	n_{ij}	\dots	n_{iq}	$n_{i\bullet}$
\vdots		\vdots	\vdots		\vdots		\vdots	\vdots
x_p		n_{p1}	n_{p2}	\dots	n_{pj}	\dots	n_{pq}	$n_{p\bullet}$
Total		$n_{\bullet 1}$	$n_{\bullet 2}$	\dots	$n_{\bullet j}$	\dots	$n_{\bullet q}$	n

Tableau de contingence : SEXE x REGION

```
> TabContEf<-table(SEXE,REGION)
```

```
> print(TabContEf) # affiche le nom des variables
```

	REGION			
SEXE	NE	NW	S	W
F	61	62	97	77
M	68	60	103	71

```
> addmargins(TabContEf)
```

	REGION				
SEXE	NE	NW	S	W	Sum
F	61	62	97	77	297
M	68	60	103	71	302
Sum	129	122	200	148	599

X : SEXE et Y : REGION

- X de type à $p = \dots\dots\dots$ modalités.
- Y de type 0 $q = \dots\dots\dots$ modalités.
- Mesures conjointes de X et Y sur $n = \dots\dots$ individus.

Distribution jointe en fréquences

- Pour $i = 1, \dots, p$ et $j = 1, \dots, q$

▶ $f_{ij} = \frac{n_{ij}}{n}$: proportion d'individus possédant la modalité x_i de la variable X et la modalité y_j de la variable Y .

▶ $f_{i\bullet} = \sum_{j=1}^q f_{ij}$: fréquence de la modalité x_i de X

▶ $f_{\bullet j} = \sum_{i=1}^p f_{ij}$: fréquence de la modalité y_j de Y

▶ $1 = \sum_{i=1}^p \sum_{j=1}^q f_{ij} = \sum_{i=1}^p f_{i\bullet} = \sum_{j=1}^q f_{\bullet j}$

- **Distribution jointe en fréquences** de X et de Y :

$$\{(x_i, y_j, f_{ij}) ; 1 \leq i \leq p, 1 \leq j \leq q\}$$

Tableau de contingence en fréquences (p lignes, q colonnes)

X	Y	y_1	y_2	\dots	y_j	\dots	y_q	Total
x_1		f_{11}	f_{12}	\dots	f_{1j}	\dots	f_{1q}	$f_{1\bullet}$
x_2		f_{21}	f_{22}	\dots	f_{2j}	\dots	f_{2q}	$f_{2\bullet}$
\vdots		\vdots	\vdots		\vdots		\vdots	\vdots
x_i		f_{i1}	f_{i2}	\dots	f_{ij}	\dots	f_{iq}	$f_{i\bullet}$
\vdots		\vdots	\vdots		\vdots		\vdots	\vdots
x_p		f_{p1}	f_{p2}	\dots	f_{pj}	\dots	f_{pq}	$f_{p\bullet}$
Total		$f_{\bullet 1}$	$f_{\bullet 2}$	\dots	$f_{\bullet j}$	\dots	$f_{\bullet q}$	1

Tableau de contingence : SEXE x REGION

```
> TabContFr<-prop.table(TabContEf)
```

```
> print(TabContFr)
```

	REGION				
SEXE	NE	NW	S	W	
F	0.1018364	0.1035058	0.1619366	0.1285476	
M	0.1135225	0.1001669	0.1719533	0.1185309	

```
> print(round(TabContFr,2))
```

	REGION				
SEXE	NE	NW	S	W	
F	0.10	0.10	0.16	0.13	
M	0.11	0.10	0.17	0.12	

```
> addmargins(round(TabContFr,2))
```

	REGION				
SEXE	NE	NW	S	W	Sum
F	0.10	0.10	0.16	0.13	0.49
M	0.11	0.10	0.17	0.12	0.50
Sum	0.21	0.20	0.33	0.25	0.99

Tableau de contingence en % : SEXE x REGION

```
> TabContPr<-100*prop.table(TabContEf)
```

```
> print(TabContPr)
```

```
      REGION
SEXE    NE    NW    S    W
  F 10.18364 10.35058 16.19366 12.85476
  M 11.35225 10.01669 17.19533 11.85309
```

```
> print(round(TabContPr,2))
```

```
      REGION
SEXE    NE    NW    S    W
  F 10.18 10.35 16.19 12.85
  M 11.35 10.02 17.20 11.85
```

```
> addmargins(round(TabContPr,2))
```

```
      REGION
SEXE    NE    NW    S    W    Sum
  F  10.18 10.35 16.19 12.85 49.57
  M  11.35 10.02 17.20 11.85 50.42
Sum 21.53 20.37 33.39 24.70 99.99
```

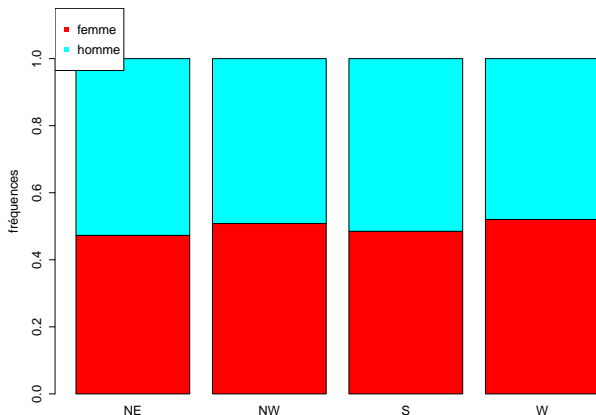
Tableau de contingence : Autre représentation

```
> library(gplots)
> balloonplot(t(TabContEf),dotsize=10,main="")
```



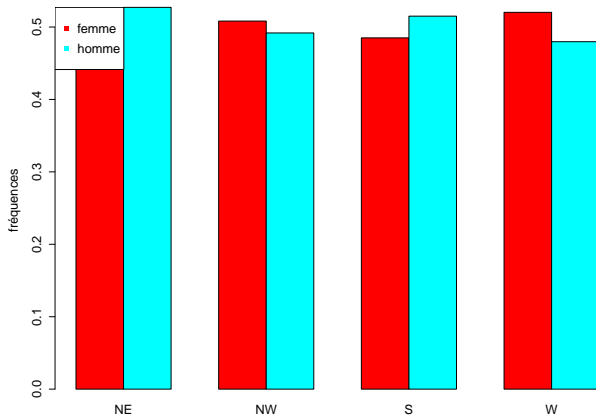
Représentations graphiques

```
> N1=nlevels(SEXE) # nombre de modalités (niveaux) du facteur Sexe  
> N2=nlevels(REGION) # nombre de modalités (niveaux) du facteur REGION  
> couleurs=rainbow(N1)  
> barplot(TabContFr, col=couleurs,2)  
> legend("topleft", legend=c("F", "H"), col=couleurs,pch=15)
```



Représentations graphiques

```
> barplot(TabContFr,beside=TRUE, col=couleurs, 2)  
> legend("topleft", legend=c("F", "H"), col=couleurs,pch=15)
```



Représentations graphiques

A ne pas faire (sauf si les modalités sont équilibrées) !

```
> couleurs=rainbow(N2)
```

```
> mosaicplot(TabContEf,col=couleurs,main="")
```

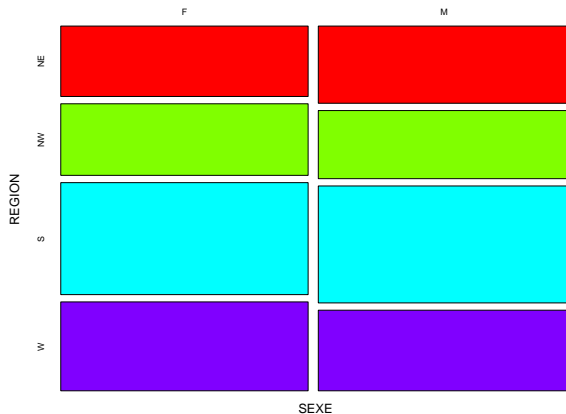


Tableau de contingence : SEXE x SALAIRE

⇒ Tableau de contingence Qualitatif x Quantitatif.

```
> Nclasse=4 # Nombre de classes
> SALAIRE<-cut(SAL_HOR,breaks=Nclasse)
> TabContEf<-table(SEXE,SALAIRE)
```

```
> print(TabContEf)
```

	SALAIRE			
SEXE	(1.9,26.2]	(26.2,50.5]	(50.5,74.8]	(74.8,99.1]
F	262	31	3	1
M	244	49	7	2

```
> addmargins(TabContEf)
```

	SALAIRE				
SEXE	(1.9,26.2]	(26.2,50.5]	(50.5,74.8]	(74.8,99.1]	Sum
F	262	31	3	1	297
M	244	49	7	2	302
Sum	506	80	10	3	599

Tableau de contingence : SEXE x SALAIRE

```
> TabContFr<-prop.table(TabContEf)
```

```
> print(TabContFr)
```

```
      SALAIRE
SEXE  (1.9,26.2] (26.2,50.5] (50.5,74.8] (74.8,99.1]
      F 0.437395659 0.051752922 0.005008347 0.001669449
      M 0.407345576 0.081803005 0.011686144 0.003338898
```

```
> print(round(TabContFr,2))
```

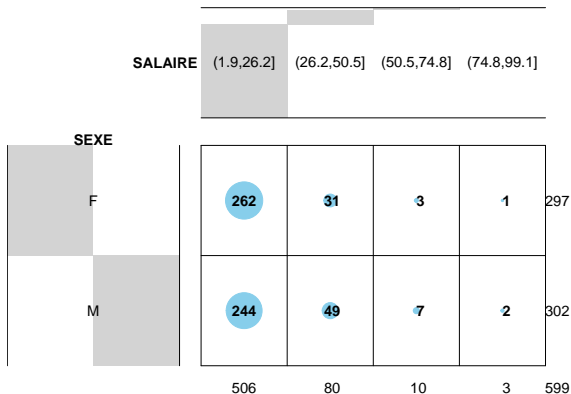
```
      SALAIRE
SEXE  (1.9,26.2] (26.2,50.5] (50.5,74.8] (74.8,99.1]
      F      0.44      0.05      0.01      0.00
      M      0.41      0.08      0.01      0.00
```

```
> addmargins(round(TabContFr,2))
```

```
      SALAIRE
SEXE  (1.9,26.2] (26.2,50.5] (50.5,74.8] (74.8,99.1] Sum
      F      0.44      0.05      0.01      0.00 0.50
      M      0.41      0.08      0.01      0.00 0.50
      Sum      0.85      0.13      0.02      0.00 1.00
```

Tableau de contingence : Autre représentation

```
> balloonplot(t(TabContEf),dotsize=10,main="")
```

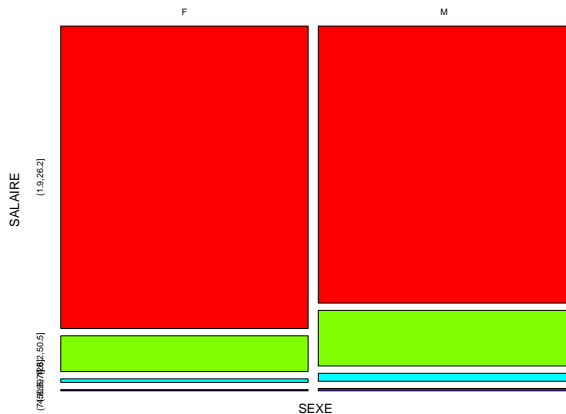


Représentations graphiques

A ne pas faire (sauf si les modalités sont équilibrées) !

```
> couleurs=rainbow(N2)
```

```
> mosaicplot(TabContEf,col=couleurs,main="")
```



Représentations graphiques

```
> boxplot(SAL_HOR ~ SEXE,xlab="Sexe",ylab="SALAIRE")  
> abline(h=mean(SAL_HOR,na.rm=T),lty=2,col="red",lwd=2)
```

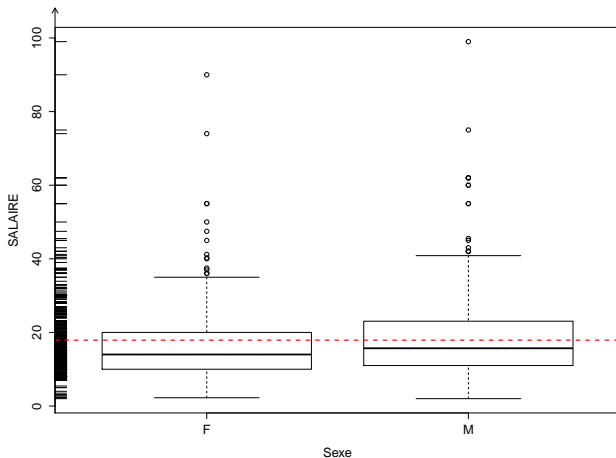


Tableau de contingence : AGE x SALAIRE

⇒ Tableau de contingence Quantitatif x Quantitatif.

Salaire	éage (ans) horaire	[16 ;32[[32 ;48[[48 ;64[[64 ;80]	Total
[2 ;26[180	156	144	26	506
[26 ;50[11	28	40	1	80
[50 ;76[0	5	4	1	10
[76 ;100]		1	0	1	1	3
Total		192	189	189	29	599

- X à $p = \dots$ classes.
- Y à $q = \dots$ classes.
- Mesures conjointes de X et Y sur $n = \dots$ individus.
- **4**

Tableau de contingence : AGE x SALAIRE

```
> NclasseS=4 # Nombre de classes : Salaire
> SALAIRE<-cut(SAL_HOR,breaks=NclasseS)
> NclasseA=4 # Nombre de classes : Age
> Age<-cut(AGE,breaks=NclasseA)
> TabContEf<-table(Age,SALAIRE) # Tableau de contingence : Effectif
> print(TabContEf)
```

	SALAIRE			
Age	(1.9,26.2]	(26.2,50.5]	(50.5,74.8]	(74.8,99.1]
(15.9,32]	180	11	0	1
(32,48]	156	28	5	0
(48,64]	144	40	4	1
(64,80.1]	26	1	1	1

```
> addmargins(TabContEf)
```

	SALAIRE				
Age	(1.9,26.2]	(26.2,50.5]	(50.5,74.8]	(74.8,99.1]	Sum
(15.9,32]	180	11	0	1	192
(32,48]	156	28	5	0	189
(48,64]	144	40	4	1	189
(64,80.1]	26	1	1	1	29
Sum	506	80	10	3	599

Tableau de contingence : AGE x SALAIRE

```
> TabContFr<-prop.table(TabContEf)    # Tableau de contingence : Frequence
```

```
> print(round(TabContFr,2))
```

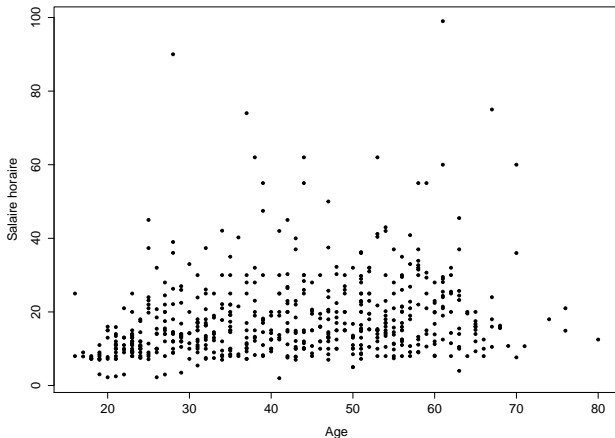
Age	SALAIRE			
	(1.9,26.2]	(26.2,50.5]	(50.5,74.8]	(74.8,99.1]
(15.9,32]	0.30	0.02	0.00	0.00
(32,48]	0.26	0.05	0.01	0.00
(48,64]	0.24	0.07	0.01	0.00
(64,80.1]	0.04	0.00	0.00	0.00

```
> addmargins(round(TabContFr,2))
```

Age	SALAIRE				Sum
	(1.9,26.2]	(26.2,50.5]	(50.5,74.8]	(74.8,99.1]	
(15.9,32]	0.30	0.02	0.00	0.00	0.32
(32,48]	0.26	0.05	0.01	0.00	0.32
(48,64]	0.24	0.07	0.01	0.00	0.32
(64,80.1]	0.04	0.00	0.00	0.00	0.04
Sum	0.84	0.14	0.02	0.00	1.00

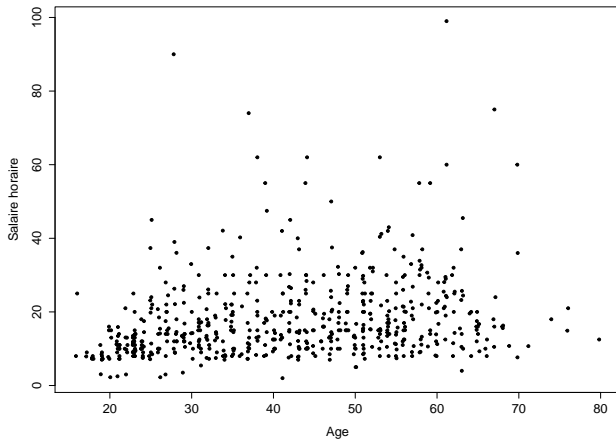
Représentation graphique plus appropriée

```
> plot(AGE,SAL_HOR,pch=20,xlab="Age",ylab="Salaire horaire",main="")
```



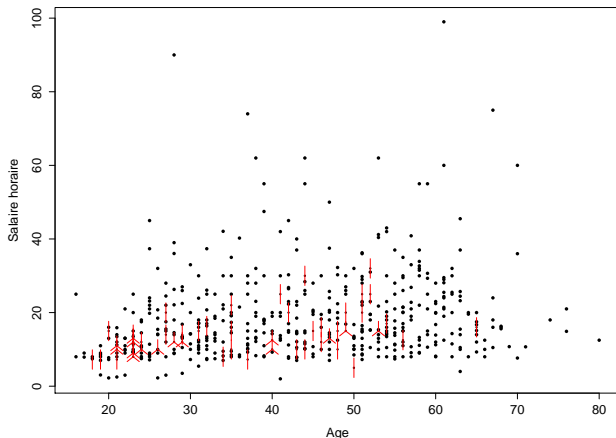
Problème des points superposés

```
> # Solution 1 : Bruitage des points  
> plot(jitter(AGE), jitter(SAL_HOR), pch=20, xlab="Age",  
       ylab="Salaire horaire", main="")
```



Problème des points superposés

```
> # Solution 2 : symboles adapt\'es  
> sunflowerplot(x=AGE,y=SAL_HOR,,xlab="Age",ylab="Salaire horaire",  
main="")
```



Distributions marginales

- **Distribution marginale de X en effectifs et en fréquences**

$$\{(x_i, n_{i\bullet}) ; 1 \leq i \leq p\} \quad \{(x_i, f_{i\bullet}) ; 1 \leq i \leq p\}$$

⇒ Dernière colonne du tableau de contingence en effectifs ou fréquences

- **Distribution marginale de Y en effectifs et en fréquences**

$$\{(y_j, n_{\bullet j}) ; 1 \leq j \leq q\} \quad \{(y_j, f_{\bullet j}) ; 1 \leq j \leq q\}$$

⇒ Dernière ligne du tableau de contingence en effectifs ou fréquences

Tableaux des effectifs/fréquences de X et de Y

X	effectif	fréquence
x_1	$n_{1\bullet}$	$f_{1\bullet}$
x_2	$n_{2\bullet}$	$f_{2\bullet}$
\vdots	\vdots	\vdots
x_i	$n_{i\bullet}$	$f_{i\bullet}$
\vdots	\vdots	\vdots
x_p	$n_{p\bullet}$	$f_{p\bullet}$
Total	$n = \sum_{i=1}^p n_{i\bullet}$	1

Dist. marginale de X
en eff. et en fréq.

Y	effectif	fréquence
y_1	$n_{\bullet 1}$	$f_{\bullet 1}$
y_2	$n_{\bullet 2}$	$f_{\bullet 2}$
\vdots	\vdots	\vdots
y_j	$n_{\bullet j}$	$f_{\bullet j}$
\vdots	\vdots	\vdots
y_q	$n_{\bullet q}$	$f_{\bullet q}$
Total	$n = \sum_{j=1}^q n_{\bullet j}$	1

Dist. marginale de Y
en eff. et en fréq.

Distributions marginales : SEXE et REGION

Y Sexe X	WE	NW	S	W	Total
Femme	61	62	97	77	297
Homme	68	60	103	71	302
Total	129	122	200	148	599

```
> margin.table(TabContEf,1)
```

SEXE

```
  F    M
297 302
```

```
> margin.table(TabContEf,2)
```

REGION

```
  NE  NW   S   W
129 122 200 148
```

Distributions marginales : SEXE et REGION

Y Sexe X	WE	NW	S	W	Total
Femme	0.1018	0.1035	0.1619	0.1285	0.495
Homme	0.1135	0.1002	0.1720	0.1185	0.504
Total	0.215	0.203	0.333	0.247	1

```
> margin.table(TabContFr,1)
```

SEXE

	F	M
	0.4958264	0.5041736

```
> margin.table(TabContFr,2)
```

REGION

	NE	NW	S	W
	0.215	0.203	0.333	0.247

Distributions conditionnelles

- Distributions conditionnelles de X sachant Y (colonne fixée) et de Y sachant X (ligne fixée)
- En effectifs, pour tout $i = 1, \dots, p$ et $j = 1, \dots, q$

▶ n_{ij} : nombre d'individus tq $X = x_i$ **et** $Y = y_j$

▶ $n_{i/j}$: nombre d'individus tq $X = x_i$ **parmi** ceux pour lesquels $Y = y_j$

$$n_{i/j} = n_{ij} \quad \text{avec } j \text{ fixé}$$

▶ $n_{j/i}$: nombre d'individus tq $Y = y_j$ **parmi** ceux pour lesquels $X = x_i$

$$n_{j/i} = n_{ij}, \quad i \text{ fixé}$$

- En fréquences, pour tout $i = 1, \dots, p$ et $j = 1, \dots, q$,

- ▶ f_{ij} : proportion d'individus tq $X = x_i$ **et** $Y = y_j$
- ▶ $f_{i/j}$: proportion d'individus pour lesquels $X = x_i$ **parmi** ceux pour lesquels $Y = y_j$.

$$f_{i/j} = \frac{n_{ij}}{n_{\bullet j}}$$

- ▶ $f_{j/i}$: proportion d'individus pour lesquels $Y = y_j$ **parmi** ceux pour lesquels $X = x_i$

$$f_{j/i} = \frac{n_{ij}}{n_{i\bullet}}$$

Distributions conditionnelles en effectifs et fréquences

- **Distribution conditionnelle en effectifs de X sachant $Y = y_j$**

$$\{(x_i, n_{i/j}) ; 1 \leq i \leq p, j \text{ fixé}\}$$

($j^{\text{ème}}$ colonne du tableau de contingence en effectifs)

- **Distribution conditionnelle en effectifs de Y sachant $X = x_i$**

$$\{(y_j, n_{j/i}) ; 1 \leq j \leq q, i \text{ fixé}\}$$

($i^{\text{ème}}$ ligne du tableau de contingence en effectifs).

- **Distribution conditionnelle en fréquences de X sachant $Y = y_j$:**

$$\{(x_i, f_{i/j}) ; 1 \leq i \leq p, j \text{ fixé}\}$$

- **Distribution conditionnelle en fréquences de Y sachant $X = x_i$:**

$$\{(y_j, f_{j/i}) ; 1 \leq j \leq q, i \text{ fixé}\}$$

$X/Y = y_j$	effectif	fréquence.
x_1	$n_{1/j} = n_{1j}$	$f_{1/j} = \frac{n_{1j}}{n_{\bullet j}}$
x_2	$n_{2/j} = n_{2j}$	$f_{2/j} = \frac{n_{2j}}{n_{\bullet j}}$
\vdots	\vdots	\vdots
x_i	$n_{i/j} = n_{ij}$	$f_{i/j} = \frac{n_{ij}}{n_{\bullet j}}$
\vdots	\vdots	\vdots
x_p	$n_{p/j} = n_{pj}$	$f_{p/j} = \frac{n_{pj}}{n_{\bullet j}}$
Total	$n_{\bullet j} = \sum_{i=1}^p n_{ij}$	1

Dist. cond. eff. et fréq. de X sachant $Y = y_j$

$Y/X = x_i$	effectif	fréquence
y_1	$n_{1/i} = n_{i1}$	$f_{1/i} = \frac{n_{i1}}{n_{i\bullet}}$
y_2	$n_{2/i} = n_{i2}$	$f_{2/i} = \frac{n_{i2}}{n_{i\bullet}}$
\vdots	\vdots	\vdots
y_j	$n_{j/i} = n_{ij}$	$f_{j/i} = \frac{n_{ij}}{n_{i\bullet}}$
\vdots	\vdots	\vdots
y_q	$n_{q/i} = n_{iq}$	$f_{q/i} = \frac{n_{iq}}{n_{i\bullet}}$
Total	$n_{i\bullet} = \sum_{j=1}^q n_{ij}$	1

Dist. cond .eff. et fréq. de Y sachant $X = x_i$

- Il y a q distributions conditionnelles de X sachant $Y = y_j$ (autant que les q modalités ou classes de Y)
- Il y a p distributions conditionnelles de Y sachant $X = x_i$ (autant que les p modalités ou classes de X)

Tableau des q distributions conditionnelles de X sachant Y

Distribution conditionnelle de X sachant $Y = y_j$ dans la colonne j

X	Y	y_1	y_2	\dots	y_j	\dots	y_q
x_1		$f_{1/1}$	$f_{1/2}$	\dots	$f_{1/j}$	\dots	$f_{1/q}$
x_2		$f_{2/1}$	$f_{2/2}$	\dots	$f_{2/j}$	\dots	$f_{2/q}$
\vdots		\vdots	\vdots		\vdots		\vdots
x_i		$f_{i/1}$	$f_{i/2}$	\dots	$f_{i/j}$	\dots	$f_{i/q}$
\vdots		\vdots	\vdots		\vdots		\vdots
x_p		$f_{p/1}$	$f_{p/2}$	\dots	$f_{p/j}$	\dots	$f_{p/q}$
Total		1	1	\dots	1	\dots	1

Tableau des p distributions conditionnelles de Y sachant X

Distribution conditionnelle de Y sachant $X = x_i$ dans la ligne i

X	Y	y_1	y_2	\dots	y_j	\dots	y_q	Total
	x_1	$f_{1/1}$	$f_{2/1}$	\dots	$f_{j/1}$	\dots	$f_{q/1}$	1
	x_2	$f_{1/2}$	$f_{2/2}$	\dots	$f_{j/2}$	\dots	$f_{q/2}$	1
	\vdots	\vdots	\vdots		\vdots		\vdots	
	x_i	$f_{1/i}$	$f_{2/i}$	\dots	$f_{j/i}$	\dots	$f_{q/i}$	1
	\vdots	\vdots	\vdots		\vdots		\vdots	
	x_p	$f_{1/p}$	$f_{2/p}$	\dots	$f_{j/p}$	\dots	$f_{q/p}$	1

Exemple : Distributions conditionnelles en effectifs de Y sachant X

Salaire Y	[2 ;26[[26,50[[50,76[[76,100[Total
Sexe X					
Femme	262	31	3	1	297
Homme	244	49	7	2	302
Total	506	80	10	3	599

- **Dist. cond. en effectifs du salaire horaire chez (sachant que) les hommes**

Parmi les hommes, il y a personnes qui gagnent entre 2 et 26 dollars.

- Sur les personnes observées, ... sont des hommes et gagnent entre 2 et 26 dollars.

Exemple : Distributions conditionnelles en effectifs de X sachant Y

Salaire Y	[2 ;26[[26,50[[50,76[[76,100[Total
Sexe X					
Femme	262	31	3	1	297
Homme	244	49	7	2	302
Total	506	80	10	3	599

- **Dist. cond. en effectifs du sexe sachant que le salaire horaire est compris entre 2 et 26 dollars.**

Parmi les ... personnes qui gagnent entre 2 et 26 dollars, il y a ... hommes.

- Sur les ... personnes observés, ... sont des hommes et gagnent entre 2 et 26 dollars.

Tableau des $q = 4$ distributions conditionnelles en fréquences du sexe X sachant le salaire horaire Y

Salaire Y	[2 ;26[[26,50[[50,76[[76,100[Total
Sexe X					
Femme	52%	39%	30%	33%	50%
Homme	48%	61%	70%	67%	50%
Total	100%	100%	100%	100%	100%

- Dist. cond. en fréquences du sexe sachant que le salaire horaire est compris entre 2 et 26 dollars.
- Parmi les ... personnes qui gagnent entre 2 et 26 dollars, il y en a ...% hommes.
- Sur les personnes observées,% sont des hommes et gagnent entre 2 et 26 dollars

Tableau des $p = 3$ distributions conditionnelles en fréquences du salaire horaire Y selon le sexe X

Salaire Y	[2 ;26[[26,50[[50,76[[76,100[Total
Sexe X					
Femme	88%	10%	1%	1%	100%
Homme	81%	16%	2%	1%	100%
dist. marg. de Y	85%	13%	2%	0%	100%

- Dist. cond. en fréquences de l'âge sachant la catégorie de personnel.
- Parmi les ... hommes, il y a ...% des personnes qui gagnent entre 2 et 26 dollars.
- Sur les ... personnes observés, ...% sont des hommes et gagnent entre 2 et 26 dollars

Moyennes, variances marginales et conditionnelles

- **UNIQUEMENT** pour variables quantitatives.
- Données brutes : calculs similaires à ceux effectués en statistique univariée après extraction des individus d'intérêt.
- Données groupées : à partir des tableaux de contingence.

Moyennes et variances marginales

- Distribution marginale de X en effectifs/fréquences

$$\{(x_i, n_{i\bullet}) ; 1 \leq i \leq p\} \quad \{(x_i, f_{i\bullet}) ; 1 \leq i \leq p\}$$

- Distribution marginale de Y en effectifs/fréquences

$$\{(y_j, n_{\bullet j}) ; 1 \leq j \leq q\} \quad \{(y_j, f_{\bullet j}) ; 1 \leq j \leq q\}$$

- Moyennes marginales \bar{x} et \bar{y}

$$\bar{x} = \frac{1}{n} \sum_{i=1}^p n_{i\bullet} x_i = \sum_{i=1}^p f_{i\bullet} x_i$$

$$\bar{y} = \frac{1}{n} \sum_{j=1}^q n_{\bullet j} y_j = \sum_{j=1}^q f_{\bullet j} y_j$$

Moyennes marginales : AGE et SALAIRE

```
> print(TabContEf)
> addmargins(TabContEf)
```

	SALAIRE				
Age	(1.9,26.2]	(26.2,50.5]	(50.5,74.8]	(74.8,99.1]	Sum
(15.9,32]	180	11	0	1	192
(32,48]	156	28	5	0	189
(48,64]	144	40	4	1	189
(64,80.1]	26	1	1	1	29
Sum	506	80	10	3	599

```
> margin.table(TabContEf,1)
Age
(15.9,32]    (32,48]    (48,64]    (64,80.1]
      192         189         189         29

> margin.table(TabContEf,2)
SALAIRE
(1.9,26.2] (26.2,50.5] (50.5,74.8] (74.8,99.1]
      506         80         10         3

## A comparer avec :
```

```
> mean(AGE)
[1] 41.84975
> mean(SAL_HOR)
[1] 17.89835
```

- Variances marginales σ_x^2 et σ_y^2

$$V(x) = \sigma_x^2 = \frac{1}{n} \sum_{i=1}^p n_{i\bullet} (x_i - \bar{x})^2 = \sum_{i=1}^p f_{i\bullet} (x_i - \bar{x})^2$$

$$V(y) = \sigma_y^2 = \frac{1}{n} \sum_{j=1}^q n_{\bullet j} (y_j - \bar{y})^2 = \sum_{j=1}^q f_{\bullet j} (y_j - \bar{y})^2$$

- Soit aussi

$$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^p n_{i\bullet} x_i^2 - (\bar{x})^2 = \sum_{i=1}^p f_{i\bullet} x_i^2 - (\bar{x})^2$$

$$\sigma_y^2 = \frac{1}{n} \sum_{j=1}^q n_{\bullet j} y_j^2 - (\bar{y})^2 = \sum_{j=1}^q f_{\bullet j} y_j^2 - (\bar{y})^2$$

Variances marginales : AGE et SALAIRE

```
> print(TabContEf)
> addmargins(TabContEf)
```

	SALAIRE				
Age	(1.9,26.2]	(26.2,50.5]	(50.5,74.8]	(74.8,99.1]	Sum
(15.9,32]	180	11	0	1	192
(32,48]	156	28	5	0	189
(48,64]	144	40	4	1	189
(64,80.1]	26	1	1	1	29
Sum	506	80	10	3	599

```
> margin.table(TabContEf,1)
Age
(15.9,32]    (32,48]    (48,64]    (64,80.1]
      192         189         189         29
> margin.table(TabContEf,2)
SALAIRE
(1.9,26.2] (26.2,50.5] (50.5,74.8] (74.8,99.1]
      506         80         10         3
## A comparer avec :

> var(AGE)
[1] 199.275
> var(SAL_HOR)
[1] 127.2247
```

Moyennes et variances conditionnelles

Pour $j = 1, \dots, q$

- Dist. cond. de X en effectifs/fréquences sachant que $Y = y_j$

$$\{(x_i, n_{i/j}) ; 1 \leq i \leq p\} \quad \{(x_i, f_{i/j}) ; 1 \leq i \leq p\}$$

avec

$$n_{i/j} = n_{ij} \quad \text{et} \quad f_{i/j} = \frac{n_{ij}}{n_{\bullet j}}$$

- Moyenne conditionnelle de X sachant que $Y = y_j$: $\bar{x}_{/j}$

$$\bar{x}_{/j} = \bar{x}_{/Y=y_j} = \frac{1}{n_{\bullet j}} \sum_{i=1}^p n_{i/j} x_i = \frac{1}{n_{\bullet j}} \sum_{i=1}^p n_{ij} x_i = \sum_{i=1}^p f_{i/j} x_i ;$$

- Variance conditionnelle de X sachant que $Y = y_j$: $\sigma_{x/j}^2$

$$\sigma_{x/j}^2 = V(x_{/Y=y_j}) = \frac{1}{n_{\bullet j}} \sum_{i=1}^p n_{i/j} (x_i - \bar{x}_{/j})^2 = \sum_{i=1}^p f_{i/j} (x_i - \bar{x}_{/j})^2 .$$

Pour $i = 1, \dots, p$

- Dist. cond. de Y en effectifs/fréquences sachant que $X = x_i$

$$\{(y_j, n_{j/i}) ; 1 \leq j \leq q\} \quad \{(y_j, f_{j/i}) ; 1 \leq j \leq q\}$$

avec

$$n_{j/i} = n_{ij} \quad \text{et} \quad f_{j/i} = \frac{n_{ij}}{n_{i\bullet}}$$

- Moyenne conditionnelle de Y sachant que $X = x_i$: $\bar{y}_{/i}$

$$\bar{y}_{/i} = \bar{y}_{/X=x_i} = \frac{1}{n_{i\bullet}} \sum_{j=1}^q n_{j/i} y_j = \frac{1}{n_{i\bullet}} \sum_{j=1}^q n_{ij} y_j = \sum_{j=1}^q f_{j/i} y_j ;$$

- Variance conditionnelle de Y sachant que $X = x_i$: $\sigma_{y/i}^2$

$$\sigma_{y/i}^2 = V(y_{/X=x_i}) = \frac{1}{n_{i\bullet}} \sum_{j=1}^q n_{j/i} (y_j - \bar{y}_{/i})^2 = \sum_{j=1}^q f_{j/i} (y_j - \bar{y}_{/i})^2 .$$

Autre écriture de la variance conditionnelle

$$\begin{aligned}\sigma_{x/j}^2 &= \frac{1}{n_{\bullet j}} \sum_{i=1}^p n_{ij} (x_i - \bar{x}_{/j})^2 \\ &= \frac{1}{n_{\bullet j}} \sum_{i=1}^p n_{ij} x_i^2 - (\bar{x}_{/j})^2 = \sum_{i=1}^p f_{i/j} x_i^2 - (\bar{x}_{/j})^2 \\ \sigma_{y/i}^2 &= \frac{1}{n_{i\bullet}} \sum_{j=1}^q n_{ji} (y_j - \bar{y}_{/i})^2 \\ &= \frac{1}{n_{i\bullet}} \sum_{j=1}^q n_{ji} y_j^2 - (\bar{y}_{/i})^2 = \sum_{j=1}^q f_{j/i} y_j^2 - (\bar{y}_{/i})^2\end{aligned}$$

Moyennes et variances conditionnelles du salaire par âge

```
> round(tapply(SAL_HOR, Age, mean),2)
(15.9,32]   (32,48]   (48,64]   (64,80.1]
   14.14     18.62     20.83     18.97
> round(tapply(SAL_HOR, Age, var),2)
(15.9,32]   (32,48]   (48,64]   (64,80.1]
   80.15    127.34    140.10    215.17

> tapply(SAL_HOR, Age, summary)
$`(15.9,32]`
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.25   9.00   12.00   14.14   16.34   90.00

$`(32,48]`
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.00  10.75   15.00   18.62   22.11   74.00

$`(48,64]`
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.00  13.00   19.00   20.83   25.72   99.00

$`(64,80.1]`
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  7.65  12.00   16.00   18.97   18.00   75.00
```

Moyennes et variances conditionnelles de l'âge par tranche de salaire

```
> round(tapply(AGE, SALAIRE, mean),2)
(1.9,26.2] (26.2,50.5] (50.5,74.8] (74.8,99.1]
  40.80      47.06      50.30      52.00

> round(tapply(AGE, SALAIRE, var),2)
(1.9,26.2] (26.2,50.5] (50.5,74.8] (74.8,99.1]
  205.08    123.86    131.12    441.00

> tapply(AGE,SALAIRE,summary)
$`(1.9,26.2]`
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.0   28.0   40.0   40.8   52.0   80.0

$`(26.2,50.5]`
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  25.00  38.75  49.00  47.06  56.25  70.00

$`(50.5,74.8]`
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  37.00  40.25  48.50  50.30  58.75  70.00

$`(74.8,99.1]`
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  28.0   44.5   61.0   52.0   64.0   67.0
```

Lien entre moyennes marginales et conditionnelles

⇒ On peut retrouver la moyenne marginale (générale) en calculant la moyenne pondérée des moyennes conditionnelles.

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{j=1}^q n_{\bullet j} \bar{x}_{/j} = \sum_{j=1}^q f_{\bullet j} \bar{x}_{/j} \\ \bar{y} &= \frac{1}{n} \sum_{i=1}^p n_{i\bullet} \bar{y}_{/i} = \sum_{i=1}^p f_{i\bullet} \bar{y}_{/i}\end{aligned}$$

Décomposition de la variance

⇒ On peut pas retrouver la variance marginale à partir des variances conditionnelles.

Variance marginale = variance des moyennes conditionnelles + moyenne des variances conditionnelles.

$$\begin{aligned}\sigma^2_{\mathbf{x}} &= \frac{1}{n} \sum_{j=1}^q n_{\bullet j} (\bar{\mathbf{x}}_{/j} - \bar{\mathbf{x}})^2 + \frac{1}{n} \sum_{j=1}^q n_{\bullet j} \sigma^2_{\mathbf{x}/j} \\ \sigma^2_{\mathbf{y}} &= \frac{1}{n} \sum_{i=1}^p n_{i\bullet} (\bar{\mathbf{y}}_{/i} - \bar{\mathbf{y}})^2 + \frac{1}{n} \sum_{i=1}^p n_{i\bullet} \sigma^2_{\mathbf{y}/i}\end{aligned}$$

Remarque

- On peut calculer les moyennes et les variances conditionnelles d'une variable quantitative sachant les modalités d'une variable qualitative.
- Mais la réciproque est fausse ! Evident !

Exemple :

SEXE X	Salaire Y	[2 ;26[[26,50[[50,76[[76,100[Total
Femme		88%	10%	1%	1%	100%
Homme		81%	16%	2%	1%	100%
dist. marg. de Y		85%	13%	2%	0%	100%

- Le salaire horaire moyen de l'ensemble des personnes observés est de 17,9 dollars $\Rightarrow \bar{y} = 17,9$
- Le salaire horaire moyen des femmes A est de 16,6 dollars $\Rightarrow \bar{y}_{/A} = 16,6$
- Le salaire horaire moyen des hommes B est de 19,17 dollars $\Rightarrow \bar{y}_{/B} = 19,17$
- $17,9 = \bar{y} = \sum_{i=1}^p f_{i\bullet} \bar{y}_{/i} = 0,4959 * 16,6 + 0,5041 * 19,17$
 \Rightarrow voir Slide 52.

Exemple :

SEXE X	Salaire Y	[2;26[[26,50[[50,76[[76,100[Total
Femme		88%	10%	1%	1%	100%
Homme		81%	16%	2%	1%	100%
dist. marg. de Y		85%	13%	2%	0%	100%

- La variance marginale du salaire horaire est $\sigma_y^2 = 127,22$.
- La variance du salaire horaire des femmes est $\sigma_{y/F}^2 = 105,84$.
- La variance du salaire horaire des hommes est $\sigma_{y/M}^2 = 145,39$.

Moyennes et variances conditionnelles du salaire horaire par sexe

```
> round(tapply(SAL_HOR, SEXE, mean),2)
```

F	M
16.60	19.17

```
> round(tapply(SAL_HOR, SEXE, var),2)
```

F	M
105.84	145.39

```
> tapply(SAL_HOR,SEXE,summary)
```

\$F

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.25	10.00	14.00	16.60	20.00	90.00

\$M

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.00	11.00	15.70	19.17	23.04	99.00