N° 28

COURS DE STATISTIQUES

. D. DACUNHA CASTELLE

Réédition de 1973 - 1974

Fascicule 1

CHAPITRE 0

STATISTIQUE DESCRIPTIVE.

A - INTRODUCTION.

La statistique descriptive est une méthode de description des données recueillies à propos de l'étude de certains phénomènes d'ordre économique, sociologique ou expérimental. La description se fait essentiellement dans deux directions.

a) Direction géométrique. Ceci consiste d'abord à classer les données et à les disposer de manière la plus parlante possible (tableaux, images, graphiques etc...). Dans un deuxième temps (cf. fascicule 3), on utilise des techniques plus élaborées, qui consistent par exemple à projeter les données supposées ou rendues numériques (à valeurs dans $\mathbb{R}^k$) sur des plans convenablement choisis, et à interpréter les figures ainsi obtenues. Cette analyse des données moderne (car nécessitant de puissants moyens de calculs) est donc assez voisine de la géométrie descriptive ou de la géométrie côtée (dessin de machines). La différence essentielle vient de ce que les règles de lecture de figures obtenues sont complexes, et amènent à des déductions de caractère probabiliste.

b) Direction numérique. Elle consiste à résumer l'ensemble des données par la valeur de certaines grandeurs numériques typiques.

Les deux directions ne sont pas évidemment exclusives, et sont toujours utilisées simultanément.

Population. Les données recueillies portent (terme démographique) sur une population, c'est-à-dire sur un certain ensemble d'unités statistiques ou d'individus, éléments de cette population qui doit donc être définie avec précision.

Caractères.

Chacun des individus de la population peut être décrit par un ou plusieurs caractères.

Par exemple pour le personnel d'une entreprise : sexe, âge, qualification, nationalité, nombre de personnes à charge, appartenance à tel syndicat...

Les caractères étudiés peuvent être qualitatifs ou quantitatifs.

Qualitatif si ses diverses modalités (modalité : différentes situations possibles du caractère) ne sont pas mesurables

ex : sexe, couleur de cheveux, qualification ...

Les modalités d'un caractère qualitatif sont les différentes rubriques d'une nomenclature établie de telle façon que l'individu figure dans une et une seule rubrique (rubriques exhaustives et incompatibles). Chaque nomenclature peut être plus ou moins détaillée.

Quantitatif si les diverses modalités sont mesurables ou repérables, c.à.d. si à chacune des modalités est attachée un nombre. Ce nombre est la variable statistique.

Variables statistiques.

Discrètes, si ses valeurs possibles sont isolées (le nombre de valeurs possibles pouvant être fini ou infini).

Continues, si ses valeurs possibles sont en nombre infini et en général à valeur dans un intervalle de $\mathbb{R}$.

Souvent, la distinction est difficile à faire entre variable continue et variable discrète.

Par exemple, toute mesure est discrète du fait d'une précision toujours limitée, alors que la nature intrinsèque de la variable (par exemple le diamètre d'une pièce) en fait une variable continue. Réciproquement, on considèrera qu'une variable qui peut prendre un très grand nombre de valeurs possibles, sera une variable continue. Par exemple :

"le salaire d'un ouvrier"

"le bénéfice annuel d'une entreprise"

"le diamètre mesuré au $^1/_{100}$ de mm d'une pièce".

Dans le cas du salaire, la plus faible unité monétaire est le centima. Ainsi, assimiler le salaire à une variable continue, c'est assimiler la plus faible unité monétaire à la précision des mesures.

Pour étudier une variable statistique continue, on définira des classes de valeurs possibles, pouvant avoir une amplitude constante ou variable.

Le nombre de classes à adopter dépend de la précision des mesures et des effectifs de la population étudiée

- trop de classes $\Rightarrow$ irrégularités accidentelles provenant du faible nombre d'individus par classe.
- pas assez de classes $\Rightarrow$ perte d'information.

On justifiera pourquoi, plus tard, il est intéressant d'obtenir dans les différentes classes des effectifs comparables.

Concluons cette introduction par une remarque importante. La statistique descriptive porte sur une population donnée. Le problème de l'échantillonnage et du sondage (cf. chapitre 1) ne se pose pas. Il n'y a pas de modèle statistique en statistique descriptive, uniquement un ensemble de données, sans structuration a priori.

NOMENCLATURE INSEE DES CATÉGORIES SOCIO-PROFESSIONNELLES

0. AGRICULTEURS EXPLOITANTS.
   0. 0. Agriculteurs exploitants

1. SALARIÉS AGRICOLES.
   1. 0. Salariés agricoles

2. PATRONS DE L'INDUSTRIE ET DU COMMERCE.
   2. 1. Industriels
   2. 2. Artisans
   2. 3. Patrons pêcheurs
   2. 6. Gros commerçants
   2. 7. Petits commerçants

3. PROFESSIONS LIBÉRALES ET CADRES SUPÉRIEURS.
   3. 0. Professions libérales
   3. 2. Professeurs; professions littéraires et scientifiques
   3. 3. Ingénieurs
   3. 4. Cadres administratifs supérieurs

4. CADRES MOYENS.
   4. 1. Instituteurs; professions intellectuelles diverses
   4. 2. Services médicaux et sociaux
   4. 3. Techniciens
   4. 4. Cadres administratifs moyens

5. EMPLOYÉS.
   5. 1. Employés de bureau
   5. 3. Employés de commerce

6. OUVRIERS.
   6. 0. Contremaîtres
   6. 1. Ouvriers qualifiés
   6. 3. Ouvriers spécialisés
   6. 5. Mineurs
   6. 6. Marins et pêcheurs
   6. 7. Apprentis ouvriers
   6. 8. Manœuvres

7. PERSONNELS DE SERVICE.
   7. 0. Gens de maison
   7. 1. Femmes de ménage
   7. 2. Autres personnels de service

8. AUTRES CATÉGORIES.
   8. 0. Artistes
   8. 1. Clergé
   8. 2. Armée et Police

9. PERSONNES NON ACTIVES.
   9. 1. Étudiants et élèves
   9. 2. Militaires du contingent
   9. 3. Anciens agriculteurs (exploitants et salariés)
   9. 4. Retirés des affaires
   9. 5. Retraités du secteur public
   9. 6. Anciens salariés du secteur privé
   9. 9. Autres personnes non actives.

B - DISTRIBUTIONS STATISTIQUES A UN CARACTERE

1 - TABLEAUX STATISTIQUES.

Soit une population de n individus, décrite par le caractère C prenant k modalités $C_1, C_2, \ldots, C_k$.

n_i est le nombre d'individus présentant la modalité C_i ou effectif de la modalité C_i.

$f_i = \frac{n_i}{n}$ est la fréquence de la modalité C_i.

Le tableau suivant décrit la population étudiée,

Modalité de C	Effectif
C_1	n_1
C_2	n_2
C_i	n_i
C_k	n_k
Total	n

ou fréquence (Total connu)

Caractère qualitatif

Même forme que celle-ci-dessus indiquée. ⟹
Exemple : Répartition des Etrangers qui vivent en France (Recensement général de 1962)

(Modalités) Nationalité	Effectif
Italiens	645 000
Belges	78 000
Allemands	46 000
Hollandais, Luxembourgeois	17 000
Espagnols	431 000
Portugais	50 000
Polonais	177 000
Autres Européens	124 000
Algériens	335 000
Marocains, Tunisiens	54 000
Autres Etrangers	194 000
Total	2 151 000

(Modalités) Nombre de pièces à rebuter par lot α	Nombre correspondant de lots ou effectif
1	2
2	9
3	14
4	20
5	18
6	15
7	9
8	6
9	4
10	2
11	1
Total	100

⟸

Caractère quantitatif discret :

Sur 100 lots, chacun de 100 pièces, on a observé le nombre de pièces défectueuses.
La variable statistique est le nombre de lots ayant donné α pièces défectueuses.

Cas d'une statistique continue.

Dans ce cas, les modalités du caractère sont définies à partir des classes de valeurs possibles (la division en classe est arbitraire).

Exemple : Distribution des ouvriers ayant travaillé toute l'année chez le même employeur selon le revenu annuel (année : 1952)

Tranche de salaire annuel net en milliers de francs	Nombre d'ouvriers
Moins de 100	1 721
de 100 à moins de 125	2 413
— 125 — — 150	4 342
— 150 — — 175	8 234
— 175 — — 200	13 300
— 200 — — 225	16 053
— 225 — — 250	16 774
— 250 — — 300	33 251
— 300 — — 350	29 211
— 350 — — 400	22 453
— 400 — — 500	24 605
— 500 — — 600	9 477
— 600 — — 800	4 093
— 800 — — 1 000	443
— 1 000 — — 1 500	125
— 1 500 — — 2 000	12
— 2 000 — — 5 000	14
Total	185 951

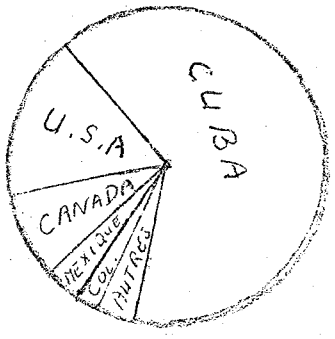
Le problème concret du choix des classes est souvent très important pour les applications numériques. Pour calculer une moyenne il n'est pas toujours souhaitable d'assimiler une classe à son "milieu".

2 - REPRESENTATION GRAPHIQUE

1. Caractères qualitatifs.

Par secteur angulaire : Exemple :
Médailles obtenues aux jeux Pan américains de Cali (1971) par pays et par tête d'habitant.

Il y a proportionalité des aires aux effectifs.

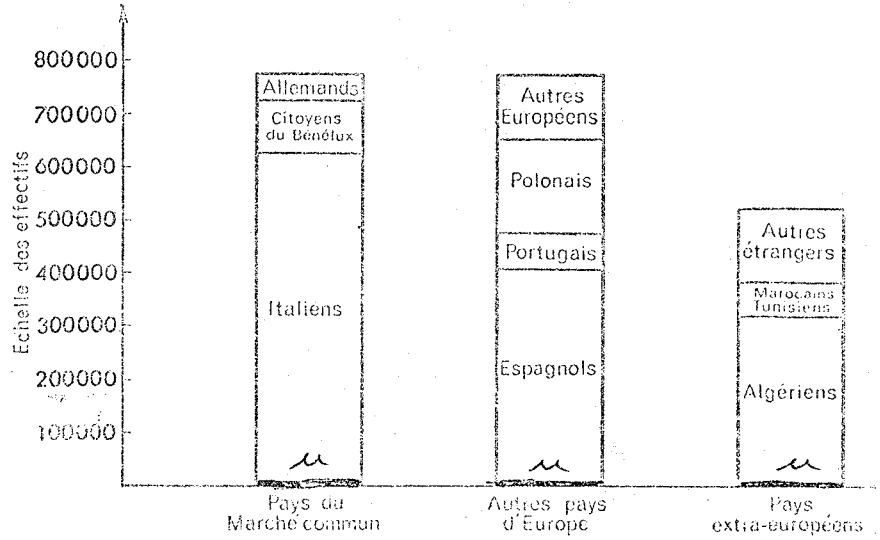


Par tuyaux d'orgue.

Exemple : répartition des étrangers en France (1962)

⇒

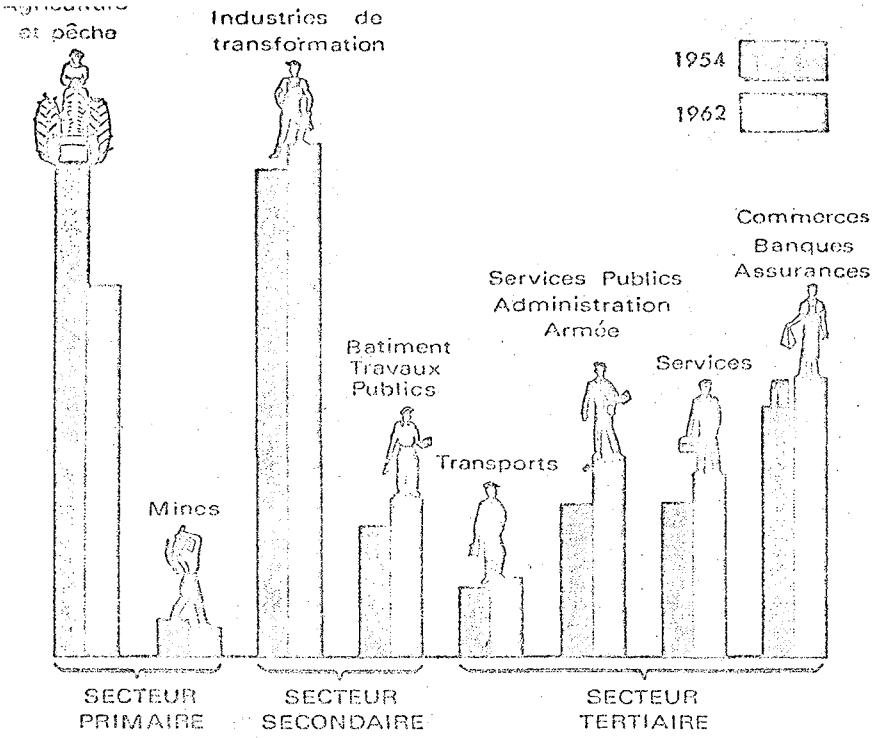
Les tuyaux d'orgue ont une base constante et une hauteur proportionnelle à l'effectif correspondant. Leur surface est donc proportionnelle à l'effectif.



La représentation par tuyaux d'orgue est particulièrement agréable quand on veut figurer une variation entre deux variables statistiques prenant les mêmes modalités

⇒

(Extrait de : Documents pour la classe I P N 1964)



Évolution de la population active française suivant la branche d'activité de 1954 à 1962.

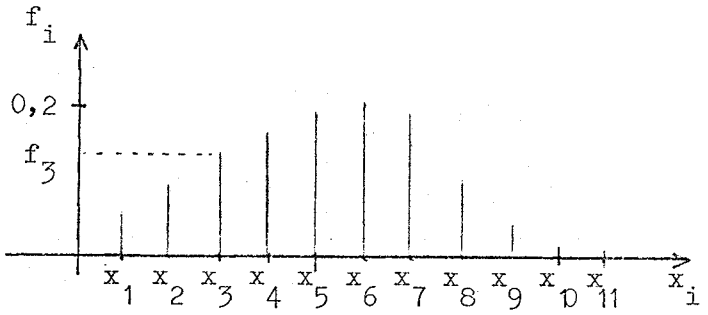
2. Caractères quantitatifs.

→ Variable discrète ou discretisée

<u>Diagramme en bâtons</u>

on figure la fréquence f_i pour la modalité x_i

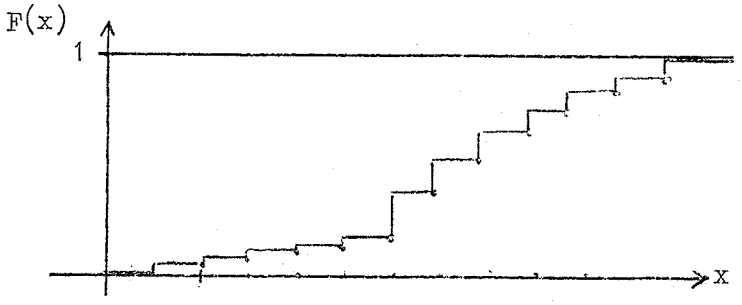
$(\Sigma f_i = 1)$



<u>Courbe cumulative ou fonction de répartition</u>

$$F(x) = \sum_{j=1}^{i} f_j$$

pour $x_i < x \leqslant x_{i+1}$

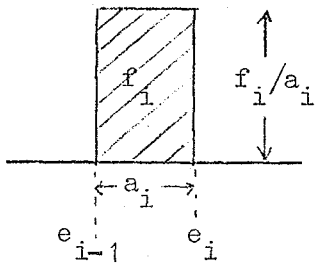


(Discontinuité aux points x_i , continuité à droite en x_i).

Dans le cas continu étudié (premier tableau, page 2), les amplitudes des différentes classes sont des multiples de 25.000 frs.

Les effectifs sont directement comparables si ils correspondent à la même amplitude.

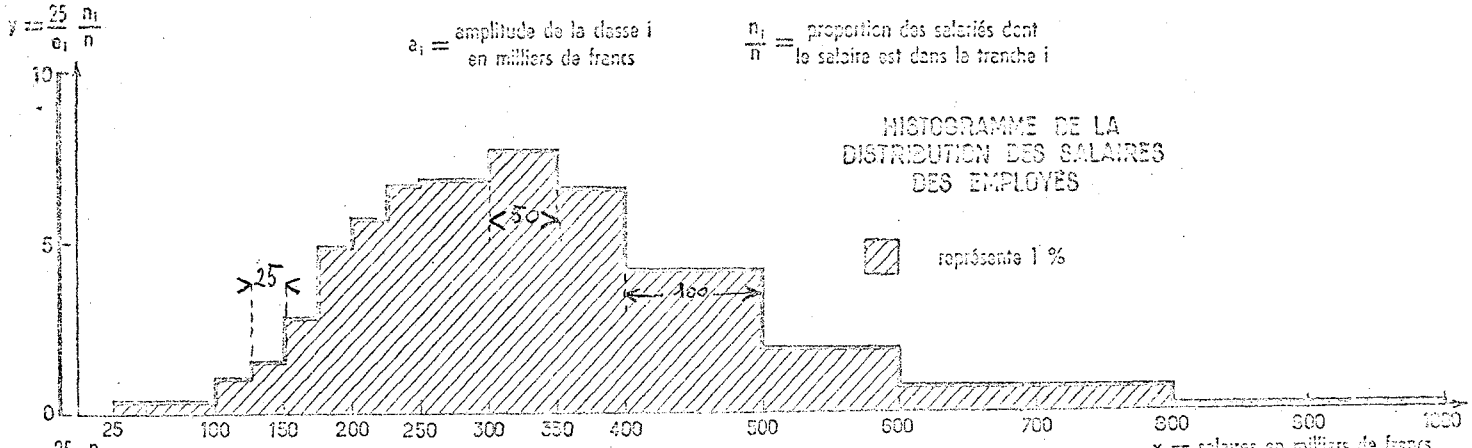
Si les amplitudes sont différentes, les surfaces figurées sur une classe devront représenter la fréquence f_i

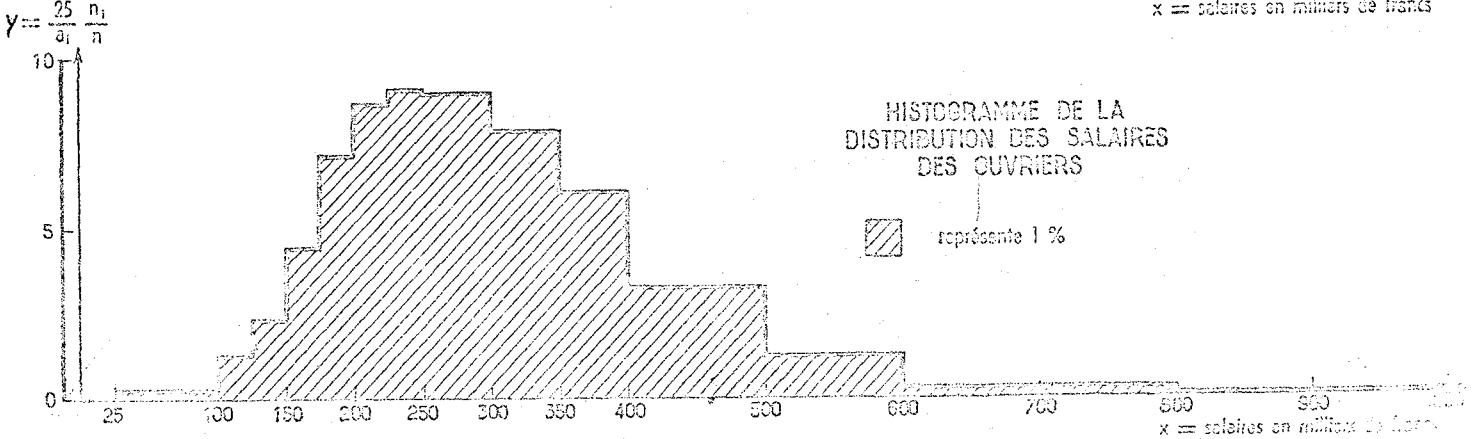


Si les amplitudes de classes diminuent, certaines irrégularités, dues à la faiblesse des effectifs, apparaissent. Cependant, si le nombre d'observations augmente, alors en faisant diminuer l'amplitude des classes, on se "rapprochera" de la courbe de densité théorique.

Si l'amplitude des classes augmente, on fait disparaître les irrégularités, en même temps, on approche de la courbe de densité théorique avec moins de précision.

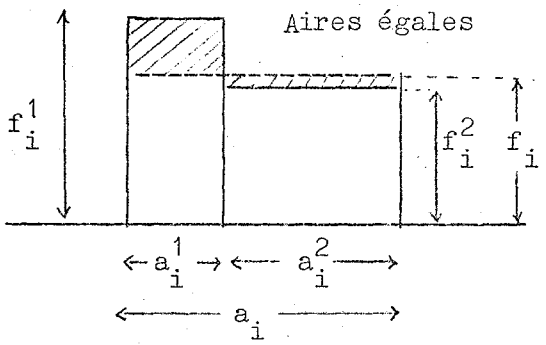
HISTOGRAMMES COMPARÉS DES DISTRIBUTIONS DES SALAIRES DES EMPLOYÉS ET OUVRIERS





Remarque : Introduction d'une classe supplémentaire ou réunion de deux classes adjacentes.

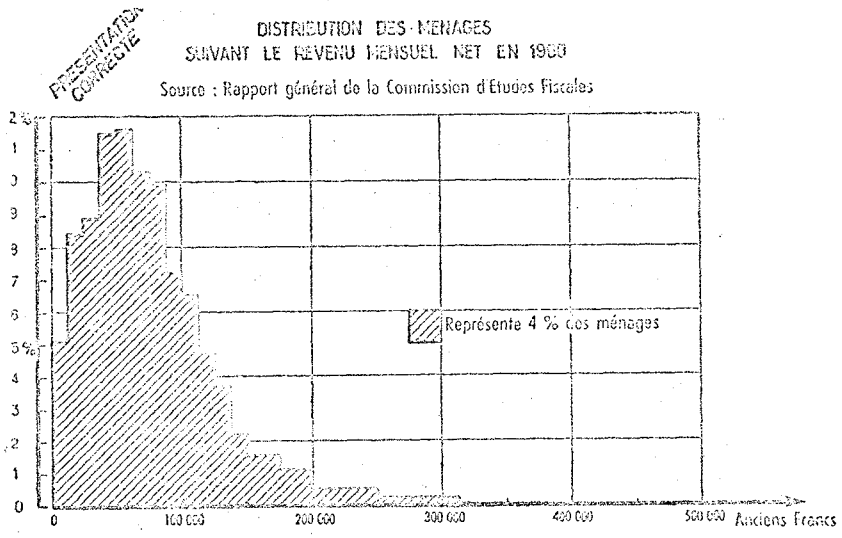
Cette opération est schématisée par le dessin suivant :

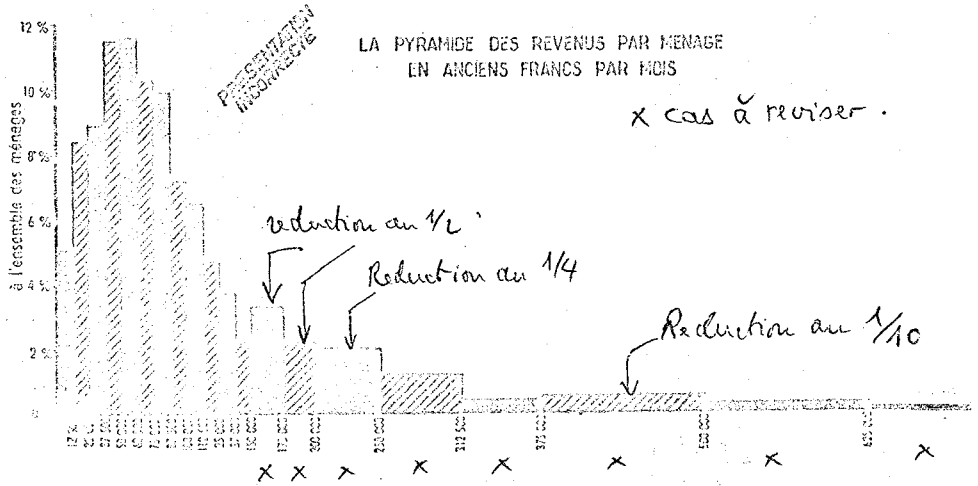


On doit avoir : $a_i^1 f_i^1 + a_i^2 f_i^2 = (a_i^1 + a_i^2) f_i \Longrightarrow f_i$.

Exemple de présentation incorrecte :

Par exemple, pour la classe [175 000, 200 000], la représentation incorrecte figure en ordonnée f_i ; les surfaces hachurées représentent donc $f_i \times a_i$.





3 - CARACTERISTIQUES DE VALEURS CENTRALES.

Elles doivent :

- être définies de façon aussi objectives que possible,
- dépendre de toutes les observations, mais pas de leur ordre (sauf dans certains problèmes spécifiques faisant intervenir le temps, que nous n'étudierons pas)
- être peu sensibles aux fluctuations de l'échantillonnage.

La Médiane :

C'est une valeur de la variable statistique qui partage en deux effectifs égaux les individus de la population.

Si F est la fonction de répartition de l'échantillon

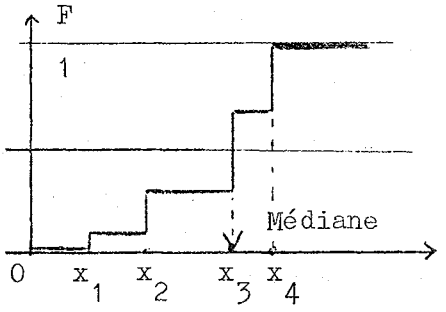
$$M \text{ médiane} \iff F(M) \geqslant \frac{1}{2} \quad 1-F(M) \leqslant \frac{1}{2} \qquad (1)$$

Si F est strictement croissante, M est unique comme solution de (1)

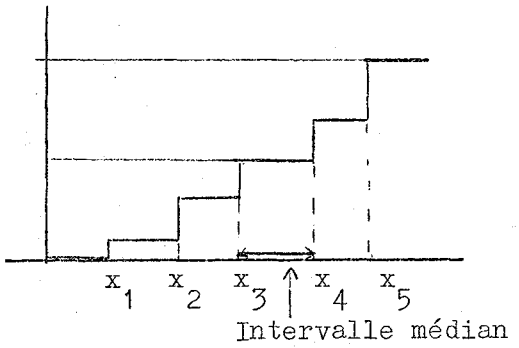
Si F est discrète, deux cas

. Aucune valeur x_i ne donne $F(x_i) = \frac{1}{2}$. On retient le x_i tel que :

$$F(x_i) < \frac{1}{2} < F(x_{i+1}) \Rightarrow$$



. Si $\exists x_i$ tel que $F(x_i) = \frac{1}{2}$ alors tout point de $[x_i, x_{i+1}[$ satisfait (1) $\Rightarrow$ On a un intervalle médian.



Règle d'obtention : Classement dans l'ordre croissant des observations x_i .
Calcul des effectifs cumulés $nF(x_i)$.

Si $nF(x_i) < 50 < nF(x_{i+1}) \rightarrow x_i$ valeur médiane
$nF(x_i) = 50 \qquad [x_i\ x_{i+1}[$ intervalle médian

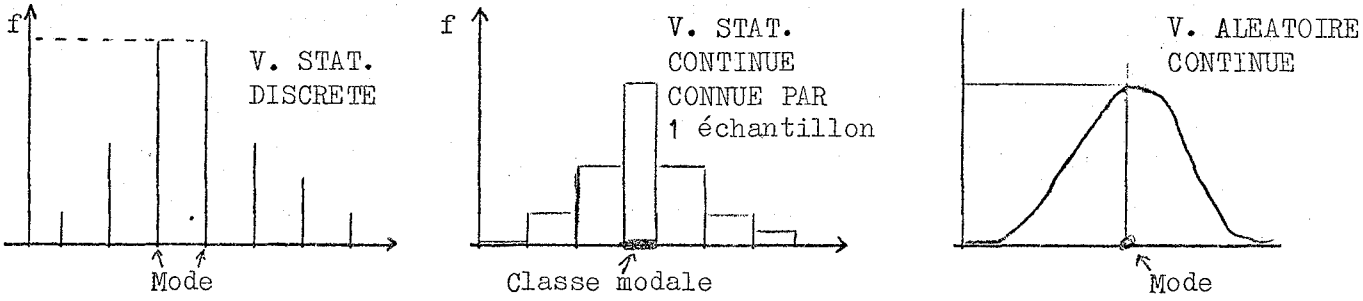
Classement des obs.	x_0	x_1	...	x_6	x_7	.....	x_{25}
Effectifs cumulés				45	63		

$\frac{n}{2} = 50$

Le Mode :

. C'est la valeur la plus "fréquente" de la variable statistique, on dit aussi valeur dominante. C'est pratiquement une valeur "centrale", quand on considère des distributions "régulières".

. Le mode d'une distribution continue à densité correspond au point où la densité prend la plus grande valeur



Moyenne : c'est la caractéristique la plus importante. Elle n'est définie que pour des variables quantitatives à valeurs dans un espace vectoriel.

. C'est : $\bar{x} = \sum_{i=1}^{k} f_i x_i$ si $\{x_1, \ldots, x_k\}$ est l'ensemble des valeurs prises par l'échantillon.

4 - CARACTERISTIQUES DE DISPERSION.

Ecart quadratique moyen ou écart-type.

. C'est $\sigma = \sqrt{\sum_i f_i |x_i - \bar{x}|^2}$

σ^2 est la variance.

5 - CARACTERISTIQUES DE FORME. (A titre d'exemple, elles sont très arbitraires).

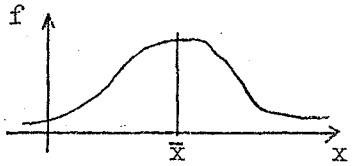
Coefficient d'assymétrie.

Si une distribution est symétrique, ces différents moments centrés d'ordre impair sont nuls.

En effet :

$$\int_R (x-\bar{x})^r f(x)\,dx = \int_{-\infty}^{\bar{x}} + \int_{\bar{x}}^{\infty}$$

et $$\int_{-\infty}^{\bar{x}} (x-\bar{x})^2 f(x)\,dx = -\int_{\bar{x}}^{\infty} (x-\bar{x})^2 f(x)\,dx$$



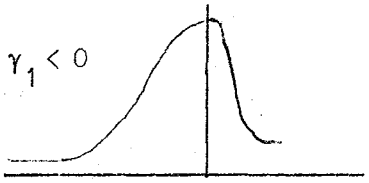
(Symétrie de f par rapport à $x = \bar{x}$)

. Fischer a proposé le coefficient :

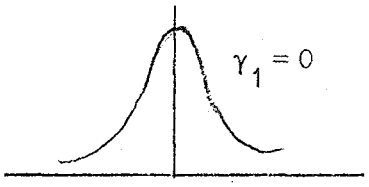
$$\boxed{\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3}} \quad \text{où} \quad \begin{cases} \mu_3 = \int (x-\bar{x})^3 f(x)\,dx \\ \mu_2 = \int (x-\bar{x})^2 f(x)\,dx \end{cases}$$

→ sans dimension (c'est pour cela que l'on divise par $\mu_2^{3/2}$)

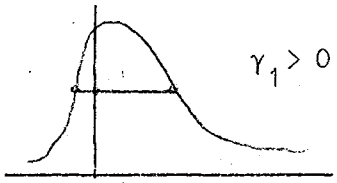
→ nul si la distribution a un axe de symétrie



Distribution dissymétrique étalée à gauche.



symétrique



Distribution dissymétrique étalée à droite.

$d = \frac{Q_1 + Q_3 - 2M}{M}$ est également un coefficient d'assymétrie, sans dimension. · ($d = 0$ si la distribution est unimodale et symétrique).

6 - CARACTERISTIQUES DE CONCENTRATION.

Considérons le tableau des salaires (page 2, B).

A chaque case (classe) associons la masse des salaires ; par exemple :

Classe $(125 < \text{salaire} < 150) \Longrightarrow 4342 \times \frac{125 \times 150}{2} = 601$ Millions

Classe $e_i \Longrightarrow S_i$ = masse des salaires de la classe e_i

Classe $e_i \Longrightarrow q_i = \frac{\sum_{j=1}^{i} S_j}{\sum_{j=1}^{k} S_j}$

Ainsi les ouvriers dont le salaire est inférieur à 200 représentent 16,5 % de l'effectif des ouvriers et se partagent 8,29 % de la masse totale des salaires.

Notons : p_i : proportion des ouvriers dont le salaire est $\leqslant e_i$

q_i : proportion de la masse salariale gagnée par les ouvriers dont le salaire est $\leqslant e_i$

Limites de classe e_i (en milliers de francs)	Effectifs n_i	Masses des salaires par classe : S_i (en millions de francs)	$p_i = F(e_i)$ en %	$q_i = \frac{\sum_{j=1}^{i} S_j}{\sum_{j=1}^{k} S_j}$ (en %)
	1 721	114		
100	2 413	273	0,93	0,20
125	4 342	601	2,22	0,66
150	8 264	1 349	4,56	1,70
175	13 300	2 494	9,00	4,01
200	16 053	3 404	16,15	8,29
225	16 774	3 972	24,79	14,14
250	33 251	9 120	33,81	20,96
300	29 211	9 461	51,69	36,62
350	22 453	8 391	67,40	52,86
400	24 005	10 623	79,47	67,27
500	9 477	5 133	92,38	85,51
600	4 093	2 717	97,48	94,33
800	443	386	99,68	98,99
1 000	125	144	99,92	99,65
1 500	12	20,3	99,99	99,90
2 000	14	37,7	99,99	99,93
5 000			100,00	100,00
Total	185 951	58 240	—	—

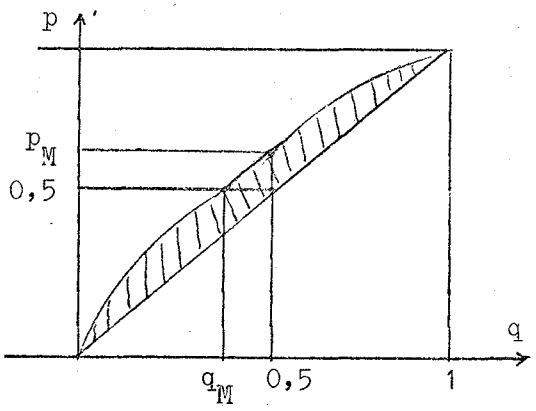
Courbe de concentration.

C'est la courbe

$$p_i = f(q_i)$$

. Comme $p \leqslant q$, elle se situe dans la moitié supérieure du carré 1×1 construit sur les axes.

. Elle est croissante



C - DISTRIBUTIONS STATISTIQUES A DEUX CARACTERES

Considérons une population de n individus décrits simultanément par deux caractères A et B, A ayant k modalités $\{A_1, A_2, \ldots, A_k\}$, B ayant p modalités $\{B_1, B_2, \ldots, B_p\}$.

On cherchera à décrire la distribution statistique associée

soit : - par des tableaux et une représentation graphique

soit : - par une description numérique.

Les caractères peuvent être qualitatifs ou quantitatifs.

I - TABLEAUX STATISTIQUES.

Notons n_{ij} = nombre d'individus présentant à la fois le caractère A_i et B_j.

Du fait que les modalités de A

de B sont incompatibles et exhaustives.

On a : $\sum_{i,j} n_{ij} = n$.

Modalités de B / Modalités de A	B_1	B_2	...	B_j	...	B_p	Totaux (Marges) horizontaux
A_1	n_{11}	n_{12}	...	n_{1j}		n_{1p}	$n_{1.}$
A_2	n_{21}	n_{22}	...	nn_{2j}		n_{2p}	$n_{2.}$
⋮							
A_i	n_{i1}	n_{i2}	...	n_{ij}		n_{ip}	$n_{i.}$
⋮							
A_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kp}	$n_{k.}$
Tableaux verticaux	$n_{.1}$	$n_{.2}$	...	$n_{.k.}$		$n_{.p}$	n

La fréquence de (A_i, B_j) est $\frac{n_{ij}}{n} = f_{ij}$.

Distributions marginales.

C'est la distribution des deux marges

distribution marginale associée à A : $n_{1.}, n_{2.}, \ldots, n_{k.}$

" " " à B : $n_{.1}, n_{.2}, \ldots\ldots n_{.p}$

C'est la distribution associée à l'un des caractères, une fois l'autre oublié.

Distributions conditionnelles.

Considérons la sous-population réalisant le caractère B_j. Sur cette sous-population, la distribution du caractère A est appelée la distribution conditionnelle de A_i, à condition que B_j soit réalisé.

On a : $f_i^j(A_i/B_j) = \frac{n_{ij}}{n_{.j}}$.

Il y a p distributions conditionnelles suivant le caractère A

k " " " " " B

Si on note $(A_i \ B_j)$ l'évènement : A_i et B_j réalisé on a :

$$P(A_i \cap B_j) = P(A_i)\ P(B_j/A_i) = P(B_j)\ P(A_i/B_j).$$

Indépendance.

Les caractères A et B sont indépendants si réciproque

$$P(A/B) = P(A) \quad (\text{et alors } P(B/A) = P(B)).$$

Alors, cela est équivalent au fait que les p distributions conditionnelles de A sont les mêmes, en particulier identiques à la distribution marginale de A. En effet : f_i^j ind de j

$$f_i^j = \frac{n_{i1}}{n_{.1}} = \ldots = \frac{n_{ij}}{n_{.j}} = \ldots = \frac{n_{ip}}{n_{.p}} = \frac{n_{i.}}{n_{..}} = f_{i.}.$$

Liaison fonctionnelle.

Le caractère A est lié fonctionnellement au caractère B si à chaque modalité B_j de B correspond une telle modalité du caractère A (on noterait : $A = f(B)$).

Cette notion n'est pas réciproque comme celle d'indépendance. Elle est particulièrement intéressante dans le cas de caractères quantitatifs.

Si f est bijective, la notion est réciproque.

A \ B	B_1	B_2	B_3	B_4	B_5
A_1	4	0	2	0	0
A_2	0	1	0	1	0
A_3	0	0	0	0	3

. Distributions conditionnelles de A dégénérées.

Il y a globalement 6 types de distributions à 2 caractères

Situation plus favorable à une description graphique :
- Les deux qualitatifs
- 1 qualitatif, l'autre quantitatif discret
- " " " continu

Situation plus favorable à la description numérique :
- 2 quantitatifs discrets
- 1 quantitatif discret, 1 continu
- 2 quantitatifs continus

II - REPRESENTATIONS GRAPHIQUES.

Une représentation graphique d'une distribution statistique à 2 caractères doit s'attacher à figurer, d'une part la distribution globale (représenter toute l'information), et si possible les distributions conditionnelles ainsi que marginales, les distributions conditionnelles étant définies par $f_{i/j} = n_{ij}/n_{ij}$.

1er cas : A et B qualitatifs.

On peut représenter la distribution globale et une famille de distributions conditionnelles (A/B_j) (ou l'inverse).

n_{ij} est "représenté" par un rectangle de base $n_{.j}$ et dont la hauteur est proportionnelle à la fréquence conditionnelle f_i^j

$$n_{ij} = n_{.j} f_i^j .$$

On a donc : → effectifs marginaux caractère B $n_{.j}$ (base rectangles)

→ effectifs du tableau (aire des rectangles)

→ fréquences conditionnelles (A/B_j) soit f_i^j.

Voir tableaux et graphiques pages 4 et 5.

2ème cas : 1 caractère qualitatif, 1 caractère quantitatif.

A désigne le caractère qualitatif

B " " " quantitatif.

On peut déjà utiliser un mode de représentation comme si les deux caractères étaient qualitatifs.

On peut également représenter les distributions conditionnelles à partir de diagrammes en bâtons (autant de diagrammes que de modalités du caractère qualitatif). Voir page 6.

Distributions conditionnelles en fréquences cumulées (Catégorie/Epoque construction) ⇒

Logements (résidences principales) suivant l'époque de construction (1) et la catégorie socio-professionnelle du chef de ménage (*Recensement général de la Population ; mars 1962. Sondage au 1/20*).

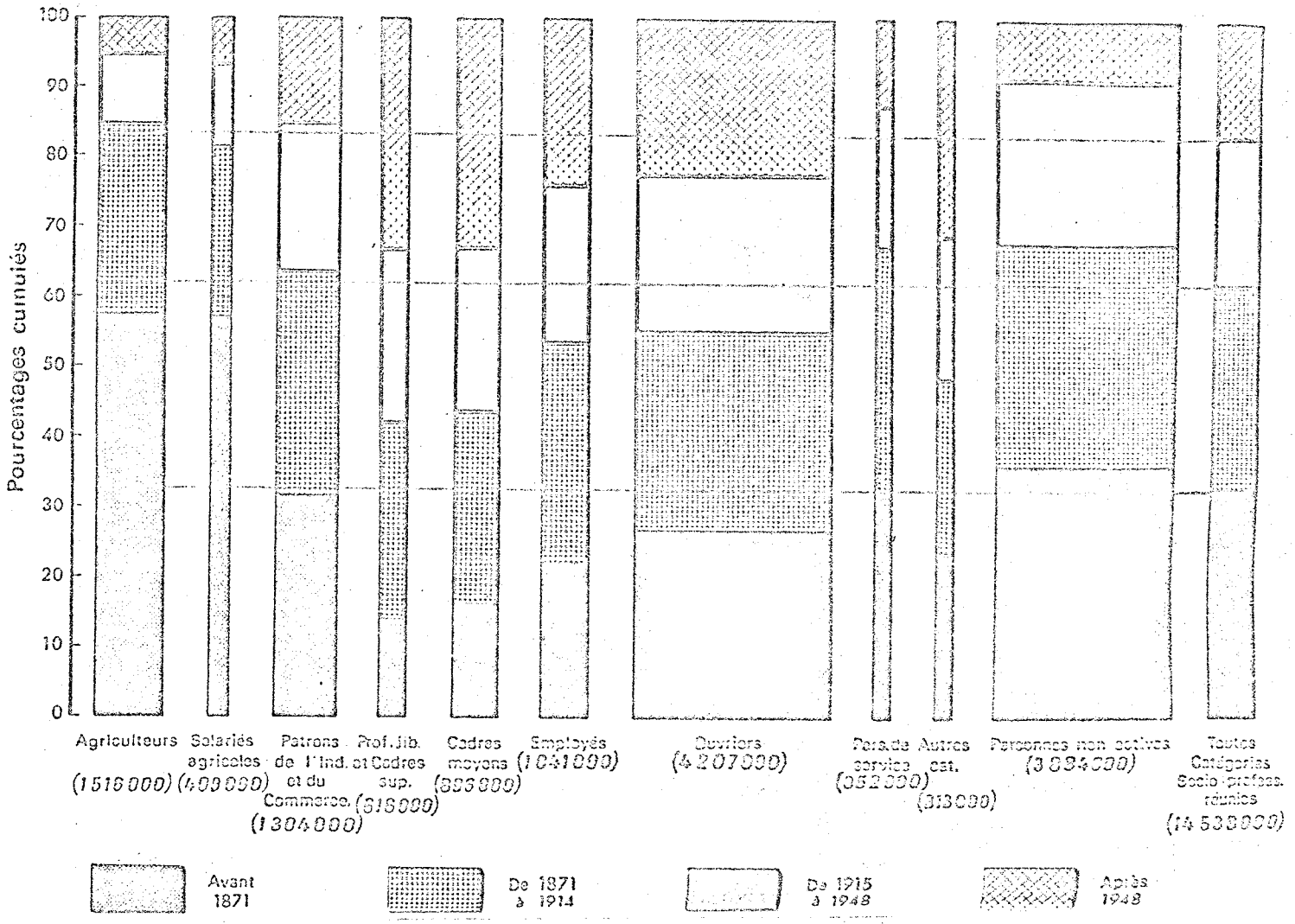
Catégorie socio-professionnelle \ Epoque de construction	Avant 1871	De 1871 à 1914	De 1915 à 1948	Après 1948	Total
Agriculteurs	873 340	410 040	158 380	74 620	1 516 380
Salariés agricoles	233 060	100 160	48 600	27 280	409 100
Patrons de l'industrie et du commerce	415 380	413 000	280 520	195 380	1 304 280
Professions libérales et Cadres supérieurs	87 120	175 660	148 760	204 440	615 980
Cadres moyens	144 560	247 860	210 640	293 180	896 240
Employés	231 760	322 700	237 800	249 180	1 041 440
Ouvriers	1 118 440	1 177 820	954 500	956 040	4 206 800
Personnel de service	112 560	123 260	72 400	43 720	351 940
Autres catégories	73 240	77 960	65 360	95 960	312 520
Personnes non actives	1 398 840	1 232 120	932 340	322 220	3 883 520
Total	4 686 300	4 280 580	3 109 300	2 462 020	14 538 200

(1) En toute rigueur, l'année de construction est un caractère quantitatif continu. Du fait de la nature des fichiers des Cartes retenues, on assimile ici chaque classe définissant une époque de construction à la modalité d'un caractère qualitatif.

Distributions conditionnelles (Epoque/Catégorie socio.) ⇒

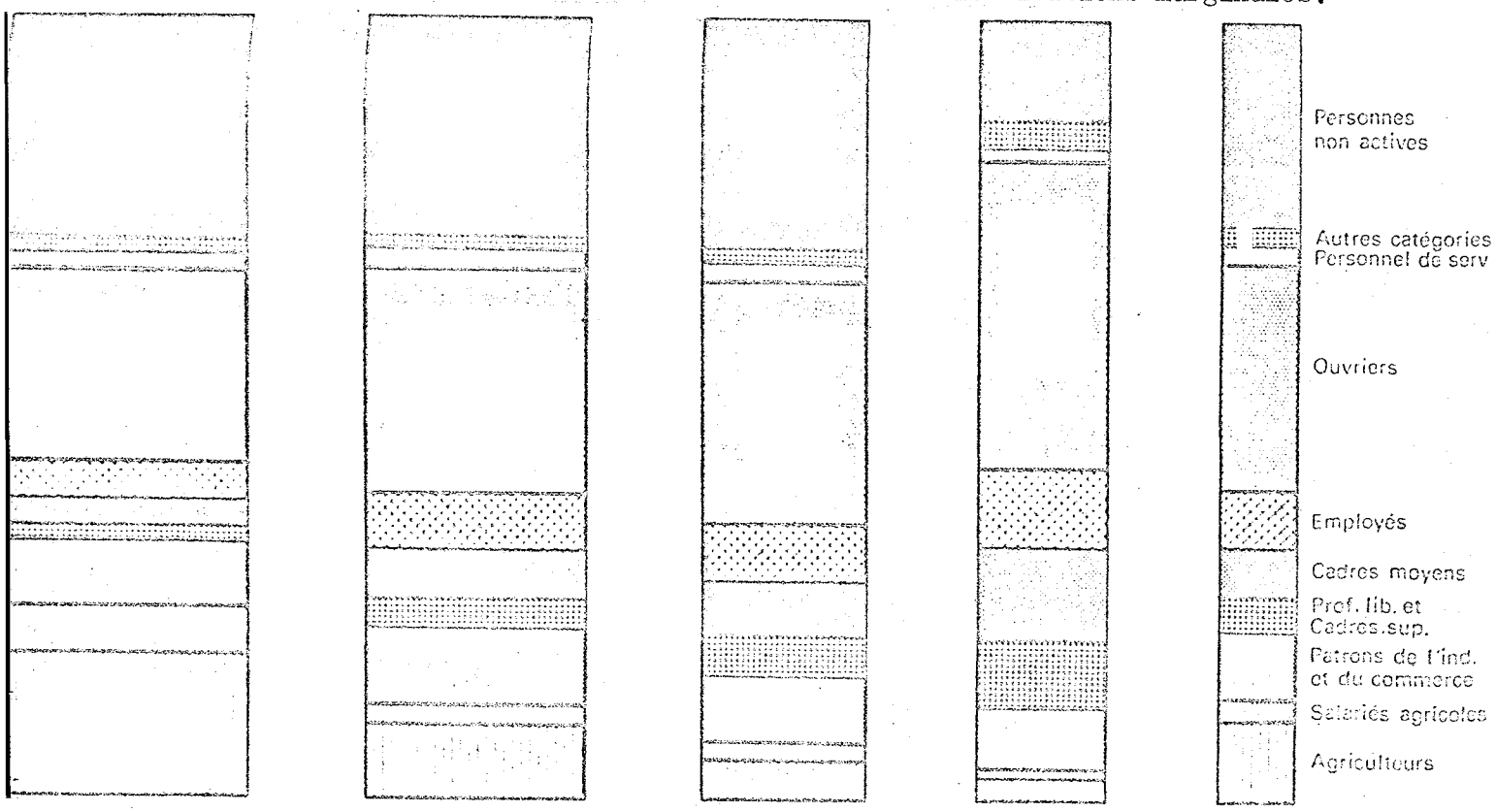
Catégorie socio-professionnelle \ Epoque de construction	Avant 1871	De 1871 à 1914	De 1915 à 1948	Après 1948
Agriculteurs	57,6	84,7	95,1	100,0
Salariés agricoles	57,0	81,5	93,4	100,0
Patrons de l'industrie et du commerce	31,8	63,5	85,0	100,0
Professions libérales et cadres supérieurs	14,1	42,6	66,8	100,0
Cadres moyens	16,1	43,8	67,3	100,0
Employés	22,3	53,3	76,1	100,0
Ouvriers	26,6	54,6	77,3	100,0
Personnel de service	32,0	67,0	87,6	100,0
Autres catégories	23,4	48,4	69,3	100,0
Personnes non actives	36,0	67,7	91,7	100,0
Toutes catégories socio-professionnelles réunies	32,2	61,7	83,1	100,0

Catégorie socio-professionnelle \ Epoque de construction	Avant 1871	De 1871 à 1914	De 1915 à 1948	Après 1948	Toutes époques de construction réunies
Agriculteurs	18,6	9,6	5,1	3,0	10,4
Salariés agricoles	23,6	11,9	6,7	4,1	13,2
Patrons de l'industrie et du commerce	32,5	21,6	15,7	12,1	22,2
Professions libérales et cadres supérieurs	34,4	25,7	20,5	20,4	26,4
Cadres moyens	37,5	31,5	27,3	32,3	32,6
Employés	42,4	39,0	34,9	42,4	39,8
Ouvriers	66,3	66,5	65,6	81,2	68,7
Personnel de service	68,7	69,4	67,9	83,0	71,1
Autres catégories	70,2	71,2	70,0	86,9	73,3
Personnes non actives	100,0	100,0	100,0	100,0	100,0



Distributions conditionnelles (Catégorie/Epoque)

Les dernières colonnes représentent les distributions marginales.

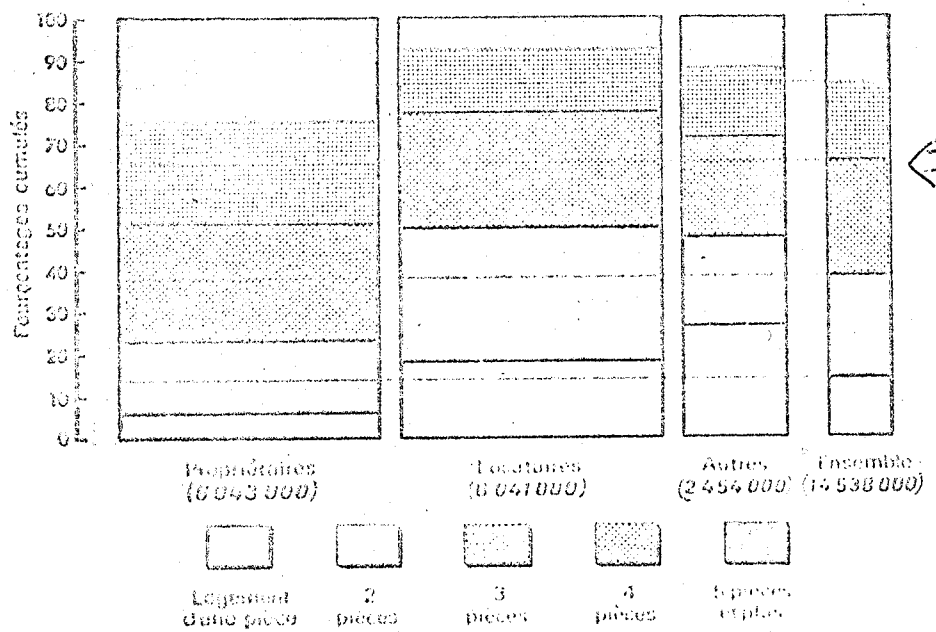


Distribution des logements suivant la catégorie socio-professionnelle du chef de ménage et l'époque de construction.

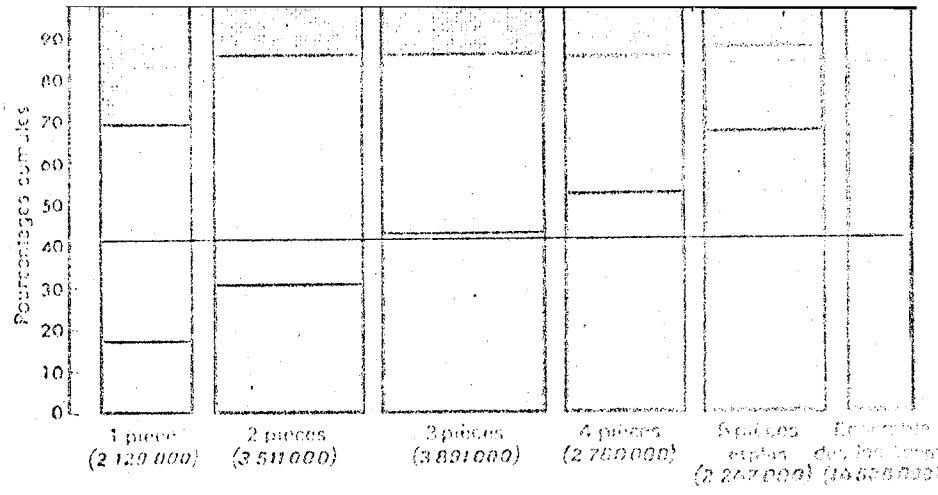
Exemple : Distribution du logement suivant le statut d'occupation et le nombre de pièces.

Statut d'occupation \ Nombre de pièces	1	2	3	4	5 et plus	Total
Propriétaires ...	371 600	1 078 420	1 639 960	1 460 700	1 491 800	6 042 480
Locataires (1)	1 099 040	1 923 280	1 673 760	868 140	477 140	6 041 360
Autres (2).....	658 540	509 560	576 720	431 440	278 100	2 454 360
Total	2 129 180	3 511 260	3 890 440	2 760 280	2 247 040	14 538 200

(1) Locataires d'un logement loué vide.
(2) Personnes logées par leur employeur (1 280 000), personnes logées à titre gracieux (670 000), locataires ou sous-locataires d'un local meublé (500 000).

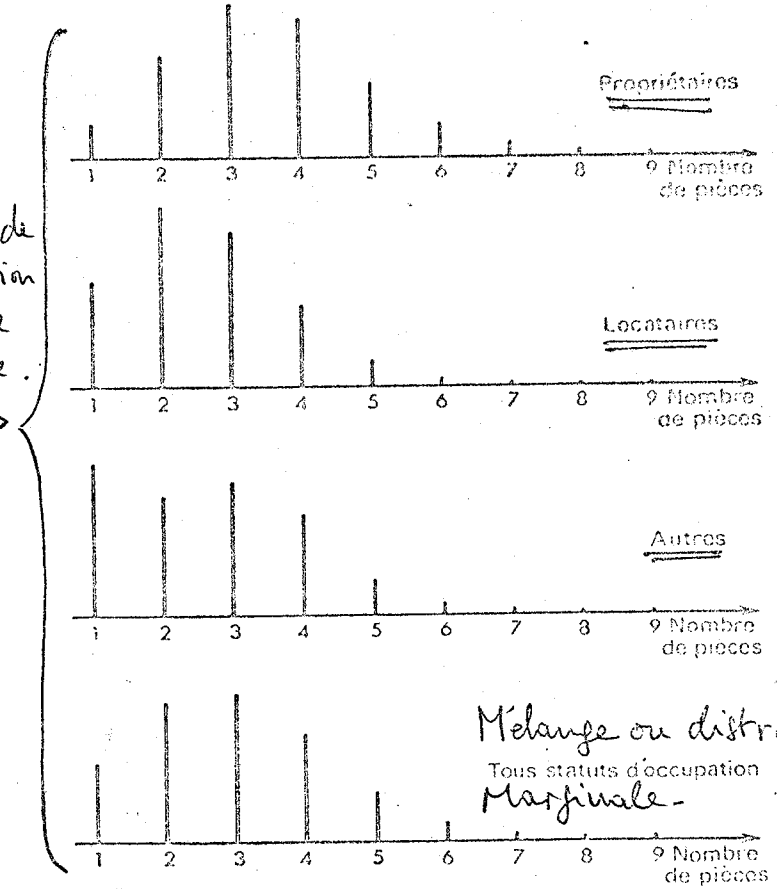


a) Représentation de la distribution globale



b) Représentation de la distribution globale et des distributions conditionnelles suivant le statut d'occupation.

Possibilité de représentation sur un axe numérique.



c) Représentation des distributions conditionnelles suivant le nombre de pièces.

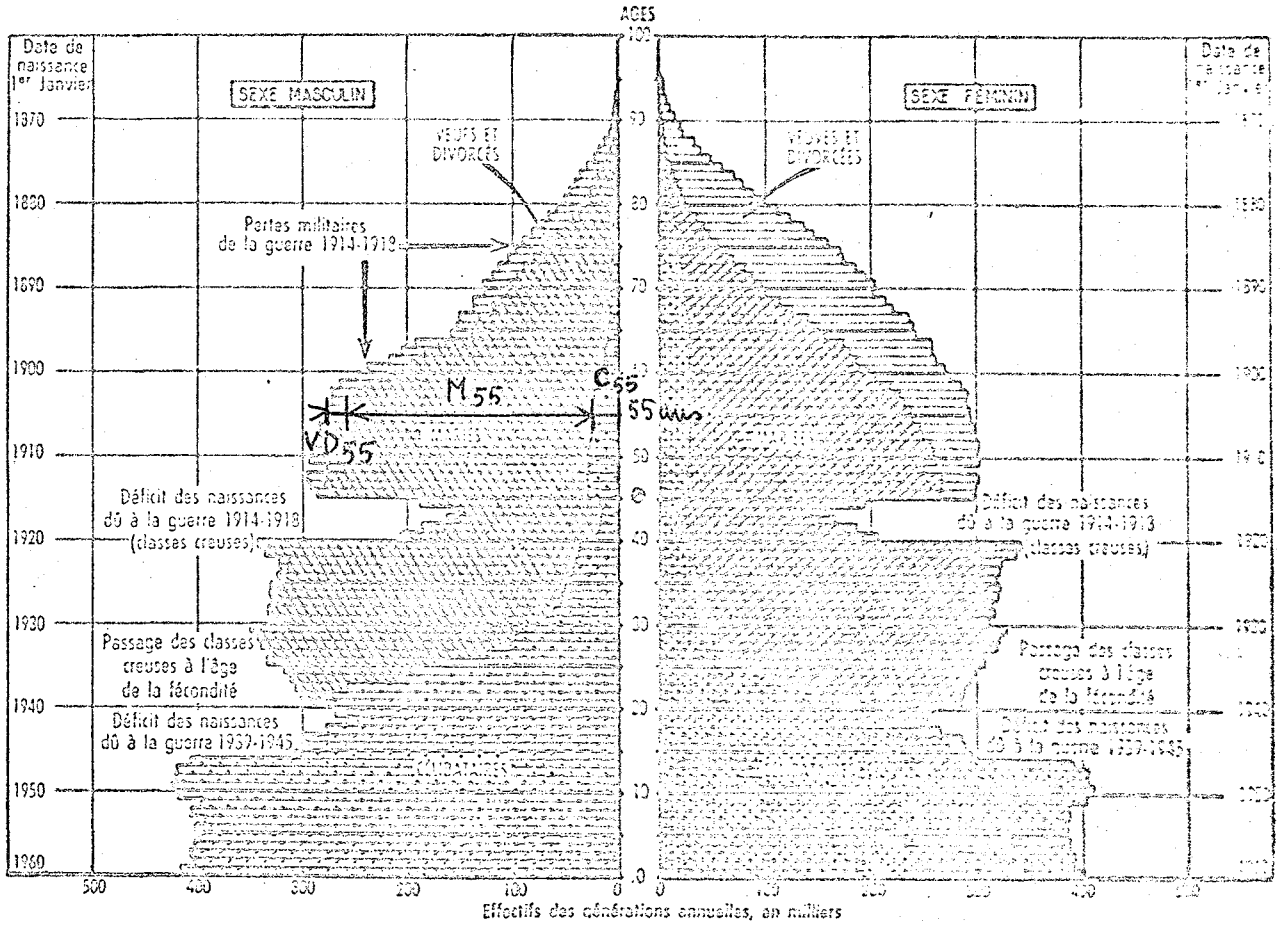
Mélange ou distributive Marginale.

Autre exemple. Distribution de la population française masculine suivant l'âge et l'état matrimonial.

1) Pyramide, où apparaît, à âge fixé les distributions conditionnelles de l'état matrimonial.

2) Distribution conditionnelle suivant l'âge pour les 3 modalités de l'état matrimonial.

PYRAMIDE DES AGES DE LA POPULATION FRANÇAISE PAR ÉTAT MATRIMONIAL - ÉVALUATION AU 1-1-1960
Source : I. N. S. E. E.

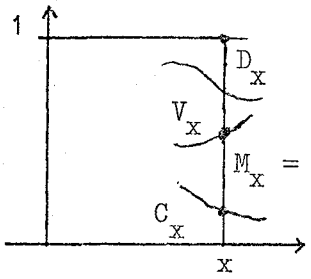


Notons C_x = effectifs des célibataires à l'âge x

M_x = " " mariés à l'âge x

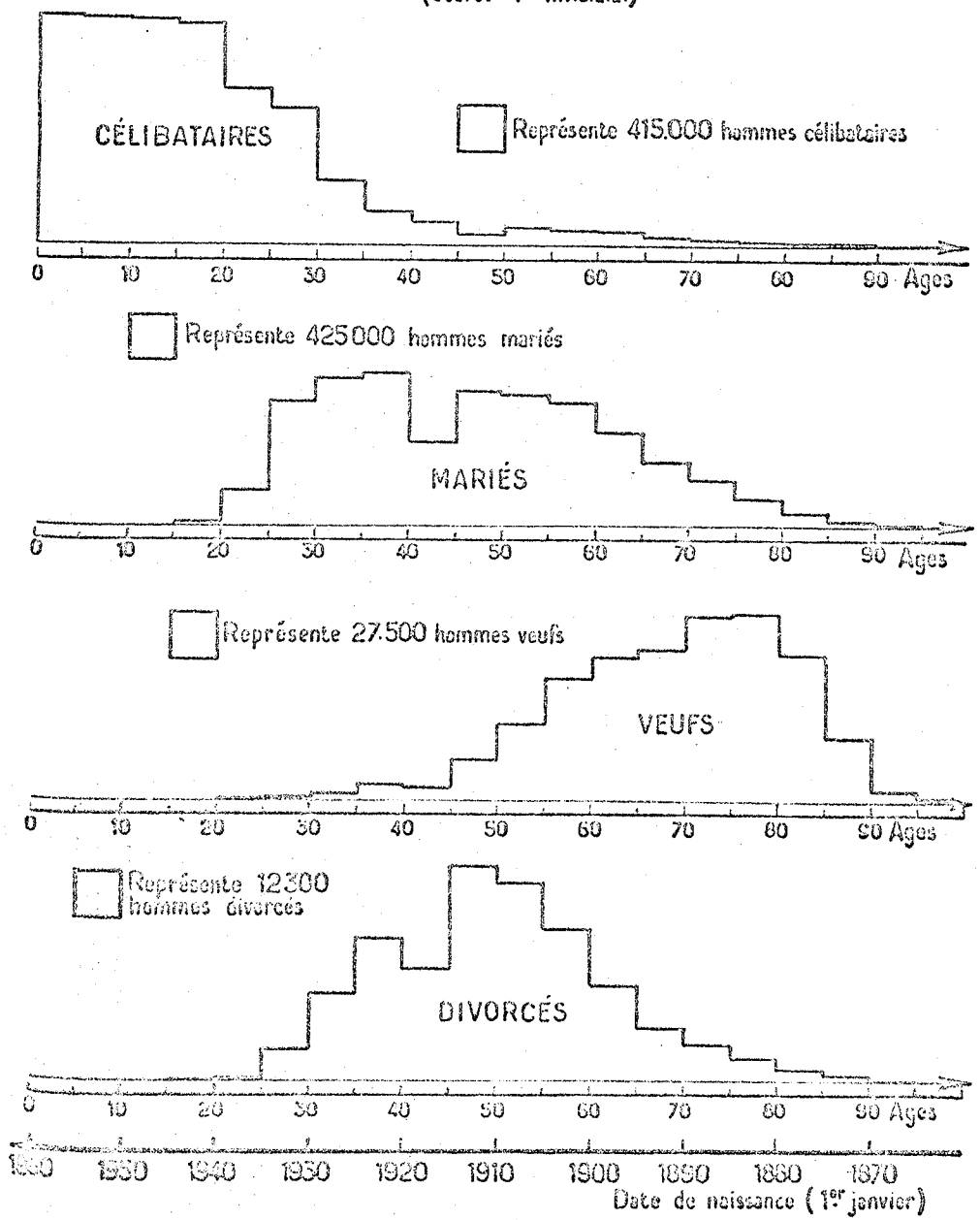
VD_x = " " veufs ou divorcés " x .

On peut considérer x comme une variable continue, et ainsi représenter les distributions des 4 modalités de l'état matrimonial pour tout x (Si on sépare V et D)

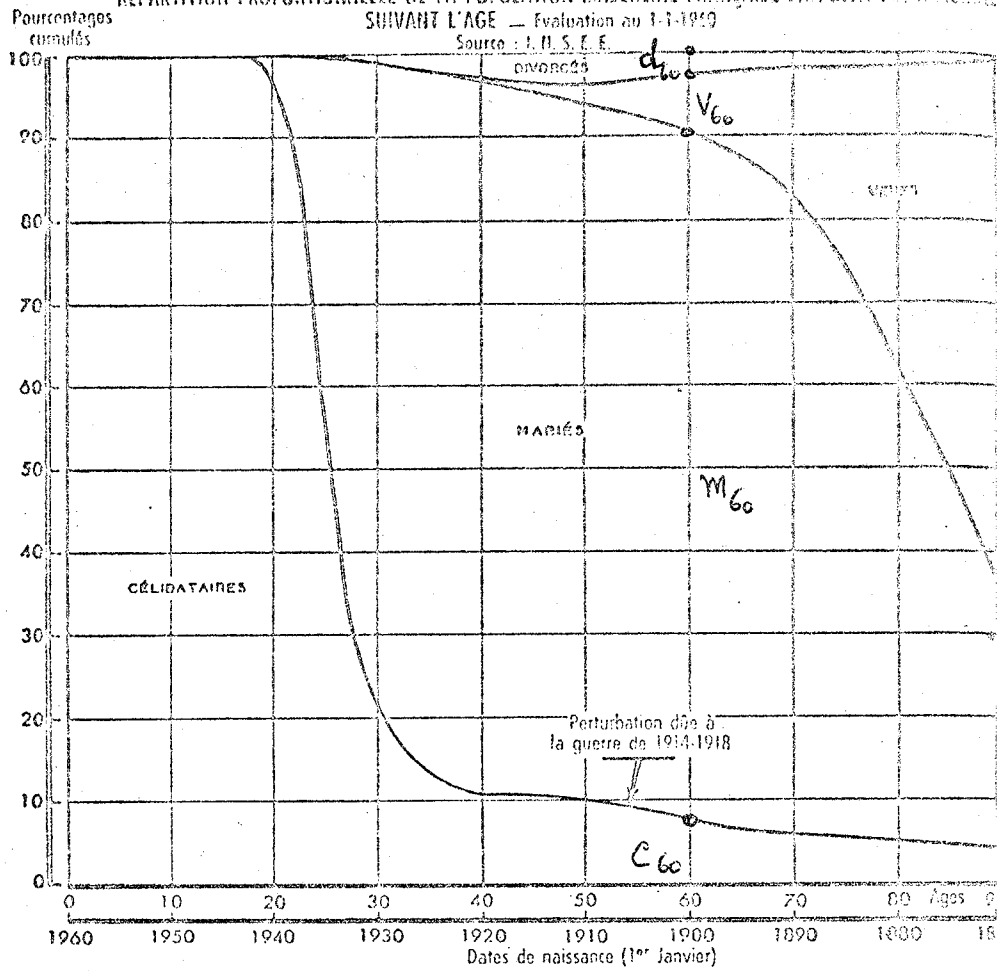


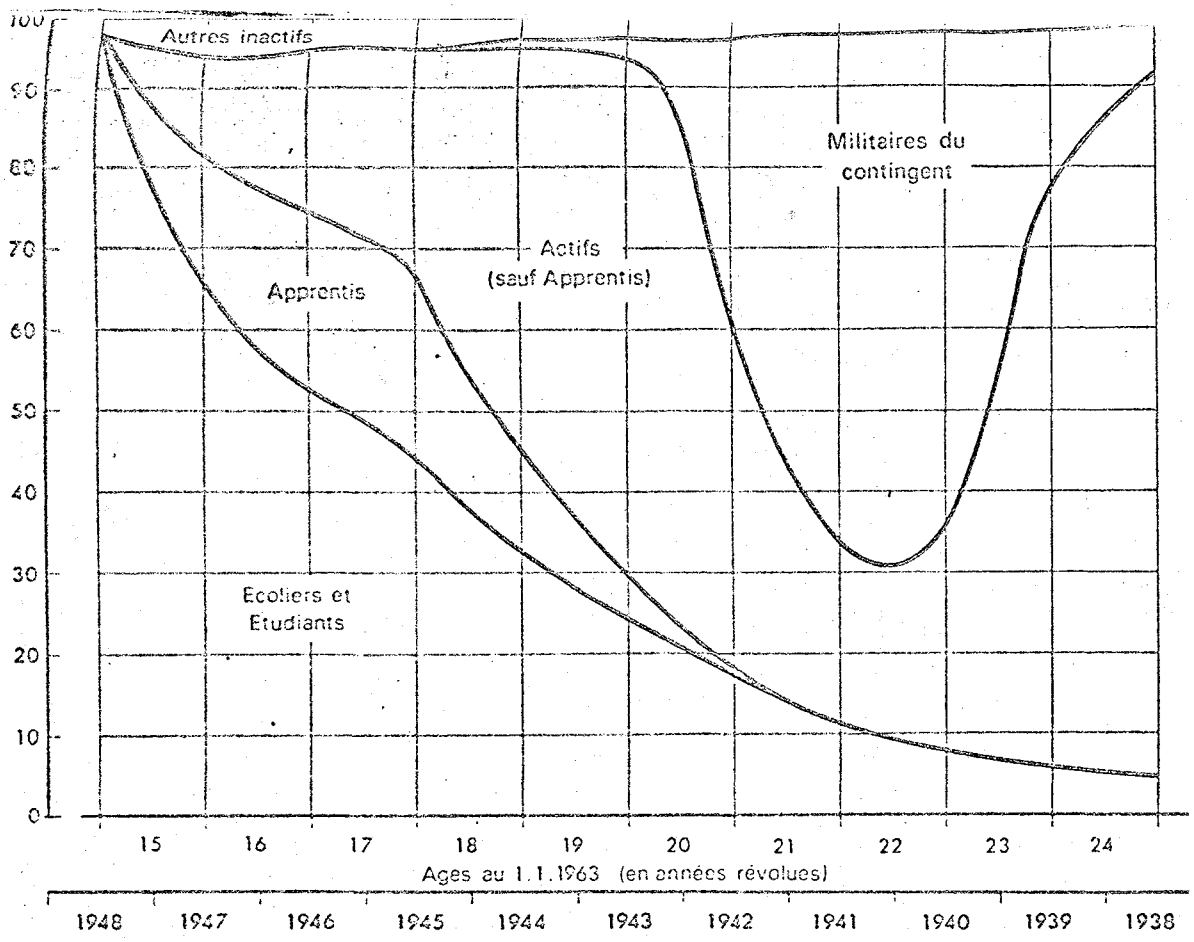
M_x = °/₀ de mariés à l'âge $x = \dfrac{M_x}{C_x+M_x+VD_x}$

Evaluation du 1er janvier 1960
(Source : I.N.S.E.E.)



RÉPARTITION PROPORTIONNELLE DE LA POPULATION MASCULINE FRANÇAISE PAR ÉTAT MATRIMONIAL
SUIVANT L'AGE — Evaluation au 1-1-1959
Source : I. N. S. E. E.





Répartition de la population masculine de 15 à 24 ans suivant le type d'activité
(Recensement général de la Population ; mars 1962).

3 derniers cas : Deux caractères quantitatifs.

On peut utiliser déjà les différents modes de représentation énoncés dans le 1^{er} et $2^{ème}$ cas.

Notons X et Y les deux variables définissant les caractères. Par exemple, si X et Y sont discrètes et n_{ij} est l'effectif (x_i, y_j), on peut représenter cet effectif par un cercle centré en (x_i, y_j) de surface % à n_{ij} .

Deux exemples sont donnés :

1) Répartition des ménages français suivant le nombre de personnes et le nombre de pièces du logement occupé.

2) Répartition des ménages suivant l'âge du chef de ménage et le nombre d'enfants de moins de 16 ans.

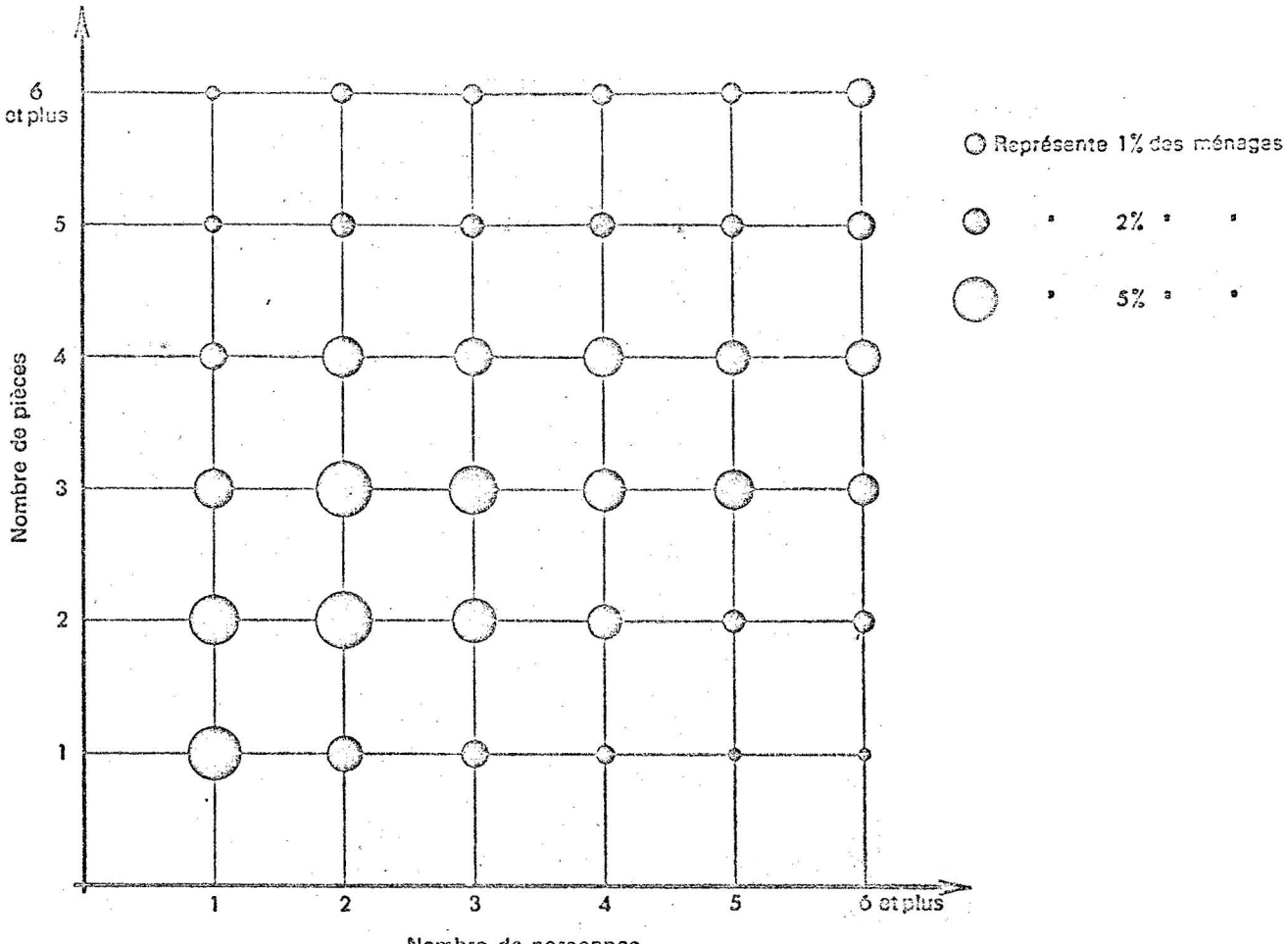
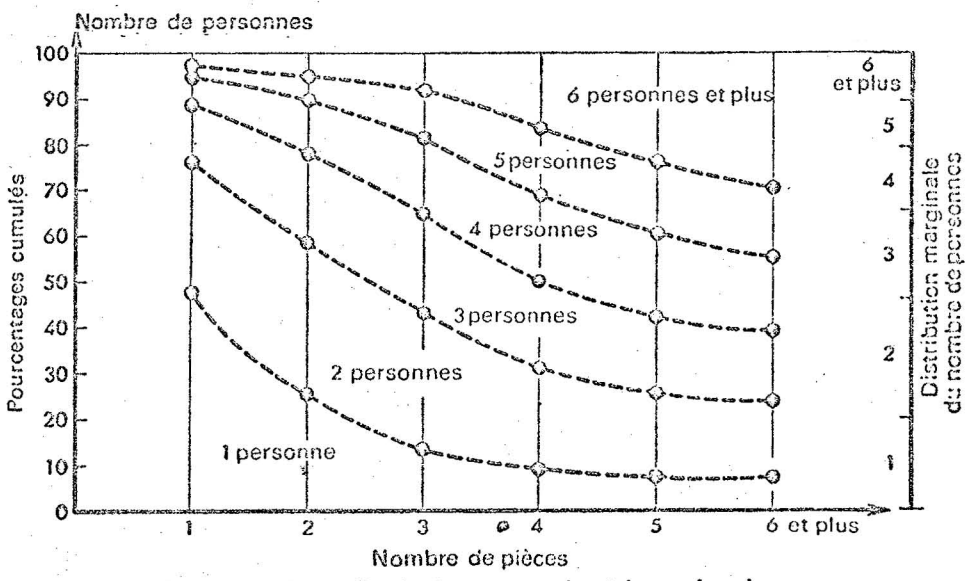
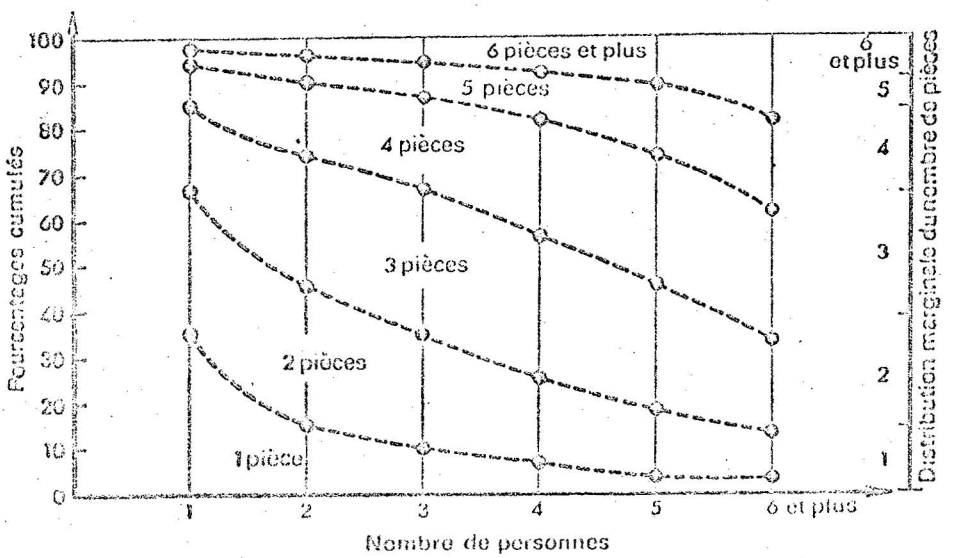


Tableau des données :
n_{ij} = nombre de familles

Nbre de pièces / Nbr. de personnes	1	2	3	4	5	6 et plus	Total
1	1 019 420	880 860	536 320	248 180	95 140	71 680	2 851 600
2	597 120	1 181 480	1 128 640	607 420	230 180	156 920	3 901 760
3	274 920	691 180	856 000	524 460	215 960	151 600	2 714 120
4	134 160	403 620	674 040	534 020	230 860	159 940	2 136 640
5	57 280	192 180	375 840	396 300	201 260	147 140	1 370 000
6 et plus	46 280	161 940	319 620	449 900	301 020	285 340	1 564 980
Total	2 129 180	3 511 260	3 890 440	2 760 280	1 274 420	972 620	14 538 200

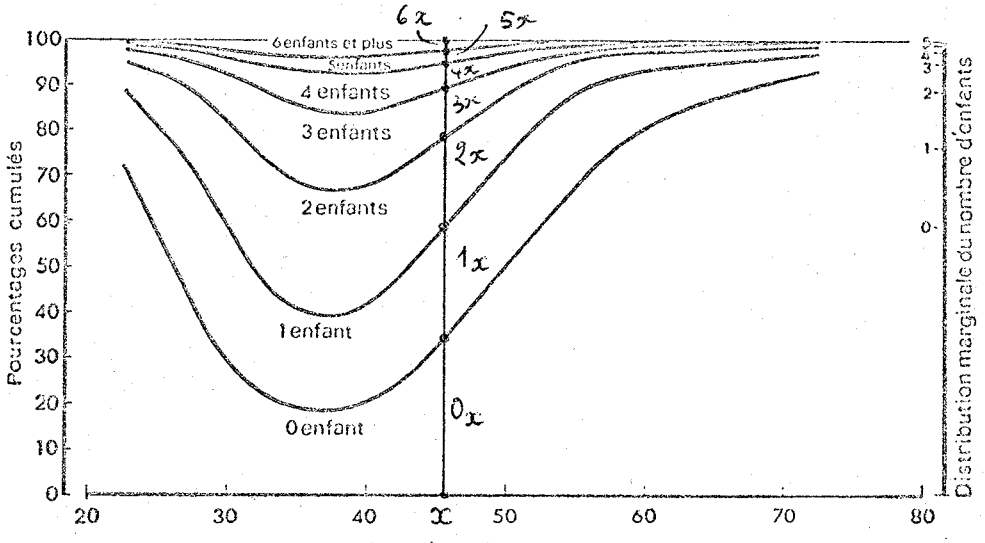


Distributions conditionnelles des logements suivant le nombre de personnes en fonction du nombre de pièces.



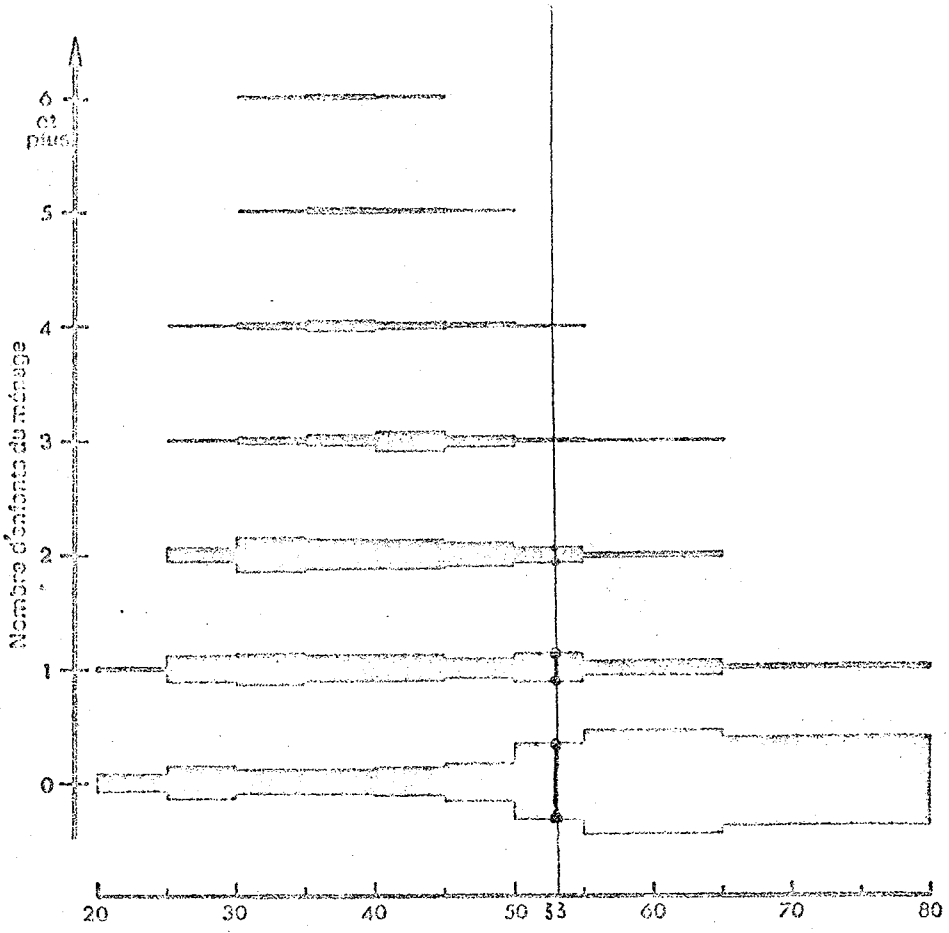
Répartition proportionnelle des ménages suivant l'âge du chef de ménage et le nombre d'enfants de 16 ans et moins (Recensement général de 1962).

Nbre d'enfants du ménage / Age du chef de ménage	0	1	2	3	4	5	6 et plus	Total
— 25 ans —	1 477	376	131	35	11	3	1	2 034
— 30 ans —	2 736	2 142	1 192	430	133	43	24	6 700
— 35 ans —	2 120	2 372	2 496	1 345	597	249	177	9 356
— 40 ans —	1 983	2 130	2 612	1 688	892	424	336	10 065
— 45 ans —	2 315	2 159	2 137	1 309	690	308	263	9 181
— 50 ans —	3 086	1 913	1 313	685	345	147	111	7 600
— 55 ans —	6 275	2 420	1 072	433	193	68	53	10 514
— 65 ans —	17 080	2 489	886	281	107	34	23	20 820
	22 139	951	406	158	55	16	5	23 730
Total	59 131	16 952	12 245	6 364	3 023	1 292	993	100 000

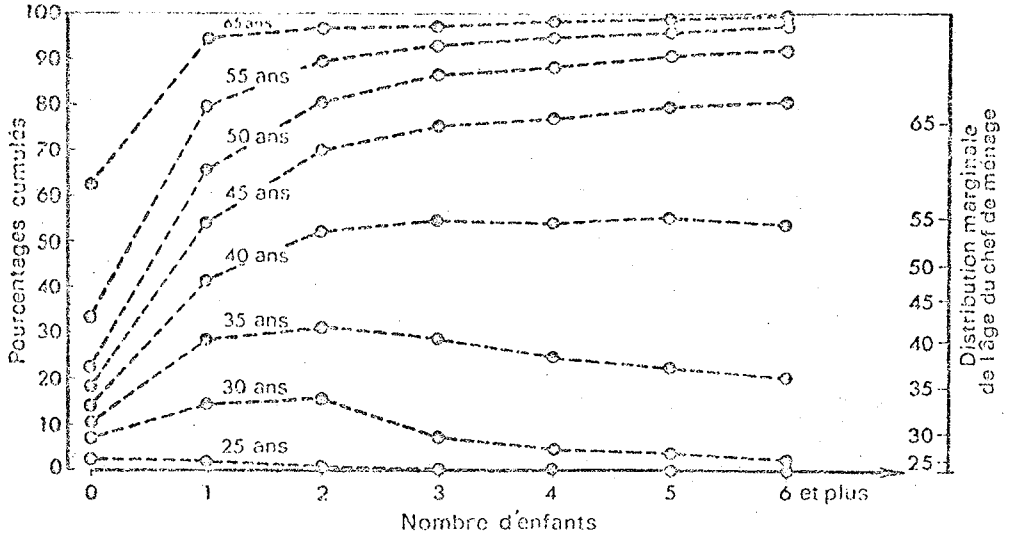


Age du chef de ménage

Distributions conditionnelles des ménages suivant le nombre d'enfants en fonction de l'âge du chef de ménage.



Distribution des ménages suivant l'âge du chef de ménage et le nombre d'enfants de 16 ans et moins.



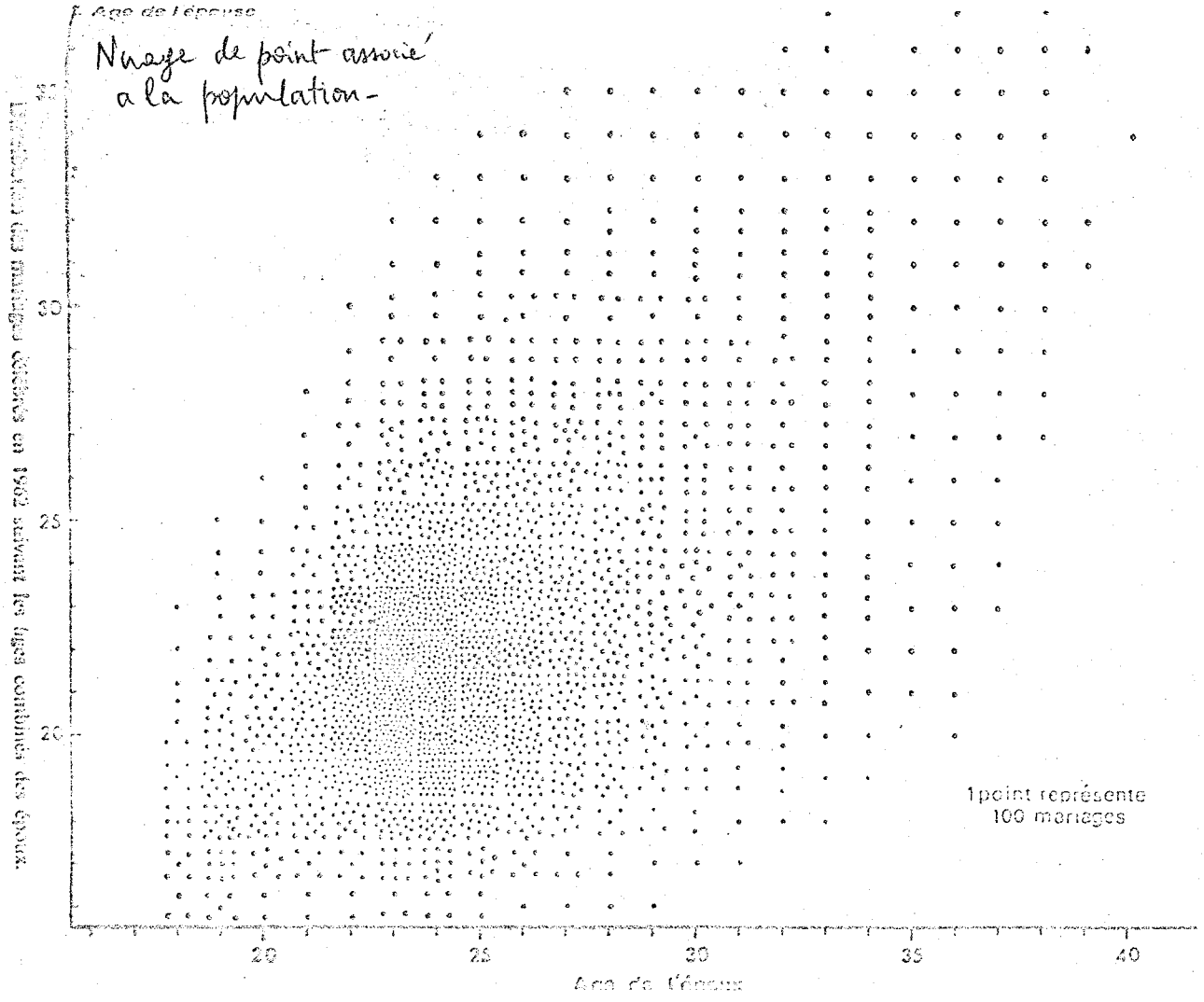
Distributions conditionnelles des ménages suivant l'âge du chef de ménage en fonction du nombre d'enfants.

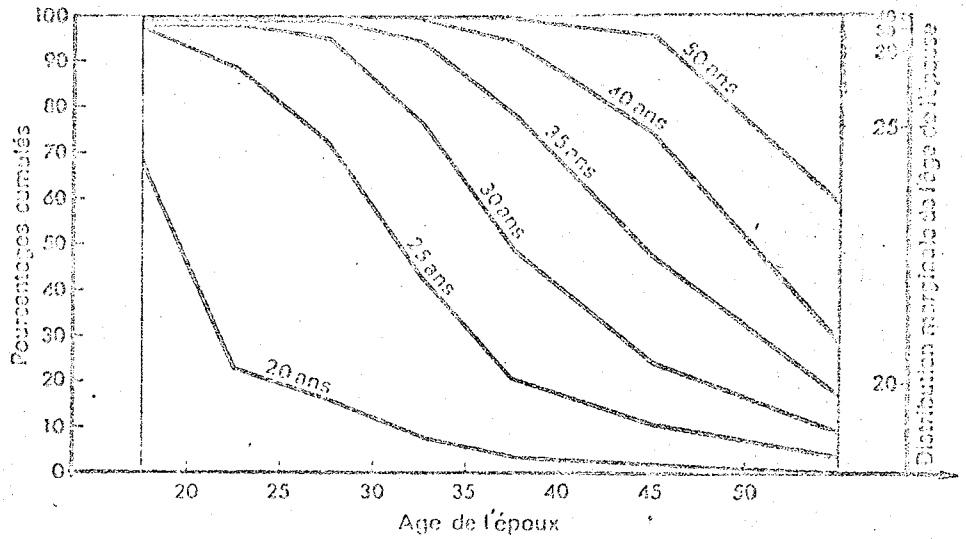
Autre exemple : Distribution des mariages suivant l'âge de l'époux et l'âge de l'épouse.

Exemple 3.

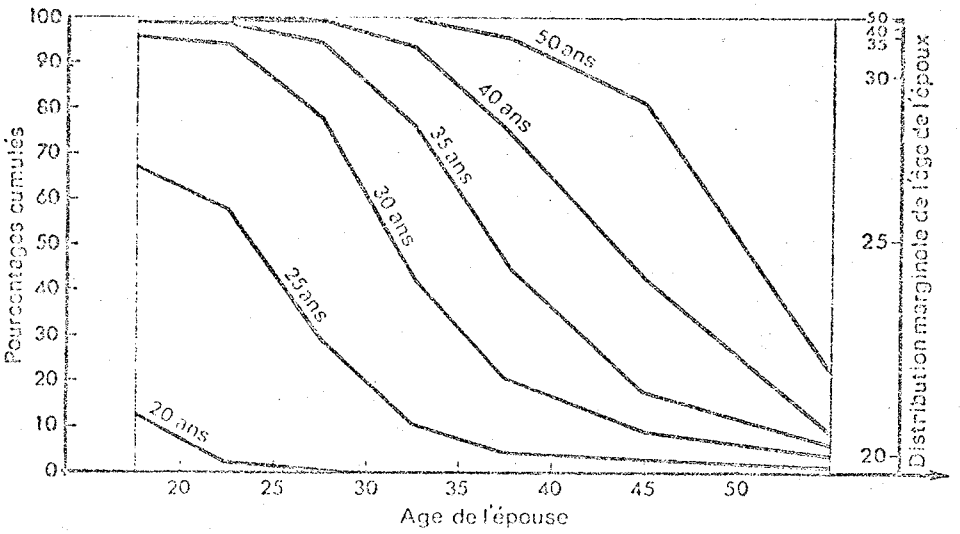
Distribution des mariages célébrés en 1962 suivant les âges combinés des époux (1).

Age de l'époux / Age de l'épouse	20	25	30	35	40	50	Total
20	6 756	3 051	180	15	3	3	0	10 017
25	29 446	84 556	13 430	1 205	168	50	10	128 885
30	15 893	54 978	22 774	3 890	651	113	14	98 313
35	1 789	8 289	7 809	4 111	1 021	244	15	23 278
40	255	1 304	1 906	2 078	1 232	362	20	7 247
50	66	283	447	733	852	697	120	3 200
Total	6	46	59	83	145	336	472	1 147
	54 190	152 507	46 695	12 115	4 072	1 807	651	272 057





Distributions conditionnelles de l'âge de l'épouse en fonction de l'âge de l'époux.

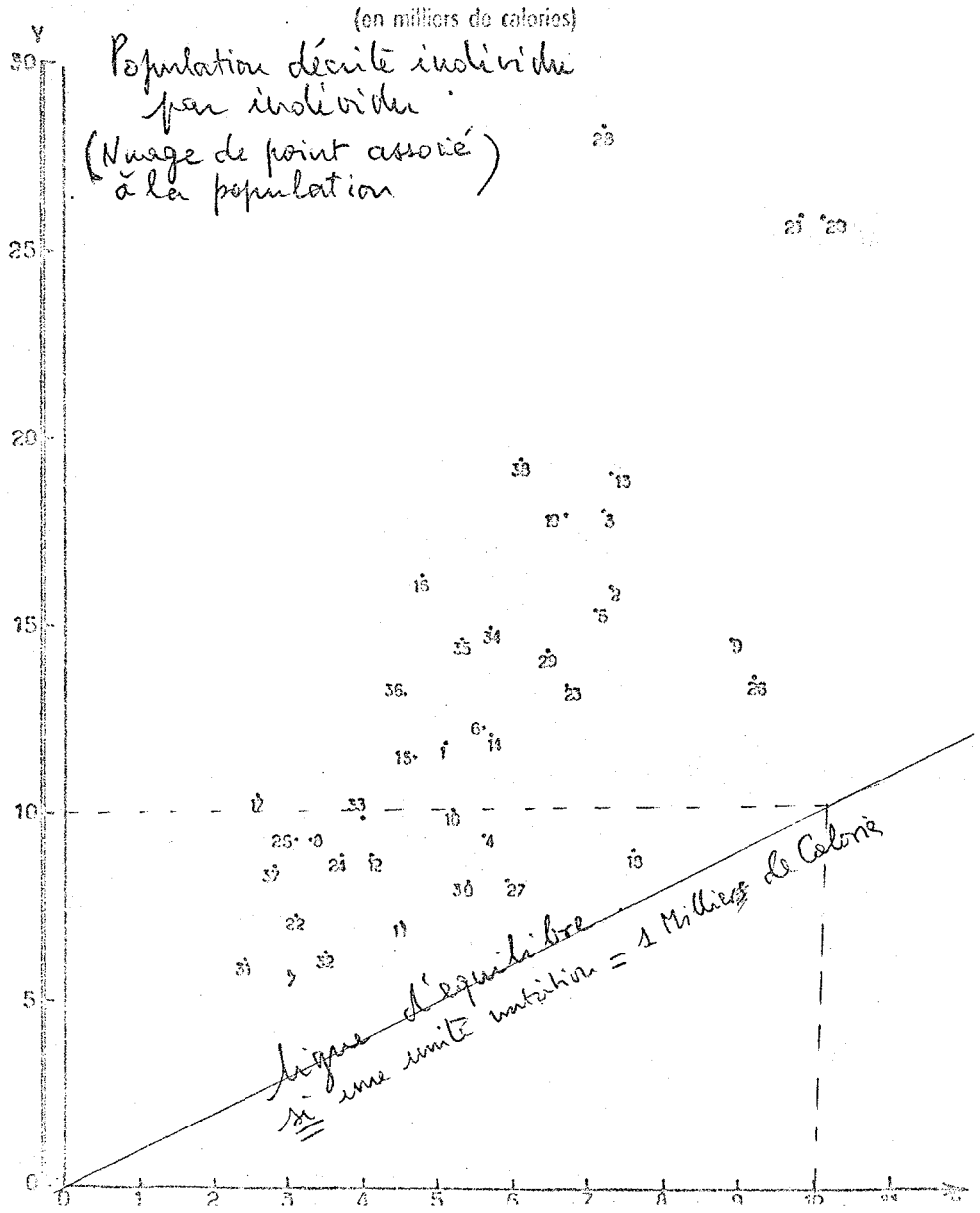


Distributions conditionnelles de l'âge de l'époux en fonction de l'âge de l'épouse.

Autre exemple : Description dans un plan de tous les individus

Ménage n°	x	y	Ménage n°	x	y	Ménage n°	x	y
1	5,1	11,9	14	5,7	12,1	27	5.9	8,2
2	7,3	16,0	15	4,7	11,5	28	7.2	20,2
3	7,2	13,0	16	4,8	10,3	29	6.4	14,3
4	5,5	9,4	17	2,6	10,5	30	5.4	8,2
5	7,1	15,4	18	7.6	9,0	31	2.4	6,1
6	5,6	12,3	19	6,7	17,9	32	3.5	6,3
7	3,0	5,8	20	10.1	25,8	33	4.0	9,9
8	3,3	9,3	21	9.8	25,8	34	5.7	14,9
9	8,9	14,6	22	3,1	7.3	35	5.3	14,6
10	5,2	10,1	23	6,7	13,4	36	4.6	13,2
11	4,5	7.1	24	3,7	8,9	37	2.8	8,6
12	4,1	8,9	25	3,1	9,3	38	6,1	19,4
13	7,3	19.0	26	9.2	13.6			

DISTRIBUTION DE 38 MÉNAGES SUIVANT LE NOMBRE D'UNITÉS DE NUTRITION (X) ET LA CONSOMMATION (Y)
(en milliers de calories)



- Consommation : nombre de milliers de calories consommées par ménage (y)
- Unités de nutrition : besoins alimentaires du ménage (calculés théoriquement suivant nombre de personnes du ménage, sexe, âge, travail effectué ...) = (x)

(Sondage effectué dans une ville du Proche Orient).

Cas particulier : Séries chronologiques.

Lorsque l'un des caractères est le temps, la série statistique s'appelle série chronologique (l'autre caractère est quelconque).

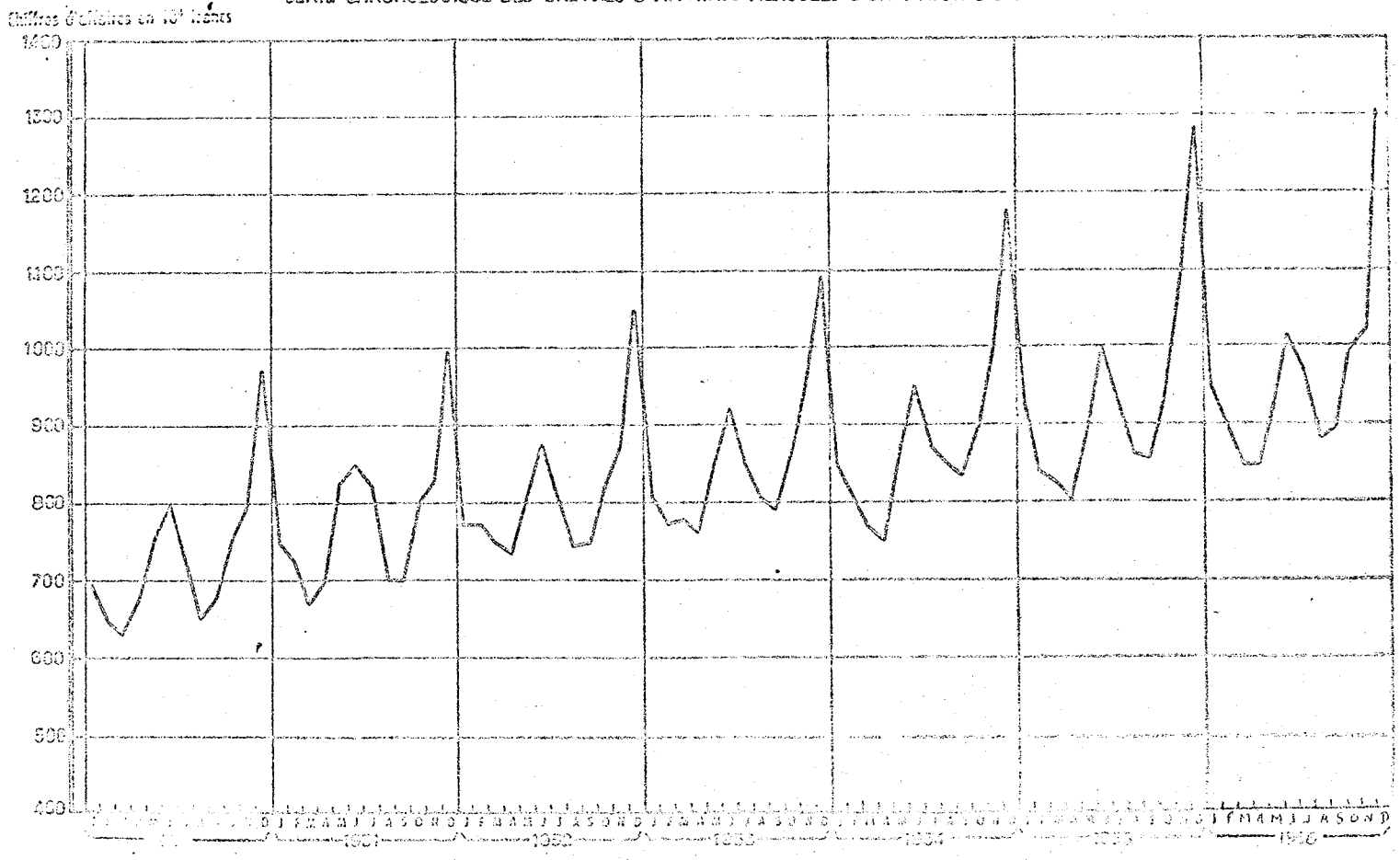
La liaison est fonctionnelle

Série des chiffres d'affaires mensuels d'un rayon d'un grand magasin (exemple cité par R. Dumas dans L'Entreprise et la Statistique; Dunod éditeur, Paris, 1954).

Unité : 10 000 francs

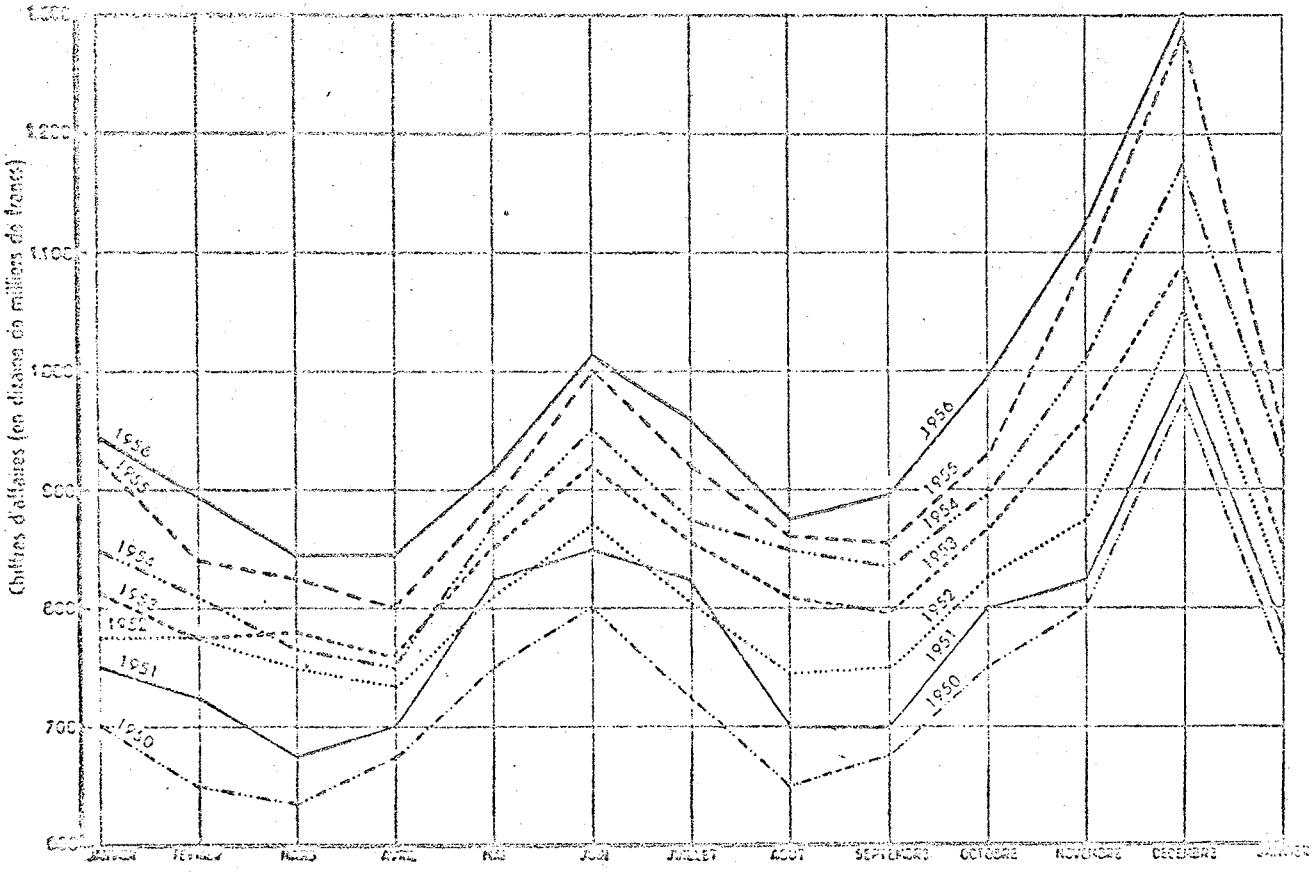
Année / Mois	1950	1951	1952	1953	1954	1955	1956
Janvier	700	750	775	815	850	925	945
Février	650	725	775	775	810	870	895
Mars	635	675	750	780	765	825	845
Avril	675	700	735	760	750	830	845
Mai	750	825	810	850	875	850	915
Juin	800	850	870	920	950	1 000	1 015
Juillet	725	825	805	855	875	920	950
Août	650	700	745	810	850	890	875
Septembre	675	700	750	795	835	865	895
Octobre	750	800	825	865	895	930	965
Novembre	800	825	875	960	1 010	1 080	1 120
Décembre	975	1 000	1 050	1 080	1 175	1 285	1 390

SÉRIE CHRONOLOGIQUE DES CHIFFRES D'AFFAIRES MENSUELS D'UN RAYON D'UN GRAND MAGASIN



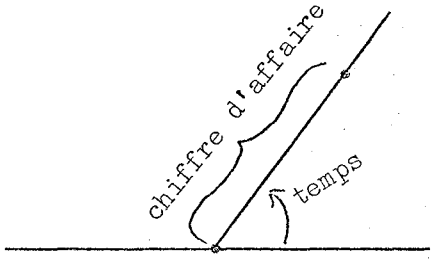
SÉRIE CHRONOLOGIQUE DES CHIFFRES D'AFFAIRES D'UN RAYON D'UN GRAND MAGASIN

Courbes annuelles superposées



Représentation polaire.

Sachant qu'une année correspond à un angle de 2π on peut représenter de façon polaire la série chronologique précédente.



Nota. Il existe bien d'autres représentations graphiques de distributions statistiques à deux caractères (cartogrammes dans la description d'unités géographiques suivant un caractère, représentation graphiques utilisant des papiers fonctionnels, graphique triangulaire...).

CHAPITRE I

Modèles statistiques. Théorie de la décision.

1. BUTS DE LA STATISTIQUE.

La démarche essentielle de la statistique consiste à préciser à partir de données expérimentales un modèle probabiliste mal connu.

Il y aura de très nombreuses façons de traduire le verbe préciser : estimer, tester, classer, décider, etc...

Par exemple, si l'on considère deux traitements par engrais d'une même plante, s'ils donnent les tailles suivantes pour 10 pousses

30	37	39	34	36	38	32	32	31
31	38	38	31	31	34	35	32	32

peut-on dire et en quel sens que l'un est meilleur que l'autre si le modèle probabiliste consiste à associer à chaque traitement une loi de probabilité, soit F_1 et F_2, a-t-on F_1 "meilleur" que F_2 en un certain sens (par exemple moyenne de $F_1 \geq$ moyenne de F_2 ou autre critère, médiane $F_1 \geq$ médiane F_2).

Le travail du statisticien est donc d'examiner, à l'aide d'outils mathématiques, des données, qui sont les résultats d'une expérience et situées dans un certain ensemble de modèles probabilistes ; soit $(\Omega, \mathcal{A}, P_\theta)_{\theta \in \Theta}$ (on a l'habitude de toujours prendre Ω, $\mathcal{A}$ identiques pour tous les modèles car en fait $(\Omega, \mathcal{A})$ représente la structure des données qui est évidemment indépendante du modèle réel et inconnu [fixé par une valeur θ_o de θ, $\theta_o \in \Theta$].

Le choix de l'ensemble $(\Omega, \mathcal{A}, P_\theta)_{\theta \in \Theta}$ des modèles relève en principe, de l'expérimentateur, de l'utilisateur

mais en pratique, le statisticien est fréquemment amené à participer au choix de la collection de modèles.

La réalisation de l'expérience elle-même relève aussi du statisticien, en ce qu'il participe à l'élaboration du plan (déroulement) de l'expérience. Ceci est dû au fait suivant, la quantité d'information , la précision et la richesse des déductions que l'on peut faire à partir du traitement de données (résultats d'une expérience) dépend de la structure mathématique de ces données. On peut gagner du temps et du matériel en organisant l'expérience d'une certaine manière, c'est là l'objet de l'étude des plans d'expériences.

De plus, il ne sert à rien d'élaborer des modèles complexes si on ne peut pas travailler dessus concrêtement. Donc même au point 1, les techniques statistiques d'extraction de l'information à partir des données sont des plus importantes. Cependant dans certains domaines comme les sciences humaines et l'économie, il est fréquent d'avoir à traiter des données brutes car l'expérience n'est pas modifiable par le statisticien.

La statistique peut être développée à partir de plusieurs points de vue :

a) Celui de la théorie de l'information (très lié à des concepts physiques de type entropie) qui a l'avantage de conduire à une théorie riche, mais l'inconvénient de reposer sur des concepts peu intuitifs pour un étudiant actuel en mathématiques. Nous ne l'aborderons pratiquement pas ici.

b) Celui de la théorie de la décision, lié à la pratique des sciences expérimentales et sociales est celui que nous choisirons.

De toute manière, une fois posés les concepts de base, les différents développements se recouvrent et l'ensemble de la théorie garde une certaine unité.

2. MODELE STATISTIQUE. STATISTIQUE.

Définition. Soit $(\Omega, \mathcal{A})$ un espace mesuré, $(P_\theta)_{\theta\in\Theta}$ une famille de probabilités sur cet espace, indexée par θ. L'ensemble $(\Omega, \mathcal{A}, P_\theta ; \theta \in \Theta)$ est appelé modèle statistique. C'est donc une famille d'espaces de probabilité. Ayant choisi un modèle statistique $(\Omega, \mathcal{A}, P_\theta ; \theta \in \Theta)$ et réalisé une expérience, le statisticien doit indiquer quelles sont les valeurs de θ qui sont non contradictoires, en un sens probabiliste à bien préciser, avec les résultats de l'expérience.

L'exemple suivant est à la fois simple et non trivial.

Exemple 1. On joue à pile ou face, et on appelle θ la probabilité de tirer pile, $1-\theta$, celle de tirer face. On peut faire n tirages (n-expériences) indépendants.

Soit X_i la v.a. qui représente la i-ème expérience : $X_i = 1$ si on a obtenu pile au i-ème coup, $X_i = 0$ sinon. Ce qu'on connait, c'est donc une suite de nombres $+1, 0$, qui représente les résultats de ces n-expériences, $(x_1, x_2 \dots x_n)$.

D'après la loi des grands nombres, on sait que :

$$P_{\theta'}\left(\left|\frac{X_1+X_2+\dots X_n}{n} - \theta\right| > \varepsilon\right) = A(\theta,\theta',\varepsilon) \text{ tend vers } 0 \text{ si } n \to +\infty$$

si et seulement si $\theta = \theta'$.

Considérons le résultat de l'expérience faite $(X_1(\omega)\dots X_n(\omega)) = (x_1, x_2, \dots x_n)$, et θ_o une valeur telle que

$$\left|\frac{x_1+x_2+\dots x_n}{n} - \theta_o\right| > \varepsilon .$$

Si θ est égal à θ_o, on aurait réalisé l'évènement

$$B(\varepsilon) = \left\{\left|\frac{X_1+X_2+\dots X_n}{n} - \theta_o\right| > \varepsilon\right\}$$

avec la probabilité $P_{\theta_o}(B\varepsilon) = A(\theta_o, \theta_o, \varepsilon) = \eta$.

Si η est petit, on admet qu'il est plus que douteux que l'on puisse se trouver dans la situation $\theta = \theta_o$.

Intuitivement, on rejette la valeur θ_o, et aussi les valeurs très voisines de θ_o .

Statistique - Image d'un modèle statistique par une statistique.

Définition : On appelle statistique sur le modèle statistique $\{\Omega, \mathcal{A}(P_\theta)_{\theta \in \Theta}\}$ toute fonction mesurable T de $(\Omega, \mathcal{A})$ dans un espace mesurable $(\Omega', \mathcal{A}')$.

Observer une statistique T, c'est observer un résultat déduit de l'expérience ; ou c'est remplacer le résultat x de l'expérience par la fonction T(x). On peut associer à T, considéré maintenant comme le résultat de l'expérience, un autre modèle statistique

$$\{\Omega', \mathcal{A}', \{T(P_\theta)\}_{\theta \in \Theta}\}$$

appelé image du 1er modèle par T. Il est fréquent qu'une statistique T donne sur le paramètre autant d'information que l'expérience : le modèle image par T du 1er modèle sera alors le modèle naturel. Nous préciserons plus tard cette idée (Chapitre 8 : Exhaustivité). Indiquons ici quelques exemples, qui seront précisés au Chapitre 8.

Exemple : Soit $X = (X_1, X_2, \ldots, X_n)$ un n-échantillon d'une famille de lois $(\nu_\theta)_{\theta \in \Theta}$ sur $\mathbb{R}^d$.

Statistique d'ordre.

Intuitivement, l'ordre des observations n'a pas d'importance puisque les X_i sont indépendantes, et on peut par exemple remplacer la suite des observations $(x_1, \ldots, x_n)$ par la suite ordonnée $(x_{\sigma(1)}, \ldots, x_{\sigma(x)})$ qui s'en déduit par la permutation σ telle que $x_{\sigma(1)} \leq x_{\sigma(2)} \leq \ldots \leq x_{\sigma(n)}$. La statistique T correspondante, appelée statistique d'ordre, a une loi qui dépend de la valeur de θ, mais il est intuitif que la loi de X conditionnée par T est indépendante de θ: si on sait que T appartient à un borélien A de $(\mathbb{R}^d)^n$ la probabilité pour que X appartienne à un borélien A de $(\mathbb{R}^d)^n$ est égale, intuitivement à la proportion des permutés des points de B qui sont dans A.

3. COMMENT FORMULER UN PROBLEME DE STATISTIQUE. VOCABULAIRE ET FORMALISME DE LA THEORIE DE LA DECISION.

Il existe un certain nombre de divisions traditionnelles de la statistique. L'une concerne la nature de l'ensemble Θ : si cet ensemble est "simple", par exemple un ouvert de R^d, on dit que l'on a un modèle paramétrique à traiter, sinon on parle de ca non-paramétriques. Nous reviendrons plus loin et plus à fond sur cette distinction.

Mais la division la plus importante consiste en la manière d'énoncer la conclusion : si celle-ci est du type "$\theta = \theta_o$", on parle de problème d'estimation. Si par contre $\Theta = \Theta_1 + \Theta_2$, et que la conclusion est du type $\theta \in \Theta_1$, on parle de problème de test. Il y a évidemment bien d'autres "conclusions" possibles ; pour préciser ce que nous entendons par conclusion, nous allons introduire le formalisme de la théorie de la décision.

Les décisions que va prendre le statisticien vont être fonction du résultat de ses expériences. Plus précisément, s'il a observé $X(\omega)$, sa décision va être une fonction $h[X(\omega)] = d(\omega)$.

Définitions. Soit $(\Omega, \mathcal{A})$ l'espace mesuré du modèle statistique et X une v.a. sur $(\Omega, \mathcal{A})$.

Soit $(A, \mathcal{A}')$ un espace mesuré, appelé espace des décisions.

Une règle de décision, d, (relative à X) est une v.a. de $(\Omega, \mathcal{A})$ dans $(A, \mathcal{A}')$ mesurable par (rapport à la tribu engendrée par X).

On va avoir besoin de classer les décisions. Pour cela, on se donne au départ une fonction de perte W : de $\Theta \times A \to R^+$, mesurable :

$W(\theta_o, d(\omega))$ est la perte qu'entraîne pour le statisticien (ou son employeur) le fait de prendre la décision $d(\omega)$ lorsque la vraie valeur du paramètre inconnu est θ_o.

Les fonctions de perte classiques sont :

$W(\theta,d) = |\theta - d|^2$ (perte quadratique qui intervient dans les problèmes d'estimation)

$W(\theta,d) = 1_{\Theta_2}(d)$ si $\theta \in \Theta_1$

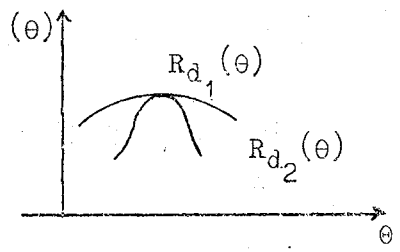
$= 1_{\Theta_1}(d)$ si $\theta \in \Theta_2$ (pénalisation de 1 pour toute mauvaise décision).

Comme on ne connait pas θ, on calcule, pour chaque valeur de θ la perte de moyenne dite risque

Définition. On appelle fonction de risque pour la décision d, la fonction R_d de $\Theta \to R^+$, définie par :

$$R_d(\theta) = E_\theta[W(\theta, d(\omega))]$$
$$= \int_\Omega W(\theta, d(\omega))\, dP_\theta(\omega).$$

L'ordre naturel sur les fonctions de domaine Θ établit une relation de pré-ordre sur les règles de décision :



une décision d_1 est dite meilleure que la décision d_2 si $R_{d_1}(\theta) \leqslant R_{d_2}(\theta)$ pour tout $\theta \in \Theta$.

Définitions. Une règle de décision est dite admissible, s'il n'en existe pas de strictement meilleure.

S'il existe un plus grand élément pour cet ordre, on parle de règle optimale.

En général, pour des raisons pratiques (calcul) ou liées au problème, on impose aux règles de décision des contraintes limitatives ; on se limite donc par avance à un ensemble $\mathcal{D}$ de règles.

Exemple : une décision d_o est dite minimax, si elle minimise le risque maximum :

$$\inf_d \sup_\theta R_d(\theta) = \sup_\theta R_{d_o}(\theta).$$

On peut ne s'intéresser qu'à ce type de décisions. Nous verrons plus tard d'autres classes de décisions.

En résumé, un problème statistique, vu dans le cadre de la théorie de la décision statistique est la donnée d'un modèle $(\Omega, \mathcal{A}, P_\theta)_{\theta\in\Theta}$; le choix d'une v.a. X qui correspond à l'expérience, d'un ensemble $(A, \mathcal{A})$ de valeurs de décisions possibles, éventuellement la donnée d'une classe $\mathcal{D}$ de v.a. à valeurs dans A, $\mathcal{B}(X)$-mesurables (ensemble des règles de décision auxquelles on se limite), d'une fonction de perte W de $\Theta \times A \to R^+$

Résoudre le problème, c'est trouver les règles de décision qui sont admissibles. Toutefois, le formalisme donné ne permet pas d'aller plus loin : c.à.d. il ne permet pas de choisir entre plusieurs règles admissibles s'il en existe. Le choix dépend alors du problème concret et du type de qualité qu'il vaut mieux imposer à la règle cherchée de ce point de vue (cf. exemple en fin du Chapitre III).

On peut généraliser la notions de règle de décision à celle de stratégie (ou stratégie aléatoire).

Une stratégie est une application (X-mesurable) $(\Omega, \mathcal{A}) \xrightarrow{S} \pi(A, \mathcal{B})$ où π est l'ensemble des probabilités sur A.

Si $B \subset A$, alors on choisit $d \in B$ avec la probabilité $S(X(\omega))(B)$.

Une règle de décision est aussi dite stratégie pure (ou non aléatoire). Elle correspond au cas où $S(X(\omega)) = \delta_{d(X(\omega))}$. δ_a mesure de Dirac en a.

Le risque d'une telle stratégie pour la fonction de perte $W(\theta, .)$ est

$$R_S(\theta) = E_\theta \left[\int_A W(\theta, u)\, S(\omega, du) \right]$$

si $S(\omega, du) = S(X(\omega))$. (Elément de $\pi(A)$).

On note $\hat{\mathcal{S}}$ l'ensemble des stratégies et $\mathcal{S}$ l'ensemble des règles de décision. Le lecteur définira sans peine, par analogue avec les décisions des stratégies admissibles, optimales, minimax.

Remarque : Soit $\phi \in \hat{\mathcal{S}}$ une stratégie admissible de risque constant ; alors ϕ est minimax .

En effet, si $\phi' \in \hat{\mathcal{S}}$, il est nécessaire que $\sup_{0 \in \Theta} R(0,\phi')$ soit supérieur ou égal à la valeur constante $k = R(0,\phi)$ $(0 \in \Theta)$ sinon ϕ' serait meilleure que ϕ et ϕ ne serait pas admissible. Donc :

$$k = R(0,\phi) = \sup_{0 \in \Theta} R(0,\phi) = \inf_{\phi' \in \hat{\mathcal{S}}} \sup_{0 \in \Theta} R(\theta,\phi')$$

Risques bayésiens. Θ est supposé mesuré.

Supposons que l'on munisse Θ d'une probabilité μ ou "loi a priori" du paramètre. Cela veut dire que l'on a a priori des idées sur les valeurs possibles de θ. Alors le risque moyen de $\phi \in \hat{\mathcal{S}}$ est :

$$r(\mu,\phi) = \int d\mu(\theta)\, R(\theta,\phi) .$$

en notant $R(\theta,\phi)$ pour $R_\theta(\phi)$.

Définitions: Soit μ une probabilité sur Θ.

Soit $\phi \in \hat{\mathcal{S}}$. On appelle risque bayésien de ϕ par rapport à la loi a priori μ, le nombre :

$$r(\mu,\phi) = \int_\Theta d\mu(\theta)\, R(\theta,\phi)$$

On dit que d est une fonction de décision bayésienne, par rapport à μ, si :

$$r(\mu,d) = \inf_{d' \in \mathcal{S}} r(\mu,d')$$

(On dit que ϕ est une stratégie bayésienne par rapport à μ, si :

$$r(\mu,\phi) = \inf_{\phi' \in \hat{\mathcal{S}}} r(\mu,\phi')$$

au lieu de dire "bayésienne par rapport à μ" on dira aussi "μ admissible".

Remarque : Il est souvent difficile de justifier théoriquement la donnée a priori d'une loi μ. C'est cependant parfois possible : dans le cadre de la théorie des jeux, si Θ est l'ensemble des jeux possibles du joueur ennemi, μ correspond à la stratégie de ce joueur. L'intérêt des notions bayésiennes est aussi technique ; dans de nombreux cas (cf. chap. 2) on sait calculer des stratégies bayésiennes, alors qu'il est difficile d'obtenir directement des stratégies admissibles. Cela donne souvent des stratégies admissibles.

Soient en effet μ une loi a priori et ϕ une stratégie bayésienne par rapport à μ. Soit ϕ' une stratégie aussi bonne que ϕ

$$\forall \theta , \quad R(\theta,\phi') \leq R(\theta,\phi)$$

D'où $r(\mu,\phi') \leq r(\mu,\phi)$.

Cela implique souvent qu'il n'existe aucun θ_o tel que $R(\theta_o,\phi') < R(\theta_o,\phi)$, c'est à dire que ϕ est admissible.

En particulier (vérifiez le) si :

- ϕ est l'unique stratégie bayésienne par rapport à μ (P_θ p.s. pour tout θ)
- Θ est dénombrable et μ charge tous les points de Π
- Θ est un ouvert de $\mathbb{R}^d$, μ charge tout ouvert inclus dans Θ (le support de μ est Θ) et la fonction $\theta \longmapsto R(\theta,\phi')$ est continue pour tout $\phi' \in \hat{\mathcal{S}}$.

Tout cela est vrai en remplaçant les stratégies (bayésiennes ou admissibles) par les fonctions de décision . (Cf. Chapitre II)

Un exemple de problème décisionnel bayésien.

Soit 2 populations de lois normales sur $\mathbb{R}^2$, respectivement $N_{\theta_1} = N(0, I)$, $N_{\theta_2} = N(m, 2I)$, I identité. On donne comme mesure a priori sur (θ_1, θ_2) la mesure (1/2, 1/2). Tracer une droite D dans le plan R^2, telle que la règle de décision qui consiste à classer un point dans la première population s'il est d'un côté de la droite, dans la deuxième sinon, soit optimale parmi ces règles.

Soit $\lambda X + \mu y = c$ la droite D avec $\lambda^2 + \mu^2 = 1$.

Soit X un point de coordonnées (X_1, X_2) la loi de $\lambda X_1 + \mu X_2$ est si θ_1 est vrai $N(0,I)$, si θ_2 est vrai $N(\langle m,a\rangle, 2I)$, si a est le vecteur (λ,μ).

Les risques sont donc

$$R(\theta_1, D) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c} e^{-x^2/2}\, dx$$

$$R(\theta_2, D) = \frac{1}{2\sqrt{\pi}} \int_{c}^{\infty} e^{-\frac{(x - \langle m,a\rangle)^2}{2}}\, dx$$

et le risque bayésien est $1/2\ |R(\theta, D) + R(\theta_2, D)|$

Minimisons d'abord en a, puis en c.

$\min_a R(\theta_2, D)$ est atteint pour $\langle m, a \rangle$ maximum, soit comme a est normé, pour $\lambda = \frac{m_1}{m_1 + m_2}$, $\mu = \frac{m_2}{m_1 + m_2}$

D est donc orthogonale à m , c est alors défini

$$e^{-c^2/2} - e^{-\frac{(c - ||m||)^2}{2}} = 0$$

soit $c = \frac{||m||}{2}$.

COMPLEMENT DE COURS

Modèle fini de la théorie de la décision

Les notations sont celles du paragraphe précédent. Supposons que Θ est un ensemble fini $\Theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$. A chaque $\phi \in \hat{\mathcal{S}}$, on peut associer un point $M_\phi \in \mathbb{R}^k$ où $M_\phi = \{R(\theta_i, \phi)\}_{1 \leq i \leq k}$

Soient $\hat{S} = \{M_\phi ; \phi \in \hat{\mathcal{S}}\}$ et $S = \{M_\phi ; \phi \in \mathcal{S}\}$.

L'ensemble $\hat{S}$ est un convexe de $\mathbb{R}^k$, car si ϕ et ϕ' sont dans $\hat{\mathcal{S}}$ et si $\alpha \in]0,1[$, $\alpha\phi + (1-\alpha)\phi' \in \hat{\mathcal{S}}$ et $M_{\alpha\phi+(1-\alpha)\phi'} = \alpha M_\phi + (1-\alpha) M_{\phi'} \in \hat{S}$.

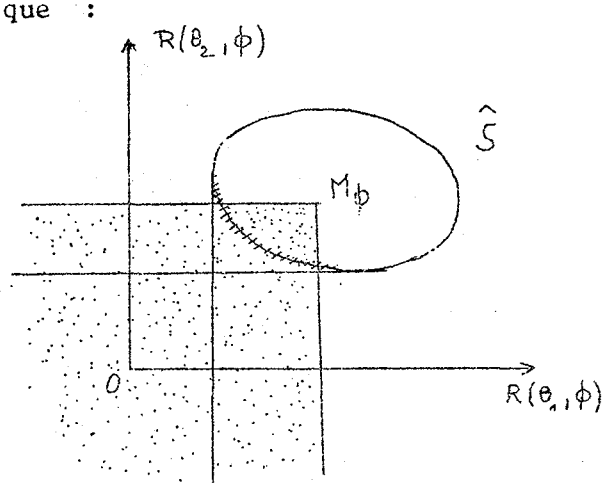
Supposons $\phi' \lesssim \phi$. Alors le point $M_{\phi'}$ est dans l'ensemble $Q[M_\phi]$, si l'on associe à tout point $M = \{x_1, x_2, \ldots, x_k\}$ de $\mathbb{R}^k$ l'ensemble $Q[M] = \{y_1, y_2, \ldots, y_k) ; y_1 \leq x_1, y_2 \leq x_2, \ldots, y_k \leq x_k\}$

Donc dire que ϕ est admissible c'est dire que :

$$Q[M_\phi] \cap \hat{S} = M_\phi$$

exemple $(k = 2)$

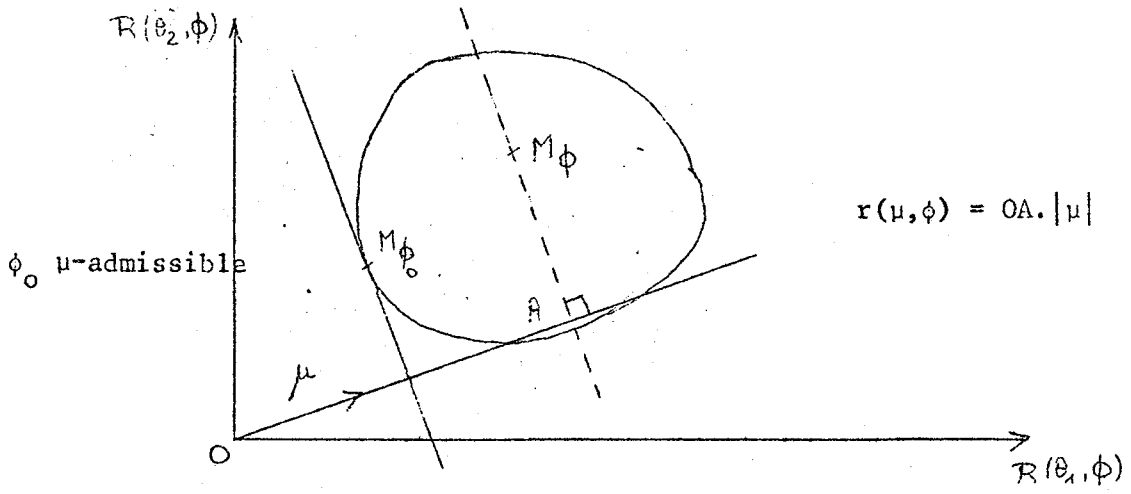
- la zone sombre est $Q(M_\phi)$
- si $\hat{S}$ est fermé, la zone hachurée est l'ensemble des M_ϕ pour ϕ admissible.

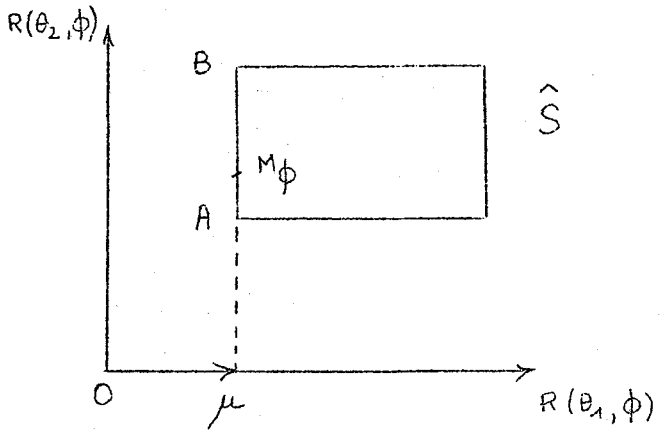


Soit μ une loi a priori sur Θ ; posons $\mu_i = \mu(\theta_i)$ $(1 \leqslant i \leqslant k)$; μ est le vecteur $(\mu_i)_{1 \leqslant i \leqslant k}$. Le risque bayésien de $\phi \in \hat{\mathcal{S}}$ est :

$$r(\mu,\phi) = \sum_{i=1}^{k} \mu_i \quad R(\theta_i,\phi) = < \mu, M_\phi >$$

Si une stratégie ϕ est μ - admissible, $\hat{S}$ est d'un seul côté de la droite passant par M_ϕ et orthogonale à μ ; on sait que si les μ_i sont tous non nuls une telle stratégie est admissible ; c'est faux dans le cas contraire comme le montre le 2ème schéma ci-dessous:



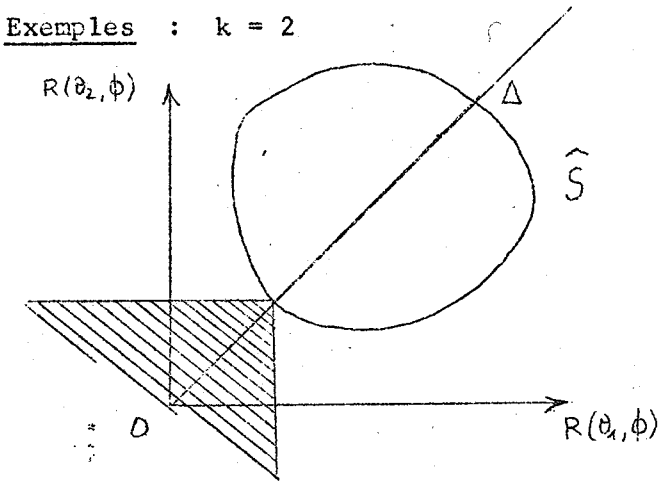


$\mu = (1,0)$

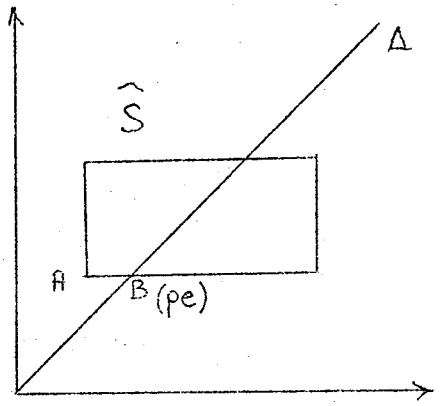
ϕ est μ - admissible, si $M_\phi \in AB$, mais n'est aussi admissible que si $M_\phi = A$.

Minimiser le risque maximun, c'est minimiser l'abscisse de M_ϕ , si M_ϕ est en dessous de la 1ère bisectrice Δ et l'ordonnée de M_ϕ , si M_ϕ est en dessus. Par conséquent, le minimum du risque maximun est le nombre $\rho = \{\sup\{\alpha ; Q[\alpha e] \cap \hat{S}\} = \emptyset\}$, où e est le vecteur de $\mathbb{R}^k$, dont toutes les coordonnées valent 1 ; si $M_\phi \in Q[\rho e] \cap \hat{S}$, alors ϕ est une stratégie minimax. En particulier, si S est compact, il existe une stratégie minimax.

Exemples : k = 2



φ est minimax, de risque constant et admissible.



φ est minimax, si $M_\phi \in AB$ Mais φ n'est admissible que si $M_\phi = A$, et de risque constant que si $M_\phi = B$.

Pour permettre de résoudre des problèmes simples dans le cas fini nous devons préciser la structure de $\hat{S}$, lorsque A est fini.

Préliminaires : Soient S_1 et S_2 2 convexes disjoints de $\mathbb{R}^k$. Il existe une forme linéaire u sur $\mathbb{R}^k$ et un nombre $c \in \mathbb{R}$, tels que $u(x) \geq c$ pour $x \in S_1$ et $u(x) \leq c$ pour $x \in S_2$. Autrement dit, il existe un hyperplan qui sépare S_1 et S_2. Nous ne montrerons pas cette propriété, qui est une forme du théorème de Hahn Banach, qu'on peut montrer élémentairement sur $\mathbb{R}^k$. On en déduit le :

lemme : Soit S un convexe de $\mathbb{R}^k$; soit Z un vecteur aléatoire de dimension k, tel que $Z \in S$ presque sûrement et que E(Z) existe ; alors $E(Z) \in S$.

Démonstration - Le théorème est vrai si k = 1, car un convexe est ici un intervalle. Supposons le vrai pour k - 1. Si E(Z) n'est pas dans S, il existe une forme linéaire u telle que $u(E(Z)) \leq u(x)$ pour tout $x \in S$. Donc :

$$u[E(Z)] = E[u(Z)] \leq u(Z) \quad \text{p.s.}$$

cela implique que u(Z) vaut E[u(Z)] (p.s.).

Soit $S' = \{x ; u(x) = E[u(Z)]\} \cap S$; c'est un convexe d'un espace euclidien à k - 1 dimensions et $Z \in S'$.

D'après l'hypothèse de récurrence, $E(Z) \in S' \subset S$.

Proposition : Si A est fini, $\hat{S}$ est l'enveloppe convexe de S (le plus petit convexe contenant S).

Démonstration : $\hat{S}$ est convexe ; il suffit donc de montrer que $\hat{S}$ est contenu dans l'enveloppe convexe de S_o de S.

Soit $A = (a_1, a_2, \ldots, a_n)$. Soit υ une probabilité sur A. Munissons $[0,1]$ de la mesure de Lebesgue P (soit E l'espérance par rapport à P). Il existe une fonction Z_υ de $[0,1]$ dans A, de loi υ ; il suffit en effet de diviser $[0,1]$ en m intervalles disjoints de longueurs respective $\upsilon(a_i)$ $(1 \leq i \leq k)$ et de poser $Z_\upsilon = a_i$ sur l'intervalle de longueur $\upsilon(a_i)$. Soit alors une stratégie $\phi \in \hat{\mathcal{S}}$; à chaque $x \in E$ correspond une mesure $\phi(x)$ sur A donc une fonction $Z_{\phi(x)}$ de $[0,1]$ dans A -

$$R(0,\phi) = \int d P_0(x) \int d \phi(x)\ (a)\ W(0,a) = \int d P_0(x)\ E\left[W(0,Z_{\phi(x)})\right]$$

$$= E\left[\int d P_0(x)\ W(0,Z_{\phi(x)})\right]$$

Pour chaque $\omega \in [0,1]$, $x \to Z_{\phi(x)}(\omega)$ est une règle de décision notée $Z_\phi(\omega)$.

$$R(0,\phi) = \int dP(\omega)\ R(0,\ Z_\phi(\omega))$$

Pour chaque $\omega \in [0,1]$, $R(\theta,Z_\phi(\omega))$ est dans S donc dans S_o.
Par suite $R(0,\phi) \in S_o$ et $\hat{S} \subset S_o$.

Conséquence : Si Θ et A sont finis et si S est compact, alors $\hat{S}$ est convexe compact ; c'est en effet l'image de $[0,1]^k \times S^k$ dans $\mathbb{R}^k$ par l'application continue

$$((\lambda_i)_{1\leq i\leq k},\ (x_i)_{1\leq i\leq k}) \longmapsto \sum_{i=1}^{k} \lambda_i x_i .$$

En particulier, si Ω est fini, S est fini et $\hat{S}$ est convexe compact.

Classes complètes de stratégies

Une classe complète de stratégies est une partie $C \subset \hat{\mathcal{S}}$ telle que si $\phi \in \hat{\mathcal{S}} \setminus C$, il existe une stratégie dans C meilleure que ϕ. Une classe essentiellement complète de stratégies est une partie $C \subset \hat{\mathcal{S}}$, telle que si $\phi \in \hat{\mathcal{S}} \setminus C$, il existe une stratégie dans $\hat{\mathcal{S}} \setminus C$ aussi bonne que ϕ. Une classe complète (resp. essentiellement complète) C est minimale, si il n'y a pas de sous ensemble strict de C qui soit une classe complète (resp. essentiellement complète).

Le théorème suivant permettra, dans les problèmes d'estimation, de ne considérer que les fonctions de décision (cf. Chapitre II).

Théorème

Supposons que A est un convexe de $\mathbb{R}^d$ et que pour chaque $\theta \in \Theta$, $a \mapsto W(\theta,a)$ est une fonction convexe continue. Si $\phi \in \hat{\mathcal{S}}$ et si pour chaque x, $\int |a|\, \phi(x, da)$ est fini, alors la fonction de décision $\psi(x) = \int a\, \phi(x, da)$ est aussi bonne que ϕ. Par suite si $\int |a|\, \phi(x, da)$ est finie pour toute $\phi \in \hat{\mathcal{S}}$, la classe $\mathcal{S}$ est essentiellement complète.

Remarque : Une fonction d'un convexe C dans $\mathbb{R}$ est convexe si :

$$\alpha \in]0, 1[,\ x \in C \text{ et } y \in C \implies f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

En particulier si $d = 1$, et si f est 2 fois dérivable, avec une dérivée seconde positive, f est convexe. Si d est quelconque, si $|\cdot|$ est la valeur absolue dans $\mathbb{R}^d$ et si $0 \in \mathbb{R}^d$, $x \mapsto |0-x|$ est convexe ; si ϕ est une fonction convexe de $\mathbb{R}_+$ dans $\mathbb{R}$, $\phi(|0-x|)$ est convexe ; en particulier $|0-x|^k$ est convexe pour $k \geq 1$.

Cas particulier : Soit α une fonction de Θ dans un ensemble convexe A de $\mathbb{R}^d$. On estime $\alpha(0)$ à l'aide de la fonction de perte quadratique $W(\theta,a) = |\alpha(0) - a|^2$. Alors pour $\phi \in \hat{\mathcal{S}}$, $\int |\alpha(0) - a|^2\, \phi(x, da)$ est fini et $\int |a|\, \phi(x, da) < \infty$.

Donc pour ces problèmes, on pourra se limiter aux fonctions de décision (l'introduction de stratégies aléatoires n'ajoute rien) : dans la théorie de l'estimation une fonction de décision est appelée estimateur.

La remarque précédente serait encore valable pour $W(\theta,a) = |\alpha(\theta) - a|^k$ avec $k \geq 1$. Par contre, dans les problèmes de test nous aurons besoin de stratégies aléatoires.
Pour démontrer le théorème nous avons besoin du lemme suivant :

Lemme : Inégalité de Jensen

Soit C un ensemble convexe non vide de $\mathbb{R}^d$. Soit f une fonction convexe continue de C dans $\mathbb{R}$. Soit Z un vecteur aléatoire de dimension d, sur un espace de probabilité $(\Omega, \mathcal{A}, P)$. On suppose que $Z \in C$, P p.s. et que Z est intégrable. Alors :

a) $E(Z) \in C$

b) $f(E(Z)) \leq E\, f(Z)$

Démonstration - Nous avons déjà vu que $E(Z) \in C$. Soit alors C_1 le convexe de $\mathbb{R}^{d+1}$ défini par :

$$C_1 = \{(x_1, \ldots, x_{d+1}) ; (x_1, \ldots, x_d) \in C, x_{d+1} > f(x_1, \ldots, x_d)\}$$

Le point $\{E(Z), f(E(Z))\}$ est un point de $\mathbb{R}^{d+1}$ qui n'est pas dans C_1. Il existe donc un vecteur $(\alpha_i)_{1\leq i\leq d+1}$ de $\mathbb{R}^{d+1}$ tel que pour tout $(x_1, \ldots, x_{d+1}) \in C_1$

$$\sum_{i=1}^{d+1} \alpha_i x_i \geq \sum_{i=1}^{d} \alpha_i E(Z_i) + \alpha_{d+1} f(E(Z))$$

(où $(Z_i)_{1\leq i\leq d}$ sont les composantes de Z).

On peut prendre x_{d+1} aussi grand que l'on veut, donc $\alpha_{d+1} \geq 0$.
Mais le point $\{(x_i)_{1\leq i\leq d} = x, f(x)\}$ appartient à l'adhérence de C_1 pour tout $x \in C$; donc pour $x = (x_i)_{1\leq i\leq d} \in C$

$$\alpha_{d+1} f(x) \geq \sum_{i=1}^{d} \alpha_i \left[E(Z_i) - x_i\right] + \alpha_{d+1} f(E(Z))$$

Par suite, p.s. :

$$\alpha_{d+1} f(Z) \geq \sum_{i=1}^{d} \alpha_i (E(Z_i) - Z_i) + \alpha_{d+1} f(E(Z))$$

* Supposons $\alpha_{d+1} > 0$. Alors en intégrant on obtient :

$$E(f(Z)) \geq f(E(Z))$$

* Si $\alpha_{d+1} = 0$, on obtient $\sum_{i=1}^{d} \alpha_i [E(Z_i) - Z_i] \leq 0$, mais l'espérance de cette v.a. étant nulle, on a $\sum_{i=1}^{d} \alpha_i (E(Z_i) - Z_i) = 0$. Si $d = 1$, cela veut dire que $Z = E(Z)$ p.s. et $f(Z) = f(E(Z)) = E(f(Z))$ p.s. Sinon, Z est un vecteur aléatoire à valeurs dans le convexe $S' = S \cap \{(x_i)_{1 \leq i \leq d} ; \sum_{i=1}^{d} \alpha_i (E(Z_i) - x_i) = 0\}$, qui est un convexe d'un espace euclidien de dimension d-1. Le théorème est donc prouvé par récurrence sur la dimension d.

Démonstration du théorème

Avec les hypothèses faites, on a pour tout $x \in E$:

$$W(\theta, \psi(x)) = W\left[\theta, \int a \, \phi(x, da)\right] \leq \int W(\theta, a) \, \phi(x, da)$$

On le voit en écrivant l'inégalité de Jensen avec la probabilité $\phi(x, \cdot)$. D'où $R(\theta, \psi) \leq R(\theta, \phi)$.

EXERCICES

① Dans le cadre de la théorie des jeux, soit $\Theta = A =]0 + \infty[$ et $W(\theta, a) = e^{-\theta a}$ quels que soient le paramètre θ élément de Θ, et l'action a choisie par le statisticien dans A, c'est à dire que le statisticien a intérêt à choisir la plus grande valeur possible pour a. Soit F la fonction de répartition d'une variable aléatoire Z à valeurs dans $]0 + \infty[$, telle que $EZ = +\infty$. F peut être considéré comme un élément de $\hat{\mathcal{S}}$. Montrer qu'aucun élément de $\mathcal{S}$, ensemble des stratégies pures, ne peut être aussi bon que la stratégie F.

4. DONNEES - ECHANTILLON - ECHANTILLONNAGE - LOIS CLASSIQUES - NOTATIONS CLASSIQUES.

En statistique descriptive nous avons vu quelques manières de décrire une population considérée dans son entier. Le modèle statistique est alors fini. En statistique mathématique ou inductive, on cherche à partir d'un certain nombre de données que l'on fabrique avec l'aide de l'expérimentateur ou soi-même, à obtenir des renseignements sur le modèle statistique.

Un échantillon est un sous-ensemble de l'ensemble population;

I - Echantillonage d'une population finie

a) Description

Considérons une population finie de N individus parmi lesquels on distingue r types distincts ; pour $1 \leqslant i \leqslant r$, soit N_i le nombre d'individus de type i. Pour déterminer les valeurs des N_i, on peut faire un recensement, c'est-à-dire observer toute la population. Mais on préfère souvent, faute de temps ou d'argent, effectuer un "sondage", c'est-à-dire prélever un échantillon de n individus de la population et déduire de cet échantillon des estimations sur les paramètres $N_1, N_2, \ldots, N_r$.

Les résultats d'un sondage seront donc une suite de n élements de $\{1, 2, \ldots, r\}$, autrement dit un point $(x_i)_{1\leqslant i\leqslant n} \in \{1,2,\ldots,r\}^n = E$. Le point $(x_i)_{1\leqslant i\leqslant n}$ représente le sondage, où le $i^{\text{ème}}$ individu examiné est de type r. A chaque sondage $x = (x_i)_{1\leqslant i\leqslant n}$, on peut associer la suite $\{n_1(x), n_2(x), \ldots, n_r(x)\}$ où pour $1 \leqslant j \leqslant r$, $n_j(x)$ est le nombre des individus examinés de type j :

$$n_j(x) = \sum_{i=1}^{n} 1_{\{j\}}(x_i)$$

On peut envisager 2 types de sondages :

b) Sondage avec remise

A chacun des n tirages, la probabilité de tirer un individu de type i est la fréquence de ce type $p_i = \frac{N_i}{N}$. Les n tirages sont indépendants entre eux. Donc E est muni d'une probabilité P définie par :

$$P(x) = P\left[(x_1, \ldots, x_n)\right] = p_1^{n_1(x)} \cdot p_2^{n_2(x)} \cdots p_r^{n_r(x)}$$

Cette probabilité dépend du paramètre $(p_1, p_2, \ldots, p_r)$.

On obtient le modèle statistique

$$\{E, \mathcal{P}(E), (P_{p_1,p_2,\ldots,p_r})_{(p_1,p_2,\ldots,p_r)\in\Pi}\}$$

où Π est le sous ensemble des points de $[0,1]^r$ dont la somme des coordonnées vaut 1.

Les n fonctions coordonnées $(X_i)_{1\leq i\leq n}$ sont des v.a. indépendantes de loi $\nu_{p_1,\ldots,p_r} = \sum_{i=1}^{r} p_i \delta_i$; X_i représente la valeur obtenue au $i^{ème}$ tirage. On a un n-échantillon de la famille de lois $\{\nu_{p_1,p_2,\ldots,p_r}\}_{(p_1,\ldots,p_r)\in\Pi}$.

Le vecteur $x \longmapsto \overline{n}(x) = \{n_i(x)\}_{1\leq i\leq r}$ est une statistique très importante. Si $\{i_1,\ldots, i_r\}$ est une suite de r nombres de $\{0, 1, \ldots, n\}$ dont la somme vaut n, les sondages x tels que $\overline{n}(x) = \{i_1, \ldots, i_r\}$ sont équiprobables ; il y a $\frac{n!}{\prod_{k=1}^{r} i_k!}$ suites de ce type.

On obtient donc :

$$P_{p_1, \ldots, p_r}(n_1 = i_1, \ldots, n_r = i_r) = \frac{n!}{\prod_{k=1}^{r} i_k!} \prod_{k=1}^{r} (p_k)^{i_k}$$

si $\sum_{k=1}^{r} i_k = n$; sinon la probabilité que $\overline{n} = (i_k)_{1\leq k\leq r}$ est nulle.

Supposons r = 2 ; alors on remplace le type "2" par la notation "0" Le modèle est alors :

$$\left\{\{0,1\}^n = \Omega\ ,\ \mathcal{P}(\Omega)\ ,\ (P_p)_{p\in[0,1]}\right\}.$$

Les v.a. $(X_i)_{1\leq i\leq n}$ sont indépendantes et leur loi est la loi de bernoulli $b(1,p) = (1-p)\,\delta_0 + p\,\delta_1$. La v.a. $n_1(x)$ a la loi binomiale $b(n,p) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \delta_k$.

c) Sondage sans remise

Contrairement au cas précédent un sondage sans remise revient à extraire de la population un bloc non ordonné de n individus.

Il y a $\binom{N}{n}$ tels blocs. La donnée du résultat d'un sondage est la donnée des nombres $(n_i)_{1\leq i\leq r}$ de tirages du sondage x où on obtient i. Il y a $\prod_{i=1}^{r}\binom{N_i}{n_i}$ tels blocs non ordonnés de n individus.

Le modèle statistique est alors $\{E, \mathcal{P}(E), (P_{N_1,\ldots,N_r})_{N_i,\ldots,N_r \in \Pi}\}$ où :

$$E = \{(n_i)_{1\leq i\leq r} \quad \text{où} \quad 0 \leq n_i \leq n\ ,\ \sum_{i=1}^{r} n_i = n\}$$

$$\Pi = \{N_1,\ldots,N_r\ ;\ 0 \leq N_i \leq N \quad \text{et} \quad \sum_{i=1}^{n} N_i = N\}$$

$$P_{N_1,\ldots,N_r}(n_1,n_2,\ldots,n_r) = \frac{\binom{N_1}{n_1}\binom{N_2}{n_2}\ldots\binom{N_r}{n_r}}{\binom{N}{n}}$$

Le loi $P_{N_1,\ldots,N_r}$ est une loi "hypergéométrique".

Remarque : Lorsque N est très grand, les premiers tirages ne modifient à peu près pas la composition de l'urne. Le tirage sans remise sera alors considéré comme un tirage avec remise (de loi plus simple).

II - Echantillonnage dans une population infinie. Forme classique du modèle statistique.

En général, dans l'étude d'un phénomène expérimental, on ne suppose rien sur la population du point de vue mathématique. Mais concrètement les données sont très souvent considérées comme une réalisation de n v.a. indépendantes et de même loi F_θ. Si ces variables sont réelles, on notera

$$(\mathbb{R}, \mathcal{B}, F_\theta)^{\otimes n}_{\theta\in\Theta} \qquad \text{le modèle}$$

le modèle statistique ainsi obtenu. L'échantillon est donné $S = (X_1,\ldots, X_n)$, et on dit aussi échantillon pour réalisation de l'échantillon $(X_1(\omega),\ldots, X_n(\omega))$. Les X_i sont donc les applications coordonnées de ce modèle.

On verra à la fin du chapitre II sur l'estimation d'autres types de fabrication d'échantillons dans des problèmes à objet

bien spécifique, et pour certains types de population. (Sondages stratifiés).

On note $\overline{X} = \dfrac{X_1 + \ldots + X_n}{n}$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Si les X_i sont réels, on note $X_{(1)} \ldots X_{(n)}$ les réarrangements par ordre croissant des X_i.

On notera $N(m, \sigma^2)$ la loi normale de moyenne m, de variance σ^2, $N_r(m, K)$ la loi normale de moyenne m dans $\mathbb{R}^r$, de covariance K. Si L est une loi $X \sim L$ signifie que la v.a. X suit la loi L.

Il sera fréquent de considérer des modèles images de $(\mathbb{R}, \mathcal{B}, F_\theta)^{\otimes n}_{\theta \in \Theta}$ par des statistiques à 1 ou de dimension, et de raisonner uniquement sur ces modèles images.

Fréquemment dans les modèles paramétriques ou non paramétriques on regardera une famille de loi du type

$$F\left(\frac{x - \theta}{\sigma}\right)$$

s'appelle alors paramètre de <u>position</u> (centrage, translation et paramètre d'<u>échelle</u>.

Ex. $\dfrac{1}{\sqrt{2\pi}} \exp \dfrac{1}{2} \dfrac{(x - \theta)^2}{\sigma} \dfrac{dx}{\sigma}$, $\dfrac{1}{\pi} \dfrac{1}{1 + \left(\frac{x - \theta}{\sigma}\right)^2} \dfrac{dx}{\sigma}$

CHAPITRE II

THEORIE ELEMENTAIRE DE L'ESTIMATION ET SONDAGES

Soit $(X_1 \dots X_n) = X$, un n-échantillon d'une v.a. de loi F. Sauf à la fin du chapitre, le modèle statistique est toujours $(\mathbb{R}, \mathcal{B}, F)^{\otimes n}_{F \in \mathcal{F}}$, où $\mathcal{F}$ est un ensemble de lois de probabilités sur $\mathbb{R}$ (nous verrons sur des exemples des v.a. à valeurs dans $\mathbb{R}^K$, $K \geq 1$).

Soit θ un paramètre réel associé à F, par exemple : moyenne, variance, médiane, fonctions de ces quantités. Notant Θ l'ensemble des valeurs de θ, nous remarquerons qu'en général, l'application $F \to \theta(F)$ n'est pas biunivoque. Faire une <u>estimation</u>, c'est au vu de l'échantillon X prendre une décision $\hat{\theta}(X)$ au sujet de la valeur de θ.

<u>Définition</u> 1. Un <u>estimateur</u> $\hat{\theta}$ de θ est une application $\hat{\theta} : X \to \Theta$; $\hat{\theta} \circ X(\omega)$ s'appelle l'<u>estimation</u> de θ.

La <u>fonction de perte</u> dans ce problème de décision, est la perte quadratique définie par

$$W(\theta, \hat{\theta}(X)) = |\hat{\theta}(X) - \theta|^2 ,$$

on suppose donc toujours $E_F|\hat{\theta}(X)|^2 < \infty$. L'estimation de θ par $\hat{\theta}$ comporte donc le risque :

$$R_{\hat{\theta}}(F) = E_F|\hat{\theta}(X) - \theta|^2 .$$

Le risque peut aussi être considéré comme une mesure de la <u>précision</u> de $\hat{\theta}$.

Lorsqu'il existe une application biunivoque $\mathcal{F} \longleftrightarrow \Theta$ (<u>modèles paramétriques</u>) et on a alors :

$$R_{\hat{\theta}}(\theta) = E_\theta|\hat{\theta}(X) - \theta|^2 .$$

La règle générale de comparaison des règles de décision donne ici : $\hat{\theta}_1$ est meilleur que θ_2 si

$$R_{\hat{\theta}_1}(F) \leq R_{\hat{\theta}_2}(F) \quad \text{pour tout} \quad F .$$

Exemple 1. Soit θ la moyenne de F,

$$\mathcal{F} = \{F,\ F \text{ probabilité sur } \mathbb{R}\}$$

$$= \pi(\mathbb{R})$$

$$\bar{X} = \frac{X_1+\ldots+X_n}{n} \quad \text{est un estimateur de } \theta ,$$

de risque $E_F|\bar{X} - \theta|^2 = \sigma^2(F)$.

Remarquons que l'on a $E_F \bar{X} = \theta(F)$. Cette propriété de centrage d'un estimateur nous amène à poser la définition suivante :

Définition 2. Un estimateur de θ est sans biais si pour tout $F \in \mathcal{F}$

$$\boxed{E_F \hat{\theta} = \theta(F).}$$

Exemple 2. Soit à estimer $\sigma^2(F)$.

$$S^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1} \quad \text{est un estimateur de } \sigma^2 . \text{ On a } E_F s^2 = \sigma^2(F).$$

s^2 est sans biais.

Exemple 3. Soit $X_1 \ldots X_n$ un n-échantillon de $N(0,\sigma^2)$, $\sigma^2 \in \mathbb{R}^+$.

Soit $T = \sum_{i=1}^{n} X_i^2$, $U = \frac{T}{n}$ est un estimateur sans biais de σ^2, $E_\sigma U = \sigma^2$.

On a $E_{\sigma^2} X_i^4 = 3\sigma^4$ (calculs aisés) et donc

$$E_{\sigma^2} T^2 = E_{\sigma^2}\Big(\sum_{i=1}^{n} X_i^4 + \Sigma\, X_i^2 X_j^2\Big)$$

$$= n(n+2)\sigma^4$$

et donc $$R_U(\sigma) = \frac{n+2}{n}\sigma^4 - \sigma^4 = \frac{2}{n}\ \sigma^4 .$$

Posons $U_c = cnU$.

Des calculs élémentaires donnent

$$E_{\sigma^2} U_c = cn\sigma^2$$

$$E_{\sigma^2}(U_c-\sigma^2)^2 = \sigma^4[n(n+2)c^2 - 2nc+1]$$

qui est minimum pour $c = \frac{1}{n+2}$ et vaut

$$\Big[1 - \Big(\frac{n}{n+2}\Big)\Big]\sigma^4 = \frac{2}{n+2}\ \sigma^4 .$$

On voit sur cet exemple qu'un estimateur sans biais peut être moins bon qu'un estimateur avec biais.

L'intérêt essentiel d'un estimateur sans biais est le suivant :

si l'on répète des estimations indépendantes du même paramètre, à l'aide du même estimateur sans biais : c.à.d. que si : $X^k = (X_1^k \dots X_n^k)$, $k \in K$ est une suite d'échantillons indépendants, de même loi, si $\hat{\theta}_k$ $(k \in K)$ est la suite des estimateurs sans biais de θ, associés aux X^k, indépendants et de même loi, alors :

$$\frac{\hat{\theta}_1+\hat{\theta}_2+\dots \hat{\theta}_k}{K} \xrightarrow{\text{p.s.}} E_F(\hat{\theta}_1) = \theta(F),$$

le risque associé à $\frac{\theta_1+\theta_2+\dots\theta_k}{K}$ valant alors $\frac{\text{var}_F(\hat{\theta}_1)}{K}$, ce qui est une bonne propriété que n'ont pas les estimateurs biaisés.

Dans le cas où n est petit ou moyen, la recherche d'estimateurs sans biais n'a pas toujours un sens. Un ensemble extrême est le suivant :

Exemple 4. (téléphoniste).

Un téléphoniste ne connait pas son standard. Il écoute une heure et enregistre x appels. Le nombre d'appels suit une loi de Poisson de paramètre inconnu θ. Pour aller voir son amie, il lui faut 2 heures. Quelle est la probabilité pour qu'il n'y ait pas d'appel pendant ces deux heures.

La probabilité pour qu'il n'y ait pas d'appel pendant ces deux heures, est égale à :

$$P_\theta(X_1 + X_2 = 0) = e^{-2\theta} .$$

On a donc à estimer $e^{-2\theta} = f(\theta)$.

Soit $\hat{f}$ un estimateur sans biais de $e^{-2\theta}$. On doit avoir

$$E_\theta(d) = e^{-2\theta} \text{ pour tout } \theta \geqslant 0 ,$$

or

$$E_\theta(d) = \sum_{n=0}^{\infty} d(n)e^{-\theta}\frac{\theta^n}{n!} = \sum_{n=0}^{\infty} e^{-\theta}(-1)^n\frac{\theta^n}{n!} \text{ pour tout } \theta \geqslant 0 ,$$

donc nécessairement $\hat{f}(n) = (-1)^n$, ce qui ne présente aucun intérêt !

Certains modes sont dit non paramétriques. Essentiellement ce sont ceux où $\mathcal{F}$ est très grand et en particulier ceux où il n'existe pas d'application biunivoque $\mathcal{F} \longleftrightarrow \theta$ (il n'y a pas d'accord général sur cette définition). C'est le cas des exemples 1, 2 et de l'exemple 5 ci-dessous par opposition au cas paramétrique (ex. 3, 6, 7).

Exemple 5. Soit $X_{(1)}, X_{(2)} \dots X_{(2n+1)}$ le réarrangement par ordre crois-

sant d'un $(2n+1)$ échantillon $X_1 \dots X_{2n+1}$. Supposons que $\mathcal{F} = \{ F$ probabilités diffuses sur $\mathbb{R}\}$. Soit μ la médiane de F, $(F(\mu) = 1/2)$. Alors $\hat{\mu} = X_{(n+1)}$ est un estimateur de μ (nous reviendrons sur ce cas au chapitre VI).

Exemple 6. Soit une loi de Bernoulli de paramètre p. $\bar{X}$ est un estimateur de p de risque $p-p^2$.

Exemple 7. Soit $N(m,\sigma^2)$ une loi normale $\theta = (m,\sigma^2)$, $\hat{\theta} = (\bar{X},s^2)$ est un estimateur de θ, le risque n'est pas défini car $\theta \in \mathbb{R}^2$. On peut prendre pour fonction de perte $\|\hat{\theta} - \theta\|^2$ pour une certaine norme euclidienne sur $\mathbb{R}^2$.

Supposons maintenant avoir un modèle paramétrique $(\Omega, \mathcal{A}, P_\theta)_{\theta\in\Theta}$, tel que l'on ait une idée a priori sur la valeur de θ traduite par une distribution de probabilité μ sur Θ, supposé doté d'une structure d'espace mesuré $(\Theta,\mathcal{C})$ μ est appelée distribution a priori du paramètre θ, et le cadre des problèmes que l'on va poser est dit bayésien. La probabilité μ sera souvent choisie pour de simples raisons de commodité mathématique, que nous exposerons plus loin, ou bien dans les meilleurs cas, en tenant compte d'expériences précédentes.

Un estimateur, noté ici d, de $f(\theta)$ sera dit estimateur de Bayes si le problème de théorie de la décision que l'on pose est le suivant, la perte est toujours $|d(X) - f(\theta)|^2$, si X est l'échantillon, mais le risque est le risque bayésien, μ-moyenne du risque ordinaire, soit

$$R_\mu(d) = \int_\Theta R_\theta(d)\ \mu(d\theta)$$
$$= \int_\Theta E_\theta |d-f(\theta)|^2\ \mu(d\theta).$$

Un estimateur d est dit μ-admissible, si $R_\mu(d) \leqslant R_\mu(d')$, pour tout autre estimateur d'.

Exemple 8.a. Soit $\Theta = \{\theta_1,\dots \theta_k\}$, $\mu = (p_1 \dots p_k)$, $p_i > 0$ pour tout $i = 1 \dots k$.

Alors si d est μ-admissible, d est admissible au sens ordinaire.

En effet $R_\mu(d) = \sum_{i=1}^{k} p_i \, E_\theta(d - f(\theta_i))^2$.

Soit d' un autre estimateur, par hypothèse $R_\mu(d) \leq R_\mu(d')$.

Donc il existe i_o tel que

$$E_{\theta_{i_o}} [d - f(\theta_{i_o})]^2 \leqslant E_{\theta_{i_o}} [d' - f(\theta_{i_o})]^2 .$$

Deux cas seulement sont alors possibles : les estimateurs d et d' ne sont pas comparables, ou alors d est meilleur que d' . d est donc admissible au sens ordinaire

Exemple 8.b. Soit $\theta = \mathbb{R}$. On suppose que $R_d(\theta) = E_\theta |d-\theta|^2$ est une fonction continue de θ (c'est un cas très fréquent), pour tout estimateur.

Soit d_o un estimateur μ-admissible. Alors d_o est admissible si μ a pour support $\mathbb{R}$ (le support de μ est le plus petit fermé F tel que $\mu(F^c) = 0$).

Supposons d_o non-admissible. Alors il existe θ_o et d tels que :

$$E_{\theta_o} |d-\theta_o|^2 < E_{\theta_o} |d_o-\theta_o|^2 \quad (1)$$

et

$$E_\theta |d-\theta|^2 \leqslant E_\theta |d_o-\theta|^2 .$$

Comme les fonctions $R_d(\theta)$ et $R_{d_o}(\theta_o)$ sont continues, l'inégalité (1) a lieu pour tous les θ d'un intervalle I qui est donc chargé par μ puisque μ est de support R .

$$\int E_\theta |d_o-\theta|^2 \, d\mu(\theta) > \int E_\theta |d-\theta|^2 \, d\mu(\theta) ,$$

d_o n'est donc pas μ-admissible, contraire à l'hypothèse.

On voit sur ces deux exemples que la méthode de Bayes est très intéressante ; bien entendu, pratiquement son caractère est très discuté, mais il faut noter qu'elle donne souvent, par des calculs simples, de très bons estimateurs.

L'interprétation suivante, quoique restrictive, a été souvent utilisée pour interprèter le point de vue bayesien, bien qu'elle repose sur des fondements expérimentaux difficiles, souvent, à justifier, comme nous l'avons vu.

On suppose $\Theta = \mathbb{R}$, $d\mu(\theta) = g(\theta)\, d\theta$, et que le modèle statistique est :

$$(\mathbb{R}, \mathcal{B}, f_\theta(x)dx)^{\otimes n}_{\theta\in\Theta} .$$

On considère θ comme une v.a. et plus précisément on considère que le couple (X,θ) admet sur $(\mathbb{R}^{n+1}, \mathcal{B}_{n+1})$ la densité $g(\theta)\, f_\theta(x)$.

$g(\theta)$ est la densité a priori de θ . C'est la densité marginale de θ , celle qu'on attribue à θ avant toute expérience.

La densité a postériori de θ est la densité de θ connaissant le résultat X de l'expérience, c.à.d. :

$$g_x(\theta) = \frac{f_\theta(x)g(\theta)}{\int_R f_\theta(x)g(\theta)d\theta} .$$

La recherche d'un estimateur bayésien de $k(\theta)$ a alors une interprétation simple : il s'agit de trouver une v.a. X-mesurable telle que :

$$\int_R E_\theta |d-k(\theta)|^2 \, g(\theta)d\theta \quad \text{soit minimale c.à.d.}$$

$$\int_{R^{n+1}} |d(x)-k(\theta)|^2 \, f_\theta(x)g(\theta)_{dx}d\theta$$

soit minimale, $k(\theta)$ étant donné.

Dans $L^2(R_n \times \bar{R}\,, \mathcal{B}_n \otimes \bar{\mathcal{B}}\,, f_\theta(x)\, g(\theta)d\theta \times dx)$, il s'agit de trouver une fonction X-mesurable, dont la distance à $k(\theta)$ soit minimale.

Nous savons, d'après le C 4 de probabilités, qu'il existe une solution et une seule à ce problème, $E^X[k(\theta)]$. L'estimateur de Bayes existe, est unique et égal à :

$$d(x) = \frac{\int k(\theta)f_\theta(x)g(\theta)d\theta}{\int f_\theta(x)g(\theta)d\theta} ,$$

d'après la formule donnant l'espérance conditionnelle dans le cas d'une densité.

Exemple 8.c. On cherche à estimer le paramètre d'une loi uniforme $\frac{1}{\theta}1_{[0,\theta]}(x)$, avec un 1-échantillon. On cherche un estimateur de Bayes associé à $g(\theta) = \theta\, e^{-\theta}$.

L'estimateur est donc, sur R^+ ,

$$d(x) = \frac{\int \theta\, e^{-\theta}\, 1_{[0,\theta]}(x)\,d\theta}{\int e^{-\theta}\, 1_{[0,\theta]}(x)\,d\theta} = x+1 = \frac{\int_x^\infty \theta\, e^{-\theta} d\theta}{\int_x^\infty e^{-\theta} dx} = \frac{x\, e^{-x} + e^{-x}}{e^{-x}} .$$

De plus $R_d(\theta) = \int_0^\theta d^2(x)\frac{dx}{\theta} - 2\int_0^\theta d(x) + \theta^2$ est une fonction continue de θ .

Exemple 8.d. Soit à estimer le paramètre θ d'une loi de Bernoulli à l'aide d'un n-échantillon. Soit S le nombre de face $(+1)$, $n-S$ le nombre de pile (0).

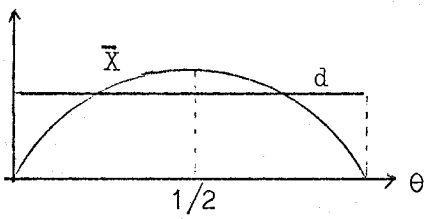
Considérons les estimateurs $\frac{S+a}{n+b}$, $E_\theta S = n\theta$, $\underset{\theta}{\text{var}}\, S = n\theta(1-\theta)$

$$E\left|\frac{S+a}{n+b} - \theta\right|^2 = \frac{1}{(n+b)^2} E(S - n\theta + a - b\theta)^2 = \frac{n\theta(1-\theta) + (a-b\theta)^2}{(n+b)^2}$$

Choisissons $a = \frac{\sqrt{n}}{2}$, $b = \sqrt{n}$ $\quad E\left(\frac{S+\frac{\sqrt{n}}{2}}{n+\sqrt{n}} - \theta\right)^2 = \frac{n\theta(1-\theta)+n(\frac{1}{2}-\theta)^2}{(n+\sqrt{n})^2} = \frac{1}{4(\sqrt{n}+1)^2}$.

L'estimateur $d = \frac{S+\sqrt{n}/2}{n+\sqrt{n}}$ a un risque constant et égal à $\frac{1}{4n(1+1/\sqrt{n})^2}$.

risque



L'estimateur $\bar{X} = \frac{S}{n}$ a un risque $\frac{\theta(1-\theta)}{n}$.

d est meilleur que $\bar{X}$ sur l'intervalle défini par $\theta(1-\theta) \leq \frac{1}{4(1+(1/\sqrt{n})^2}$

dont la taille est de l'ordre de $1/\sqrt{n}$. On peut voir que d est un estimateur de Bayes.

Considérons les densités définies pour $a, b > 0$ par

$$f_{x,y}(\theta) = \frac{1}{B(x,y)}\, \theta^{x-1}(1-\theta)^{y-1}$$

où $$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

$$= \frac{(x-1)!\,(y-1)!}{(x+y-1)!} \text{ , si } x, y \in \mathbb{N} .$$

On a alors, comme estimateur de Bayes pour la densité $f_{x,y}(\theta)$

$$d_{x,y} = \frac{\int_0^1 \theta^{S+x}\,(1-\theta)^{n-S+y-1} d\theta}{\int_0^1 \theta^{S+x-1}(1-\theta)^{n-S+y-1} d\theta}$$

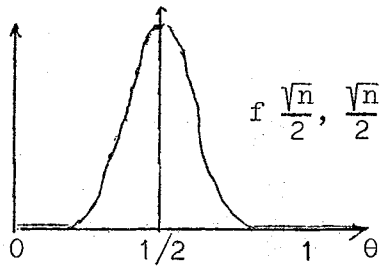
$$= \frac{\Gamma(S+x+1)\Gamma(n-S+y)}{\Gamma(n+x+y+1)} \quad \frac{\Gamma(n+x+y)}{\Gamma(S+x)\Gamma(n-S+y)}$$

$$= \frac{S+x}{n+x+y} \quad .$$

Pour $x = \frac{\sqrt{n}}{2}$, $y = \frac{\sqrt{n}}{2}$, il vient

$$d_{x,y} = \frac{S+\sqrt{n}/2}{S+\sqrt{n}}$$

$$= d$$



d est donc de Bayes.

Comme sa mesure charge tout $(0,1)$ d est admissible, et comme le risque de d est constant, tout autre estimateur a un risque supérieur à cette constante, au moins en un point. Donc d est <u>minimax</u>. Il est intuitif que le maximum du risque est atteint pour $1/2$ (maximum d'incertitude) et donc que, la loi $f\ \frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}$ qui "concentre" tout le risque de Bayes autour du point $1/2$ (car très pointue) est très bonne pour les valeurs voisines de $1/2$. Si l'on prend comme mesure de Bayes la mesure de Lebesgue $(x = y = 1)$, l'estimateur est $\frac{S+1}{n+2}$ dont le risque vaut $\frac{np(1-p)+(1-p)^2}{(n+2)^2}$ qui vaut $\frac{1}{4(n+2)}$ pour $p = \frac{1}{2}$.

Cet estimateur est également admissible et biaisé. On a

$$\frac{np(1-p)+(1-p)^2}{(n+2)^2} \leq \frac{p(1-p)}{n} \quad .$$

Si $$n(1-p) \leq (2n+4)p$$

soit $\frac{n}{3n+4} \leq p$, donc pour n grand, pour $p \geq \frac{1}{3}$. Cet estimateur est meilleur que $\bar{X}$, mais il est mauvais au voisinage de 0 (si $S = 0$, il donne $\frac{1}{n+2}$, si $S = n$ il donne $\frac{n+1}{n+2}$, alors que $\bar{X}$ donne les vraies valeurs. Nous n'avons pas montré que $\bar{X}$ était admissible !.

On a donc un moyen de fabrication d'estimateurs minimax : on cherche une mesure de Bayes telle que l'estimateur associé soit de risque constant.

Cet exemple est très intéressant en ce qu'il conduit à un résultat peu intuitif. (On pourrait croire que $\frac{S}{n}$ est le meilleur dans tous les cas possibles par raison de symétrie, notamment non minimax).

Le lecteur pourra cependant se persuader à l'aide de ces quelques exemples que le problème du choix entre plusieurs estimateurs admissibles n'est pas simple, pour n petit, même si le modèle est simple (lois de Bernoulli ou de Gauss), et que la pratique du statisticien est un élément important du choix.

Nous verrons plus tard une théorie plus élaborée de l'estimation en introduisant certaines classes d'estimateurs et aussi les problèmes relatifs aux échantillons de taille n grand et au comportement asymptotique.

E cet effet on se limitera souvent à rechercher un estimateur dans une classe particulière, par exemple

Définition 3. Un estimateur $\hat{\theta}$ sans biais de θ est dit de variance minimum si il est de risque minimum parmi les estimateurs sans biais ; autrement dit si $E_\theta \hat{\theta}_1 = \theta$ implique

$$E_\theta|\hat{\theta}_1-\theta|^2 \geq E_\theta|\hat{\theta}-\theta|^2 .$$

Il arrive fréquemment comme nous le verrons plus tard qu'il existe des estimateurs de variance minimum.

Remarquons enfin, que l'on peut avoir des stratégies d'estimation de type aléatoire, mais elles sont peu utilisées.

2. ESTIMATEURS CONSISTANTS. LA METHODE DU MAXIMUM DE VRAISEMBLANCE.

a) Le problème est d'étudier des estimateurs ayant un bon comportement asymptotique.

Soit donc une suite (T_n) d'estimateurs de $g(\theta)$ dans un modèle $(\Omega, \mathcal{A}, P_\theta)_{\theta\in\Theta}$.

En fait, on considère une suite de modèles $\otimes^n(\mathbb{R}, \mathcal{B}, F_\theta)_{\theta\in\Theta}$, associés à des échantillons de taille croissante et pour chaque n on a un estimateur de $g(\theta)$ mais pour pouvoir écrire la définition, il vaut mieux considérer tous les estimateurs définis sur le même espace.

Définition. Une suite T_n d'estimateurs est dite consistante si

$$\lim_{n\to\infty} T_n = g(\theta) \qquad P_\theta - \text{p.s} \qquad \text{pour tout } \theta \in \Theta .$$

Exemple 1. Soit θ la moyenne de F_θ. D'après la loi p.s. des grands nombres

$$\frac{X_1+\ldots+X_n}{n} = T_n \xrightarrow[P_\theta \text{ p.s.}]{} \theta$$

(T_n) est donc une suite consistante d'estimateurs.

De même $\dfrac{\sum_{i=1}^n (X_i-T_n)^2}{n} \xrightarrow[\text{p.s.}]{} \sigma^2(\theta)$ et donc $S_n = \dfrac{\Sigma(X_i-T_n)^2}{n}$ est une suite d'estimateurs consistants de la variance.

Mais cette définition mathématique maniée sans précaution, a peu de sens concret. Si on pose

$$T'_n = 0 \qquad \text{si} \quad n < 10^6$$
$$T'_n = T_n \qquad \text{si} \quad n \geq 10^6$$

alors T'_n est également une suite d'estimateurs consistants de la moyenne mais son impact concret est nul (car concrètement on a n grand mais $n << 10^6$).

Il faudrait donc si l'on veut donner un intérêt autre que mathématique, préciser la définition de manière à pouvoir choisir raisonnablement entre plusieurs suites d'estimateurs consistants. Nous ne le ferons pas, nous contentant d'étudier des méthodes dont on démontre par ailleurs la qualité.

La méthode du maximum de vraisemblance.

Soit un modèle régulier sur $(\mathbb{R}^n, \mathfrak{B}_n)$ défini par des probabilités du type $(p_\theta dF)_{\theta\in\Theta}$, où p_θ sont les densités de la loi F_θ par rapport à une mesure dF. On a donc

$$dP_\theta(X_1 \dots X_n) = p_\theta(x_1)\dots p_\theta(x_n)\, dF(X_1 \dots X_n).$$

Soit $x = (x_1 \dots x_n)$ la valeur observée, estimer θ par la méthode du maximum de vraisemblance, c'est calculer θ tel que $g_{n,\theta}(x) = p_\theta(x_1)\dots p_\theta(x_n)$ soit maximum. Donc $\hat\theta$ est l'estimateur défini (s'il existe) par :

$$\boxed{g_{n,\hat\theta_n(x)}(x) = \sup_{\theta\in\Theta} g_{n,\theta}(x)}$$

Dans la pratique, on utilise souvent cette méthode même pour des échantillons de taille $n \# 20$, et concrètement les résultats ne sont pas mauvais. Pour chaque n, on a un estimateur $\hat\theta_n$. On démontre que sous des hypothèses très larges

$$\hat\theta_n \xrightarrow[p.s.]{} \theta \quad \text{si } \Theta \text{ est une partie de } \mathbb{R}^d.$$

Remarquons que si Θ est un ouvert de $\mathbb{R}^d$ et si $g_{n,\theta}$ est dérivable on a

$$\frac{d}{d\theta} g_{n,\hat\theta(n)}(x) = 0$$

soit aussi

$$\frac{d}{d\theta} \log g_{n,\hat\theta(n)}(x) = 0$$

ou

$$\boxed{\sum_1^n \frac{d}{d\theta} \log p_{\hat\theta(n)}(x_j) = 0}$$

équation dite du maximum de vraisemblance, qui dans ce cas est une condition nécessaire pour $\hat\theta_n$.

Exemple 1. Famille exponentielle. On a

$$\log p_\theta(x) = \langle\theta, T(x)\rangle - \Phi(\theta) = \Sigma\, \theta_\ell T_\ell(x) - \Phi(\theta)$$

où $\theta \in \mathbb{R}^k$, $T : \mathbb{R} \to \mathbb{R}^k$, $T\circ X$ est donc un vecteur aléatoire à K dimension, et Φ une fonction normative, supposée convexe et définie dans un voisinage ouvert de 0, et

$$\exp \Phi(\theta) = \int \exp \langle\theta, T(x)\rangle\, p_\theta(x) dF(x)$$

(ces densités seront étudiées en détail plus tard).

et donc l'équation de vraisemblance s'écrit

$$\sum_1^n T_\ell(x_i) = n \frac{\partial \Phi}{\partial \theta_\ell} \qquad \ell = 1 \dots d ,$$

puisque Θ est un ouvert d'intérieur non vide de $\mathbb{R}^d$. Soit encore

$$\Phi'(\theta) = \frac{\sum_1^n T(x_i)}{n}$$

$$= S_n(x)$$

Comme Φ est convexe par rapport à chaque variable (cf. définition de Φ) il existe en général une solution $\hat{\theta}$ qui correspond effectivement à un maximum de $g_{n,\theta}(x)$.

Remarquons que comme

$$E_\theta\, T(x_i) = \Phi'(\theta)$$

on a $$\lim S_n(x) = \Phi'(\theta)$$

et donc, puisque Φ' est monotone on a $\hat{\theta}_n(x) \to \theta$, donc l'estimateur est consistant.

Donc, pour une famille exponentielle, s'il existe un estimateur du maximum de vraisemblance $\hat{\theta}_n$ pour θ , on a

$$\Phi'(\hat{\theta}_n) = \frac{\sum_1^n T(x_i)}{n}$$

et de plus cet estimateur est alors consistant.

<u>Exemple 1.a</u>. Soit une loi de Bernoulli B_θ . Par rapport à la mesure de masse 1 sur chaque point de $\{0,1\}^n$

$$g_{n,\theta}(x) = \theta^{\sum_1^n x_j} (1-\theta)^{n - \sum_1^n x_j} .$$

Soit $T = \Sigma\, X_i$

$$\frac{d}{d\theta}\, g_{n,\theta}(x) = \frac{t}{\theta} - \frac{n-t}{1-\theta}$$

l'équation du maximum de vraisemblance donne donc

$$\hat{\theta} = \frac{t}{n} ,$$

et on vérifie sans difficulté avec la dérivée seconde qu'il s'agit bien d'un maximum. On a donc ici $\hat{\theta} = \bar{x}$.

<u>Exemple 1.b</u>. Si $F_\theta = N(\mu, \sigma^2)$

$$\log g_{n,\theta}(x) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{j=1}^n (x_j - \mu)^2$$

L'équation du maximum de vraisemblance donne ($d = 2$, θ et à 2 dimensions)

$$\frac{\partial \log g_{n,\theta}(x)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{j=1}^{n} (x_j - \mu) = 0$$

$$\frac{\partial \log g_{n,\theta}(x)}{\partial \sigma} = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{\sigma^3} \sum_{j=1}^{n} (x_j - \mu)^2 = 0$$

d'où $\mu = \frac{\sum_{j=1}^{n} x_j}{n} = \bar{x}$

$$\sigma^2 = \frac{1}{n} \sum_{j=1}^{n} (x_i - \bar{x})^2 .$$

On vérifie, en utilisant l'inégalité $\log x \leq \frac{x^2-1}{2}$ pour $x > 0$ que $\hat{\theta} = (\bar{x} , \frac{1}{n} \sum_{1}^{n} (x_i - \bar{x})^2)$ donne bien un maximum de $\log g_{n,\theta}(x)$.

Exemple 1.c. Si X a une loi uniforme sur $[a,b]$, $\theta = \{a,b\}$

$$g_{n,\theta}(x) = \pi_{i=1}^{n} 1_{[a,b]}(x_i)$$

alors p_θ n'est pas dérivable mais il est facile d'étudier directement le maximum qui vaut $\frac{1}{(b-a)^n}$ si $a = \inf_j x_j$, $b = \sup_j x_j$.

Nous allons maintenant étudier un problème qui concerne un autre type d'échantillonnage.

3. Estimation de la moyenne dans une population non homogène (sondage). Echantillonnage en classes (stratifié).

La population est divisée en M classes sociales. Soit N la population totale, on connait N_j nombre d'individus de la classe j, j = 1 ... M. On veut estimer la moyenne d'un caractère défini sur la population ; par exemple le % du revenu dépensé par tête pour l'alimentation.

On peut imaginer que dans certaines classes la variance de cette variable est plus petite que dans d'autres. En particulier que certaines classes nombreuses sont particulièrement homogènes.

Dans ces conditions si on a une idée de ces variances, on a intérêt à faire un échantillonnage d'un type particulier.

La classe j contient N_j individus, $N_j/N = p_j$. Si on fait un échantillonnage de taille n_j dans la classe, la moyenne de l'échantillon $(X_j^1, \ldots, X_j^{n_j})$ est $\hat{\mu}_j = \frac{X_j^1 + \ldots + X_j^{n_j}}{n_j}$, c'est un estimateur de la moyenne m du caractère X (% du revenu...) pour la classe j, c'est-à-dire de la moyenne de la loi conditionnelle de $X/X \in j$, soit μ_j.

Supposons maintenant faire les 2 types d'échantillonnage suivants :

a) tirer n individus au sort dans la population entière (N supposé infini) soit $X_1 \ldots X_n$ les v.a. % du revenu ... des n individus tirés. On en déduit une estimation classique de m, $\bar{X} = \frac{X_1 + \ldots + X_n}{n}$.

b) tirons n_j individus dans la classe j (supposée de taille infinie), les n_j étant soumis à la contrainte

$$\sum_{j=1}^{n} n_j = n$$

A cet échantillon, on associe un nouvel estimateur de m,

$$\hat{\mu} = \sum_{i=1}^{M} p_j \hat{\mu}_j$$

$$\text{On a } E\hat{\mu} = \sum_{i=1}^{M} p_j E\hat{\mu}_j$$

$$= \sum_{i=1}^{M} p_j \mu_j$$

Or on a, pour une v.a. X_ℓ quelconque tirée de la population totale

$$EX_\ell = \sum_{j=1}^{n} E\ (X_\ell \cap X_\ell \in j)$$

$$= \sum_{j=1}^{n} E\ (X_\ell / X_\ell \in j)\ P\ (X_\ell \in j)$$

$$= \sum_{j=1}^{n} p_j\ \mu_j$$

Donc $\hat{\mu}$ estimateur de m, constitué de la moyenne pondérée des $\hat{\mu}_j$ est un estimateur sans biais.

c) Comparons maintenant les variances de ces estimateurs.

$\sigma^2\ (\bar{X}) = \frac{1}{n}\ \sigma^2$ si σ^2 est la variance de X_ℓ, élément quelconque de la population.

$$\sigma^2\ (\hat{\mu}) = \sum_{j=1}^{M} p_j^{\,2}\ \sigma^2\ (\hat{\mu}_j)$$

(variance de la somme de v.a. indépendantes)

$$= \sum_{j=1}^{M} p_j^{\,2}\ \frac{\sigma_j^{\,2}}{n_j}$$

si $\sigma_j^{\,2}$ est la variance d'un élément de la classe j.

Comparons ces variances. On a

$$E\ (X_\ell - m)^2 = \sum_{j=1}^{M} E\ [(X_\ell - m)^2 \cap (X_\ell \in j)]$$

$$= \sum_{j=1}^{M} E\ [(X_\ell - m)^2 / X_\ell \in j]\ p_j$$

$$= \sum_{j=1}^{M} E\ (X_\ell - \mu_j + \mu_j - m\ /\ X_\ell \in j)^2\ p_j$$

$$= \sum_{j=1}^{M} E\ [(X_\ell^j - \mu_j) + (\mu_j - m)]^2\ p_j$$

Si X_ℓ^j est élément quelconque de la classe j,
$E\ (X_\ell^j - \mu_j) = 0$,

$\mu_j - m$ est une constante, donc $E\ (X_\ell^j - \mu_j)\ (\mu_j - m) = 0$ et par suite on a

$$\sigma^2 = E\ (X_\ell - m)^2$$

$$= \sum_{j=1}^{M} p_j\ \sigma_j^{\,2} + p_j\ (\mu_j - m)^2$$

Interprétation :

Le premier terme correspond à la partie de la variance σ^2 due à la variabilité σ_j à l'intérieur du groupe j, le 2ème terme à la variabilité $(\mu_j - m)^2$ entre le groupe j et la population générale (carré de la "distance" du groupe j à la population générale). On a donc :

$$\sigma^2\ (\bar{X}) = \frac{1}{n}\ \left[\sum_{j=1}^{M} p_j\ \sigma_j^{\,2} + p_j\ (\mu_j - m)^2\right]$$

On vérifiera aisément, si M = 2 que

$$\frac{1}{n_1}\ p_1^{\,2}\ \sigma_1^{\,2} + \frac{1}{n_2}\ p_2^{\,2}\ \sigma_2^{\,2} \leqslant \frac{1}{n_1 + n_2}\ (p_1\ \sigma_1^{\,2} + p_2\ \sigma_2^{\,2})$$

implique que $\hat{\mu}$ est meilleur que $\bar{X}$. Montrons que l'on peut toujours choisir les n_j pour qu'il en soit ainsi et déterminons le choix optimal des n_j. Remarquons d'abord que si l'on prend l'estimateur "naturel" qui consiste à prendre les n_j proportionnels aux p_j soit $\frac{n_j}{n} \neq p_j$, $n_j = [np_j]$ éventuellement corrigé d'une unité pour que la somme fasse 1, ($[np_j]$ signifiant le plus grand entier $\leqslant np_j$) on a

$$\sigma^2\ (\hat{\mu}_1) \neq \sum_{j=1}^{M} p_j^{\,2}\ \frac{\sigma_j^{\,2}}{p_j}\ n$$

$$\neq \sum_{j=1}^{M} p_j\ \sigma_j^{\,2}$$

et donc

$$\sigma^2 (\overline{X}) \neq \sigma^2 (\hat{\mu}_1) + \sum_{j=1}^{M} p_j (m - \mu_j)^2$$

ce qui montre déjà que $\hat{\mu}_1$ est meilleur que $\overline{X}$.

Pour avoir le choix optimal des n_j il faut minimiser, par le meilleur choix possible des n_j ,

$$\sum_{j=1}^{M} p_j^2 \frac{\sigma_j^2}{n_j} \quad \text{sous la contrainte} \quad \sum_{1}^{M} n_j = n \ .$$

Nous laissons faire le calcul pour M = 2 par une méthode directe. Pour M quelconque employons la méthode de Lagrange.

Soit à minimiser

$$L (n_1 \ldots n_M) = \sum_{j=1}^{M} p_j^2 \frac{\sigma_j^2}{n_j} + \lambda^2 \left(\sum_{1}^{M} n_j - n \right)$$

Ecrivons $\left\{ \frac{\partial L}{\partial n_j} = 0 \right.$ $\qquad j = 1 \ldots M$

Soit $\left\{ \frac{p_j^2 \sigma_j^2}{n_j^2} = \lambda^2 \right.$ $\qquad j = 1 \ldots M$

ou $\left\{ \frac{p_j \sigma_j}{\lambda} = n_j \right.$ $\qquad j = 1 \ldots M$

et en additionnant,

$$n = \frac{1}{\lambda} \sum_{j=1}^{M} p_j \sigma_j$$

d'où le choix optimal des n_j

$$n_j = \frac{p_j \sigma_j}{\sum_{j=1}^{M} p_j \sigma_j} n \qquad j = 1 \ldots M$$

(on vérifie aisément que cet extremum de L est un minimum car lorsque $n_j \to 0$, $L \to \infty$).

Le choix optimal (en nombres entiers) est donc de prendre

$$n_j = \left[\frac{p_j \sigma_j n}{\sum_j p_j \sigma_j} \right]$$

c'est-à-dire de prendre les n_j proportionnels aux $p_j \sigma_j$, proportions de la classe j pondérée par sa variance. Par exemple, et c'est le cas pratique, on augmentera les proportions dans le sondage des individus d'une classe peu représentée dans la population mais à forte variance (i.e. classes aisées dans ce problème socio-économique) et on diminuera au X d'une classe nombreuse à revenu peu dispersé. On augmentera ainsi la précision du résultat obtenu sur la population entière.

CHAPITRE III

INTRODUCTION A LA THEORIE DES TESTS ET DES REGIONS DE CONFIANCE

1. Position du problème.

Soit $X = (X_1 \dots X_n)$ un n-échantillon d'un modèle statistique $(\Omega, \mathcal{A}, P_\theta)_{\theta\in\Theta}$.

Soit H_o une hypothèse faite sur le modèle et H_1 l'alternative de H_o. Soit $\Theta_o = \{\theta \in \Theta, \text{"}H_o \text{ est vraie dans } (\Omega, \mathcal{A}, P_\theta)\}$.

$\Theta_1 = \{\theta \in \Theta, \text{"}H_1 \text{ est vraie dans } (\Omega, \mathcal{A}, P_\theta)\}$. Donc $\Theta = \Theta_o \cup \Theta_1$. Faire un test de l'hypothèse H_o contre l'alternative H_1 c'est, ayant étudié une réalisation $X_1(\omega), \dots, X_n(\omega)$ de l'échantillon, énoncer la

conclusion "H_o est vraie" (c.à.d. $\theta \in \Theta_o$)

ou "H_1 est vraie" (c.à.d. $\theta \in \Theta_1$).

(Remarquons qu'en fait, dans beaucoup de problèmes concrets, on n'a pas $\Theta = \Theta_o \cup \Theta_1$, on teste H_o contre H_1, c'est-à-dire on regarde qui de H_o ou de H_1 a le "plus de chances" d'être vraie, mais $\Theta_o \cup \Theta_1$ est simplement une partie de Θ, H_o et H_1 étant deux hypothèses s'excluant mais H_1 n'est pas l'alternative à H_o et on étudie souvent plusieurs hypothèses H_1).

Exemple 1. Test sur le paramètre d'une loi de Bernoulli.

Soit $(\Omega, \mathcal{A}, P_\theta)_{\theta\in\Theta} = (\{0,1\}, \mathcal{P}\{0,1\}, B_\theta)^{\otimes n}_{\theta\in]0,1[}$ où B_θ est la loi de Bernoulli

$$P(X_i = 0) = 1 - \theta, \quad P(X_i = 1) = \theta .$$

a) On peut tester $\theta = \frac{1}{2}$ contre $\theta \neq \frac{1}{2}$

b) $\theta > \frac{1}{2}$ contre $\theta \leqslant \frac{1}{2}$

c) $\theta = \frac{1}{2}$ contre $\theta > \frac{1}{2}$ (voir la remarque).

Exemple 2. Test sur la variance d'une loi $N(0,\sigma^2)$.

$$(\Omega,\mathcal{A},P_\sigma)_{\sigma^2\in\mathbb{R}^+} = (\mathbb{R},\mathcal{B},N(0,\sigma^2))^{\otimes n}_{\sigma^2\in\mathbb{R}^+}$$

a) On peut tester $\theta = 1$ contre $\theta \neq 1$

b) $\theta \geqslant 2$ contre $\theta < 2$

c) $\theta = 1$ contre $\theta > 2$

d) $\theta \in]2,4[$ contre $\theta \geqslant 4$.

Exemple 3. Test (non paramétrique) sur la médiane.

$$(\Omega,\mathcal{A},P_\theta)_{\theta\in\Theta} = (\mathbb{R},\mathcal{B},F)^{\otimes n}_{\theta\in\pi_c(\mathbb{R})} ,$$

où $\pi_c(\mathbb{R})$ est l'ensemble des lois de probabilités continues sur $\mathbb{R}$. On peut tester

a) $\mu \geqq \frac{1}{2}$ contre $\mu < \frac{1}{2}$

b) $\mu = \frac{1}{2}$ contre $\mu \neq \frac{1}{2}$.

Exemple 4. Test (non paramétrique) d'homogénéité.

Soient $(X_1 \ldots X_n)$ et $(Y_1 \ldots Y_m)$ deux échantillons indépendants définis respectivement $(\mathbb{R},\mathcal{B},F)^{\otimes n}_{F\in\pi(\mathbb{R})}$ et $(\mathbb{R},\mathcal{B},F_1)^{\otimes m}_{F\in\pi(\mathbb{R})}$. Le modèle global est $(\mathbb{R}^{n+m},\mathcal{B}_{n+m},(\overset{n}{\underset{1}{\otimes}}F)\otimes(\overset{m}{\underset{1}{\otimes}}F_1))_{F\times F_1\in\pi^2(\mathbb{R})}$ F et F_1 lois de probabilités sur $\mathbb{R}$; donc $\theta = \{F,F_1\} \in \pi^2(\mathbb{R})$, ensemble des couples de lois de probabilité sur $\mathbb{R}$. Un test d'homogénéité est un test portant sur l'hypothèse : les deux échantillons proviennent de la même loi, soit $H_o : F = F_1$ contre $H_1 : F \neq F_1$ (donc θ_o est la diagonale de $\pi^2(\mathbb{R})$).

Exemple 5. Test non paramétrique d'indépendance.

Avec les mêmes notations que dans le cas précédent, on veut tester si les échantillons $(X_1 \ldots X_n)$ et $(Y_1 \ldots Y_n)$ proviennent de v.a. indépendantes, X_k et Y_k étant mesurées simultanément. Le modèle est donc $(\mathbb{R}^{2n},\mathcal{B}_{2n},\overset{n}{\underset{1}{\otimes}}F)$. On teste l'hypothèse

"$F = F_1 \otimes F_2$" contre "P n'est pas du type $F_1 \otimes F_2$",

où F_1 et $F_2 \in \pi(\mathbb{R})$.

Exemple 6. Test du caractère strictement aléatoire.

Soit $(X_1(\omega),\ldots,X_n(\omega))$ une réalisation de n v.a.r. dont on ne sait pas si elles sont indépendantes. Problème : répondre à la question :

il existe une loi de probabilité F telle que $(X_1 \dots X_n)$ puissent être considérées comme la réalisation d'un n-échantillon constitué de v.a. indépendantes) de là F .

Le modèle à prendre est donc :

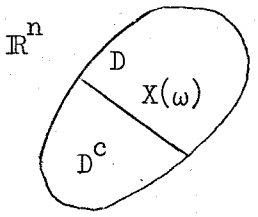
$(\mathbb{R}^n, \mathcal{B}_n, F)_{F \in \pi(\mathbb{R}^n)}$, où $\pi(\mathbb{R}^n)$ est l'ensemble des probabilités sur $\mathbb{R}^n$. On dispose donc d'un échantillon de taille 1 pour répondre à la question H_o : $F = \bigotimes_1^n G$, où G est une loi de probabilité sur $\mathbb{R}$. Dans certains problèmes on doit préciser H_o : $F = \bigotimes_1^n \lambda$ où λ est la mesure de Lebesgue sur $[0,1]$. Par exemple, si on a fabriqué à l'aide d'un ordinateur n nombres et que l'on veuille vérifier que ses nombres peuvent être considérés comme n nombres "réels" tirés au sort suivant la loi uniforme sur $[0,1]$ (cette spécification de F s'appelle test d'ajustement) et indépendemment.

On a l'habitude de classer les tests en 2 catégories. On parle de test <u>paramétrique</u> lorsque Θ est un ensemble de $\mathbb{R}$ ou de $\mathbb{R}^k$ les exemples 1 et 2 sont des tests paramétriques. On parle de test <u>non paramétrique</u> dans les autres cas. Bien entendu, cette division n'est pas claire mathématiquement puisque un test paramétrique apparaitrait comme un cas particulier de test non paramétrique et que souvent on peut, à l'aide d'une bijection "peu naturelle" identifier Θ à un sous-ensemble de $\mathbb{R}$ ou de $\mathbb{R}^n$.

Intuitivement les tests sont non-paramétriques si l'ensemble Θ est une partie de $\pi(\mathbb{R})$ ensemble des probabilités sur $\mathbb{R}$ (ou $\pi(\mathbb{R}^n)$) qui ne se représente pas "naturellement" comme un ensemble simple de $\mathbb{R}$ ou $\mathbb{R}^k$. C'est le cas des tests 3, 4, 5, 6.

2. <u>Forme de la règle de décision</u>.

Ayant observé $(X_1(\omega), \dots X_n(\omega))$ on doit répondre "H_o est vraie" ou "H_o est fausse" statistiquement, c'est-à-dire avec un risque d'erreur dans la réponse qui va être étudiée plus bas. Ceci étant, les règles de décision utilisées seront dans les bons cas du type suivant :



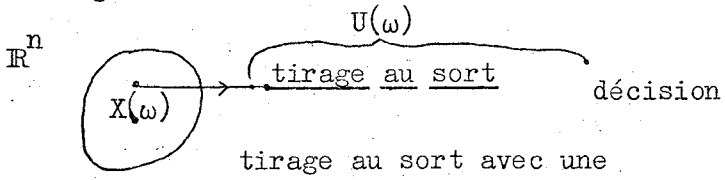
décision : "on rejette H_o"

"on accepte H_o".

Une règle décision est une partition (D,D^c) de $\mathbb{R}^n$. Si $X(\omega) \in D$ on rejette H_o, si $X(\omega) \in D^c$ on accepte H_o.

Cependant, on ne peut pas toujours se contenter de ce type de règle, car nous verrons que même dans certains cas simples, il existe un type de règles plus compliquées mais meilleures, comme nous l'avons vu au chapitre I.

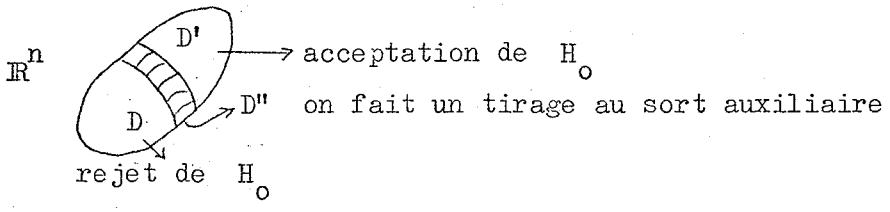
Le schéma général sera le suivant :



tirage au sort avec une probabilité $p(X(\omega))$ de rejeter H_o, une probabilité $[1-p(X(\omega))]$ d'accepter H_o.

A chaque valeur observée $X(\omega)$, on associe une probabilité p et on effectue un tirage au sort, (c'est-à-dire on réalise une v.a. de Bernoulli U de toute manière indépendante de X une fois choisie p) et suivant la valeur de U, on choisira H_o ou on rejettera H_o. p s'appelle la fonction critique du test, c'est donc une fonction $\mathbb{R}^n \to [0,1]$.

Concrètement la situation sera très souvent la suivante



D et D' sont en général des ouverts, D" leur frontière commune.

3. Pertes et risques dans un problème de test.

La fonction de perte que l'on prendra est la suivante : soit d la décision prise, d est à valeurs dans l'ensemble à 2 éléments (H_o,H_1), la perte vaut 1 si la décision prise est mauvaise, 0 si la décision prise est la bonne.

Si $(\Omega, \mathcal{A}, P_\theta)_{\theta\in\Theta}$ est le modèle, si $H_o = \{\theta \in \Theta_o\}$ et si le test est défini par une région de rejet D et une région d'acceptation D^c, on a une fonction de perte :

$W : \Omega \times \Theta \to \{0,1\}$ définie par

$w(\omega,\theta) = 1_D(X(\omega))$ si $\theta \in \Theta_o$

$w(\omega,\theta) = 1_{D^c}(X(\omega))$ si $\theta \in \Theta_1$.

Le risque est comme d'habitude l'espérance de la perte, soit :

$$R_D(\theta) = E_\theta w(\omega,\theta)$$
$$= P_\theta(D) \quad \text{si} \quad \theta \in \Theta_o$$
$$= P_\theta(D^c) \quad \text{si} \quad \theta \in \Theta_1 .$$

Remarque. Le choix de la fonction de perte très spéciale 1_D sur Θ_o, 1_{D^c} sur Θ_1, peut paraître arbitraire. On pourrait pénaliser les erreurs grossières, c'est-à-dire lorsque Θ_o est par exemple un bout de $\mathbb{R}^d$ dire qu'accepter à tort H_o est peu grave si il est voisin de Θ_o, et est très grave si θ est loin de Θ_o et donc réintroduire une fonction de perte de type quadratique.

En fait, de même que dans la théorie de l'estimation, le fait de travailler avec une fonction de perte quadratique simplifiait l'exposé de la théorie sans perdre grand-chose sur la puissance d'application, le fait de se limiter ici à cette fonction de perte simplifie l'exposition et il n'y a pas grande difficulté à changer de point de vue. De plus le paragraphe suivant justifiera beaucoup plus profondément le choix de cette fonction de perte, qui aura alors un sens expérimental très précis.

La relation $(D,D^c) > (D',D'^c)$ si et seulement si

$$R_D(\theta) \leq R_{D'}(\theta) \quad \text{pour tout} \quad \theta \in \Theta$$

définit une relation d'ordre partiel sur les tests. On a donc une notion de test admissible et une notion de test optimal (cf. Chap. I) et on pourrait dérouler la théorie des tests exactement comme celle de l'estimation. Mais nous allons voir par la suite que ce point de vue n'est pas en général, celui que l'on choisit.

Etudions maintenant perte et risque dans le cas général d'un test avec tirage au sort auxiliaire U

$$\begin{cases} w(\theta,\omega,\omega') = 1 & \text{si } U(\omega') = H_o \text{ et si } \theta \notin \Theta_o \\ w(\theta,\omega,\omega') = 1 & \text{si } U(\omega') = H_1 \text{ et si } \theta \in \Theta_o \\ w(\theta,\omega,\omega') = 0 & \text{dans les autres cas (c.à.d. quand il n'y a pas d'erreur de faite).} \end{cases}$$

On a donc

$$\begin{aligned} R_d(\theta) &= E\, w(\theta,\omega,\omega') \\ &= E\, E^X(\theta,\omega,\omega') \\ &= \int_\Omega p(X(\omega)\, dP_\theta(\omega)) \end{aligned}$$

si $\theta \in \Theta_o$, et

$$R_d(\theta) = \int [1-p(X(\omega))]\, dP_\theta(\omega) \quad \text{si } \theta \notin \Theta_o .$$

Exemple. Soit à tester $\theta = \frac{1}{4}$ contre $\theta = \frac{3}{4}$ pour une loi de Bernoulli à l'aide d'un $(2n+1)$ échantillon. Le modèle est donc $(\{0,1\}^{2n+1}, (\{0,1\}^{2n+1}, \overset{2n+1}{\otimes} B_\theta)_{\theta\in\{\frac{1}{4},\frac{3}{4}\}}$ où B_θ est la loi de Bernoulli de paramètre θ.

Pour des raisons de symétrie (justifiées au chap. VII), on peut se limiter aux régions D de $\{0,1\}^{2n+1}$ définies par $X_1(\omega)+\ldots+X_{2n+1}(\omega) \in E$ où E est une partie de $[0,2n+1] \subset \mathbb{N}$. La décision sera donc

$$\begin{cases} \frac{1}{4} & \text{si } S \in E^c, \quad S = \sum_{i=1}^{n} X_i \\ \frac{3}{4} & \text{si } S \in E . \end{cases}$$

On a donc

$$R(\tfrac{1}{4}) = \sum_{k\in E} \binom{2n+1}{k} (\tfrac{3}{4})^{2n+1-k} (\tfrac{1}{4})^k$$

$$R(\tfrac{3}{4}) = \sum_{k\in E^c} \binom{2n+1}{k} (\tfrac{1}{4})^{2n+1-k} (\tfrac{3}{4})^k .$$

Pour $k < n$ on a $(\frac{3}{4})^{2n+1-2k} > (\frac{1}{4})^{2n+1-2k}$ et donc pour minimiser le risque il faut prendre $[0,n-1] \in E$

$$[n+1,\ 2n+1] \in E^c .$$

Mais que se passe-t-il pour la valeur n ?

Si on choisit par exemple $\frac{3}{4}$, on augmente le risque de $\binom{2n+1}{n}(\frac{1}{4})^{n+1}(\frac{3}{4})^n$ dans le cas où $\theta = \frac{1}{4}$ et on ne l'augmente pas dans le cas où $\theta = \frac{3}{4}$. Par contre si l'on fait un tirage au sort auxiliaire, par exemple avec $p = \frac{1}{2}$,

on augmente les deux risques, mais on diminue le risque maximum, comme on le vérifiera par un calcul simple.

Les deux règles de décision ne sont donc pas comparables.

4. Désymétrisation du risque et formulation classique de la théorie des tests.

Dans les faits le problème se pose souvent de la manière suivante. Soit à tester l'innocuité d'une pilule. Il est beaucoup plus grave de rejeter à tort l'hypothèse que la pilule a des effets dangereux que d'accepter à tort cette hypothèse. De manière générale, ceci étant particulièrement vrai dans le domaine des sciences expérimentales. La démarche est la suivante. Soit une expérience associée à un modèle à tester contre l'alternative (ou une autre hypothèse exclusive de H_o) H_1 . Le risque de rejeter à tort H_o est considéré comme plus grave que celui d'accepter à tort H_o , la deuxième conclusion n'amenant que la poursuite du travail de recherche, alors que la première amène à l'application de ce travail (voir exemple ci-dessus). Dans beaucoup de problèmes, il est un type de risque maximal que l'on accepte de courir dans une certaine situation : un test sur la solidité d'un pont doit donner une certitude très forte, autrement dit on ne peut prendre de forts risques en rejetant l'hypothèse "le pont n'est pas solide".

Bien entendu, la notion de risque dépend beaucoup du domaine considéré. La quantification d'évènement négligeable est très variable. Par exemple chacun sait, dans notre monde, que la probabilité qu'une chaise soit dans un état non solide est de l'ordre de 10^{-34} et ceci n'empêche pas les physiciens de s'asseoir.

Le risque admissible à courir pour une hypothèse donnée n'est pas le même pour la ruine de l'assureur, le joueur en bourse, le joueur de belote ou le biologiste qui cherche à prouver une certaine loi. Ce problème de quantification du risque acceptable relève donc du domaine concret, et de ce fait ce choix est quelquefois assez subjectif.

Définition. Soit le test d'une hypothèse H_o contre une hypothèse H_1 effectué sur un modèle statistique $(\mathbb{R}^n, \mathfrak{B}_n, (P_\theta)_{\theta\in\Theta})$, H_o étant associée à Θ_o, H_1 à Θ_1. Le test est effectué à l'aide d'une région de rejet D et d'une région d'acceptation D^c.

On appelle risque de $1^{ère}$ espèce $P_\theta(D)$ pour $\theta \in \Theta_o$: c'est la probabilité de rejeter à tort l'hypothèse H_o. On appelle risque de 2^e espèce $P_\theta(D^c)$ pour $\theta \in \Theta_1$. C'est la probabilité d'accepter à tort H_o (nous verrons que l'on utilise plutôt $P_\theta(D) = 1 - P_\theta(D^c)$ que le risque de 2^e espèce.

Le niveau d'un test est $\alpha = \sup_{\theta\in\Theta_o} P_\theta(D)$, c'est le maximum du risque de $1^{ère}$ espèce

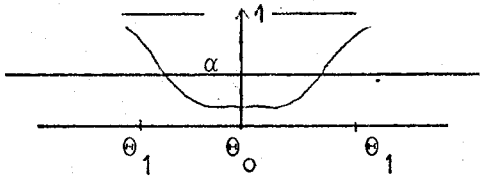
α est le risque maximum que l'on accepte de courir dans le problème de test ; on dit aussi que le test est significatif au niveau α.

La puissance du test est la fonction

$$\theta \to [0,1]$$

$$\theta \to P_\theta(D)$$

(fin de la définition).



test de niveau α avec fonction puissance.

On introduit alors une nouvelle relation d'ordre sur les tests.

Cette relation de pré-ordre ne compare que des tests de niveau inférieur ou égal à un niveau donné.

Définition. Soient D et D' deux tests (de région de rejet D et D' respectivement) de la même hypothèse H_o, contre la même alternative et de niveau $\leq \alpha$, α donné. D est dit plus puissant que D' si $P_\theta(D) \geqslant P_\theta(D')$ pour tout $\theta \in \Theta_1$. Il s'agit donc de la relation d'ordre partiel adapté au point de vue symétrisé.

(Etant sûr de courir un risque acceptable de $1^{ère}$ espèce, l'expérimentation peut comparer deux tests).

Dans une classe donnée de tests de niveau $\leq \alpha$, un test maximal pour cet ordre est dit test uniformément plus puissant (U.M.P. en abrégé d'anglais) plutôt qu'optimal. Nous noterons cette propriété par U.P.P..

Nous pouvons étendre facilement toutes ces notions au cas de test avec tirage au sort auxiliaire.

Soit X l'échantillon et $p(X)$ la probabilité de choisir D dans le tirage au sort auxiliaire après observation de X. Le niveau du test est :

$$\alpha = \sup_{\theta \in \Theta_o} \int_{\mathbb{R}^n} p(X(\omega))\ dP_\theta(\omega)$$

et la puissance la fonction

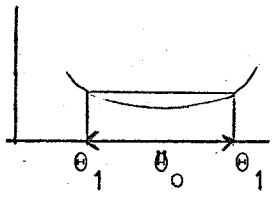
$$\theta \to \int_{\mathbb{R}^n} p(X(\omega))\ dP_\theta(\omega).$$

Mathématiquement un des intérêts des tests avec tirage au sort auxiliaire sera de donner de bons théorèmes d'existence (et de construire) des tests à niveau α donné (et non à niveau $\leq \alpha$).

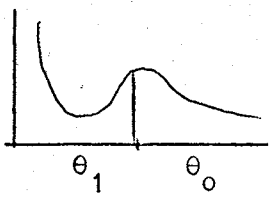
5. Test sans biais.

On dit qu'un test est sans biais si

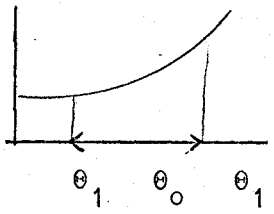
$$P_{\theta_o}(D) \leq P_{\theta_1}(D) \quad \text{pour tout} \quad \theta_o \in \Theta_o\ ,\ \theta_1 \in \Theta_1$$



test sans biais



test biaisé



test biaisé

Si le test est à tirage au sort auxiliaire de fonction critique p, le test est sans biais si

$$E_{\theta_o} p \leq E_{\theta_1} p \quad \text{pour tout} \quad \theta_o \in \Theta_o\ ,\ \theta_1 \in \Theta_1\ .$$

Il est pratiquement exclu de s'intéresser concrètement aux tests biaisés.

6. Test et intervalles de confiance. Dualité pour les tests $(\theta = a)$ contre $(\theta \neq a)$.

A) REGIONS DE CONFIANCE.

Il s'agit d'une troisième méthode de formulation des décisions en statistiques. Dans la théorie des tests la réponse est oui ou non. En estimation c'est $\theta = \theta_o$. En région de confiance c'est $\theta \in S$ où S est un

sous-ensemble de Θ .

Soit $\mathcal{X}$ les valeurs de l'échantillon (en général $\mathcal{X} = \mathbb{R}^n$). On dit qu'une famille $(S_x)_{x \in \mathcal{X}}$ de parties de Θ est une famille de <u>région de confiance</u> au niveau $1-\alpha$ pour la fonction $g(\theta)$ si l'on a

$$\inf_{\theta \in \Theta} P_\theta\{x,\ g(\theta) \in S_x\} \geq 1-\alpha$$

autrement dit : à chaque x , on associe S_x , tel que pour P_θ , c'est-à-dire si θ est la valeur "vraie", la probabilité que $g(\theta)$ soit dans la région S_x déterminée par la valeur x de l'échantillon soit au moins $1-\alpha$.

On peut alors définir un pré-ordre sur les régions de confiance. Une famille (S_x) est dite <u>plus courte ou meilleure</u> qu'une famille S'_x (<u>au niveau</u> $1-\alpha$) si on a :

pour tout θ_1 , θ_2 , $g(\theta_1) \neq g(\theta_2)$

$$P_{\theta_2}\{(x, g(\theta_1)) \in S(x)\} \leq P_{\theta_2}\{x,\ g(\theta_1) \in S'_x\}.$$

C'est une définition très naturelle qui signifie que S_x sépare mieux les valeurs distinctes $g(\theta_1)$ et $g(\theta_2)$ que S'_x . Cette définition dissymétrique est évidemment choisie en rapport avec la théorie des tests.

Enfin par analogie avec la théorie des tests, on dit que des régions de confiance sont <u>sans biais</u> si et seulement si pour tout θ_1, θ_2 tels que $g(\theta_1) \neq g(\theta_2)$ on a :

$$P_{\theta_2}\{x,\ g(\theta_1) \in S_x\} \leq 1-\alpha \ .$$

B) <u>FAMILLES DE TESTS ET FAMILLES DE REGIONS DE CONFIANCE</u>.

Considérons une famille de tests T_a de $(\theta = a)$ contre $(\theta \neq a)$ (famille de tests <u>bilatères</u>) et soit D_a leur région de rejet. Posons $S_x = \{a,\ x \in D_a^c\}$, (on suppose l'application $(a,x) \to 1_{D_a}(x)$ mesurable). Inversement soit S_x une famille de régions de confiance. Définissons une famille de tests D_a par $D_a = \{x,\ a \in S_x^c\}$. Les deux définitions sont duales : <u>l'espace Θ du paramètre et l'espace de l'échantillon jouent un rôle symétrique</u>, puisque

$$S_x = \{a,\ x \in D_a^c\} \iff D_a = \{x,\ a \in S_x^c\}.$$

Donc il y a correspondance biunivoque entre famille de tests bilatères et familles de régions de confiance.

Si la famille de tests est au niveau α, on a $P_a(D_a) = \alpha$. Le niveau de la famille de régions de confiance S_x est défini par

$$\inf_\theta P_a\{x,\ a \in S_x\} = \inf_\theta P_a(D_a^c) = 1-\alpha .$$

Montrons enfin que les relations de pré-ordre sur les familles de régions de confiance et de tests sont les mêmes.

Dire que D_a est meilleure que D'_a pour tout a (D'_a étant une autre famille de tests au niveau α) c'est dire que $P_b(D_a) > P_b(D'_a)$ pour tout $b \neq a$ (D_a est plus puissant que D'_a), donc

$$P_b\{x,\ a \in S_x^c\} \geq P_b\{x,\ a \in S_x'^c\}$$

ou

$$P_b\{x,\ a \in S_x\} \leq P_b\{x,\ a \in S'_x\}$$

qui est la relation de pré-ordre sur les familles S_x et S'_x. La réciproque est immédiate.

Exemple. Considérons la famille de tests de $\{\theta = \theta_o\}$ contre $\{\theta \neq \theta_o\}$, où θ est la moyenne d'une loi normale de variance 1, tests au niveau α.

On verra qu'il existe alors un test sans biais de la forme

$$D_{\theta_o} = \{T < c_1 \cup T > c_2\},$$

où $T = \sum_1^n X_i$, $X_1 \ldots X_n$ étant l'échantillon, avec de plus $P_{\theta_o}(T < c_1 \cup T > c_2) = \alpha$. Cette équation ne suffisant pas à déterminer c_1 et c_2, choisissons les symétriques par rapport à $n\,\theta_o$, soit

$$c_1 = n\,\theta_o - \sqrt{n}\, b(\theta_o),\quad c_2 = n\,\theta_o + \sqrt{n}\, b(\theta_o).$$

La région de test s'écrit donc :

$$\{|\sum_{i=1}^n X_i - n\,\theta_o| > \sqrt{n}\, b(\theta_o)\} = D_{\theta_o},$$

où b est déterminé par

$$P_{\theta_o}\{|\sum_{i=1}^n X_i - n\,\theta_o| > \sqrt{n}\, b(\theta_o)\} = \alpha .$$

La loi de ΣX_i est $N(n\,\theta_o, n)$, d'où .

$$\frac{1}{\sqrt{n}\,\sqrt{2n}} \int_{n\theta_o - b\sqrt{n}}^{n\theta_o + b\sqrt{n}} \exp \frac{(x-n\,\theta_o)^2}{2\pi}\, dx = 1-\alpha$$

soit $\frac{1}{\sqrt{2\pi}} \int_{-b}^{b} e^{-\frac{x^2}{2}} dx = 1-\alpha$.

$b(\theta_o)$ est donc une constante qui se lit simplement sur la table de la loi normale. La région de test est un intervalle de l'espace des valeurs de l'échantillon.

Etudions maintenant la famille des régions de confiance associées à ces régions de tests.

On ne considèrera pas la valeur $(x_1 \dots x_n)$ de l'échantillon mais seulement la valeur $\sum_{i=1}^{n} x_i = t$ de la statistique $\sum_{1}^{n} X_i$.

On a $S_x = \{\theta,\ x \in D_\theta^c\}$

$S_t = \{\theta,\ t \in D_\theta^c\}$

(puisque D_θ ne "porte" que sur T).

Donc $S_t = \{\theta,\ |t-n\theta| \leqslant \sqrt{n}\, b\}$

$\{S_t\}_{t\in R}$ est donc une famille d'intervalles de confiance. Si α est donné, b est calculé par $\frac{2}{\sqrt{2\pi}} \int_0^b e^{-\frac{x^2}{2}} dx = 1-\alpha$, si t est la valeur observée, l'intervalle de confiance au niveau $1-\alpha$ est donc $[\frac{t}{n} - \frac{b}{\sqrt{n}}, \frac{t}{n} + \frac{b}{\sqrt{n}}]$. On voit que la taille de l'intervalle décroît à la vitesse $\frac{1}{\sqrt{n}}$.

CHAPITRE IV

LA STATISTIQUE NON PARAMETRIQUE

I. Introduction.

Les statistiques descriptives utilisent souvent le terme de paramètre dans un contexte où il n'y a pas de modèle. Les paramètres comme la médiane servent à décrire l'échantillon ou les données. Il en est de même en statistique mathématique (ou inductive). Très souvent, on n'a pas d'idées précises sur le modèle (on : que ce soit l'expérimentateur ou le statisticien). Fixer arbitrairement un modèle de manière à pouvoir employer une méthode mathématique (test ou estimation) bien connue, relève souvent plus d'une démarche idéologique ("couverture" rationnelle d'une théorie mal justifiée) que d'une démarche scientifique. On désigne par statistique non-paramétrique deux sortes de techniques a priori assez différentes.

D'abord toutes celles qui opèrent sur un modèle très général et qui en particulier dans le cas des tests n'ont pas d'équivalent dans le domaine classique (ex : test du caractère "aléatoire" de l'échantillon). D'autre part, l'utilisation de statistiques dont la loi est indépendante du modèle sous des hypothèses très larges (free-statistics en anglais, les traductions françaises ne sont pas très sûres : libres ou universelles), ce sont essentiellement des statistiques fondées sur la notion d'ordre, comme nous le verrons ci-dessous. Les hypothèses sur le modèle sont très générales (loi continue ou quelquefois symétrique).

Parmi les avantages de la statistique non paramétrique, citons son caractère général (indépendant du modèle), le fait que très souvent, elle procède plus du classement des données que de leur mesure (on dit en américain

qu'elle est bien adaptée aux données sales c.à.d. entachées d'erreurs de mesures), ces caractères sont dits caractères de robustesse.

Par contre, il n'est pas possible de calculer des puissances en un sens simple. Disons cependant que de manière très générale, surtout pour les petits échantillons, quand on emploie une méthode non paramétrique sur un échantillon de source paramétrique (simulation) ou que l'on calcule mathématiquement la puissance d'un test non-paramétrique dans un modèle classique (gaussien, exponentiel etc...), les méthodes non-paramétriques, si elles sont moins bonnes que les paramétriques, leur sont cependant comparables (puissance souvent supérieure à la moitié de la paramétrique). Concluons en disant que trop souvent ce sont la tradition de modélisation abusive, le goût d'utilisation de "belles techniques mathématiques", le caractère ennuyeux des théories non-paramétriques qui font utiliser à tort des théories paramétriques. De ce point de vue, nous mettons en garde le lecteur. La place réservée dans ce cours aux techniques non paramétriques est insuffisante par rapport à leur importance, surtout dans le type d'application ou la modélisation est très arbitraire (par exemple : les Sciences Humaines). Pour terminer, remarquons que nous avons classé à part le test du χ^2, bien qu'il soit non-paramétrique. Ceci essentiellement parce qu'il n'est pas fondé sur les statistiques d'ordre et que sa justification n'est qu'asymptotique.

2. Statistiques d'ordre.

Soit $X_1 \dots X_n$ un n-échantillon de v.a. de loi F sur $\mathbb{R}$. On notera encore par F la fonction de répartition des X et l'on supposera F continue. L'échantillon ordonné $X_{(1)} \dots X_{(n)}$ est l'ensemble des X réordonné par ordre croissant, c.à.d. $X_{(1)} = \inf_{1 \leqslant i \leqslant n} X_i$, $X_{(2)} = 2^e X_i$ par ordre croissant, $(X_{(n)} =$ plus grand X_i.

Une statistique T est dite universelle ou libre si sa loi est indépen-

dante de la loi F de l'échantillon. L'exemple fondamental de statistique universelle T à valeurs dans $\mathbb{R}^n$, est celui de la suite $U_1, \ldots, U_n$; où $U_1 = F(X_1), \ldots U_n = F(X_n)$. Les U_i sont alors des v.a. indépendantes, de loi uniforme sur $[0,1]$:

$$P(a < F(X_i) < b) = P(F^{-1}(a) < X_i < F^{-1}(b))$$
$$= b-a \quad .$$

Donnons quelques interprétations évidentes des $X_{(i)}$.

a) $X_{(n)}$ intervient dans les problèmes de sécurité.

b) $X_{(1)}$ intervient dans les problèmes de sécurité, ou de fiabilité

c) La médiane $(X(\frac{n+1}{2})$ si n est impair, n'importe quel nombre entre $X_{(\frac{n}{2})}$ et $X_{(\frac{n}{2})+1}$ si n est pair, est une mesure de position ou de tendance moyenne (tendance dite centrale ou principale).

d) $\frac{X_{(1)}+X_{(n)}}{2}$ est aussi une mesure de la tendance centrale.

e) La plage $X_{(n)}-X_{(1)}$ est une mesure de dispersion.

f) La transformation $(X_1 \ldots X_n) \to (X_{(1)} \ldots X_{(n)}) = Y$ n'est évidemment pas biunivoque. Elle l'est sur chaque $\mathbb{R}^n_\sigma \subset \mathbb{R}^n$ où $Y = \sigma X$, (c.à.d. $(X_{(1)} \ldots X_{(n)}) = (X_{\sigma(1)} \ldots X_{\sigma(n)})$), σ permutation <u>donnée</u>. Dans ce domaine le jacobien de $Y = \sigma X$ est 1. On en déduit aisément que la densité de $(X_{(1)} \ldots X_{(n)})$ est $n! \prod_{i=1}^{n} f_X(y_i)$; f densité de X (puisque il y a $n!$ permutations σ). $y_1 < y_2 \ldots < y_n$

Les lois des $X_{(r)}$ (marginales) sont immédiates à calculer, par un argument combinatoire, par exemple si F a une densité f :

$$f_r(y_r) = \frac{n!}{(r-1)!(n-r)!} F^{r-1}(y_r)[1-F(y_r)]^{n-r} f(y_r).$$

<u>Exercices</u>.

a) Montrer que la densité de $X_{(r)}, X_{(s)}$ est donnée par

$$f_{r,s}(x,y) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!}$$
$$\times [F(x)^{r-1}(F(y)-F(x))^{s-r-1}(1-F(y))^{n-s}]$$
$$\times f(x)\, f(y), \qquad x < y, \quad r < s .$$

b) Trouver la densité de la médiane μ :

si $n = 2m+1$ on a $f_{\mu}(x) = f_{m}(x)$

si $n = 2m$ on prendra $\mu = \frac{X_{(m)}+X_{(n+1)}}{2}$ et ensuite partir de $f_{m,m+1}(x,y)$ et faire le changement de variable $u = \frac{x+y}{2}$, $v = y$

il vient $f_{\mu}(u) = \frac{(2m)!2}{(m-1)!^2} \int_{u}^{\infty} F(2u-v)^{m-1}$
$\times (1-F(v))^{m-1} f(2u-v) f(v)\, dv$.

c) Calculer la densité de la plage (range) $X_{(n)}-X_{(1)}$.

d) Montrer que $E|U_r|^k = \frac{n!(r+k-1)!}{(n+k)!(r-1)!}$

$$\mathrm{cov}(U_r, U_s) = \frac{r(n-s+1)}{(n+1)^2(n+2)} \qquad r < s$$

corrélation $(U_{(1)}, U_{(n)}) = \frac{1}{n}$.

Citons un résultat de convergence en loi dû à Smirnov, trop compliqué pour être démontré ici mais qui donne une idée des approximations utilisables lorsque n est grand.

<u>Théorème</u>. Supposons que $n \to \infty$, $\frac{r(n)}{n} = p$ et soit π défini par $F(\pi) = p$. Alors

$$\sqrt{\frac{n}{p(1-p)}}\ f(\pi)\ (X_{(r)} - \pi) \xrightarrow[\text{loi}]{} N(0,1)$$, où $f(\pi)$ est une constante.

3. <u>Runs</u>.

Soit un échantillon tiré d'une population à 2 caractères A et B . On note par un mot la réalisation de cet échantillon de taille n, par exemple, $X_1(\omega) \ldots X_n(\omega) = A\ A\ B\ A\ B\ B\ B \ldots A\ B\ A$.

Un <u>run</u> est une suite de mêmes symboles

$$A\ A\ B\ B\ B\ A\ B\ A = (AA)(BBB)(A)(B)(A)$$

intervenant successivement dans le mot et de longueur maximale parmi les

suites ayant cette propriété (précédée et suivie soit d'un symbole différent, soit de rien). Nombres de runs et longueur des runs sont des statistiques exploitables, notamment pour tester le caractère aléatoire d'un échantillon donné, c'est-à-dire pour vérifier que l'échantillon peut être considéré comme tiré au sort dans une population infinie, chacun des individus représentant la réalisation d'une v.a. X_i dans une suite $X_1 \ldots X_n$ de v.a. indépendantes et équidistribuées. L'hypothèse alternative n'étant souvent pas détaillée, on peut perfectionner la technique des runs dans le cas des caractères quantitatifs en considérant les runs croissants et décroissants (cf. plus loin).

a) Nombre total de runs

n taille échantillon

n_1 nombre de A
n_2 nombre de B
} fixés

r_1 nombre de A-runs

r_2 nombre de B-runs

$R = r_1 + r_2$.

Problème : trouver la distribution de R lorsque l'hypothèse H_o : "on a tiré au hasard et indépendamment n individus dans une population infinie" est vérifiée.

Lemme 1. Le nombre de configurations distinctes de n objets indiscernables dans r boites discernables est $\binom{n-1}{r-1}$.

Lemme 2. Soient R_1, R_2 les v.a. nombres de A-runs et B-runs, la loi du couple R_1, R_2 est donnée par

$$P(R_1 = r_1 \cap R_2 = r_2) = \frac{C\binom{n_1-1}{r_1-1}\binom{n_2-1}{r_2-1}}{\binom{n_1+n_2}{n_1}}$$

lorsque $r_1 = 1, 2, \ldots, n_1$, $r_2 = 1, 2, \ldots, n_2$

$r_1 = r_2$, ou $r_1 = r_2 \pm 1$

$C = 2$ si $r_1 = r_2$, $C = 1$ si $r_1 = r_2 \pm 1$.

Lemme 3.

$$P(R_1 = r_1) = \frac{\binom{n_1-1}{r_1-1}\binom{n_2+1}{r_1}}{\binom{n}{n_1}}$$

$$P(R_1+R_2=r) = \frac{2\binom{n_1-1}{r/2-1}\binom{n_2-1}{r/2-1}}{\binom{n}{n_1}}, \quad r \text{ pair}$$

$$= \frac{\binom{n_1-1}{(r-1)/2}\binom{n_2-1}{(r-3)/2}+\binom{n_1-1}{r-3/2}\binom{n_2-1}{r-1/2}}{\binom{n}{n_1}} \quad \text{si } r \text{ impair}$$

et $2 \leqslant r \leqslant n$.

Ex : $n_1 = 5$, $n_2 = 4$

$P(R = 9) = 0.008$ $\qquad$ $P(R = 8) = 0.063$

$P(R = 2) = 0.0016$ $\qquad$ $P(R = 3) = 0.056$.

Si on rejette pour $R \in \{2,9\}$, on trouve comme niveau significatif 0.024

$R \in \{2,3,8,9\}$, $\alpha = 0.143$. (cf. Tables).

Etude asymptotique.

On a, si $\lambda = \frac{n_1}{n}$, et si $n \to \infty$ avec λ constant,

$$\lim E \frac{R}{n} = 2\lambda(1-\lambda)$$

$$\lim \text{Var} \frac{R}{\sqrt{n}} = 4\lambda^2(1-\lambda)^2$$

et un calcul pénible montre que

$$\frac{R - 2\lambda(1-\lambda)}{2\sqrt{n}\,\lambda(1-\lambda)} \to N(0,1).$$

Remarques. 1. On ne sait évidemment rien sur la puissance, l'alternative à H_o étant trop large !

2. Si on rejette H_o on a intérêt à regarder s'il n'y a pas une tendance unilatérale par exemple "les A plus grands que les B " et à faire alors un test unilatéral car cette tendance indique une tendance à un nombre anormalement petit de runs (les grandes valeurs étant sans signification).

3. On peut aussi remplacer ce test par un test fondé sur la distribution du run le plus long (un run trop long est suspect, un plus long

(un run trop long est suspect, un plus long run trop court peut indiquer une périodicité).

4. Dans le cas où les v.a. observées sont quantitatives c.à.d. réelles, on utilise souvent la méthode dit du run up and down. Le run démarre chaque fois que la v.a. observée est plus grande que la précédente (run up ou croissant) si la suite précédente était du type down (décroissante) et réciproquement

Ex. : 7 8 2 3 4 5 s'écrit

$+ - + + + \Longrightarrow$ 3 runs $(1,+)$ $(1,-)$ $(3,+)$.

Des calculs compliqués donnent la distribution du nombre de runs up and down sous H_o et le comportement asymptotique, $\dfrac{R-(2n-1)/3}{[(16n-29)/90]^{1/2}} \sim N(0,1)$.

Le test de H_o a pour région de rejet les <u>trop petites</u> valeurs de R.

4. <u>Tests d'ajustement</u>.

L'hypothèse H_o dans un tel test est la suivante, l'échantillon est la réalisation de n v.a. indépendantes équidistribuées, de loi F_o fixée. On ne discute ni de l'indépendance ni de l'équidistribution, c.à.d. que l'on admet que l'échantillon est du type "classique". On discute F_o

$$H_o : F = F_o$$

$$H_1 : F \neq F_o .$$

<u>Remarque importante</u> : très souvent le test d'ajustement consiste à prendre pour F_o une loi, par exemple, gaussienne dont on a <u>estimé</u> les paramètres m et σ^2. Si l'estimation s'est faite à partir de l'échantillon, les méthodes de tests d'ajustement à $F(m,\sigma^2)$ exposées ci-dessous ne sont <u>plus valables</u>. On doit utiliser des résultats plus élaborés.

<u>Remarque</u>. Comme nous l'avons déjà dit, le test du χ^2 est reporté à un chapitre ultérieur, c'est typiquement un test d'ajustement.

a) Distribution empirique.

On pose $\hat{F}_n(x) = \frac{k}{n}$ si $X_{(k)} \leqslant x < X_{(k+1)}$

$= 0$ si $x < X_{(1)}$

$= 1$ si $x > X_{(n)}$

(où $X_{(j)}$ est l'échantillon réordonné).

On a donc

$$P(\hat{F}_n(x) = \frac{k}{n}) = \binom{n}{j}(F(x))^k(1-F(x))^{n-k}$$

pour $k = 0 \ldots n$, si F est la fonction (inconnue) de répartition de X.

En particulier $E\,\hat{F}(x) = F(x)$

$$\text{Var}\,\hat{F}(x) = \frac{F(x)(1-F(x))}{n} .$$

La loi large des grands nombres montre que $\hat{F}_n(x) \to F(x)$ en probabilité.

Un résultat beaucoup plus fort est le suivant :

Théorème de Glivenko-Cantelli.

$\hat{F}_n(x) \to F(x)$ uniformément p.s. c'est-à-dire

$\forall\, \varepsilon > 0,\ P(\sup|\hat{F}_n(x) - F(x)| > \varepsilon) \to 0$.

Nous laisserons la démonstration de côté. Par contre une application immédiate du théorème de limite centrale donne

$$\sqrt{n}\ \frac{(\hat{F}_n(x)-F(x))}{\sqrt{F(x)(1-F(x))}} \to N(0,1).$$

b) Les tests de Kolmogorov-Smirnov.

On pose, pour F continue,

$$D_n = \sup_x |\hat{F}_n(x) - F(x)|$$

$$D_n^+ = \sup_x \hat{F}_n(x) - F(x)$$

$$D_n^- = \sup F(x) - F_n(x) ;$$

$$D_n = \max_{0 \leqslant i \leqslant n} \sup_{X(i) \leqslant x < X_{i+1}} |\frac{i}{n} - F(x)|$$

en posant $X_{(0)} = -\infty$, $X_{(n+1)} = \infty$

d'où $$D_n = \max_i |\frac{i}{n} - \inf_{X(i) \leqslant x < X_{i+1}} F(x)|$$

$$= \max_{1 \leqslant i \leqslant n} |\frac{i}{n} - F(X_{(i)})|$$

$$= \max_{1 \leqslant i \leqslant n} |\frac{i}{n} - U_{(i)}| .$$

Donc la loi de D_n est indépendante de F. De même pour D_n^+ et D_n^-. Le calcul de D_n est pénible. On a

$$P(D_n < \frac{1}{2n} + v) = \begin{cases} 0 \text{ si } v \leqslant 0 \\ \int_{1/2n-v}^{1/2n+v} \cdots \int_{2n-1/2n-v}^{2n-1/2n+v} f(u_1 \ldots u_n)\, du_n \ldots du \\ 1 \text{ si } v \geqslant \frac{2n-1}{2n} \end{cases}$$

où $f(u_1 \ldots u_n) = n!$ si $0 < u_1 < \ldots < u_n < 1$

$= 0$ sinon

est la densité de $U_{(1)} \ldots U_{(n)}$.

<u>Démonstration</u> : $0 < \frac{1}{2n} + v < 1$ ou $-\frac{1}{2n} < v < \frac{2n-1}{2n}$ car $0 \leqslant D_n \leqslant 1$.

$$P(D_n < \frac{1}{2n} + v) = P\{\sup_x |\hat{F}_n(x) - x| < \frac{1}{2n} + v\}$$

$$= P(\bigcap_{i=0}^{n} [|\frac{i}{n} - x| < \frac{1}{2n} + v,\ U_{(i)} < x < U_{i+1}]).$$

Soit $A_i = \{\frac{2i-1}{2n} - v \leqslant U_{(i)} < U_{(i+1)} < \frac{2i+1}{2n} + v\}$

$$P(D_n < \frac{1}{2n} + v) = P(\bigcap_{i=0}^{n} A_i)$$

$$= P(\bigcap_{i=0}^{n-1} A_i \cap A_{i+1})$$

$$= P[\bigcap_{i=0}^{n-1} (\frac{2i+1}{2n} - v \leqslant X_{i+1} < \frac{2i+1}{2n} + v)]$$

d'où le résultat.

Des tables donnent $P(D_n > D_{n,\alpha}) = \alpha$, on en déduit un test d'ajustement.

<u>Théorème limite de Kolmogorov et Smirnov</u>.

Si F est continue,

$$\lim_{n\to\infty} P(D_n < \frac{z}{\sqrt{n}}) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2 z^2}.$$

Cette fonction est bien tabulée. L'approximation est très bonne pour $n \geqslant 35$ (de l'ordre de 0,01).

<u>Théorème</u>. On a, pour F continue

$$P(D_n^+ < t) = \begin{cases} 0 & t \leqslant 0 \\ \int_{1-t}^{1} \int_{\frac{n-1}{n}-t}^{u_n} \cdots \int_{\frac{2}{n}-t}^{u_3} \int_{\frac{1}{n}-t}^{u_2} f(u)\, du_1 \ldots du_n & \text{pour } 0 < t < 1 \\ 1 & t \geqslant 1. \end{cases}$$

<u>Démonstration</u>. $P(D_n^+ < t) = P[\bigcap_{1 \leqslant i \leqslant n} (\frac{i}{n} - X_{(i)} < t)]$

$$= P[\bigcap_{i=1}^{n} (X_{(i)} > \frac{i}{n} - t]$$ C.Q.F.D.

Propriété limite :

$$\lim_{n\to\infty} P(D_n^+ < \frac{z}{\sqrt{n}}) = 1 - e^{-2z^2}$$

et $$4nD_n^{+2} \xrightarrow{\text{loi}} \chi^2(2) \quad \text{(cf. chapitre sur le } \chi^2\text{)}.$$

5. Statistiques fondées sur les rangs des statistiques d'ordre.

A) Définition des rangs.

Soit $\sigma(\omega)$ la permutation définie par

$$X_{(j)}(\omega) = X_{\sigma(\omega)(k)}(\omega)$$

pour tout j, $j = 1 \ldots n$.

j s'appelle le rang de X_k dans l'échantillon $X_1 \ldots X_n$. On notera donc $\sigma(k)$ ce rang, qui est une v.a. à valeurs dans $\{1 \ldots n\}$, on peut encore la définir par :

$$\sigma(k) = \sum_{i\neq k} 1_{(X_k > X_i)} + 1 .$$

La distribution de $\sigma(k)$ est uniforme sur $\{1 \ldots n\}$. On a :

$$P[(X_k < x) \cap \sigma(k) = j] = \frac{1}{n} P(X_k < x / \sigma(k)=j)$$

$$= \frac{1}{n} P(X_{(j)} < x).$$

Supposons que X ait une loi continue F . On a :

$$E\, X_k\, \sigma(X_k) = \sum_{j=1}^{n} \frac{j\, E\, X_{(j)}}{n}$$

$$= \sum_{j=1}^{n} \frac{jn!}{(j-1)!(n-j)!} \int_{\mathbb{R}} x\, F(x)^{j-1} (1-F(x))^{n-j} dF(x),$$

d'après la forme vue plus haut de la loi de $X_{(j)}$.

Des calculs élémentaires donnent :

$$\sum_{j=1}^{n} j\, E\, X_{(j)} = n(n-1)\, E\, X_j\, F(X_j) + n\, E\, X_j$$

$$E\, \sigma(k) = \frac{n+1}{2} \quad , \quad \text{var}\, \sigma(k) = \frac{n^2-1}{12}$$

et $$\rho(X_k, \sigma(k)) = \sqrt{\frac{12(N-1)}{N+1}}\ \frac{E\, X\, F(X) - 1/2\, E\, X}{\sqrt{\text{Var } X}} .$$

Remarques.

1. Cette corrélation est assez variable et montre que les rangs ne

représentent pas très fidèlement l'échantillon, mais son sens n'est pas très clair. Nous y reviendrons.

2. On dit qu'il y a des noeuds lorsque l'on observe plusieurs valeurs égales des variables. Ceci est fréquent lorsque l'on utilise des rangs pour des lois F discrètes, ce que l'on fait couramment. Pour traiter les noeuds on peut :

a) faire un tirage au sort pour déterminer le rang entre valeurs égales. Cette méthode a l'avantage d'étendre la validité de la théorie aux lois discrètes.

b) Affecter à toutes les variables concernées le rang moyen. Ceci fausse les lois des statistiques concernées (par exemple : cela diminue la variance de $\sigma(k)$) ; il existe d'autres procédures.

B) Tests de la médiane.

La médiane joue, en non paramétrique, le rôle que joue la moyenne pour certaines lois paramétriques comme la gaussienne (où la moyenne est la médiane) et d'autres.

Les tests non paramétriques que nous allons définir seront utilement comparés aux tests paramétriques. Rappelons que la médiane μ est définie comme l'un des nombres tels que

$$P(X \geqslant \mu) = P(X \leqslant \mu) = 1/2 \ .$$

Le problème est de tester H_o : "$\mu = \mu_o$" contre H_1 : "$\mu \neq \mu_o$" (test bilatère) ou contre H_1' : "$\mu > \mu_o$" test unilatère.

On supposera $P(X = \mu_o) = 0$.

a) Le test du signe.

Si H_o est vrai, la v.a. $S_n = \sum_{i=1}^{n} 1_{(X_i > \mu)}$ suit une loi binômiale $B(n, 1/2)$. Les grandes et les petites valeurs de S_n conduiront donc au rejet de l'hypothèse H_o (dans le cas bilatère). De manière précise si α est donné, on peut déterminer des entiers $k_1(\alpha)$ et $k_2(\alpha)$ tels que

$$k_1(\alpha) = \inf\{h,\ P(S_n \geqslant h) \leqslant \tfrac{\alpha}{2}\}$$

$$k_2(\alpha) = \sup\{h,\ P(S_n \leqslant h) \leqslant \tfrac{\alpha}{2}\}$$

et la région de rejet sera

$$(S_n \geqslant k_1(\alpha)) \cup (S_n \leq k_2(\alpha)).$$

De même dans le cas unilatère la région de rejet sera $(S_n \geqslant k(\alpha))$ avec

$$k(\alpha) = \inf\{h,\ \sum_{k=h}^{n} \frac{1}{2^n} \binom{n}{k} \leqslant \alpha\}.$$

On sait que pour n grand

$$\frac{S_n - n/2}{1/2\sqrt{n}} \sim N(0,1),$$

on peut donc approximer la loi de S_n par une loi normale.

Cependant, comme on prend des v.a. discrètes, on a intérêt pour tel unilatère par exemple, à choisir $k(\alpha)$ de la manière suivante

$$k(\alpha) = [1/2 + 1/2\sqrt{n}\ z_\alpha + 1/2\ n]$$

où z_α est défini par : $\frac{1}{\sqrt{2\pi}} \int_{z_\alpha}^{\infty} e^{-x^2/2} dx = \alpha$ (la correction 1/2 venant du caractère discret de la binômiale, on laisse au lecteur le soin de la justifier).

On peut calculer la puissance de ce test sur des modèles paramétriques. Si H_o est rejeté, et si on appelle p la probabilité que $X > \mu$, on a, par exemple pour le test unilatère :

$$\pi(p) = \sum_{k(\alpha)}^{n} \binom{n}{k} p^k (1-p)^{n-k}.$$

Exemple numérique : $\alpha = 0{,}05$. Alors il n'existe pas de test de niveau significatif α, la valeur la plus proche, par exemple pour $n = 16$ est $k = 12$, qui donne

$$\sum_{12}^{16} \binom{16}{k} \frac{1}{2^{16}} = 0{,}038.$$

Si on veut $\alpha = 0{,}05$, il faut, comme à l'habitude pour les lois discrètes, faire un tirage au sort auxiliaire. Si on prend $\alpha = 0{,}038$, la loi gaussienne $N(29{,}04\ ;\ 1)$ on peut tester $H_o : \mu_o = 28$ contre $H_1 : \mu_o > 28$. On a alors $p = 0{,}85$ (c.à.d. on est très loin de H_o).

La puissance du test binômial $\pi(0,85)$ est $0,92$. La puissance du test de la moyenne (cf. chapitre sur les gaussiennes) est

$$\pi'(29,04) = P(\bar{X} > 28,44)$$
$$= 0,99.$$

La différence est considérable mais on est dans un cas vraiment défavorable car l'écart à la moyenne est de l'ordre de σ.

b) Intervalle de confiance et test d'ordre sur la médiane.

On a, sous H_o,

$$P(X_{(k_1)} \leqslant \mu_o < X_{(k_2)}) = P(U_{(k_1)} < \tfrac{1}{2} < U_{(k_2)})$$
$$= \int_{u<\frac{1}{2}<v} g_{k_1,k_2,n}(u,v)\,du\,dv = c(n,k_1,k_2).$$

Si $g_{k_1,k_2,n}$ est la densité des statistiques d'ordre numéro k_1 et k_2 d'un échantillon de loi uniforme. Si α est donné, il existe des choix (non uniques) de k_1 et k_2 tels que

$$c(n,k_1,k_2) \geqslant 1-\alpha$$

et aussi voisin que possible de $1-\alpha$. On en déduit :

a) un intervalle de confiance au niveau α pour μ., l'intervalle

$$[X_{(k_1)}, X_{(k_2)}].$$

b) un test : H_o est rejeté si

$$\mu \notin [X_{(k_1)}, X_{(k_2)}].$$

Pour n grand, on utilise l'approximation normale, pour

$$S_n = \sum_{j=1}^{n} 1_{(-\infty,\mu_o)}(X_j)$$

qui est telle, sous H_o, que

$$\frac{S_n - n/2}{1/2\sqrt{n}} \sim N(0,1).$$

Ayant calculé z_α comme précédemment, on calcule les solutions entières de l'équation approchée,

$$\frac{2y-n}{\sqrt{n}} = \pm z_\alpha ,$$

soient y_1, y_2 puis m tel que $X_{(m)} \leqslant \mu_o \leqslant X_{(m+1)}$ et si $m \notin [y_1,y_2]$

on rejette H_o . Le lecteur détaillera aisément cette procédure.

c) Cas de 2 échantillons appariés.

Soient 2 échantillons appariés de v.a. X et Y, non nécessairement indépendants. $(X_1 \dots X_n)$ et $(Y_1 \dots Y_n)$, sous la forme $(X_1,Y_1)\dots(X_n,Y_n)$.

Soit $Z_i = X_i - Y_i$, la loi de Z_i étant supposée continue. On teste ici, la médiane des différences est 0 (et non la différence des médianes, si X et Y sont des v.a. indépendantes ou simplement symétriques, l'égalité des médianes équivaut à une médiane des différences, nulle). On est ramené au problème du test de signe déjà vu.

d) Test de Wilcoxon ou test signe et rang.

Soit μ_o la médiane des X_i . N'utiliser que les signes est appauvrir beaucoup l'information donnée par l'échantillon. Soit $D_i = X_i - \mu_o$. On ordonne les $|D_i|$ par ordre croissant et on note T^+ (resp. T^-) la somme des rangs des $D_i \geqslant 0$ (resp. des $D_i \leqslant 0$), $T^+ + T^- = \frac{n(n+1)}{2}$. On s'intéresse à la loi de T^+ (qui fait intervenir les valeurs des D_i , pas seulement leurs signes). On suppose, et il est important de le retenir, que la loi F des X est symétrique par rapport à la médiane μ_o , c'est-à-dire que : $F(\mu_o - x) = 1 - F(\mu_o + x)$, $x > 0$ et donc que $P(D_i > a) = P(D_i < a)$, la loi des D_i étant symétrique au sens ordinaire.

Comme nous l'avons déjà dit, la médiane, dans le cas symétrique doit être considérée comme le paramètre naturel de centrage et un test sur la valeur de la médiane est un test de centrage. Notant

$$Z_{(i)} = 1_{(D_{(i)} > 0)}$$

on a

$$T^+ = \sum_{i=1}^{n} i\, Z_{(i)}.$$

Il est clair que sous H_o , $P(Z_{(i)} = 1) = \frac{1}{2}$. On en déduit : $E_{H_o} T^+ = \frac{n(n+1)}{4}$ et $\mathrm{Var}_{H_o} T^+ = \frac{n(n+1)(2n+1)}{24}$.

Quand H_o est fausse, un calcul simple montre que

$$P(Z_{(i)} = 1) = n\binom{n-1}{i-1}\int_0^\infty [F(u+\mu) - F(-u-\mu)]^{i-1} \times [1-F(u+\mu) + F(-u-\mu)]^{n-i} dF(u+\mu).$$

On a alors $\text{Var } T^+ = 4 \sum_{i=1}^{n} \frac{i^2 p_i(1-p_i)}{n^2(n+1)^2}$ où $p_i = P(Z_{(i)} = 1)$, et comme $p_i(1-p_i) \leqslant \frac{1}{4}$ on a $\text{Var } \frac{T^+}{n(n+1)} \to 0$ dans tous les cas.

Si $\mu > \mu_o$, il est facile de voir que $p_i > \frac{1}{2}$ et si $\mu \neq \mu_o$ on a $p_i \neq \frac{1}{2}$ en général.

Le test unilatère sera donc défini par une région de rejet du type

$$\frac{2T^+}{N(N+1)} - \frac{1}{2} \geqslant k$$

et le test bilatère par

$$\left\{\frac{2T^+}{N(N+1)} - \frac{1}{2} > k_1\right\} \quad \left\{\frac{2T^+}{N(N+1)} - \frac{1}{2} < k_2\right\}.$$

Les valeurs de la loi de T^+ sont tabulées.

On vérifie que pour n grand (> 20)

$$\frac{4T^+ - n(n+1)}{\sqrt{\frac{2}{3}\, n(n+1)(2n+1)}} \sim N(0,1).$$

On peut évidemment employer cette technique pour le cas de 2 échantillons appariés et aussi pour tester la symétrie d'une loi (mais ce n'est pas très bon).

Conclusion : l'intérêt des tests de signe est qu'ils ne nécessitent pas des mesures très précises. Le test de signe est bien adapté à des v.a. qualitatives (0,1).

L'efficacité asymptotique relative de ces tests par rapport au test de Student pour une loi normale est de 0,64 pour le test de signe et de 0,96 pour le test signe et rang ce qui est bon.

C) Comparaison de 2 échantillons.

a) On ne suppose plus les 2 échantillons appariés (de manière naturéelle ou artificielle) et en particulier on ne suppose plus leurs tailles égales. De plus, on suppose toujours ici les 2 échantillons indépendants :

soit $X_1 \dots X_m$ et $Y_1 \dots Y_n$. On teste $H_o : F_X = F_Y$, c'est-à-dire les échantillons sont issus de populations identiques en loi (et que l'on peut donc considérer comme identiques tout court). Si les populations sont supposées normales, on sait bien tester "$m_X = m_Y$" ou "$\sigma_X = \sigma_Y$", assez mal "$m_X = m_Y$ et $\sigma_X = \sigma_Y$". Dans les problèmes non paramétriques, utilisables dès que l'on a quelques doutes, sur les modèles paramétriques, on fait l'hypothèse que les lois F sont continues (quoique on puisse appliquer, avec quelques précautions les résultats à des lois discrètes).

Les alternatives classiques à H_o sont :

$H_1 : F_Y \neq F_X$

$H_2 : F_Y(x) \geq F_X(x)$ pour tout x où Y est plus dispersé vers $+\infty$ que X.

Nous étudierons plus systématiquement comme dans le cas paramétrique, les hypothèses sur le centrage (en anglais : location) et sur l'échelle, à savoir :

H_{C_1} : "$F_Y(x) = F(x-\theta_o)$ pour un certain $\theta_o \neq 0$ "

H_{C_2} : "il existe θ, $F_Y(x) = F(x-\theta)$"

et de même

H_{E_1} : "$F_Y(x) = F_X(\sigma_o x)$" pour un certain $\sigma_o > 0$

H_{E_2} : "il existe σ tel que $F_Y(x) = F_X(\sigma x)$".

Une autre hypothèse, dite de Lehman qui peut intervenir dans des problèmes de fiabilité, est

H_L : "$F_Y(x) = F_K(x)^k$, k entier"

(Y est un maximum de k v.a. X).

Nous allons d'abord passer en revue l'application des tests déjà vus.

a) Test de runs (v.a. nulles)

On classe les X et les Y par ordre croissant pour obtenir un écrit symboliquement sous la forme

X X Y X X X Y Y X Y Y X ...

On compte le nombre (total) de runs R, s'il est trop petit, cela indique souvent que X a tendance à être plus grand que Y, ou réciproquement ; la région de rejet de H_o contre H_1 est donc $R < c_\alpha$, R étant tabulé. (cf. paragraphes précédents, échantillons de tailles m n).

b) Test de Kolmogorov et Smirnov.

Soient $\hat{F}_X$ et $\hat{F}_Y$ les distributions empiriques de X et Y. Si H_o est vraie,

$$D_{m,n} = \max|\hat{F}_X - \hat{F}_Y|$$

suit une loi indépendante de F, (immédiat en passant aux variables uniformes U_i et V_j associées aux X et aux Y).

Le calcul de la loi de $D_{m,n}$ est assez finaud et utilise profondément les propriétés des manches aléatoires arrêtées sur certaines barrières (cf. Cours de Probabilités). Un théorème de Smirnov donne la distribution asymptotique de

$$\lim_{\substack{m,n\to\infty \\ \frac{m}{n} = \text{cte}}} P\left[\sqrt{\frac{mn}{m+n}}\ D_{m,n} < z\right].$$

On trouve que c'est celle du $\sqrt{n}\, D_n$ déjà vu.

On peut aussi étudier $D^+_{m,n} = \max(\hat{F}_X - F_Y)$ et tester H_o contre H_2 Ce sont des tests très bons, applicables moyennant les précautions habituelles aux cas discrets.

c) Le test de la médiane.

Soit à tester H_o : "$F_X = F_Y$" contre une alternative H_1 à préciser, où F_X (resp. F_Y) est la loi de X (resp. Y). On a un m-échantillon $X_1 \dots X_m$ de X, un n-échantillon $Y_1 \dots Y_n$ de Y

$$N = n+m, \quad Z = (Z_1 \dots Z_{m+n}) = (X_1 \dots X_m, Y_1 \dots Y_n).$$

Soit $\hat{\mu}$ la médiane empirique de Z,

$$\hat{\mu} = Z_{(\frac{N+1}{2})} \quad \text{si } N \text{ est impair}$$

$$\hat{\mu} \in [Z_{(N/2)}, Z_{(\frac{N+2}{2})}], \quad N \text{ pair},$$

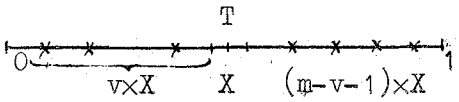
(pour se fixer les idées on peut prendre le milieu de l'intervalle).

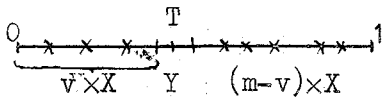
Soit $V = \sum_{i=1}^{n} 1_{(X_i<T)}$. Sous H_o, et si F est continue, posant $U = F(Z)$, on montre aisément que la loi de V est indépendante de F, c'est celle que l'on obtiendrait pour des variables de la loi uniforme sur $(0,1)$.

Les schémas ci-dessous décrivent la situation

$$P(V = v) = \left[m \binom{m-1}{v} \binom{n}{\frac{N-1}{2}-v} + n \binom{m}{v} \binom{n-1}{\frac{N-1}{2}-v}\right] \times \int_0^1 x^{\frac{N-1}{2}} (1-x)^{\frac{N-1}{2}} dx$$

$$= \frac{\binom{m}{v}\binom{n}{\frac{N-1}{2}-v}}{\binom{N}{\frac{N-1}{2}}}$$





Dans le cas où N est pair, on fait le même calcul en remplaçant dans les formules $\frac{N-1}{2}$ par $\frac{N}{2}$ (cf. interprétation plus haut).

Remarquons que sous H_o, $\frac{V}{m} \to \frac{1}{2}$ (m grand).

Le test contre l'alternative générale H_1 "$F_X = F_Y$" a donc pour région de rejet

$$\{V > c_1(\alpha) \cup V < c_2(\alpha)\}$$

c_1, c_2 convenablement choisis.

Le test devient très bon quand on prend pour alternative,

$$H_1' : F_Y(x) = F_X(x-\theta) \quad \text{pour un} \quad \theta \neq 0 .$$

On a alors :

$$H_1'' : \theta > 0 \implies \text{région de rejet} \quad V \geqslant c_\alpha$$

$$H_1'' : \theta < 0 \implies \text{région de rejet} \quad V \leqslant c_\alpha .$$

On a $\underset{H_o}{E}\, V = m \frac{\frac{N-1}{2}}{N}$ (cas impair)

et $\underset{H_o}{\mathrm{Var}}\, V = \frac{mn\,(N+1)}{N^2}$.

Si $N \to \infty$, $m \to \infty$ et $\frac{m}{N} = cte = \lambda$, la loi hypergéométrique de V , se comporte comme une binômiale (N,λ) et est donc a peu près normale une fois normée.

On utilise la statistique

$$V' = \frac{V - m/2}{(mn/N^3)^{1/2}} \sim N(0,1) .$$

Une remarque importante est la suivante :

$$V' \sim \frac{NV - m\frac{N-1}{2}}{\sqrt{mn(\frac{N-1}{2})^2/N}}$$

$$\sim \frac{V/m - W/n}{\sqrt{(\frac{V+W}{N})\ \frac{1-V+W}{N}\ \frac{N}{nm}}}$$

où

$$W = \sum_1^T 1_{(Y_i<T)} ,$$

soit

$$V' \sim \frac{V/m - W/n}{\sqrt{\hat{p}(1-\hat{p})(1/m+1/n)}}$$

où $\hat{p}$ est l'estimateur du nombre de v.a. $X < T$, donc $\hat{p} \neq 1/2$ sous H_o

Sous cette forme, on a une forte analogie avec le test comparant 2 proportions supposées normales et indépendantes. On peut aussi, évidemment obtenir par cette méthode un intervalle de confiance pour θ . On suppose que θ représente la médiane de Y (et donc que celle de X est nulle). On choisit $c_1(\alpha)$ et $c_2(\alpha)$ tels que

$$P(c_1(\alpha) < U < c_2(\alpha)) \geqslant 1-\alpha \quad (c_1,\ c_2 \text{ entiers}).$$

On calcule la médiane T de Z . On utilise la dualité tests-intervalles de confiance (cf. Chap. II).

Soit H_θ l'hypothèse "la loi de $Y-\theta$ est celle de X ".

H_θ est acceptée si et seulement si

$$X_{(c_1+1)} < Y_{(\frac{N-1}{2} - c_1)} + \theta$$

et

$$X_{(c_2)} > Y_{(\frac{N-1}{2} - c_2+1)} - \theta$$

l'intervalle de confiance est donc par dualité

$$\left[Y_{\left(\frac{N-1}{2} - c_2 + 1\right)} - X_{(c_2)},\ Y_{\left(\frac{N-1}{2} - c_1\right)} - X_{(c_1+1)} \right].$$

d) Le test U (Mann-Whitney).

Les notations sont les mêmes qu'en c). Soit

$$D_{ij} = 1_{(Y_j > X_i)}, \quad i = 1 \dots m \ , \ j = 1 \dots n$$

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ij} .$$

On pose $\pi = P(Y < X)$.

Si l'hypothèse H_o est vraie, on a $\pi = 1/2$. On a $\pi > 1/2$ si Y a "tendance" a être plus petite que X.

Il est donc clair que sous H_o la loi de U est indépendante de celle de X. Elle est assez difficile à calculer et est tabulée. L'alternative générale H_1 "$F_X \neq F_Y$" donne une région de rejet du type :

$$\left| U - \frac{mn}{2} \right| > C ,$$

fondée sur l'idée que $\underset{H_o}{E}\, U = \frac{mn}{2}$. On a de plus $\underset{\pi}{E}\, U = mn\,\pi$, ce qui explique la force du test et $\operatorname{var} \frac{U}{mn} \to 0$ pour toute hypothèse sur (X,Y).

Si on teste H_o contre H_2 $F_X(x) > F_Y(x)$ $\left(\pi > \frac{1}{2}\right)$ et la région de rejet est du type $U > \frac{mn}{2} + C$.

Sous H_o, $\dfrac{U - mn/2}{\sqrt{mn(N+1)/12}} \sim N(0,1)$

ce qui donne une bonne approximation (dès que $n, m \geqslant 6$).

Ce test présente l'avantage d'être utilisable pour des v.a. discrètes, en présence de noeuds. On pose alors

$$D_{ij} = \begin{cases} 1 & X_i > Y_j \\ 0 & X_i = Y_j \\ -1 & X_i < Y_j \end{cases}$$

$$\underset{H}{E}\, U = mn(\pi^+ - \pi^-) \quad \text{où} \quad \pi^+ = P(X > Y), \quad \pi^- = P(X < Y)$$

et

$$\underset{H_o}{E}\, U = 0 .$$

Des calculs simples permettent d'obtenir $\underset{H_o}{\operatorname{Var}}\, U$ en fonction du

nombre de noeuds.

U permet aussi d'obtenir un intervalle de confiance dans une estimation où F_Y est choisie du type $F_X(x-\theta)$. On calcule les différences $y_j - x_i = d_{ij}$. On cherche le nombre U de différences $< \theta$ ($H_{\theta=0}$ est rejeté (au niveau $\alpha/2$) si le nombre de telles différences est $< k(\alpha/2)$). On ordonne alors ces différences. En appliquant le principe de dualité test-intervalle de confiance, on montrera qu'un intervalle de confiance pour θ est tel que $\theta \geqslant (k+1)^{\text{ème}}$ (différence d_{ij}) (les d_{ij} ordonnées par ordre croissant).

Même raisonnement pour la borne supérieure de l'intervalle de confiance qui est donc

$$(d_{ij})_{(k(\frac{\alpha}{2}))}, (d_{ij})_{(N-k(\frac{\alpha}{2}))}$$

Dans ces conditions

$$(F_Y(x) = F_X(x-\theta)$$

le test U est très bon. Comparé au test de Student, sur une population quelconque, son efficacité asymptotique est $> 0,83$ et sur une population normale $> 0,96$.

e) Le test de Wilcoxon.

Il s'agit d'une présentation (légèrement) différente de d). Nous gardons les notations précédentes. Posons

$$R_i = 1 \text{ si } Z_{(i)} \text{ est un } X$$
$$= 0 \text{ si } Z_{(i)} \text{ est un } Y$$

analogue en statistiques non paramétriques de ce que nous appellerons contraste dans le cas des modèles gaussiens d'analyse de variance, est constitué des combinaisons linéaires $\sum_{i=1}^{N} a_i R_i = a(R)$. Si l'on pose en particulier $W = \sum_{i=1}^{N} i R_i$ on obtient le test de Wilcoxon.

Si U désigne la statistique introduite précédemment au d), nous laissons au lecteur le soin de vérifier que $W = U + \frac{m}{2}(m+1)$; en particu-

lier W est symétrique autour de $\underset{H_o}{E}\, W = \frac{m(N+1)}{2}$

$$\min W = \frac{m(m+1)}{2} \quad , \quad \max W = \frac{m(2N-m+1)}{2} \quad .$$

Les valeurs extrêmes de W constituent une région de rejet.

f) Tests sur l'échelle.

Supposons maintenant avoir à tester "H_o contre H_1" : il existe $\theta \neq 1$, $F_Y(x) = F_X(\theta x)$, où θ doit être interprété par exemple, comme $\frac{\sigma_Y}{\sigma_X}$ rapport des dispersions de Y et de X .

On interprète donc θ comme une variable d'échelle (cf. chapitre I). Le test gaussien analogue est le test F pour des moyennes inconnues mais égales. Le test de Mood consiste à choisir comme statistique $M = \sum_{i=1}^{N} (i - \frac{N+1}{2})^2 R_i$ qui pondère les grandes valeurs (absolues) par de fortes masses et caractérise donc la dispersion. Si M est trop grand, on rejettera H_o (X plus dispersé que Y), si M est trop petit on rejettera H_o (X moins dispersé que Y).

Le test S de Siegel-Tukey utilise

$$S = \sum_{i=1}^{N} a_i Z_i \quad \text{où} \quad a_i = 2i \ , \ 1 \leqslant i \leqslant N/2$$

$$= 2(N-i) + 2, \ \frac{N}{2} < i \leqslant N \quad \text{si } N \text{ est pair}$$

et $a_i = 2i-1, \ 1 \leqslant i \leqslant N/2 \ , \ N$ impair

$$= 2(N-i) + 1, \ N/2 < i < N \ .$$

Son avantage est d'avoir la même distribution que la statistique W du test de Wilcoxon.

CHAPITRE V

ESTIMATIONS ET TESTS EN VARIABLE GAUSSIENNE

I. Lois de certaines variables aléatoires définies à partir de variables gaussiennes.

a) χ^2 et χ'^2 .

Définition. Une v.a. Y est une variable $\chi'^2(n,\lambda)$ si

$$Y = \sum_{i=1}^{n} X_i^2$$

où les X_i sont des v.a. gaussiennes indépendantes de moyenne μ_i et de variance 1 :

$$X_i \sim N(\mu_i, 1)$$

avec $\Sigma \mu_i^2 = \lambda$. λ est appelé paramètre de non-centralité. On note

$$\chi^2(n) \sim \chi'^2(n,0).$$

Proposition. La loi de Y ne dépend en effet que de λ .

Soient $X_i = N_i + \mu_i$

$Z_i = T_i + \nu_i$ où N_i et T_i sont indépendantes de loi $N(0,1)$ et $\Sigma \mu_i^2 = \Sigma \nu_i^2$.

Soit G une transformation orthogonale de $\mathbb{R}^n$ telle que $G(\mu) = \nu$ (il en existe une puisque $\|\mu\| = \|\nu\|$) alors :

$$\|GX\|^2_{\mathbb{R}^n} = \|X\|^2_{\mathbb{R}^n} \quad \text{et} \quad \|G(N+\mu)\|^2 = \|G(N)\|^2 + 2 < GN, G\mu > + \|G(\mu)\|^2$$

si on pose : $T' = GN$; les T'_i sont des v.a. indépendantes de loi $N(0,1)$ car $G\,{}^tG = I$, (G orthogonale cf. Cours C 4) et

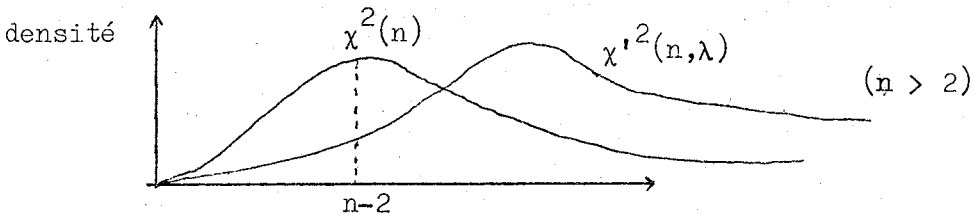
$$< GN , G\mu > = < N,\mu >$$

donc ΣX_i^2 a même loi que ΣY_i^2 , puisque $(T_1 \dots T_n)$ et $(T'_1 \dots T'_n)$ ont même loi, et $\Sigma X_i^2 = \|T'\|^2 + < T',\nu > + \|\nu\|^2$, $\Sigma Y_i^2 = \|T\|^2 + < T,\nu > + \|\nu\|^2$.

Remarque. On peut aussi calculer la fonction caractéristique de Y et constater qu'elle ne dépend que de n et λ :

$$\Phi(t) = \frac{1}{(\sqrt{1-2it})^n} \; e^{\frac{it}{1-2it}\sum_1^n \mu_j^2}$$

Distribution d'un χ'^2 .



On calcule :

$$EY = \Sigma\, EX_i^2 = n + \lambda$$

$$\text{Var } Y = 2n + 4\lambda$$

on démontre que :

$P(Y < \rho)$ pour un ρ fixé est une fonction strictement croissante de λ (paramètre de non centralité).

b) Variables F et F'.

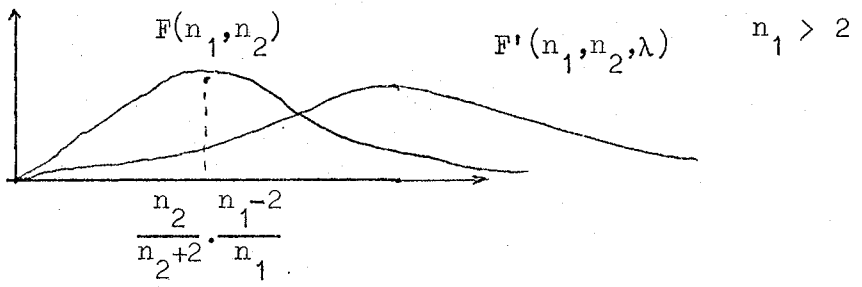
Si U_1 et U_2 sont des v.a. indépendantes de loi $\chi'^2(n_1,\lambda)$ et $\chi^2(n_2)$

$$\frac{U_1/n_1}{U_2/n_2} = F'(n_1, n_2, \lambda)$$

est une variable F' de Fischer de paramètres n_1, n_2 et λ .

Si $\lambda = 0$ on a une variable $F(n_1,n_2)$.

Densités



On vérifie, comme plus haut :

$$P(F'_{n_1,n_2,\lambda} < \rho)$$

est une fonction strictement croissante de λ .

c) Variable t de Student.

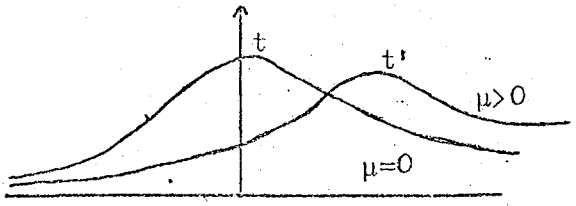
Si X et U sont des v.a. indépendantes de loi respectivement $N(\mu,1)$ et $\chi^2(n)$, la variable $\frac{X}{\sqrt{U/n}}$ a une distribution t' de Student :

$$t'(n,\mu) = \frac{X}{\sqrt{U/n}}$$

si $\mu = 0$

$$t'(n,0) = t(n).$$

Densités



II. THEOREME DE COCHRAN.

Théorème. Soient $Y_1 \ldots, Y_i \ldots, Y_n$ des v.a. gaussiennes $N(\eta_i,\gamma)$ indépendantes et $Q_1, Q_2, \ldots Q_s$ des formes quadratiques des $(Y_i)_{i\in 1\ldots n}$ telles que :

$$\sum_1^n Y_i^2 = Q_1 + Q_2 + \ldots + Q_s$$

soit n_j = rang Q_j. Si $\sum_1^s n_j = n$ alors les Q_j sont des v.a. indépendantes de loi $\gamma\, \chi'^2(n_j,\lambda_j)$ où

$$\lambda_j = Q_j(\eta_1 \ldots \eta_n).$$

Application. Estimation de l'espérance et de la variance d'une loi gaussienne.

Soit $\{Y_i,\ i = 1\ldots n\}$ un échantillon de variables $N(\mu,\sigma^2)$ alors :

$$\Sigma\, Y_i^2 = n\bar{Y}^2 + \sum_{i=1}^n (Y_i-\bar{Y})^2 \quad \text{où} \quad \bar{Y} = \frac{1}{n}\sum_1^n Y_i$$

les rangs des deux formes quadratiques étant respectivement 1 et n-1 le théorème de Cochran s'applique :

$\Sigma(Y_i-\bar{Y})^2$ est indépendante de $\bar{Y}$ et distribuée comme un $\gamma\, \chi^2_{n-1}$;

en effet $E(Y_i-\bar{Y}) = 0$ le 2e paramètre est donc nul.

Comme $E\, \chi^2_{n-1} = n-1$

$$S = \frac{1}{n-1}\,\Sigma\,(Y_i-\bar{Y})^2 \quad \text{est estimateur sans biais de } \gamma,$$

comme on l'a vu au chapitre II.

Pour la démonstration du théorème démontrons la

Proposition (algèbre).

Si $\sum_{i=1}^{n} y_i^2 = Q_1 + \ldots + Q_s$ où Q_j est une forme quadratique en $(y_i)_{i=1\ldots n}$ de rang n_j, la condition

$$n = \sum_{j=1}^{s} n_j$$

est nécessaire et suffisante pour qu'il existe une transformation orthogonale A telle que

$$z = Ay, \quad z = (z_1 \ \ldots \ z_n)$$

et que

$$Q_1 = \sum_{1}^{n_1} z_i^2$$

$$Q_2 = \sum_{n_1+1}^{n_2+n_1} z_i^2$$

$$\vdots$$

$$Q_s = \sum_{n_1+\ldots+n_{s-1}+1}^{n_1+\ldots+n_s} z_i^2 \ .$$

Démonstration. Condition nécessaire. Si une telle transformation existe alors $\sum_{1}^{n} y_i^2 = \sum_{1}^{n_1+\ldots+n_s} z_i^2$ dans un espace euclidien, le rang d'une forme quadratique étant invariant par changement de base, on a

$$n = n_1 + \ldots + n_s \ .$$

Condition suffisante. Le rang de Q_j est n_j, il existe donc des formes linéaires z_α^j en y_i $(i = 1 \ \ldots \ n)$ telle que

$$Q_j = \sum_{\alpha} \delta_\alpha^j \, z_\alpha^j \qquad \text{où } \delta_\alpha = +1 \text{ ou } -1$$

$$\text{et } \alpha = n_1 + \ldots + n_{j-1}+1, \ldots, n_1 + \ldots + n_j \ .$$

Si $\sum_{j=1}^{s} n_j = n$ on a donc n formes linéaires z_α^j définissant une matrice A dans la base de y_i, soit :

$$[z] = A[y] \qquad ([z], [y] \text{ vecteurs colonnes}).$$

Soit D la matrice diagonale des δ_α^j, $\alpha = 1 \ \ldots \ n$. Alors

$$\sum_{1}^{s} Q_j = \sum_{j=1}^{s} \sum_{\alpha} \delta_\alpha^j \, z_\alpha^{j2} = {}^t[z] \, D \, [z] = {}^t[y] \, {}^tA \, D \, A[y]$$

mais on a aussi

$$\sum^{s} Q_i = \Sigma \, y_i^2 = {}^t[y][y] \ ,$$

la matrice ${}^tADA$ est symétrique donc l'égalité des 2 formes quadratiques entraîne

$${}^tADA = I$$

donc A est régulière ;
démontrons que $D = I$ en effet si $\delta_\beta^{j_o} = -1$ comme $[y] = A^{-1}[z]$ en prenant $z_\alpha^j = 0$ pour $(\alpha,j) \neq (\beta,j_o)$ on a

$$\sum_{i=1}^{n} y_i^2 = \sum_{j=1}^{s} \sum_{\alpha=1}^{n_j} \delta_\alpha^j z_\alpha^{j2}$$

$$= \delta_\beta^{j0}$$

$$= -1 \quad \text{ce qui est impossible.}$$

Donc $D = I$ et ${}^tAA = I$ soit A est orthogonale cqfd.

Remarque. La condition $\sum_{j=1}^{s} n_j = n$ implisque en particulier que les formes Q_j sont positives.

Démonstration du théorème de Cochran.

Condition nécessaire. Si les Q_j sont des v.a. $\sim \chi'^2$ indépendantes de dimension n_j, il est immédiat que $\sum_{j=1}^{s} Q_j \sim \chi'^2$ de dimension $\sum_{j=1}^{s} n_j$ et comme $\sum_{j=1}^{s} Q_j = \sum_{i=1}^{n} y_i^2$ on a $n = \sum_{j=1}^{s} n_j$.

Condition suffisante. Supposons $n = \sum_{j=1}^{s} n_j$. Soit A la transformation orthogonale définie dans le lemme précédent. Les v.a. $Z_\alpha^j = z_\alpha^j (Y_1 \dots Y_n)$ sont des v.a. gaussiennes indépendantes. On en déduit que les Q_j sont des v.a. $\sim \gamma\, \chi'^2$ indépendantes. En effet si $[Z] = (Z_\alpha^j)$ $\alpha = 1 \dots n_j$, $j = 1 \dots s$
On a $\operatorname{cov}[Z] = E(([Z] - E[Z]) \otimes ([Z] - E[Z])) = {}^tA \operatorname{cov} Y\, A = {}^tA\, \gamma\, I\, A = \gamma I$ (cf. cours probabilité chap. 4).

On a le coefficient de non centralité de Q_j par $\lambda_j = \sum_\alpha (E\, Z_\alpha^j)^2$
or $Q_j(Y_1 \dots Y_n) = \sum_{\alpha=1}^{n_j} Z_\alpha^{j\,2}$. Z_j^α est la $(\alpha,j)^{\text{ème}}$ composante de $A[Y]$
donc $E\, Z_j^\alpha = E(A[Y])_{(\alpha,j)}$

$$\sum_{\alpha=1}^{n_j} (E\, Z_j^\alpha)^2 = \sum_{\alpha=1}^{n_j} [(E\, A[Y])_{(\alpha,j)}^2]$$

$$\sum_{\alpha=1}^{n_j} (A[y])_{(\alpha,j)}^2 = Q_j(y_1 \dots y_n)$$

donc $\sum_{\alpha=1}^{n_j} (E\, Z_j^{\alpha})^2 = Q_j(E\, Y_1 \ldots E\, Y_n)$ puisque E et A commutent. cqfd.

Remarque. Le théorème est encore valable si on remplace $\sum_{j=1}^{s} n_j = n$ par $\sum_{j=1}^{s} n_j \leq n$.

III. EXEMPLES DE TESTS CONSTRUITS AVEC DES V.A. GAUSSIENNES.

1. Comparaison des moyennes de 2 v.a. gaussiennes indépendantes de même variance.

On se propose de comparer les moyennes de 2 v.a. X et Y avec $X \sim N(\mu,\sigma^2)$ et $Y \sim N(\nu,\sigma^2)$.

On veut tester $H_o : \{\mu = \nu\}$ contre $H_1 \quad \{\mu \neq \nu\}$.

Si on a un n-échantillon du premier type de v.a. soit $X_1 \ldots X_n$ et un m-échantillon du 2e type de modèle est

$$\left(\mathbb{R}, \mathcal{B}, \frac{1}{\sqrt{2\pi}\,\sigma} \exp \frac{(x-\mu)^2}{2\sigma^2}\right)^{\otimes n} \otimes \left(\mathbb{R}, \mathcal{B}, \frac{1}{\sqrt{2\pi}\,\sigma} \exp \frac{(y-\nu)^2}{2\sigma^2}\right)^{\otimes n}, (\mu,\nu) \in \mathbb{R}^2$$

et il s'agit d'un test particulier d'homogénéité à savoir est-ce que les v.a. ont même loi ?

Si $m = n$ (on dit que l'échantillonnage est équilibré), en groupant arbitrairement les v.a. X_i et Y_j on peut prendre comme modèle

$$\bigotimes_1^n \left(\mathbb{R}, \mathcal{B}, \frac{1}{\sqrt{2\pi}\sqrt{2\sigma}} \exp \frac{(z-\theta)^2}{4\sigma^2}\right)_{\theta \in \mathbb{R}} \qquad \text{où} \quad \theta = \mu - \nu .$$

a) On connait σ^2 .

Prenons d'abord $m = n = 1$ et étudions complètement le problème dans ce cas.

$$U = X - Y$$
$$\sim N(\mu-\nu,\ 2\sigma^2)$$
$$\sim N(\theta,\ 2\sigma^2)$$

donc

$$V = \frac{U}{2\sigma^2}$$
$$\sim N\left(\frac{\theta}{2\sigma^2},\ 1\right).$$

On a un test défini par la région de rejet $D = \{\omega, |V(\omega)| > \rho\}$. Si α est le niveau désiré, alors α est déterminé par

$$P_{\theta=0}(|V| > \rho) = 2(1 - F(\rho))$$

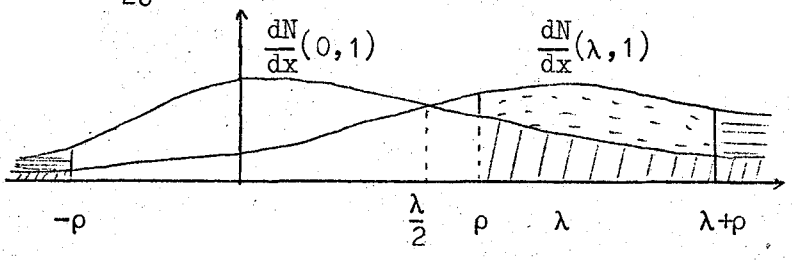
où F est la fonction de répartition tabulée d'une loi $N(0,1)$.

Etudions la puissance de ce test. On a

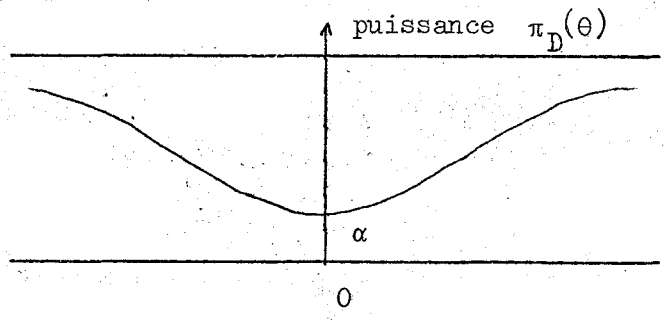
$$P_\theta(|V| > \rho) = P(|N + \frac{\theta}{2\sigma^2}| > \rho)$$

où $N \sim N(0,1)$.

Posant $\lambda = \frac{\theta}{2\sigma^2}$, on voit que la puissance vaut $2 - F(\rho-\lambda) - F(\rho+\lambda)$



La puissance est représentée par l'aire en pointillé, c'est une fonction paire, croissante de $|\theta|$



Remarquons que le test est sans biais.

Rappel : <u>Définition. Un test est dit sans biais si</u> $P_\theta(D) > P_{\theta'}(D)$ <u>pour tout</u> $\theta \in \Theta_1$ (H_o <u>fausse</u>), $\theta' \in \Theta_o$ (H_o <u>vraie</u>)

Ici $\Theta_o = \{0\}$.

Supposons maintenant disposer pour X d'un n-échantillon $(X_1 \dots X_n)$ et pour Y d'un m-échantillon $(Y_1 \dots Y_m)$. Posons comme d'habitude

$$\bar{X} = \frac{X_1+\dots+X_n}{n}, \quad \bar{Y} = \frac{Y_1+\dots+Y_m}{m}$$

$$U = \bar{X} - \bar{Y} \sim N(\mu - \nu, \frac{\sigma^2}{n} + \frac{\sigma^2}{m})$$

$$V = \frac{\bar{X} - \bar{Y}}{\sigma(\bar{n}^1+\bar{m}^1)^{1/2}} \sim N(\lambda,1)$$

avec $\lambda = \frac{\mu-\nu}{\sigma(\bar{n}^1+\bar{m}^1)^{1/2}}$. La courbe de puissance précédente reste valable. On voit comment la puissance $\pi_D(\theta)$ croît quand n et m croissent ce qui est évidemment normal.

b) Supposons maintenant σ^2 inconnu.

Dans le test à faite, on ne s'intéresse pas à σ^2, puisque celui-ci ne figure pas dans l'hypothèse. Il faut donc faire un test valable pour tout σ^2.

Si on pose $s = \frac{1}{n+m-2}\left[\sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{j=1}^{m}(X_j-\bar{Y})^2\right]$, on a $E_{(\mu,\nu,\sigma^2)}s = \sigma^2$. s est donc un estimateur sans biais de σ^2.

$\bar{X} - \bar{Y}$ n'a pas une loi indépendante de σ^2 puisque $\bar{X} - \bar{Y} \sim N(\mu - \nu, (\frac{1}{n} + \frac{1}{m})\sigma^2)$, mais

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{s(\frac{1}{n}+\frac{1}{m})}}$$ a une loi qui est

la loi d'un quotient de 2 v.a. indépendantes d'après Cochran, $\bar{X} - \bar{Y}$ et s. Alors Z a une loi $t'(n+m-2, \lambda)$, soit $t(n+m-2, \lambda)$ qui est tabulée.

Le test est alors construit.

Si H_o est vraie, soit $\lambda = 0$, Z est centree. Intuitivement, il y a donc une grande probabilité de trouver Z' près de 0 si H_o est vraie. On prend tout pour D la région $(|Z'| > \rho)$.

Ce test est sans biais. ρ est déterminé par $\alpha = P_{\theta=0}(|Z'| > \rho)$. (Si $\lambda = 0$, on a vu la loi de Z qui est $t(n+m-2)$).

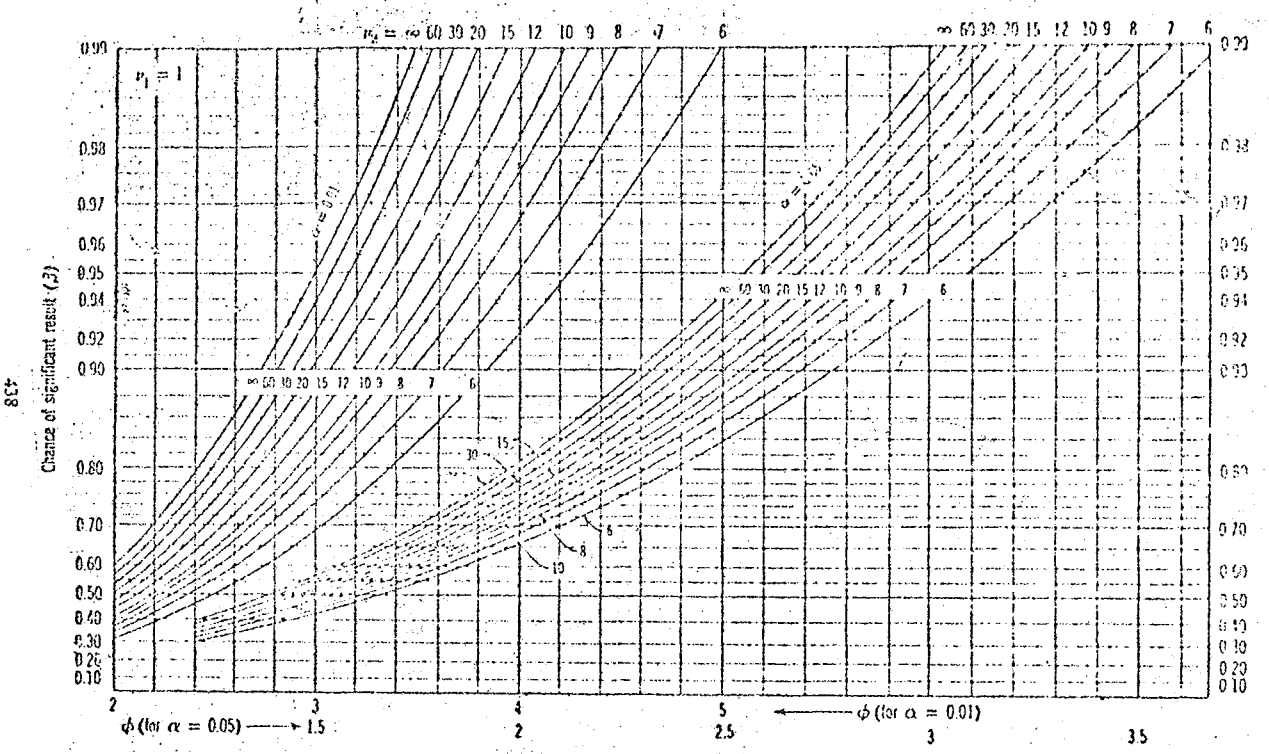
Z^2 est une variable $F'_{1,n+m-2,(\lambda)^2}$ (quotient de deux χ^2 (un $\chi^2(1,(\lambda)^2)$ est un $\chi^2(n+m-2)$ centré), et comme $P(Z^2 > \rho^2)$ est une fonction strictement croissante de $(\lambda)^2$ (voir plus haut), on a bien

$$P_o(Z^2 > \rho^2) < P_{(\mu-\nu)^2}(Z^2 > \rho^2) \quad \text{pour} \quad \nu \neq \mu .$$

Remarque. Nous verrons plus loin les qualités de ce test (quant à l'optimalité).

Etude de la puissance du test ainsi obtenu

PEARSON AND HARTLEY CHARTS* FOR THE POWER OF THE F-TEST



* By E. S. Pearson and H. O. Hartley in *Biometrika*, Vol. 38, pp. 115–122 (1951). Reproduced with the kind permission of the authors and the editor.

Ces courbes donnent la puissance d'un test F, $P(F'_{1,\nu_2,\lambda} > \rho^2)$ pour des niveaux $\alpha = 0{,}01$ et $\alpha = 0{,}05$, c.à.d. que ρ^2 est calculé pour que le niveau soit le bon.

On a porté en abcisse $\phi = \frac{\sqrt{\lambda}}{\sqrt{2}}$, $(\nu_2 = n+m-2,\ \lambda = (\mu-\nu)^2)$.

Supposons que la valeur de l'échantillon donne une valeur supérieure à ρ^2 (pour Z'^2) c.à.d. qu'il faudrait rejeter H_o : il est important de s'assurer que dans la région où on l'applique le test est assez puissant.

Exemple. On sait (par une estimation, par exemple, ou quelquefois par des expériences antérieures) que Φ est de l'ordre de $2{,}5$. Si $\nu_2 = 6$ la puissance du test de niveau $0{,}01$ est $0{,}5$ c'est-à-dire que la probabilité d'avoir conclu juste est $0{,}5$, ce n'est pas très sérieux. Il faut continuer l'étude. La situation n'est pas très probante. Le lecteur réfléchira à cette

situation courante qui explique l'importance de l'étude de la puissance.

Remarque. Si le degré de liberté du χ^2 du dénominateur augmente la puissance du test aussi : d'où un intérêt d'avoir de grands échantillons.

Le cas $\nu_2 = \infty$ porté sur la figure correspond au carré d'une variable gaussienne, en effet $\lim_n \frac{1}{n} \Sigma X_i^2 = 1$ (loi des grands nombres) on a donc aussi le courbe de puissance du test a).

Autre exemple de test :

Comparaison des variances de deux lois normales.

On dispose de deux échantillons indépendants

$$(X_i)_{i=1\ldots n} \text{ de loi } N(\mu, \gamma_1)$$
$$(Y_i)_{i=1\ldots m} \text{ de loi } N(\nu, \gamma_2)$$

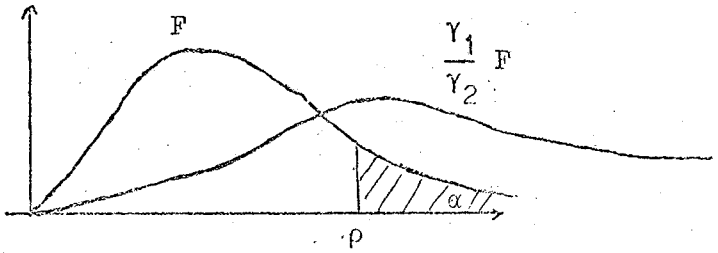
alors

$$s_X = \sum_1^n (X_i - \bar{X})^2 \sim \gamma_1 \chi^2_{n-1}$$
$$s_Y = \sum_1^m (Y_j - \bar{Y})^2 \sim \gamma_2 \chi^2_{m-1}$$

donc

$$U = \frac{s_X/n-1}{s_Y/m-1} \sim \frac{\gamma_1}{\gamma_2} F_{n-1,m-1} .$$

a) Supposons qu'on veuille tester $\gamma_1 = \gamma_2$ contre $\gamma_1 > \gamma_2$. Les densités de F et $\frac{\gamma_1}{\gamma_2} F$ ont l'allure suivante :



on propose donc le test

$$D = \text{rejet de } H_o \text{ si } U > \rho$$

où ρ est calculé par $P[F_{n-1,m-1} > \rho] = \alpha$ (niveau désiré).

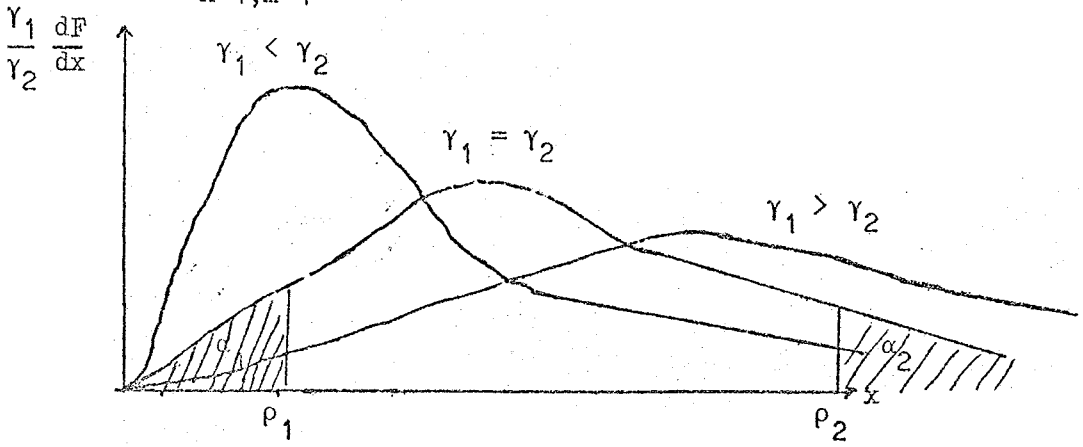
Puissance du test.

$$P[U > \rho] = P[\frac{\gamma_1}{\gamma_2} F_{n-1,m-1} > \rho] = P[F > \frac{\gamma_2}{\gamma_1} \rho]$$

si $\frac{\gamma_2}{\gamma_1} < 1$ c.à.d. H_o fausse, la puissance est supérieure à α et pour avoir

une idée de sa valeur on utilise une table donnant la fonction de répartition d'une variable F .

b) Supposons maintenant qu'on veuille tester $\gamma_1 = \gamma_2$ contre $\gamma_1 \neq \gamma_2$ il faut donc exclure les valeurs de U où les variables $\frac{\gamma_1}{\gamma_2}F$ pour $\gamma_1 \neq \gamma_2$ sont trop probables, c.à.d. les petites et les grandes valeurs de U .(nous écrivons F pour $F_{n-1,m-1}$ dans la suite)



on va exclure H_o si

$$U < \rho_1 \quad \text{ou} \quad U > \rho_2$$

$$D = [U < \rho_1] \; [U > \rho_2].$$

Pour obtenir le niveau α il faut que

$$P[F < \rho_1] + P[F > \rho_2] = \alpha$$

mais il reste une indétermination puisqu'il y a deux bornes :

on peut par exemple choisir

$$P[F < \rho_1] = \alpha/2 \quad \text{et} \quad P[F > \rho_2] = \alpha/2$$

on peut aussi chercher un test sans biais c.à.d. tel que

$$P[\tfrac{1}{\lambda} F < \rho_1] + P[\tfrac{1}{\lambda} F > \rho_2]$$

soit minimum pour $\lambda = 1$. Si g est la densité d'une variable F

$$P[F < \lambda\rho_2] + P[F > \lambda\rho_2] = \int_0^{\lambda\rho_1} g(x)dx + \int_{\lambda\rho_2}^{\infty} g(x)dx$$

la condition "sans biais" s'écrit :

$$\rho_1 g(\rho_1) - \rho_2 g(\rho_2) = 0$$

or $$g(x) = \frac{\frac{p}{q}}{\beta(\frac{p}{2},\frac{q}{2})} \frac{(\frac{p}{q}x)^{\frac{p}{2}-1}}{(1+\frac{p}{q}x)^{\frac{p+q}{2}}}$$ pour une $F_{p,q}$

d'où une résolution possible en ρ_1, ρ_2 .

En fait ce test est intéressant si on garde l'hypothèse H_o et dans ce cas le caractère "sans biais" est secondaire. En effet, pour conclure au rejet, il faut encore s'assurer qu'on est dans une région où la puissance est raisonnable : c.à.d. que, ayant une idée du rapport $\frac{\gamma_1}{\gamma_2}$ on calcule la puissance en ce point, mais dans ce cas le test unilatère précédent est plus justifié (puisque on a déjà une estimation de $\frac{\gamma_1}{\gamma_2}$).

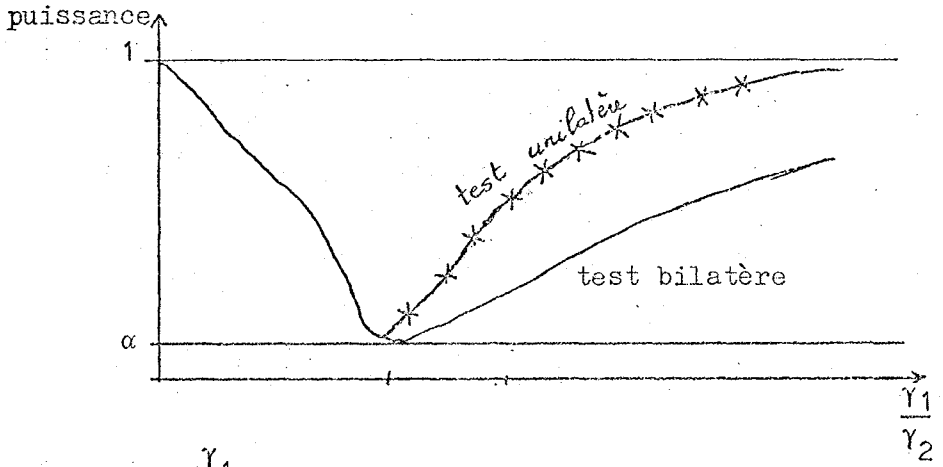
Remarque pour le calcul de ρ_1 et ρ_2 . Les tables donnent ρ tel que

$$P[F_{n,m} > \rho] = \alpha$$

pour des niveaux α classiques. L'autre borne s'obtient en remarquant que

$$P[F_{n,m} < \rho] = P[F_{m,n} > \frac{1}{\rho}] .$$

Nous n'avons fait qu'amorcer ici la discussion. On voit que le choix entre le test unilatère ($\gamma_1 = \gamma_2$ contre $\gamma_1 > \gamma_2$) et le test initial ($\gamma_1 = \gamma_2$ contre $\gamma_1 \neq \gamma_2$) est très technique et dépend du crédit que l'on accorde à l'estimation de $\frac{\gamma_1}{\gamma_2}$).



Pour les valeurs $\frac{\gamma_1}{\gamma_2}$ un peu plus grande que 1 le choix est délicat et nous reviendrons dessus plus loin du point de vue théorique (concrètement bien que démarrant avec l'idée de faire un test bilatère il peut être mauvais de rester sur ce point de vue).

Les moyennes μ et ν interviennent ici comme paramètres fantômes, elles sont été éliminées d'entrée.

Nous n'aborderons pas dans le cours l'aspect théorique de l'élimination de ces paramètres fantômes, nous contentant de voir sur divers exemples comment D est fabriqué de manière à ce qu'ils ne figurent pas.

Intervalles de confiance.

Nous ne détaillerons pas ici la forme des régions de confiance possible. A l'aide des statistiques introduites pour les tests, le lecteur le fera aisément. Indiquons simplement deux petits résultats numériques fort utiles. Soit un n-échantillon de $N(\mu,\sigma^2)$, Φ la fonction de répartition de $N(0,1)$. On a $\Phi^{-1}(0,025) = -1,96 \# -2$

$$\Phi^{-1}(0,015) = -3.$$

Par exemple si σ est connu les intervalles de confiance pour μ seront

a) bilatère

$$\bar{X} - \frac{\sigma}{\sqrt{n}} \left|\Phi^{-1}(\tfrac{\alpha}{2})\right|, \ \bar{X} + \frac{\sigma}{\sqrt{n}} \left|\Phi^{-1}(\tfrac{\alpha}{2})\right|$$

b) unilatère à droite

$$\left]\bar{X} + \frac{\sigma}{\sqrt{n}} \left|\Phi^{-1}(\alpha)\right|, \infty\right[.$$

APPENDICE ET RESUME

A) Notes sur quelques lois intervenant souvent à propos des problèmes sur les gaussiennes notamment.

La loi Gamma $\Gamma(a,\lambda)$ a pour densité

$$\gamma_{a,\lambda}(x) = \frac{1}{\Gamma(a)} \lambda^a e^{-\lambda x} x^{a-1} 1_{\mathbb{R}^+}(x)$$

où $$\Gamma(a) = \int_0^\infty e^{-x} x^{a-1} dx$$

$\Gamma(a+1) = \Gamma(a)$, $\Gamma(n+1) = n !$

Pour $a = 1$, on obtient la loi exponentielle ordinaire.

On a $$\int_0^\infty e^{-tx} \gamma_{a,\lambda}(x)dx = (1 + \frac{t}{\lambda})^{-a}$$

et $$\Gamma(a,\lambda) * \Gamma(b,\lambda) = \Gamma(a+b, \lambda)$$

$$\Gamma(\tfrac{n}{2}, 1/2) = \chi^2(n).$$

Soient X et Y deux variables indépendantes $X \sim \Gamma(a,\lambda)$, $Y \sim \Gamma(b,\lambda)$.

Alors X/Y a pour loi $\beta(a,b)$ de densité

$$g_{a,b}(x) = \frac{\Gamma(a+b)}{\Gamma(a)\ \Gamma(b)} \ \frac{x^{a-1}}{(1+x)^{a+b}} \ 1_{\mathbb{R}^+}(x) .$$

Si $U \sim \beta(\frac{n_1}{2}, \frac{n_2}{2})$, $\frac{n_2}{n_1} U \sim F(n_1, n_2)$.

B) RESUME. Statistique à utiliser :

sur la moyenne μ

a) σ connu $\sqrt{n}\,\dfrac{\bar{X}-\mu}{\sigma} \sim N(0,1)$

b) σ inconnu $\sqrt{n}\,\dfrac{\bar{X}-\mu}{s} \sim T(n-1)$.

Statistique à utiliser sur σ

a) μ connu $\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \sim \chi^2(n)$

b) μ inconnu $\dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$.

Pour 2 échantillons on a

$$\bar{X}-\bar{Y} \sim N\left(\mu-\mu', \frac{\sigma^2}{n}+\frac{\sigma'^2}{n}\right)$$

$$\frac{(n-1)s_X^2}{\sigma^2} - \frac{(n'-1)s_Y^2}{\sigma'^2} \sim \chi^2(n+n'-2)$$

$$\frac{\sigma'^2}{\sigma^2}\,\frac{s_X^2}{s_Y^2} \sim F(n-1,\, n'-1)$$

et si $\sigma^2 = \sigma'^2$

$$\frac{\bar{X}-\bar{Y}-(\mu-\mu')}{\left[(n-1)s_X^2+(n'-1)s_Y^2\right]^{1/2}}\;\frac{\sqrt{n+n'-2}}{\sqrt{\frac{1}{n}+\frac{1}{n'}}} \sim T(n+n'-2).$$

CHAPITRE VI

LE TEST DU χ^2

Il s'agit d'une méthode non paramétrique. Nous l'avons présentée à part car elle n'est pas fondée sur des considérations portant sur les statistiques d'ordre, elle n'a de justification qu'asymptotique. Ceci étant c'est sûrement le test statistique le plus employé et son emploi correct présente de nombreuses difficultés. Nous expliquons dans ce chapitre le fondement du test (loi multinomiale) et les conditions d'applications les plus communes, sans démontrer les théorèmes intéressants, concernant les rapports test-estimation, quant au choix convenable du nombre de degrés de liberté.

1. PRELIMINAIRES. Convergence d'une loi multinomiale vers un χ^2.

Soient $p_1 \dots p_r$, r nombres strictement positifs tels que $\sum_{i=1}^{r} p_i = 1$. Soient X_j des v.a. indépendantes telles que $P(X_j = i) = p_i$, $j = 1 \dots n$; $i = 1 \dots r$. On pose

$$Z_i^n = \sum_{j=1}^{n} 1_{\{i\}}(X_j), \quad Z^n = (Z_1^n, \dots, Z_r^n).$$

On a : $P(Z^n = (n_1, \dots n_r)) = \frac{n!}{n_1! \dots n_r!} p_1^{n_1} \dots p_r^{n_r}$, où $n_1, \dots, n_r \in \mathbb{N}$, $\sum_{i=1}^{r} n_i = n$, par un raisonnement combinatoire élémentaire. Z^n est donc une v.a. portée par hyperplan $\sum_{i=1}^{r} n_i = n$ de $\mathbb{N}^r$. La loi de Z^n, notée $M_d(n, p_1, \dots, p_r)$ est appelée loi multinomiale dégénérée. Si $M_d(n_1, p_1, \dots, p_r)$ et $M_d(n_2, p_1, \dots, p_r)$ sont les lois de 2 v.a. indépendantes Z^{n_1} et Z^{n_2} alors $M_d(n_1+n_2, p_1, \dots, p_r)$ est la loi de $Z^{n_1} + Z^{n_2}$.

Posons $Y^n = (Y_1^n, \dots, Y_r^n)^t$

$$= \left(\frac{Z_1^n - np_1}{\sqrt{np_1}}, \dots, \frac{Z_r^n - np_r}{\sqrt{np_r}} \right)$$

et $\sqrt{p} = (\sqrt{p_1} \dots \sqrt{p_r})^t$.

Posons $V_j = \left(\frac{1}{\sqrt{p_i}}\, 1_{\{i\}}(X_j)\right)_{i=1\ldots r}$

Les V_j forment une suite de n vecteurs équidistribués indépendants. Comme

$$E\left(\frac{1}{\sqrt{p_h}}\, 1_{(h)}(X_j) - \sqrt{p_h}\right)\left(\frac{1}{\sqrt{p_k}}\, 1_{(k)}(X_j) - \sqrt{p_k}\right) = \delta^h_h - \sqrt{p_h p_k}\,,$$

on a $\operatorname{cov} V_j = I_r - {}^t\sqrt{p}\,\sqrt{p}$ où I_r est la matrice identité sur $\mathbb{R}^r$.

D'après le théorème de convergence vers une loi normale dans $\mathbb{R}^r$ (cf. cours de Probabilités, chapitre 6) on a :

$$\frac{\sum_{j=1}^{n} (V_j - \sqrt{p})}{\sqrt{n}} \xrightarrow{\text{loi}} N(0, I_r - {}^t\sqrt{p}\,\sqrt{p})$$

et comme l'application $\mathbb{R}^n \to \mathbb{R}$ définie par $x \to \|x\|$ est continue, on a

$$\left\| \frac{\sum_{1}^{n} V_j - \sqrt{p}}{\sqrt{n}} \right\|^2 \xrightarrow{\text{loi}} \text{loi} \quad \|N(0, I_r - {}^t\sqrt{pp})\|^2$$

(si T est une fonction $\mathbb{R}^n \to \mathbb{R}^k$ continue et si $X \xrightarrow{\text{loi}} Y$ alors $T(X) \xrightarrow{\text{loi}} T(Y)$ puisque $\int f \circ (T \circ X) = \int (f \circ T)(X)$ et que $f \circ T \in C(\mathbb{R}^n)$ si f est continue, bornée de $\mathbb{R}^k$ dans $\mathbb{R}$).

Il reste à calculer la loi du carré de la norme d'une v.a. $Z \sim N(0, I_r - {}^t\sqrt{pp})$. La norme est invariante par un changement de base orthonormée. Soit B une telle matrice orthonormée du type $(\sqrt{p}/C)$, ($\sqrt{p}$ est acceptable pour une première colonne puisque $\|\sqrt{p}\|^2 = 1$). On a $BZ \sim N(0, B(I_r - {}^t\sqrt{pp})\,{}^tB)$ et $B(I_r - {}^t\sqrt{pp})\,{}^tB = B\,{}^tB - (\sqrt{p}/C\ {}^t\sqrt{p}\ \sqrt{p}\ {}^t(\sqrt{p}/C))$ or $(\sqrt{p}/C)\begin{pmatrix} {}^t\sqrt{p} \\ {}^tC \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & I_{r-1} \end{pmatrix}$ puisque B est orthogonale. Soit $C\ {}^t\sqrt{p} = 0$ et donc $(\sqrt{p}/C)({}^t\sqrt{p}\ \sqrt{p}) = (\sqrt{p}/C)$ et donc

$$B(I_r - {}^t\sqrt{pp})\,{}^tB = \begin{pmatrix} 1 & 0 \\ 0 & I_{r-1} \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 0 \\ 0 & I_{r-1} \end{pmatrix}$$

donc la loi de $\|Z\|^2$ est un χ^2_{r-1}.

2. Mesure de la distance de deux lois. (Cf. aussi Chap. IV test de Kolmogorov et Smirnov).

Commençons par rappeler la démonstration du lemme suivant :

Soit un n-échantillon $X_1, \ldots, X_n$ de v.a. à valeurs dans $\mathbb{R}$ (ou $\mathbb{R}^n$).

Si A est un borélien posons $F_n(A) = \frac{n_A}{n}$ où n_A = nombre $(X_j \in A,\ 1 \leq j \leq n)$. Alors $F_n \to F$, où F est la loi de X.

F_n est en effet (c'est trivial) une probabilité sur $\mathbb{R}$ (dite répartition ou loi empirique).

On a $n_A = \sum_{j=1}^{n} 1_A(X_j)$.

Les v.a. $1_A(X_j)$ sont indépendantes, de même loi, et $E1_A(X_j) = F(A)$. La loi forte des grands nombres assure alors que $\frac{n_A}{n} \to F(A)$ p.s. d'où le lemme.

Il y a de très nombreuses manières de mesurer la distance de 2 lois F_n et F. Les mesures théoriques classiques sont intéressantes. Dans ce chapitre nous définirons des "distances" $d(F_n,F)$ avec l'idée suivante. Les lois F_n. que nous considérons convergent sous une hypothèse convenable H_o vers la loi F. Dans un problème d'estimation on choisira F (loi à estimer) de manière que $d(F_n,F)$ soit minimale, F_n étant la loi empirique associée à l'échantillon. Dans un problème de test, F sera fixée, si H_o hypothèse à tester est vraie, alors on aura $d(F_n,F)$ petite, si H_o est fausse on aura $d(F_n,F)$ grande. Ceci étant dit, les "distances" élémentaires (qui ne sont pas des vraies distances) sont fondées sur la loi multinomiale. Soit $(p_i)_{i=1\ldots r}$ une loi à r valeurs a_i, et soit $X_1 \ldots X_n$ un échantillon prenant ces mêmes r valeurs a_i. Soit n_i = nombre $(X_j = a_i)$. La loi empirique est $(\frac{n_i}{n})_{i=1\ldots r}$. On mesure sa distance à la loi $(p_i)_{i=1\ldots r}$ par

$$\chi^2 = \sum_{i=1}^{r} \frac{(\frac{n_i}{n} - p_i)^2\, n}{p_i}$$

$$= \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$$

ou par

$$d_{\text{Hellinger}} = \cos^{-1} \sum_{i=1}^{r} \frac{n_i}{n} p_i$$

$$d_{\text{Kullback}} = \Sigma\ p_i \log\left[\frac{p_i}{n_i}\, n\right] \quad \text{etc...}\ .$$

Pour estimer les nombres p_i qui caractérisent la vraie loi, on décide (arbitrairement mais on peut justifier asymptotiquement) de minimiser

ces distances. Pour ce faire, on suppose que p_i est une fonction bien dérivable d'un paramètre θ variant dans un bon "ouvert", du type

$$(\theta < x_i < 1 \ , \ \sum_{i=1}^{r} x_i = 1).$$

Par exemple dans le cas du χ^2 , on a en dérivant par rapport à p_i ,

$$\frac{\partial \chi^2}{\partial p_i} = 0 \implies \hat{p}_i = \frac{n_i}{n} .$$

L'estimateur obtenu $\frac{n_i}{n} \xrightarrow[(n\to\infty)]{} p_i$ vraie valeur, la suite d'estimateurs fabriquée ainsi est donc consistante. Nous laissons au lecteur le soin d'étudier les autres cas en précisant les hypothèses.

Si on a des lois F quelconques on se ramène aux techniques précédentes en approximant F par une loi à r valeurs, en choisissant une partition $A_1 \dots A_r$ de $\mathbb{R}$. On estime $p_i = \int_{A_i} dF_i$ en le comparant à $\frac{n_i}{n}$, où $n_i = \sum_{j=1}^{n} 1_{A_i}(X_j)$.

3. Les tests du χ^2 .

a) Test sur la loi multinomiale.

Soit $X_1 \dots X_n$ un n-échantillon (n grand) d'une v.a. à r-valeurs de loi $(\pi_1(\theta),\dots \pi_2(\theta))$, où $\theta \in \Theta$. Soit $p_1 = \pi_1(\theta_o),\dots, p_r = \pi_r(\theta_o)$. $0 < p_i < 1$. Soit à tester H_o :

" $\pi_1(\theta) = p_1 \ ,\dots, \pi_r(\theta) = p_r$ " contre H_1 :

" $\pi_1(\theta) \neq p_1$ ou ... ou $\pi_r(\theta) \neq p_r$ ".

On sait, en reprenant les notations du § 1, que d'après la loi p.s. des grands nombres, on a

$$\frac{Z^n}{n} \to (\pi_1(\theta),\dots \pi_r(\theta)) \quad P_\theta \text{ p.s.}$$

si $P_\theta = \otimes^n (\pi_1(\theta),\dots, \pi_r(\theta))$.

Donc $$\frac{\|Z^n - n\,p\|^2}{n} = n \left\| \frac{Z^n}{n} - p \right\|^2$$

$$\sim C^2 n \quad \text{avec} \quad C^2 = \Sigma |\pi_1(\theta) - p_i|^2$$

$$= \|\pi(\theta) - p\|^2 \quad \text{si} \quad \pi(\theta) \neq p .$$

Posons toujours $Y^n = (Y_1^n \dots Y_2^n)$

avec $$Y_i^n = \frac{Z_i^n - np_i}{\sqrt{n}\sqrt{p_i}} .$$

Le raisonnement ci-dessus montre que si $\pi(\theta) \neq p$, donc $\|Y^n\|^2 \sim C'n \to \infty$ si H_o est fausse et on a vu, que si $\pi(\theta) = p$, on avait par contre $\|Y^n\|^2 \to \chi^2(r-1)$.

On en déduit un test. On commence par assimiler pour n grand la distribution de Y^n (dans le cas où H_o est vraie) à celle d'un $\chi^2(r-1)$.

Soit α le niveau donné. Définissons d, par $P(\chi^2(r-1) \geq d) = \alpha$. Le test est défini par $D = (\|Y^n\|^2 \geq d)$. La puissance d'un tel test est évidemment très compliquée à calculer puisqu'elle dépend de $\pi_1(\theta), \dots, \pi_r(\theta)$ et aussi un peu de n. Pour cette raison, on est amené à introduire dans l'étude des tests asymptotiques de ce type (c.à.d. de test où la loi de la statistique utilisée est confondue avec la loi limite pour $n = \infty$ dans le cas où l'hypothèse est vraie) d'autres notions plus adaptées comme celle d'efficacité.

Nous ne développerons pas dans ce cours ces notions, nous contentant d'exposer le pourquoi et la méthode des tests les plus importants, sans rechercher leurs qualités dans l'ensemble des tests asymptotiques. Les χ^2 sont tabulés.

b) Tests d'ajustements.

Soit F une loi quelconque sur un ensemble E, on a un n-échantillon $X_1, \dots, X_n$.

Problème : tester H_o "la loi de X est F" contre H_1 "la loi de X n'est pas F".

Un tel problème est dit test d'ajustement (on essaie de voir si la loi F s'ajuste bien à la répartition des X_j).

A cet effet, on fait une partition de E en $A_1, \dots, A_r$ et on con-

sidère les variables $X_{j,i} = 1_{\{i\}}X_j$, $Z_i^n = \sum_{j=1}^{n} X_{j,i}$, $Z^n = (Z_1^n, \ldots, Z_r^n)$.

Z^n a une loi multinomiale $M_d(p_1 \ldots p_r)$ avec $p_i = P(X_j \in A_i)$. Le principe du test d'ajustement est alors le suivant : remplacer l'hypothèse H_o par l'hypothèse H'_o (plus faible), Z^n a une loi multinomiale $M_d(p_1 \ldots p_r)$. (Si H_o est vraie, H'_o est vrai a fortiori, l'implication contraire étant fausse). On teste alors H'_o comme au paragraphe précédent contre H_1 : Z^n a une loi $M_d(p'_1 \ldots p'_r)$, $p' \neq p$.

Ce test d'ajustement très simple dépend évidemment beaucoup du partage $A_1 \ldots A_r$ (on prend en général $r \geqslant 7$ quand c'est possible).

c) Comparaison de plusieurs échantillons de lois multinomiales. Tests d'homogénéité.

Soient m échantillons $(X_{1,i}, \ldots, X_{n_i,i})$ $i = 1 \ldots m$, de lois multinomiales associées à m populations. Problème : ces échantillons proviennent-ils d'une même population ?

Nous allons étudier le problème pour $m = 2$ sur l'exemple de groupes sanguins relevant de 2 populations (d'après Rao). On a les fréquences des groupes O, A, B, AB .

Soit p , q la fréquence des gènes A et B pour la première population, p', q' les mêmes fréquences pour la deuxième population. A-t-on $p = p'$? On a

	O	A	B	AB	TOTAL
1ère population	121	120	79	33	353
2ème population	118	95	121	30	364

Les classes sont toutes faites. Le problème se pose donc directement comme problème sur les multinomiales.

Posant $\pi_{i,k}$, $i = 1, 2$, $k = 1, 2, 3, 4$

les probabilités pour un individu de la population i d'appartenir au groupe k , il faut tester si $\pi_{i,k} = \pi_{i',k}$ pour tout k . On estime $\pi_{i,k}$ par $\frac{n_{i,k}}{n_i}$ où n_i est le nombre total d'individus de la population i , et $n_{i,k}$

le nombre de ceux du groupe k .

Supposons vérifiée l'hypothèse H_o : "homogénéité des populations". On a alors pour la population globale $n_k = n_{1,k} + n_{2,k}$ individus dans la classe k et l'on estime la probabilité π_k d'être dans la classe k par $\frac{n_k}{n}$. On calcule alors la somme des distances χ^2 (des 2 populations), distances à la répartition estimée sous H_o , soit

$$\chi^2 = \sum_{i=1}^{2} \sum_{k} \frac{\left(\frac{n_{i,k}}{n_i} - \frac{n_k}{n}\right)^2 n}{\frac{n_k}{n}}$$

$$= \sum_{i} \sum_{k} \frac{(n\, n_{i,k} - n_i n_k)^2}{n_i^2\, n_k}$$

Si l'on avait pas fait d'estimation c'est-à-dire si la loi était une loi a priori on aurait la somme de deux χ^2 $(k-1)$ indépendants soit un χ^2 $(2k-2)$. Mais, on démontre que chaque estimation d'un paramètre fait perdre un degré de liberté du χ^2 (en fait puisque l'on estime la loi a priori à partir de l'échantillon, on se place intuitivement près de la loi multinomiale la plus voisine). Ceci est valable à condition que les paramètres estimés soient fonctionnellement indépendants, c'est-à-dire qu'il n'existe une fonction (implicite) des paramètres identiquement nulle. Ici $\sum_k \pi_k = 1$ et on a donc estimé $k-1$ paramètres, il reste donc $2(k-1) - (k-1)$ degrés de liberté. Si l'on avait comparé entre elles h populations le nombre de degré de liberté serait donc $h(k-1) - (k-1) = (h-1)(k-1)$. Dans l'exemple concret $\chi^2 = 11{,}73$, ce qui pour 3 degrés de liberté amène au rejet de l'hypothèse au niveau 5 % , paramètre π_k . Le degré de liberté du χ^2 est donc $2(k-1) - k-1 = k-1$. Dans l'exemple étudié, on a un $\chi^2(3)$ qui vaut 11,73 et qui amène à rejeter l'hypothèse d'homogénéité des populations au niveau 5 %

Remarquons que dans certains problèmes on considère une des populations comme population test, et donc connue, non interprétée comme échantillon. On regarde alors si une autre population s'ajuste à celle-ci et l'on obtient un test sur une loi fixée, sans estimation donc à $(k-1)$ degrés de

liberté pour 2 populations et à $(h-1)(k-1)$ degré de liberté pour h population dont une fixe ! Mais alors l'hypothèse testée ne compare pas entre elles les $(h-1)$ populations, elle les compare séparément à H_o .

d) Test d'indépendance et tables de contingence.

On se propose de tester l'indépendance de 2 caractères constatés ou mesurés sur une même population par exemple la forme du pollen et la couleur de certaines fleurs obtenues par croisement (Bateson).

Soient F et C ces deux caractères i $(i = 1, 2)$ ils peuvent prendre les valeurs k ($k = 1,2$ long et rond) et ℓ ($\ell = 1,2$ violet, rouge). On a (table de contingence).

Forme pollen \ couleur	violet	rouge	TOTAL
long	296	27	323
rond	19	85	104
Total	315	112	427

Problème : tester l'hypothèse H_o : F et C sont indépendants.

Notons $n_{i,k}$ le nombre d'individus ayant les valeurs k, ℓ de caractères. On pose $n_\ell = \sum_k n_{\ell,k}$, $n_k = \sum_i n_{\ell,k}$. On a $\Sigma\, n_\ell = n$, $\Sigma\, n_k = n$.

Supposons que la probabilité inconnue d'être dans la classe (ℓ,k) soit $\pi_{\ell,k}$. La probabilité d'obtenir l'échantillonnage donné (loi multinomiale) vaut

$$n!\ \prod_\ell \prod_k \frac{\pi_{\ell,k}^{\,n_{\ell,k}}}{n_{\ell,k}!} = p_{\ell,k} \ .$$

L'indépendance se traduit par $\pi_{\ell,k} = \pi_\ell \pi_k$ où π_ℓ est la probabilité de l'état ℓ ,

$$\pi_\ell = \sum_k \pi_{\ell,k} \ , \quad \pi_k = \sum_\ell \pi_{\ell,k}$$

$$\Sigma\, \pi_\ell = 1 \ , \ \Sigma\, \pi_k = 1 \ .$$

On peut écrire $p_{\ell,k}$ sous la forme

$$p_{\ell,k} = n!\ \prod_\ell \frac{\pi_\ell^{\,n_\ell}}{n_\ell!} \prod_k \frac{\pi_k^{\,n_k}}{n_k!}\ \frac{\prod n_\ell! n_k!}{n!}$$

$$\times \prod \frac{1}{n_{\ell,k}!}\ \left(\frac{\pi_{\ell,k}}{\pi_\ell \pi_k}\right)^{n_{\ell,k}}$$

S'il y a indépendance, on obtient simplement l'expression

$$p_{\ell,k} = n!\ \prod_\ell \frac{\pi_\ell^{\,n_\ell}}{n_\ell!} \prod_k \frac{\pi_k^{\,n_k}}{n_k!}\ \frac{\prod n_\ell! n_k! / n_{\ell k}!}{n!}$$

La quantité $\frac{\prod n_\ell! n_k! / n_{\ell k}!}{n!}$ représente exactement (et pour toute loi multinomiale) la loi conditionnelle des $n_{\ell k}$ donnée les lois marginales n_ℓ et n_k.

Si l'on connaît à l'avance les π_k et les π_ℓ on fera le test d'indépendance suivant. On calcule

$$\chi^2 = \sum_k \sum_\ell \frac{(n_{k\ell} - n\pi_k\pi_\ell)^2}{n\pi_k\pi_\ell}$$

et on fait simplement un test d'ajustement sur la loi $\pi_k\pi_\ell$. Donc si k peut prendre les valeurs $1 \ldots m$ et ℓ les valeurs $1 \ldots m'$, le nombre de degré de liberté à choisir est $mm' - 1$ pour faire le test. Mais on peut améliorer la méthode de la manière suivante. On peut d'abord isoler les χ^2 qui correspondent aux déviations de l'échantillon par rapport aux marginales. Soit

$$\chi_1^2 = \sum_\ell \frac{(n_\ell - n\pi_\ell)^2}{n\pi_\ell}, \quad \chi_2^2 = \sum_k \frac{(n_k - n\pi_k)^2}{n\pi_k}$$

et pour $\chi_3^2 = \chi^2 - \chi_1^2 - \chi_2^2$. Comme χ_3^2 a $mm'-1$ degré de liberté $\chi_1^2(m-1)$ et χ_2^2 $m'-1$, χ_3^2 en a $mm'-1 - m+1 - m'+1 = (m-1)(m'-1)$.

χ_3^2 mesure la déviation (la distance) de l'échantillon à l'indépendance, une fois corrigées les déviations normales aux marginales et c'est donc χ_3^2 que l'on utilisera pour faire le test.

Remarque : on raisonne quelquefois en disant que $\Sigma \pi_i = 1$, $\Sigma \pi_k = 1$ et donc que le nombre de $(\pi_i\pi_k)$ intervenant dans la loi multinomiale à tester est $(r - 1)(s - 1)$, mais ce raisonnement n'est pas très clair.

Dans l'exemple précédent si il y a indépendance des caractères et si les rapports dans chaque classe valent 3 on obtient

$\chi_1^2 = 0,0945$ 1^e degré de liberté
$\chi_2^2 = 0,3443$ " " "
$\chi_3^2 = 221,6833$ pour 1 degré de liberté.

Donc en fait seul χ_3^2 importe.

Supposons maintenant que les marginales π_k, π_ℓ ne soient pas connues. On veut les estimer, par exemple par la méthode de maximum de vraisemblance et l'on trouve $\hat{\pi}_\ell = \frac{n_\ell}{n}$, $\hat{\pi}_k = \frac{n_k}{n}$

$$\chi^2 = \sum_k \sum_\ell \frac{(n_{ij} - n\frac{n_\ell n_k}{n\,n})^2}{\frac{n_\ell n_k}{n}} .$$

Quel degré de liberté faut-il choisir pour faire le test. Nous avons estimé $(m-1)+(m'-1)$ paramètres π_ℓ, π_k. Le nombre total de paramètres est $mm'-1$ L'estimation faite conduit à diminuer d'autant le degré de liberté du χ^2 Intuitivement parce que les paramètres inconnus $\pi_{k,\ell}$ sont liés par les relations $\sum_\ell \pi_{k\ell} = \hat{\pi}_k$ et $\sum_k \pi_{k\ell} = \hat{\pi}_\ell$ une fois l'estimation faite et donc que le nombre de paramètres à ajuster après estimation est $mm'-1-(m-1)-(m'-1)$, (la dernière des relations $\sum_\ell \pi_{k\ell} = \hat{\pi}_k$ se déduit des autres puisque $\sum_k \sum_\ell \pi_{k\ell} = \sum_k \hat{\pi}_k = 1$).

Ce raisonnement intuitif peut être rendu rigoureux en utilisant le résultat énoncé plus haut. Nous laisserons cette démonstration de côté, en insistant sur le fait que dans tous les cas un test d'indépendance amène un χ^2 à $(m-1)(m'-1)$ degrés de liberté.

Dans le cas numérique étudié, on trouve sur π_k, π_ℓ ne sont pas spécifiés $\chi^2 = 218,8722$, ce qui est à peu près la valeur obtenue en utilisant des marginales connues à l'avance.

De toute manière, il faut bien voir la valeur relative de ces tests. Si la taille du nombre d'individus dans une des cases π_{ik} est petite cette case peut fausser le χ^2 et il faut faire appel à des méthodes plus fines mais encore basées sur le χ^2.