Statistiques et Informatique Appliquées à la Sociologie

UV SOC23A1

Présentation du cours 2001/2002

Organisation matérielle

Cours magistral : 7 ou 8 heures
mercredi 13h45-14h45 - A222
Travaux dirigés de statistiques : 1 h par quinzaine
- mardi 8h15 à 9h15 - sem A - Gr. 1 - A221
- mardi 8h15 à 9h15 - sem B - Gr. 2 - A221

Travaux dirigés d'informatique : 2 h par quinzaine. Effectif max. : 20
- mercredi 8h15-10h15 - sem A - A203
- mercredi 8h15-10h15 - sem B - A203
- mercredi 13h45-15h55 - sem A - A203
- mercredi 13h45-15h55 - sem B - A203
- vendredi 8h15-10h15 - sem A - A203

Contrôle des connaissances : (contrôle continu)
Examen écrit (1 heure) en janvier
Un dossier

Bibliographie

- Bloss et Grosseti, Introduction aux méthodes statistiques en sociologie, PUF, Coll. Le Sociologue

- G. Mialaret. Statistiques appliquées aux sciences humaines. PUF

- Colin, Lavoie, Delisle. Initiation aux méthodes quantitatives en Sc. Humaines CDR. No 310-LAV V2098/A

- P. Rateau, Méthode et statistique expérimentales en sciences humaines, Ellipses

Contenu Pour le premier semestre : CM essentiellement consacré aux statistiques.

Statistiques descriptives : résumer, donner une vue synthétique d'un ensemble de données : représentations graphiques, paramètres de position et de dispersion
Statistiques "mathématiques" : notion de distribution théorique.

Documents fournis :
Transparents du CM et polycopiés de TD
Egalement accessibles sur Internet (dans le courant du semestre) (au format .pdf lisible par Acrobat Reader) :

`http://geai.univ-brest.fr/~carpenti/`

Pourquoi faut-il étudier les statistiques ?

Les statistiques sont-elles utiles au sociologue ?

Les statistiques, il y a des calculatrices et des logiciels pour faire cela. Oui, mais ...

Introduction – Vocabulaire

Collecte des données

Sur qui?

Population

Individu statistique, unité statistique, sujet

A propos de quoi?

Attribut, caractère, variable statistique
Modalités d'une variable : exhaustives - exclusives
Champ ou domaine de variation

Nature d'une variable statistique

Variables nominales – échelle nominale

Variables ordinales – échelle ordinale

Variables numériques discrètes ou continues
Echelles d'intervalles, échelles de rapports

Recueil et présentation d'un ensemble de données

Individus : s_1, s_2, ..., s_N
Variables étudiées : X, Y, ...
Tableau protocole :

	Variable X	Variable Y
Individu 1	x_1	y_1
Individu 2	x_2	y_2
...	...	...
Individu N	x_N	y_N

Recensement ou tri à plat : **tableau d'effectifs**

Modalités	Effectifs	Fréquences
Modalité a_1	n_1	f_1
Modalité a_2	n_2	f_2
...	...	
Modalité a_k	n_k	f_k
	N	1 (ou 100%)

Fréquences :

$$f_i = \frac{\text{effectif de la modalité } a_i}{\text{effectif total}}$$

Variable regroupée en classes :

Classes	Effectifs	Fréquences
$[a_1, a_2[$	n_1	f_1
$[a_2, a_3[$	n_2	f_2
...	...	
$[a_k, a_{k+1}[$	n_k	f_k
	N	1 (ou 100%)

Tableau de données chronologiques

Exemple : étude d'une même variable à deux moments différents.

CSP	1962	1990	Ec. abs.	Ec. rel.	Coeff.
Agriculteurs	11	38	27	245 %	3.45
Sal. agricoles	1	5	4	400 %	5
Patrons	37	75	38	103 %	2.03
...	...	...	...	...	...
Total	197	894	697	354 %	4.54

Ecart absolu entre deux effectifs d'une même modalité.
Ecart relatif ou taux de variation :

$$\text{Ecart relatif} = \frac{\text{Effectif 1990} - \text{Effectif 1962}}{\text{Effectif 1962}}$$

Coefficient multiplicateur : 1 + Ecart relatif

Taux de variation annuel moyen entre 1962 et 1990 :

$$\text{Coeff. moyen} = 3.45^{\frac{1}{28}} = 1.045$$

$$\text{Taux moyen} = 4.5\%$$

Etude conjointe de deux variables ou tri croisé : **tableau de contingence**

$X \setminus Y$	b_1	b_2	$\ldots$	b_j	$\ldots$	b_l	
a_1	n_{11}	n_{12}		n_{1j}	$\ldots$	n_{1l}	
a_2	n_{21}	n_{22}			$\ldots$		
$\ldots$							
a_i	n_{i1}	n_{i2}		n_{ij}	$\ldots$		
$\ldots$							
a_k	n_{k1}	n_{k2}			$\ldots$	n_{kl}	
	$N_{\cdot 1}$						N

Exemple

	Hommes	Femmes	Total
Comédie	90	150	240
Drame	50	90	140
Variétés	160	160	320
	300	400	700

Deux (voire trois) manières de calculer des fréquences :

Par ligne

	Hommes	Femmes	Total
Comédie	37.5%	62.5%	100%
Drame	35.7%	64.3%	100%
Variétés	50%	50%	100%
	42.8%	57.2%	100%

Par colonne

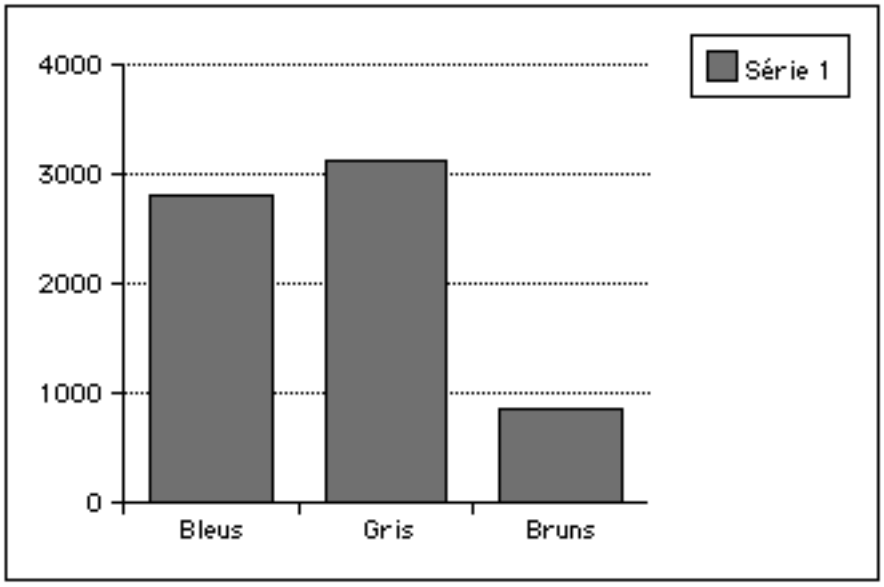
	Hommes	Femmes	Total
Comédie	30%	37.5%	34.3%
Drame	17%	22.5%	20%
Variétés	53%	40%	45.7%
	100%	100%	100%

Représentations graphiques

Variables nominales

Mod.	Eff.	Freq.	%
Bleus	2811	0,41	41%
Gris	3132	0,46	46%
Bruns	857	0,13	13%
Total	6800	1	100%

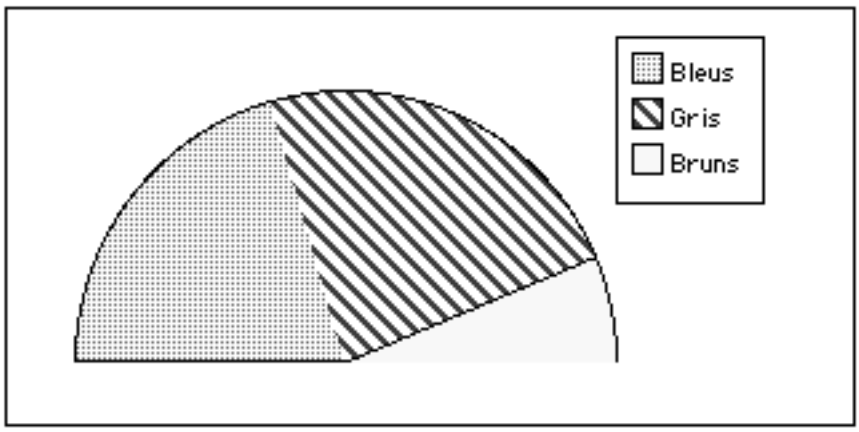
Diagrammes à bandes



Diagrammes circulaires ou semi-circulaires

Méthode de construction : construire un tableau de proportionnalité

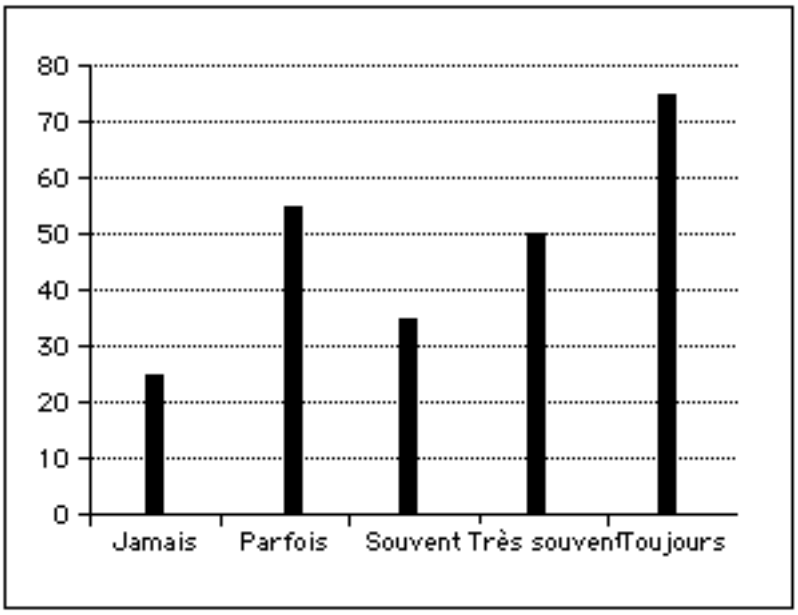
Modalités	Bleu	Gris	Bruns
Fréq.	0,41	0,46	0,13
Angle	74	83	23



Variable ordinale : diagramme en bâtons

On a demandé à 240 sujets s'ils fermaient à clef la porte de leur appartement.

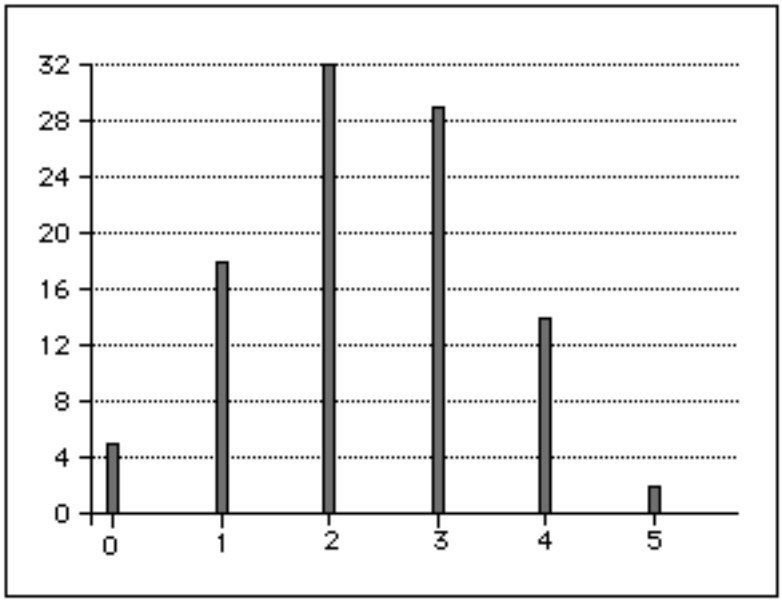
	Jam.	Parf.	Souv.	T. souv.	Tjrs
Eff.	25	55	35	50	75



Variable numérique discrète : même type de construction ; graduation régulière sur l'axe des abscisses.

Exemple : nombre de garçons dans des familles de 5 enfants

Mod.	0	1	2	3	4	5
Eff.	5	18	32	29	14	2



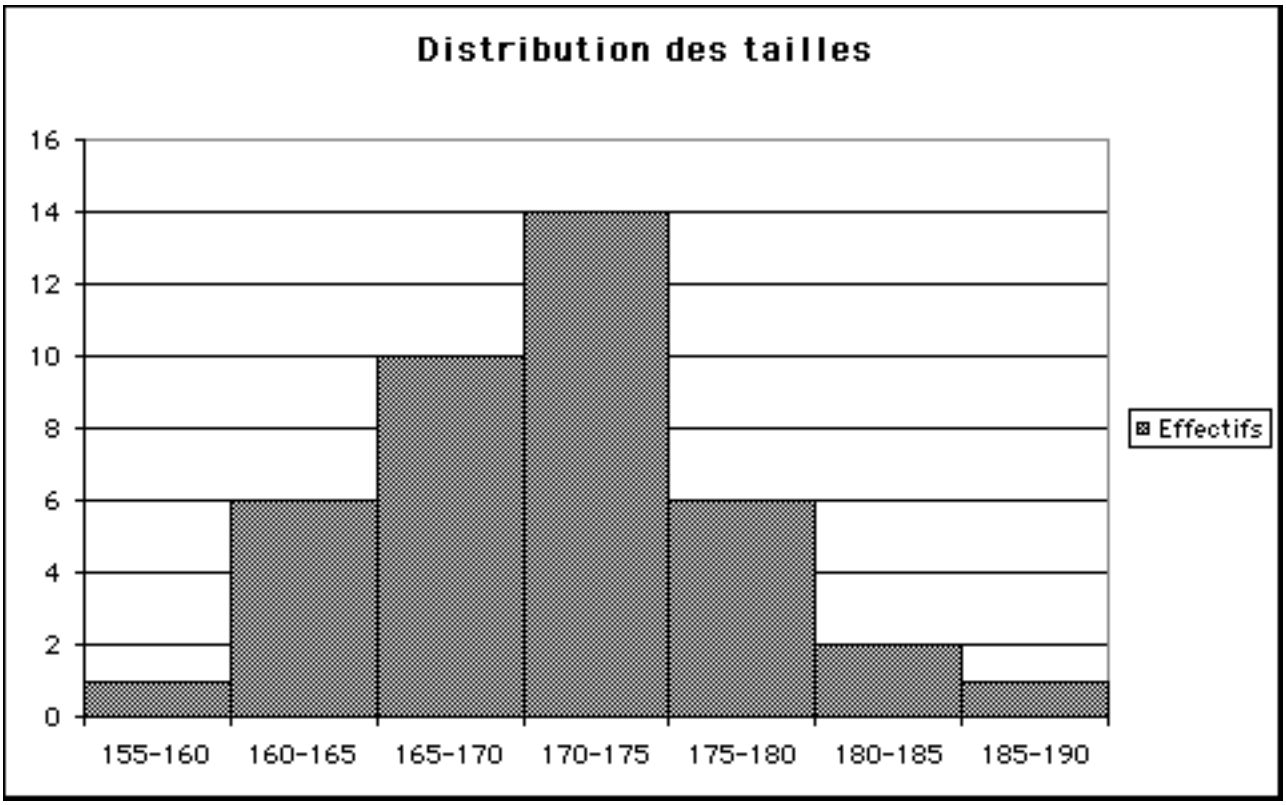
Variable numérique regroupée en classes : histogramme.

Un histogramme est formé de rectangles adjacents :
– dont la base est proportionnelle à l'amplitude de la classe
– dont l'aire est proportionnelle à l'effectif de la classe

Classes de même amplitude

Classe	Eff.
$[155,160[$	1
$[160,165[$	6
$[165,170[$	10
$[170,175[$	14
$[175,180[$	6
$[180,185[$	2
$[185,190[$	1

Hauteur des rectangles : effectifs des classes

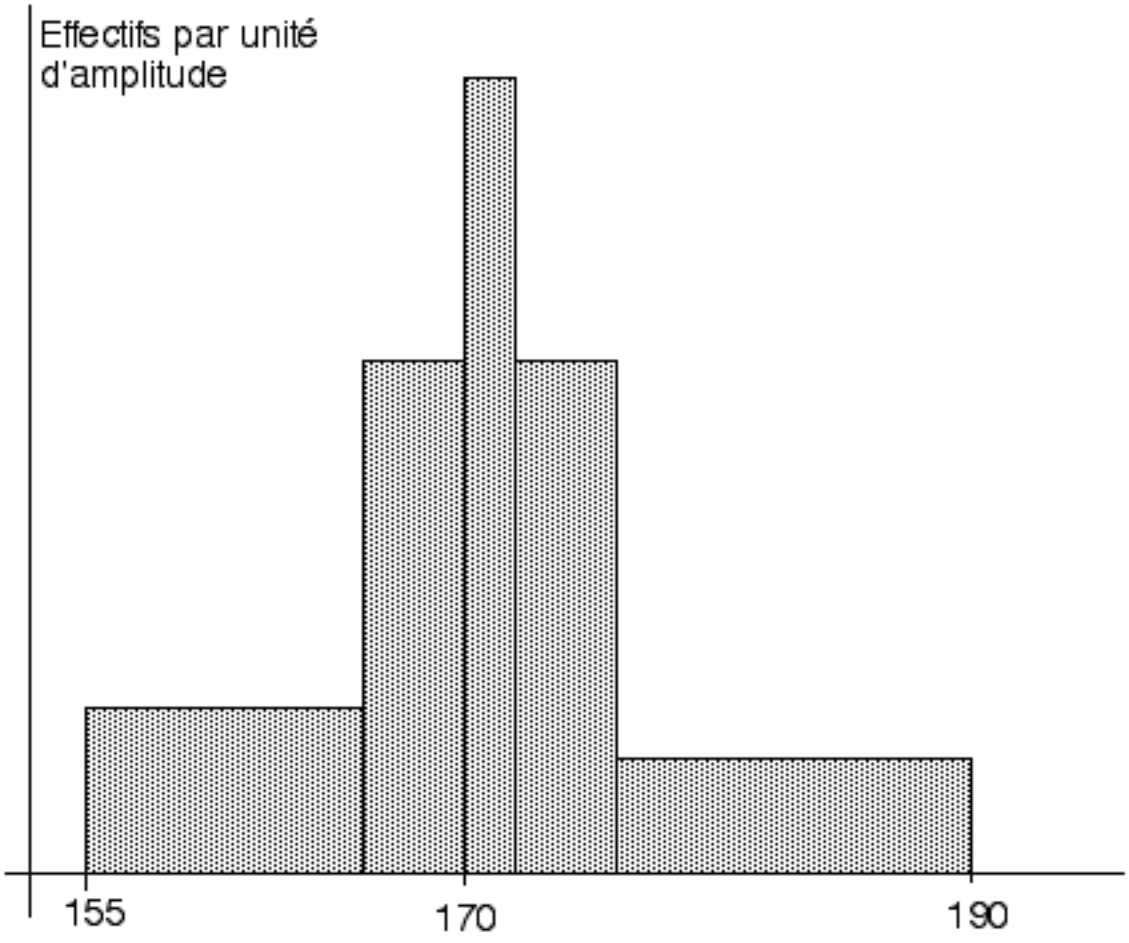


Classes d'amplitudes différentes

$$\text{densité} = \frac{\text{effectif}}{\text{amplitude}}$$

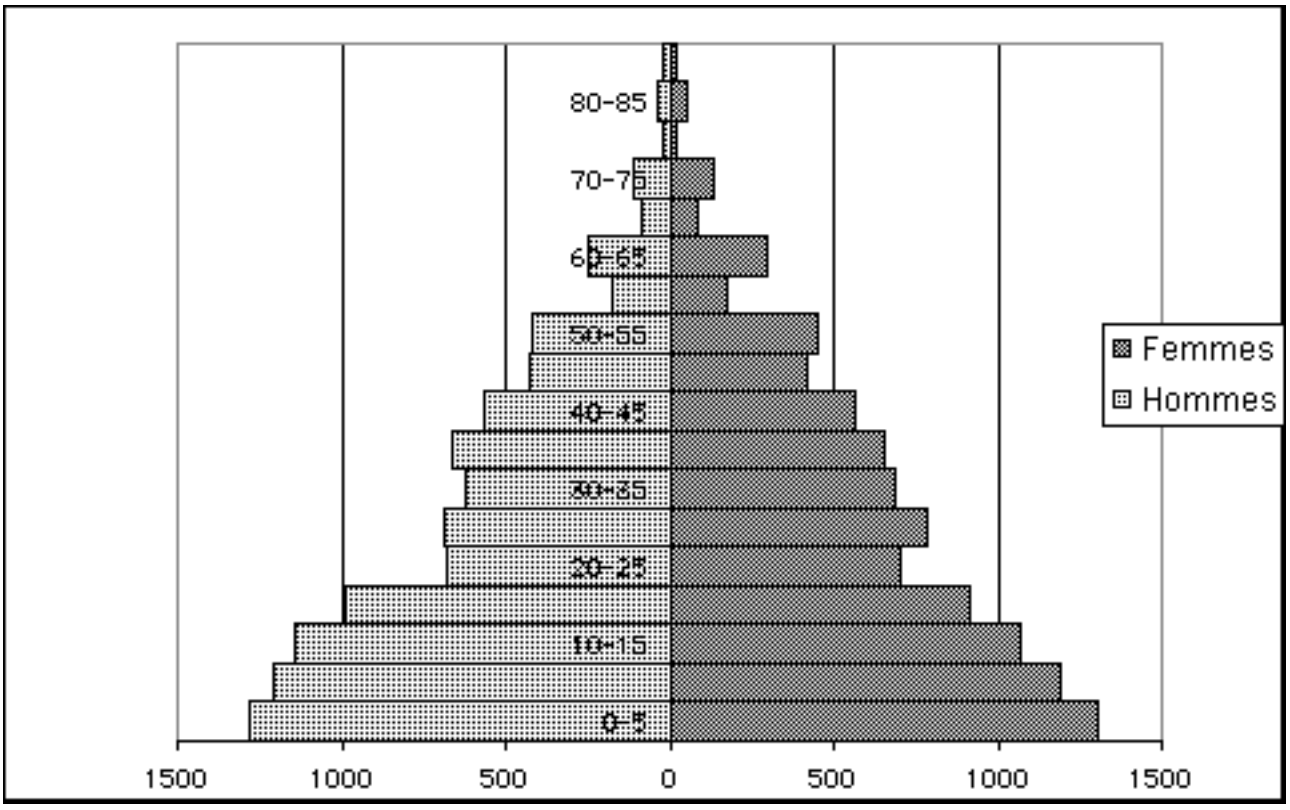
Hauteur des rectangles : densités des classes

Classe	Eff.	Amplitude	Densité
$[155, 166[$	8	11	0.73
$[166, 170[$	9	4	2.25
$[170, 172[$	7	2	3.5
$[172, 176[$	9	4	2.25
$[176, 190]$	7	14	0.5



Représentation de plusieurs variables sur un même graphique

Exemple : pyramide des âges



Caractéristiques de position

Mode, classe modale

Mode d'une série statistique (nominale, ordinale ou numérique) : modalité correspondant à l'effectif le plus élevé.
N.B. Une série statistique peut admettre plusieurs modes.

Classe modale d'une série statistique regroupée en classes : classe qui a la plus forte densité.
N.B. : c'est la classe correspondant au rectangle de hauteur maximale dans l'histogramme.

Médiane

Variable ordinale ou numérique.

Les individus sont *classés par valeurs croissantes de la variable*. La médiane est la valeur du caractère observée sur l'individu "médian", à savoir :

– Si N est impair, la médiane est la modalité observée sur l'individu de rang $\frac{N+1}{2}$

– Si N est pair et si le caractère est numérique, la médiane est la moyenne des modalités observées sur les individus de rangs $\frac{N}{2}$ et $\frac{N}{2}+1$.

Moyenne arithmétique

Caractère numérique.

– Calcul à partir d'un tableau protocole

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

– Calcul à partir d'un tableau d'effectifs

$$\mu = \frac{1}{N} \sum_{i=1}^{k} n_i a_i = \sum_{i=1}^{k} f_i a_i$$

Exemple :

Mod.	Effect.	$n_i a_i$
0	5	0
1	18	18
2	32	64
3	29	87
4	14	56
5	2	10
Total	100	235

$$\mu = 2,35$$

– Cas d'une variable répartie en classes

On considère que la masse de chaque classe est concentrée au centre $c_i = \frac{a_i + a_{i+1}}{2}$ de la classe.

Exemple :

Classes	Effect.	Centres	$n_i c_i$
$[155, 166[$	8	160,5	1284
$[166, 170[$	9	168	1512
$[170, 172[$	7	171	1197
$[172, 176[$	9	174	1566
$[176, 190[$	7	183	1281
	40		6840

$$\mu = \frac{6840}{40} = 171$$

Caractéristiques de dispersion

Etendue

$x_1, x_2, \ldots, x_n$: valeurs observées d'une variable statistique numérique.

$$x_{max} = Max(x_1, x_2, \ldots, x_n)$$

$$x_{min} = Min(x_1, x_2, \ldots, x_n)$$

L'étendue de la variable est :

$$e = x_{max} - x_{min}$$

Quartiles

Soit une série statistique numérique de médiane M.

Premier quartile Q_1 : médiane de la série des observations strictement inférieures à M.

Deuxième quartile Q_2 : médiane M de la série complète.

Troisième quartile Q_3 : médiane de la série des observations strictement supérieures à M.

L'écart interquartile est défini par :

$$Iq = Q_3 - Q_1$$

Représentation graphique permettant de visualiser l'étendue et les quartiles : *boîte à moustaches.*

Généralisation : déciles, centiles...

Variance et écart type

Définition : La variance est la moyenne des carrés des écarts à la moyenne.

– A partir d'un tableau protocole

$$V = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

– A partir d'un tableau d'effectifs

$$V = \frac{1}{N}\sum_{i=1}^{k} n_i(a_i - \mu)^2 = \sum_{i=1}^{k} f_i(a_i - \mu)^2$$

L'écart type est donné par : $\sigma = \sqrt{V}$.

Calcul pratique

"Moyenne des carrés moins carré de la moyenne"

$$V = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2$$

$$V = \left(\frac{1}{N} \sum_{i=1}^{k} n_i a_i^2 \right) - \mu^2$$

Remarques
Cas d'une variable répartie en classes : utiliser les centres de classes.
Unités, effet d'un changement d'origine ou d'unités.

Organisation des calculs

– Cas d'une variable discrète

Mod.	Effect.	$n_i a_i$	$n_i a_i^2$
0	5	0	0
1	18	18	18
2	32	64	128
3	29	87	261
4	14	56	224
5	2	10	50
Total	100	235	681

$$\mu = 2.35 \ ; \ V = 6.81 - 2.35^2 = 1.29 \ ; \ \sigma = 1.13$$

– Cas d'une variable répartie en classes

Classes	Effect.	Centres	$n_i c_i$	$n_i c_i^2$
$[155, 166[$	8	160.5	1 284	206 082
$[166, 170[$	9	168	1 512	254 016
$[170, 172[$	7	171	1 197	204 687
$[172, 176[$	9	174	1 566	272 484
$[176, 190]$	7	183	1 281	234 423
	40		6 840	1 171 692

$$\mu = 171$$

$$V = \frac{1171692}{40} - 171^2 = 29292.3 - 29241 = 51.3$$

$$\sigma = \sqrt{51.3} = 7,16 \text{ cm}$$

Analyses bivariées - Introduction

Jusqu'à présent : études portant sur une seule variable.

Etude simultanée de deux variables nominales :

Analyse croisée de deux variables (par ex. questionnaire d'enquête)

– Loisir préféré et sexe

– Opinion sur l'immigration et sensibilité politique

Question posée : les deux variables sont-elles *indépendantes* ou *dépendantes* ?

Outil : analyse d'un tableau de contingence à l'aide d'un test du χ^2.

Etude de la liaison entre deux variables numériques

– Population d'étudiants. Variables : note de février et note de juin. Lien éventuel ?

– Population de sujets : lien entre taille et poids

Question posée : Y a-t-il un lien, une *corrélation* entre les deux variables ?

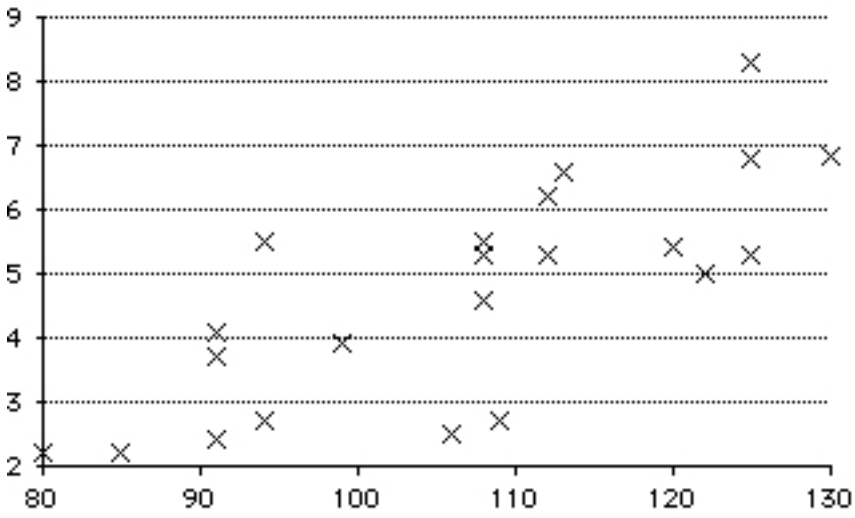
Outil : étude de la corrélation linéaire entre les deux variables

Corrélation linéaire

Données : deux variables numériques définies sur la même population

	X	Y
s_1	x_1	y_1
s_2	x_2	y_2
...	...	...

Nuage de points : points (x_i, y_i)



Covariance des variables X et Y

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

ou

$$Cov(X,Y) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\ \overline{y}$$

Coefficient de corrélation de Bravais Pearson

$$r = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

Remarques

– Formules analogues avec corrélation, écarts types, . . .corrigés

– Il existe des relations non linéaires

– Corrélation n'est pas causalité

Régression linéaire

Rôle "explicatif" de l'une des variables par rapport à l'autre. Les variations de Y peuvent-elles (au moins en partie) être expliquées par celles de X ? Peuvent-elles être prédites par celles de X ?

Modèle permettant d'estimer Y connaissant X

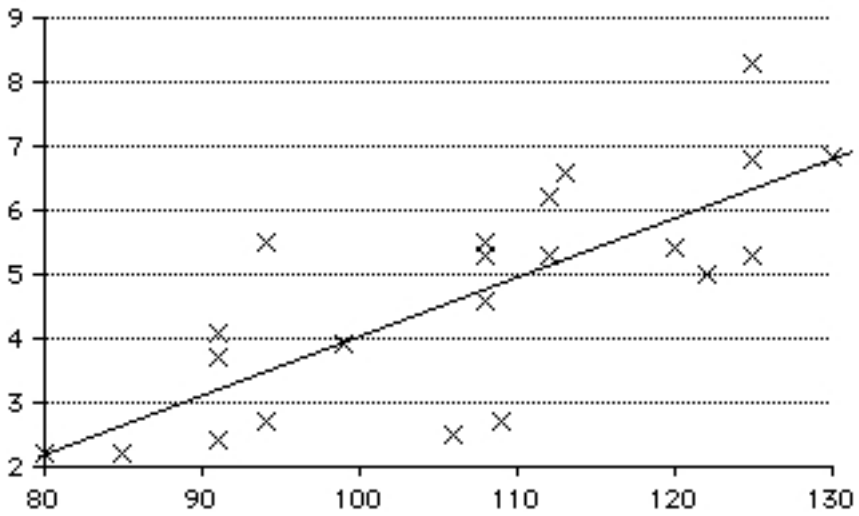
Droite de régression de Y par rapport à X :

La droite de régression de Y par rapport à X a pour équation :

$$y = ax + b$$

avec :

$$a = \frac{Cov(X,Y)}{\sigma^2(X)} \quad ; \quad b = \overline{y} - a\overline{x}$$



Comparaison des valeurs observées et des valeurs estimées

Valeurs estimées : $\widehat{y}_i = ax_i + b$
Erreur (ou résidu) : $e_i = y_i - \widehat{y}_i$

On montre que :

$$\sigma^2(Y) = \sigma^2(\widehat{Y}) + \sigma^2(E)$$

avec :

$$\frac{\sigma^2(E)}{\sigma^2(Y)} = 1 - r^2 \quad ; \quad \frac{\sigma^2(\widehat{Y})}{\sigma^2(Y)} = r^2$$

$\sigma^2(\widehat{Y})$: variance *expliquée* (par la variation de X, par le modèle)

$\sigma^2(E)$: variance *perdue* ou *résiduelle*

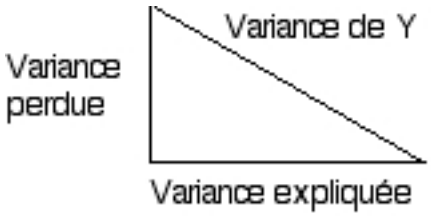
r^2 : part de la variance de Y qui est expliquée par la variance de X. *Coefficient de détermination*

Exemple : $r = 0.86$

$r^2 = 0.75$; $1 - r^2 = 0.25$; $\sqrt{1 - r^2} = 0.5$.

• La part de la variance de Y expliquée par la variation de X est de 75%.

• L'écart type des résidus est la moitié de l'écart type de Y.



Etude de la liaison entre deux caractères qualitatifs

Exemple : préférences des publics masculin et féminin.
Effectifs observés

	H	F	Total
Comédie	90	75	165
Drame	50	45	95
Variétés	160	80	240
Total	300	200	500

Goûts dépendants du sexe ?

Exemple : $\frac{240}{500} = 48\%$. 48% des personnes interrogées préfèrent les variétés. Si les deux variables étaient parfaitement indépendantes, on devrait retrouver :
– 48% des hommes, c'est-à-dire 144 hommes préférant les variétés
– 48% des femmes, c'est-à-dire 96 femmes préférant les variétés.

Effectifs attendus (ou théoriques) si indépendance :

$$\text{Dans chaque case : effectif} = \frac{\text{total ligne} \times \text{total colonne}}{\text{total général}}$$

	H	F
Comédie	99	66
Drame	57	38
Variétés	144	96

Problème : évaluer la distance entre les deux tableaux ?

Calcul de la "distance" du χ^2 définie par :

$$\chi^2_{obs} = \sum_{i,j} \frac{(n_{ij} - t_{ij})^2}{t_{ij}}$$

Calcul pratique :

Mod.	n_{ij}	t_{ij}	$\frac{(n_{ij} - t_{ij})^2}{t_{ij}}$
H.C.	90	99	0.82
H.D	...		0.86
H.V.			1.78
F.C.			1.23
F.D.			1.29
F.V.			2.67
			8.64

Quelle interpétation peut-on donner de χ^2_{obs} ?

– Coefficient de contingence de Pearson :

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

– Coefficient Phi de Cramér :

$$\Phi_c = \sqrt{\frac{\chi^2}{N(k-1)}}$$

où N est l'effectif total et k le minimum des deux valeurs nombre de lignes, nombre de colonnes.

– En fait, la véritable interprétation de χ^2 est faite en statistiques inférentielles, à l'aide d'un test du χ^2.

Test du χ^2

– 500 personnes : échantillon

– 2 sources de variation : effet du sexe, hasard

– *Si seul le hasard est en cause*, la distance suit une *loi du χ^2 à 2 ddl.*

– On se fixe un seuil de 5% (par exemple)

– *Si seul le hasard est en cause*, on a seulement 5% de chances d'observer un χ^2 supérieur à la *valeur critique* $\chi^2_c = 5.991$.

– Or on a observé : $\chi^2_{obs} = 8.64$.

– Conclusion : différence de goûts selon le sexe.

Résumé

Tableau de contingence : effectifs observés n_{ij}
Totaux par ligne : $n_{i.}$ par colonne : $n_{.j}$
Total général : N ou $n_{..}$
l lignes et c colonnes

Effectifs théoriques : tableau (t_{ij}) avec :

$$t_{ij} = \frac{n_{i.}\, n_{.j}}{n_{..}} = \frac{total\ \ ligne \times total\ \ colonne}{total\ \ général}$$

Distance du χ^2 :

$$\chi^2_{obs} = \sum_{i,j} \frac{(n_{ij} - t_{ij})^2}{t_{ij}}$$

Test proprement dit :

– Hypothèses :
H_0 : Les variables sont indépendantes.
H_1 : Les variables sont dépendantes.

– On fixe un seuil α (5%, 1%, . . .)

– Lecture de la table : valeur critique χ^2_{crit} pour le seuil α et $(l-1)(c-1)$ ddl

– Intervalles d'acceptation et de rejet

– Comparaison de χ^2_{obs} et de χ^2_{crit}

– Conclusion :
  – Si $\chi^2_{obs} < \chi^2_{crit}$, indépendance acceptée
  – Si $\chi^2_{obs} > \chi^2_{crit}$, indépendance rejetée ;
  les variables dépendent l'une de l'autre.

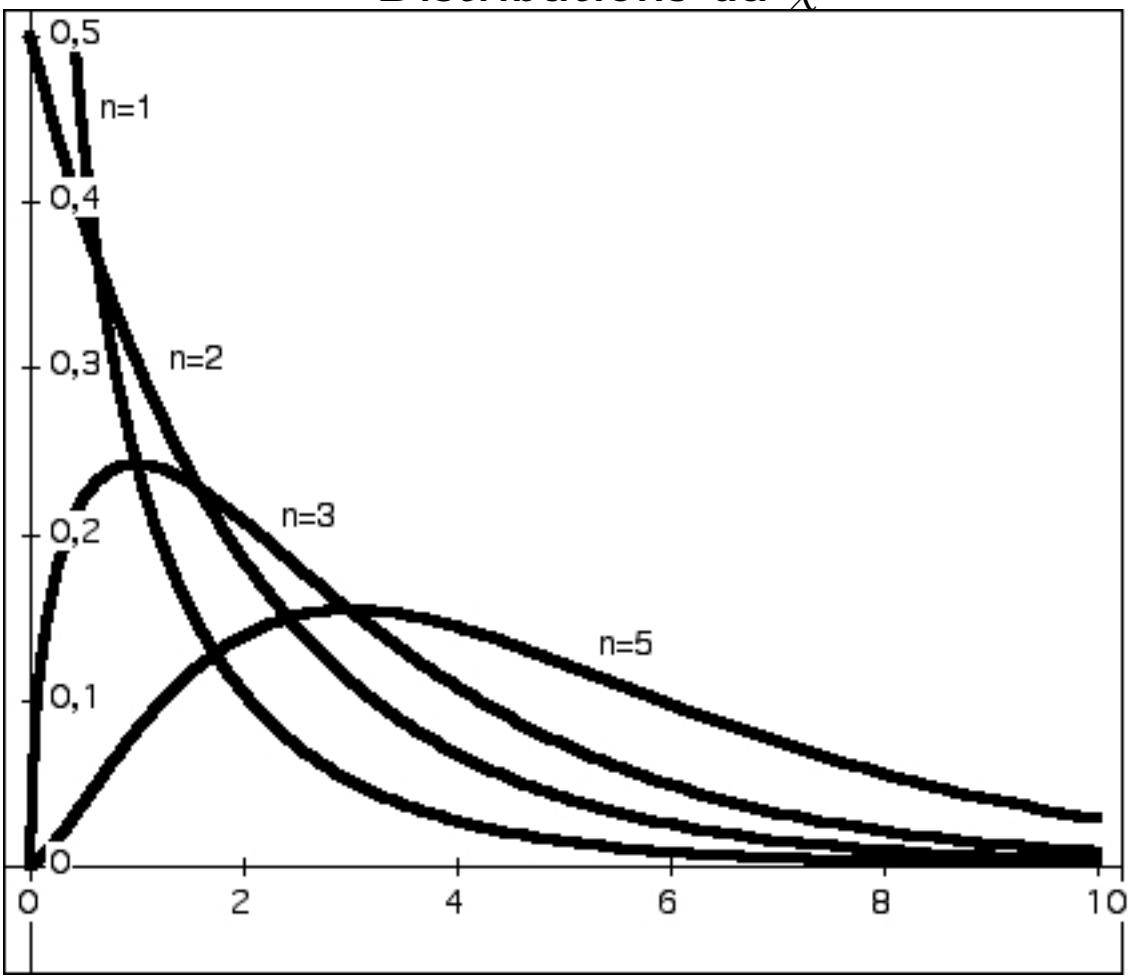
Remarques

– Condition sur les effectifs théoriques minimaux ; regroupement éventuel des modalités

– Correction de Yates (tableau 2×2)

$$\chi^2_{corr} = \sum_{i,j} \frac{(|n_{ij} - t_{ij}| - 0.5)^2}{t_{ij}}$$

Distributions du χ^2



Une autre utilisation du χ^2 : adéquation à une loi théorique

Confronter un tableau d'effectifs "théoriques" à un tableau d'effectifs observés.

Exemple : Etude de Durkheim sur le suicide selon la saison (entre 1835 et 1843).

Pour un échantillon de 1000 suicides :

Print.	Eté	Autom.	Hiver	Total
283	306	210	201	1000

Fréquences et effectifs théoriques :

Print.	Eté	Autom.	Hiver	Total
25%	25%	25 %	25 %	100%
250	250	250	250	1000

Contributions au χ^2 :

	Print.	Eté	Autom.	Hiver	χ^2
Obs.	283	306	210	201	
Théo.	250	250	250	250	
Ctri	4.356	12.544	6.40	9.604	32.904

Test proprement dit :

H_0 : Le tableau de fréquences théoriques correspond aux fréquences dans la population
H_1 : Les fréquences dans la population sont différentes de celles du tableau de fréquences théoriques

Statistique de test : χ^2 à 3 ddl.
Pourquoi 3 ? : 4 observations. Total théorique (1000) calculé à partir des observations. $4 - 1 = 3$.

Seuil : 1%. Valeur critique : $\chi^2_{crit} = 11.35$

Règle de décision :
– Si $\chi^2_{obs} \leq 11.345$, on accepte H_0.
– Si $\chi^2_{obs} > 11.345$, on refuse H_0 et on accepte H_1.

Conclusion :
Ici, $\chi^2_{obs} = 32.9$. On accepte donc H_1. Autrement dit, le taux de suicides varie avec la saison.