L'ANALYSE DES DONNEES

Samuel MAYOL

UTILISATIONS

La statistique est l'ensemble des méthodes scientifiques à partir desquelles sont recueillies, présentées, résumées et analysées les données.

Deux niveaux:

Stat. Descriptive sert à décrire des ensembles nombreux et dégager l'essentiel de l'information qui en résulte.

Stat. Théorique permet d'énoncer et d'élaborer des lois.

I - TABLEAUX STATISTIQUES

- **1) CAS D'UNE SEULE VARIABLE**

- Le tableau brut se présente sous la forme suivante:

Individu	variable
1	x_1
2	x_2
.	.
.	.
.	.
n	x_n

- Le nombre d'individus observé étant en général important, le tableau précédent ne permet pas d'analyser l'information obtenue. Il est donc nécessaire de créer un tableau plus synthétique où les observations identiques (possédant la même modalité) ont été regroupées.

modalité	effectif
C_1	n_1
C_2	n_2
.	.
.	.
.	.
C_k	n_k

- Pour une variable qualitative, les modalités sont susceptibles d'être classées en groupes d'individus.
- Pour une variable quantitative, les modalités ne sont pas classées
- Exemple: le sexe est une variable qualitative alors que l'âge ne l'est pas. Néanmoins toutes les variables peuvent être transformées en variable qualitative, par regroupement. Ainsi, le variable âge peut permettre une classification des individus en classes d'âge.

2) CAS DE DEUX VARIABLES

Individu	variable1	variable2
1	x_1	y_1
2	x_2	y_2
.	.	.
.	.	.
.	.	.
n	x_n	y_n

Le tableau brut se présente sous la forme suivante

On désire créer un tableau appelé tableau de contingence donnant le nombre d'individus possédant simultanément la modalité i de variable1 et la modalité j de variable2 qui se présentera sous la forme suivante:

		Variable2					
		D_1	. .	D_j	. .	D_r
variable1	C_1	n_{11}	. .	n_{1j}	. .	n_{1r}
	.	.		.		.
	.	.		.		.
	C_i	n_{i1}	. .	n_{ij}	. .	n_{ir}
	.	.		.		.
	.	.		.		.
	C_k	n_{k1}	. .	n_{kj}	. .	n_{kr}

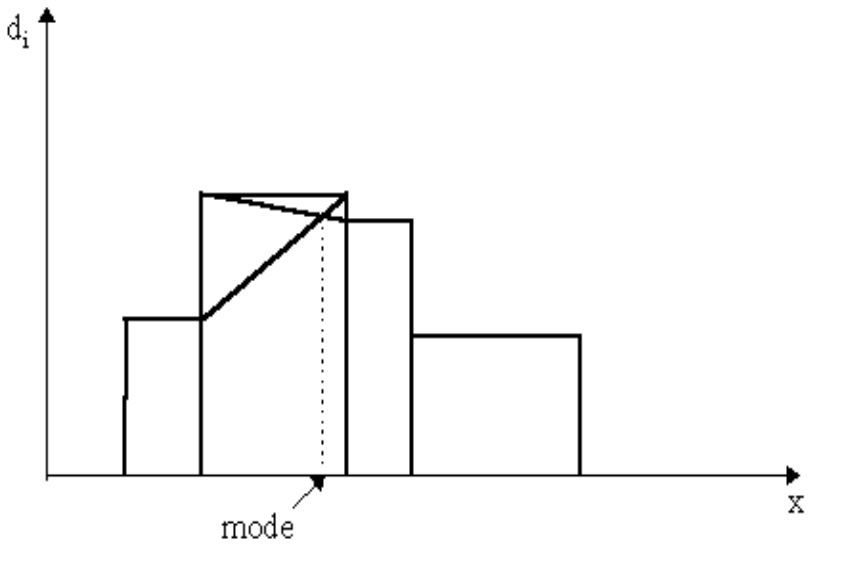
II - TENDANCE CENTRALE

1) MODE

C'est la valeur observée d'effectif maximum.

Variable discrète: classer les données par ordre croissant. Celle d'effectif maximum donne le mode.

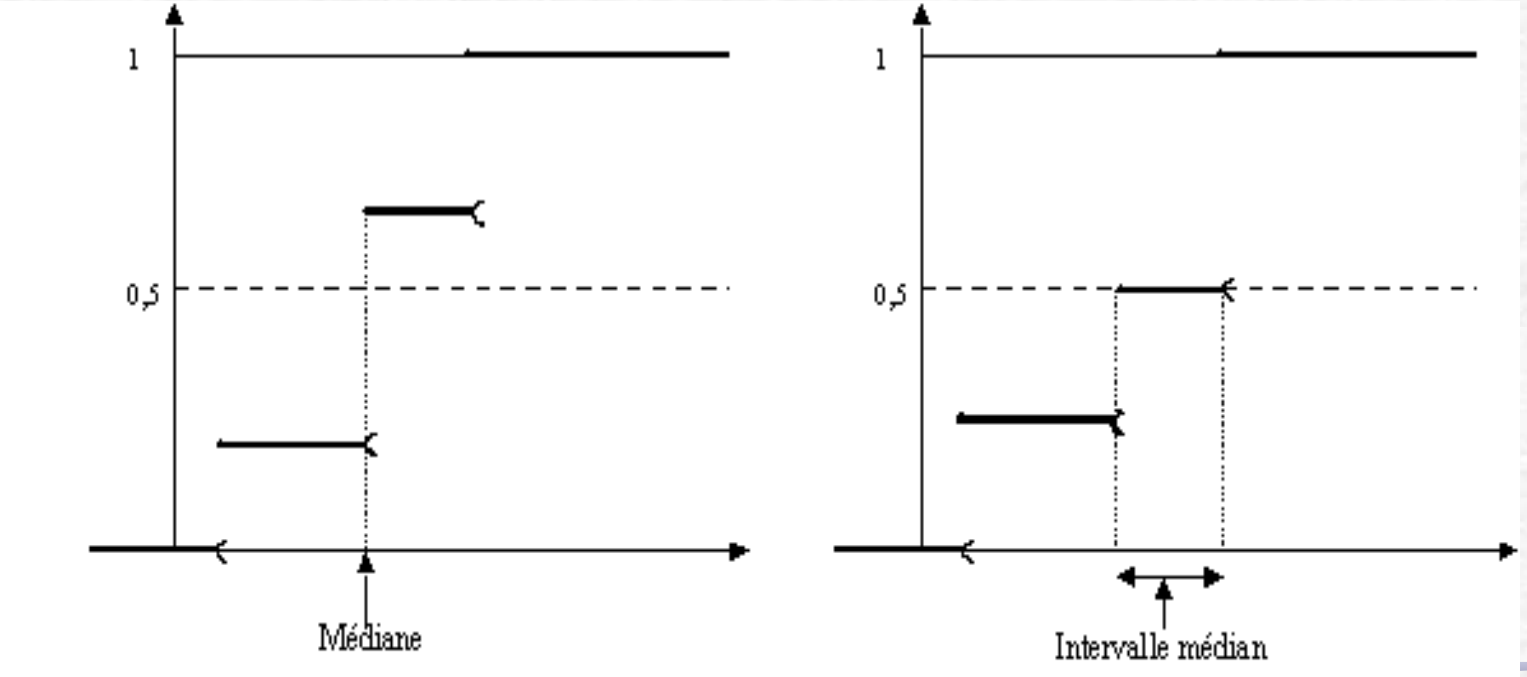
Il est fortement conseillé d'utiliser le diagramme en bâtons pour déterminer le mode. En effet, deux valeurs consécutives xi , xi+1 peuvent avoir le même effectif maximum; on parlera d'intervalle modal [xi , xi+1]. Il peut aussi y avoir un mélange de deux populations qui conduit à un diagramme en bâtons où apparaissent deux bosses; on considérera deux modes. Il est déconseillé, sauf raison explicite, d'envisager plus de deux modes.



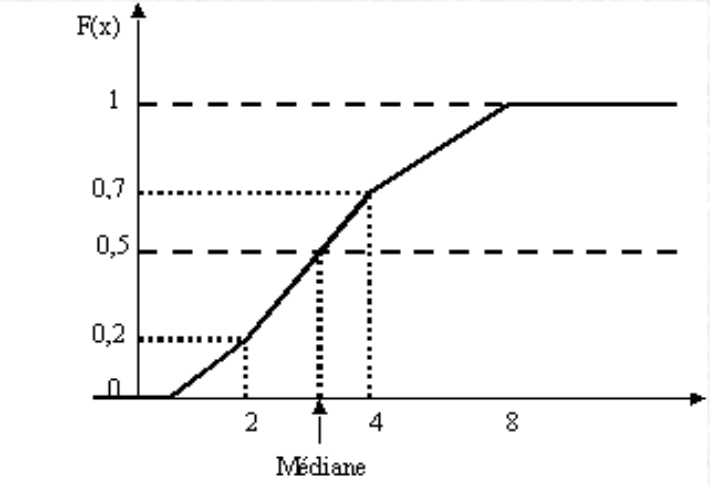
Variable classée: la classe modale correspond à la classe ayant l'effectif maximum. Il est fortement conseillé d'utiliser l'histogramme pour déterminer le mode. Comme pour le cas discret, on peut avoir deux classes modales. Toutes les valeurs de la classe pouvant à priori se réaliser, on ne se contentera pas de déterminer la classe modale. Une des valeurs de cette classe sera le mode. Certains auteurs préconisent par simplicité de prendre le centre de la classe modale. Il est préférable cependant de tenir compte des classes adjacentes de la manière suivante:

2) MEDIANE

- Les valeurs étant rangées par ordre croissant, c'est la valeur de la variable qui sépare les observations en deux groupes d'effectifs égaux.
- Variable discrète: la détermination peut s'obtenir à partir du tableau statistique en recherchant la valeur de la variable correspondant à une fonction cumulée égale à n/2 (effectif cumulé) ou 1⁄2 (fréquence cumulée). Il est encore plus facile de lire sur les graphiques cumulatifs les abscisses des points d'ordonnée n/2 (effectif cumulé) ou 1⁄2 (fréquence cumulée). Si tout un intervalle a pour image n/2 (1⁄2 pour la fréquence), on parlera d'intervalle médian (on peut prendre le milieu de l'intervalle comme médiane)



- Variable classée: l'abscisse du point d'ordonnée n/2 (1⁄2 pour la fréquence)se situe en général à l'intérieur d'une classe. Pour obtenir une valeur plus précise de la médiane, on procède à une interpolation linéaire. La valeur de la médiane peut être lue sur le graphique ou calculée analytiquement.



d'où la valeur de la médiane.

De manière générale, si a et b sont les bornes de la classe contenant la médiane, F(a) et F(b) les valeurs de la fréquence cumulée croissante en a et b, alors

$$\frac{\text{Mé} - 2}{4 - 2} = \frac{0,5 - 0,2}{0,7 - 0,2}$$

$$\text{Mé} = a + (b - a) \times \frac{0,5 - F(a)}{F(b) - F(a)}$$

3) MOYENNE ARITHMETIQUE

- Si xi sont les observations d'une variable discrète ou les centres de classe d'une variable classée,

la moyenne arithmétiq ue $\bar{x}$ est égale à $\sum_{i=1}^{k} \frac{n_i x_i}{n} = \sum_{i=1}^{k} f_i x_i$

La moyenne arithmétique est un paramètre de tendance centrale plus utilisé que les autres de par ses propriétés algébriques:

a) Pour plusieurs populations d'effectifs n1, n2, ..., nk, de moyennes respectives

moyenne globale = moyenne des moyennes

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{n}$$

b) La moyenne arithmétique conserve les changements d'échelle et d'origine

$$x : (x_i, n_i) \rightarrow y : (y_i = a x_i + b, n_i)$$

$$\bar{x} \rightarrow \bar{y} = a\bar{x} + b$$

4) QUANTILES

Ce sont des caractéristiques de position.

Il y a 1 médiane Mé qui sépare les observations en 2 groupes d'effectifs égaux

3 quartiles Q1, Q2, Q3 qui séparent les observations en 4 groupes d'effectifs égaux

9 déciles D1, D2, ..., D9 qui séparent les observations en 10 groupes d'effectifs égaux

99 centiles C1, C2, ..., C99 qui séparent les observations en 100 groupes d'effectifs égaux

La détermination de ces caractéristiques est identique à celle de la médiane.

Les quartiles sont obtenus lorsqu'on a cumulé 25, 50, 75% de la population

Les déciles sont obtenus lorsqu'on a cumulé 10, 20,...., 90% de la population

Les centiles sont obtenus lorsqu'on a cumulé 1, 2,...., 99% de la population

Remarque: la notion de déciles et de centiles n'a de sens que s'il y a beaucoup d'observations et donc essentiellement pour une variable classée.

III - DISPERSION

Comme leur nom l'indique, ces caractéristiques essayent de synthétiser par une seule valeur numérique la dispersion de toutes les valeurs observées.

1) ÉTENDUE

C'est la différence entre la plus grande et la plus petite observation

2) INTERVALLE INTER-QUARTILE

C'est la différence entre le troisième et le premier quartile

3) VARIANCE ET ECART-TYPE

Si xi sont les observations d'une variable discrète ou les centres de classe d'une variable classée, la variance

V est égale à $\sum_{i=1}^{k} \frac{n_i (x_i - \bar{x})^2}{n} = \sum_{i=1}^{k} f_i (x_i - \bar{x})^2$

On a aussi $V = \sum_{i=1}^{k} \frac{n_i x_i^2}{n} - \bar{x}^2$

c.à.d. moyenne des carrés - carré de la moyenne

- On utilise plus couramment l'écart-type qui est la racine carrée de la variance et qui a l'avantage d'être un nombre de même dimension que les données (contrairement à la variance qui en est le carré)
- La variance est un paramètre de dispersion plus utilisé que les autres de par ses propriétés algébriques:

a) Pour plusieurs populations d'effectifs $n_1, n_2, \ldots, n_k$, de moyennes respectives $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$, de variances $V_1, V_2, \ldots, V_k$

Variance globale = variance des moyennes + moyenne des variances

$$V = \sum_{i=1}^{k} \frac{n_i}{n}(\bar{x}_i - \bar{\bar{x}})^2 + \sum_{i=1}^{k} \frac{n_i}{n} V_i$$

où $\bar{\bar{x}}$ représente la moyenne des moyennes

b) changement d'échelle et d'origine

$$x : (x_i, n_i) \rightarrow y : (y_i = ax_i + b, n_i)$$

$$V_x \rightarrow V_y = a^2 V_x$$

4) COEFFICIENT DE VARIATION

$$CV = \frac{\sigma}{\bar{x}}$$

C'est un coefficient qui permet de relativiser l'écart-type en fonction de la taille des valeurs. Il permet ainsi de comparer la dispersion de séries de mesures exprimées dans des unités différentes

IV - LA CORRELATION

- Lorsqu'on observe deux variables quantitatives sur les mêmes individus, on peut s'intéresser à une liaison éventuelle entre ces deux variables.
- Le coefficient de corrélation r permet de mesurer un lien entre deux variables continues X et Y, sans que celui-ci ne soit forcément un lien de causalité. Il est calculé avec la formule.

$$r_{XY} = \frac{COV(X;Y)}{\sqrt{(var(X)\ var(Y)}}$$

Où la variance de X, var (X), est égale à la moyenne de X moins la moyenne de X au carré, et la covariance de X et Y, cov (X, Y) est égale à la moyenne de XY moins le produit de la moyenne de X et de la moyenne de Y.

L'interprétation de la valeur de r est la suivante

r proche de 0 = pas de lien entre X et Y
r proche de 1 = lien fort et dans le même sens
r proche de -1 = lien fort en sens contraire

Exemple
On mesure sur 7 enfants leur âge et le nombre d'heures passées par jour devant la télévision. Observe t-on un lien ?

X = AGE	5	6	7	13	15	18	10
Y = TEMPS	2	2	2	4	3	3	2

Les calculs sont les suivants

X	5	6	7	13	15	18	10
Y	2	2	2	4	3	3	2
X^2	25	36	49	169	225	324	100
Y^2	4	4	4	16	9	9	4
XY	10	12	14	52	45	54	20

Les moyennes sont

X	10,57
Y	2,57
X^2	132,57
Y^2	7,14
XY	29,57

D’où

VAR (X) = $132,57-10,57^2 = 20,85$
VAR (Y) = $7,14-2,57^2 = 0,54$
COV (X,Y) = 29,57 - 10,57x2,57 = 2,4
r = 0,72

La corrélation est donc assez forte, et le temps passé devant la télévision semble augmenter avec l’âge. Cependant, il faut savoir si ce résultat est significatif, c’est à dire si ce coefficient de corrélation est significativement différent de 0. Le test z de Fisher permet de le savoir.
Dans ce test, on cherche à savoir si le coefficient de corrélation es□□t différent ou non d’une certaine valeur $r_0 = 0$

Pour effectuer ce test, on calcule la quantité $(Z-Z_0)\sqrt{(n-3)}$ où n est le nombre d'individus et

$$Z0 = 0{,}5\mathrm{Ln}\left(\frac{1+r_0}{1-r_0}\right)$$

et

$$Z0 = 0{,}5\mathrm{Ln}\left(\frac{1+r}{1-r}\right)$$

Si cette quantité est supérieure à 2, les deux coefficients de corrélation r_0 et r sont significativement différents l'un de l'autre.

Ici Z = 1,8

Le coefficient de corrélation 0,72 n'est donc pas significativement différent de 0, ce qui signifie que le lien entre l'âge et le temps passé devant la télévision n'est pas significatif sur cet échantillon. La raison à cela est bien sûr la taille beaucoup trop faible de l'échantillon

IV - LA REGRESSION

Faire une régression linéaire entre une variable à expliquer et des Variables explicatives X1, X2, X3, …, c'est trouver la meilleure équation du type

$Y = a1X1+a2X2+a3X3+\ldots+anXn+b$

Le t de Student permet de savoir si cette variable est significative. Elle l'est lorsque la valeur absolue de t est supérieure à 1,96.

VI - LE KHI DEUX

Le test du khi 2, noté χ^2 , permet de savoir si un lien observé entre deux variables est significatif. Pour utiliser ce test, les variables doivent êtres qualitatives.

Testons par exemple le lien entre le genre des consommateurs et la possession d'un vêtement de marque Dyna, spécialisée dans le sport.

	Possèdent un Dyna	Ne possèdent pas de Dyna	Total
Hommes	10	40	50
Femmes	20	30	50
Total	30	70	100

Ce tableau indique, par exemple, que 20 femmes possèdent un Dyna

La proportion de femmes qui possèdent un Dyna est donc de 40%.
Chez les hommes, elle est de 20%.
Les femmes semblent donc plus susceptibles d'acheter cette marque.
Toutefois cette différence observée entre hommes et femmes pourrait uniquement résulter de fluctuations engendrées par un échantillon trop petit.
Il faut donc vérifier que cette différence n'est pas due à l'échantillonnage.

$$\chi^2 = n \left\{ \left(\sum_{\text{cases}} \frac{\text{case}2}{\sum \text{ligne} \times \sum \text{colonne}} \right) - 1 \right\}$$

Où

« case » = effectif de chacune des 4 cases du tableau

« ligne » et « colonne » = ligne et colonne du tableau où se trouve cette case

n = nombre total d'individus

$$\chi^2 = 100 \left\{ \left(\frac{40^2}{50X70} + \frac{30^2}{50X70} + \frac{10^2}{50X30} + \frac{20^2}{50X30} \right) - 1 \right\}$$

$$\chi^2 = 4{,}76$$

Le lien entre deux variables est statistiquement significatif quand Le χ^2 calculé est supérieur au χ^2 critique, qui dépend de la taille du tableau donné.

Quand les deux variables ont chacune deux modalités, comme c'est Le cas dans cet exemple, le degré de liberté v est égal à (nombre de lignes -1) X (nombre de colonnes -1)

Ici (2-1)X(2-1) = 1

Avec un v=1, on obtient un χ^2 critique = 3,84

χ^2 observé = 4,76 > χ^2critique = 3,84

Le lien observé ici est donc significatif malgré la petite taille de l'échantillon.

Le test du χ^2 permet également de vérifier s'il y a accord entre des effectifs issus de la réalité et des effectifs supposés à partir d'une théorie.
Dans ce cas on calcul les effectifs théorique de chaque phénomène observé.

Le calcul du χ^2 s'effectue selon la formule

$$\chi^2 \text{ cal} = \sum \frac{(\text{effectif théorique - effectif observé})^2}{\text{Effectif théorique}}$$

Exemple : les données suivantes concernent la consommation d'un jus de fruit auprès d'un échantillon de 220 femmes.

	0-1 enfant	2 enfants	+2 ans	Total	%
Oui , regulièrement	5	30	8	43	19,5
Oui, occasionnellement	20	50	45	115	52,3
Non	12	40	10	62	28,2
Total	37	120	63	220	100

	0-1 enfant	2 enfants	+2 ans	Total
Oui , regulièrement	8*	23	12	43
Oui, occasionnellement	19	63*	33	115
Non	19	34	18*	62
Total	37	120	63	220

* 37x19,5% = 8

* 120x52,3% = 63

* 63x28,2% = 18

$$\chi^2 = \frac{(8-5)^2}{8} + \frac{(23-30)^2}{23} + \frac{(12-8)^2}{12} + \frac{(19-20)^2}{19} + \frac{(63-50)^2}{63} + \frac{(33-45)^2}{33} +$$

$$\frac{(10-12)^2}{10} + \frac{(34-40)^2}{34} + \frac{(18-10)^2}{18} = 14{,}79$$

Le nombre de degrés de libertés ici est égal à 4.
La lecture de la table du de Pearson indique que la probabilité d'indépendance est inférieure à 1%. Le nombre d'enfants explique bien la propension à acheter des jus de fruits.

VII - L'ANALYSE EN COMPOSANTE PRINCIPALE (ACP)

Cette méthode réalise un exercice de sémantique en identifiant Des groupes de variables associées à des réponses similaires.

Elle permet une radioscopie des données.

Elle n'opère que sur des variables quantitatives (échelles ou numériques)

L'analyse en composantes principales est une technique de statistique descriptive qui calcule les axes principaux du nuage des observations regroupées selon les modalités de la variable à évaluer.
Les cartes factorielles sont des représentations graphiques issues de cette analyse. Ce sont des graphiques en deux dimensions, chacune représentant un des axes.
Les critères d'évaluation et les modalités de la variable à évaluer sont positionnées sur ce graphique. L'interprétation de leur position relative donnera des indications sur les rapport qu'entretiennent les variables.

Le pourcentage de variance qui caractérise chacun des axes est indiqué au coté de chacun des axes. La fonction Choisir les axes permet de modifier les composantes de la carte.

L'analyse en composantes principales suppose le calcul préalable des coefficients de corrélations entre deux couples de critères. On peut également présenter ces coefficients sous la forme de la Matrice des corrélations ou du Diagramme des corrélations (graphique dans lequel les variables corrélées sont reliées par un trait dont l'épaisseur est fonction du coefficient de corrélation).

Analyse en composantes principales

Variables :VITESSE, CONFORT, SECURITE.

VITESSE	Nb. cit.	Fréq.
Non réponse	25	12,5%
Pas du tout	7	3,5%
Plutôt non	17	8,5%
Moyennement	61	30,5%
Plutôt oui	45	22,5%
Tout à fait	45	22,5%
TOTAL OBS.	**200**	**100%**

Moyenne = 3,59 Ecart-type = 1,09

SECURITE	Nb. cit.	Fréq.
Non réponse	25	12,5%
Pas du tout	4	2,0%
Plutôt non	14	7,0%
Moyennement	69	34,5%
Plutôt oui	78	39,0%
Tout à fait	10	5,0%
TOTAL OBS.	**200**	**100%**

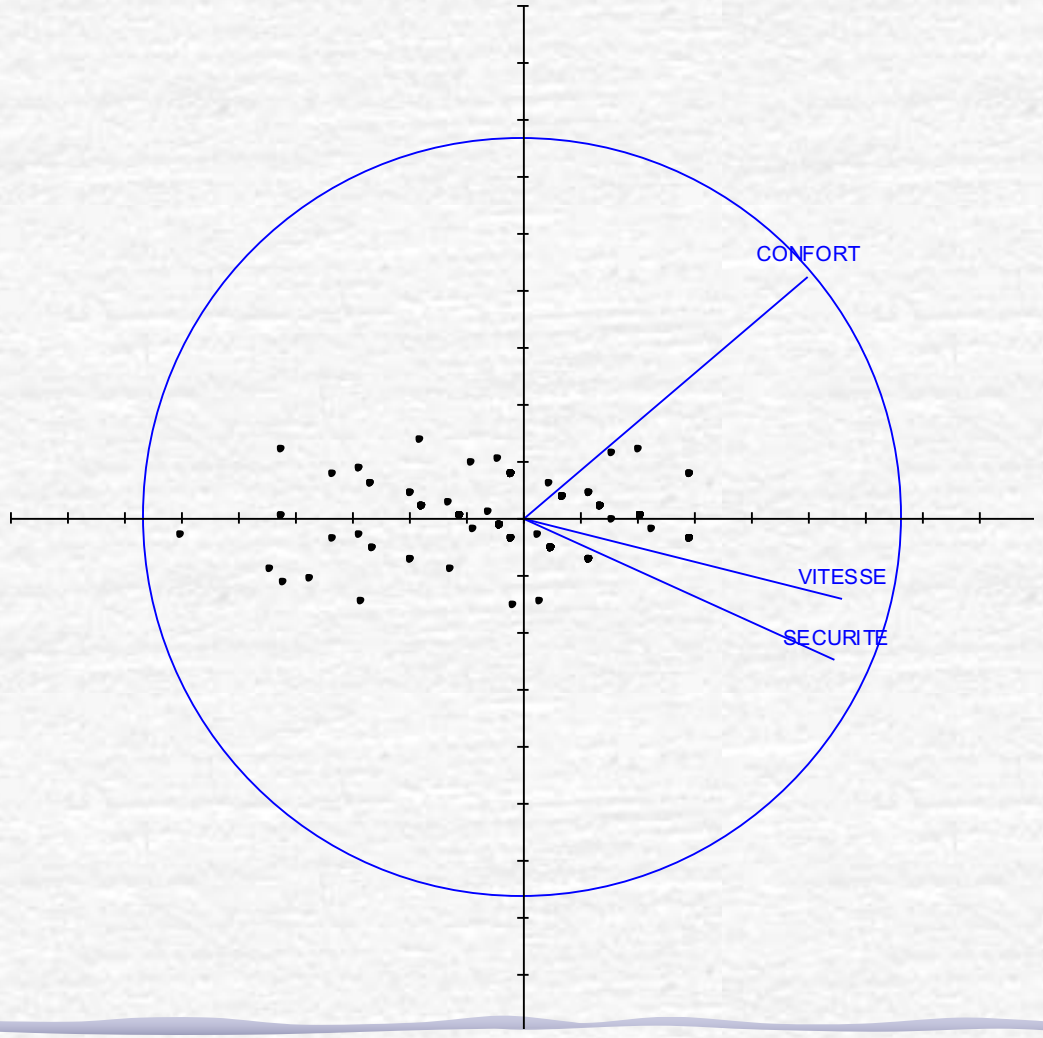
Moyenne = 3,43 Ecart-type = 0,81

CONFORT	Nb. cit.	Fréq.
Non réponse	25	12,5%
Pas du tout	5	2,5%
Plutôt non	11	5,5%
Moyennement	85	42,5%
Plutôt oui	69	34,5%
Tout à fait	5	2,5%
TOTAL OBS.	**200**	**100%**

Moyenne = 3,33 Ecart-type = 0,76

Analyse en composantes principales

Variables :VITESSE, CONFORT, SECURITE.



La carte montre les positions des 3 critères et les coordonnées des 175 observations.

25 observations ne sont pas prises en compte (non-réponse à au moins un des critères).

86.2% de la variance est expliquée par les deux axes représentés.

Chaque observation est représentée par un point.

Analyse en composantes principales

Variables :VITESSE, CONFORT, SECURITE.

	VITESSE	CONFORT	SECURITE
VITESSE	1,00		
CONFORT	0,47	1,00	
SECURITE	0,58	0,42	1,00

La carte montre les positions des 3 critères et les coordonnées des 175 observations.

25 observations ne sont pas prises en compte (non-réponse à au moins un des critères).

86.2% de la variance est expliquée par les deux axes représentés.

Chaque observation est représentée par un point.

Analyse en composantes principales

Variables :VITESSE, CONFORT, SECURITE.

	Axe 1 (+66.0%)		Axe 2 (+20.2%)	
CONTRIBUTIONS POSITIVES	VITESSE	+36,0%	CONFORT	+69,0%
	SECURITE	+34,0%		
CONTRIBUTIONS NEGATIVES			SECURITE	-23,0%
			VITESSE	-7,0%

La carte montre les positions des 3 critères et les coordonnées des 175 observations.

25 observations ne sont pas prises en compte (non-réponse à au moins un des critères).

86.2% de la variance est expliquée par les deux axes représentés.

Chaque observation est représentée par un point.

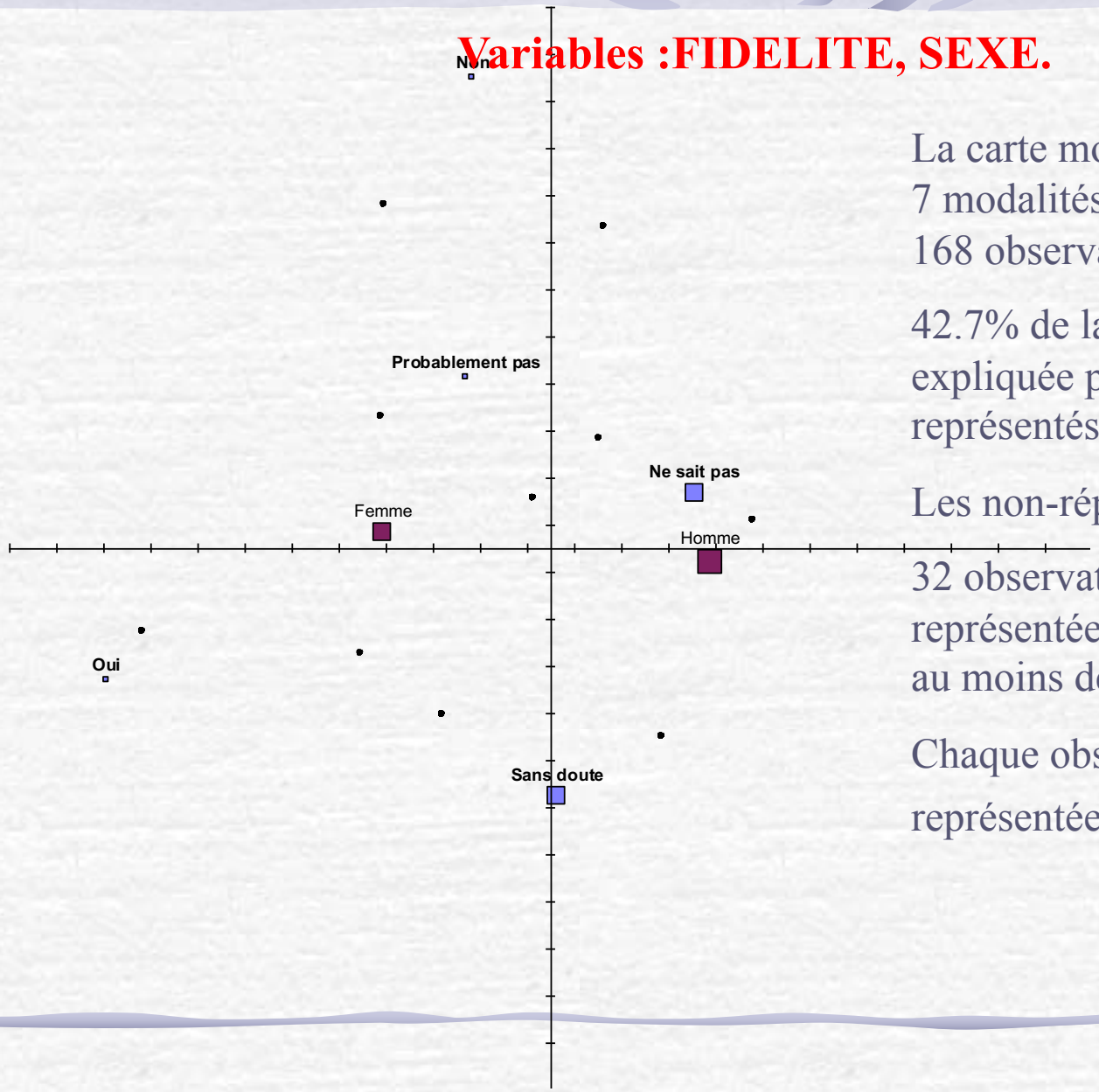
VII - L'ANALYSE FACTORIELLE DES CORRESPONDANCES (AFC)

Le résultat de cette analyse est un graphique en deux dimensions, due l'on appelle la carte factorielle et qui donne les projections des modalités suivant les axes les plus représentatifs. Le pourcentage de la variance expliquée par chacun des axes est mentionnée en regard de chacun des axes. La fonction permet de choisir les deux facteurs (composantes) qui seront utilisées sur l'axe horizontal et l'axe vertical de la carte factorielle.

Le pourcentage de variance expliquée qui caractérise chacun des axes factoriels est indiqué sous les axes.

Analyse des correspondances multiples

Variables :FIDELITE, SEXE.



La carte montre les positions des 7 modalités et les coordonnées des 168 observations.

42.7% de la variance est expliquée par les deux axes représentés.

Les non-réponses ont été ignorées.

32 observations ne sont pas représentées (non-réponse à l'une au moins des questions).

Chaque observation est représentée par un point.

Analyse des correspondances multiples

Variables :FIDELITE, SEXE.

	F1	F2	F3	F4	F5	S1	S2
F1	15	-	-	-	-	5	10
F2	-	52	-	-	-	27	25
F3	-	-	19	-	-	9	10
F4	-	-	-	17	-	8	9
F5	-	-	-	-	65	37	28
S1	5	27	9	8	37	86	-
S2	10	25	10	9	28	-	82

La carte montre les positions des 7 modalités et les coordonnées des 168 observations.

42.7% de la variance est expliquée par les deux axes représentés..

Le tableau est le tableau des effectifs (tableau de Burt) pour les 7 modalités.

Les non-réponses ont été ignorées.

F1 : Oui F2 : Sans doute F3 : Probablement pas F4 : Non F5 : Ne sait pas S1 : Homme S2 : Femme

Analyse des correspondances multiples

Variables :FIDELITE, SEXE.

	Axe 1 (+22.7%)		Axe 2 (+20.0%)	
CONTRIBUTIONS POSITIVES	Homme	+24.3%	Non	+46.7%
	Ne sait pas	+14.7%	Probablement pas	+6.9%
	Sans doute	+0.0%	Ne sait pas	+2.4%
CONTRIBUTIONS NEGATIVES	Oui	-32.6%	Sans doute	-40.1%
	Femme	-25.6%	Oui	-3.3%
	Probablement pas	-1.6%	Homme	-0.3%

Le tableau donne, pour les 2 premiers axes factoriels, les contributions relatives (positives et négatives) des modalités.

Les non-réponses ont été ignorées.