Analyse Statistique

Bernard Clément, PhD

Département de mathématiques appliquées

École Polytechnique Montréal

1988

TABLE DES MATIÈRES

CHAPITRE 4: VARIABLES ET VECTEURS ALÉATOIRES

CHAPITRE 5: DISTRIBUTIONS DISCRÈTES

CHAPITRE 6: DISTRIBUTIONS CONTINUES

TABLE DES MATIÈRES

CHAPITRE 7: ESTIMATION

CHAPITRE 8: TESTS D'HYPOTHÈSES

CHAPITRE 9: L'ANALYSE DE RÉGRESSION

LISTE DES FIGURES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES TABLES

FORMULAIRE

LISTE DES TABLEAUX

LISTE DES EXEMPLES

LISTE DES EXEMPLES

EXEMPLES D'UTILISATION DE SAS

CHAPITRE 1

EXEMPLES DE SITUATIONS D'ANALYSE DE DONNÉES

1.0 SOMMAIRE

On présente des situations concrètes où une analyse statistique est nécessaire afin de motiver l'étude de l'analyse de données et des modèles stochastiques. Les éléments communs de ces exemples sont identifiés. Dans la dernière partie du chapitre on propose différentes classifications pour les variables.

1.1 COMPARAISON DE DEUX FILS

Un nouveau type de fil a été développé et on espère que sa tension de rupture sera supérieure à celui en usage présentement. Afin de mettre à l'épreuve cette proposition, on prélève vingt-cinq observations de tension de rupture sur chacun des deux types de fil:

ancien fil					nouveau fil				
150	155	146	150	148	152	154	154	154	153
148	152	150	149	154	153	154	152	150	152
147	154	155	153	152	152	153	154	152	153
154	150	151	151	153	154	157	150	156	154
151	153	149	152	150	150	152	149	154	151

Les données sont représentées à la figure 1.1.

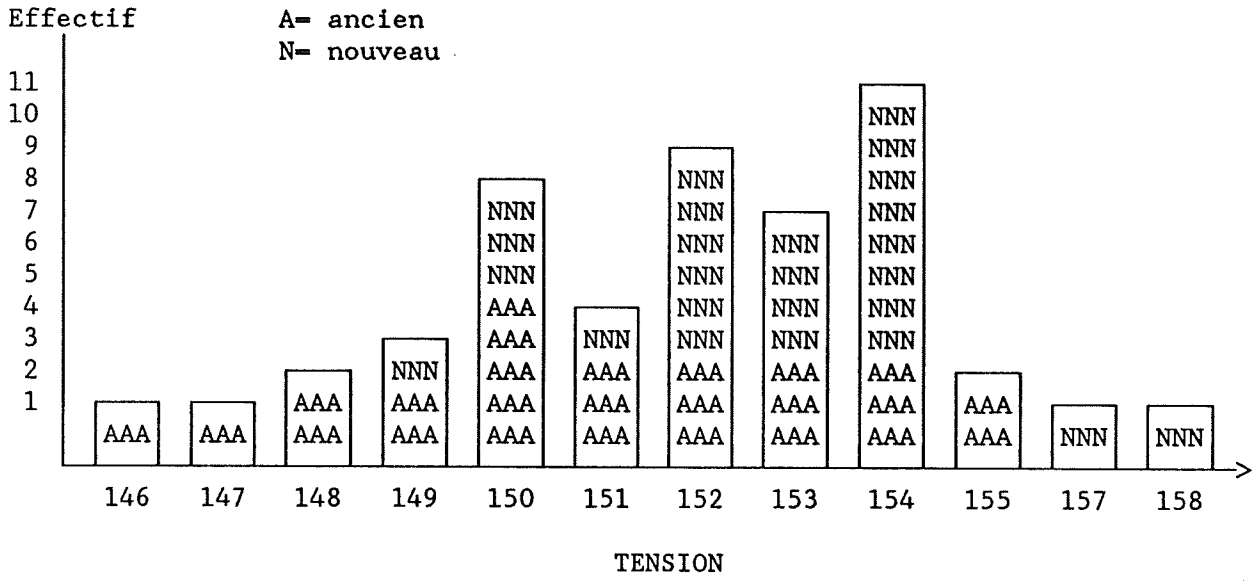


Figure 1.1: données de la tension de rupture

La comparaison sera faite par l'intermédiaire de la différence entre les moyennes arithmétiques. Les calculs donnent une tension moyenne de 151.08 pour le fil ancien et de 152.76 pour le nouveau fil. La différence entre les moyennes est de 152.76 - 151.08 = 1.68

Cette différence est-elle suffisante pour conclure que le nouveau fil a une tension de rupture supérieure à celle de l'ancien fil?

Comment élaborer un critère de décision qui tienne compte:

- de la dispersion dans les données de chaque type de fil
- du nombre d'observations dans chaque série de données
- des erreurs de décision pour:

  . déclarer des tensions de rupture différentes entre les deux fils alors qu'il n'y en a pas

  . ou déclarer que les tensions sont les mêmes alors qu'elles sont différentes.

Comment fait-on pour déterminer le nombre d'observations dans des études tout en tenant compte de la dispersion dans les données et les erreurs de décisions?

1.2 NOMBRE DE PANNES

Le tableau suivant est le résultat d'une compilation du nombre de pannes selon trois équipes de travail identifiées J (jour), S (soir) et N (nuit) travaillant sur des machines de type A,B,C,D.

nombre de pannes

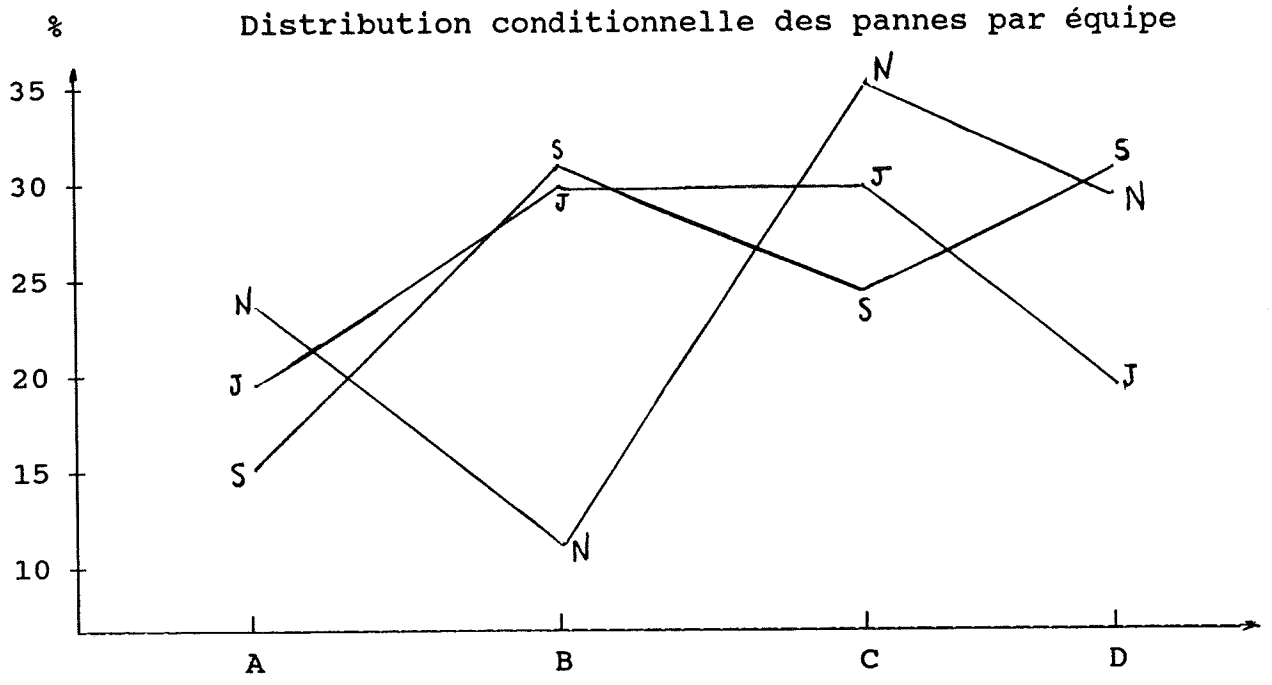
	machine				
équipe	A	B	C	D	total
J	10	15	15	10	50
S	10	20	15	20	65
N	20	10	30	25	85
total	40	45	60	55	200

La distribution des pannes est-elle différente d'une équipe de travail à l'autre?

La distribution des pannes est-elle différente d'une machine à l'autre?

Y a t-il une liaison entre la variable équipe et la variable machine?

Comment évaluer l'assurance de nos conclusions si les données représentent un échantillon de 200 pannes choisies au hasard parmi toutes les pannes?



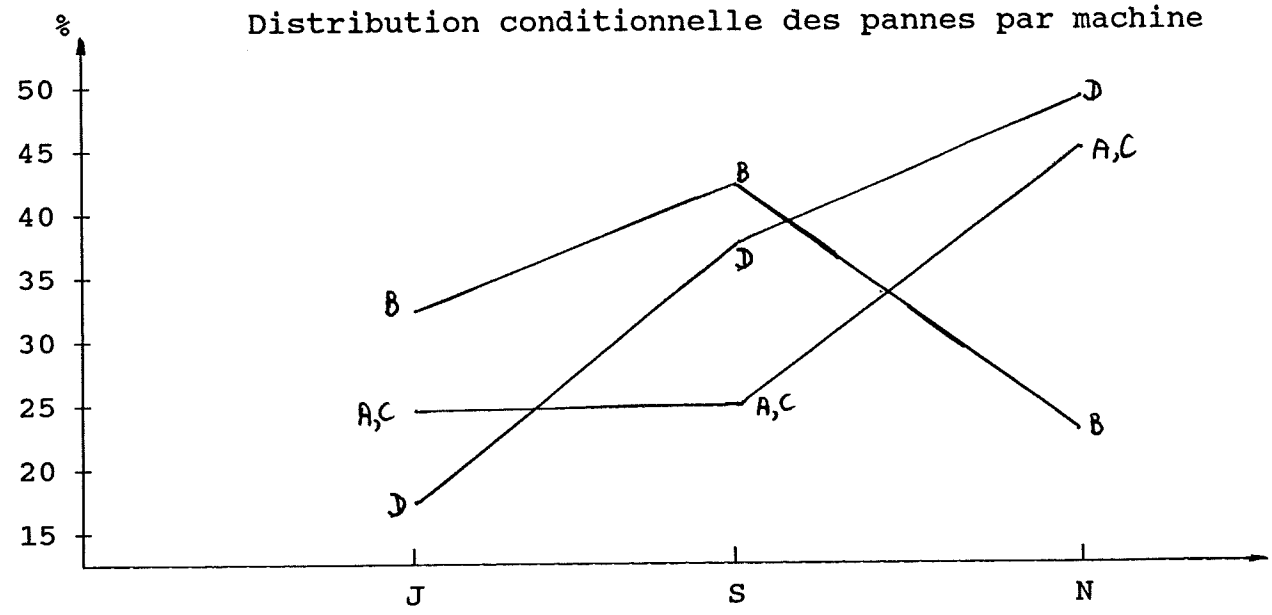


Figure 1.2: distribution des pannes

1.3 CONSOMMATION D'EAU

On veut lier la consommation d'eau potable pour des municipalités résidentielles avec le nombre d'habitants de la ville. On dispose de peu de données, soit

X	5682	11338	36464	16380	40922	25462
Y	2.44	6.14	39.53	10.32	46.04	21.39

où X représente le nombre d'habitants de la ville
Y représente la consommation en millions de mètres cu

Les données sont représentées par un graphique à la figure 1.3.

. Peut-on résumer ces données par une équation $Y = f(X)$?

. Quelle forme d'équation f doit-on choisir?

. Si on choisit une équation de la forme $Y = \beta_0 + \beta_1 * X$, comment estimer les paramètres β_0 et β_1?
Quelle est la précision des estimations obtenues?

. Quelle est la précision de l'équation calculée?

. Comment évaluer l'assurance de nos conclusions?

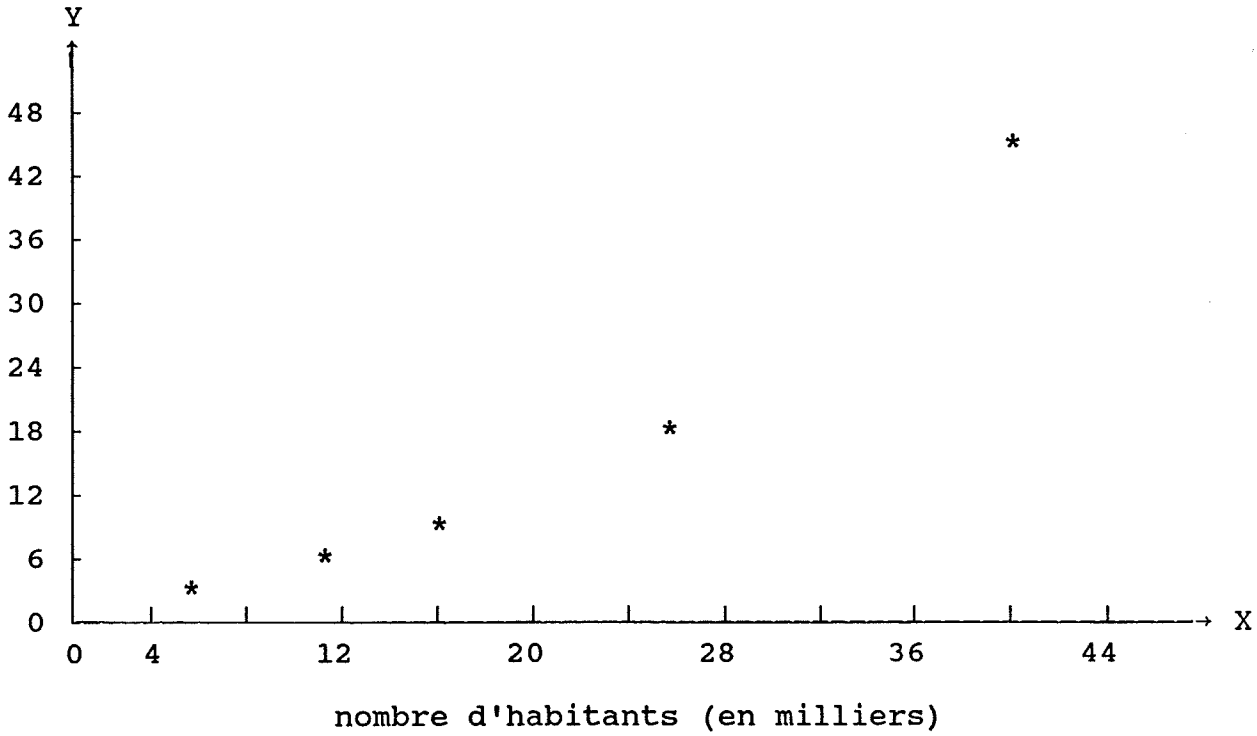


Figure 1.3: consommation d'eau en fonction du nombre d'habitants

1.4 COMPOSITION DU BÉTON

Le tableau de données a été constitué pour étudier l'influence de la composition du béton sur la chaleur dégagée (cal/gr) 28 jours après la prise.

X1	X2	X3	X4	Y
7	26	6	60	78.5
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

où X1 = % de $3C_AO.AL_2O_3$
X2 = % de $3C_AO.SIO_2$
X3 = % de $4C_AO.AL_2.F_{E2}O_3$
X4 = % de $2C_AO.S_IO_2$
Y = chaleur dégagée

Peut-on déterminer une équation de prédiction

$$Y = f(X1, X2, X3, X4)$$

reliant la chaleur dégagée Y avec la composition X1, X2, X3, X4 du béton?

Quelle forme de fonction f doit-on choisir?

Pour une équation de la forme

$$Y = \beta_0 + \beta_1 * X1 + \beta_2 * X2 + \beta_3 * X3 + \beta_4 * X4$$

Comment estimer les paramètres β_0, β_1, β_2, β_3, β_4?

Quelle est la précision des paramètres estimés?

Quelle est la valeur de prédiction de l'équation?

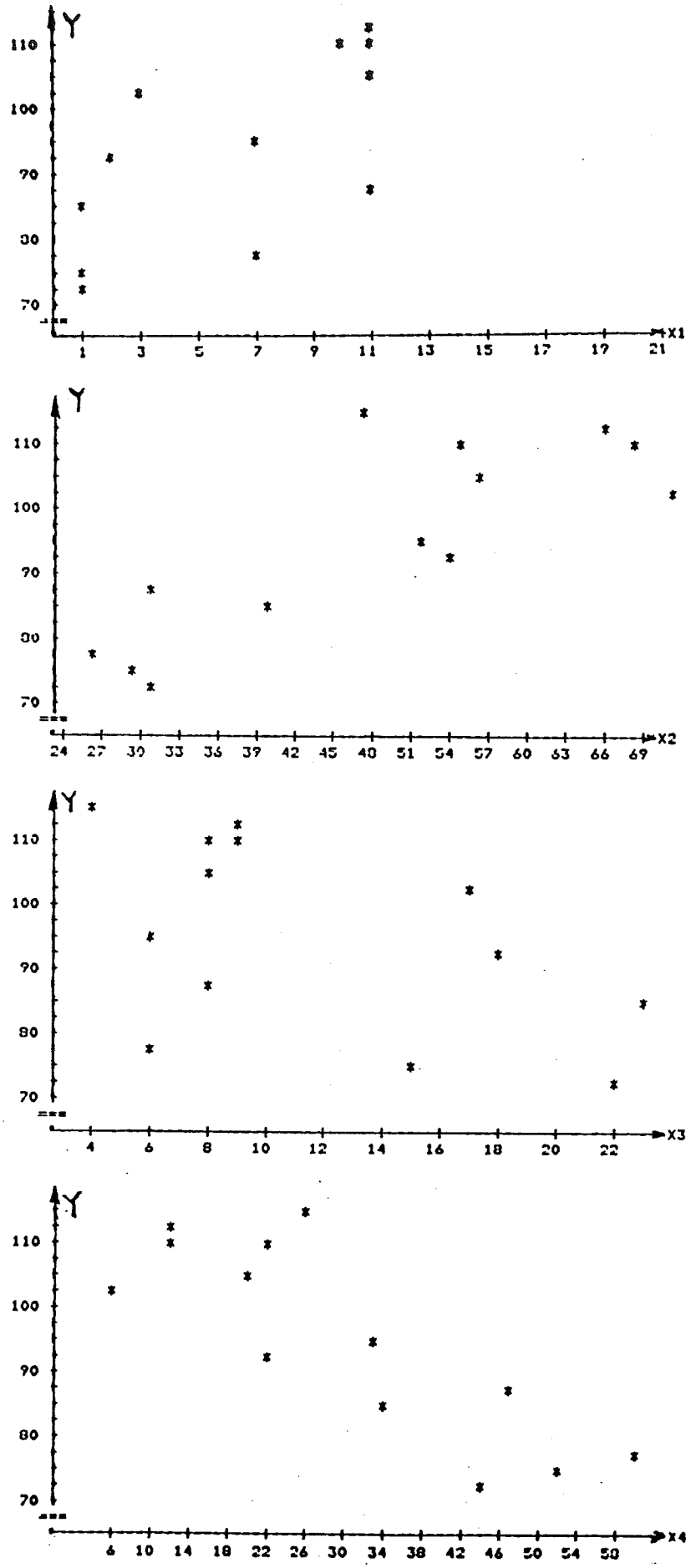


Figure 1.4: diagramme de dispersion conjointe

1.5 USURE DE PISTON

On veut comparer l'usure de quatre marques de piston identifiés A, B, C, D utilisées avec cinq types d'huile 1, 2, 3, 4, 5. On a mesuré la perte de poids (grammes) du piston après 10 heures de fonctionnement et constitué le tableau de données et calculé les moyennes par marque et par type d'huile.

	type d'huile					
marque	1	2	3	4	5	moyenne
A	1.641	1.782	1.570	1.493	1.672	1.6316
B	1.306	1.568	1.240	1.415	1.291	1.3640
C	1.149	1.223	1.068	1.118	1.004	1.1124
D	1.025	1.919	1.982	1.812	2.015	1.7506
moyenne	1.2802	1.6230	1.4650	1.4595	1.4955	1.46465

Y a-t-il une différence significative entre les marques?

Y a-t-il une différence significative entre les huiles?

Comment élaborer un critère de décision qui tienne compte de la dispersion des données, du nombre d'observations et des erreurs de décision?

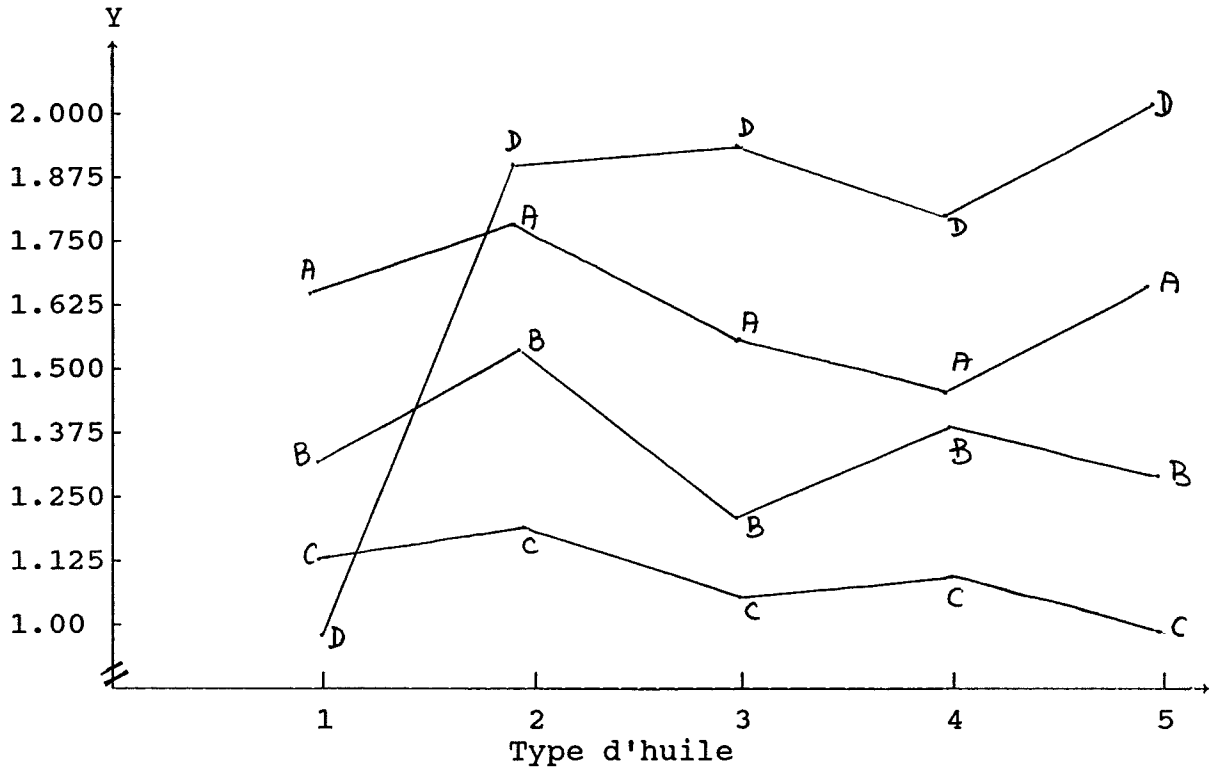


Figure 1.5: usure de piston

1.6 CARTES DE CONTRÔLE DE QUALITÉ

Le contrôle de la qualité est souvent la responsabilité de l'ingénieur et, parmi les techniques de ce champ d'activité, il y a le contrôle statistique des fabrications au moyen de cartes de contrôle. Celles-ci permettent d'établir si un procédé de fabrication est sous contrôle, c'est à dire si les variations de certaines caractéristiques se situent à l'intérieur de certaines limites dites de contrôle.

Dans cet exemple on a mesuré le poids de 4 contenants à 50 intervalles réguliers et on a tracé deux cartes de contrôle présentées à la figure 1.6: la première, pour la moyenne de poids et la deuxième, pour la variation de poids.

#	X_1	X_2	X_3	X_4	$\bar{X}$	R	#	X_1	X_2	X_3	X_4	$\bar{X}$	R
1	476	478	473	459	472	19	26	450	441	444	443	444	9
2	485	454	456	454	462	31	27	454	451	455	460	455	9
3	451	452	458	473	458	22	28	456	463	-	445	455	18
4	465	492	482	467	476	27	29	447	446	431	433	439	16
5	469	461	452	465	462	17	30	447	443	438	453	445	15
6	459	485	447	460	463	38	31	440	454	459	470	456	30
7	450	463	488	455	464	38	32	480	472	475	472	475	8
8	461	478	464	441	461	37	33	449	451	453	453	454	14
9	456	458	439	448	450	19	34	454	455	452	447	452	8
10	459	462	495	500	479	41	35	474	467	477	451	467	26
11	443	453	457	458	453	15	36	459	457	465	444	456	21
12	470	450	478	471	467	28	37	465	475	456	468	466	19
13	457	456	460	457	458	4	38	458	450	451	452	452	8
14	434	424	428	438	431	14	39	447	417	449	445	440	32
15	460	444	450	463	454	19	40	453	442	456	453	451	14
16	467	476	485	474	476	18	41	471	467	461	455	464	16
17	471	469	487	476	476	18	42	462	454	462	468	462	14
18	473	452	449	449	456	24	43	474	471	471	463	470	11
19	477	511	495	508	498	34	44	461	454	468	452	459	16
20	458	437	452	447	448	21	45	473	453	465	475	466	22
21	427	443	457	485	453	58	46	474	455	486	490	476	35
22	491	473	466	459	470	32	47	466	471	482	474	473	16
23	471	472	472	481	474	10	48	447	454	476	486	466	39
24	443	460	462	479	461	36	49	473	488	482	475	480	15
25	461	476	478	454	467	24	50	460	450	461	445	454	16

où $\bar{X} = (X_1 + X_2 + X_3 + X_4)/4$ $\quad R = X_{max} - X_{min}$

$X_{max} = MAX(X_1, X_2, X_3, X_4)$ $\quad X_{min} = MIN(X_1, X_2, X_3, X_4)$

Comment établir des cartes de contrôle de qualité présentées à la figure 1.6?

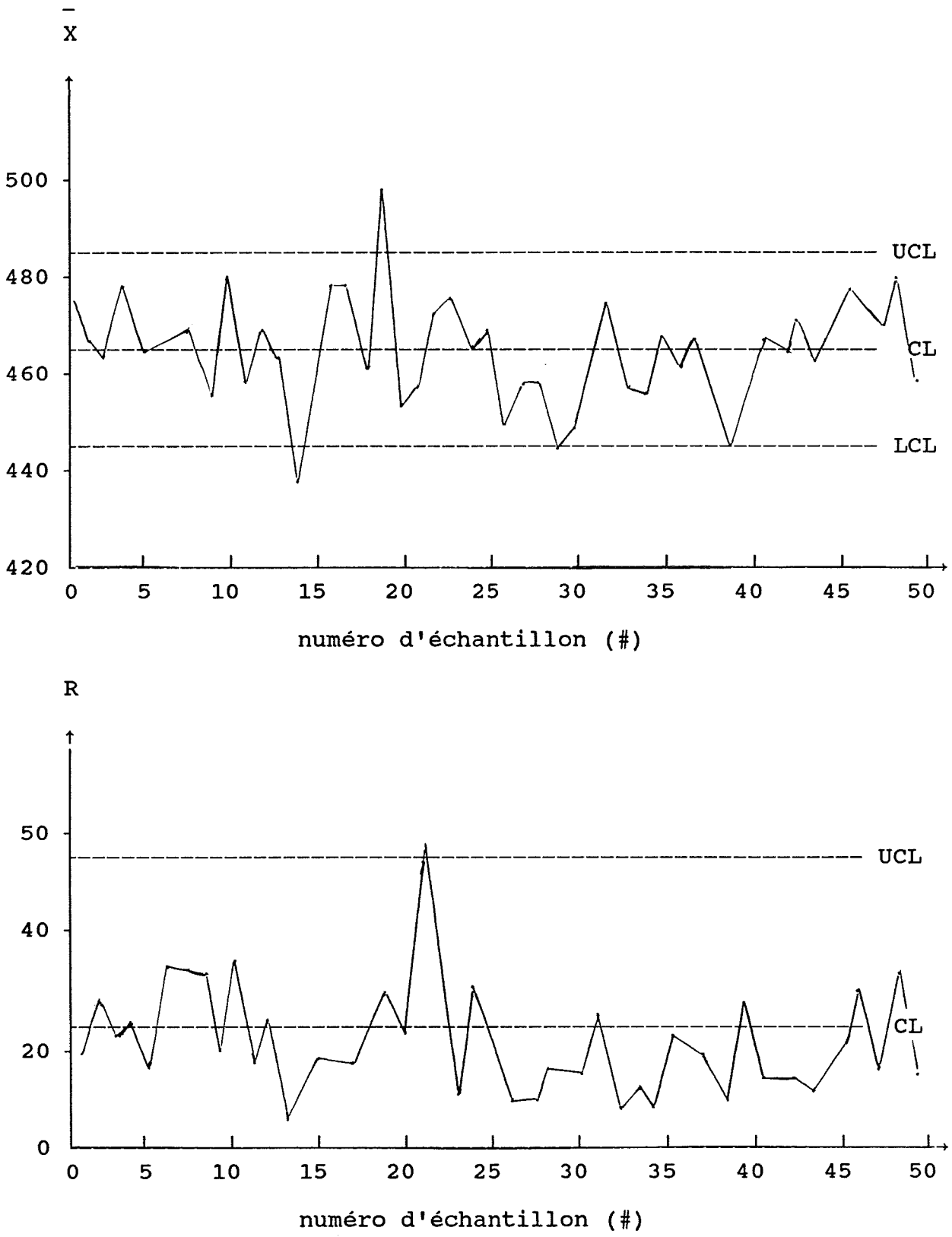


Figure 1.6: cartes de contrôle de qualité

1.7 CONTRÔLE DE QUALITÉ PAR ÉCHANTILLONNAGE

La base de cette technique repose sur l'idée que l'on peut contrôler la qualité d'un lot d'objets par l'inspection d'une partie du lot seulement. Le plus souvent, la méthode d'échantillonnage est la seule possible, soit à cause des coûts d'inspection ou encore parce que l'examen des objets est destructif.

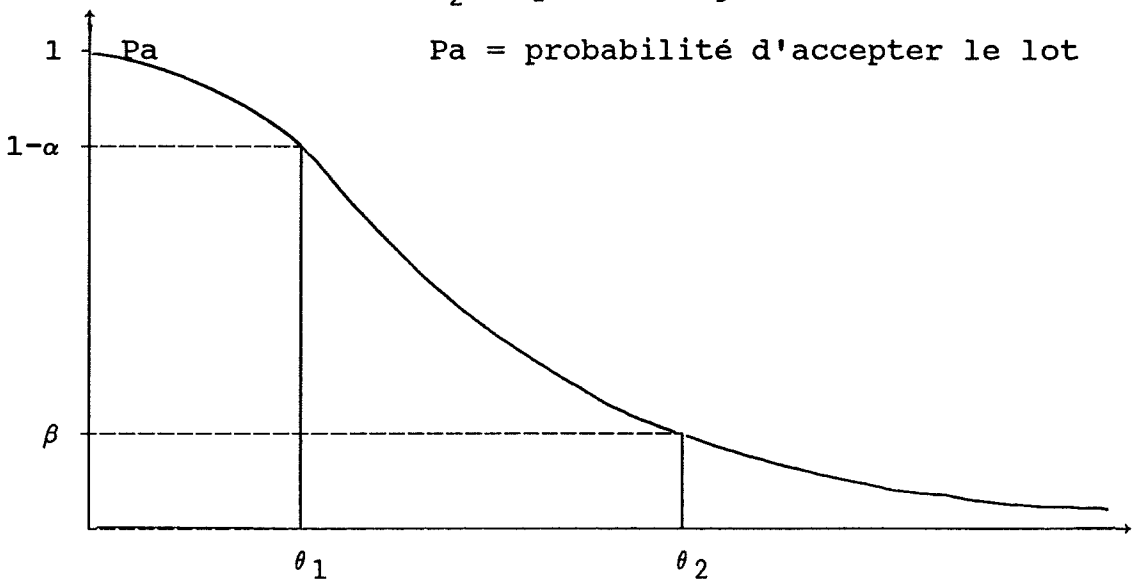
Un lot de N (connu) objets contient M (inconnu) objets défectueux. Sur la base d'un échantillon de n (à déterminer) objets choisis au hasard on doit décider d'accepter ou rejeter le lot selon le critère suivant: si le nombre d'objets défectueux est inférieur ou égal à c (à déterminer) il est accepté, sinon il est refusé.

Un plan d'échantillonnage est défini par le triplet (N, n, c) ainsi que la donnée de deux points sur la COURBE CARACTÉRISTIQUE du plan. Cette courbe donne la probabilité d'acceptation du lot en fonction de la proportion véritable (θ = M/N) d'objets défectueux dans le lot. L'allure d'une telle courbe est présentée sur la figure 1.7. Les deux points choisis sont

$(\theta_1, 1 - \alpha)$ où α = risque du producteur
θ_1 = qualité acceptable

(θ_2, β) où β = risque du consommateur
θ_2 = qualité rejetable

Pa = probabilité d'accepter le lot



Comment déterminer le plan d'échantillonnage (n, c) en spécifiant les deux points $(\theta_1, 1 - \alpha)$ et (θ_2, β)?

Figure 1.7: courbe caractéristique d'un plan d'échantillonnage

1.8 ANALYSE D'UN SONDAGE

Une maison de sondage a conduit une étude sur la popularité de trois candidats dans une élection qui sera tenue dans les prochaines semaines. On a fait des interviews téléphoniques sur un échantillon de 100 personnes choisies scientifiquement selon une méthode d'échantillonnage. En plus de leur demander leur intentions de vote, on a retenu les informations suivantes pour chaque répondant: âge, langue, sexe.

Une première analyse de la popularité des 3 candidats a donné:

candidat	effectif	pourcentage
1	31	31
2	39	39
3	25	25
indécis	5	5

La différence de 8% entre le candidat 1 et le candidat 2 est-elle vraiement significative si l'on tient compte des aléas de l'échantillonage?

Comment tenir compte des variables âge, langue et sexe sur les intentions de vote?

Comment déterminer le nombre d'observations à prendre dans un sondage?

Comment évaluer la qualité de nos conclusions à l'ensemble de tous les voteurs?

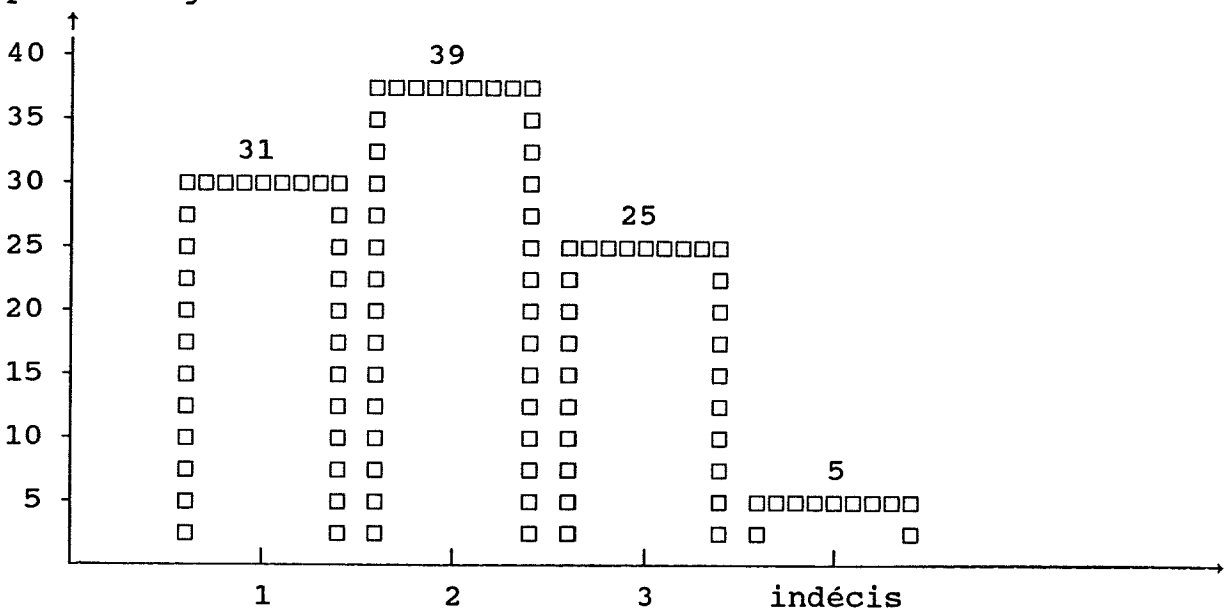


Figure 1.8: distribution des intentions de vote

1.9 QUELQUES ÉLÉMENTS COMMUNS DE CES EXEMPLES

L'examen de ces exemples permet de dégager les éléments suivants:

des données: généralement elles constituent une petite partie d'un ensemble beaucoup plus vaste appelé POPULATION; typiquement le nombre d'observations est petit par rapport à la taille de la population. Les données présentent de la VARIABILITÉ, attribuable au procédé de fabrication, au mesurage et à l'échantillonnage.

des variables: elles se distinguent par

- leur type: qualitative ou numérique
- leur rôle: à expliquer ou explicative

des objectifs:

réduction:	décrire et résumer les données par des méthodes numériques et graphiques
comparaison:	entre deux ou plusieurs groupes pour décider s'ils présentent des différences réelles compte tenu de la variabilité
liaison:	entre deux variables quantitatives ou deux variables qualitatives
équation:	entre une variable quantitative et une ou plusieurs variables quantitatives
décision:	prise dans des conditions d'incertitude à cause de l'échantillonnage et de la variabilité des données.

1.10 CLASSIFICATION DES VARIABLES

Le terme VARIABLE est employé pour indiquer une propriété ou une caractéristique que l'on peut constater ou compter ou mesurer.

Par exemple on peut

- constater la couleur d'un objet
- constater si l'objet est défectueux ou non
- compter le nombre de défauts de l'objet
- mesurer le poids de l'objet

On peut classer les variables de plusieurs façons selon le point de vue adopté et le rôle qu'elles jouent dans les analyses. On a aussi proposé plusieurs d'autres classifications selon l'échelle de mesure, les valeurs qu'elles peuvent prendre.

types de classification

échelle de mesure	valeur	rôle
nominale	qualitative	explicative
ordinale	entière	expliquée
intervalle	réelle	
rapport		

Nous ne donnerons pas une définition formelle de chacun de ces termes mais plutôt des exemples puisés parmi les situations exposées au début du chapitre.

Variable qualitative nominale

- marque de piston
- nom de compagnie
- catégorie socio-professionnelle
- lieu d'habitation
- type d'huile
- langue

Variable qualitative ordinale

- classe d'âge
- rang d'un vin dans une classification

Variable quantitative intervalle

- température
- dates de calendrier

Variable quantitative ratio

- usure d'un piston
- % de $3C_AO.AL_2O_3$
- nombre de pannes
- teneur d'un minerai
- revenu
- poids d'un objet
- nombre d'habitants dans une ville

	QUALITATIVE		QUANTITATIVE	
	nominale	ordinale	discrète	continue
exemple	couleur objet	rang dans classification	nombre de défauts objet	poids objet
action	identification	classement	comptage	mesurage
opérations mathématiques	non	non	oui	oui

Variable continue

Une variable est continue si elle peut prendre toutes les valeurs entre deux nombres réels. Par exemple la taille d'un individu peut être 1.72 ou 1.73 mais toutes les valeurs intermédiaires comme 1.7246 sont possibles. D'un point de vue pratique toutes les variables continues sont mesurées avec un certain niveau de discrétisation à cause de la précision des appareils de mesure.

- tension de rupture
- usure d'un piston
- âge d'un individu

Variable discrète

Une variable est discrète si elle n'est pas continue; par exemple des comptages définissent des variables discrètes car seuls les nombres 0, 1, 2, ... sont des valeurs possibles.

- nombre de pannes
- nombre d'habitants

Variable expliquée / Variable explicative

L'analyse des données dispose d'un grand nombre de méthodes et elles sont en partie dictées par

- la nature des variables
- le nombre de variables
- le rôle des variables

Une variable est dite EXPLICATIVE si elle sert à en expliquer une autre appelée variable EXPLIQUÉE. Les termes de variables INDÉPENDANTES et variables DÉPENDANTES sont aussi employés mais nous réservons cette terminologie pour la théorie des probabilités.

exemple	variable expliquée	variable explicative
fil	force de rupture	type de fil
sondage	popularité candidat	.âge répondant .langue du répondant
piston	usure du piston	.marque du piston .type d'huile
eau	consommation d'eau	nombre d'habitants
béton	force de compression	composition du béton

1.11 CODAGES NUMÉRIQUES

Pour des fins de traitement informatique des données, on code souvent numériquement des informations qualitatives. Par exemple, les réponses possibles d'une question pourraient être codées de la manière suivante:

1 = oui 2 = non 3 = ne sait pas

mais tout autre codage numérique peut être utilisé.

Variables indicatrices

Il s'agit de variables nominales à deux modalités qui indique la présence ou l'absence d'une propriété. Par exemple si l'objet examiné est défectueux ou non, ou encore, si la personne interrogée est un homme ou une femme. Il est utile de coder une telle variable de la façon suivante:

1 = présence de la propriété ou caractéristique
0 = absence de la propriété ou caractéristique

Par exemple

1 = objet est non-défectueux
0 = objet est défectueux

le codage 0-1 est pratique car si on observe une série de valeurs $x_1, x_2, x_3, \ldots, x_n$ on a

$$\sum_{\alpha=1}^{n} x_\alpha = \text{nombre de valeurs égales à 1}$$

Variables multinominales

Il s'agit de variables nominales ayant trois modalités et plus. Par exemple, on pourrait coder la religion d'une personne interrogée de la façon suivante:

1 = catholique 2 = protestante 3 = juive

4 = musulmane 5 = autre

Cette codification numérique est arbitraire. Il est préférable de coder en introduisant autant de variables indicatrices que le nombre de modalités de la variable qualitative. Par exemple

	variables indicatrices				
religion	I_1	I_2	I_3	I_4	I_5
catholique	1	0	0	0	0
protestante	0	1	0	0	0
juive	0	0	1	0	0
musulmane	0	0	0	1	0
autre	0	0	0	0	1

En additionnant l'ensemble des valeurs d'une variable indicatrice, on obtient des comptages. Par exemple

$\sum_{\alpha=1}^{n} I_{1\alpha}$ = nombre de personnes interrogées de religion catholique

et $\frac{1}{n} \sum_{\alpha=1}^{n} I_{1\alpha}$ = proportion de personnes interrogées de religion catholique

Notons que le codage proposé introduit une relation linéaire entre les variables indicatrices car

$$I_1 + I_2 + I_3 + I_4 + I_5 = 1$$

Le passage d'une variable multimodale à un ensemble de variables indicatrices associées à chacune des modalités s'appelle passage à la FORME DISJONCTIVE COMPLÈTE.

Le codage et l'utilisation des variables indicatrices est aussi utile pour les variables continues lors de la construction des tableaux d'effectifs qui seront expliqués au chapitre 2.

CHAPITRE 2

ANALYSE STATISTIQUE DESCRIPTIVE

2.0 SOMMAIRE

Le premier objectif de l'analyse des données est la description des données en vue d'en extraire les caractéristiques essentielles et l'information pertinente. Cet objectif peut être une fin en soi ou être une étape en vue d'une prise de décision. On distingue les:

Indicateurs numériques

Paramètres descriptifs pour extraire les indicateurs de:

- tendance centrale
- dispersion
- forme
- décision

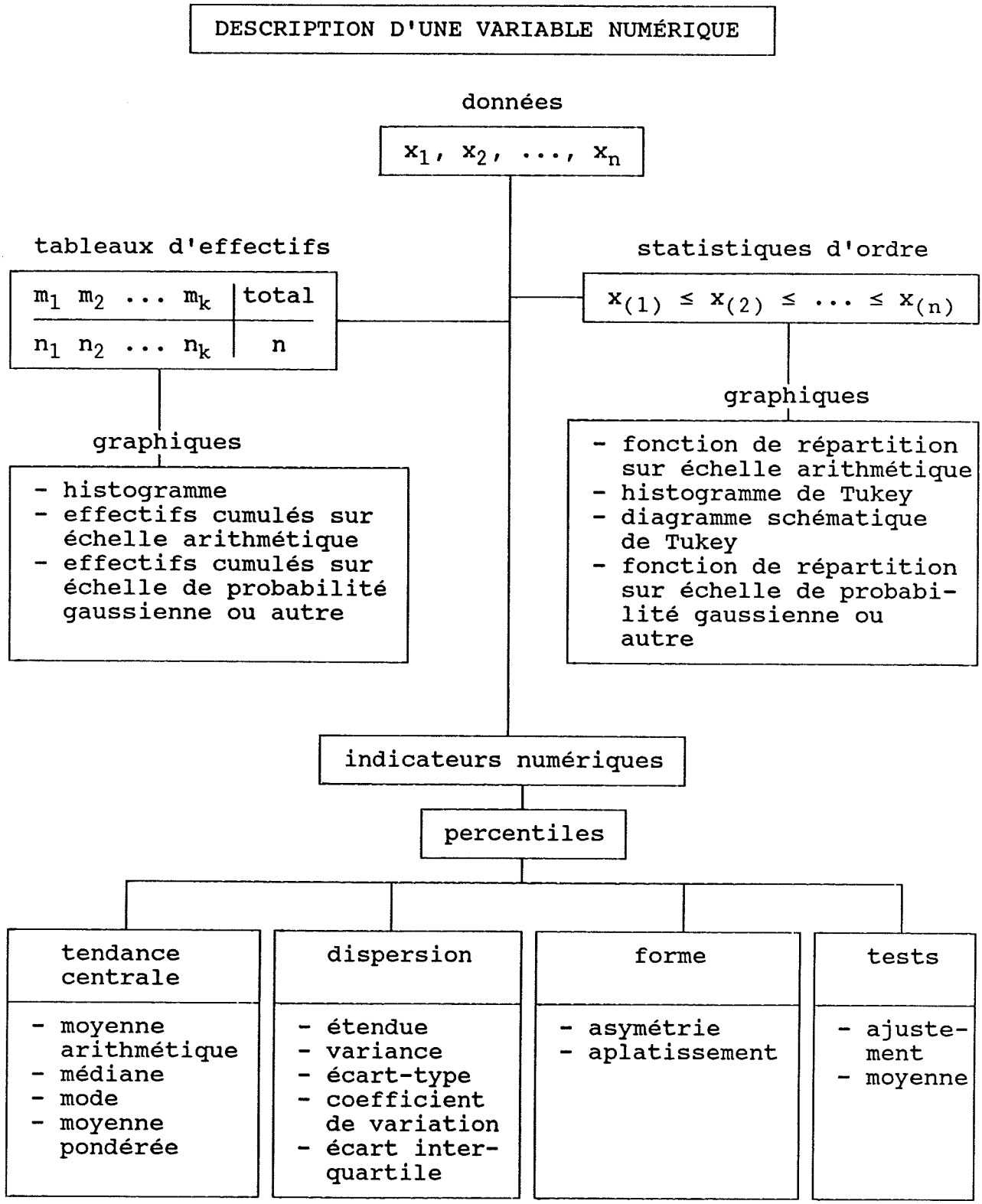
Descripteurs graphiques

Graphiques permettant de visualiser les données:

- histogramme
- histogramme de Tukey
- diagramme schématique
- effectifs cumulés sur échelle arithmétique et sur échelle de probabilité gaussienne

Le tableau de la page suivante donne une vue d'ensemble des opérations pour décrire une variable numérique.

Tableau 2.1:


DESCRIPTION D'UNE VARIABLE NUMÉRIQUE
données
$x_1, x_2, \ldots, x_n$
tableaux d'effectifs
m_1 m_2 ... m_k | total
n_1 n_2 ... n_k | n
graphiques
- histogramme
- effectifs cumulés sur échelle arithmétique
- effectifs cumulés sur échelle de probabilité gaussienne ou autre
statistiques d'ordre
$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$
graphiques
- fonction de répartition sur échelle arithmétique
- histogramme de Tukey
- diagramme schématique de Tukey
- fonction de répartition sur échelle de probabilité gaussienne ou autre
indicateurs numériques
percentiles
tendance centrale
- moyenne arithmétique
- médiane
- mode
- moyenne pondérée
dispersion
- étendue
- variance
- écart-type
- coefficient de variation
- écart inter-quartile
forme
- asymétrie
- aplatissement
tests
- ajuste-ment
- moyenne

2.1 DESCRIPTION D'UNE VARIABLE NUMÉRIQUE

Données: $x_1, x_2, \ldots, x_n$

représentent une série de valeurs numériques d'une variable quantitative. Sous des conditions précices, les valeurs peuvent représenter une réalisation de l'observation d'un ensemble plus vaste appelé la POPULATION. Dans ce cas l'ensemble des valeurs est appelé un ÉCHANTILLON de taille n .

Exemple 2.1: les données représentent 165 mesures de la force (psi) de compression d'un certain type de béton.

4855	3705	3125	3120	4125	4410	4415	4580	3680	4670
3490	3815	4315	3545	3695	3675	3705	4450	4410	4810
4315	4155	4095	4225	3565	4135	4585	4330	4315	3865
3465	3305	3740	3670	4375	4645	5200	4100	4090	4015
4170	4510	4730	4330	3070	3260	3870	3870	3530	4150
3235	3040	3620	4395	4150	5005	3805	3600	4135	3290
4080	3310	4345	3920	3825	4480	4540	4580	3730	4700
4060	3570	4335	3410	4085	4290	4320	4495	4195	4195
4080	4120	4535	4030	4660	4010	3320	4050	3360	3535
3450	4330	3725	4020	4150	3940	3605	4000	4315	3280
4200	3375	4055	3670	4900	4390	4015	4505	4890	4190
3390	4030	4610	4345	4070	4505	4485	4290	3980	5110
4320	4335	3760	4020	4540	4015	3985	4140	4115	4475
3800	4110	4245	4320	3720	3845	3920	3660	4600	3935
3320	4210	3610	4125	4590	3985	3590	3585	3995	3890
4015	4490	3195	3365	3890	3975	4095	4540	5420	5200
3885	4270	4010	4730	4565					

Statistiques d'ordre: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

est la série des données en ordre croissant

$x_{(1)}$ est la plus petite valeur (minimum)
$x_{(2)}$ est la deuxième plus petite valeur
.
.
.
$x_{(n)}$ est la plus grande valeur (maximum)

Fonction de répartition échantillonnalle: $F_n(x)$ définie par

$$F_n(x) = \begin{cases} 0 & x < x_{(1)} \\ \dfrac{\alpha}{n} & x_{(\alpha)} \leq x < x_{(\alpha+1)} \quad \alpha = 1,2,\ldots,n-1 \\ 1 & x_{(n)} \leq x \end{cases} \qquad (2.1)$$

Cette fonction est de type escalier avec des discontinuités (sauts) de 1/n à chacune des valeurs $x_{(\alpha)}$.

Exemple (suite):

Valeur	Nombre	Rang	$F_n(x)$	Valeur	Nombre	Rang	$F_n(x)$
3040	1	1	0.006	3670	2	35	0.212
3070	1	2	0.012	3675	1	36	0.218
3120	1	3	0.018	3680	1	37	0.224
3125	1	4	0.024	3695	1	38	0.230
3195	1	5	0.030	3705	2	40	0.242
3235	1	6	0.036	3720	1	41	0.248
3260	1	7	0.042	3725	1	42	0.255
3280	1	8	0.048	3730	1	43	0.261
3290	1	9	0.055	3740	1	44	0.267
3305	1	10	0.061	3760	1	45	0.273
3310	1	11	0.067	3800	1	46	0.279
3320	2	13	0.079	3805	1	47	0.285
3360	1	14	0.085	3815	1	48	0.291
3365	1	15	0.091	3825	1	49	0.297
3375	1	16	0.097	3845	1	50	0.303
3390	1	17	0.103	3865	1	51	0.309
3410	1	18	0.109	3870	2	53	0.321
3450	1	19	0.115	3885	1	54	0.327
3465	1	20	0.121	3890	2	56	0.339
3490	1	21	0.127	3920	2	58	0.352
3530	1	22	0.133	3935	1	59	0.358
3535	1	23	0.139	3940	1	60	0.364
3545	1	24	0.145	3975	1	61	0.370
3565	1	25	0.152	3980	1	62	0.376
3570	1	26	0.158	3985	2	64	0.388
3585	1	27	0.164	3995	1	65	0.394
3590	1	28	0.170	4000	1	66	0.400
3600	1	29	0.176	4010	2	68	0.412
3605	1	30	0.182	4015	4	72	0.436
3610	1	31	0.188	4020	2	74	0.448
3620	1	32	0.194	4030	2	76	0.461
3660	1	33	0.200	4050	1	77	0.467

suite

valeur	nombre	rang	$F_n(x)$	valeur	nombre	rang	$F_n(x)$
4055	1	78	0.473	4395	1	127	0.770
4060	1	79	0.479	4410	2	129	0.782
4070	1	80	0.485	4415	1	130	0.788
4080	2	82	0.497	4450	1	131	0.794
4085	1	83	0.503	4475	1	132	0.800
4090	1	84	0.509	4480	1	133	0.806
4095	2	86	0.521	4485	1	134	0.812
4100	1	87	0.527	4490	1	135	0.818
4110	1	88	0.533	4495	1	136	0.824
4115	1	89	0.539	4505	2	138	0.836
4120	1	90	0.545	4510	1	139	0.842
4125	2	92	0.558	4535	1	140	0.848
4135	2	94	0.570	4540	3	143	0.867
4140	1	95	0.576	4565	1	144	0.873
4150	3	98	0.594	4580	2	146	0.885
4155	1	99	0.600	4585	1	147	0.891
4170	1	100	0.606	4590	1	148	0.897
4190	1	101	0.612	4600	1	149	0.903
4195	2	103	0.624	4610	1	150	0.909
4200	1	104	0.630	4645	1	151	0.915
4210	1	105	0.636	4660	1	152	0.921
4225	1	106	0.642	4670	1	153	0.927
4245	1	107	0.648	4700	1	154	0.933
4270	1	108	0.655	4730	2	156	0.945
4290	2	110	0.667	4810	1	157	0.952
4315	4	114	0.691	4855	1	158	0.958
4320	3	117	0.709	4890	1	159	0.964
4330	3	120	0.727	4900	1	160	0.970
4335	2	122	0.739	5005	1	161	0.976
4345	2	124	0.752	5110	1	162	0.982
4375	1	125	0.758	5200	2	164	0.994
4390	1	126	0.764	5420	1	165	1.000

Remarque: On peut définir un concept de rang qui tient compte des valeurs ex aequo dans une série. Par exemple, la valeur 4080 apparaît deux fois et correspond aux rangs 81 et 82. On pourrait donc, si on le désire, lui attribué le rang moyen de 81.5.

Percentile (quantile, fractile)

Le t-ième PERCENTILE où $(0 < t < 100)$ d'une série de valeurs $x_1, x_2, \ldots, x_n$ est une valeur y qui a t% des observations plus petites à elle et (1-t) % des observations plus grandes. Cette définition peut se formaliser de la manière suivante.

Soit $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

les statistiques d'ordre, et posons

$$p = \frac{t}{100} \quad , \quad np = j + f$$

où $j = [np]$ est la partie entière de np

$f = np - [np]$ est la partie fractionnaire de np

Le t-ième percentile y est défini par:

$$y = (1-\omega)\, x_{(j)} + \omega\, x_{(j+1)} \qquad (2.2)$$

$$= x_{(j)} + \omega\, (x_{(j+1)} - x_{(j)})$$

où ω est un poids entre 0 et 1, $0 \leq \omega \leq 1$

Le t-ième percentile est une valeur interpolée entre $x_{(j)}$ et $x_{(j+1)}$. On a proposé plusieurs choix pour le poids ω:

Choix	Poids	Interprétation
1	$\omega = f$	moyenne pondérée visant $x_{(j)}$ ($x_{(0)} = x_{(1)}$ par convention)
2	$\omega = \begin{cases} 0 & \text{si } f \leq 0.5 \\ 1 & \text{si } f \geq 0.5 \end{cases}$	valeur la plus près de $x_{(j)}$ ou $x_{(j+1)}$
3	$\omega = \begin{cases} 0 & \text{si } f = 0 \\ 1 & \text{si } f > 0 \end{cases}$	fonction de répartition échantillonnale
4	$\omega = (n+1)p - [(n+1)p]$ et $j = [(n+1)p]$	même que le choix 1 avec (n+1) plutôt que n
5	$\omega = \begin{cases} 1/2 & \text{si } f = 0 \\ 1 & \text{si } f > 0 \end{cases}$	fonction de répartition échantillonnale modifiée
6	$\omega = \begin{cases} 1/2 & \text{si } f = 0.5 \\ 1 & \text{si } f = 0 \end{cases}$	pour faire coincider la définition du 50ième percentile avec celle de médiane.

Le choix d'un ω donnera des valeurs légèrement différentes pour le t-ième percentile y. D'autre part certains percentiles portent des noms particuliers. Par exemple si

t = 25	y s'appelle le PREMIER QUARTILE et il est généralement noté par Q_1
t = 50	y s'appelle le DEUXIÈME QUARTILE
t = 75	y s'appelle le TROISIÈME QUARTILE et il est généralement noté Q_3

Exemple 2.2: suite de l'exemple 2.1

On obtient les percentiles suivants avec chacun des choix de la fonction de poids ω:

t	20	37	50	97
Choix 1	3660	3975.25	4082.5	4905.25
Choix 2	3660	3975	4082.5	4900
Choix 3	3660	3980	4085	5005
Choix 4	3662	3977.1	4085	5007.1
Choix 5	3665	3980	4085	5110
Choix 6	3610	3980	4085	5005

Tableau d'effectifs (ou fréquences)

Le tableau d'effectifs (ou fréquences) est une construction par laquelle on substitue aux données brutes $x_1, x_2, \ldots, x_n$ une nouvelle série de valeurs

$$m_1, m_2, \ldots, m_k$$

Ces valeurs sont les points milieux d'une série d'intervalles contigus définis sur l'ensemble des valeurs. Chacune de ces valeurs m_α est affectée du nombre de valeurs dans chacun des intervalles correspondants. En général, on choisit des intervalles de longueurs égales. Le tableau d'effectifs se présente sous la forme suivante:

points milieux	m_1	m_2 ...	m_k	Total
effectifs	n_1	n_2 ...	n_k	n

où m_α: point milieu de l'intervalle $[a_\alpha, b_\alpha[$
n_α: nombre de valeurs (EFFECTIF) dans l'intervalle $[a_\alpha, b_\alpha[$

L'idée de base est de remplacer un grand nombre de valeurs par un petit nombre de valeurs équidistantes. Il s'agit en fait d'un codage de la variable et la perte d'information est largement compensée par une meilleure compréhension des données. De par sa construction même, on peut obtenir plusieurs tableaux différents selon le choix du nombre d'intervalles, la longueur des intervalles et la position du premier intervalle. Le nombre d'intervalles k est entre 5 et 20 selon le tableau suivant:

n	k
<20	-
20-50	5-7
50-100	6-10
100-200	7-12
>200	10-20

Quantités dérivées du tableau d'effectifs:

$$\sum_{\alpha=1}^{i} n_\alpha \quad : \text{effectif cumulatif,} \quad i = 1,2,\ldots,k$$

$$\frac{n_\alpha}{n} \quad : \text{fréquence ou fréquence relative} \qquad (2.3)$$

$$\frac{1}{n} \sum_{\alpha=1}^{i} n_\alpha \quad : \text{fréquence cumulative,} \quad i = 1,2,\ldots,k$$

Remarque: Le terme fréquence est aussi employé pour désigner un effectif; le contexte permet de discerner le sens exact du terme employé.

Exemple 2.3: suite de l'exemple 2.1

L'étendue entre la plus grande valeur 5420 et la plus petite valeur 3040 étant 2380 on a décidé de construire un tableau d'effectifs en utilisant des intervalles de longueur 200 en plaçant la limite supérieure du premier intervalle à 3100. Il faut alors 13 intervalles dont les points milieux seront situés aux valeurs 3000, 3200, ..., 5400 . On obtient alors le tableau d'effectifs 2.2

Tableau 2.2: tableau d'effectifs

POINT MILIEU (m_α)	EFFECTIF (n_α)	EFFECTIF CUMULATIF	POURCENTAGE	POURCENTAGE CUMULATIF
3000	2	2	1.2	1.2
3200	7	9	4.2	5.5
3400	12	21	7.3	12.7
3600	17	38	10.3	23.0
3800	18	56	10.9	33.9
4000	30	86	18.2	52.1
4200	24	110	14.5	66.7
4400	26	136	15.8	82.4
4600	17	153	10.3	92.7
4800	6	159	3.6	96.4
5000	2	161	1.2	97.6
5200	3	164	1.8	99.4
5400	1	165	0.6	100.0

2.2 INDICATEURS DE TENDANCE CENTRALE

On veut identifier le "centre" (tendance centrale) de l'ensemble des valeurs et dégager une valeur typique; plusieurs quantités ont été proposées:

Moyenne arithmétique: généralement notée $\bar{x}$ et définie par

$$\bar{x} = \frac{1}{n} \sum_{\alpha=1}^{n} x_\alpha \qquad (2.4)$$

Mode: valeur la plus fréquente; en général, le mode n'est pas unique.

Médiane: valeur au centre des statistiques d'ordre;

$$\tilde{x} = \begin{cases} x_{(k)} & k = (n+1)/2, \ n \text{ impair} \\ \dfrac{x_{(k)} + x_{(k+1)}}{2} & k = n/2, \quad n \text{ pair} \end{cases} \qquad (2.5)$$

La médiane correspond au 50ième percentile défini selon le choix 6.

Remarques:

(1) la moyenne est la plus employée
(2) le mode est très peu utilisé
(3) la médiane jouit d'une propriété de robustesse: elle demeure inchangée si on modifie les valeurs extrêmes. Par exemple, les valeurs

1, 3, 4, 7, 12

ont pour médiane $\tilde{x} = 4$ et pour moyenne $\bar{x} = 5.4$.

Par contre si on remplace 12 par 12 + c où c > 0 on a

$$\tilde{x} = 4 \qquad \text{et} \qquad \bar{x} = 5.4 + c/5$$

La médiane est donc utile lorsque l'on veut éliminer l'effet de valeurs extrèmes douteuses.

La médiane est la solution du problème d'extremum suivant:

$$\min_a \sum_{\alpha=1}^{n} |x_\alpha - a| = \sum_{\alpha=1}^{n} |x_\alpha - \tilde{x}|$$

Moyenne pondérée

On peut calculer d'autres moyennes appelées moyennes pondérées; elles sont définies par:

$$\sum_{\alpha=1}^{n} w_\alpha x_\alpha \tag{2.6}$$

où w_α représente un poids associé à x_α et vérifiant

$$w_\alpha \geq 0 , \qquad \sum_{\alpha=1}^{n} w_\alpha = 1$$

Cas particuliers:

(a) $w_\alpha = 1/n$, $\sum_{\alpha=1}^{n} w_\alpha x_\alpha = \bar{x}$ moyenne arithmétique

(b)
$$w_\alpha = \begin{cases} \dfrac{1}{n-2} & \alpha = 2, \ldots, n-1 \\ 0 & \alpha = 1, n \end{cases}$$

Alors
$$\sum_{\alpha=1}^{n} w_\alpha x_{(\alpha)} = \sum_{\alpha=2}^{n-1} \frac{1}{n-2} x_{(\alpha)}$$

est la moyenne des observations après élimination de la plus petite et de la plus grande valeur

(c)
$$w_\alpha = \begin{cases} 1 & \alpha = (n+1)/2 \ , \quad n \text{ impair} \\ 0 & \text{autrement} \end{cases}$$

$\sum_{\alpha=1}^{n} w_\alpha x_{(\alpha)} = \tilde{x}$ médiane d'un nombre impair de valeurs

(d)
$$w_\alpha = \begin{cases} 1/2 & \alpha = k, k+1 \qquad k = n/2 \ , \quad n \text{ pair} \\ 0 & \text{autrement} \end{cases}$$

$\sum_{\alpha=1}^{n} w_\alpha x_{(\alpha)} = \tilde{x}$ médiane d'un nombre pair de valeurs

(e) Soient deux séries de valeurs

$x_1, x_2, \ldots, x_{n1}$ de moyenne $\bar{x}_1 = \frac{1}{n_1} \sum_{\alpha=1}^{n_1} x_\alpha$

$x_1', x_2', \ldots, x_{n2}'$ de moyenne $\bar{x}_2 = \frac{1}{n_2} \sum_{\alpha=1}^{n_2} x_\alpha'$

alors la moyenne générale $\bar{x}$ de l'ensemble des deux séries est une moyenne pondérée:

En effet

$$\bar{x} = \left(\sum_{\alpha=1}^{n_1} x_\alpha + \sum_{\alpha=1}^{n_2} x_\alpha' \right) / (n_1 + n_2) = w_1 \bar{x}_1 + w_2 \bar{x}_2$$

où $w_1 = n_1/(n_1 + n_2)$ et $w_2 = n_2/(n_1 + n_2)$

(f) Si la série initiale $x_1, x_2, \ldots, x_n$ a été remplacée par un tableau d'effectifs

points milieux	m_1	m_2	...	m_k	TOTAL
effectifs	n_1	n_2	...	n_k	n

alors la moyenne $\bar{x}$ peut se calculer approximativement par une moyenne pondérée des points milieux

$$\bar{x} = \frac{1}{n} \sum_{\alpha=1}^{n} x_\alpha \approx \frac{1}{n} \sum_{\alpha=1}^{k} n_\alpha m_\alpha = \sum_{\alpha=1}^{k} \frac{n_\alpha}{n} m_\alpha = \bar{m}$$

En général, $\bar{m}$ constitue une bonne approximation de $\bar{x}$.

Exemple 2.4: suite de l'exemple 2.1

$$\bar{x} = \frac{1}{165} \sum_{\alpha=1}^{165} x_\alpha = 4066.3$$

$\tilde{x}$ = médiane = $x_{(83)}$ = 4085 ou 4082.5 (selon le choix adopté (voir page 2-7).

mode = 4015 et 4315

$$\bar{m} = \sum_{\alpha=1}^{13} \frac{n_\alpha}{n} m_\alpha = 4069.1$$

2.3 INDICATEURS DE DISPERSION

Il s'agit de quantités pour chiffrer la variabilité car les mesures de position seules ne fournissent pas un résumé suffisamment complet des données.

Étendue: ('RANGE')

$$x_{(n)} - x_{(1)} \tag{2.7}$$

est la différence entre la plus grande valeur et la plus petite valeur. Elle sert pour guider la construction d'un tableau d'effectifs et elle est aussi employée en contrôle statistique de la qualité.

Variance:

$$\frac{\sum_{\alpha=1}^{n}(x_\alpha - \bar{x})^2}{n-1} = s^2 \tag{2.8}$$

est la moyenne des écarts quadratiques $(x_\alpha - \bar{x})^2$. Le diviseur n-1 est associé au concept de DEGRÉ DE LIBERTÉ. La série des écarts:

$$x_1 - \bar{x} \ , \ x_2 - \bar{x} \ , \ \ldots, \ x_n - \bar{x}$$

vérifie

$$\sum_{\alpha=1}^{n} (x_\alpha - \bar{x}) = 0$$

et seulement (n-1) écarts sont linéairement indépendants.

Remarque: pour le calcul de la variance à partir d'un tableau d'effectifs, on a:

$$s_x^2 \approx \frac{1}{n-1} \sum_{\alpha=1}^{k} n_\alpha \, (m_\alpha - \bar{m})^2 = s_m^2$$

Écart-type: ('STANDARD DEVIATION')

$$s = \sqrt{s^2} \tag{2.9}$$

racine carrée de la variance. Cette mesure s'exprime dans l'unité de mesure de x .

Coefficient de variation:

$$cv = (s/\bar{x})*100 \tag{2.10}$$

est une mesure de dispersion relative et elle est généralement exprimée en pourcentage.

Exemple 2.5: suite de l'exemple 2.1

étendue = 5420 - 3040 = 2380
variance = s^2 = 216461
écart-type = s = 465.25
coefficient de variation = cv = 11.44%

Identité utile:

$$\sum_{\alpha=1}^{n} (x_\alpha - a)^2 = \sum_{\alpha=1}^{n} (x_\alpha - \bar{x})^2 + n(\bar{x} - a)^2$$

Cette identité a deux conséquences remarquables:

(i) $$\min_a \sum_{\alpha=1}^{n} (x_\alpha - a)^2 = \sum_{\alpha=1}^{n} (x_\alpha - \bar{x})^2$$

C'est le principe des moindres carrés et il sera employé comme méthode d'estimation de paramètres.

(ii) $$\sum_{\alpha=1}^{n} (x_\alpha - \bar{x})^2 = \sum_{\alpha=1}^{n} x_\alpha^2 - n\bar{x}^2$$

en posant a = 0

Cette équation est utile pour le calcul de la variance.

Écart-interquartile: $Q_3 - Q_1 = IQ$ (2.11)
('INTERQUARTILE RANGE')

où Q_1 est le 25ième percentile
Q_3 est le 75ième percentile

Cette mesure est utilisée dans la construction du diagramme schématique de Tukey.

Exemple 2.6: suite de l'exemple 2.1

$$Q_3 - Q_1 = 4345 - 3721 = 624$$

Écart-type de la moyenne: $s/\sqrt{n}$ (2.12)
('STANDARD ERROR OF THE MEAN')

est une mesure de dispersion associée à la moyenne $\bar{x}$. Nous verrons son utilité pour les règles de décision basées sur $\bar{x}$.

Exemple 2.7: suite de l'exemple 2.1

$$s/\sqrt{n} = 465.25/\sqrt{165} = 36.22$$

Valeurs Z ou centrées-réduites:
('Z SCORES')

$$z_\alpha = \frac{x_\alpha - \bar{x}}{s} \qquad \alpha = 1, 2, \ldots, n$$

sont des valeurs obtenues de la série originelle $x_1, x_2, \ldots, x_n$ et qui ont les propriétés suivantes:

- sans unité
- moyenne nulle: $\bar{z} = 0$
- écart-type unitaire: $s_z = 1$

Les données originelles peuvent s'écrire:

$$x_\alpha = \bar{x} + z_\alpha s$$

2.4 INDICATEURS DE FORME

Les indicateurs de forme sont associés à la distribution de l'ensemble des valeurs.

Coefficient d'asymétrie:
('SKEWNESS')

$$\frac{1}{n} \sum_{\alpha=1}^{n} z_\alpha^3 = b_1 \qquad (2.13)$$

où z_α sont les valeurs centrées-réduites de la série originelle.

Lorsque les valeurs sont distribuées symétriquement de chaque côté de la moyenne $\bar{x}$ on a $b_1 = 0$

Si $b_1 > 0$ la distribution est asymétrique vers la droite.

Si $b_1 < 0$ la distribution est asymétrique vers la gauche.

Exemple 2.8: suite de l'exemple 2.1

$$b_1 = 0.06$$

Coefficient d'aplatissement:
('KURTOSIS')

$$\frac{1}{n}\sum_{\alpha=1}^{n} z_\alpha^4 - 3 = b_2 \qquad (2.14)$$

est un coefficient qui mesure l'épaisseur des extrémités de la distribution.

On soustrait la valeur 3 pour fin de comparaison avec le modèle gaussien (normal). Pour des données générées par ce modèle, le coefficient b_2 est nul ou près de zéro.

Exemple 2.9: suite de l'exemple 2.1

$b_2 = -0.096$

Remarque: les définitions de b_1 et b_2 varient quelque peu selon les auteurs. Par exemple, le progiciel SAS calcule b_1 et b_2 selon les formules:

$$b_1 = \frac{n}{(n-1)(n-2)} \sum_{\alpha=1}^{n} z_\alpha^3$$

$$b_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} z_\alpha^4 - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

2.5 INDICATEURS POUR DÉCISIONS *

Certaines quantités servent de base à la prise de décisions. Parmi celles-ci, on note les indicateurs pour décider:

- si la moyenne théorique (population) est égale à une constante.
- si la forme de la distribution théorique de laquelle proviennent les données est gaussienne.

Statistique de Student:

$$t = \frac{\bar{x}}{s/\sqrt{n}} \qquad (2.15)$$

est employée pour décider si la moyenne de la population de laquelle provient l'échantillon est zéro. La procédure de décision dépend de la probabilité de dépasser la valeur de t . Cette probabilité est calculée selon une loi dite de STUDENT.

---

* (utile au chapitre 8 des tests d'hypothèses)

Statistique de Shapiro-Wilk: $$W = b^2/(n-1)s^2 \qquad (2.16)$$

où s est l'écart-type des données

$$b = \sum_{\alpha=1}^{k} a_{n-\alpha+1}(x_{(n-\alpha+1)} - x_{(\alpha)})$$

$k = n/2$ si n est pair
$k = (n-1)/2$ si n est impair

Un tableau ($n \leq 50$) des coefficients $a_{n-\alpha+1}$ sera donné dans le chapitre des tests d'hypothèses.

La valeur de W est comprise entre 0.7 (approximativement) et 1 et on l'emploie pour tester l'hypothèse que les données proviennent d'une population gaussienne.

Statistique de Kolmogorov: $$D = \max_{\alpha} \left| F_n(x_\alpha) - \Phi\left[\frac{x_\alpha - \bar{x}}{s}\right] \right| \qquad (2.17)$$

où $F_n(.)$ est la fonction de répartition échantillonnale et $\Phi(.)$ est la fonction de répartition du modèle gaussien centré réduit

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} t^2\right] dt \qquad (2.18)$$

La valeur de D est utilisée pour tester l'hypothèse que les données proviennent d'une population gaussienne.

2.6 GRAPHIQUES

Histogramme est un diagramme en barres verticales (ou horizontales) représentant un tableau d'effectifs. Dans un histogramme vertical, on place sur l'axe horizontal les points milieux et sur l'axe vertical on note l'effectif correspondant.

Exemple 2.10: données de l'exemple 2.1

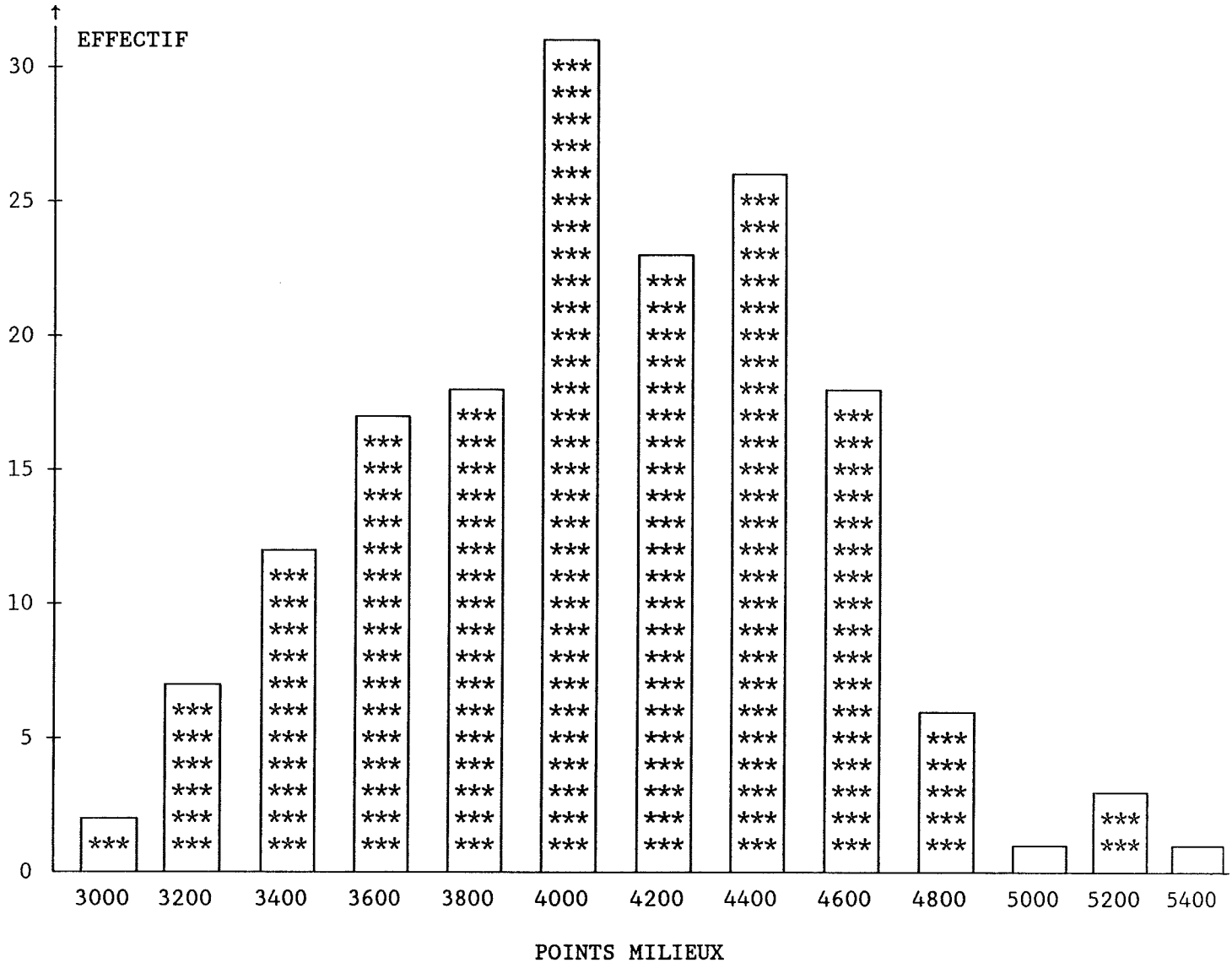


Figure 2.1: histogramme

Histogramme de Tukey
('STEM AND LEAF')

L'histogramme de Tukey est une variante de l'histogramme dans lequel on extrait de chaque nombre une partie principale appelée tige ('stem') et une partie secondaire appelée feuille ('leaf').

La partie principale sert de base à l'histogramme de Tukey et la partie secondaire sert à calculer les effectifs. En comparaison d'un histogramme usuel, celui de Tukey retient plus d'information. Pour obtenir un nombre de tiges entre 5 et 20, il faut quelquefois multiplier les unités par une puissance de 10 .

Exemple 2.11: histogramme de Tukey, données de l'exemple 2.1

Tige	Feuille	Nombre
54	2	1
52	00	2
50	01	2
48	1590	4
46	01467033	8
44	111567788991344468889	21
42	012479911112223333344799	24
40	011111122335567888999011222334555579999	38
38	00124677899223478889	20
36	00126777890022346	17
34	15693346789	11
32	368901226679	12
30	47229	5

----+----+----+----+----+----+----+----+
Multipliez les données par 100

Par exemple la ligne

32 368901226679

représente les nombres

3230, 3260, 3280, 3290, 3300, 3310, 3320, 3320, 3360, 3360, 3370, 3390

Diagramme schématique
('BOX PLOT')

Le diagramme schématique est un diagramme proposé par Tukey. Il s'agit d'une boîte et de deux droites ('box and whiskers') construit de la façon suivante:

- d'un rectangle dont les limites sont situées au 75ième percentile (Q_3) et au 25ième percentile (Q_1)

- la médiane est représentée par une ligne pleine à l'intérieur de la boîte

- la moyenne est identifiée avec le signe plus (+)

- du rectangle on trace deux droites de longueur

$$MAX(x_{(1)}, Q_1 - 1.5*IQ)$$

pour la droite inférieure, et de longueur

$$MIN(x_{(n)}, Q_3 + 1.5*IQ)$$

pour la droite supérieure où IQ est l'écart-interquartile $IQ = Q_3 - Q_1$

- toute valeur située dans $(Q_1 - 3*IQ, Q_1 - 1.5*IQ)$ ou $(Q_3 + 1.5*IQ, Q_3 + 3*IQ)$ est indiquée par 0. Ces valeurs apparaissent une fois sur 20, en moyenne, pour des données provenant d'une population gaussienne

- toute valeur inférieure à $Q_1 - 3*IQ$ ou supérieure à $Q_3 + 3*IQ$ est indiquée par un astérisque (*). Ces valeurs apparaissent une fois sur 200, en moyenne, pour des données provenant d'une population gaussienne

- le diagramme schématique est particulièrement utile pour comparer plusieurs séries de données tel qu'illustré sur le graphique suivant sur l'exemple de la comparaison des deux types de fil provenant du chapitre 1.

Exemple 2.12: Tension du fil par type

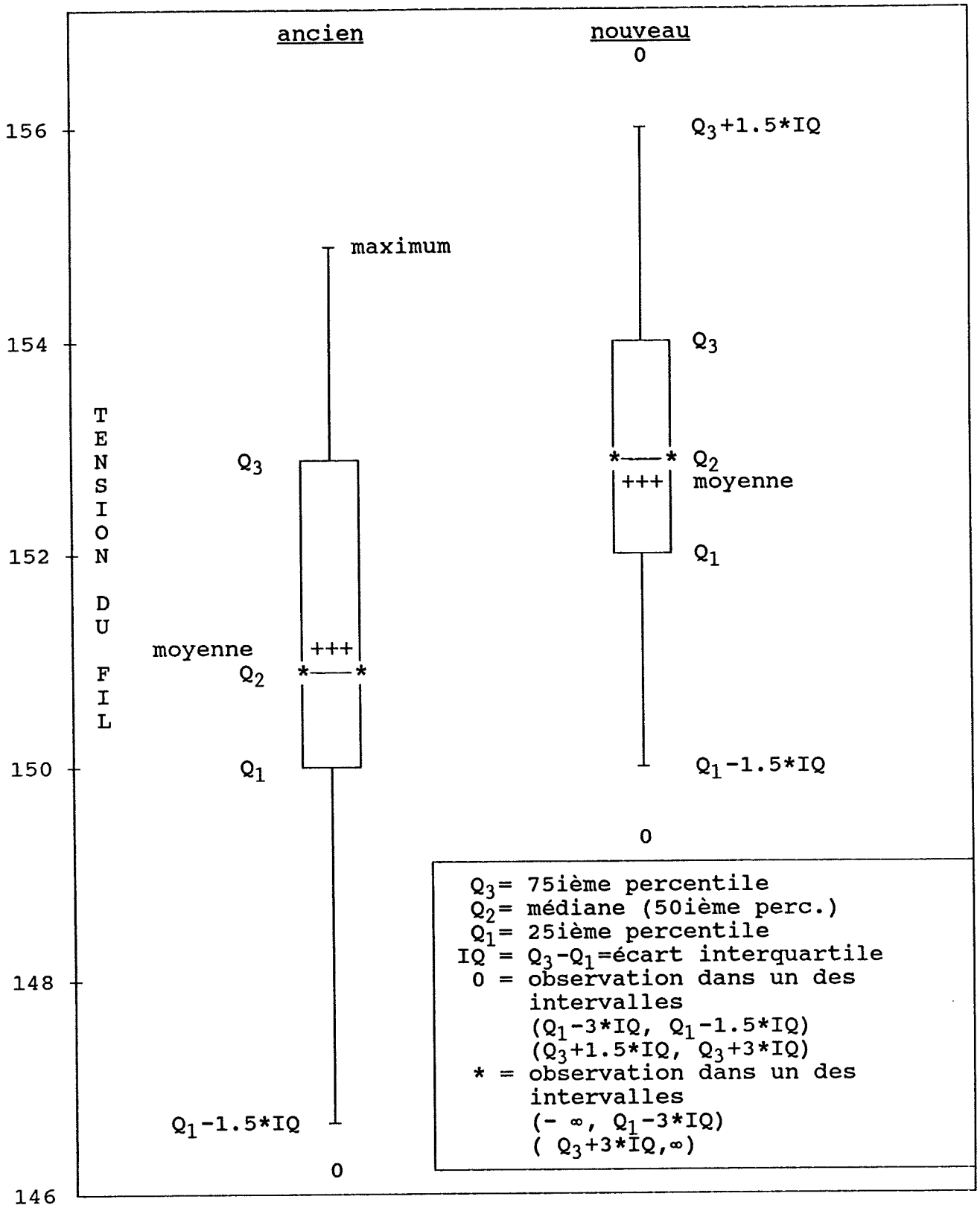


Figure 2.2: diagramme schématique

Graphique du pourcentage cumulatif
('CUMULATIVE PLOT')

On associe à la limite supérieure de chaque intervalle, le pourcentage cumulatif jusqu'à cette limite. C'est le graphique donnant le pourcentage cumulatif du nombre d'observations inférieures ou égales à chacune des limites supérieures des intervalles du tableau des effectifs. On relie par des segments de droites l'ensemble de ces points.

Pourcentage cumulatif Exemple 2.13

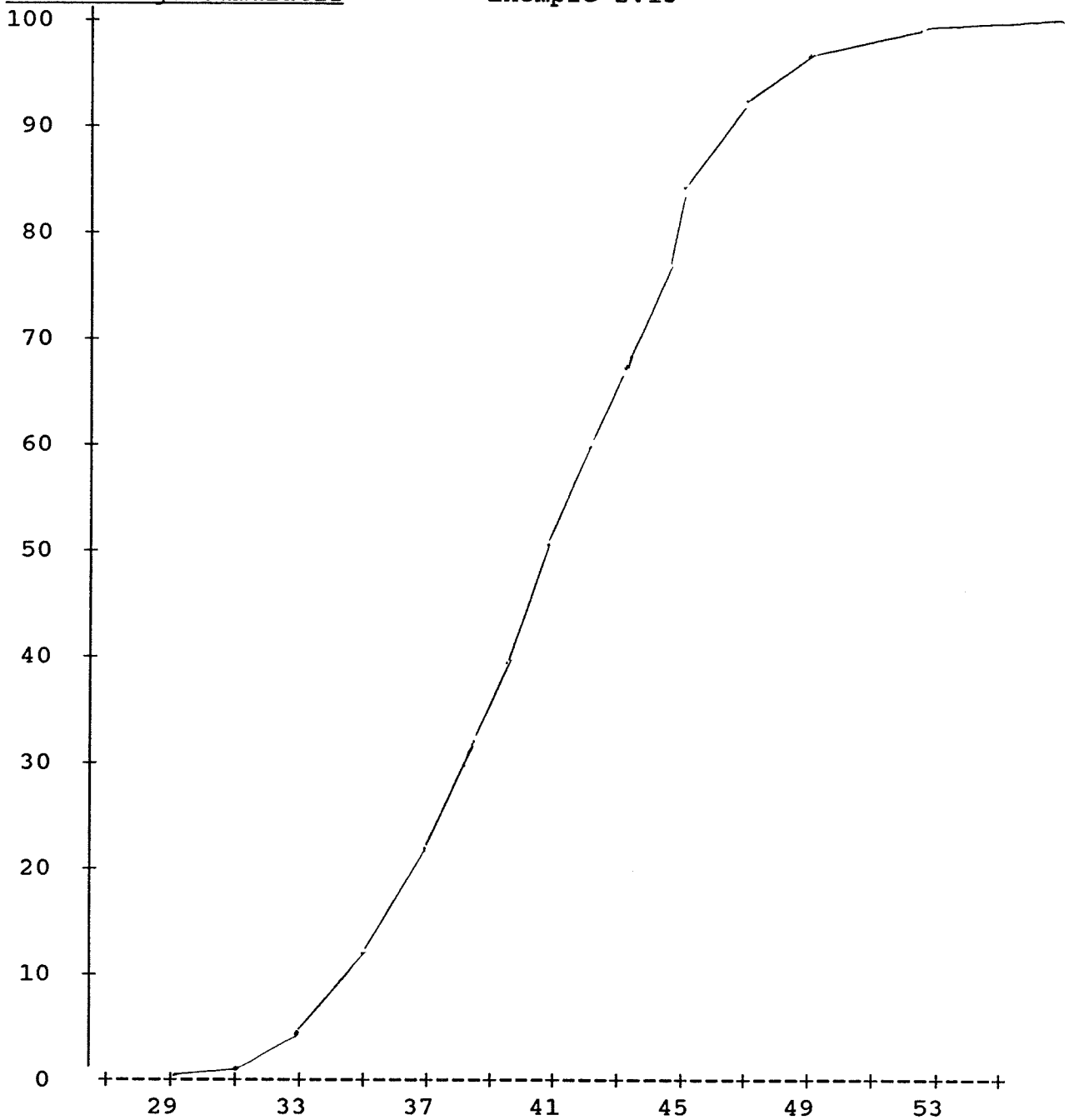


Figure 2.3: pourcentage cumulatif

Effectif cumulatif sur une échelle gaussienne
('NORMAL PROBABILITY PLOT')

C'est la représentation des données sur une échelle gaussienne, pour décider avec un examen visuel, si les données proviennent d'un modèle gaussien.

On fait le graphique des points $(x_{(i)}, y_i)$ où

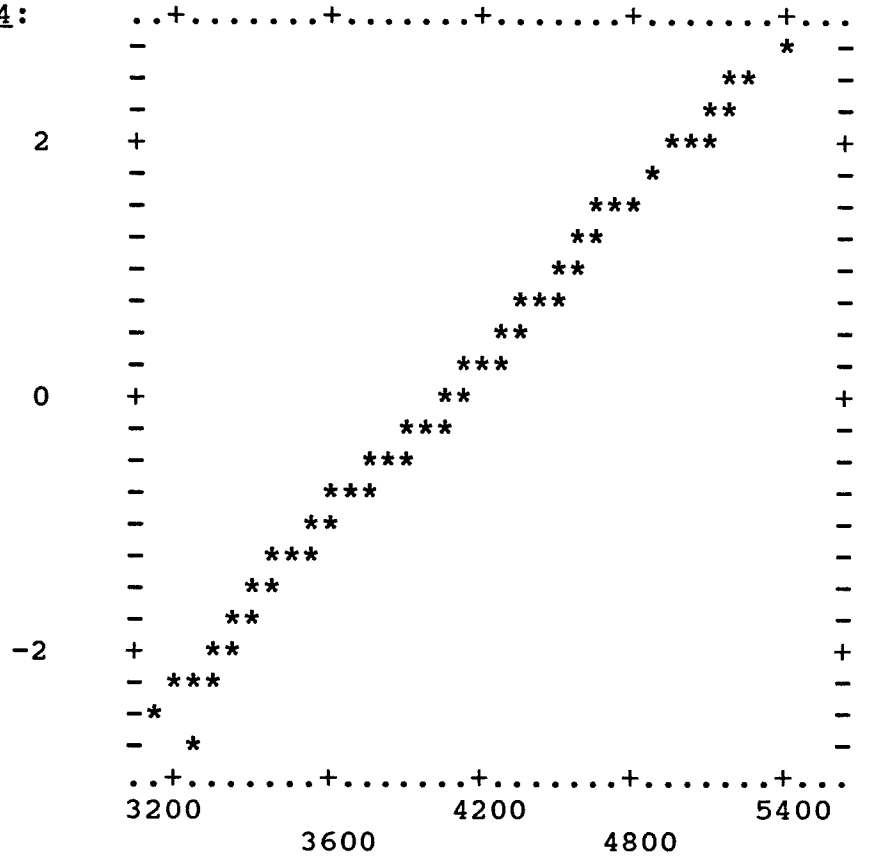
$$y_i = \Phi^{-1}\left[\frac{i - 0.375}{n + 0.250}\right] \quad i = 1,2,\ldots,n \qquad (2.19)$$

Φ^{-1} étant la réciproque de la fonction gaussienne définie en (2.18).

Remarque: plusieurs autres choix ont été proposés comme argument de Φ^{-1}: i/n+1, (i-1/2)/n

Ce type de graphique est un cas particulier (mais important) de ce qu'il est maintenant convenu d'appeler les graphiques percentiles-percentiles où on met en relation les percentiles expérimentaux avec les percentiles calculés par une distribution théorique.

Exemple 2.14:



DES VALEURS PROVENANT D'UNE DISTRIBUTION GAUSSIENNE SERAIENT SITUÉES SUR UNE DROITE

Figure 2.4: échelle gaussienne

2.7 DESCRIPTION DE DEUX VARIABLES NUMÉRIQUES

On définit les statistiques élémentaires de deux variables numériques X et Y par les quantités suivantes.

Données: (x_α, y_α) $\alpha = 1, 2, \ldots, n$

Diagramme de dispersion conjointe: ("Scattergram") graphique des données dans un système d'axes rectangulaires

Moyennes:

$$\bar{x} = \frac{1}{n}\sum_{\alpha=1}^{n} x_\alpha = \text{moyenne de x}$$

$$\bar{y} = \frac{1}{n}\sum_{\alpha=1}^{n} y_\alpha = \text{moyenne de y}$$

Variances:

$$s_x^2 = \frac{1}{n-1}\sum_{\alpha=1}^{n} (x_\alpha - \bar{x})^2 = \text{variance de x}$$

$$s_y^2 = \frac{1}{n-1}\sum_{\alpha=1}^{n} (y_\alpha - \bar{y})^2 = \text{variance de y}$$

Covariance:

$$s_{xy} = \frac{1}{n-1}\sum_{\alpha=1}^{n} (x_\alpha - \bar{x})(y_\alpha - \bar{y}) = \text{covariance de X et Y} \qquad (2.20)$$

elle mesure la variation conjointe (covariation) de x et y

Coefficient de corrélation:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \qquad -1 \le r_{xy} \le 1 \qquad (2.21)$$

est une mesure de liaison LINÉAIRE entre x et y . On note aussi r_{xy} par r s'il n'y a pas d'ambiguité.

Le coefficient r s'appelle le COEFFICIENT DE CORRÉLATION de Pearson.

On montre que:

$$r = \pm 1 \quad <----> \quad y = \beta_0 + \beta_1 x$$

Droite de moindres carrés

C'est une droite, calculée par le principe des moindres carrés, et ayant pour but la modélisation du couple de variables (X,Y) par l'équation:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X$$

où $$\hat{\beta}_1 = r\frac{s_y}{s_x}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (2.22)$$

La notation ^ est employée pour indiquer qu'il s'agit d'estimations calculées à partir des données. Une mesure de variation des points autour de la droite est obtenue par

$$\text{ÉCART-TYPE RÉSIDUEL} = s_y\sqrt{1-r^2}$$

Nous verrons la justification de tous ces calculs dans le chapitre sur la régression.

Exemple 2.15: $\bar{x} = \bar{y} = 3$ $s_x = s_y = 1$

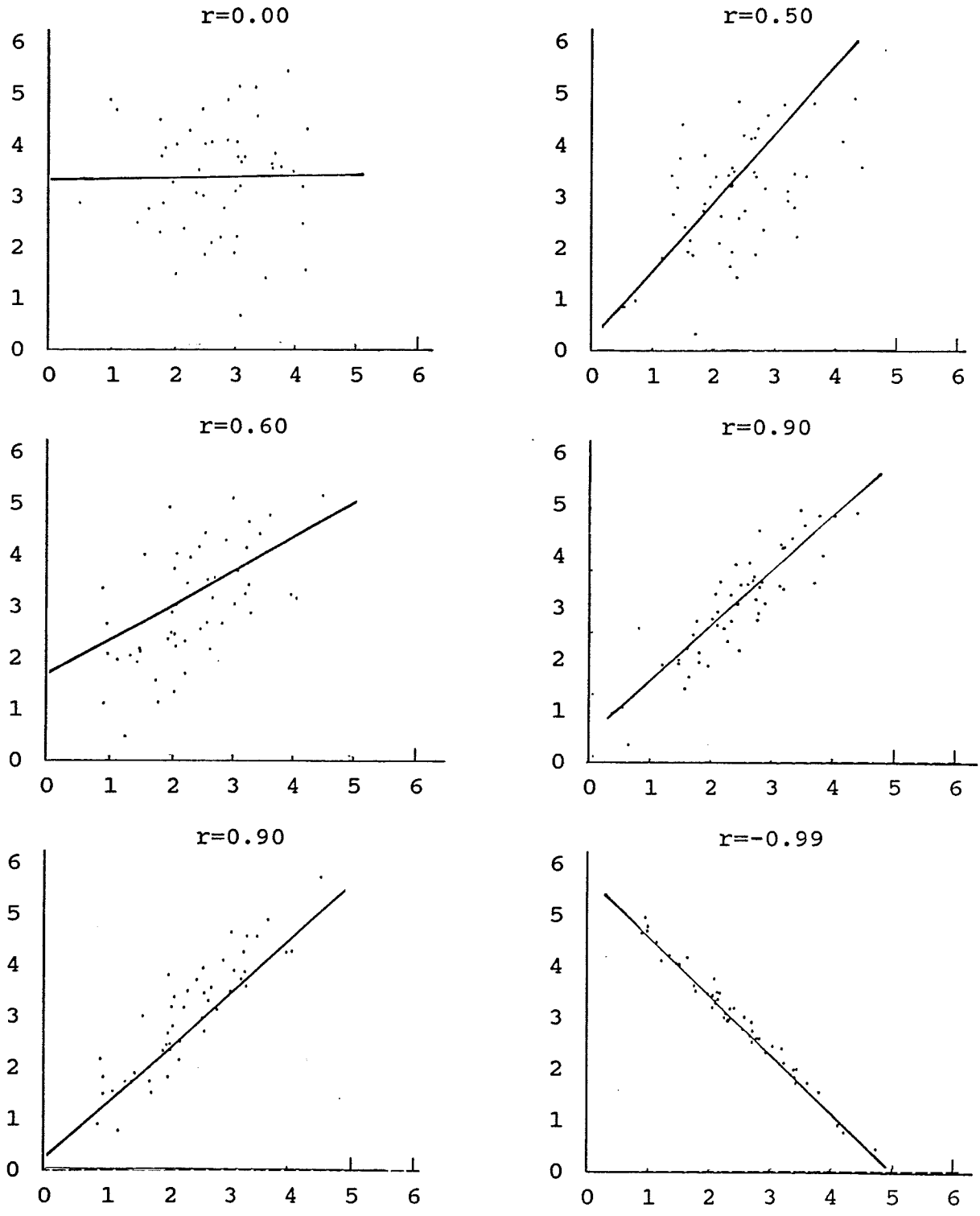


figure 2.5: exemples de diagrammes de dispersion conjointe

2.8 DESCRIPTION DE DEUX VARIABLES QUALITATIVES

Soient (X,Y) deux variables qualitatives et notons par $A_1, A_2, \ldots, A_r$ les r modalités de X et $B_1, B_2, \ldots, B_c$ les c modalités de Y .

À partir de n observations (x_α, y_α) sur le couple (X, Y) on peut constituer un TABLEAU D'EFFECTIFS CONJOINTS, appelé aussi TABLEAU DE CONTINGENCE

effectifs conjoints (n_{ij})

X \ Y	B_1	B_2	...	B_j	...	B_c	TOTAL
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1c}	n_{1+}
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2c}	n_{2+}
.	.	.		.		.	.
.	.	.		.		.	.
.	.	.		.		.	.
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ic}	n_{i+}
.	.	.		.		.	.
.	.	.		.		.	.
.	.	.		.		.	.
A_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rc}	n_{r+}
TOTAL	n_{+1}	n_{+2}	...	n_{+j}	...	n_{+c}	n

où n_{ij} = EFFECTIF CONJOINT de (A_i, B_j)

$$n_{i+} = \sum_j n_{ij} = \text{effectif de } A_i, \quad i = 1,2,\ldots,r \qquad (2.23)$$

$$n_{+j} = \sum_i n_{ij} = \text{effectif de } B_j, \quad j = 1,2,\ldots,c$$

$$n = \Sigma\Sigma\, n_{ij} = \text{effectif total}$$

On définit

$$f_{ij} = n_{ij}/n = \text{fréquence conjointe de } (A_i, B_j)$$

$$f_{i+} = n_{i+}/n = \text{fréquence marginale de } A_i$$

$$f_{+j} = n_{+j}/n = \text{fréquence marginale de } B_j \qquad (2.24)$$

$$\frac{f_{ij}}{f_{+j}} = \frac{n_{ij}}{n_{+j}} = \text{fréquence conditionnelle de } A_i \text{ si } Y = B_j$$

$$\frac{f_{ij}}{f_{i+}} = \frac{n_{ij}}{n_{i+}} = \text{fréquence conditionnelle de } B_j \text{ si } X = A_i$$

Indépendance

Une question fondamentale qui dérive de l'étude d'un tel tableau est de décider si les variables sont liées. Une procédure statistique pour décider de cette question est basée sur la quantité D^2 proposée par K. Pearson.

$$D^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \quad \text{où } e_{ij} = \frac{n_{i+} * n_{+j}}{n} \qquad (2.25)$$

Cette quantité sera justifiée au chapitre des tests d'hypothèses et elle repose sur le concept d'INDÉPENDANCE qui sera défini au chapitre des probabilités.

Mesures de liaison entre les variables

La quantité D^2 permet de mesurer le degré de liaison entre X et Y . Deux indicateurs, parmi plusieurs, sont ainsi définis à partir de D^2:

$$\text{COEFFICIENT PHI} = \sqrt{D^2/n} \qquad (2.26)$$

$$\text{COEFFICIENT DE CONTINGENCE} = C = \sqrt{D^2/(D^2+n)} = \sqrt{\Phi^2/(1+\Phi^2)} \qquad (2.27)$$

Les coefficients PHI et C numériquement voisins, jouent pour des variables qualitatives le même rôle que r joue pour des variables numériques.

Exemple 2.16:

Le tableau des effectifs conjoints croisant la variable équipe et la variable machine a été calculé dans le tableau ci-joint.

Dans chaque cellule du tableau on retrouve dans l'ordre:

effectif observé n_{ij}
effectif calculé sous l'hypothèse d'indépendance e_{ij}
déviation = $n_{ij} - e_{ij}$
contribution à la statistique du D^2, $(n_{ij} - e_{ij})^2/e_{ij}$
pourcentage du grand total
pourcentage du total rangée
pourcentage du total colonne

	MACHINE				
ÉQUIPE	A	B	C	D	TOTAL
JOUR	10	15	15	10	50
	10.0	11.2	15.0	13.8	
	0.0	3.8	0.0	-3.8	
	0.0	1.3	0.0	1.0	
	5.00	7.50	7.50	5.00	25.00
	20.00	30.00	30.00	20.00	
	25.00	33.33	25.00	18.18	
SOIR	10	20	15	20	65
	13.0	14.6	19.5	17.9	
	-3.0	5.4	-4.5	2.1	
	0.7	2.0	1.0	0.3	
	5.00	10.00	7.50	10.00	32.50
	15.38	30.77	23.08	30.77	
	25.00	44.44	25.00	36.36	
NUIT	20	10	30	25	85
	17.0	19.1	25.5	23.4	
	3.0	-9.1	4.5	1.6	
	0.5	4.4	0.8	0.1	
	10.00	5.00	15.00	12.50	42.50
	23.53	11.76	35.29	29.41	
	50.00	22.22	50.00	45.45	
TOTAL	40	45	60	55	200
	20.00	22.50	30.00	27.50	100.00

D^2	12.022
PHI	0.245
COEFFICIENT DE CONTINGENCE	0.238

2.9 DESCRIPTION DE PLUSIEURS VARIABLES NUMÉRIQUES

C'est très souvent le cas que les données contiennent des informations sur plusieurs variables numériques. Il est utile de représenter les données à l'aide d'une matrice.

matrice des données

variable

	1	2	...	P
1ière observation:	x_{11}	x_{12}	...	x_{1p}
2ième observation:	x_{21}	x_{22}	...	x_{2p}
...	...	...	...	...
nième observation:	x_{n1}	x_{n2}	...	x_{np}

Les descriptions élémentaires des données sont formées des:

<u>moyennes</u>: $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p$

<u>variances</u>: $s_1^2, s_2^2, \ldots, s_p^2$

et de la <u>MATRICE DE CORRÉLATION</u> entre chaque paire de variables:

variable

variable	1	2	...	p
1	1	r_{12}	...	r_{1p}
2	r_{21}	1	...	r_{2p}
.	.	.	...	.
.	.	.	...	.
.	.	.	...	.
p	r_{p1}	r_{p2}	...	1

Cette matrice est symétrique: $r_{ij} = r_{ji}$

$r_{ii} = 1$

La grande majorité des analyses multidimensionnelles a pour point de départ une matrice de corrélation.

2.10 UTILISATION DE SAS

Le progiciel SAS possède plusieurs procédures (PROC) pour faire la description des données.

procédure	sortie
CHART	- histogrammes horizontaux - histogrammes verticaux - diagrammes circulaires ('pie chart') - diagrammes en blocs (3 dimensions) - diagrammes en étoiles
CORR	- moyennes et écart-types de variables - coefficients de corrélation de Pearson - test de la nullité des coefficients
FREQ	- tableaux d'effectifs unidimensionnels - tableaux d'effectifs multidimensionnels - mesures d'association pour les tableaux en deux dimensions
MEANS	- caractéristiques de position, dispersion et de forme - test de Student pour la moyenne
PLOT	- graphiques en deux dimensions (scattergramme)
PRINT	- impression des données d'un fichier SAS
SUMMARY	- identique à MEANS mais pour créer automatiquement un fichier SAS contenant les caractéristiques de position, dispersion; il n'y a pas de sortie imprimée
TABULATE	- tableaux complexes de statistiques descriptives spécialement pour des classifications et hiérarchies; les tableaux peuvent avoir trois dimensions; rangée, colonne, page.
UNIVARIATE	- caractéristiques de position, forme, dispersion - percentiles - test d'ajustement à une loi gaussienne - diagramme schématique de Tukey - histogramme de Tukey - graphique sur échelle gaussienne

```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++
+   EXEMPLE  : DESCRIPTION DE VARIABLES NUMERIQUES     +
+                                                      +
+   PROCEDURE: MEANS, UNIVARIATE                       +
++++++++++++++++++++++++++++++++++++++++++++++++++++++++

DATA FORCE;
     INPUT X @@;   /* MODE INPUT AVEC PLUSIEURS OBSERVATIONS
                                          PAR LIGNE */;
     TITLE '165 DONNEES DE FORCE DE COMPRESSION (PSI)';
     LIST;CARDS;
4855 3705 3125 3120 4125 4410 4415 4580 3680 4670
3490 3815 4315 3545 3695 3675 3705 4450 4410 4810
4315 4155 4095 4225 3565 4135 4585 4330 4315 3865
3465 3305 3740 3670 4375 4645 5200 4100 4090 4015
4170 4510 4730 4330 3070 3260 3870 3870 3530 4150
3235 3040 3620 4395 4150 5005 3805 3600 4135 3290
4080 3310 4345 3920 3825 4480 4540 4580 3730 4700
4060 3570 4335 3410 4085 4290 4320 4495 4195 4195
4080 4120 4535 4030 4660 4010 3320 4050 3360 3535
3450 4330 3725 4020 4150 3940 3605 4000 4315 3280
4200 3375 4055 3670 4900 4390 4015 4505 4890 4190
3390 4030 4610 4345 4070 4505 4485 4290 3980 5110
4320 4335 3760 4020 4540 4015 3985 4140 4115 4475
3800 4110 4245 4320 3720 3845 3920 3660 4600 3935
3320 4210 3610 4125 4590 3985 3590 3585 3995 3890
4015 4490 3195 3365 3890 3975 4095 4540 5420 5200
3885 4270 4010 4730 4565
 ;
PROC MEANS DATA=FORCE;
     TITLE 'DESCRIPTION SOMMAIRE';

PROC MEANS DATA=FORCE N MEAN STD MIN MAX RANGE
                              CV SKEWNESS KURTOSIS;
     TITLE 'DESCRIPTION AVEC MOTS-CLES CHOISIS';

PROC UNIVARIATE DATA=FORCE FREQ NORMAL PLOT PCTLDEF=1;

     TITLE1 'DESCRIPTION DETAILLEE AVEC LISTE ORDONNEE,';
     TITLE2 'PERCENTILES, HISTOGRAMME DE TUKEY,';
     TITLE3 'DIAGRAMME SCHEMATIQUE ET ECHELLE GAUSSIENNE';
```

++++++++++++++++++++++++++
+ OUTPUT DE PROC MEANS +
++++++++++++++++++++++++++

DESCRIPTION SOMMAIRE

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE
X	165	4066.3030303	465.25330662	3040.0000000	5420.000000

++++++++++++++++++++++++
+ OUTPUT DE PROC MEANS +
++++++++++++++++++++++++

DESCRIPTION AVEC MOTS-CLES CHOISIS

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE
X	165	4066.3030303	465.25330662	3040.0000000	5420.000000

VARIABLE	RANGE	C.V.	SKEWNESS	KURTOSIS
X	2380.00000000	11.442	0.06015519	-0.09589499

++++++++++++++++++++++++++++++
+ OUTPUT DE PROC UNIVARIATE +
++++++++++++++++++++++++++++++

DESCRIPTION DETAILLEE AVEC LISTE ORDONNEE,
PERCENTILES, HISTOGRAMME DE TUKEY,
DIAGRAMME SCHEMATIQUE ET ECHELLE GAUSSIENNE

UNIVARIATE

VARIABLE=X

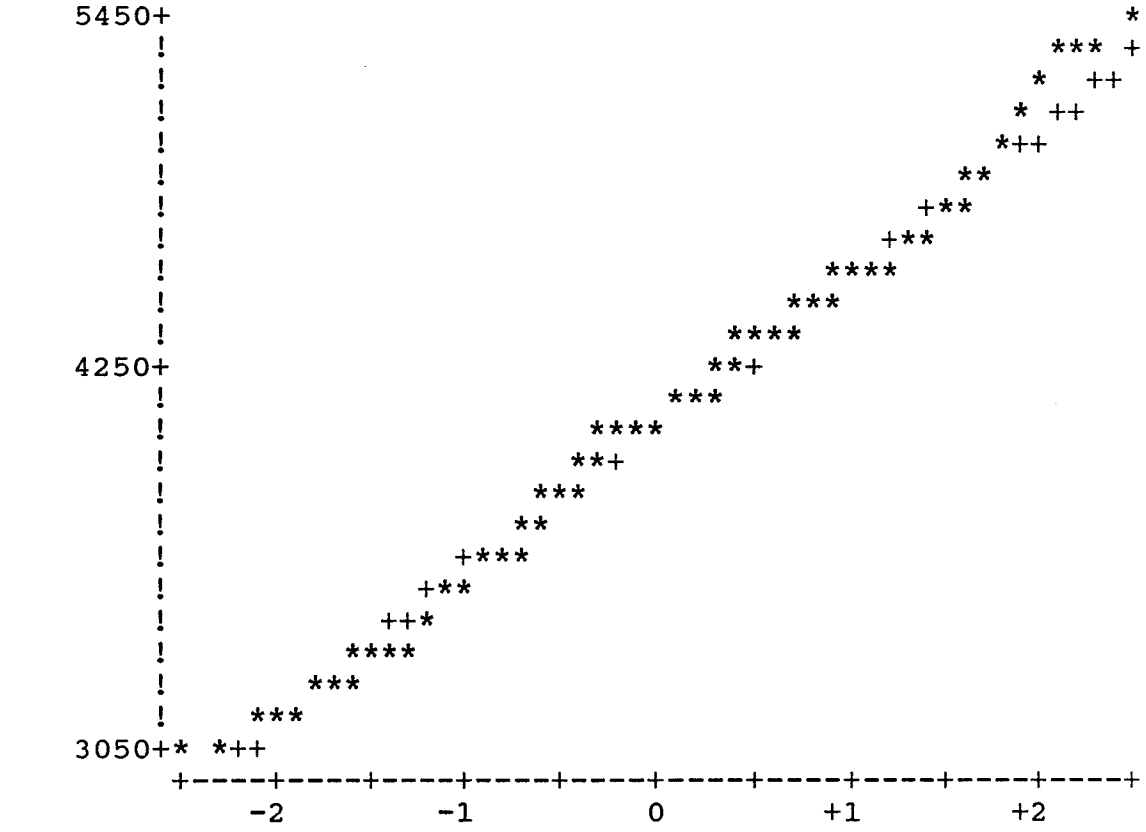
MOMENTS

N	165	SUM WGTS	165
MEAN	4066.3	SUM	670940
STD DEV	465.253	VARIANCE	216461
SKEWNESS	0.0601552	KURTOSIS	-0.095895
USS	2763744900	CSS	35499545
CV	11.4417	STD MEAN	36.2199
T:MEAN=0	112.267	PROB>!T!	0.0001
SGN RANK	6847.5	PROB>!S!	0.0001
NUM ^= 0	165		
D:NORMAL	0.0585737	PROB>D	>.15

QUANTILES(DEF=1)				EXTREMES	
100% MAX	5420	99%	5200	LOWEST	HIGHEST
75% Q3	4345	95%	4790	3040	5005
50% MED	4082.5	90%	4595	3070	5110
25% Q1	3721.25	10%	3382.5	3120	5200
0% MIN	3040	5%	3282.5	3125	5200
		1%	3059.5	3195	5420
RANGE	2380				
Q3-Q1	623.75				
MODE	4015				

```
STEM LEAF                          #                                  BOXPLOT
  54 2                             1                                     0
  53
  52 00                            2                                     !
  51 1                             1                                     !
  50 0                             1                                     !
  49 0                             1                                     !
  48 159                           3                                     !
  47 033                           3                                     !
  46 01467                         5                                     !
  45 001344468889                 12                                     !
  44 111578899                     9                                     !
  43 11112223333344799            17                                +------+
  42 0124799                       7                                !      !
  41 01122233455557999            17                                !      !
  40 011111122335567888999        21                                *--+--*
  39 223478889                     9                                !      !
  38 00124677899                  11                                !      !
  37 0022346                       7                                +------+
  36 0012677789                   10                                   !
  35 3346789                       7                                   !
  34 1569                          4                                   !
  33 01226679                      8                                   !
  32 3689                          4                                   !
  31 229                           3                                   !
  30 47                            2                                   !
     ----+----+----+----+-
  MULTIPLY STEM.LEAF BY 10**+02
```

NORMAL PROBABILITY PLOT



FREQUENCY TABLE

VALUE	COUNT	PERCENTS CELL	PERCENTS CUM	VALUE	COUNT	PERCENTS CELL	PERCENTS CUM	VALUE	COUNT	PERCENTS CELL	PERCENTS CUM
3040	1	1	0.6	3890	2	56	33.9	4475	1	132	80.0
3070	1	2	1.2	3920	2	58	35.2	4480	1	133	80.6
3120	1	3	1.8	3935	1	59	35.8	4485	1	134	81.2
3125	1	4	2.4	3940	1	60	36.4	4490	1	135	81.8
3195	1	5	3.0	3975	1	61	37.0	4495	1	136	82.4
3235	1	6	3.6	3980	1	62	37.6	4505	2	138	83.6
3260	1	7	4.2	3985	2	64	38.8	4510	1	139	84.2
3280	1	8	4.8	3995	1	65	39.4	4535	1	140	84.8
3290	1	9	5.5	4000	1	66	40.0	4540	3	143	86.7
3305	1	10	6.1	4010	2	68	41.2	4565	1	144	87.3
3310	1	11	6.7	4015	4	72	43.6	4580	2	146	88.5
3320	2	13	7.9	4020	2	74	44.8	4585	1	147	89.1
3360	1	14	8.5	4030	2	76	46.1	4590	1	148	89.7
3365	1	15	9.1	4050	1	77	46.7	4600	1	149	90.3
3375	1	16	9.7	4055	1	78	47.3	4610	1	150	90.9
3390	1	17	10.3	4060	1	79	47.9	4645	1	151	91.5
3410	1	18	10.9	4070	1	80	48.5	4660	1	152	92.1
3450	1	19	11.5	4080	2	82	49.7	4670	1	153	92.7
3465	1	20	12.1	4085	1	83	50.3	4700	1	154	93.3
3490	1	21	12.7	4090	1	84	50.9	4730	2	156	94.5
3530	1	22	13.3	4095	2	86	52.1	4810	1	157	95.2
3535	1	23	13.9	4100	1	87	52.7	4855	1	158	95.8
3545	1	24	14.5	4110	1	88	53.3	4890	1	159	96.4
3565	1	25	15.2	4115	1	89	53.9	4900	1	160	97.0
3570	1	26	15.8	4120	1	90	54.5	5005	1	161	97.6
3585	1	27	16.4	4125	2	92	55.8	5110	1	162	98.2
3590	1	28	17.0	4135	2	94	57.0	5200	2	164	99.4
3600	1	29	17.6	4140	1	95	57.6	5420	1	165	100.0
3605	1	30	18.2	4150	3	98	59.4				
3610	1	31	18.8	4155	1	99	60.0				
3620	1	32	19.4	4170	1	100	60.6				
3660	1	33	20.0	4190	1	101	61.2				
3670	2	35	21.2	4195	2	103	62.4				
3675	1	36	21.8	4200	1	104	63.0				
3680	1	37	22.4	4210	1	105	63.6				
3695	1	38	23.0	4225	1	106	64.2				
3705	2	40	24.2	4245	1	107	64.8				
3720	1	41	24.8	4270	1	108	65.5				
3725	1	42	25.5	4290	2	110	66.7				
3730	1	43	26.1	4315	4	114	69.1				
3740	1	44	26.7	4320	3	117	70.9				
3760	1	45	27.3	4330	3	120	72.7				
3800	1	46	27.9	4335	2	122	73.9				
3805	1	47	28.5	4345	2	124	75.2				
3815	1	48	29.1	4375	1	125	75.8				
3825	1	49	29.7	4390	1	126	76.4				
3845	1	50	30.3	4395	1	127	77.0				
3865	1	51	30.9	4410	2	129	78.2				
3870	2	53	32.1	4415	1	130	78.8				
3885	1	54	32.7	4450	1	131	79.4				

```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
+  EXEMPLE  : RECODAGE D'UNE VARIABLE NUMERIQUE EN CLASSES,    +
+             CALCUL D'UN TABLEAU D'EFFECTIFS,                 +
+             TRACAGE D'UN HISTOGRAMME VERTICAL                +
+  PROCEDURE: FREQ, CHART                                      +
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

DATA FORCE2;
            SET FORCE;

       /*RECODAGE DE LA VARIABLE X */;

     IF         X < 3100 THEN M=3000;
     IF 3100 <= X < 3300 THEN M=3200;
     IF 3300 <= X < 3500 THEN M=3400;
     IF 3500 <= X < 3700 THEN M=3600;
     IF 3700 <= X < 3900 THEN M=3800;
     IF 3900 <= X < 4100 THEN M=4000;
     IF 4100 <= X < 4300 THEN M=4200;
     IF 4300 <= X < 4500 THEN M=4400;
     IF 4500 <= X < 4700 THEN M=4600;
     IF 4700 <= X < 4900 THEN M=4800;
     IF 4900 <= X < 5100 THEN M=5000;
     IF 5100 <= X < 5300 THEN M=5200;

PROC FREQ DATA=FORCE2;
     TABLES M;
     TITLE 'TABLEAU D''EFFECTIFS';

PROC CHART DATA=FORCE2;
     VBAR M / TYPE=FREQ
              MIDPOINTS= 3000 TO 5400 BY 200;
     TITLE 'HISTOGRAMME VERTICAL';
```

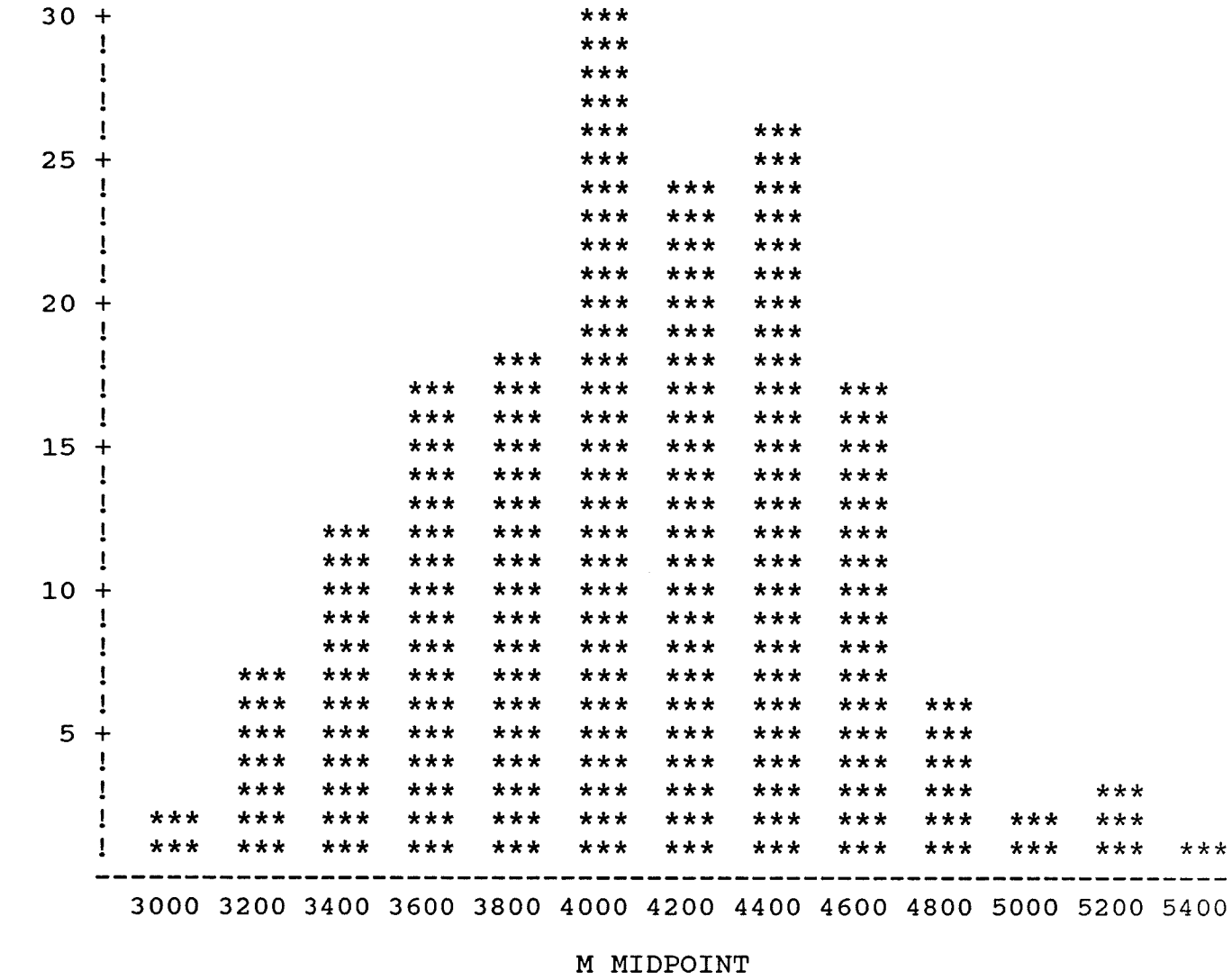
```
                        ++++++++++++++++++++++++
                        + OUTPUT DE PROC FREQ +
                        ++++++++++++++++++++++++
                         TABLEAU D'EFFECTIFS
```

M	FREQUENCY	PERCENT	CUMULATIVE FREQUENCY	CUMULATIVE PERCENT
3000	2	1.2	2	1.2
3200	7	4.2	9	5.5
3400	12	7.3	21	12.7
3600	17	10.3	38	23.0
3800	18	10.9	56	33.9
4000	30	18.2	86	52.1
4200	24	14.5	110	66.7
4400	26	15.8	136	82.4
4600	17	10.3	153	92.7
4800	6	3.6	159	96.4
5000	2	1.2	161	97.6
5200	3	1.8	164	99.4
5400	1	0.6	165	100.0

****
* OUTPUT DE PROC CHART *
****

HISTOGRAMME VERTICAL

FREQUENCY BAR CHART

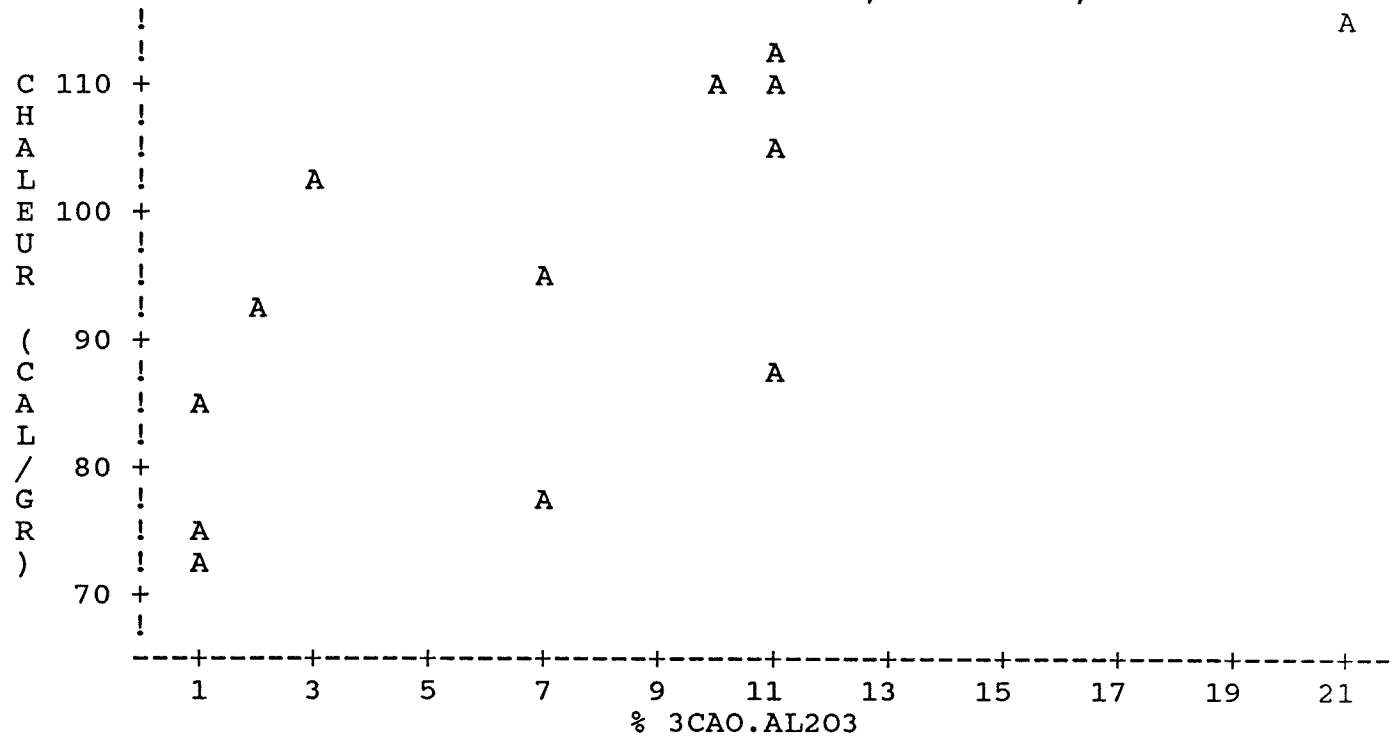


```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++
+  EXEMPLE  : TRACAGE DE SCATERGRAMMES,  CALCUL D'UNE +
+             MATRICE DE CORRELATION                  +
+  PROCEDURE: PLOT, CORR                              +
++++++++++++++++++++++++++++++++++++++++++++++++++++++++

DATA BETON;
     INPUT X1-X5;           /* DECLARATION IMPLICITE DU NOM DES
                               5 VARIABLES: X1 X2 X3 X4 X5  */ ;
     LABEL X1=% 3CAO.AL2O3
           X2=% 3CAO.SIO2
           X3=% DE 4CAO.AL2.FE2O3
           X4=% DE 2CAO.SIO2
           X5=CHALEUR (CAL/GR);
     LIST; CARDS;
 7 26  6 60  78.5
 1 29 15 52  74.3
11 56  8 20 104.3
11 31  8 47  87.6
 7 52  6 33  95.9
11 55  9 22 109.2
 3 71 17  6 102.7
 1 31 22 44  72.5
 2 54 18 22  93.1
21 47  4 26 115.9
 1 40 23 34  83.8
11 66  9 12 113.3
10 68  8 12 109.4
;
PROC PLOT;
     PLOT X5*(X1 X2 X3 X4) / VPOS=20 ;
     TITLE 'SCATTERGRAMME DE X5 VERSUS X1 X2 X3 X4';

PROC CORR;
     TITLE 'MATRICE DE CORRELATION';
```

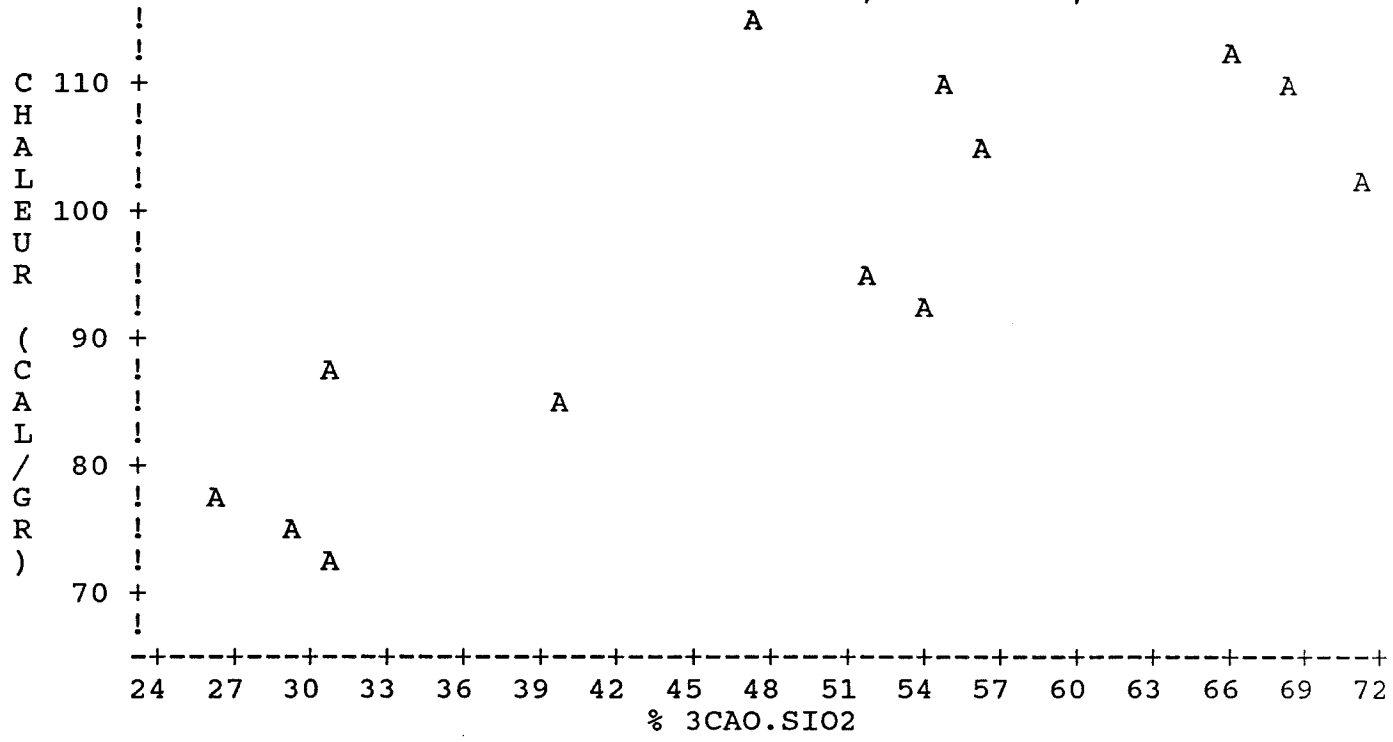
++++++++++++++++++++++++
+ OUTPUT DE PROC PLOT +
++++++++++++++++++++++++

SCATTERGRAMME DE X5 VERSUS X1 X2 X3 X4

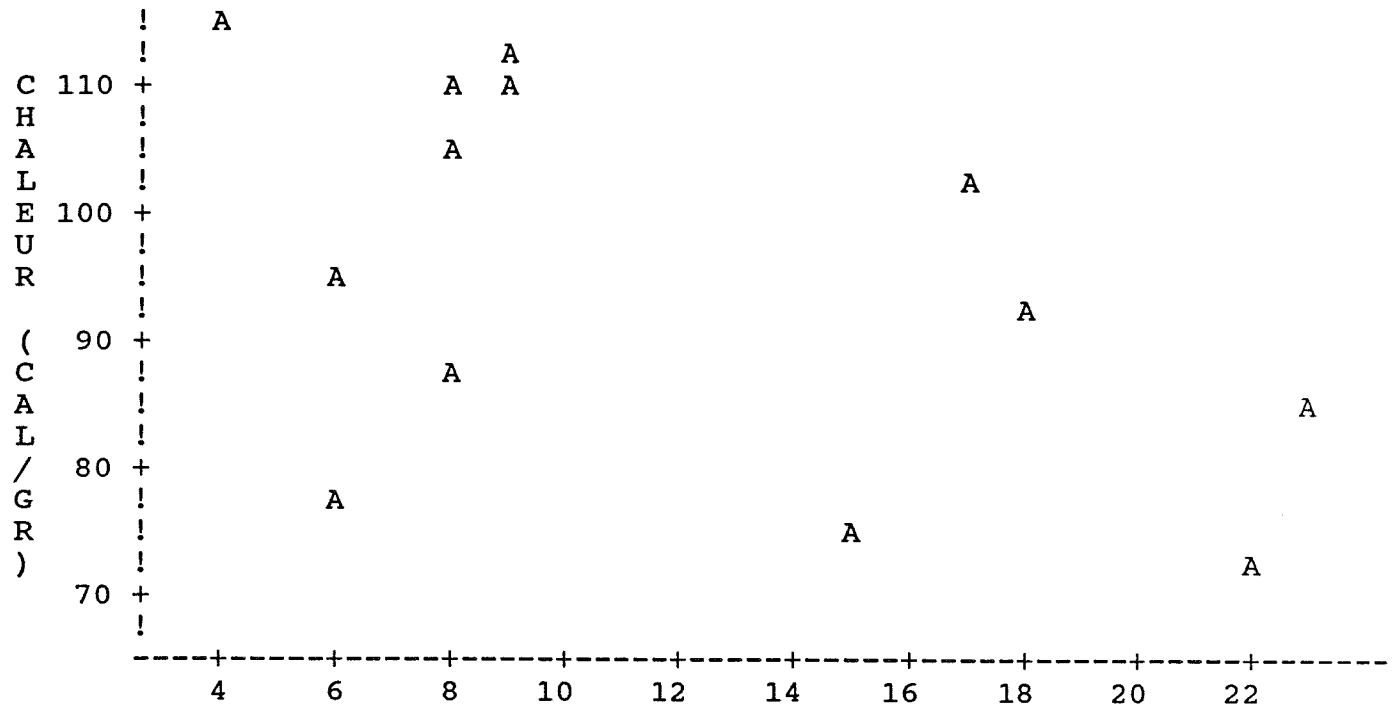
PLOT OF X5*X1 LEGEND: A = 1 OBS, B = 2 OBS, ETC.



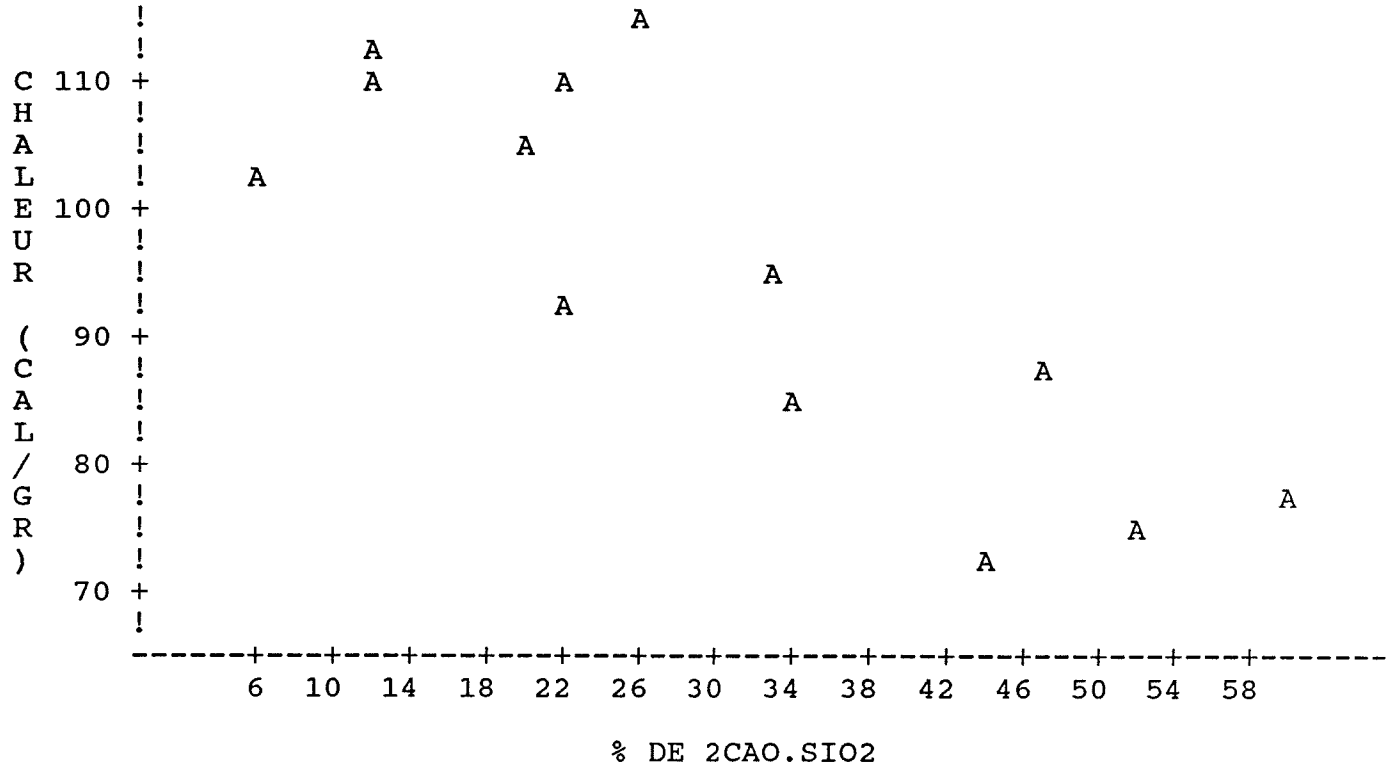
PLOT OF X5*X2 LEGEND: A = 1 OBS, B = 2 OBS, ETC.



PLOT OF X5*X3 LEGEND: A = 1 OBS, B = 2 OBS, ETC.



PLOT OF X5*X4 LEGEND: A = 1 OBS, B = 2 OBS, ETC.



++++++++++++++++++++++++
+ OUTPUT DE PROC CORR +
++++++++++++++++++++++++

MATRICE DE CORRELATION

VARIABLE	N	MEAN	STD DEV	SUM	MINIMUM	MAXIMUM
X1	13	7.46154	5.88239	97.000	1.00000	21.0000
X2	13	48.15385	15.56088	626.000	26.00000	71.0000
X3	13	11.76923	6.40513	153.000	4.00000	23.0000
X4	13	30.00000	16.73818	390.000	6.00000	60.0000
X5	13	95.42308	15.04372	1240.500	72.50000	115.9000

PEARSON CORRELATION COEFFICIENTS / PROB > !R! UNDER H0:RHO=0 / N = 13

	X1	X2	X3	X4	X5
X1	1.00000	0.22858	-0.82413	-0.24545	0.73072
% 3CAO.AL2O3	0.0000	0.4526	0.0005	0.4189	0.0046
X2	0.22858	1.00000	-0.13924	-0.97295	0.81625
% 3CAO.SIO2	0.4526	0.0000	0.6501	0.0001	0.0007
X3	-0.82413	-0.13924	1.00000	0.02954	-0.53467
% DE 4CAO.AL2.FE2O3	0.0005	0.6501	0.0000	0.9237	0.0598
X4	-0.24545	-0.97295	0.02954	1.00000	-0.82131
% DE 2CAO.SIO2	0.4189	0.0001	0.9237	0.0000	0.0006
X5	0.73072	0.81625	-0.53467	-0.82131	1.00000
CHALEUR (CAL/GR)	0.0046	0.0007	0.0598	0.0006	0.0000

```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
+   EXEMPLE : CALCUL D'UN TABLEAU D'EFFECTIFS CROISES,      +
+             CALCUL DE MESURES D'ASSOCIATION ENTRE         +
+             DEUX VARIABLES QUALITATIVES                   +
+  PROCEDURE: FREQ                                          +
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
```

```
DATA PANNES;
     INPUT EQUIPE $ MACHINE $ NOBS @@;
     LIST;CARDS;
          JOUR A 10 JOUR B 15 JOUR C 15 JOUR D 10
          SOIR A 10 SOIR B 20 SOIR C 15 SOIR D 20
          NUIT A 20 NUIT B 10 NUIT C 30 NUIT D 25
          ;
PROC FREQ DATA=PANNES;
     TABLES EQUIPE*MACHINE / EXPECTED DEVIATION CELLCHI2 CHISQ;
     WEIGHT NOBS;
     TITLE1 'TABLEAU DE CONTINGENCE';
     TITLE2 'ET LES MESURES D''ASSOCIATION';
```

++++++++++++++++++++++++
+ OUTPUT DE PROC FREQ +
++++++++++++++++++++++++

TABLEAU DE CONTINGENCE
ET LES MESURES D'ASSOCIATION
TABLE OF EQUIPE BY MACHINE

EQUIPE / MACHINE FREQUENCY! EXPECTED ! DEVIATION! CELL CHI2! PERCENT ! ROW PCT ! COL PCT !	A	B	C	D	TOTAL
JOUR	10	15	15	10	50
	10.0	11.3	15.0	13.8	
	0.0	3.8	0.0	-3.8	
	0	1.25	0	1.02273	
	5.00	7.50	7.50	5.00	25.00
	20.00	30.00	30.00	20.00	
	25.00	33.33	25.00	18.18	
NUIT	20	10	30	25	85
	17.0	19.1	25.5	23.4	
	3.0	-9.1	4.5	1.6	
	.529412	4.35376	.794118	.112968	
	10.00	5.00	15.00	12.50	42.50
	23.53	11.76	35.29	29.41	
	50.00	22.22	50.00	45.45	
SOIR	10	20	15	20	65
	13.0	14.6	19.5	17.9	
	-3.0	5.4	-4.5	2.1	
	.692308	1.97543	1.03846	.252622	
	5.00	10.00	7.50	10.00	32.50
	15.38	30.77	23.08	30.77	
	25.00	44.44	25.00	36.36	
TOTAL	40	45	60	55	200
	20.00	22.50	30.00	27.50	100.00

STATISTICS FOR TABLE OF EQUIPE BY MACHINE

STATISTIC	DF	VALUE	PROB
CHI-SQUARE	6	12.022	0.061
LIKELIHOOD RATIO CHI-SQUARE	6	12.800	0.046
MANTEL-HAENSZEL CHI-SQUARE	1	0.780	0.377
PHI		0.245	
CONTINGENCY COEFFICIENT		0.238	
CRAMER'S V		0.173	
SAMPLE SIZE = 200			

2.11 CONCEPTS GÉOMÉTRIQUES EN ANALYSE DES DONNÉES(*)

Tableau des données

Un ensemble de n individus (terme générique pour désigner une unité d'observation) sur lesquels on mesure p variables (quantitatives) peut se représenter par un tableau de n lignes et p colonnes en associant les lignes avec les individus et les colonnes avec les variables.

Tableau des données

		Variable 1	2	...	j	...	p
	1	x_{11}	x_{12}	...	x_{1j}	...	x_{1p}
	2	x_{21}	x_{22}	...	x_{2j}	...	x_{2p}
	.						
	.						
	.						
individu	i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ip}
	.						
	.						
	.						
	n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{np}

où x_{ij} est la valeur observée de la variable j sur l'individu i

On peut donc naturellement associer au tableau, une matrice X de dimension nxp

$$X = [x_{ij}]$$

On aurait pu choisir la matrice X^T, transposée de X

$$X^T = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \cdot & & & \\ \cdot & & & \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{bmatrix}$$

---

* Cette section n'est pas nécessaire pour la suite.

pour représenter le tableau initial des données car il s'agit de la même information. Il n'y a aucune raison fondamentale de choisir X plutôt que X^T: il faut adopter arbitrairement une convention et s'y conformer par la suite. Pour l'exposition de certains concepts la matrice X^T est plus intéressante. Nous avons retenu la matrice X car c'est le choix le plus souvent rencontré dans les ouvrages de références et de plus elle est compatible avec l'ordre employé pour lire les données avec les progiciels d'analyse de données comme SAS et BMDP.

Première interprétation: points (ou vecteurs) des individus dans l'espace des variables R^p.

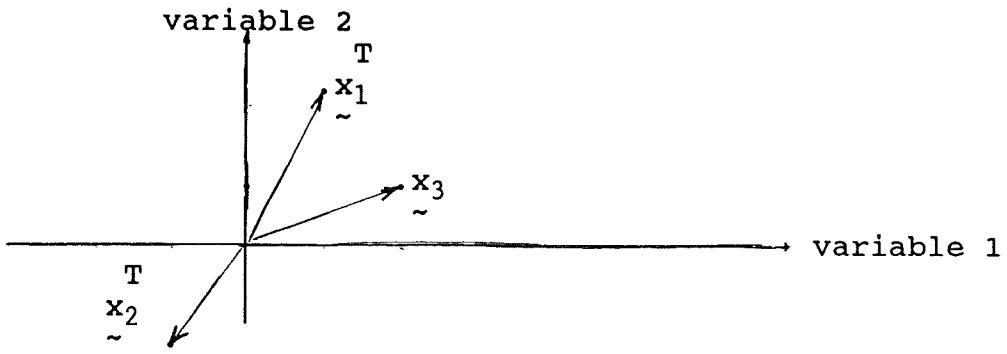
La matrice X peut s'interpréter géométriquement de deux points de vue selon ses lignes ou selon ses colonnes. Les lignes de la matrice X sont des points (vecteurs) dans l'espace des variables R^p.

$$X = \begin{bmatrix} \underset{\sim}{x}_1^T \\ \vdots \\ \underset{\sim}{x}_n^T \end{bmatrix} \qquad \underset{\sim}{x}_n^T = (x_{i1}, \ldots, x_{ip})$$

où $\underset{\sim}{x}_i$ est un vecteur px1 représentant le i-ième individu

Remarque: on utilise la représentation colonne pour les vecteurs. Une fois de plus ce choix est arbitraire mais l'important est de choisir une convention et de s'y conformer par la suite

Exemple:
$$X_{3x2} = \begin{bmatrix} 1 & 2.0 \\ -1 & -1.5 \\ 2 & 1.0 \end{bmatrix} = \begin{bmatrix} \underset{\sim}{x}_1^T \\ \underset{\sim}{x}_2^T \\ \underset{\sim}{x}_3^T \end{bmatrix}$$



L'ensemble des individus représenté par les n vecteurs $\underset{\sim}{x}_1^T, \ldots, \underset{\sim}{x}_n^T$ génère un nuage de points dans R^P. Plusieurs questions relatives à l'analyse des données se situent dans cet espace: le centre de gravité du nuage, la forme du nuage, les axes d'inertie (variabilité) du nuage, la classification et les regroupements d'individus, la comparaison de plusieurs nuages associés à plusieurs tableaux de données etc.

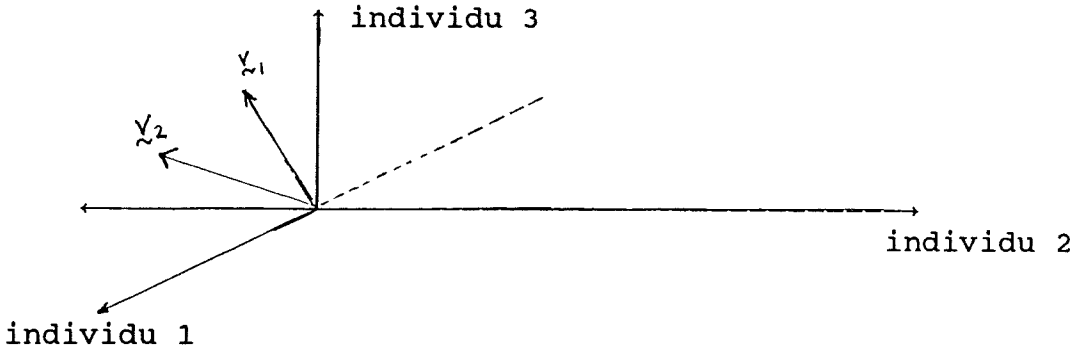
Deuxième interprétation: vecteurs dans l'espace des individus R^n

La deuxième interprétation nous est fournie en considérant les variables comme des points-vecteurs dans l'espace des individus R^n (appelé aussi espace échantillonnal)

$$X = [\underset{\sim}{v}_1, \underset{\sim}{v}_2, \ldots, \underset{\sim}{v}_p]$$

où $\underset{\sim}{v}_j = (x_{1j}, \ldots, x_{nj})^T$ représente la j-ième vecteur-variable dans R^n.

Exemple: (suite)
$$X_{3x2} = \begin{bmatrix} 1 & 2.0 \\ -1 & -1.5 \\ 2 & 1.0 \end{bmatrix} = [\underset{\sim}{v}_1, \underset{\sim}{v}_2]$$



Il est très intéressant de penser aux variables comme des vecteurs dans l'espace des individus car plusieurs concepts statistiques, comme la moyenne, la variance et la corrélation peuvent s'interpréter en termes de concepts géométriques.

Plusieurs techniques de l'analyse des données utilisent surtout cette dernière représentation car les questions à résoudre s'expriment par l'analyse des variables: dispersion des variables, liaison entre des paires de variables, prédiction d'une variable à partir de d'autres variables etc.

Interprétation géométrique des statistiques élémentaires

Le tableau initial $X = [x_{ij}]$ peut se résumer en partie par:

les moyennes: $\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_p$

les variances: $s_1^2, s_2^2, \ldots, s_p^2$

les covariances: $s_{11}, \ldots, s_{1p}, \ldots, s_{p1}, \ldots, s_{pp}$

les corrélations: $r_{11}, \ldots, r_{1p}, \ldots, r_{p1}, \ldots, r_{pp}$

où $$\bar{v}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{v}_j)^2 \qquad j=1,2,\ldots,p$$

$$s_{jj'} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{v}_j)\ (x_{ij'} - \bar{v}_{j'})$$

$$r_{jj'} = \frac{s_{jj'}}{s_j\ s_{j'}} \qquad j,j' = 1,\ldots p$$

On peut associer des concepts géométriques usuels à chacune de ces quantités en se plaçant dans l'espace des individus.

Les moyennes

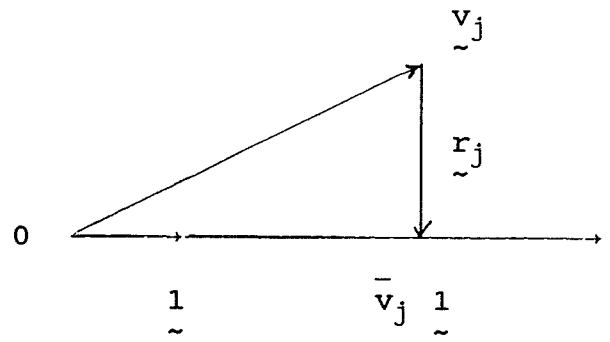
Notons par $\underset{\sim}{1} = [1,1\ldots,1]^T$ le vecteur nx1 équiangle avec les axes de coordonnées de l'espace des individus. Le vecteur $\underset{\sim}{1}$ est de longueur (euclidienne) $\sqrt{n}$ et $\frac{1}{\sqrt{n}}\underset{\sim}{1}$ est un vecteur unitaire.

Le vecteur projection (voir rappel ci-bas) de la j-ième variable $\underset{\sim}{v}_j$ est donc le vecteur

$$\underset{\sim}{v}_j^T \left(\frac{1}{\sqrt{n}}\underset{\sim}{1} \right) \frac{1}{\sqrt{n}}\underset{\sim}{1} = \frac{1}{n}\left(\underset{\sim}{v}_j^T \underset{\sim}{1} \right)\underset{\sim}{1}$$

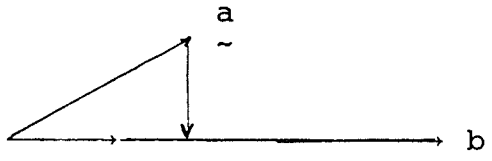
$$= \frac{1}{n}\left(\sum_{i=1}^{n} x_{ij} \right)\underset{\sim}{1}$$

$$= \bar{v}_j \underset{\sim}{1}$$



Rappel: le vecteur projection du vecteur $\underset{\sim}{a}$ sur le vecteur $\underset{\sim}{b}$ est

$$(\underset{\sim}{a}^T \underset{\sim}{b}) \frac{\underset{\sim}{b}}{\sqrt{\underset{\sim}{b}^T\underset{\sim}{b}}}$$



On peut écrire l'équation de décomposition de $\underset{\sim}{v}_j$

$$\underset{\sim}{v}_j = \bar{v}_j \ \underset{\sim}{1} + \underset{\sim}{r}_j$$

où

$$\underset{\sim}{r}_j = \underset{\sim}{v}_j - \bar{v}_j \ \underset{\sim}{1}$$

est le <u>vecteur résiduel</u>. Les vecteurs $\bar{v}_j \ \underset{\sim}{1}$ et $\underset{\sim}{r}_j$ sont orthogonaux car

$$(\bar{v}_j \underset{\sim}{1})^T \ \underset{\sim}{r}_j = \bar{v}_j \ \underset{\sim}{1}^T \ \underset{\sim}{v}_j - \bar{v}_j^2 \ \underset{\sim}{1}^T \underset{\sim}{1}$$

$$= \bar{v}_j \ n\bar{v}_j - \bar{v}_j^2 \ n = 0$$

<u>Exemple</u> (suite)

$$X = \begin{bmatrix} 1 & 2.0 \\ -1 & -1.5 \\ 2 & 1.0 \end{bmatrix}$$

$$\bar{v}_1 = 2/3 \qquad \bar{v}_2 = 0.5$$

$$\bar{v}_1 \ \underset{\sim}{1} = \begin{bmatrix} 2/3 \\ 2/3 \\ 2/3 \end{bmatrix} \qquad \bar{v}_2 \ \underset{\sim}{1} = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}$$

$$\underset{\sim}{r}_1 = \begin{bmatrix} 1/3 \\ -5/3 \\ 4/3 \end{bmatrix} \qquad \underset{\sim}{r}_2 = \begin{bmatrix} 1.5 \\ -2.0 \\ 0.5 \end{bmatrix}$$

Les vecteurs résiduels $\underset{\sim}{r}_1, \ldots, \underset{\sim}{r}_p$ sont obtenus de $\underset{\sim}{v}_1, \ldots, \underset{\sim}{v}_p$ par une projection sur le vecteur équiangle $\underset{\sim}{1}$. La moyenne de la de la j-ième variable $\bar{v}_j$ est la projection du j-ième vecteur-variable $\underset{\sim}{v}_j$ sur le vecteur équiangle $\underset{\sim}{1} = (1, \ldots, 1)^T$. Le vecteur-projection $\bar{v}_j \ \underset{\sim}{1}$ de la j-ième variable a une longueur égale à $\sqrt{n} \ |\bar{v}_j|$.

Les variances

La variance de la j-ième variable $\underset{\sim}{v}_j$ est

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{v}_j)^2$$

$$= \frac{1}{n-1} (\underset{\sim}{v}_j - \bar{v}_j \underset{\sim}{1})^T (\underset{\sim}{v}_j - \bar{v}_j \underset{\sim}{1})$$

$$= \frac{1}{n-1} \underset{\sim}{r}_j^T \underset{\sim}{r}_j$$

$$= \frac{1}{n-1} ||\underset{\sim}{r}_j||^2$$

où $||\underset{\sim}{r}_j||$ est la norme euclidienne du vecteur résiduel $\underset{\sim}{r}_j$.

Une variance est donc, à une constante multiplicative près, le carré de la norme du vecteur-variable résiduel $\underset{\sim}{r}_j$ après projection du vecteur-variable $\underset{\sim}{v}_j$ sur le vecteur $\bar{v}_j \underset{\sim}{1}$

Exemple $$X = \begin{bmatrix} 1 & 2.0 \\ -1 & -1.5 \\ 2 & 1.0 \end{bmatrix}$$

$$||\underset{\sim}{r}_1||^2 = \frac{1}{9} + \frac{25}{9} + \frac{16}{9} = \frac{42}{9} = 2*s_1^2$$

$$||\underset{\sim}{r}_2||^2 = (1.5)^2 + (-2.0)^2 + (0.5)^2 = 6.50 = 2*s_1^2$$

Les corrélations

La matrice des variances-covariances S est la matrice d'ordre p définie par

$$S = [s_{jj'}]$$

où
$$s_{jj'} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{v}_j)(x_{ij'} - \bar{v}_{j'})$$

$$= \frac{1}{n-1} \underset{\sim}{r}_j^T \underset{\sim}{r}_{j'}$$

La covariance $s_{jj'}$ est, à une constante multiplicative près, le produit scalaire entre les vecteurs résiduels $\underset{\sim}{r}_j$ et $\underset{\sim}{r}_{j'}$. Notons par $\theta_{jj'}$ l'angle entre les vecteurs $\underset{\sim}{r}_j$ et $\underset{\sim}{r}_{j'}$. On a

$$\text{Cos}(\theta_{jj'}) = \frac{\underset{\sim}{r}_j^T \underset{\sim}{r}_{j'}}{\sqrt{||\underset{\sim}{r}_j||^2 ||\underset{\sim}{r}_{j'}||^2}}$$

$$= \frac{(n-1)\, s_{jj'}}{\sqrt{(n-1)s_{jj} * (n-1)s_{j'j'}}}$$

$$= \frac{s_{jj'}}{\sqrt{s_{jj} * s_{j'j'}}}$$

$$= r_{jj'}$$

Donc le coefficient de corrélation entre la variable j et la variable j' est le cosinus de l'angle entre le vecteur-variable résiduel $\underset{\sim}{r}_j$ et le vecteur-variable $\underset{\sim}{r}_{j'}$

Exemple
(suite)

$$X = \begin{bmatrix} 1 & 2.0 \\ -1 & -1.5 \\ 2 & 1.0 \end{bmatrix}$$

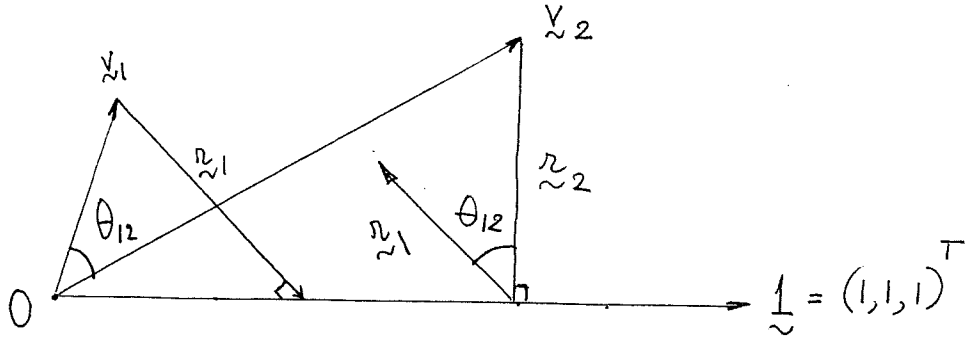
$$\underset{\sim}{r}_1 = \begin{bmatrix} -1/3 \\ -5/3 \\ 4/3 \end{bmatrix} \qquad \underset{\sim}{r}_2 \begin{bmatrix} 1.5 \\ -2.0 \\ 0.5 \end{bmatrix}$$

$$||\underset{\sim}{r}_1|| = \sqrt{42/9} \qquad ||\underset{\sim}{r}_2|| = \sqrt{6.50}$$

$$\underset{\sim}{r}_1^T \underset{\sim}{r}_2 = 10.5/3$$

$$\text{Cos } (\theta_{12}) = \frac{10.5/3}{\sqrt{42/9} * \sqrt{6.5}} = 0.635 = r_{12}$$

$\theta_{12} = 50.5$ degrés est l'angle entre les vecteurs-variables $\underset{\sim}{v}_1$ et $\underset{\sim}{v}_2$.



plan déterminé par $\underset{\sim}{1}$ et $\underset{\sim}{V}_2$
dans l'espace des individus R^3

CHAPITRE 3

PROBABILITÉS

3.0 SOMMAIRE

Ce chapitre contient une brève introduction à la théorie des probabilités dont l'objectif est de traiter des situations et des événements impliquant l'incertitude et le hasard. Parmi les applications à l'analyse des données, il faut noter la modélisation du processus d'échantillonage et l'évaluation de risques dans la prise de décisions. On définit les espaces de probabilité de façon axiomatique et on présente des résultats qui en découlent: les concepts de probabilité conditionnelle, loi de multiplication, indépendance, probabilité totale et formule de Bayes.

3.1 QUELQUES RÈGLES ET FORMULES DE DÉNOMBREMENT

Nous rappelons quelques formules de dénombrement et utiles pour le calcul des probabilités.

Règle de multiplication pour les choix successifs

Si un choix est effectué en deux étapes et qu'il y a k_1 façons de réaliser la première et k_2 façons de réaliser la seconde, il y a $k_1 * k_2$ façons de réaliser le choix. Par exemple, le nombre de choix possibles dans un menu comprenant 5 entrées, 3 plats principaux et 4 desserts est de 5*3*4=60.

Arrangement de n objets distincts

Une liste ordonnée de $k \leq n$ objets utilisés parmi les n est un arrangement; si tous les objets sont employés ($k = n$) l'arrangement est dit permutation. Le nombre d'arrangements de n objets pris k à la fois est de

$$n(n-1) \ldots (n-k+1) = n!/(n-k)! \qquad (3.1)$$

où $k! = k(k-1)(k-2)\ldots 3.2.1$; $0!=1$

Le nombre de permutations de n objets est n!

Par exemple, le nombre d'échantillons (sous-ensembles) sans remise de taille k pris dans une population de n individus est un arrangement si l'ordre du tirage est pris en compte. Si les tirages sont effectués avec remise, le nombre d'échantillons est de n^k.

Combinaison de n objets distincts

Le résultat du choix de k objets parmi n est une combinaison de n objets pris k à la fois si l'ordre des tirages n'est pas pris en compte. Le nombre de combinaisons de n objets pris k à la fois sera noté

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad (3.2)$$

Par exemple, le nombre d'échantillons sans remise de taille k d'une population de n individus est $\binom{n}{k}$ si l'on ne tient pas compte de l'ordre des tirages.

Si l'échantillonnage est effectué avec remise et sans tenir compte de l'ordre, le nombre de possibilités est de $\binom{n+k-1}{k}$

Tableau 3.1: nombre de sous-ensembles de k objets distincts choisis parmi n.

tirage (échantillonnage) \ ordre importe	oui	non
sans remise	$\frac{n!}{(n-k)!}$	$\frac{n!}{k!(n-k)!} = \binom{n}{k}$
avec remise	n^k	$\binom{n+k-1}{k}$

3.2 ESPACES DE PROBABILITÉS

Événements

On constate que l'on peut classer les événements de toutes sortes en:

- événements certains
- événements impossibles
- événements incertains.

Les événements incertains, forment la très grande majorité, et ne le sont pas tous au même degré. Le degré de réalisation d'un événement incertain est représenté par un nombre entre 0 et 1 appelé la probabilité de l'événement. Un événement impossible a une probabilité de 0 et un événement sûr a une probabilité de 1.

Les événements sont associés à des expériences dont les résultats ne sont pas prévisibles d'avance et appelées EXPÉRIENCES ALÉATOIRES. Afin de traiter une expérience on associe un ensemble qui sera ultérieurement muni d'une mesure de probabilité pour devenir un ESPACE DE PROBABILITÉS. Voici quelques expériences auxquelles on a associé un espace noté Ω.

EXPÉRIENCE	ESPACE
1. jet répété 3 fois d'une pièce de monnaie	$\Omega_1 = \left\{ \begin{array}{l} \text{PPP, PPF, PFP, FPP,} \\ \text{FFP, FPF, PFF, FFF} \end{array} \right\}$ où P ≡ pile F ≡ face
2. jet paire de dés à jouer	$\Omega_2 = \left\{ \begin{array}{llll} (1,1), & (1,2), & \ldots, & (1,6) \\ (2,1), & (2,2), & \ldots, & (2,6) \\ \ldots, & \ldots, & \ldots, & \ldots \\ (6,1), & (6,2), & \ldots, & (6,6) \end{array} \right\}$ où (i,j) signifie un résultat i sur le 1er dé et un résultat j sur le 2ième dé.
3. tirages <u>avec</u> remise de deux articles dans un lot de N articles dont M sont défectueux	$\Omega_3 = \{(i,j): 1 \leq i,j \leq N\}$
4. tirage <u>sans</u> remise de deux articles dans un lot de N articles dont M sont défectueux	$\Omega_4 = \left\{ \begin{array}{l} (i,j): 1 \leq i \leq N \\ \qquad\quad 1 \leq j \leq N \\ \qquad\quad j \neq i \end{array} \right\}$
5. tirage avec remise d'articles dans un lot jusqu'à l'obtention d'un article défectueux	$\Omega_5 = \{1, 2, \ldots\}$ où i = numéro du tirage pour obtenir le premier article défectueux
6. LOTO 6/49	$\Omega_6 = \left\{ \begin{array}{l} (i_1,i_2,i_3,i_4,i_5,i_6): \\ 1 \leq i_k \leq 49, k = 1,2,\ldots 6 \\ i_1 \neq i_2 \neq i_3 \neq i_4 \neq i_5 \neq i_6 \end{array} \right\}$
7. - durée ampoule électrique - tension de rupture fil - usure d'un piston - force de compression d'un béton	Ω_7 = R = ensemble de nombres réels
8. - nombre d'articles défectueux dans un échantillon de taille n pris sans remise dans un lot de N articles ($n \leq N$)	$\Omega_8 = \{0, 1, 2, \ldots, n\}$
9. - nombre de défauts dans une plaque d'acier de 1 m^2	$\Omega_9 = \{0, 1, 2, \ldots\}$

À une expérience aléatoire on associe un espace ou ensemble ou référentiel Ω comprenant tous les résultats possibles

$$\Omega = \{\omega_1, \omega_2, \ldots\}$$

Les éléments ω_α de Ω sont des ÉVÉNEMENTS ALÉATOIRES ÉLÉMENTAIRES et les ÉVÉNEMENTS ALÉATOIRES COMPOSÉS E sont les sous-ensembles de Ω formé par des réunions d'événements aléatoires élémentaires.

$$E \subset \Omega$$

Exemple 3.1 jet d'une paire de dés

$\Omega = \{(i,j): \ 1 \le i \le 6, \quad 1 \le j \le 6\}$

Ω contient 36 événements élémentaires, $\omega_{ij} = \{(i,j)\}$

Voici quelques exemples d'événements composés

E_1 = somme des 2 dés égale 7
$= \{(i,j): \ i + j = 7\}$
$= \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$

E_2 = somme des 2 dés égale 2, 3 ou 12
$= \{(i,j): \ i + j = 2, 3, 12\}$
$= \{(1,1), (1,2), (2,1) (6,6)\}$

E_3 = somme des 2 dés est comprise entre 2 et 12
$= \{(i,j): \ 2 \le i + j \le 12\}$
$= \{(1,1), (1,2), \ldots, (6,6)\}$
$= \Omega$ = événement certain

E_4 = somme des 2 dés est supérieure à 12
$= \phi$ = ensemble vide = événement impossible

On peut classer les événements de Ω selon le nombre d'éléments qu'ils contiennent. Ainsi on a

$$\binom{36}{k} = \frac{36!}{k!(36-k)!}$$

événements contenant k éléments de Ω
$k = 0, 1, \ldots, 36$

et le nombre total d'événements de Ω est

$$\sum_{k=0}^{36} \binom{36}{k} = \sum_{k=0}^{36} \binom{36}{k} 1^k \, 1^{36-k} = (1 + 1)^{36} = 2^{36}$$

$$= 68\ 719\ 476\ 736$$

En général le nombre d'événements élémentaires et composés d'un espace fini Ω de cardinalité n est 2^n.

Exemple 3.2 LOTO 6/49

$$\Omega = \left\{ (i_1, i_2, i_3, i_4, i_5, i_6): \begin{array}{c} 1 \leq i_\alpha \leq 49, \alpha = 1,2, \ldots,6 \\ i_1 \neq i_2 \neq i_3 \neq i_4 \neq i_5 \neq i_6 \end{array} \right\}$$

Le calcul du nombre d'événements élémentaires associés aux lots gagnants donne:

Événements	Nombre d'événements élémentaires
$E_1 \equiv \Omega$	$\binom{49}{6} = \frac{49!}{6!43!} = 13\ 983\ 816$
$E_2 \equiv$ 6 sur 6	$\binom{6}{6}\binom{43}{0} = 1$
$E_3 \equiv$ 5 sur 6 +	$\binom{6}{5}\binom{1}{1}\binom{42}{0} = 6$
$E_4 \equiv$ 5 sur 6	$\binom{6}{5}\binom{1}{0}\binom{42}{1} = 252$
$E_5 \equiv$ 4 sur 6	$\binom{6}{4}\binom{43}{2} = 13545$
$E_6 \equiv$ 3 sur 6	$\binom{6}{3}\binom{43}{3} = 246820$

La justification de ces calculs repose sur les règles et formules de dénombrement vues à la section 3.1. Les 49 numéros de 1 à 49 sont classés en deux catégories: les 6 numéros gagnants et les autres. Les tirages sont effectués sans remise et sans tenir compte de l'ordre. Dans la catégorie 5 sur 6+ et 5 sur 6 le calcul demande de séparer les numéros en trois catégories: les 6 numéros gagnants, le numéro complémentaire et les autres.

Exemple 3.3 tension de rupture fil métallique

$\Omega = R^+$, la droite réelle positive

E_1 = tension de rupture plus grande que 150

$E_1 = (150, \infty)$

E_2 = tension de rupture entre 130 et 150 inclusivement

$E_2 = [130, 150]$

Dans cet exemple l'espace Ω contient une infinité non-dénombrable d'événements élémentaires que sont les nombres réels. Comme nous le verrons par la suite, les phénomènes aléatoires à valeurs numériques sont très importants pour les applications.

Algèbre des événements

Soient A, B deux événements quelconques de Ω . On définit:

A se réalise	si le résultat de l'expérience aléatoire est un des événements élémentaires qui le composent
$A = B$	s'il contiennent les mêmes événements élémentaires
$A \cup B$	l'événement réunion de A et B; il se réalise si au moins un des deux événements A ou B se réalise
$A \cap B$	l'événement intersection de A et B; il se réalise si les deux événements A et B se réalisent conjointement
A'	l'événement contraire de A; il se réalise quand A ne se réalise pas
$A \subset B$	B est réalisé chaque fois que A l'est
ϕ	événement impossible (jamais réalisé)
Ω	événement certain (toujours réalisé)
$A \cap B = \phi$	les événements A et B sont INCOMPATIBLES

Une suite d'événements $A_1, A_2, \ldots, A_k$ s'appelle une PARTITION de Ω si

$$A_\alpha \cap A_\beta = \phi \qquad \alpha \neq \beta$$
$$\bigcup_{\alpha=1}^{k} A_\alpha = \Omega \tag{3.3}$$

Voici quelques identités utiles faciles à vérifier à l'aide de diagrammes de Venn:

$$A \cup A' = \Omega \ , \quad A \cap A' = \phi$$
$$A = (A \cap B) \cup (A \cap B')$$
$$A \cup B = A \cup (B \cap A')$$
$$A \cup B = (A \cap B') \cup (A' \cap B) \cup (A \cap B) \tag{3.4}$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$(A \cup B)' = A' \cap B'$$
$$(A \cap B)' = A' \cup B'$$

Mesures de probabilités

Depuis Kolmogorov (1933) on présente la probabilité sous la forme axiomatique, représentation idéalisée de phénomènes observables. On introduit, sans les définir, des concepts fondamentaux et on énonce des propositions qui décrivent les propriétés qui les gouvernent. Ces concepts et propriétés ne sont pas arbitraires mais sont inspirés par des notions expérimentales. Pour que la théorie ait une utilité pratique, il faut que les résultats démontrés de la théorie puissent être confrontés à des faits observés.

Définition d'une mesure de probabilité

Soit Ω un ensemble. L'application notée P est une MESURE DE PROBABILITÉ (ou simplement PROBABILITÉ) si elle est une fonction réelle définie pour tout événement A et telle que

$$P : A \longrightarrow R$$

(1) $0 \leq P(A) \leq 1$

(2) $P(\Omega) = 1$ (3.5)

(3) Si $A_1 \cap A_2 = \phi$ alors

$$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

En résumé, la mesure de probabilité P est positive, normée et additive. Ces propriétés s'inspirent de propriétés analogues que possèdent les FRÉQUENCES RELATIVES.

Définition d'une fréquence relative

Soit une expérience aléatoire et Ω l'espace associé et A un événement de Ω .

Répétons l'expérience n fois et notons par n(A) le nombre de fois que A s'est réalisé dans la suite des n répétions. On définit

$f(A) = n(A)/n$ = fréquence relative de A. (3.6)

On observe que

(1) $0 \leq f(A) \leq 1$

(2) $f(\Omega) = 1$

et que si A_1 et A_2 sont deux événements incompatibles

(3) $f(A_1 \text{ ou } A_2) = f(A_1) + f(A_2)$

Remarques:

- f(A) dépend de la séquence de réalisation et une autre séquence de n répétions donne, en général, une valeur différente à f(A)
- la valeur de f(A) tend à se stabiliser quand n devient grand
- f(A) est un fait observable dans la réalité tandis que P(A) représente un postulat (modèle)

Pour une expérience aléatoire Ω on peut définir plusieurs applications P qui soient des mesures de probabilités. La théorie ne dit rien sur le choix de P. Un ESPACE DE PROBABILITÉS est formé par le couple (Ω,P) où Ω est un référentiel et P une mesure de probabilités.

Décider si une mesure P est une modélisation valable pour les applications consiste à confronter P(A) avec f(A). D'une manière générale, l'ensemble des procédures permettant la confrontation entre les modèles probabilistes et les données empiriques ou expérimentales est l'ANALYSE STATISTIQUE composée de deux sortes de procédures: estimation des paramètres et tests d'hypothèses.

Autres définitions de la probabilité

$$\text{classique:} \quad P(A) = \frac{\text{nombre de cas favorables à A}}{\text{nombre de cas possibles}}$$

$$\text{statistique:} \quad P(A) = \lim_{n\to\infty} n(A)/n$$

Ces deux autres définitions de la probabilité présentent des difficultés. Par exemple la définition classique ne s'applique que pour des espaces finis et présuppose les modèles d'égales probabilités pour les événements élémentaires. La définition statistique n'a jamais été justifiée formellement.

La définition axiomatique de probabilité est une notion purement mathématique définie de façon abstraite et la théorie décrète que chaque événement est affecté d'un nombre appelé probabilité de l'événement. Connaître la valeur de la probabilité d'un événement ou savoir s'il existe des moyens pour l'obtenir ne sont pas des questions relevant de la théorie. L'objectif fondamental de la théorie est de développer des moyens de calculer la probabilité de certains événements A après avoir postulé un espace de probabilités (Ω,P).

Il y a trois questions essentielles reliées à l'étude et l'utilisation des probabilités:

. quel est le sens d'une affirmation telle par exemple la probabilité de l'événement est 0.90; la réponse est généralement donnée en termes de fréquences relatives;

. comment utiliser des probabilités connues pour calculer la probabilité de d'autres événements.

. l'estimation des paramètres de modèles de probabilité à l'aide de données expérimentales.

3.3 CLASSIFICATION DES ESPACES

On peut classer les espaces de probabilités selon le cardinal de Ω.

(A) Ω contient un nombre fini d'événements élémentaires

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$$

Une mesure de probabilité P souvent proposée est l'ÉQUIPROBABILITÉ

$$P(\{\omega_\alpha\}) = 1/k \qquad \alpha = 1,\ldots,k \tag{3.7}$$

Cette mesure repose sur le principe intuitif de symétrie selon lequel tous les événements élémentaires sont également vraisemblables. L'équiprobabilité est une règle assez générale dans les jeux de hasard telles les arrivées de pile ou face dans le jet d'une pièce de monnaie.

Dans ce cas les problèmes de calculs de probabilité se ramènent à des problèmes de dénombrement, c'est-à-dire compter le nombre d'événements élémentaires qui réalisent A .

Exemple 3.4: Loto 6/49, suite de l'exemple 3.2

Si on suppose le modèle d'équiprobabilité sur Ω on a

$$P(A) = \frac{\text{card}(A)}{\text{card}(\Omega)} = \frac{\text{card}(A)}{13\ 983\ 816}$$

où card(A) est le nombre d'événements élémentaires de A.

Ainsi les probabilités des événements associés aux lots gagnants sont:

Événement	card(A)	Probabilité
6 sur 6	1	0.000 000 07
5 sur 6+	6	0.000 000 43
5 sur 6	252	0.000 018 02
4 sur 6	13545	0.000 968 41
3 sur 6	246820	0.017 650 40

(B) Ω contient un nombre dénombrable d'événements élémentaires

$$\Omega = \{\omega_1, \omega_2, \ldots\}$$

Une mesure de probabilité sur Ω peut se définir sur les événements élémentaires en posant

$$p_\alpha = P(\{\omega_\alpha\})$$

$$\sum_{\alpha=1}^{\infty} p_\alpha = 1 \qquad (3.8)$$

Dans les applications on rencontre souvent

$$\Omega = \{1, 2, 3, \ldots\}$$

et $$\Omega = \{0, 1, 2, \ldots\}$$

Il s'agit de VARIABLE ALÉATOIRE DISCRÈTE dont voici deux exemples.

Exemple 3.5: Soit X le nombre de jeux ou essais avant d'obtenir le chiffre 1 avec un dé à jouer. On peut poser

$$X = \Omega = \{1, 2, \ldots\}$$

où le chiffre représente le nombre de jeux effectués. En supposant l'équiprobabilité de chaque face, on peut proposer le modèle

$$P(\{k\}) = \left(\frac{5}{6}\right)^{k-1} \frac{1}{6} \qquad k = 1, 2, \ldots$$

On vérifie facilement que cette mesure de probabilité satisfait l'équation (3.8).

Cette fonction de probabilité s'appelle loi de probabilité GÉOMÉTRIQUE puisque les probabilités sont générées par une progression géométrique de raison 5/6.

Exemple 3.6: Soit X le nombre de défauts sur une plaque d'acier ayant une surface de s mètres carrés

$$X = \Omega = \{0,1,2,\ldots\}$$

Sous certaines hypothèses qui seront exposées plus loin sous le nom de PROCESSUS DE POISSON on obtient le modèle de probabilité

$$P[\{x\}] = P[X=x] = \frac{\exp(-\lambda s)(\lambda s)^x}{x!} \qquad (3.9)$$

$$x = 0,1,2,\ldots \qquad \lambda > 0$$

Le paramètre λ s'appelle l'intensité du processus et représente le nombre moyen de défauts par mètres carrés. Cette mesure de probabilité s'appelle loi de POISSON et sera étudiée au chapitre 5.

(C) Ω contient un nombre non-dénombrable d'événements élémentaires

Dans les applications Ω est un intervalle de nombres réels ou l'ensemble R des nombres réels ou encore R^n. Les définitions et propriétés introduites jusqu'ici pour les espaces finis ou dénombrables s'étendent mais au prix d'un effort important de formalisation mathématique.

Une des difficultés réside dans le fait que la mesure de probabilité ne peut pas se définir sur tous les sous-ensembles de Ω mais sur une classe de sous-ensembles possédant une propriété d'additivité. Sur la droite réelle cette classe est formée de l'ensemble des intervalles semi-ouverts [a,b[et tous les ensembles obtenus par réunion, intersection et passage au complémentaire.

Dans la pratique on parle de VARIABLE ALÉATOIRE CONTINUE si $\Omega = R$ et la mesure de probabilité est définie par une fonction DENSITÉ DE PROBABILITÉ $f_X(x)$ ayant les propriétés suivantes

$$f_X(x) \geq 0$$
$$\int_{-\infty}^{\infty} f_X(x)dx = 1 \qquad (3.10)$$

La mesure de probabilité d'un événement A est définie par

$$P(A) = \int_A f_X(x)dx \qquad (3.11)$$

où A appartient à la classe définie plus haut.

Voici un exemple d'une famille de densité de probabilité définie par deux paramètres μ et σ .

Exemple 3.7: densité gaussienne ou normale

$$f_X(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right] \qquad (3.12)$$

$$-\infty < x < \infty , \quad -\infty < \mu < \infty , \quad \sigma > 0$$

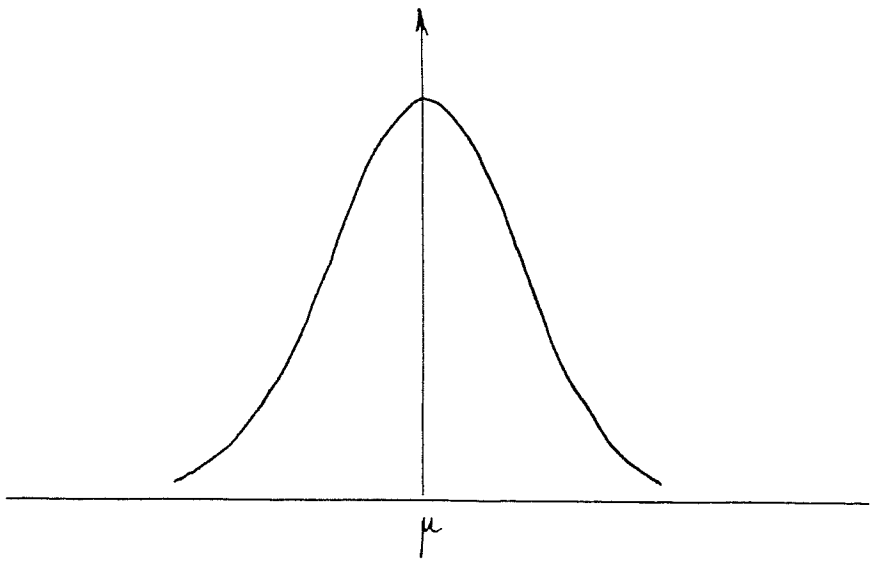


Figure 3.1: distribution gaussienne

3.4 CONSÉQUENCES DES AXIOMES DE PROBABILITÉS

Les axiomes (1) - (2) - (3) d'une mesure de probabilités conduisent aux conséquences suivantes que nous ne démontrons pas.

Proposition 3.1

(a) Si $A_\alpha \cap A_\beta = \phi, \quad \alpha \neq \beta$ alors

$$P\left[\bigcup_{\alpha=1}^{k} A_\alpha\right] = \sum_{\alpha=1}^{k} P(A_\alpha) \qquad k = 2,3,\ldots$$

(b) Si $A \subset B$ alors $P(A) \leq P(B)$

(c) $P(A') = 1 - P(A)$

En particulier $P(\phi) = 0$

(d) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(e) $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

(f) $$P\left[\bigcup_{\alpha=1}^{k} A_\alpha\right] = \sum_{\alpha=1}^{k} P(A_\alpha) - \sum_{\alpha<\beta} P(A_\alpha \cap A_\beta) + \sum_{\alpha<\beta<\gamma} P(A_\alpha \cap A_\beta \cap A_\gamma) - \ldots + (-1)^{k-1} P\left[\bigcap_{\alpha=1}^{k} A_\alpha\right]$$

Remarque: La propriété (a), appelée additivité finie, est une conséquence immédiate de l'axiome (3). Dans certains types de calculs de probabilités, il est nécessaire de postuler un axiome d'additivité dénombrable beaucoup plus fort que (3) soit:

(3)' $A_\alpha \cap A_\beta = \phi \qquad \alpha \neq \beta \qquad \alpha,\beta = 1,2,\ldots$

$$P\left[\bigcup_{\alpha=1}^{\infty} A_\alpha\right] = \sum_{\alpha=1}^{\infty} P(A_\alpha) \tag{3.13}$$

Exemple 3.8: Une firme de génie-conseil a soumissionné sur trois projets. Définissons les événements:

A_α : le projet α est obtenu $\alpha = 1, 2, 3$

et une mesure de probabilité par le tableau

événement	A_1	A_2	A_3	A_1 et A_2
probabilité	0.22	0.25	0.28	0.11

événement	A_1 ou A_3	A_2 et A_3	A_1 et A_2 et A_3
probabilité	0.45	0.07	0.01

On demande de

(a) définir l'espace Ω

(b) calculer la probabilité d'obtenir au moins un projet

(c) calculer la probabilité d'obtenir le troisième projet seulement

Solution:

(a) Le diagramme de Venn permet de visualiser les événements A_1, A_2, A_3 ainsi que les 8 événements élémentaires ω_α qui composent $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8\}$

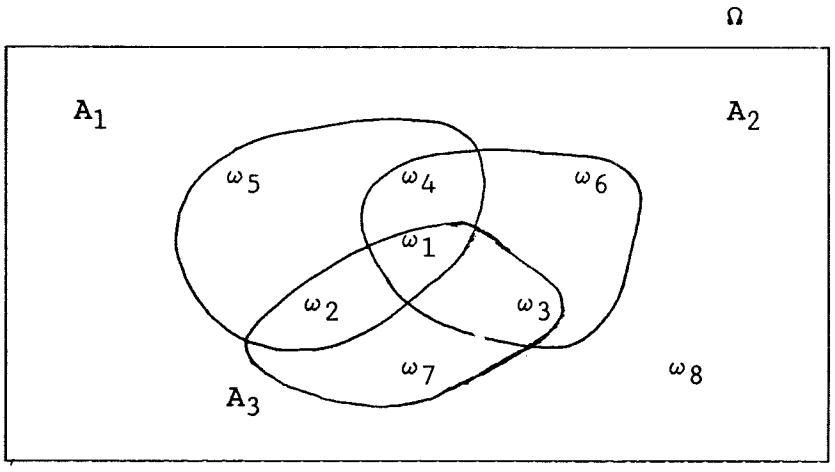


Figure 3.2: diagramme de Venn

où $\omega_1 = A_1 \cap A_2 \cap A_3$ $\omega_2 = A_1 \cap A_2' \cap A_3$

$\omega_3 = A_1' \cap A_2 \cap A_3$ $\omega_4 = A_1 \cap A_2 \cap A_3'$

$\omega_5 = A_1 \cap A_2' \cap A_3'$ $\omega_6 = A_1' \cap A_2 \cap A_3'$

$\omega_7 = A_1' \cap A_2' \cap A_3$ $\omega_8 = A_1' \cap A_2' \cap A_3'$

Les événements A_1, A_2, A_3 et leurs intersections peuvent s'écrire:

$$A_1 = \{\omega_1, \omega_2, \omega_4, \omega_5\} \quad A_2 = \{\omega_1, \omega_3, \omega_4, \omega_6\}$$

$$A_3 = \{\omega_1, \omega_2, \omega_3, \omega_7\}$$

$$A_1 \cap A_2 = \{\omega_1, \omega_4\}\ ,\ A_1 \cap A_3 = \{\omega_1, \omega_2\}$$

$$A_2 \cap A_3 = \{\omega_1, \omega_3\}$$

Du tableau $P[\{\omega_1\}] = 0.01$ et, en utilisant l'additivité de la mesure P

$$P[\{\omega_4\}] = P(A_1 \cap A_2) - P[\{\omega_1\}] = 0.11 - 0.01 = 0.10$$

$$P[\{\omega_3\}] = P(A_2 \cap A_3) - P[\{\omega_1\}] = 0.07 - 0.01 = 0.06$$

À l'aide de la propriété (d) de la proposition 3.1

$$P(A_1 \cap A_3) = P(A_1) + P(A_3) - P(A_1 \cup A_3)$$

$$= 0.22 + 0.28 - 0.45 = 0.05$$

On obtient ainsi les probabilités des autres événements élémentaires

$$P[\{\omega_2\}] = P(A_1 \cap A_3) - P[\{\omega_1\}] = 0.05 - 0.01 = 0.04$$

$$P[\{\omega_5\}] = P(A_1) - P(\{\omega_1, \omega_2, \omega_4\})$$

$$= 0.22 - P(\{\omega_1\}) - P(\{\omega_2\}) - P(\{\omega_4\})$$

$$= 0.22 - 0.01 - 0.04 - 0.10 = 0.07$$

$$P[\{\omega_6\}] = P(A_2) - P[\{\omega_1, \omega_3, \omega_4\}]$$

$$= 0.25 - 0.01 - 0.06 - 0.10 = 0.08$$

$$P[\{\omega_7\}] = P(A_3) - P[\{\omega_1, \omega_2, \omega_3\}]$$

$$= 0.28 - 0.01 - 0.04 - 0.06$$

$$= 0.17$$

$$P[\{\omega_8\}] = P(\Omega) - \sum_{\alpha=1}^{7} P[\{\omega_\alpha\}] = 0.47$$

La mesure de probabilité implicitement définie par le tableau initial a été reconstituée sur les événements élémentaires ω_α

événement élémentaire	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
mesure de probabilité	0.01	0.04	0.06	0.10	0.07	0.08	0.17	0.47

(b) P(au moins un des trois projets est obtenu)

$$= P(A_1 \cup A_2 \cup A_3) = P(\{\omega_1, \omega_2, \ldots, \omega_7\})$$

$$= \sum_{\alpha=1}^{7} P(\{\omega_\alpha\}) = 0.53$$

(c) P(troisième projet seulement est obtenu)

$$= P(A_1' \cap A_2' \cap A_3) = P(\{\omega_7\}) = 0.17$$

3.5 PROBABILITÉ CONDITIONNELLE ET INDÉPENDANCE

Sachant qu'un événement B s'est réalisé, on veut évaluer la probabilité de réalisation d'un autre événement A . Cette probabilité notée $P(A|B)$ sera appelée PROBABILITÉ CONDITIONNELLE de A étant donné B réalisé.

Définition de la probabilité conditionnelle

$$P(A|B) = P(A \cap B)/P(B) \quad \text{avec} \quad P(B) > 0 \qquad (3.14)$$

On peut justifier cette définition en examinant le comportement des fréquences relatives. Considérons n répétitions d'une expérience associée aux événements A et B et posons

$n(B)$: le nombre de fois que B est réalisé

$n(A \cap B)$: le nombre de fois que A et B se sont réalisés conjointement

on a

$f(A \cap B) = n(A \cap B)/n$: fréquence relative de $A \cap B$

$f(B) = n(B)/n$: fréquence relative de B

La fréquence relative de A parmi les résultats où B s'est réalisée est

$$f(A|B) = n(A \cap B)/n(B) = \frac{n(A \cap B)/n}{n(B)/n} = f(A \cap B)/f(B)$$

En fait l'espace de probabilité initial (Ω, P) définit un nouvel espace de probabilité $(B, P(.|B))$ pour tout événement B tel que $P(B) > 0$. Il est important de remarquer que $P(.|B)$ est une mesure de probabilité satisfaisant aux axiomes (1) - (2) - (3)

$$P(\Omega|B) = P(\Omega \cap B|B) = P(B|B) = P(B)/P(B) = 1$$

Si $A_1 \cap A_2 = \phi$

$$P(A_1 \cup A_2|B) = P((A_1 \cup A_2) \cap B)/P(B)$$

$$= [P(A_1 \cap B) + P(A_2 \cap B)]/P(B)$$

$$= P(A_1|B) + P(A_2|B)$$

Si $P(A) > 0$ et $P(B) > 0$ on peut écrire

$$P(A \cap B) = P(B)\ P(A|B)$$
$$P(A \cap B) = P(A)\ P(B|A) \qquad (3.15)$$

appelée LOI DE MULTIPLICATION.

Il peut arriver que la réalisation d'un événement B n'ait aucune influence sur la réalisation d'un événement A . Cette possibilité est très importante pour le calcul de certaines probabilités et en analyse statistique.

<u>Définition de l'indépendance de deux événements</u>

Deux événements A et B sont INDÉPENDANTS si

$$P(A \cap B) = P(A)P(B) \qquad (3.16)$$

Il suit immédiatement que

$$P(A|B) = P(A)$$
$$\text{et} \quad P(B|A) = P(B) \qquad (3.17)$$

Remarques

(a) Notons que si A et B sont incompatibles ils ne sont pas indépendants car

$P(A \cap B) = P(\phi) = 0$ et

$P(A)P(B) \neq 0$ en général

(b) De même si $A \subset B$ ils ne sont pas indépendants car

$P(A \cap B) = P(A) \neq P(A)P(B)$ à moins que $P(B) = 1$

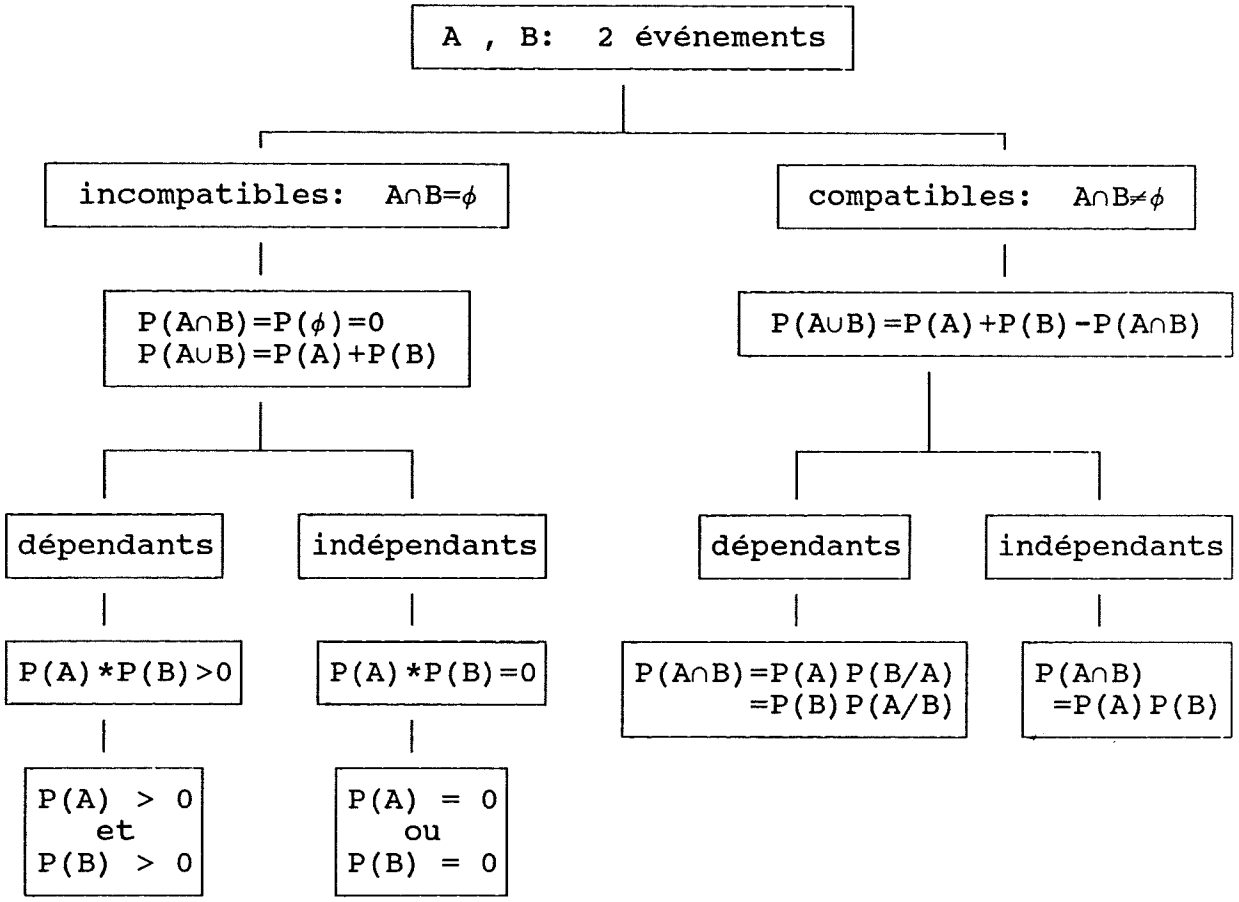
(c) Si A et B sont indépendants alors

A et B' sont indépendants

A' et B sont indépendants

A' et B' sont indépendants

(d) Le schéma suivant fait ressortir les implications et relations entre les notions d'indépendance et d'incompatibilité de deux événements A et B.



La notion d'indépendance de deux événements A et B est liée à la mesure de probabilité P comme le montre les trois exemples suivants.

Exemple 3.9 Soit un lot de 10 articles dont 4 sont défectueux. On tire 2 articles au hasard avec remise du premier article dans le lot avant le deuxième tirage.

Considérons:

A : article défectueux au premier tirage
B : article défectueux au deuxième tirage

et assumons une mesure d'équiprobabilité P . Montrons que A et B sont indépendants ce qui correspond à l'idée intuitive que l'on se fait de A et B puisque le deuxième tirage ne dépend aucunement de ce qui s'est passé lors du premier tirage.

Supposons les articles numérotés

1, 2, 3, 4 : articles défectueux

5, 6, ..., 10 : articles non-défectueux

On peut décrire les événements A et B et l'espace de cette expérience par

$\Omega = \{\omega_{ij} = (i,j) : 1 \leq i,j \leq 10\}$

$card(\Omega) = 100$ où card (Ω) dénote le cardinal de Ω

$P[\{\omega_{ij}\}] = 0.01$ en adoptant le modèle d'équiprobabilité

$A = \{\omega_{ij} = (i,j) : 1 \leq i \leq 4 , 1 \leq j \leq 10\}$

$card(A) = 40$ $P(A) = 0.40$

$B = \{\omega_{ij} = (i,j) : 1 \leq i \leq 10 , 1 \leq j \leq 4\}$

$card(B) = 40$ $P(B) = 0.40$

$A \cap B = \{\omega_{ij} = (i,j) : 1 \leq i,j \leq 4\}$

$card\ (A \cap B) = 16$ $P(A \cap B) = 0.16$

On a $P(A \cap B) = 0.16 = 0.40*0.40 = P(A)*P(B)$

Donc les événements A et B sont indépendants.

Exemple 3.10 Reconsidérons l'exemple 3.9 mais où les tirages sont effectués sans remise

$\Omega = \{\omega_{ij} = (i,j) : \ 1 \le i,j \le 10 , \quad i \ne j\}$

$card(\Omega) = 90$

$P[\{\omega_{ij}\}] = 1/90$ selon le modèle d'équiprobabilité

$A = \{\omega_{ij} = (i,j) : \ 1 \le i \le 4 , \ 1 \le j \le 10 , \ i \ne j\}$

$card(A) = 36 \qquad P(A) = 36/90 = 0.40$

$B = \{\omega_{ij} = (i,j) : \ 1 \le i \le 10 , \ 1 \le j \le 4 , \ i \ne j\}$

$card(B) = 36 \qquad P(B) = 36/90 = 0.40$

$A \cap B = \{\omega_{ij} = (i,j) : \ 1 \le i \le 4, \ 1 \le j \le 4, \quad i \ne j\}$

$card(A \cap B) = 12$

$P(A \cap B) = 12/90 = 0.1333 \ne 0.16 = 0.04*0.04 = P(A)*P(B)$

donc A et B sont dépendants comme on pouvait le prévoir.

Exemple 3.11 Reconsidérons l'exemple en effectuant les tirages avec remise mais en adoptant une mesure de probabilités différente de celle employée à l'exemple (3.9).

$\Omega = \{\omega_{ij} = (i,j) : \ 1 \le i, \ j \le 10\}$

$A = \{\omega_{ij} : \ i = 1, 2, 3, 4 ; \quad j = 1, 2, \ldots, 10\}$

$B = \{\omega_{ij} : \ i = 1, 2, 3, \ldots, 10; \ j = 1, 2, 3, 4\}$

Choisissons la mesure de probabilité suivante:

$$P[\{\omega_{ij}\}] = \begin{cases} 1/48 & i,j = 1,2,3,4 \\ 1/144 & i = 1,2,3,4 \text{ et } j = 5, \ldots, 10 \\ & i = 5, \ldots, 10 \text{ et } j = 1,2,3,4 \\ 1/108 & i,j = 5,6, \ldots, 10 \end{cases}$$

On obtient

$P(A \cap B) = 16/48 = 1/3 , \qquad P(A) = P(B) = 5/12$

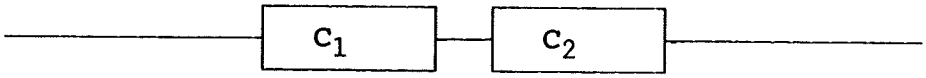
et $P(A)*P(B) = 1/4 \ne 1/3 = P(A \cap B)$

Les événements A et B sont dépendants même si les tirages sont effectuées avec remise. La mesure de probabilité seule détermine donc si deux événements sont indépendants ou non.

Exemple 3.12 Fiabilité d'un système

Un système est un arrangement de composants mécaniques et/ou électroniques et dont la fonction est de produire ou réaliser une certaine fonction. De tels systèmes sont soumis à des bris de composants et peuvent tomber en panne à des temps imprévisibles. On définit la FIABILITÉ d'un système comme la probabilité que le système soit en opération après t unités de temps. La théorie de la fiabilité a développé des modèles pour représenter ces fonctions. Notre propos ici n'est d'aborder le vaste sujet de la fiabilité mais plutôt d'illustrer le calcul de la fiabilité d'un système si on connait la fiabilité de chacun des composants qui le composent. Il s'agit d'une application des règles de la théorie des probabilités.

Un système de deux composants C_1 et C_2 placés en série est opérant si les deux composants fonctionnent. On peut représenter un tel système selon le schéma



Soit F_S la fiabilité d'un système ainsi formé et F_1, F_2 les fiabilités respectives des composants C_1 et C_2. Si on suppose l'indépendance de C_1 et C_2 on a, selon la règle de multiplication,

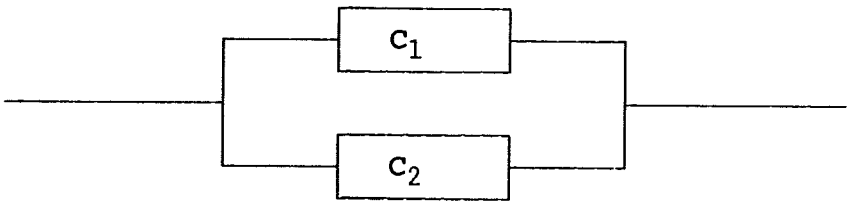
$$F_S = F_1 * F_2$$

D'une manière générale, pour n composants $C_1, C_2, \ldots, C_n$ placés en série et indépendants

$$F_S = \prod_{\alpha=1}^{n} F_\alpha \qquad (3.18)$$

où F_α représente la fiabilité du composant C_α et F_S la fiabilité du système.

Un système formé de 2 composants C_1 et C_2 en parallèle est opérant si au moins un composant fonctionne. On peut représenter un tel système selon le schéma.



Soit F_p la fiabilité d'un tel système et F_1, F_2 les fiabilités de C_1 , C_2 respectivement. Si on suppose l'indépendance de C_1 et C_2 , on a

$$F_p = F_1 + F_2 - F_1 * F_2 = 1 - (1-F_1)(1-F_2)$$

par la règle d'addition. Plus généralement, si un système est formé de n composants C_1, C_2, ..., C_n placés en parallèle et indépendants

$$F_p = 1 - \prod_{\alpha=1}^{n} (1-F_\alpha) \qquad (3.19)$$

où F_p est la fiabilité du système et F_α est la fiabilité du composant C_α .

Les règles précédentes permettent d'évaluer la fiabilité d'un système mixte formés des composants placés en série et en parallèle. La méthode consiste à décomposer le système en sous-systèmes pour lesquels on applique les deux règles précédentes. Un exemple d'un tel arrangement est représenté à la figure 3.3.

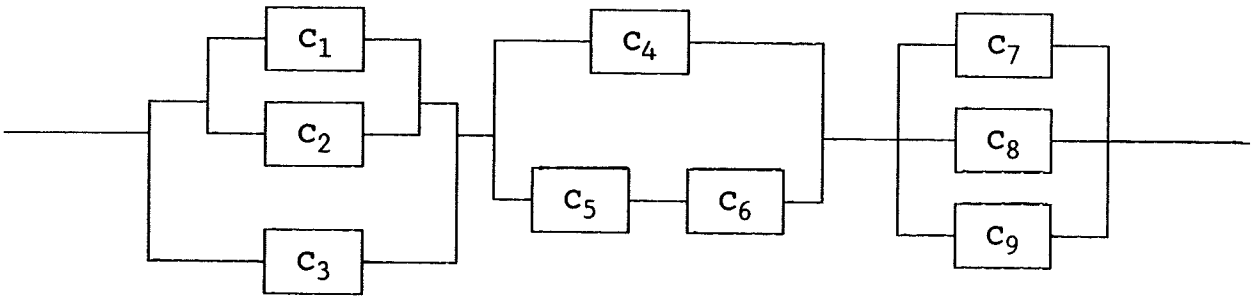


Figure 3.3: exemple d'un système

Notons par F_α la fiabilité du composant C_α $\alpha = 1, \ldots, 9$

On peut décomposer le système en 3 sous-systèmes placés en série:

sous-système A formé de C_1, C_2, C_3

sous-système B formé de C_4, C_5, C_6

sous-système C formé de C_7, C_8, C_9

Soient F_A, F_B, F_C les fiabilités des sous-systèmes A, B, C respectivement. En supposant l'indépendance des composants et en appliquant les règles d'addition et multiplication on a:

$$F_A = 1 - (1-F_1)(1-F_2)(1-F_3)$$

$$F_B = 1 - [(1-F_4)(1-F_5F_6)]$$

$$F_C = 1 - (1-F_7)(1-F_8)(1-F_9)$$

et la fiabilité F du système est égale a

$$F = F_A * F_B * F_C$$

Par exemple, si tous les composants C_α ont une fiabilité $F_\alpha = 0.95$ après 100 heures de fonctionnement, on obtient:

$$F_A = 1 - 0.05*0.05 * 0.05 = 0.999875$$

$$F_B = 1 - [0.05 * (1 - 0.95 * 0.95)] = 0.995125$$

$$F_C = 1 - 0.05 * 0.05 * 0.05 = 0.999875$$

et $F = 0.994876$

3.6 FORMULE DE BAYES

Proposition 3.2: règle de Bayes

Soient A et B deux événements ayant des probabilités non nulles

$$P(A) > 0 , \quad P(B) > 0$$

La règle de multiplication établit une équation entre les probabilités conjointes et probabilités conditionnelles

$$P(A \cap B) = P(A)P(B|A)$$

$$P(A \cap B) = P(B)P(A|B)$$

Si on élimine $P(A \cap B)$ on obtient la relation suivante appelée RÈGLE DE BAYES:

$$P(A|B) = P(A)P(B|A)/P(B) \qquad (3.20)$$

Cette règle sert de base à un ensemble de techniques appelées statistiques Bayesiennes.

Dans ce contexte

A est une estimation ou hypothèse a priori (avant les données)

B est l'information fournie par des données

P(A) est la probabilité a priori

P(B|A) est la probabilité des données selon A

P(A|B) est la probabilité a posteriori c'est-à-dire une réévaluation de P(A) compte tenu des données

Proposition 3.3: formule des probabilités totales

Considérons une partition A_α, $\alpha = 1,\ldots,n$ de Ω

$$A_\alpha \cap A_\beta = 0 \qquad \alpha \neq \beta \qquad \bigcup_{\alpha=1}^{n} A_\alpha = \Omega$$

Tout événement B peut s'écrire $B = \bigcup_{\alpha=1}^{n} (A_\alpha \cap B)$

dont les événements $A_\alpha \cap B$ sont disjoints. Il suit

$$P(B) = \sum_{\alpha=1}^{n} P(A_\alpha \cap B) = \sum_{\alpha=1}^{n} P(A_\alpha)P(B|A_\alpha) \qquad (3.21)$$

Cette dernière équation s'appelle FORMULE DES PROBABILITÉS TOTALES. En particulier:

$$P(B) = P(A)\ P(B|A) + (1-P(A))\ P(B|A')$$
$$P(B|A') + P(A)[P(B|A) - B(B|A')]$$

Proposition 3.4: formule de Bayes

En combinant les règles précédentes on obtient la FORMULE DE BAYES:

$$P(A_k|B) = \frac{P(A_k)P(B|A_k)}{\sum_{\alpha=1}^{n} P(A_\alpha)P(B|A_\alpha)} \qquad k = 1,2,\ldots,n \qquad (3.22)$$

Exemple 3.13

On achète de trois fournisseurs 1, 2 et 3 des articles dans les proportions respectives: 0.15, 0.80 et 0.05. Par expérience on a observé que le pourcentage d'articles défectueux était de 2, 1 et 4% respectivement. Ces informations sont résumées dans le tableau ci-joint.

Fournisseur	Proportion des achats	Proportion des articles défectueux
1	0.15	0.02
2	0.80	0.01
3	0.05	0.04

On demande

(a) de calculer la probabilité qu'un article choisi au hasard soit défectueux

(b) calculez la probabilité qu'il provienne du fournisseur 3 si un article choisi au hasard est défectueux.

Solution: définissons les événements.

B : article défectueux

A_α : article provient du fournisseur $\alpha = 1, 2, 3$

On a

$$P(A_1) = 0.15 \quad P(A_2) = 0.80 \quad P(A_3) = 0.05$$

$$P(B|A_1) = 0.02 \quad P(B|A_2) = 0.01 \quad P(B|A_3) = 0.04$$

Donc

$$P(B) = 0.15 * 0.02 + 0.80 * 0.01 + 0.05 * 0.04$$

$$= 0.003 + 0.008 + 0.002 = 0.013$$

est la probabilité qu'un article choisi au hasard soit défectueux. La probabilité que l'article défectueux provienne du fournisseur 3 est:

$$P(A_3|B) = P(A_3)P(B|A_3)/P(B) = (0.05 * 0.04)/0.013$$

$$= 0.002/0.013 = 0.154$$

3.7 EXERCICES

3.1 Un boulon est choisi au hasard dans une boîte de 10 000. Il peut être affecté de trois types de défauts A,B,C. Le tableau suivant résume les probabilités associées à ces trois types de défauts:

A	0.0100
B	0.0050
C	0.0075
A et B	0.0025
A et C	0.0030
B et C	0.0020
A et B et C	0.0010

Quelle est la probabilité que le boulon possède au moins un de ces défauts? Faites un diagramme de Venn et définissez l'espace des événements élémentaires.

3.2 Les trois options les plus populaires d'un certain type de nouvelle voiture sont:

A : La transmission automatique

B : La servodirection

C : La radio

une analyse des ventes a montré que les acheteurs ont choisi:

Option	% des ventes
A	70
B	75
C	80
A ou B	80
A ou C	85
B ou C	90
A ou B ou C	95

calculez les probabilités des événements suivants:

(a) l'acheteur choisit une des trois options

(b) l'acheteur choisit la radio seulement

(c) l'acheteur ne choisit aucune des trois options

(d) l'acheteur choisit exactement une des trois options.

3.3 Une étude sur la relation entre les revenus des conducteurs (F = faible, M = moyen et E = élevé) et les marques d'automobiles dénotés A, B, C, nous donne le tableau suivant:

Revenu / Marque	F	M	E
A	0.10	0.13	0.02
B	0.20	0.12	0.08
C	0.10	0.15	0.10

Cette table montre, par exemple, que la probabilité qu'un chauffeur à faible revenu préfère la marque A est de 0.10 (équivalent à $P(F \cap A)$). La probabilité qu'un chauffeur achète une voiture de la compagnie A est de 0.25 et la probabilité qu'un chauffeur ait un faible revenu est de 0.40. Utilisez cette table pour calculer les probabilités suivantes:

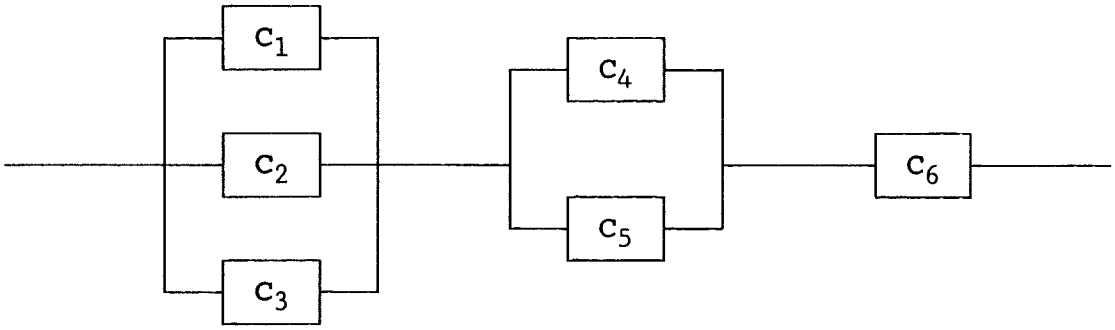
(a) $P(B|E)$ (b) $P(M|C)$ (c) $P(A|M)$

(d) $P(M|A)$ (e) $P[(M \cap B)|C]$ (f) $P[(F \cup M)|C]$

3.4 Un nouveau système de freinage est composé de trois sous-systèmes: électrique, hydraulique et mécanique. Les fiabilités (probabilité de bon fonctionnement durant une période de temps déterminée) de ces trois sous-systèmes sont 0.995, 0.993 et 0.994 respectivement. Calculez la fiabilité du système si les trois sous-systèmes sont placés en série et opèrent indépendamment.

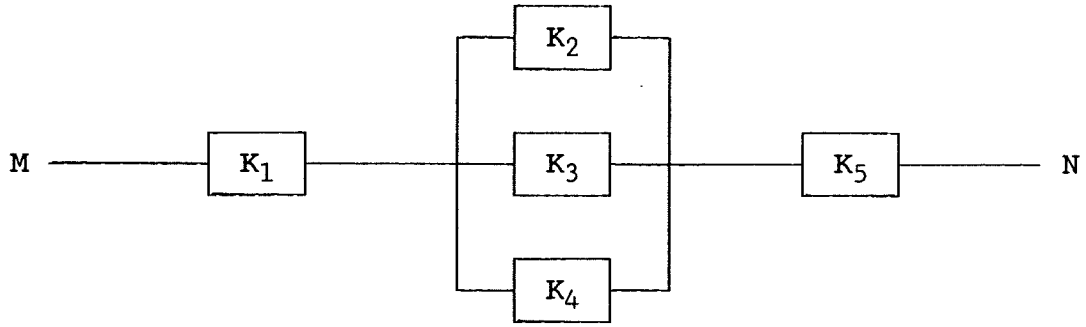
3.5 Le diagramme représente un système formé de 6 composants C1, C2, C3, C4, C5, C6 opérant indépendamment. La fiabilité de chaque composant est:

composant	C1	C2	C3	C4	C5	C6
fiabilité	0.8	0.9	0.9	0.8	0.9	0.95



quelle est la fiabilité du système?

3.6 Le diagramme d'un circuit constitué de 5 commutateurs K_i ($i = 1, 2, \ldots, 5$)



Le fait que les différents commutateurs soient ouverts ou fermés constitue des événements indépendants dont les probabilités sont:

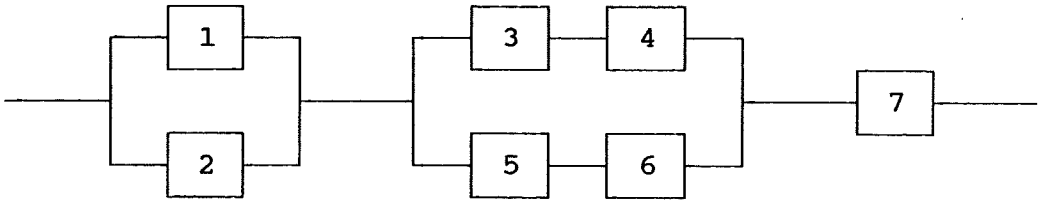
commutateurs ouverts (courant passe)	K_1	K_2	K_3	K_4	K_5
probabilité	0.6	0.4	0.7	0.9	0.5

Déterminez la probabilité que ce courant ne se rende pas au point N .

3.7 On considère le système de composants connectés tel qu'illustré sur le diagramme suivant:

Les composants 1 et 2 sont connectés en parallèle donc ce sous-système fonctionne si l'un ou l'autre de ces composants fonctionne; les composants 3 et 4 sont reliés en série; de fait ce sous-système est inopérant si l'un ou l'autre ou les deux composants sont défectueux. Si les parties du montage sont indépendantes les unes des autres et ont une probabilité d'opération de 0.9, calculez la probabilité que le montage global soit opérant?

3.8 Pour le montage suivant, les composants opèrent avec une probabilité de 0.9. Trouvez la probabilité que le montage global soit opérant. Si on ajoute un autre composant en parallèle au composant 7, quel sera la probabilité de fonctionnement du montage?



3.9 Deux villes A et B sont séparées par 3 feux de circulation chacun ayant un cycle d'une minute; les feux verts ayant des durées respectives de 40, 30 et 20 secondes. En supposant que vous respectiez les lois de la circulation et l'indépendance des trois feux de circulation. Trouvez la probabilité de faire le trajet entre A et B:

(a) sans arrêt

(b) avec 1 arrêt exactement

(c) avec 2 arrêts exactement

(d) avec au moins 1 arrêt

3.10 Une centrale hydroélectrique possède deux génératrices et à cause de l'entretien et du bris occasionnel, les génératrices peuvent être hors d'usage. Définissons les événements:

A: la première génératrice est hors d'usage
B: la deuxième génératrice est hors d'usage

Par expérience, on estime les probabilités de ces événements à:

$$P(A) = 0.01 \quad \text{et} \quad P(B) = 0.02$$

Au cours de l'été, une température supérieure à 30 degrés Celsius est notée T et sa probabilité est $P(T) = 0.30$.

Dans ces conditions, cela entraîne une demande accrue de courant pour la climatisation, et la capacité de la centrale à faire face à cette demande est:

satisfaisante (S): si les deux génératrices fonctionnent et la température est inférieure à 30

faible (F): si une des deux génératrices est hors d'usage et la température est supérieure à 30 degrés

marginale (M): autrement

Les événements A, B et T sont mutuellement indépendants.

(a) déterminez l'espace de tous les résultats en terme de A, B, T
(b) exprimez les événements S, F, M et en termes A, B, T
(c) calculez la probabilité, qu'il y ait exactement une génératrice hors d'usage
(d) calculez P(S), P(F), P(M)

3.11 La réalisation d'un projet de construction requiert une série de travaux soit: excavation (E), fondation (F), structure (S). Les probabilités qu'ils soient complétés selon l'échéancier sont respectivement de 0.8, 0.7 et 0.9. On assume que les événements sont mutuellement indépendants. Calculez les probabilités des événements:

- projet complété selon l'échéancier (K)
- excavation complétée et au moins une des deux autres étapes complétée en temps (G)
- une des trois étapes complétée en temps (H)

3.12 Les rebus provenant d'une usine sont traités en trois étapes: primaire, secondaire et tertiaire. À l'étape primaire, le traitement peut être bon, incomplet ou raté; à l'étape secondaire et tertiaire, il peut être bon ou raté. On suppose l'indépendance mutuelle de tous les événements et l'équiprobabilité des résultats à chaque étape.

(a) définir l'espace de tous les résultats possibles
(b) calculez la mesure de probabilité sur l'espace
(c) calculez la probabilité d'avoir au moins deux étapes de bonnes

(d) répondre aux questions (b) et (c) selon les probabilités suivantes de chaque étape:

	bon	incomplet	raté
primaire	0.8	0.1	0.1
secondaire	0.7	0.0	0.3
tertiaire	0.5	0.0	0.5

3.13 La quantité d'eau emmagasinée dans un réservoir peut être représentée par trois états: rempli (R), à moitié rempli (M) et vide (V). À cause du caractère aléatoire du débit d'eau entrant dans le réservoir ainsi que du débit sortant pour satisfaire la demande, la quantité d'eau emmagasinée peut changer d'un état à l'autre durant chaque saison. Les probabilités de transition (conditionnelles) d'un état à l'autre entre le début et la fin d'une saison sont:

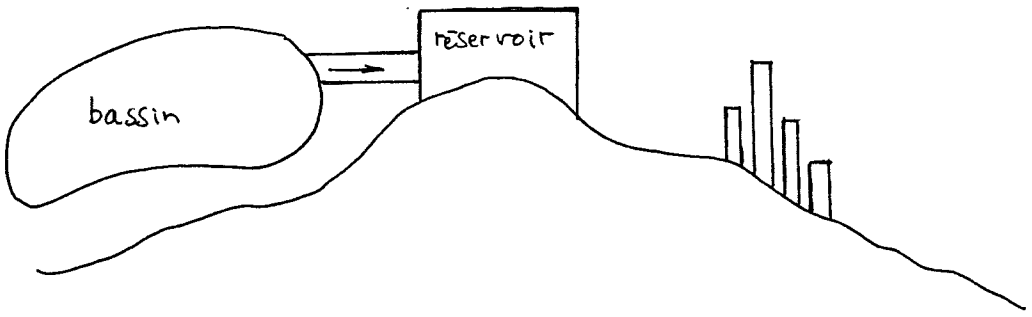
	fin		
début	V_f	M_f	R_f
V_d	0.4	0.5	0.1
M_d	0.3	0.3	0.4
R_d	0.1	0.7	0.2

Par exemple: $P(M_f | V_d) = 0.5$

Si au début de la première saison, P(V) = 0.1 , P(M) = 0.7, P(R) = 0.2 , calculez les probabilités que le réservoir

(a) soit rempli à la fin de la première saison
(b) ne soit pas vide à la fin de la première saison
(c) soit rempli à la fin de la deuxième saison
(d) ne soit pas vide à la fin de la deuxième saison.
(e) déterminez les probabilités de chaque état après 3 saisons.

3.14 Le système d'approvisionnement en eau d'une ville consiste en un réservoir et un pipeline acheminant l'eau d'un bassin situé plus loin selon la figure 3.4.



La quantité d'eau du réservoir est variable car elle dépend des précipitations. D'autre part, la consommation d'eau fluctue selon les jours. Définissons les événements suivants et leurs probabilités:

A: quantité d'eau disponible du bassin est faible
B: quantité d'eau du réservoir est faible
C: consommation d'eau est faible

$P(A) = 0.20 \qquad P(B) = 0.15 \qquad P(C) = 0.50$

On estime que:

$P(A'|C') = 0.75 \qquad P(B|A) = 0.50$

$P(A \cap B|C') = 0.50 * P(A \cap B)$

B et C sont indépendants

Si la consommation est forte, on observe une pénurie d'eau (E) quand la quantité d'eau du bassin est faible ou la quantité d'eau du réservoir est faible. Calculez la probabilité de E .

3.15 Le temps requis pour compléter un projet de construction dépend de la possibilité de grève. Les probabilités conditionnelles de délai sont de:

événement	$D\|A\cap B$	$D\|A\cap B'$	$D\|A'\cap B$	$D\|A'\cap B'$
probabilité	1.00	0.80	0.40	0.05

où A: menuisiers en grève, B: plombiers en grève, D: délai.

Par expérience on estime qu'il y a:

60% de chance que les plombiers tombent en grève si les menuisiers le sont déjà

30% de chance que les menuisiers tombent en grève si les plombiers le sont déjà

10% de chance de grève chez les plombiers.

(a) Calculez la probabilité d'un délai
(b) s'il y a eu délai, calculez la probabilité que
  (i) les menuisiers et plombiers aient été en grève
  (ii) les menuisiers seuls aient été en grève
  (iii) les plombiers seuls aient été en grève

3.16 Un camionneur travaillant au déblaiement de la neige peut faire, en une journée de 8 heures, entre 1 et 20 voyages inclusivement. Pour différentes raisons imprévisibles le nombre de voyages peut varier d'un jour à l'autre. Les informations accumulées permettent de dire que:

- le nombre de journées où il effectue moins de 6 voyages est égal à celui où il en fait plus de 15;
- il y a six fois plus de journées où il fait de 6 à 10 voyages que de journées où il en fait moins de 6;
- il y a douze fois plus de journées où il fait de 11 à 15 voyages que de journées où il en fait moins de 6;
- sur $\{1,2,3,4,5\}$ et sur $\{16,17,18,19,20\}$ il y a équiprobabilité des événements élémentaires.

(a) Décrire l'espace Ω des résultats et une mesure de probabilité qui reflète les informations.

(b) Calculez la probabilité qu'il effectue entre 8 et 17 voyages.

3.17 La pollution de l'air dans une ville est principalement causée par les industries et les gaz d'échappement des autos. On définit les événements

A: contrôler la pollution industrielle

B: contrôler la pollution des autos

et C: diminuer le degré de pollution en dessous d'un niveau acceptable.

On donne $P(A) = 0.75$, $P(B) = 0.60$ et on admet que A et B sont indépendants. De plus on sait que $P(C|A \cap B') = 0.80$ et $P(C|A' \cap B) = 0.80$

(a) Calculez $P(C)$, $P(A \cap B'|C')$ et $P(B'|C')$.

(b) Dites en mots ce que représente $P[(A' \cap B)|C']$.

3.18 Deux routes 1 et 2 se rejoignent pour former la route 3 selon la figure 3.5. Les routes 1 et 2 ont la même capacité (nombre de voies) et la route 3 a une capacité plus grande. Aux heures de pointe, la probabilité que la circulation soit excessive est de 0.10 sur la route 1 et de 0.30 sur la route 2. De plus, sachant qu'elle est excessive sur la route 2, elle l'est aussi sur la route 1 avec probabilité 1/3. Posons

A_1, A_2, A_3: la circulation est excessive sur les routes 1,2,3.

(a) Calculez la probabilité que la circulation soit excessive aux heures de pointe.

(i) sur les routes 1 et 2

(ii) sur la route 2 sachant qu'il l'est sur la route 1

(iii) sur la route 1 seulement

(iv) sur la route 2 seulement

(v) sur la route 1 ou sur la route 2

(vi) ni sur la route 1 ni sur la route 2.

(b) Sur la route 3, la circulation est excessive

- sûrement si elle l'est sur la route 1 et sur la route 2
- avec probabilité 0.15 si elle l'est sur la route 2 seulement
- avec probabilité 0.10 si elle ne l'est pas ni sur la route 1 ni sur la route 2

Calculez la probabilité que la circulation soit excessive

(i) sur la route 3

(ii) sur la route 1 sachant qu'elle est excessive sur la route 3.

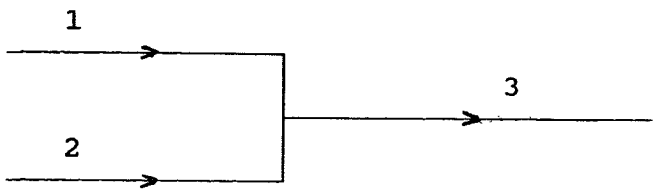
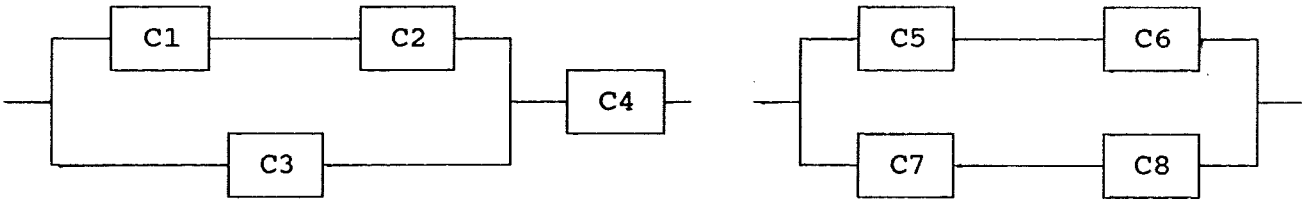


Figure 3.5: exercice 3.18

3.19 Deux systèmes A et B sont formés chacun de 4 composants ayant tous une fiabilité de p (0 <p <1)



Système A Système B

(a) Calculez la fiabilité du système A

(b) Calculez la fiabilité du système B

(c) Quel système est le plus fiable?

3.20 Une méthode pour différencier la roche de type granitique de la roche de type basaltique est d'examiner une partie du spectre infrarouge provenant des rayons solaires qui sont reflétés sur la surface de la roche étudiée. Posons R_1, R_2 et R_3 les mesures de l'intensité du spectre de trois différentes longueurs d'ondes. Quand les mesures sont prises par avion, différentes valeurs de R_1 sont obtenues. Des survols au-dessus de plusieurs régions dont la composition en roche est connue, ont fourni les tableaux des probabilités conditionnelles suivantes:

	$R_1 < R_2 < R_3$	$R_1 < R_3 < R_2$	$R_3 < R_1 < R_2$
Granite	0.60	0.25	0.15
Basalte	0.10	0.20	0.70

où par exemple, $P(R_1 < R_2 < R_3 \mid \text{granite}) = 0.60$

Supposons que, pour une région donnée, la probabilité pour une roche choisie au hasard est

$$P(\text{granite}) = 0.25 \qquad P(\text{basalte}) = 0.75$$

(a) Calculez $P(R_1 < R_2 < R_3)$

Suggestion: utilisez la formule des probabilités totales

(b) Calculez $P[\text{granite} \mid R_1 < R_2 < R_3]$ et

$$P[\text{basalte} \mid R_1 < R_2 < R_3]$$

(c) Si $R_1 < R_2 < R_3$, allez-vous classifier la roche comme granite ou basalte?

(d) Si $R_1 < R_3 < R_2$, comment allez-vous classifier la roche?

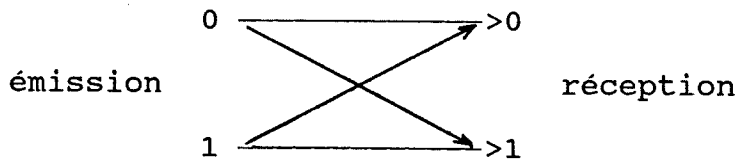
(e) Si $R_3 < R_1 < R_2$, comment allez-vous classifier la roche?

(f) En utilisant la règle de classification définie par (c)-(d)-(e) quelle est la probabilité de faire une classification erronée?

(g) Soit $p = P(\text{granite})$, $0 < p < 1$. Pour quelles valeurs de p on classifiera toujours la roche comme granite peu importe les résultats des spectres.

3.21 Un des problèmes qui contribua à la création de la théorie des probabilités au 17 ième siècle, fut posé à Pascal par un noble français, le Chevalier de Méré. Il a trouvé expérimentalement que la probabilité d'obtenir au moins un "6" en jetant quatre (4) fois un dé est supérieure à celle d'obtenir au moins un double 6 en lançant vingt-quatre (24) fois deux (2) dés; le résultat lui semblait irraisonnable. Avait-il raison?

3.22 Un système de communication simple est représenté par le schéma suivant:



À cause du "bruit" le signal émis est quelquefois mal reçu. Définissons les événements suivants:

E_0: "0" émis $\qquad$ E_1: "1" émis

R_0: "0" reçu $\qquad$ R_1: "1" reçu

On sait que $P(R_0 \mid E_0) = 0.7$ et $P(R_1 \mid E_1) = 0.8$ et que le "0" est émis 60% du temps. Calculez les probabilités des événements suivants:

(a) R_1, $E_0 \mid R_1$

(b) d'une erreur de transmission

3.23 Le jeu de "craps" se joue avec deux dés de la façon suivante:

- si le lanceur obtient 7 ou 11: il gagne
- si le lanceur obtient 2, 3 ou 12: il perd
- si le lanceur obtient 4, 5, 6, 8, 9, 10: il doit lancer de nouveau les dés jusqu'à ce qu'il obtienne 7 ou le chiffre initial (4, 5, 6, 8, 9, 10)

  (i) si le 7 apparaît avant le chiffre initial: il perd

  (ii) si le chiffre initial apparaît avant le 7: il gagne

Posons

G: le lanceur gagne

G_k: le lanceur gagne après k jets, k=1,2,...

$G_k(s)$: le lanceur gagne après k jets avec un total de s au 1er jet, s=4,5,6,8,9,10

Montrez que

(a) $P(G_1) = 2/9$

(b) $P(G_2(4)) = P(G_2(10)) = (3/36)^2$

$P(G_2(5)) = P(G_2(9)) = (4/36)^2$

$P(G_2(6)) = P(G_2(8)) = (5/36)^2$

(c) $P(G_k(s)) = (3/36)^2 \ (27/36)^{k-2}$ k=2,3,...;s=4,10

$P(G_k(s)) = (4/36)^2 \ (26/36)^{k-2}$ k=2,3,...;s=5,9

$P(G_k(s)) = (5/36)^2 \ (25/36)^{k-2}$ k=2,3,...;s=6,8

(d) $G = \bigcup_{k=1}^{\infty} G_k$, $G_i \cap G_j = \phi$, $i \neq j$

(e) $P(G) = 244/495 = 0.493$

3.24 Un appareil est formé de deux composants A et B. La probabilité que le composant A soit remplacé à cause d'une panne est de 0.1 et dans ces circonstances le composant B sera remplacé avec probabilité 0.2. D'autre part si A n'est pas remplacé, B le sera avec probabilité 0.7. L'appareil peut être en panne sans que A et B soient en panne.

(a) Décrire l'espace de probabilité de l'état de l'appareil.

(b) Quel pourcentage des pannes de l'appareil nécessite le remplacement de

(i) A seulement

(ii) B seulement

(iii) A et B

3.25 Un réseau de distribution d'électricité est composé de trois stations A,B,C. Par expérience, les probabilités conditionnelles d'une panne du réseau sont:

i	1	2	3	4	5	6	7	8
$P(D\|\omega_i)$	0.50	0.20	0.30	0.10	0.15	0.10	0.05	0.01

où D: panne du réseau

$\omega_1 = A \cap B \cap C$ $\omega_2 = A \cap B \cap C'$ $\omega_3 = A \cap B' \cap C$ $\omega_4 = A' \cap B \cap C$

$\omega_5 = A \cap B' \cap C'$ $\omega_6 = A' \cap B \cap C'$ $\omega_7 = A' \cap B' \cap C$ $\omega_8 = A' \cap B' \cap C'$

A: station A surchargée B: station B surchargée
C: station C surchargée

D'autre part, les probabilités de l'état du réseau sont:

i	1	2	3	4	5	6	7	8
$P(\omega i)$	0.01	0.02	0.03	0.02	0.05	0.05	0.05	?

(A) Calculez la probabilité

(a) d'une panne
(b) que la surcharge de la station A soit la cause d'une panne.
(c) que la surcharge de la station B soit la cause d'une panne.
(d) que la surcharge de la station C soit la cause d'une panne.

(e) que la surcharge de deux stations ou plus soit la cause d'une panne.
(f) que la cause d'une panne soit attribuable à aucune surcharge des stations A, B et C.

(B) Classez les causes d'une panne en ordre de priorité.

3.26 Démontrez les résultats suivants:
(a) Si $P(B|A') = P(B|A)$ alors A et B sont indépendants.
(b) La probabilité de réalisation d'exactement un des événements A,B est égale à:
$$P(A) + P(B) - 2P(A \cap B)$$
(c) La probabilité de réalisation d'exactement un des événements A,B,C est égale à:
$$P(A) + P(B) + P(C) - 2P(A \cap B) - 2P(A \cap C) - 2P(B \cap C) + 3P(A \cap B \cap C)$$

3.27 Le choix d'un diagnostic médical est fonction des probabilités a priori relatives à chacune des maladies possibles ainsi que des informations apportées par les caractéristiques du malade. Un échantillon de grande taille a permis de déterminer le tableau qui suit concernant des malades ayant subi une radio thoracique.

Maladies	$P(M_i)$	$P(H\|M_i)$	$P(A\|M_i)$
M_1	0.30	0.9	0.1
M_2	0.20	0.7	0.6
M_3	0.15	0.8	0.2
M_4	0.05	0.6	0.3
M_5	0.01	0.5	0.5

H = homme A= âgé de moins de 40 ans

Calculez en supposant l'indépendance de l'âge et de sexe de la personne, les probabilités suivantes:

(a) $P(H \cap A' | M_i)$, $P(H \cap A' \cap M_i)$, $P(H \cap A')$, $P(M_i | H \cap A')$

(b) Classez les maladies (M_i) par ordre de probabilité décroissante chez un homme âgé de plus de 40 ans.

3.28 Des essais effectués sur un nouveau alcool-test ont permis d'établir que:

(i) 5 fois sur 100, l'alcool-test s'est révélé positif alors qu'une personne n'était pas en état d'ébriété.

(ii) 90 fois sur 100, l'alcool-test s'est révélé positif alors qu'une personne était réellement en état d'ébriété.

(iii) 1% des personnes contrôlées sont réellement en état d'ébriété.

On définit les événements E,A,B:

E: être en état d'ébriété,

A: l'alcool-test est positif,

B: l'alcool-test est négatif.

Calculez:

(a) $P(A)$

(b) $P(E|A)$

(c) Calculez $P(E|A)$ si dans (i) plus haut, on change 5/100 par 1/100.

3.29 Un appareil est composé de deux unités de type A et de trois unités de type B. Il fonctionne si au moins une unité de type A et au moins deux unités de type B sont en bonne condition. De plus, les unités sont indépendantes. On pose

A_k: la k-ième unité de type A est en bonne condition $k = 1,2$

B_j: la j-ième unité de type B est en bonne condition $j = 1,2,3$

C: l'appareil fonctionne

$p_1 = P(A_k)$ $\qquad$ $p_2 = P(B_j)$

(a) Exprimez C en fonction de A_k et B_j.

(b) Calculez $P(C)$ en fonction de p_1 et p_2.

(c) Evaluez $P(C)$ pour $p_1 = 0.50, 0.90, 0.99$ et $p_2 = 0.50, 0.90, 0.99$.

3.30 Un échantillon de 100 articles est choisi sans remise d'un lot de 1000. Calculez la probabilité que l'échantillon contienne seulement de bons articles si on suppose que le nombre d'articles défectueux dans le lot est entre 0 et 5 inclusivement et que toutes les possibilités sont également probables.

3.31 Un lot contient 1% d'articles défectueux. Quelle serait la taille n d'un échantillon à prélever de telle sorte que la probabilité soit au moins de 0.95 de trouver au moins un article défectueux.

3.32 Un appareil est formé de deux composants, A et B, susceptibles de tomber en panne. Les composants sont placés en parallèle et ne sont pas indépendants. On estime à:

0.20, la probabilité d'une panne du composant A;

0.80, la probabilité d'une panne du composant B si le composant A est en panne;

0.40, la probabilité d'une panne du composant B si le composant A n'est pas en panne.

a) Calculez la probabilité d'une panne

(i) du composant B;

(ii) de l'appareil;

(iii) du composant A si le composant B est en panne;

(iv) d'exactement un composant.

b) Afin d'augmenter la fiabilité de l'appareil, on installe un troisième composant, C, de telle sorte que les composants A, B et C sont placés en parallèle. La probabilité que le composant C tombe en panne est de 0.2 et cela indépendamment de l'état (panne ou non-panne) des composants A et B.

(i) Calculez la probabilité que l'appareil formé des composants A,B,C tombe en panne.

(ii) Etant donné que l'appareil est en fonctionnement, quelle est la probabilité que le composant C soit en panne?

3.33 Dans une usine de fabrication de composants électroniques on assure le contrôle de qualité à l'aide de trois tests:

. chaque composant est assujetti au test numéro 1

. si le composant réussit le test numéro 1, il est soumis au test numéro 2

. si le composant réussit le test numéro 2, il subit le test numéro 3

. dès qu'un composant échoue à l'un des tests, on le retourne pour réparation

Définissons les événements:

A_i: le composant échoue le test numéro i, i=1,2,3

Par expérience on estime que:

$$P(A_1)=0.10, \quad P(A_2|A_1')=0.05, \quad P(A_3|A_1' \cap A_2')=0.02$$

(a) Montrez que les événements élémentaires de l'espace Ω sont:

$$\omega_1=A_1 \qquad \omega_2=A_1' \cap A_2$$

$$\omega_3=A_1' \cap A_2' \cap A_3 \qquad \omega_4=A_1' \cap A_2' \cap A_3'$$

(b) Calculez la probabilité de chacun des événements élémentaires

(c) Soit R l'événement dénotant que le composant doit être réparé

(i) exprimez R en fonction de A_1, A_2, A_3

(ii) calculez la probabilité de R

(iii) calculez $P(A_1' \cap A_2|R)$

(d) On teste trois composants et on définit les événements

R_k: le k-ième composant doit être réparé k=1,2,3

B: au moins un des trois composants réussit les trois tests

On suppose que les événements R_k sont indépendants:

(i) exprimez B en fonction de R_1, R_2, R_3

(ii) calculez P(B)

3.34 Un appareil électronique est formé de deux composants A et B. La probabilité de remplacer A est de 0.50 si l'appareil est en panne. Une panne du composant A endommage le composant B et la probabilité de remplacer le composant B est de 0.70. Toutefois, si le composant A ne doit pas être remplacé, la probabilité de remplacer le composant B est de 0.10. Calculez la probabilité de:

(a) remplacer A et B,
(b) remplacer A seulement,
(c) remplacer B seulement,
(d) remplacer aucun des deux composants.

3.8 RÉPONSES EXERCICES

3.1 0.0160

3.2 (a) 0.95 (b) 0.15 (c) 0.05 (d) 0.30

3.3 (a) 0.40 (b) 0.428 (c) 0.325 (d) 0.52 (e) 0
(f) 0.714

3.4 0.982107

3.5 0.929138

3.6 0.7054

3.7 0.9981

3.8 0.85883 , 0.94472

3.9 (a) 1/9 (b) 7/18 (c) 7/18 (d) 8/9

3.10 (c) 0.0296 (d) 0.67914, 0.02072, 0.29366

3.11 0.504, 0.776, 0.092

3.12 (c) 0.4167

3.13 (a) 0.33 (b) 0.73 (c) 0.253 (d) 0.739
(e) (0.2733, 0.4635, 0.26313)

3.14 0.35

3.15 (a) 0.118 (b) 0.254, 0.1356, 0.2373

3.16 (a) $p_i = 1/100$ $i = 1,2,3,4,5,16,17,18,19,20$
$p_i = 6/100$ $i = 6,7,8,9,10$
$p_i = 12/100$ $i = 11,12,13,14,15$

3.17 (a) 0.19, 0.316, 0.842

3.18 (a) 0.10, 1, 0, 0.20, 0.3, 0.70 (b) 0.20 , 0.50

3.19 (a) $p(1-(1-p)(1-p^2))$

(b) $1 - (1-p^2)^2$

3.20 (a) 0.225 (b) 2/3, 1/3 (c) granite (d) basalte
(e) basalte (f) 0.175 (g) $p > 14/17$

3.21 1 dé: 0.5177, 2 dés: 0.4914

3.22 (a) 0.50, 0.36
(b) 0.26

3.24 (b)i 0.08 (b)ii 0.63 (b)iii 0.02

3.25 (a) 0.0427 (b) 0.1756 (c) 0.1171 (d) 0.0585
(e) 0.4684 (f) 0.1803

3.27 (a)

i	1	2	3	4	5	
$P(H\cap A'\mid M_i)$	0.81	0.28	0.64	0.42	0.25	
$P(H\cap A'\cap M_i)$	0.243	0.056	0.096	0.021	0.0025	0.4185=P(H∩A')
$P(M_i\mid H\cap A')$	0.5806	0.1338	0.2294	0.0502	0.0059	

(b) $M_1 > M_3 > M_2 > M_5 > M_4$

3.28 (a) 0.054 (b) 0.1667 (c) 0.476

3.29 (c)

p_1 \ p_2	0.50	0.90	0.99
0.50	0.375	0.729	0.7498
0.90	0.495	0.9623	0.9897
0.99	0.4995	0.9719	0.9996

3.30 0.78

3.31 $n \geq 299$

3.32 (a) i - 0.50 ii - 0.16 iii - 0.32 iv - 0.72
(b) i - 0.968 ii - 0.1735

3.33 (b)

	ω_1	ω_2	ω_3	ω_4
Prob.	0.10	0.045	0.171	0.8379

(c) ii - 0.1621 iii - 0.2776
(d) ii - 0.9957

3.34 (a) 0.35 (b) 0.15 (c) 0.05 (d) 0.45

CHAPITRE 4

VARIABLES ET VECTEURS ALÉATOIRES

4.0 SOMMAIRE

On constate, en consultant les exemples du chapitre 3, que beaucoup de phénomènes aléatoires sont déjà à valeurs numériques. S'ils ne le sont pas on est presque toujours intéressé par certaines caractéristiques numériques associées au résultat plutôt qu'au résultat lui-même. Cette idée sera formalisée dans ce chapitre sous le nom de VARIABLE ALÉATOIRE. Les différentes caractérisations des variables aléatoires sont définies ainsi que les caractéristiques dérivées de la variable. Enfin le concept de loi conjointe sera défini pour traiter les vecteurs aléatoires.

4.1 VARIABLE ALÉATOIRE

Définition:

Soit (Ω, P) un espace de probabilité et X une application de Ω dans l'ensemble R des nombres réels. Toute fonction réelle X définie sur Ω à valeurs réelles est appelée une VARIABLE ALÉATOIRE:

$$X : \Omega \longrightarrow R$$

Remarques:

- Il s'agit d'une fonction mais la tradition veut que X soit appelée variable.
- À chaque résultat ω de l'expérience Ω on associe un nombre réel $X(\omega)$ et cette valeur varie d'un résultat à l'autre puisque ω dépend du hasard.
- Lorsque Ω est continu il faut imposer une certaine condition à X mais cette condition est toujours réalisée dans les applications.
- On notera, en général, les variables aléatoires avec des lettres majuscules: X,Y,... et leurs valeurs par des lettres minuscules: x,y,...

. Si $\Omega = R$, la fonction identité sur Ω, $I_\Omega = X$ définit elle même une variable aléatoire:

$$\Omega = R \xrightarrow{I_\Omega = X} R$$

. Si X est une variable aléatoire et

$$g : R \longrightarrow R$$

est une fonction réelle alors $g(X)$ est une variable aléatoire sur Ω

$$\Omega \xrightarrow{X} R \xrightarrow{g} R$$

$$g \circ X = g(X)$$

Définition d'une fonction de répartition:

La fonction $F_X(x)$ réelle définie par

$$F_X(x) = P[X \leq x] \qquad x \in R$$

$$= P[\{\omega \in \Omega : X(\omega) \leq x\}] \qquad (4.1)$$

s'appelle FONCTION DE RÉPARTITION de la variable aléatoire X. Cette fonction $F_X(.)$ est induite de la mesure de probabilité P de l'espace (Ω,P). En pratique on peut considerer $(R, F_X(.))$ comme un nouvel espace de probabilité.

La fonction F_X possède trois propriétés importantes:

(a) $F_X(x_1) \leq F_X(x_2)$ tout $x_1 \leq x_2$
(non décroissante)

(b) $F_X(-\infty) = 0, \quad F_X(+\infty) = 1$ (4.2)

(c) $\lim_{n\to\infty} F_X(x + 1/n) = F_X(x)$
(continuité à droite)

Exemple 4.1: jet de deux dés à jouer

Reprenons l'exemple 3.2 du chapitre 3 où

$$\Omega = \{\omega = (i,j): \ 1 \leq i,j \leq 6\}$$

représente l'espace associé au jet de 2 dés.

Soit la variable aléatoire $X(\omega) = X(i,j) = i + j$ le total des deux dés. Utilisons la mesure d'équiprobabilité

$$P[\{\omega\}] = 1/36$$

Cette mesure de probabilité induit une fonction de répartition $F_X(x)$ définie par

x	2	3	4	5	6	7	8	9	10	11	12
$F_X(x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$

où par exemple

$$F_X(4) = P[X \leq 4]$$

$$= P[\{(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)\}]$$

$$= 6/36 = 1/6$$

puisque chaque couple (i,j) a une égale probabilité $1/36$ selon la mesure P. On peut aussi obtenir la probabilité associée à une valeur x de X. Par exemple

$$P(X=4) = P[\{(1,3),(3,1),(2,2)\}] = 3/36$$

$$= F_X(4) - F_X(3)$$

Deux classes de fonctions de répartition sont intéressantes pour les applications:

- les fonctions de répartition en escalier
- les fonctions de répartition dérivables

$F_X(.)$ de type escalier: variables discrètes

Les variables aléatoires associées à ce type de fonction sont appelées VARIABLES ALÉATOIRES DISCRÈTES. En particulier lorsque Ω est un ensemble fini

$$\Omega = \{\omega_1, \ldots, \omega_n\}$$

toute variable définie sur Ω est discrète et l'ensemble des valeurs distinctes $\{x_1, \ldots, x_k\}$, $(k \leq n)$ de X définit une partition A_α de Ω où

$$A_\alpha = \{\omega: \; X(\omega) = x_\alpha\} \qquad \alpha = 1,2,\ldots,k$$

La fonction

$$p_X(x_\alpha) = P(A_\alpha) \tag{4.3}$$

est appelée MASSE DE PROBABILITÉ (ou LOI) de la variable X .

L'équation (4.1) implique les propriétés suivantes:

$$\text{(a)} \quad p_X(x_\alpha) \geq 0 \qquad i = 1,\ldots,k$$

$$\text{(b)} \quad \sum_{\alpha=1}^{k} p_X(x_\alpha) = 1 \tag{4.4}$$

$$\text{(c)} \quad F_X(x) = \sum_{x_\alpha \leq x} p_X(x_\alpha)$$

L'exemple 4.1 est représentatif de ce type de variables.

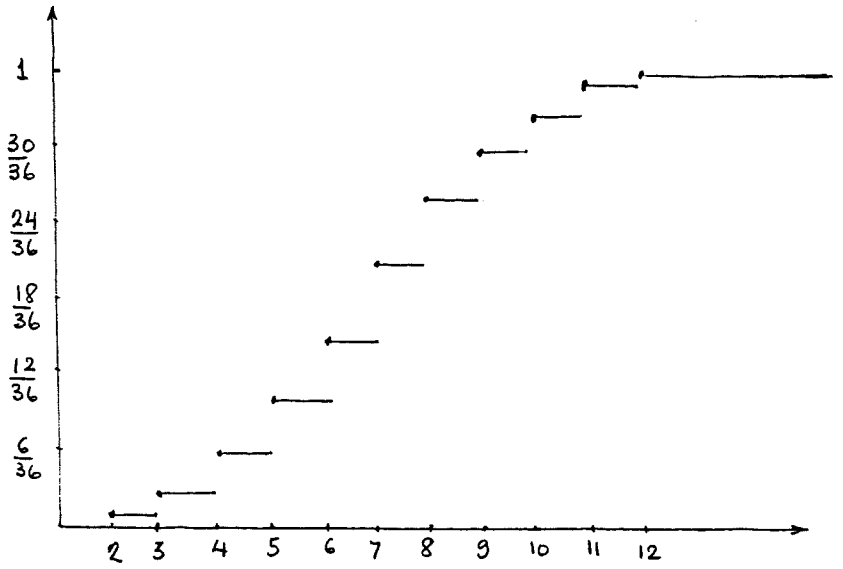


Figure 4.1: fonction de répartition de l'exemple 4.1

$F_X(.)$ Dérivable: variables continues

$F_X(x)$ est absolument continue s'il existe une fonction $f_X(x)$ telle que

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du \tag{4.5}$$

La fonction $f_X(x)$ est appelée DENSITÉ DE PROBABILITÉ et la variable X est appelée VARIABLE ALÉATOIRE CONTINUE. La définition de $F_X(.)$ vue en (4.1) implique les propriétés suivantes:

$$\begin{aligned} &(a) \quad f_X(x) \geq 0 \\ &(b) \quad \int_{-\infty}^{\infty} f_X(x)\,dx = 1 \end{aligned} \tag{4.6}$$

D'autre part il est évident que

$$\begin{aligned} f_X(x) &= \frac{d}{dx} F_X(x) \\ P[a \leq X \leq b] &= \int_a^b f_X(x)\,dx = F_X(b) - F_X(a) \\ f_X(x) &\approx P[x \leq X \leq x+\Delta x]/\Delta x, \qquad \Delta x \text{ petit} \end{aligned} \tag{4.7}$$

Exemple 4.2: durée d'un composant

Soit X la durée (heures) d'un composant électronique jusqu'à ce qu'il tombe en panne. Un modèle souvent utilisé pour représenter la fonction de répartition $F_X(x)$ est

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \quad \lambda \geq 0 \end{cases} \tag{4.8}$$

Le paramètre λ est une caractéristique spécifique du phénomène étudié et on montrera dans la prochaine section que $1/\lambda$ représente une moyenne. La densité de probabilité $f_X(x)$ est

$$f_X(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geq 0 \end{cases}$$

et on lui donne le nom de DISTRIBUTION EXPONENTIELLE de paramètre λ .

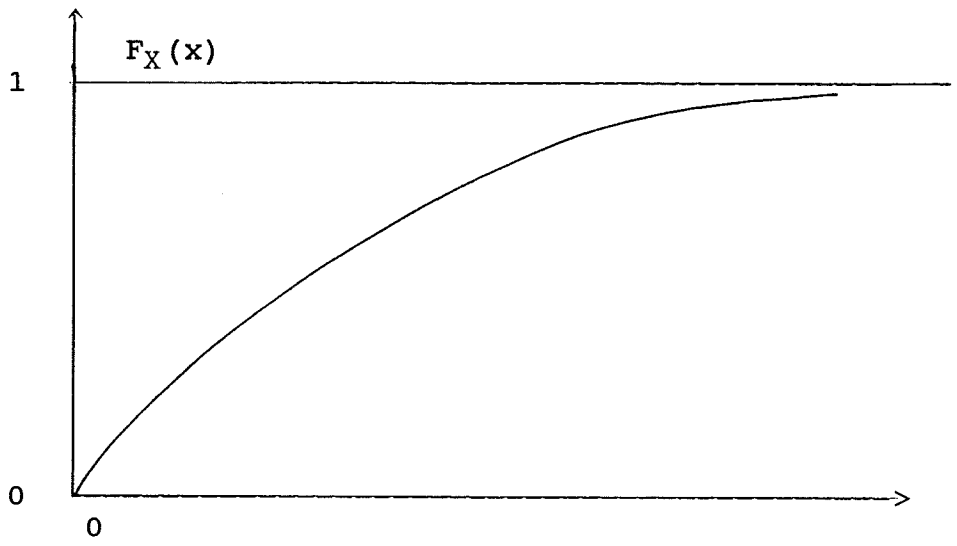


Figure 4.2: fonction de répartition d'une distribution exponentielle

4.2 INDICATEURS ASSOCIÉES À UNE VARIABLE ALÉATOIRE

La loi de probabilité ou distribution d'une variable contient toute l'information relative à celle-ci. Toutefois il est très utile de la résumer par quelques nombres dont les percentiles et les moments. Parmi ces derniers il faut noter la moyenne, la variance, le coefficient d'asymétrie et le coefficient d'aplatissement.

Percentiles

Soit X une variable aléatoire continue de densité $f_X(x)$ et p tel que $0 < p < 1$. Le p-ième PERCENTILE de X, noté x_p est cette valeur unique de X telle que

$$\int_{-\infty}^{x_p} f_X(x)dx = p \qquad 0 < p < 1 \qquad (4.9)$$

D'une manière équivalente

$$x_p = F_X^{-1}(p) \qquad (4.10)$$

où $F_X(.)$ est la fonction de répartition de X . Les percentiles où $p = 0.25, 0.50, 0.75$ s'appellent premier, deuxième et troisième quartile. Le deuxième quartile s'appelle aussi la MEDIANE.

Remarques:

- La notation x_p n'est pas universelle et par la suite nous dérogerons à notre convention pour quatre distributions d'importance: gaussienne centrée-réduite, student, khi-deux et Fisher.

- La définition de percentiles pour le cas d'une variable discrète présente des difficultés (unicité) puisque F_X^{-1} n'existe pas. Le problème est de même nature que celui rencontré lors de l'étude des distributions expérimentales du chapitre 2.

- Le concept de percentile permet de résoudre les équations de probabilité de la forme:

$$P[X \leq a] = \alpha$$

où a est inconnu et α est spécifié entre 0 et 1. Nous aurons à résoudre de telles équations dans les procédures d'estimation par intervalles et les tests d'hypothèses.

Exemple 4.3: distribution exponentielle

Il faut trouver la valeur x_p telle que

$$\int_0^{x_p} \lambda e^{-\lambda x} dx = p$$

$$1 - e^{-\lambda x} = p$$

$$x_p = (-1/\lambda) \ln(1 - p) \qquad (4.11)$$

Moments:

Soit $k = 1,2,\ldots$ un entier quelconque; le k-ième MOMENT de X par rapport à l'origine noté μ'_k est défini par

$$\mu'_k = \begin{cases} \sum_{x_\alpha} x_\alpha^k \, p_X(x_\alpha) & \text{si X est discrète} \\ \int_{-\infty}^{\infty} x^k f_X(x) dx & \text{si X est continue} \end{cases} \qquad (4.12)$$

Exemple 4.4: distribution exponentielle

$$\mu'_k = \int_0^{\infty} x^k \lambda e^{-\lambda x} dx$$

En intégrant par parties on établit la relation de récurrence

$$\mu'_k = \frac{k}{\lambda} \mu'_{k-1}$$

et par application répétée on obtient

$$\mu'_k = \frac{k!}{\lambda^k} \qquad k = 0,1,2,\ldots \qquad (4.13)$$

Moyenne:

En particulier μ'_1, aussi noté μ ou μ_X ou E(X), s'appelle la MOYENNE de X et c'est un indicateur de position centrale de la distribution

$$\mu'_1 = \mu = \mu_X = E(X) \qquad (4.14)$$

Par exemple, une variable distribuée selon une loi exponentielle de paramètre λ a une moyenne

$$\mu = 1/\lambda$$

Notons la définition et le résultat suivant: une distribution est SYMÉTRIQUE par rapport à un nombre a si

$$\begin{aligned} p_X(a + x) &= p_X(a - x) \qquad \text{pour tout } x \\ f_X(a + x) &= f_X(a - x) \qquad \text{pour tout } x \end{aligned} \tag{4.15}$$

où $p_X(.)$ et $f_X(.)$ représentent une masse ou densité selon que la variable est discrète ou continue. Pour une telle distribution il suit que

$$\mu = E(X) = a \tag{4.16}$$

Exemple 4.5: suite de l'exemple 4.1

La masse de probabilité est symétrique par rapport à $x = 7$ d'où

$$\mu = E(X) = 7$$

Moments centrés

Soit $k = 1,2,...$ un entier quelconque; le k-ième MOMENT CENTRÉ de X noté μ_K par rapport à la moyenne μ est défini par

$$\mu_k = \begin{cases} \Sigma(x_\alpha - \mu)^k p_X(x_\alpha) & \text{si X est discrète} \\ \int_{-\infty}^{\infty} (x - \mu)^k f_X(x)dx & \text{si X est continue} \end{cases} \tag{4.17}$$

En particulier, tout moment centré d'ordre impair d'une distribution symétrique est zéro.

Variance

En particulier μ_2, aussi noté σ_X^2 ou VAR(X), s'appelle la la VARIANCE et représente un indicateur de dispersion

$$\mu_2 = \sigma_x^2 = VAR(X) \tag{4.18}$$

Pour le calcul de la variance il est plus commode d'employer l'équation

$$\sigma_X^2 = \mu_2' - \mu_X^2 \tag{4.19}$$

Exemple 4.6: distribution exponentielle

$$\sigma_X^2 = 2/\lambda^2 - 1/\lambda^2 = 1/\lambda^2 \qquad (4.20)$$

en utilisant les équations (4.13) et (4.19).

Écart-type

Une mesure de variabilité exprimée dans l'unité de X est l'ÉCART-TYPE noté σ_x ou ET(x) et est définie par

$$\sigma_x = \sqrt{\sigma_x^2} = ET(X) \qquad (4.21)$$

Exemple 4.7: suite des exemples 4.1 et 4.5

$$\sigma_x^2 = 2^2 * 1/36 + 3^2 * 2/36 + \ldots + 12^2 * 1/36 - (7)^2 = 35/6$$

$$\sigma_x = \sqrt{5.83} = 2.415$$

Coefficient d'asymétrie

Comme son nom l'indique, il sert à mesurer le degré de non symétrie de la distribution; il est noté β_1 et défini par

$$\beta_1 = \mu_3/\sigma^3 \qquad (4.22)$$

En particulier toute distribution symétrique a un coefficient β_1 zéro

$$\beta_1 = 0$$

Le coefficient β_1 est un nombre pur et indépendant de l'unité employée.

Exemple 4.8 distribution exponentielle

Utilisant les équations (4.13), (4.20) et (4.22) on obtient

$$\sigma = 1/\lambda$$

$$\mu_3 = \int_0^\infty (x - 1/\lambda)^3 \lambda e^{-\lambda x} dx$$

Si on développe $(x-1/\lambda)^3$ et on utilise l'équation (4.13) il vient

$$\mu_3 = 6/\lambda^3 - 3(1/\lambda)(2/\lambda^2) + 3(1/\lambda^2)(1/\lambda) - 1/\lambda^3$$
$$= 2/\lambda^3$$

Alors

$$\beta_1 = (2/\lambda^3)/(1/\lambda^3) = 2$$

Coefficient d'aplatissement

Il est noté β_2 et sert à mesurer l'importance des extrémités (disons $(-\infty,\mu-\sigma)\cup(\mu+\sigma,\infty)$) de la distribution par rapport à sa partie centrale et défini par

$$\beta_2 = \frac{\mu_4}{\sigma^4} - 3 \qquad (4.23)$$

On soustrait la valeur 3 pour fin de comparaison avec le modèle gaussien ou normal puisque dans ce cas $\mu_4/\sigma^4 = 3$.

Exemple 4.9: distribution exponentielle

Par définition

$$\mu_4 = \int_0^\infty (x - 1/\lambda)^4 \lambda e^{-\lambda x} dx$$

$$= \int_0^\infty (x^4 - 4x^3(1/\lambda) + 6x^2(1/\lambda^2) - 4x(1/\lambda^3) + 1/\lambda^4) e^{-\lambda x} dx$$

$$= \mu_4' - (4/\lambda)\mu_3' + (6/\lambda^2)\mu_2' - (4/\lambda^3)\mu_1' + 1/\lambda^4$$

Utilisant l'équation (4.13)

$$\mu_k' = k!/\lambda^k$$

on obtient

$$\mu_4 = 9/\lambda^4$$

et finalement

$$\beta_2 = \mu_4/\sigma^4 - 3 = (9/\lambda^4)/(1/\lambda^4) - 3 = 6$$

C'est une distribution dont les extrémités sont plus importantes que celles d'une loi normale.

Distinction importante

Dans les applications statistiques, estimation et tests d'hypothèses, il est impératif de distinguer les indicateurs associés à un modèle et les indicateurs dérivant des données. La notation et le contexte permettent de situer si les quantités dérivent du modèle ou des données.

Modèle		Données
X		$x_1, x_2, \ldots, x_n$
loi $f_X(.)$		tableau d'effectifs
p:probabilité		f:fréquence
$F_X(.)$	répartition	$F_n(x)$
μ	moyenne	$\bar{x}$
σ	écart-type	s
β_1	coefficient d'asymétrie	b_1
β_2	coefficient d'aplatissent	b_2
x_p	percentile	Y_p

4.3 TRANSFORMATIONS

Soit X une variable aléatoire de masse $p_X(x)$ (ou densité $f_X(x)$) et soit g une application de R dans R . Alors $g(X) = Y$ est une nouvelle variable aléatoire dont on peut calculer la loi en faisant certaines hypothèses sur la nature de X et la transformation g . On peut établir les propositions suivantes.

Proposition 4.1:

Soit X une variable discrète. Alors la masse de probabilité de $Y = g(X)$, notée $p_Y(.)$ peut se calculer selon la formule

$$p_Y(y) = \sum_{A_x} p_X(x) \qquad (4.24)$$

où $A_x = \{x: g(x) = y\}$

et $p_X(.)$ est la fonction de masse de X

Exemple 4.10: transformation d'une variable discrète

Soit X une variable discrète avec la masse de probabilité

X	-1	1	2	3
$p_X(x)$	1/4	1/4	1/3	1/6

Soit la transformation

$$g(X) = X^2 = Y$$

Alors

$$p_Y(1) = p_X(-1) + p_X(1) = 1/2$$

$$p_Y(4) = p_X(2) = 1/3$$

$$p_Y(9) = p_X(3) = 1/6$$

D'où

Y	1	4	9
$p_Y(y)$	1/2	1/3	1/6

Proposition 4.2:

Soit X une variable continue, f_X sa fonction de densité et g une transformation continue, croissante et dérivable. Alors

$$f_Y(y) = f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y) \qquad (4.25)$$

où $f_Y(y)$ est la densité de $Y = g(X)$ et $g^{-1}(.)$ est la fonction réciproque de g .

Remarque: si g est décroissante on remplace $\frac{d}{dy} g^{-1}(y)$ par

$$- \frac{d}{dy} g^{-1}(y)$$

Exemple 4.11: loi gaussienne centrée-réduite

Nous avons déjà défini à l'exemple 3.7 la distribution gaussienne de paramètres μ et σ:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad x \in R$$

Effectuons la transformation affine

$$y = g(x) = (x - \mu)/\sigma$$

dont la réciproque est

$$x = g^{-1}(y) = \mu + \sigma y$$

Alors

$$\frac{d}{dy} g^{-1}(y) = \sigma$$

et la densité de $Y = (X - \mu)/\sigma$ est

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} y^2\right] \qquad y \in R$$

une loi gaussienne de paramètres $\mu = 0$ et $\sigma = 1$ appelée gaussienne centrée-réduite. Nous étudierons ses propriétés plus loin.

Exemple 4.12: distribution log-normale

Soit X une variable positive dont le logarithme $\ell nX = Y$ suit une distribution normale de moyenne ξ et d'écart-type τ:

$$f_Y(y) = \frac{1}{\tau\sqrt{2\pi}} \exp\left[\frac{1}{2} \left(\frac{y-\xi}{\tau} \right)^2 \right]$$

Déterminons la densité de probabilité de X. On a

$$X = e^Y = g(Y)$$

$$g^{-1}(x) = y = \ell nx$$

$$\frac{d}{dx} g^{-1}(x) = \frac{1}{x}$$

et finalement

$$f_X(x) = \frac{1}{\tau\sqrt{2\pi}} \frac{1}{x} \exp\left[-\frac{1}{2} \left(\frac{\ell nx-\xi}{\tau} \right)^2 \right]$$

Cette distribution sera étudiée au chapitre 6.

Proposition 4.3:

Soit X une variable continue et g une transformation continue et dérivable. Alors

$$F_Y(y) = F_X(g^{-1}(y) \tag{4.26}$$

où $F_Y(.)$ est la fonction de répartition de $Y = g(X)$

et $F_X(.)$ est la fonction de répartition de X

Exemple 4.13: distribution khi-deux avec un degré de liberté

Soit X une variable gaussienne centrée-réduite

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x^2\right] \qquad x \in R$$

et posons $\Phi_X(x) = \int_{-\infty}^{x} f_X(t)dt$ sa fonction de répartition.

Soit $y = g(x) = x^2$ la transformation considérée et $Y = g(X) = X^2$ la nouvelle variable aléatoire correspondante. On a pour $y > 0$

$$\begin{aligned} F_Y(y) &= \Phi_X(\{x: X^2 \le y\}) \\ &= \Phi_X(\{x: -\sqrt{y} \le X \le \sqrt{y}\}) \\ &= \Phi_X(\sqrt{y}) - \Phi_X(-\sqrt{y}) \end{aligned}$$

La densité de Y est

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = (1/2\sqrt{y})\, \Phi'_X(\sqrt{y}) + (1/2\sqrt{y})\, \Phi'_X(-\sqrt{y}) \\ &= (1/2\sqrt{y})[f_X(\sqrt{y}) + f_X(\sqrt{y})] \\ &= (1/2\sqrt{2\pi})y^{(1/2)-1}\left[\exp\left(-\frac{1}{2}y\right) + \exp\left(-\frac{1}{2}y\right)\right] \\ &= (1/\sqrt{2\pi})y^{(1/2)-1}\left[\exp\left(-\frac{1}{2}y\right)\right] \end{aligned}$$

Cette distribution s'appelle DISTRIBUTION KHI-DEUX avec 1 degré de liberté et ses propriétés seront étudiées au chapitre 6.

Exemple 4.14: Soit $F_X(.)$ la fonction de répartition d'une variable continue X et Y une nouvelle variable définie par:

$$Y = F_X(X)$$

Alors, Y est distribuée uniformément dans l'intervalle [0,1].

C'est-à-dire:

$$F_Y(y) = \begin{cases} 0 & y<0 \\ y & 0\leq y\leq 1 \\ 1 & y\geq 1 \end{cases}$$

En effet,

$$\begin{aligned} F_Y(y) &= P[Y\leq y] = P[F_X(X)\leq y] \\ &= P[X \leq F_X^{-1}(y)] \\ &= F_X[F_X^{-1}(y)] \\ &= y \end{aligned}$$

Ce résultat est employé en simulation stochastique pour définir des fonctions génératrices de nombres aléatoires selon diverses distributions $F_X(x)$ en utilisant un générateur d'une distribution uniforme.

Espérance mathématique

Soit X une variable aléatoire et $y = g(x)$ une transformation de R dans R . On définit L'ESPÉRANCE MATHÉMATIQUE de $g(X)$, notée $E(g(X))$, par

$$E(g(X)) = \begin{cases} \sum_{x_\alpha} g(x_\alpha)p_X(x_\alpha) & \text{si X est discrète} \\ \int_{-\infty}^{\infty} g(x)f_X(x)dx & \text{si X est continue} \end{cases} \qquad (4.27)$$

Remarques:

. La définition proposée est en fait une conséquence de la définition de la moyenne de la nouvelle variable $Y = g(X)$ et des équations de passage pour obtenir la loi de probabilité de Y à partir de celle de X:

$$E(Y) = \sum_y y\, p_Y(y) = \sum_x g(x)\, p_X(x) = E(g(Y))$$

si la variable X est discrète et

$$E(Y) = \int_{-\infty}^{\infty} y\, f_Y(y)dy = \int_{-\infty}^{\infty} g(x)\, f_X(x)dx = E(g(X))$$

si la variable X est continue.

. Les différentes caractéristiques: les moments, la moyenne, la variance sont des cas particuliers. Par exemple

$$g(X) = X , \qquad E(g(X)) = \mu_X = \text{la moyenne de } X$$

$$g(X) = (X - \mu_X)^2 , \qquad E(g(X)) = \sigma_X^2 = \text{la variance de } X$$

. Le concept d'espérance mathématique peut être employé comme critère de décision dans des situations d'incertitude.

Exemple 4.15: MINI-LOTO

On peut définir l'espace de probabilité suivant pour décrire la loterie MINI LOTO

$$\Omega = \left\{ (x_1,x_2,x_3,x_4,x_5,x_6): \; x_j = 0,1,2,\ldots,9 \;\; j = 1,2,3,4,5,6 \right\}$$

Le nombre de billets distincts est 10^6 et notons par $(x_1^*, x_2^*, x_3^*, x_4^*, x_5^*, x_6^*)$ le résultat du tirage. La structure des lots ainsi que le nombre de billets correspondants est donné dans le tableau

Lot	billet	nombre
50 000$	$(x_1^*, x_2^*, x_3^*, x_4^*, x_5^*, x_6^*)$	1
5 000$	$(y_1, x_2^*, x_3^*, x_4^*, x_5^*, x_6^*)$ $y_1 \neq x_1^*$	9
250$	$(y_1, y_2, x_3^*, x_4^*, x_5^*, x_6^*)$ $y_2 \neq x_2^*$	90
50$	$(y_1, y_2, y_3, x_4^*, x_5^*, x_6^*)$ $y_3 \neq x_3^*$	900
5$	$(y_1, y_2, y_3, y_4, x_5^*, x_6^*)$ $y_4 \neq x_4^*$	9 000

Si on pose X = lot à gagner on a

$$E(X) = (1/10^6)[50\,000 + 9*5\,000 + 90*250 + 900*50 + 9000*5]$$

$$= 0.2075\$$$

C'est la valeur moyenne ou espérance mathématique de la MINI LOTO. En comparaison du prix d'un billet de 0.50$, on est donc perdant, en moyenne, en jouant à cette loterie.

Propriétés de l'espérance mathématique

L'espérance mathématique est une fonctionnelle linéaire: si X_1 et X_2 sont deux variables aléatoires et α un nombre réel quelconque

$$E(X_1 + X_2) = E(X_1) + E(X_2) \qquad (4.28)$$

$$E(\alpha X_1) = \alpha E(X_1)$$

De plus si $$Y = \beta_0 + \beta_1 X$$

Alors $$E(Y) = \beta_0 + \beta_1 E(X)$$

$$VAR(Y) = \beta_1^2 VAR(X)$$

$$ET(Y) = \beta_1 ET(X)$$

4.4 COUPLE DE VARIABLES ALÉATOIRES DISCRÈTES

Considérons deux variables aléatoires discrètes X et Y définies sur le même espace de probabilité (Ω, P). Soit $(x_1, x_2, \ldots, x_r)$ les valeurs prises par X sur les événements de la partition induite $(A_1, A_2, \ldots, A_r)$ et $(Y_1, Y_2, \ldots, Y_c)$ les valeurs prises par Y sur les événements de la partition induite $(B_1, B_2, \ldots, B_c)$. Le comportement du coupe (X, Y) se ramène à l'étude du comportement sur les événements $A_\alpha \cap B_\beta$ dont l'ensemble constitue une partition de Ω. Par exemple si $r = 3$ et $c = 2$, on a le tableau à double entrées

Partition de Ω

X \ Y	Y_1	Y_2	
x_1	$A_1 \cap B_1$	$A_1 \cap B_2$	A_1
x_2	$A_2 \cap B_1$	$A_2 \cap B_2$	A_2
x_3	$A_3 \cap B_1$	$A_3 \cap B_2$	A_3
	B_1	B_2	Ω

A l'intersection de la ligne α et de la colonne β on trouve l'événement $A_\alpha \cap B_\beta$ correspondant à $X = X_\alpha$ et $Y = Y_\beta$. Notons que

$$\bigcup_{\beta=1}^{c} (A_\alpha \cap B_\beta) = A_\alpha \cap \left(\bigcup_{\beta=1}^{c} B_\beta \right) = A_\alpha$$

$$\bigcup_{\alpha=1}^{r} (A_\alpha \cap B_\beta) = \left(\bigcup_{\alpha=1}^{r} A_\alpha \right) \cap B_\beta = B_\beta$$

La fonction $p_{X,Y}(., .)$ définie par

$$p_{X,Y}(x_\alpha, y_\beta) = P(A_\alpha \cap B_\beta) = P(X = x_\alpha, Y = y_\beta) \qquad (4.29)$$

s'appelle MASSE DE PROBABILITÉ CONJOINTE du couple (X, Y). Elle vérifie les conditions suivantes:

$$\begin{aligned} &0 \le p_{X,Y}(x, y) \le 1 \\ &\sum_x \sum_y p_{X,Y}(x, y) = 1 \end{aligned} \qquad (4.30)$$

Le tableau suivant exhibe la loi conjointe de (X, Y) pour $r = 3$ et $c = 2$.

Distribution de (X,Y) $\quad r = 3 \quad c = 2$

X \ Y	y_1	y_2	$P_X(\ .\ )$
x_1	$P_{X,Y}(x_1,y_1)$	$P_{X,Y}(x_1,y_2)$	$P_X(x_1)$
x_2	$P_{X,Y}(x_2,y_1)$	$P_{X,Y}(x_2,y_2)$	$P_X(x_2)$
x_3	$P_{X,Y}(x_3,y_1)$	$P_{X,Y}(x_3,y_2)$	$P_X(x_3)$
$P_y(\ .\ )$	$P_Y(y_1)$	$P_Y(y_2)$	1

On définit les LOIS MARGINALES:

$$\begin{aligned} p_X(x) &= \sum_y p_{X,Y}(x,y) \\ p_Y(y) &= \sum_x p_{X,Y}(x,y) \end{aligned} \tag{4.31}$$

et les LOIS CONDITIONNELLES:

$$\begin{aligned} p_{X|Y=y}(x) &= p_{X,Y}(x,y)/p_Y(y) \\ p_{Y|X=x}(y) &= p_{X,Y}(x,y)/p_X(x) \end{aligned} \tag{4.32}$$

en accord avec la définition de probabilité conditionnelle. D'autre part, l'équation (4.32) permet d'écrire:

$$\begin{aligned} p_{X,Y}(x,y) &= p_X(x)p_{Y|X=x}(y) \\ &= p_Y(y)p_{X|Y=y}(x) \end{aligned} \tag{4.33}$$

Exemple 4.16: Le tableau suivant donne la distribution conjointe du couple de variables (X,Y) = (HUMIDITÉ, TEMPERATURE) établit dans une région donnée sur une période de plusieurs années.

X \ Y	1	2	3	4	5	6	$p_X(x)$
1	0.01	0.02	0.02	0.03	0.02	0.01	0.11
2	0.02	0.04	0.06	0.08	0.05	0.01	0.26
3	0.02	0.05	0.08	0.11	0.07	0.02	0.35
4	0.01	0.03	0.06	0.05	0.03	0.02	0.20
5	0.00	0.01	0.02	0.03	0.01	0.01	0.08
$p_Y(y)$	0.06	0.15	0.24	0.30	0.18	0.07	1.00

Signification des codages

X(%)
1 = 0 - 20
2 = 20 - 40
3 = 40 - 60
4 = 60 - 80
5 = 80 - 100

Y(°C)
1 = 0 ou moins
2 = 0 - 5
3 = 5 - 10
4 = 10 - 15
5 = 15 - 20
6 = 20 ou plus

lois conditionnelles $p_{X|Y=y}(x)$

X \ Y	1	2	3	4	5	6
1	0.167	0.133	0.083	0.100	0.111	0.143
2	0.333	0.267	0.250	0.267	0.278	0.143
3	0.333	0.333	0.333	0.366	0.389	0.286
4	0.167	0.200	0.250	0.167	0.167	0.285
5	0.000	0.067	0.084	0.100	0.055	0.143
Total	1.000	1.000	1.000	1.000	1.000	1.000

lois conditionnelles $p_{Y|X=x}(y)$

X \ Y	1	2	3	4	5	6	total
1	0.091	0.182	0.182	0.272	0.182	0.091	1.00
2	0.077	0.154	0.231	0.308	0.192	0.038	1.00
3	0.057	0.143	0.229	0.314	0.200	0.057	1.00
4	0.050	0.150	0.300	0.250	0.150	0.100	1.00
5	0.000	0.125	0.250	0.375	0.125	0.125	1.00

Indépendance entre deux variables aléatoires

Intuitivement, la variable aléatoire X est indépendante de la variable aléatoire Y si connaître la valeur x prise par X n'a pas d'incidence sur la distribution conditionnelle de Y . En d'autres termes toutes les distributions conditionnelles $p_{Y|X=x}(y)$ sont identiques à la distribution marginale $p_Y(y)$ de Y . Par symétrie il faut aussi que toutes les distributions conditionnelles $p_{X|Y=y}$ soient identiques à la distribution marginale $p_X(x)$ de X . Nous proposons donc la définition suivante:

Définition (version 1): Deux variables aléatoires discrètes (X,Y) sont INDÉPENDANTES si

$$p_{X|Y=y}(x) = p_X(x) \quad \text{pour tout } (x,y)$$
$$p_{Y|X=x}(y) = p_Y(y) \quad \text{pour tout } (x,y) \qquad (4.34)$$

L'équation (4.33) permet de reformuler cette définition par une seule équation.

Définition (version 2): Deux variables aléatoires discrètes (X,Y) sont INDÉPENDANTES si

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) \quad \text{pour tout } (x,y) \qquad (4.35)$$

Exemple 4.17: suite de l'exemple 4.16

On constate que le couple (X,Y) ne définit pas deux variables indépendantes puisque toutes les lois conditionnelles sont différentes. Cela est aussi confirmé par la distribution conjointe où par exemple

et $p_X(x=1)p_Y(y=1) = 0.11 * 0.06 = 0.066 \neq 0.01 = p_{X,Y}(x=1,y=1)$

4.5 COUPLE DE VARIABLES ALÉATOIRES CONTINUES

On considère deux variables aléatoires X et Y sur le même espace probabilisé (Ω,P). On définit la FONCTION DE RÉPARTITION CONJOINTE $F_{X,Y}(x,y)$:

$$F_{X,Y}(x,y) = P[X \leq x, \quad Y \leq y] \tag{4.36}$$

On dira qu'il s'agit d'un couple de variables continues s'il existe une fonction $f_{X,Y}(x,y)$ appelée DENSITÉ CONJOINTE telle que:

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v)\,dv\,du \tag{4.37}$$

La densité conjointe $f_{X,Y}(x,y)$ peut s'obtenir de la fonction de répartition conjointe $F_{X,Y}(x,y)$

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) \tag{4.38}$$

Toute fonction de densité conjointe $f_{X,Y}(x,y)$ vérifie les conditions suivantes:

$$\begin{aligned} & f_{X,Y}(x,y) \geq 0 \\ & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1 \end{aligned} \tag{4.39}$$

Par analogie avec le cas discret, on définit les concepts suivants.

Densités marginales

$$\begin{aligned} \text{de X:} \quad f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \\ \text{de Y:} \quad f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx \end{aligned} \tag{4.40}$$

Densités conditionnelles

$$\begin{aligned} &\text{de X étant donné Y=y :} \quad f_{X|Y=y}(x) = f_{X,Y}(x,y)/f_Y(y) \\ &\text{de Y étant donné X=x :} \quad f_{Y|X=x}(y) = f_{X,Y}(x,y)/f_X(x) \end{aligned} \tag{4.41}$$

Indépendance: deux variables continues sont indépendantes si:

$$f_{X,Y}(x,y) = f_X(x)\ f_Y(y) \qquad \text{tout } (x,y)$$

Exemple 4.18: Soit (X,Y) deux variables aléatoires continues et définissons la densité suivante:

$$f_{X,Y}(x,y) = \begin{cases} c(5 - x - 0.5y) & 0 \le x \le 2,\ 0 \le y \le 2 \\ & c > 0 \\ 0 & \text{ailleurs} \end{cases}$$

Le lecteur vérifiera les résultats:

- $c = 1/14$
- $f_X(x) = (9 - 2x)/14$
- $f_Y(y) = (8 - y)/14$
- $f_{X|Y=y}(x) = (5 - 0.5y - x)/(8 - y)$
- $f_{Y|X=x}(y) = (5 - 0.5y - x)/(9 - 2x)$
- les variables ne sont pas indépendantes

$$F_{X,Y}(x,y) = \begin{cases} 0 \quad \text{si} \quad x<0 \text{ et tout } y \\ 0 \quad \text{si} \quad y<0 \text{ et tout } x \\ c(5xy - 0.25xy^2 - 0.50x^2y) \quad \text{si} \quad 0<x<2 \text{ et } 0<y<2 \\ c(8y - 0.50y^2) \quad \text{si} \quad x \ge 2 \text{ et } 0<y<2 \\ c(9x - x^2) \quad \text{si} \quad 0<x<2 \text{ et } y \ge 2 \\ 1 \quad \text{si} \quad x \ge 2 \text{ et } y \ge 2 \end{cases}$$

4.6 INDICATEURS ASSOCIÉES À UN COUPLE

La loi de probabilité conjointe contient toute l'information relative à un couple de variables. Il est commode de résumer le comportement du couple par quelques nombres dont les moments d'ordre 1 et 2.

Espérance

Soit (X, Y) un couple de variables et $p_{X,Y}$ ou $f_{X,Y}$ sa masse ou densité conjointe selon le cas discret ou continu. Soit h(x,y) une fonction à valeur réelle et Z = h(X,Y) une transformation. L'ESPÉRANCE MATHÉMATIQUE DE Z est définie par

$$E(h(X,Y)) = \begin{cases} \sum_x \sum_y h(x,y)\ p_{X,Y}(x,y) \\ \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y)\ f_{X,Y}(x,y)\ dxdy \end{cases} \qquad (4.42)$$

Des choix particuliers de h définissent les quantités suivantes:

Moyennes

$$\text{de } X : \mu_X = E(X)$$

$$\text{de } Y : \mu_Y = E(Y)$$

Variances

$$\text{de } X : \sigma_X^2 = E[(X - \mu_X)^2] = VAR(X)$$

$$\text{de } Y : \sigma_Y^2 = E[(Y - \mu_Y)^2] = VAR(Y)$$

Covariance

$$\sigma_{XY} = COV(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] \qquad (4.43)$$

Coefficient de corrélation

$$\rho_{XY} = COV(X,Y)/\sigma_X\sigma_Y \qquad (4.44)$$

La démonstration des propositions suivantes est laissée au lecteur.

Proposition 4.4:

(a) $VAR(\alpha X + \beta) = \alpha^2 VAR(X)$ tout (α, β)

(b) $COV(X,Y) = E(XY) - E(X)E(Y)$

(c) $VAR(aX + bY) = a^2\ VAR(X) + b^2\ VAR(Y) + 2ab\ COV(X,Y)$

(d) $-1 \leq \rho_{XY} \leq 1$

(e) Si $Y = \alpha X + \beta$ alors $\rho_{XY} = -1$ si $\alpha < 0$
et $\rho_{XY} = 1$ si $\alpha > 0$

(f) Si $\rho_{XY} = \pm 1$ alors $Y = \alpha X + \beta$

Proposition 4.5:

Si X,Y sont des variables aléatoires indépendantes alors:

(a) $E(XY) = E(X)E(Y)$

(b) $COV(X,Y) = 0$

(c) $\rho_{XY} = 0$

(d) $VAR(X + Y) = VAR(X) + VAR(Y)$

Remarques

- Le coefficient de corrélation $\rho(X,Y)$ est principalement une mesure de dépendance linéaire $Y = \alpha X + \beta$ entre deux variables

- Une corrélation zéro entre deux variables n'implique pas l'indépendance en général; la corrélation entre X et $Y = X^2$ est zéro si la distribution de X est symétrique par rapport à 0 car

$$E(XY) = E(X^3) = 0$$
$$E(X) = 0$$
$$COV(X,Y) = E(XY) - E(X)E(Y) = 0$$
$$\rho_{XY} = 0$$

Par exemple, la masse de probabilité conjointe suivante

X \ Y	0	1	4	$p_X(x)$
-2	0	0	1/5	1/5
-1	0	1/5	0	1/5
0	1/5	0	0	1/5
1	0	1/5	0	1/5
2	0	0	1/5	1/5
$p_Y(y)$	1/5	2/5	2/5	1

est telle que $Y = X^2$, $p_X(x)$ est symétrique par rapport à 0 mais les variables sont dépendantes puisque

$$p_{X,Y}(x,y) \neq p_X(x)p_Y(y) \quad \text{pour tout } (x,y)$$

Le coefficient de corrélation est erratique comme mesure de dépendance fonctionnelle entre deux variables; la valeur du coefficient ρ_{XY} peut être faible ou élevée lorsque

$$Y = g(X) \quad \text{où} \quad g(X) \neq \alpha X + \beta$$

dépendamment de la distribution conjointe de (X,Y); en d'autres termes on ne peut utiliser le coefficient ρ_{XY} comme un indicateur de la forme de la fonction g (s'il en existe une!) entre X et Y .

Exemple 4.19 suite de l'exemple 4.16

$$\mu_X = 1*0.11 + 2*0.26 + 3*0.35 + 4*0.20 + 5*0.08$$
$$= 2.88$$

$$\mu_Y = 1*0.06 + 2*0.15 + 3*0.24 + 4*0.30 + 5*0.18 + 6*0.07$$
$$= 3.60$$

$$\sigma_X^2 = 1^2*0.11 + 2^2*0.26 + 3^2*0.35 + 4^2*0.20 + 5^2*0.08 - \mu_X^2$$
$$= 9.50 - (2.88)^2 = 1.2056$$
$$\sigma_X = 1.098$$

$$\sigma_Y^2 = 1^2*0.06 + 2^2*0.15 + 3^2*0.24 + 4^2*0.30 + 5^2*0.18 + 6^2*0.07 - \mu_Y^2$$
$$= 14.64 - (3.60)^2 = 1.68$$
$$\sigma_Y = 1.296$$

$$E(XY) = 1*1*0.01 + 1*2*0.02 + \ldots + 5*6*0.01$$
$$= 10.45$$

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y = 10.45 - 2.88*3.60 = +0.082$$

$$\rho_{XY} = \sigma_{XY}/\sigma_X\sigma_Y = 0.082/(1.098*1.296) = 0.058$$

La corrélation linéaire entre la variable température et la variable humidité est faible.

Exemple 4.20: suite de l'exemple 4.18

$$\mu_X = E(X) = \int_0^2 x\,(9-2x)/14\,dx = 19/21$$

$$\mu_Y = E(Y) = \int_0^2 y\,(8-y)/14\,dy = 20/21$$

$$E(X^2) = \int_0^2 x^2\,(9-2x)/14\,dx = 8/7$$

$$\sigma_X^2 = E(X^2) - \mu_X^2 = 143/441$$

$$\sigma_X = \sqrt{143/21}$$

$$E(Y^2) = \int_0^2 y^2\,(8-y)/14\,dy = 26/21$$

$$\sigma_Y^2 = E(Y^2) - \mu_Y^2 = 146/441$$

$$\sigma_Y = \sqrt{146/21}$$

$$E(XY) = \int_0^2 \int_0^2 xy\,(5 - x - y/2)/14\,dxdy = 6/7$$

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y = 6/7 - (19/21)*(20/21) = -2/441$$

$$\rho_{XY} = \sigma_{XY}/\sigma_X*\sigma_Y = -2/(\sqrt{143})(\sqrt{146}) = -0.0138$$

Distinction importante

Dans les applications statistiques il est impératif de distinguer les caractéristiques associées à un modèle et les caractéristiques dérivant des données. La notation et le contexte permettent de situer si les quantités dérivent du modèle ou des données.

Modèle		Données
(X,Y)		(x_α, y_α) $\alpha = 1,\ldots,n$
loi $f_{X,Y}(x,y)$		tableau d'effectifs conjoints
probabilité		fréquence
μ_X, μ_Y	moyennes	$\bar{x}$, $\bar{y}$
σ_X, σ_Y	écarts-types	s_x, s_y
σ_{XY}	covariance	s_{XY}
ρ_{XY}	corrélation	r_{XY}

4.7 VECTEURS ALÉATOIRES

Le concept de distribution conjointe d'un couple de variables aléatoires peut se généraliser à un nombre quelconque de variables. Nous présentons les définitions dans le cas où toutes les variables sont continues et nous laissons au lecteur le soin de les adapter lorsque toutes les variables sont discrètes.

Définition

Soit (Ω,P) un espace de probabilité. Une application $X = (X_1,X_2,\ldots,X_p)$ de Ω dans R^p est appelée un VECTEUR ALÉATOIRE à p dimensions

$$X : \Omega \longrightarrow R^p$$

Chaque composante de X est une variable aléatoire

$$X_\alpha : \Omega \longrightarrow R \qquad \alpha = 1,2,\ldots,p$$

La mesure de probabilité P définit une fonction de R^p dans R appelée FONCTION DE RÉPARTITION CONJOINTE, notée $F_{X_1,X_2,\ldots,X_p}(x_1,x_2,\ldots,x_p)$.

$$F_{X_1,X_2,\ldots,X_p}(x_1,x_2,\ldots,x_p)$$

$$= P[X_1 \leq x_1, X_2 \leq x_2, \ldots, X_p \leq x_p] \tag{4.45}$$

On dit que X est un VECTEUR CONTINU si ses composantes sont des variables aléatoires continues. Dans ce cas on peut caractériser le vecteur par une DENSITÉ DE PROBABILITÉ CONJOINTE, notée $f_{X_1,X_2,\ldots,X_p}(x_1,x_2,\ldots,x_p)$ telle que

$$f_{X_1,X_2,\ldots,X_p}(x_1,x_2,\ldots,x_p) \geq 0$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{X1,X2,\ldots,Xp}(x_1, x_2,\ldots,x_p) \prod_{\alpha=1}^{p} dx_\alpha = 1 \tag{4.46}$$

$$F_{X_1,X_2,\ldots,X_p}(x_1,x_2,\ldots,x_p)$$

$$= \int_{-\infty}^{xp}\cdots\int_{-\infty}^{x2}\int_{-\infty}^{x1} f_{X1,X2,\ldots,Xp}(u_1, u_2, \ldots, u_p) \prod_{\alpha=1}^{p} du_\alpha \tag{4.47}$$

On peut obtenir de la densité conjointe, la DENSITÉ MARGINALE de la composante X_α de X en intégrant toutes les autres composantes sur R^{p-1}

$$f_{X\alpha}(x) =$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X_1,X_2,\ldots,X_p}(u_1,u_2,\ldots,u_p) \prod_{\beta\neq\alpha} du_\beta \qquad (4.48)$$

De même on peut obtenir la DENSITÉ CONJOINTE MARGINALE d'un couple de composantes (X_α, X_β) de X en intégrant toutes les autres composantes sur R^{p-2}.

Les caractéristiques de base d'un vecteur aléatoire sont fournies par les moments d'ordre 1 et 2, soient les moyennes, les variances, les covariances et les corrélations.

Moyennes

$$\mu_\alpha = E(X_\alpha) \qquad \alpha = 1,2,\ldots,p$$
$$E(X) = (\mu_1,\mu_2,\ldots,\mu_p) \qquad (4.49)$$

Matrice de covariance

$$\sigma_{\alpha\beta} = E((X_\alpha - \mu_\alpha)(X_\beta - \mu_\beta)) \qquad \alpha,\beta = 1,2,\ldots,p$$
$$= \text{covariance entre } X_\alpha \text{ et } X_\beta$$
$$\sigma_{\alpha\alpha} = E((X_\alpha - \mu_\alpha)^2) \qquad \alpha = 1,2,\ldots,p \qquad (4.50)$$
$$= \text{variance de } X_\alpha = \sigma_\alpha^2$$

$$\Sigma = [\sigma_{\alpha\beta}] \quad \text{matrice symétrique } p \times p \qquad (4.51)$$

appelée MATRICE DE COVARIANCE dont la diagonale principale contient les variances de chacune des variables. La corrélation entre X_α et X_β est:

$$\rho_{\alpha\beta} = \sigma_{\alpha\beta} / \sqrt{\sigma_{\alpha\alpha}\ \sigma_{\beta\beta}} \qquad \alpha\beta = 1,2,\ldots,p \qquad (4.52)$$

On note que $\rho_{\alpha\alpha} = 1$ et $\rho_{\alpha\beta} = \rho_{\beta\alpha}$

On peut rassembler toutes les corrélations dans une matrice $p \times p$ symétrique appelée MATRICE DE CORRÉLATION.

$$
\begin{array}{cc}
 & \text{variable} \\
 & \begin{array}{cccccc} 1 & 2 & & \cdots & & p \end{array} \\
\text{variable} & \begin{array}{c} 1 \\ 2 \\ \vdots \\ p \end{array} \left[\begin{array}{cccccc} 1 & & & & \text{symétrique} & \\ \rho_{21} & 1 & & & & \\ \vdots & & & & & \\ \rho_{p1} & \rho_{p2} & & \cdots & & 1 \end{array} \right]
\end{array}
\tag{4.53}
$$

D'un point de vue théorique et comme point de départ de plusieurs techniques d'analyse statistique multidimensionnelle on suppose souvent la loi gaussienne (normale) à plusieurs dimensions. Cette loi (distribution) généralise la distribution gaussienne unidimensionnelle de l'exemple 3.7. Nous ne donnerons pas ici l'équation de définition de la densité multinormale car cela dépasse les besoins immédiats de ces notes axées sur des techniques unidimensionnelles. Le besoin de traiter les concepts de vecteurs aléatoires est de permettre de fixer le langage mathématique de l'ÉCHANTILLONNAGE et ses applications au raisonnement statistique. En particulier, le concept de variables indépendantes est essentiel pour la suite.

<u>Définition</u>

Les variables $X_1, X_2, \ldots, X_p$ sont INDÉPENDANTES si

$$
f_{X_1, X_2, \ldots, X_p}(x_1, x_2, \ldots, x_p) = \prod_{\alpha=1}^{p} f_{X\alpha}(x_\alpha) \tag{4.54}
$$

où $f_{X_1, X_2, \ldots, X_p}(.)$ est la densité conjointe de $(X_1, X_2, \ldots, X_p)$

et $f_{X\alpha}(.)$ est la densité marginale de x_α.

Nous consignons pour fins de référence, les résultats concernant les caractéristiques de combinaisons linéaires de variables aléatoires ainsi que pour une fonction quelconque.

Proposition 4.6:

Soient $X_1, X_2, \ldots, X_p$ des variables aléatoires et $a_1, a_2, \ldots, a_p$ des nombres réels. Alors

(a) $$E\left(\sum_{\alpha=1}^{p} a_\alpha X_\alpha\right) = \sum_{\alpha=1}^{p} a_\alpha E(X_\alpha) \tag{4.55}$$

(b) $$VAR\left(\sum_{\alpha=1}^{p} a_\alpha X_\alpha\right) = \sum_{\alpha=1}^{p} a_\alpha^2 \, VAR(X_\alpha) + \sum_{\alpha \neq \beta} \sum a_\alpha a_\beta \, COV(X_\alpha, X_\beta) \tag{4.56}$$

(c) Si les variables sont 2 à 2 indépendantes

$$VAR\left(\sum_{\alpha=1}^{p} a_\alpha X_\alpha\right) = \sum_{\alpha=1}^{p} a_\alpha^2 \, VAR(X_\alpha) \tag{4.57}$$

(d) Cas particulier

Soient $X_1, X_2, \ldots, X_p$, 2 à 2 indépendantes, identiquement distribuées, de moyenne μ et variance σ^2. Alors

$$E\left[\sum_{\alpha=1}^{p} \frac{1}{p} X_\alpha\right] = E(\bar{X}) = \mu \tag{4.58}$$

$$VAR(\bar{X}) = \sigma^2/p \tag{4.59}$$

(e) Si $g(X_1, X_2, \ldots, X_p)$ possède des dérivées secondes alors

$$E(g(X_1, X_2, \ldots, X_p)) \approx g(E(X_1), \ldots, E(X_\alpha)) \tag{4.60}$$

$$VAR(g(X_1, \ldots, X_p) \approx \sum_{\alpha=1}^{p} (\partial g/\partial x_\alpha)^2 \, VAR(X_\alpha) + \sum_{\alpha \neq \beta} \sum (\partial g/\partial x_\alpha)(\partial g/\partial x_\beta) \, COV(X_\alpha, X_\beta) \tag{4.61}$$

(formule dite de propagation des erreurs)

En particulier, si $p = 1$

$$E(g(X)) \approx g(E(X))$$

$$VAR(g(X)) \approx \left[\frac{dg(x)}{dx}\bigg|_{x=E(X)}\right]^2 VAR(X)$$

Exemple 4.21

Un réservoir d'eau illustré à la figure 4.3 est approvisionné de trois sources: deux ruisseaux A et B et les précipitations directes C. Chaque source dépend des pluies P dans le bassin entourant le réservoir. On a établi les équations

$$A = 0.2*P + 0.3$$

$$B = 0.15*P + 0.4$$

$$C = 0.03*P$$

exprimant A,B,C (en millions de gallons) en fonction de P. Trois sources contribuent à vider le réservoir: la consommation municipale (M), l'évaporation (EV) et l'irrigation (I). Des études hydrologiques et autres ont permis d'établir les caractéristiques suivantes pour les trois prochains mois:

Variable	Moyenne	Variance
$P(p_0)$	15.0	2.0
I(mg)	1.5	0.3
M(mg)	1.0	0.1
EV(mg)	2.5	0.4

On croit que les variables P, I, M et E sont 2 à 2 indépendantes.

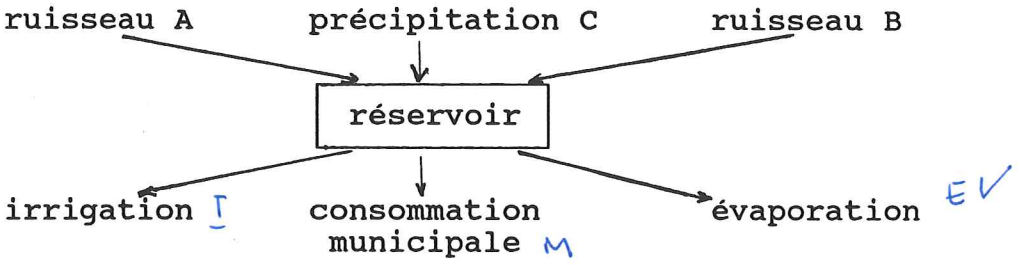


Figure 4.3: exemple 4.21

Questions

(a) Soit T la quantité d'eau entrant dans le réservoir durant les trois prochains mois. Calculez la moyenne et la variance de T .

(b) Au début de la période de trois mois, le réservoir contient 30 mg. Soit S le niveau du réservoir à la fin de la période de trois mois. Calculez la moyenne et la variance de S .

Solutions

(a) $T = A + B + C = (0.2*P + 0.3) + (0.15*P + 0.4) + 0.03*P$

$= 0.38*P + 0.7$

Donc $E(T) = 0.38*E(P) + 0.7 = 0.38*15 + 0.7 = 6.4$

$VAR(T) = (0.38)^2 \ VAR(P) = (0.38)^2*2 = 0.29$

(b) $S = 30 + T - (I + M + EV)$

Donc $E(S) = 30 + E(T) - E(I) - E(M) - E(EV)$

$= 30 + 6.4 - 1.5 - 1 - 2.5 = 31.4$

$VAR(S) = VAR(T) + VAR(I) + VAR(M) + VAR(EV)$

$= 0.29 + 0.3 + 0.1 + 0.4 = 1.09$

Exemple 4.22

Dans une étude de pollution par le bruit, le niveau du bruit au point C résulte de deux sources émettrices en A et B selon la figure 4.4. L'intensité du bruit en A et B est donnée dans le tableau ci-joint

Source	Moyenne	Coefficient de variation
I_A	1000	10%
I_B	2000	15%

De plus les deux sources sont positivement corrélées avec une corrélation de 0.5. L'intensité du bruit décroît avec la distance de la source émettrice selon l'équation

$$I(d) = I/(d+1)^2$$

où I est l'intensité générée à la source et I(d) est l'intensité à une distance d de la source. Soit I_c, l'intensité au point C .

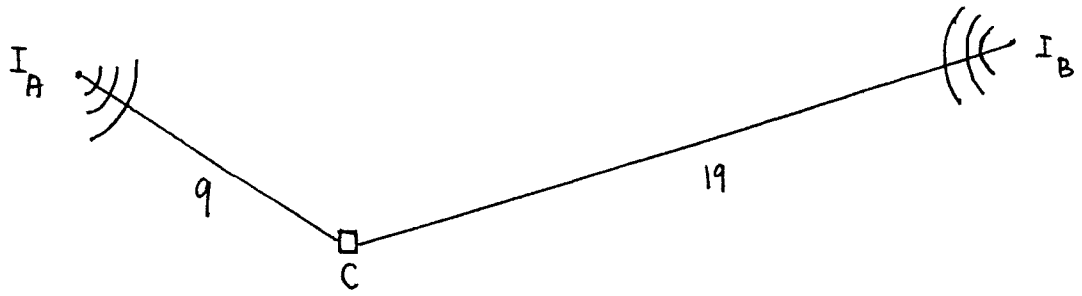


Figure 4.4: exemple 4.22

Questions

(a) Calculez la moyenne et la variance de I_c

(b) Une mesure de l'intensité souvent employée est le décibel D défini par

$$D = 40 \ln (2I)$$

Calculez la moyenne et l'écart-type de D_c .

Solutions

(a) $I_C = I_A/(9+1)^2 + I_B/(19+1)^2 = 0.01*I_A + 0.0025*I_B$

On a $ET(I_A) = E(I_A)*0.10 = 100$

$ET(I_B) = E(I_B)*0.15 = 300$

$COV(I_A,I_B) = CORR(I_A,I_B)*ET(I_A)*ET(I_B)$
$= 0.5*100*300 = 15000$

où ET(.) designe l'écart-type.

Alors $E(I_C) = 0.01*E(I_A) + 0.0025*E(I_B)$

$= 0.01*1000 + 0.0025*2000 = 15$

$VAR(I_C) = (0.01)^2 VAR(I_A) + (0.0025)^2 VAR(I_B)$

$+ 2(0.01)(0.0025)COV(I_A, I_B)$

$= (0.01)^2(100)^2 + (0.0025)^2(300)^2$

$+ 2(0.01)(0.0025)15000 = 2.3125$

(b) $D_C = 40 \ln (2*I_C)$. En utilisant les équations (4.60) et (4.61) on a:

$E(D_C) \approx 40 \ln (2E(I_C)) \approx 136.05$

$VAR(D_C) \approx (dD_C/dI_C)^2 VAR(I_C)$

$\approx (40 * 2/E(I_C))^2 VAR(I_C)$

$\approx (40 * 2/15)^2(37/16) = 65.78 = (8.11)^2$

$ET(D_C) = 8.11$

4.8 RAISONNEMENT STATISTIQUE

La démarche statistique est schématisée par:

- un phénomène montrant une certaine régularité statistique
- une description du phénomène par une (ou plusieurs) variable(s) aléatoire X dont la distribution (densité ou répartition) n'est pas complètement connue
- la possibilité d'observer de façon répétée la variable aléatoire du modèle
- l'utilisation de ces observations pour affiner sa connaissance de la distribution

Posons:

X : variable aléatoire associée au phénomène

F_X : fonction de répartition de X dépendant de certains paramètres inconnus

$(x_1,\ldots,x_n)$: n observations de X

$(X_1,\ldots,X_n)$: n variables aléatoires indépendantes et identiquement distribuées à X

$F_{X\alpha}(x) = F_X(x)$ tout x , $\alpha = 1,\ldots,n$

Alors

X : est la variable parente ou source (à toutes fins pratique infinie) qui génère les données; on dit aussi POPULATION

$(X_1,\ldots,X_n)$: est le vecteur aléatoire appelé ÉCHANTILLON ALÉATOIRE extrait d'une POPULATION X dont

$(x_1,\ldots,x_n)$: est une réalisation (ou échantillon ou observation); c'est UNE réalisation du vecteur $(X_1,\ldots,X_n)$

$\{(x_1,\ldots,x_n)\}$: est l'ensemble des valeurs prises par $(X_1,\ldots,X_n)$ ou ESPACE DES OBSERVATIONS (INDIVIDUS) et la dimension de cet espace est

n : TAILLE de l'échantillon

En général l'information à extraire des observations passe par une fonction

$$Y = g(X_1, \ldots, X_n)$$

appelée STATISTIQUE dont

$$g(x_1, x_2, \ldots, x_n)$$

est UNE réalisation. Les statistiques souvent utilisées sont:

$\bar{X} = \frac{1}{n} \sum_{\alpha=1}^{n} X_\alpha$: la moyenne empirique

$S_X^2 = \sum_{\alpha=1}^{n} \frac{(X_\alpha - \bar{X})^2}{n - 1}$: la variance empirique

$S_{XY} = \sum_{\alpha=1}^{n} \frac{(X_\alpha - \bar{X})\,(Y_\alpha - \bar{Y})}{n - 1}$: la covariance empirique

$r = S_{XY}/S_X S_Y$: la corrélation empirique

Beaucoup d'autres statistiques sont aussi proposées pour traiter les applications. Les concepts, critères et procédures de l'estimation statistique et des tests d'hypothèses reposent exclusivement sur le concept fondamental de DISTRIBUTION D'ÉCHANTILLONNAGE c'est-à-dire

> la loi de probabilité de $Y = g(X_1, \ldots, X_n)$ où $(X_1, \ldots, X_n)$ est un échantillon aléatoire (variables indépendantes et identiquement distribuées).

La détermination de la loi d'échantillonnage de Y dépend de la fonction g(.) et de la distribution de X .

$$(X_1, X_2, \ldots, X_n) \xrightarrow{\quad g \quad} Y$$

D'une manière générale, le problème n'est pas facile à résoudre, mais les réponses, dans plusieurs cas importants, seront fournies aux chapitres des distributions. D'autre part, il n'est pas toujours nécessaire de connaître exactement la loi d'échantillonnage et des approximations sont généralement suffisantes.

Nous allons illustrer, à l'aide d'un exemple simple, le raisonnement statistique et la terminologie introduite dans cette section.

Exemple 4.23

Soit la population X dont la distribution est définie par:

X	1	2	3	4	5	6
$p_X(x;\theta)$	$(1-\theta)/6$	1/6	1/6	1/6	1/6	$(1+\theta)/6$

La distribution dépend d'un paramètre inconnu θ compris entre -1 et +1

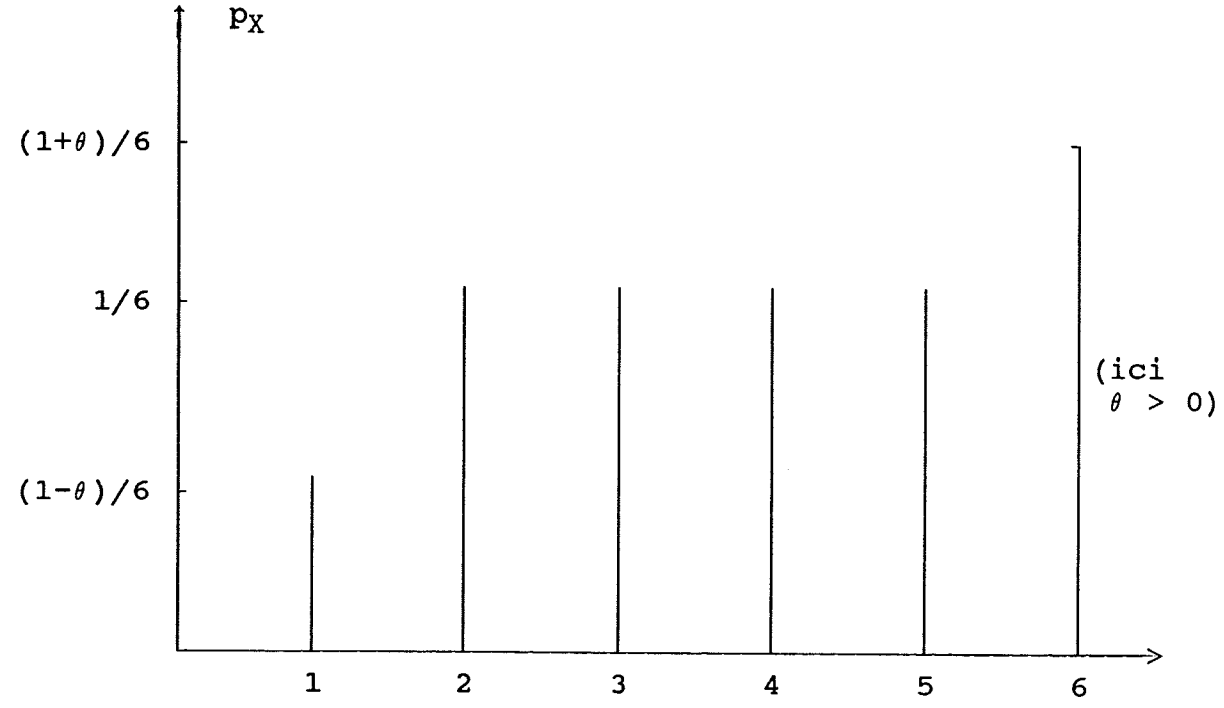


Figure 4.5: exemple 4.23

La variable X pourrait représenter, par exemple, le résultat sur un dé à jouer et on pourrait s'intéresser au problème de l'estimation de θ ou encore de décider si le dé est balancé ($\theta = 0$) ou non ($\theta \neq 0$). On calcule

$$E(X) = 3.5 + (5/6)\theta$$

$$VAR(X) = [105 - 25\theta^2]/36$$

Selon le type de problème envisagé (estimation ou test) on choisit une statistique

$$Y = g(X_1, \ldots, X_n)$$

basée sur un échantillon aléatoire de taille n .

Supposons que n = 2 et que la statistique choisie est

$$Y = (X_1 + X_2)/2 = \bar{X}$$

Quelle est la loi d'échantillonnage de Y ?

Solution

On a

$$p_{X\alpha}(x;\theta) = p_X(x;\theta) \qquad \alpha = 1,2$$

et puisque les variables X_1, X_2 sont indépendantes

$$p_{X_1,X_2}(x_1,x_2;\theta) = p_X(x_1;\theta)p_X(x_2;\theta) \qquad (2)$$

où (x_1,x_2) parcourt la liste de tous les 36 échantillons distincts

$$\{(x_1,x_2) : x_1, x_2 = 1, 2, \ldots, 6\}$$

On obtient

$$p_{X1,X2}(x_1,x_2;\theta) = \begin{cases} (1-\theta)^2/36 & x_1 = x_2 = 1 \\ (1-\theta)/36 & x_1 = 1,\ x_2 = 2,3,4,5 \\ & x_1 = 2,3,4,5,\ x_2 = 1 \\ (1-\theta)(1+\theta)/36 & x_1 = 1,\ x_2 = 6 \\ & x_1 = 6,\ x_2 = 1 \\ 1/36 & x_1, x_2 = 2, 3, 4, 5 \\ (1+\theta)/36 & x_1 = 6,\ x_2 = 2, 3, 4, 5 \\ & x_1 = 2, 3, 4, 5,\ x_2 = 6 \\ (1+\theta)^2/36 & x_1 = 6,\ x_2 = 6 \end{cases}$$

La loi de probabilité de $Y = \bar{X}$ est

$$p_{\bar{X}}(\bar{x};\theta) = \sum_{A(\bar{x})} p_{X1,X2}(x_1,x_2;\theta)$$

où $$A(\bar{x}) = \{(x_1,x_2) : (x_1+x_2)/2 = \bar{x}\}$$

et $$\bar{x} = 1,\ 3/2,\ 2,\ 5/2,\ \ldots,\ 11/2,\ 6$$

On peut identifier les éléments de $A(\bar{x})$ à l'aide du graphique de la figure 4.6.

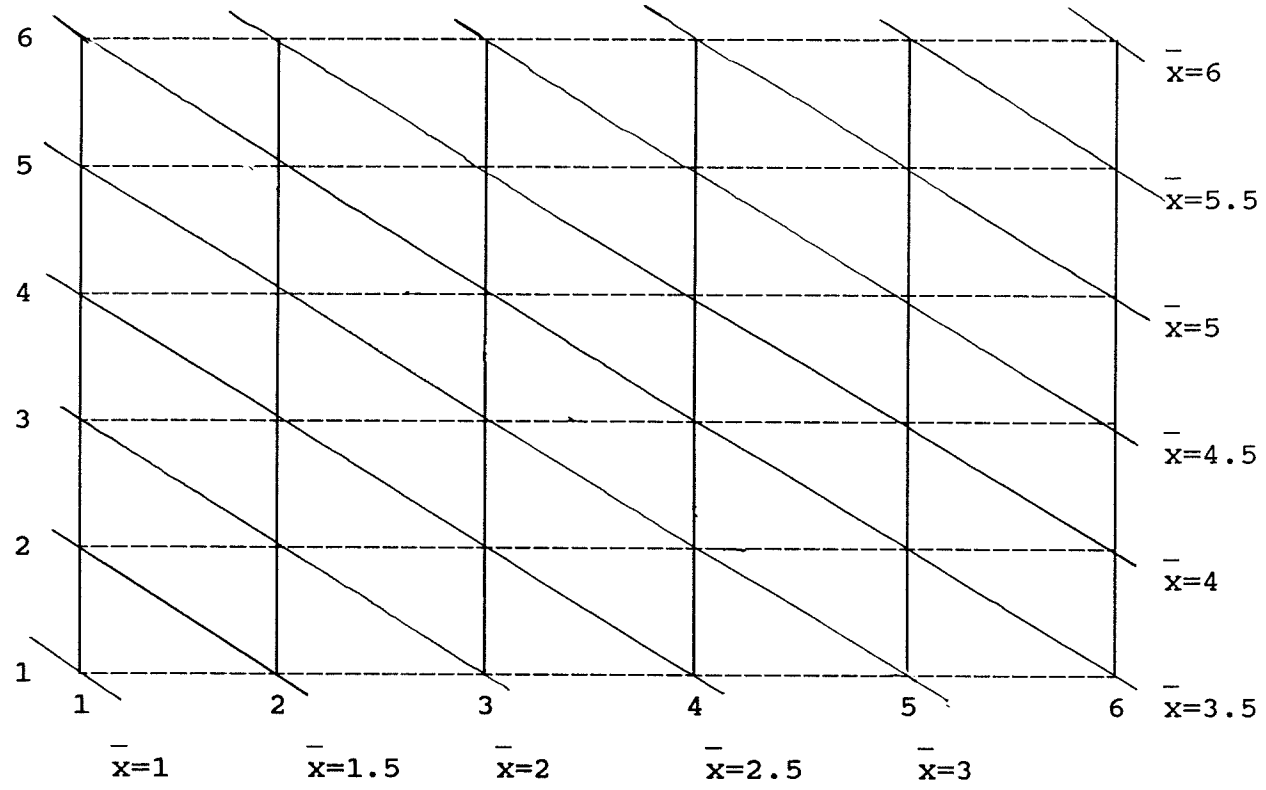


Figure 4.6: exemple 4.23 suite

Les valeurs de (x_1, x_2) telles que $\bar{x} = c$ sont situées sur des droites d'équation:

$$x_1 + x_2 = 2c$$

Par exemple

$$A(1) = \{(1,1)\}$$

$$A(1.5) = \{(1,2), (2,1)\}$$

et ainsi de suite. La distribution de $\bar{X}$, $p_{\bar{X}}(\bar{x};\theta)$ est définie par le tableau suivant où $a=2(1-\theta)$ et $b=2(1+\theta)$ et est représentée à la figure 4.7.

$\bar{X}$	1	1.5	2	2.5	3
$36\ p_{\bar{X}}(x;\theta)$	$(1-\theta)^2$	a	a+1	a+2	a+3

$\bar{X}$	3.5	4	4.5	5	5.5	6
$36\ p_{\bar{X}}(x;\theta)$	$2(1-\theta^2)+4$	b+3	b+2	b+1	b	$(1+\theta)^2$

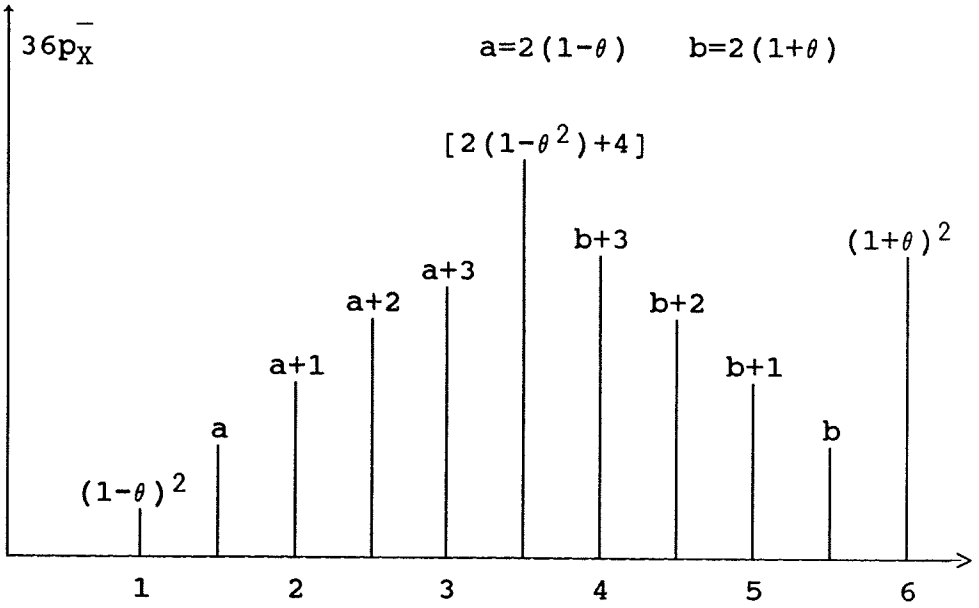


Figure 4.7: loi d'échantillonnage de $\bar{X}$ avec n = 2, exemple 4.23

On calcule

$$E(\bar{X}) = 3.5 + (5/6)\theta = E(X)$$

$$VAR(\bar{X}) = [105 - 25\theta^2]/72 = VAR(X)/2$$

On constate que X et $\bar{X}$ ont la même moyenne tandis que $\bar{X}$ a une variance deux fois plus petite que celle de X. Cela est un cas particulier des formules (4.58) et (4.59) avec p=2. Ces formules sont très importantes et seront utilisées souvent dans la suite.

Il nous a été possible d'obtenir la distribution d'échantillonnage car cet exemple est relativement simple puisque n = 2 et la fonction g(.) est linéaire. Le problème se complique pour des fonctions non-linéaires et pour des tailles échantillonnales n de plus en plus grandes. Ce chapitre nous a fourni quelques réponses et dans le prochain chapitre la distribution exacte ou approximative sera donnée pour certaines transformations. Parmi les transformations d'importance pour les applications on note les sommes, les sommes de carrés ainsi que les quotients de variables aléatoires indépendantes et identiquement distribuées.

4.9 EXERCICES

4.1 Une variable aléatoire possède une densité constante sur l'intervalle (-c,c) et sa variance est 1. Calculez la constante c, la moyenne, la médiane, la fonction de répartition et les percentiles suivants: 0.05, 0.10, 0.25, 0.95.

4.2 Le pourcentage A d'un certain additif dans un carburant détermine son prix de vente. Si A est inférieur à 70%, le carburant se vend 0.52/litre et s'il est supérieur à 70%, le carburant se vend 0.58/litre. Calculez le revenu moyen/litre en supposant que A est distribué uniformément.

4.3 Une variable X a une densité de probabilité $f_X(x)$ définie par:

$$f_X(x) = \begin{cases} c\,x & 0 \le x \le 2 \qquad c > 0 \\ c\,(4 - x) & 2 \le x \le 4 \\ 0 & \text{ailleurs} \end{cases}$$

Calculez la constante c, la moyenne, la variance, l'écart-type, la fonction de répartition et les p-ième percentiles: 0.05, 0.25, 0.50, 0.75, 0.99.

4.4 Une variable aléatoire X a une densité de probabilité $f_X(x)$ définie par:

$$f_X(x) = \begin{cases} c\,\exp(-x) & x \ge 0 \qquad c > 0 \\ 0 & x < 0 \end{cases}$$

Calculez la constante c, la moyenne, la variance, l'écart-type, la fonction de répartition, la médiane, le 90-ième percentile et le 99-ième percentile.

4.5 Un manufacturier d'appareils de télévision offre une garantie d'un an sur la lampe-écran. Il estime que le temps (ans) avant la première panne est une variable T dont la densité de probabilité $f_T(t)$ est définie par:

$$f_T(t) = \begin{cases} 1/4\,\exp\,(-t/4) & t \ge 0 \\ 0 & t < 0 \end{cases}$$

(a) Quel pourcentage des appareils seront réparés durant la période de garantie?

(b) Si une vente rapporte un profit de 200$ et que le coût de réparation est de 200$, quel est le profit moyen réalisé par le manufacturier?

4.6 Le temps (heures) avant la première panne d'une lampe-écran d'un appareil de télévision est une variable T dont la fonction de repartition $F_T(t)$ est définie par:

$$F_T(t) = \begin{cases} 1 - \exp(-ct) & t \geq 0 \qquad c > 0 \\ 0 & t < 0 \end{cases}$$

où c est une constante dépendant du manufacturier.

(a) Calculez la densité de probabilité de T .

(b) On sait que 50% des lampes-écrans tombent en panne en moins de 1500 heures. Quelle est la probabilité que la première panne arrive après 3000 heures?

4.7 Le moteur d'une voiture neuve est garanti pour 1 an. La durée de vie (ans) T du moteur est une variable exponentielle de moyenne 3 ans. Le profit réalisé par la vente de la voiture est de 1000$ et le coût de réparation est possiblement >250 pour que E(P) = 0.

(a) Quel est le profit moyen?

(b) Jusqu'à quelle limite pourrait-on étendre la garantie sans perdre de l'argent?

(c) Refaire les calculs (a) et (b) avec une moyenne de 2 ans et une moyenne de 4 ans.

4.8 Un manufacturier d'appareils de télévision offre une garantie de un an. Le temps d'utilisation avant la première panne est une variable exponentielle avec une moyenne de 20 000 heures. Il en coûte 300$ pour fabriquer l'appareil, 150$ pour le réparer et il est vendu 400$. Quel est le profit moyen du manufacturier si l'on suppose que les appareils sont en usage continu (e.g. dans les aérogares)?

4.9 Une molécule dans un gaz possède une vitesse V qui est une variable aléatoire avec densité

$$f_V(v) = \begin{cases} a\, v \exp(-v^2) & v \geq 0 \\ 0 & v < 0 \end{cases} \quad , \quad a > 0$$

(a) Déterminez la constante a

(b) Trouvez la fonction de répartition de V

(c) L'énergie cinétique E d'une molécule est donnée par $E = mV^2/2$. Calculez $P[E < 8m]$.

4.10 Un ingénieur doit régler une machine automatique qui produit r objets à l'heure. La proportion d'articles défectueux θ s'accroit avec r . Il y a un profit de \$1.00 pour chaque article non-défectueux produit et une perte de \$20.00 pour chaque article défectueux. On sait que la proportion de défectueux obéit à la loi

$$f(\theta) = \begin{cases} (0.001)r\, \theta^{0.001r-1} & 0 < \theta < 1 \\ 0 & \text{ailleurs} \end{cases}$$

(a) Si on fixe r et θ, montrez que le profit par heure est égal à $r(1 - 21\theta)$

(b) Si on fixe r, montrez que le profit moyen par heure est égal à

$$\frac{1000r - 20r^2}{1000 + r}$$

(c) Montrez que le profit par heure moyen maximum est atteint avec r = 25.

4.11 Un lot de 10 articles contient 3 articles défectueux. On tire (sans remise) les articles un à la fois et on examine à chaque tirage si l'article est défectueux ou non. Soit X la variable aléatoire représentant le nombre d'articles tirés afin d'obtenir un deuxième article défectueux; trouvez la fonction de masse de probabilité $p_X(x)$, sa moyenne et son écart-type.

4.12 Une boite de N articles contient D (D < N) articles défectueux. On tire les articles un à la fois et on vérifie si l'article est défectueux ou non. Trouvez la probabilité que la γ-ième ($1 \leq \gamma \leq D$) article défectueux soit trouvé au n-ième tirage ($1 \leq n \leq N$).

4.13 Une famille de densité de probabilité utilisée pour représenter la distribution des revenus, la taille des villes, la taille des entreprises, etc... s'appelle la loi de Pareto définie par:

$$f_X(x;k,\theta) = \begin{cases} 0 & x < \theta \qquad k > 0 \\ \dfrac{k\theta^k}{x^{k+1}} & x \geq \theta \end{cases}$$

(a) Déterminez une expression explicite (sans signe d'intégration) pour la fonction de répartition $F(x;k,\theta)$

(b) Calculez $P[2 < X < 3]$ si $k = 2$ et $\theta = 1$

(c) Déterminez une expression explicite pour le p-ième percentile x_p

(d) Calculez la médiane si $k = 2$ et $\theta = 1$

(e) Si $k > 1$, déterminez une expression pour la moyenne et calculez sa valeur pour $\theta = 1$ et $k = 2$

(f) Si $k > 2$, déterminez une expression pour l'écart-type.

4.14 Le temps d'attente (en minutes) pour être servi à un guichet est une variable aléatoire continue T ayant une fonction de densité $f_T(t)$:

$$f_T(t) = \begin{cases} 0 & t < 0 \\ 1/2 & 0 \leq t < 1 \\ \dfrac{3}{2t^4} & t \geq 1 \end{cases}$$

(a) Déterminez la fonction de répartition F_T

(b) Calculez $P[T > 2 \mid T > 1]$

(c) Calculez la moyenne et la médiane de T

(d) Calculez l'écart-type de T

4.15 Une variable X a une densité de probabilité $f_X(x)$ définie par:

$$f_X(x) = \begin{cases} \frac{3}{2} x^2 & -c < x < c \\ 0 & \text{ailleurs} \end{cases}$$

(a) Déterminez la constante c

(b) Calculez la fonction de répartition $F_X(x)$

(c) Calculez la variance

(d) Calculez le 90^e percentile de la distribution.

4.16 On prend au hasard un point à l'intérieur d'une sphère de rayon r. La probabilité que ce point appartienne à une région sphérique est proportionnelle au volume de cette région. Soit X la distance du point choisi au centre de la sphère.

(a) Déterminez la fonction de répartition de X.

(b) Trouvez la fonction de densité de X.

4.17 Une variable aléatoire X est distribuée selon la loi DOUBLE EXPONENTIELLE si sa densité $f_X(x;\alpha,\beta)$ s'écrit:

$$f_X(x;\alpha,\beta) = \begin{cases} \frac{1}{2\beta} \exp\left[\frac{\alpha-x}{\beta}\right] & x \geq \alpha \\ \frac{1}{2\beta} \exp\left[\frac{x-\alpha}{\beta}\right] & x \leq \alpha \end{cases}$$

où α et β sont deux paramètres tels que $-\infty<\alpha<\infty$, $\beta > 0$

(a) Déterminez la fonction de répartition $F_X(x;\alpha,\beta)$

(b) Calculez la moyenne et l'écart-type de X

(c) Déterminez une expression pour le p-ième percentile de X et calculez l'écart interquartile

(d) Calculez le coefficient d'aplatissement β_2 de X

4.18 Soit X une variable aléatoire continue telle que

$$P[\ X > x] = (\mu x + 1)\ \exp(-\mu x) \qquad x \geq 0, \quad \mu > 0$$

(a) Déterminez la fonction de répartition $F_X(x;\mu)$

(b) Déterminez la fonction de densité $f_X(x;\mu)$

(c) Calculez la moyenne de X

(d) Calculez l'écart-type de X

4.19 Soit $p_{X,Y}(x,y)$ une fonction de probabilité conjointe définie par le tableau suivant:

x \ y	0	1	2
-1	1/16	1/16	2/16
0	1/16	1/16	1/16
1	1/16	2/16	6/16

(a) Calculez $p_{X(x)}$, $P_{Y|x=0}(y)$, $P_{Y|x=1}(y)$, $P_{X|y=2}(x)$.

(b) Calculez $P[X = -1,\ 0 \leq Y < 2]$

(c) Les variables X et Y sont-elles indépendantes?

(d) Calculez le coefficient de corrélation.

4.20 Un vecteur aléatoire V = (X,Y) a pour fonction de masse de probabilité p_V

$$p_V(x,y) = \begin{cases} \dfrac{\theta^x}{x!}\ \exp(-\theta)\ \binom{x}{y}\ p^y\ (1-p)^{x-y} & \text{si} \quad \begin{matrix} x = 0,1,2,\ldots \\ y = 0,1,\ldots,x \end{matrix} \\ 0 & \text{ailleurs} \end{cases}$$

(a) Déterminez la loi de probabilité marginale de X et la loi de probabilité conditionnelle de Y étant donné X = x. Précisez $p_{Y|x=0}(y)$.

(b) Déterminez la loi marginale de Y.

(c) Calculez les moyennes et écart-types de X et Y.

(d) X et Y sont-elles indépendantes?

4.21 On désigne par X un nombre choisi au hasard sur l'intervalle [0,1] et par la suite un nombre Y est choisi au hasard sur l'intervalle [0,x] où x est la réalisation de X.

(a) Déterminez la densité marginale f_X et la densité conditionnelle $f_{Y|X=x}(y)$

(b) Calculez P [X+Y > 1].

(c) Déterminez $f_Y(y)$.

(d) Calculez les moyennes, les écart-types et le coefficient de corrélation entre X et Y.

4.22 Pour chacune des densités conjointes $f_{X,Y}(x,y)$ calculez la constante c , les densités marginales, les moyennes, les variances et le coefficient de corrélation. Est-ce que les variables sont indépendantes?

(a) $$f_{X,Y}(x,y) = \begin{cases} c/1000 & 0 \le x \le 1000 \\ & 0 \le y \le 10 \\ 0 & \text{ailleurs} \end{cases}$$

(b) $$f_{X,Y}(x,y) = \begin{cases} c & 0 \le x \le 1 \ , 0 \le y \le 1, \\ & x \le y \\ 0 & \text{ailleurs} \end{cases}$$

(c) $$f_{X,Y}(x,y) = \begin{cases} c(5-x-y/2) & 0 \le x \le 2 \\ & 0 \le y \le 2 \\ 0 & \text{ailleurs} \end{cases}$$

(d) $$f_{X,Y}(x,y) = \begin{cases} cxy\ \exp(-x^2-y^2) & x,y \ge 0 \\ 0 & \text{ailleurs} \end{cases}$$

4.23 Deux variables aléatoires ont une densité conjointe de probabilité

$$f_{X,Y}(x,y) = \begin{cases} k(x^2 + y^2) & 0 \le x \le a \ , \ 0 \le y \le b \\ 0 & \text{ailleurs} \end{cases}$$

(a) Calculez la constante k

(b) Calculez P[0 < X < a/2 , 0 < Y < b/2]

(c) Calculez les densités marginales

(d) Les variables sont-elles indépendantes ?

4.24 Soit X_1, X_2, X_3 trois variables aléatoires indépendantes de moyenne μ et de variance σ^2 et

$$Y = X_1 + X_2$$

$$Z = aX_2 + bX_3$$

Calculez le coefficient de corrélation entre Y et Z.

4.25 Soit X une variable dénotant l'heure de la journée où une marchandise est envoyée et Y l'heure de la journée où la marchandise est reçue. La densité conjointe de (X,Y) est définie par

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{288} & 0 \le x < y \le 24 \\ 0 & \text{ailleurs} \end{cases}$$

(a) Calculez la densité marginale de X et celle de Y.

(b) Calculez les densités conditionnelles $X|Y=y$ et $Y|X=x$.

(c) Calculez la probabilité que la réception de la marchandise ait lieu au plus tard 6 heures après son envoi.

4.26 On note par X la pression du pneu avant droit et Y la pression du pneu avant gauche d'une voiture. On suppose que la densité conjointe de (X,Y) est définie par:

$$f_{x,y}(x,y) = \begin{cases} c(x+y) & 20 \le x \le 30, \quad 20 \le y \le 30 \\ 0 & \text{ailleurs} \end{cases}$$

où c est une constante.

(a) Calculez la constante c

(b) Quelle est la probabilité que la pression du pneu avant droit soit inférieure à la pression du pneu avant gauche.

(c) Les variables sont-elles indépendantes?

4.27 La densité conjointe $f_{X,Y}$ de deux variables X et Y est définie par

$$f_{X,Y}(x,y) = \begin{cases} 24xy & 0 \le x \le 1, \quad 0 \le y \le 1, \quad 0 \le x+y \le 1 \\ 0 & \text{autrement} \end{cases}$$

(a) Calculez $P[X + Y \le 0.5]$

(b) Calculez la densité marginale de X

(c) Les variables sont-elles indépendantes?

4.28 Soit X une variable aléatoire discrète ayant la distribution suivante:

X	1	2	3
$P_X(x)$	0.2	0.5	0.3

(a) Calculez la moyenne μ et la variance σ^2 de X.

(b) Donnez la liste de tous les échantillons distincts de taille n = 3 (il y en a 27).

(c) Calculez la distribution d'échantillonnage de

$$\bar{X} = 1/3 \ (X_1 + X_2 + X_3)$$

(d) Calculez la moyenne et la variance de $\bar{X}$

(e) Vérifiez que $VAR(\bar{X}) = VAR(X) / 3$

4.29 Dans une banque, un caissier électronique permet de retirer des billets de 50 $ ou 100 $ à l'aide d'une carte magnétique. Il se peut aussi que le client ne puisse retirer d'argent si le compte n'est pas approvisionné ou si le client a fait une erreur de manipulation. Le nombre de clients X utilisant l'appareil dans un intervalle de 5 minutes est une variable aléatoire dont la masse de probabilité $p_X(x)$ est

x	0	1	2
$p_X(x)$	0.30	0.50	0.20

Le montant total Y retiré en 5 minutes est une variable dont la masse de probabilité conditionnelle est:

$$p_{Y|X=1}(y) = \begin{cases} 0.1 & \text{si } y=0 \\ 0.7 & \text{si } y=1 \\ 0.2 & \text{si } y=2 \end{cases}$$

(a) Montrez que

y	0	50	100	150	200
$P_{Y\|X=0}(y)$	1	0	0	0	0
$P_{Y\|X=2}(y)$	0.01	0.14	0.53	0.28	0.04

(b) Les variables aléatoires X et Y sont-elles indépendantes? Justifiez votre réponse.

(c) Calculez la probabilité P[X=1,Y=100].

(d) Calculez la probabilité P[Y=0].

(e) Calculez le nombre moyen de clients utilisant l'appareil en une heure.

4.30 Deux signaux X et Y indépendants sont émis selon une distribution uniforme durant un intervalle de temps T (fixe). Calculez la probabilité que la réception soit brouillée si un brouillage apparaît dès que la différence de temps entre les deux signaux est inférieure ou égale à b ou $0 < b < T$.

4.31 Soit la densité conjointe de probabilité:

$$f_{X,Y}(x,y) = c\,xy\,\exp\left[-\frac{(x+y)}{\lambda}\right] \quad x,y \geq 0$$

Montrez que:

(a) $c = \lambda^{-4}$

(b) $F_{X,Y}(x,y) = \left[1 - e^{-x/\lambda} - \frac{x}{\lambda} e^{-x/\lambda}\right]\left[1 - e^{-y/\lambda} - \frac{y}{\lambda} e^{-y/\lambda}\right]$

(c) $P[X+Y \geq \lambda] = \frac{8}{3e} = 0.98$

4.32 La densité de probabilité conjointe de deux variables (X,Y) est définie par:

$$f_{X,Y}(x,y) = \begin{cases} \frac{3}{4} & -1 \le x \le 1,\ x^2 \le y \le 1 \\ 0 & \text{ailleurs} \end{cases}$$

(a) Calculez

(i) les densités marginales de X et Y.

(ii) le coefficient de corrélation entre X et Y.

(b) Les variables X et Y sont-elles indépendantes? Justifiez.

4.33 On considère l'expérience de jeter deux tétraèdres distincts dont les faces sont numérotées 1, 2, 3, 4. Soit X_1 la variable dénotant le numéro de la face sur laquelle repose le premier tétraèdre et X_2 la variable dénotant le numéro de la face sur laquelle repose le second tétraèdre.

(a) Calculez la masse de probabilité conjointe de (X_1,X_2).

(b) Calculez la masse de probabilité conjointe de (X_1,Y) ou $Y = MAX(X_1,X_2)$.

(c) Calculez les masses de probabilités marginales de X_1 et Y.

(d) Les variables X_1 et Y sont-elles indépendantes?

4.34 Deux procédés de fabrication indépendants produisent des cylindres évidés et des pistons pour un assemblage. Le diamètre extérieur des pistons est représenté par une variable X distribuée uniformément sur l'intervalle [98.5, 100.5] tandis que le diamètre intérieur des cylindres est représenté par une variable Y distribuée uniformément sur l'intervalle [99, 101]. Calculez la probabilité d'effectuer l'insertion du piston dans le cylindre.

4.35 Une variable continue X suit une distribution LOGISTIQUE de paramètres de localisation α et d'échelle β a pour fonction de répartition $F_X(x;\alpha,\beta)$ définie par:

$$F_X(x;\alpha,\beta) = 1/(1+\exp\ (\ -\ \left[\frac{x-\alpha}{\beta}\right]\ ))$$

$$-\infty<x<\infty \quad , \quad -\infty<\alpha<\infty \quad , \quad \beta>0$$

(a) Déterminez la densité de probabilité $f_X(x;\alpha,\beta)$

(b) Montrez que:

$$f_X(x;\alpha,\beta) = \frac{1}{\beta}\ F_X(x;\alpha,\beta)\ (1\ -\ F_X(x;\alpha,\beta))$$

(c) Montrez que la variable $Y=(x-\alpha)/\beta$ est distribuée selon la loi logistique de paramètre de localisation 0 et de paramètre d'échelle 1.

(d) On montre que $E(Y)=0$ et $ET(Y)=\pi/\sqrt{3}$. Déterminez $E(X)$ et $ET(X)$.

(e) Montrez que le p-ième percentile $(0<p<1)$ x_p de X est:

$$x_p = \alpha + \beta\ \ell n\ \left[\frac{p}{1-p}\right]$$

4.36 La quantité de chaleur Q dégagée par un conducteur de résistance R (ohn) traversé par un courant I (amp) durant un temps T (minutes) est:

$$Q = 0.24\ I^2RT$$

Les variables I,R,T sont indépendantes 2 à 2 et leurs moyennes et écart-types sont donnés dans le tableau:

variable	moyenne	écart-type
I	10	0.1
R	30	0.2
T	10	0.5

Calculez la moyenne et l'écart-type de Q.

4.37 <u>Inégalité de Tchebycheff et le rapport signal-bruit</u>

Une inégalité très célèbre due à Tchebycheff s'énonce comme suit:

$$P[\ |X-\mu| \leq \epsilon\ ] \geq 1\ -\ \frac{\sigma^2}{\epsilon^2}$$

pour toute variable aléatoire X de moyenne $\mu=E(X)$ et d'écart-type $ET(X)=\sigma$. En particulier:

$$P[\ |X-\mu| \leq 4.5\sigma\ ] \geq 0.95$$

$$P[\ |X-\mu| \leq 10\sigma\ ] \geq 0.99$$

(a) Démontrez que l'inégalité de Tchebycheff peut se formuler ainsi:

$$P\left[\ \left|\frac{X-\mu}{\mu}\right| \leq \delta\ \right] \geq 1 - \frac{1}{\delta^2}\,\frac{\sigma^2}{\mu^2}$$

pour toute variable X de moyenne $\mu \neq 0$.

(b) Déduire de l'inégalité précédente que l'erreur relative de l'estimation de μ par X sera petite avec une probabilité élevée si le quotient $|\mu|/\sigma$ est grand. Le quotient $|\mu|/\sigma$ est appelé le rapport signal-bruit et sa réciproque $\sigma/|\mu|$ est appelée le coefficient de variation.

(c) A partir de quelles valeurs du rapport signal-bruit peut-on affirmer que:

$$P\left[\ \left|\frac{X-\mu}{\mu}\right| < \delta\ \right] = 1-\alpha$$

pour $\alpha = 0.10,\ 0.05,\ 0.01$ et $\delta = 0.1,\ 0.5$.

(d) Calculez le rapport signal-bruit pour les distributions suivantes:

(i) uniforme (constante) sur l'intervalle [a,b].

(ii) exponentielle de paramètre λ.

4.10 RÉPONSES EXERCICES

4.1 (a) $c = \sqrt{3}$, (b) $E(X) = 0$

(c) $F_X(x) = 0$ si $x < -\sqrt{3}$

$= (x+\sqrt{3})/2\sqrt{3}$ si $-\sqrt{3} \leq x \leq \sqrt{3}$

$= 1$ si $x \geq \sqrt{3}$

(d) -1.5588, -1.3856, -0.866, 1.5588

4.2 0.538 $

4.3 (a) c=0,25 (b) E(X)=2 (c) $\sigma 2=2/3$

(d) $F_X(x) = 0$ si $x \leq 0$

$= x^2/8$ si $0 \leq x \leq 2$

$= 1 - \frac{1}{8}(4-x)^2$ si $2 \leq x \leq 4$

$= 1$ si $x \geq 4$

(e)

p	0.05	0.25	0.50	0.75	0.99
x_p	0.14	1.41	2.00	2.59	3.71

4.4 (a) c=1 (b) 1 (c) 1 (d) $-\ln(1-p)$

(e)

p	0.50	0.90	0.99
x_p	0.69	2.30	4.60

4.5 (a) 0.2212 (b) 155,76 $

4.6 (a) ce^{-ct} (b) c=0.00046 (c) 0.2516

4.7 (a) 929,13 $ (b) à l'infini (c) 901.63 $, 944.70 $

4.8 46.80 $

4.9 (a) a=2 (b) $1-\exp(-v^2)$ (c) $1-e^{-16}$

4.11 $p_X(x) = \frac{1}{120}(x-1)(10-x)$ $x = 2,3,\ldots,9$

$E(X) = 5.5$, $ET(X) = 6.35$

4.12 $$p_X(x) = \frac{\binom{N_1}{\gamma-1}\binom{N-N_1}{x-\gamma}}{\binom{N}{x-1}} \quad \frac{N_1-\gamma+1}{N-x+1}$$

4.13 (a) $F_X(x;k,\theta) = 1-(\theta/x)^k \qquad x \geq \theta$

(b) 5/36

(c) $x_p = \theta(1-p)^{-1/k}$

(d) $\sqrt{2}$

(e) $E(X) = \binom{k}{k-1}\theta \qquad k > 1$

(f) $\sigma = \frac{\sqrt{k}}{(k-1)\sqrt{k-2}}\theta \qquad k > 2$

4.14 (a) $F_T(t) = 0 \qquad t<0$

$= \frac{1}{2}t \qquad 0\leq t<1$

$= 1 - \frac{1}{2t^3} \qquad t\geq 1$

(b) 1/8 (c) 1 (d) $\sqrt{\frac{2}{3}}$

4.15 (a) 1

(b) $$F_X(x) = \begin{cases} 0 & x < -1 \\ \frac{x^3+1}{2} & -1 \leq x \leq 1 \\ 1 & x \geq 1 \end{cases}$$

(c) $E(X) = 0$, $VAR(X) = 3/5$ (d) 0.928

4.16 (i) $F_X(x) = (x/r)^3 \qquad 0 \leq x \leq r$

(ii) $f_X(x) = \frac{3}{r^3}x^2 \qquad 0 \leq x \leq r$

4.17 (a) $F_X(x) = \frac{1}{2} \exp\left[- \frac{(\alpha - x)}{\beta} \right]$ si $x \leq \alpha$

$= 1 - \frac{1}{2} \exp\left[- \frac{(x-\alpha)}{\beta} \right]$ si $x \geq \alpha$

(b) $E(X) = \alpha$, $ET(X) = \beta\sqrt{2}$

(c) $x_p = \alpha + \beta \ln(2p)$ si $0 \leq p \leq 0.5$

$= \alpha - \beta \ln(2(1-p))$ si $0.5 \leq p \leq 1$

$IQ = x_{0.75} - x_{0.25} = 1.386\beta$

(d) 3

4.18 (a) $F_X(x;\mu) = 1 - e^{-\mu x}(1+\mu x)$ $x \geq 0$

(b) $f_X(x;\mu) = \mu^2 x\, e^{-\mu x}$

(c) $E(X) = 2/\mu$

(d) $ET(X) = \sqrt{2/\mu}$

4.19 (a)
$p_X(x) = 4/16$ si $x = -1$
$= 3/16$ si $x = 0$
$= 9/16$ si $x = 1$

$p_{Y|X=0}(y) = 1/3$ si $y = 0,1,2$

$p_{Y|X=1}(y)$
$= 1/9$ si $y = 0$
$= 2/9$ si $y = 1$
$= 6/9$ si $y = 2$

$p_{Y|X=2}(x)$
$= 2/9$ si $x = -1$
$= 1/9$ si $x = 0$
$= 6/9$ si $x = 1$

(b) 1/8 (c) non (d) 0.201

4.20 (a) $p_X(x) = \theta^x e^{-\theta}/x!$ $x=0,1,2,\ldots$

(b) $p_Y(y) = (p\theta)^y e^{-\theta p}/y!$ $y=0,1,2,\ldots$

$p_{Y|X=x}(y) = \binom{x}{y} p^y (1-p)^{x-y}$ $y=0,1,2,\ldots,x$

(c) $E(X) = \theta$ $E(Y) = \theta p$
$VAR(X) = \theta$ $VAR(Y) = \theta p$

(d) Non

4.21 (a) $f_X(x) = 1 \qquad 0 \le x \le 1$

$f_{Y|X=x}(y) = \frac{1}{x} \qquad 0 \le y < x$

(b) 0.3069

(c) $f_Y(y) = -\ln y \qquad 0 < y < 1$

(d) $E(X) = \frac{1}{2} \qquad ET(X) = 1/\sqrt{12}$

$E(Y) = \frac{1}{4} \qquad ET(Y) = \sqrt{7/12}$

$\rho = 0.65$

4.22 (a) $c = 1/10$, $f_X(x) = 1/1000 \quad 0 < x < 1000$
$= 0 \quad$ ailleurs

$f_Y(y) = 1/10 \quad 0 < y < 10$
$= 0 \quad$ ailleurs

$E(X) = 500 \qquad VAR(X) = 8333.33$

$E(Y) = 5 \qquad VAR(Y) = 8.33$

$\rho = 0 \qquad$ variables indépendantes

(b) $c = 2 \qquad f_X(x) = 2(1-x) \quad 0 \le x \le 1$
$= 0 \quad$ ailleurs

$f_Y(y) = 2y \quad 0 \le y \le 1$
$= 0 \quad$ ailleurs

$E(X) = 1/3 \qquad VAR(X) = 1/18$

$E(Y) = 2/3 \qquad VAR(Y) = 1/18$

$\rho = 0.50 \qquad$ variables dépendantes

(c) voir réponses à l'exemple 4.18

(d) $c = 4 \qquad f_X(x) = 2x\, e^{-x2} \quad x \ge 0$
$= 0 \quad$ ailleurs

$f_Y(y) = 2y\, e^{-y2} \quad y \ge 0$
$= 0 \quad$ ailleurs

$E(X) = E(Y) = \sqrt{\pi/2}$

$VAR(X) = VAR(Y) = (4-\pi)/4$

$\rho = 0$ variables indépendantes

4.23 (a) $h = 3/ab(a^2+b^2)$

(b) 1/16

(c) $f_X(x) = \frac{kb}{3}(b^3+3x^2)$ $0 \leq x \leq a$

$= 0$ ailleurs

$f_Y(y) = \frac{ka}{3}(a^2+3y^2)$ $0 \leq y \leq b$

(d) non

4.24 $\rho = a/\sqrt{2(a^2+b^2)}$

4.25 (a) $f_X(x) = (24-x)/288$ $0 \leq x \leq 24$
$= 0$ ailleurs

$f_Y(y) = y/288$ $0 \leq y \leq 24$
$= 0$ ailleurs

(b) $f_{X|Y=y}(x) = 1/y$ $0 \leq x \leq y$
$= 0$ ailleurs

$f_{Y|X=x}(y) = 1/24-x$ $x \leq y \leq 24$
$= 0$ ailleurs

(c) 0.4375

4.26 (a) $c = 1/5000$ b) 0.50 (c) non

4.27 (a) 1/16

(b) $f_X(x) = 12x(1-x)^2$ $0 \leq x \leq 1$
$= 0$ ailleurs

(c) non

4.28 (a) $E(X) = 2.1$ $VAR(X) = 0.49$

(d) $E(\bar{X}) = 2.1$ $VAR(\bar{X}) = 0.16333$

4.29 (b) non (c) 0.1 (d) 0.352 (e) 10.8

4.30 $1 - \left[1 - \frac{b}{t}\right]^2$

4.31 (a) i $f_X(x) = \frac{3}{4}(1-x^2)$ $-1 \le x \le 1$

$= 0$ ailleurs

$f_Y(y) = \frac{3}{2}\sqrt{y}$ $0 \le y \le 1$

$= 0$ ailleurs

(a) ii 0

(b) non

4.33 (a) $p_{X1,X2}(x_1,x_2) = \frac{1}{16}$ $x_1 = 1,2,3,4$, $x_2 = 1,2,3,4$

(b)

y \ x_1	1	2	3	4
1	1/16	0	0	0
2	1/16	2/16	0	0
3	1/16	1/16	3/16	0
4	1/16	1/16	1/16	4/16

(c) $p_{X1}(x_1) = \frac{1}{4}$ $x = 1,2,3,4$

$p_Y(y)$ $= 1/16$ $y = 1$
$= 3/16$ $y = 2$
$= 5/16$ $y = 3$
$= 7/16$ $y = 4$

(d) non

4.34 0.8125

4.36 7 200 , 6.997

4.37 (c)

δ \ α	0.10	0.05	0.01
0.1	31.6	44.7	100
0.5	6.3	8.9	20

(d) 0.58 (b-a/a+b) , 1

CHAPITRE 5

DISTRIBUTIONS DISCRÈTES

5.0 SOMMAIRE

Dans ce chapitre nous présentons plusieurs familles de distributions discrètes d'un usage utile pour les applications. Ces distributions sont exposées à partir d'une fonction de masse contenant un ou plusieurs paramètres; les principales caractéristiques de ces distributions sont obtenues.

5.1 DISTRIBUTION HYPERGÉOMÉTRIQUE

La distribution hypergéométrique intervient lorsque l'on échantillonne une population finie dont les éléments appartiennent à deux classes exclusives et exhaustives, par exemple un lot dont les articles sont classés défectueux ou non-défectueux. La distribution hypergéométrique est à la base de tous les plans d'échantillonnage et de contrôle de qualité en cours de réception.

Considérons un lot de N articles dont $D \leq N$ sont défectueux. Un échantillon sans remise de taille n est obtenu et on note par X le nombre d'articles défectueux dans l'échantillon. La masse de probabilité $p_X(x;N,D,n)$ est définie par

$$p_X(x;N,D,n) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} \qquad \text{où } k_1 \leq x \leq k_2 \qquad (5.1)$$

$$k_1 = \text{Max } (0,\ n - N + D) \qquad k_2 = \text{Min } (n,D)$$

Il y a $\binom{N}{n}$ échantillons de taille n dont $\binom{D}{x}\binom{N-D}{n-x}$ contiennent x défectueux. Si on suppose que tous les échantillons sont également probables, on obtient l'équation (5.1).

On montre que

$$\begin{aligned} E(X) &= n\theta \\ Var(X) &= n\theta(1 - \theta)\, N(1 - f)/(N - 1) \\ \beta_1 &= n(1 - 2\,)(1 - 2f)/(N - 2)\sqrt{Var(X)} \\ \beta_2 &= 3[(N - 1)(N + 6)]/(N - 2)(N - 3) \\ &\quad + \alpha(N - 1)N(N + 1)/(N - n)(N - 2)(N - 3) \end{aligned} \qquad (5.2)$$

où $\theta = D/N$ = fraction défectueux dans le lot
$f = n/N$ = fraction échantillonnale
$\alpha = [1 - 6N(\theta(1 - 0) + f)(1 - f)/(N + 1)]/n\theta(1 - \theta)$

On notera la fonction de répartition de la distribution par

$$PROBHYPR(N,D,n,x = \sum_{k=k_1}^{x} \binom{D}{k} \binom{N - D}{n - k} \Big/ \binom{N}{n} \qquad (5.3)$$

ce qui est aussi la notation adoptée par SAS.

Exemple 5.1: Un lot de N=40 articles est soumis à une inspection partielle à partir d'un échantillon de n=10 articles choisis au hasard. Le lot est accepté si l'échantillon contient au plus trois articles défectueux et il est retourné au producteur autrement. Calculez la probabilité

(a) de retourner le lot au producteur si le nombre d'articles défectueux dans le lot est de 8.

(b) d'accepter le lot si le nombre d'articles défectueux dans le lot est de 20.

Solution: Le nombre X d'articles défectueux dans l'échantillon est distribué selon une loi hypergéométrique de paramètres N=40, D=8 ou 20 selon le cas, n=10.

$$P[\text{rejeter lot}] = P(X \geq 4] = \sum_{k=4}^{10} \binom{8}{k} \binom{32}{10-k} \Big/ \binom{40}{10}$$
$$= 0.089$$
$$P[\text{accepter lot}] = P(X \leq 3] = \sum_{k=0}^{3} \binom{20}{k} \binom{20}{10-k} \Big/ \binom{40}{10}$$
$$= 0.136$$

APPLICATION: Plan d'échantillonnage simple

Un lot de N articles contient un nombre $D \leq N$ d'articles défectueux. Un plan d'échantillonnage simple consiste à prélever un échantillon de n articles et de rejeter le lot si l'échantillon contient plus de c articles défectueux. On précise le plan en spécifiant deux points sur la courbe caractéristique définie comme la probabilité d'accepter le lot en fonction de $\theta = D/N$ la proportion de défectueux dans le lot. Les points choisis sont $(\theta_1, 1 - \alpha)$ et (θ_2, β) où $\theta_1 < \theta_2$, $0 < \alpha, \beta < 1$, $0 < \beta < 1-\alpha$

α = risque du producteur
= probabilité de rejeter un lot de qualité acceptable θ_1
β = risque du consommateur
= probabilité d'accepter un lot de qualité rejetable θ_2

Déterminer un plan consiste à trouver les valeurs de n et c en fonction de $(\theta_1, 1 - \alpha)$ et (θ_2, β). Le nombre d'articles défectueux X dans l'échantillon de taille n suit une distribution hypergéométrique de paramètres (N, D, n). Le lot est accepté si $X \leq c$ et la probabilité correspondante est

$$P[X \leq c] = \sum_{x=k_1}^{c} \binom{D}{x} \binom{N-D}{n-x} \Big/ \binom{N}{n}$$

$$= \text{PROBHYPR}(N,D,n,c) \qquad (5.4)$$

Le plan doit satisfaire les deux équations suivantes:

$$\text{PROBHYPR}(N, D = \theta_1 N, n, c) = 1 - \alpha$$
$$\text{PROBHYPR}(N, D = \theta_2 N, n, c) = \beta \qquad (5.5)$$

où $\theta_1 < \theta_2$ et $0 < \alpha < 1$, $0 < \beta < 1-\alpha$

Le système (5.5) contient deux inconnues (n, c). En général, une solution exacte n'existe pas puisque la fonction est discontinue. On se satisfait d'une solution approchée où les égalités sont remplacées par des quasi égalités. D'autre part, la résolution des équations par une méthode analytique présente des difficultés puisque la fonction hypergéométrique est malaisée à manipuler. Une méthode de résolution consiste à remplacer la fonction hypergéométrique par des approximations obtenues de la distribution binomiale, Poisson ou gaussienne.

Il existe une autre méthode de résolution des équations basée sur une double itération de n et c en commençant avec n = 1 et c = 0. Cette méthode a l'avantage de pouvoir se programmer et un programme utilisant le système SAS sera présenté à la section 5.6.

5.2 DISTRIBUTION BINOMIALE

Il existe de nombreuses expériences aléatoires où les résultats peuvent être classés en deux catégories exclusives et exhaustives. Par exemple:

- les articles produits dans une ligne d'assemblage et soumis à un contrôle de qualité peuvent être classés défectueux ou non
- les eaux usées à la sortie d'une usine peuvent être ou ne pas être acceptables selon des normes de qualité
- dans une analyse de sols par carottage, chaque essai peut résulter dans la rencontre de grosses pierres ou pas
- le débit annuel de pointe (maximal) d'une rivière peut ou ne pas excéder un débit fixé.

Les conditions d'opération de l'expérience constituent des ESSAIS DE BERNOULLI si

(a) Chaque essai n'a que deux résultats possibles:

- réalisation d'un certain événement E
- non-réalisation de l'événement E

(b) La probabilité de réalisation de E est constante pour tous les essais

$$\theta = P(E) \qquad 0 < \theta < 1$$

(c) Les essais sont indépendants.

La variable d'intérêt dans ces circonstances est

X = nombre de fois où E s'est réalisé dans la suite des n essais (échantillon de taille n)

Les valeurs de X sont {0,1,2,...,n} et les hypothèses (a)-(b)-(c) conduisent à la fonction de masse de probabilité pour X:

$$P[X = x] = p_X(x;\theta,n) = \binom{n}{x} \theta^x (1 - \theta)^{n - x} \qquad (5.6)$$

où $0 < \theta < 1$ et $x = 0,1,2,...,n$

La distribution porte le nom de BINOMIALE puisque les termes de la loi sont ceux du développement du binôme de Newton (1642 - 1727).

La distribution fut obtenue par Jacques Bernoulli (1654-1705) dans son traité 'Ars Conjecdandi' publié en 1713 mais les coefficients binomiaux étaient connus de Blaise Pascal (1623 - 1662). Les paramètres sont n = taille de l'échantillon et θ, la probabilité de réalisation de E. Le paramètre θ est généralement inconnu et doit être estimé avec des données.

La justification de l'équation (5.1) est la suivante:

si X = x alors $\theta^x(1 - \theta)^{n - x}$

est la probabilité de x réalisations de E et (n - x) réalisations de E' dans une séquence particulière de E et E'. Cette probabilité est la même quelque soit l'ordre dans laquelle les x réalisations de E se sont faites. Le nombre de telles séquences est $\binom{n}{x}$ correspondant au nombre de façons x possibles de placer les x réalisations de E parmi les n essais.

Cette distribution de probabilité est une distribution d'échantillonnage telle que définie au chapitre 4. En effet, notons par X_α la variable aléatoire définie par

$$X_\alpha = \begin{cases} 1 & \text{si E s'est réalisée au x-ième essai} \\ 0 & \text{si E ne s'est pas réalisé au x-ième essai} \end{cases}$$

Alors

X_α	0	1
$Px_\alpha\ (x_\alpha;\theta)$	$1 - \theta$	θ

pour $\alpha = 1,2,\ldots,n$

Les variables X_α sont indépendantes, identiquement distribuées et la variable binomiale X est définie par

$$X = \sum_{\alpha=1}^{n} X_\alpha$$

Il est facile de montrer que

$$\begin{aligned} E(X_\alpha) &= \theta \\ VAR(X_\alpha) &= \theta(1 - \theta) \end{aligned} \qquad (5.7)$$

et donc

$$\begin{aligned} E(X) &= n\theta \\ VAR(X) &= n\theta(1 - \theta) \end{aligned} \qquad (5.8)$$

Notons d'autre part que la variable $\frac{X}{n} = \bar{X}$ utilisée dans les applications est telle que

$$E(\bar{X}) = \theta$$
$$VAR(\bar{X}) = \frac{\theta(1 - \theta)}{n} \tag{5.9}$$

Le résultat (5.9) peut aussi s'obtenir directement de l'équation (5.6).

La fonction de répartition sera notée par:

$$PROBBNML(\theta, n, x) = \sum_{k=0}^{x} \binom{n}{k} \theta^k (1 - \theta)^{n - k} \tag{5.10}$$

et est disponible dans le système SAS sous cette appellation. On peut obtenir les termes de masse d'une variable binomiale à l'aide de l'équation

$$\binom{n}{x} \theta^x (1 - \theta)^{n-x} = PROBBNML(\theta,n,x) - PROBBNML(\theta,n,x-1) \tag{5.11}$$

Les coefficients d'asymétrie (β_1) et d'aplatissement (β_2) de la distribution sont:

$$\beta_1 = (1 - 2\theta)/\sqrt{n\theta(1 - \theta)}$$
$$\beta_2 = (1 - 6\theta(1 - \theta))/n\theta(1 - \theta) \tag{5.12}$$

Quelques exemples de distributions binomiales sont illustrés sur la figure 5.2.

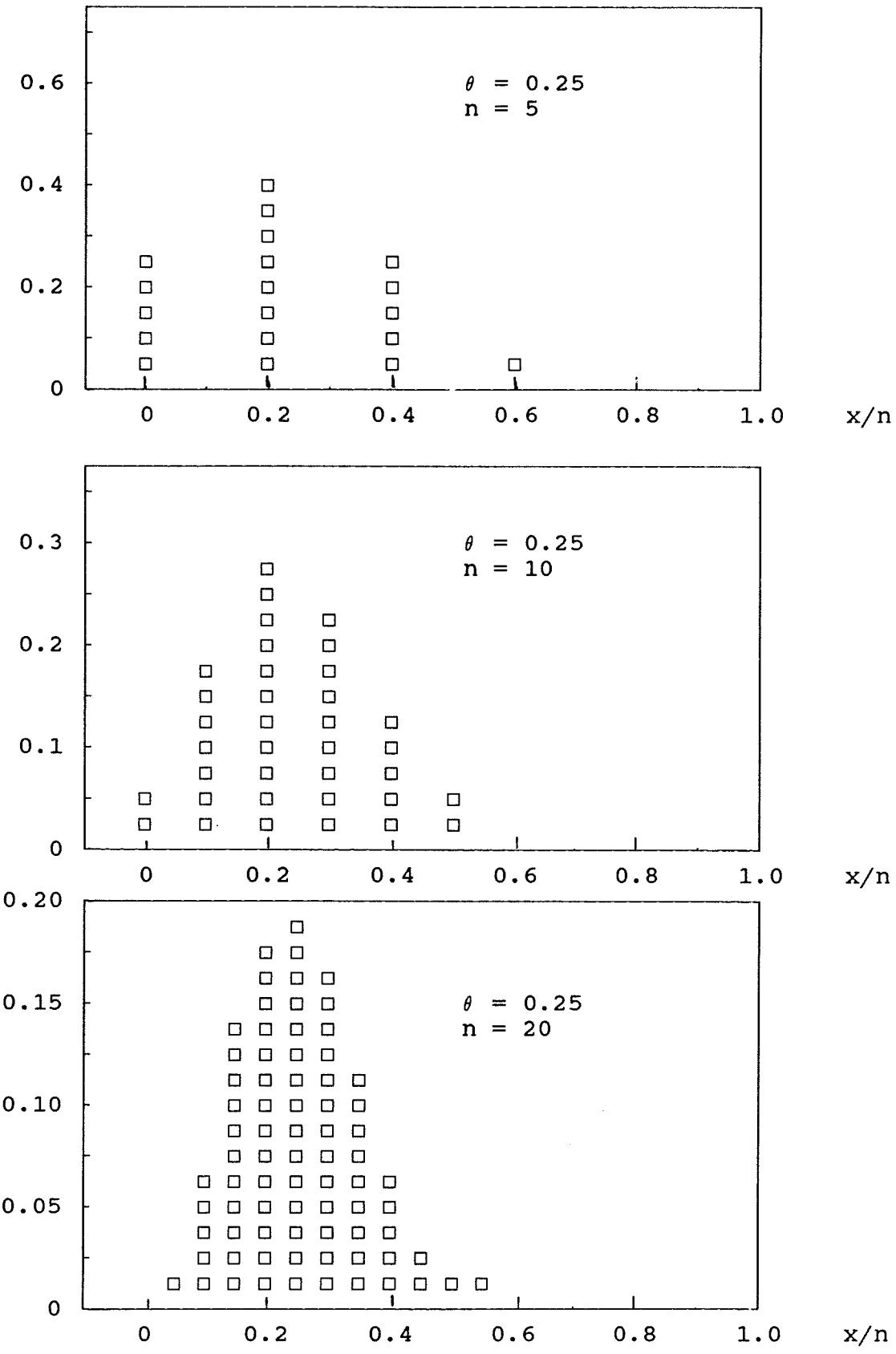


Figure 5.1: distribution binomiale

La distribution est symétrique seulement si $\theta = 1/2$. Une table sommaire de la fonction PROBBNML(θ,n,x) est fournie à la section 5.7 dans laquelle n = 2(1)20, θ = 0.05(0.05)0.50. Nous verrons plus loin une approximation de la fonction PROBBNML(θ,n,x) à l'aide de la fonction de répartition $\Phi(.)$ d'une variable normale centrée- réduite. Les relations de la distribution binomiale avec la distribution hypergéométrique et la loi de Poisson seront étudiées à la section 5.4.

Exemple 5.2A

Dans un projet de construction de route, on utilise cinq machines à niveler. La fiabilité de ces appareils est estimée à 0.06 sur une période de 900 heures à l'aide d'une certaine distribution. Si on suppose que les pannes sont indépendantes d'une machine à l'autre, calculez

(a) la probabilité que deux machines tombent en panne dans moins de 900 heures

(b) la probabilité d'au moins une machine en panne dans moins de 900 heures.

Solution

Posons X = nombre de machines en panne dans moins de 900 heures.

Selon les hypothèses du problème, X suit une distribution binomiale de paramètres n = 5 et θ = 0.06

Donc

(a) $P[X = 2] = \binom{5}{2} (0.06)^2(0.94)^3 = 0.7339$

(b) $P[X \geq 1] = 1 - P[X = 0] = 1 - (0.94)^5 = 0.2661$

Exemple 5.2B

Dans le design d'un ouvrage de contrôle d'une rivière, le débit de pointe annuel Q est une considération importante. Soit θ la probabilité que le débit excède un niveau Q_0.
Quelle est la probabilité que le débit de pointe excède le niveau Q_0 au moins une fois durant une période de n ans? Déterminez le nombre d'années n minimum afin que l'événement se réalise au moins une fois $(Q > Q_0)$ avec une probabilité d'au moins P. Calculez n pour les couples (θ,P) des valeurs suivantes:

θ = 0.1, 0.3, 0.5, 0.7, 0.9
P = 0.5, 0.7, 0.9, 0.99

Solution

Posons X = nombre de débits de pointe annuels excédant un niveau Q_0.

Si on suppose que les débits de pointe annuels sont indépendants d'une année à l'autre, X est alors distribué selon une loi binomiale de paramètres n et θ. On a

$$P[X \geq 1] = 1 - P[X = 0] = 1 - (1 - \theta)^n$$

On désire que

$$P[X \geq 1] \geq P$$

Donc

$$n \geq \frac{\ln (1-P)}{\ln (1-\theta)}$$

P \ θ	0.1	0.3	0.5	0.7	0.9
0.50	6.58	1.94	1.00	0.50	0.30
0.70	11.43	3.37	1.74	1.00	0.52
0.90	21.85	6.45	3.32	1.91	1.00
0.99	43.71	12.91	6.64	3.82	2.00

5.3 DISTRIBUTION DE POISSON

Plusieurs problèmes d'intérêt pour les ingénieurs impliquent la réalisation ou non d'événements dans le temps et/ou l'espace. Par exemple:

- des imperfections peuvent se placer n'importe où sur une plaque d'acier
- des secousses sismiques peuvent survenir n'importe où et n'importe quand dans des régions
- des accidents de traffic peuvent arriver en tout temps sur une route.

De tels problèmes impliquant le temps ou l'espace pourraient se modéliser à l'aide d'essais de Bernoulli en divisant le temps ou l'espace en petits intervalles et en assumant que l'événement se réalise ou non dans chaque intervalle. Pour tenir compte de la possibilité que l'événement se réalise plus d'une fois dans l'intervalle de temps (ou d'espace), on propose des hypothèses (a)-(b)-(c)-(d) définissant un PROCESSUS DE POISSON.

(a) Un événement peut se réaliser au hasard dans le temps (ou l'espace).

(b) Les réalisations de l'événement dans un intervalle de temps (ou d'espace) sont indépendantes des réalisations de l'événement dans tout autre intervalle de temps disjoint du premier.

(c) La probabilité de réalisation de l'événement dans un petit intervalle de temps (ou d'espace) Δt est égale à $\lambda\Delta t$, où $\lambda > 0$ est une constante appelée l'intensité ou cadence du processus.

(d) La probabilité de plus d'une réalisation durant un petit intervalle de temps (ou d'espace) Δt est négligeable.

Une des variables d'intérêt d'un tel processus est

X_t = nombre de réalisations de l'événement durant un intervalle de temps de longueur t.

Notons $p_X(x;t)$ la fonction de masse de probabilité de X_t et $X_{\Delta t}$ le nombre de réalisations de l'événement durant un intervalle de temps Δt.

Proposition 5.1: sous les hypothèses (a)-(b)-(c)-(d) ci-haut, la masse de probabilité de X_t est

$$p_X(x;t) = \frac{(\lambda t)^x}{x!} \exp(-\lambda t) \quad , \; x=0,1,2,\ldots$$

où λ est le nombre moyen de réalisations par unité de temps.

Démonstration

Les hypothèses (c) et (d) s'écrivent

$$P[X_{\Delta t} = 1] = \lambda \Delta t$$

$$P[X_{\Delta t} = k] = h((\Delta t)^k) \quad k \geq 2$$

$$P[X_{\Delta t} = 0] = 1 - \lambda \Delta t - h((\Delta t)^2)$$

où $h(\Delta t)$ représente une fonction telle que

$$\lim_{\Delta t \to 0} h(\Delta t) = 0$$

Considérons l'événement $\{X_{t+\Delta t} = x\}$ où $x \geq 1$ et sa décomposition

$$\left\{X_{t+\Delta t} = x\right\} = \bigcup_{j=0}^{x} \left\{X_t = x - j \text{ et } X_{\Delta t} = j\right\}$$

Alors

$$P[X_{t+\Delta t} = x] = \sum_{j=0}^{x} P[X_t = x-j]P[X_{\Delta t} = j]$$

$$\begin{aligned} p_X(x;\, t+\Delta t) &= P[X_t=x]P[X_{\Delta t}=0] + P[X_t=x-1]P[X_{\Delta t}=1] + h((\Delta t)^2) \\ &= p_X(x;t)\,[1 - \lambda\Delta t - h((\Delta t)^2)] \\ &\quad + p_X(x-1;t)\,[\lambda \Delta t] + h((\Delta t)^2) \end{aligned} \tag{5.13}$$

Cette dernière équation peut se mettre sous la forme

$$\frac{p_X(x;\, t+\Delta t) - p_X(x;\, t)}{\Delta t} = -\lambda p_X(x;\, t) + \lambda p_X(x-1;\, t) + h(\Delta t) \tag{5.14}$$

En passant à la limite lorsque $\Delta t \to 0$ on obtient l'équation différentielle:

$$\frac{d}{dt} p_X(x;\ t) = -\lambda p_X(x;\ t) + \lambda p_X(x - 1;\ t) \qquad (5.15)$$

avec la condition initiale

$$p_X(x;\ 0) = 0 \qquad (5.16)$$

Pour $x = 0$ l'équation (5.15) se réduit à

$$\frac{d}{dt} p_X(0;\ t) = -\lambda p_X(0\ ;\ t) \qquad (5.17)$$

avec la condition initale

$$p_X(0;\ 0) = 1 \qquad (5.18)$$

La solution de (5.17) et (5.18) est

$$p_X(0;\ t) = \exp(-\lambda t)$$

Pour $x \geq 1$ la solution de (5.15) et (5.16) est

$$p_X(x;\ t) = \frac{(\lambda t)^x}{x!} \exp(-\lambda t) \qquad (5.19)$$

Le paramètre λ est l'intensité du processus et puisque

$$E(X_t) = \lambda t \qquad (5.20)$$

on a

$$\lambda = E(X_t)/t \qquad (5.21)$$

Donc, λ est le nombre moyen de réalisations de l'événement par unité de temps (ou d'espace)

Si l'intervalle de temps est fixé et on pose $\theta = \lambda t$, le nombre de réalisations X de l'événement a pour masse de probabilité

$$p_X(x;\ \theta) = (\theta^x/x!)\ \exp(-\theta) \qquad \theta > 0, x = 0,1,2,... \qquad (5.22)$$

Historiquement, la distribution fut développée par Simon Denis Poisson (1781 - 1840) dans un ouvrage sur l'application des probabilités aux jugements en matière civile et criminelle publié en 1837. Les processus de Poisson et la distribution de Poisson sont des modèles riches d'applications. Voici quelques exemples:

- désintégrations radioactives
- distributions spatiales des missiles V2 à Londres durant la Deuxième Guerre Mondiale
- comptages de bactéries et cellules dans le sang
- nombre de mauvaises interconnections de lignes téléphoniques.

Une table sommaire de la fonction de répartition de Poisson

$$\text{POISSON}(\theta, x) = \sum_{k=0}^{x} \frac{e^{-\theta}\theta^{x}}{x!} \tag{5.23}$$

est fournie à la section 5.7 pour certaines valeurs de θ.

Notons les principales caractéristiques numériques de la distribution de Poisson

$$\begin{aligned} E(X) &= \theta \\ VAR(X) &= \theta \\ \beta_1 &= 1/\sqrt{\theta} \\ \beta_2 &= 3 + (1/\theta) \end{aligned} \tag{5.24}$$

Exemple 5.3 design d'une baie pour virage à gauche sur une rue avec feux de circulation

Supposons que le cycle des feux de circulation est de t minutes et l'on désire que la baie puisse accomoder, en moyenne, N virages à gauche à l'heure et cela avec une probabilité de $1 - \alpha$. Déterminez la longueur de la baie en nombre k de longueurs de voitures en faisant l'hypothèse que les voitures sont toutes de même longueur. Déterminez les valeurs de k pour les combinaisons suivantes de t, N et α

$$\begin{aligned} t &= 0.5,\ 0.8,\ 1.0,\ 1.5,\ 2 \\ N &= 50,\ 100,\ 150,\ 200 \\ \alpha &= 0.10,\ 0.01 \end{aligned}$$

Solution

Posons X_t le nombre de voitures effectuant un virage à gauche durant une période de t minutes. En supposant que les voitures effectuant un virage à gauche est un processus de Poisson, l'intensité moyenne du processus est N/60 par minute. Donc X_t suit une distribution de Poisson de paramètre $\theta = Nt/60$. La probabilité qu'il n'y ait pas plus de k voitures durant une période de t minutes est:

$$P[X_t \leq k] = \sum_{\alpha=0}^{k} \left(\frac{Nt}{60}\right)^{\alpha} \frac{1}{\alpha!} \exp\left(-\frac{Nt}{60}\right) = \text{POISSON}\left(\theta = \frac{Nt}{60},\ k\right) \tag{5.25}$$

Il faut donc trouver la plus petite valeur de k telle que

$$\text{POISSON}\left(\theta = \frac{Nt}{60},\ k\right) = 1 - \alpha \tag{5.26}$$

La résolution de l'équation peut s'obtenir par des essais et le tableau ci-joint donne la liste des valeurs de k satisfaisant (5.26)

Tableau 5.1: exemple 5.3

$\alpha = 0.10$

N \ t	0.5	0.8	1	1.5	2
50	1	2	2	3	3
100	2	3	3	5	6
150	3	4	5	6	8
200	3	5	6	8	10

$\alpha = 0.01$

N \ t	0.5	0.8	1	1.5	2
50	2	3	3	4	5
100	3	5	5	7	8
150	4	6	7	9	11
200	5	7	8	11	13

5.4 COMPARAISON DES DISTRIBUTIONS: HYPERGÉOMÉTRIQUE, BINOMIALE, POISSON

Les distributions hypergéométriques, binomiale et Poisson présentent des ressemblances au niveau des fonctions de probabilités lorsque des paramètres sont convenablement choisis. Nous illustrons ce phénomène avec trois exemples et concluons en donnant des règles d'approximations.

CAS NO 1: Hypergéométrique: taille du lot = 1000 = N
nombre défectueux = 100 = D
taille échantillon = 100 = n
proportion défectueux = 0.10 = D/N

Binomiale: proportion défectueux = 0.10 = θ
taille échantillon = 100 = n

Poisson : paramètre = 10 = λ = nθ

Tableau 5.1: cas no 1

	FONCTION DE RÉPARTITION			FONCTION DE MASSE		
x	HYPERG.	BINOMIALE	POISSON	HYPERG.	BINOMIALE	POISSON
0	0.000015	0.000027	0.000045	0.000015	0.000027	0.000045
1	0.000198	0.000322	0.000499	0.000183	0.000295	0.000454
2	0.001319	0.001945	0.002769	0.001121	0.001623	0.002270
3	0.005789	0.007836	0.010336	0.004470	0.005892	0.007567
4	0.018866	0.023711	0.029253	0.013077	0.015875	0.018917
5	0.048807	0.057577	0.067086	0.029942	0.033866	0.037833
6	0.104685	0.117156	0.130141	0.055878	0.059579	0.063055
7	0.192087	0.206051	0.220221	0.087402	0.088895	0.090079
8	0.309033	0.320874	0.332820	0.116946	0.114823	0.112599
9	0.444981	0.451290	0.457930	0.135947	0.130416	0.125110
10	0.583966	0.583156	0.583040	0.138985	0.131865	0.125110
11	0.710161	0.703033	0.696776	0.126194	0.119878	0.113736
12	0.812745	0.801821	0.791556	0.102585	0.098788	0.094780
13	0.887910	0.876123	0.864464	0.075165	0.074302	0.072908
14	0.937833	0.927427	0.916542	0.049923	0.051304	0.052077
15	0.968036	0.960109	0.951260	0.030203	0.032682	0.034718
16	0.984750	0.979401	0.972958	0.016714	0.019292	0.021699
17	0.993241	0.989993	0.985722	0.008491	0.010592	0.012764
18	0.997214	0.995419	0.992813	0.003973	0.005427	0.007091
19	0.998930	0.998021	0.996546	0.001717	0.002602	0.003732
20	0.999617	0.999192	0.998412	0.000687	0.001171	0.001866

CAS NO 2: Hypergéométrique: taille du lot = 1000
nombre défectueux = 100
taille échantillon = 20

Binomiale: proportion défectueux = 0.10
taille échantillon = 20

Poisson: paramètre = 2

Tableau 5.2: cas no 2

	FONCTION DE RÉPARTITION			FONCTION DE MASSE		
x	HYPERG.	BINOMIALE	POISSON	HYPERG.	BINOMIALE	POISSON
0	0.11900	0.12158	0.13534	0.11900	0.12158	0.13534
1	0.38915	0.39175	0.40601	0.27015	0.27017	0.27067
2	0.67722	0.67693	0.67668	0.28807	0.28518	0.27067
3	0.86905	0.86705	0.85712	0.19182	0.19012	0.18044
4	0.95851	0.95683	0.94735	0.08945	0.08977	0.09022
5	0.98956	0.98875	0.98344	0.03105	0.03192	0.03608
6	0.99789	0.99761	0.99547	0.00832	0.00886	0.01203
7	0.99965	0.99958	0.99890	0.00176	0.00197	0.00343
8	0.99995	0.99994	0.99976	0.00030	0.00035	0.00085
9	0.99999	0.99999	0.99995	0.00004	0.00005	0.00019
10	1.00000	1.00000	0.99999	0.00000	0.00000	0.00003

CAS NO 3: Hypergéométrique: taille du lot = 1000
nombre défectueux = 100
taille échantillon = 10

Binomiale: proportion défectueux = 0.10
taille échantillon = 10

Poisson: paramètre = 1

Tableau 5.3: cas no 3

	FONCTION DE RÉPARTITION			FONCTION DE MASSE		
x	HYPERG.	BINOMIALE	POISSON	HYPERG.	BINOMIALE	POISSON
0	0.34693	0.34868	0.36788	0.34693	0.34868	0.36788
1	0.73630	0.73610	0.73576	0.38936	0.38742	0.36787
2	0.93076	0.92981	0.91970	0.19446	0.19371	0.18394
3	0.98767	0.98720	0.98101	0.05691	0.05739	0.06131
4	0.99848	0.99837	0.99634	0.01080	0.01116	0.01532
5	0.99987	0.99985	0.99941	0.00139	0.00148	0.00306
6	0.99999	0.99999	0.99992	0.00012	0.00013	0.00051
7	1.00000	1.00000	0.99999	0.00000	0.00000	0.00007
8	1.00000	1.00000	1.00000	0.00000	0.00000	0.00000
9	1.00000	1.00000	1.00000	0.00000	0.00000	0.00000
10	1.00000	1.00000	1.00000	0.00000	0.00000	0.00000

L'examen de ces trois exemples et des analyses comparatives des fonctions de masse et des fonctions de répartition permettent de dégager les règles suivantes:

(a) On peut approcher une distribution hypergéométrique de paramètres (N,D,n) par une distribution binomiale de paramètres (n, θ) où $\theta = D/N$ si $n/N \leq 0.10$

(b) On peut approcher une distribution binomiale de paramètres (n,θ) par une distribution de Poisson de paramètre $\lambda = n\theta$ si $n \geq 10$ et $\theta \leq 0.10$

(c) D'une manière générale, la précision des approximations précédentes est plus grande pour les fonctions de répartition que pour les fonctions de masse.

Les règles (a) et (b) se traduise par la proposition 5.2.

Proposition 5.2

(a) $$\lim_{\substack{N\to\infty \\ D/N = \theta \text{ fixe}}} \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

(b) $$\lim_{\substack{n\to\infty \\ n\theta=\lambda \text{ fixe}}} \binom{n}{x} \theta^x (1-\theta)^{n-x} = \frac{e^{-\lambda}\ \lambda^x}{x!}$$

5.5 DISTRIBUTION GÉOMÉTRIQUE

Supposons que l'on effectue, à intervalles de temps régulier, des essais de Bernoulli et que θ $(0 < \theta < 1)$ représente la probabilité de réalisation de l'événement: on dit alors qu'il s'agit d'un processus de Bernoulli.

On s'intéresse au temps X de la première réalisation de l'événement. Si X = x alors la première réalisation s'est effectuée au x ième essai (temps) et à tous les x - 1 essais précédents l'événement ne s'est pas réalisé. La masse de probabilité de X est donc

$$p_X(x;\ \theta) = P[X = x] = \theta(1 - \theta)^{x-1} \qquad (5.27)$$

$$x = 1,2,\ldots;\ 0 < \theta < 1$$

connue sous le nom de distribution GÉOMÉTRIQUE puisque les valeurs de la fonction sont en progression géométrique.

La fonction de répartition associée sera notée $GEO(x;\theta)$

$$\begin{aligned} GEO(x;\ \theta) &= \sum_{\alpha=1}^{x} \theta(1 - \theta)^{\alpha-1} \\ &= 1 - (1 - \theta)^{x}\ ,\quad x = 1,\ 2,\ \ldots \end{aligned} \qquad (5.28)$$

La moyenne et la variance de la distribution sont:

$$\begin{aligned} E(X) &= 1/\theta \\ VAR(X) &= (1 - \theta)/\theta^2 \end{aligned} \qquad (5.29)$$

La distribution géométrique possède la propriété suivante:

$$P[X = x_1 + x_2 \,|\, X > x_2] = P[X = x_1] \quad \text{pour tout } (x_1, x_2) \qquad (5.30)$$

Cette propriété caractérise uniquement la distribution géométrique parmi toutes les distributions discrètes définies sur les entiers naturels et est dite propriété de non-vieillissement.

La variable X est aussi appelée le temps de la première récurrence et, puisqu'il y a indépendance entre les réalisations, on peut dire que X est le temps entre deux réalisations consécutives du processus. La valeur moyenne de X s'appelle PÉRIODE DE RÉCURRENCE MOYENNE.

Exemple 5.4

Lors de la construction de certains ouvrages tels des barrages, grands édifices, plates formes en haute mer, etc.. on doit tenir compte de la réalisation possible d'événements associés au caractère aléatoire de certaines variables, tels le débit d'une rivière, la vitesse du vent, l'intensité d'une secousse sismique, etc... Prenons, par exemple, le cas des crues printanières et faisons l'hypothèse que les débits excédant un niveau Q_0, est un processus de Bernoulli. On veut déterminer, avec quelle période de récurrence moyenne $r(a,\alpha)$ on doit concevoir l'ouvrage si on veut s'assurer, avec une probabilité de $(1 - \alpha)$, que l'on contrôle toutes les crues durant une période de a année. Calculez les valeurs de $r(a,\alpha)$ pour

$$a = 1, 2, 5, 10, 20, 30, 50$$
$$\alpha = 0.50, 0.25, 0.10, 0.05$$

Solution

Notons par X le nombre de crues excédant la capacité de l'ouvrage durant la période de a années et θ la probabilité d'une telle crue durant une année quelconque. Avec l'hypothèse d'un processus de Bernoulli on sait que X suit une distribution binomiale de paramètres $n = a$ et $\theta = 1/r$ où r est la période de récurrence moyenne du processus. L'ouvrage contrôle toutes les crues si $X = 0$ et on veut en être sûr avec probabilité $1 - \alpha$. Donc

$$P[X = 0] = 1 - \alpha$$

$$\left(1 - \frac{1}{r}\right)^a = 1 - \alpha$$

et alors

$$r = \frac{1}{1-(1-\alpha)^{1/a}} \qquad (5.31)$$

Le tableau 5.4 donne les valeurs de r en fonction de a et α

Tableau 5.4: exemple 5.4

a \ 1-α	0.50	0.75	0.90	0.95
1	2.00	4.00	10.00	20.00
2	3.41	7.46	19.49	39.49
3	7.73	17.89	47.96	97.98
10	14.93	35.26	95.41	195.46
20	29.36	70.02	190.32	390.41
30	43.78	104.78	285.24	585.37
50	72.64	174.06	475.06	975.29

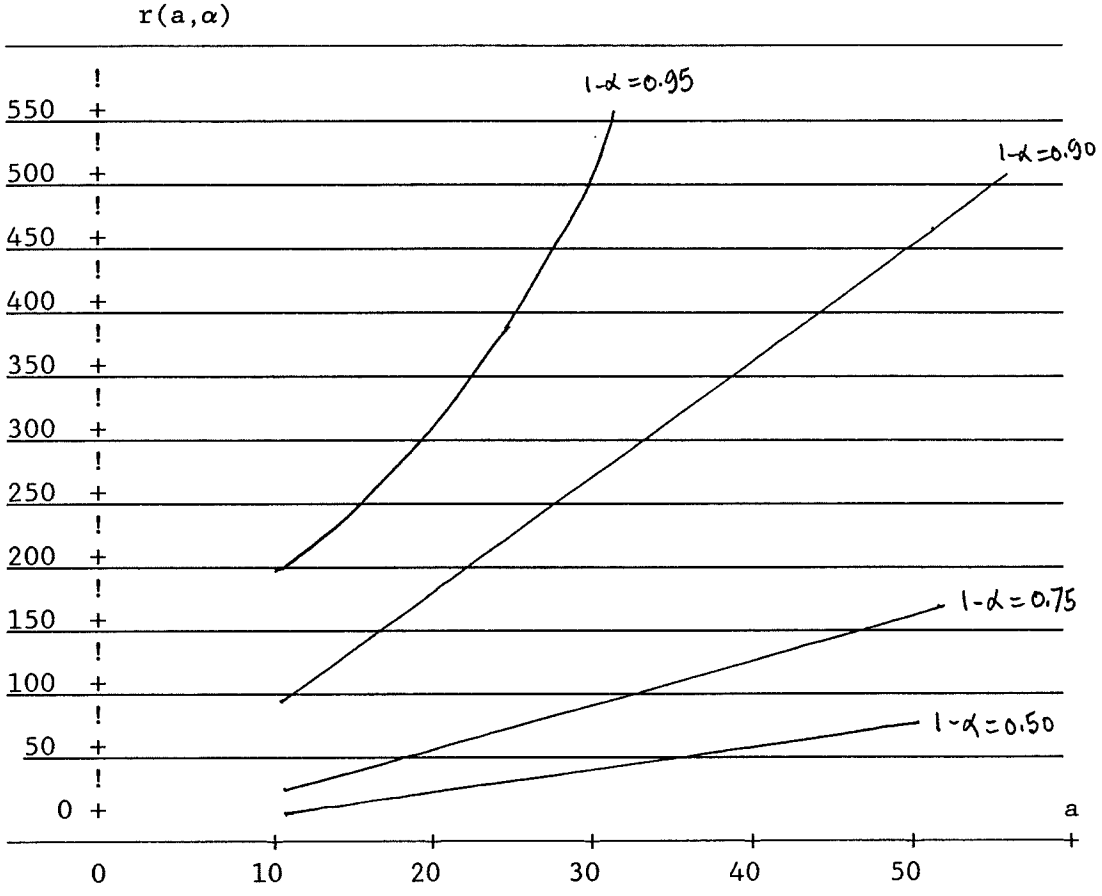


Figure 5.2: exemple 5.4

On constate, par exemple, que si la vie utile est fixée à 50 ans, on doit, pour être sûr à 90%, concevoir l'ouvrage sur la base d'événements ayant une période de récurrence moyenne de 475 ans.

Le graphique précédent peut être utilisé pour évaluer le risque d'un événement excédent la capacité de l'ouvrage durant sa vie. Par exemple, si l'ouvrage est construit pour les événements ayant une période de récurrence moyenne de 50 ans et si la vie de l'ouvrage est fixée à 10 ans, alors il y a une probabilité d'environ 0.20 que l'événement se réalise.

Si on pose r = a on a

$$P[X \geq 1] = 1 - \left(1 - \frac{1}{a}\right)^a \simeq 1 - e^{-1} = 0.632 \ , \ a \geq 10$$

Donc si la construction de l'ouvrage est basée sur des événements sur une période de récurrence moyenne égale à la vie utile, la probabilité est de 0.63 qu'il y ait au moins une réalisation de l'événement durant cette période. Le résultat est valable quelque soit la période de récurrence au-delà de 10 ans.

5.6 UTILISATION DE SAS: PLAN D'ÉCHANTILLONNAGE SIMPLE

Un lot de taille NL contient une proportion THETA (inconnue) d'articles défectueux. On veut prélever un échantillon de taille NE (à déterminer) et on décidera d'accepter le lot si l'échantillon contient C (à déterminer) articles défectueux ou moins. Si l'échantillon contient plus de C articles défectueux le lot sera rejeté et retourné au producteur.

Paramètres spécifiés par l'utilisateur:

NL: taille du lot (si disponible)
s'emploie seulement avec la distribution hypergéométrique
(échantillonnage sans remise)

THETA1: proportion de défectueux acceptable (entre 0 et 1)
(aussi notée AQL 'acceptable quality limit')

ALPHA: risque du producteur (entre 0 et 1) correspondant à THETA1

THETA2: proportion de défectueux rejetable (entre 0 et 1)
(aussi notée RQL 'rejectable quality limit' ou LTPD 'lot tolerance percent defective')

BETA: risque du consommateur (entre 0 et 1) correspondant à THETA2, THETA2 < THETA1, BETA < 1 - ALPHA.

Paramètres déterminés par le programme:

NE: taille de l'échantillon à prélever (entier)

C: constante (entier) critique conduisant au rejet du lot si l'échantillon contient plus de C défectueux

La détermination de NE et C est faite afin de satisfaire les équations suivantes:

$$P(THETA1,NL,NE,C) = 1 - ALPHA$$

$$P(THETA2,NL,NE,C) = BETA$$

où P(THETA,NL,NE,C) représente la probabilité d'observer au plus C articles défectueux dans un échantillon de taille NE prélevé dans un lot de taille NL.

Le calcul de NE et C se fait dans le DATA TAILLE après avoir précisé les paramètres NL, THETA1, ALPHA, THETA2, BETA dans le DATA PARAM. La méthode utilisée est une double itération sur NE et C.

La probabilité d'accepter le lot est calculée par la fonction PROBAC avec la distribution hypergéométrique (sans remise) ou la distribution binomiale (avec remise) selon le cas.

L'utilisateur dispose de deux modèles de probabilités:

hypergéométrique: PROBHYPR
binomiale: PROBBNML

Le choix de la distribution est dicté par les facteurs suivants:

- échantillonnage avec ou sans remise
- connaissance de NL ou non
- la valeur de THETA

La courbe d'efficacité (caractéristique) du plan est calculée et tracée dans le DATA COURBE en employant les valeurs de NE et C trouvées dans le DATA TAILLE.

Programme

```
DATA PARAM;
     INPUT  NUM PLAN $ NL THETA1 ALPHA THETA2 BETA  ;

     LABEL
               NUM='NUMERO DU PLAN'
              PLAN='IDENTIFICATION PLAN'
                NL='TAILLE LOT'
            THETA1='QUALITE ACCEPTABLE'
             ALPHA='RISQUE PRODUCTEUR'
            THETA2='QUALITE REJETABLE'
              BETA='RISQUE CONSOMMATEUR'  ;
     CARDS;
               1  A  0.02 0.05 0.08 0.10

DATA TAILLE;
           SET PARAM;              *récupération des paramètres;
           NE=0; C=-1              *initialisation;
           DO;                     *itération;
                DEUX: C+1;
                  UN: NE+1;

                      D=NL*THETA2;
                      LAMBA=NE*THETA2;

                      PROBAC=PROBHYPR(NL,D,NE,C);
                     *PROBAC=PROBBNML(THETA2,NE,C);

                      IF PROBAC > BETA THEN GO TO UN;

                      D=NL*THETA1;
                      LAMBA=NE*THETA1;

                      PROBAC=PROBHYPR(NL,D,NE,C)
                     *PROBAC=PROBBNML(THETA1,NE,C);

                      IF PROBAC < 1 - ALPHA THEN GO TO DEUX;

           END;
           RETAIN NE C NL THETA1 ALPHA THETA2 BETA  ;
FILE PRINT  NOTITLES;

     PUT / @17  'PLAN D-ECHANTILLONNAGE SIMPLE'
         / @10  '-----------------------------------------'
         / @10  'TAILLE LOT (NL)             =' @39 NL
         //@10  'QUALITE ACCEPTABLE (THETA1)=' @39 THETA1
         / @10  'RISQUE PRODUCTEUR (ALPHA)  =' @39 ALPHA
         / @10  'QUALITE REJETABLE (THETA2) =' @39 THETA2
         / @10  'RISQUE CONSOMMATEUR (BETA) =' @39 BETA
         /
         / @10  'LE PLAN SATISFAISANT LES CONDITIONS CI-HAUT EST:
```

```
      PUT /
          / @10  'TAILLE ECHANTILLON (NE)     =' @39 NE
          / @10  'NOMBRE MAXIMUM DEFECTUEUX DANS ECHANTILLON'
          / @10  'POUR ACCEPTER LE LOT (C)    =' @39 C
          //@10  '================================================';

DATA COURBE;
          SET TAILLE;
             DO K=1 TO 40;
                    THETA=K*THETA2/30;
                    D=THETA*NL;
                    LAMBA=NE*THETA;

                    PROBAC=PROBHYPR(NL,D,NE,C);
                   *PROBAC=PROBBNML(THETA,NE,C);

             OUTPUT;
             END;

PROC PRINT; VAR THETA PROBAC;
     TITLE 'Courbe caractéristique du plan optimal
           d'échantillonnage';
PROC PLOT;
           PLOT PROBAC*THETA='*'/VAXIS = 0 TO 1 BY 0.05;
                                HAXIS = 0 TO 0.15 BY 0.01;
```

PLAN D-ECHANTILLONNAGE SIMPLE

TAILLE LOT (NL) = 10000
QUALITE ACCEPTABLE (THETA1)= 0.02
RISQUE PRODUCTEUR (ALPHA) = 0.05
QUALITE REJETABLE (THETA2) = 0.08
RISQUE CONSOMMATEUR (BETA) = 0.1
LE PLAN SATISFAISANT LES CONDITIONS CI-HAUT EST:
TAILLE ECHANTILLON (NE) = 98
NOMBRE MAXIMUM DEFECTUEUX DANS ECHANTILLON
POUR ACCEPTER LE LOT (C) = 4

OBS	THETA	PROBAC
1	0.00266	1.00000
2	0.00533	0.99984
3	0.00800	0.99896
4	0.01066	0.99621
5	0.01333	0.99022
6	0.01600	0.98016
7	0.01866	0.96422
8	0.02133	0.94206
9	0.02400	0.91466
10	0.02666	0.88014
11	0.02933	0.84003
12	0.03200	0.79693
13	0.03466	0.74852
14	0.03733	0.69748
15	0.04000	0.64689
16	0.04266	0.59388
17	0.04533	0.54133
18	0.04800	0.49192
19	0.05066	0.44253
20	0.05333	0.39560
21	0.05600	0.35312
22	0.05866	0.31207
23	0.06133	0.27428
24	0.06400	0.24104
25	0.06666	0.20974
26	0.06933	0.18164
27	0.07200	0.15746
28	0.07466	0.13516
29	0.07733	0.11554
30	0.08000	0.09895
31	0.08266	0.08393
32	0.08533	0.07091
33	0.08800	0.06009
34	0.09066	0.05042
35	0.09333	0.04216
36	0.09600	0.03538
37	0.09866	0.02940
38	0.10133	0.02435
39	0.10400	0.02026
40	0.10666	0.01668

Graphique de la courbe caractéristique

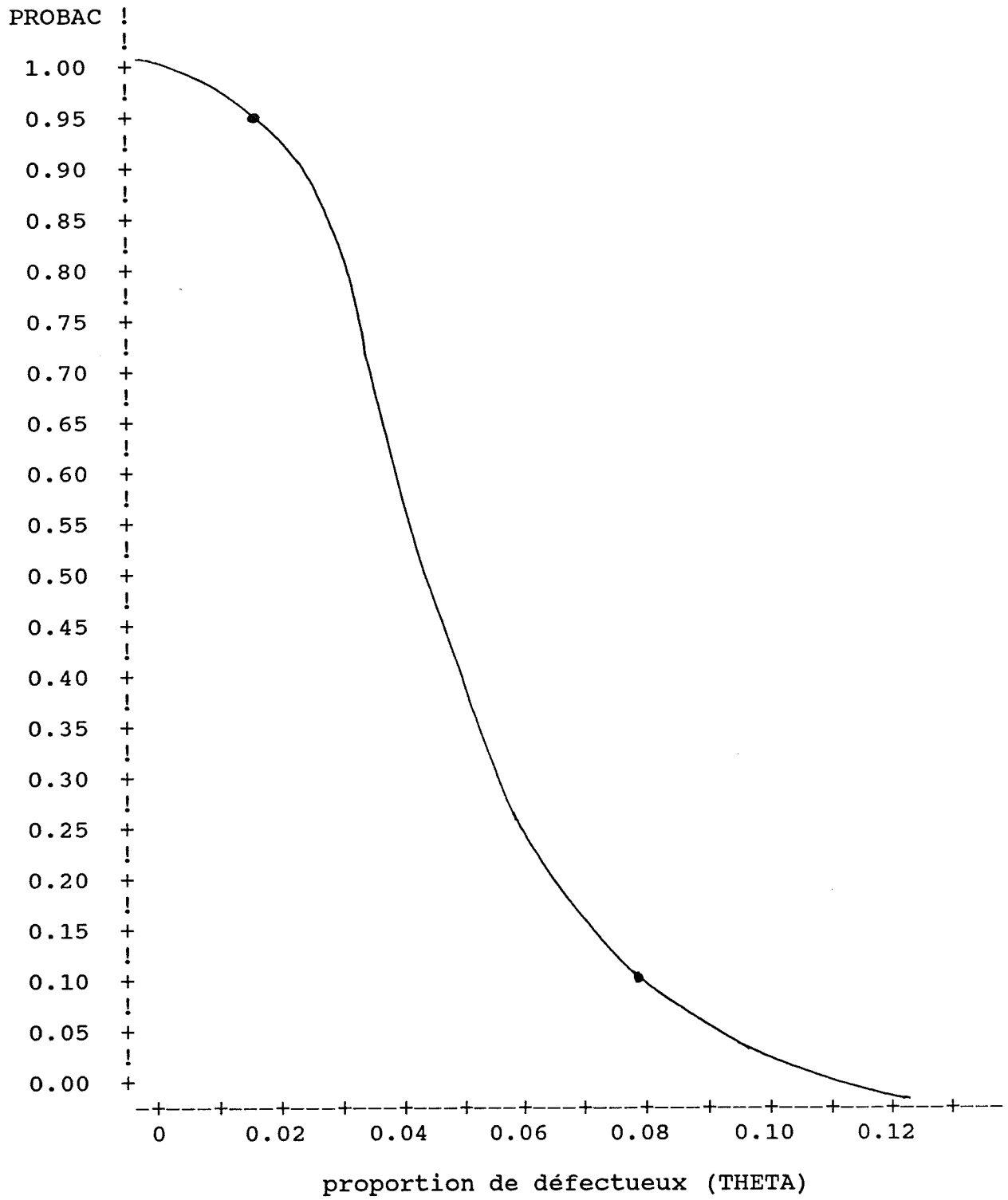


Figure 5.3: courbe caractéristique

5.7 TABLES

Le tableau 5.5 donne les valeurs de la fonction de répartition PROBBNML(θ,n,x) de la distribution binomiale de paramètres (n,θ):

$$\text{PROBBNML}(\theta,n,x) = \sum_{k=0}^{x} \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

Pour les valeurs suivantes des paramètres

$$n = 2(1)\ 20$$
$$\theta = 0.05(0.05)0.50$$

Pour les valeurs de $\theta > 0.50$ on peut employer la relation suivante:

$$\text{PROBBNML}(\theta,n,x) = 1 - \text{PROBBNML}(1-\theta,n,n-x-1)$$

et pour obtenir la valeur individuelle d'un terme de la loi binomiale, on procède par différence

$$\binom{n}{x} \theta^x (1-\theta)^{n-x} = \text{PROBBNML}(\theta,n,x) - \text{PROBBNML}(\theta,n,x-1)$$

Le tableau 5.6 donne les valeurs de la fonction de répartition d'une loi de Poisson de paramètre λ:

$$\text{POISSON}(\theta,x) = \sum_{k=0}^{x} \frac{e^{-\theta}\theta^x}{x!}$$

où θ = 0.01,0.05,0.10(0.10)2.0(0.20)3.0 (0.50)8.0
9.0,10.0,12.0,14.0,16.0,18.0,20.0

Tableau 5.5: fonction PROBBNML(θ,n,x)

n	x	θ .05	.10	.15	.20	.25	.30	.35	.40	.45	.50
2	0	.903	.810	.722	.640	.563	.490	.423	.360	.303	.250
	1	.998	.990	.978	.960	.938	.910	.878	.840	.798	.750
3	0	.857	.729	.614	.512	.422	.343	.275	.216	.166	.125
	1	.993	.972	.939	.896	.844	.784	.718	.648	.575	.500
	2	1.000	.999	.997	.992	.984	.973	.957	.936	.909	.875
4	0	.815	.656	.522	.410	.316	.240	.179	.130	.092	.063
	1	.986	.948	.890	.819	.738	.652	.563	.475	.391	.313
	2	1.000	.996	.988	.973	.949	.916	.874	.821	.759	.687
	3		1.000	.999	.998	.996	.992	.985	.974	.959	.938
5	0	.774	.590	.444	.328	.237	.168	.116	.078	.050	.031
	1	.977	.919	.835	.737	.633	.528	.428	.337	.256	.188
	2	.999	.991	.973	.942	.896	.837	.765	.683	.593	.500
	3	1.000	1.000	.998	.993	.984	.969	.946	.913	.869	.813
	4			1.000	1.000	.999	.998	.995	.990	.982	.969
6	0	.735	.531	.377	.262	.178	.118	.075	.047	.028	.016
	1	.967	.886	.776	.655	.534	.420	.319	.233	.164	.109
	2	.998	.984	.953	.901	.831	.744	.647	.544	.442	.344
	3	1.000	.999	.994	.983	.962	.930	.883	.821	.745	.656
	4		1.000	1.000	.998	.995	.989	.978	.959	.931	.891
	5				1.000	1.000	.999	.998	.996	.992	.984
7	0	.698	.478	.321	.210	.133	.082	.049	.028	.015	.008
	1	.956	.850	.717	.577	.445	.329	.234	.159	.102	.063
	2	.996	.974	.926	.852	.756	.647	.532	.420	.316	.227
	3	1.000	.997	.988	.967	.929	.874	.800	.710	.608	.500
	4		1.000	.999	.995	.987	.971	.944	.904	.847	.773
	5			1.000	1.000	.999	.996	.991	.981	.964	.938
	6					1.000	1.000	.999	.998	.996	.992
8	0	.663	.430	.272	.168	.100	.058	.032	.017	.008	.004
	1	.943	.813	.657	.503	.367	.255	.169	.106	.063	.035
	2	.994	.962	.895	.797	.679	.552	.428	.315	.220	.145
	3	1.000	.995	.979	.944	.886	.806	.706	.594	.477	.363
	4		1.000	.997	.990	.973	.942	.894	.826	.740	.637
	5			1.000	.999	.996	.989	.975	.950	.912	.855
	6				1.000	1.000	.999	.996	.991	.982	.965
	7						1.000	1.000	.999	.998	.996
9	0	.630	.387	.232	.134	.075	.040	.021	.010	.005	.002
	1	.929	.775	.599	.436	.300	.196	.121	.071	.039	.020
	2	.992	.947	.859	.738	.601	.463	.337	.232	.150	.090
	3	.999	.992	.966	.914	.834	.730	.609	.483	.361	.254
	4	1.000	.999	.994	.980	.951	.901	.828	.733	.621	.500
	5		1.000	.999	.997	.990	.975	.946	.901	.834	.746
	6			1.000	1.000	.999	.996	.989	.975	.950	.910
	7					1.000	1.000	.999	.996	.991	.980
	8							1.000	1.000	.999	.998

Tableau 5.5: suite

n	x	θ .05	.10	.15	.20	.25	.30	.35	.40	.45	.50
10	0	.599	.349	.197	.107	.056	.028	.013	.006	.003	.001
	1	.914	.736	.544	.376	.244	.149	.086	.046	.023	.011
	2	.988	.930	.820	.678	.526	.383	.262	.167	.100	.055
	3	.999	.987	.950	.879	.776	.650	.514	.382	.266	.172
	4	1.000	.998	.990	.967	.922	.850	.751	.633	.504	.377
	5		1.000	.999	.994	.980	.953	.905	.834	.738	.623
	6			1.000	.999	.996	.989	.974	.945	.898	.828
	7				1.000	1.000	.998	.995	.988	.973	.945
	8						1.000	.999	.998	.995	.989
	9							1.000	1.000	1.000	.999
11	0	.569	.314	.167	.086	.042	.020	.009	.004	.001	.000
	1	.898	.697	.492	.322	.197	.113	.061	.030	.014	.006
	2	.985	.910	.779	.617	.455	.313	.200	.119	.065	.033
	3	.998	.981	.931	.839	.713	.570	.426	.296	.191	.113
	4	1.000	.997	.984	.950	.885	.790	.668	.533	.397	.274
	5		1.000	.997	.988	.966	.922	.851	.753	.633	.500
	6			1.000	.998	.992	.978	.950	.901	.826	.726
	7				1.000	.999	.996	.988	.971	.939	.887
	8					1.000	.999	.998	.994	.985	.967
	9						1.000	1.000	.999	.998	.994
	10								1.000	1.000	1.000
12	0	.540	.282	.142	.069	.032	.014	.006	.002	.001	.000
	1	.882	.659	.443	.275	.158	.085	.042	.020	.008	.003
	2	.980	.889	.736	.558	.391	.253	.151	.083	.042	.019
	3	.998	.974	.908	.795	.649	.493	.347	.225	.134	.073
	4	1.000	.996	.976	.927	.842	.724	.583	.438	.304	.194
	5		.999	.995	.981	.946	.882	.787	.665	.527	.387
	6		1.000	.999	.996	.986	.961	.915	.842	.739	.613
	7			1.000	.999	.997	.991	.974	.943	.888	.806
	8				1.000	1.000	.998	.994	.985	.964	.927
	9						1.000	.999	.997	.992	.981
	10							1.000	1.000	.999	.997
	11									1.000	1.000
13	0	.513	.254	.121	.055	.024	.010	.004	.001	.000	.000
	1	.865	.621	.398	.234	.127	.064	.030	.013	.005	.002
	2	.975	.866	.692	.502	.333	.202	.113	.058	.027	.011
	3	.997	.966	.882	.747	.584	.421	.278	.169	.093	.046
	4	1.000	.994	.966	.901	.794	.654	.501	.353	.228	.133
	5		.999	.992	.970	.920	.835	.716	.574	.427	.291
	6		1.000	.999	.993	.976	.938	.871	.771	.644	.500
	7			1.000	.999	.994	.982	.954	.902	.821	.709

Tableau 5.5: suite

n	x	θ .05	.10	.15	.20	.25	.30	.35	.40	.45	.50
13	8				1.000	.999	.996	.987	.968	.930	.867
	9					1.000	.999	.997	.992	.980	.954
	10						1.000	1.000	.999	.996	.989
	11								1.000	.999	.998
	12									1.000	1.000
14	0	.488	.229	.103	.044	.018	.007	.002	.001	.000	.000
	1	.847	.585	.357	.198	.101	.047	.021	.008	.003	.001
	2	.970	.842	.648	.448	.281	.161	.084	.040	.017	.006
	3	.996	.956	.853	.698	.521	.355	.220	.124	.063	.029
	4	1.000	.991	.953	.870	.742	.584	.423	.279	.167	.090
	5		.999	.988	.956	.888	.781	.641	.486	.337	.212
	6		1.000	.998	.988	.962	.907	.816	.692	.546	.395
	7			1.000	.998	.990	.969	.925	.850	.741	.605
	8				1.000	.998	.992	.976	.942	.881	.788
	9					1.000	.998	.994	.982	.957	.910
	10						1.000	.999	.996	.989	.971
	11							1.000	.999	.998	.994
	12								1.000	1.000	.999
15	0	.463	.206	.087	.035	.013	.005	.002	.000	.000	.000
	1	.829	.549	.319	.167	.080	.035	.014	.005	.002	.000
	2	.964	.816	.604	.398	.236	.127	.062	.027	.011	.004
	3	.995	.944	.823	.648	.461	.297	.173	.091	.042	.018
	4	.999	.987	.938	.836	.686	.515	.352	.217	.120	.059
	5	1.000	.998	.983	.939	.852	.722	.564	.403	.261	.151
	6		1.000	.996	.982	.943	.869	.755	.610	.452	.304
	7			.999	.996	.983	.950	.887	.787	.654	.500
	8			1.000	.999	.996	.985	.958	.905	.818	.696
	9				1.000	.999	.996	.988	.966	.923	.849
	10					1.000	.999	.997	.991	.975	.941
	11						1.000	1.000	.998	.994	.982
	12								1.000	.999	.996
	13									1.000	1.000
16	0	.440	.185	.074	.028	.010	.003	.001	.000	.000	.000
	1	.811	.515	.284	.141	.063	.026	.010	.003	.001	.000
	2	.957	.789	.561	.352	.197	.099	.045	.018	.007	.002
	3	.993	.932	.790	.598	.405	.246	.134	.065	.028	.011
	4	.999	.983	.921	.798	.630	.450	.289	.167	.085	.038
	5	1.000	.997	.976	.918	.810	.660	.490	.329	.198	.105
	6		.999	.994	.973	.920	.825	.688	.527	.366	.227
	7		1.000	.999	.993	.973	.926	.841	.716	.563	.402
	8			1.000	.999	.993	.974	.933	.858	.744	.598
	9				1.000	.998	.993	.977	.942	.876	.773

Tableau 5.5: suite

n	x	θ .05	.10	.15	.20	.25	.30	.35	.40	.45	.50
16	10					1.000	.998	.994	.981	.951	.895
	11						1.000	.999	.995	.985	.962
	12							1.000	.999	.997	.989
	13								1.000	.999	.998
	14									1.000	1.000
17	0	.418	.167	.063	.023	.008	.002	.001	.000	.000	.000
	1	.792	.482	.252	.118	.050	.019	.007	.002	.001	.000
	2	.950	.762	.520	.310	.164	.077	.033	.012	.004	.001
	3	.991	.917	.756	.549	.353	.202	.103	.046	.018	.006
	4	.999	.978	.901	.758	.574	.389	.235	.126	.060	.025
	5	1.000	.995	.968	.894	.765	.597	.420	.264	.147	.072
	6		.999	.992	.962	.893	.775	.619	.448	.290	.166
	7		1.000	.998	.989	.960	.895	.787	.641	.474	.315
	8			1.000	.997	.988	.960	.901	.801	.663	.500
	9				1.000	.997	.987	.962	.908	.817	.685
	10					.999	.997	.988	.965	.917	.834
	11					1.000	.999	.997	.989	.970	.928
	12						1.000	.999	.997	.991	.975
	13							1.000	1.000	.998	.994
	14									1.000	.999
	15										1.000
18	0	.397	.150	.054	.018	.006	.002	.000	.000	.000	.000
	1	.774	.450	.224	.099	.039	.014	.005	.001	.000	.000
	2	.942	.734	.480	.271	.135	.060	.024	.008	.003	.001
	3	.989	.902	.720	.501	.306	.165	.078	.033	.012	.004
	4	.998	.972	.879	.716	.519	.333	.189	.094	.041	.015
	5	1.000	.994	.958	.867	.717	.534	.355	.209	.108	.048
	6		.999	.988	.949	.861	.722	.549	.374	.226	.119
	7		1.000	.997	.984	.943	.859	.728	.563	.391	.240
	8			.999	.996	.981	.940	.861	.737	.578	.407
	9			1.000	.999	.995	.979	.940	.865	.747	.593
	10				1.000	.999	.994	.979	.942	.872	.760
	11					1.000	.999	.994	.980	.946	.881
	12						1.000	.999	.994	.982	.952
	13							1.000	.999	.995	.985
	14								1.000	.999	.996
	15									1.000	.999
	16										1.000

Tableau 5.5: suite

n	x	θ .05	.10	.15	.20	.25	.30	.35	.40	.45	.50
19	0	.377	.135	.046	.014	.004	.001	.000	.000	.000	.000
	1	.755	.420	.198	.083	.031	.010	.003	.001	.000	.000
	2	.933	.705	.441	.237	.111	.046	.017	.005	.002	.000
	3	.987	.885	.684	.455	.263	.133	.059	.023	.008	.002
	4	.998	.965	.856	.673	.465	.282	.150	.070	.028	.010
	5	1.000	.991	.946	.837	.668	.474	.297	.163	.078	.032
	6		.998	.984	.932	.825	.666	.481	.308	.173	.084
	7		1.000	.996	.977	.923	.818	.666	.488	.317	.180
	8			.999	.993	.971	.916	.815	.667	.494	.324
	9			1.000	.998	.991	.967	.913	.814	.671	.500
	10				1.000	.998	.989	.965	.912	.816	.676
	11					1.000	.997	.989	.965	.913	.820
	12						.999	.997	.988	.966	.916
	13						1.000	.999	.997	.989	.968
	14							1.000	.999	.997	.990
	15								1.000	.999	.998
	16									1.000	1.000
20	0	.358	.122	.039	.012	.003	.001	.000	.000	.000	.000
	1	.736	.392	.176	.069	.024	.008	.002	.001	.000	.000
	2	.925	.677	.405	.206	.091	.035	.012	.004	.001	.000
	3	.984	.867	.648	.411	.225	.107	.044	.016	.005	.001
	4	.997	.957	.830	.630	.415	.238	.118	.051	.019	.006
	5	1.000	.989	.933	.804	.617	.416	.245	.126	.055	.021
	6		.998	.978	.913	.786	.608	.417	.250	.130	.058
	7		1.000	.994	.968	.898	.772	.601	.416	.252	.132
	8			.999	.990	.959	.887	.762	.596	.414	.252
	9			1.000	.997	.986	.952	.878	.755	.591	.412
	10				.999	.996	.983	.947	.872	.751	.588
	11				1.000	.999	.995	.980	.943	.869	.748
	12					1.000	.999	.994	.979	.942	.868
	13						1.000	.998	.994	.979	.942
	14							1.000	.998	.994	.979
	15								1.000	.998	.994
	16									1.000	.999
	17										1.000

Tableau 5.6: fonction POISSON(θ,x)

θ

x	.01	.05	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	.990	.951	.905	.819	.741	.670	.607	.549	.497	.449	.407
1	1.000	.999	.995	.982	.963	.938	.910	.878	.844	.809	.772
2		2.000	1.000	.999	.996	.992	.986	.977	.966	.953	.937
3				1.000	1.000	.999	.998	.997	.994	.991	.987
4						1.000	1.000	1.000	.999	.999	.998
5									1.000	1.000	1.000

θ

x	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	.368	.333	.301	.273	.247	.223	.202	.183	.165	.150	.135
1	.736	.699	.663	.627	.592	.558	.525	.493	.473	.434	.406
2	.920	.900	.879	.857	.833	.809	.783	.757	.731	.704	.677
3	.981	.974	.966	.957	.946	.934	.921	.907	.891	.875	.857
4	.996	.995	.992	.989	.986	.981	.976	.970	.964	.956	.947
5	.999	.999	.998	.998	.997	.996	.994	.992	.990	.987	.983
6	1.000	1.000	1.000	1.000	.999	.999	.999	.998	.997	.997	.995
7					1.000	1.000	1.000	1.000	.999	.999	.999
8									1.000	1.000	1.000

θ

x	2.2	2.4	2.6	2.8	3.0	3.5	4.0	4.5	5.0	5.5	6.0
0	.111	.091	.074	.061	.050	.030	.018	.011	.007	.004	.002
1	.355	.308	.267	.231	.199	.136	.092	.061	.040	.027	.017
2	.623	.570	.518	.469	.423	.321	.238	.174	.125	.088	.062
3	.819	.779	.736	.692	.647	.537	.433	.342	.265	.202	.151
4	.928	.904	.877	.848	.815	.725	.629	.532	.440	.358	.285
5	.975	.964	.951	.935	.916	.858	.785	.703	.616	.529	.446
6	.993	.988	.983	.976	.966	.935	.889	.831	.762	.686	.606
7	.998	.997	.995	.992	.988	.973	.949	.913	.867	.809	.744
8	1.000	.999	.999	.998	.996	.990	.979	.960	.932	.894	.847
9		1.000	1.000	.999	.999	.997	.992	.983	.968	.946	.916
10				1.000	1.000	.999	.997	.993	.986	.975	.957
11						1.000	.999	.998	.995	.989	.980
12							1.000	.999	.998	.996	.991
13								1.000	.999	.998	.996
14									1.000	.999	.999
15										1.000	.999
16											1.000

Tableau 5.6: suite

	θ										
x	6.5	7.0	7.5	8.0	9.0	10.0	12.0	14.0	16.0	18.0	20.0
0	.002	.001	.001								
1	.011	.007	.005	.003	.001						
2	.043	.030	.020	.014	.006	.003	.001				
3	.112	.082	.059	.042	.021	.010	.002				
4	.224	.173	.132	.100	.055	.029	.008	.002			
5	.369	.301	.241	.191	.116	.067	.020	.006	.001		
6	.527	.450	.378	.313	.207	.130	.046	.014	.004	.001	
7	.673	.599	.525	.453	.324	.220	.090	.032	.010	.003	.001
8	.792	.729	.662	.593	.456	.333	.155	.062	.022	.007	.002
9	.877	.830	.776	.717	.587	.458	.242	.109	.043	.015	.005
10	.933	.901	.862	.816	.706	.583	.347	.176	.077	.030	.011
11	.966	.947	.921	.888	.803	.697	.462	.260	.127	.055	.021
12	.984	.973	.957	.936	.876	.792	.576	.358	.193	.092	.039
13	.993	.987	.978	.966	.926	.864	.682	.464	.275	.143	.066
14	.997	.994	.990	.983	.959	.917	.772	.570	.368	.208	.105
15	.999	.998	.995	.992	.978	.951	.844	.669	.467	.287	.157
16	1.000	.999	.998	.996	.989	.973	.899	.756	.566	.375	.221
17		1.000	.999	.998	.995	.986	.937	.827	.659	.469	.297
18			1.000	.999	.998	.993	.963	.883	.742	.562	.381
19				1.000	.999	.997	.979	.923	.812	.651	.470
20					1.000	.998	.988	.952	.868	.731	.559
21						.999	.994	.971	.911	.799	.644
22						1.000	.997	.983	.942	.855	.721
23							.999	.991	.963	.899	.787
24							.999	.995	.978	.932	.843
25							1.000	.997	.987	.955	.888
26								.999	.993	.972	.922
27								.999	.996	.983	.948
28								1.000	.998	.990	.966
29									.999	.994	.978
30									.999	.997	.987
31									1.000	.998	.992
32										.999	.995
33										1.000	.997

5.8 EXERCICES

5.1. Des articles produit en série contiennent en moyenne 2% d'articles défectueux. À chaque heure un échantillon de 50 articles est prélevé et on arrête la production si l'on trouve plus de 2 articles défectueux. Quelle est la probabilité que la production soit arrêtée avec ce plan d'échantillonnage?

5.2. Des lots de 25 appareils sont assujettis au plan d'échantillonnage suivant: un échantillon de 5 appareils est choisi sans remise et le lot est refusé si 3 ou plus sont défectueux; autrement le lot (diminué des appareils défectueux de l'échantillon) est accepté. Si on suppose que le lot contient 4 appareils défectueux calculez

(a) la probabilité que le lot soit accepté

(b) la taille moyenne des lots acceptés

(c) une approximation de la probabilité calculée en (a) à l'aide de la loi binomiale.

5.3. Un acheteur reçoit des lots de 25 articles et utilise le plan d'échantillonnage suivant: un échantillon de taille n est prélevé sans remise et le lot est refusé s'il contient au moins un article défectueux. Quelle est la taille de l'échantillon si l'acheteur veut être sûr à 0.95 de rejeter un lot contenant 7 articles défectueux?

5.4. Supposons que seulement 20% de tous les conducteurs arrêtent complètement à une intersection sous un signal d'arrêt pour les quatres directions, lorsqu'aucun autre véhicule n'est visible. Quelle est la probabilité que sur 20 chauffeurs choisis au hasard arrivant à une intersection.

(a) au plus 5 s'arrêtent complètement

(b) exactement 5 s'arrêtent complètement

(c) au moins 5 s'arrêtent complètement

(d) combien sur les vingt prochains chauffeurs espérez-vous qu'ils s'arrêtent complètement?

5.5. Dans une collection de 20 roches, 10 sont de type basalte et 10 sont de type granite. Cinq roches sont choisies au hasard (sans remise) pour des fins d'analyses chimiques. Soit X le nombre de roches de type basalte dans l'échantillon.

(a) Précisez la loi de probabilité de X et ses paramètres.

(b) Calculez la probabilité que l'échantillon contienne seulement des roches de même type.

5.6. Un manufacturier et un acheteur s'entendent sur la procédure suivante. L'acheteur inspecte 20 articles pris au hasard dans le lot; s'il trouve 1 article défectueux ou moins l'acheteur paie le lot 1 100$, s'il en trouve 5 ou plus il le paie 200$, autrement le lot coûte 600$. Si la proportion d'articles défectueux dans le lot est de 0.20, calculez le prix moyen d'un lot.

5.7. Dans une aérogare cinq radars sont en opération et chaque radar a une probabilité de 0.9 de détecter un avion. Les radars opèrent indépendamment les uns des autres.

(a) Calculez la probabilité de détecter un avion par au moins 4 radars.

(b) Sachant qu'au moins 3 radars ont détecté un avion quelle est la probabilité que les 5 radars aient détecté cet avion?

(c) Combien de radars sont nécessaires si on veut que la probabilité de détecter un avion par au moins un radar soit de 0.9999?

5.8. La probabilité qu'un transistor d'un certain type fonctionne pendant plus de 500 heures est égale à 0.2. Si on teste 20 transistors de ce type, quelle est la probabilité que

(a) exactement 4 transistors fonctionnent pendant plus de 500 heures

(b) au plus 6 des 20 transistors fonctionnent pendant plus de 500 heures

(c) entre 4 et 6 (inclusivement) transistors fonctionnent pendant plus de 500 heures.

5.9. Un examen est composé de 10 questions de 1 point chacune. Chaque réponse vaut 1 ou 0 selon quelle est jugée bonne ou erronnée. Le professeur estime que les étudiants ont une probabilité de 0.85 de réussir chaque question. La note de passage a été fixée à 7. Calculez la probabilité que:

(a) un étudiant réussisse l'examen

(b) au plus un étudiant d'un groupe de 20 échoue l'examen.

5.10. Soit X une variable distribuée selon une loi binomiale de paramètres (n,θ).

(a) Montrez que $Y = n - X$ est distribuée selon une loi binomiale de paramètres $(n, 1-\theta)$

(b) Soit $$b(x;n,\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

et $$PROBBNML(\theta,n,x) = \sum_{k=0}^{x} b(k;n,\theta)$$

Montrez que

(i) $b(x;n,1-\theta) = b(n-x;n,\theta)$

(ii) $PROBBNML(1-\theta,n,x) = 1-PROBBNML(\theta,n,n-x-1)$

(c) Utilisez les résultats précédents ainsi que la table 5.5 pour évaluer les probabilités suivantes d'une variable X distribuée selon une loi binomiale de paramètres $(n=15, \theta=0.75)$

(i) $P[X \leq 10]$
(ii) $P[X = 10]$
(iii) $P[X > 8]$

5.11. La probabilité d'observer pendant une minute aucune imperfection sur une plaque soumise à un procédé d'électrolyse continuel est de 0.82. Si on suppose que les imperfections se réalisent selon un processus de Poisson, calculez la probabilité d'observer:

(a) aucune imperfection pendant 5 minutes

(b) au plus 2 imperfections pendant 5 minutes

(c) au moins 3 imperfections pendant 1 minute

(d) entre 2 et 4 imperfections pendant 2 minutes

(c) Déterminez une expression mathématique $H(n_o, c)$ représentant la probabilité qu'au moins une personne ayant fait des réservations ne trouve pas de place sur l'envolée.

(d) Supposons que $n_o = 100$ pour un type d'envolée. La compagnie veut déterminer une politique de réservation (valeur de c) de telle sorte que la probabilité qu'au moins une personne ne trouve pas de place sur l'envolée ne dépasse pas 0.10. Déterminez la valeur de c.

5.16 Des clients se présentent à un comptoir à raison de 2 par minute en moyenne. Calculez la probabilité que dans chacune de 2 périodes disjointes de 2 minutes le nombre de clients soit:

(a) égal à 2 (b) 2 ou plus (c) au plus 3

5.17 Certaines pièces essentielles au bon fonctionnement d'une machine tombent en panne à raison de 1 pièce par 5 semaines. Il y a 2 pièces de rechange en inventaire et le nouvel approvisionnement arrivera dans 9 semaines. Quelle est la probabilité que, durant les 9 prochaines semaines, la production cesse pendant 1 semaine ou plus à cause d'un manque de pièces de rechange?

5.18 La demande quotitienne d'un certain produit est une variable de Poisson de moyenne 20. Combien d'articles devrait-on stocker si on veut répondre à la demande avec une probabilité d'au moins 0.99?

5.19 Des clients se présentent à un comptoir au rythme moyen de 60 à l'heure. Quelle est la probabilité que:

(a) 5 minutes se soient écoulées depuis que le dernier client est arrivé

(b) 3 minutes se soient écoulées depuis que l'avant dernier client soit arrivé

(c) il n'y ait aucun client durant les 2 prochaines minutes étant donné qu'il y a eu un client ou plus pendant les deux dernières minutes.

5.12. Une opératrice de la compagnie de téléphone place en moyenne, 24 appels interurbains par heure; les appels arrivent suivant un processus de Poisson. Calculez la probabilité

(a) qu'elle ne place aucun appel durant une période de 5 minutes

(b) qu'elle place au moins 6 appels durant une période de 10 minutes

5.13. Le nombre de pannes durant une semaine d'un appareil est une variable de type Poisson. Durant la dernière année (52 semaines) il y a eu 26 pannes. Calculez les probabilités des événements suivants:

(a) Aucune panne durant deux semaines consécutives.

(b) Au plus 2 pannes durant 5 semaines consécutives.

5.14. Le nombre de pannes X d'un appareil pendant une période de t heures est une variable de Poisson de moyenne 0.8t. Une compagnie fait la location de l'appareil à 400$ l'heure et le répare au coût de $100X^2$$.

(a) Montrez que $E(X^2) = 0.8t + 0.64t^2$

(b) Exprimez le profit moyen pour une période de t heures en fonction de t

(c) Pour quelle valeur de t ce profit est-il maximum et que vaut-il?

5.15. Une compagnie d'aviation a constaté que 4% des personnes qui font des réservations pour ses envolées ne se présentent pas au départ. La compagnie a donc décidé d'accepter un nombre de réservations $n_o + c$ plus grand que le nombre de sièges disponibles n_o. Notons par X le nombre de personnes ayant fait des réservations et se présentant au départ d'une envolée et par Y le nombre de personnes ayant fait des réservations et ne se présentant pas au départ d'une envolée.

(a) Quelle est la loi de probabilité de X? Précisez ses paramètres.

(b) Quelle est la loi de probabilité de Y? Précisez ses paramètres.

5.20 Le nombre d'imperfections sur le fini extérieur d'une voiture nouvellement fabriquée est une variable de Poisson avec un taux moyen de 0.1 par m^2. On refait le fini extérieur au coût de 500$ si on décèle trois imperfections ou plus lors du contrôle final de qualité. Le profit brut est de 1 000$ pour chaque voiture fabriquée et la surface totale d'une voiture est approximativement 10 m^2.

(a) Calculez le profit net moyen.

(b) Il serait possible de réduire à 0.05 le nombre moyen d'imperfections par m^2 mais le profit brut serait alors de (1 000 - C)$. Pour quelle valeur maximale de C, le nouveau procédé serait-il plus rentable que le premier?

5.21 Une compagnie d'assurance estime à 0.002 la probabilité d'une réclamation couvrant un certain type d'accident. La compagnie possède 1 000 polices d'assurance couvrant ce type d'accident. Quelle est la probabilité de recevoir au plus 3 réclamations? Au moins 5?

5.22 Un commerçant reçoit un lot de 100 appareils. Pour sauver du temps, il décide d'utiliser le plan d'échantillonnage suivant: il choisit deux appareils, au hasard et sans remise, et il décide d'accepter le lot si les deux appareils choisis ne sont pas défectueux. Soit X la variable aléatoire qui représente le nombre d'appareils défectueux dans l'échantillon.

(a) Précisez la loi de probabilité de X, ainsi que les paramètres de la loi.

(b) Si le lot contient 2 appareils défectueux, calculez la probabilité que le lot soit accepté.

(c) Calculez la probabilité en (b) à l'aide d'une loi binomiale.

(d) Calculez la probabilité en (c) à l'aide d'une loi de Poisson.

5.23 Le nombre moyen de pannes d'un appareil est de 10 pour une période d'opération de 10 000 heures. Calculez la probabilité que l'appareil soit encore en opération après une période d'utilisation de 100 heures.

5.24 Le nombre de composants X tombant en panne d'un appareil complexe est distribué selon une loi de Poisson de paramètre λ. La durée, la réparation D est reliée au nombre de composants en panne selon l'équation

$$D = D_0 \ (1-e^{-\alpha X})$$

où D_0 et α sont des constantes. Calculez la durée moyenne des réparations.

5.25 Le nombre moyen d'articles non-conformes produits par un procédé de fabrication est de 6 par période de 25 minutes selon un processus de Poisson. On considère une heure de production subdivisée en 12 périodes de 5 minutes. Posons

X: le nombre d'articles non-conformes produits durant une période de 5 minutes

Z: le nombre de périodes parmi les 12 pour lesquelles aucun article non-conforme est produit

Y: le nombre de périodes de 5 minutes requises pour obtenir une première période pendant laquelle aucun article non-conforme est produit.

(a) Précisez la distribution de X et son(ses) paramètre(s).

(b) Précisez la distribution de Y et son(ses) paramètre(s).

(c) Précisez la distribution de Z et son(ses) paramètre(s).

(d) A quelle période, en moyenne, aucun article non-conforme sera produit pour la première fois?

(e) Quelle est la probabilité, que dans exactement 2 des 12 périodes, on observe aucun article non-conforme?

(f) Quelle est la probabilité, que l'on ait produit exactement 2 articles non-conformes durant une période de 5 minutes, étant donné qu'au plus 4 articles non-conformes ont été produits durant la même période?

5.26 On se propose d'étudier la proportion θ d'articles non-conformes aux spécifications pour un lot d'articles manufacturés. On décide de prendre un échantillon avec remise $(X_1, X_2, \ldots, X_{20})$.

où

$$X_i = \begin{cases} 1 & \text{si l'article est non-conforme au i-ième tirage} \quad i=1, 2, \ldots, 20 \\ 0 & \text{autrement} \end{cases}$$

(a) On note par X le nombre d'articles non-conformes dans l'échantillon

(i) précisez la masse de probabilité $p_X(x)$

(ii) précisez la masse de probabilité $p_X(x)$ si les tirages sont effectués sans remise et la taille du lot est 1000

(b) Si les tirages sont effectués avec remise et $\theta = 0.25$

(i) calculez $P[X=10]$

(ii) calculez $P[X \geq 10]$ en employant une approximation basée sur une distribution de Poisson

(iii) calculez $P[X=10]$ en employant une approximation basée sur une distribution normale.

5.27 La finale entre deux équipes (disons A et B) se joue en un maximum de 7 matchs (sans match nul). L'équipe qui gagne 4 matchs est déclarée gagnante. On fait les hypothèses suivantes:

. les parties sont indépendantes
. la probabilité que l'équipe gagne une partie est θ
. l'équipe A est au moins aussi bonne que l'équipe B, de sorte que $0.50 \leq \theta < 1$

Soit X le nombre de parties pour déterminer le gagnant. Montrez que la masse de probabilité $p_X(x,\theta)$ de X est

$$p_X(x,\theta) = \begin{cases} \theta^4 + (1-\theta)^4 & \text{si } x=4 \\ 4\theta(1-\theta)\,[\theta^3 + (1-\theta)^3] & \text{si } x=5 \\ 10\,\theta^2(1-\theta)^2\,[\theta^2 + (1-\theta)^2] & \text{si } x=6 \\ 20\,\theta^3(1-\theta)^3 & \text{si } x=7 \end{cases}$$

5.28 Un lot de très grande taille contient une proportion θ d'articles défectueux. On tire un premier échantillon de taille n_1 et on

. accepte le lot si l'échantillon contient aucun article défectueux.

. rejette le lot si l'échantillon contient deux articles défectueux ou plus.

. tire un second échantillon de taille n_2 si le nombre d'articles défectueux dans l'échantillon de taille n_1 est 1 et on accepte le lot si le nombre total d'articles défectueux n'excède pas 2 dans l'échantillon de taille $n_1 + n_2$.

Calculez la probabilité que:

(a) le premier échantillon contient aucun article défectueux.

(b) le premier échantillon contient un article défectueux et le second échantillon contient 0 ou 1 article défectueux.

(c) le premier échantillon contient deux articles défectueux et le second contient aucun article défectueux.

(d) l'on accepte le lot.

5.29 Le nombre de pannes majeures de moteur de camions est de deux en moyenne par jour selon un processus de Poisson. Ce type de panne requiert les services d'un mécanicien durant une journée entière. Combien de mécaniciens une compagnie de location de camions doit-elle employer afin d'assurer, avec une probabilité de 0.95, qu'un mécanicien est disponible pour réparer chaque moteur en panne?

5.30 Une centrale nucléaire laisse échapper une quantité de gaz radioactif deux fois par mois en moyenne. Calculez la probabilité qu'il s'écoule au moins trois mois avant la première émission de gaz? Quel est le temps moyen avant d'observer la première émission?

5.31 Par expérience, 80% des imprimantes utilisées pour les ordinateurs personnels fonctionnent sans ajustement; les autres ont besoin d'être ajustées. Un marchand vend dix imprimantes durant un mois.

(a) Calculez la probabilité qu'au moins neuf des imprimantes vendues fonctionnent sans ajustement.

(b) Pendant cinq mois, le marchand a vendu dix imprimantes par mois. Quelle est la probabilité qu'au moins neuf imprimantes fonctionnent sans ajustement pour chacun des cinq mois?

5.32 Dans un stock de vingt microprocesseurs, trois sont égratignés et inutilisables. L'égratignure ne peut être détecter à l'oeil nu. On en choisit cinq au hasard et sans remise que l'on installe sur un appareil électronique.

(a) Trouvez la masse de probabilité du nombre X de microprocesseurs sélectionnés ayant une égratignure.

(b) Calculez E(X) et VAR(X).

(c) Calculez la probabilité de sélectionner seulement des microprocesseurs sans égratignure.

(d) Calculez la probabilité de sélectionner au moins un microprocesseur avec une égratignure.

5.33 Une centrale nucléaire perd une quantité détectable de gaz radioactif, en moyenne, deux fois par mois.

(a) Calculez la probabilité qu'il y ait au plus quatre émissions durant un mois.

(b) Quel est le nombre moyen d'émissions durant une période de trois mois?

Si on observe 12 émissions ou plus durant une période de trois mois, pensez-vous que c'est une raison suffisante de mettre en doute la moyenne de 2 par mois? Expliquez en calculant $P[Y \geq 12]$ où Y est le nombre d'émissions radioactives durant trois mois.

5.34 Lorsqu'un programme est soumis à un ordinateur central, le temps d'attente avant qu'il soit mis en mode exécution est basé sur les ressources demandées. Par expérience, on sait qu'un programme soumis est mis en exécution après une minute d'attente avec une probabilité de 0.25. Dans une journée, on soumet cinq programmes avec un intervalle de temps assez grand pour assurer l'indépendance. Soit X le nombre de programmes exécutés en dedans d'une minute.

(a) Calculez la moyenne et l'écart-type de X.

(b) Calculez la probabilité qu'aucun des programmes soit en mode exécution à l'intérieur d'une minute.

(c) Cinq programmes sont soumis deux jours consécutifs. Quelle est la probabilité qu'il n'y ait aucun programme en mode d'exécution à l'intérieur d'une minute durant ces deux jours?

5.35 Un nouveau type de frein est à l'étude et on pense que ces freins pourront durer au moins 100 000 km pour 90% des véhicules qui les utiliseront. Un laboratoire a simulé la conduite de 100 automobiles utilisant ces freins. Soit X est le nombre d'automobiles ayant besoin de changer les freins avant la durée de 100 000 km.

(a) Quelle est la distribution de X? la moyenne de X?

(b) Quelle autre distribution connue peut être employer pour approcher la distribution de X?

(c) On admet que 90% est surestimé si on doit changer les freins sur 17 automobiles ou plus avant 100 000 km. Quelle est la probabilité d'observer cet événement en admettant pour être correct le pourcentage de 90%.

5.36 Dans un jeu vidéo, le joueur tente de capturer un trésor caché derrière une porte parmi cinq. La position du trésor varie au hasard et à n'importe quel moment il peut être derrière n'importe quelle porte. Lorsque le joueur frappe à une porte le trésor lui appartient s'il est derrière la porte. Sinon, le joueur doit retourner à son point de départ et revenir frapper à une autre porte en passant par un dangereux labyrinthe. La partie est terminée lorsque le joueur s'approprie le trésor. Soit X le nombre d'essais que doit faire le joueur pour s'approprier le trésor.

(a) Quelle est la distribution de X?

(b) Calculez $P[X \leq 3]$ et $P[X > 5]$.

(c) Calculez le nombre moyen d'essais nécessaire pour capturer le trésor?

5.37 Un fabricant de progiciels pour micro-ordinateurs offre à ses clients un service de consultation téléphonique. Le service est disponible de 9h00 à 17h00, les jours ouvrables. On sait par expérience que le nombre d'appels reçus chaque jour est distribué selon une loi de Poisson de moyenne 56. Calculez la probabilité:

(a) que le premier appel soit reçu avant 9h15.

(b) qu'au moins un appel soit reçu après 16h00.

(c) de recevoir deux appels entre 13h00 et 13h05.

(d) de recevoir pas plus de cinq appels durant chaque période d'une heure commençant à 9h00.

5.9 RÉPONSES EXERCICES

5.1 0.784

5.2 (a) 0.9838 (b) 23.14 (c) 0.9682

5.3 n = 8

5.4 (a) 0.804 (b) 0.174 (c) 0.370 (d) 4

5.5 (b) 0.0325 (c) 486,50 $

5.7 (a) 0.9185 (b) 0.5956 (c) 4

5.8 (a) 0.219 (b) 0.913 (c) 0.502

5.9 (a) 0.95 (b) 0.736

5.10 (a) Bin $(n,1-\theta)$ (c) 0.314, 0.166, 0.004

5.11 (a) 0.368 (b) 0.919 (c) 0.002 (d) 0.06

5.12 (a) 0.135 (b) 0.215

5.13 (a) 0.368 (b) 0.544

5.14 (c) t = 2.5

5.15 (a) Bin $(n_0+c, \theta = 0.96)$ (b) Bin $(n_0+c, \theta = 0.04)$
(d) c = 2

5.16 (a) 0.02 (b) 0.825 (c) 0.187

5.17 0.217 , 5.18 , c = 31

5.19 (a) 0.0067 (b) 0.149 (c) 0.135

5.20 (a) 959,85 $ (b) c < 32,96 $

5.21 (a) 0.857 (b) 0.053

5.22 (b) 0.9602 (c) 0.9604 (d) 0.9608

5.23 0.000045

5.24 $E(D) = D0\ [1 - e^{\lambda(e^{-\alpha}-1)}]$

5.25 (b) 0.301 (d) 4ième période (e) 0.1665
(f) 0.886

5.26 (b) 0.01 , 0.032 , 0.0079

5.28 $(1-\theta)^{n_1} + n_1\theta(1-\theta)^{n_1+n_2-1} + \frac{n_1(n_1-1)}{2}\theta^2(1-\theta)^{n_1+n_2-1} + n_1 n_2 \theta^2 (1-\theta)^{n_1+n_2-2}$

5.29 5

5.30 (a) 0.0025 (b) 15 jours

5.31 (a) 0.376 (b) 0.0074

5.32 (b) 0.75, 0.50 (c) 7/11 (d) 4/11

5.33 (a) 0.947 (b) 6 (c) oui

5.34 (a) 1.25, $\sqrt{15/4}$ (b) = 0.2373 (c) 0.0563

5.35 (a) bin(n = 100, θ = 0.10) (b) Poisson (λ = 10)
(c) 0.027

5.36 (a) géométrique θ = 1/5 (b) 0.4888 (c) 5

5.37 (a) 0.8262 (b) 0.9991 (c) 0.0949 (d) 0.0906

CHAPITRE 6

DISTRIBUTIONS CONTINUES

6.0 SOMMAIRE

On définit et présente les propriétés des distributions continues utiles pour l'analyse statistique et le calcul de probabilités. Le célèbre théorème central-limite de la théorie des probabilités est énoncé et appliqué à l'échantillonnage.

6.1 DISTRIBUTION EXPONENTIELLE

Soit N_x le nombre de réalisations d'un processus de Poisson d'intensité λ au temps x. On sait que la masse de probabilité de N_x est

$$p_{Nx}(n;\lambda) = (\lambda x)^n \exp(-\lambda x)/n! \ , \ n = 0,1,2,... \quad \lambda > 0$$

Soit X le temps de la première réalisation du processus. On a

$$P[X > x] = P[N_x = 0] = \exp(-\lambda x)$$

La fonction de répartition de X est

$$F_X(x; \lambda) = P[X \leq x] = 1 - \exp(-\lambda x)$$

et sa densité

$$f_X(x; \lambda) = \lambda \exp(-\lambda x) \ , \quad x \geq 0 \ , \quad \lambda > 0 \ . \qquad (6.1)$$

Cette distribution représente aussi le temps d'attente entre deux réalisations consécutives d'un processus de Poisson. La distribution est aussi employée en théorie de la fiabilité.

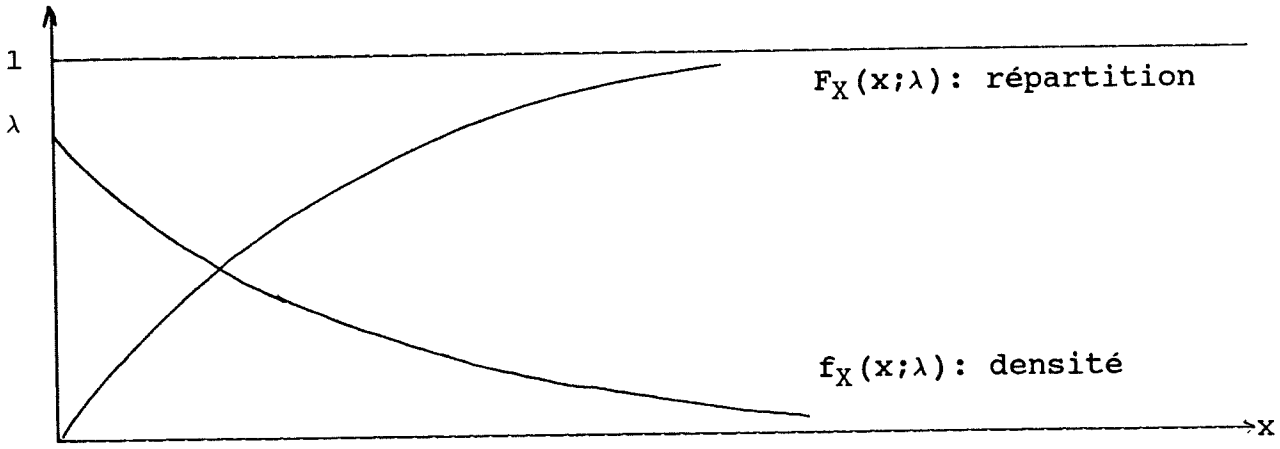


Figure 6.1: distribution exponentielle

Propriété caractéristique

La distribution exponentielle possède une propriété de non-vieillissement analogue à celle de la distribution géométrique.

$$P[X \geq x_1 + x_2 \mid X \geq x_2] = P[X \geq x_1] \quad \text{pour tout } (x_1, x_2) \qquad (6.2)$$

Cette propriété est une caractéristique unique de la distribution exponentielle parmi toutes les distributions continues définies sur les nombres réels positifs.

Moments et percentiles

$$\begin{aligned} E(X) &= 1/\lambda \\ VAR(X) &= 1/\lambda^2 \\ \beta_1 &= 2 \\ \beta_2 &= 6 \\ x_p &= (-1/\lambda)\ell n(1-p) \qquad 0 < p < 1 \end{aligned} \qquad (6.3)$$

(p-ième percentile)

Exemple 6.1

Un système de contrôle d'une centrale nucléaire utilise une génératrice d'urgence utilisant des moteurs diesels et nécessite au moins un moteur pour sa mise en marche. La durée des moteurs suit une distribution exponentielle avec une moyenne de 15 ans et chaque moteur fonctionne ou non indépendamment des autres. Déterminez combien de moteurs on devrait utiliser si on veut que la génératrice fonctionne durant les deux premières années avec une probabilité de 0.9999.

Solution:

La durée X (ans) d'un moteur suit une distribution exponentielle de moyenne $1/\lambda = 15$ c'est-à-dire avec paramètre $\lambda = 1/15$. La probabilité qu'un moteur fonctionne au moins 2 ans est

$$P[X \geq 2] = \int_2^{\infty} (1/15)\exp(-x/15)dx$$

$$= \exp(-2/15) = 0.875$$

Soit n le nombre de moteurs diesels et Y le nombre de moteurs diesels en opération après deux ans; Y suit une distribution binomiale de paramètres n et $\theta = 0.875$.

On veut trouver la plus petite valeur de n telle que

$$P[Y \geq 1] = 0.9999$$

d'où

$$P[Y = 0] = 0.0001$$

$$(0.125)^n = 0.0001$$

$$n = \log(0.0001)/\log(0.125) = 4.429$$

Il faut au moins 5 moteurs diesels pour être sûr à 0.9999 que la génératrice fonctionne durant les deux prochaines années.

Exemple 6.2: éléments de la théorie de la fiabilité

Un appareil ou système est mis en service au temps x=0 et on note par X le temps d'attente jusqu'à la première panne (bris). La probabilité que l'appareil fonctionne jusqu'au temps x est appelée la FONCTION DE FIABIITÉ (ou FIABILITÉ) et on la note par R(x). On a

$$R(x) = P[X>x] = 1-F_X(x)$$

Puisqu'aucun appareil ne peut fonctionner continuellement sans jamais tomber en panne et les propriétés de la fonction de répartition $F_X(x)$ conduisant à affirmer que:

$$R(0)=1,\ R(\infty)=0,\ R(x) \text{ non croissante}$$

Les objectifs de la théorie de la fiabilité et de ces applications sont de:

. proposer et analyser des modèles de fonctions R(x);

. développer les méthodes pour augmenter la fiabilité de systèmes en:

- réduisant la complexité des systèmes,
- augmentant la fiabilité des composants des systèmes,
- utilisant la redondance de composants placés en parallèle ou en attente,
- utilisant la maintenance curative (réparation) et préventive (remplacement).

. maximiser la fiabilité de systèmes en tenant compte de contraintes de poids, taille, coût et de temps.

On définit le TAUX DE PANNES INSTANTANÉ $\lambda(x)$ par:

$$\lambda(x) = -\frac{1}{R(x)}\frac{d}{dx}R(x)$$

et on montre que:

$$\lambda(x) = \frac{f_X(x)}{R(x)} = \frac{f_X(x)}{1-f_X(x)}$$

$\lambda(x)$ représente la probabilité conditionnelle que l'appareil tombe en panne en temps x étant donné qu'il a fonctionné jusqu'à x.

D'autre part, l'expression générale de la fiabilité en terme du taux de panne instantané est:

$$R(x) = \exp\left[-\int_0^x \lambda(t)\,dt \right]$$

Le graphique de $\lambda(x)$ en fonction du temps x s'appelle la COURBE DE MORTALITÉ et elle peut schématiquement se représenter selon la figure 6.2.

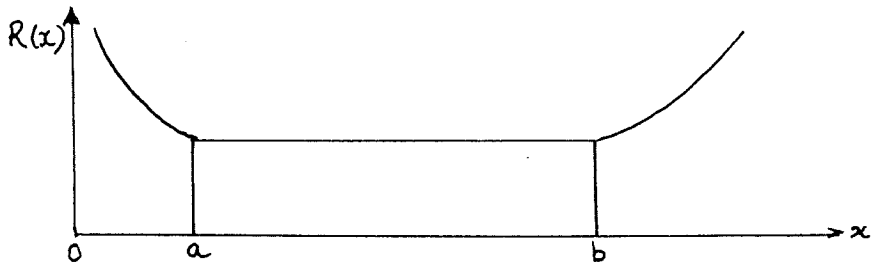


Fig. 6.2: courbe de mortalité

On distingue généralement trois périodes:

- (o,a): période de mortalité infantile avec un taux de mortalité élevé.
- (a,b): période de vie utile où le taux de mortabilité est généralement constant et où les pannes sont l'effet du hasard (pas d'usure).
- (b,∞): période d'usure où les pannes sont attribuables au vieillissement de l'appareil ou système.

On définit la MORTALITÉ $m(x)$ par

$$m(x) = -\frac{dR(x)}{dx} = \lambda(x)\,\exp\left[-\int_0^x \lambda(t)\,dt \right]$$

et elle représente la probabilité (inconditionnelle) que le composant tombe en panne au temps x. On a

$$m(x) = f_X(x) \quad \text{et} \quad \lambda(x) = \frac{m(x)}{R(x)}$$

La quantité:

$$\int_0^\infty x\, m(x)\, dx = \int_0^\infty R(x)\,dx = MTTF$$

s'appelle le temps moyen jusqu'à la première panne ("mean time to faiture") dans le cas d'un composant simple qui n'est pas réparé mais remplacé quand il tombe en panne. Dans le cas de système de k composants réparables, on définit plutôt le concept de temps moyen entre deux pannes consécutives ("mean time between faiture") noté MTBF et on le calcule par la moyenne harmonique (voir exercice 2.14) du temps moyen $MTTF_j$; de la j-ième composante:

$$MTBF = \frac{1}{\sum_{j=1}^{k} \frac{1}{MTTF_j}}$$

Cas particulier: modèle exponentiel

Durant la période de vie utile où l'on assume généralement un taux de mortalité constant, le modèle exponentiel avec sa propriété de non-vieillissement (voir (6.2)) conduit aux expressions suivantes:

$R(x) = e^{-\lambda x}$
$\lambda(x) = \lambda$
$m(n) = \lambda e^{-\lambda x}$
$MTTF = 1/\lambda$
$MTBF = 1/\lambda$ (cas d'un seul composant)

Exemple numérique

Un appareil a un taux de panne constant de 5×10^{-6} par heure.

(a) Quelle est la fiabilité de l'appareil pour une période d'opération de 100 heures?

(b) Combien, dans un lot de 10 000 appareils, seront en panne après une période d'opération de 100 heures?

(c) Quel est le temps d'attente moyen de la première panne (MTTF) ainsi que le temps moyen entre deux pannes (MTBF)?

(d) Quelle est la fiabilité de l'appareil pour x = MTBF/20, MTBF/10, MTBF/2, MTBF?

Solution

(a) En employant le modèle exponentiel, on a, selon la donnée du problème,

$$\lambda = 5*10^{-6}/h, \quad R(x) = \exp(-5*10^{-6}x)$$

$$R(100) = \exp(-5*10^{-6}*10^{2}) = \exp(-5*10^{-4}) = 0.9995$$

(b) Sur un lot de 10 000 appareils, après 100 heures d'opération, il y aura en moyenne:

$$10\ 000 * 0.005 = 5$$

appareils en panne.

(c) On considère l'appareil comme formé d'un seul composant et alors

$$MTTF = MTBF = 1/\lambda = 200\ 000 \text{ heures}$$

(d) La fiabilité de l'appareil est:

x	MTBF/20	MTBF/10	MTBF/2	MTBF
R(x)	0.99	0.90	0.61	0.37

6.2 DISTRIBUTION GAMMA

Une variable aléatoire X est distribuée selon une distribution gamma de paramètres (α,β) si sa densité f_X est de la forme:

$$f_X(x;\ \alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\ x^{\alpha-1}\exp\left[-\frac{x}{\beta}\right] \qquad x \geq 0 \quad \alpha,\beta > 0 \qquad (6.4)$$

où $$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}\ e^{-x}\ dx$$

est la fonction gamma.

Les propriétés les plus importantes de la fonction gamma $\Gamma(.)$ sont:

(a) $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)\ , \quad \alpha > 1$

(b) $\Gamma(n) = (n-1)!$ si n est un entier

(c) $\Gamma(1/2) = \sqrt{\pi}$

La figure 6.3 illustre quelques exemples de la distribution gamma.

Cas particuliers de la distribution gamma

(a) $\alpha = 1,\ \beta = 1/\lambda$: distribution exponentielle

(b) $\alpha = n/2$, n entier, $\beta = 2$: distribution khi-deux, avec n degrés de liberté

(c) $\alpha = n$, n entier : distribution Erlang

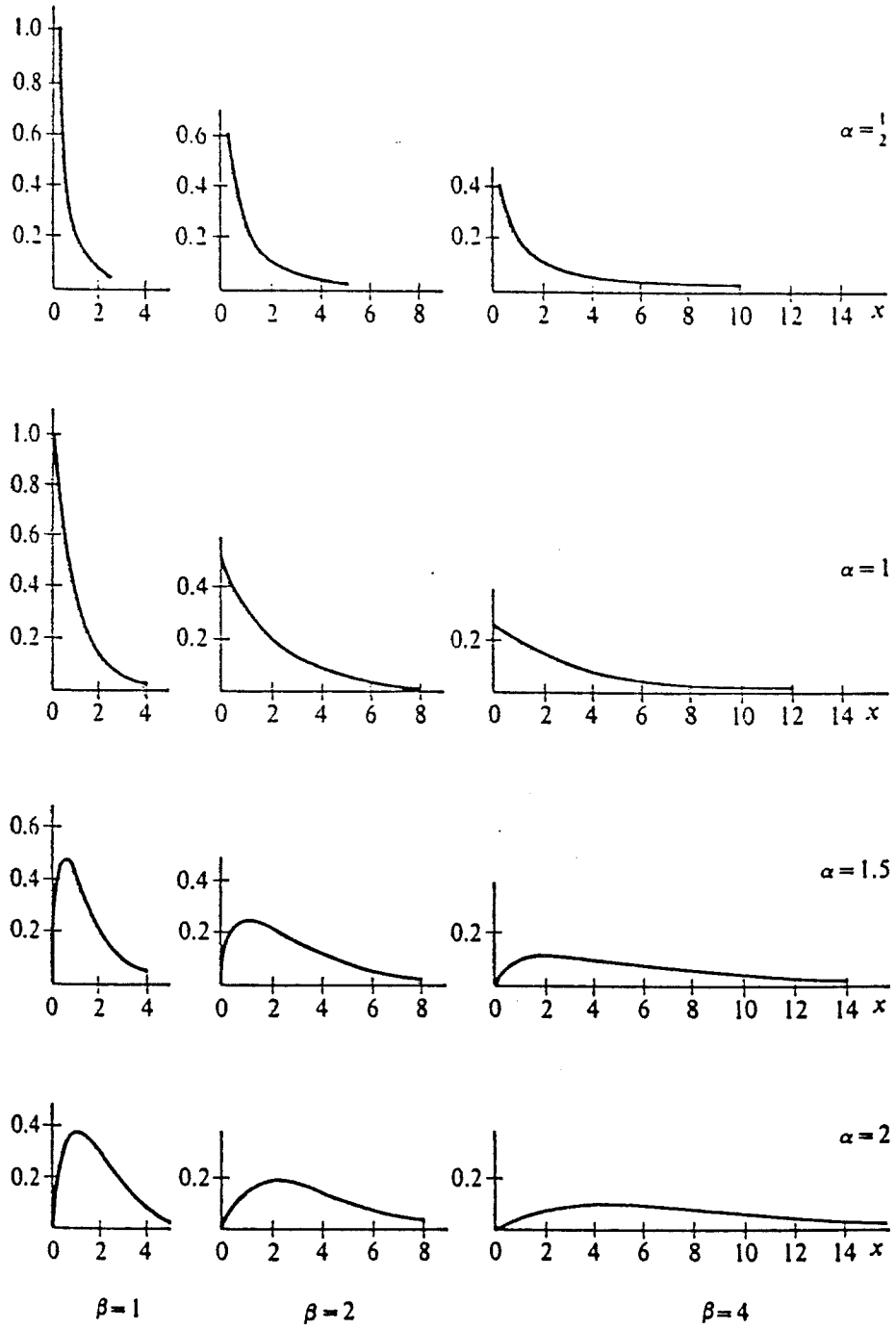


Figure 6.3: distribution gamma

Applications

(a) Soit X_k le temps de la k-ième (k = 1,2,...) réalisation d'un processus de Poisson d'intensité λ . Notons par N_x le nombre de réalisations au temps x . Alors

$$p_{Nx}(n;\lambda) = (\lambda x)^n \exp(-\lambda x)/n! \ , \quad n = 0,1,2,\ldots; \quad \lambda > 0$$

et

$$P[X_k > x] = P[N_x \leq k-1]$$

$$= \sum_{n=0}^{k-1} (\lambda x)^n \exp(-\lambda x)/n! \ = \ 1 - F_{Xk}(x;k,\lambda)$$

où $F_{Xk}(x;k,\lambda)$ désigne la fonction de répartition de X_k et $f_{Xk}(x;k,\lambda)$ sa densité. Il suit que

$$f_{Xk}(x;k,\lambda) = \frac{d}{dx} F_{Xk}(x;k,\lambda) = \frac{d}{dx} P[X_k \leq x]$$

$$= \frac{d}{dx} (1 - P[X_k > x])$$

$$= \frac{d}{dx} \left[1 - \sum_{n=0}^{k-1} (\lambda x)^n/n! \ \exp(-\lambda x)\right]$$

$$= \lambda (\lambda x)^{k-1} \exp(-\lambda x)/(k-1)!$$

Donc X_k a une distribution gamma de paramètres $\alpha = k$ et $\beta = 1/\lambda$.

(b) Soit T_k le temps entre la (k-1)-ième réalisation et la k-ième réalisation du processus de Poisson. On sait que T_k suit une distribution exponentielle de paramètre λ c'est-à-dire

$$f_{Tk}(x;\lambda) = (\lambda) \exp(-\lambda x) \ , \quad x \geq 0 \ , \quad k = 1,2,\ldots,$$

D'autre part les variables $T_1, T_2, \ldots, T_k$ sont indépendantes et le temps de la k-ième réalisation est

$$X_k = T_1 + T_2 + \ldots + T_k$$

En conséquence la somme de k variables indépendantes, de distribution expontielle avec paramètre λ est une distribution gamma de paramètres $\alpha = k$ et $\beta = 1/\lambda$.

Résumons ces résultats en une proposition.

Proposition 6.1

(a) Le temps de la k-ième réalisation X_k d'un processus de Poisson d'intensité λ est une variable gamma de paramètres $\alpha=k$ et $\beta=1/\lambda$.

(b) La somme de k variables indépendantes et de distribution exponentielle avec paramètre λ est une variable de distribution gamma avec paramètres $\alpha=k$ et $\beta=1/\lambda$.

Moments

Les principales caractéristiques numériques de la distribution gamma sont

$$\begin{aligned} E(X) &= \alpha\beta \\ VAR(X) &= \alpha\beta^2 \\ \beta_1 &= 2/\sqrt{\alpha} \\ \beta_2 &= 6/\alpha \end{aligned} \qquad (6.5)$$

Fonction de répartition

Notons par $F_X(x;\alpha,\beta)$ la fonction de répartition d'une distribution gamma de paramètres α et β. On a

$$F_X(x;\alpha,\beta) = \int_0^x f_X(t;\alpha,\beta)dt \qquad (6.6)$$

où $f_X(t;\alpha,\beta)$ est définie par l'équation (6.4)

Effectuons le changement de variable $u = t/\beta$ dans (6.6). Il vient la relation

$$F_X(x;\alpha,\beta) = F_Y(x/\beta;\alpha,1) \quad \text{où} \quad Y = X/\beta \qquad (6.7)$$

permettant d'évaluer la fonction de répartition d'une distribution gamma (α,β) à l'aide d'une distribution gamma $(\alpha,1)$. En d'autres termes, si X est distribuée gamma (α,β) alors $Y = X/\beta$ est distribuée gamma $(\alpha,1)$. Le paramètre β d'une distribution gamma est un paramètre d'échelle tandis que α est un paramètre de forme.

Cas particulier ou $\alpha = n$ = entier et $\beta = 1$

On a

$$F_X(x;n,1) = \int_0^x t^{n-1}\exp(-t)/\Gamma(n)\ dt$$

En effectuant une intégration par parties on obtient la relation de récurrence:

$$F_X(x;n,1) = -x^{n-1}\exp(-x)/\Gamma(n) + F_X(x;n-1,1) \quad , \quad n \geq 2 \qquad (6.8)$$

et pour $n = 1$ on a

$$F_X(x;1,1) = 1 - \exp(-x) \qquad (6.9)$$

Il suit de (6.8) et (6.9) une formule explicite pour $F_X(x;n,1)$

$$F_X(x;n,1) = 1 - \sum_{k=0}^{n-1} x^k\exp(-x)/k! \qquad (6.10)$$

Notation de SAS - percentiles

Le système SAS utilise la notation PROBGAM(x,α) pour désigner la fonction de répartition d'une distribution gamma (α,1) et GAMINV(p,α) pour sa réciproque.

$$\text{PROBGAM}(x,\alpha) = F_X(x;\alpha,1) \quad , \quad \alpha > 0 \quad , \quad x > 0 \qquad (6.11)$$

$$\text{GAMINV}(p,\alpha) = x_{p,\alpha} \qquad (6.12)$$

où $F_X(x_{p,\alpha};\alpha,1) = p \quad , \quad 0 < p < 1$

C'est-à-dire $x_{p,\alpha}$ est le p-ième percentile d'une distribution gamma de paramètres (α,1). Une approximation de $x_{p,\alpha}$ est donnée par

$$x_{p,\alpha} \simeq \alpha \left(1 - \frac{0.11111}{\alpha} + \frac{0.33333}{\sqrt{\alpha}} z_{1-p} \right)^3 \qquad (6.13)$$

où z_{1-p} est le p-ième percentile d'une distribution gaussienne centrée réduite.

D'autre part si $x_{p,\alpha,\beta}$ représente le p-ième percentile d'une distribution gamma (α,β) on a:

$$x_{p,\alpha,\beta} = \beta * x_{p,\alpha}$$

en employant l'équation (6.7)

Exemple 6.3

Depuis l'ouverture d'une autoroute il y a 20 ans, on a observé 40 accidents mortels. On suppose que ceux-ci suivent un processus de Poisson. Calculez la probabilité que le temps d'attente depuis maintenant jusqu'au troisième accident mortel, soit entre 15 et 20 mois.

Solution:

On peut estimer l'intensité du processus à

$$\lambda = 40/20 = 2 \quad \text{par an ou } 1/6 \text{ par mois}$$

Nous avons vu que le temps (mois) d'attente du troisième accident mortel suit une distribution gamma de paramètres $\alpha = 3$, $\beta = 1/(1/6) = 6$.

La probabilité cherchée est donc

$$\begin{aligned}
P[15 \le X \le 20] &= F_X(20;3,6) - F_X(15;3,6) \\
&= F_Y(20/6;3,1) - F_Y(15/6;3,1) \quad \text{où } Y=X/6 \\
&= [1 - \sum_{k=0}^{2} [(20/6)^k \exp(-20/6)]/(k)!] \\
&\quad - [1 - \sum_{k=0}^{2} [(15/6)^k \exp(-15/6)]/(k)!] \\
&= \sum_{k=0}^{2} [(15/6)^k \exp(-15/6)]/(k)! \\
&\quad - \sum_{k=0}^{2} [(20/6)^k \exp(-20/6)]/(k)! \\
&= \exp(-15/6)[1 + 15/6 + (1/2!)(15/6)^2] \\
&\quad - \exp(-20/6)[1 + 20/6 + (1/2!)(20/6)^2] \\
&= 0.5438 - 0.3528 \\
&= 0.1910
\end{aligned}$$

6.3 DISTRIBUTION GAUSSIENNE (NORMALE)

Une variable aléatoire X est distribuée selon une loi GAUSSIENNE (ou normale) si sa densité est de la forme:

$$f_X(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad -\infty < x < \infty \qquad (6.14)$$

où μ et σ sont deux paramètres tels que: $-\infty < \mu < \infty$, $\sigma > 0$

Il s'agit en fait d'une classe de distributions indexée par les paramètres μ et σ. Nous utiliserons la notation

$$X \sim N(\mu,\sigma^2)$$

pour désigner une variable gaussienne avec paramètres μ et σ.

L'allure de la fonction $f_X(x;\mu,\sigma)$ est représentée sur la figure 6.4.

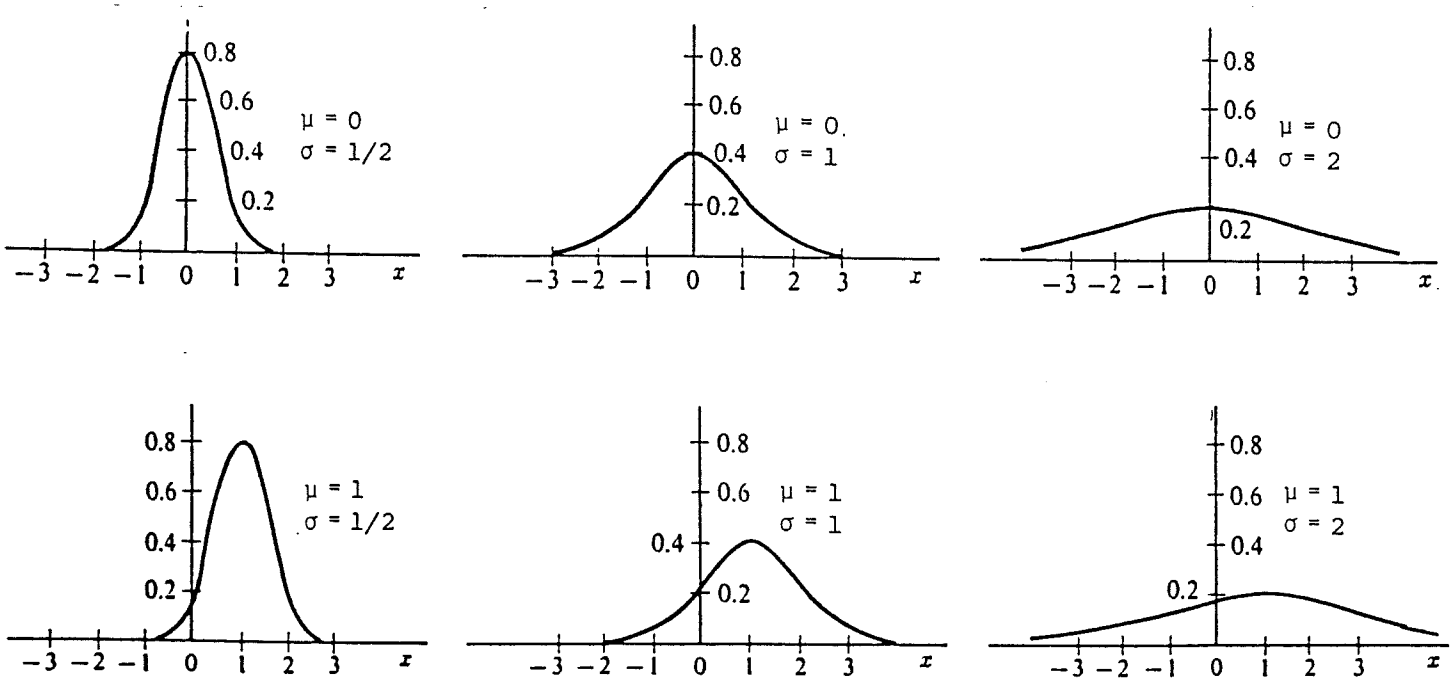


Figure 6.4: distribution gaussienne

Une analyse de la densité et un examen des graphiques, nous permet de constater que la distribution:

- est symétrique par rapport au paramètre μ et par conséquent μ est la moyenne
- est de plus en plus dispersée autour de μ au fur et à mesure que le paramètre σ augmente: en fait σ est l'écart-type de la distribution.

La distribution gaussienne est sans aucun doute, le modèle le plus utilisé au niveau de la théorie et des applications de la probabilité et de l'analyse statistique. Elle a une longue tradition historique rattachée à De Moivre (1667-1754), Laplace (1749-1827) et Gauss (1777-1855).

La distribution de très nombreux caractères numériques dans les sciences physiques, biologiques et sociales semblent suivre une distribution gaussienne. C'est en fait, dans un rôle d'approximation que la distribution gaussienne se révèle la plus intéressante et la plus utile. De nombreuses distributions mêmes discrètes peuvent être approchées par une distribution gaussienne. De plus, des sommes et moyennes de variables aléatoires sont approximativement distribuées selon le modèle gaussien comme nous le verrons lors de la présention du célèbre théorème central-limite à la section suivante.

Moments

$$E(X) = \mu$$

$$VAR(X) = \sigma^2$$

$$\beta_1 = 0$$

$$\beta_2 = 0$$

Gaussienne centrée-réduite: $\mu = 0, \sigma = 1$

La distribution gaussienne avec $\mu = 0$ et $\sigma = 1$ s'appelle la distribution gaussienne centrée-réduite. Nous désignons (sauf avis contraire) une variable ainsi distribuée par la lettre Z. La fonction de répartition correspondante sera notée par $\Phi(.)$.

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} t^2\right] dt \qquad -\infty < z < \infty \qquad (6.15)$$

Tous les calculs de probabilités et la détermination de percentiles de toutes les distributions gaussiennes utilisent la fontion $\Phi(z)$. Cette fonction a été évaluée numériquement et une table est fournie à la section 6.11. La table donne les valeurs de $\Phi(z)$ pour z entre 0.00 et 3.49 . Pour les valeurs négatives de z on utilise la relation

$$\Phi(-z) = 1 - \Phi(z) \qquad (6.16)$$

La figure 6.5 montre la fonction de densité et la fonction de répartition $\Phi(.)$ d'une gaussienne centrée-réduite.

Notation de SAS

$$\Phi(z) = \text{PROBNORM}(z)$$

$$\Phi^{-1}(p) = \text{PROBIT}(p) \qquad 0 < p < 1$$

Approximations

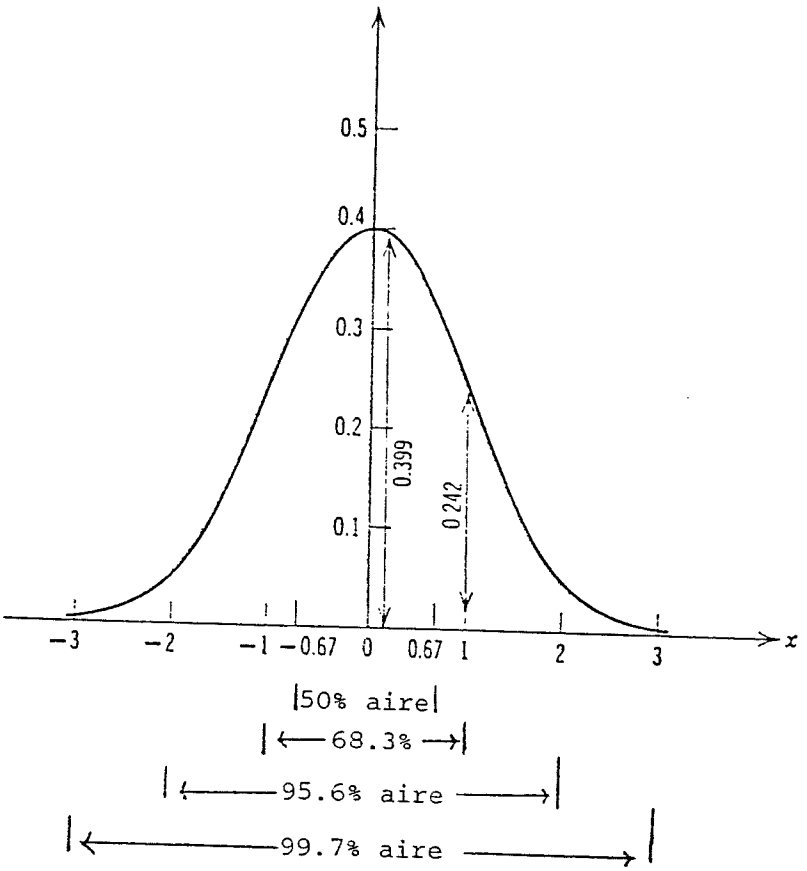
$$\Phi(z) \approx [1 + \exp(-1.5976z(1 + 0.04417z^2))]^{-1} \qquad (6.17)$$

$$\Phi^{-1}(p) \approx \left[c - \frac{2.30753 + 0.27061c}{1 + 0.99229c + 0.04481c^2} \right] \qquad (6.18)$$

où $c = (-2*\ln(1 - p))^{1/2} \qquad 0.5 \le p < 1$

et $\Phi^{-1}(p) = -\Phi^{-1}(1-p) \qquad 0 < p \le 0.5 \qquad (6.19)$

La précision de ces approximations est excellente; l'erreur est de ± 0.0001 pour (6.17) et ± 0.003 pour (6.18)



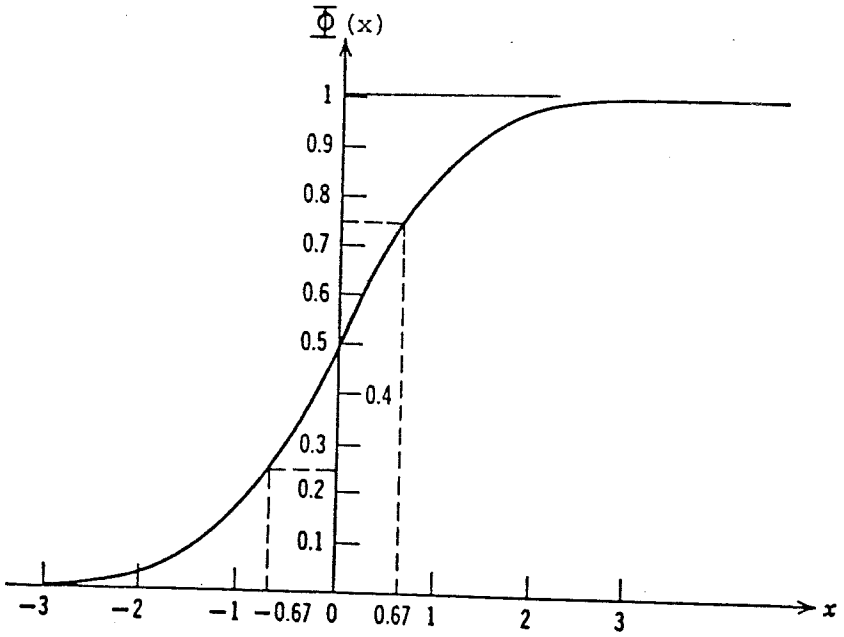
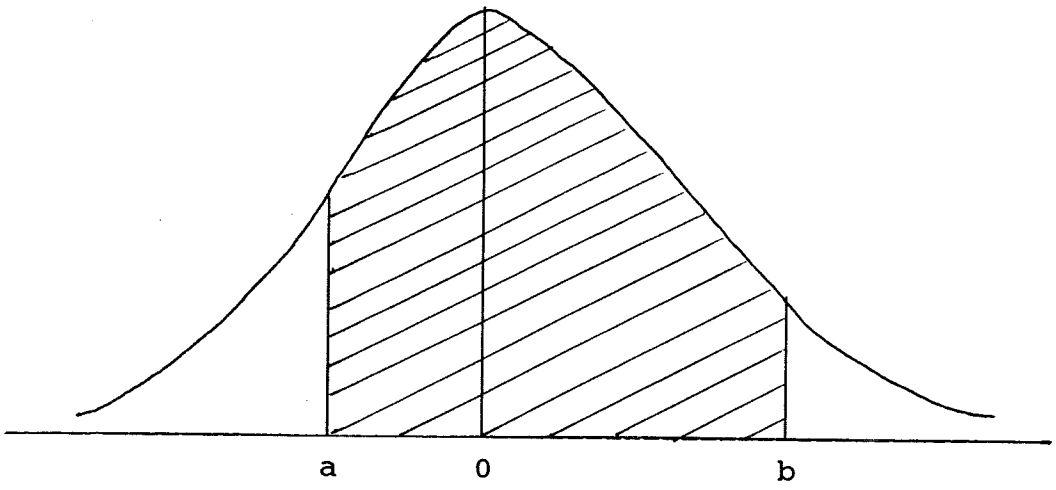


Figure 6.5: distribution gaussienne centrée-réduite

On évalue les probabilités associées à une variable gaussienne centrée-réduite en consultant la table de la fonction Φ .



$$P[a \leq Z \leq b] = \Phi(b) - \Phi(a)$$

Figure 6.6: évaluation des probabilités

Par exemple

$$P[0 \leq Z \leq 1] = \Phi(1) - \Phi(0)$$
$$= 0.8413 - 0.5000 = 0.3413$$
$$P[-2 \leq Z \leq 1] = \Phi(1) - \Phi(-2)$$
$$= \Phi(1) - (1 - \Phi(2))$$
$$= \Phi(1) + \Phi(2) - 1$$
$$= 0.8413 + 0.4772 - 1$$
$$= 0.8185$$
$$P[\,|Z| \leq 1] = P[-1 \leq Z \leq 1]$$
$$= \Phi(1) - \Phi(-1)$$
$$= \Phi(1) - (1 - \Phi(1))$$
$$= 2*\Phi(1) - 1$$
$$= 2*\ 0.8413 - 1$$
$$= 0.6826$$

Percentiles d'une distribution N(0,1)

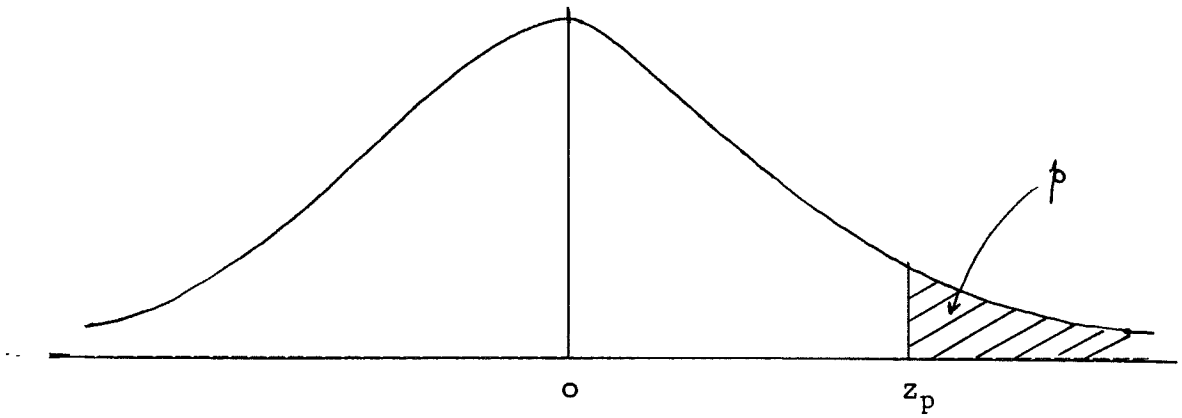


Figure 6.7: définition de z_p

On notera par z_p le (1-p)-ième percentile d'une distribution N(0, 1); z_p est la valeur telle que

$$P[Z > z_p] = p$$

$$P[Z \leq z_p] = 1 - p$$

$$\Phi(z_p) = 1 - p$$

$$z_p = \Phi^{-1}(1 - p) \qquad (6.19)$$

Remarque:

Cette notation n'est pas conforme à celle du chapitre 4 pour noter les percentiles où on utilise l'indice pour représenter la surface de gauche. Ici, l'indice p dans z_p réfère à la surface de droite. Cette dernière convention correspond à un usage très répandu.

Si $0 \leq p \leq 3.49$, on obtient de la table de la fonction Φ, les valeurs z_p correspondantes. Le tableau suivant donne une liste des percentiles d'une distribution N(0,1) pour certaines valeurs particulières de p .

p	0.50	0.25	0.10	0.05	0.025
z_p	0	0.675	1.282	1.645	1.960

p	0.01	0.005	0.001	0.0005	0.00005
z_p	2.326	2.576	3.090	3.291	3.891

Si $0.5 < p < 1$, alors $z_p < 0$ et par la symétrie de la distribution gaussienne on a:

$$z_p = - z_{1-p} \tag{6.20}$$

Cette équation est valable pour $0 < p < 1$.

L'équation (6.19) est à la base de la résolution des équations en probabilité pour une variable $Z \sim N(0,1)$:

$$P[Z \leq a] = 1 - \alpha \tag{6.21}$$

$$P[a_1 \leq Z \leq a_2] = 1 - \alpha \tag{6.22}$$

où α est fixé entre 0 et 1 et a, a_1, a_2 sont les inconnues à déterminer.

Ainsi $a = z_\alpha$ pour (6.21). D'autre part, l'équation (6.22) contient deux inconnues a_1 et a_2 et une solution consiste en:

$$P[Z \geq a_2] = \alpha/2$$

$$P[Z \leq a_1] = \alpha/2$$

Donc $a_2 = z_{\alpha/2}$ et $a_1 = z_{1-(\alpha/2)} = -z_{\alpha/2}$

Par exemple

$$P[a_1 \leq Z \leq a_2] = 0.95$$

possède la solution $a_2 = z_{0.025} = 1.96$ et $a_2 = -1.96$

<u>Proposition 6.2</u>

(a) Soit $E(X) = \mu$, $VAR(X) = \sigma^2$, $Z = (X-\mu)/\sigma$

alors $E(Z) = 0$, $VAR(Z) = 1$

(b) Si $X \sim N(\mu, \sigma^2)$ alors $Z \sim N(0,1)$

(c) Si $X \sim N(\mu, \sigma^2)$ alors

$P[a \leq X \leq b] = \Phi[(b-\mu)/\sigma] - \Phi[(a-\mu)/\sigma)]$

Ce résultat (b) a déjà été démontré au chapitre 4. La transformation affine

$$X \longrightarrow (X-\mu)/\sigma = Z$$

est une opération de changement d'origine et d'échelle et la variable Z est la forme centrée-réduite obtenue de X .

Exemple 6.4: Supposons $X \sim N(\mu = 100, \sigma^2 = 64)$.

Calculez $P[92 \leq X \leq 110]$

Solution:
$$P[92 \leq X \leq 110] = \Phi[(110-100)/8] - \Phi[(92-100)/8]$$
$$= \Phi(1.25) - \Phi(-1)$$
$$= \Phi(1.25) - 1 + \Phi(1)$$
$$= 0.8944 - 1 + 0.8413$$
$$= 0.7357$$

Proposition 6.3:

Le (1-p)ième percentile, notée x_p, d'une variable X distribuée $N(\mu,\sigma^2)$ est

$$x_p = \mu + \sigma z_p \tag{6.23}$$

En effet, par définition
$$P[X \leq x_p] = 1 - p$$
$$\Phi[(x_p - \mu)/\sigma] = 1 - p$$
$$(x_p - \mu)/\sigma = \Phi^{-1}(1-p) = z_p$$
$$x_p = \mu + z_p\sigma$$

La proposition 6.3 permet de résoudre des équations de la forme:

$$P[a_1 \leq X \leq a_2] = 1 - \alpha$$

où $0 \leq \alpha \leq 1$ et $X \sim N(\mu,\sigma^2)$. Comme il existe une infinité de couples (a_1,a_2) satisfaisant cette équation, on peut obtenir une solution unique en posant

$$P[X > a_2] = P[X < a_1] = \alpha/2$$

Alors
$$a_2 = \mu + \sigma z_{\alpha/2} \quad \text{et} \quad a_1 = \mu - \sigma z_{\alpha/2}$$

On peut montrer que cette solution donne l'intervalle (a_1,a_2) le plus court.

Proposition 6.4

$$\text{Soit} \quad L = \sum_{\alpha=1}^{n} a_\alpha X_\alpha \qquad a_\alpha \in R$$

une combinaison linéaire de variables indépendantes et gaussiennes: $X_\alpha \sim N(\mu_\alpha, \sigma_\alpha^2)$ $\alpha = 1,2,\ldots,n$

Alors $L \sim N(\mu_L, \sigma_L^2)$

où
$$\mu_L = \sum_{\alpha=1}^{n} a_\alpha \mu_\alpha \tag{6.24}$$

$$\sigma_L^2 = \sum_{\alpha=1}^{n} a_\alpha^2 \sigma_\alpha^2 \tag{6.25}$$

Les formules (6.24) et (6.25) sont valides quelle que soit la distribution de X_α et ont été vues au chapitre 4.

Cas particuliers

(a) Si $X_\alpha \sim N(\mu,\sigma^2)$ et $a_\alpha = 1$, $\alpha = 1,2,\ldots,n$, alors

$$\sum_{\alpha=1}^{n} X_\alpha \sim N(n\mu, n\sigma^2) \tag{6.26}$$

(b) Si $X_\alpha \sim N(\mu,\sigma^2)$ et $a_\alpha = 1/n$, $\alpha = 1,2,\ldots,n$, alors

$$\sum_{\alpha=1}^{n} \frac{1}{n} X_\alpha = \bar{X} \sim N(\mu, \sigma^2/n) \tag{6.27}$$

Ce dernier résultat peut se mettre sous la forme:

$$\frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}} \sim N(0,1) \tag{6.28}$$

(c) Soient $X_\alpha \sim N(\mu_x, \sigma_x^2) \quad \alpha = 1,2,\ldots,n_1$

$Y_\alpha \sim N(\mu_y, \sigma_y^2) \quad \alpha = 1,2,\ldots,n_2$

$$\bar{X} = \frac{1}{n_1} \sum_{\alpha=1}^{n_1} X_\alpha \qquad \bar{Y} = \frac{1}{n_2} \sum_{\alpha=1}^{n_2} Y_\alpha$$

et supposons l'indépendance des variables $X_1,\ldots,X_{n1}$, $Y_1,\ldots,Y_{n2}$. Alors

$$\bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}\right)$$

c'est-à-dire

$$\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}}} \sim N(0,1) \qquad (6.29)$$

Les résultats (6.28) et (6.29) sont des distributions d'échantillonnage pour la moyenne et la différence de moyennes pour le cas de populations gaussiennes et elles seront employées aux chapitres 7 et 8.

Exemple 6.5

Une compagnie de construction sait par expérience que la durée (en jours) nécessaire pour compléter un certain type de construction suit une distribution gaussienne approximativement. Un projet de ce type est proposé et tenant compte de ses ressources humaines, la compagnie estime à une chance sur 20 la probabilité de compléter le projet en moins de 20 jours et à une chance sur 100 de compléter le projet en plus de 50 jours.

(a) Déterminez la moyenne et l'écart-type de la distribution.

(b) La densité d'une variable gaussienne est non-nulle sur $(-\infty,0)$; déterminez la probabilité correspondante et dites si le modèle gaussien est raisonable compte tenu de cette probabilité.

Solution

(a) Soit X la durée en jours. Selon la donnée du problème $X \sim N(\mu,\sigma^2)$. On a

$$P[X \leq 20] = 0.05$$

$$P[X \geq 50] = 0.01$$

Les équations peuvent s'écrire

$$\Phi[(20 - \mu)/\sigma] = 0.05$$

$$1 - \Phi[(50 - \mu)/\sigma] = 0.01$$

Donc

$$(20 - \mu)/\sigma = \Phi^{-1}(0.05) = z_{0.95} = -z_{0.05} = -1.645$$

$$(50 - \mu)/\sigma = \Phi^{-1}(0.99) = z_{0.01} = 2.33$$

$$\mu - 1.645\sigma = 20$$
$$\mu + 2.33\sigma = 50$$

On obtient $\mu = 32.42$ et $\sigma = 7.55$

(b) $P[X \leq 0] = \Phi[(0-32.42)/7.55] = \Phi(-4.29) = 0.000004$

D'un point de vue strictement logique $x>0$ et tout modèle de distribution de X devrait assigner une probabilité zéro à l'événement $X \leq 0$. La probabilité trouvée est suffisamment proche de zéro pour conclure que le modèle gaussien est raisonnable.

Exemple 6.6

Supposons que la force agissant sur une colonne d'un édifice de plusieurs étages est la somme du poids de la structure, d'une charge variable (meubles, occupation humaine) et de la force du vent. On assume que ces trois variables sont gaussiennes, indépendantes et de paramètres selon le tableau:

charge (kg)	moyenne	coefficient de variation (%)
structure	14	10
charge variable	6	20
vent	2	30

(a) Déterminez la distribution et les paramètres de la charge totale

(b) Si la force de la colonne est une variable gaussienne indépendante de la charge totale, de moyenne 50% plus grande que la charge totale moyenne et a un coefficient de variation de 5% , calculez la probabilité que la colonne s'affaisse sous la charge.

Solution Notons par

(a) S_1 la charge de la structure
S_2 la charge variable
S_3 la charge des vents
S la charge totale
F la force de la colonne.

Selon les hypothèses

$$S_1 \sim N(\mu_1 = 14 , \sigma_1^2 = (1.4)^2)$$

$$S_2 \sim N(\mu_2 = 6 , \sigma_2^2 = (1.2)^2)$$

$$S_3 \sim N(\mu_3 = 2 , \sigma_3^2 = (0.6)^2)$$

on a $S = S_1 + S_2 + S_3 \sim N(\mu_S, \sigma_S^2)$

où $\mu_S = 14 + 6 + 2 = 22$

$$\sigma_S^2 = (1.4)^2 + (1.2)^2 + (0.6)^2 = 3.76 = (1.94)^2$$

puisque les variables S_1, S_2, S_3 sont indépendantes

(b) $F \sim N(\mu_F, \sigma_F^2)$

et $\mu_F = 1.5 * 22 = 33$

$\sigma_F = 0.05 * 33 = 1.65$

On cherche $P[F < S] = P[F - S < 0]$

La variable $D = F - S$ est distribuée $N(\mu_D, \sigma_D^2)$

puisque les variables F et S sont gaussiennes et indépendantes.

$$\mu_D = \mu_F - \mu_S = 33 - 22 = 11$$

$$\sigma_D^2 = \sigma_F^2 + \sigma_S^2 = (1.65)^2 + (1.94)^2 = 6.48 = (2.15)^2$$

Donc $P[D < 0] = \Phi[(0-11)/2.55] = \Phi(-4.32) \approx 0.0000$

Exemple 6.7

Soient $X_i \sim N(\mu = 100, \sigma^2 = 225)$ $i = 1,2,\ldots,n$

Pour quelle valeur de n

$$P[|\bar{X}-100| < 5] = 0.95$$

Solution: l'écart-type de X est 15. On a

$$P[|\bar{X}-100| < 5] = P[|\bar{X}-100|/(15/\sqrt{n}) < 5/(15/\sqrt{n})]$$
$$= P[|Z| < \sqrt{n}/3]$$
$$= \Phi(\sqrt{n}/3) - \Phi(-\sqrt{n}/3)$$
$$= 2\,\Phi(\sqrt{n}/3) - 1$$

Donc $\Phi(\sqrt{n}/3) = 0.975$

$\sqrt{n}/3 = \Phi^{-1}(0.975) = 1.96$

$n = 35$ en arrondissant à l'entier supérieur.

Exemple 6.8 Soient $X_i \sim N(\mu_x = 50, \sigma_x^2 = 50)$ $i = 1,2,\ldots,25$

et $Y_i \sim N(\mu_y = 60, \sigma_y^2 = 64)$ $i = 1,\ldots,16.$

Calculez la probabilité que $|\bar{Y} - \bar{X}| > a$ pour

$a = 1, 5, 10, 15, 20$

Solution: La distribution de $\bar{Y} - \bar{X}$ est $N(\mu, \sigma^2)$

où $\mu = \mu_y - \mu_x = 60 - 50 = 10$

$$\sigma^2 = \sigma_x^2/n_1 + \sigma_y^2/n_2 = 50/25 + 64/16 = 6$$

$$\begin{aligned} P(a) &= P[|\bar{Y}-\bar{X}| > a] = 1 - P[-a < \bar{Y}-\bar{X} < a] \\ &= 1 - [\Phi((a-10)/\sqrt{6}) - \Phi((-a-10)/\sqrt{6}))] \\ &= 1 - \Phi((a-10)/\sqrt{6}) - 1 + \Phi((a+10)/\sqrt{6}) \\ &= \Phi((a+10)/\sqrt{6}) - \Phi((a-10)/\sqrt{6}) \\ &= \Phi((4.08 + a/\sqrt{6})) - \Phi((a-10)/\sqrt{6}) \\ &\approx 1 - \Phi((a-10)/\sqrt{6}) \end{aligned}$$

$$\begin{aligned} P(a=1) &= 1 - \Phi(-9/\sqrt{6}) = 0.9999 \\ P(a=5) &= 1 - \Phi(-5/\sqrt{6}) = 0.9793 \\ P(a=10) &= 1 - \Phi(0/\sqrt{6}) = 0.5000 \\ P(a-15) &= 1 - \Phi(5/\sqrt{6}) = 0.0207 \\ P(a=20) &= 1 - \Phi(10/\sqrt{6}) = 0.0000 \end{aligned}$$

6.4 THÉORÈME CENTRAL-LIMITE

La théorie des probabilités est dominée par un résultat central énonçant la convergence vers une distribution gaussienne d'une somme de variables indépendantes. Le résultat peut même s'étendre à une somme de variables non indépendantes sous réserves de conditions particulières. L'intérêt remarquable de ce résultat réside dans le fait que la tendance vers la distribution gaussienne

- ne dépend pas de la distribution des variables composant la somme
- ne nécessite pas un grand nombre de variables dans la somme: 30 variables ou moins dans certains cas sont généralement suffisantes.

L'application pratique du résultat est de permettre de faire des approximations pour le calcul des probabilités, en particulier de pouvoir remplacer des distributions de probabilités malaisées à manipuler (e.g. distribution binomiale) par une distribution gaussienne. Nous n'énoncerons pas le résultat dans toute sa généralité mais seulement pour le cas de variables indépendantes identiquement distribuées. Ce cas est suffisant pour les applications envisagées.

Proposition 6.5: Central-limite

Soient X_α $\alpha = 1,2,\ldots,n$ des variables aléatoires indépendantes de moyenne μ et de variance σ^2

$$\text{Soit } Y_n = \left(\sum_{\alpha=1}^{n} X_\alpha - n\mu \right) / \sigma\sqrt{n} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

Alors la fonction de répartition de Y_n tend vers la fonction de répartition d'une distribution gaussienne centrée-réduite:

$$\lim_{n\to\infty} F_{Y_n}(x) = \Phi(x) \qquad \text{pour tout } x \qquad (6.30)$$

La démonstration demande un développement mathématique assez long et dépasse notre intérêt immédiat.

Remarques:

. d'un point de vue pratique

$$F_{Yn}(x) \approx \Phi(x) \qquad \text{pour} \quad n \geq 30$$

. le résultat (6.27) peut se formuler de plusieurs manières équivalentes; approximativement pour $n \geq 30$

$$\sum_{\alpha=1}^{n} X_\alpha \sim N(n\mu, n\sigma^2) \tag{6.31}$$

$$\bar{X} = (1/n) \sum_{\alpha=1}^{n} X_\alpha \sim N(\mu, \sigma^2/n) \tag{6.32}$$

$$(\bar{X}-\mu)/(\sigma/\sqrt{n}) \sim N(0,1) \tag{6.33}$$

Exemple 6.9

Une calculatrice électronique d'un certain type contient quatre circuits imprimés. Soit p_i la probabilité qu'une calculatrice reçue pour réparation nécessite la réparation de i circuits où i = 1, 2, 3, 4 et $p_1 = 1/2$, $p_2 = 1/4$ et $p_3 = p_4 = 1/8$. Si un lot de 100 calculatrices est reçu pour réparation, calculez la probabilité qu'au plus 200 circuits soient brisés.

Solution:

Notons par X_α le nombre de circuits brisés sur la calculatrice $\alpha = 1, 2, \ldots, 100$. Le nombre total de circuits brisés est $\sum_{\alpha=1}^{100} X_\alpha$. On a

$$E(X_\alpha) = 1*1/2 + 2*1/4 + 3*1/8 + 4*1/8 = 15/8 = \mu$$

$$VAR(X_\alpha) = E(X_\alpha^2) - (E(X_\alpha))^2 = 71/64 = \sigma^2$$

$$E\left[\sum_{\alpha=1}^{100} X_\alpha\right] = 1500/8 = 187.5$$

$$VAR\left[\sum_{\alpha=1}^{100} X_\alpha\right] = 100*71/64 = 110.74$$

et

$$P\left[\sum_{\alpha=1}^{100} X_\alpha \leq 200\right] \approx \Phi((200-187.5)/\sqrt{110.74})$$

$$= \Phi(1.19) = 0.883$$

Application: approximation d'une distribution binomiale par une distribution gaussienne

On a vu au chapitre 4 qu'une variable binomiale X de paramètres θ et n pouvait s'exprimer à l'aide d'une somme

$$X = \sum_{\alpha=1}^{n} X_\alpha$$

où les variables X_α sont indépendantes à valeurs 0 ou 1 et ou $\theta = P[X_\alpha = 1]$. Il suit donc, en appliquant le théorème central-limite à ce cas particulier avec $\mu = \theta$ et $\sigma^2 = \theta(1-\theta)$, que

X suit approximativement distribution $N(n\theta, n\theta(1-\theta))$

ou encore

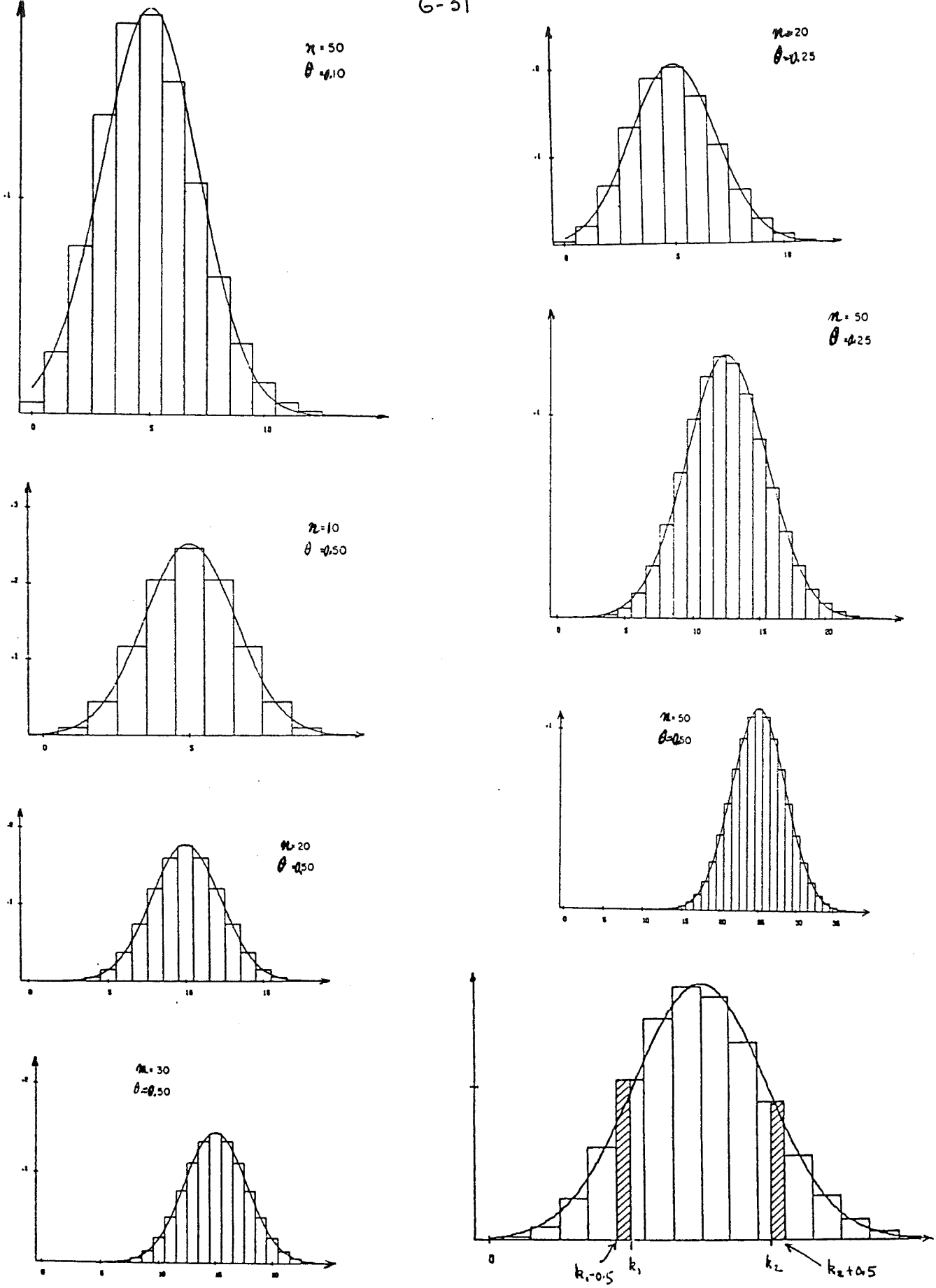
$\bar{X}$ suit approximativement distribution $N\left(\theta, \frac{\theta(1-\theta)}{n}\right)$

ou encore si $MIN(n\theta, n(1-0)) \geq 5$, alors

$$P[k_1 \leq X \leq k_2] = \Phi[(k_2+0.5-n\theta)/\sqrt{n\theta(1-\theta)}] - \Phi[(k_1-0.5-n\theta)/\sqrt{n\theta(1-\theta)}] \qquad (6.34)$$

où $0 < k_1 \leq k_2 < n$

La formule (6.34) contient un terme ± 0.5 appelé correction pour la continuité donnant une approximation plus précise. La figure 6.8 illustre la comparaison entre la distribution binomiale et la distribution gaussienne pour différentes valeurs de n et θ. On constate que la précision de l'approximation est meilleure pour des valeurs de θ voisines de 1/2 et se détériore lorsque θ s'approche de 0 ou 1 (n fixe).



correction pour la continuité

Figure 6.8: approximation d'une distribution binomiale par une distribution normale

Exemple 6.10

Un dé à jouer est lancé 6000 fois et on note à chaque fois si le résultat "1" est observé. Calculez la probabilité que le résultat "1" soit observé entre 950 et 1030 fois inclusivement.

Solution:

Soit X le nombre de fois que le "1" est observé après 6000 essais. On sait que X suit une distribution binomiale de paramètres n = 6000 et θ = 1/6 sous l'hypothèse d'un dé bien balancé. La probabilité cherchée est

$$P[950 \leq X \leq 1030] = \sum_{k=950}^{1030} \binom{6000}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{6000-k}$$

L'évaluation de cette somme est laborieuse et exige une grande précision dans les calculs. On peut l'évaluer en utilisant une approximation par une distribution gaussienne de paramètres.

$$\mu = n\theta = 6000 * 1/6 = 1000$$

$$\sigma^2 = n\theta(1-\theta) = 6000 * 1/6 * 5/6 = 833.33 = (28.87)^2$$

d'où $P[950 \leq X \leq 1030] = \Phi[(1030+0.5-1000)/28.87]$

$$- \Phi[(950-0.5-1000)/28.87]$$

$$= \Phi(1.06) - \Phi(-1.75)$$

$$= 0.8554 - 0.0401$$

$$= 0.8153$$

6.5 DISTRIBUTION KHI-DEUX

Cette distribution ainsi que les distributions de Student et Fisher qui seront vues dans les prochaines sections sont utilisées dans les procédures statistiques appliquées à des variables gaussiennes. Il importe surtout de savoir reconnaître ces distributions et utiliser les tables correspondantes.

Définition

Soient $Z_\alpha \sim N(0,1)$ $\alpha = 1,\ldots,n$. La variable

$$\chi^2_n = \sum_{\alpha=1}^{n} Z_\alpha^2$$ est appelée variable KHI-DEUX à n degrés de liberté. La densité de la variable est

$$f_{\chi^2_n}(x) = \frac{1}{2^{n/2}\,\Gamma\left(\frac{n}{2}\right)}\, x^{(n/2)-1}\, \exp(-x/2) \qquad x \geq 0 \qquad (6.35)$$

On constate que c'est un cas particulier d'une variable gamma de paramètres $\alpha = n/2$ et $\beta = 2$. La figure 6.9 illustre la fonction de densité pour plusieurs valeurs de n .

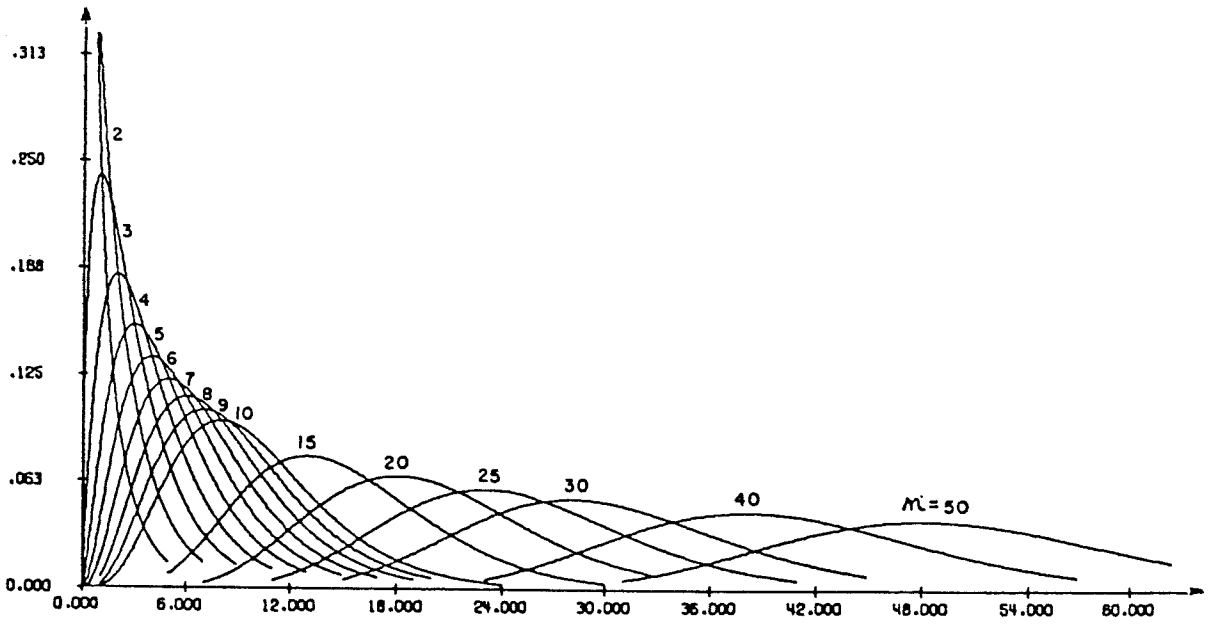


Figure 6.9: distribution khi-deux

Propriétés

$$E(\chi^2_n) = n$$
$$VAR(\chi^2_n) = 2n$$
$$\beta_1 = \sqrt{8/n} \qquad (6.36)$$
$$\beta_2 = (12/n)$$

Loi d'addition

Si χ^2_{n1} et χ^2_{n2} sont deux variables khi-deux indépendantes alors

$$\chi^2_{n1} + \chi^2_{n2} = \chi^2_{n1+n2}$$

Percentiles

On notera par $\chi^2_{n,\alpha}$ le $(1-\alpha)$-ième percentile d'une variable χ^2_n:

$$P[\chi^2_n > \chi^2_{n,\alpha}] = \alpha \qquad 0 < \alpha < 1 \qquad (6.37)$$

Une table des valeurs $\chi^2_{n,\alpha}$ est fournie à la section 6.11 pour certaines valeurs de n et α. Par exemple, pour n = 20 et $\alpha = 0.05$ on lit $\chi^2_{20,0.05} = 31.41$.

Pour une valeur de n supérieure à 100, on peut utiliser l'approximation de Wilson-Hilferty:

$$\chi^2_{n,\alpha} \approx n\left[z_\alpha\sqrt{\frac{2}{9n}} + 1 - \frac{2}{9n}\right]^3 \qquad (6.38)$$

ou encore celle de Fisher:

$$\chi^2_{n,\alpha} \approx \frac{1}{2}\left(z_\alpha + \sqrt{2n - 1}\right)^2$$

où z_α est le $(1-\alpha)$-ième percentile d'une distribution N(0,1). Par exemple pour n = 100 et $\alpha = 0.05$, la formule (6.38) donne

$$\chi^2_{100,0.05} = 100[1.645\sqrt{2/9*100} + 1 - 2/9*100]^3$$
$$= 124.34$$

correspondant à la valeur exacte fournie par la table.

Proposition 6.6

(a) Soit $X_\alpha \sim N(\mu,\sigma^2)$, $\alpha = 1,2,\ldots,n$

Alors $\sum_{\alpha=1}^{n} [(X_\alpha-\mu)/\sigma]^2$ suit une distribution χ^2_n

Le résultat est immédiat puisque $(X_\alpha-\mu)/\sigma = Z_\alpha$ suit une distribution N(0,1)

(b) Soit $X_\alpha \sim N(\mu,\sigma^2)$, $\alpha = 1,2,\ldots,n$

$$S^2 = \frac{1}{n-1} \sum_{\alpha-1}^{n} (X_\alpha-\bar{X})^2$$

Alors

$$(n-1)S^2/\sigma^2 = \sum_{\alpha=1}^{n} [(X_\alpha-\bar{X})/\sigma]^2$$ suit une distribution

khi-deux avec (n-1) degrés de liberté.

On peut écrire l'identité suivante:

$$\sum_{\alpha=1}^{n} \left(\frac{X_\alpha-\mu}{\sigma}\right)^2 = \sum_{\alpha=1}^{n} \left(\frac{X_\alpha-\bar{X}}{\sigma}\right)^2 + \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$$

D'après (a) le membre de gauche suit une distribution χ^2_n.

D'autre part on sait que $(\bar{X}-\mu)/(\sigma/\sqrt{n})$ suit une distribution N(0,1),donc son carré suit une distribution χ^2_1. En admettant l'indépendance statistique de $[(\bar{X}-\mu)/(\sigma/\sqrt{n})]^2$ et $\sum_{\alpha=1}^{n} [(X_\alpha-\bar{X})/\sigma]^2$,le résultat paraît plausible. Le résultat (b) sera employé pour étudier la distribution de $\bar{X}$ lorsque σ est inconnu et pour des questions d'inférence concernant le paramètre σ^2 aux chapitres 7 et 8.

6.6 DISTRIBUTION DE STUDENT

Student est le pseudonyme de W.S. Gosset (1876-1937).

Définition

Soit Z une variable de distribution gaussienne N(0,1) et U une variable de distribution khi-deux avec n degrés de liberté et indépendante de Z . La variable

$$T_n = Z/\sqrt{U/n}$$

est appelée une variable de Student à n degrés de liberté. La densité d'une telle variable est:

$$f_{Tn}(x) = [\Gamma(n+1/2)/\sqrt{n\pi}\Gamma(n/2)]\ (1 + x^2/n)^{-(n+1)/2} \qquad (6.39)$$

La figure 6.10 illustre la fonction pour diverses valeurs de n . Lorsque n = 1 la distribution porte le nom de CAUCHY.

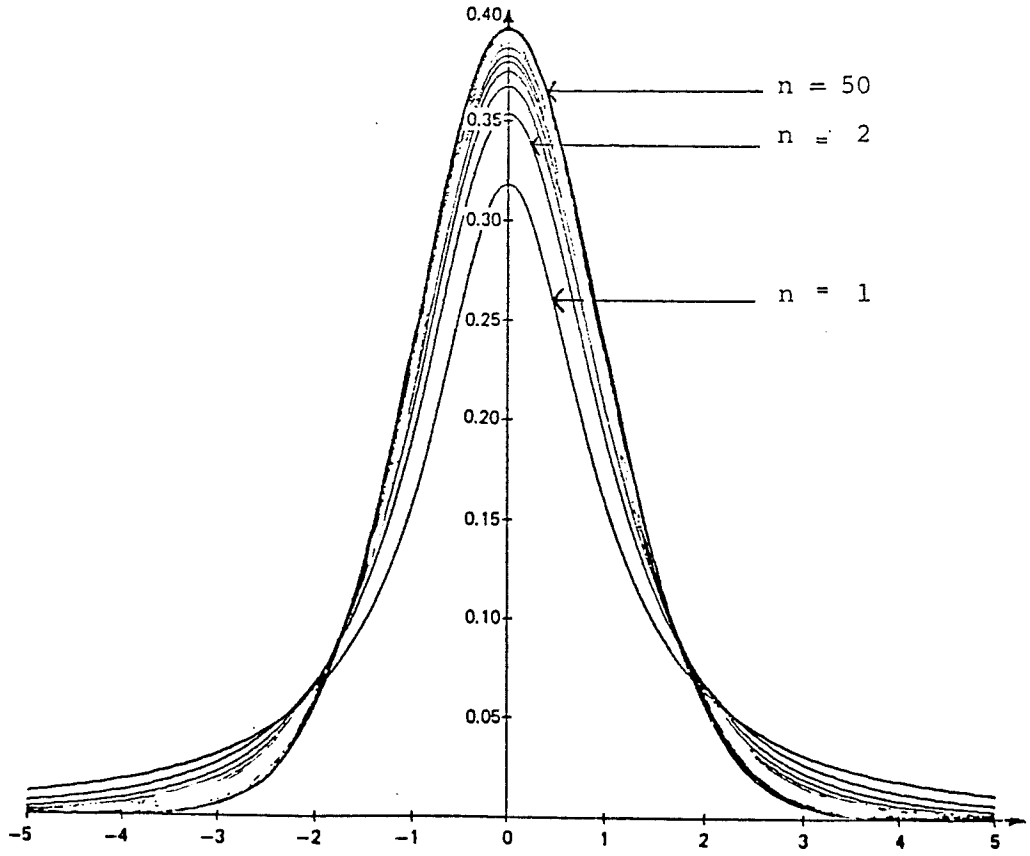


Figure 6.10: distribution de Student

Propriétés

$$\begin{aligned} E(T_n) &= 0 && n > 1 \\ VAR(T_n) &= n/n-2 && n > 2 \\ \beta_1 &= 0 && n > 3 \\ \beta_2 &= 6/(n-4) && n > 4 \end{aligned} \qquad (6.40)$$

Percentiles

On notera par $t_{n,\alpha}$ le $(1-\alpha)$-ième percentile d'une variable de Student T_n:

$$P[T_n > t_{n,\alpha}] = \alpha \qquad 0 < \alpha < 1 \qquad (6.41)$$

Une table des valeurs de $t_{n,\alpha}$ est fournie à la section 6.11 pour certaines valeurs de n et α . Par exemple pour $n = 10$ et $\alpha = 0.05$ on lit $t_{10,0.05} = 1.812$. Lorsque $n \to \infty$ la distribution converge vers une distribution gaussienne centrée-réduite

$$\lim_{n\to\infty} f_{Tn}(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} x^2\right]$$

On peut donc employer l'approximation suivante pour les percentiles d'une variable T_n:

$$t_{n,\alpha} \approx z_\alpha \qquad n \geq 30 \qquad (6.42)$$

La distribution est utile en inférence statistique impliquant la moyenne, la différence de deux moyennes et dans les modèles de régression.

Proposition 6.7

(a) Soit X_α un échantillon de taille n provenant d'une distribution $N(\mu,\sigma^2)$. Posons:

$$\bar{X} = (1/n) \sum_{\alpha=1}^{n} X_\alpha \quad \text{et} \quad S^2 = \{1/(n-1)\} \sum_{\alpha=1}^{n} (X_\alpha-\bar{X})^2$$

Alors $(\bar{X}-\mu)/(S/\sqrt{n})$ suit une distribution de Student avec (n-1) degrés de liberté.

En effet on peut écrire

$$(\bar{X}-\mu)/(S/\sqrt{n}) = [(\bar{X}-\mu)/(\sigma/\sqrt{n})]/(S/\sigma)$$

$$= [(\bar{X}-\mu)/(\sigma/\sqrt{n})]/\sqrt{[(n-1)S^2/\sigma^2]/(n-1)} = T_{n-1}$$

Puisque $(\bar{X}-\mu)/(\sigma/\sqrt{n}) \sim N(0,1)$

$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$

et en admettant l'indépendance statistique de $(\bar{X}-\mu)/(\sigma/\sqrt{n})$ et $(n-1)S^2/\sigma^2$, le résultat suit de la définition d'une variable Student.

(b) Soit $X_\alpha \sim N(\mu_x,\sigma^2)$, $\alpha = 1,2,\ldots,n_1$

$Y_\alpha \sim N(\mu_y,\sigma^2)$, $\alpha = 1,2,\ldots,n_2$

Posons

$$\bar{X} = (1/n_1) \sum_{\alpha=1}^{n_1} X_\alpha \ , \qquad \bar{Y} = (1/n_2) \sum_{\alpha=1}^{n_2} Y_\alpha$$

$$S_x^2 = \frac{1}{n_1-1} \sum_{\alpha=1}^{n_1} (X_\alpha-\bar{X})^2 \qquad S_y^2 = \frac{1}{n_2-1} \sum_{\alpha=1}^{n_2} (Y_\alpha-\bar{Y})^2$$

$$S2 = \frac{[(n_1-1)S_x^2 + (n_2-1)S_y^2]}{n_1+n_2-2}$$

Alors

$$\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (6.43)$$

suit distribution de Student avec $(n_1 + n_2 - 2)$ degrés de liberté.

La justification de ce résultat est une conséquence des énoncés suivants:

$U_1 = [(n_1-1)S_x^2]/\sigma^2$ suit une distribution χ^2_{n1-1}

$U_2 = [(n_2-1)S_y^2]/\sigma^2$ suit une distribution χ^2_{n2-1}

U_1 et U_2 sont indépendantes

$[(n_1+n_2-2)]S^2/\sigma^2 = U_1 + U_2$ suit une distribution $\chi^2_{n1+n2-2}$

$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{S\sqrt{(1/n_1 + 1/n_2)}} = \frac{(\bar{X}-\bar{Y}) - [(\mu_x-\mu_y)/\sqrt{(\sigma^2/n_1)+(\sigma^2/n_2)}]}{\sqrt{[(n_1+n_2-2)S^2/\sigma^2]/(n_1+n_2-2)}}$$

$$= Z/\sqrt{(U_1+U_2)/(n_1+n_2-2)}$$

Puisque $Z \sim N(0,1)$ et $U_1 + U_2 \sim \chi^2_{n1+n2-2}$, le résultat suit de la définition d'une variable de Student.

6.7 DISTRIBUTION F DE FISHER - SNEDECOR

R.A. Fisher (1890-1962).

Définition

Soit U_1 une variable de distribution χ^2_{n1} et U_2 une variable indépendante de distribution χ^2_{n2}. La variable

$$F_{n1,n2} = (U_1/n_1)/(U_2/n_2)$$

est appelée une variable F de Fisher-Snedecor à n_1 degrés de liberté au numérateur et n_2 degrés de liberté au dénominateur. On remarque que le carré d'une variable de Student à n degrés de liberté est un F particulier puisque

$$T_n^2 = (\chi_1^2/1)/(\chi_n^2) = F_{1,n}$$

La densité d'une variable F de Fisher est

$$f_{n1,n2}(x) = \frac{\Gamma((n_1+n_2)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)} \left[\frac{n_1}{n_2}\right]^{n1/2} \frac{x^{n1/2 - 1}}{(1+(n_1/n_2)x)^{(n1+n2)/2}} \qquad (6.44)$$

$$x > 0$$

La figure 6.11 illustre la fonction pour diverses valeurs de n_1 et n_2.

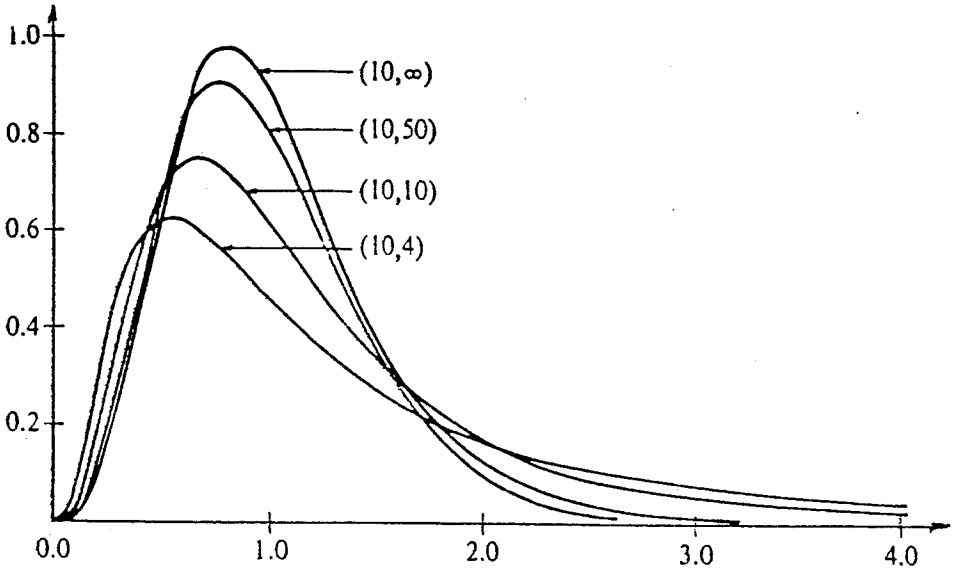


Figure 6.11: distribution de Fisher-Snedecor

Propriétés

$$E(F_{n1,n2}) = n_2/(n_2-2) \qquad n_2 > 2$$

$$VAR(F_{n1,n2}) = [2\, n_2^2(n_1+n_2-2)]/[n_1(n_2-2)^2(n_2-4)] \qquad n_2 > 4$$

$$\beta_1 = (2n_1+n_2-2)\sqrt{[8(n_2-4)/(n_1+n_2-2)n_1}\ /\ (n_2-6) \qquad n_2 > 6$$

$$\beta_2 = \frac{12((n_2-2)^2(n_2-4) + n_1(n_1+n_2-2)(5n_2-22))}{n_1(n_2-6)(n_2-8)(n_1+n_2-2)} \qquad n_2 > 8$$

Percentiles

On notera par $F_{n1,n2,\alpha}$ le $(1-\alpha)$-ième percentile d'une variable $F_{n1,n2}$ c-à-d

$$P[F_{n1,n2} > F_{n1,n2,\alpha}] = \alpha \qquad (6.45)$$

Une table des valeurs $F_{n1,n2,\alpha}$ est fournie à la section 6.11. Par exemple si $n_1 = 10$, $n_2 = 15$, $\alpha = 0.05$, on lit $F_{10,15,0.05} = 2.54$. De par la définition d'une variable F on peut établir l'équation suivante

$$F_{n1,n2,1-\alpha} = 1/F_{n2,n1,\alpha}$$

Pour des valeurs de n_1, n_2 supérieures à celles de la table on peut employer l'approximation suivante:

$$F_{n1,n2,\alpha} \approx \exp\left[1/n_2 - 1/n_1 + z_\alpha\sqrt{(2/n_1 + 2/n_2)}\ \right] \qquad (6.46)$$

où z_α est le $(1-\alpha)$-ième percentile d'une distribution N(0,1).

Proposition 6.8

Soit $X_\alpha \sim N(\mu_x, \sigma_x^2)$, $\alpha = 1,2,\ldots,n_1$

$Y_\alpha \sim N(\mu_y, \sigma_y^2)$, $\alpha = 1,2,\ldots,n_2$

et supposons l'indépendance des variables X_α, Y_α

Posons

$$\bar{X} = (1/n_1) \sum_{\alpha=1}^{n_1} X_\alpha \qquad \bar{Y} = (1/n_2) \sum_{\alpha=1}^{n_2} Y_\alpha$$

$$S_x^2 = \frac{1}{n_1-1} \sum_{\alpha=1}^{n_1} (X_\alpha - \bar{X})^2 \qquad S_y^2 = \frac{1}{n_2-1} \sum_{\alpha=1}^{n_2} (Y_\alpha - \bar{Y})^2$$

Alors $\dfrac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2}$ suit une distribution $F_{n1-1,n2-1}$.

Le résultat est une conséquence des énoncés suivants:

$U_1 = (n_1-1)S_x^2/\sigma_x^2$ suit une distribution χ^2_{n1-1}

$U_2 = (n_2-1)S_x^2/\sigma_y^2$ suit une distribution χ^2_{n2-1}

U_1 et U_2 sont indépendantes

et

$$(U_1/(n_1-1))/(U_2/(n_2-1)) = \frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} = F_{n1-1,n2-1}$$

de par la définition d'une variable F de Fisher.

Le résultat sera utilisé lors de l'inférence concernant deux variances dans les chapitres 7 et 8.

6.8 AUTRES DISTRIBUTIONS

De nombreuses distributions ont été développées pour les applications ou à la suite de développements théoriques. Nous présentons trois de ces modèles: log-normale, Weibull, bêta.

Distribution log-normale: $LN(\xi, \tau^2)$

Une variable X est distribuée selon le modèle log-normale si ℓnX suit une distribution normale (gaussienne). La densité de probabilité s'écrit:

$$f_X(x;\xi,\tau) = \frac{1}{\tau\sqrt{2\pi}} \frac{1}{x} \exp\left[-\frac{1}{2}\left(\frac{\ell nx-\xi}{\tau}\right)^2\right] \qquad (6.47)$$

$$x > 0, \quad -\infty < \xi < \infty, \quad \tau > 0$$

L'allure de la densité est illustrée à la figure (6.12) pour quelques valeurs des paramètres ξ et τ. Les principales caractéristiques de la distribution sont:

$$\mu = E(X) = \exp[\xi + \frac{1}{2}\tau^2]$$

$$\sigma = ET(X) = \mu\sqrt{e^{\tau^2}-1} = \mu\eta \qquad (6.48)$$

où $$\eta = \sqrt{e^{\tau^2}-1} = \frac{\sigma}{\mu}$$

est le coefficient de variation de X.

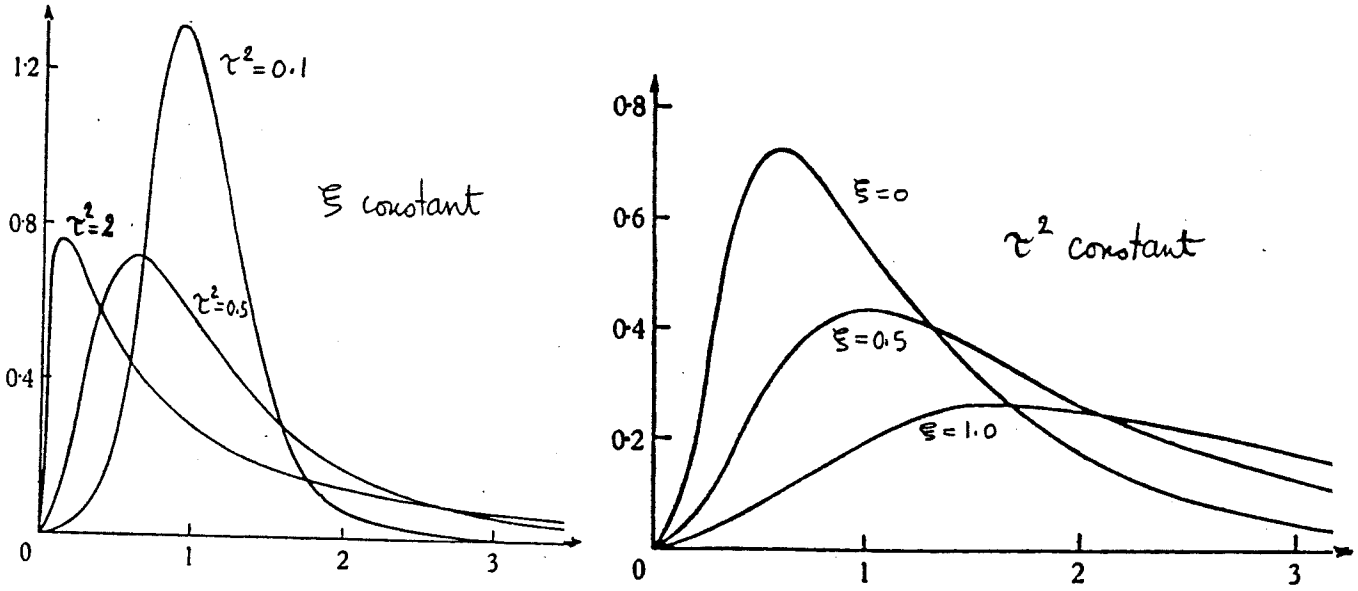


Figure 6.12: distribution log-normale

On peut déterminer ξ et τ en fonction de μ et σ

$$\tau^2 = \ell n(1+\eta^2) = \ell n\left[1+\left(\frac{\sigma}{\mu}\right)^2\right] \tag{6.49}$$

$$\xi = \ell n\ \mu - \frac{1}{2}\tau^2$$

La fonction de répartition F_X peut s'exprimer à l'aide de la fonction de répartition d'une variable gaussienne centrée-réduite

$$F_X(x) = \Phi\left[\frac{\ell nx-\xi}{\tau}\right] \tag{6.50}$$

Le $(1-\alpha)$-ième percentile x_α s'exprime en fonction du $(1-\alpha)$-ième percentile z_α d'une variable gaussienne centrée-réduite:

$$x_\alpha = \exp(\xi + z_\alpha\ \tau) \tag{6.51}$$

En particulier, la médiane est:

$$x_{0.50} = \exp(\xi)$$

La distribution log-normale est par certains aspects, un modèle beaucoup plus réaliste que la distribution normale dans un grand nombre d'applications. Très souvent les variables observées dans les applications sont de par leur nature à valeurs positives, ce qui exclut théoriquement la distribution normale comme modèle plausible. D'autre part, sous certaines conditions sur les paramètres, le modèle log-normal présente une allure semblable au modèle normal. En effet si

$$\eta = \frac{\sigma}{\mu} \leq 0.30 \quad \text{(disons)}$$

Alors

$$\tau^2 = \ell n(1+\eta^2) \approx \eta^2 = \frac{\sigma^2}{\mu^2} \tag{6.52}$$

et la distribution log-normale $LN(\xi,\tau^2)$ peut-être remplacée par une distribution normale $N(\mu,\sigma^2)$. La distribution a été employée pour étudier les précipitations, les débits de rivière, le volume de traffic aérien, la résistance de certains matériaux, la distribution de la taille de particules dans les agrégats et est un modèle pour l'étude de la distribution de produits de variables aléatoires.

Exemple 6.11

Selon des relevés hydrologiques, la précipitation annuelle sur un bassin de drainage est distribuée selon un modèle log-normal avec une moyenne de 1800 mm et un écart-type de 360 mm.

(a) Déterminez les paramètres de la distribution normale correspondante.

(b) Calculez les percentiles suivants: 50, 90, 95, 99 correspondant à des périodes de récurrences de 2, 10, 20 et 100 ans respectivement.

(c) Calculez la probabilité que la précipitation annuelle:

(i) dépasse 1000 mm

(ii) soit comprise entre 800 et 2000 mm

(d) Refaites les calculs de (b) et (c) en employant un modèle normal correspondant et comparez les résultats avec ceux du modèle normal.

Solution

(a) Notons par X la précipitation annuelle. On a

$$\mu = E(X) = 1800$$

$$\sigma = ET(X) = 360$$

$$\frac{\sigma}{\mu} = \eta = 0.20$$

Donc les paramètres de la distribution normale sont:

$$\tau^2 = \ell n(1 + \eta^2) = 0.039 = (0.198)^2 \approx (0.20)^2$$

$$\xi = \ell n\, \mu - \frac{\tau^2}{2} = \ell n(1800) - 0.02 = 7.48$$

(b) Le $(1-\alpha)$ième percentile x_α est, d'après l'équation (6.51)

$$x_\alpha = \exp(\xi + z_\alpha\, \tau) = \exp(7.48 + z_\alpha * 0.198)$$

D'où

α	0.50	0.10	0.05	0.01
x_α	1772.24	2283.4	2454.58	2808.90

(c) $P[X \geq 1000] = 1 - P[X \leq 1000]$

$$= 1 - \Phi\left[\frac{\ell n(1000) - 7.48}{0.20}\right]$$

$$= 1 - \Phi\ (2.85)$$

$$= 0.9978$$

$$P[800 \leq X \leq 1000\ ] = \Phi\left[\frac{\ell n(2000) - 7.48}{0.20}\right] - \Phi\left[\frac{\ell n(800) - 7.48}{0.20}\right]$$

$$= \Phi\ (0.60) - \Phi(-3.98) = 0.73$$

(d) Puisque $\eta = \frac{\sigma}{\mu} = 0.20 \leq 0.30$ on peut employer la distribution normale $N(\mu, \sigma^2)$ pour approcher la distribution de X.

Les percentiles avec le modèle $N(\mu = 1800,\ \sigma^2 = (360)^2)$ sont

α	0.50	0.10	0.05	0.01
x_α	1800	2260.8	2392	2637

$$P[X \geq 1000] = 1 - P[X \leq 1000] = 1 - \Phi\left[\frac{1000-1800}{360}\right]$$

$$= 1 - \Phi(-2.22) = 0.9868$$

$$P[800 \leq X \leq 2000] = \Phi\left[\frac{2000-1800}{360}\right] - \Phi\left[\frac{800-1800}{360}\right]$$

$$= \Phi\ (0.55) - \Phi(-2.78)$$

$$= 0.7088 - 0.0027$$

$$= 0.7061$$

Distribution Weibull: $W(\alpha,\beta)$

Une variable X est distribuée selon le modèle de Weibull avec paramètres (α,β) si sa fonction de densité est de la forme:

$$f_X(x;\alpha,\beta) = \frac{\beta}{\alpha}\left[\frac{x}{\alpha}\right]^{\beta-1} \exp\left[-\left[\frac{x}{\alpha}\right]^{\beta}\right] \qquad (6.53)$$

$$x > 0, \quad \alpha > 0, \quad \beta > 0$$

En fait, la variable $\left[\frac{X}{\alpha}\right]^{\beta} = Y$ est distribuée selon une distribution exponentielle

$$f_Y(y) = e^{-y} \quad , \qquad y > 0$$

et on peut donc considérer la distribution de Weibull comme une généralisation de la distribution exponentielle. La figure 6.13 illustre l'allure de la densité de Weibull pour diverses valeurs de β; β est un paramètre de forme et α est un paramètre d'échelle.

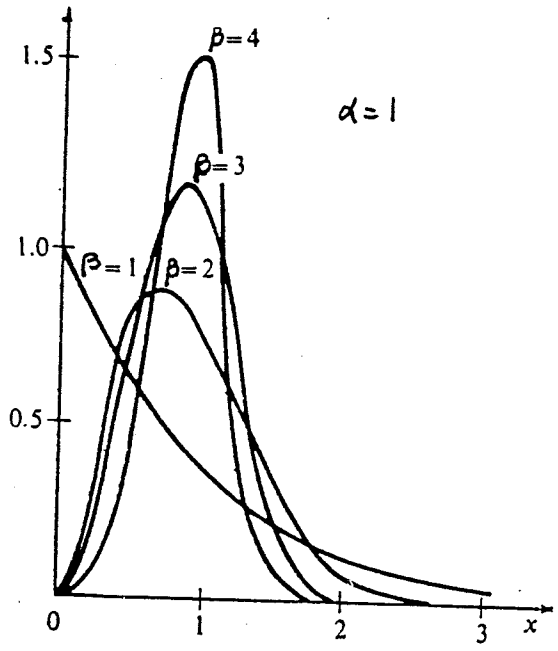


Figure 6.13: distribution de Weibull

La distribution de Weibull a été employée pour étudier la distribution de la force de rupture de matériaux, en fiabilité, en contrôle de qualité et en hydrologie pour l'étude de la distribution des valeurs extrêmes (inondations).

Les caractéristiques principales de la distribution sont:

$$\mu = E(X) = \alpha\Gamma\left[\frac{1}{\beta} + 1\right] \approx \alpha\left[1 - \frac{0.577}{\beta} + \frac{0.989}{\beta^2}\right]$$

$$\sigma = ET(X) = \alpha^2 \left[\Gamma\left[\frac{2}{\beta} + 1\right] - \Gamma^2\left[\frac{1}{\beta} + 1\right]\right] \tag{6.54}$$

$$\approx 1.2825 \frac{\alpha}{\beta}$$

Pour les valeurs de β dans le voisinage de 3.6, la distribution Weibull a une forme semblable à une distribution normale. La fonction de répartition $F_X(x;\alpha,\beta)$ est

$$F_X(x;\alpha,\beta) = 1 - \exp\left[-\left[\frac{x}{\alpha}\right]^{\beta}\right] \tag{6.55}$$

et permet d'obtenir l'expression suivante pour le $(1-\gamma)$ième percentile x_γ:

$$F_X(x_\gamma;\alpha,\beta) = 1 - \gamma$$

$$x_\gamma = \alpha(-\ell n\gamma)^{1/\beta} \tag{6.56}$$

Exemple 6.12

Une étude a montré que la distribution du débit annuel de pointe (maximal) d'une rivière est une distribution de Weibull de moyenne 1960 m^3/s et d'écart-type 1200 m^3/s.

(a) Déterminez les paramètres α et β de la distribution de Weibull.

(b) Calculez les périodes de récurrence de 2, 10, 20, 50 ans.

Solution

(a) On estime les paramètres α et β en résolvant les équations (méthode des moments)

(i) $$\alpha\left[1 + 0.577*\frac{1}{\beta} + 0.989*\frac{1}{\beta^2}\right] = 1960$$

(ii) $$1.282\ \frac{\alpha}{\beta} = 1200$$

De (ii) on a

$$\alpha = \frac{1200}{1.282}*\beta = 936.04*\beta$$

et remplaçant cette valeur dans (i) on obtient

$$936.04*\beta^2 + 540.09*\beta - 274.26 = 0$$

La seule racine positive est $\beta = 0.325$ et la valeur de α correspondante est

$$\alpha = 936.04*0.325 = 304.1$$

La densité de probabilité du débit annuel de pointe X s'écrit

$$f_X(x) = \frac{0.325}{304.1}\left(\frac{x}{304.2}\right)^{0.325-1} \exp\left[-\left(\frac{x}{304.1}\right)^{0.325}\right]$$

(b) Par définition, la période de récurrence de période r années est cette valeur x_r telle que

$$F_X(x_r) = 1 - \frac{1}{r}$$

c'est-à-dire

$$x_r = F_X^{-1}\left[1 - \frac{1}{r}\right]$$

donc

$$x_\gamma = \alpha\left[-\ell n\left[\frac{1}{r}\right]\right]^{1/\beta} \qquad \text{où } \gamma = \frac{1}{r}$$

d'après l'équation (6.56)

r	2	5	20	50	100
$\gamma = \frac{1}{r}$	0.50	0.10	0.05	0.02	0.01
x_γ	98	3959	8896	20223	33406

Distribution bêta: $Be(\alpha,\beta)$

Une variable X est distribuée selon le modèle bêta avec paramètres (α,β) si sa fonction de densité s'écrit:

$$f_X(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \qquad (6.57)$$

$$0 < x < 1 \quad , \quad \alpha < 0 \quad , \quad \beta > 0$$

L'allure de la distribution est illustrée à la figure (6.14) pour quelques valeurs de α et β.

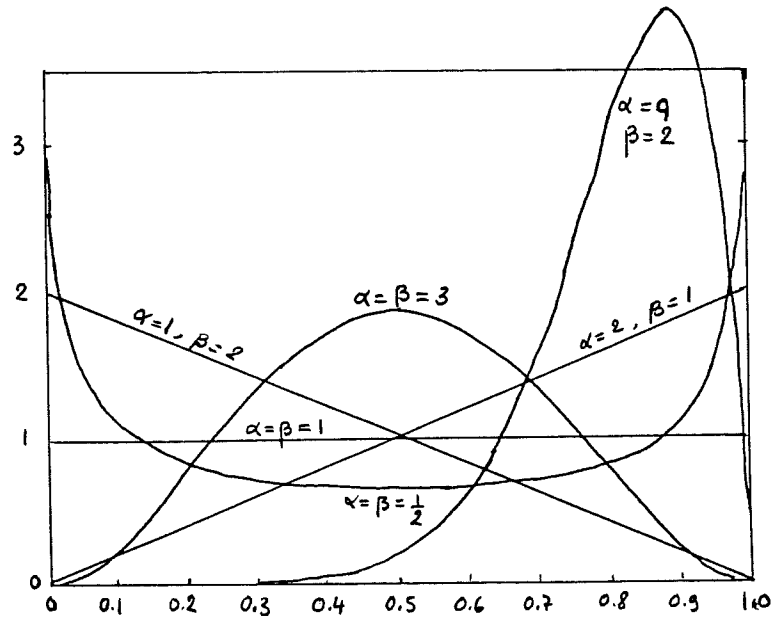


Figure 6.14: distribution bêta

Les principales caractéristiques de la distribution sont:

$$\mu = E(X) = \frac{\alpha}{\alpha+\beta}$$

$$\sigma = ET(X) = \frac{1}{(\alpha+\beta)} \sqrt{\frac{\alpha\beta}{1+\alpha+\beta}} \qquad (6.58)$$

Si $\alpha=\beta$ alors la distribution est symétrique par rapport à x=0.5. En particulier si $\alpha=\beta=1$ la distribution est appelée UNIFORME puisque la densité est constante. La distribution est surtout employée comme modèle théorique lorsqu'on étudie une variable aléatoire X dont les valeurs sont contenues dans un intervalle de nombres réels disons [a,b]. On peut alors postuler comme plausible la distribution bêta pour la variable.

$$Y = \frac{X-a}{b-a}$$

Exemple 6.13: distribution bêta

Dans une certaine ville, la proportion des rues devant être réparées chaque année suit une distribution bêta de moyenne 1/3 et de variance 2/63.

(a) Déterminez les paramètres α et β de la distribution bêta.

(b) Quelle est la probabilité que plus de cinquante pourcent des rues devront être réparées dans une année.

Solution

(a) L'équation (6.58) et les informations du problème permettent d'écrire

(i) $$\mu = \frac{\alpha}{\alpha+\beta} = \frac{1}{3}$$

(ii) $$\sigma^2 = \frac{1}{(\alpha+\beta)^2} \quad \frac{\alpha\beta}{(1+\alpha+\beta)} = \frac{2}{63}$$

De (i) on a $\beta=2\alpha$ et en remplaçant cette expression dans (ii) on obtient

$$\frac{2\,\alpha^2}{(3\alpha)^2(1+3\alpha)} = \frac{2}{63}$$

Donc $\alpha = 2$ et $\beta = 4$

et la densité de probabilité de X s'écrit:

$$f_X(x) = 20x(1-x)^3 \qquad 0 \le x \le 1$$

(b) $$P[X \ge 0.50] = \int_{0.5}^{1} f_X(x)dx$$

$$= \int_{0.5}^{1} 20x(1-x)^3 dx$$

$$= 0.1875$$

6.9 CHOIX D'UNE DISTRIBUTION

Les distributions jouent deux rôles principaux en analyse statistique: représentation et échantillonnage.

REPRÉSENTATION

En général, on utilisera une distribution contenant 1 ou 2 paramètres inconnus pour représenter la population. Dans le cas unidimensionnel d'un seul caractère étudié, deux types de distributions sont souvent postulées: Bernoulli et normale.

Bernoulli

Les éléments de la population sont classés en deux catégories exclusives et exhaustives, e.g. défectueux ou non-défectueux. La distribution correspondante est:

X	1	0
probabilité	θ	$1-\theta$

$0 < \theta < 1$

où 1 représente les éléments de la première catégorie, par exemple des articles défectueux,

0 représente les éléments de l'autre catégorie, par exemple des articles non défectueux,

θ est la proportion des éléments de la première catégorie et représente la probabilité de choisir un élément de la première catégorie:

$$P[X = 1] = \theta$$

Normale

Lorsque le caractère étudié X est continu et peut prendre des valeurs réelles on utilise souvent la distribution gaussienne

$$X \sim N(\mu, \sigma^2)$$

Autres types de populations continues

Lorsque le modèle gaussien ne constitue pas un modèle réaliste on peut utiliser d'autres distributions telle la gamma, la log-normale et la bêta. Occasionnellement, on emploie des distributions ayant 3 ou 4 paramètres.

Ces modèles sont utiles pour représenter des variables ayant un caractère d'asymétrie prononcé.

Pour guider le choix de la distribution, on peut utiliser la figure 6.15 qui situe, les densités de probabilités usuelles dans un système d'axes déterminés par les coefficients d'asymétrie (β_1) et d'aplatissement (β_2). La figure 6.16 présente un résumé des relations entre les différentes distributions.

ÉCHANTILLONNAGE

Pour étudier le comportement d'échantillonnage de quantités définies à partir des observations (variables aléatoires indépendantes) telles $\bar{X}$, S^2, $\bar{X}-\bar{Y}$, etc...
L'estimation de paramètres et les tests d'hypothèses repose sur ces distributions et le tableau ci-joint résume les principaux résultats que nous avons vu jusqu'à maintenant.

Population: Bernoulli (θ)

X = nombre d'éléments de catégorie 1 dans l'échantillon de taille n

$\bar{X}$ = X/n = proportion d'éléments de catégorie 1 dans l'échantillon de taille n

Distribution d'échantillonnage

- X suit une distribution hypergéométrique (N,D,n) si l'échantillonnage est sans remise dans une population de taille N dont $D = \theta * N$ sont de catégorie 1.
- X suit une distribution binomiale (n,θ) si l'échantillonnage est avec remise.
- $\bar{X}$ suit approximativement une distribution gaussienne:

$$N(\theta, \theta(1-\theta)/n) \quad \text{si} \quad n\theta > 5 \quad \text{et} \quad n(1-\theta) > 5$$

Population: $N(\mu, \sigma^2)$

$$\bar{X} = \frac{1}{n} \sum_{\alpha=1}^{n} X_\alpha \ , \qquad S^2 = \frac{1}{n-1} \sum_{\alpha=1}^{n} (X_\alpha - \bar{X})^2$$

$$L = \sum_{\alpha=1}^{n} a_\alpha X_\alpha \ , \qquad \mu_L = \mu \sum_{\alpha=1}^{n} a_\alpha \ , \qquad \sigma_L^2 = \sigma^2 \sum_{\alpha=1}^{n} a_\alpha^2$$

Distributions d'échantillonnage

- $\bar{X} \sim N(\mu, \sigma^2/n)$
- $L \sim N(\mu_L, \sigma_L^2)$
- $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$
- $(\bar{X}-\mu)/(S/\sqrt{n}) \sim$ Student$(n-1)$

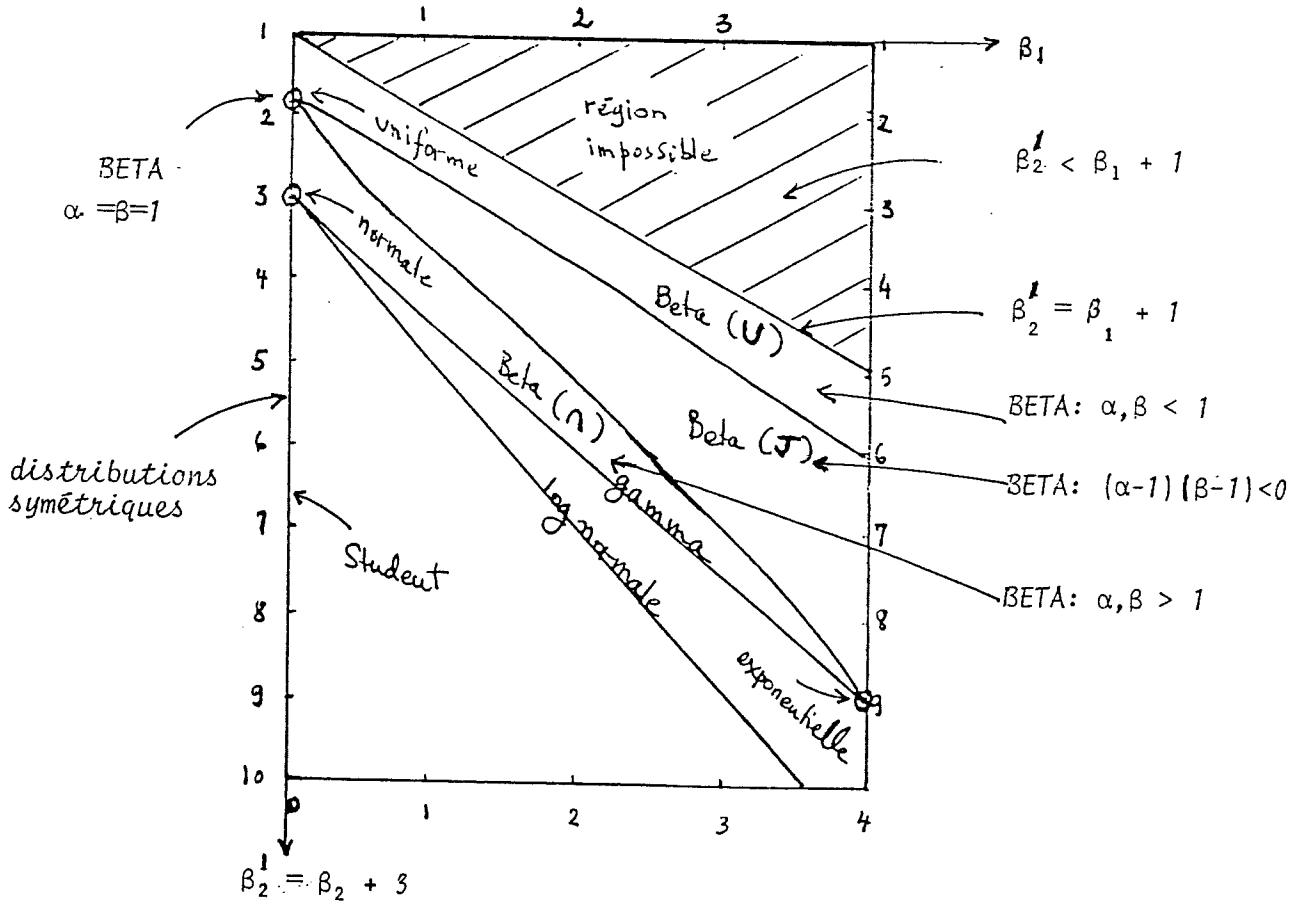


Figure 6.15: distributions selon les coefficients β_1, β_2

DISTRIBUTIONS DE PROBABILITÉS

NOM	PARA-MÈTRES	DOMAINE	MASSE/ DENSITÉ	MOYENNE	VARIANCE
BERNOULLI	$0 \le \theta \le 1$	0,1	$\theta^x(1-\theta)^{1-x}$	θ	$\theta(1-\theta)$
BINOMIALE	$n=1,2,\ldots$ $0 \le \theta \le 1$	$0,1,..,n$	$\binom{n}{x}\theta^x(1-\theta)^{n-x}$	$n\theta$	$n\theta(1-\theta)$
GÉOMÉ-TRIQUE	$0 < \theta < 1$	$1,2,\ldots$	$\theta(1-\theta)^{x-1}$	$\frac{1}{\theta}$	$\frac{1-\theta}{\theta^2}$
PASCAL	$0 < \theta < 1$ $r = 1,2,..$	$r, r+1,$ $\ldots\ldots$	$\binom{x-1}{r-1}\theta^r\ (1-\theta)^{x-r}$	$\frac{r}{\theta}$	$\frac{r(1-\theta)}{\theta^2}$
HYPERGÉO-MÉTRIQUE	$N=1,2,\ldots$ $n=1,\ldots N$ $D=1,2,\ldots N$	$0,1,\ldots,n$	$\dfrac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}$	$n\frac{D}{N}$	$n\frac{D}{N}\left(1-\frac{D}{N}\right)\left(\frac{N-n}{N-1}\right)$
POISSON	$\lambda > 0$	$0,1,2\ldots$	$\frac{\lambda^x}{x!}\exp(-\lambda)$	λ	λ
UNIFORME	α,β $\alpha < \beta$	$[\alpha,\beta]$	$\frac{1}{\beta-\alpha}$	$\frac{\alpha+\beta}{2}$	$\frac{(\beta-\alpha)^2}{12}$
EXPONEN-TIELLE	$\lambda > 0$	$[0,\infty)$	$\lambda\exp(-\lambda x)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
GAMMA	$r > 0$ $\lambda > 0$	$[0,\infty)$	$\frac{\lambda^r}{\Gamma(r)}x^{r-1}\exp(-\lambda x)$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$
WEIBULL	$\alpha>0$ $\beta>0$	$[0,\infty)$	$\frac{1}{\alpha}\beta x^{\beta-1}\exp\left[-\frac{1}{\alpha}x^{\beta}\right]$	$\alpha\Gamma\left(1+\frac{1}{\beta}\right)$	$\alpha^2\left[\Gamma\left(1+\frac{2}{\beta}\right)-\Gamma^2\left(1+\frac{1}{\beta}\right)\right]$

DISTRIBUTIONS DE PROBABILITÉS

NOM	PARA-MÈTRES	DOMAINE	MASSE/ DENSITÉ	MOYENNE	VARIANCE
GAUS-SIENNE $N(\mu,\sigma^2)$	$-\infty<\mu<\infty$ $\sigma>0$	$(-\infty,\infty)$	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$	μ	σ^2
LOG NOR-MALE $LN(\xi,\tau^2)$	$\infty<\xi<\infty$ $\tau>0$	$[0,\infty)$	$\frac{1}{\tau\sqrt{2\pi}} \frac{1}{x} \exp\left[-\frac{1}{2}\left(\frac{\ln x-\xi}{\tau}\right)\right]^2$	$\exp\left[\mu+\frac{\sigma^2}{2}\right]$	$\exp\left[2\xi+\tau^2\right]*\left[e^{\tau^2}-1\right.$
BETA $Be(\alpha,\beta)$	$\alpha>0$ $\beta>0$	$[0,1]$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(1+\alpha+\beta)}$
KHI-DEUX X^2_ν	$\nu=1,2,..$	$[0,\infty)$	$\frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} \exp[-x/2)$	ν	2ν
STUDENT T_ν	$\nu=1,2,..$	$(-\infty,\infty)$	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	0	$\frac{\nu}{\nu-2}$ $(\nu>2)$
FISHER F_{ν_1,ν_2}	$\nu_1=1,2,..$ $\nu_2=1,2,..$	$[0,\infty)$	$C_{\nu 1,\nu 2} \frac{x^{\nu 1/2-1}}{\left[1+\frac{\nu_1}{\nu_2} x\right]^{\frac{\nu_1+\nu_2}{2}}}$	$\frac{\nu_2}{\nu_2-2}$	$\frac{2\nu_2^2(\nu_1\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$

$$C_{\nu 1,\nu 2} = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}$$

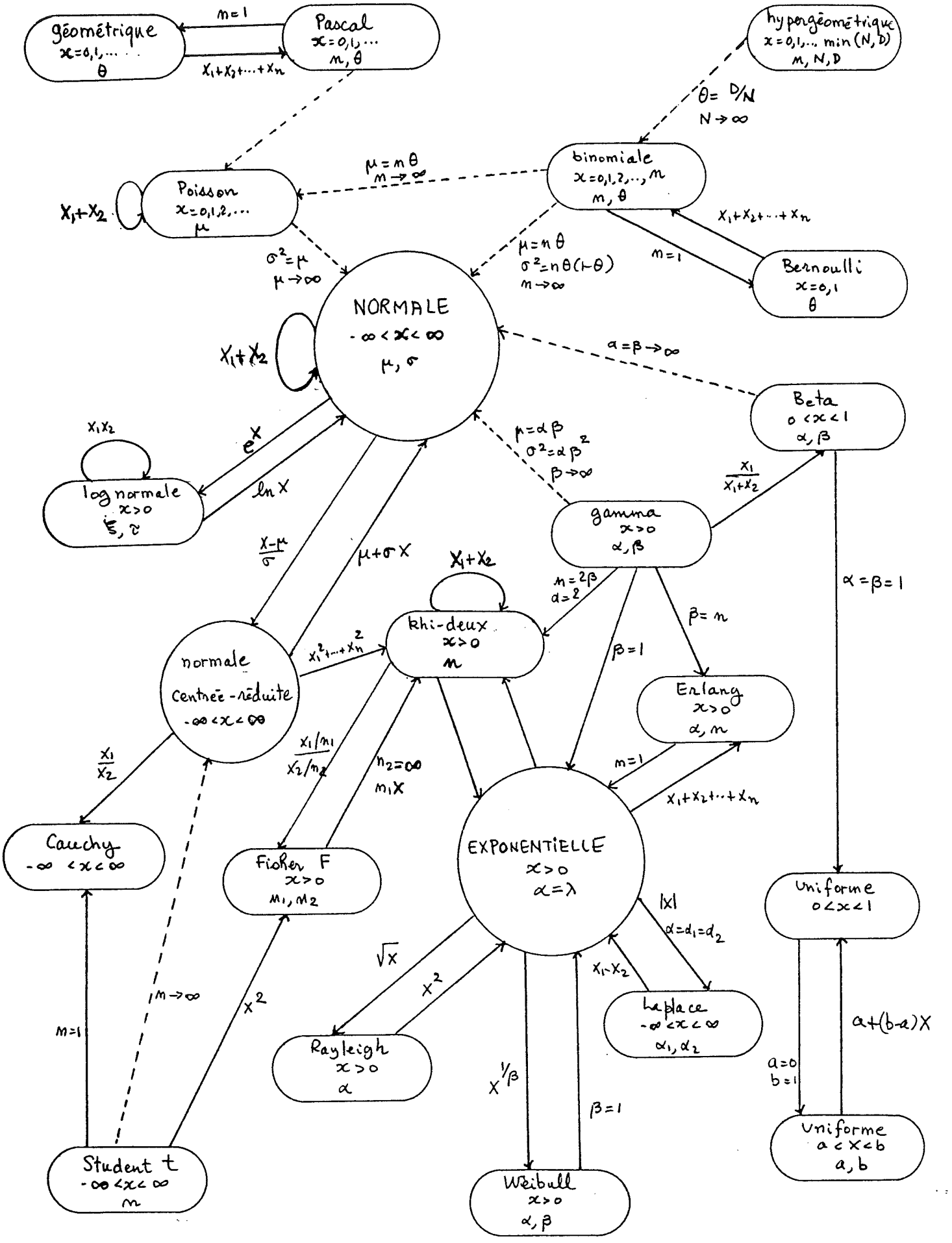


Figure 6.15 : relations entre les distributions

6.10 FONCTIONS DE PROBABILITÉS EN SAS

Le progiciel d'analyse statistique SAS dispose de plus de 130 fonctions dont les fonctions de répartition de probabilité et leurs réciproques. Le tableau donne la nomenclature adaptée par SAS pour employer ces fonctions à l'intérieur d'un programme SAS.

distribution	fonction de répartition	répartition réciproque (0<p<1)
hypergéométrique (N,D,n)	PROBHYPR(N,D,n,x)	-
binomiale (n,θ)	PROBBNML(θ,n,x)	-
Poisson(λ)	POISSON(λ,x)	-
binomiale négative (n,θ)	PROBNEGB(θ,n,x)	-
gamma (α,1)	PROBGAM(x,α)	GAMINV(p,α)
normale (0,1)	PROBNORM(x)	PROBIT(p)
bêta (α,β)	PROBBETA(x,α,β)	BETAINV(p,α,β)
khi-deux (n)	PROBCHI(x,n)	voir remarque (b)
Student (n)	PROBT(x,n)	voir remarque (c)
Fisher (n_1,n_2)	PROBF(x,n_1,n_2)	voir remarque (d)

Remarques

(a) N, D, n, θ, λ, n_1, n_2 α, β sont les paramètres usuels associés aux différentes distributions de probabilité et $0 < p < 1$

(b) $\chi^2_{\nu,\alpha} = 2*\text{GAMINV}(1-\alpha, \frac{\nu}{2})$

(c) $t_{\nu,\alpha} = \sqrt{\frac{\nu\ x_{1-2\alpha}}{1 - x_{1-2\alpha}}}$ où $x_{1-2\alpha} = \text{BETAINV}(1-2\alpha, \frac{1}{2}, \frac{\nu}{2})$

(d) $F_{\nu 1,\nu 2,\alpha} = \frac{\nu_2\ x_{1-\alpha}}{\nu_1\ (1-x_{1-\alpha})}$ où $x_{1-\alpha} = \text{BETAINV}(1-\alpha, \frac{\nu_1}{2}, \frac{\nu_2}{2})$

Le progiciel contient aussi des fonctions permettant la génération d'échantillons pseudo-aléatoires selon les distributions de probabilité:

NORMAL: gaussienne centrée-réduite

RANNOR: gaussienne centrée-réduite

RANBIN: binomiale

RANCAU: Cauchy

RANEXP: exponentielle

RANGAM: gamma

RANTBL: tableau de probabilités spécifiées.

On consultera le manuel de base de SAS, SAS User's Guide: Basics Version 5, sur la procédure d'appel de ces fonctions génératrices d'échantillons.

6.11 EXEMPLES D'UTILISATION DE SAS

```
++++++++++++++++++++++++++++++++++++++++++++++++
+  EXEMPLE  : GENERATION DE DONNEES PROVENANT  +
+             D'UNE DISTRIBUTION NORMALE       +
+  FONCTION : RANNOR                           +
+  PROCEDURE: MEANS, CHART                     +
++++++++++++++++++++++++++++++++++++++++++++++++
```

```
TITLE1 'ECHANTILLON DE 1000 OBSERVATIONS D''UNE POPULATION';
TITLE2 'NORMALE DE MOYENNE 100 ET D''ECART-TYPE 15';

DATA NORMAL;
      DO N=1 TO 1000;
         X=100 + 15*RANNOR(53429);
         OUTPUT;
      END;
PROC MEANS DATA=NORMAL;
     VAR X;
PROC CHART DATA=NORMAL;
     VBAR X / MIDPOINTS=55 TO 145 BY 3;
```

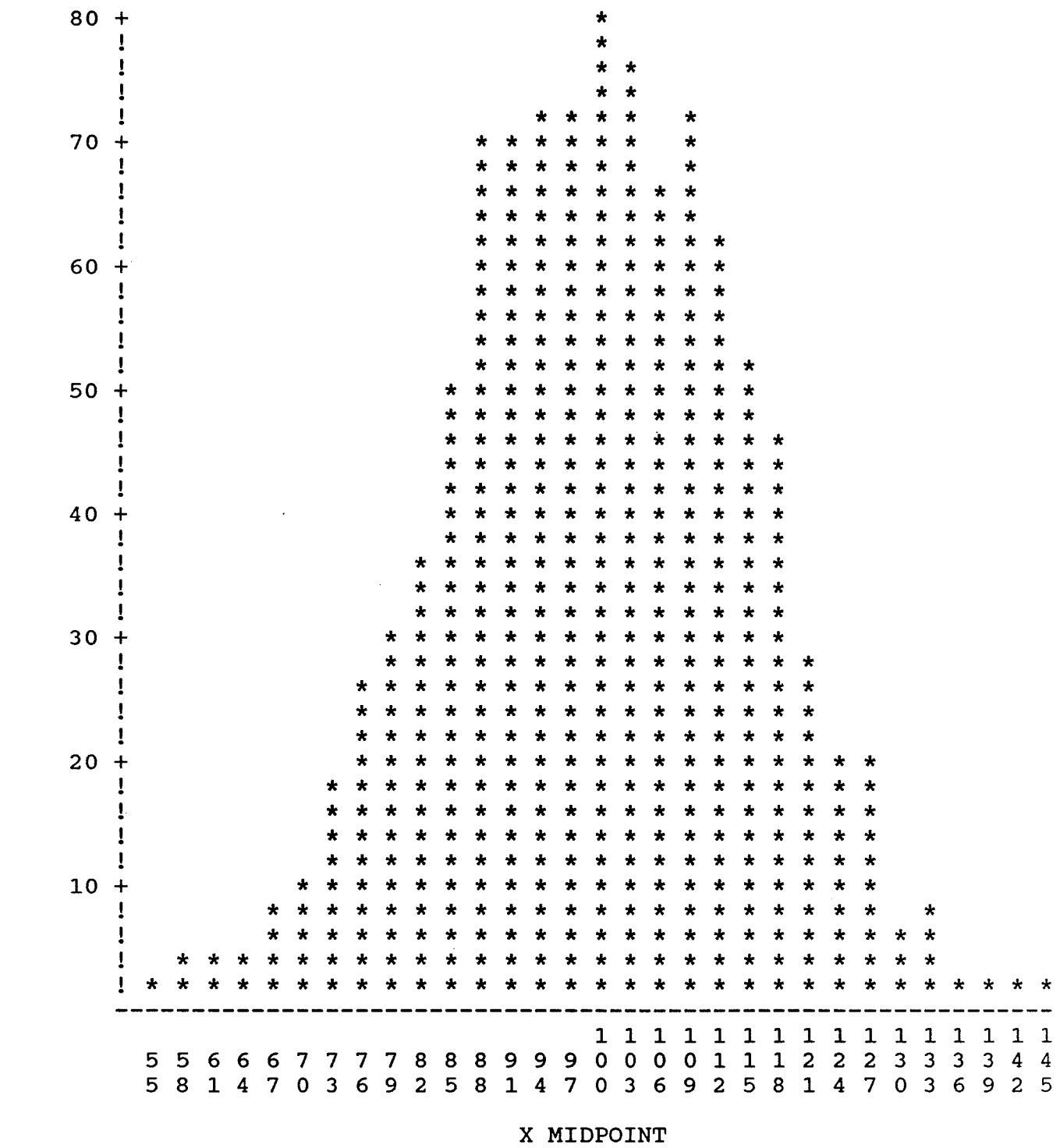
```
                        ++++++++++++++++++++++++
                        + OUTPUT DE PROC MEANS +
                        ++++++++++++++++++++++++
```

ECHANTILLON DE 1000 OBSERVATIONS D'UNE POPULATION
NORMALE DE MOYENNE 100 ET D'ECART-TYPE 15

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE
X	1000	99.61597340	14.90927629	38.41557776	148.2110117

++++++++++++++++++++++++
+ OUTPUT DE PROC CHART +
++++++++++++++++++++++++

FREQUENCY BAR CHART

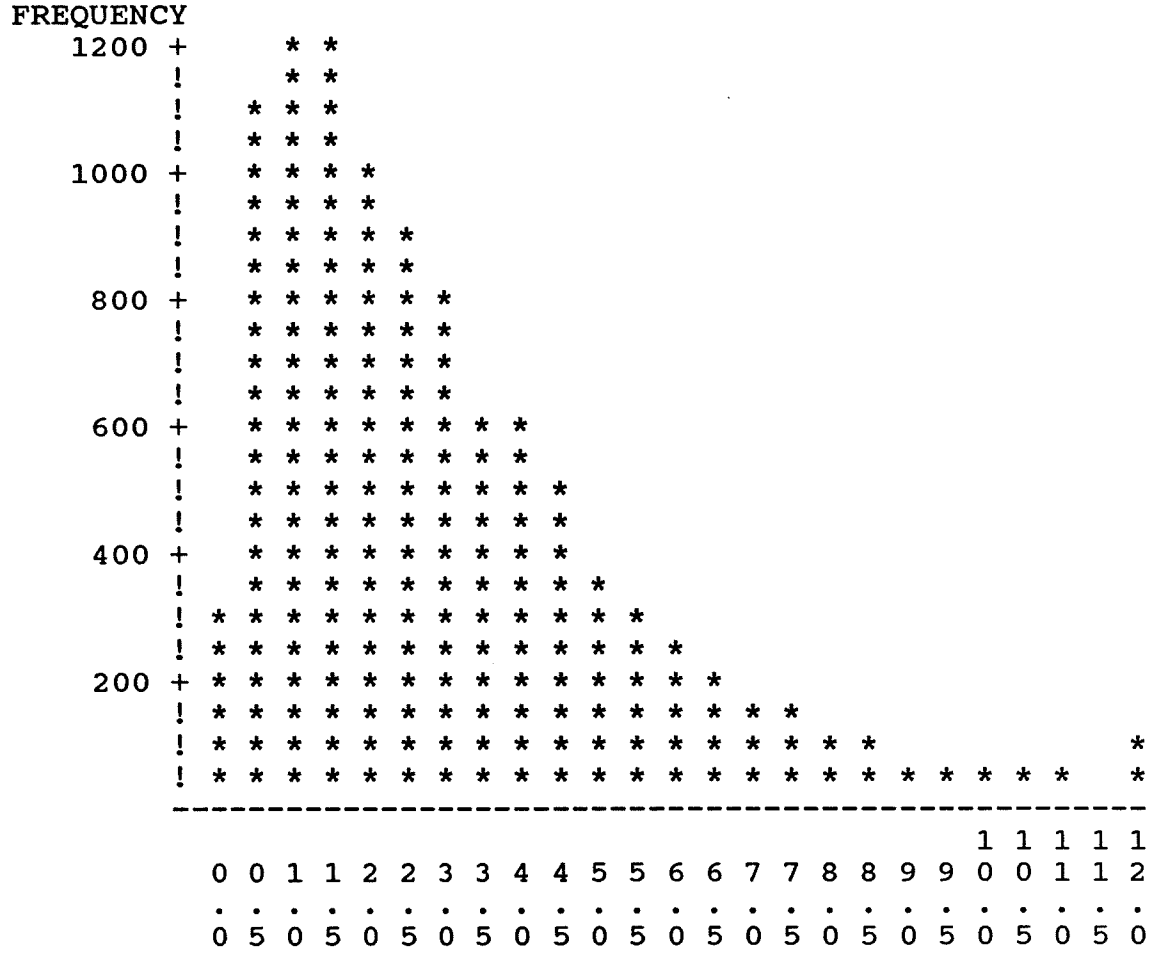


```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
+  EXEMPLE :  ILLUSTRATION DU THEOREME LIMITE-CENTRALE:  +
+             1000 ECHANTILLONS DE TAILLE N=10 D'UNE     +
+             DISTRIBUTION GAMMA (1.5, 2)                +
+  FONCTION : RANGAM                                     +
+  PROCEDURE: MEANS, CHART                               +
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

TITLE1 'ILLUSTRATION DU THEOREME CENTRAL-LIMITE';
TITLE2 '1000 ECHANTILLONS DE TAILLE N=10';
TITLE3 'POPULATION GAMMA DE PARAMETRES';
TITLE4 'ALPHA=1.5 ET BETA=2';

DATA GAMMA;
     DO N=1 TO 10000;
        X=2*RANGAM(322145,1.5);
        OUTPUT;
     END;
PROC CHART DATA=GAMMA;
     VBAR X / MIDPOINTS=0 TO 12.0 BY 0.5;
     HBAR X / MIDPOINTS=0 TO 12.0 BY 0.5;

DATA ECHANT10;
     DO ECHANT=1 TO 1000;
        DO N=1 TO 10;
           X=2*RANGAM(322145,1.5);
           OUTPUT;
        END;
     END;

PROC MEANS DATA=ECHANT10 NOPRINT;
     OUTPUT OUT=MOY10 MEAN=MOY;
     VAR X;
     BY ECHANT;

PROC CHART DATA=MOY10;
     HBAR MOY / MIDPOINTS=0.75 TO 5.95 BY 0.2;

PROC MEANS DATA=MOY10;
     VAR MOY;
```

```
++++++++++++++++++++++++
+ OUTPUT DU PROGRAMME +
++++++++++++++++++++++++
```

ILLUSTRATION DU THEOREME CENTRAL-LIMITE
1000 ECHANTILLONS DE TAILLE N=10
POPULATION GAMMA DE PARAMETRES
ALPHA=1.5 ET BETA=2

FREQUENCY BAR CHART



FREQUENCY BAR CHART

MIDPOINT X		FREQ	CUM. FREQ	PERCENT	CUM. PERCENT
	!				
0.0	!******	295	295	2.95	2.95
0.5	!**********	1077	1372	10.77	13.72
1.0	!**********	1213	2585	12.13	25.85
1.5	!**********	1183	3768	11.83	37.68
2.0	!**********	987	4755	9.87	47.55
2.5	!**********	908	5663	9.08	56.63
3.0	!**********	780	6443	7.80	64.43
3.5	!**********	592	7035	5.92	70.35
4.0	!**********	587	7622	5.87	76.22
4.5	!**********	476	8098	4.76	80.98
5.0	!*******	357	8455	3.57	84.55
5.5	!******	321	8776	3.21	87.76
6.0	!*****	236	9012	2.36	90.12
6.5	!****	197	9209	1.97	92.09
7.0	!***	156	9365	1.56	93.65
7.5	!***	138	9503	1.38	95.03
8.0	!**	92	9595	0.92	95.95
8.5	!**	78	9673	0.78	96.73
9.0	!*	65	9738	0.65	97.38
9.5	!*	54	9792	0.54	97.92
10.0	!*	42	9834	0.42	98.34
10.5	!*	39	9873	0.39	98.73
11.0	!*	25	9898	0.25	98.98
11.5	!	19	9917	0.19	99.17
12.0	!**	83	10000	0.83	100.00
	--------+-------+--------+				
	400 800 1200				

FREQUENCY

FREQUENCY BAR CHART

MIDPOINT MOY		FREQ	CUM. FREQ	PERCENT	CUM. PERCENT
	!				
0.75	!	0	0	0.00	0.00
0.95	!	1	1	0.10	0.10
1.15	!	2	3	0.20	0.30
1.35	!	2	5	0.20	0.50
1.55	!****	21	26	2.10	2.60
1.75	!*****	27	53	2.70	5.30
1.95	!*********	44	97	4.40	9.70
2.15	!**********	62	159	6.20	15.90
2.35	!******************	92	251	9.20	25.10
2.55	!********************	101	352	10.10	35.20
2.75	!********************	98	450	9.80	45.00
2.95	!**********************	108	558	10.80	55.80
3.15	!******************	92	650	9.20	65.00
3.35	!****************	78	728	7.80	72.80
3.55	!****************	78	806	7.80	80.60
3.75	!**********	51	857	5.10	85.70
3.95	!*********	43	900	4.30	90.00
4.15	!********	41	941	4.10	94.10
4.35	!****	21	962	2.10	96.20
4.55	!***	14	976	1.40	97.60
4.75	!**	11	987	1.10	98.70
4.95	!*	4	991	0.40	99.10
5.15	!*	4	995	0.40	99.50
5.35	!*	4	999	0.40	99.90
5.55	!	0	999	0.00	99.90
5.75	!	1	1000	0.10	100.00
5.95	!	0	1000	0.00	100.00

```
    ----+---+---+---+---+--
       20  40  60  80  100
```

FREQUENCY

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE
MOY	1000	2.99882197	0.76300121	1.03687094	5.68460329

6.12 TABLES

L'utilisation fréquente des distributions: gaussienne centrée-réduite, Student, Khi-deux et Fisher demande l'accès à une tabulation car leurs fonctions de répartitions sont complexes à évaluer. Les tables suivantes ont été retenues:

(a) Table 6.1 de la fonction de répartition $\Phi(z)$ d'une variable gaussienne (normale) centrée-réduite

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} t^2\right] dt \qquad o \leq z \leq 3.49$$

(b) Table 6.2 du $(1-\alpha)$ième percentile $t_{\nu,\alpha}$ de la distribution de Student avec ν degrés de liberté pour les valeurs suivantes de (ν,α):

ν = 1(1)30(5)50(10)100(20)200,∞

α = 0.45, 0.35, 0.25, 0.15, 0.10, 0.05, 0.025, 0.01, 0.005, 0.0005

(c) Table 6.3 du $(1-\alpha)$ième percentile $\chi^2_{\nu,\alpha}$ de la distribution khi-deux avec ν degrés de liberté pour les valeurs suivantes de (ν,α):

ν = 1(1)30(5)50(10)100(20)200

α = 0.995, 0.99, 0.975, 0.95, 0.90, 0.80, 0.70, 0.60, 0.50, 0.40, 0.30, 0.20, 0.10, 0.05, 0.025, 0.01, 0.005

(d) Table 6.4 du $(1-\alpha)$ième percentile $F_{\nu 1,\nu 2,\alpha}$ de la distribution de Fisher avec

$\nu 1$ = degrés de liberté au numérateur et
$\nu 2$ = degrés de liberté au dénominateur

Les valeurs de $(\nu 1,\nu 2,\alpha)$ sont:

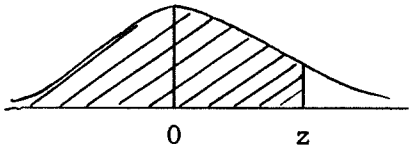
$\nu 1$ = 1(1)20(5)30(10)50(50)200

$\nu 2$ = 1(1)30(2)50(10)100(25)200(100)500,1000

α = 0.05, 0.025, 0.01, 0.005

Table 6.1

GAUSSIENNE CENTRÉE-RÉDUITE



$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}t^2\right] dt$$

$$\Phi(-z) = 1 - \Phi(z)$$

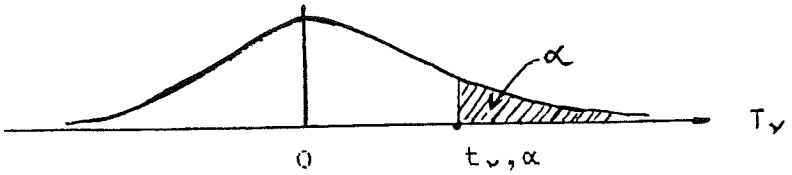
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.8788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

$$z_\alpha = \Phi^{-1}(1-\alpha) \qquad z_{1-\alpha} = -z_\alpha$$

α	0.10	0.05	0.025	0.01	0.005	0.001
z_α	1.282	1.645	1.960	2.326	2.576	3.090

Table 6.2

STUDENT



ν \ α	0.45	0.35	0.25	0.15	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.158	0.510	1.000	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.142	0.445	0.816	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.137	0.424	0.765	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.134	0.414	0.741	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.408	0.727	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.131	0.404	0.718	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.402	0.711	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.130	0.399	0.706	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.398	0.703	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.397	0.700	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.396	0.697	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.395	0.695	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.394	0.694	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.393	0.692	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.393	0.691	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.392	0.690	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.128	0.392	0.689	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.392	0.688	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.127	0.391	0.688	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.391	0.687	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.391	0.686	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.390	0.686	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.390	0.685	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	0.127	0.390	0.685	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.390	0.684	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.390	0.684	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.127	0.389	0.684	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.127	0.389	0.683	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.389	0.683	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.127	0.389	0.683	1.055	1.310	1.697	2.042	2.457	2.750	3.646
35	0.127	0.388	0.682	1.052	1.306	1.690	2.030	2.438	2.724	3.591
40	0.126	0.388	0.681	1.050	1.303	1.684	2.021	2.423	2.704	3.551
45	0.126	0.388	0.680	1.049	1.301	1.679	2.014	2.412	2.690	3.520
50	0.126	0.388	0.679	1.047	1.299	1.676	2.009	2.403	2.678	3.496
60	0.126	0.387	0.679	1.045	1.296	1.671	2.000	2.390	2.660	3.460
70	0.126	0.387	0.678	1.044	1.294	1.667	1.994	2.381	2.648	3.435
80	0.126	0.387	0.678	1.043	1.292	1.664	1.990	2.374	2.639	3.416
90	0.126	0.387	0.677	1.042	1.291	1.662	1.987	2.368	2.632	3.402
100	0.126	0.386	0.677	1.042	1.290	1.660	1.984	2.364	2.626	3.390
120	0.126	0.386	0.677	1.041	1.289	1.658	1.980	2.358	2.617	3.373
140	0.126	0.386	0.676	1.040	1.288	1.656	1.977	2.353	2.611	3.361
160	0.126	0.386	0.676	1.040	1.287	1.654	1.975	2.350	2.607	3.352
180	0.126	0.386	0.676	1.039	1.286	1.653	1.973	2.547	2.603	3.345
200	0.126	0.386	0.676	1.039	1.286	1.653	1.972	2.345	2.601	3.340
∞	0.126	0.385	0.674	1.036	1.282	1.645	1.960	2.326	2.576	3.291

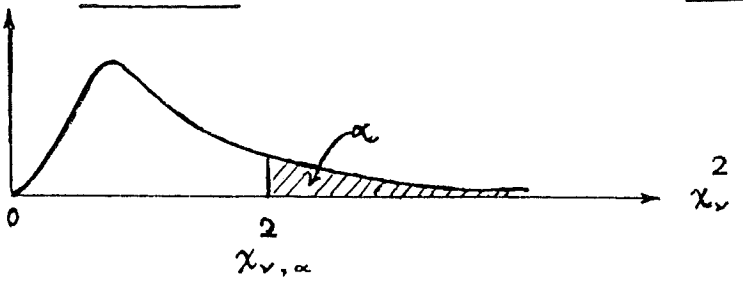
$$P[\ T_\nu > t_{\nu,\alpha}\ ] = \alpha$$

$t_{\nu,\alpha}$ = $(1-\alpha)$ ième percentile d'une variable de Student T_ν avec ν degrés de liberté

Exemple: $t_{20,0.05} = 1.725$

Table 6.3

KHI-DEUX



ν \ α	0.995	0.99	0.975	0.95	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.40	0.05	0.025	0.01	0.005	0.0005
1	0.0001	0.0002	0.001	0.004	0.016	0.064	0.148	0.275	0.455	0.708	1.074	1.642	2.706	3.841	5.024	6.635	7.879	12.116
2	0.010	0.020	0.051	0.103	0.211	0.446	0.713	1.022	1.386	1.833	2.408	3.219	4.605	5.991	7.378	9.210	10.597	15.202
3	0.072	0.115	0.216	0.352	0.584	1.005	1.424	1.869	2.366	2.946	3.665	4.642	6.251	7.815	9.348	11.345	12.838	17.730
4	0.207	0.297	0.484	0.711	1.064	1.649	2.195	2.753	3.357	4.045	4.878	5.989	7.779	9.488	11.143	13.277	14.860	19.997
5	0.412	0.554	0.831	1.145	1.610	2.343	3.000	3.655	4.351	5.132	6.064	7.289	9.236	11.070	12.833	15.086	16.750	22.105
6	0.676	0.872	1.237	1.635	2.204	3.070	3.828	4.570	5.348	6.211	7.231	8.558	10.645	12.592	14.449	16.812	18.548	24.103
7	0.989	1.239	1.690	2.167	2.833	3.822	4.671	5.493	6.346	7.283	8.383	9.803	12.017	14.067	16.013	18.475	20.278	26.018
8	1.344	1.646	2.180	2.733	3.490	4.594	5.527	6.423	7.344	8.351	9.524	11.030	13.362	15.507	17.535	20.090	21.955	27.868
9	1.735	2.088	2.700	3.325	4.168	5.380	6.393	7.357	8.343	9.414	10.656	12.242	14.684	16.919	19.023	21.666	23.589	29.665
10	2.156	2.558	3.247	3.940	4.865	6.179	7.267	8.295	9.342	10.473	11.781	13.442	15.987	18.307	20.483	23.209	25.188	31.420
11	2.603	3.053	3.816	4.575	5.578	6.989	8.148	9.237	10.341	11.530	12.899	14.631	17.275	19.675	21.920	24.725	26.757	33.137
12	3.074	3.571	4.404	5.226	6.304	7.807	9.034	10.182	11.340	12.584	14.011	15.812	18.549	21.026	23.337	26.217	28.300	34.821
13	3.565	4.107	5.009	5.892	7.042	8.634	9.926	11.129	12.340	13.636	15.119	16.985	19.812	22.362	24.736	27.688	29.819	36.478
14	4.075	4.660	5.629	6.571	7.790	9.467	10.821	12.078	13.339	14.685	16.222	18.151	21.064	23.685	26.119	29.141	31.319	38.109
15	4.601	5.229	6.262	7.261	8.547	10.307	11.721	13.030	14.339	15.733	17.322	19.311	22.307	24.996	27.488	30.578	32.801	39.719
16	5.142	5.812	6.908	7.962	9.312	11.152	12.624	13.983	15.338	16.780	18.418	20.465	23.542	26.296	28.845	32.000	34.267	41.308
17	5.697	6.408	7.564	8.672	10.085	12.002	13.531	14.937	16.338	17.824	19.511	21.615	24.769	27.587	30.191	33.409	35.718	42.879
18	6.265	7.015	8.231	9.390	10.865	12.857	14.440	15.893	17.338	18.868	20.601	22.760	25.989	28.869	31.526	34.805	37.156	44.434
19	6.844	7.633	8.907	10.117	11.651	13.716	15.352	16.850	18.338	19.910	21.689	23.900	27.204	30.144	32.852	36.191	38.582	45.973
20	7.434	8.260	9.591	10.851	12.443	14.578	16.266	17.809	19.337	20.951	22.775	25.038	28.412	31.410	34.170	37.566	39.997	47.498
21	8.034	8.897	10.283	11.591	13.240	15.445	17.182	18.768	20.337	21.991	23.858	26.171	29.615	32.671	35.479	38.932	41.401	49.011
22	8.643	9.542	10.982	12.338	14.041	16.314	18.101	19.729	21.337	23.031	24.939	27.301	30.813	33.924	36.781	40.289	42.796	50.511
23	9.260	10.196	11.689	13.091	14.848	17.187	19.021	20.690	22.337	24.069	26.018	28.429	32.007	35.172	38.076	41.638	44.181	52.000
24	9.886	10.856	12.401	13.848	15.659	18.062	19.943	21.652	23.337	25.106	27.096	29.553	33.196	36.415	39.364	42.980	45.559	53.479
25	10.520	11.524	13.120	14.611	16.473	18.940	20.867	22.616	24.337	26.143	28.172	30.675	34.382	37.652	40.646	44.314	46.928	54.947
26	11.160	12.198	13.844	15.379	17.292	19.820	21.792	23.579	25.336	27.179	29.246	31.795	35.563	38.885	41.923	45.642	48.290	56.407
27	11.808	12.879	14.573	16.151	18.114	20.703	22.719	24.544	26.336	28.214	30.319	32.912	36.741	40.113	43.195	46.963	49.645	57.858
28	12.461	13.565	15.308	16.928	18.939	21.588	23.647	25.509	27.336	29.249	31.391	34.027	37.916	41.337	44.461	48.278	50.993	59.300
29	13.121	14.256	16.047	17.708	19.768	22.475	24.577	26.475	28.336	30.283	32.461	35.139	39.087	42.557	45.722	49.588	52.336	60.735
30	13.787	14.953	16.791	18.493	20.599	23.364	25.508	27.442	29.336	31.316	33.530	36.250	40.256	43.773	46.979	50.892	53.672	62.162
35	17.192	18.509	20.569	22.465	24.797	27.836	30.178	32.282	34.336	36.475	38.859	41.778	46.059	49.802	53.203	57.342	60.275	69.199
40	20.707	22.164	24.433	26.509	29.051	32.345	34.872	37.134	39.335	41.622	44.165	47.269	51.805	55.758	59.342	63.691	66.766	76.095
45	24.311	25.901	28.366	30.612	33.350	36.884	39.585	41.995	44.335	46.761	49.452	52.729	57.505	61.656	65.410	69.957	73.166	82.876
50	27.991	29.707	32.357	34.764	37.689	41.449	44.313	46.864	49.335	51.892	54.723	58.164	63.167	67.505	71.420	76.154	79.490	89.561
60	35.534	37.485	40.482	43.188	46.459	50.641	53.809	56.620	59.335	62.135	65.227	68.972	74.397	79.082	83.298	88.379	91.952	102.695
70	43.275	45.442	48.758	51.739	55.329	59.898	63.346	66.396	69.334	72.358	75.689	79.715	85.527	90.531	95.023	100.425	104.215	115.578
80	51.172	53.540	57.153	60.391	64.278	69.207	72.915	76.188	79.334	82.566	86.120	90.405	96.578	101.879	106.629	112.329	116.321	128.261
90	59.196	61.754	65.647	69.126	73.291	78.558	82.511	85.993	89.334	92.761	96.524	101.054	107.565	113.145	118.136	124.116	128.299	140.782
100	67.328	70.065	74.222	77.929	82.358	87.945	92.129	95.808	99.334	102.946	106.906	111.667	118.498	124.342	129.561	135.807	140.169	153.167
120	83.852	86.923	91.573	95.705	100.624	106.806	111.419	115.465	119.334	123.289	127.616	132.806	140.233	146.567	152.211	158.950	163.648	177.603
140	100.655	104.034	109.137	113.659	119.029	125.758	130.766	135.149	139.334	143.604	148.269	153.854	161.827	168.613	174.648	181.840	186.847	201.683
160	117.679	121.346	126.870	131.756	137.546	144.783	150.158	154.856	159.334	163.898	168.876	174.828	183.311	190.516	196.915	204.530	209.824	225.481
180	134.884	138.820	144.741	149.969	156.153	163.868	169.588	174.580	179.334	184.173	189.446	195.743	204.704	212.304	219.044	227.056	232.620	249.048
200	152.241	156.432	162.728	168.279	174.835	183.003	189.049	194.319	199.334	204.434	209.985	216.609	226.021	233.994	241.058	249.445	255.264	272.423

$$P[\ \chi^2_\nu > \chi^2_{\nu,\alpha}\ ] = \alpha$$

$\chi^2_{\nu,\alpha}$ = (1-α)ième percentile d'une variable

khi-deux avec ν degrés de liberté

<u>Exemple</u>: $X^2_{20,\ 0.05} = 31.41$

Table 6.4 : FISHER

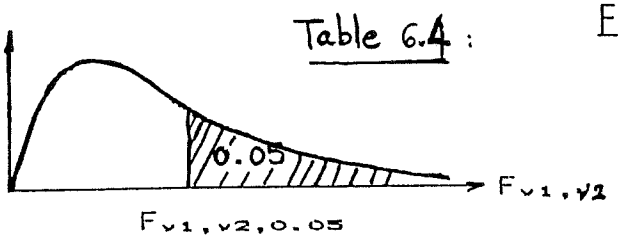


Table du 95ième percentile

ν_1 = degrés de liberté du numérateur

ν_2 = degrés de liberté du dénominateur

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	40	50	100	150	200
1	161	200	216	225	230	234	237	239	241	242	243	244	245	245	246	246	247	247	248	248	249	250	251	252	253	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70	8.69	8.68	8.67	8.67	8.66	8.63	8.62	8.59	8.58	8.55	8.54	8.54
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86	5.84	5.83	5.82	5.81	5.80	5.77	5.75	5.72	5.70	5.66	5.65	5.65
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62	4.60	4.59	4.58	4.57	4.56	4.52	4.50	4.46	4.44	4.41	4.39	4.39
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94	3.92	3.91	3.90	3.88	3.87	3.83	3.81	3.77	3.75	3.71	3.70	3.69
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51	3.49	3.48	3.47	3.46	3.44	3.40	3.38	3.34	3.32	3.27	3.26	3.25
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22	3.20	3.19	3.17	3.16	3.15	3.11	3.08	3.04	3.02	2.97	2.96	2.95
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01	2.99	2.97	2.96	2.95	2.94	2.89	2.86	2.83	2.80	2.76	2.74	2.73
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85	2.83	2.81	2.80	2.79	2.77	2.73	2.70	2.66	2.64	2.59	2.57	2.56
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72	2.70	2.69	2.67	2.66	2.65	2.60	2.57	2.53	2.51	2.46	2.44	2.43
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62	2.60	2.58	2.57	2.56	2.54	2.50	2.47	2.43	2.40	2.35	2.33	2.32
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53	2.51	2.50	2.48	2.47	2.46	2.41	2.38	2.34	2.31	2.26	2.24	2.23
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46	2.44	2.43	2.41	2.40	2.39	2.34	2.31	2.27	2.24	2.19	2.17	2.16
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40	2.38	2.37	2.35	2.34	2.33	2.28	2.25	2.20	2.18	2.12	2.10	2.10
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35	2.33	2.32	2.30	2.29	2.28	2.23	2.19	2.15	2.12	2.07	2.05	2.04
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31	2.29	2.27	2.26	2.24	2.23	2.18	2.15	2.10	2.08	2.02	2.00	1.99
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.27	2.25	2.23	2.22	2.20	2.19	2.14	2.11	2.06	2.04	1.98	1.96	1.95
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23	2.21	2.20	2.18	2.17	2.16	2.11	2.07	2.03	2.00	1.94	1.92	1.91
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.18	2.17	2.15	2.14	2.12	2.07	2.04	1.99	1.97	1.91	1.89	1.88
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.22	2.20	2.18	2.16	2.14	2.12	2.11	2.10	2.05	2.01	1.96	1.94	1.88	1.86	1.84
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.20	2.17	2.15	2.13	2.11	2.10	2.08	2.07	2.02	1.98	1.94	1.91	1.85	1.83	1.82
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.18	2.15	2.13	2.11	2.09	2.08	2.06	2.05	2.00	1.96	1.91	1.88	1.82	1.80	1.79
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.15	2.13	2.11	2.09	2.07	2.05	2.04	2.03	1.97	1.94	1.89	1.86	1.80	1.78	1.77
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.14	2.11	2.09	2.07	2.05	2.04	2.02	2.01	1.96	1.92	1.87	1.84	1.78	1.76	1.75
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.12	2.09	2.07	2.05	2.03	2.02	2.00	1.99	1.94	1.90	1.85	1.82	1.76	1.74	1.73
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17	2.13	2.10	2.08	2.06	2.04	2.02	2.00	1.99	1.97	1.92	1.88	1.84	1.81	1.74	1.72	1.71
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.09	2.06	2.04	2.02	2.00	1.99	1.97	1.96	1.91	1.87	1.82	1.79	1.73	1.70	1.69
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.08	2.05	2.03	2.01	1.99	1.97	1.96	1.94	1.89	1.85	1.81	1.77	1.71	1.69	1.67
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.06	2.04	2.01	1.99	1.98	1.96	1.95	1.93	1.88	1.84	1.79	1.76	1.70	1.67	1.66
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14	2.10	2.07	2.04	2.01	1.99	1.97	1.95	1.94	1.92	1.91	1.85	1.82	1.77	1.74	1.67	1.64	1.63
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12	2.08	2.05	2.02	1.99	1.97	1.95	1.93	1.92	1.90	1.89	1.83	1.80	1.75	1.71	1.65	1.62	1.61
36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11	2.07	2.03	2.00	1.98	1.95	1.93	1.92	1.90	1.88	1.87	1.81	1.78	1.73	1.69	1.62	1.60	1.59
38	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.99	1.96	1.94	1.92	1.90	1.88	1.87	1.85	1.80	1.76	1.71	1.68	1.61	1.58	1.57
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.97	1.95	1.92	1.90	1.89	1.87	1.85	1.84	1.78	1.74	1.69	1.66	1.59	1.56	1.55
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.03	1.99	1.96	1.94	1.91	1.89	1.87	1.86	1.84	1.83	1.77	1.73	1.68	1.65	1.57	1.55	1.53
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.95	1.92	1.90	1.88	1.86	1.84	1.83	1.81	1.76	1.72	1.67	1.63	1.56	1.53	1.52
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.15	2.09	2.04	2.00	1.97	1.94	1.91	1.89	1.87	1.85	1.83	1.82	1.80	1.75	1.71	1.65	1.62	1.55	1.52	1.51
48	4.04	3.19	2.80	2.57	2.41	2.29	2.21	2.14	2.08	2.03	1.99	1.96	1.93	1.90	1.88	1.86	1.84	1.82	1.81	1.79	1.74	1.70	1.64	1.61	1.54	1.51	1.49
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.99	1.95	1.92	1.89	1.87	1.85	1.83	1.81	1.80	1.78	1.73	1.69	1.63	1.60	1.52	1.50	1.48
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.89	1.86	1.84	1.82	1.80	1.78	1.76	1.75	1.69	1.65	1.59	1.56	1.48	1.45	1.44
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.93	1.89	1.86	1.84	1.81	1.79	1.77	1.75	1.74	1.72	1.66	1.62	1.57	1.53	1.45	1.42	1.40
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.91	1.88	1.84	1.82	1.79	1.77	1.75	1.73	1.72	1.70	1.64	1.60	1.54	1.51	1.43	1.39	1.38
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.90	1.86	1.83	1.80	1.78	1.76	1.74	1.72	1.70	1.69	1.63	1.59	1.53	1.49	1.41	1.38	1.36
100	3.94	3.09	2.70	2.45	2.31	2.19	2.10	2.03	1.97	1.93	1.89	1.85	1.82	1.79	1.77	1.75	1.73	1.71	1.69	1.68	1.62	1.57	1.52	1.48	1.39	1.36	1.34
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.96	1.91	1.87	1.83	1.80	1.77	1.75	1.73	1.71	1.69	1.67	1.66	1.59	1.55	1.49	1.45	1.36	1.33	1.31
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82	1.79	1.76	1.73	1.71	1.69	1.67	1.66	1.64	1.58	1.54	1.48	1.44	1.34	1.31	1.29
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.84	1.80	1.77	1.74	1.72	1.69	1.67	1.66	1.64	1.62	1.56	1.52	1.46	1.41	1.32	1.28	1.26
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.82	1.78	1.75	1.72	1.70	1.68	1.66	1.64	1.62	1.61	1.54	1.50	1.43	1.39	1.30	1.26	1.23
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.77	1.74	1.71	1.69	1.66	1.64	1.62	1.61	1.59	1.53	1.48	1.42	1.38	1.28	1.23	1.21
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.80	1.76	1.73	1.70	1.68	1.65	1.63	1.61	1.60	1.58	1.52	1.47	1.41	1.36	1.26	1.22	1.19

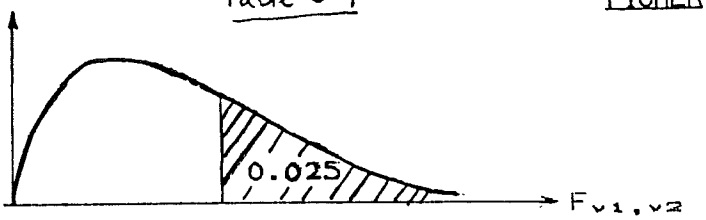
$$P[\ F_{\nu 1,\nu 2} > F_{\nu 1,\nu 2,0.05}\ ] = 0.05$$

$F_{\nu 1,\nu 2,0.05}$ = 95ième percentile d'une variable $F_{\nu 1,\nu 2}$

avec ν_1 degrés de liberté au numérateur
ν_2 degrés de liberté au dénominateur

Exemple: $F_{10,25,0.05} = 2.24$

Table 6.4 FISHER



$F_{\nu 1,\nu 2,0.025}$

Table du 97.5ième percentile

ν_1 = degrés de liberté du numérateur

ν_2 = degrés de liberté du dénominateur

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	40	50	100	150	200
1	548	800	864	900	922	937	948	957	963	969	973	977	980	983	985	987	989	990	992	993	998	1001	1006	1008	1013	1015	1016
2	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.5	39.5	39.5	39.5	39.5	39.5	39.5
3	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.4	14.3	14.3	14.3	14.3	14.2	14.2	14.2	14.2	14.2	14.1	14.1	14.0	14.0	14.0	13.9	13.9
4	12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.79	8.75	8.71	8.68	8.66	8.63	8.61	8.59	8.58	8.56	8.50	8.46	8.41	8.38	8.32	8.30	8.29
5	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.57	6.52	6.49	6.46	6.43	6.40	6.38	6.36	6.34	6.33	6.27	6.23	6.18	6.14	6.08	6.06	6.05
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.41	5.37	5.33	5.30	5.27	5.24	5.22	5.20	5.18	5.17	5.11	5.07	5.01	4.98	4.92	4.89	4.88
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.71	4.67	4.63	4.60	4.57	4.54	4.52	4.50	4.48	4.47	4.40	4.36	4.31	4.28	4.21	4.19	4.18
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.24	4.20	4.16	4.13	4.10	4.08	4.05	4.03	4.02	4.00	3.94	3.89	3.84	3.81	3.74	3.72	3.70
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.91	3.87	3.83	3.80	3.77	3.74	3.72	3.70	3.68	3.67	3.60	3.56	3.51	3.47	3.40	3.38	3.37
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.66	3.62	3.58	3.55	3.52	3.50	3.47	3.45	3.44	3.42	3.35	3.31	3.26	3.22	3.15	3.13	3.12
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.47	3.43	3.39	3.36	3.33	3.30	3.28	3.26	3.24	3.23	3.16	3.12	3.06	3.03	2.96	2.93	2.92
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.32	3.28	3.24	3.21	3.18	3.15	3.13	3.11	3.09	3.07	3.01	2.96	2.91	2.87	2.80	2.78	2.76
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.20	3.15	3.12	3.09	3.05	3.03	3.00	2.98	2.96	2.95	2.88	2.84	2.78	2.74	2.67	2.65	2.63
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.09	3.05	3.01	2.98	2.95	2.92	2.90	2.88	2.86	2.84	2.78	2.73	2.67	2.64	2.56	2.54	2.53
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	3.01	2.96	2.92	2.89	2.86	2.84	2.81	2.79	2.77	2.76	2.69	2.64	2.59	2.55	2.47	2.45	2.44
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.93	2.89	2.85	2.82	2.79	2.76	2.74	2.72	2.70	2.68	2.61	2.57	2.51	2.47	2.40	2.37	2.36
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.87	2.82	2.79	2.75	2.72	2.70	2.67	2.65	2.63	2.62	2.55	2.50	2.44	2.41	2.33	2.30	2.29
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.81	2.77	2.73	2.70	2.67	2.64	2.62	2.60	2.58	2.56	2.49	2.44	2.38	2.35	2.27	2.24	2.23
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.76	2.72	2.68	2.65	2.62	2.59	2.57	2.55	2.53	2.51	2.44	2.39	2.33	2.30	2.22	2.19	2.18
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.72	2.68	2.64	2.60	2.57	2.55	2.52	2.50	2.48	2.46	2.40	2.35	2.29	2.25	2.17	2.14	2.13
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.68	2.64	2.60	2.56	2.53	2.51	2.48	2.46	2.44	2.42	2.36	2.31	2.25	2.21	2.13	2.10	2.09
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.65	2.60	2.56	2.53	2.50	2.47	2.45	2.43	2.41	2.39	2.32	2.27	2.21	2.17	2.09	2.06	2.05
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.62	2.57	2.53	2.50	2.47	2.44	2.42	2.39	2.37	2.36	2.29	2.24	2.18	2.14	2.06	2.03	2.01
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.59	2.54	2.50	2.47	2.44	2.41	2.39	2.36	2.35	2.33	2.26	2.21	2.15	2.11	2.02	2.00	1.98
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.56	2.51	2.48	2.44	2.41	2.38	2.36	2.34	2.32	2.30	2.23	2.18	2.12	2.08	2.00	1.97	1.95
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.54	2.49	2.45	2.42	2.39	2.36	2.34	2.31	2.29	2.28	2.21	2.16	2.09	2.05	1.97	1.94	1.92
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.51	2.47	2.43	2.39	2.36	2.34	2.31	2.29	2.27	2.25	2.18	2.13	2.07	2.03	1.94	1.91	1.90
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.49	2.45	2.41	2.37	2.34	2.32	2.29	2.27	2.25	2.23	2.16	2.11	2.05	2.01	1.92	1.89	1.88
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.48	2.43	2.39	2.36	2.33	2.30	2.27	2.25	2.23	2.21	2.14	2.09	2.03	1.99	1.90	1.87	1.86
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.46	2.41	2.37	2.34	2.31	2.28	2.26	2.23	2.21	2.20	2.12	2.07	2.01	1.97	1.88	1.85	1.84
32	5.53	4.15	3.56	3.22	3.00	2.84	2.71	2.62	2.54	2.48	2.43	2.38	2.34	2.31	2.28	2.25	2.22	2.20	2.18	2.16	2.09	2.04	1.98	1.93	1.85	1.82	1.80
34	5.50	4.12	3.53	3.19	2.97	2.81	2.69	2.59	2.52	2.45	2.40	2.35	2.31	2.28	2.25	2.22	2.20	2.17	2.15	2.13	2.06	2.01	1.95	1.90	1.82	1.78	1.77
36	5.47	4.09	3.50	3.17	2.94	2.78	2.66	2.57	2.49	2.43	2.37	2.33	2.29	2.25	2.22	2.20	2.17	2.15	2.13	2.11	2.04	1.99	1.92	1.88	1.79	1.76	1.74
38	5.45	4.07	3.48	3.15	2.92	2.76	2.64	2.55	2.47	2.41	2.35	2.31	2.27	2.23	2.20	2.17	2.15	2.13	2.11	2.09	2.01	1.96	1.90	1.85	1.76	1.73	1.71
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.13	2.11	2.09	2.07	1.99	1.94	1.88	1.83	1.74	1.71	1.69
42	5.40	4.03	3.45	3.11	2.89	2.73	2.61	2.51	2.43	2.37	2.32	2.27	2.23	2.20	2.16	2.14	2.11	2.09	2.07	2.05	1.98	1.92	1.86	1.81	1.72	1.69	1.67
44	5.39	4.02	3.43	3.09	2.87	2.71	2.59	2.50	2.42	2.36	2.30	2.26	2.22	2.18	2.15	2.12	2.10	2.07	2.05	2.03	1.96	1.91	1.84	1.80	1.70	1.67	1.65
46	5.37	4.00	3.42	3.08	2.86	2.70	2.58	2.48	2.41	2.34	2.29	2.24	2.20	2.17	2.13	2.11	2.08	2.06	2.04	2.02	1.94	1.89	1.82	1.78	1.69	1.65	1.63
48	5.35	3.99	3.40	3.07	2.84	2.69	2.56	2.47	2.39	2.33	2.27	2.23	2.19	2.15	2.12	2.09	2.07	2.05	2.02	2.01	1.93	1.88	1.81	1.77	1.67	1.64	1.62
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.26	2.22	2.18	2.14	2.11	2.08	2.06	2.03	2.01	1.99	1.92	1.87	1.80	1.75	1.66	1.62	1.60
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.22	2.17	2.13	2.09	2.06	2.03	2.01	1.98	1.96	1.94	1.87	1.82	1.74	1.70	1.60	1.56	1.54
70	5.25	3.89	3.31	2.97	2.75	2.59	2.47	2.38	2.30	2.24	2.18	2.14	2.10	2.06	2.03	2.00	1.97	1.95	1.93	1.91	1.83	1.78	1.71	1.66	1.56	1.52	1.50
80	5.22	3.86	3.28	2.95	2.73	2.57	2.45	2.35	2.28	2.21	2.16	2.11	2.07	2.03	2.00	1.97	1.95	1.92	1.90	1.88	1.81	1.75	1.68	1.63	1.53	1.49	1.47
90	5.20	3.84	3.26	2.93	2.71	2.55	2.43	2.34	2.26	2.19	2.14	2.09	2.05	2.02	1.98	1.95	1.93	1.91	1.88	1.86	1.79	1.73	1.66	1.61	1.50	1.46	1.44
100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	2.12	2.08	2.04	2.00	1.97	1.94	1.91	1.89	1.87	1.85	1.77	1.71	1.64	1.59	1.48	1.44	1.42
125	5.15	3.80	3.22	2.89	2.67	2.51	2.39	2.30	2.22	2.15	2.10	2.05	2.01	1.97	1.94	1.91	1.89	1.86	1.84	1.82	1.74	1.68	1.61	1.56	1.45	1.40	1.38
150	5.13	3.78	3.20	2.87	2.65	2.49	2.37	2.28	2.20	2.13	2.08	2.03	1.99	1.95	1.92	1.89	1.87	1.84	1.82	1.80	1.72	1.67	1.59	1.54	1.42	1.38	1.35
200	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18	2.11	2.06	2.01	1.97	1.93	1.90	1.87	1.84	1.82	1.80	1.78	1.70	1.64	1.56	1.51	1.39	1.35	1.32
300	5.07	3.73	3.16	2.83	2.61	2.45	2.33	2.23	2.16	2.09	2.04	1.99	1.95	1.91	1.88	1.85	1.82	1.80	1.77	1.75	1.67	1.62	1.54	1.48	1.36	1.31	1.28
500	5.05	3.72	3.14	2.81	2.59	2.43	2.31	2.22	2.14	2.07	2.02	1.97	1.93	1.89	1.86	1.83	1.80	1.78	1.76	1.74	1.65	1.60	1.52	1.46	1.34	1.28	1.25
1000	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20	2.13	2.06	2.01	1.96	1.92	1.88	1.85	1.82	1.79	1.77	1.74	1.72	1.64	1.58	1.50	1.45	1.32	1.26	1.23

$$P[\ F_{\nu 1,\nu 2} > F_{\nu 1,\nu 2,0.025}\ ] = 0.025$$

$F_{\nu 1,\nu 2,0.975}$ = 97.5ième percentile d'une variable $F_{\nu 1,\nu 2}$
avec ν_1 degrés de liberté au numérateur
ν_2 degrés de liberté au dénominateur

Exemple: $F_{10,25,0.025} = 2.61$

Table 6.4 FISHER

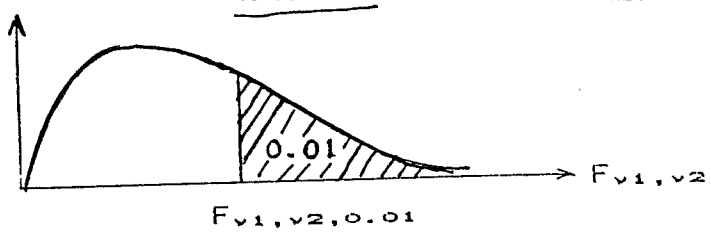


Table du 99ième percentile

ν_1 = degrés de liberté du numérateur

ν_2 = degrés de liberté du dénominateur

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	40	50	100	150	200
1	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6083	6106	6126	6143	6157	6170	6181	6192	6201	6209	6240	6261	6287	6303	6334	6345	6350
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	27.1	27.0	26.9	26.9	26.8	26.8	26.8	26.7	26.7	26.6	26.5	26.4	26.4	26.2	26.2	26.2
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.5	14.4	14.3	14.2	14.2	14.2	14.1	14.1	14.0	14.0	13.9	13.8	13.7	13.7	13.6	13.5	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.96	9.89	9.82	9.77	9.72	9.68	9.64	9.61	9.58	9.55	9.45	9.38	9.29	9.24	9.13	9.09	9.08
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.66	7.60	7.56	7.52	7.48	7.45	7.42	7.40	7.30	7.23	7.14	7.09	6.99	6.95	6.93
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.41	6.36	6.31	6.28	6.24	6.21	6.18	6.16	6.06	5.99	5.91	5.86	5.75	5.72	5.70
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.61	5.56	5.52	5.48	5.44	5.41	5.38	5.36	5.26	5.20	5.12	5.07	4.96	4.93	4.91
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	5.05	5.01	4.96	4.92	4.89	4.86	4.83	4.81	4.71	4.65	4.57	4.52	4.41	4.38	4.36
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.65	4.60	4.56	4.52	4.49	4.46	4.43	4.41	4.31	4.25	4.17	4.12	4.01	3.98	3.96
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.34	4.29	4.25	4.21	4.18	4.15	4.12	4.10	4.01	3.94	3.86	3.81	3.71	3.67	3.66
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.10	4.05	4.01	3.97	3.94	3.91	3.88	3.86	3.76	3.70	3.62	3.57	3.47	3.43	3.41
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.91	3.86	3.82	3.78	3.75	3.72	3.69	3.66	3.57	3.51	3.43	3.38	3.27	3.24	3.22
14	8.88	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.75	3.70	3.66	3.62	3.59	3.56	3.53	3.51	3.41	3.35	3.27	3.22	3.11	3.08	3.06
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.61	3.56	3.52	3.49	3.45	3.42	3.40	3.37	3.28	3.21	3.13	3.08	2.98	2.94	2.92
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55	3.50	3.45	3.41	3.37	3.34	3.31	3.28	3.26	3.16	3.10	3.02	2.97	2.86	2.83	2.81
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46	3.40	3.35	3.31	3.27	3.24	3.21	3.19	3.16	3.07	3.00	2.92	2.87	2.76	2.73	2.71
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37	3.32	3.27	3.23	3.19	3.16	3.13	3.10	3.08	2.98	2.92	2.84	2.78	2.68	2.64	2.62
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.24	3.19	3.15	3.12	3.08	3.05	3.03	3.00	2.91	2.84	2.76	2.71	2.60	2.57	2.55
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.18	3.13	3.09	3.05	3.02	2.99	2.96	2.94	2.84	2.78	2.69	2.64	2.54	2.50	2.48
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.24	3.17	3.12	3.07	3.03	2.99	2.96	2.93	2.90	2.88	2.79	2.72	2.64	2.58	2.48	2.44	2.42
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.07	3.02	2.98	2.94	2.91	2.88	2.85	2.83	2.73	2.67	2.58	2.53	2.42	2.38	2.36
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	3.02	2.97	2.93	2.89	2.86	2.83	2.80	2.78	2.69	2.62	2.54	2.48	2.37	2.34	2.32
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03	2.98	2.93	2.89	2.85	2.82	2.79	2.76	2.74	2.64	2.58	2.49	2.44	2.33	2.29	2.27
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	3.06	2.99	2.94	2.89	2.85	2.81	2.78	2.75	2.72	2.70	2.60	2.54	2.45	2.40	2.29	2.25	2.23
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	3.02	2.96	2.90	2.86	2.81	2.78	2.75	2.72	2.69	2.66	2.57	2.50	2.42	2.36	2.25	2.21	2.19
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.99	2.93	2.87	2.82	2.78	2.75	2.71	2.68	2.66	2.63	2.54	2.47	2.38	2.33	2.22	2.18	2.16
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.96	2.90	2.84	2.79	2.75	2.72	2.68	2.65	2.63	2.60	2.51	2.44	2.35	2.30	2.19	2.15	2.13
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.93	2.87	2.81	2.77	2.73	2.69	2.66	2.63	2.60	2.57	2.48	2.41	2.33	2.27	2.16	2.12	2.10
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84	2.79	2.74	2.70	2.66	2.63	2.60	2.57	2.55	2.45	2.39	2.30	2.25	2.13	2.09	2.07
32	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02	2.93	2.86	2.80	2.74	2.70	2.65	2.62	2.58	2.55	2.53	2.50	2.41	2.34	2.25	2.20	2.08	2.04	2.02
34	7.44	5.29	4.42	3.93	3.61	3.39	3.22	3.09	2.98	2.89	2.82	2.76	2.70	2.66	2.61	2.58	2.54	2.51	2.49	2.46	2.37	2.30	2.21	2.16	2.04	2.00	1.98
36	7.40	5.25	4.38	3.89	3.57	3.35	3.18	3.05	2.95	2.86	2.79	2.72	2.67	2.62	2.58	2.54	2.51	2.48	2.45	2.43	2.33	2.26	2.18	2.12	2.00	1.96	1.94
38	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.92	2.83	2.75	2.69	2.64	2.59	2.55	2.51	2.48	2.45	2.42	2.40	2.30	2.23	2.14	2.09	1.97	1.93	1.90
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66	2.61	2.56	2.52	2.48	2.45	2.42	2.39	2.37	2.27	2.20	2.11	2.06	1.94	1.90	1.87
42	7.28	5.15	4.29	3.80	3.49	3.27	3.10	2.97	2.86	2.78	2.70	2.64	2.59	2.54	2.50	2.46	2.43	2.40	2.37	2.34	2.25	2.18	2.09	2.03	1.91	1.87	1.85
44	7.25	5.12	4.26	3.78	3.47	3.24	3.08	2.95	2.84	2.75	2.68	2.62	2.56	2.52	2.47	2.44	2.40	2.37	2.35	2.32	2.22	2.15	2.07	2.01	1.89	1.84	1.82
46	7.22	5.10	4.24	3.76	3.44	3.22	3.06	2.93	2.82	2.73	2.66	2.60	2.54	2.50	2.45	2.42	2.38	2.35	2.33	2.30	2.20	2.13	2.04	1.99	1.86	1.82	1.80
48	7.19	5.08	4.22	3.74	3.43	3.20	3.04	2.91	2.80	2.71	2.64	2.58	2.53	2.48	2.44	2.40	2.37	2.33	2.31	2.28	2.18	2.12	2.02	1.97	1.84	1.80	1.78
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.63	2.56	2.51	2.46	2.42	2.38	2.35	2.32	2.29	2.27	2.17	2.10	2.01	1.95	1.82	1.78	1.76
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.44	2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.10	2.03	1.94	1.88	1.75	1.70	1.68
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59	2.51	2.45	2.40	2.35	2.31	2.27	2.23	2.20	2.18	2.15	2.05	1.98	1.89	1.83	1.70	1.65	1.62
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.48	2.42	2.36	2.31	2.27	2.23	2.20	2.17	2.14	2.12	2.01	1.94	1.85	1.79	1.65	1.61	1.58
90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52	2.45	2.39	2.33	2.29	2.24	2.21	2.17	2.14	2.11	2.09	1.99	1.92	1.82	1.76	1.62	1.57	1.55
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.43	2.37	2.31	2.27	2.22	2.19	2.15	2.12	2.09	2.07	1.97	1.89	1.80	1.74	1.60	1.55	1.52
125	6.84	4.78	3.94	3.47	3.17	2.95	2.79	2.66	2.55	2.47	2.39	2.33	2.28	2.23	2.19	2.15	2.11	2.08	2.05	2.03	1.93	1.85	1.76	1.69	1.55	1.50	1.47
150	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53	2.44	2.37	2.31	2.25	2.20	2.16	2.12	2.09	2.06	2.03	2.00	1.90	1.83	1.73	1.66	1.52	1.46	1.43
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.34	2.27	2.22	2.17	2.13	2.09	2.06	2.03	2.00	1.97	1.87	1.79	1.69	1.63	1.48	1.42	1.39
300	6.72	4.68	3.85	3.38	3.08	2.86	2.70	2.57	2.47	2.38	2.31	2.24	2.19	2.14	2.10	2.06	2.03	1.99	1.97	1.94	1.84	1.76	1.66	1.59	1.44	1.38	1.35
500	6.69	4.65	3.82	3.36	3.05	2.84	2.68	2.55	2.44	2.36	2.28	2.22	2.17	2.12	2.07	2.04	2.00	1.97	1.94	1.92	1.81	1.74	1.63	1.57	1.41	1.34	1.31
1000	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.27	2.20	2.15	2.10	2.06	2.02	1.98	1.95	1.92	1.90	1.79	1.72	1.61	1.54	1.38	1.32	1.28

$$P[\ F_{\nu 1,\nu 2} > F_{\nu 1,\nu 2,0.01}\ ] = 0.01$$

$F_{\nu 1,\nu 2,0.01}$ = 99ième percentile d'une variable $F_{\nu 1,\nu 2}$

avec ν_1 degrés de liberté au numérateur

ν_2 degrés de liberté au dénominateur

Exemple: $F_{10,25,0.01} = 3.13$

Table 6.4 FISHER

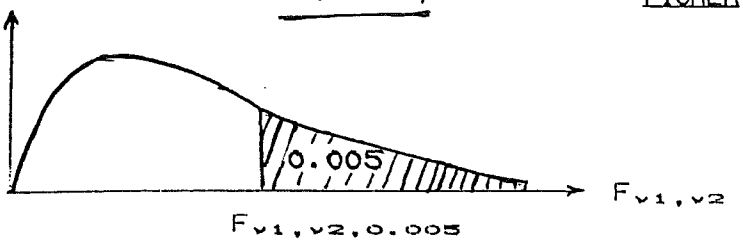


Table du 99.5ième percentile

ν_1 = degrés de liberté du numérateur

ν_2 = degrés de liberté du dénominateur

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	40	50	100	150	200
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199
3	55.6	49.8	47.5	46.2	45.4	44.8	44.4	44.1	43.9	43.7	43.5	43.4	43.3	43.2	43.1	43.0	42.9	42.9	42.8	42.8	42.6	42.5	42.3	42.2	42.0	42.0	41.9
4	31.3	26.3	24.3	23.2	22.5	22.0	21.6	21.4	21.1	21.0	20.8	20.7	20.6	20.5	20.4	20.4	20.3	20.3	20.2	20.2	20.0	19.9	19.8	19.7	19.5	19.4	19.4
5	22.8	18.3	16.5	15.6	14.9	14.5	14.2	14.0	13.8	13.6	13.5	13.4	13.3	13.2	13.1	13.1	13.0	13.0	12.9	12.9	12.8	12.7	12.5	12.5	12.3	12.2	12.2
6	18.6	14.5	12.9	12.0	11.5	11.1	10.8	10.6	10.4	10.3	10.1	10.0	9.95	9.88	9.81	9.76	9.71	9.66	9.62	9.59	9.45	9.36	9.24	9.17	9.03	8.98	8.95
7	16.2	12.4	10.9	10.1	9.52	9.16	8.89	8.68	8.51	8.38	8.27	8.18	8.10	8.03	7.97	7.91	7.87	7.83	7.79	7.75	7.62	7.53	7.42	7.35	7.22	7.17	7.15
8	14.7	11.0	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.10	7.01	6.94	6.87	6.81	6.76	6.72	6.68	6.64	6.61	6.48	6.40	6.29	6.22	6.09	6.04	6.02
9	13.6	10.1	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.31	6.23	6.15	6.09	6.03	5.98	5.94	5.90	5.86	5.83	5.71	5.62	5.52	5.45	5.32	5.28	5.26
10	12.8	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.75	5.66	5.59	5.53	5.47	5.42	5.38	5.34	5.31	5.27	5.15	5.07	4.97	4.90	4.77	4.73	4.71
11	12.2	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	5.32	5.24	5.18	5.10	5.05	5.00	4.96	4.92	4.89	4.86	4.74	4.65	4.55	4.49	4.36	4.31	4.29
12	11.8	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.99	4.91	4.84	4.77	4.72	4.67	4.63	4.59	4.56	4.53	4.41	4.33	4.23	4.17	4.04	3.99	3.97
13	11.4	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.72	4.64	4.57	4.51	4.46	4.41	4.37	4.33	4.30	4.27	4.15	4.07	3.97	3.91	3.78	3.74	3.71
14	11.1	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.51	4.43	4.36	4.30	4.25	4.20	4.16	4.12	4.09	4.06	3.94	3.86	3.76	3.70	3.57	3.53	3.50
15	10.8	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.33	4.25	4.18	4.12	4.07	4.02	3.98	3.95	3.91	3.88	3.77	3.69	3.58	3.52	3.39	3.35	3.33
16	10.6	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	4.18	4.10	4.03	3.97	3.92	3.87	3.83	3.80	3.76	3.73	3.62	3.54	3.44	3.37	3.25	3.20	3.18
17	10.4	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	4.05	3.97	3.90	3.84	3.79	3.75	3.71	3.67	3.64	3.61	3.49	3.41	3.31	3.25	3.12	3.07	3.05
18	10.2	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	3.94	3.86	3.79	3.73	3.68	3.64	3.60	3.56	3.53	3.50	3.38	3.30	3.20	3.14	3.01	2.96	2.94
19	10.1	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	3.84	3.76	3.70	3.64	3.59	3.54	3.50	3.46	3.43	3.40	3.29	3.21	3.11	3.04	2.91	2.87	2.85
20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.76	3.68	3.61	3.55	3.50	3.46	3.42	3.38	3.35	3.32	3.20	3.12	3.02	2.96	2.83	2.78	2.76
21	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77	3.68	3.60	3.54	3.48	3.43	3.38	3.34	3.31	3.27	3.24	3.13	3.05	2.95	2.88	2.75	2.71	2.68
22	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70	3.61	3.54	3.47	3.41	3.36	3.31	3.27	3.24	3.21	3.18	3.06	2.98	2.88	2.82	2.69	2.64	2.62
23	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64	3.55	3.47	3.41	3.35	3.30	3.25	3.21	3.18	3.15	3.12	3.00	2.92	2.82	2.76	2.62	2.58	2.56
24	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.50	3.42	3.35	3.30	3.25	3.20	3.16	3.12	3.09	3.06	2.95	2.87	2.77	2.70	2.57	2.52	2.50
25	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54	3.45	3.37	3.30	3.25	3.20	3.15	3.11	3.08	3.04	3.01	2.90	2.82	2.72	2.65	2.52	2.47	2.45
26	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49	3.40	3.33	3.26	3.20	3.15	3.11	3.07	3.03	3.00	2.97	2.85	2.77	2.67	2.61	2.47	2.43	2.40
27	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45	3.36	3.28	3.22	3.16	3.11	3.07	3.03	2.99	2.96	2.93	2.81	2.73	2.63	2.57	2.43	2.38	2.36
28	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41	3.32	3.25	3.18	3.12	3.07	3.03	2.99	2.95	2.92	2.89	2.77	2.69	2.59	2.53	2.39	2.35	2.32
29	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48	3.38	3.29	3.21	3.15	3.09	3.04	2.99	2.95	2.92	2.88	2.86	2.74	2.66	2.56	2.49	2.36	2.31	2.29
30	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.25	3.18	3.11	3.06	3.01	2.96	2.92	2.89	2.85	2.82	2.71	2.63	2.52	2.46	2.32	2.28	2.25
32	9.09	6.28	5.17	4.56	4.17	3.89	3.68	3.52	3.39	3.29	3.20	3.12	3.06	3.00	2.95	2.90	2.86	2.83	2.80	2.77	2.65	2.57	2.47	2.40	2.26	2.22	2.19
34	9.01	6.22	5.11	4.50	4.11	3.84	3.63	3.47	3.34	3.24	3.15	3.07	3.01	2.95	2.90	2.85	2.81	2.78	2.75	2.72	2.60	2.52	2.42	2.35	2.21	2.16	2.14
36	8.94	6.16	5.06	4.46	4.06	3.79	3.58	3.42	3.30	3.19	3.10	3.03	2.96	2.90	2.85	2.81	2.77	2.73	2.70	2.67	2.56	2.48	2.37	2.30	2.17	2.12	2.09
38	8.88	6.11	5.02	4.41	4.02	3.75	3.54	3.39	3.26	3.15	3.06	2.99	2.92	2.87	2.82	2.77	2.73	2.70	2.66	2.63	2.52	2.44	2.33	2.27	2.12	2.08	2.05
40	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	3.03	2.95	2.89	2.83	2.78	2.74	2.70	2.66	2.63	2.60	2.48	2.40	2.30	2.23	2.09	2.04	2.01
42	8.78	6.03	4.94	4.34	3.95	3.68	3.48	3.32	3.19	3.09	3.00	2.92	2.86	2.80	2.75	2.71	2.67	2.63	2.60	2.57	2.45	2.37	2.26	2.20	2.06	2.00	1.98
44	8.74	5.99	4.91	4.31	3.92	3.65	3.45	3.29	3.16	3.06	2.97	2.89	2.83	2.77	2.72	2.68	2.64	2.60	2.57	2.54	2.42	2.34	2.24	2.17	2.03	1.97	1.95
46	8.70	5.96	4.88	4.28	3.90	3.62	3.42	3.26	3.14	3.03	2.94	2.87	2.80	2.75	2.70	2.65	2.61	2.58	2.54	2.51	2.40	2.32	2.21	2.14	2.00	1.95	1.92
48	8.66	5.93	4.85	4.25	3.87	3.60	3.40	3.24	3.11	3.01	2.92	2.85	2.78	2.72	2.67	2.63	2.59	2.55	2.52	2.49	2.37	2.29	2.19	2.12	1.97	1.92	1.90
50	8.63	5.90	4.83	4.23	3.85	3.58	3.38	3.22	3.09	2.99	2.90	2.82	2.76	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.35	2.27	2.16	2.10	1.95	1.90	1.87
60	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.82	2.74	2.68	2.62	2.57	2.53	2.49	2.45	2.42	2.39	2.27	2.19	2.08	2.01	1.86	1.81	1.78
70	8.40	5.72	4.66	4.08	3.70	3.43	3.23	3.08	2.95	2.85	2.76	2.68	2.62	2.56	2.51	2.47	2.43	2.39	2.36	2.33	2.21	2.13	2.02	1.95	1.80	1.74	1.71
80	8.33	5.67	4.61	4.03	3.65	3.39	3.19	3.03	2.91	2.80	2.72	2.64	2.58	2.52	2.47	2.43	2.39	2.35	2.32	2.29	2.17	2.08	1.97	1.90	1.75	1.69	1.66
90	8.28	5.62	4.57	3.99	3.62	3.35	3.15	3.00	2.87	2.77	2.68	2.61	2.54	2.49	2.44	2.39	2.35	2.32	2.28	2.25	2.13	2.05	1.94	1.87	1.71	1.65	1.62
100	8.24	5.59	4.54	3.96	3.59	3.33	3.13	2.97	2.85	2.74	2.66	2.58	2.52	2.46	2.41	2.37	2.33	2.29	2.26	2.23	2.11	2.02	1.91	1.84	1.68	1.62	1.59
125	8.17	5.53	4.49	3.91	3.54	3.28	3.08	2.93	2.80	2.70	2.61	2.54	2.47	2.42	2.37	2.32	2.28	2.24	2.21	2.18	2.06	1.98	1.86	1.79	1.63	1.56	1.53
150	8.12	5.49	4.45	3.88	3.51	3.25	3.05	2.89	2.77	2.67	2.58	2.51	2.44	2.38	2.33	2.29	2.25	2.21	2.18	2.15	2.03	1.94	1.83	1.76	1.59	1.53	1.49
200	8.06	5.44	4.41	3.84	3.47	3.21	3.01	2.86	2.73	2.63	2.54	2.47	2.40	2.35	2.30	2.25	2.21	2.18	2.14	2.11	1.99	1.91	1.79	1.71	1.54	1.48	1.44
300	8.00	5.39	4.36	3.80	3.43	3.17	2.97	2.82	2.69	2.59	2.51	2.43	2.37	2.31	2.26	2.21	2.17	2.14	2.10	2.07	1.95	1.87	1.75	1.67	1.50	1.43	1.39
500	7.95	5.35	4.33	3.76	3.40	3.14	2.94	2.79	2.66	2.56	2.48	2.40	2.34	2.28	2.23	2.19	2.14	2.11	2.07	2.04	1.92	1.84	1.72	1.64	1.46	1.39	1.35
1000	7.91	5.33	4.30	3.74	3.37	3.11	2.92	2.77	2.64	2.54	2.45	2.38	2.32	2.26	2.21	2.16	2.12	2.09	2.05	2.02	1.90	1.81	1.69	1.61	1.43	1.36	1.31

$$P[\ F_{\nu 1,\nu 2} > F_{\nu 1,\nu 2,0.005}\ ] = 0.005$$

$F_{\nu 1,\nu 2,0.005}$ = 99.5ième percentile d'une variable $F_{\nu 1,\nu 2}$

avec ν_1 degrés de liberté au numérateur

ν_2 degrés de liberté au dénominateur

Exemple: $F_{10,25,0.005} = 3.54$

6.13 EXERCICES

6.1 La durée de vie X d'un certain type de pile-sèche est une variable gaussienne de moyenne 500 jours et d'écart-type 50 jours. Quelle fraction des piles survivront 580 jours? Moins de 450 jours?

6.2 Un certain type d'ampoule a une durée représentée par une variable gaussienne de moyenne 2500 heures et d'écart-type 75 heures. Résoudre les équations suivantes:

(a) $P[X \leq a] = 0.05$

(b) $P[b \leq X \leq c] = 0.95$
Remarque: puisqu'il n'y a pas de solution unique on choisira l'intervalle le plus court.

(c) $P[X \geq d] = 0.99$

6.3 Les appels à un central téléphonique arrivent suivant un processus de Poisson avec une intensité λ par minute. On sait par expérience que la probabilité de recevoir 1 appel durant une minute est égale au triple de celle de recevoir 0 appel durant la même période.

(a) Soit X le nombre d'appels reçus durant une minute. Précisez la loi de probabilité de X et calculez $P[1 \leq X \leq 4]$.

(b) Soit Y le nombre d'appels reçus durant trois minutes. Précisez la loi de probabilité de Y et calculez $P[Y \geq 4]$.

(c) Soit W_1 le temps (en minutes) d'attente jusqu'au premier appel à partir d'un instant $t = 0$. Précisez la loi de probabilité de W_1 et calculez $P[W_1 \leq 1]$.

(d) Soit W_2 le temps (en minutes) d'attente entre le premier et le deuxième appel. Précisez la loi de probabilité de W_2 et calculez $P[W_2 > 1]$.

(e) Soit W le temps d'attente jusqu'au deuxième appel à partir de l'instant $t = 0$. Précisez la loi de probabilité de W ainsi que ses paramètres.

(f) On considère 100 périodes consécutives de une minute et on note par U le nombre de ces périodes où aucun appel n'a été reçu. Précisez la loi de probabilité de U et calculez $P[U \leq 1]$.

6.4 La dureté Rockwell d'un métal est déterminée en appliquant un poinçon dur à une surface de métal et en mesurant la profondeur de la pénétration. On suppose que la dureté Rockwell X d'un alliage suit une distribution gaussienne de moyenne 70 et d'écart-type 3.

(a) Quelle est la probabilité que la dureté soit comprise entre 65 et 75 inclusivement.

(b) Pour quelle valeur de c $P[70-c \leq X \leq 70+c] = 0.95$

(c) Calculez le 5^e et le 99^e percentille de X.

(d) Soit Y le nombre d'alliages sur 10 dont la dureté est entre 65 et 75. Calculez la valeur moyenne de Y.

(e) Soit Z le nombre d'alliages sur 10 dont la dureté est inférieure à 73. Calculez $P[Z \leq 8]$.

(f) On choisit au hasard 100 spécimens et on note par W le nombre dont la dureté se situe entre 65 et 75. Quelle est la loi de probabilité de W? Précisez ses paramètres.

(g) Calculez la probabilité que $W \geq 90$ en utilisant une approximation basée sur une distribution gaussienne.

6.5 Un manufacturier produit des boulons dont le diamètre est une variable gaussienne de moyenne 1.21 cm et d'écart-type 0.02 cm. Un assemblage exige un diamètre entre 1.20 et 1.25.

(a) Quel est le pourcentage des boulons à l'extérieur des limites?

(b) À quelle valeur devrait-on fixer la moyenne afin de réduire le plus possible ce pourcentage si l'écart-type demeure inchangé?

(c) À quelle valeur peut-on fixer l'écart-type pour réduire le pourcentage à 0.10 si la moyenne reste inchangée à 1.21?

6.6 La durée de vie d'un certain composant électronique est distribuée normalement avec une moyenne de 95 heures et un écart-type de 6 heures. Cinq de ces composants sont utilisés dans un circuit et tous ces composants sont nécessaires pour faire fonctionner le circuit.

(a) Quelle est la probabilité qu'un composant dure au moins 100 heures?

(b) Quelle est la probabilité que le circuit opère plus de 90 heures?

6.7 Le pH du sol pris dans certaines régions géographiques suit une distribution gaussienne avec une moyenne de 6 et un écart-type de 0.10. Quelle est la probabilité que le pH soit:

(a) entre 5.3 et 6.15

(b) dépasse 6.1

(c) soit au plus 5.95

6.8 La force de compression X d'un certain type de béton peut se représenter par une distribution gaussienne de moyenne 4200 psi et d'écart-type 400 psi.

(a) Calculez les probabilités suivantes:

(i) $P[3000 \leq X \leq 4500]$
(ii) $P[X \geq 3268]$
(iii) $P[X \leq 5000]$

(b) Résoudre les équations suivantes:

(i) $P[X \leq x_1] = 0.75$
(ii) $P[X \geq x_2] = 0.90$

(c) La construction d'un ouvrage de grande dimension nécessite 3000 m^3 de béton. Le béton est livré par des bétonneuses de 3 m^3 et dans chaque livraison une petite quantité de béton est soumise à un test. Soit Y le nomtre de test ne dépassant pas 3268 psi.

(i) Précisez la distribution de Y et ses paramètres.
(ii) Quelle est la valeur moyenne de Y?

6.9 La consommation journalière d'électricité (en millions de Kwh) est une variable X distribuée selon une loi gamma de paramètres $\alpha = 3$ et $\beta = 2$. D'autre part, la capacité de production est de 12 millions de Kwh.

(a) Calculez la probabilité que la demande excède la capacité

(i) dans une journée
(ii) pour deux journées consécutives
(iii) au plus deux journées quelconques durant une semaine

(b) Quelle devrait être la capacité de production afin de satisfaire la demande avec une probabilité de 0.95?

6.10 Deux marques d'un appareil sont disponibles dont la durée de vie en heure est une variable gaussienne avec les paramètres.

marque	X	Y
moyenne	40	48
variance	9	36

Lequel devrait-on acheter pour avoir une durée d'au moins 30 heures?

6.11 Soient X, Y, Z et W des variables indépendantes gaussiennes dont les paramètres sont:

variable	X	Y	Z	W
moyenne	2	2	4	4
variance	3	4	4	2

On considère la nouvelle variable T définie par

$$T = X + 2Y + Z + W$$

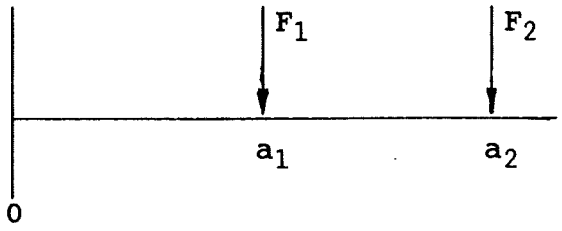
(a) Précisez la loi de probabilité de T et ainsi que ses paramètres

(b) Calculez la probabilité $P[10 \leq T \leq 14]$

(c) Répondez à la question (a) pour la variable

$$S = X - 2Y + Z - 3W$$

6.12 Deux charges F_1 et F_2 sont appliquées à une barre encastrée tel qu'illustré. Le moment fléchissant à "0" dû aux charges est $a_1F_1 + a_2F_2$.



(a) Les charges F_1 et F_2 sont des variables aléatoires indépendantes de moyenne 2 et 4 et d'écart-type 0.5 et 1.0 respectivement. Si a_1=5 et a_2=10, quelle est la moyenne du moment fléchissant ainsi que son écart-type?

(b) Si les deux charges sont dépendantes avec une corrélation de 0.5, quelle est l'écart-type du moment fléchissant?

(c) Si F_1 et F_2 suivent des distributions normales, quelle est la probabilité que le moment fléchissant dépasse la valeur de 75?

(d) Supposons que les deux charges soient placées à des distances aléatoires A_1 et A_2 de moyenne 5 et 10 respectivement et d'écart-type de 0.5. Les distances A_1 et A_2 et les charges F_1,F_2 sont des variables indépendantes. Quelle est la moyenne et l'écart-type du moment fléchissant?

6.13 Soient X_1 et X_2 deux variables indépendantes telles que:

$$X_1 \sim N(8,36) \qquad X_2 \sim N(9,25)$$

Calculez

(a) $P[\ 2X_1 - X_2 \leq 20]$

(b) $P[\ X_2 \leq X_1]$

(c) $P[\ X_1 + 2X_2 \geq 40]$

6.14 Un procédé de fabrication produit 10% d'articles défectueux. On prélève un échantillon de 200 articles au hasard et on note par X le nombre d'articles défectueux de l'échantillon. Utilisez une approximation par une loi gaussienne pour calculer les probabilités suivantes:

(a) $P[X \leq 20]$ (b) $P[X = 20]$

(c) $P[15 \leq X \leq 25]$ (d) $P[X > 25]$

6.15 Dans un contrôle de qualité en cours de réception on doit prélever un échantillon de taille n d'un lot contenant 10% de défectueux. Calculez la taille nécessaire afin que:

$$P[0.05 \leq X/n \leq 0.15] = 0.95$$

où X représente le nombre d'articles défectueux dans l'échantillon.

6.16 La probabilité qu'un disjoncteur électrique fonctionne dans des conditions normales est de 0.98. Un échantillon aléatoire de 1 000 disjoncteurs a été vérifié et 27 sont défectueux. Calculez la probabilité d'observer 27 ou plus disjoncteurs défectueux.

6.17 Il a été établi que la durée X d'un appareil suit une distribution gamma:

$$f_X(x) = x\, e^{-x} \qquad x > 0$$

On considère un échantillon aléatoire de 49 de ces appareils.

(a) Quelle est la distribution de probabilité approximative de

$$\overline{X} = \sum_{i=1}^{49} X_i/49$$

(b) Calculez $P[1.8 \leq \overline{X} \leq 2.1]$

6.18 Supposons que le gain d'un joueur est une variable aléatoire X uniforme sur [-3, 3] .

(a) Calculez approximativement $P\left[\sum_{i=1}^{100} X_i \geq 200\right]$

où X_i est le gain du i-ième jeu i=1,2,...,100

(b) Trouvez A tel que $P\left[\sum_{1=1}^{100} X_i > A\right] \geq 0.95$

(c) Trouvez n tel que $P\left[\sum_{i=1}^{n} X_i \leq 180\right] \geq 0.95$

6.19 La durée de vie en heures d'une ampoule électrique est une variable aléatoire normale de moyenne 200 et d'écart-type 40. On tire un échantillon aléatoire de 100 ampoules. Soit Y le nombre d'ampoules dont la durée de vie est supérieure à 240 et $\bar{X}$ la durée moyenne des 100 ampoules.

(a) Déterminez la distribution de probabilité de Y et ses paramètres.

(b) Déterminez la distribution de probabilité de $\bar{X}$ et ses paramètres

(c) Calculez $P[\bar{X} \leq 188]$

(d) On a observé $\bar{X} = 188$. Seriez-vous prêt à mettre en doute que la durée moyenne n'est pas 200 à la suite du calcul fait en (c)?

6.20 Une ville compte 10 000 unités d'habitation et deux usines. La demande d'eau potable en gallons requise dans une journée est une variable dont on connait les caractéristiques suivantes:

Unité	Variable	Moyenne	Ecart-type	Distribution
habitation	Q_i $i=1,2,\ldots,10\,000$	50	20	inconnue
usine 1	U_1	10 000	2 000	normale
usine 2	U_2	25 000	5 000	normale

Posons $Q_D = \sum_{i=1}^{10\,000} Q_i$ la demande domestique

$Q_T = Q_D + U_1 + U_2$ la demande totale

On suppose l'indépendance des variables Q_i, U_1, U_2.

(a) Calculez la moyenne et l'écart-type de Q_D et de Q_T

(b) Calculez, à l'aide d'une approximation utilisant le théorème central-limite, la valeur a telle que $P[Q_D \geq a] = 0.01$.

(c) Quelle devrait-être la capacité de l'usine de filtration si l'on veut satisfaire la demande totale avec une probabilité de 0.98?

6.21 Une compagnie d'assurances générales a N assurés, dont 5% auront un sinistre durant l'année. Le montant versé à un sinistré X_i est une variable de moyenne μ et d'écart-type σ. On suppose l'indépendance mutuelle des X_i et que tous les clients sont assurés pour la même valeur et paient une prime k. Soit S le montant total déboursé par la compagnie aux sinistrés et notons par P son profit.

(a) Quelle est la moyenne et la variance de S?

(b) Quelle est approximativement la distribution de probabilité de S?

(c) Exprimez P en fonction de S et précisez la distribution (approximative) de P ainsi que sa moyenne et variance.

(d) Déterminez le montant de la prime k pour que le profit de la compagnie dépasse P_o avec une probabilité de 0.999. On exprimera k en fonction de P_o, μ, σ et N.

6.22 La demande quotidienne d'énergie électrique en Kwh durant les mois d'hiver pour une maison utilisant le chauffage électrique est une variable de moyenne 200 et d'écart-type 20. Dans une ville de banlieue comptant 5000 maisons, 10% utilisent le chauffage électrique. Notons par D la demande totale d'énergie électrique pour les maisons chauffées à l'électricité.

(a) Déterminez la loi de probabilité de D et ses paramètres.

(b) Calculez une valeur de d telle que $P[D \geq d] = 0.01$

6.23 La durée T d'un certain composant électronique suit une loi exponentielle de densité

$$f_T(t) = \frac{1}{\mu} \exp(-t/\mu) \qquad t \geq 0 \quad \mu > 0.$$

(a) Calculez la probabilité que la durée moyenne $\bar{X}$ de 36 composants dépasse 125 heures si $\mu = 100$.

(b) Combien de composants doit-on avoir afin que la différence entre $\bar{X}$ et μ en valeur absolue n'excède pas 10 avec une probabilité de 0.95?

6.24 Un procédé de fabrication produit 1% de défectueux. Un échantillon de 100 est prélevé et on demande de calculer la probabilité $P(X/100 \leq 0.03)$ où X est le nombre d'articles défectueux dans l'échantillon.

6.25 Veuillez compléter le tableau suivant:

percentile	0.025	0.05	0.50	0.95
variable khi-deux 8 degrés de liberté				
variable Student 12 degrés de liberté				
variable Fisher 10 degrés de liberté numérateur 15 degrés de liberté dénominateur			□□□□□□ □□□□□□ □□□□□□ □□□□□□ □□□□□□	

6.26 Soient X_1, X_2, X_3, X_4 des variables indépendantes provenant d'une population gaussienne centrée-réduite. Posons

$$\bar{X} = \frac{1}{24} \sum_{i=1}^{4} X_i$$

$$U = (X_1-\bar{X})^2 + (X_2-\bar{X})^2 + (X_3-\bar{X})^2 + (X_4-\bar{X})^2$$

$$V = (X_1^2+X_2^2)/(X_3^2+X_4^2)$$

$$W = X_1 / \sqrt{(X_2^2+X_3^2)/2}$$

Déterminez les valeurs u_1, v_1, v_2, w_1 satisfaisant aux équations suivantes:

(a) $P[U \leq u_1] = 0.95$

(b) $P[V \geq v_1] = 0.05$
$P[V \leq v_2] = 0.05$

(c) $P[-w_1 \leq W \leq w_1] = 0.80$

6.27 Les variables Z_1, Z_2,..., Z_{10} forment un échantillon aléatoire d'une population gaussienne N(0,1). On considère les variables X, Y, W

$$X = \sum_{\alpha=1}^{6} Z_\alpha^2 \qquad Y = \sum_{\alpha=7}^{10} Z_\alpha^2 \qquad W = X/Y$$

Calculez la probabilité

$$P[0.24 \leq W \leq 13.80]$$

6.28 Soient $Z_1, Z_2, Z_3, Z_4, Z_5, Y_1, Y_2, Y_3$ des variables gaussiennes indépendantes telles que

$$Z_i \sim N(0,1) \quad i = 1,2,3,4,5$$
$$Y_j \sim N(5,25) \quad j = 1,2,3$$

Posons

$$W = \sum_{j=1}^{3} \left[\frac{Y_i - \bar{Y}}{5}\right]^2 \qquad \bar{Y} = \frac{1}{3} \sum_{i=1}^{3} Y_j$$

$$V = \sum_{j=1}^{3} \left[\frac{Y_i - 5}{5}\right]^2 \qquad U = \sum_{i=1}^{5} Z_i^2$$

Résoudre les équations suivantes:

(a) $P[W \leq a] = 0.95$

(b) $P\left[\frac{U}{V} \geq b\right] = 0.01$

6.29 Soient $X_1, X_2, \ldots, X_6$ des variables indépendantes gaussiennes telles que

$$X_i \sim N(0,1) \quad i = 1,2,3$$
$$X_i \sim N(12,4) \quad i = 4,5,6$$

On pose

$$\bar{X} = \frac{1}{3}(X_1 + X_2 + X_3)$$

$$H = (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + (X_3 - \bar{X})^2$$

$$G = \sum_{i=4}^{6} \left[\frac{X_i - 12}{2}\right]^2 \qquad J = X_1 / \sqrt{G/3}$$

Résoudre les équations suivantes:

(a) $P[H \leq a] = 0.95$

(b) $P[H/G \leq b] = 0.01$

(c) $P[-c \leq J \leq c] = 0.90$

6.30 Une variable X est distribuée $N(\mu,\sigma^2)$. On définit la déviation moyenne D par:

$$D = E[\ |x-\mu|\ ]$$

Montrez que $D = \sigma \left(\frac{2}{\pi} \right)^{0.5} \approx 0.8\ \sigma$

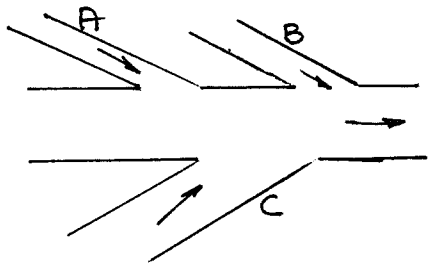
6.31 Un club de personnes décide de mettre sur pied un casino d'un soir dont les profits seront versés à une oeuvre de charité. Les organisateurs ont décidé

- de demander à leurs membres de défrayer les coûts fixes
- d'admettre 1000 joueurs seulement, chacun avec une même mise initiale θ (en milliers de dollars) au début de la soirée
- de choisir des jeux tels que le gain brut X_i (en milliers de dollars) du i-ème joueur soit distribué uniformément sur l'intervalle $(0, \frac{3}{2}\theta)$.

(a) Soit Y le gain brut total des 1000 joueurs. Précisez la distribution approximative de Y ainsi que ses paramètres.

(b) Déterminez le montant θ que doit payer chaque joueur afin que le profit net (en milliers de dollars) du casino soit supérieur à 50 avec une probabilité de 0.95.

6.32 Une voie rapide de circulation automobile à trois voies d'accès, disons A, B, C (voir figure).



Le nombre de voitures accédant à la voie rapide durant une période de 1 heure est définie par des variables notées X_A, X_B, X_C ayant les caractéristiques suivantes:

	X_A	X_B	X_C
moyenne	800	1000	600
écart-type	40	50	30

Notons par X le nombre total de voitures accédant à la voie rapide durant une période de 1 heure.

(a) Calculez

(i) la moyenne de X

(ii) l'écart-type de X en supposant que les variables X_A, X_B, X_C sont 2 à 2 indépendantes

(iii) la probabilité que X soit comprise entre 2300 et 2500 si l'on suppose que les variables X_A, X_B, X_C sont indépendantes et distribuées normalement

(iv) la probabilité que X soit supérieure à 2 500 sous les mêmes hypothèses que (iii)

(b) Soit Y le nombre de fois que $X \geq 2\,500$ durant 100 périodes de 1 heure

(i) précisez la distribution de Y et ses paramètres

(ii) calculez, en utilisant une approximation basée sur la distribution normale, la probabilité que la variable Y soit supérieure ou égale à 10.

(c) Si l'on suppose que les variables X_A, X_B, X_C sont distribuées normalement et que les coefficients de corrélation sont:

$$\begin{aligned} 0.5 &\text{ entre } X_A \text{ et } X_B \\ 0.8 &\text{ entre } X_A \text{ et } X_C \\ -0.5 &\text{ entre } X_B \text{ et } X_C \end{aligned}$$

Calculez la moyenne et l'écart-type de X.

6.33 Déterminez la taille n d'un échantillon aléatoire provenant d'une population $N(\mu, \sigma^2)$ de telle sorte que

$$P\left[\left| \frac{\bar{X} - \mu}{S/\sqrt{n}} \right| < 2 \right] = 0.95$$

où

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad , \quad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

6.34 Une variable X de distribution log-normale satisfait les équations suivantes:

$$P[X \leq 10] = 0.0062$$
$$P[X \geq 100] = 0.6915$$

Déterminez ses paramètres, sa moyenne, sa variance, sa médiane et le 90-ième percentile.

6.35 Un industriel vend un article au prix fixe v. Il rembourse le prix d'achat à tout acheteur qui constate que le poids W de l'article est inférieur à un poids donné w_0 et il récupère l'article dont la valeur de la matière première utilisable est p. La distribution de W est normale de moyenne μ et d'écart-type σ. Un réglage adéquat permet de fixer μ à n'importe quelle valeur désirable mais σ n'est pas réglable. Le prix de revient R est une fonction du poids de l'article:

$$R = \alpha + \beta W$$

(a) Déterminez l'expression du bénéfice B(W).

(b) Déterminez le bénéfice moyen $B(\mu)$.

(c) Trouvez la valeur μ_0 qui maximise $B(\mu)$.

(d) <u>Application numérique</u>

$\nu = 10$ \$	$w_0 = 8$ kg	$\sigma = 1$ kg
$\alpha = 3$	$\beta = 0.05$	p = 2 \$

6.36 Le temps moyen entre deux pannes consécutives est de 100 heures. Si les pannes sont distribuées selon un processus de Poisson, calculez la probabilité:

(a) d'au moins 1 panne dans des périodes de 1 100 et 1 000 heures.

(b) d'exactement 1 panne dans des périodes de 1 100 et 1 000 heures.

6.37 Une variable X suit une distribution de Rayleigh et a pour fonction de densité:

$$f_X(x) = \frac{x}{\alpha^2} e^{-x^2/2\alpha^2} \qquad x \geq 0$$

$$= 0 \qquad x < 0$$

(a) Montrez que la fonction de répartition est:

$$F_X(x) = 1 - e^{-x^2/2\alpha^2}$$

(b) Déterminez une expression pour le p-ième percentile.

(c) Montrez que la moyenne et l'écart-type sont:

$$E(X) = \alpha \left(\frac{\pi}{2} \right)^{0.5}$$

$$ET(X) = \alpha \left(\frac{4-\pi}{2} \right)^{0.5}$$

6.38 Soit $X_1, X_2, \ldots, X_n$ un échantillon aléatoire provenant d'une population $N(\mu, \sigma^2)$. Posons

$$\bar{X}_k = \frac{1}{k} \sum_{i=1}^{k} X_i \qquad S_k^2 = \frac{1}{k-1} \sum_{i=1}^{k} (X_i - \bar{X}_k)^2$$

$$\bar{X}_{n-k} = \frac{1}{n-k} \sum_{i=k+1}^{n} X_i \qquad S_{n-k}^2 = \frac{1}{n-k-1} \sum_{i=k+1}^{n} (X_i - \bar{X}_{n-k})^2$$

Quelle est la distribution de:

(a) $\frac{1}{2} (\bar{X}_k + \bar{X}_{n-k}) = V$

(b) $\frac{1}{\sigma^2} \left[(k-1) S_k^2 + (n-k-1) S_{n-k}^2 \right] = V$

(c) $S_k^2 / S_{n-k}^2 = W$

6.39 On veut investir un capital de N dollars dans un ensemble de n actions dont les rendements (%) sont représentés par n variables indépendantes $X_1, X_2, \ldots, X_n$ de moyennes $\mu_1, \mu_2, \ldots, \mu_n$ et variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$. On investit un total de N_j dollars dans l'action j donnant un rendement R pour le portefeuille:

$$R = \sum_{i=1}^{n} N_j X_j \quad , \quad \sum_{j=1}^{n} N_j = N$$

(a) Déterminez le rendement moyen et l'écart-type associé à un portefeuille.

(b) On veut distribuer un capital de 10 000 $ entre deux actions de rendement:

$$\mu_1 = 10\% \qquad \sigma_1 = 2\%$$
$$\mu_2 = 20\% \qquad \sigma_2 = 12\%$$

Si on veut maintenir un écart-type minimum, quel montant N_1, N_2 doit-on investir dans chacune des actions?

6.40 La quantité X (en millions de litres) de bière vendue pendant une semaine durant l'été est distribuée selon une loi gamma de moyenne 18 et d'écart-type 6.

(a) Déterminez les paramètres α, β de la distribution.

(b) Calculez les probabilités:

(i) $P[X \leq 7]$
(ii) $P[X \geq 24]$
(iii) $P[14 \leq X \leq 32]$

6.14 RÉPONSES EXERCICES

6.1 (a) 0.55 (b) 0.1587

6.2 (a) a = 2376.61 (b) b = 2352.97 , c = 2647.02
(c) d = 2325.25

6.3 (a) 0.765 (b) 0.979 (c) 0.9502 (d) 0.0478
(e) gamma ($\alpha = 2$, $\beta = 1/3$) (f) 0.0377

6.4 (a) 0.905 (c) c = 5.88 (c) 65.08 et 76.99
(d) 9.05 (e) 0.492 (f) bin (n = 100, $\theta = 0.905$)

6.5 (a) 33.1% (b) 1.225 (c) 0.0078

6.6 (a) 0.7967 (b) 0.321

6.7 (a) 0.933 (b) 0.16 (c) 0.3085

6.8 (a) 0.7721, 0.99, 0.9772 (b) 4468, 4712
(c) bin (n = 1000, $\theta = 0.99$), 990

6.9 (a) 0.062, 0.0038, 0.991 (b) 12.55

6.10 On achète Y.

6.11 (a) N(14,5) (b) 0.288 (c) N(-12,41)

6.12 (a) 50, 10.30 (b) 10.31 (c) 0.008
(d) 50, 10.55

6.13 (a) 0.84 (b) 0.45 (c) 0.12

6.14 (a) 0.548 (b) 0.095 (c) 0.806 (d) 0.145

6.15 118

6.16 0.071

6.17 (a) N(2,2/49) (b) 0.524

6.18 (a) 0 (b) -28.49 (c) 3992

6.19 (a) bin (n = 100, $\theta = 0.16$) (b) N(200,16)
(c) 0.0013 (d) oui

6.20 (a) 500 000, 2 000 (b) 504 660 (c) 546 800

6.21 (a) $\frac{N}{20}\mu$, $\frac{N}{20}\sigma^2$ (b) Normale

(c) $N\left(Nk - \frac{N}{20}\mu , \frac{N}{20}\sigma^2\right)$ (d) $\frac{\mu}{20} + \frac{Po}{N} + \frac{3.09\sigma}{\sqrt{20N}}$

6.22 (a) N(100 000, 200 000) (b) 101 042

6.23 (a) 0.067 (b) 384

6.24 0.9816

6.25 khi-deux: 2.18, 2.73, 7.34, 15.51

T_{12}: -2.178, -1.782, 0, 1.782

$F_{10,15}$: 0.284, 0.351, X, 2.54

6.26 (a) 7.815 (b) 19, 0.053 (c) 1.886

6.27 0.975

6.28 (a) 5.991 (b) 47

6.29 (a) 5.99 (b) 0.0067 (c) 2.353

6.31 (a) $N(750\theta, 187.5\theta^2)$ (b) 219,80 $

6.32 (a) 2400, 70.71, 0.8414, 0.08
(b) Bin (n = 100, θ = 0.0792), 0.2802
(c) 2400, 86.14, 0.7540

6.33 n = 61

6.34 X ~ LN (ξ = 1.42, τ^2 = 0.032)
E(X) = 4.22, ET(X) = 0.76
médiane = 4.14, $x_{0.90}$ = 4.31

6.35 (a) $B(W) = p-\alpha-\beta W$ si $W < W_0$

$= v-\alpha-\beta W$ si $W > W_0$

(b) $B(\mu) = v - \alpha - \beta\mu + (p+\alpha-2v)\,\Phi\left(\frac{W_0-\mu}{\sigma}\right)$

(c) $\mu = W_0 + \sigma\left[\ln\left(\frac{p+\alpha-2v}{\beta\sigma^2\sqrt{2\pi}}\right)^2\right]^{1/2}$

(d) μ = 11.09

6.36 (a) 0.01, 0.632, 0.995
(b) 0.0099, 0.368, 0.0045

6.37 (b) $x_p = \alpha\sqrt{2\ell n(1-p)}$

6.38 (a) $V \sim N\left(\mu, \frac{\tau^2}{4}\left(\frac{1}{k} + \frac{1}{n-k}\right)\right)$

(b) $U \sim X^2_{n-2}$

(c) $W \sim F_{k-1,n-k-1}$

6.39 (a) $E(R) = \Sigma N_j \mu_j$, $VAR(R) = \Sigma N_j^2 \sigma_j^2$

(b) $N_1 = 9\ 730\ \$$, $N_2 = 270\ \$$

6.40 (a) $\alpha = 9$, $\beta = 2$
(b) 0.01, 0.155, 0.719

CHAPITRE 7

ESTIMATION

7.0 SOMMAIRE

Ce chapitre est consacré au problème de l'estimation statistique des paramètres de distributions et des paramètres apparaissant dans les équations liant une variable expliquée et un ensemble de variables explicatives. Les propriétés des estimateurs et les principales méthodes d'estimation sont présentés ainsi que le problème du calcul de la taille échantillonnale. Des exemples d'application illustrent les principales méthodes.

7.1 DÉFINITION DU PROBLÈME

On convient de distinguer trois types de problèmes en analyse statistique des données:

- la description des données
- l'estimation des paramètres
- les tests d'hypothèses

Les deux derniers types de problèmes constituent la partie inférentielle de l'analyse statistique: généralisation de l'échantillon à la population. Ce chapitre est consacré à l'estimation et les tests seront vus au chapitre suivant.

Nous employons le terme paramètre pour désigner

- les paramètres des distributions: $N(\mu,\sigma^2)$, $Gamma(\alpha,\beta)$, $exponentielle(\lambda)$, $bêta(\alpha,\beta)$, $Poisson(\lambda)$, $binomiale(n,\theta)$, hypergéométrique(N,D,n) ou autre.

- les paramètres dérivés des distributions: la moyenne, l'écart-type, coefficient d'asymétrie, coefficient d'aplatissement, le coefficient de corrélation entre deux variables, les coefficients de corrélation entre plusieurs variables, etc...

. les paramètres apparaissant dans des équations liant une variable expliquée, disons Y, et un ensemble de variables explicatives $X_1, X_2, \ldots, X_p$. Ces équations sont de la forme générale suivante:

$$Y = \varphi(X_1, X_2, \ldots, X_p; \beta_0, \beta_1, \ldots, \beta_k) + \varepsilon$$

où φ est une fonction de forme connue, $\beta_0, \beta_1, \ldots, \beta_k$ sont des paramètres inconnus et ε est le terme d'erreur. Nous exposerons, lors de l'étude de la méthode des moindres carrés, la provenance et l'utilité de ce type de modèle statistique.

Les modèles (distributions et/ou équations) sont des classes de fonctions paramètrées suffisamment riches pour qu'ils puissent représenter mathématiquement les phénomènes réels. Cela pose le problème de l'estimation des paramètres à partir des données engendrées par l'observation ou l'échantillonnage du phénomène. La classe de modèles à choisir dépend de l'objectif de l'étude qui a conduit à la collecte des données et est la responsabilité de celui qui en fait l'analyse.

On dispose d'un échantillon aléatoire $X_1, X_2, \ldots, X_n$ (variables aléatoires indépendantes et identiquement distribuées) et on désire estimer un (ou plusieurs) paramètre θ_o inconnu. On distingue deux sortes d'estimateur:

. ESTIMATEUR PONCTUEL de θ, noté $\hat{\theta}$

$\hat{\theta}$ est une fonction $\hat{\theta} = H(X_1, \ldots, X_n)$ à valeurs réelles satisfaisant certaines propriétés. Lorsque l'on dispose des données $x_1, x_2, \ldots, x_n$ la valeur de la fonction H constitue une valeur numérique

$$\hat{\theta} = H(x_1, \ldots, x_n)$$

estimant θ_o. Il est important de constater que $\hat{\theta}$ l'estimateur est une variable aléatoire et est donc distribuée selon une loi de probabilité d'échantillonnage. Il fait donc du sens de se référer à la moyenne de $\hat{\theta}$ et la variance de $\hat{\theta}$. C'est d'ailleurs en se référant à cette loi d'échantillonnage que nous proposerons de définir les propriétés des estimateurs.

Il est aussi important de distinguer l'estimateur $\hat{\theta}$ et sa réalisation $\hat{\theta}$ (abus de notation!). C'est la distinction entre une fonction et la valeur de la fonction; le contexte permet de distinguer ce dont il s'agit.

. ESTIMATEUR PAR INTERVALLE

Comme son nom l'indique, il s'agit d'un intervalle, disons (a,b), calculé de telle sorte que $\theta_o \varepsilon (a,b)$ avec une probabilité assez élevée. Cette probabilité est appelée coefficient de confiance et représente une fréquence à long terme que l'intervalle ainsi calculé est de la classe des intervalles qui contiennent le paramètre θ_o. Le coefficient de confiance est généralement supérieur à 0.90 et est fixé par l'utilisateur. Les valeurs de a et b dépendent de l'échantillon $X_1, X_2, \ldots, X_n$ et du coefficient de confiance.

7.2 PROPRIÉTÉS DES ESTIMATEURS PONCTUELS

Soit $\hat{\theta} = H(X_1,\ldots,X_n)$ un estimateur ponctuel et désignons par $f_{\hat{\theta}}(\theta)$ sa distribution d'échantillonnage (densité disons).

Nous utilisons la notation $\hat{\theta}$ pour désigner un estimateur et notons par θ_o la "vraie" valeur du paramètre θ. On juge de la qualité d'un estimateur selon les propriétés suivantes dérivant de la distribution d'échantillonnage illustrée à la figure 7.1.

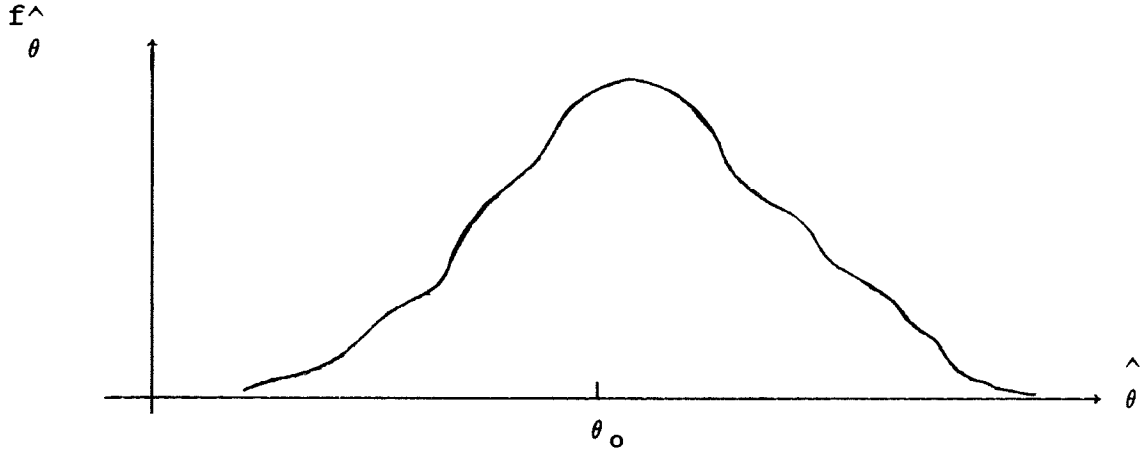


Figure 7.1: distribution d'échantillonnage de $\hat{\theta}$

Biais ou erreur systématique

L'estimateur $\hat{\theta}$ ne commet pas D'ERREUR SYSTÉMATIQUE ou est SANS BIAIS pour le paramètre θ_o si la valeur moyenne de $\hat{\theta}$ est égale à θ_o c'est-à-dire,

$$E(\hat{\theta}) = \int_{\theta} \hat{\theta} f_{\hat{\theta}}(\hat{\theta}) d\hat{\theta} = \theta_o \qquad (7.1)$$

où E(.) dénote l'espérance mathématique selon la distribution $f_{\hat{\theta}}(.)$. Le BIAIS d'un estimateur est défini par

$$BIAIS(\hat{\theta}) = E(\hat{\theta}) - \theta_o \qquad (7.2)$$

On dit aussi que le biais est une erreur systématique puisque la différence entre $\hat{\theta}$ et θ_0 ne s'annulle pas en moyenne. Il est donc généralement désirable de choisir des estimateurs sans biais. La propriété de non-biais est surtout importante pour des échantillons de très petite taille (disons $n \leq 10$) car, sous certaines conditions, les estimateurs biaisés (biais $\neq 0$) voient leurs biais tendre vers zéro lorsque $n \to \infty$.

Exemple 7.1: Estimation de la moyenne d'une population

Soit $\theta = \mu = E(X)$ la moyenne d'une population X. Estimons θ par

$$\hat{\theta} = H(X_1,\ldots,X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

D'après l'équation (4.57) du chapitre 4, on a

$$E(\hat{\theta}) = E(\bar{X}) = \mu = \theta$$

Donc $\bar{X}$ constitue un estimateur sans biais de la moyenne. Ce résultat est vrai quelque soit la distribution de X.

Exemple 7.2: Estimation de la variance d'une population $N(\mu,\sigma^2)$ avec moyenne inconnue

Soit $\theta = \sigma^2 = VAR(X)$ la variance d'une population. Estimons θ par

$$\hat{\theta}^2 = H(X_1,\ldots,X_n) = \frac{1}{n-1}\sum_{i=1}^{n} (X_i-\bar{X})^2 = S^2$$

À la section 6.5 nous avons vu que $\frac{(n-1)}{\sigma^2} S^2$ suit la distribution χ^2_{n-1} et que $E(\chi^2_{n-1}) = n-1$. Donc

$$E(\hat{\theta}) = E(S^2) = E\left[\frac{n-1}{\sigma^2} S^2 \frac{\sigma^2}{n-1}\right]$$
$$= \frac{\sigma^2}{n-1} E(\chi^2_{n-1}) = \sigma^2$$

C'est pour avoir un estimateur sans biais de σ^2 que nous utilisons S^2 plutôt que l'estimateur

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 = V^2 = \left(\frac{n-1}{n}\right)S^2$$

En effet, cet estimateur est biaisé puisque

$$E(V^2) = E\left(\frac{(n-1)S^2}{n}\right) = \frac{(n-1)}{n}E(S^2)$$

$$= \frac{n-1}{n}\sigma^2 = \sigma^2 - \sigma^2/n$$

$$\text{BIAIS}(V^2) = -\sigma^2/n$$

Cet estimateur sous-estime σ^2 avec une erreur systématique de σ^2/n. On note que le biais tend vers zéro lorsque $n\to\infty$ illustrant une remarque faite ci-haut.

Exemple 7.3: Estimation de la variance d'une population $N(\mu,\sigma^2)$ avec moyenne μ connue

Soit $\theta = \sigma^2$ le paramètre à estimer et proposons l'estimateur

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$$

Montrons que $\hat{\theta}$ est sans biais. En effet

$$\frac{n\hat{\theta}}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$$ suit une loi χ^2_n d'après la section 6.5.

Donc

$$E(\hat{\theta}) = E\left[\frac{n\hat{\theta}}{\sigma^2}\frac{\sigma^2}{n}\right] = \frac{\sigma^2}{n}E(\chi^2_n) = \sigma^2$$

Exemple 7.4: Estimation de l'écart-type σ d'une population $N(\mu,\sigma^2)$ avec moyenne μ inconnue

Posons $\theta = \sigma$ le paramètre à estimer et proposons l'estimateur

$$\hat{\theta} = S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

On peut montrer que

$$E(S) = k_n\sigma \qquad \text{où } k_n = \sqrt{\frac{2}{n-1}}\,\frac{\Gamma(n/2)}{\Gamma((n-1)/2)} \tag{7.3}$$

Le tableau 7.1 donne quelques valeurs de k_n.

Tableau 7.1: Valeurs de k_n

n	2	3	4	5	10	15	20	30
k_n	0.80	0.89	0.92	0.94	0.97	0.982	0.986	0.991

Donc S n'est pas un estimateur sans biais de σ.

En contrôle de qualité, un autre estimateur basé sur l'étendue est proposé. Il est conseillé de ne l'employer que pour de petits échantillons seulement (disons $n \leq 10$). Il s'agit de

$$\hat{\sigma} = \frac{R}{d_n}$$

où $$R = MAX(X_1,X_2,\ldots,X_n) - MIN(X_1,X_2,\ldots,X_n)$$
$$= X_{(n)} - X_{(1)}$$

On montre que

$$E(R) = d_n\sigma$$

où $$d_n = \int_{-\infty}^{\infty}[1 - (1 - \Phi(x))^n - (\Phi(x))^n]dx \tag{7.4}$$

Il suit

$$E(\hat{\sigma}) = E\left[\frac{R}{d_n}\right] = \sigma$$

La valeur de d_n a été calculée de telle sorte que R/d_n constitue une estimation sans biais de σ.

Le tableau 7.2 donne quelques valeurs de d_n

Tableau 7.2: Valeurs de d_n

n	2	3	4	5	6	7	8	9
d_n	1.128	1.693	2.059	2.326	2.534	2.704	2.847	2.970

n	10	11	12	13	14	15	20	25
d_n	3.078	3.173	3.258	3.360	3.407	3.472	3.735	3.931

Précision ou erreur d'échantillonnage

La précision d'un estimateur est définie comme étant la propriété de donner des valeurs ayant peu de dispersion. On convient donc de mesurer la précision d'un estimateur à l'aide de sa variance:

$$VAR(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2 \qquad (7.5)$$

$$= \int_{-\infty}^{\infty} (\hat{\theta} - E(\hat{\theta}))^2 \, f_{\hat{\theta}}(\hat{\theta}) d\hat{\theta}$$

ou de son écart-type

$$ET(\hat{\theta}) = \sqrt{VAR(\hat{\theta})} \qquad (7.6)$$

D'une manière générale, nous verrons que les estimateurs utilisés dans les applications vérifient

$$ET(\hat{\theta}) = \frac{c}{\sqrt{n}} \qquad (7.7)$$

où c est une constante qui ne dépend pas de n.

Erreur quadratique moyenne

Une mesure de précision globale tenant compte du biais et de la variance est l'erreur quadratique moyenne notée $EQM(\hat{\theta})$. Elle est définie par

$$EQM(\hat{\theta}) = E[(\hat{\theta} - \theta_o)^2] \tag{7.8}$$

$$= \int_{-\infty}^{\infty} (\hat{\theta} - \theta_o)^2 f_{\hat{\theta}}(\hat{\theta}) d\hat{\theta}$$

On peut décomposer l'erreur quadratique moyenne selon

$$EQM(\hat{\theta}) = E[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta_o)^2]$$

$$= E[(\hat{\theta} - E(\hat{\theta}))^2] + E[(E(\hat{\theta}) - \theta_o)^2]$$

$$+ 2E[(\hat{\theta} - E(\hat{\theta}))][(E(\hat{\theta}) - \theta_o)]$$

Mais

$$E[(E(\hat{\theta}) - \theta_o)]^2 = [E(\hat{\theta}) - \theta_o]^2 = (BIAIS(\hat{\theta}))^2$$

et

$$E[\hat{\theta} - E(\hat{\theta})][(E(\hat{\theta}) - \theta_o)] = [E(\hat{\theta}) - E(\hat{\theta})][E(\hat{\theta}) - \theta_o] = 0$$

donc

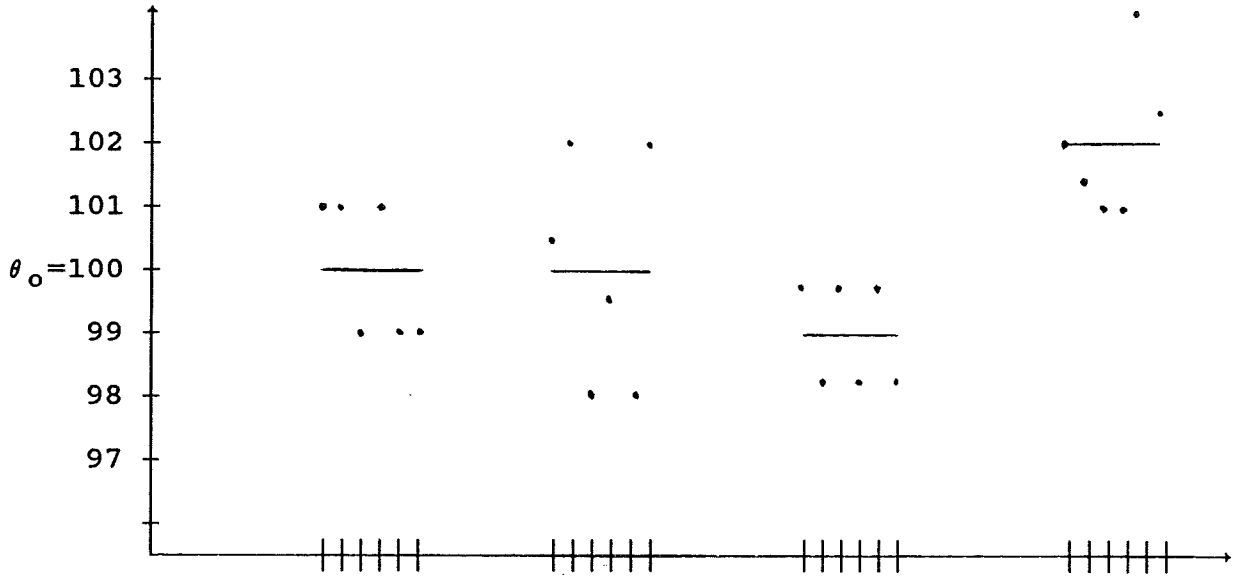
$$EQM(\hat{\theta}) = VAR(\hat{\theta}) + (BIAIS(\hat{\theta}))^2 \tag{7.9}$$

En particulier

$$EQM(\hat{\theta}) = VAR(\hat{\theta}) \quad \text{si } E(\hat{\theta}) = \theta_o \tag{7.10}$$

Les différentes définitions concernant les propriétés des estimateurs sont illustrées à la figure 7.2. On dispose de quatre estimateurs $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4$ pour estimer un paramètre dont la valeur serait égale à 100 disons et on illustre un comportement hypothétique des estimateurs pour ~~six~~ 4 échantillons de taille n.

Figure 7.2 Comportement de 4 estimateurs



Echantillon de taille n	123456	123456	123456	123456
estimateur	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
valeurs hypothétiques	101,101 99,101 99,99	100.5,102 98,99.5 98,102	99.8,98.2 99.8,98.2 99.8,98.2	102,101.5 101,101 104,102.5
moyenne	100	100	99	102
biais	0	0	-1	2
variance	6/5	16.5/5	3.84/5	6.5/5
Erreur quadratique moyenne	6/5	16.5/5	8.84/5	26.5/5

Exemple 7.5: Estimation de la moyenne μ d'une population de variance σ^2

On a vu que l'estimateur $\hat{\theta} = \bar{X}$ est sans biais pour le paramètre μ

$$E(\bar{X}) = \mu$$

Donc

$$EQM(\bar{X}) = VAR(\bar{X}) = E(\bar{X} - \mu)^2 = \sigma^2/n \qquad (7.11)$$

et

$$ET(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

où ET(.) désigne l'écart-type

Exemple 7.6: Estimation du paramètre θ d'une population Bernoulli

La population est définie par une masse de probabilité

$$P_X(x;\theta) = \theta^x(1 - \theta)^{1-x} \quad x = 0,1 \quad , \quad 0 < \theta < 1$$

La moyenne et la variance de cette population est

$$\mu = E(X) = 0*(1-\theta) + 1 * \theta = \theta$$

$$\sigma^2 = VAR(X) = (0 - \theta)^2*(1 - \theta) + (1 - \theta)^2*\theta = \theta(1 - \theta)$$

Proposons d'estimer θ par $\bar{X}$. Alors

$$E(\bar{X}) = \theta$$

et

$$EQM(\bar{X}) = VAR(\bar{X}) = E[\bar{X} - \theta]^2 = \frac{\sigma^2}{n} = \frac{\theta(1 - \theta)}{n} \qquad (7.12)$$

Exemple 7.7: Estimation de la variance σ^2 de $N(\mu,\sigma^2)$

Nous proposons d'estimer $\theta = \sigma^2$ à l'aide de la variance échantillonnale

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \hat{\theta}$$

Nous avons vu à l'exemple 7.2 que S^2 est un estimateur sans biais. D'autre part S^2 est une variable χ^2_{n-1} à une constante constante multiplicative près

$$S^2 = \frac{\sigma^2}{n-1} \frac{n-1}{\sigma^2} \quad S^2 = \left[\frac{\sigma^2}{n-1}\right] \chi^2_{n-1}$$

Alors

$$EQM(S^2) = VAR(S^2)$$

$$= \frac{\sigma^4}{(n-1)^2} VAR(\chi^2_{n-1}) = \frac{2\sigma^4}{n-1} \qquad (7.13)$$

$$ET(S^2) = \sqrt{\frac{2}{n-1}}\, \sigma^2$$

Exemple 7.8: Estimation de l'écart-type σ de $N(\mu,\sigma^2)$

À l'exemple 7.4 nous avons proposé l'écart-type échantillonnal $S = \hat{\theta}$ pour estimer l'écart-type $\sigma = \theta$. Sous hypothèse de normalité, on peut montrer que

$$VAR(S) = (1 - k_n^2)\sigma^2 \approx \frac{\sigma^2}{2n} \qquad (7.14)$$

ou k_n est définie par l'équation (7.3)

Donc

$$ET(S) \approx \frac{\sigma}{\sqrt{2n}} \qquad (7.15)$$

En résumé, on juge la qualité d'un estimateur $\hat{\theta}$ par ses deux premiers moments relativement à sa distribution d'échantillonnage. Idéalement, on recherche (s'il existe) l'estimateur sans biais à variance minimale, c'est le "meilleur" estimateur.

Nous allons maintenant exposer les différentes méthodes d'estimation:

- méthode de vraisemblance maximale
- méthode des moments
- méthode des moindres carrés
- méthode des intervalles de confiance.

7.3 MÉTHODE DE VRAISEMBLANCE MAXIMALE

Principe

La méthode de vraisemblance maximale est faite pour estimer les paramètres des distributions et suppose donc la connaissance de la forme de la distribution. Elle peut s'employer pour des distributions unidimentionnelles ou multidimensionnelles ainsi que pour des paramètres unidimentionnels ou multidimentionnels. Nous allons présenter la méthode pour le cas des distributions unidimentionnelles.

Soit X une variable aléatoire continue (discrète) et $f_X(x;\theta_1,\theta_2,\ldots,\theta_p)$ sa densité (masse) de probabilité. La forme de la fonction $f_X(.)$ est connue mais on suppose les paramètres $\theta_1,\theta_2,\ldots,\theta_p$ inconnus. Soit $x_1,x_2,\ldots,x_n$ la réalisation d'un échantillon aléatoire $X_1,X_2,\ldots,X_n$ et définissons la fonction $L(\theta_1,\theta_2,\ldots,\theta_p)$ appelée FONCTION DE VRAISEMBLANCE par

$$L(\theta_1,\theta_2,\ldots,\theta_p) = \prod_{i=1}^{n} f_X(x_i;\theta_1,\theta_2,\ldots,\theta_p) \qquad (7.16)$$

La fonction de vraisemblance L(.) est définie dans un domaine appelé l'espace des paramètres et l'objectif de la méthode d'estimation consiste à rechercher la valeur des paramètres qui maximise la fonction L(.), d'où son nom. Voici quelques exemples de fonction de vraisemblance.

Exemple 7.9: distribution de Bernoulli

On a $f_X(x;\theta) = \theta^x(1-\theta)^{1-x}$, $x = 0,1$, $0 < \theta < 1$

$$L(\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i} = \theta^{n\bar{x}}(1-\theta)^{n(1-\bar{x})} \qquad (7.17)$$

Exemple 7.10: distribution de Poisson

On a $f_X(x;\theta) = \exp(-\theta)\dfrac{\theta^x}{x!}$, $x = 0,1,2,\ldots$, $\theta > 0$

$$L(\theta) = \prod_{i=1}^{n} \exp(-\theta) \frac{\theta^{x_i}}{x_i!} = \exp(-n\theta)\ \theta^{n\bar{x}} \prod_{i=1}^{n} \left[\frac{1}{x_i!}\right] \qquad (7.18)$$

Exemple 7.11: distribution gaussienne $N(\mu,\sigma^2) = N(\theta_1,\theta_2)$

$$\text{On a } f_X(x;\theta_1,\theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} \exp\left[-\frac{1}{2\theta_2} (x - \theta_1)^2 \right]$$

$$-\infty < x < \infty\ ,\quad -\infty < \theta_1 < \infty\ ,\quad \theta_2 > 0$$

$$L(\theta_1,\theta_2) = \prod_{i=1}^{n} f_X(x_i;\theta_1,\theta_2)$$

$$= (2\pi\theta_2)^{-n/2} \exp\left[-\frac{1}{2\theta_2} \sum_{i=1}^{n} (x_i - \theta_1)^2\right] \qquad (7.19)$$

Définition des estimateurs à vraisemblance maximale

Les estimateurs à vraisemblance maximale de $\theta_1,\theta_2,\ldots,\theta_p$ notés $\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_p$ sont les valeurs de $\theta_1,\theta_2,\ldots,\theta_p$ qui maximisent la fonction de vraisemblance $L(\theta_1,\theta_2\ldots,\theta_p)$ c'est-à-dire

$$L(\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_p) \geq L(\theta_1,\theta_2,\ldots\theta_p)$$

dans l'espace des paramètres.

Remarques

- Pour les distributions usuelles, les estimateurs sont uniques.
- La recherche du maximum de L(.) peut se faire par le moyen usuel de l'annulation de la dérivée lorsque la fonction est dérivable et que $\theta_1,\ldots,\theta_p$ varient sur un continum.
- Il peut arriver que le maximum soit à la frontière de l'espace des paramètres auquel cas, l'annulation de la dérivée ne peut s'appliquer.

. Il est plus commode de résoudre le problème de la recherche du maximum avec le logarithme de L(.)

$$\ell n\ L(\theta_1,\ldots,\theta_p) = \ell n\Big(\prod_{i=1}^{n} f_X(x_i;\theta_1,\ldots,\theta_p)\Big)$$

$$= \sum_{i=1}^{n} \ell n\ f_X(x_i;\theta_1,\ldots,\theta_p) \qquad (7.20)$$

La fonction logarithme étant monotone, les mêmes valeurs $\hat{\theta}_1,\ldots,\hat{\theta}_p$ maximisent $\ell n\ L(.)$ et $L(.)$

. Les étapes qui conduisent à l'obtention des estimateurs à vraisemblance maximale sont

- Définition de la fonction $f_X(x;\theta_1,\ldots,\theta_p)$
- Calcul de la fonction de vraisemblance

$$L(\theta_1,\ldots,\theta_p) = \prod_{i=1}^{n} f_X(x_i;\theta_1,\ldots,\theta_p)$$

basée sur un échantillon de taille n, $x_1,x_2,\ldots,x_n$

- Calcul du logarithme de L(.)

$$\ell nL(\theta_1,\ldots,\theta_p) = \sum_{i=1}^{n} \ell nf_X(x_i;\theta_1,\ldots,\theta_p)$$

- Établissement des équations à résoudre

$$\frac{\partial}{\partial\theta_\alpha}\ \ell n\ L(\theta_1,\ldots,\theta_p) = 0 \qquad \alpha = 1,\ldots,p \qquad (7.21)$$

- Résolution des équations

$$\hat{\theta}_\alpha = \hat{\theta}_\alpha(x_1,x_2,\ldots,x_n) \qquad \alpha = 1,\ldots,p \qquad (7.22)$$

Certaines distributions e.g. bêta, gamma, Weibull conduisent à un système d'équations difficile à résoudre analytiquement et il faut alors recourir à des méthodes numériques itératives.

Applications

Nous allons illustrer la méthode avec quelques exemples, et proposer un tableau résumant l'application de la méthode à plusieurs distributions.

Exemple 7.12: distribution de Bernoulli

La fonction de vraisemblance a été calculée à l'équation (7.17)

$$L(\theta) = \theta^{\Sigma x_i} (1 - \theta)^{n - \Sigma x_i}$$

$$\ell n\ L(\theta) = \left[\sum_{i=1}^{n} x_i\right] \ell n\ \theta + \left[n - \sum_{i=1}^{n} x_i\right] \ell n\ (1 - \theta)$$

$$\frac{d}{d\theta}\ \ell n\ \ L(\theta) = \sum_{i=1}^{n} x_i \left[\frac{1}{\theta}\right] + \left[n - \sum_{i=1}^{n} x_i\right] \left[\frac{-1}{1 - \theta}\right] = \frac{n(\bar{x}-\theta)}{\theta(1-\theta)}$$

L'équation $\frac{d}{d\theta}\ \ell n\ L(\theta) = 0$ a pour solution

$$\hat{\theta} = \bar{x} \qquad (7.23)$$

On peut vérifié que cette solution correspond à un maximum de la fonction $L(\theta)$.

Exemple 7.13: distribution gaussienne $N(\theta_1, \theta_2)$

$$\theta_1 = \mu\ ,\ \theta_2 = \sigma^2$$

La fonction de vraisemblance est, d'après l'équation (7.19)

$$L(\theta_1, \theta_2) = (2\pi\theta_2)^{-n/2} \exp\left[-\frac{1}{2\theta_2} \sum_{i=1}^{n} (x_i - \theta_1)^2\right]$$

$$\ell n\ L(\theta_1, \theta_2) = -\left[\frac{n}{2}\right] \ell n\ (2\pi) - \left[\frac{n}{2}\right] \ell n(\theta_2) - \frac{1}{2\theta_2} \sum_{i=1}^{n} (x_i - \theta_1)^2$$

$$\frac{\partial}{\partial \theta_1} \ell n\, L(\theta_1, \theta_2) = \frac{1}{\theta_2} \sum_{i=1}^{n} (x_i - \theta_1) = 0 \qquad (7.24)$$

$$\frac{\partial}{\partial \theta_2} \ell n\, L(\theta_1, \theta_2) = -\frac{n}{2}\frac{1}{\theta_2} + \frac{1}{2\theta_2^2} \sum_{i=1}^{n} (x_i - \theta_1)^2 = 0 \qquad (7.25)$$

La solution de (7.24) est

$$\hat{\theta}_1 = \bar{x}$$

et, remplaçant cette valeur dans l'équation (7.25), on obtient

$$\hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\theta}_1)^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (7.26)$$

Le tableau 7.3 résume les estimateurs à vraisemblance maximale pour le cas de distributions usuelles.

Tableau 7.3: estimateurs à vraisemblance maximale

Nom	distribution	paramètres	estimateur
Bernoulli	$f_X(x;\theta) = \theta^x(1-\theta)^{1-x}$ $x = 0,1$	$0<\theta<1$	$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$
Binomiale	$f_X(x;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ $x = 0,1,2,\ldots,n$ n connu	$0<\theta<1$	$\hat{\theta} = \frac{1}{n}\frac{1}{N}\sum_{i=1}^{N} x_i = \frac{\bar{x}}{n}$ N = taille de l'échantillon
Géométrique	$f_X(x;\theta) = \theta(1-\theta)^{x-1}$ $x = 1,2,\ldots$	$0<\theta<1$	$\hat{\theta} = \frac{n}{\Sigma x_i} = \frac{1}{\bar{x}}$
Poisson	$f_X(x;\theta) = \frac{e^{-\theta}\,\theta^x}{x!}$ $x = 0,1,2,\ldots$	$\theta>0$	$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$
Exponentielle	$f_X(x;\theta) = \theta e^{-\theta x}$ $x > 0$	$\theta>0$	$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$
Gamma	$f_X(x;\theta) = \frac{1}{\theta^\alpha \Gamma(\alpha)} x^{\alpha-1}e^{-x/\theta}$ $x > 0$ α connu	$\theta>0$	$\hat{\theta} = \frac{1}{\alpha}\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\bar{x}}{\alpha}$
Rectangulaire	$f_X(x;\theta) = \frac{1}{\theta}$ $0 < x < \theta$	$\theta>0$	$\hat{\theta} = \mathrm{MAX}(x_1,\ldots,x_n)$

Tableau 7.3: estimateurs à vraisemblance maximale (suite)

Nom	distribution	paramètres	estimateur
Normale σ connu $\theta = \mu$	$f_X(x;\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x-\theta)^2\right]$	$\theta \in R$	$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$
Normale μ connu $\theta = \sigma^2$	$f_X(x;\theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left[-\frac{1}{2\theta}(x-\mu)^2\right]$	$\theta_1 \in R$ $\theta_2 > 0$	$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$
Normale $\theta_1=\mu$ $\theta_2=\sigma^2$	$f_X(x;\theta_1,\theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} \exp\left[-\frac{1}{2\theta_2}(x-\theta_1)^2\right]$		
		$\theta_1 \in R$ $\theta_2 > 0$	$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Log-normale	$f_X(x;\theta_1,\theta_2) = \frac{1}{x\sqrt{2\pi\theta_2}} \exp\left[-\frac{1}{2\theta_2}(\ell n\, x - \theta_1)^2\right]$		
		$\theta_1 > 0$ $\theta_2 > 0$	$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} \ell n x_i$ $\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(\ell n x_i - \hat{\theta}_1)^2$

Propriétés des estimateurs à vraisemblance maximale

(a) Les estimateurs à vraisemblance maximale sont quelquefois biaisés mais il est presque toujours possible de les corriger en multipliant par une constante appropriée.

(b) Les estimateurs à vraisemblance maximale sont invariants pour des transformations quelconques des paramètres. Soit $\hat{\theta}$ l'estimateur à vraisemblance maximale de θ, $h(\theta)$ une fonction quelconque de θ et $h(\theta)$ l'estimateur à vraisemblance maximale de $h(\theta)$. Alors

$$h(\theta) = h(\hat{\theta}).$$

(c) Si $\hat{\theta}$ est l'estimateur à vraisemblance maximale de θ et $f_X(x;\theta)$ est la loi de probabilité de X, alors $\hat{\theta}$ est sous certaines conditions de régularité et $n \geq 30$ disons, approximativement distribuée selon une loi gaussienne

$$\hat{\theta} \sim N\left(\theta, \frac{1}{nB^2}\right) \tag{7.27}$$

où

$$B^2 = \int \left(\frac{d}{d\theta} \ell n\, f_X(x;\theta) \right)^2 f_X(x;\theta)\, dx \tag{7.28}$$

$$B^2 = E\left[\left(\frac{d}{d\theta} \ell n\, f_X(x;\theta) \right)^2 \right]$$

7.4 MÉTHODE DES MOMENTS

Cette méthode d'estimation est basée sur le fait que les moments expérimentaux calculés à partir d'une série d'observations convergent vers les moments de la distribution lorsque la taille échantillonnale augmente indéfiniment. On pose donc, en principe, l'égalité des deux types de moments pour établir un système d'équations dont la solution permet d'obtenir des estimateurs. Cette méthode donne en général, des estimateurs moins précis que ceux de la méthode de vraisemblance maximale. Elle peut être utilisée lorsque la méthode de vraisemblance conduit à des équations difficiles à résoudre.

Principe

Soit X distribuée selon $f_X(x;\theta_1,\ldots,\theta_p)$ et $\theta_1,\theta_2,\ldots,\theta_p$ les paramètres. On définit les moments théoriques μ_r de la distribution par

$$\mu_r = \int x^r \, f_X(x;\theta_1,\ldots,\theta_p)dx$$

$$= h_r(\theta_1,\theta_2,\ldots,\theta_p) \qquad r = 1,2,\ldots \qquad (7.29)$$

Le moment d'ordre r est une fonction connue $h_r(.)$ de paramètres inconnus $\theta_1,\theta_2,\ldots,\theta_p$.

D'autre part, à l'aide d'observations $x_1,x_2,\ldots,x_n$ provenant de la distribution, on calcule les moments expérimentaux m_r par

$$m_r = \frac{1}{n} \sum_{i=1}^{n} x_i^r \qquad r = 1,2,\ldots \qquad (7.30)$$

La méthode consiste à poser le système d'équations

$$h_r(\theta_1,\theta_2,\ldots,\theta_p) = m_r \qquad r = 1,2,\ldots,p$$

dont la solution exprime les paramètres $\theta_1,\theta_2,\ldots,\theta_p$ en fonction des moments $m_1,m_2,\ldots,m_r$

$$\hat{\theta}_\alpha = g_\alpha(m_1,m_2,\ldots,m_r) \qquad \alpha = 1,2,\ldots,p \qquad (7.31)$$

APPLICATIONS

Exemple 7.14: distribution normale $N(\mu,\sigma^2) = N(\theta_1,\theta_2)$

Puisqu'il y a deux paramètres à estimer, le système sera composé de deux équations. On a

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

$$\mu_1 = \int x\, f_X(x;\theta_1,\theta_2)dx = \theta_1$$

$$\mu_2 = \int x^2\, f_X(x;\theta_1,\theta_2)dx = \theta_2 + \theta_1^2$$

où $$f_X(x;\theta_1,\theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} \exp\left[-\frac{1(x-\theta_1)^2}{2\theta_2}\right]$$

Il faut résoudre pour θ_1 et θ_2, le système

$$\theta_1 = m_1 = \bar{x}$$

$$\theta_2 + \theta_1^2 = m_2$$

dont la solution est

$$\hat{\theta}_1 = m_1 = \bar{x}$$

$$\hat{\theta}_2 = m_2 - \hat{\theta}_1^2 \qquad (7.32)$$

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

Les estimateurs coïncident avec les estimateurs à vraisemblance maximale de l'exemple 7.13.

Exemple 7.15: distribution gamma GAMMA(α,β) = GAMMA(θ_1,θ_2)

La méthode de vraisemblance maximale conduit à des équations non-linéaires devant être résolues par des méthodes itératives. La méthode des moments donne des équations non-linéaires que l'on peut les résoudre de façon analytique.

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

D'après l'équation (6.5)

$$\mu_1 = \int x\, f_X(x;\theta_1,\theta_2)dx = \theta_1\theta_2$$

$$\mu_2 = \int x^2\, f_X(x;\theta_1,\theta_2)dx = \theta_1\theta_2^2(1+\theta_1)$$

où $$f_X(x;\theta_1,\theta_2) = \frac{1}{\theta_2^{\theta_1}\,\Gamma(\theta_1)}\, x^{\theta_1-1} \exp\left[-\frac{x}{\theta_2}\right]$$

Le système à résoudre est

$$\theta_1\theta_2 = m_1 = \bar{x}$$

$$\theta_1\theta_2^2(1+\theta_1) = m_2$$

dont la solution est:

$$\hat{\theta}_2 = \frac{m_2-\bar{x}^2}{\bar{x}} \qquad \hat{\theta}_1 = \frac{\bar{x}^2}{m_2-\bar{x}^2} \tag{7.33}$$

Exemple 7.16 distribution BETA(α,β) = BETA(θ_1,θ_2)

D'après l'équation (6.57)

$$\mu_1 = \int_0^1 x\, f_X(x;\theta_1,\theta_2)dx = \frac{\theta_1}{\theta_1 + \theta_2}$$

$$\mu_2 = \int_0^1 x^2\, f_X(x;\theta_1,\theta_2)dx = \frac{\theta_1\theta_2}{(\theta_1+\theta_2)^2(\theta_1+\theta_2+1)} + \frac{\theta_1^2}{(\theta_1+\theta_2)^2}$$

où $$f_X(x;\theta_1,\theta_2) = \frac{\Gamma(\theta_1 + \theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)} x^{\theta_1-1} (1 - x)^{\theta_2-1}$$

Le système à résoudre est

$$\frac{\theta_1}{\theta_1 + \theta_2} = m_1 = \bar{x}$$

$$\frac{\theta_1\theta_2}{(\theta_1+\theta_2)^2(\theta_1+\theta_2+1)} + \frac{\theta_1^2}{(\theta_1+\theta_2)^2} = m_2$$

dont la solution est

$$\hat{\theta}_1 = \frac{(\bar{x} - m_2)\bar{x}}{m_2 - \bar{x}^2}$$

$$\hat{\theta}_2 = \frac{(\bar{x} - m_2)(1 - \bar{x})}{m_2 - \bar{x}^2} \qquad (7.34)$$

7.5 MÉTHODE DES MOINDRES CARRÉS

Modèles statistiques

Un très grand nombre de méthodes d'analyses statistiques peuvent se formuler à l'aide des modèles statistiques de la forme:

$$Y = \varphi(X_1, X_2, \ldots, X_k; \beta_o, \ldots, \beta_p) + \varepsilon \qquad (7.35)$$

où Y représente une variable à expliquer (ou dépendante),
φ est une fonction mathématique de forme connue
$X_1, X_2, \ldots, X_k$ sont des variables explicatives
$\beta_o, \ldots, \beta_p$ sont des paramètres à estimer
ε est un terme d'erreur tel que

$$E(\varepsilon) = 0 \quad , \quad VAR(\varepsilon) = \sigma^2 \qquad (7.36)$$

Lorsque les variables explicatives varient sur un continum, on dit que le modèle (7.35) est un MODÈLE DE RÉGRESSION. Si les variables $X_1, X_2, \ldots, X_k$ sont des indicatrices (0-1), le modèle (7.35) est un MODÈLE D'ANALYSE DE VARIANCE.

Ces derniers servent à analyser et décomposer les variations de la variable Y produites par l'effet de facteurs contrôlés $X, \ldots X_k$. C'est un domaine riche d'applications dans la planification des expériences industrielles et scientifiques.

Un modèle est dit LINÉAIRE dans les paramètres $\beta_o, \beta_1, \ldots, \beta_P$ si

$$\varphi(X_1, \ldots, X_k; \beta_0, \ldots, \beta_p) = \sum_{j=0}^{p} \beta_j \varphi_j(X_1, \ldots, X_k) \qquad (7.37)$$

où $\varphi_j(.)$ sont des fonctions ne contenant aucun paramètre. Autrement, le modèle est dit NON-LINÉAIRE. Toutefois, parmi ces derniers, certains peuvent devenir linéaires après avoir effectué des transformations sur les variables $X_1, \ldots, X_k, Y$. L'observation du modèle génère des données

$$(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) \quad i = 1, \ldots, n > p+1 \qquad (7.38)$$

et les principales questions à résoudre sont:

- l'estimation des paramètres $\beta_o, \beta_1, \ldots, \beta_P$
- l'estimation de la variance σ^2
- les tests d'hypothèses concernant les paramètres $\beta_o, \ldots, \beta_p$.
- l'analyse de la qualité de représentation du modèle

L'estimation de $\beta_o,\beta_1,\ldots,\beta_p$ se fait en adoptant le PRINCIPE DES MOINDRES CARRÉS. On constitue la fonction objective $S(\beta_o,\ldots,\beta_p)$ dont on cherche un minimum dans l'espace des paramètres

$$S(\beta_o,\beta_1,\ldots,\beta_p) = \sum_{i=1}^{n} [y_i - \varphi(x_{i1},\ldots,x_{ik};\beta_o,\ldots,\beta_p)]^2$$

Une condition nécessaire est fournie par le système d'équations:

$$\frac{\partial}{\partial \beta_\alpha} S(\beta_o,\beta_1,\ldots,\beta_p) = 0 \quad \alpha = 0,1,2,\ldots,p \tag{7.39}$$

dans les inconnues $\beta_o,\beta_1,\ldots,\beta_p$.

Cas particuliers

Exemple 7.17: modèle sans variable explicative

$$Y = \beta_o + \varepsilon \tag{7.40}$$

$$S(\beta_o) = \sum_{i=1}^{n} (y_i - \beta_o)^2$$

$$\frac{d}{d\beta_o} S(\beta_o) = \sum_{i=1}^{n} 2(y_i - \beta_o)(-1) = 0$$

Dont la solution est:

$$\hat{\beta}_o = \bar{y} \tag{7.41}$$

Exemple 7.18: modèle de régression linéaire simple

$$Y = \beta_o + \beta_1 X + \varepsilon \tag{7.42}$$

$$S(\beta_o,\beta_1) = \sum_{i=1}^{n} (y_i - \beta_o - \beta_1 x_i)^2$$

$$\frac{\partial S}{\partial \beta_o} = \sum_{i=1}^{n} 2(y_i - \beta_o - \beta_1 x_i)(-1) = 0$$

$$\frac{\partial S}{\partial \beta_1} = \sum 2(y_i - \beta_o - \beta_1 x_i)(-x_i) = 0$$

Dont la solution est:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad (7.43)$$

$$\hat{\beta}_o = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (7.44)$$

Une étude détaillée du modèle sera faite au chapitre 9.

Exemple 7.19: modèle de régression simple non-linéaire

$$Y = \beta_o + \beta_1 e^{\beta_2 X} + \varepsilon \qquad (7.45)$$

est non-linéaire dans les paramètres $\beta_o, \beta_1, \beta_2$

$$S(\beta_o, \beta_1, \beta_2) = \sum_{i=1}^{n} \left[y_i - \beta_o - \beta_1 e^{\beta_2 x_i} \right]^2$$

L'annulation des dérivées partielles de S par rapport à $\beta_o, \beta_1, \beta_2$ conduit au système d'équations non-linéaires dans les inconnues $\beta_o, \beta_1, \beta_2$

$$n\beta_o + \beta_1 \sum_{i=1}^{n} e^{\beta_2 x_i} = \sum_{i=1}^{n} y_i$$

$$\beta_o \sum_{i=1}^{n} e^{\beta_2 x_i} + \beta_1 \sum_{i=1}^{n} e^{2\beta_2 x_i} = \sum_{i=1}^{n} y_i e^{\beta_2 x_i}$$

$$\beta_o \sum_{i=1}^{n} x_i e^{\beta_2 x_i} + \beta_1 \sum_{i=1}^{n} x_i e^{2\beta_2 x_i} = \sum_{i=1}^{n} y_i x_i e^{\beta_2 x_i}$$

La résolution peut se faire par des méthodes numériques itératives.

Exemple 7.20: modèle de régression polynômiale

$$Y = \beta_o + \beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p + \varepsilon \qquad (7.46)$$

$$S(\beta_o,\beta_1,\ldots,\beta_p) = \sum_{i=1}^{n} \left[y_i - \sum_{\alpha=0}^{n} \beta_\alpha x_i^\alpha \right]^2$$

$$\frac{\partial S}{\partial \beta_k} = \sum_{i=1}^{n} 2 \left[y_i - \sum_{\alpha=0}^{p} \beta_\alpha x_i^\alpha \right] \left[-x_i^k \right] = 0 \qquad k = 0,1,2,\ldots,p$$

est un système d'équations linéaires dans les paramètres $\beta_o,\beta_1,\ldots,\beta_p$. Il s'écrit:

$$\begin{bmatrix} n & \Sigma x_i & \Sigma x_i^2 & \ldots & \Sigma x_i^p \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i^3 & \ldots & \Sigma x_i^{p+1} \\ \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^4 & \ldots & \Sigma x_i^{p+2} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \Sigma x_i^p & \Sigma x_i^{p+1} & \ldots & \ldots & \Sigma x_i^{2p} \end{bmatrix} \begin{bmatrix} \beta_o \\ \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{bmatrix} = \begin{bmatrix} \Sigma y_i \\ \Sigma y_i x_i \\ \Sigma y_i x_i^2 \\ \cdot \\ \cdot \\ \cdot \\ \Sigma y_i x_i^p \end{bmatrix} \qquad (7.47)$$

La résolution de tels systèmes exige le recours à un progiciel comme le système SAS. Dans le système plusieurs procédures traitent de la régression dont REG,NLIN et GLM.

Exemple 7.21: modèle de régression multiple

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \qquad (7.48)$$

$$S(\beta_o,\ldots,\beta_p) = \sum_{i=1}^{n} \left(y_i - \sum_{i=1}^{p} \beta_\alpha x_{i\alpha} \right)^2$$

$$\frac{\partial S}{\partial \beta_k} = \sum_{i=1}^{n} 2 \left(y_i - \sum_{\alpha=0}^{p} \beta_\alpha x_{i\alpha} \right)(-x_{ik}) = 0 \quad , \quad k = 0,1,2,\ldots,p$$

est un système d'équations linéaires pouvant s'écrire

$$\begin{bmatrix} n & \Sigma x_{i1} & \Sigma x_{i2} & \ldots & \Sigma x_{ip} \\ \Sigma x_{i1} & \Sigma x_{i1}^2 & \Sigma x_{i2}x_{i1} & \ldots & \Sigma x_{ip}x_{i1} \\ \Sigma x_{i2} & \Sigma x_{i1}x_{i2} & \Sigma x_{i2}^2 & \ldots & \Sigma x_{ip}x_{i2} \\ \vdots & \vdots & \vdots & & \vdots \\ \Sigma x_{ip} & \Sigma x_{i1}x_{ip} & \Sigma x_{i2}x_{ip} & \ldots & \Sigma x_{ip}^2 \end{bmatrix} \begin{bmatrix} \beta_o \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \Sigma y_i \\ \Sigma y_i x_{i1} \\ \Sigma y_i x_{i2} \\ \vdots \\ \Sigma y_i x_{ip} \end{bmatrix} \qquad (7.49)$$

Notons que le modèle de régression polynômiale est un cas particulier du modèle de régression multiple où $X_k = X^k$ $k = 0,1,\ldots,p$.

Exemple 7.22: modèle pour comparer deux groupes de données

Nous avons soulevé au chapitre 1, le problème de la comparaison de la force de rupture de deux sortes de fils, disons A et B. Nous pouvons formuler ce problème à l'aide d'un modèle statistique avec variables indicatrices. Posons

$$X_1 = \begin{cases} 1 & \text{si l'observation provient du fil A} \\ 0 & \text{autrement} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{si l'observation provient du fil B} \\ 0 & \text{autrement} \end{cases}$$

La force de rupture Y peut s'écrire

$$Y = \begin{cases} \beta_o + \beta_1 + \varepsilon & \text{si le fil est de type A} \\ \beta_o + \beta_2 + \varepsilon & \text{si le fil est de type B} \end{cases}$$

où encore

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad (7.50)$$

$$S(\beta_o,\beta_1,\beta_2) = \sum_{i=1}^{n} (y_i - \beta_o - \beta_1 x_{i1} - \beta_2 x_{i2})^2$$

Notons que le modèle est un cas particulier du modèle de régression multiple avec p = 2 où les valeurs de (X_{i1},X_{i2}) sont (1,0) ou (0,1) selon le cas.

Le système d'équations à résoudre est

$$\begin{bmatrix} n & \Sigma x_{i1} & \Sigma x_{i2} \\ \Sigma x_{i1} & \Sigma x_{i1}^2 & \Sigma x_{i1}x_{i2} \\ \Sigma x_{i2} & \Sigma x_{i1}x_{i2} & \Sigma x_{i2}^2 \end{bmatrix} \begin{bmatrix} \beta_o \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \Sigma y_i \\ \Sigma y_i x_{i1} \\ \Sigma y_i x_{i2} \end{bmatrix} \qquad (7.51)$$

Mais

$$\sum_{1}^{n} x_{i1} = n_1 \qquad \sum_{i=1}^{n} x_{i2} = n_2$$

$$\sum_{i=1}^{n} x_{i1}^2 = n_1 \qquad \sum_{i=1}^{n} x_{i2}^2 = n_1$$

$$\sum_{i=1}^{n} x_{1i}x_{i2} = 0 \qquad \sum_{i=1}^{n} y_i = n\bar{y}$$

$$\sum_{i=1}^{n} x_{i1}y_i = n_1\bar{y}_1 \qquad \sum_{i=1}^{n} x_{i2}y_i = n_2\bar{y}_2$$

où n_1 est le nombre d'observations des données du fil A
n_2 est le nombre d'observations des données du fil B

$\bar{y}_1$: moyenne des observations des données du fil A

$\bar{y}_2$: moyenne des observations des données du fil B

$\bar{y}$: moyenne de toutes les observations.

Le système (7.51)

$$\begin{aligned} n\beta_o + n_1\beta_1 + n_2\beta_2 &= n\bar{y} && \text{(a)} \\ n_1\beta_o + n_1\beta_1 &= n_1\bar{y}_1 && \text{(b)} \\ n_2\beta_o + n_2\beta_2 &= n_2\bar{y}_2 && \text{(c)} \end{aligned} \qquad (7.52)$$

On constate que le système est de rang 2 et ne possède pas de solution unique. On peut utiliser les équations (b)-(c) et la condition "naturelle" suivante:

$$n_1\beta_1 + n_2\beta_2 = 0 \qquad \text{(d)}$$

La solution de (b)-(c)-(d) donne

$$\begin{aligned} \hat{\beta}_o &= \bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n} \\ \hat{\beta}_1 &= \bar{y}_1 - \hat{\beta}_o = \bar{y}_1 - \bar{y} \\ \hat{\beta}_2 &= \bar{y}_2 - \hat{\beta}_o = \bar{y}_2 - \bar{y} \end{aligned} \qquad (7.53)$$

Application numérique: exemple de comparaison de deux types de fil du chapitre 1

$n = 50 \qquad n_1 = 25 \qquad n_2 = 25$

$\bar{y} = 151.92 \qquad \bar{y}_1 = 151.08 \qquad \bar{y}_2 = 152.76$

$\hat{\beta}_o = \bar{y} = 151.92 \qquad \hat{\beta}_1 = \bar{y}_1 - \bar{y} = -0.84$

$\hat{\beta}_2 = \bar{y}_2 - \bar{y} = 152.76 - 151.92 = 0.84$

Exemple 7.23: modèle de classification simple

On généralise l'exemple précédent à plus de deux groupes de données. Par exemple, on peut faire la comparaison de la quantité d'hydrocarbures émise par cinq marques de voitures. On introduit des variables indicatrices 0-1 pour chacun des groupes de données:

$$X_\alpha = \begin{cases} 1 & \text{si l'observation provient du } \alpha\text{-ième groupe} \\ 0 & \text{autrement} \end{cases} \qquad (7.54)$$

$\alpha = 1,2,\ldots,k$ où k est le nombre de groupes.

Le modèle proposé est

$$y_i = \begin{cases} \beta_o + \beta_1 + \varepsilon_i & : \text{1er groupe} \\ \beta_o + \beta_2 + \varepsilon_i & : \text{2-ième groupe} \\ \cdots\cdots\cdots\cdots & \\ \beta_o + \beta_k + \varepsilon_i & : \text{k-ième groupe} \end{cases} \qquad (7.55)$$

ou plus simplement

$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i \qquad (7.56)$$

D'une manière analogue à l'exemple précédent on impose une équation additionnelle afin d'obtenir une solution unique

$$n_1\beta_1 + n_2\beta_2 + \ldots + n_k\beta_k = 0 \qquad (7.57)$$

où $n_1, n_2, \ldots, n_k$ sont les tailles échantillonnales respectives de chacun des groupes. La solution obtenue par le principe des moindres carrés est

$$\begin{aligned} &\hat{\beta}_o = \bar{y} \\ &\hat{\beta} = \bar{y}_k - \bar{y} \quad \alpha = 1,2,\ldots,k \end{aligned} \qquad (7.58)$$

où $\bar{y}$ est la moyenne de toutes les observations et $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$ sont les moyennes de chacun des groupes. Une analyse complète du modèle sera faite au chapitre 8.

Remarque: On peut écrire d'une manière simplifiée le modèle (7.55) en introduisant des indices doubles. Posons

$$y_{ij} = \text{j-ième observation du iième groupe} \qquad i = 1,2,\ldots,k;\ j = 1,2,..,n_i \tag{7.59}$$

On a alors

$$y_{ij} = \beta_o + \beta_i + \varepsilon_{ij} \tag{7.60}$$

Les modèles d'analyse de variance sont généralement présentés sous cette forme.

Exemple 7.24: modèle de classification double

Ils sont employés pour faire l'analyse de données obtenues en faisant varier deux facteurs contrôlés. Par exemple, nous avons déjà présenté au chapitre 1 des données obtenues représentant la perte de poids de pistons de quatre marques différentes utilisées avec cinq types d'huile. Dans cet exemple, les deux facteurs sont la marque de piston et le type d'huile. Le premier facteur est à quatre modalités (A,B,C,D) et le deuxième facteur contient cinq modalités (1,2,3,4,5) .

D'une manière générale les données d'expériences où deux facteurs sont croisés peuvent se représenter par

$$y_{ijk} = \mu_o + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \qquad i = 1,\ldots,I;\ j = 1,\ldots,J;\ k = 1,2,\ldots,n_{ij} \tag{7.61}$$

où μ_o: effet général

α_i: effet du 1-er facteur - modalité i

β_j: effet de 2-ième facteur - modalité j

γ_{ij}: effet interaction des deux facteurs

Le modèle contient 1 + I + J + IJ paramètres et afin de résoudre les équations, on impose les contraintes naturelles suivantes

$$\sum_{i=1}^{I} \alpha_i = 0 \qquad \sum_{i=1}^{I} \gamma_{ij} = 0 \quad j = 1,\ldots,J$$

$$\sum_{j=1}^{J} \beta_j = 0 \qquad \sum_{j=1}^{J} \gamma_{ij} = 0 \quad i = 1,\ldots,I \tag{7.62}$$

Le principe des moindres carrés conduit aux estimateurs suivants si $n_{ij} = n$ pour tout i,j

$$\begin{aligned}
\hat{\theta}_o &= \bar{y}_{...} \\
\hat{\alpha}_i &= \bar{y}_{i..} - \bar{y}_{...} \\
\hat{\beta}_j &= \bar{y}_{.j.} - \bar{y}_{...} \\
\hat{\gamma}_{ij} &= \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}
\end{aligned} \tag{7.63}$$

où
$$\bar{y}_{...} = \frac{1}{nIJ} \sum_i \sum_j \sum_k y_{ijk} \qquad \bar{y}_{i..} = \frac{1}{nJ} \sum_j \sum_k y_{ijk}$$
$$\bar{y}_{.j.} = \frac{1}{nI} \sum_i \sum_k y_{ijk} \qquad \bar{y}_{ij.} = \frac{1}{n} \sum_k y_{ijk} \tag{7.64}$$

<u>Propriétés des estimateurs de moindres carrés</u>

(a) Si la variable à expliquer est normalement distribuée

$$Y \sim N(\varphi(X_1,\ldots,X_k;\beta_o,\beta_1,\ldots,\beta_p),\sigma^2)$$

alors les estimateurs de moindres carrés sont identiques aux estimateurs à vraisemblance maximale.

En effet, la fonction de vraisemblance s'écrit

$$L(\beta_o,\beta_1,..,\beta_p,\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \varphi(x_{i1},..,x_{ik};\beta_{o1},..,\beta_p))^2\right]$$

et donc

$$\max_{\beta_o,..,\beta_p} L(\beta_{o1},..,\beta_p,\sigma^2) = \min_{\beta_{o1},..,\beta_p} \sum_{i=1}^{n} (y_i - \varphi(x_{i1},..,x_{ik};\beta_o,..,\beta_p))^2$$

(b) Si le modèle est linéaire alors les estimateurs de moindres carrés sont des fonctions linéaires des observations $y_1,\ldots,y_n$. En effet, posons

$$\varphi(x_{i1},\ldots,x_{ik};\beta_{o1},\ldots,\beta_p) = \sum_{\alpha=0}^{p} \beta_\alpha \varphi_\alpha(x_{i1},\ldots,x_{ik})$$

$$= \sum_{\alpha=0}^{p} \beta_\alpha \, z_{i\alpha} \qquad \text{où} \qquad z_{i\alpha} = g_\alpha(x_{i1},\ldots,x_{ik})$$

$$\underset{\sim}{\beta} = (\beta_o,\ldots,\beta_p)' : \quad (p+1) \times 1$$

$$\underset{\sim}{Y} = (y_1,\ldots,y_n)' : \quad n \times 1$$

$$\underset{\sim}{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)' : \quad n \times 1$$

$$Z = (Z_{i\alpha}) \qquad : \quad n \times (p+1)$$

alors les équations de moindres carrés à résoudre (7.39) s'écrivent:

$$(Z'Z) \underset{\sim}{\beta} = Z'\underset{\sim}{Y} \tag{7.65}$$

dont la solution est

$$\underset{\sim}{\hat{\beta}} = (Z'Z)^{-1} Z' \underset{\sim}{Y} = C \underset{\sim}{Y} \tag{7.66}$$

où $C = (Z'Z)^{-1} Z'$ est une matrice $(p+1) \times n$ de valeurs connues.

(c) Si le modèle est linéaire alors les estimateurs de moindres carrés sont sans biais et possèdent la plus petite variance dans la classe des estimateurs linéaires (théorème de Gauss-Markov). On peut écrire

$$\underset{\sim}{Y} = Z \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

donc

$$\hat{\underset{\sim}{\beta}} = (Z'Z)^{-1} Z' \underset{\sim}{Y} = (Z'Z)^{-1} Z'(Z \underset{\sim}{\beta} + \underset{\sim}{\varepsilon})$$

$$\hat{\underset{\sim}{\beta}} = (Z'Z)^{-1} (Z'Z) \underset{\sim}{\beta} + (Z'Z)^{-1}Z'\underset{\sim}{\varepsilon}$$

$$\hat{\underset{\sim}{\beta}} = \underset{\sim}{\beta} + (Z'Z)^{-1}Z'\underset{\sim}{\varepsilon}$$

et alors $\quad E(\hat{\underset{\sim}{\beta}}) = \underset{\sim}{\beta}$

La matrice de variance-covariance de $\hat{\underset{\sim}{\beta}}$ est

$$\begin{aligned}
VAR(\hat{\underset{\sim}{\beta}}) &= E\,[(\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta})(\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta})'] \\
&= E\,[((Z'Z)^{-1} Z' \underset{\sim}{\varepsilon})\,(Z'Z)^{-1} Z'\underset{\sim}{\varepsilon})'] \\
&= E\,[(Z'Z)^{-1} Z' \underset{\sim}{\varepsilon}\,\underset{\sim}{\varepsilon}'\, Z\,(Z'Z)^{-1}] \\
&= (Z'Z)^{-1} Z'\, E\,(\underset{\sim\sim}{\varepsilon\varepsilon'})\, Z\,(Z'Z)^{-1} \\
&= (Z'Z)^{-1} Z'\, \sigma^2\, I_n\, Z\,(Z'Z)^{-1} \\
&= (Z'Z)^{-1}\,(Z'Z)\,(Z'Z)^{-1}\,\sigma^2 \\
&= (Z'Z)^{-1}\,\sigma^2 \qquad (7.67)
\end{aligned}$$

où I_n est la matrice identité d'ordre n

7.6 MÉTHODE DES INTERVALLES DE CONFIANCE

Définition du problème

Soit $X_1, X_2, \ldots, X_n$ un échantillon aléatoire provenant de la distribution $f_X(x;\theta)$. L'estimation du paramètre θ par un intervalle consiste à déterminer deux nombres $a = a(X_1,\ldots,X_n)$ et $b = b(X_1,\ldots,X_n)$ tels que

$$P[\ a(X_1,\ldots,X_n) \leq \theta \leq b(X_1,\ldots,X_n)\ ] = 1 - \alpha \qquad (7.68)$$

où $1 - \alpha$ $(0 < \alpha < 1)$ s'appelle un COEFFICIENT DE CONFIANCE. Il est spécifié d'avance et choisi généralement près de 1, par exemple 0.90, 0.95, 0.99. Ce coefficient représente la fréquence à long terme de la méthode à générer des intervalles contenant le paramètre.

L'équation (7.68) ne possède pas une solution unique à moins d'ajouter une autre condition. Généralement, on substitue à (7.68) les deux équations suivantes:

$$P[\ \theta \geq b(X_1,\ldots,X_n)\ ] = \frac{\alpha}{2} \qquad (7.69)$$

$$P[\ \theta \leq a(X_1,\ldots,X_n)\ ] = \frac{\alpha}{2} \qquad (7.70)$$

Cette solution a souvent l'avantage de produire l'intervalle ayant la plus petite longueur parmi tous les intervalles dont le coefficient de confiance est $1 - \alpha$.

La méthode générale pour obtenir les valeurs de a et b repose sur la connaissance d'un estimateur $\hat{\theta} = \hat{\theta}(X_1,\ldots,X_n)$ de θ ainsi que de sa loi d'échantillonnage $f_{\hat{\theta}}(\hat{\theta})$. On détermine une fonction de $\hat{\theta}$ et θ, disons $H(\hat{\theta},\theta)$, telle que sa loi de probabilité ne dépend d'aucun paramètre inconnu. Il est donc possible de déterminer (en consultant des tables généralement) deux nombres $h_1 = h_1(\alpha)$ et $h_2 = h_2(\alpha)$ tels que

$$P[\ h_1 \leq H(\hat{\theta},\theta) \leq h_2\ ] = 1 - \alpha$$

L'intervalle de confiance (a,b) sera obtenu de cette dernière équation en isolant θ

$$P[\ a \leq \theta \leq b\ ] = 1 - \alpha$$

On voit alors que a et b sont des fonctions de α par l'intermédiaire de h_1 et h_2 et de $X_1, X_2, \ldots, X_n$ par l'intermédiaire de $\hat{\theta}$.

Applications

Exemple 7.25: intervalle de confiance pour la moyenne μ d'une distribution normale $N(\mu, \sigma^2)$ avec σ connu.

On sait que $\hat{\mu} = \bar{X}$ est l'estimateur de μ et que

$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. La quantité $H(\hat{\mu}, \mu) = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ est distribuée selon une distribution N(0,1). L'équation

$$P[\, h_1 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq h_2 \,] = 1 - \alpha$$

possède la solution $h_1 = -z_{\alpha/2}$ et $h_2 = z_{\alpha/2}$, où $z_{\alpha/2}$ est le $(1-\alpha/2)$-ième percentile d'une distribution normale centrée-réduite.

Donc

$$P\left[-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \right] = 1 - \alpha$$

ou encore

$$P\left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] = 1 - \alpha$$

L'intervalle de confiance pour μ est donc:

$$\mu: \bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \qquad (7.71)$$

Application numérique

La durée de 20 ampoules a donné $\bar{x} = 1014$ (heures) et on sait que l'écart-type σ est égal à 100. Un intervalle de confiance à 95% pour la moyenne véritable du lot de toutes les ampoules de même type est

$$\mu: \quad 1014 \pm 1.96 * \frac{100}{\sqrt{20}}$$

$$\mu: \quad 1014 \pm 43.8$$

Exemple 7.26: intervalle de confiance pour la moyenne μ d'une distribution normale $N(\mu,\sigma^2)$ avec σ inconnu.

L'estimateur de μ est $\hat{\mu} = \bar{X}$ et $(\bar{X} - \mu)\sqrt{n}/s$ est distribué selon une loi de Student avec n-1 degrés de liberté. L'équation

$$P\left[h_1 \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq h_2 \right] = 1 - \alpha$$

possède la solution $h_1 = -t_{n-1,\alpha/2}$ et $h_2 = t_{n-1,\alpha/2}$,

$$P\left[\bar{X} - t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} \right] = 1 - \alpha$$

L'intervalle de confiance pour μ est donc:

$$\mu: \quad \bar{X} \pm t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} \qquad (7.72)$$

Application numérique

La durée moyenne d'un échantillon de 20 ampoules a donné $\bar{x} = 1014$ (heures) avec un écart-type $s = 100$ (heures). Un intervalle de confiance à 95% pour la moyenne véritable du lot d'ampoules est

$$\mu: \quad 1014 \pm 2.09 * \frac{100}{\sqrt{20}}$$

Exemple 7.27: intervalle de confiance pour la moyenne μ d'une distribution quelconque avec un échantillon de taille supérieure à 30.

L'estimateur de μ est $\hat{\mu} = \bar{X}$ et le théorème central-limite nous assure que

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim N(0,1)$$

approximativement.

On en déduit l'intervalle de confiance approximatif

$$\mu: \quad \bar{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}} \tag{7.73}$$

Exemple 7.28: intervalle de confiance pour la variance σ^2 d'une distribution normale $N(\mu,\sigma^2)$ avec μ inconnu.

L'estimateur ponctuel de σ^2 est S^2 et on sait que

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Donc

$$P\left[\chi^2_{n-1,1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1,\alpha/2}\right] = 1 - \alpha$$

L'intervalle de confiance pour σ^2 est donc:

$$\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} \qquad (7.74)$$

Application numérique

La durée de 20 ampoules a donné une durée moyenne de 1014 et une variance de 1000. Un intervalle de confiance à 95% pour σ^2 est

$$\frac{19 * 1000}{32.85} \leq \sigma^2 \leq \frac{19 * 1000}{8.91}$$

$$578.4 \leq \sigma^2 \leq 2132.4$$

Exemple 7.29: intervalle de confiance approximatif pour l'écart-type d'une distribution quelconque avec un échantillon de taille supérieure à 30.

On peut montrer que

$$\frac{S - \sigma}{\sigma/\sqrt{2n}} \sim N(0,1)$$

approximativement. Donc

$$P\left[-z_{\alpha/2} \leq \frac{S - \sigma}{\sigma/\sqrt{2n}} \leq z_{\alpha/2} \right] = 1 - \alpha$$

L'intervalle de confiance pour σ est:

$$\frac{S}{1 + \frac{z_{\alpha/2}}{\sqrt{2n}}} \leq \sigma \leq \frac{S}{1 - \frac{z_{\alpha/2}}{\sqrt{2n}}} \qquad (7.75)$$

Application numérique: n=40, $s=\sqrt{1000}=31.62$, $\alpha=0.05$

$$\frac{31.62}{1+\frac{1.96}{\sqrt{80}}} \leq \sigma \leq \frac{31.62}{1-\frac{1.96}{\sqrt{80}}}$$

$$25.93 \leq \sigma \leq 40.49$$

Exemple 7.30: Intervalle de confiance approximatif pour le paramètre θ d'une distribution Bernoulli

$$p_X(x;\theta) = \theta^x(1 - \theta)^{1-x}, \quad 0 < \theta < 1, \quad x = 0,1$$

L'estimateur de θ est $\hat{\theta} = \bar{X}$ et le théorème central-limite permet d'écrire

$$\hat{\theta} = \bar{X} \sim N\left(\theta, \frac{\theta(1 - \theta)}{n}\right)$$

Donc

$$P\left[-z_{\alpha/2} \leq \frac{\bar{X} - \theta}{\sqrt{\frac{\theta(1 - \theta)}{n}}} \leq z_{\alpha/2} \right] = 1 - \alpha$$

$$P\left[\bar{X} - z_{\alpha/2}\sqrt{\frac{\theta(1 - \theta)}{n}} \leq \theta \leq \bar{X} + z_{\alpha/2}\sqrt{\frac{\theta(1 - \theta)}{n}} \right] = 1 - \alpha$$

Puisque θ est inconnu, on remplace θ par $\bar{X}$ pour obtenir

$$\theta = \bar{X} \pm z_{\alpha/2}\sqrt{\frac{\bar{X}(1 - \bar{X})}{n}} \qquad (7.76)$$

Application numérique

$$n = 1000, \ \bar{x} = 0.30, \ 1 - \alpha = 0.95$$

$$\theta = 0.30 \pm 1.96\sqrt{\frac{0.30 * 0.70}{1000}}$$

$$\theta = 0.30 \pm 0.03$$

Exemple 7.31: intervalle de confiance approximatif pour l'estimateur à vraisemblance maximale ($n \geq 30$)

Soit θ un paramètre et $\hat{\theta}$ son estimateur à vraisemblance maximale. On sait que $\hat{\theta}$ suit asymtotiquement ($n \to \infty$) une distribution normale $N\left(\theta, \frac{1}{nB^2}\right)$ où B est définie par l'équation (7.28). Un intervalle de confiance avec coefficient $(1-\alpha)$ pour θ est donc:

$$\theta : \hat{\theta} \pm z_{\alpha/2} \frac{1}{B\sqrt{n}} \tag{7.77}$$

D'une manière générale, B dépend de θ et possiblement de d'autres paramètres inconnus. Les paramètres doivent être estimés et l'évaluation de B est faite avec ces estimations. Illustrons l'équation (7.77) pour le cas d'une distribution de Poisson.

Exemple 7.32: intervalle de confiance approximatif pour le paramètre θ d'une distribution de Poisson.

$$f_X(x;\theta) = \frac{e^{-\theta}\ \theta^x}{x!} \qquad x = 0,1,2,\ldots;\ \theta > 0$$

On a vu que $\hat{\theta} = \bar{X}$ est l'estimateur à vraisemblance maximale. Calculons B.

$$\ell n\ f_X(x;\theta) = -\theta + x\ \ell n\ \theta - \ell n(x!)$$

$$\frac{d}{d\theta}\ \ell n\ f_X(x;\theta) = -1 + x\ \frac{1}{\theta} = \frac{x - \theta}{\theta}$$

$$E\left[\frac{d}{d\theta}\ \ell n\ f_X(x;\theta)\right]^2 = E\left[\frac{X - \theta}{\theta}\right]^2 = \frac{1}{\theta^2}\ E(X - \theta)^2$$

$$= \frac{1}{\theta^2}\ VAR(X) = \frac{1}{\theta^2}\ \theta = \frac{1}{\theta} = B^2$$

donc

$$\theta : \hat{\theta} \pm z_{\alpha/2} \sqrt{\frac{\hat{\theta}}{n}} \tag{7.78}$$

7.7 FORMULAIRE DES INTERVALLES DE CONFIANCE

$1 - \alpha$ = coefficient de confiance; * intervalle approximatif

Paramètre	Conditions	Intervalle de confiance
moyenne (μ)	distribution normale $N(\mu,\sigma^2)$ σ connu	$\bar{x} \pm z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$
moyenne (μ)	distribution normale $N(\mu,\sigma^2)$ σ inconnu $s^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$	$\bar{x} \pm t_{n-1,\alpha/2} \dfrac{s}{\sqrt{n}}$
moyenne (μ)	distribution quelconque $n \geq 30$	$\bar{x} \pm z_{\alpha/2} \dfrac{s}{\sqrt{n}}$ (*)
différence de 2 moyennes: $\mu_1 - \mu_2$	distributions normales $N(\mu_1,\sigma_1^2), N(\mu_1,\sigma_2^2)$ σ_1, σ_2 connus	$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$
différence de 2 moyennes: $\mu_1 - \mu_2$	distributions normales $N(\mu_1,\sigma^2), N(\mu_2,\sigma^2)$ σ^2 inconnu $s_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ $s_1^2 = \dfrac{1}{n_1-1} \sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)^2$ $s_2^2 = \dfrac{1}{n_2-1} \sum_{i=1}^{n_2} (x_{i2} - \bar{x}_2)^2$	$(\bar{x}_1 - \bar{x}_2) \pm t_{\nu,\alpha/2}\, s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$ $\nu = n_1 + n_2 - 2$

Paramètre	Conditions	Intervalle de confiance
différence de 2 moyennes: $\mu_1 - \mu_2$	distributions normales $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$ $n_1, n_2 \geq 30$ σ_1, σ_2 inconnus	$(\bar{x}_1 - \bar{x}_2) \pm t_{\nu, \alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ (*)
	$a = s_1^2/n_1,\ b = s_2^2/n_2$	$\nu = \frac{(a+b)^2}{a^2/(n_1-1) + b^2/(n_2-1)}$
variance (σ^2)	distribution normale $N(\mu, \sigma^2)$ μ connu $s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$	$\frac{ns^2}{\chi^2_{n, \alpha/2}} \leq \sigma^2 \leq \frac{ns^2}{\chi^2_{n, 1-\alpha/2}}$
variance (σ^2)	distribution normale $N(\mu, \sigma^2)$ μ inconnu $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$	$\frac{(n-1)s^2}{\chi^2_{n-1, \alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{n, 1-\alpha/2}}$
écart-type (σ)	distribution quelconque (μ, σ^2) μ inconnu, $n \geq 30$	$\frac{s}{1 + \frac{z_{\alpha/2}}{\sqrt{2n}}} \leq \sigma \leq \frac{s}{1 - \frac{z_{\alpha/2}}{\sqrt{2n}}}$ (*)

Paramètre	Conditions	Intervalle de confiance
quotient de 2 variances: σ_1^2 / σ_2^2	distributions normales $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ μ_1, μ_2 inconnus	$\frac{s_1^2}{s_2^2} a \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2}{s_2^2} b$ $a = F_{n2-1, n1-1, 1-\alpha/2}$ $b = F_{n2-1, n1-1, \alpha/2}$
proportion θ	distribution Bernoulli	$\bar{x} \pm z_{\alpha/2} \sqrt{\frac{\bar{x}(1-\bar{x})}{n}}$ (*)
différence 2 proportions: $\theta_1 - \theta_2$	distributions Bernoulli $n_1, n_2 \geq 30$	$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\bar{x}_1(1-\bar{x}_1)}{n_1} + \frac{\bar{x}_2(1-\bar{x}_2)}{n_2}}$ (*)
θ	$f(x;\theta)$ $\hat{\theta}$ est. vraisemblance maximale, $n \geq 30$ $B^2 = E\left[\frac{d}{d\theta} \ln f_x(x;\theta)\right]^2$	$\hat{\theta} \pm z_{\alpha/2} \frac{1}{\sqrt{n\,B}}$ (*)

7.8 CALCUL DU NOMBRE D'OBSERVATIONS

Définition du problème

Soit θ un paramètre à estimer et $\hat{\theta}$ un estimateur. Combien d'observations n doit-on prendre si on veut estimer θ avec une certaine précision? Pour répondre à cette question, on doit fournir les informations suivantes:

(a) la précision désirée E

sous la forme de la différence maximum E entre θ et $\hat{\theta}$

$$|\theta - \hat{\theta}| < E$$

(b) l'écart-type σ de la population

sinon, au moins une information sur l'étendue et la forme de la distribution qui peut servir à calculer σ comme nous le verrons plus loin.

(c) le coefficient de confiance $1 - \alpha$

est un coefficient qui a la même interprétation que celui des intervalles de confiance. En général, la valeur de $1 - \alpha$ est supérieure à 0.90.

Les informations (a)-(b)-(c) se traduisent par l'équation

$$P[\ |\hat{\theta} - \theta| < E\ ] = 1 - \alpha \qquad (7.79)$$

Supposons que (approximativement)

$$\hat{\theta} \sim N\left(\theta, \frac{1}{nB^2}\right)$$

Il suit que

$$P\left[\ \sqrt{n}\ B\ |\ \hat{\theta} - \theta\ |\ < E\ \sqrt{n}\ B\ \right] = 1 - \alpha$$

et puisque $\sqrt{n}\, B\, (\hat{\theta} - \theta) \sim N(0,1)$ on a

$$E\sqrt{n}\, B = z_{\alpha/2}$$

$$n = \left[\frac{z_{\alpha/2}}{E}\right]^2 \frac{1}{B^2} \qquad (7.80)$$

La valeur de B dépend du type de paramètre θ que l'on veut estimer et les exemples 7.33 et 7.34 présentent la solution de deux cas particuliers importants.

Exemple 7.33: Estimation de la moyenne θ d'une population $N(\theta, \sigma^2)$

$$f_x(x;\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left[\frac{x-\theta}{\sigma}\right]^2\right]$$

$$\frac{d}{d\theta} \ln f_x(x;\theta) = \left[\frac{x-\theta}{\sigma^2}\right]$$

$$E\left[\frac{d}{d\theta} \ln f_x(x;\theta)\right]^2 = \frac{1}{\sigma^4} E\left[X-\theta\right]^2 = \frac{1}{\sigma^2}$$

$$B^2 = \frac{1}{\sigma^2}$$

La formule (7.80) devient $$n = \left[\frac{z_{\alpha/2}\ \sigma}{E}\right]^2 \qquad (7.81)$$

Remarques

(a) La valeur de n dépend du quotient σ/E. Si on ne connaît pas la valeur de σ, on peut exprimer E en termes de σ, $E = f\sigma$ où $f > 0$ est une constante appropriée. L'équation (7.81) s'écrit alors

$$n = \left[\frac{z_{\alpha/2}}{f}\right]^2 \qquad (7.82)$$

Le tableau 7.4 a été calculé selon cette formule.

Tableau 7.4: taille échantillonnale pour estimer une moyenne

$1-\alpha$ \ f	0.1	0.25	0.50	1.00
0.90	271	44	11	3
0.95	384	62	16	4
0.99	664	106	27	7

(b) Quelquefois la précision désirée E est exprimée en pourcentage E* de la moyenne μ

Alors $$E = \frac{\mu E^*}{100}$$

où E^* est un nombre entre 0 et 100. Alors, l'équation de définition de n est

$$P\left[|\bar{X} - \mu| < \frac{\mu E^*}{100} \right] = 1 - \alpha$$

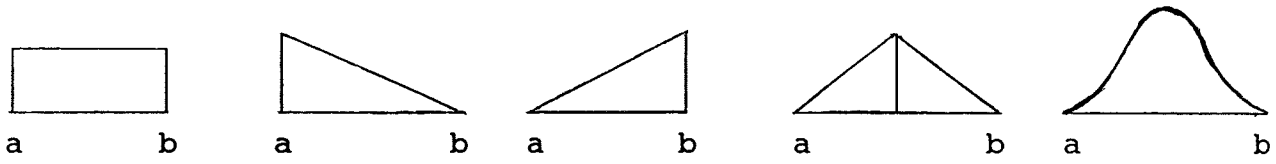
dont la solution est

$$n = \left(\frac{z_{\alpha/2}}{E^*} \right)^2 V^2 \qquad (7.83)$$

où $V = 100 \frac{\sigma}{\mu}$ est le coefficient de variation exprimé en pourcentage.

(c) Lorsque σ est inconnu

Si aucune estimation de σ est disponible, on peut utiliser une des formes suivantes de distributions pour avoir une approximation de σ. Dans ce cas de doute, on utilisera la distribution rectangulaire. Cette méthode exige de l'utilisateur qu'il puisse préciser la valeur de a et b.



$\sigma = (b-a)/3.5, \sigma = (b-a)/4.2, \sigma = (b-a)/4.2, \sigma = (b-a)/4.9, \sigma = (b-a)/6$

Figure 7.3: relation entre σ et la forme de quelques distributions

Exemple 7.34: Estimation d'une proportion θ $(0 < \theta < 1)$

On sait que $\hat{\theta} = \bar{X}$ et la distribution de $\bar{X}$ est approximativement

$$\bar{X} \sim N\left(\theta, \frac{\theta(1 - \theta)}{n}\right)$$

On trouve

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \theta(1 - \theta) \qquad (7.84)$$

Si on possède une estimation préalable de θ, on peut alors calculer n. Sinon, on suppose $\theta = 0.5$ correspondant au pire cas puisque $\theta(1 - \theta) \leq 0.25$ quelque soit θ. On obtient alors

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \frac{1}{4} \qquad (7.85)$$

Le tableau 7.5 a été calculé à l'aide de l'équation (7.84)

Tableau 7.5: taille échantillonnale pour estimer une proportion

$1 - \alpha = 0.90$

E \ θ	0.1	0.2	0.3	0.4	0.5
0.01	2436	4330	5683	6495	6765
0.02	609	1083	1421	1624	1692
0.03	271	481	632	722	752
0.05	98	174	228	260	271

$1 - \alpha = 0.95$

E \ θ	0.1	0.2	0.3	0.4	0.5
0.01	3458	6147	8068	9220	9604
0.02	865	1537	2017	2305	2401
0.03	385	683	897	1025	1068
0.05	139	246	323	369	385

C H A P I T R E 8

TESTS D'HYPOTHÈSES

8.0 SOMMAIRE

Ce chapitre est consacré au problème de la construction de procédures statistiques afin de mettre à l'épreuve des hypothèses statistiques concernant la nature des données. Les principaux tests sont présentés: tests d'ajustement, tests d'indépendance, tests sur une ou plusieurs moyennes, tests sur une ou plusieurs variances, tests sur une ou plusieurs proportions. Pour chaque test on présente: les conditions d'applications, la procédure du test et un exemple. L'expression de la fonction caractéristique et le calcul de la taille échantillonnale sont présentés pour certains tests.

8.1 CLASSIFICATION DES TESTS

Les hypothèses statistiques sont des affirmations concernant l'une ou l'autre des situations suivantes:

- la forme d'une distribution: tests d'ajustement
- la valeur d'un paramètre d'une distribution: moyenne, variance, proportion
- l'égalité des paramètres de deux ou plusieurs distributions: moyennes, variances, proportions
- la nullité de certains paramètres dans les modèles statistiques de régression et d'analyse de variance
- l'indépendance entre deux variables qualitatives ou quantitatives
- l'indépendance d'une série d'observations

La liste précédente recouvre les cas pratiques les plus souvent rencontrés mais elle n'est pas exhaustive. Les procédures de tests statistiques sont parmi les plus développées de l'ensemble des méthodes d'analyse statistique de données.

On peut encore classer les tests portant sur les paramètres selon la sorte de paramètre ainsi que le nombre d'échantillons impliqués. La majorité des tests usuels reposent sur des distributions d'échantillonnage telles la gaussienne centrée réduite (N(0,1)), la khi-deux (χ^2_ν), la Student (T_ν) et la Fisher (F). D'autres tests reposent sur des distributions spéciales et nécessitent l'accès à une tabulation correspondante. Le tableau qui suit donne un panorama des tests, la distribution d'échantillonnage employée et le nom du test s'il y a lieu.

Tests paramétriques

nombre d'échantillons indépendants \ paramètre	moyenne	variance	proportion
1	N(0,1) Student	khi-deux	N(0,1)
2	N(0,1) Student	Fisher	khi-deux
3 ou plus	Fisher	Bartlett Hartley	khi-deux

Tests d'ajustement

général: khi-deux de Pearson, Kolmogorov - Smirnov

normalité: Liliefors, Shapiro - Wilk

Tests d'indépendance

2 variables qualitatives: test du khi-deux dans un tableau de contingence

2 variables quantitatives: coefficient de corrélation simple

plusieurs variables quantitatives: coefficient de corrélation multiple

8.2 CONCEPTS DE BASE

Soit X une variable aléatoire et $f_x(x:\theta)$ une famille de distributions de X paramétrée avec θ variant dans un espace Ω. Une hypothèse paramétrique est une déclaration concernant la valeur de θ exprimée sous la forme suivante:

$$H_N: \theta = \theta_0 \qquad (8.1)$$

dite HYPOTHÈSE NULLE. Un TEST STATISTIQUE est une procédure basée sur un échantillon aléatoire $X_1, X_2, \ldots, X_n$ tirée de la population X et permettant de conclure à l'une ou l'autre de deux décisions possibles concernant H_N:

- rejeter H_N
- ne pas rejeter H_N

On dit aussi procédure de mise à l'épreuve puisqu'il s'agit de confronter une hypothèse avec des données et de décider si celles-ci sont en accord ou non avec H_N.

Il y a deux types de décisions erronnées que l'on peut commettre lors de l'exécution d'un test statistique soit:

- rejeter H_N quand H_N est vraie: ERREUR DE PREMIÈRE ESPÈCE
- ne pas rejeter H_N quand H_N est fausse: ERREUR DE DEUXIÈME ESPÈCE

La région de rejet de H_N s'appelle RÉGION CRITIQUE et elle définit un sous-espace A de R^n.

si $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$ appartient à A on rejette H_N

si $\underset{\sim}{x}$ n'appartient pas à A on ne rejette pas H_N.

Définir un test statistique est donc équivalent à la spécification d'une région critique.

Toute procédure de test statistique est contrôlée par les probabilités de commettre ces erreurs

$$P\,[\text{rejeter } H_N \text{ alors que } \theta = \theta_0\,] = \alpha \qquad (8.2)$$

$$P\,[\text{ne pas rejeter } H_N \text{ quand } \theta = \theta_1 \neq \theta_0] = \beta \qquad (8.3)$$

On appelle α le RISQUE DE PREMIÈRE ESPÈCE ou le SEUIL DU TEST et β le RISQUE DE DEUXIÈME ESPÈCE à $\theta = \theta_1$. On convient de définir le test OPTIMAL comme étant celui qui, parmi tous les tests de seuil α, a le plus petit risque de deuxième espèce β.

On note que la valeur de β dépend de la valeur choisie θ_1 de θ et si l'on fait varier θ_1 on obtient une fonction

$$\beta = \beta(\theta) \tag{8.4}$$

appelée la FONCTION CARACTÉRISTIQUE DU TEST.

En particulier $\beta(\theta_0) = 1 - \alpha$ (8.5)

décision basée sur $(X_1, X_2, .., X_n)$ \ Statut de l'hypothèse nulle H_N	H_N vraie	H_N fausse
rejet de H_N	erreur de 1ère espèce α = P (erreur 1ère espèce) = risque 1ère espèce	bonne décision $1 - \beta$ = puissance
non rejet de H_N	bonne décision $1 - \alpha$	erreur de 2è espèce β = P (erreur 2è espèce) = risque 2è espèce

8.3 RÉSULTATS GÉNÉRAUX POUR LES TESTS PARAMÉTRIQUES (*)

Une hypothèse est dite SIMPLE si elle détermine complètement la distribution de probabilité de X; dans le cas contraire elle est dite HYPOTHÈSE COMPOSÉE.

Par exemple, si la distribution de X est $N(\theta_1, \theta_2^2)$

$$f_x(x;\theta_1,\theta_2) = \frac{1}{\theta_2\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\theta_1}{\theta_2}\right)^2\right] \qquad (8.6)$$

alors

H: $\theta_1 = 10$, $\theta_2 = 1$ est une hypothèse simple

H: $\theta_1 < 10$, $\theta_2 = 1$ est une hypothèse composée

La détermination des procédures statistiques optimales pour des hypothèses simples est explicitée dans le résultat fondamental de NEYMAN-PEARSON.

Proposition 8.1: résultat de Neyman-Pearson

Soit X distribuée selon $f_x(x;\theta)$ et H_N, H_A les deux hypothèses simples:

$$H_N: \theta = \theta_0 \qquad H_A: \theta = \theta_1$$

Soit $L(\theta) = \prod_{i=1}^{n} f_x(x_i;\theta)$ la fonction de vraisemblance de θ.

Alors la région A* définie par

$$A^* = \left\{ (x_1,..x_n): \frac{L(\theta_1)}{L(\theta_0)} \geq c \right\} \qquad (8.7)$$

possède la propriété optimale suivante:

Parmi toutes les régions A ayant le même risque de première espèce α, A* possède le plus petit risque de deuxième espèce β.

Démonstration

Posons $dx = \prod_{i=1}^{n} dx_i$

(*) Cette section n'est pas absolument nécessaire pour la suite du chapitre.

Par définition du risque de première espèce

$$\int_{A^*} L(\theta_0)\, dx = \int_{A} L(\theta_0)\, dx = \alpha$$

Donc
$$\int_{A^* \cap A'} L(\theta_0)\, dx = \int_{A^{*\prime} \cap A} L(\theta_0)\, dx \qquad (8.8)$$

puisque l'intégrale sur la partie commune $A^* \cap A$ peut être éliminée de deux membres de la première équation.

D'autre part

$$\beta^* = \int_{A^{*\prime}} L(\theta_1)\, dx = 1 - \int_{A^*} L(\theta_1)\, dx$$

et
$$\beta = \int_{A'} L(\theta_1)\, dx = 1 - \int_{A} L(\theta_1)\, dx$$

La différence entre β^* et β peut s'écrire

$$\beta^* - \beta = \int_{A} L(\theta_1)\, dx - \int_{A^*} L(\theta_1)\, dx$$

$$= \int_{A' \cap A^*} L(\theta_1)\, dx - \int_{A \cap A^{*\prime}} L(\theta_1)\, dx$$

Mais dans $A' \cap A^*$: $L(\theta_1) \geq c\, L(\theta_0)$

et dans $A \cap A^{*\prime}$: $L(\theta_1) \leq c\, L(\theta_0)$

Donc

$$\beta^* - \beta \leq \int_{A' \cap A^*} c\, L(\theta_0)\, dx - \int_{A \cap A^{*\prime}} c\, L(\theta_0)\, dx$$

$$\leq c \left[\int_{A' \cap A^*} L(\theta_0)\, dx - \int_{A \cap A^{*\prime}} L(\theta_0)\, dx \right]$$

≤ 0 puisque d'après (8.8) le terme entre crochets est nul.

Exemple 8.1: distribution exponentielle

$$f_x(x;\theta) = \theta\, e^{-\theta x} \qquad x \geq 0,\ \theta > 0$$

$$H_N: \theta = \theta_0 \qquad H_A: \theta = \theta_1 < \theta_0$$

Solution

$$L(\theta) = \prod_{i=1}^{n} f_x(x_i;\theta) = \theta^n \exp\left[-\theta \sum_{i=1}^{n} x_i\right]$$

D'après le résultat de Neyman-Pearson la meilleure région critique A* est définie par:

$$A^* = \left\{ \underset{\sim}{x} = (x_1, x_2, \ldots, x_n): \frac{L(\theta_1)}{L(\theta_0)} \geq c \right\}$$

Mais

$$\frac{L(\theta_1)}{L(\theta_0)} = \left(\frac{\theta_1}{\theta_0}\right)^n \exp\left[-(\theta_1-\theta_0) \sum_{1=1}^{n} x_i\right]$$

et

$$\frac{L(\theta_1)}{L(\theta_0)} \geq c \qquad \text{s'écrit}$$

$$\sum_{i=1}^{n} x_i \geq \frac{1}{(\theta_0-\theta_1)} \ln\left[c\left(\frac{\theta_1}{\theta_0}\right)^n\right] = c_1$$

ou encore

$$\bar{x} \geq c_2 = nc_1$$

Donc $A^* = \{ (x_1 \ldots, x_n): \bar{x} \geq c_2 \}$ est la meilleure région critique et la valeur de c_2 est fixée pour obtenir un test de seuil α.

Exemple 8.2: distribution normale à variance connue

$$f_x(x;\theta,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2\right] \qquad \sigma \text{ connu}$$

$$H_N: \theta = \theta_0 \qquad H_A: \theta = \theta_1 < \theta_0$$

Solution

$$L(\theta) = \prod_{i=1}^{n} f_x(x_i;\theta,\sigma) = \sigma^{-n}(2\pi)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\theta)^2\right]$$

On a

$$\frac{L(\theta_1)}{L(\theta_0)} = \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta_1)^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta_0)^2\right]$$

et $\dfrac{L(\theta_1)}{L(\theta_0)} \geq c$ donne

$$\exp\left[\frac{1}{2\sigma^2}\left(\Sigma(x_i-\theta_0)^2 - \Sigma(x_i-\theta_1)^2\right)\right] \geq c$$

$$\Sigma(x_i-\theta_0)^2 - \Sigma(x_i-\theta_1)^2 \geq 2\sigma^2 \ell nc = c_1$$

$$2(\theta_1-\theta_0)\Sigma x_i \geq c_1 + (\theta_1^2 - \theta_0^2)n$$

$$\bar{x} \leq \frac{c_1 + (\theta_1^2 - \theta_0^2)n}{2n(\theta_1-\theta_0)} = c_3$$

Donc $A^* = \{ \bar{x}: \bar{x} \leq c_3 \}$ est la région critique optimale

La constante c_3 est déterminée de telle sorte que A^* soit de seuil α. On a

$$\frac{\bar{X}-\theta_0}{\sigma/\sqrt{n}} \leq \frac{c_3-\theta_0}{\sigma/\sqrt{n}} \qquad \text{sous } H_N$$

et puisque $\dfrac{\bar{X}-\theta_0}{\sigma/\sqrt{n}}$ est une variable normale centrée-réduite

$$\frac{c_3-\theta_0}{\sigma/\sqrt{n}} = z_{1-\alpha} = -z_\alpha$$

et finalement

$$c_3 = \theta_0 - z_\alpha \sigma/\sqrt{n} \qquad (8.9)$$

<u>Principe du quotient du maximum de vraisemblance</u>

Lorsque les hypothèses sont composées, il n'existe pas, en général, de meilleur test. On a développé le principe du quotient du maximum de vraisemblance qui permet de générer des régions critiques.

Les hypothèses composées sont très souvent rencontrées dans le cadre de paramètres multidimensionnels $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ d'une distribution de probabilité $f_x(x; \theta_1, \ldots, \theta_p)$. Les paramètres $(\theta_1, \ldots, \theta_p)$ varient dans un espace de dimension p disons Ω. L'hypothèse nulle H_N est composée et de la forme

$$H_N: \quad (\theta_1, \ldots, \theta_p) \in \omega$$

où ω est un sous-espace de Ω de dimension r où $1 \leq r \leq p$.

Soit $f_x(x; \theta_1, \ldots, \theta_p) = f_x(x; \underset{\sim}{\theta})$ la distribution de probabilité de X et

$$L(\underset{\sim}{\theta}) = \prod_{i=1}^{n} f_x(x_i; \underset{\sim}{\theta})$$

la fonction de vraisemblance de $\underset{\sim}{\theta}$. Notons par $\hat{\underset{\sim}{\theta}}_\Omega$ la valeur de $\underset{\sim}{\theta}$ qui rend $L(\underset{\sim}{\theta})$ maximum dans l'espace Ω et $\hat{\underset{\sim}{\theta}}_\omega$ la valeur de $\underset{\sim}{\theta}$ qui rend $L(\underset{\sim}{\theta})$ maximum dans l'espace ω.

On définit le rapport

$$\Lambda = \frac{L(\hat{\underset{\sim}{\theta}}_\omega)}{L(\hat{\underset{\sim}{\theta}}_\Omega)} \tag{8.10}$$

appelé le QUOTIENT DU MAXIMUM DE VRAISEMBLANCE de H_N

On remarque que Λ est une variable aléatoire comprise entre 0 et 1

$$0 \leq \Lambda \leq 1 \tag{8.11}$$

La valeur de Λ mesure le degré de vraisemblance de H_N et peut servir de statistique pour tester H_N.

<u>Proposition 8.2: test du quotient de vraisemblance</u>

On rejette H_N si $\Lambda \leq \Lambda_0$ (8.12)

où Λ_0 est choisie pour obtenir un test de seuil α.

<u>Exemple 8.3</u>: distribution normale $N(\mu, \sigma^2) = N(\theta_1, \theta_2)$

$$f_x(x; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} \exp\left[-\frac{1}{2\theta_2^2}(x-\theta_1)^2\right]$$

H_N: $\theta_1 = \theta_0, \quad \theta_2 > 0$

$$L(\theta_1,\theta_2) = \prod_{i=1}^{n} f_x(x_i, \theta_1, \theta_2)$$

$$= \theta_2^{-n/2} (2\pi)^{-n/2} \exp\left[-\frac{1}{2\theta_2^2} \Sigma(x_i - \theta_1)^2\right]$$

Ici $\Omega = [(\theta_1,\theta_2) : \theta_1 \in R, \quad \theta_2 > 0]$

$\omega = [(\theta_1,\theta_2) : \theta_1 = \theta_0, \quad \theta_2 > 0]$

$$L(\underset{\sim}{\theta}_\omega) = \max_{\omega} L(\theta_1,\theta_2) = \max_{\theta_2} L(\theta_0,\hat{\theta}_2)$$

$$= \hat{\theta}_2^{-n/2} (2\pi)^{-n/2} \exp\left[-\frac{1}{2\hat{\theta}_2^2} \sum_{i=1}^{n} (x_i - \theta_0)^2\right]$$

où $\hat{\theta}_2^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \theta_0)^2$

Donc $L(\underset{\sim}{\hat{\theta}}_\omega) = (2\pi)^{-n/2} \hat{\theta}_2^{-n/2} \exp(-n/2)$

D'autre part

$$L(\underset{\sim}{\hat{\theta}}_\Omega) = \max_{\Omega} L(\theta_1,\theta_2) = L(\theta_1 = \bar{x}, \tilde{\theta}_2)$$

$$= (\tilde{\theta}_2)^{-n/2} (2\pi)^{-n/2} \exp\left[-\frac{1}{2\tilde{\theta}_2} \sum_{i=1}^{n} (x_i - \bar{x})^2\right]$$

où $\tilde{\theta}_2^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$

Donc

$$L(\underset{\sim}{\hat{\theta}}_\Omega) = (2\pi)^{-n/2} \tilde{\theta}_2^{-n/2} \exp(-n/2)$$

Alors

$$\Lambda = \frac{\hat{\theta}_2^{-n/2}}{\tilde{\theta}_2^{-n/2}} = \left[\frac{\Sigma (x_i-\theta_0)^2}{\Sigma (x_i-\bar{x})^2} \right]^{-n/2}$$

$$\Lambda^{2/n} = \frac{\Sigma (x_i-\bar{x})^2}{\Sigma (x_i-\theta_0)^2} = \frac{\Sigma(x_i-\bar{x})^2}{\Sigma(x_i-\bar{x})^2 + n(\bar{x}-\theta_0)^2}$$

$$= \frac{1}{1 + \dfrac{n(\bar{X}-\theta_0)^2}{\Sigma(x_i-\bar{X})^2}} = \frac{1}{1 + \dfrac{T^2}{n-1}}$$

où $T = \dfrac{(\bar{X}-\theta_0)}{s/\sqrt{n}}$ suit une distribution de Student et

$s^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i-\bar{x})^2$ est la variance échantillonnale

Le test du quotient du maximum de vraisemblance peut s'écrire sous la forme suivante:

On rejette H_N si $T = \dfrac{|\bar{X} - \theta_0|}{s/\sqrt{n}} \geq c$

où $c = t_{n-1,\alpha/2}$

Proposition 8.3: résultat de Wilks

Sous certaines conditions de régularité, la distribution de $-2 \ell n \Lambda$ suit approximativement une distribution khi-deux avec ν degrés de liberté où $\nu = p-r$

8.4 TEST SUR UNE MOYENNE: VARIANCE CONNUE

Soit $\mu = E(X)$ la moyenne d'une population X. Nous voulons tester l'hypothèse nulle H_N définie par

$$H_N: \mu = \mu_0$$

à l'aide d'un échantillon de taille n, $X_1, X_2, \ldots, X_n$ tirée de la population X. Nous allons étudier le problème selon diverses hypothèses de base concernant la distribution de X et le type de contre hypothèse (alternative).

Cas A: $X \sim N(\mu, \sigma^2)$

$H_A: \mu = \mu_1 > \mu_0$ (alternative unilatérale à droite)

La région critique optimale selon le lemme de Neyman-Pearson est:

$$\text{rejeter } H_N \text{ si } \bar{X} > c \tag{8.13}$$

On peut écrire (8.13) sous la forme:

$$\text{rejeter } H_N \text{ si } \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > \frac{c-\mu_0}{\sigma/\sqrt{n}} \tag{8.14}$$

Puisque la variable

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

est distribuée selon une loi N(0,1), un test de seuil α donne

$$\frac{c - \mu_0}{\sigma/\sqrt{n}} = z_\alpha \tag{8.15}$$

et

$$c = \mu_0 + z_\alpha \, \sigma/\sqrt{n} \tag{8.16}$$

où z_α est le 100 $(1-\alpha)$ième percentile d'une loi N(0,1)

Fonction caractéristique (ou d'efficacité)

La probabilité $\beta(\mu)$ d'accepter H_N lorsque la moyenne de la population est μ

$$\beta(\mu) = P\left[\bar{X} < \mu_0 + z_\alpha\, \sigma/\sqrt{n}\right]$$

$$= P\left[\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_\alpha + \left(\frac{\mu_0-\mu}{\sigma}\right)\sqrt{n}\right]$$

$$= \Phi(z_\alpha - \delta\sqrt{n}) \qquad (8.17)$$

$$= \beta(\alpha, \delta, n)$$

où $$\delta = \frac{\mu-\mu_0}{\sigma} \qquad (8.18)$$

est appelé le PARAMÈTRE DE NON-CENTRALITÉ

Cette fonction a été tracée à la figure 8.1 pour:

α = 0.01, 0.05
n = 1(1)10, 15, 20, 30, 40, 50, 75, 100
$-1 \leq \delta \leq 3$

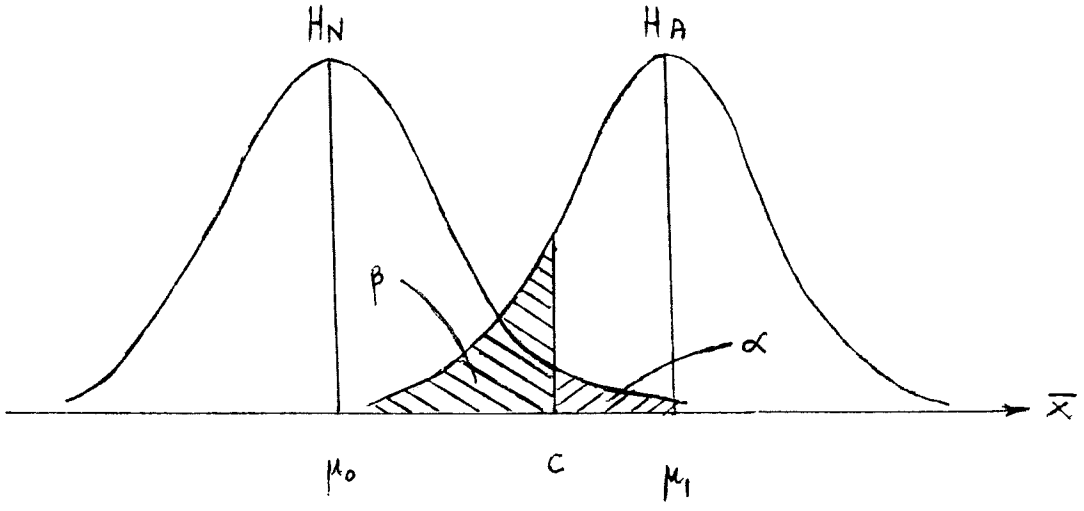
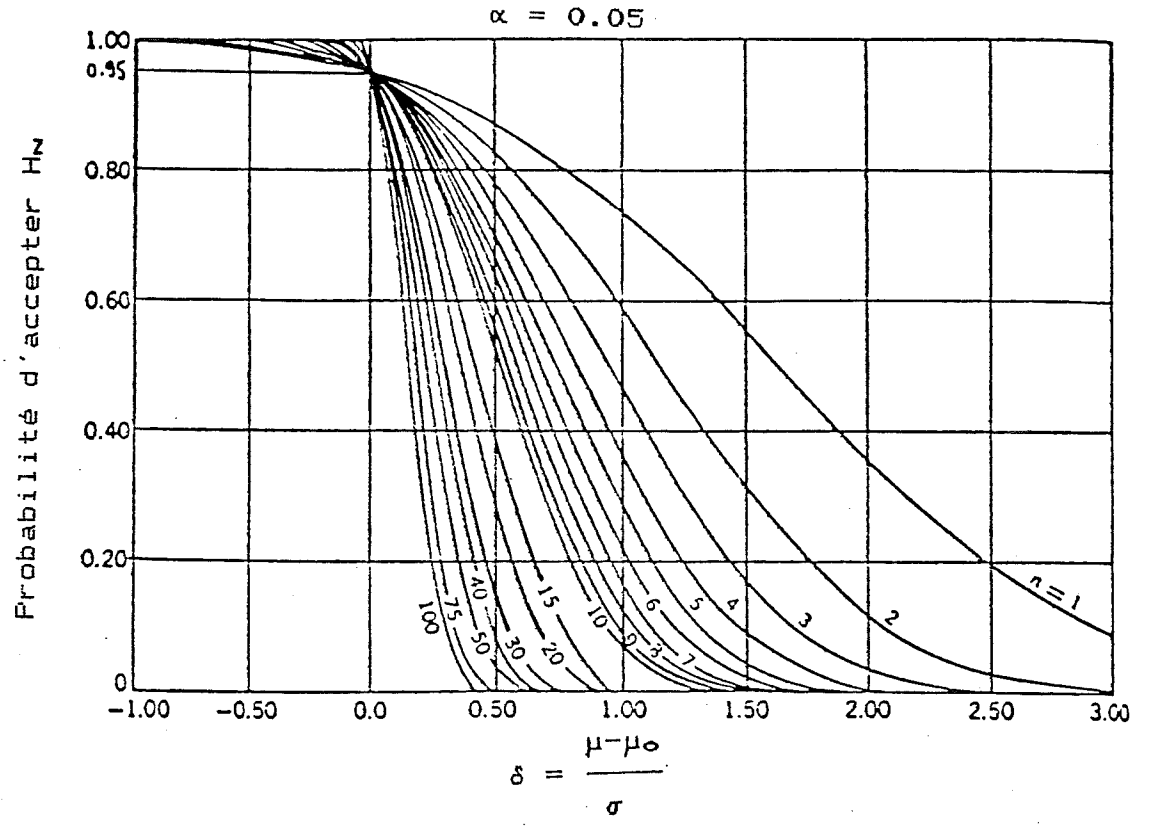
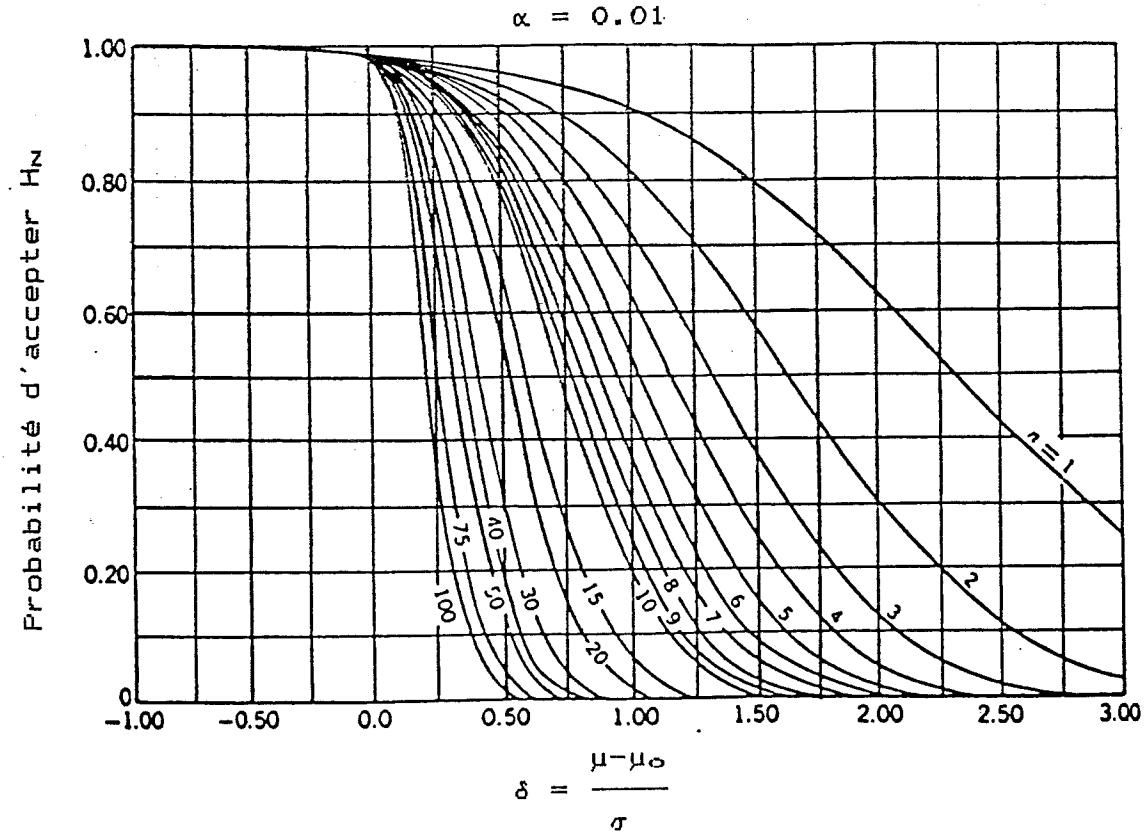


fig 8.0 : risque α et risque β

Figure 8.1 Courbe caractéristique du test H_n: $\mu = \mu_0$
distribution normale avec σ connu
et alternative unilatérale





Calcul de n

Si on fixe la valeur du risque de deuxième expèce à β pour une valeur μ_1 on a

$$\beta(\mu_1) = \Phi(z_\alpha - \left(\frac{\mu_1-\mu_0}{\sigma}\right)\sqrt{n}\,) = \beta$$

$$z_\alpha - \left(\frac{\mu_1-\mu_0}{\sigma}\right)\sqrt{n} = \Phi^{-1}(\beta) = z_{1-\beta} = -z_\beta$$

et si on isole n de cette dernière équation, on obtient

$$n = (z_\alpha + z_\beta)^2 \left(\frac{\sigma}{\mu_1-\mu_0}\right)^2 \qquad (8.19)$$

Exemple 8.4:

Un acier d'un alliage spécial a une tension de rupture moyenne de 25800 psi et un écart-type de 300 psi. Un changement dans la composition de l'alliage devrait augmenter la tension moyenne sans changer l'écart-type. Si le nouvel alliage ne produit aucun changement dans la tension moyenne, on veut pouvoir le conclure avec une probabilité de 0.99. D'autre part, si la tension moyenne est augmentée de 250 psi, on veut que le risque de ne pas détecter ce changement soit de 0.10.

(a) Définir l'hypothèse à tester, son alternative, le seuil du test et le risque de deuxième espèce.

(b) Posez les équations définissant les risques de 1ère espèce et de 2e espèce et les résoudre afin de trouver la taille de l'échantillon à prélever ainsi que la région critique (rejet) du test.

(c) Calculez la fonction caractéristique du test à 25800, 26000, 26050 et 26100.

(d) Un échantillon de 19 observations a donné $\bar{x} = 25970$. Quelle est la conclusion?

Solution

On a

$$H_N: \mu = \mu_0 = 25800$$

$$H_A: \mu > 25800$$

$$\sigma = 300 \ , \quad \alpha = 0.01$$

$$\beta(\mu_1 = 25800 + 250) = \beta(\mu_1 = 26050) = 0.10$$

$$P\left[\bar{X} > c \text{ si } \mu = 25800\right] = 0.01$$

$$P\left[\bar{X} < c \text{ si } \mu = 26050\right] = 0.10$$

D'après les équations (8.19) et (8.16)

$$n = (2.33 + 1.28)^2 \ (300/250)^2$$

$$= 18.77$$

Donc

$$n = 19$$

$$c = 25800 + 2.33 * 300/\sqrt{19}$$

$$= 25960.36$$

La fonction caractéristique est

$$\beta(\mu) = \Phi\left[z_\alpha - \left[\frac{\mu-\mu_0}{\sigma}\right] \sqrt{n}\right]$$

$$= \Phi\left[2.33 - \left[\frac{\mu-25800}{300}\right] \sqrt{19}\right]$$

μ	25800	26000	26050	26100
$\beta(\mu)$	0.99	0.28	0.10	0.02

D'autre part puisque $\bar{X} = 25970 > c$ on rejette H_N

Cas B: $X \sim N(\mu, \sigma^2)$

H_A: $\mu = \mu_1 < \mu_0$ (alternative unilatérale à gauche)

La région critique de H_N est:

$$\text{rejeter } H_N \text{ si } \bar{X} < c \tag{8.20}$$

et en fixant le seuil du test à α on a:

$$c = \mu_0 - z_\alpha \, \sigma/\sqrt{n} \tag{8.21}$$

Cette région critique est optimale au sens du lemme de Neyman-Pearson.

Fonction caractéristique

$$\begin{aligned}
\beta(\mu) &= P\left[\bar{X} > \mu_0 - z_\alpha \, \sigma/\sqrt{n}\right] \\
&= P\left[\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > -z_\alpha + \left(\frac{\mu_0-\mu}{\sigma}\right)\sqrt{n}\right] \\
&= 1 - \Phi\left(-z_\alpha - \left(\frac{\mu_0-\mu}{\sigma}\right)\sqrt{n}\right) \qquad (8.22) \\
&= \Phi(z_\alpha + \delta\sqrt{n}) \\
&= \beta(\alpha, \delta, n)
\end{aligned}$$

où δ est défini par (8.18)

Les graphiques 8.1 peuvent être utilisés à condition de faire une réflexion des courbes autour de l'axe vertical au point 0.

Calcul de n

La formule (8.19) est aussi valable pour le cas B.

Cas C: $X \sim N(\mu, \sigma^2)$

$H_A: \mu = \mu_1 \neq \mu_0$ (alternative bilatérale)

La région critique est de la forme:

$$\text{rejeter } H_N \text{ si } \quad \bar{X} < c_1 \text{ ou } \bar{X} > c_2 \tag{8.23}$$

Si on partage également le risque de première espèce α sur chaque morceau de la région critique on obtient

$$\begin{aligned} c_1 &= \mu_0 - z_{\alpha/2} * \sigma/\sqrt{n} \\ c_2 &= \mu_0 + z_{\alpha/2} * \sigma/\sqrt{n} \end{aligned} \tag{8.24}$$

D'une manière équivalente la région critique peut s'écrire:

$$\text{rejeter } H_N \text{ si } \quad \frac{|\bar{X}-\mu_0|}{\sigma/\sqrt{n}} > z_{\alpha/2} \tag{8.25}$$

Fonction caractéristique

$$\beta(\mu) = P[c_1 < \bar{X} < c_2]$$

$$= \Phi\left[\frac{c_2-\mu}{\sigma/\sqrt{n}}\right] - \Phi\left[\frac{c_1-\mu}{\sigma/\sqrt{n}}\right]$$

et en remplaçant c_1 et c_2 par l'équation (8.24)

$$\begin{aligned} \beta(\mu) &= \Phi\left[z_{\alpha/2} - \left[\frac{\mu-\mu_0}{\sigma}\right]\sqrt{n}\right] - \Phi\left[-z_{\alpha/2} - \left[\frac{\mu-\mu_0}{\sigma}\right]\sqrt{n}\right] \\ &= \Phi(z_{\alpha/2} - \delta\sqrt{n}) + \Phi(z_{\alpha/2} + \delta\sqrt{n}) - 1 \qquad (8.26) \\ &= \beta(\alpha, \delta, n) \end{aligned}$$

où $\delta = \dfrac{\mu-\mu_0}{\sigma}$

Cette fonction est tracée à la figure 8.2 pour les valeurs suivantes:

$$\alpha = 0.05,\ 0.01$$

$$n = 1(1)\ 10,\ 15,\ 20,\ 30,\ 40,\ 50,\ 75,\ 100$$

$$0 \leq \delta \leq 4$$

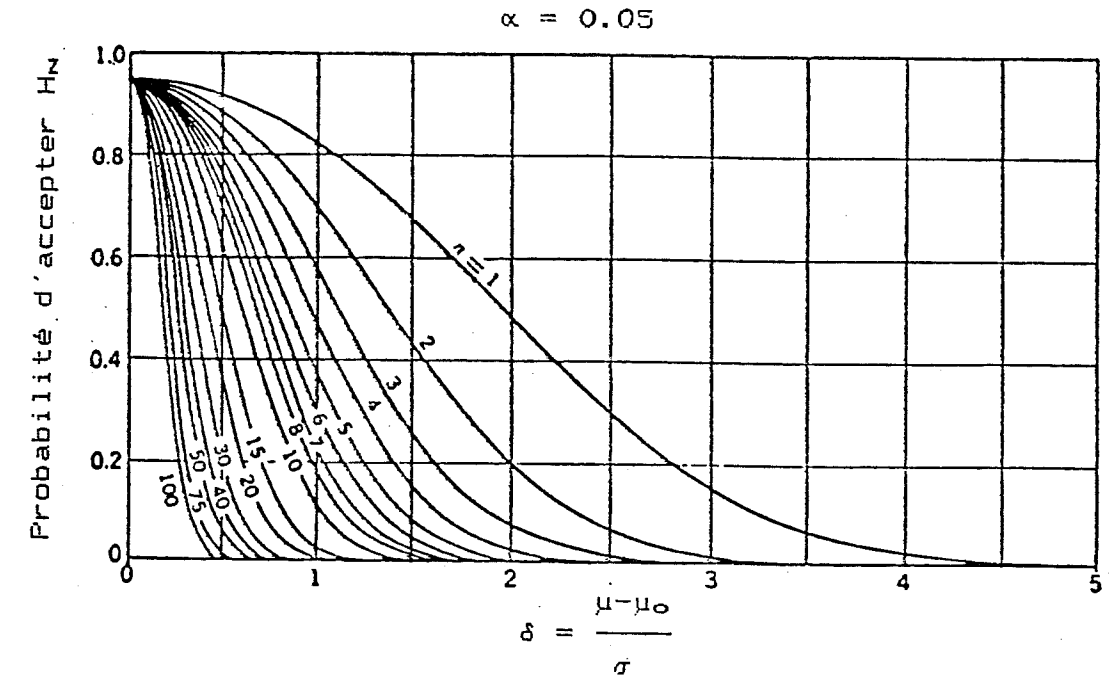
Pour les valeurs de $\delta < 0$ on utilise la relation

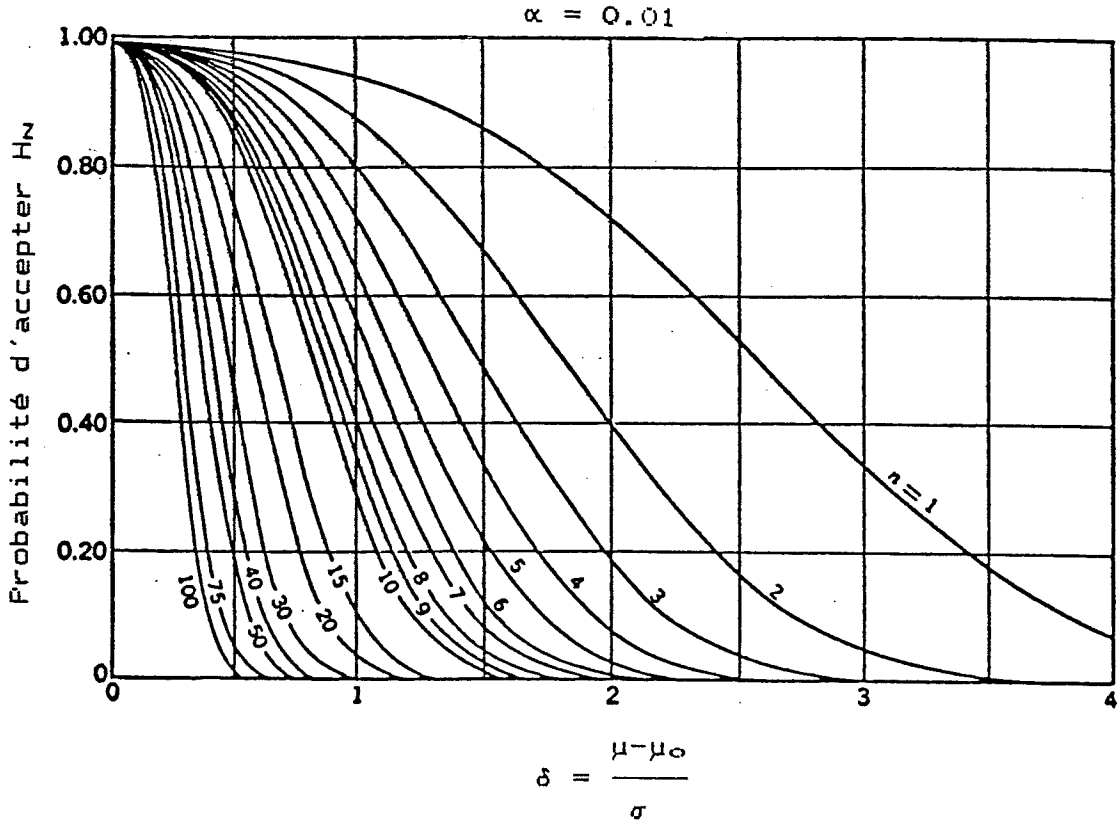
$$\beta(\alpha, -\delta, n) = \beta(\alpha, \delta, n)$$

Figure 8.2 Courbe caractéristique du test H_N: $\mu = \mu_0$

Distribution normale avec σ connu

Alternative bilatérale





Calcul de n

On doit résoudre pour n, l'équation (8.26) en fixant une valeur β au risque de deuxième espèce pour $\mu = \mu_1$. On a

$$\beta(\mu_1) = \Phi\ (z_{\alpha/2} - \delta\sqrt{n}) + \Phi\ (z_{\alpha/2} + \delta\sqrt{n}) - 1$$

où $$\delta = \frac{\mu_1 - \mu_0}{\sigma}$$

Si l'on suppose que $\mu_1 < \mu_0$ alors $\delta < 0$ et

$$\Phi\ (z_{\alpha/2} - \delta\sqrt{n}) - 1 \approx 0$$

$$\beta(\mu_1) \approx \Phi\ (z_{\alpha/2} + \delta\sqrt{n}) = \beta$$

Donc $$z_{\alpha/2} + \delta\sqrt{n} = \Phi^{-1}(\beta) = z_{1-\beta} = -z_\beta$$

et finalement

$$n = (z_{\alpha/2} + z_\beta)^2 \left[\frac{\sigma}{\mu_1 - \mu_0}\right]^2 \qquad (8.27)$$

Si $\mu_1 > \mu_0$ un raisonnement analogue conduit aussi à l'équation (8.27).

Cas D: $X \sim D(\mu, \sigma^2)$ où D est une distribution quelconque de moyenne μ et variance σ^2, $n \geq 30$

Si on examine les cas A, B et C on se rend compte que la détermination des régions critiques reposent sur le fait que

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

est distribuée selon une loi N(0,1).

D'autre part le théorème central-limite nous assure de ce résultat d'une manière approximative sous la condition que $n \geq 30$. C'est donc dire que les tests précédents sont approximativement valables si $n \geq 30$.

8.5 TEST SUR UNE MOYENNE: VARIANCE INCONNUE

Cas A: alternative unilatérale

$$H_N: \mu = \mu_0$$

$$H_A: \mu > \mu_0$$

Nous avons vu (section 8.3) que le meilleur est:

$$\text{rejeter } H_N \text{ si } \bar{X} > c \tag{8.28}$$

$$\text{ou } \frac{\bar{X}-\mu_0}{S/\sqrt{n}} > \frac{c-\mu_0}{S/\sqrt{n}}$$

Puisque $\frac{\bar{X}-\mu_0}{S/\sqrt{n}}$ est une variable de Student avec n-1 degrés de liberté sous l'hypothèse H_N

$$\frac{c-\mu_0}{S/\sqrt{n}} = t_{n-1,\alpha} \tag{8.29}$$

où α est le seuil du test, $t_{n-1,\alpha}$ est le 100 $(1-\alpha)$ième percentile d'une distribution de Student et S est l'écart-type échantillonnal.

Fonction caractéristique

$$\beta(\mu) = P\left[\frac{\bar{X}-\mu_0}{S/\sqrt{n}} < t_{n-1,\alpha}\right] \tag{8.30}$$

On a

$$W = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} + \frac{\mu-\mu_0}{\sigma/\sqrt{n}} \frac{\sigma}{S}$$

$$= U + \left(\frac{\mu-\mu_0}{\sigma}\right)\sqrt{n}\sqrt{\frac{n-1}{(n-1)S^2/\sigma^2}}$$

$$= U + \left[\frac{\mu-\mu_0}{\sigma}\right] \sqrt{n}\,\sqrt{n-1/V}$$

On sait que U suit une distribution de Student, T_{n-1}

V suit une distribution khi-deux, χ^2_{n-1}

La variable W suit une distribution de probabilité dite distribution de STUDENT NON-CENTRALE avec paramètre de non-centralité $\delta\sqrt{n}$ où

$$\delta = (\mu-\mu_0)/\sigma \qquad (8.31)$$

L'expression analytique exacte de (8.30) est malaisée à manipuler. On constate que cette fonction dépend de $\alpha, (\mu-\mu_0)/\sigma$ ainsi que de n

$$\beta(\mu) = \beta(\alpha, \delta, n)$$

Cette fonction a été calculée et tracée à la figure 8.3 pour les valeurs suivantes des paramètres α, n et δ

$$\alpha = 0.05,\ 0.01$$

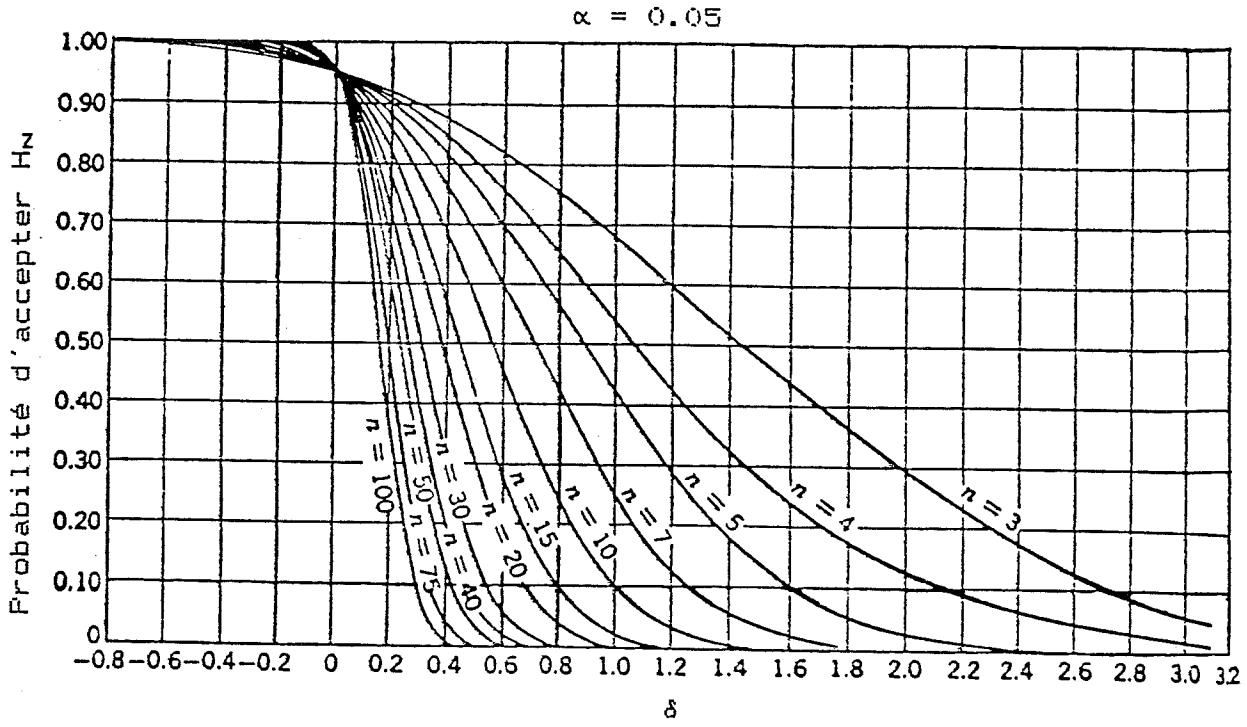
$$n = 3,4,5,5,7,10,15,20,30,40,50,75,100$$

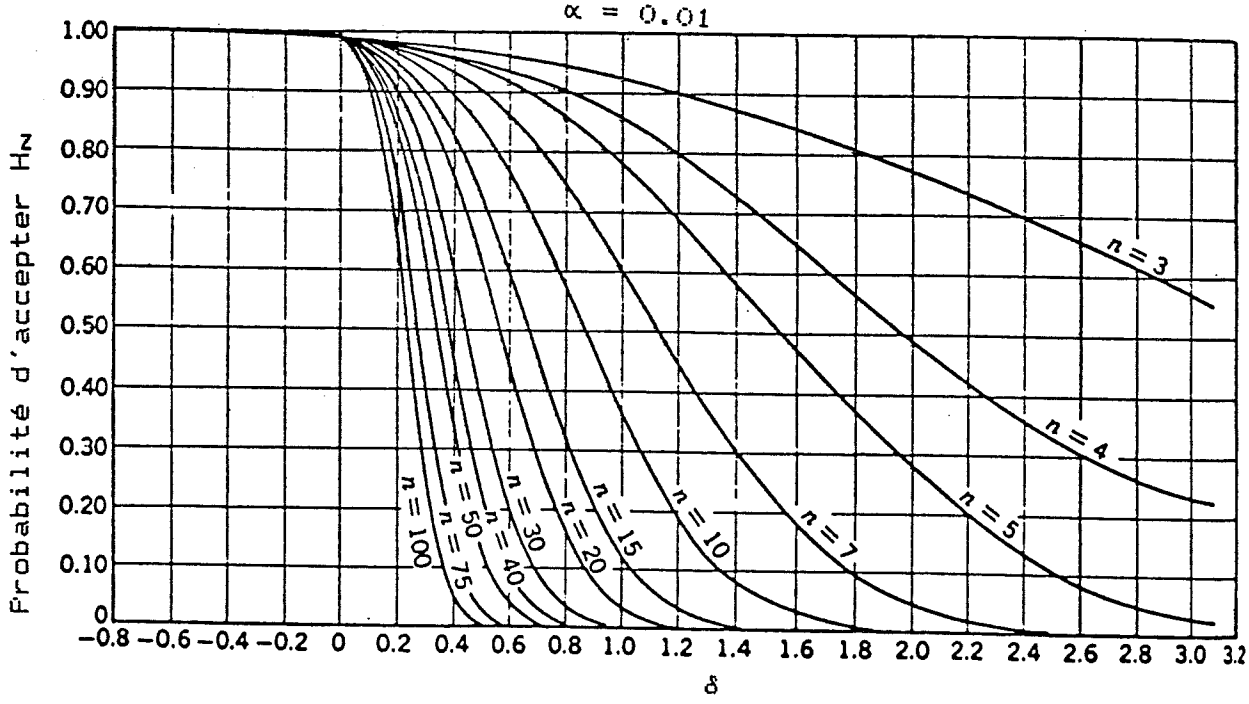
$$-0.8 \leq \delta \leq 3.2$$

Figure 8.3 Courbe caractéristique de H_N: $\mu = \mu_0$

Distribution normale, σ inconnu

Alternative unilatérale





Approximation de $\beta(\mu)$

La distribution de probabilité de W est malaisée mais on peut développer l'approximation suivante pour la fonction caractéristique.

D'après l'équation (8.30) on a

$$\beta(\mu) = P\left[\bar{X} - t_{n-1,\alpha}\, S/\sqrt{n} < \mu_0\right]$$

La variable $U = \bar{X} - t_{n-1,\alpha}\, S/\sqrt{n}$

est une fonction de $\bar{X}$ et S et on a vu au chapitre 6 que:

$$E(\bar{X}) = \mu \ , \quad VAR(\bar{X}) = \sigma^2/n$$

$$E(S) \approx \sigma \ , \quad VAR(S) \approx \sigma^2/2n$$

Alors U suit approximativement une distribution normale de paramètres

$$E(U) \approx \mu - t_{n-1,\alpha}\, \sigma/\sqrt{n}$$

$$VAR(U) \approx \frac{\sigma^2}{n} + \frac{t^2_{n-1,\alpha}}{n}\frac{\sigma^2}{2n} = \frac{\sigma^2}{n}\left(1 + \frac{t^2_{n-1,\alpha}}{2n}\right)$$

Donc

$$\beta(\mu) \approx P\left[\frac{U-E(U)}{\sqrt{VAR(U)}} < \frac{\mu_0 - E(U)}{\sqrt{VAR(U)}}\right]$$

$$\approx \Phi\left(\frac{\mu_0 - \mu + \dfrac{\sigma\, t_{n-1,\alpha}}{\sqrt{n}}}{\sigma/\sqrt{n}\,\sqrt{1 + \dfrac{t^2_{n-1,\alpha}}{2n}}}\right)$$

$$\approx \Phi\left(\frac{t_{n-1,\alpha} - \delta\sqrt{n}}{\sqrt{1 + \dfrac{t^2_{n-1,\alpha}}{2n}}}\right) \qquad (8.32)$$

où δ est définie par (8.31)

Cas B: alternative bilatérale

Le test de H_N: $\mu=\mu_0$ contre H_A: $\mu\neq\mu_0$ est défini par la région critique de seuil α:

$$\text{rejeter } H_N \text{ si } \frac{|\bar{X}-\mu_0|}{S/\sqrt{n}} > t_{n-1,\alpha/2} \qquad (8.33)$$

La fonction caractéristique $\beta(\mu)$ est:

$$\beta(\mu) = P\left[\frac{|\bar{X}-\mu_0|}{S/\sqrt{n}} < t_{n-1,\alpha/2}\right] \qquad (8.34)$$

$$= \beta(\alpha,\delta,n)$$

où $\delta = \mu-\mu_0/\sigma$

et elle est tracée à la figure 8.4 pour les valeurs suivantes:

$\alpha = 0.01,\ 0.05$

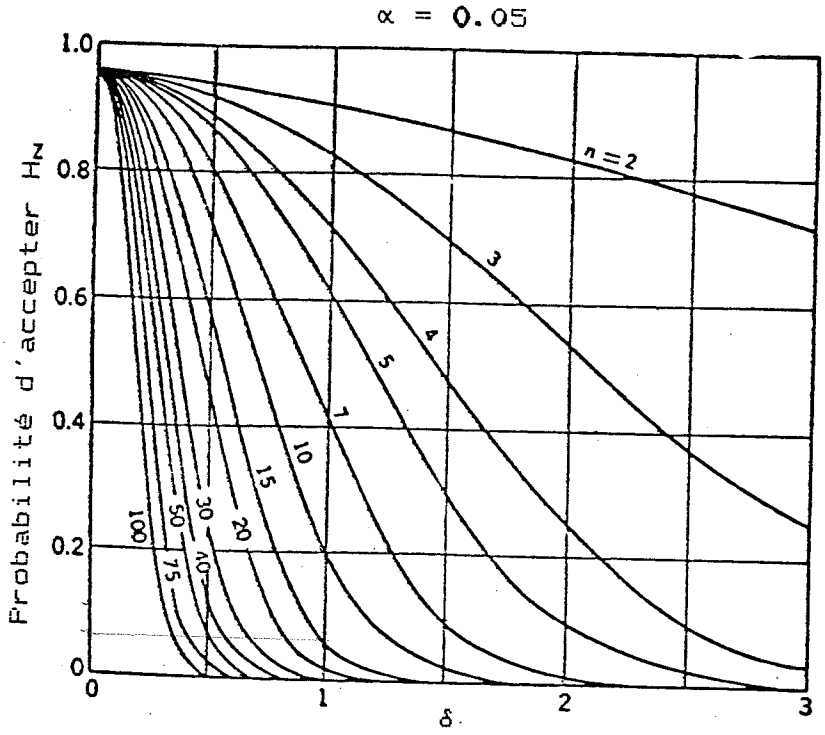
$n = 2,3,4,5,7,10,15,20,30,40,50,75,100$

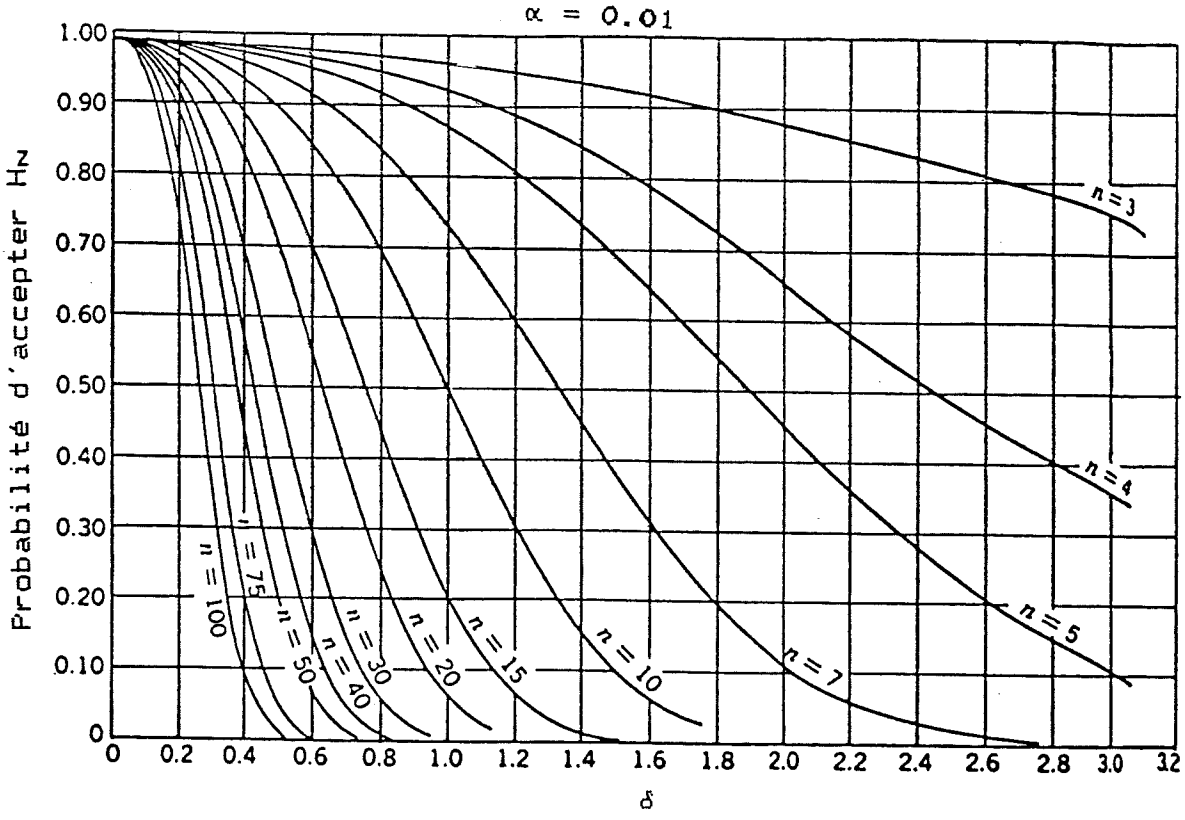
$0 \leq \delta \leq 3$

Figure 8.4 Courbe caractéristique de H_N: $\mu = \mu_0$

Distribution normale, σ inconnu

Alternative bilatérale





On peut évaluer $\beta(\mu)$ approximativement par:

$$\beta(\mu) \approx \Phi\left[\frac{t_{n-1,\alpha/2} + \delta\sqrt{n}}{\sqrt{1 + t^2_{n-1,\alpha/2}/2n}}\right] + \Phi\left[\frac{t_{n-1,\alpha/2} - \delta\sqrt{n}}{\sqrt{1 + t^2_{n-1,\alpha/2}/2n}}\right] - 1 \qquad (8.35)$$

La formule (8.35) s'obtient par une démarche analogue à celle utilisée pour obtenir la formule (8.34).

Calcul de n

Pour obtenir le nombre d'observations pour conduire un test sur une moyenne lorsque l'écart-type est inconnu il faut résoudre l'équation (8.34) ou (8.35) selon qu'il s'agit d'un test unilatéral ou bilatéral. On doit préciser la valeur du risque de deuxième espèce β à une valeur particulière μ de l'alternative. Le tableau 8.1 donne les valeurs de n

$$n = n(\delta, \alpha, \beta) \qquad (8.36)$$

pour les valeurs suivantes de δ, α et β

$\delta = \mu - \mu_0/\sigma = 0.15(0.05)\ 1.00(0.10)\ 2.5(0.5)\ 4$

$\beta = 0.01,\ 0.05,\ 0.1,\ 0.2,\ 0.5$

$\alpha = 0.005,\ 0.01,\ 0.025,\ 0.05$

Tableau 8.1

Nombre d'observations pour test

$H_N : \mu = \mu_0$

écart-type σ inconnu

unilatéral / bilatéral	$\alpha = 0.005$ / $\alpha = 0.01$					$\alpha = 0.01$ / $\alpha = 0.02$					$\alpha = 0.025$ / $\alpha = 0.05$					$\alpha = 0.05$ / $\alpha = 0.1$					
β	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	β
0.15																				122	0.15
0.20										139					99					70	0.20
0.25					110					90				128	64			139	101	45	0.25
0.30				134	78				115	63			119	90	45		122	97	71	32	0.30
0.35			125	99	58			109	85	47		109	88	67	34		90	72	52	24	0.35
0.40		115	97	77	45		101	85	66	37	117	84	68	51	26	101	70	55	40	19	0.40
0.45		92	77	62	37	110	81	68	53	30	93	67	54	41	21	80	55	44	33	15	0.45
0.50	100	75	63	51	30	90	66	55	43	25	76	54	44	34	18	65	45	36	27	13	0.50
0.55	83	63	53	42	26	75	55	46	36	21	63	45	37	28	15	54	38	30	22	11	0.55
0.60	71	53	45	36	22	63	47	39	31	18	53	38	32	24	13	46	32	26	19	9	0.60
0.65	61	46	39	31	20	55	41	34	27	16	46	33	27	21	12	39	28	22	17	8	0.65
0.70	53	40	34	28	17	47	35	30	24	14	40	29	24	19	10	34	24	19	15	8	0.70
0.75	47	36	30	25	16	42	31	27	21	13	35	26	21	16	9	30	21	17	13	7	0.75
0.80	41	32	27	22	14	37	28	24	19	12	31	22	19	15	9	27	19	15	12	6	0.80
0.85	37	29	24	20	13	33	25	21	17	11	28	21	17	13	8	24	17	14	11	6	0.85
0.90	34	26	22	18	12	29	23	19	16	10	25	19	16	12	7	21	15	13	10	5	0.90
0.95	31	24	20	17	11	27	21	18	14	9	23	17	14	11	7	19	14	11	9	5	0.95
1.00	28	22	19	16	10	25	19	16	13	9	21	16	13	10	6	18	13	11	8	5	1.00
1.1	24	19	16	14	9	21	16	14	12	8	18	13	11	9	6	15	11	9	7		1.1
1.2	21	16	14	12	8	18	14	12	10	7	15	12	10	8	5	13	10	8	6		1.2
1.3	18	15	13	11	8	16	13	11	9	6	14	10	9	7		11	8	7	6		1.3
1.4	16	13	12	10	7	14	11	10	9	6	12	9	8	7		10	8	7	5		1.4
1.5	15	12	11	9	7	13	10	9	8	6	11	8	7	6		9	7	6			1.5
1.6	13	11	10	8	6	12	10	9	7	5	10	8	7	6		8	6	6			1.6
1.7	12	10	9	8	6	11	9	8	7		9	7	6	5		8	6	5			1.7
1.8	12	10	9	8	6	10	8	7	7		8	7	6			7	6				1.8
1.9	11	9	8	7	6	10	8	7	6		8	6	6			7	5				1.9
2.0	10	8	8	7	5	9	7	7	6		7	6	5			6					2.0
2.1	10	8	7	7		8	7	6	6		7	6				6					2.1
2.2	9	8	7	6		8	7	6	5		7	6				6					2.2
2.3	9	7	7	6		8	6	6			6	5				5					2.3
2.4	8	7	7	6		7	6	6			6										2.4
2.5	8	7	6	6		7	6	6			6										2.5
3.0	7	6	6	5		6	5	5			5										3.0
3.5	6	5	5			5															3.5
4.0	6																				4.0

$\delta = \frac{\mu - \mu_0}{\sigma}$ — δ

Exemple 8.5: test d'une moyenne avec variance inconnue

Un échantillon aléatoire de taille n = 10 provenant d'une population normale a donné:

0.983, 1.005, 0.998, 0.986, 0.991
1.002, 0.996, 0.983, 0.994, 1.002

Testez $H_N: \mu = 1.000$

Contre $H_A: \mu \neq 1.000$

Solution

On calcule $\bar{X} = 0.994 \qquad s = 0.008055$

$$t = \frac{|0.994 - 1.000|}{0.00805/\sqrt{10}} = 2.355$$

Si on utilise un seuil $\alpha = 0.05$, la table de la distribution de Student avec n-1=9 degrés de liberté donne

$$t_{9,0.025} = 2.26$$

et on rejette H_N.

Combien d'observations devrait-on avoir si on veut faire un test avec α=0.05 et β=0.05 afin de détecter un écart δ=1?

D'après le tableau 8.1 il faut n=16 observations.

8.6 TEST SUR UNE VARIANCE

Soit σ^2=VAR(X) la variance d'une population X. Nous voulons tester l'hypothèse nulle:

$$H_N: \sigma^2 = \sigma_0^2$$

à l'aide d'un échantillon aléatoire $X_1, X_2, \ldots, X_n$. Pour construire le test il est nécessaire de faire certaines hypothèses de base sur la distribution de X et le type de contre-hypothèse envisagée. Nous supposons une population normale

$$X \sim N(\mu, \sigma^2)$$

avec moyenne μ inconnue. Notons par S^2 la variance échantillonnale

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

Cas A: $H_A: \sigma^2 < \sigma_0^2$

La région critique du test est:

$$\text{rejeter } H_N \text{ au seuil } \alpha \text{ si } \frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{n-1,1-\alpha} \qquad (8.37)$$

L'équation (8.32) résulte du fait que $(n-1)S^2/\sigma_0^2$ est distribuée selon une distribution khi-deux avec n-1 degrés de liberté.

Fonction caractéristique

$$\beta(\sigma^2) = P\left[\frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{n-1,1-\alpha}\right]$$

$$= P\left[\frac{(n-1)S^2}{\sigma^2} > \frac{\sigma_0^2}{\sigma^2}\chi^2_{n-1,1-\alpha}\right]$$

$$= P\left[\chi^2_{n-1} > \frac{1}{\lambda^2}\chi^2_{n-1,1-\alpha}\right] \qquad (8.38)$$

$$= \beta(\alpha,\lambda,n)$$

où $\lambda = \sigma/\sigma_0$ (8.39)

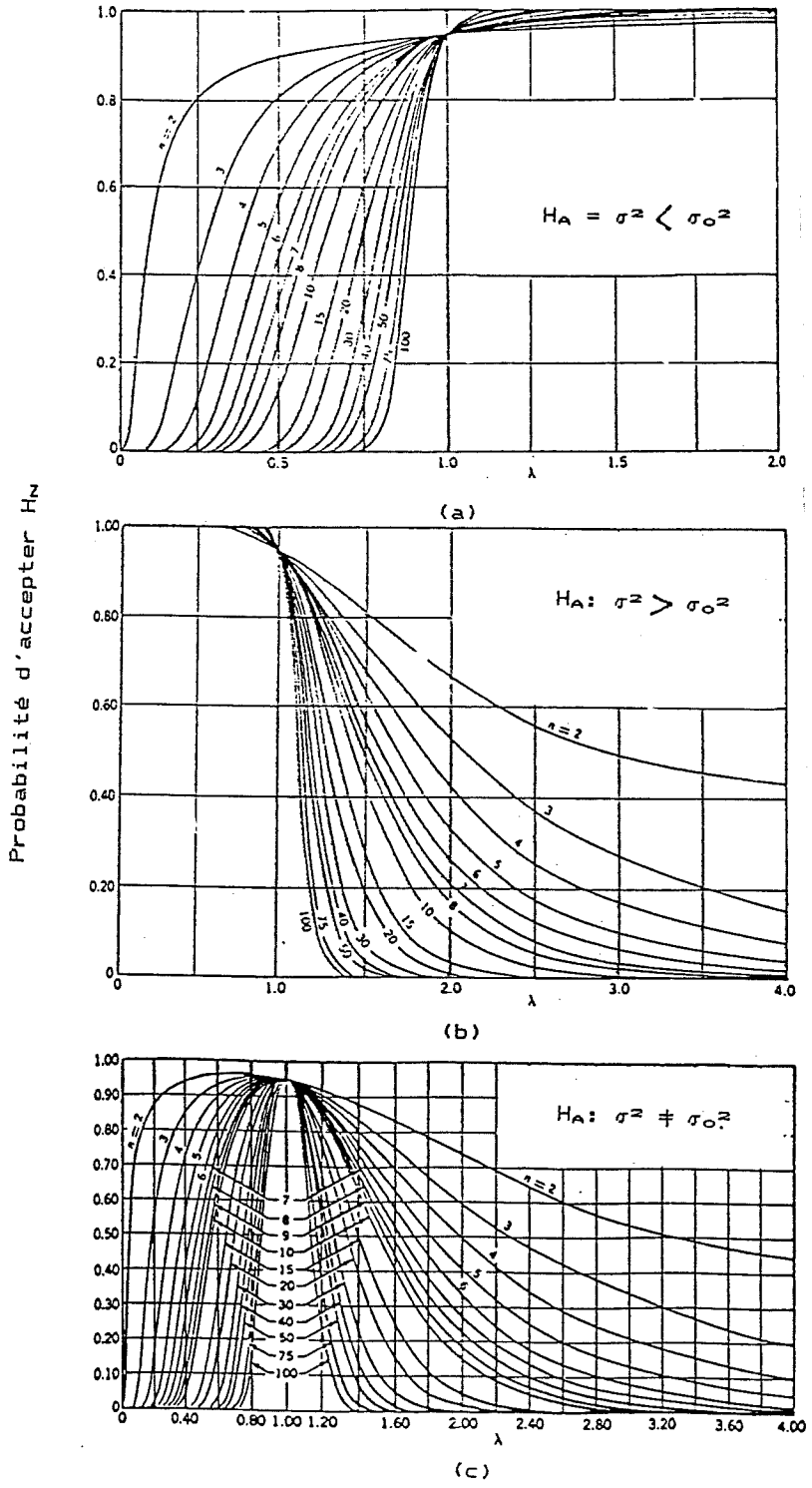
Le graphique de $\beta(\alpha=0.05,\lambda,n)$ est tracé à la figure 8.5 pour

$0 \leq \lambda \leq 4$

n=2(1)10,15,20,30,40,50,75,100

Figure 8.5 Courbe caractéristique du test de H_N: $\sigma^2=\sigma_0^2$

Distribution normale, $\alpha=0.05$, $\lambda=\sigma/\sigma_0$



Calcul de n

On doit résoudre l'équation (8.39) en fixant le risque de deuxième espèce à β pour une valeur de σ_1.

$$\beta(\sigma_1^2) = P\left[\chi^2_{n-1} > \frac{1}{\lambda^2}\chi^2_{n-1,1-\alpha}\right] = \beta$$

$$\lambda^2 = \frac{\chi^2_{n-1,1-\alpha}}{\chi^2_{n-1,\beta}} \qquad (8.40)$$

On peut résoudre (8.40) en consultant la table des percentiles d'une distribution khi-deux. Par exemple, si on fixe

$$\lambda=0.75, \quad \alpha=0.05, \quad \beta=0.10$$

la figure 8.5 donne n $\approx$ 50 et la table de la distribution khi-deux à ν=50 donne

$$\chi^2_{50,0.95} = 34.764$$

$$\chi^2_{50,0.10} = 63.167$$

$$\frac{\chi^2_{50,0.95}}{\chi^2_{50,0.10}} = \frac{34.764}{63.167} = 0.55 = (0.74)^2$$

La taille de l'échantillon est donc n=ν+1=51

On peut résoudre d'équation (8.40) pour n en remplaçant les percentiles de la distribution khi-deux par les approximations vues à l'équation (6.38)

$$\chi^2_{n-1,1-\alpha} \approx 1/2\left(-z_\alpha + \sqrt{2n-3}\right)^2$$

$$\chi^2_{n-1,\beta} \approx 1/2\left(z_\beta + \sqrt{2n-3}\right)^2$$

On obtient

$$n = \frac{3}{2} + \frac{1}{2}\left(\frac{z_\alpha + \lambda\, z_\beta}{\lambda-1}\right)^2 \qquad (8.41)$$

Par exemple si

$$\lambda=0.75, \quad \alpha=0.05, \quad \beta=0.10$$

$$n = \frac{3}{2} + \frac{1}{2}\left[\frac{1.645 + 0.75*1.28}{1-0.75}\right]^2$$

$$= 55.78$$

La taille de l'échantillon est 56.

Cas B: $H_A: \sigma^2 > \sigma_0^2$, $\frac{\sigma}{\sigma_0} = \lambda > 1$

Région critique de seuil α:

$$\frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{n-1,\alpha} \qquad (8.42)$$

Fonction caractéristique

$$\beta(\sigma^2) = P\left[\chi^2_{n-1} < \frac{1}{\lambda^2}\chi^2_{n-1,\alpha}\right] \qquad (8.43)$$

$$= \beta(\alpha,\lambda,n)$$

Le graphique de β est tracé à la figure 8.5.

Calcul de n

Un raisonnement analogue au cas A permet de développer la formule suivante identique à 8.41:

$$n = \frac{3}{2} + \frac{1}{2}\left[\frac{z_\alpha + \lambda\, z_\beta}{\lambda-1}\right]^2 \qquad (8.44)$$

Cas C: $H_A: \sigma^2 \neq \sigma_0^2$

Région critique de seuil α:

$$\frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{n-1,\alpha/2} \text{ ou } < \chi^2_{n-1,1-\alpha/2} \qquad (8.45)$$

Fonction caractéristique

$$\beta(\sigma^2) = P\left[\chi^2_{n-1} < (1/\lambda^2)\chi^2_{n-1,\alpha/2}\right] + P\left[\chi^2_{n-1} > \frac{1}{\lambda^2}\chi^2_{n-1,1-\alpha/2}\right] - 1 \qquad (8.46)$$

Le graphique de cette fonction est tracé à la figure 8.5

Calcul de n

On peut estimer n à l'aide de la courbe caractéristique ou employer la formule (8.41) en remplaçant α par $\alpha/2$.

Exemple 8.6 : test sur une variance

La variance d'un échantillon de 30 observations a donné

$$s^2 = 0.0349$$

Est-ce que cette variance dévie de $\sigma_0^2 = 0.02$?

Solution H_N: $\sigma^2 = \sigma_0^2 = 0.02$

H_A: $\sigma^2 > \sigma_0^2$

La statistique du test est

$$\chi^2_{29} = \frac{29*0.0349}{0.02} = 50.605$$

On rejette H_N au seuil α=0.05 puisque

$$\chi^2_{29,0.05} = 47.557$$

8.7 TEST SUR UNE PROPORTION

Supposons que l'on observe k articles défectueux dans un échantillon de taille n tiré d'un lot (population). Est-ce que la proportion d'articles défectueux θ dans le lot excède une valeur spécifiée, disons θ_0? On peut formuler la question à l'aide d'un test de l'hypothèse nulle:

$$H_N: \theta = \theta_0 \qquad (8.47)$$

contre $\qquad H_A: \theta > \theta_0$

On représente la population par une variable aléatoire X de type Bernoulli

X	0	1
$p_X(x)$	$1-\theta$	θ

où $\theta=P(X=1)$ est la probabilité qu'un article choisi au hasard soit défectueux et donc θ représente la proportion de défectueux dans le lot.

Soit Y la variable aléatoire représentant le nombre d'articles défectueux dans un échantillon $X_1, X_2, \ldots, X_n$ de taille n tiré de la population X. On a vu au chapitre 5 que la distribution de

$$Y = \sum_{i=1}^{n} X_i$$

est

- hypergéométrique $(N, D=\theta N, n)$ si les tirages sont effectués sans remise et N est la taille du lot
- binomiale (n,θ) si les tirages sont effectués avec remise ou si les tirages sont effectués sans remise et que $n/N \leq 0.1$

Test exact de H_N

$$\text{rejeter } H_N \text{ si } Y \geq c + 1 \qquad (8.48)$$

La constante c est déterminée selon le seuil α choisi et conduit à l'équation suivante en employant la distribution binomiale

$$P\ (Y>c) = \sum_{x=c+1}^{n} \binom{n}{x} \theta_0^x\ (1-\theta_0)^{n-x} = \alpha \qquad (8.49)$$

Il n'est pas sûr que l'on puisse déterminer une valeur c afin de satisfaire l'équation (8.49) exactement à cause du caractère discontinu de la fonction. Une solution analytique est malaisée mais on peut résoudre par une méthode itérative.

D'autre part si on veut concevoir un plan d'échantillonnage (n,c), il faut résoudre simultanément l'équation (8.49) et l'équation (8.50)

$$\beta(\theta_1)=P\ [\ Y \leq c\ ] = \sum_{x=0}^{c} \binom{n}{x} \theta_1^x\ (1-\theta_1)^{n-x} = \beta \qquad (8.50)$$

Nous avons déjà présenté au chapitre 5 un programme SAS faisant la solution des équations (8.49) et (8.50) par une méthode itérative. Nous allons présenter maintenant une solution analytique basée sur l'approximation de la distribution binomiale par une distribution normale.

Test de H_N à l'aide de la distribution normale

En employant une approximation normale, les équations (8.49) et (8.50) deviennent

$$P\ [\ Y \geq c+1\ ] \approx 1 - \Phi\left[\frac{c + 0.5 - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}\right] = \alpha \qquad (8.51)$$

$$P\ [\ Y \leq c\ ] \approx \Phi\left[\frac{c + 0.5 - n\theta_1}{\sqrt{n\theta_1(1-\theta_1)}}\right] = \beta \qquad (8.52)$$

La constante c est

$$c = n\theta_0 - 0.5 + z_\alpha\sqrt{n\theta_0(1-\theta_0)} \qquad (8.53)$$

$$c = n\theta_1 - 0.5 - z_\beta\sqrt{n\theta_1(1-\theta_1)} \qquad (8.54)$$

En isolant n des équations (8.53) et (8.54) on obtient:

$$n = \frac{1}{(\theta_1-\theta_0)^2}\left[z_\alpha\sqrt{\theta_0(1-\theta_0)} + z_\beta\sqrt{\theta_1(1-\theta_1)}\ \right]^2 \qquad (8.55)$$

La fonction caractéristique du test est:

$$\beta(\theta) = P\ [\ Y \leq c\ ]$$

$$= \Phi\left[\frac{c+0.5-n\theta}{\sqrt{n\theta(1-\theta)}}\right]$$

En remplaçant c dans cette dernière équation par (8.55) on a:

$$\beta(\theta) = \Phi\left[\left(z_\alpha - \sqrt{n}\,\frac{(\theta-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}}\right)\sqrt{\frac{\theta_0(1-\theta_0)}{\theta(1-\theta)}}\right] \qquad (8.56)$$

Par exemple si

$$\alpha=0.01 \quad \theta_0=0.05 \quad \beta=0.05 \quad \theta_1=0.10$$

on obtient

$$n=401.2 \simeq 402 \quad \text{et} \quad c=29.78 \approx 30$$

alors que le programme SAS du chapitre 5 et la distribution binomiale donnent

$$n=425 \quad \text{et} \quad c=32$$

avec

$$P\ [\ Y \geq 32;\ n=425,\ \theta_0=0.05\ ] = 0.00909 \approx \alpha$$
$$p\ [\ Y \leq 31;\ n=425,\ \theta_1=0.10\ ] = 0.04875 \approx \beta$$

<u>Cas d'une alternative bilatérale</u>

$$H_A:\ \theta \neq \theta_0 \qquad (8.57)$$

La région critique est de la forme:

$$\text{rejeter } H_N \text{ si} \quad Y \leq c_1-1 \quad \text{ou } Y \geq c_2+1$$

Une approximation basée sur la distribution normale donne

$$c_1 = n\theta_0+0.5 - z_{\alpha/2}\sqrt{n\theta_0(1-\theta_0)}$$
$$c_2 = n\theta_0-0.5 + z_{\alpha/2}\sqrt{n\theta_0(1-\theta_0)} \qquad (8.58)$$

où α est le seuil du test.

Le calcul de la taille échantillonnale n s'obtient de l'équation (8.55) avec α remplacé par $\alpha/2$.

8.8 TEST D'ÉGALITÉ DE DEUX MOYENNES

Soient $X_1, X_2, \ldots, X_n$, un échantillon aléatoire provenant d'une population $N(\mu_x, \sigma_x^2)$ et $Y_1, Y_2, \ldots, Y_n$ un échantillon aléatoire provenant d'une autre population $N(\mu_y, \sigma_y^2)$. On veut tester l'hypothèse nulle

$$H_N: \mu_x - \mu_y = 0$$

La procédure de test statistique dépend des hypothèses de base concernant la connaissance où non des variances σ_x^2, σ_y^2, de l'indépendance où non des deux échantillons ainsi que de l'alternative H_A envisagée. Nous distinguerons quatre cas.

Cas A: σ_x^2, σ_y^2 connues

Traitons le cas de l'alternative unilatérale

$$H_A: \mu_x - \mu_y < 0 \tag{8.59}$$

Les calculs de la région critique, de la taille échantillonnale et la fonction caractéristique sont basés sur le fait que

$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2/n_1 + \sigma_y^2/n_2}} \sim N(0,1) \tag{8.60}$$

selon l'équation (6.29).

Posons $$Z = (\bar{X} - \bar{Y}) / \sqrt{\sigma_X^2/n_1 + \sigma_y^2/n_2} \tag{8.61}$$

La région critique de seuil α est:

$$\text{rejeter } H_N \text{ si } Z < -z_\alpha \tag{8.62}$$

Remarque

On peut tester l'hypothèse nulle

$$H_N: \mu_x - \mu_y = a \tag{8.63}$$

où a est une valeur spécifiée contre l'alternative

$$H_A: \mu_x - \mu_y < a \tag{8.64}$$

en employant le rapport critique

$$Z_a = \frac{(\bar{X} - \bar{Y}) - a}{\sqrt{\sigma_x^2/n_1 + \sigma_y^2/n_2}} \qquad (8.65)$$

dans l'équation (8.62)

Fonction caractéristique

Posons

$$\Delta = \mu_x - \mu_y \qquad (8.66)$$

$$\delta = \frac{\Delta}{\sqrt{\sigma_x^2/n_1 + \sigma_y^2/n_2}} \qquad (8.67)$$

La fonction caractéristique est par définition la probabilité d'accepter H_N en fonction de Δ et des paramètres, $n_1, n_2, \sigma_x, \sigma_y$

$$\begin{aligned} P [Z > -z_\alpha] &= 1 - P [Z < -z_\alpha] \\ &= 1 - P [Z - \delta < -z_\alpha - \delta] \\ &= 1 - \Phi (-z_\alpha - \delta) \\ &= \Phi(z_\alpha + \delta) \\ &= \beta(\alpha, \delta, n_1, n_2, \sigma_x, \sigma_y) \end{aligned} \qquad (8.68)$$

Calcul de la taille échantillonnale

Posons $n_1 = n_2 = n$ et fixons α et δ dans l'équation (8.68) de telle sorte que

$$\beta(\alpha, \delta, n, \sigma_x, \sigma_y) = \beta$$

Alors

$$z_\alpha + \frac{\Delta\sqrt{n}}{\sqrt{\sigma_x^2 + \sigma_y^2}} = \Phi^{-1}(\beta) = z_{1-\beta} = -z_\beta$$

et en isolant n, on obtient:

$$n = (z_\alpha + z_\beta)^2 \left[\frac{\sigma_x^2 + \sigma_y^2}{\Delta^2}\right] \qquad (8.69)$$

Le cas $n_1=n$ et $n_2=\lambda^2 n$ où λ est spécifié conduit à l'équation (8.69) avec σ_y remplacé par σ_y/λ

Cas d'alternative bilatérale

$$H_A: \mu_x - \mu_y \neq 0$$

Dans ce cas la région critique de seuil α est:

$$\text{rejeter } H_N \text{ si } |Z| > z_{\alpha/2} \qquad (8.70)$$

où Z est définie par (8.61)

La fonction caractéristique du test est:

$$\beta(\alpha, \Delta, n_1, n_2, \sigma_x, \sigma_y)$$

$$= \Phi(z_{\alpha/2} + \delta) + \Phi(z_{\alpha/2} - \delta) - 1 \qquad (8.71)$$

où δ est définie par (8.67)

Le calcul de la taille échantillonnale se fait avec l'équation (8.69) avec α remplacé par $\alpha/2$

Cas B: $\sigma_x^2 = \sigma_y^2 = \sigma^2$, σ^2 inconnue

Le développement du test est basé sur le fait que

$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim T_{n1+n2-2} \qquad (8.72)$$

où $$S_p^2 = \frac{(n_1-1)\, S_x^2 + (n_2-1) S_y^2}{n_1+n_2-2}$$

$$S_x^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2 \quad , \quad S_y^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

Posons
$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{1/n_1 + 1/n_2}} \tag{8.73}$$

Cas d'alternative unilatérale

$$H_A : \mu_x - \mu_y = \Delta < 0$$

Posons
$$\nu = n_1 + n_2 - 2$$

Test: on rejette H_N au seuil α si $T < - t_{\nu,\alpha}$ (8.74)

Fonction caractéristique

$$\beta(\Delta) = P\left[T > - t_{\nu,\alpha} \; ; \; \mu_x - \mu_y = \Delta \right]$$

$$= P\left[(\bar{X}-\bar{Y}) > -t_{\nu,\alpha} S_p\sqrt{1/n_1 + 1/n_2} \; ; \; \mu_x-\mu_y=\Delta \right]$$

$$= P\left[(\bar{X}-\bar{Y}) + t_{\nu,\alpha} \; S_p\sqrt{1/n_1 + 1/n_2} > 0 \; ; \; \mu_x-\mu_y=\Delta \right] \tag{8.75}$$

Soit
$$U = \bar{X}-\bar{Y} + t_{\nu,\alpha} \; S_p\sqrt{1/n_1 + 1/n_2}$$

Alors $E(U) = \Delta + t_{\nu,\alpha} \; S_p\sqrt{1/n_1 + 1/n_2}$

$$VAR(U) = \sigma^2(1/n_1 + 1/n_2) \; (1 + t^2_{\nu,\alpha}/2\nu)$$

$$\beta(\Delta) = P\,[\,U > 0\,] = 1-P\,[\,U < 0\,]$$

$$\approx 1- \Phi\left[\frac{-E(U)}{\sqrt{VAR(U)}}\right]$$

$$\approx \Phi\left[\frac{E(U)}{\sqrt{VAR(U)}}\right]$$

$$\approx \Phi\left[\frac{t_{\nu,\alpha} + \lambda}{\sqrt{1+t^2_{\nu,\alpha}/2\nu}}\right] \qquad (8.76)$$

où

$$\lambda = \frac{\Delta}{\sigma\sqrt{1/n_1 + 1/n_2}} \qquad (8.77)$$

Calcul de n

Posons $n_1 = n_2 = n$ et $\beta(\Delta,\alpha,n) = \beta$

Alors $$n \approx 2(z_\beta + z_\alpha)^2\ (\sigma/\Delta)^2 \qquad (8.78)$$

Le tableau 8.2 a été calculé à l'aide de l'équation (8.78) en employant l'approximation

$$\frac{t_{\nu,\alpha} + \lambda}{\sqrt{1+t^2_{\nu,\alpha}/2\nu}} \approx z_\alpha + \lambda$$

Tableau 8.2 Nombre d'observations (n) pour test

$H_N : \mu_x - \mu_y = 0$

deux échantillons de taille n

même écart-type σ inconnu

unilatéral / bilatéral	$\alpha = 0.005$ / $\alpha = 0.01$					$\alpha = 0.01$ / $\alpha = 0.02$					$\alpha = 0.025$ / $\alpha = 0.05$					$\alpha = 0.05$ / $\alpha = 0.1$					
β	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	ρ
0.05																					0.05
0.10																					0.10
0.15																					0.15
0.20																				137	0.20
0.25															124					88	0.25
0.30										123					87					61	0.30
0.35					110					90					64				102	45	0.35
0.40					85					70				100	50			108	78	35	0.40
0.45				118	68				101	55			105	79	39		108	86	62	28	0.45
0.50				96	55			106	82	45		106	86	64	32		88	70	51	23	0.50
0.55			101	79	46		106	88	68	38		87	71	53	27	112	73	58	42	19	0.55
0.60		101	85	67	39		90	74	58	32	104	74	60	45	23	89	61	49	36	16	0.60
0.65		87	73	57	34	104	77	64	49	27	88	63	51	39	20	76	52	42	30	14	0.65
0.70	100	75	63	50	29	90	66	55	43	24	76	55	44	34	17	66	45	36	26	12	0.70
0.75	88	66	55	44	26	79	58	48	38	21	67	48	39	29	15	57	40	32	23	11	0.75
0.80	77	58	49	39	23	70	51	43	33	19	59	42	34	26	14	50	35	28	21	10	0.80
0.85	69	51	43	35	21	62	46	38	30	17	52	37	31	23	12	45	31	25	18	9	0.85
0.90	62	46	39	31	19	55	41	34	27	15	47	34	27	21	11	40	28	22	16	8	0.90
$\frac{\Delta}{\sigma}$ 0.95	55	42	35	28	17	50	37	31	24	14	42	30	25	19	10	36	25	20	15	7	0.95
1.00	50	38	32	26	15	45	33	28	22	13	38	27	23	17	9	33	23	18	14	7	1.00
1.1	42	32	27	22	13	38	28	23	19	11	32	23	19	14	8	27	19	15	12	6	1.1
1.2	36	27	23	18	11	32	24	20	16	9	27	20	16	12	7	23	16	13	10	5	1.2
1.3	31	23	20	16	10	28	21	17	14	8	23	17	14	11	6	20	14	11	9	5	1.3
1.4	27	20	17	14	9	24	18	15	12	8	20	15	12	10	6	17	12	10	8	4	1.4
1.5	24	18	15	13	8	21	16	14	11	7	18	13	11	9	5	15	11	9	7	4	1.5
1.6	21	16	14	11	7	19	14	12	10	6	16	12	10	8	5	14	10	8	6	4	1.6
1.7	19	15	13	10	7	17	13	11	9	6	14	11	9	7	4	12	9	7	6	3	1.7
1.8	17	13	11	10	6	15	12	10	8	5	13	10	8	6	4	11	8	7	5		1.8
1.9	16	12	11	9	6	14	11	9	8	5	12	9	7	6	4	10	7	6	5		1.9
2.0	14	11	10	8	6	13	10	9	7	5	11	8	7	6	4	9	7	6	4		2.0
2.1	13	10	9	8	5	12	9	8	7	5	10	8	6	5	3	8	6	5	4		2.1
2.2	12	10	8	7	5	11	9	7	6	4	9	7	6	5		8	6	5	4		2.2
2.3	11	9	8	7	5	10	8	7	6	4	9	7	6	5		7	5	5	4		2.3
2.4	11	9	8	6	5	10	8	7	6	4	8	6	5	4		7	5	4	4		2.4
2.5	10	8	7	6	4	9	7	6	5	4	8	6	5	4		6	5	4	3		2.5
3.0	8	6	6	5	4	7	6	5	4	3	6	5	4	4		5	4	3			3.0
3.5	6	5	5	4	3	6	5	4	4		5	4	4	3		4	3				3.5
4.0	6	5	4	4		5	4	4	3		4	4	3			4					4.0

$\frac{\Delta}{\sigma}$

Cas d'alternative bilatérale

$$H_A: \mu_x - \mu_y = \Delta \neq 0$$

on rejette H_N au seuil α si $| T | > t_{\nu,\alpha/2}$ (8.79)

La fonction caractéristique est:

$$\beta(\Delta) = P [| T | < t_{\nu,\alpha/2} ; \mu_x - \mu_y = \Delta]$$

$$\approx \Phi \left[\frac{t_{\nu,\alpha/2} + \lambda}{\sqrt{1+t^2_{\nu,\alpha/2\nu}}} \right] + \Phi \left[\frac{t_{\nu,\alpha/2} - \lambda}{\sqrt{1+t^2_{\nu,\alpha/2\nu}}} \right] - 1 \qquad (8.80)$$

le calcul de n s'obtient de l'équation (8.78) avec α remplacé par $\alpha/2$

Exemple 8.7: données sur la tension de rupture de deux types de fils du chapitre 1

Les calculs donnent: $H_A: \mu_N - \mu_A > 0$

type	n	moyenne ($\bar{x}$)	écart-type (s)
ancien	25	151.08	2.465
nouveau	25	152.76	1.877

Si on admet l'égalité des variances $\sigma_1^2 = \sigma_2^2$ qui sera confirmée par un test à la section suivante, on calcule

$$S_p^2 = \frac{24 * (2.465)^2 + 24 * (1.877)^2}{48}$$

$$= 4.799 = (2.191)^2$$

$$T = \frac{152.76 - 151.08}{2.191 \sqrt{\frac{1}{25} + \frac{1}{25}}} = 2.711$$

Au seuil $\alpha = 0.01$ on rejette H_N puisque $t_{48,0.005} = 2.68$

La tension moyenne de rupture du fil nouveau est donc supérieure à celle du fil ancien.

Cas C: σ_1^2, σ_2^2 inconnus

Le développement du test est basé sur le résultat suivant:

$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{8.81}$$

suit approximativement une distribution de Student avec ν degrés de liberté

où

$$\nu = \frac{(a + b)^2}{a^2/(n_1-1) + b^2/(n_2-1)} \tag{8.82}$$

$$a = S_1^2/n_1 \quad , \quad b = S_2^2/n_2$$

On montre que

$$\text{Min}(n_1-1, n_2-1) \leq \nu \leq n_1 + n_2 - 2$$

La région critique du test est

Alternative	Région critique de seuil α
$H_A: \mu_x - \mu_y < 0$	$T < - t_{\nu,\alpha}$
$H_A: \mu_x - \mu_y > 0$	$T > t_{\nu,\alpha}$
$H_A: \mu_x - \mu_y \neq 0$	$\lvert T \rvert > t_{\nu,\alpha/2}$

où

$$T = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{8.83}$$

Exemple 8.8 : comparaison de deux types de béton

	béton I (kg/cm^2)	béton II (kg/cm^2)
	284	318
	311	318
	290	312
	280	-
n	4	3
$\bar{x}$	291.25	316.00
s	13.79	3.46
s^2	190.25	12.00

L'hypothèse d'égalité des variances semble à première vue très douteuse et nous utiliserons la procédure avec variances inégales. On calcule

$$a = s_1^2/n_1 = 47.56$$

$$b = s_2^2/n_2 = 4$$

$$\nu = \frac{(a + b)^2}{a^2/(n_1-1) + b^2/(n_2-1)} = \frac{(51.56)^2}{754.06 + 8} = 3.49$$

$$T = \frac{316 - 291.25}{\sqrt{47.56 + 4}} = 3.447$$

Au seuil $\alpha = 0.05$ on rejette l'hypothèse d'égalité des tension de rupture moyenne des bétons puisque

$$t_{3,0.05} = 2.353$$

$$t_{4,0.05} = 2.132$$

Cas D: échantillons pairés

Il arrive, dans certaines circonstances expérimentales, que l'on ne puisse pas faire, comme dans les cas A,B,C étudiés précédemment, l'hypothèse de base que les deux échantillons sont indépendants. Cette situation se présente lorsque les deux échantillons sont définis sur des mêmes unités expérimentales. On identifie cette situation sous le nom de MESURES RÉPÉTÉES. À titre d'exemple supposons que l'on veuille étudier si l'effet d'un traitement médical permet d'abaisser la pression artérielle (disons). On obtient alors deux séries de mesures chez n individus:

$x_1, x_2 .., x_n$: pression artérielle avant traitement

$Y_1, Y_2, .., Y_n$: pression artérielle après traitement

Il y a une liaison naturelle pour le couple (x_i, y_i) i=1,..n et l'hypothèse de base d'indépendance entre les variables X et Y n'est pas réaliste.

On veut tester l'hypothèse nulle

$$H_N: \mu_x - \mu_y = 0$$

afin de détecter si le traitement a un effet significatif (rejet de H_N) ou non (non rejet de H_N).

Pour construire un test, on considère la variable

$$D_i = X_i - Y_i \qquad i=1,\ldots,n \tag{8.84}$$

et sous l'hypothèse de normalité des variables X_i et Y_i on a

$$\bar{D} = \bar{X} - \bar{Y} \sim N(\mu_x - \mu_y, \sigma^2/n)$$

où $$\sigma^2 = VAR(X_i - Y_i) = \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y \tag{8.85}$$

ρ est le coefficient de corrélation entre X et Y

Le problème est alors ramené à un test de la nullité d'une moyenne avec variance σ^2 inconnue. Cette situation a été étudiée à la section précédente et on peut résumer le test dans le tableau suivant:

Alternative	Région critique de seuil α
$H_A: \mu_x - \mu_y < 0$	$T < -t_{n-1,\alpha}$
$H_A: \mu_x - \mu_y > 0$	$T > t_{n-1,\alpha}$
$H_A: \mu_y - \mu_x \neq 0$	$\lvert T \rvert > t_{n-1,\alpha/2}$

où $$T = \frac{\bar{D}}{s_D/\sqrt{n}} \qquad \bar{D} = \bar{X} - \bar{Y} \tag{8.86}$$

$$s_D^2 = \frac{1}{n-1} \sum_{i=1}^{n} (D_i - \bar{D})^2$$

Exemple 8.9: échantillons pairés

Les données représentent le nombre moyen de journées de travail perdues à cause d'accidents de travail dans 10 usines avant et après un programme de sécurité au travail. Le programme a-t-il été efficace?

Usine	Avant	Après	Différence
1	4.5	3.6	0.9
2	7.3	6.0	1.3
3	4.6	4.4	0.2
4	12.4	11.9	0.5
5	3.3	3.5	-0.2
6	5.7	5.1	0.6
7	8.3	7.7	0.6
8	3.4	2.9	0.5
9	2.6	2.4	0.2
10	1.7	1.1	0.6

Solution

$$\bar{D} = 0.52 \qquad s_D = 0.408$$

$$T = \frac{0.52}{0.408/\sqrt{10}} = 4.03$$

et $$t_{9,0.01} = 2.821$$

On rejette H_N et on peut conclure que le programme a été efficace.

8.9 TEST D'ÉGALITÉ DE DEUX VARIANCES

On dispose de deux échantillons indépendants provenant de deux populations normales avec moyennes inconnues

$$X_i \sim N(\mu_x, \sigma_1^2) \qquad i=1,2,..,n_1$$

$$Y_i \sim N(\mu_y, \sigma_2^2) \qquad i=1,2,..,n_2$$

et on veut tester l'hypothèse nulle

$$H_N: \sigma_1^2 = \sigma_2^2 \tag{8.87}$$

Posons

$$S_1^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$$

$$S_2^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

On sait que $(n_i-1)S_i^2/\sigma_i^2$ suit une distribution $\chi^2_{n_i-1}$ $(i=1,2)$ et puisque les deux échantillons sont indépendants le rapport F

$$F = \frac{S_1^2}{S_2^2} \tag{8.88}$$

suit une distribution de Fisher avec n_1-1 et n_2-1 degrés de liberté si H_N est vraie. On peut formuler le test suivant

Alternative	Région critique de seuil α
$H_A: \sigma_1^2 < \sigma_2^2$	$F < F_{n1-1,n2-1,\,1-\alpha}$
$H_A: \sigma_1^2 > \sigma_2^2$	$F > F_{n1-1,n2-1,\alpha}$
$H_A: \sigma_1^2 \neq \sigma_2^2$	$F < F_{n1-1,n2-1,1-\alpha/2}$ ou $F > F_{n1-1,n2-1,\alpha/2}$

Remarque pour le test bilatéral

Étant donné la symétrie du problème on peut utiliser le rapport critique F

$$F' = \frac{\text{plus grande variance}}{\text{plus petite variance}} \qquad (8.89)$$

qui est distribuée selon une distribution de Fisher (ν_1, ν_2)

où

$$\nu_1 = \begin{cases} n_1 - 1 \text{ si} & s_1^2 > s_2^2 \\ n_2 - 1 \text{ si} & s_2^2 > s_1^2 \end{cases}$$

$$\nu_2 = \begin{cases} n_2 - 1 \text{ si} & s_1^2 > s_2^2 \\ n_1 - 1 \text{ si} & s_2^2 > s_1^2 \end{cases}$$

Le test devient

$$\text{rejeter } H_N \text{ si } F' > F_{\nu 1, \nu 2, \alpha/2}$$

Exemple 8.10: exemple des fils du chapitre 1

Lors de l'exemple 8.7 nous avons fait l'hypothèse de base que $\sigma_1^2 = \sigma_2^2 = \sigma^2$ lors de l'exécution du test d'égalité des moyennes. Rappelons les calculs

type	n	moyenne $(\bar{x})$	écart-type (s)
ancien	25	151.08	2.465
nouveau	25	152.76	1.877

Testons l'hypothèse

$$H_N: \quad \sigma_1^2 = \sigma_2^2$$

On a

$$F' = \frac{(2.465)^2}{(1.877)^2} = 1.725$$

Au seuil $\alpha = 0.05$, $F_{24,24,0.025} = 2.31$ (par interpolation) et donc on ne rejette pas H_N

8.10 TEST D'ÉGALITÉ DE PLUSIEURS PROPORTIONS

On dispose d'échantillons indépendants provenant de k populations de type Bernoulli (θ_j) et on veut tester l'hypothèse nulle

$$H_N: \theta_1 = \theta_2 = \ldots = \theta_k = \theta \qquad (8.90)$$

contre l'alternative H_A: $\theta_j \neq \theta_{j'}$ pour au moins un couple $j \neq j'$

On peut représenter les données dans un tableau

Population Bernoulli	θ_1	θ_2	...	θ_k	Total
nombre succès	x_1	x_2	...	x_k	$\sum_{j=1}^{k} x_j = x$
nombre d'échecs	n_1-x_1	n_2-x_2	...	n_k-x_k	$n-x$
taille de l'échantillon	n_1	n_2	...	n_k	$\sum_{j=1}^{k} n_j = n$

Chaque variable X_j suit une distribution binomiale de paramètres n_j et θ_j

$$X_j \sim \text{bin}(n_j, \theta_j) \qquad j=1,2,\ldots,k$$

Le test de H_N sera développé en utilisant une approximation normale pour chacune des distributions binomiales:

$$X_j \sim N(n_j\theta_j,\ n_j\theta_j(1-\theta_j)) \qquad j=1,2,\ldots,k$$

Estimation de θ sous H_N

La méthode de vraisemblance maximale conduit à l'estimation de θ par:

$$\hat{\theta} = \frac{\sum_{j=1}^{k} x_j}{\sum_{j=1}^{k} n_j} = \frac{x}{n} \qquad (8.91)$$

Justification du test

Si l'hypothèse H_N est vraie

$$X_j \sim N(n_j\theta,\ n_j\theta(1-\theta))$$

et, en remplaçant θ par $\hat{\theta}$,

$\dfrac{(X_j - n_j\hat{\theta})^2}{n_j\hat{\theta}(1-\hat{\theta})}$ est une variable khi-deux avec 1 degré de liberté approximativement

La statistique

$$D^2 = \sum_{j=1}^{k} \frac{(x_j - n_j\hat{\theta})^2}{n_j\hat{\theta}(1-\hat{\theta})} \qquad (8.92)$$

est une mesure de l'écart entre la distribution observée $(x_1, x_2, \ldots, x_k)$ et la distribution théorique estimée $(n_1\hat{\theta},\ n_2\hat{\theta},\ \ldots,\ n_k\hat{\theta})$. La construction du test est basé sur la distribution d'échantillonnage de D^2.

Distribution d'échantillonnage de D^2

D^2 suit approximativement une distribution χ^2_{k-1}

à condition que: $n_j\hat{\theta} \geq 5$

Test

On rejette H_N au seuil α si

$$D^2 > \chi^2_{k-1,\alpha} \qquad (8.93)$$

Exemple 8.11: test d'égalité de trois proportions

Trois matériaux A,B,C ont été soumis à des essais de tractions et on note si le matériau est demeuré intact ou s'est déformé

Matériau	A	B	C	Total
déformé	41	27	22	90
intact	79	53	78	210
nombre	120	80	100	300

On calcule

$$\hat{\theta} = \frac{90}{300} = 0.30$$

$$D^2 = \frac{(41-36)^2}{25.2} + \frac{(27-24)^2}{16.8} + \frac{(22-30)^2}{21} = 4.575$$

On ne rejette pas l'hypothèse d'égalité au seuil de 0.05 puisque la valeur critique est 5.99

$$\chi^2_{2,0.05} = 5.99$$

Remarques

(a) Le test est aussi connu sous le nom de test d'homogénéité de k populations binomiales.

(b) Si k = 2 le test peut se faire par l'intermédiaire de la statistique

$$Z = \frac{X_1/n_1 - X_2/n_2}{\sqrt{\hat{\theta}(1-\hat{\theta})\,[1/n_1+1/n_2]}} \qquad (8.94)$$

qui est approximativement distribuée selon une distribution gaussienne centrée-réduite.

Puisque

$$D^2 = \frac{(x_1 - n_1\hat{\theta})^2}{n_1\hat{\theta}(1-\hat{\theta})} + \frac{(x_2 - n_2\hat{\theta})^2}{n_2\hat{\theta}(1-\hat{\theta})} = Z^2$$

on peut reformuler la procédure du test avec Z et tenir compte de diverses formes d'alternatives:

Alternative	Région critique de seuil α
$H_A: \theta_1 < \theta_2$	$Z < -z_\alpha$
$H_A: \theta_1 > \theta_2$	$Z > z_\alpha$
$H_A: \theta_1 \neq \theta_2$	$\lvert Z \rvert > z_{\alpha/2}$

où Z est définie par l'équation (8.94)

(c) Posons pour $j = 1,2,\ldots,k$.

$$n_{ij} = \begin{cases} x_j & \text{si } i=1 \\ n_j - x_j & \text{si } i=2 \end{cases}$$

$$e_{ij} = \begin{cases} n_j\hat{\theta} & \text{si } i=1 \\ n_j(1-\hat{\theta}) & \text{si } i=2 \end{cases}$$

Alors

$$\frac{(n_{1j}-e_{1j})^2}{e_{1j}} + \frac{(n_{2j}-e_{2j})^2}{e_{2j}} = \frac{(x_j - n_j\hat{\theta})^2}{n_j\hat{\theta}(1-\hat{\theta})}$$

et

$$D^2 = \sum_{j=1}^{k} \frac{(x_j - n_j\hat{\theta})^2}{n_j\hat{\theta}(1-\hat{\theta})} = \sum_{i=1}^{2}\sum_{j=1}^{k} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \qquad (8.95)$$

L'expression de D^2 définie par l'équation (8.95) est commode car elle peut se généraliser pour l'étude des tableaux de données r x c de rangées et k colonnes:

- tableaux d'effectifs conjoints croisant deux variables qualitatives ayant r et c modalités respectivement.

- tableaux d'effectifs de c échantillons indépendants provenant de populations multinomiales ayant r modalités.

8.11 TEST D'INDÉPENDANCE ENTRE DEUX VARIABLES QUALITATIVES

Soient (X,Y) deux variables qualitatives et notons par A_1, A_2, ..., A_r les r modalités de X et B_1, B_2, ..., B_c les c modalités de Y .

À partir des n observations sur le couple (X, Y) on peut constituer un TABLEAU D'EFFECTIFS CONJOINTS, appelé aussi TABLEAU DE CONTINGENCE

Effectifs conjoints (n_{ij})

X \ Y	B_1	B_2	...	B_j	...	B_c	TOTAL
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1c}	n_{1+}
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2c}	n_{2+}
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ic}	n_{i+}
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
A_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rc}	n_{r+}
TOTAL	n_{+1}	n_{+2}	...	n_{+j}	...	n_{+c}	n

où n_{ij} = EFFECTIF CONJOINT de (A_i, B_j)

$n_{i+} = \sum_j n_{ij}$ = effectif de A_i, $i = 1, 2, \ldots, r$

$n_{+j} = \sum_i n_{ij}$ = effectif de B_j, $j = 1, 2, \ldots, c$

$n = \Sigma\Sigma n_{ij}$ = effectif total

On définit

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

$f_{i+} = \frac{n_{i+}}{n}$ = fréquence marginale de A_i

$f_{+j} = \frac{n_{+j}}{n}$ = fréquence marginale de B_j

Modèle

Le modèle général de probabilités associé à un tel tableau de données est

$$p_{ij} = P [A_i \cap B_j] \quad i=1,2,...,r; \; j=1,2,...c$$

Les probabilités marginales de A_i et B_j sont

$$p_{i+} = \sum_{j=1}^{c} p_{ij} = P(A_i)$$

$$p_{+j} = \sum_{i=1}^{r} p_{ij} = P(B_j)$$

Une hypothèse intéressante et que l'on peut vouloir tester à l'aide du tableau de contingence est celle de l'indépendance des variables X et Y.

L'hypothèse nulle d'indépendance H_N est

$$H_N: \; p_{ij} = p_{i+} * p_{+j} \quad \text{tout } (i,j)$$

et la contre-hypothèse H_A est formée de la négation de H_N.

Test

Notons par Z_{ij} le nombre d'observations de l'échantillon de taille n ayant la modalité conjointe $A_i \cap B_j$. Alors Z_{ij} suit une distribution binomiale de paramètres n et p_{ij}:

$$Z_{ij} \sim Bin(n, p_{ij})$$

et si l'on admet l'hypothèse d'indépendance

$$Z_{ij} \sim Bin(n, p_{i+} * p_{+j})$$

Donc

$$E(Z_{ij}) = n \, p_{i+} * p_{+j}$$

est la valeur moyenne de Z_{ij}.

Les paramètres p_{i+} et p_{+j} sont inconnus et la méthode de vraisemblance maximale conduit aux estimations suivantes:

$$\hat{p}_{i+} = \frac{n_{i+}}{n} = f_{i+} \qquad \hat{p}_{+j} = \frac{n_{+j}}{n} = f_{+j}$$

Alors $$e_{ij} = n \hat{p}_{i+} * \hat{p}_{+j} = \frac{n_{i+} * n_{+j}}{n} \qquad (8.96)$$

est une estimation de l'effectif moyen attendus sous l'hypothèse d'indépendance. Le test repose sur la comparaison du tableau d'effectifs observés (n_{ij}) avec le tableau d'effectifs attendus (e_{ij}).

La statistique proposée par K. Pearson est

$$D^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(n_{ij}-e_{ij})^2}{e_{ij}} \qquad (8.97)$$

dont la distribution d'échantillonnage est de type khi-deux avec $\nu = (r-1)(c-1)$ degrés de liberté.

La justification pour la formule des degrés de liberté vient du fait que les espaces de paramètres sont:

$$\Omega = \left[(p_{11}, \ldots, p_{rc}): \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} = 1 \right] \text{ modèle général}$$

$$\omega = \left[(p_{11}, \ldots, p_{rc}): p_{ij} = p_{i+} * p_{+j}, \sum_{i=1}^{r} p_{i+}=1, \sum_{j=1}^{c} p_{+j}=1 \right]$$

modèle d'indépendance

et que la dimension de Ω est rc-1
la dimension de ω est (r-1) + (c-1)

ν = dimension de Ω - dimension ω = rc-1 - (r+c-2)
= (r-1) (c-1)

<u>Test</u>: on rejette H_N au seuil α si $D^2 > \chi^2_{\nu,\alpha}$

où $\nu=(r-1)(c-1)$ et D^2 est définie par (8.97)

<u>Condition</u>: $e_{ij} \geq 5$ pour tout (i,j)

Exemple 8.12 : exemple pannes du chapitre 1

machine (Y)

équipe(X)	A	B	C	D	total
J	10	15	15	10	50
S	10	20	15	20	65
N	20	10	30	25	85
total	40	45	60	55	200

Testons l'indépendance de X et Y.

Le tableau des effectifs attendus e_{ij} est

	A	B	C	D
J	10.0	11.3	15.0	13.8
S	13.0	14.6	19.5	17.9
N	17.0	19.1	25.5	23.4

et la statistique de K. Pearson est

$D^2 = 12.02$ avec 6 degrés de liberté.

Au seuil $\alpha = 0.05$, $\chi^2_{6,0.05} = 12.59$

On ne rejette pas l'hypothèse d'indépendance.

8.12 TESTS D'AJUSTEMENT

Le problème

Le problème de l'ajustement des données à une distribution théorique est une question fondamentale qui a contribué de façon significative au développement de la théorie des méthodes statistiques. Le développement de la théorie n'est pas encore terminé et, même de nos jours, de nombreux travaux sont publiés sur cette question.

Il existe deux raisons pour lesquelles on veut faire un test d'ajustement

- Résumer les données par une distribution convenablement choisie dans un esprit de description compacte définie par une forme de fonction connue et de quelques paramètres.

- S'assurer que les hypothèses de distribution sur lesquelles reposent plusieurs tests statistiques et intervalles de confiance (en particulier l'hypothèse de normalité dans les modèles statistiques) sont justifiées

Les méthodes employées sont:

- l'usage des coefficients d'asymétrie (b_1) et d'aplatissement (b_2) et du diagramme (β_1, β_2) afin de sélectionner une famille de distributions

- l'usage des graphiques percentiles-percentiles où l'on met en correspondance les percentiles expérimentaux et des percentiles dérivés d'une distribution théorique. Le cas le plus important sans doute est celui de la représentation des données sur une échelle gausso-arithmétique pour examiner l'hypothèse de normalité

- les tests statistiques: khi-deux de Pearson, Kolmogorov - Smirnov, Liliefors, Shapiro - Wilk

Formulation

Un test d'ajustement est une procédure statistique permettant de répondre à la question suivante: l'échantillon provient-il d'une distribution spécifique $f_X(x;\theta)$?. On formule la question sous la forme d'une HYPOTHÈSE dite NULLE H_N

$$H_N: \quad X \sim f_x(x;\theta) \tag{8.98}$$

où $f_X(.)$ est la distribution spécifiée et θ représente un ou plusieurs paramètres connus ou inconnus.

Il faut comprendre que le test n'est pas de "démontrer" que X suit la loi $f_X(.)$, mais de s'assurer la compatibilité de l'échantillon avec la loi. En fait, l'unique façon de démontrer que X suit une loi $f_X(.)$ est d'avoir la connaissance d'un échantillon de taille infinie.

Test khi-deux de Pearson

Le test d'ajustement de K. Pearson est probablement le premier test statistique à avoir été proposé et il est à l'origine de l'introduction de la très importante distribution khi-deux en statistique. L'idée de base du test est de faire la comparaison d'une distribution expérimentale, définie par un tableau d'effectifs, avec une distribution théorique choisie dans une classe de fonctions paramétrées. La comparaison est faite par l'intermédiaire d'une statistique mesurant la "distance" entre les deux distributions. Si cette distance est "petite", on admet l'adéquation entre la distribution expérimentale et la distribution théorique. Sinon le modèle théorique est rejeté.

Le test sera exposé pour le cas d'une variable continue. Soit $X_1, X_2, \ldots, X_n$ un échantillon de taille n provenant d'une variable X de loi $f_X(x;\theta)$. On calcule un tableau d'effectifs $n_1, n_2, \ldots, n_k$ où n_i est le nombre d'observations dans l'intervalle (a_i, a_{i+1}) $i = 1,2,\ldots,k$ où $a_1, a_2, \ldots, a_{k+1}$ sont les limites des intervalles contigus convenablement choisis.

Tableau d'effectifs

intervalles	(a_1,a_2)	(a_2,a_3)	$\ldots(a_i,a_{i+1})$	$\ldots(a_k,a_{k+1})$	Total
effectifs	n_1	n_2	$\ldots$ n_i $\ldots$	n_k	n

Dans le cas d'une variable discrète, les intervalles sont remplacés par les valeurs distinctes (généralement entières) de la variable.

En assumant l'hypothèse H_N, que la variable X suit la distribution $f_X(x;\theta)$, on calcule des effectifs attendus e_i selon les formules:

$$e_i = np_i \qquad i = 1,2,\ldots,k \tag{8.99}$$

$$p_i = \int_{a_i}^{a_{i+1}} f_X(x;\theta)\,dx \tag{8.100}$$

La statistique du test, proposé par Pearson, est la distance du khi-deux

$$D^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} \qquad (8.101)$$

On rejette l'hypothèse que l'échantillon provient de la distribution $f_X(x;\theta)$ si $D^2 > c$ où c est une constante définie par la distribution d'échantillonnage de D^2 et le seuil (α) choisi selon l'équation (8.2). Cette distribution d'échantillonnage est précisée dans la proposition suivante.

Proposition: La variable D^2 définie par l'équation (8.6) est asymptotiquement $(n\to\infty)$ distribuée selon une distribution khi-deux χ^2_ν où $\nu = k - 1 - p$ et p = nombre de paramètres estimés. La distribution χ^2_ν peut être employée pour des échantillons finis si $e_i \geq 5$ pour tout i.

Test du khi-deux de K. Pearson

On rejette H_N au seuil α si $D^2 > \chi^2_{\nu,\alpha}$ (8.102)

Exemple 8.13: test d'une distribution de Poisson

On a compté le nombre (X) de défauts de fabrication sur une plaque d'acier de 1 m^2 et proposé une distribution de Poisson pour X:

$$H_N: \quad p_X(x;\theta) = \frac{e^{-\theta}\,\theta^x}{x!}, \quad x = 0,1,2,\ldots \quad \theta > 0$$

Un échantillon de 60 plaques d'acier a donné le tableau d'effectifs

Tableau des effectifs

X	0	1	2	3	≥4	Total
n_i	32	15	9	4	0	60

On estime le paramètre θ avec la moyenne échantillonnale

$$\hat{\theta} = \bar{x} = \frac{32*0 + 15*1 + 9*2 + 4*3}{60} = 0.75$$

Avec cette valeur on calcule le tableau des probabilités p_i et des effectifs attendus e_i

X	0	1	2	3	≥4	Total
p_i	0.472	0.354	0.133	0.033	0.008	1
e_i	28.32	21.24	7.98	1.98	0.48	60

Les effectifs 1.98 et 0.48 étant inférieur à 5, on effectue un regroupement des trois dernières classes

X	0	1	≥2	Total
n_i	32	15	13	60
e_i	28.32	21.24	10.44	60
$\frac{(n_i-e_i)^2}{e_i}$	0.48	1.83	0.63	2.94

on obtient $D^2 = 2.94$ avec $\nu = 3-1-1 = 1$ degré de liberté. Il n'y a pas lieu de rejeter le modèle d'une distribution de Poisson puisque selon la table, $\chi^2_{1,0.05} = 3.84$

Exemple 8.14: test d'une distribution normale

Les données représentent la force de rupture (psi) de 100 bouteilles en vitre.

265	215	299	234	260	268	220	307	263	205
197	286	274	243	231	267	281	265	214	318
271	264	187	208	300	276	258	242	317	346
280	242	260	321	228	250	290	258	267	293
277	283	235	308	260	223	294	281	254	265
200	235	246	328	296	276	264	269	235	290
283	272	265	280	334	231	263	248	176	221
265	262	271	245	301	280	274	253	287	258
275	274	254	278	250	337	274	260	248	261
278	230	265	270	298	257	210	280	269	251

Une première analyse descriptive a donné: minimum=176, maximum=346, étendue=170, moyenne=264.06, écart-type=32.02, coefficient asy.=-0.13, coeff. aplat.=0.52

Tableau d'effectifs

intervalles	$]-\infty,190[$	$[190,210[$	$[210,230[$	$[230,250[$	$[250,270[$
effectifs	2	5	6	16	30

intervalles	$[270,290[$	$[290,310[$	$[310,330[$	$[330,\infty[$	Total
effectifs	24	10	4	3	100

Le tableau suivant présente une méthode systématique d'organisation des calculs pour l'exécution du test khi-deux.

Limite des intervalles a_i	Effectifs n_i	$z_i = \frac{a_i - 264.06}{32.02}$	$\Phi(z_i)$	p_i	e_i	$\frac{(n_i - e_i)^2}{e_i}$
	2			0.010	1	
190		-2.31	0.010			1.25
	5			0.036	3.6	
210		-1.69	0.046			
	6			0.099	9.9	1.54
230		-1.06	0.145			
	16			0.185	18.5	0.34
250		-0.44	0.330			
	30			0.245	24.5	1.23
270		0.19	0.575			
	24			0.216	21.6	0.27
290		0.81	0.791			
	10			0.133	13.3	0.82
310		1.43	0.924			
	4			0.056	5.6	
330		2.06	0.980			0.05
	3			0.020	2	
Total	100	-	-	1	100	5.50

Nous avons dû regrouper les deux classes extrêmes afin de satisfaire aux conditions d'application du test. La statistique de Pearson donne 5.50 avec $\nu = 7-1-2 = 4$ degrés de liberté puisqu'il reste 7 intervalles après le regroupement et que deux paramètres de la distribution normale ont été estimés avec les données.
Avec un seuil $\alpha = 0.05$, l'hypothèse d'une distribution normale n'est pas rejetée puisque $\chi^2_{4,0.05} = 9.49$. Cette conclusion sera confirmée avec le test d'ajustement de Shapiro - Wilk conçu spécialement pour tester la normalité.

Test de Kolmogorov - Smirnov

Le test du khi-deux de Pearson repose sur la construction d'un tableau d'effectifs ce qui exige un échantillon assez grand (disons ≥ 50). Par contre, le test de Kolmogorov - Smirnov peut être appliqué quel que soit la taille de l'échantillon. L'idée de base du test repose sur une statistique mesurant une distance entre la fonction de répartition théorique $F_X(x)$ et la fonction de répartition empirique $F_n(x)$.

Description du test

Soit $F_X(x)$ la fonction de répartition théorique de X et posons l'hypothèse

$$H_N: \text{ X distribuée selon } F_X(x)$$

Soit $x_1, x_2, \ldots, x_n$, les valeurs d'un échantillon de taille n provenant de X et définissons la fonction de répartition échantillonnale $F_n(x)$

$$F_n(x) = \begin{cases} 0 & x < x_{(1)} \\ \frac{1}{n} & x_{(1)} \leq x < x_{(2)} \\ \vdots & \vdots \\ \frac{k}{n} & x_{(k)} \leq x < x_{(k+1)} \\ \vdots & \vdots \\ 1 & x_{(n)} \leq x \end{cases} \qquad (8.103)$$

où $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ sont les valeurs ordonnées de l'échantillon.

La statistique de Kolmogorov - Smirnov est définie par

$$K = \max_{x} \mid F_X(n) - F_n(x) \mid \qquad (8.104)$$

et représente l'écart maximal entre les deux distributions. La règle de décision du test est:

$$\text{rejeter } H_N \text{ au seuil } \alpha \text{ si } K > k_{n,\alpha} \qquad (8.105)$$

Les constantes $k_{n,\alpha}$ représentent le 100 $(1-\alpha)^{\text{-ième}}$ percentile de la distribution d'échantillonnage de la statistique K. Le tableau 8.3 présente les valeurs de $k_{n,\alpha}$.

Tableau 8.3 des valeurs critiques $k_{n,\alpha}$ du test de Kolmogorov - Smirnov

	α		
n	0.10	0.05	0.01
1	0.950	0.975	0.995
2	0.776	0.842	0.929
3	0.636	0.708	0.829
4	0.565	0.624	0.734
5	0.509	0.563	0.669
6	0.468	0.519	0.617
7	0.436	0.483	0.576
8	0.410	0.454	0.542
9	0.381	0.430	0.513
10	0.369	0.409	0.489
11	0.352	0.391	0.468
12	0.338	0.375	0.449
13	0.325	0.361	0.432
14	0.314	0.349	0.418
15	0.304	0.338	0.404
16	0.295	0.327	0.392
17	0.286	0.318	0.381
18	0.279	0.309	0.371
19	0.271	0.301	0.361
20	0.265	0.294	0.352
21	0.259	0.287	0.344
22	0.253	0.281	0.337
23	0.247	0.275	0.330
24	0.242	0.269	0.323
25	0.238	0.264	0.317
26	0.233	0.259	0.311
27	0.229	0.254	0.305
28	0.225	0.250	0.300
29	0.221	0.246	0.295
30	0.218	0.242	0.290
31	0.214	0.238	0.285
32	0.211	0.234	0.281
33	0.208	0.231	0.277
34	0.205	0.227	0.273
35	0.202	0.224	0.269
36	0.199	0.221	0.265
37	0.196	0.218	0.262
38	0.194	0.215	0.258
39	0.191	0.213	0.255
40	0.189	0.210	0.252
≥ 41	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

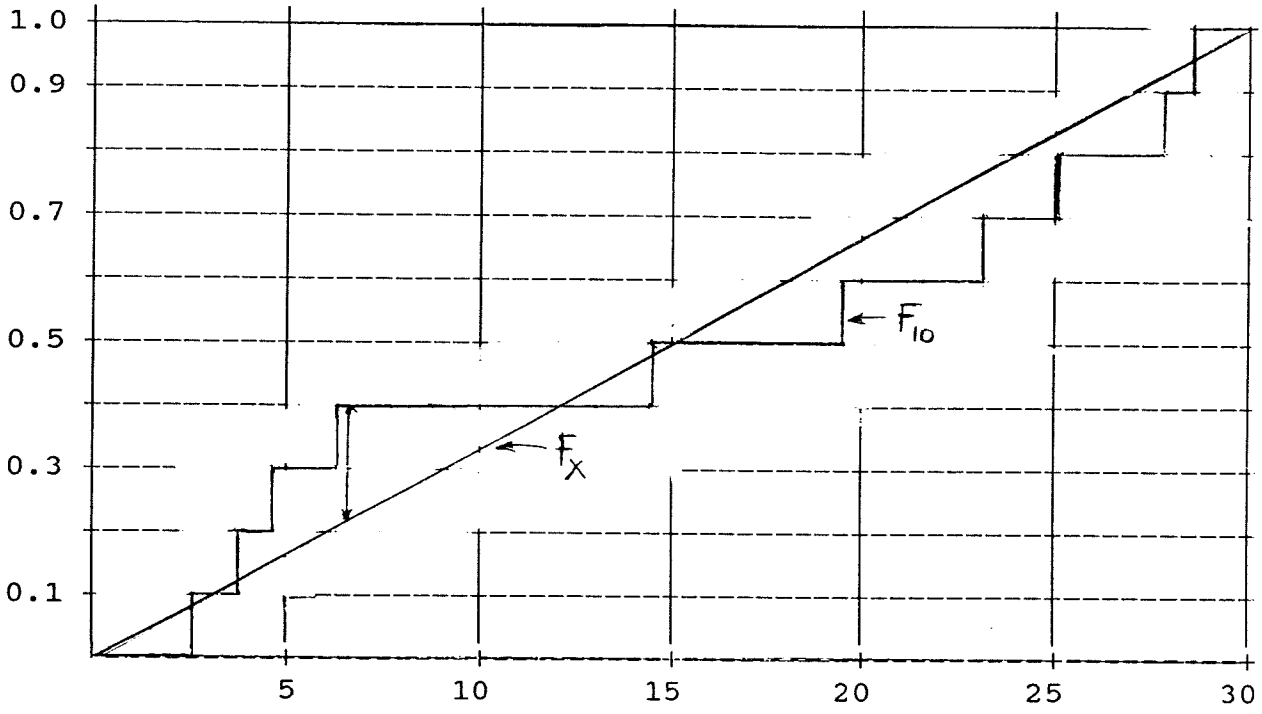
Exemple 8.14: test d'une distribution uniforme

Les données 4.8, 14.8, 28.2, 23.1, 4.4, 28.7, 19.5, 2.4, 25.0, 6.2 proviennent-elles d'une distribution uniforme sur l'intervalle [0,30]?

Posons H_N: $F_X(x) = \begin{cases} 0 & x < 0 \\ x/30 & 0 \le x \le 30 \\ 1 & x \ge 30 \end{cases}$

où $F_X(.)$ représente la répartition d'une distribution uniforme sur [0,30]. On calcule la fonction de répartition empirique $F_{10}(x)$

$$F_{10}(x) = \begin{cases} 0 & x < 2.4 \\ 1/10 & 2.4 \le x < 4.4 \\ 2/10 & 4.4 \le x < 4.8 \\ 3/10 & 4.8 \le x < 6.2 \\ 4/10 & 6.2 \le x < 14.8 \\ 5/10 & 14.8 \le x < 19.5 \\ 6/10 & 19.5 \le x < 23.1 \\ 7/10 & 23.1 \le x < 25.0 \\ 8/10 & 25.0 \le x < 28.2 \\ 9/10 & 28.2 \le x < 28.7 \\ 1 & 28.7 \le x \end{cases}$$



On calcule

$$K = \max_{0 \leq x \leq 30} |F_X(x) - F_{10}(x)| = 0.19$$

et puisque $k_{10,0.05} = 0.409$ on ne rejette pas le modèle d'une distribution uniforme.

Modification de Liliefors

- Le test de Kolmogorov - Smirnov repose sur l'hypothèse de base que $F_X(.)$ est continue et complètement spécifiée sans paramètres inconnus. Dans les autres cas, les valeurs de $k_{n,\alpha}$ de la table sont conservatrices.

- Pour tester le cas particulier d'une distribution normale avec paramètres inconnus, on doit utiliser une table modifiée qui a été calculée par Liliefors tenant en compte l'estimation des paramètres de la distribution théorique.

$$H_N: \quad X \text{ distribuée } N(\mu,\sigma^2) \quad \mu,\sigma \text{ inconnus}$$

On calcule $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

$$L = \max_z |\Phi(z) - S_n(z)| \qquad (8.106)$$

où $\Phi(.)$ est la fonction de répartition d'une distribution normale centrée réduite $S_n(z)$ est la fonction de répartition empirique définie pour les données centrées-réduites

$$z_i = \frac{x_i - \bar{x}}{s} \qquad i = 1,2,\ldots,n$$

On rejette H_N si $L > \ell_{n,\alpha}$

où $\ell_{n,\alpha}$ est le 100 $(1-\alpha)^{\text{-ième}}$ percentile de la statistique L de Liliefors. Les valeurs $\ell_{n,\alpha}$ sont donnés dans le tableau 8.4.

Tableau 8.4

Table des valeurs $\ell_{n,\alpha}$ du test de Liliefors

	α		
n	0.10	0.05	0.01
4	0.352	0.381	0.417
5	0.315	0.337	0.405
6	0.294	0.319	0.364
7	0.276	0.300	0.348
8	0.261	0.285	0.331
9	0.249	0.271	0.311
10	0.239	0.258	0.294
11	0.230	0.249	0.284
12	0.223	0.242	0.275
13	0.214	0.234	0.268
14	0.207	0.227	0.261
15	0.201	0.220	0.257
16	0.195	0.430	0.250
17	0.189	0.206	0.245
18	0.184	0.200	0.239
19	0.179	0.195	0.235
20	0.174	0.190	0.231
25	0.158	0.173	0.200
30	0.144	0.161	0.187
≥ 31	$\frac{0.805}{\sqrt{n}}$	$\frac{0.886}{\sqrt{n}}$	$\frac{1.031}{\sqrt{n}}$

Test de Shapiro-Wilk

L'hypothèse de normalité est sans aucun doute, celle qui est la plus souvent testée dans le cadre des modèles statistiques

$$Y = \varphi(X_1,\ldots,X_k;\beta_0,\beta_1,\ldots,\beta_p) + \varepsilon$$

$$\varepsilon \sim N(0,\sigma^2)$$

En effet, les tests statistiques et les intervalles de confiance sur les paramètres $\beta_0,\ldots,\beta_p$ reposent sur l'hypothèse de normalité du terme d'erreur ε.

De nombreux tests ont été développés en regard de cette question mais les études comparatives entre les divers tests ont montré que le test de Shapiro-Wilk est le meilleur pour détecter des écarts à la distribution normale.

Description du test

$$H_N: \quad X \sim N(\mu,\sigma^2), \quad \mu,\sigma \text{ inconnus}$$

Soit $X_1,X_2,\ldots,X_n$ un échantillon aléatoire et $X_{(1)} \leq X_{(2)} \leq \ldots X_{(n)}$ les statistiques d'ordre. La statistique de Shapiro-Wilk W est définie par

$$W = \frac{\left[\sum_{i=1}^{n} a_{n,i}\, X_{(i)}\right]^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}, \qquad \begin{array}{l} a \leq W \leq 1 \\ a \approx 0.70 \end{array} \qquad (8.107)$$

où les $a_{n,i}$ sont des coefficients calculés par Shapiro-Wilk. Les coefficients $a_{n,i}$ sont donnés dans le tableau 8.3 et vérifient la relation suivante:

$$\begin{cases} a_{n,i} = 0 & \text{si n est impair et i= n+1/2} \\ a_{n,n-i+1} = -a_{n,i} \end{cases}$$

La statistique W est essentiellement le carré du coefficient de corrélation entre $X_{(i)}$ et $Z_{(i)}$ où les $Z_{(i)}$ sont les statistiques

d'ordre provenant d'une distribution normale centrée-réduite. En effet, si $X_{(i)}$ provient d'une distribution $N(\mu,\sigma^2)$ on a

$$E(X_{(i)}) = \mu + \sigma E(Z_{(i)}) = \mu + \sigma a_{n,i}$$

où

$$Z_{(i)} = \Phi^{-1}\left[\frac{i-3/8}{n-1/4}\right]$$

Donc le graphique des percentiles $X_{(i)}$ en fonction des percentiles $Z_{(i)}$ donne une droite en moyenne.

test: on rejette l'hypothèse de normalité si $W < w_{n,\alpha}$ où $w_{n,\alpha}$ est α-ième percentile de la distribution d'échantillonnage de W. Les valeurs $w_{n,\alpha}$ pour $\alpha = 0.01, 0.05, 0.10$ et $n \leq 50$ sont données sur la figure 8.6.

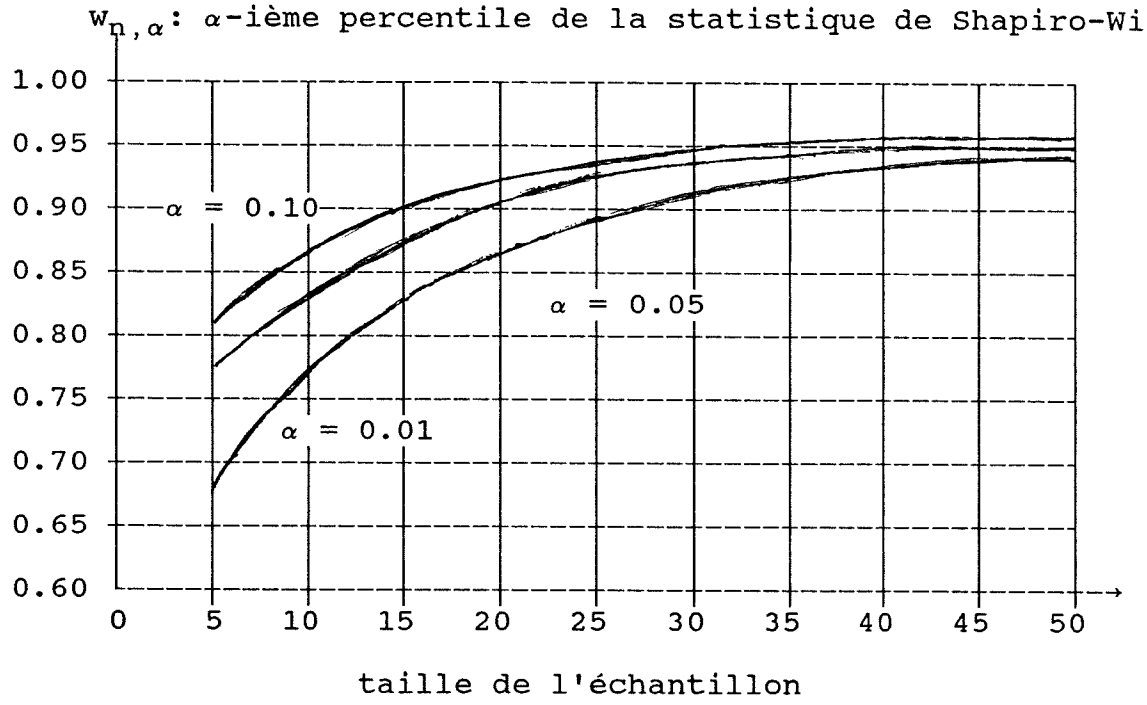


Figure 8.6: percentiles $\omega_{n,\alpha}$ de la statistique de Shapiro-Wilk

La procédure UNIVARIATE de SAS avec l'option NORMAL effectue le test de Shapiro-Wilk pour des échantillons de taille $n \leq 50$ et le test de Kolmogorov-Smirnov si $n \geq 51$. Le programme 2D du progiciel BMDP effectue le test de Shapiro-Wilk pour $3 \leq n \leq 2000$ en spécifiant l'option WSTAT dans le paragraphe PRINT.

Exemple 8.15: test de Shapiro-Wilk avec petit échantillon

Les données sont:

$$x_1 = 6,\ x_2 = 1,\ x_3 = -4,\ x_4 = 8,\ x_5 = -2,$$
$$x_6 = 5,\ x_7 = 0$$
$$\bar{x} = 2,\quad \Sigma(x_i - \bar{x})^2 = 118$$

Les statistiques d'ordre sont:

$$x_{(1)} = -4,\ x_{(2)} = -2,\ x_{(3)} = 0,$$
$$x_{(4)} = 1,\ x_{(5)} = 5,\ x_{(6)} = 6,$$
$$x_{(7)} = 8$$

Les coefficients de Shapiro-Wilk avec n = 7 provenant du tableau 8.3

$$a_{7,1} = 0.6233,\ a_{7,2} = 0.3031,\ a_{7,3} = 0.1401,\ a_{7,4} = 0.0000$$
$$a_{7,5} = -a_{7,3},\ a_{7,6} = -a_{7,2},\ a_{7,7} = -a_{7,1}$$

$$\sum_{i=1}^{7} n_{7,i}\, x_{(i)} = 0.6233(8+4) + 0.3031(6+2) + 0.1401(5+0) = 10.6049$$

$$W = \frac{(10.6069)^2}{118} = 0.9530$$

On ne rejette pas l'hypothèse de normalité au seuil $\alpha = 0.05$ puisque $w_{7,0.05} = 0.82$.

Exemple 8.16: test de Shapiro-Wilk avec n = 100, suite de l'exemple 8.2

La statistique W a été calculée avec le progiciel BMDP.

On obtient $W = 0.9785$

et $P[\ W \leq 0.9785\ ] = 0.45$

On ne rejette pas l'hypothèse de normalité.

Tableau 8.5

Coefficients $a_{n,i}$ du test de Shapiro-Wilk

i\n	2	3	4	5	6	7	8	9	10
1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6032	0.5888	0.5739
2	----	.0000	.1677	.2413	.2806	.3031	.3164	.3244	.3201
3	----	----	----	.0000	.0873	.1401	.1743	.1976	.2141
4	----	----	----	----	----	.0000	.0361	.0947	.1224
5	----	----	----	----	----	----	----	.0000	.0399

i\n	11	12	13	14	15	16	17	18	19	20
1	0.5601	0.5475	0.5350	0.5231	0.5150	0.5056	0.4968	0.4886	0.4808	0.4734
2	.3315	.3323	.3323	.3318	.3306	.3299	.3273	.3253	.3232	.3211
3	.2260	.2347	.2412	.2460	.2495	.2521	.2540	.2553	.2561	.2565
4	.1429	.1586	.1707	.1802	.1878	.1939	.1988	.2027	.2059	.2085
5	.0695	.0922	.1099	.1240	.1353	.1447	.1524	.1587	.1641	.1686
6	.0000	.0303	.0539	.0727	.0880	.1005	.1109	.1197	.1271	.1334
7	----	----	.0000	.0240	.0433	.0593	.0725	.0837	.0932	.1013
8	----	----	----	----	.0000	.0196	.0359	.0496	.0612	.0711
9	----	----	----	----	----	----	.0000	.0163	.0303	.0422
10	----	----	----	----	----	----	----	----	.0000	.0140

i\n	21	22	23	24	25	26	27	28	29	30
1	0.4643	0.4590	0.4542	0.4493	0.4450	0.4407	0.4366	0.4328	0.4291	0.4254
2	.3185	.3156	.3126	.3098	.3069	.3043	.3018	.2992	.2968	.2944
3	.2578	.2571	.2563	.2554	.2543	.2533	.2522	.2510	.2499	.2487
4	.2119	.2131	.2139	.2145	.2148	.2151	.2152	.2151	.2150	.2148
5	.1736	.1764	.1787	.1807	.1822	.1836	.1848	.1857	.1864	.1870
6	.1399	.1443	.1480	.1512	.1539	.1563	.1584	.1601	.1616	.1630
7	.1092	.1150	.1201	.1245	.1283	.1316	.1346	.1372	.1395	.1415
8	.0804	.0878	.0941	.0997	.1046	.1089	.1128	.1162	.1192	.1219
9	.0530	.0618	.0696	.0764	.0823	.0876	.0923	.0965	.1002	.1036
10	.0263	.0368	.0459	.0539	.0610	.0672	.0728	.0778	.0822	.0862
11	.0000	.0122	.0228	.0321	.0403	.0476	.0540	.0598	.0650	.0697
12	----	----	.0000	.0107	.0200	.0284	.0358	.0424	.0483	.0537
13	----	----	----	----	.0000	.0094	.0178	.0253	.0320	.0381
14	----	----	----	----	----	----	.0000	.0084	.0159	.0277
15	----	----	----	----	----	----	----	----	.0000	.0076

Tableau 8.5 suite

i\n	31	32	33	34	35	36	37	38	39	40
1	0.4220	0.4188	0.4156	0.4127	0.4096	0.4068	0.4040	0.4015	0.3989	0.3064
2	.2021	.2898	.2876	.2854	.2834	.2813	.2704	.2774	.2753	.2737
3	.2473	.2463	.2451	.2439	.2427	.2415	.2403	.2391	.2380	.2368
4	.2145	.2141	.2137	.2132	.2127	.2121	.2116	.2110	.2104	.2098
5	.1874	.1878	.1880	.1882	.1883	.1883	.1883	.1881	.1880	.1878
6	.1641	.1651	.1660	.1667	.1673	.1678	.1683	.1686	.1689	.1691
7	.1443	.1449	.1463	.1475	.1487	.1496	.1505	.1513	.1520	.1526
8	.1243	.1265	.1284	.1301	.1317	.1331	.1344	.1336	.1366	.1376
9	.1066	.1093	.1118	.1140	.1160	.1179	.1196	.1211	.1225	.1237
10	.0899	.0931	.0961	.0988	.1013	.1036	.1056	.1073	.1092	.1108
11	.0739	.0777	.0812	.0844	.0873	.0900	.0924	.0947	.0967	.0986
12	.0585	.0629	.0669	.0706	.0739	.0770	.0798	.0824	.0848	.0870
13	.0435	.0485	.0530	.0572	.0610	.0643	.0677	.0706	.0733	.0759
14	.0289	.0344	.0395	.0441	.0484	.0523	.0559	.0592	.0622	.0651
15	.0144	.0206	.0262	.0314	.0361	.0404	.0444	.0481	.0515	.0546
16	.0000	.0068	.0131	.0187	.0239	.0287	.0331	.0372	.0409	.0444
17	----	----	.0000	.0062	.0119	.0172	.0220	.0264	.0305	.0343
18	----	----	----	----	.0000	.0057	.0110	.0158	.0203	.0244
19	----	----	----	----	----	----	.0000	.0053	.0101	.0146
20	----	----	----	----	----	----	----	----	.0000	.0049

i/n	41	42	43	44	45	46	47	48	49	50
1	0.3940	0.3917	0.3894	0.3872	0.3850	0.3830	0.3808	0.3789	0.3770	0.3751
2	.2719	.2701	.2684	.2667	.2651	.2635	.2620	.2604	.2589	.2574
3	.2337	.2345	.2334	.2323	.2313	.2302	.2291	.2281	.2271	.2260
4	.2091	.2085	.2078	.2072	.2065	.2058	.2052	.2045	.2038	.2032
5	.1876	.1874	.1871	.1868	.1865	.1862	.1859	.1855	.1851	.1847
6	.1693	.1694	.1695	.1695	.1695	.1695	.1695	.1693	.1692	.1691
7	.1531	.1535	.1539	.1542	.1545	.1548	.1550	.1551	.1553	.1554
8	.1384	.1392	.1398	.1405	.1410	.1415	.1420	.1423	.1427	.1430
9	.1249	.1259	.1269	.1278	.1286	.1293	.1300	.1306	.1312	.1317
10	.1123	.1136	.1149	.1160	.1170	.1180	.1189	.1197	.1205	.1212
11	.1004	.1020	.1035	.1049	.1062	.1073	.1085	.1095	.1105	.1113
12	.0891	.0909	.0927	.0943	.0959	.0972	.0986	.0998	.1010	.1020
13	.0782	.0804	.0824	.0842	.0860	.0876	.0892	.0906	.0919	.0932
14	.0677	.0701	.0724	.0745	.0765	.0783	.0801	.0817	.0832	.0846
15	.0575	.0602	.0628	.0651	.0673	.0694	.0713	.0731	.0748	.0764
16	.0476	.0506	.0534	.0560	.0584	.0607	.0628	.0648	.0667	.0685
17	.0379	.0411	.0442	.0471	.0497	.0522	.0546	.0568	.0588	.0608
18	.0283	.0318	.0352	.0383	.0412	.0439	.0465	.0489	.0511	.0532
19	.0188	.0227	.0263	.0296	.0328	.0357	.0385	.0411	.0436	.0459
20	.0094	.0136	.0175	.0211	.0245	.0277	.0307	.0335	.0361	.0386
21	.0000	.0045	.0087	.0126	.0163	.0197	.0229	.0259	.0288	.0314
22	----	----	.0000	.0042	.0081	.0118	.0153	.0185	.0215	.0244
23	----	----	----		.0000	.0039	.0076	.0111	.0143	.0174
24	----	----	----		----	----	.0000	.0037	.0071	.0104
25	----	----	----		----	----	----	----	.0000	.0035

8.13 TEST D'ÉGALITÉ DE K MOYENNES

Le test d'égalité de deux moyennes vu à la section 8.8 peut se généraliser au cas de plusieurs (3 ou plus) moyennes. L'étude se fait à l'aide du modèle de classification simple exposé au chapitre 7. Nous abordons dans cette section un cas particulier d'un ensemble de techniques connues sous le nom de MODÈLES D'ANALYSE DE LA VARIANCE. Ces méthodes permettent d'établir si un groupe de variables qualitatives a une influence significative sur une variable continue. L'analyse principale est présentée sous la forme d'un tableau d'analyse de la variance où la variation totale de la variable est décomposée selon chacune des sources de variation associées aux variables qualitatives.

Modèle de classification simple: une variable de classification

$$y_{ij} = \beta_0 + \beta_i + \varepsilon_{ij} \qquad (8.108)$$

$$i=1,2,\ldots,k; \quad j=1,2,\ldots,n_i$$

où

y_{ij} : j-ième observation du i-ième groupe
k: nombre de groupes ou échantillons indépendants associés aux k modalités de la classification
n_i : nombre d'observations du i-ième groupe
β_0 : effet général
β_i : effet différentiel du i-ième groupe
ε_{ij} : terme d'erreur

En plus de l'équation (8.108) on ajoute l'hypothèse de base

$$\varepsilon_{ij} \sim N(0,\sigma^2) \qquad (8.109)$$

nécessaire pour effectuer les tests statistiques et calculer les intervalles de confiance sur les paramètres $\beta_0,\beta_1,\ldots,\beta_k$. Les équations (8.108) et (8.109) peuvent se reformuler en une seule équation

$$y_{ij} \sim N(\beta_0 + \beta_i,\sigma^2) \qquad (8.110)$$

On peut faire l'hypothèse supplémentaire suivante concernant les coefficients différentiels β_i soit

$$\sum_{i=1}^{k} \beta_i = 0 \qquad (8.111)$$

En effet, l'équation (8.111) est naturelle puisque si

$$\sum_{j=1}^{k} \beta_j = k\,\bar{\beta} \neq 0$$

il suffit de poser

$$\beta_0' = \beta_0 + \bar{\beta} \qquad \beta_i' = \beta_i - \bar{\beta}$$

et alors

$$y_{ij} = \beta_0+\beta_i+\varepsilon_{ij} = \beta_0'+\beta_i'+\varepsilon_{ij}$$

avec

$$\sum_{i=1}^{k} \beta_i' = 0$$

Nous admettrons donc à partir de maintenant l'équation (8.111).

La principale question soulevée par le modèle de classification simple est de savoir si la variable de classification a un effet significatif sur la variable expliquée y. En d'autres termes, les k échantillons indépendants proviennent-ils de k populations de moyennes μ_i différentes?

$$\mu_i = \beta_0 + \beta_i \qquad i=1,2,\ldots,k$$

Cette question peut se formuler par l'intermédiaire de la mise à l'épreuve de l'hypothèse nulle

$$H_N: \beta_1 = \beta_2 = \ldots = \beta_k = 0 \qquad (8.112)$$

La contre-hypothèse H_A est formée de la négation de H_N, soit

$$H_A: \beta_i \neq 0 \text{ pour au moins un } i$$

La procédure du test est basée sur une équation de décomposition de la variabilité de Y, appelée équation d'analyse de la variance.

Équation d'analyse de la variance

Posons
$$\bar{y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \qquad i=1,2,\ldots,k \qquad (8.113)$$

$$n = \sum_{i=1}^{k} n_i$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} = \sum_{i=1}^{k} \frac{n_i}{n} \bar{y}_{i.} \qquad (8.114)$$

$$SC_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij}-\bar{y})^2 \quad \text{appelée variation totale} \qquad (8.115)$$

$$SC_E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 \quad \text{appelée variation intra-groupe} \qquad (8.116)$$

$$SC_M = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y})^2 \quad \text{appelée variation inter-groupe} \qquad (8.117)$$

$$= \sum_{i=1}^{k} n_i (\bar{y}_{i.} - \bar{y})^2$$

on a

$$y_{ij} - \bar{y} = (y_{ij}-\bar{y}_{i.}) + (\bar{y}_{i.}-\bar{y})$$

Si on élève au carré et additionne pour toutes les valeurs de i et j on obtient

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij}-\bar{y})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij}-\bar{y}_{i.})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_{i.}-\bar{y})^2$$

$$+ 2 \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij}-\bar{y}_{i.})\,(\bar{y}_{i.}-\bar{y})$$

Mais la somme des produits croisés s'annulle de par la définition de $\bar{y}_{i.}$ et $\bar{y}$. Donc

$$SC_T = SC_M + SC_E \qquad (8.118)$$

appelée équation d'analyse de la variance du modèle de classification simple. À chacune des sommes de carrés est associée un degré de liberté représentant le nombre de termes linéairement indépendants contribuant à la somme. On constate que

SC_T possède n-1 degrés de liberté
SC_E possède $\Sigma(n_i-1) = n-k$ degrés de liberté
SC_M possède k-1 degrés de liberté

et parallèlement à l'équation (8.118) correspond une équation pour les degrés de liberté:

$$n-1 = (k-1) + (n-k) \qquad (8.119)$$

On définit les carrés moyens intra-groupe et inter-groupe par

$$CM_E = SC_E/(n-k) \qquad (8.120)$$

$$CM_M = SC_M/(k-1) \qquad (8.121)$$

et on présente les principales quantités, sommes de carrés, degrés de liberté et carrés moyens dans un tableau sommaire appelé tableau d'analyse de la variance.

Tableau d'analyse de la variance

Source	Somme de carrés	Degrés de liberté	Carrés moyens	F
inter-groupe (modèle)	SC_M	k-1	CM_M	$\frac{CM_M}{CM_E}$
intra-groupe (erreur)	SC_E	n-k	CM_E	-
total	SC_T	n-1	-	-

Calculs

Les calculs des sommes de carrés sont généralement effectués à l'aide des identités:

$$SC_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{\left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \right)^2}{n}$$

$$SC_M = \sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n_i} y_{ij} \right)^2}{n_i} - \frac{\left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \right)^2}{n}$$

$$SC_E = SC_T - SC_M$$

Test de H_N

Admettant l'hypothèse de base d'une distribution gaussienne pour la variable Y, on montre que

SC_M est distribuée χ^2_{k-1}

SC_E est distribuée χ^2_{n-k}

SC_M et SC_E sont indépendantes.

Alors le quotient

$$F = \frac{SC_M/(k-1)}{SC_E/(n-k)} = \frac{CM_M}{CM_E} \qquad (8.122)$$

est distribué selon une distribution de Fisher (k-1, n-k). On a donc le test suivant:

on rejette H_N au seuil α si $F > F_{k-1,n-k,\alpha}$ (8.123)

Exemple 8.17: exemple usure de piston du chapitre 1

Il s'agit d'une expérience où deux facteurs contrôlés (variables explicatives) soit la marque de piston (A,B,C,D) et le type d'huile (1,2,3,4,5) produisent une perte de poids (mesurée en grammes) du piston et représentée par une variable Y. Les données sont classées selon une classification double (cf exemple 7.24) selon les deux facteurs contrôlés. Pour illustrer seulement, nous allons traiter les données selon le modèle de classification simple, comme si l'un des facteurs était constant. Nous avons choisi de considérer la marque de piston comme étant constante et de faire l'analyse selon le modèle de classification simple définie par le type d'huile. Nous employons donc le modèle de classification simple

$$y_{ij} = \beta_0 + \beta_i + \varepsilon_{ij} \qquad i = 1,2,3,4,5; \qquad j = 1,2,3,4$$

où y_{ij} est la perte de poids mesurée en grammes du j-ième piston utilisé avec l'huile de type i

Tableau des données

type d'huile	observations	moyennes
1	1.641, 1.306, 1.149, 1.025	1.28025
2	1.782, 1.568, 1.223, 1.919	1.6230
3	1.570, 1.240, 1.068, 1.982	1.4650
4	1.493, 1.415, 1.118, 1.812	1.4595
5	1.672, 1.291, 1.004, 2.015	1.4955

On a

$$\sum_{i=1}^{5}\sum_{j=1}^{4} y_{ij} = 29.293 \qquad \sum_{i=1}^{5}\sum_{j=1}^{4} y_{ij}^2 = 44.948$$

$$\sum_{i=1}^{5} \frac{\left(\sum_{j=1}^{4} y_{ij}\right)^2}{4} = 43.144$$

$$SC_T = 44.948 - \frac{(29.293)^2}{20} = 2.044$$

$$SC_M = 43.144 - \frac{(29.293)^2}{20} = 0.240$$

$$SC_E = 2.044 - 0.240 = 1.804$$

et le tableau d'analyse de variance

Source	Somme de carrés	Degrés de liberté	Carrés moyens	F
inter-groupe	0.240	4	0.06	0.50
intra-groupe	1.804	15	0.120	-
total	2.044	19	-	-

Puisque $F_{4,15,0.05} = 3.05$ on ne rejette pas l'hypothèse nulle

$$H_N: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

et il semble que le type d'huile n'a pas un effet significatif sur la perte de poids du piston.

8.14 FORMULAIRE DES PRINCIPAUX TESTS

(α=seuil du test)

Hypothèse nulle	Statistique pour le test	Alternative	Critère de rejet	Fonction caractéristique
$H_N: \mu=\mu_0$ σ^2 connue population normale	$Z_0 = \dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	$H_A: \mu\neq\mu_0$	$\lvert Z_0\rvert > z_{\alpha/2}$	éq.(8.26) fig. 8.2
		$H_A: \mu<\mu_0$	$Z_0 < -z_\alpha$	éq.(8.22) fig. 8.1
		$H_A: \mu>\mu_0$	$Z_0 > z_\alpha$	éq.(8.17) fig. 8.1
$H_N: \mu=\mu_0$ σ^2 inconnue population normale	$T_0 = \dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}$	$H_A: \mu\neq\mu_0$	$\lvert T_0\rvert > t_{n-1,\alpha/2}$	fig.8.4
		$H_A: \mu<\mu_0$	$T_0 < -t_{n-1,\alpha}$	fig.8.3
		$H_A: \mu>\mu_0$	$T_0 > t_{n-1,\alpha}$	fig.8.3

Remarque: test employé pour le cas de différences pairées avec $x_i = x_{1i} - x_{2i}$

$H_N: \mu_1-\mu_2=0$ σ_1^2, σ_2^2 connues populations normales	$Z_0 = \dfrac{(\bar{x}_1-\bar{x}_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$	$H_A: \mu_1-\mu_2\neq 0$	$\lvert Z_0\rvert > z_{\alpha/2}$	éq.(8.71)
		$H_A: \mu_1-\mu_2<0$	$Z_0 < -z_\alpha$	éq.(8.68)
		$H_A: \mu_1-\mu_2>0$	$Z_0 > z_\alpha$	
$H_N: \mu_1-\mu_2=0$ $\sigma_1^2=\sigma_2^2=\sigma^2$ inconnues populations normales	$T_0 = \dfrac{(\bar{x}_1-\bar{x}_2)}{S_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$	$H_A: \mu_1-\mu_2\neq 0$	$\lvert T_0\rvert > t_{n1+n2-2,\alpha/2}$	éq.(8.80)
		$H_A: \mu_1-\mu_2<0$	$T_0 < -t_{n1+n2-2,\alpha}$	éq.(8.76)
		$H_A: \mu_1-\mu_2>0$	$T_0 > t_{n1+n2-2,\alpha}$	

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

H_N : $\mu_1 - \mu_2 = 0$ — $T_0 = \dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$

$\sigma_1^2 \neq \sigma_2^2$ inconnues
populations
normales

H_A : $\mu_1 - \mu_2 \neq 0$	$\lvert T_0 \rvert > t_{\nu,\alpha/2}$
H_A : $\mu_1 - \mu_2 < 0$	$T_0 < -t_{\nu,\alpha}$
H_A : $\mu_1 - \mu_2 > 0$	$T_0 > t_{\nu,\alpha}$

$$a = s_1^2/n_1, \quad b = s_2^2/n_2 \qquad \nu = \frac{(a+b)^2}{a^2/(n_1-1) + b^2/(n_2-1)}$$

H_N : $\theta = \theta_0$
population
Bernoulli
$n\theta_0 \geq 5$

$$Z_0 = \frac{\hat{\theta} - \theta_0}{\sqrt{\dfrac{\theta_0(1-\theta_0)}{n}}}$$

$$\hat{\theta} = \frac{x}{n} = \bar{x}$$

H_A : $\theta \neq \theta_0$	$\lvert Z_0 \rvert > z_{\alpha 12}$	
H_A : $\theta < \theta_0$	$Z_0 < -z_\alpha$	
H_A : $\theta > \theta_0$	$Z_0 > z_\alpha$	éq.(8.56)

H_N : $\theta_1 - \theta_2 = 0$
populations
Bernoulli

$$Z_0 = \frac{\hat{\theta}_1 - \hat{\theta}_2}{ET(\hat{\theta}_1 - \hat{\theta}_2)}$$

H_A : $\theta_1 \neq \theta_2$	$\lvert Z_0 \rvert > z_{\alpha/2}$
H_A : $\theta_1 < \theta_2$	$Z_0 < -z_\alpha$
H_A : $\theta_1 > \theta_2$	$Z_0 > z_\alpha$

$$\hat{\theta} = \frac{n_1\hat{\theta}_1 + n_2\hat{\theta}_2}{n_1+n_2} \qquad ET(\hat{\theta}_1 - \hat{\theta}_2) = \sqrt{\hat{\theta}(1-\hat{\theta})(1/n_1 + 1/n_2)}$$

$H_N: \sigma^2=\sigma_0^2$	$\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$	$H_A^2: \sigma^2\neq\sigma_0^2$	$\chi_0^2>\chi^2_{n-1,\alpha/2}$	éq.(8.46) fig.8.5
population normale			ou $\chi_0^2<\chi^2_{n-1,1-\alpha/2}$	
		$H_A: \sigma^2<\sigma_0^2$	$\chi_0^2<\chi^2_{n-1,1-\alpha}$	éq.(8.39) fig.8.5
		$H_A: \sigma^2>\sigma_0^2$	$\chi_0^2>\chi^2_{n-1,\alpha}$	éq.(8.43) fig.8.5

$H_N: \sigma_1^2=\sigma_2^2$	$F_0 = \frac{S_1^2}{S_2^2}$	$H_A: \sigma_1^2\neq\sigma_2^2$	$F_0>F_{n1-1,n2-1,\alpha/2}$
populations normales			ou $F_0<F_{n1-1,n2-1,1-\alpha/2}$
		$H_A: \sigma_1^2<\sigma_2^2$	$F_0<F_{n1-1,n2-1,1-\alpha}$
		$H_A: \sigma_1^2>\sigma_2^2$	$F_0>F_{n1-1,n2-1,\alpha}$

Test d'ajustement du Khi-deux

H_N: X suit loi $f_x(x:\theta)$ $\quad D^2 = \sum_{i=1}^{k} \frac{(n_i-e_i)^2}{e_i}$ $\quad H_A$: non H_N $\quad D^2>\chi^2_{\nu,\alpha}$

e.g. $N(\mu,\sigma^2)$
Poisson
etc...

où n_i = effectifs observés
e_i = effectifs sous H_N
p = nombre paramètres estimés
k = nombre de classes (après regroupement s'il y a lieu)
ν = k-p-1

données: tableau d'effectifs

Condition d'application: $e_i \geq 5$

Test d'indépendance entre deux variables qualitatives

H_N : $P[X = A_i , Y=B_j] = P[X=A_i]*P[Y=B_j]$
$i=1,\ldots,r$
$j=1,\ldots,c$

Données: tableau de contingence

$$D^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij}-e_{ij})^2}{e_{ij}}$$

n_{ij} = effectifs observés
e_{ij} = effectifs sous H_N
ν = (r-1)(c-1) = degrés de liberté

on rejette H_N au seuil α si:

$$D^2 > \chi^2_{\nu,\alpha}$$

condition d'application: $e_{ij} \geq 5$

Modèle de classification simple

$y_{ij} = \mu+\beta_i+\varepsilon_{ij}$; $i=1,2,\ldots k$; $j=1,2,\ldots n_i$; $\Sigma\beta_i=0$

$\varepsilon_{ij} \sim N(0,\sigma^2)$ $\quad$ H_N : $\beta_1=\beta_2=\ldots=\beta_k=0$

Tableau d'analyse de la variance

Source	Somme de carrés	Degrés de liberté	Carrés moyens	F
inter-groupe (modèle)	SC_M	k-1	CM_M	$\frac{CM_M}{CM_E}$
intra-groupe (erreur)	SC_E	n-k	CM_E	-
totale	SC_T	n-1	-	-

$$SC_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij}-\bar{y})^2 \qquad SC_M = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (\bar{y}_i-\bar{y})^2 \qquad n = \sum_{i=1}^{k} n_i$$

$$SC_E = SC_T - SC_\mu \qquad F = \frac{SC_M/k-1}{SC_E/n-k} = \frac{CM_M}{CM_E}$$

Test: on rejette H_N si $F > F_{k-1,n-k,\alpha}$

8.15 UTILISATION DE SAS: EXEMPLES

L'exécution d'un test statistique repose sur le calcul d'une statistique admettant l'hypothèse nulle et sur la comparaison de la statistique avec un percentile de la distribution d'échantillonnage de la statistique. La très grande majorité des tests statistiques ont pour distributions d'échantillonnage une des quatre lois de probabilités:

- gaussienne centrée-réduite
- Student avec ν degrés de liberté
- khi-deux avec ν degrés de liberté
- Fisher avec (ν_1,ν_2) degrés de liberté

Après avoir choisi un seuil α on compare la statistique ($>$ ou $\leq$) avec le $100(1-\alpha)$ -ième percentile $z_\alpha, t_{\nu,\alpha}, \chi^2_{\nu,\alpha}$ ou $F_{\nu 1,\nu 2,\alpha}$ selon le cas. Si la statistique calculée dépasse le percentile, on rejette l'hypothèse nulle, sinon on l'accepte.

Lorsqu'on utilise un progiciel statistique comme SAS pour exécuter un test statistique, on procède d'une manière différente. Le programme calcule la probabilité

$$P [W > w_0]$$

où W est la statistique du test et w_0 est la valeur calculée de la statistique avec les données. L'utilisateur décide de

- rejeter l'hypothèse nulle si $P [W > w_0] < \alpha$
- ne pas rejeter l'hypothèse si $P [W > w_0] > \alpha$

où α est le seuil qu'il a choisi.

Remarque: la région critique peut aussi être de la forme $W < W_0$ ou encore $|W| < W_0$ selon le test employé.

Plusieurs procédures du progiciel SAS effectuent des tests statistiques automatiquement ou par requêtes explicites. Notons pour fins de références les tests statistiques associés à quelques unes des procédures de SAS à vocation statistique.

Procédure	Test
ANOVA	. test d'égalité de k moyennes
CORR	. nullité du coefficient de corrélation
FREQ	. test d'indépendance entre les deux variables d'un tableau bidimensionnel
MEANS	. test de la nullité d'une moyenne . test d'égalité de moyennes pour deux échantillons pairés
REG	. test de signification d'un modèle de régression . test de nullité des coefficients du modèle
TTEST	. test d'égalité de moyennes de deux échantillons indépendants . test d'égalité de deux variances
UNIVARIATE	. test d'ajustement à une loi normale . test de la nullité de la moyenne

```
++++++++++++++++++++++++++++++++++++++++++++++++++++
+   EXEMPLE:  TEST DE NORMALITE DE SHAPIRO-WILK,  +
+             TRACAGE DES DONNEES SUR ECHELLE     +
+             DE PROBABILITE  GAUSSIENNE          +
+  PROCEDURE: UNIVARIATE                          +
++++++++++++++++++++++++++++++++++++++++++++++++++++

DATA EXEMPLE;
       INPUT X @@;
          LIST;CARDS;
              6 1 -4 8 -2 5 0
;
PROC UNIVARIATE DATA=EXEMPLE PLOT NORMAL;
        TITLE1 'DONNEES EXEMPLE 8.16';
        TITLE2 'ECHELLE GAUSSIENNE';
        TITLE3 'TEST DE SHAPIRO-WILK';
```

++++++++++++++++++++++++
+ OUTPUT DU PROGRAMME +
++++++++++++++++++++++++

DONNEES EXEMPLE 8.16
ECHELLE GAUSSIENNE
TEST DE SHAPIRO-WILK

UNIVARIATE

VARIABLE=X

MOMENTS

N	7	SUM WGTS	7
MEAN	2	SUM	14
STD DEV	4.43471	VARIANCE	19.6667
SKEWNESS	0.0481563	KURTOSIS	-1.53663
USS	146	CSS	118
CV	221.736	STD MEAN	1.67616
T:MEAN=0	1.1932	PROB>!T!	0.277828
SGN RANK	5.5	PROB>!S!	0.294507
NUM ^= 0	6		
W:NORMAL	0.953084	PROB<W	0.728

QUANTILES(DEF=4)

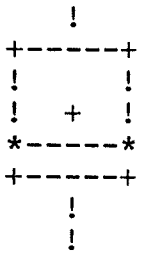
100%	MAX	8	99%	8
75%	Q3	6	95%	8
50%	MED	1	90%	8
25%	Q1	-2	10%	-4
0%	MIN	-4	5%	-4
			1%	-4
RANGE		12		
Q3-Q1		8		
MODE		-4		

EXTREMES

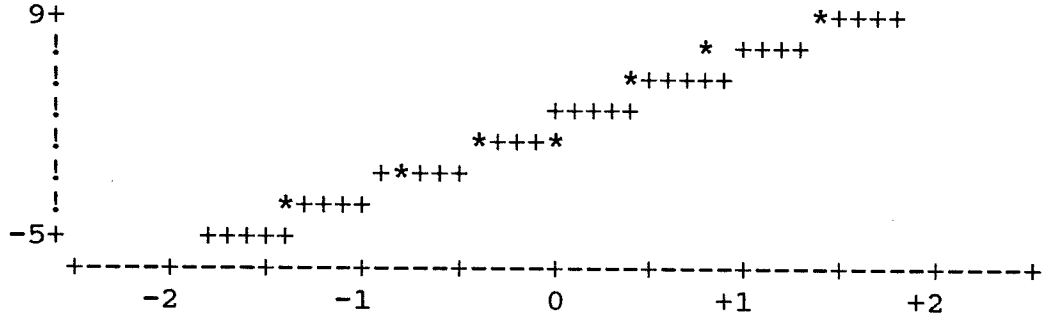
LOWEST	HIGHEST
-4	0
-2	1
0	5
1	6
5	8

STEM	LEAF	#
8	0	1
6	0	1
4	0	1
2		
0	0	1
-0	0	1
-2	0	1
-4	0	1

BOXPLOT



NORMAL PROBABILITY PLOT



```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
+  EXEMPLE : TEST SUR UNE MOYENNE AVEC VARIANCE INCONNUE      +
+                                                             +
+ PROCEDURE: MEANS                                           +
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

DATA EXEMPLE;
     INPUT X @@;
       Y = X - 1;
     LIST;CARDS;
     0.983 1.005 0.998 0.986 0.991
     1.002 0.996 0.983 0.994 1.002
;
PROC MEANS DATA=EXEMPLE N MEAN STD T PRT;
     VAR Y ;
     TITLE1 'DONNEES EXEMPLE 8.5';
     TITLE2 'TEST DE MOYENNE MU=1';
     TITLE3 'VARIANCE INCONNUE';
```

```
++++++++++++++++++++++++
+ OUTPUT DU PROGRAMME +
++++++++++++++++++++++++

  DONNEES EXEMPLE 8.5
  TEST DE MOYENNE MU=1
   VARIANCE INCONNUE
```

VARIABLE	N	MEAN	STANDARD DEVIATION	T	PR>!T!
Y	10	-0.00600000	0.00805536	-2.36	0.0429

```
++++++++++++++++++++++++++++++++++++++++++++++++++
+    EXEMPLE : TEST D'EGALITE DE DEUX MOYENNES  +
+              TEST D'EGALITE DE DEUX VARIANCES +
+              CAS D'ECHANTILLONS INDEPENDANTS  +
+  PROCEDURE : TTEST                            +
++++++++++++++++++++++++++++++++++++++++++++++++++
```

```
DATA EXEMPLE;
     INPUT TENSION @@;
       IF _N_ <= 25 THEN TYPE='ANCIEN';
       ELSE              TYPE='NOUVEAU';
     LIST;CARDS;
150 155 146 150 148 148 152 150 149 154 147 154 155 153 152
154 150 151 151 153 151 153 149 152 150
152 154 154 154 153 153 154 152 150 152 152 153 154 152 153
154 157 150 156 154 150 152 149 154 151
;
PROC TTEST DATA=EXEMPLE;
     CLASS TYPE;
     VAR TENSION;
     TITLE1 'DONNEES EXEMPLE 8.7';
     TITLE2 'COMPARAISON DE DEUX TYPES DE FILS';
     TITLE3 'TEST D''EGALITE DE DEUX MOYENNES';
     TITLE4 'TEST D''EGALITE DE DEUX VARIANCES';
```

```
++++++++++++++++++++++++
+ OUTPUT DU PROGRAMME +
++++++++++++++++++++++++
```

DONNEES EXEMPLE 8.7
COMPARAISON DE DEUX TYPES DE FILS
TEST D'EGALITE DE DEUX MOYENNES
TEST D'EGALITE DE DEUX VARIANCES

TTEST PROCEDURE

VARIABLE: TENSION

TYPE	N	MEAN	STD DEV	STD ERROR
ANCIEN	25	151.08000000	2.46508959	0.49301792
NOUVEA	25	152.76000000	1.87705443	0.37541089

VARIANCES	T	DF	PROB > !T!
UNEQUAL	-2.7111	44.8	0.0095
EQUAL	-2.7111	48.0	0.0093

FOR H0: VARIANCES ARE EQUAL, F'= 1.72 WITH 24 AND 24 DF
PROB > F'= 0.1891

```
++++++++++++++++++++++++++++++++++++++++++++++++++
+    EXEMPLE : TEST D'EGALITE DE DEUX MOYENNES  +
+              TEST D'EGALITE DE DEUX VARIANCES +
+              CAS D'ECHANTILLONS INDEPENDANTS  +
+  PROCEDURE : TTEST                            +
++++++++++++++++++++++++++++++++++++++++++++++++++
```

```
DATA EXEMPLE;
     INPUT FORCE @@;
       IF _N_ <= 4  THEN BETON='1';
       ELSE              BETON='2';
     LIST;CARDS;
284 311 290 280 318 318 312
;
PROC TTEST DATA=EXEMPLE;
     CLASS BETON;
     VAR FORCE;
     TITLE1 'DONNEES EXEMPLE 8.8';
     TITLE2 'COMPARAISON DE DEUX TYPES DE BETON';
     TITLE3 'TEST D''EGALITE DE DEUX MOYENNES';
     TITLE4 'TEST D''EGALITE DE DEUX VARIANCES';
```

```
++++++++++++++++++++++++
+ OUTPUT DU PROGRAMME +
++++++++++++++++++++++++
```

DONNEES EXEMPLE 8.8
COMPARAISON DE DEUX TYPES DE BETON
TEST D'EGALITE DE DEUX MOYENNES
TEST D'EGALITE DE DEUX VARIANCES

TTEST PROCEDURE

VARIABLE: FORCE

BETON	N	MEAN	STD DEV	STD ERROR
1	4	291.25000000	13.79311422	6.89655711
2	3	316.00000000	3.46410162	2.00000000

VARIANCES	T	DF	PROB > !T!
UNEQUAL	-3.4467	3.5	0.0337
EQUAL	-2.9712	5.0	0.0311

FOR H0: VARIANCES ARE EQUAL, F'= 15.85 WITH 3 AND 2 DF
PROB > F'= 0.1198

```
++++++++++++++++++++++++++++++++++++++++++++++++++
+  EXEMPLE :   TEST D'EGALITE DE DEUX MOYENNES +
+              CAS D'ECHANTILLONS PAIRES       +
+  PROCEDURE : MEANS                           +
++++++++++++++++++++++++++++++++++++++++++++++++++
```

```
DATA EXEMPLE;
     INPUT X Y @@;
        D = X - Y ;
     LIST;CARDS;
4.5 3.6 7.3 6.0 4.6 4.4 12.4 11.9 3.3 3.5 5.7 5.1
8.3 7.7 3.4 2.9 2.6 2.4 1.7 1.1
;
PROC MEANS DATA=EXEMPLE N MEAN STD T PRT;
     VAR D;
     TITLE1 'DONNEES EXEMPLE 8.9';
     TITLE2 'TEST D''EGALITE DE DEUX MOYENNES';
     TITLE3 'CAS D''ECHANTILLONS PAIRES';
```

```
++++++++++++++++++++++++
+ OUTPUT DU PROGRAMME +
++++++++++++++++++++++++
```

DONNEES EXEMPLE 8.9
TEST D'EGALITE DE DEUX MOYENNES
CAS D'ECHANTILLONS PAIRES

VARIABLE	N	MEAN	STANDARD DEVIATION	T	PR>!T!
D	10	0.52000000	0.40770360	4.03	0.0030

```
++++++++++++++++++++++++++++++++++++++++++++++
+  EXEMPLE  : TEST D'EGALITE DE PROPORTIONS +
+                                           +
+  PROCEDURE: FREQ                          +
++++++++++++++++++++++++++++++++++++++++++++++

DATA EXEMPLE;
     INPUT MAT $ ETAT $ NOBS @@;
     LIST;CARDS;
A DEFORME 41 B DEFORME 27 C DEFORME 22
A INTACT  79 B INTACT  53 C INTACT  78
;
PROC FREQ DATA=EXEMPLE ;
     WEIGHT NOBS;
     TABLES ETAT*MAT / EXPECTED CELLCHI2 CHISQ
                       NOROW NOCOL;
     TITLE1 'DONNEES EXEMPLE 8.11';
     TITLE2 'TEST D''EGALITE DE TROIS PROPORTIONS';
```

++++++++++++++++++++++++
+ OUTPUT DU PROGRAMME +
++++++++++++++++++++++++

DONNEES EXEMPLE 8.11
TEST D'EGALITE DE TROIS PROPORTIONS

TABLE OF ETAT BY MAT

ETAT MAT

FREQUENCY! EXPECTED ! CELL CHI2! PERCENT !	A	B	C	TOTAL
DEFORME	41	27	22	90
	36.0	24.0	30.0	
	.694444	0.375	2.13333	
	13.67	9.00	7.33	30.00
INTACT	79	53	78	210
	84.0	56.0	70.0	
	.297619	.160714	.914286	
	26.33	17.67	26.00	70.00
TOTAL	120	80	100	300
	40.00	26.67	33.33	100.00

STATISTICS FOR TABLE OF ETAT BY MAT

STATISTIC	DF	VALUE	PROB
CHI-SQUARE	2	4.575	0.101
LIKELIHOOD RATIO CHI-SQUARE	2	4.727	0.094
MANTEL-HAENSZEL CHI-SQUARE	1	3.668	0.055
PHI		0.123	
CONTINGENCY COEFFICIENT		0.123	
CRAMER'S V		0.123	

SAMPLE SIZE = 300

```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
+   EXEMPLE : TEST D'INDEPENDANCE DANS UN TABLEAU DE       +
+             CONTINGENCE                                  +
+  PROCEDURE: FREQ                                         +
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

DATA PANNES;
     INPUT EQUIPE $ MACHINE $ NOBS @@;
     LIST;CARDS;
         JOUR A 10 JOUR B 15 JOUR C 15 JOUR D 10
         SOIR A 10 SOIR B 20 SOIR C 15 SOIR D 20
         NUIT A 20 NUIT B 10 NUIT C 30 NUIT D 25
         ;
PROC FREQ DATA=PANNES;
     TABLES EQUIPE*MACHINE / EXPECTED DEVIATION CELLCHI2 CHISQ;
     WEIGHT NOBS;
     TITLE1 'TABLEAU DE CONTINGENCE';
     TITLE2 'TEST D''INDEPENDANCE';
```

++++++++++++++++++++++++
+ OUTPUT DE PROC FREQ +
++++++++++++++++++++++++

TABLEAU DE CONTINGENCE
TEST D'INDEPENDANCE

TABLE OF EQUIPE BY MACHINE

EQUIPE / MACHINE FREQUENCY EXPECTED DEVIATION CELL CHI2 PERCENT ROW PCT COL PCT	A	B	C	D	TOTAL
JOUR	10	15	15	10	50
	10.0	11.3	15.0	13.8	
	0.0	3.8	0.0	-3.8	
	0	1.25	0	1.02273	
	5.00	7.50	7.50	5.00	25.00
	20.00	30.00	30.00	20.00	
	25.00	33.33	25.00	18.18	
NUIT	20	10	30	25	85
	17.0	19.1	25.5	23.4	
	3.0	-9.1	4.5	1.6	
	.529412	4.35376	.794118	.112968	
	10.00	5.00	15.00	12.50	42.50
	23.53	11.76	35.29	29.41	
	50.00	22.22	50.00	45.45	
SOIR	10	20	15	20	65
	13.0	14.6	19.5	17.9	
	-3.0	5.4	-4.5	2.1	
	.692308	1.97543	1.03846	.252622	
	5.00	10.00	7.50	10.00	32.50
	15.38	30.77	23.08	30.77	
	25.00	44.44	25.00	36.36	
TOTAL	40	45	60	55	200
	20.00	22.50	30.00	27.50	100.00

STATISTICS FOR TABLE OF EQUIPE BY MACHINE

STATISTIC	DF	VALUE	PROB
CHI-SQUARE	6	12.022	0.061
LIKELIHOOD RATIO CHI-SQUARE	6	12.800	0.046
MANTEL-HAENSZEL CHI-SQUARE	1	0.780	0.377
PHI		0.245	
CONTINGENCY COEFFICIENT		0.238	
CRAMER'S V		0.173	

SAMPLE SIZE = 200

```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++
+  EXEMPLE    : TEST D'EGALITE DE PLUSIEURS MOYENNES    +
+               AVEC UN MODELE DE CLASSIFICATION SIMPLE +
+  PROCEDURE : ANOVA                                     +
++++++++++++++++++++++++++++++++++++++++++++++++++++++++


DATA EXEMPLE;
     INPUT HUILE PERTE @@;
     LIST;CARDS;
1  1.641  1  1.306  1  1.149  1  1.025
2  1.782  2  1.568  2  1.223  2  1.919
3  1.570  3  1.240  3  1.068  3  1.982
4  1.493  4  1.415  4  1.118  4  1.812
5  1.672  5  1.291  5  1.004  5  2.015
;
PROC ANOVA DATA=EXEMPLE ;
     CLASS HUILE;
     MODEL PERTE = HUILE ;
     TITLE1 'DONNEES EXEMPLE 8.17';
     TITLE2 'TEST D''EGALITE DE PLUSIEURS MOYENNES';

                        ++++++++++++++++++++++
                        + OUTPUT DU PROGRAMME +
                        ++++++++++++++++++++++

                          DONNEES EXEMPLE 8.17
                   TEST D'EGALITE DE PLUSIEURS MOYENNES

                      ANALYSIS OF VARIANCE PROCEDURE

                         CLASS LEVEL INFORMATION

                      CLASS      LEVELS     VALUES

                      HUILE           5     1 2 3 4 5

                  NUMBER OF OBSERVATIONS IN DATA SET = 20

                      ANALYSIS OF VARIANCE PROCEDURE
```

DEPENDENT VARIABLE: PERTE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
MODEL	4	0.24022580	0.06005645
ERROR	15	1.80405875	0.12027058
CORRECTED TOTAL	19	2.04428455	
MODEL F =	0.50		PR > F = 0.7367

R-SQUARE	C.V.	ROOT MSE	PERTE MEAN
0.117511	23.6780	0.34680050	1.46465000

SOURCE	DF	ANOVA SS	F VALUE	PR > F
HUILE	4	0.24022580	0.50	0.7367

CHAPITRE 9: L'ANALYSE DE RÉGRESSION

CHAPITRE 9

L'ANALYSE DE RÉGRESSION

9.0 SOMMAIRE

L'analyse de régression fait l'ajustement de modèles statistiques à des observations sur plusieurs variables quantitatives. Les principaux résultats concernant l'estimation des paramètres, l'exécution des tests statistiques, l'analyse des résidus et les différents problèmes rencontrés dans cette analyse sont présentés. Les différentes procédures du système SAS dédiées à la régression sont illustrées et commentées à l'aide d'exemples.

9.1 MÉTHODE

Une analyse de régression typique contient les étapes suivantes:

L'identification des variables:

- **une variable à expliquer notée Y;**
- **des variables explicatives notées X_1, X_2,•••, X_k**

  les variables sont généralement continues mais il suffit que les variables soient numériques pour effectuer les calculs.

La proposition d'un modèle de régression de la forme:

$$Y = \varphi(X_1, X_2, \ldots, X_k; \beta_0, \beta_1, \ldots, \beta_p) + \varepsilon \qquad (9.1)$$

où ϕ est la fonction de régression définie sur les variables explicatives

ε est un terme d'erreur représentant l'ensemble des effets sur Y

non explicitement identifiés par les variables explicatives

$\beta_0, \beta_1, \ldots, \beta_p$

sont des paramètres statistiques dont les valeurs seront déterminées par calcul

avec le principe de moindres carrés

- La spécification des <u>hypothèses de base:</u>

  la fonction φ est de forme connue et très souvent on la suppose linéaire dans les paramètres $\beta_0, \beta_1 \ldots, \beta_p$ qui eux sont inconnus;

  le terme d'erreur suit une distribution gaussienne $N(0,\sigma^2)$ où σ^2 est un paramètre inconnu.

- La collecte d'observations conjointes $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$ $i=1,2,\ldots,n$ sur l'ensemble des variables.

- L'estimation des paramètres du modèle par le principe des moindres carrés ou autre.

- L'analyse de l'adéquation du modèle à l'aide de tests statistiques, des valeurs prédites et l'analyse des résidus.

- L'examen de problèmes particuliers comme l'identification des points influents, la détection de variables explicatives colinéaires et l'exploration de d'autres modèles par les méthodes de sélection de variables.

<u>Remarques générales</u>

- On n'est pas toujours certain d'avoir correctement identifié l'ensemble des variables explicatives qui ont présumément un effet sur la variable à expliquer; toutes les variables qui ne sont pas identifiées comme telles sont alors confondues avec le terme d'erreur.

- La spécification de la fonction de régression est cruciale: si elle est incorrectement spécifiée, l'analyse risque de donner de piètres résultats; dans certaines applications, l'état des connaissances sur le phénomène est suffisament avancé et la forme de la fonction est connue. D'autre part, l'analyse de régression est souvent employée pour faire avancer l'état des connaissances sur le phénomène. On fait alors souvent l'hypothèse d'un modèle linéaire dans les paramètres, ce qui constitue une hypothèse de travail plausible sous certaines conditions.

- Le terme d'erreur est un postulat de l'impossibilité de prédire parfaitement la variable Y avec toutes les variables explicatives; cela tient compte du fait que l'identification de celles-ci n'est pas toujours facile à faire et aussi qu'il peut y avoir des erreurs de mesures dans les variables.

- L'hypothèse d'une distribution normale pour le terme d'erreur est nécessaire lors de l'exécution des tests statistiques et le calcul d'intervalles de confiance pour les paramètres; ces formules sont développées à l'aide de cette hypothèse de base.

- Le nombre de paramètres est généralement supérieur au nombre de variables explicatives: $p \geq k$.

- Le nombre d'observations n doit être plus grand que le nombre de paramètres: p+1 paramètres de tendance centrale $\beta_0,\ldots,\beta_p$ et un paramètre de dispersion σ^2. Il n'y a pas de règle bien établie concernant le nombre d'observations n par rapport au nombre de paramètres mais la règle $n > 2p$ semble conservatrice.

- L'estimation des paramètres $\beta_0,\beta_1,\ldots,\beta_p$ est généralement faite par la méthode des moindres carrés. Cette méthode peut se justifier comme principe dérivant de la méthode de vraisemblance maximale lorsque le terme d'erreur est normalement distribué. (Voir propriétés des moindres carrés au chapitre 7.) Toutefois la méthode des moindres carrés n'est pas résistante lorsque les observations contiennent des données aberrantes et celles-ci ne sont pas toujours faciles à identifier s'il y a un grand nombre de variables explicatives. Plusieurs autres techniques d'estimation ont été proposées afin de remplacer la méthode des moindres carrés lorsque cette dernière fait défaut. Ces méthodes appelées robustes sont équivalentes à la méthode des moindres carrés pondérés.

- L'examen des résidus est une étape essentielle de l'analyse de régression car cela permet de s'assurer a posteriori, de l'indépendance des observations, de la normalité du terme d'erreur ainsi que la constance du paramètre de dispersion σ^2.

- L'analyse de régression n'est pas toujours couronnée de succès et cela peut s'expliquer par plusieurs raisons reliées à la violation des hypothèses de base.

Classification des modèles de régression

Les modèles de régression sont classés selon le nombre de variables explicatives, le caractère de linéarité ou non de la fonction φ relativement aux paramètres ainsi que la forme de cette dernière.

Le modèle est dit SIMPLE s'il contient une seule variable explicative (k=1) et MULTIPLE s'il contient plusieurs variables explicatives ($k \geq 2$).

Le modèle est dit LINÉAIRE DANS LES PARAMÈTRES si la fonction φ peut se décomposer de la manière suivante:

$$\varphi(X_1,\ldots,X_k,\beta_0\ldots,\beta_p) = \sum_{j=0}^{p} \beta_j \varphi_j(X_1,X_2,\ldots,X_k) \qquad (9.2)$$

L'équation (9.1) peut alors s'écrire:

$$Y = \sum_{j=0}^{p} \beta_j Z_j + \varepsilon \qquad (9.3)$$

où

$$Z_j = \varphi_j(X_1,X_2,\ldots,X_k) \qquad (9.4)$$

sont de nouvelles variables définies à l'aide des variables $X_1,X_2,\ldots,X_k$ et des fonctions φ_j ne contenant aucun paramètre inconnu. L'avantage fondamental d'avoir un modèle linéaire est qu'il en résulte un système d'équations linéaires lors de l'étape de l'estimation des paramètres par le principe des moindres carrés.

Il est important de noter que le caractère de linéarité est relatif aux paramètres et non pas aux variables explicatives. La démarcation entre modèles linéaires et ceux qui ne le sont pas n'est pas toujours évidente. En effet, il existe des modèles qui ne sont pas linéaires à première vue, mais qui le deviennent après une transformation sur les variables. D'ailleurs l'exploration de transformations sur les variables fait partie de la démarche d'une analyse de régression. La transformation de variables est souvent une solution à plusieurs problèmes rencontrés dans ce type d'analyse: non linéarité, violation des hypothèses de base, recherche de meilleurs modèles etc.

Notons que si on transforme la variable Y ainsi que la fonction φ dans l'équation (9.1), le terme d'erreur est aussi transformé. Toutefois on réfère toujours au terme d'erreur comme étant la partie additive non expliquée de tout modèle de régression même s'il y a eu transformation des variables pour une raison ou une autre.

Si les variables explicatives interviennent ainsi que leurs puissances entières on dit que le modèle est POLYNOMIAL; la plus grande puissance rencontrée s'appelle le DEGRE du modèle.

Voici plusieurs modèles de régression souvent rencontrés:

- modèle sans variable explicative:

$$Y = \beta_0 + \varepsilon$$

- modèle de régression linéaire simple:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- modèle de régression linéaire simple quadratique:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

- modèle de régression linéaire polynomial de degré k:

$$Y = \beta_0 + \beta_1 X + .., + \beta_k X^k + \varepsilon$$

- modèle de régression linéaire multiple:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...,+ \beta_p X_p + \varepsilon$$

- modèle de régression linéaire quadratique à deux variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \varepsilon$$

- exemples de modèles de régression non-linéaires et qu'on peut transformer en modèle linéaire:

$$Y = \beta_0 \exp(\beta_1 X) + \varepsilon$$

$$Y = \frac{1}{\beta_0 + \beta_1 X + \varepsilon}$$

$$Y = \prod_{\alpha=0}^{p} \exp(\beta_\alpha X_\alpha)$$

. exemples de modèles non-linéaires et qu'on ne peut pas transformer en modèle linéaire:

$$Y = \beta_0 + \beta_1 \exp(\beta_2 X) + \varepsilon$$

$$Y = \beta_1 \ln(\beta_2 X + \beta_3) + \varepsilon$$

$$\ln Y = \beta_0 + \beta_1 \ln (\beta_2 X_1^{\beta 3} + (1-\beta_2) X_2^{\beta 3}) + \varepsilon$$

$$Y = \beta_0 \Phi (\beta_1 + \beta_2 x) + \varepsilon$$

$$Y = \begin{cases} \beta_0 + \beta_1 x + \beta_2 x^2 & \text{si } x < \beta_3 \\ \beta_4 & \text{si } x < \beta_3 \end{cases}$$

9.2 RÉGRESSION LINÉAIRE SIMPLE

Le modèle s'écrit

$$Y = \beta_0 + \beta_1 X + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2) \qquad (9.5)$$

$Y|_{X=x} = \beta_0 + \beta_1 x$

D'une manière équivalente $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$; rappelons que l'hypothèse de normalité n'est vraiment utile qu'à l'étape des tests et dans le calcul des intervalles de confiance.

Estimation des paramètres β_0 et β_1

Le principe des moindres carrés consiste à rechercher le minimum de la fonction auxiliaire $S(\beta_0, \beta_1)$ définie par

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 \qquad (9.6)$$

où (x_i, y_i) $i = 1,2,\ldots,n$ est l'ensemble des observations. Une condition nécessaire pour l'obtention du minimum est l'annulation des dérivées partielles de S par rapport aux paramètres β_0 et β_1. On obtient, après arrangement, le système d'équations linéaires d'inconnues β_0 et β_1:

$$\begin{aligned} n\beta_0 + \left[\sum_{i=1}^{n} x_i\right]\beta_1 &= \sum_{i=1}^{n} y_i \\ \left[\sum_{i=1}^{n} x_i\right]\beta_0 + \left[\sum_{i=1}^{n} x_i^2\right]\beta_1 &= \sum_{i=1}^{n} y_i x_i \end{aligned} \qquad (9.7)$$

La solution du système d'équations (9.7) notée $\hat{\beta}_0$ et $\hat{\beta}_1$ est

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned} \qquad (9.8)$$

Il y a plusieurs façons d'exprimer $\hat{\beta}_0$ et $\hat{\beta}_1$, en particulier comme combinaisons linéaires, ce qui est utile pour déterminer les distributions d'échantillonnage de ces estimateurs:

$$\begin{aligned} \hat{\beta}_1 &= \sum_{i=1}^{n} c_i y_i \\ \hat{\beta}_0 &= \sum_{i=1}^{n} d_i y_i \end{aligned} \qquad (9.9)$$

où

$$c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$d_i = 1/n - c_i \bar{x} \qquad (9.10)$$

(voir exercice 2.2)

L'équation de prédiction est:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$= \bar{y} + \hat{\beta}_1 (x - \bar{x}) \qquad (9.11)$$

Estimation du paramètre de dispersion σ^2

La méthode d'estimation des moments suggère la formule suivante pour l'estimation du paramètre de dispersion σ^2:

$$\hat{\sigma}^2 = \frac{1}{n-s} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (9.12)$$

où $\hat{y}_i = \hat{E}(Y_i)$ est l'estimation de la moyenne $E(Y_i)$ aux valeurs $x_{i1}, x_{i2}, \ldots, x_{in}$ des variables explicatives et s est le nombre de paramètres estimés dans $E(Y_i)$. Le fait de diviser par n - s plutôt que n permet d'obtenir une estimation sans biais.

Appliquons la formule (9.12) au cas du modèle de régression linéaire simple. La valeur de prédiction

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{9.13}$$

constitue une estimation sans biais de la moyenne (à démontrer). L'expression (9.12) peut alors se mettre sous la forme:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{9.14}$$

Définissons la somme des carrés SCE par le minimum atteint de la fonction auxiliaire $S(\beta_0, \beta_1)$:

$$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = S(\hat{\beta}_0, \hat{\beta}_1) \tag{9.15}$$

L'équation (9.14) devient:

$$\hat{\sigma}^2 = \frac{SCE}{n-2} \tag{9.16}$$

Analyse de la variance

L'analyse de la variance est une technique par laquelle on décompose la variation totale SCT définie par:

$$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{9.17}$$

selon une partie expliquée par le modèle $(\beta_0+\beta_1 x)$ et une partie attribuable au terme d'erreur (ε). On a

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

$$= (\hat{\beta}_0+\hat{\beta}_1 x_i - \bar{y}) + (y_i - \hat{y}_i)$$

$$= \hat{\beta}_1 (x_i - \bar{x}) + (y_i - \hat{y}_i) \quad \text{en vertu de l'équation (9.8).}$$

En élevant ~~Lorsqu'un élève~~ au carré et en additionnant ~~on additionne~~ pour toutes les observations

$$SCT = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$+ 2\,\hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \hat{y}_i)$$

Mais $$\hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \hat{y}_i) = \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x}))$$

$$= \hat{\beta}_1 \left[\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]$$

$$= 0$$

en vertu de la solution des équations de moindres carrés (9.8). Définissons la somme des carrés attribuable au modèle SCM par:

$$SCM = \hat{\beta}_1^2 \left| \sum_{i=1}^{n} (x_i - \bar{x})^2 \right| = \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \qquad (9.18)$$

L'équation fondamentale d'analyse de la variance du modèle de régression linéaire simple est donc

$$SCT = SCM + SCE \qquad (9.19)$$

où SCT est définie par (9.17), SCM est définie par (9.18) et SCE est définie par (9.15).

Distributions d'échantillonnage

Les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_0$ peuvent s'exprimer comme des combinaisons linéaires:

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i \quad \text{où} \quad c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{9.24}$$

$$\hat{\beta}_0 = \sum_{i=1}^{n} d_i Y_i \quad \text{où} \quad d_i = \frac{1}{n} - \bar{x} c_i \tag{9.25}$$

On montre facilement les résultats suivants concernant les coefficients c_i et d_i (exercice 2.2):

$$\sum_{i=1}^{n} c_i = 0, \qquad \sum_{i=1}^{n} c_i^2 = \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{1}{a^2}$$

$$\sum_{i=1}^{n} d_i = 1, \qquad \sum_{i=1}^{n} d_i^2 = \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{1}{n} + \frac{\bar{x}^2}{a^2} \tag{9.26}$$

Puisque les variables Y_i sont normalement distribuées

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

on obtient, en utilisant les formules concernant les combinaisons linéaires de variables normales, les résultats:

Cette quantité mesure la proportion de variation totale en y expliquée par la variable x.

La deuxième quantité est le rapport F défini par

$$F = \frac{CMM}{CME} = (n-2)\, R^2/(1-R^2) \qquad (9.21)$$

qui sert pour effectuer un test d'hypothèse sur la nullité du coefficient de régression β_1.

Test d'hypothèse

On est particulièrement intéressé à tester l'hypothèse nulle

$$H_N: \beta_1 = 0 \qquad (9.22)$$

dans le contexte du modèle de régression linéaire simple car cela permet de décider si la variable x a un pouvoir explicatif vis-à-vis de la variable y. Sous l'hypothèse de base d'une distribution gaussienne pour les erreurs ε_i on montre que:

$$\frac{(n-2)\,CME}{\sigma^2} \sim \chi^2_{n-2}$$

$$\frac{CMM}{\sigma^2} \sim \chi^2_1 \quad \text{si } H_N \text{ est vraie} \qquad (9.23)$$

et que les carrés moyens CMM et CME sont indépendants. Il s'ensuit donc que $F \sim F_{1,n-2}$ si H_N est vraie et on a le test suivant:

on rejette H_N au seuil α si $F > F_{1,n-2,\alpha}$

où F est définie par (9.21)

Distributions d'échantillonnage

Les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_0$ peuvent s'exprimer comme des combinaisons linéaires:

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i \quad \text{où} \quad c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{9.24}$$

$$\hat{\beta}_0 = \sum_{i=1}^{n} d_i Y_i \quad \text{où} \quad d_i = \frac{1}{n} - \bar{x} c_i \tag{9.25}$$

On montre facilement les résultats suivants concernant les coefficients c_i et d_i (exercice 2.2):

$$\sum_{i=1}^{n} c_i = 0, \quad \sum_{i=1}^{n} c_i^2 = \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{1}{a^2}$$

$$\sum_{i=1}^{n} d_i = 1, \quad \sum_{i=1}^{n} d_i^2 = \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{1}{n} + \frac{\bar{x}^2}{a^2} \tag{9.26}$$

Puisque les variables Y_i sont normalement distribuées

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

on obtient, en utilisant les formules concernant les combinaisons linéaires de variables normales, les résultats:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{a^2}\right)$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma/a} \sim N(0,1) \qquad (9.27)$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{a^2}\right]\right)$$

$$\frac{\hat{\beta}_0 - \beta_0}{\sigma\left[\frac{1}{n} + \frac{x^2}{a^2}\right]^{1/2}} \sim N(0,1) \qquad (9.28)$$

De plus

$$\frac{(\hat{\beta}_0+\hat{\beta}_1 x) - (\beta_0+\beta_1 x)}{\sigma\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{a^2}\right]^{1/2}} \sim N(0,1) \qquad (9.29)$$

En général, le paramètre σ est inconnu et il est estimé par $\hat{\sigma}$. Si on remplace σ par $\hat{\sigma}$ dans (9.27), (9.28) et (9.29) on obtient:

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/a} \sim T_{n-2} \qquad (9.30)$$

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\left[\frac{1}{n} + \frac{\bar{x}^2}{a^2}\right]^{1/2}} \sim T_{n-2} \qquad (9.31)$$

$$\frac{(\hat{\beta}_0+\hat{\beta}_1 x)-(\beta_0+\beta_1 x)}{\hat{\sigma}\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{a^2}\right]^{1/2}} \sim T_{n-2} \qquad (9.32)$$

Ces résultats proviennent du fait que la quantité $\frac{(n-2)\hat{\sigma}^2}{\sigma^2}$ suit une distribution khi-deux avec (n-2) degré de liberté et de la définition d'une variable de Student.

Intervalles de confiance pour les paramètres

Les résultats (9.30), (9.31) et (9.32) nous permettent d'obtenir les intervalles de confiance pour les paramètres β_1, β_0, $\beta_0+\beta_1 x$

$$\beta_1 : \hat{\beta}_1 \pm t_{n-2,\alpha/2}\ \hat{\sigma}/a \,, \qquad (9.33)$$

$$\beta_0 : \hat{\beta}_0 \pm t_{n-2,\alpha/2}\ \hat{\sigma}\left[\frac{1}{n} + \frac{\bar{x}^2}{a^2}\right]^{1/2} \qquad (9.34)$$

$$\beta_0+\beta_1 x: \quad \hat{\beta}_0+\hat{\beta}_1 x \pm t_{n-2,\alpha/2} \; \hat{\sigma}\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{a^2} \right]^{1/2} \qquad (9.35)$$

où

$$a^2 = \sum_{i=1}^{n} (x_i-\bar{x})^2$$

Notons que (9.34) est un cas particulier de (9.35) avec x=0.

Intervalle de prédiction pour la variable Y

Il est utile de calculer des intervalles de prédiction pour la variable Y pour une valeur donnée de x. On prédit Y par

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x + \varepsilon$$

de moyenne

$$E(\hat{Y}) = \beta_0 + \beta_1 x$$

et variance

$$VAR(\hat{Y}) = VAR(\hat{\beta}_0 + \hat{\beta}_1 x) + VAR(\varepsilon)$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{a^2} \right] + \sigma^2$$

$$= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{a^2} \right]$$

On en déduit le résultat suivant concernant l'intervalle de prédiction pour Y à x:

$$Y: \hat{\beta}_0+\hat{\beta}_1 x \pm t_{n-2,\alpha/2}\,\hat{\sigma}\left[1+\frac{1}{n}+\frac{(x-\bar{x})^2}{a^2}\right]^{1/2} \qquad (9.36)$$

Exemple 9.1: données sur la consommation d'eau du chapitre 1

Les données sont

POP(x)	5.682	11.338	16.380	25.462	36.464	40.922
CONS(y)	2.44	6.14	10.32	21.39	39.53	46.04

où POP représente le nombre d'habitants en milliers

CONS représente la consommation en millions de mètres cubes.

On propose d'expliquer la consommation Y à l'aide du modèle

$$CONS = \beta_o + \beta_1 * POP + \varepsilon$$

Tous les calculs sont basés sur les sommes

$$n=6 \qquad \sum_{i=1}^{n} x_i = 136.248 \qquad \sum_{i=1}^{n} y_i = 125.86$$

$$\sum_{i=1}^{n} x_i^2 = 4081.6866 \qquad \sum_{i=1}^{n} y_i^2 = 4289.9902 \qquad \sum_{i=1}^{n} x_i y_i = 4122.6240$$

Le système d'équations à résoudre est alors:

$$6\,\beta_0 + 136.248\,\beta_1 = 125.86$$

$$136.248\,\beta_0 + 4081.6866\,\beta_1 = 4122.6240$$

dont la solution $\hat{\beta}_0$ et $\hat{\beta}_1$ peut se calculer par

$$\hat{\beta}_1 = \frac{\Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n}}{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}$$

$$= \frac{4122.6240 - \frac{136.248 * 125.86}{6}}{4081.6866 - \frac{(136.248)^2}{6}}$$

$$= 1.28$$

$$\hat{\beta}_0 = \bar{y} - 1.28\ \bar{x} = -8.095$$

L'équation de prédiction est donc

$$\hat{Y} = -8.095 + 1.28\ X$$

Les calculs des sommes de carrés du tableau d'analyse de la variance sont généralement effectués à l'aide des identités algébriques:

$$SCT = \sum_{i=1}^{n} y_i^2 - \frac{\left[\sum_{i=1}^{n} y_i\right]^2}{n} = 1649.867$$

$$SCM = \hat{\beta}_1 \left[\sum_{i=1}^{n} x_i y_i - \frac{\left[\sum_{i=1}^{n} x_i\right]\left[\sum_{i=1}^{n} y_i\right]}{n} \right] = 1619.006$$

$$SCE = SCT - SCM = 30.861$$

Tableau d'analyse de la variance

Source	Somme de carrés	Degrés de liberté	Carrés moyens	F
modèle	1619.006	1	1619.006	209.846
erreur	30.861	4	7.715	-
totale	1649.867	5	-	-

On rejette l'hypothèse de la nullité du coefficient de régression puisque $F_{1,4,0.005} = 31.3$. Globalement le modèle linéaire semble satisfaisant avec un coefficient de détermination (R^2) de 0.98, ce qui est assez élevé.

Les intervalles de confiance à 95% pour les coefficients β_0 et β_1 du modèle sont:

$$\beta_1: \quad 1.28 \pm 2.776 * 2.77 * 0.032$$

$$1.28 \pm 0.25$$

$$\beta_0: \quad -8.095 \pm 2.776 * 2.305 * 0.83$$

$$-8.095 \pm 0.53$$

Les valeurs observées, prédites, résiduelles, les intervalles de confiance à 95% pour la moyenne de Y ainsi que l'intervalle de prédiction à 95% pour Y sont donnés dans le tableau 9.2.

Tableau 9.2

i	observée	prédite	résiduelle	int. confiance moyenne de Y		int.prédiction pour Y	
1	2.44	-0.82	3.26	-6.05	4.41	-10.14	8.50
2	6.14	6.42	-0.28	2.21	10.63	-2.36	15.20
3	10.32	12.875	-2.55	9.36	10.38	4.40	21.35
4	21.39	24.50	-3.11	21.28	27.72	16.14	32.86
5	39.53	38.59	0.94	33.97	43.20	29.60	47.57
6	46.04	44.29	1.74	38.83	49.76	34.84	53.75

Analyse des résidus

L'analyse de tout modèle statistique de la forme:

$$Y_i = \varphi(x_{i1}, x_{i2}, \ldots, x_{ik}; \beta_0, \ldots, \beta_1) + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2) \qquad i=1,2,\ldots n$$

repose sur les hypothèses de base suivante:

- la fonction φ est correctement spécifiée;
- les variables explicatives $X_1 \ldots X_k$ sont non-stochastiques et mesurées sans erreur;
- les erreurs ε_i sont non-corrélées à moyenne 0 et à variance constante;
- les erreurs ε_i sont distribuées selon une loi normale.

Il est difficile de s'assurer de la validité de ces hypothèses avant d'effectuer l'estimation des paramètres et de calculer le tableau d'analyse de la variance. À la suite de cette étape et si le modèle développé semble satisfaisant (R^2 élevé, test F significatif), on fait l'analyse des résidus pour s'assurer si certaines hypothèses de base ne sont pas violées.

On définit les quantités:

$$r_i = y_i - \hat{y}_i \qquad \text{: le résidu} \qquad (9.37)$$

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \qquad \text{: le "levier"} \qquad (9.38)$$

l'estimation de l'écart-type du résidu r_i

$$\hat{ET}(r_i) = \hat{\sigma} \ (1 - h_i)^{1/2} \qquad (9.39)$$

le résidu "studentisé" rs_i :

$$rs_i = \frac{r_i}{\hat{ET}(r_i)} \qquad (9.40)$$

La quantité h_i est applicable pour le cas du modèle linéaire simple seulement. C'est une mesure de l'importance du point x_i dans la détermination de la droite de moindres carrés. Dans le cas du modèle linéaire multiple, l'expression de h_i est différente.

L'analyse des résidus permet de "vérifier" a posteriori les hypothèses de base et elle est faite à l'aide des graphiques suivants:

. Le graphique des résidus studentisés $rs_{(i)}$ ordonnés sur une échelle de probabilité gaussienne. On trace les points:

$$\left[rs_{(i)}, \ \Phi^{-1}\left(\frac{i-0.375}{n+0.250}\right) \right] \qquad i=1,2,\ldots,n$$

Φ^{-1} étant la fonction réciproque gaussienne.

L'allure des points devrait présenter l'aspect d'une ligne droite illustrée à la figure 9.1(a). Ce graphique permet de s'assurer de l'hypothèse de normalité nécessaire pour l'exécution des tests et le calcul des intervalles de confiance.

Remarque: Certains auteurs utilisent (i-0.5)/n ou (i/(n+1)) pour tracer ce graphique.

. L'analyse des "grands" résidus: l'hypothèse de normalité nous permet d'affirmer que:

$$|rs_i| \leq 2 \quad \text{pour 95\% des points}$$

$$|rs_i| > 3 \quad \text{pour 0.3\% des points}$$

Les grands résidus peuvent s'expliquer par la juxtaposition d'une ou plusieurs des raisons suivantes:

.. distribution normale non vérifiée;

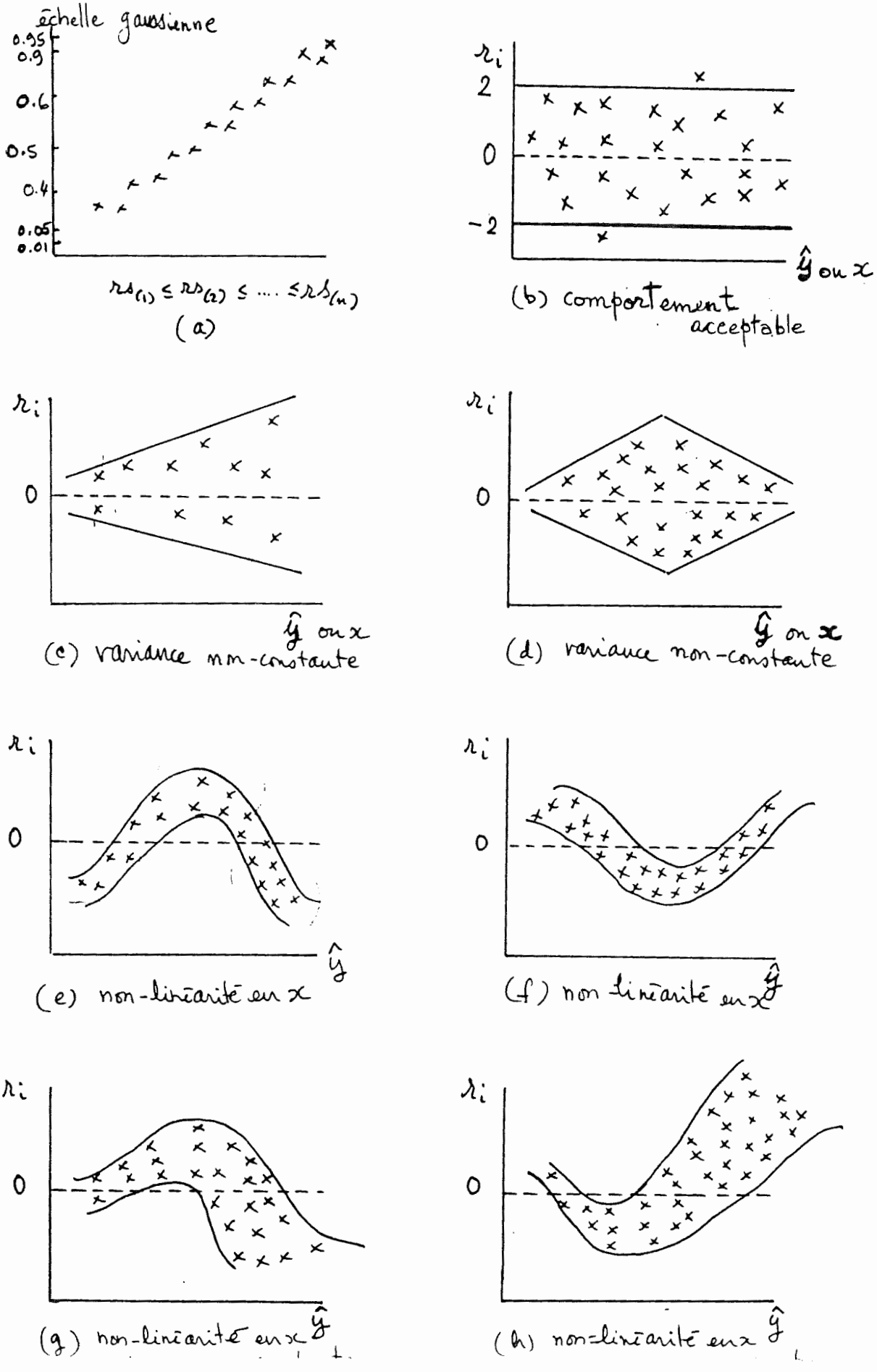
.. valeur aberrante du point (x_i, y_i);

.. fonction φ incorrectement spécifiée.

. Le graphique des résidus r_i ou rs_i en fonction de la prédiction $\hat{y}_i$ permet de vérifier si la fonction φ est correctement spécifiée. L'allure du graphique doit présenter l'aspect d'une bande horizontale tel qu'illustré sur la figure 9.1(b). Les autres cas illustrés à la figure 9.1(c)-(h) sont des indications de non-linéarité en x et/ou de variance non constante.

. Le graphique des résidus r_i ou rs_i en fonction de la variable explicative x. Ce graphique permet de vérifier si l'hypothèse d'une variance constante est plausible et si la fonction φ est correctement spécifiée. Les figures 9.1(b)-(c)-(d) correspondent à des anomalies.

Figure 9.1: Analyse graphique des résidus



Exemple 9.2: suite de l'exemple 9.1

Tableau 9.3

i	r_i	h_i	$ET(r_i)$	rs_i
1	3.26	0.4601	2.04	1.60
2	-0.28	0.2975	2.33	-0.12
3	-2.55	0.2072	2.47	-1.03
4	-3.11	0.1743	2.52	-1.23
5	0.94	0.3582	2.22	0.42
6	1.74	0.5025	1.96	0.89

Nous avons tracé les graphiques suivants:

. figure 9.2a: les points expérimentaux et la droite de moindres carrés;

. figure 9.2b les résidus en fonction de la variable explicative;

. figure 9.2c: les résidus en fonction de la valeur prédite $\hat{y}$;

. figure 9.2d: les résidus studentisés sur échelle de probabilité gaussienne.

Un examen de ces graphiques indique que:

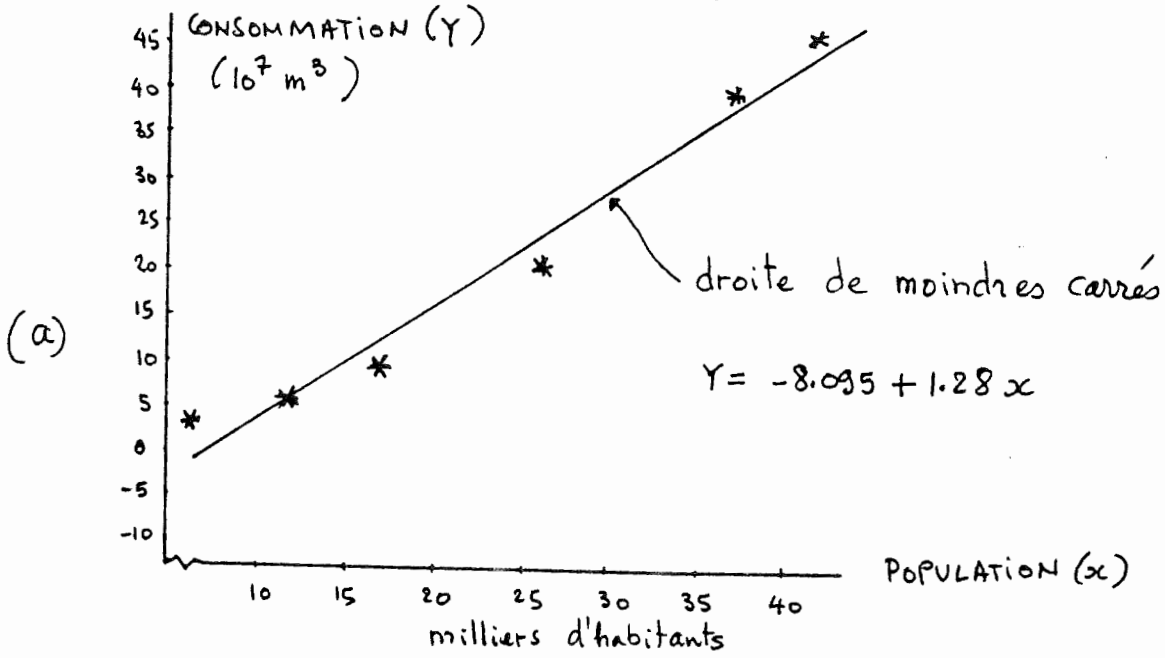
. le modèle semble satisfaisant avec un R^2 de 0.98;

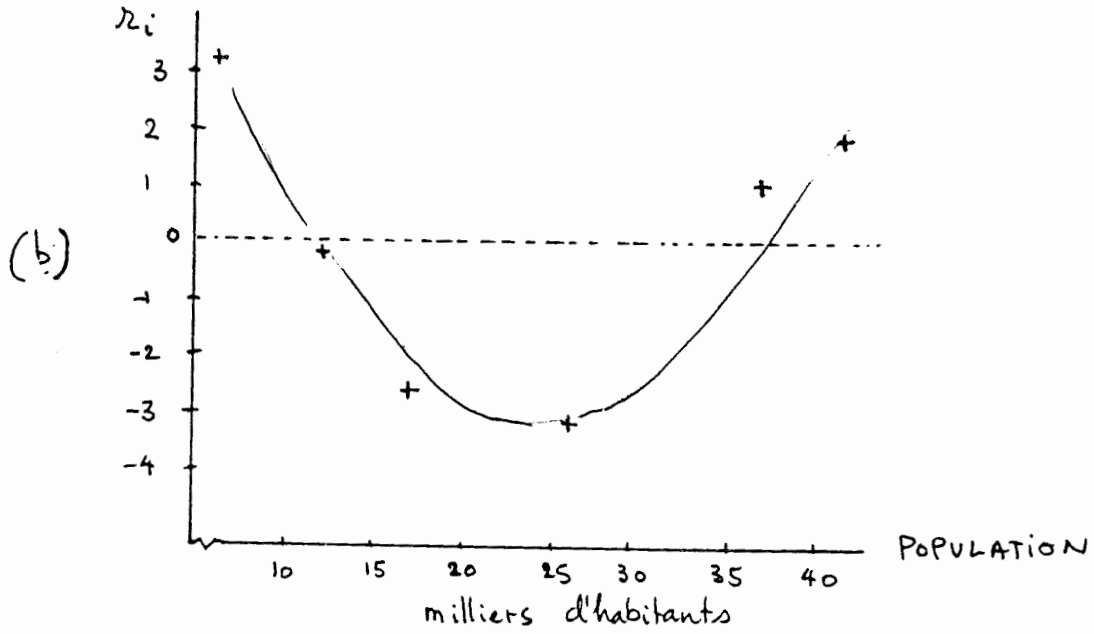
. l'hypothèse de normalité est vérifiée;

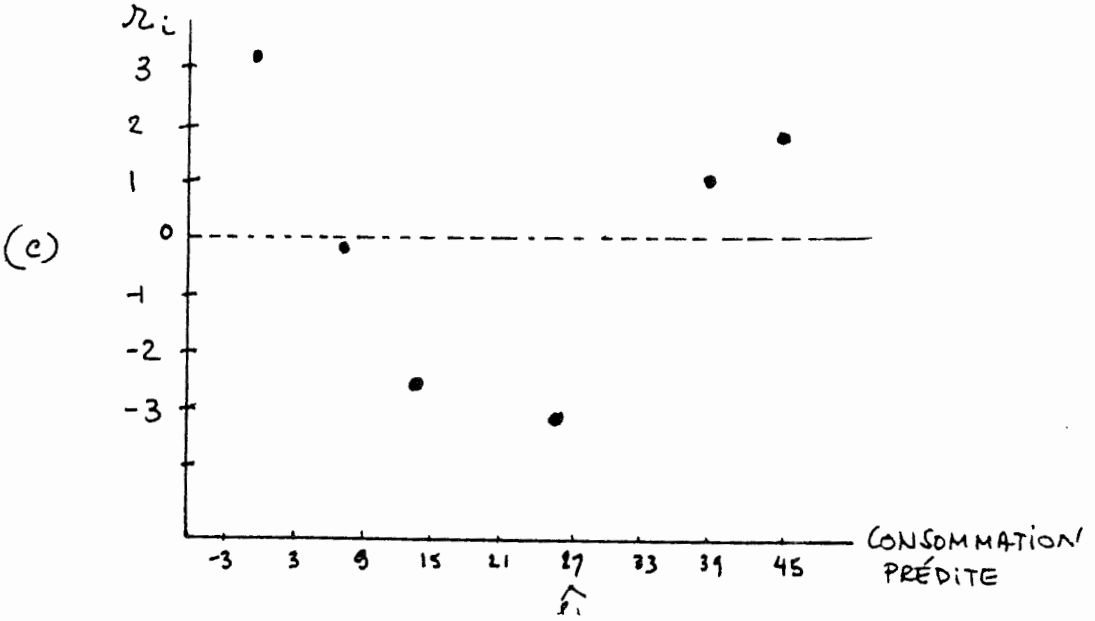
. la variance σ^2 est constante;

. il semble que l'on pourrait améliorer le modèle légèrement en utilisant un polynôme du second degré tel qu'indiqué à la figure 9.2b.

Figure 9.2: Analyse des résidus, exemple 9.1







Lorsque l'analyse du modèle proposé présente l'un ou l'autre des diagnostics suivants:

- R^2 faible et/ou test F non-significatif;
- Analyse des résidus non satisfaisante à cause de points aberrants, de l'hypothèse de normalité violée et/ou de variance non-constante

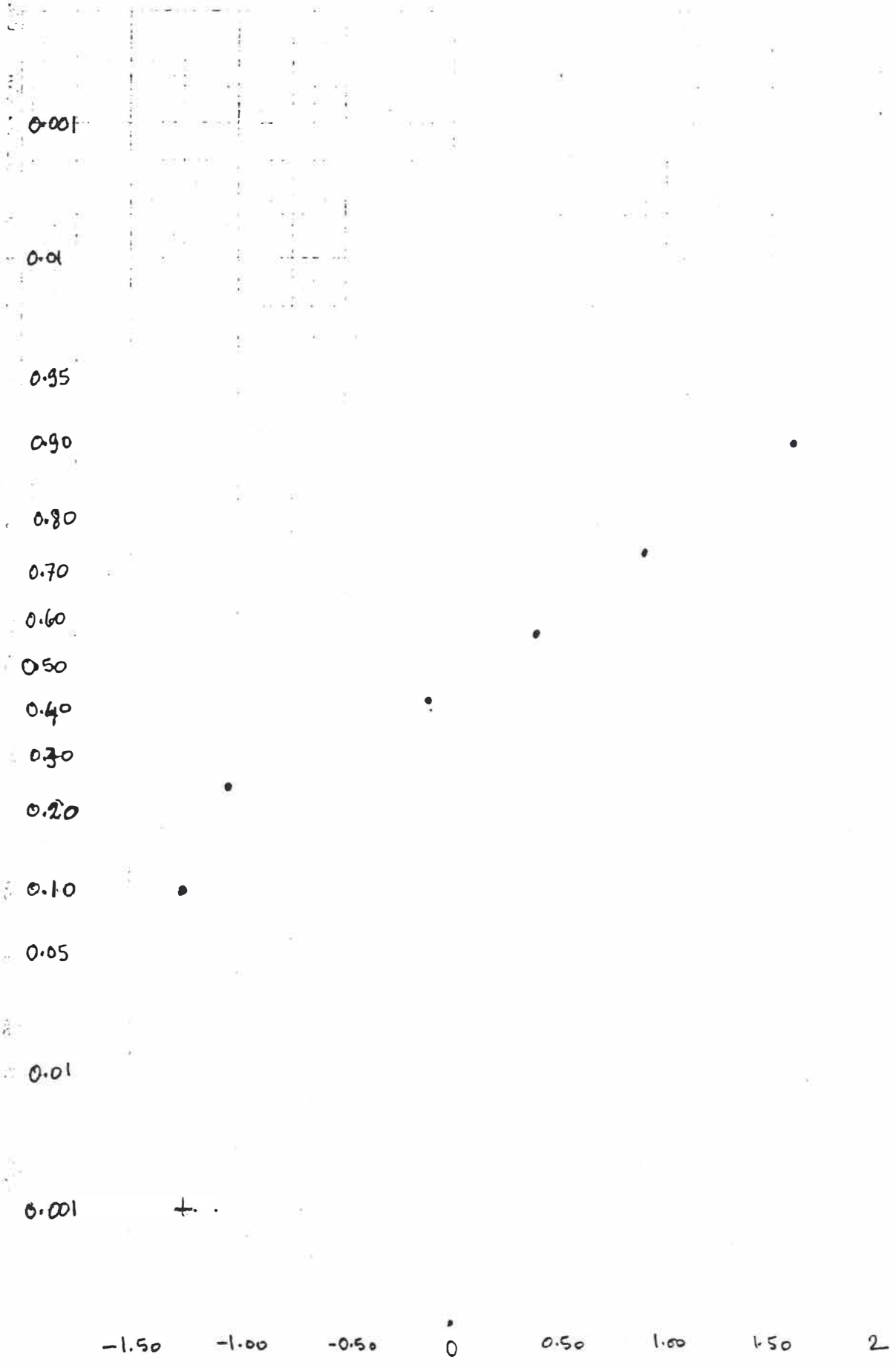
on doit alors proposer des correctifs afin d'obtenir de meilleurs résultats. Certaines anomalies constatées d'un modèle sont plus importantes que d'autres. Par ordre d'importance il y a:

- la fonction φ incorrectement spécifiée;
- l'hétérogénéité de la variance;
- la dépendance des observations y_i;
- la non-normalité du terme d'erreur.

Les éléments d'une stratégie exploratoire et corrective sont:

- la recherche de "meilleures" fonctions φ;
- l'ajout de d'autres variables explicatives dans la fonction φ;
- examen des transformations sur les variables $X_1, X_2, \ldots, X_k$ et/ou Y;
- élimination de un, ou plusieurs points aberrants ou l'utilisation d'une méthode d'estimation robuste;
- utilisation de la technique des moindres carrés pondérés.

Figure 9.2d: Résidus studentisés ~~sur~~ échelle de probabilité gaussienne



9.3 TRANSFORMATIONS

Il y a plusieurs raisons pour lesquelles on effectue des transformations sur les variables: stabilisation de la variance, rendre linéaire certains modèles, amélioration de l'ajustement et obtention d'une distribution normale.

Transformations sur Y pour stabiliser la variance

Lorsque l'hypothèse d'homogénéité de la variance est suspecte on peut rechercher une transformation sur la variable Y disons

$$Y' = h(Y) \qquad (9.41)$$

de telle sorte que VAR(Y') soit constante. Le tableau 9.4 résume quelques cas souvent rencontrés.

Tableau 9.4: transformations pour stabiliser la variance de Y

Situation	Relation entre σ^2 = VAR(Y) et μ=E(Y)	Transformation
Y est un comptage distribué selon une loi de Poisson	$\sigma^2 \alpha \mu$	$Y'=\sqrt{Y}$ si $Y>0$ ou $Y'=\sqrt{Y+\sqrt{Y+1}}$ si $Y=0$
L'étendue de Y est très grande; plusieurs ordres de grandeurs pour Y	$\sigma^2 \alpha \mu^2$	$Y'=\log Y$ si $Y>0$ ou $Y'=\log(Y+1)$ si $Y=0$
Beaucoup de valeurs de Y près de zéro ainsi que de très grandes valeurs pour Y	$\sigma^2 \alpha \mu^4$	$Y'=1/Y$ si $Y>0$ ou $Y'=1/(Y+1)$ si $Y=0$
$0 \leq Y \leq 1$ et Y est une variable binomiale	$\sigma^2 \alpha \mu(1-\mu)$	$Y'=\sin^{-1}(\sqrt{Y})$

Exemple 9.3: nombre de superviseurs - nombre d'employés

Dans une étude de 27 compagnies on a relevé le nombre de superviseurs (SUP) et le nombre d'employés (EMP)

EMP	247	267	294	311	358	423	438	450	534
SUP	32	37	30	49	44	47	68	56	62

EMP	615	627	630	688	697	700	709	850	980
SUP	100	97	84	80	78	106	88	128	130

EMP	999	1022	1015	1021	1025	1200	1250	1500	1650
SUP	109	114	117	97	160	180	112	210	135

Le premier modèle proposé est:

$$SUP = \beta_0 + \beta_1 * EMP + \varepsilon$$

$$\varepsilon \sim N(0,\sigma^2)$$

Les calculs donnent:

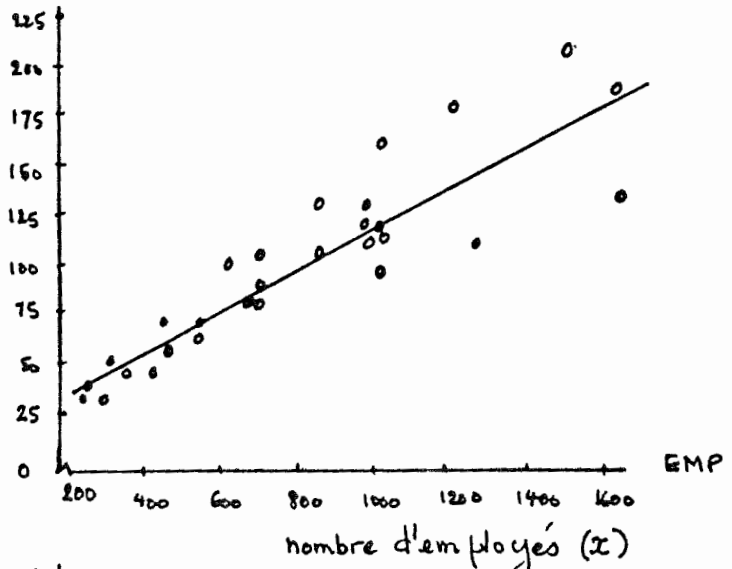
$$SUP = 14.45 + 0.105*EMP$$

$$R^2 = 0.776 \qquad F = 86.54$$

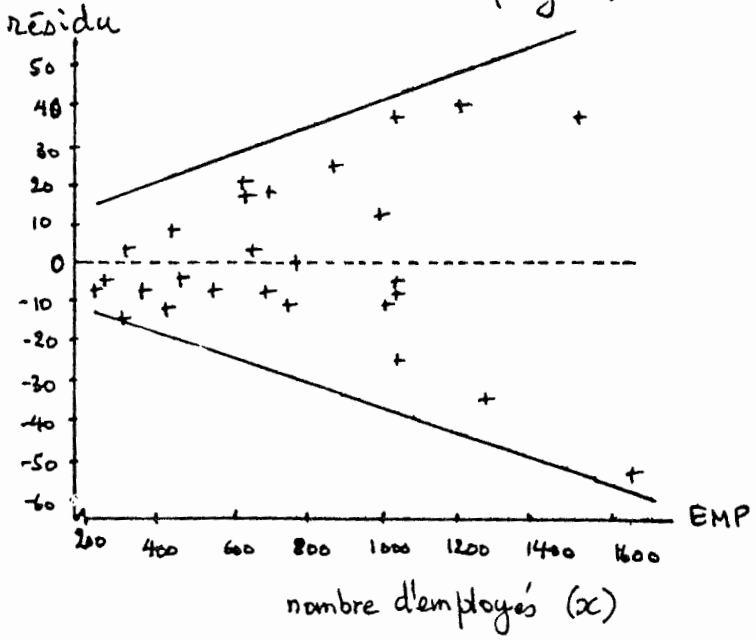
Le test de la nullité de β_1 est significatif (rejeté) et le coefficient β_0 n'est pas significatiment différent de zéro. Cependant, l'examen des points expérimentaux et des résidus sur la figure 9.3(a),(b) et (c) indique que l'hypothèse d'une variance constante pour Y n'est pas valide et par conséquent, on doit rejeter le modèle.

Figure 9.3: Analyse des résidus, exemple 9.3

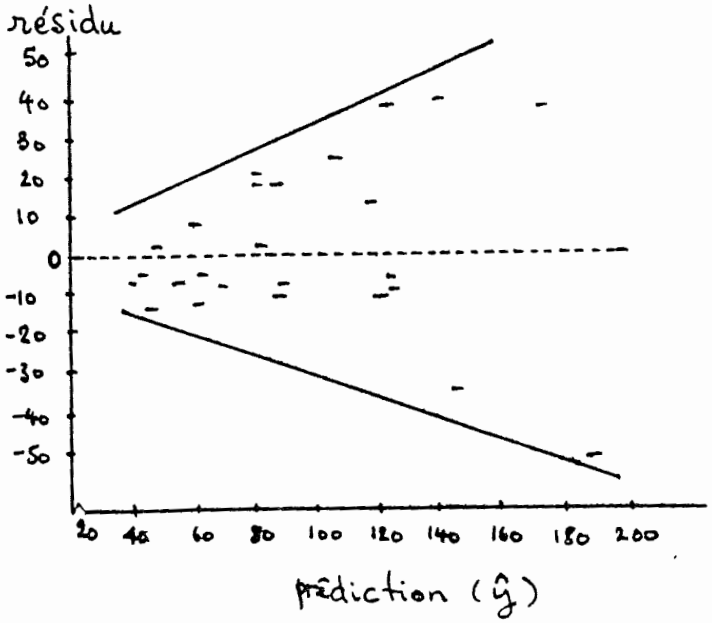
(a)



(b)



(c)



A l'examen de la figure 9.3 il semble que la variance de SUP, soit proportionnelle à la moyenne de SUP. Selon le tableau 9.4 nous allons proposer un deuxième modèle en utilisant la nouvelle variable

$$\text{RACSUP} = \sqrt{\text{SUP}}$$

$$\text{RACSUP} = \beta_0 + \beta_1 * \text{EMP} + 2$$

$$\varepsilon \sim N(0, \sigma^2) \qquad (9.44)$$

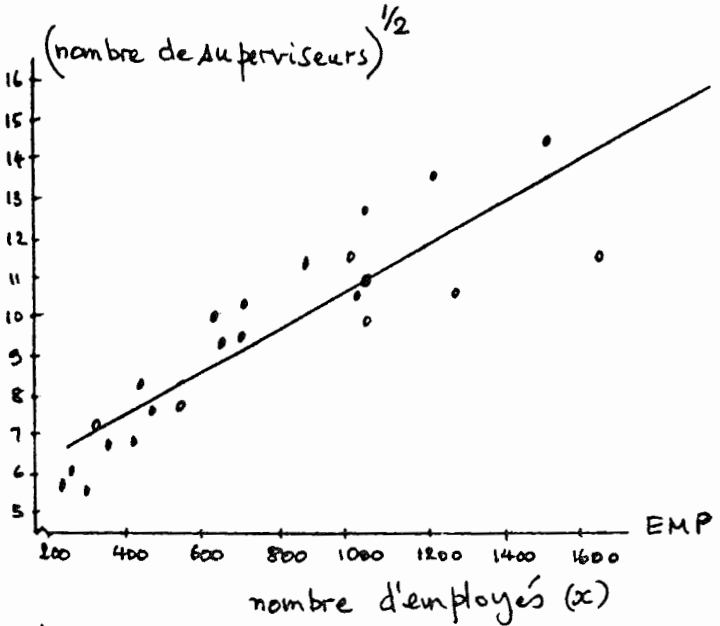
Les calculs donnent:

$$\text{RACSUP} = 5.265 + 0.055 * \text{EMP}$$
$$R^2 = 0.791 \qquad F = 94.69 \qquad (9.45)$$

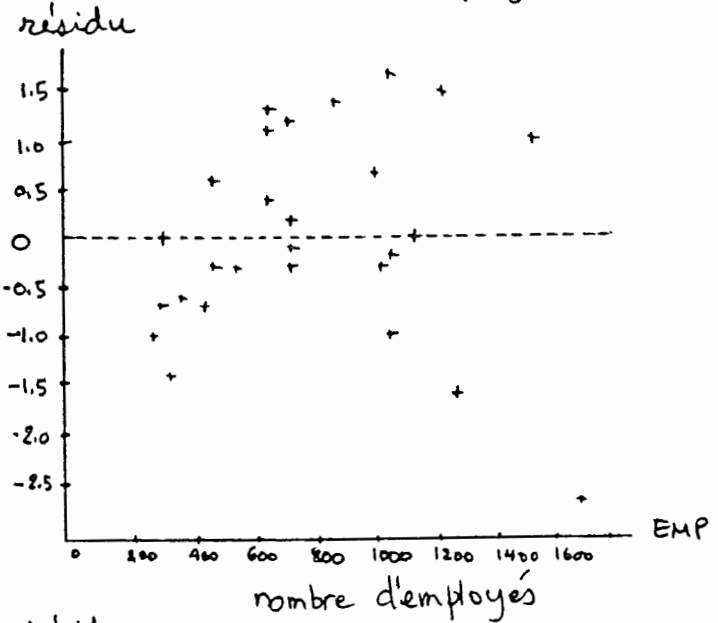
Les tests de nullité de β_0 et β_1 sont tous les deux significatifs à 0.0001. La figure 9.4 indique un comportement des résidus plus acceptable (bande horizontale). Toutefois, on pourrait peut-être améliorer le modèle (plus grand R^2) en effectuant aussi une transformation sur EMP. En effet, il semble y avoir encore une tendance des résidus à se placer en forme de croissant parabolique. Une transformation logarithmique de EMP semble intéressante puisque l'étendue de cette variable est grande. L'examen de cette possibilité est laissé en exercice.

Figure 9.4: Analyse des résidus, exemple 9.3 (Suite)

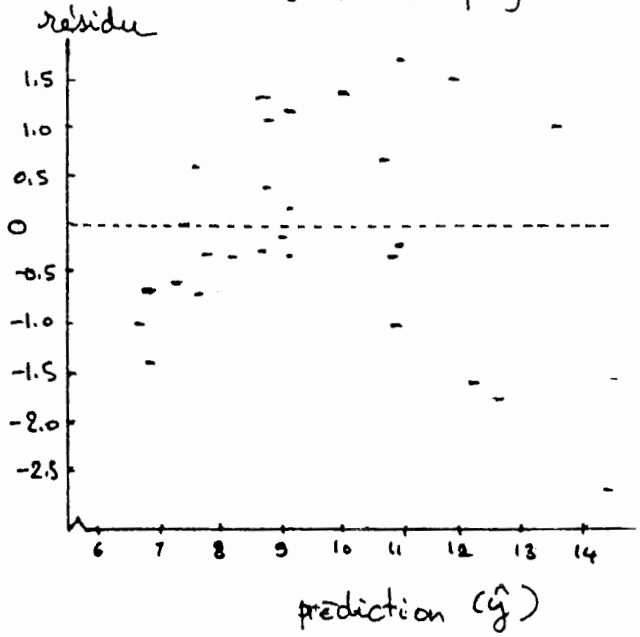
(a)



(b)



(c)



Transformations sur X et/ou Y pour rendre linéaires certains modèles

L'existence d'une relation linéaire entre deux variables est un fait assez exceptionnel. Néanmoins l'utilité d'un modèle linéaire en X est plus générale qu'il peut sembler à première vue. En effet, même si la fonction φ est non-linéaire sur l'ensemble du domaine de X, elle peut se comporter d'une manière approximativement linéaire sur certaines portions du domaine. D'autre part certains modèles non-linéaires en X peuvent, après transformation des variables, prendre une forme linéaire. Le tableau 9.5 présente des cas importants.

On détecte la non-linéarité (en x) de la fonction φ en traçant les graphiques de

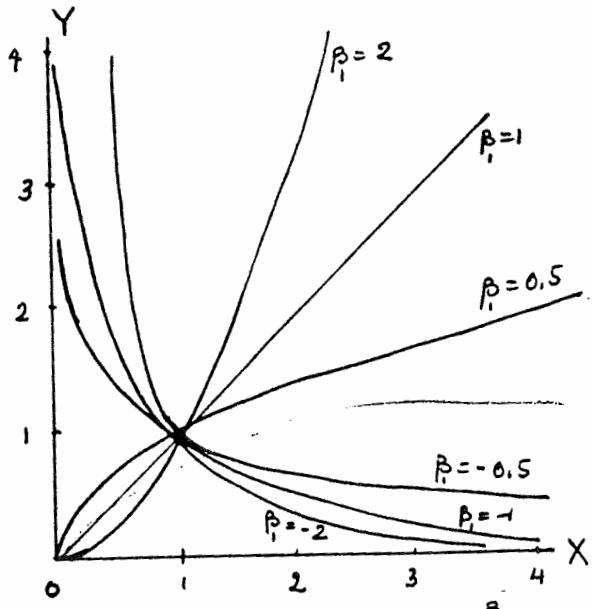
- y_i versus x_i
- r_i versus $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Le choix spécifique d'une transformation dépend de la nature des variables, du degré de connaissance du phénomène étudié et d'autres considérations. La comparaison des points expérimentaux (x_i, y_i) avec les figures 9.5 (a) - (f) peut aider à guider le choix d'une transformation. On peut aussi faire le graphique des points (x_i, y_i) sur différents échelles: logarithmique-arithmétique, arithmétique-logarithmique et logarithmique-logarithmique. Cela aide à prendre une décision sur le choix d'une transformation appropriée.

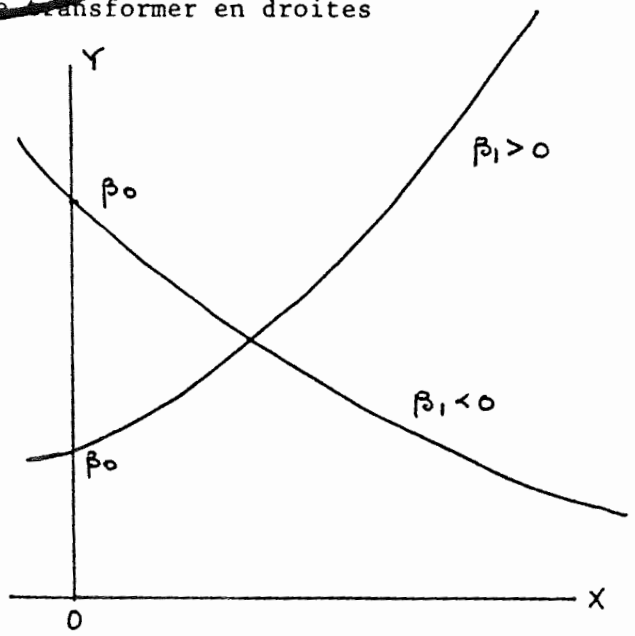
Tableau 9.5: Transformations

Fonction originelle	Transformation	Forme linéaire $Y'=\beta_0+\beta_1 X'$
$Y=\beta_0' X^{\beta_1}$	$Y'=\ell nY$ $X'=\ell nX$ $\beta_0=\ell n\beta_0'$	figure 9.5a
$Y=\beta_0' \exp(\beta_1 X)$	$Y'=\ell nY$ $X'=X$ $\beta_0=\ell n\beta_0$	figure 9.5b
$Y=\beta_0+\beta_1 \ell nX$	$Y'=Y$ $X'=\ell nX$	figure 9.5c
$Y=\frac{1}{\beta_0+\beta_1 X}$	$Y'=1/Y$ $X'=X$	figure 9.5d
$Y=\beta_0+\beta_1 (\frac{1}{X})$	$Y'=Y$ $X'=1/X$	figure 9.5e
$Y=\frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}$ $(0 < Y<1)$ (logistique)	$Y'=\ell n\left(\frac{Y}{1-Y}\right)$ $X'=X$	figure 9.5f

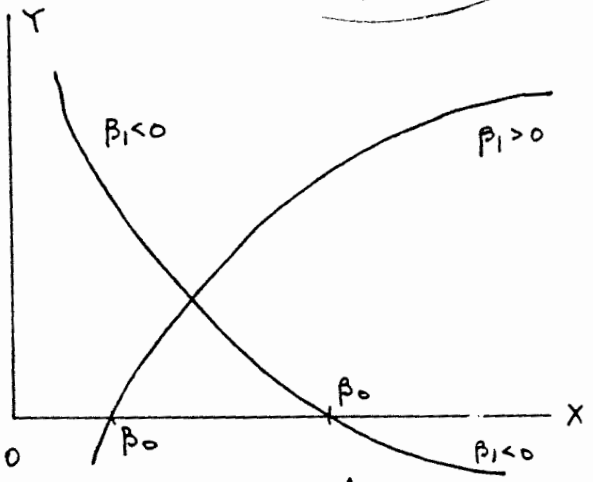
Figure 9.5: Fonctions pouvant se transformer en droites



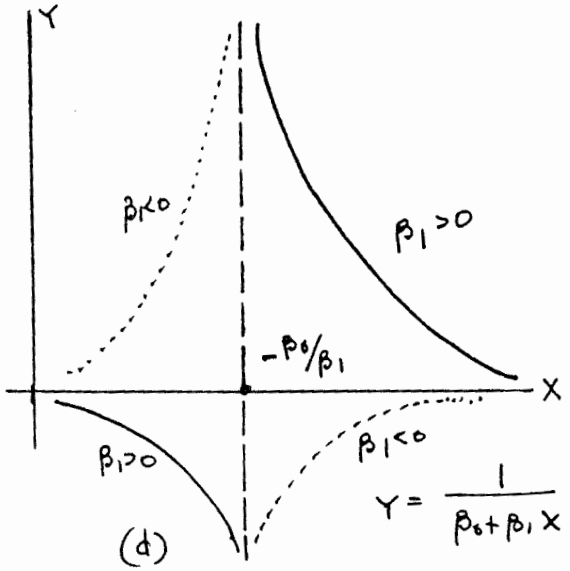
(a) $Y = \beta_0' X^{\beta_1}$



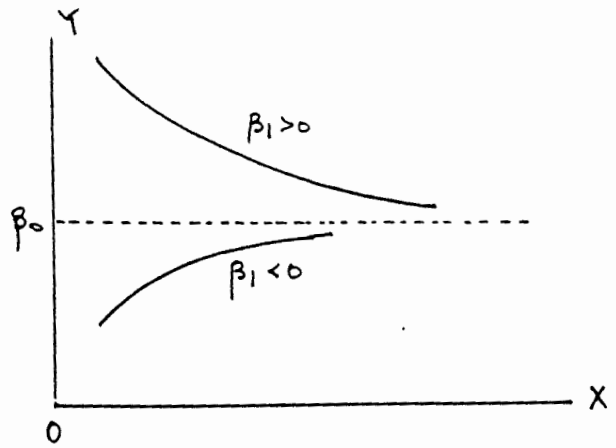
(b) $Y = \beta_0' e^{\beta_1 X}$



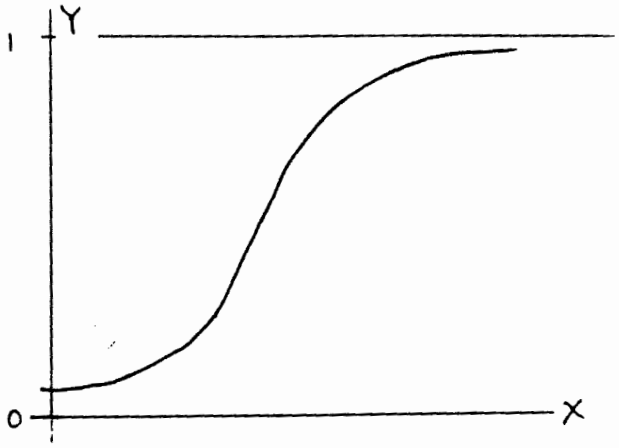
(c) $Y = \beta_0 + \beta_1 \ln X$



(d) $Y = \dfrac{1}{\beta_0 + \beta_1 X}$



(e) $Y = \beta_0 + \beta_1 \left(\frac{1}{X}\right)$



(f) $Y = \dfrac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$

Exemple 9.4: données sur la température (TEMP) d'ébullition de l'eau (en °F) et la pression (PRES) barométrique (en po de Hg) prises dans la chaîne de montagnes de l'Himalaya.

TEMP	180.6	181.0	181.9	182.4	183.2	184.1
PRES	15.376	15.919	15.928	16.235	16.385	16.817
TEMP	184.6	185.6	186.0	188.5	188.8	
PRES	16.881	17.062	17.221	18.507	18.356	
TEMP	189.5	190.6	191.1	193.6	195.6	
PRES	18.869	19.386	19.490	20.212	21.605	
TEMP	196.4	197.0	199.5	200.6	202.5	
PRES	21.928	21.892	23.030	23.726	24.697	
TEMP	208.4	210.2	210.8			
PRES	27.972	28.559	29.211			

Dans une première analyse on se propose d'ajuster le modèle

$$\text{PRES} = \beta_0 + \beta_1 * \text{TEMP} + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2) \qquad (9.46)$$

On obtient les résultats suivants:

$$\hat{\text{PRES}} = -65.43 + 0.445 * \text{TEMP} \qquad (9.47)$$

$\hat{\sigma}^2 = 0.137$ $\quad R^2 = 0.995$ $\quad F = 2896.1$

i	PRES	$\hat{PRES}$	Résidu(r_i)	Résidu studentisé(rs_i)
1	15.376	15.056	0.320	0.916
2	15.919	15.234	0.685	1.954
3	15.928	15.635	0.293	0.8315
4	16.235	15.858	0.377	1.067
5	16.385	16.214	0.171	0.481
6	16.817	16.615	0.202	0.566
7	16.881	16.838	0.043	0.120
8	17.062	17.284	-0.223	-0.618
9	17.221	17.462	-0.241	-0.671
10	18.507	18.576	-0.069	-0.191
11	18.356	18.710	-0.354	-0.978
12	18.869	19.022	-0.153	-0.422
13	19.386	19.512	-0.126	-0.347
14	19.490	19.735	-0.245	-0.675
15	20.212	20.849	-0.637	-1.757
16	21.605	21.740	-0.135	-0.3735
17	21.928	22.097	-0.169	-0.467
18	21.892	22.364	-0.472	-1.309
19	23.030	23.478	-0.448	-1.253
20	23.726	23.968	-0.242	-0.681
21	24.697	24.815	-0.118	-0.335
22	27.972	27.444	0.528	1.567
23	28.559	28.246	0.313	0.946
24	29.211	28.514	0.697	2.124

On constate, à l'examen de la figure 9.6 (a)-(b), que les résidus sont placés en forme de croissant parabolique et cela suggère l'inadéquation du modèle proposé.

Dans une deuxième tentative on propose

$$100*\text{LOGPRES} = \beta_0 + \beta_1*\text{TEMP} + \varepsilon$$

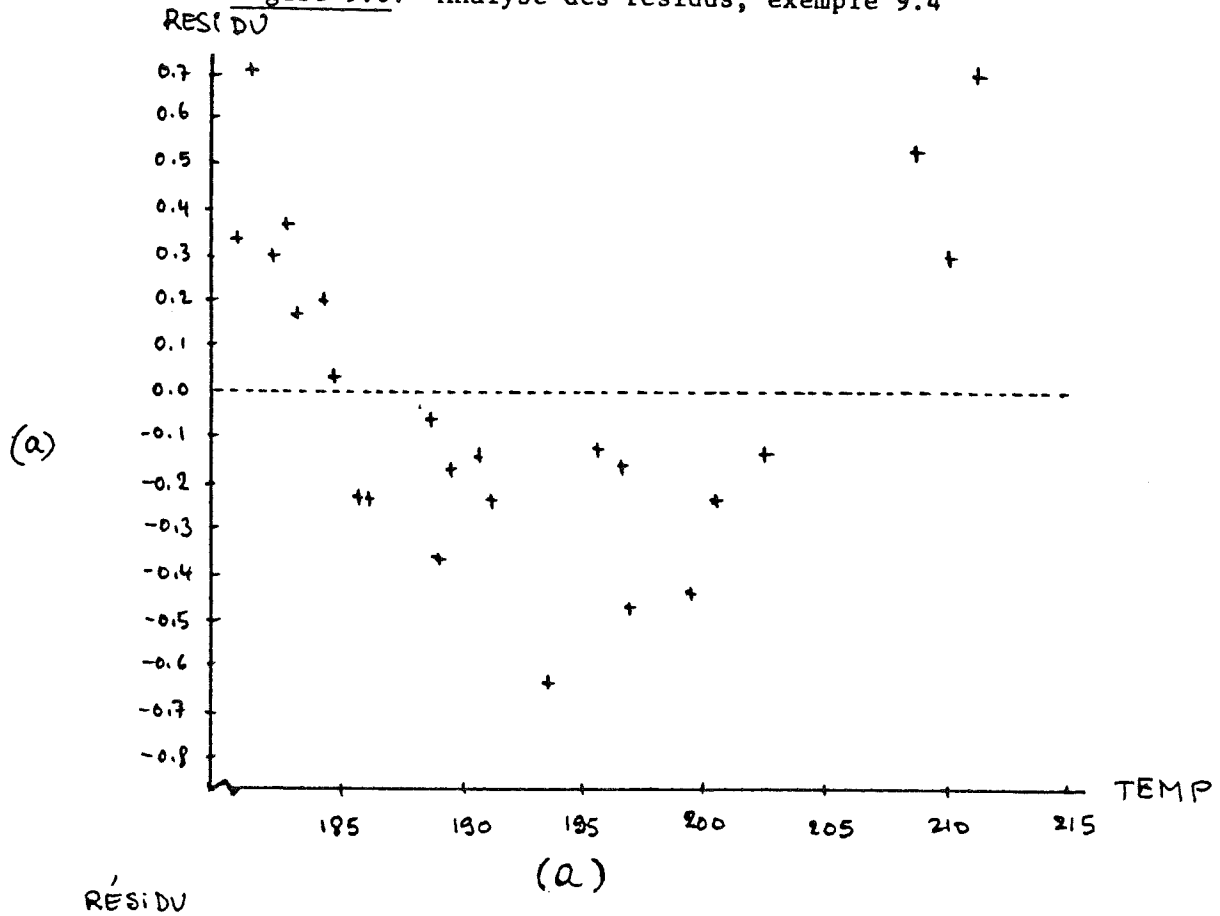
$$\varepsilon \sim N(0,\sigma^2) \qquad (9.48)$$

Les calculs donnent

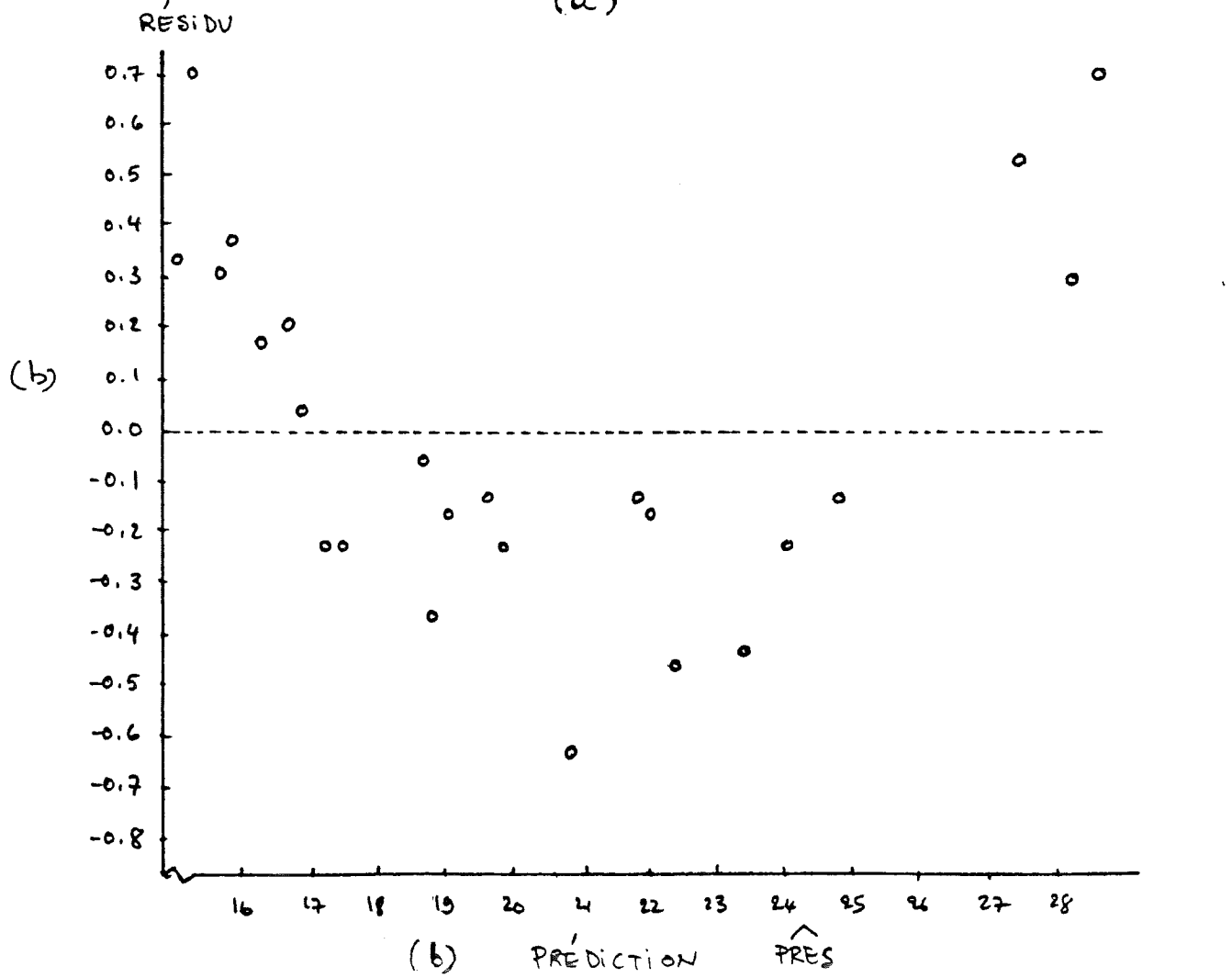
$$100*\text{LOGPRES} = -103.57 + 2.09*\text{TEMP} \qquad (9.49)$$

$\hat{\sigma}^2 = 0.76$ $\qquad R^2 = 0.998$ $\qquad F = 11495.37$

Figure 9.6: Analyse des résidus, exemple 9.4



(a)



(b)

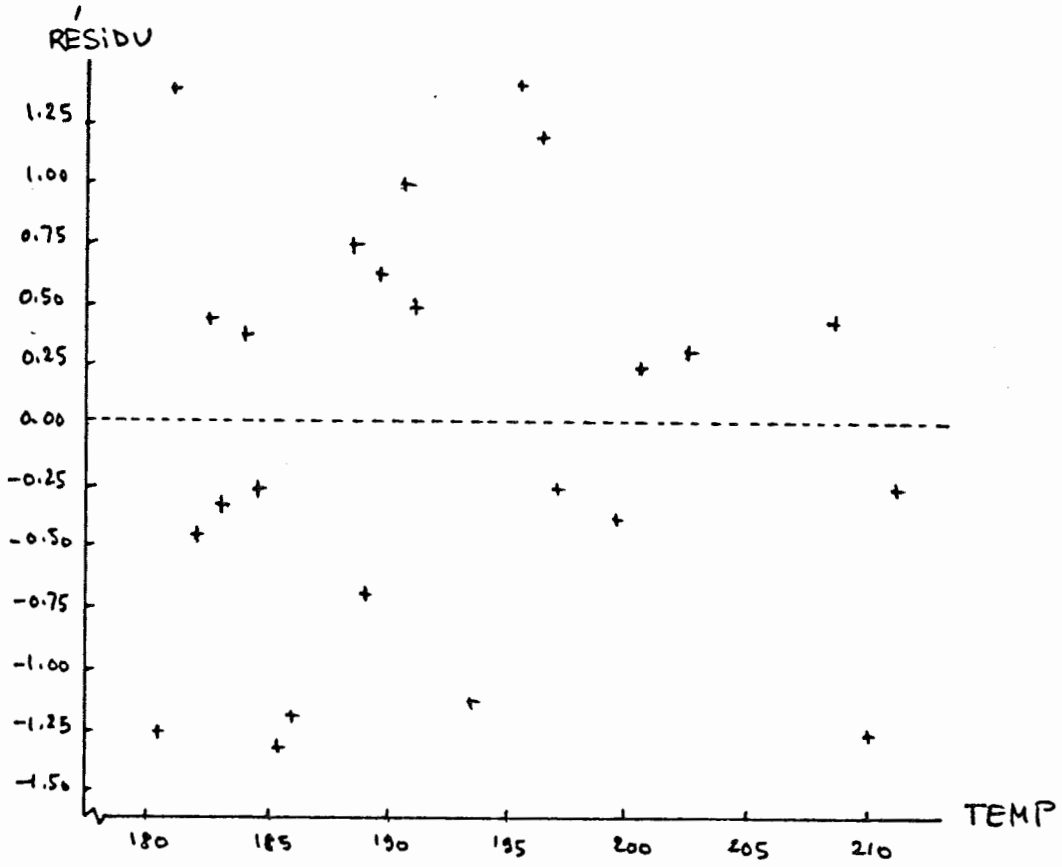
(b) PRÉDICTION $\widehat{\text{PRES}}$

i	LONGPRES (*100)	LONGPRES (*100)	Résidu (r_i)	Résidu Studentisé(rs_i)
1	273.3	274.5	-1.23	-1.49
2	276.8	275.3	1.40	1.70
3	276.8	277.2	-0.42	-0.51
4	278.7	278.3	0.44	0.53
5	279.6	280.0	-0.40	-0.38
6	282.2	281.8	-0.265	0.48
7	282.6	282.9	-1.29	-0.32
8	283.7	285.0	-1.20	-1.53
9	284.6	285.8	0.77	-1.42
10	291.8	291.0	-0.68	0.90
11	291.0	291.7	0.61	-0.80
12	293.8	293.1	1.01	0.71
13	296.5	295.4	0.50	1.18
14	297.0	296.5	-1.10	0.58
15	300.6	301.7	1.38	-1.28
16	307.3	305.9	1.19	1.62
17	308.8	307.6	-0.23	1.40
18	308.6	308.8	-0.40	-0.27
19	313.7	314.1	0.28	-0.47
20	317.7	316.4	0.31	0.33
21	320.7	320.4	0.41	0.37
22	333.1	332.7	-1.28	0.52
23	335.2	336.5	-0.28	-1.64
24	337.5	337.7		-0.36

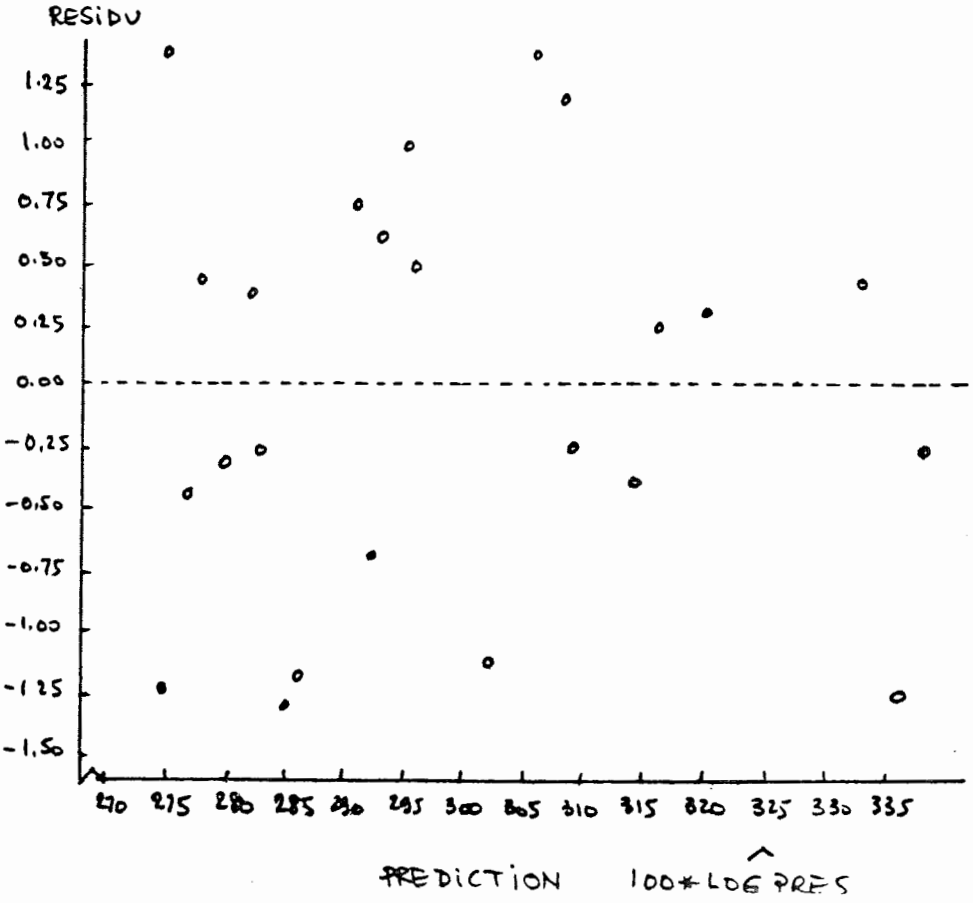
L'examen de la figure 9.7 (a),(b) ne révèle aucune tendance anormale des résidus.

Figure 9.7: Analyse des résidus suite, exemple 9.4

(a)



(b)



9.4 RÉGRESSION LINÉAIRE MULTIPLE

Le modèle

Le modèle de régression linéaire multiple à déjà été présenté au chapitre 7 (exemple 7.21). Il s'écrit

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \qquad (9.50)$$

où Y est la variable à expliquer
$X_1, X_2, \ldots X_p$ sont des variables explicatives
$\beta_o, \ldots, \beta_p$ sont des paramètres appelés coefficients de régression partiels
ε est le terme d'erreur tel que

$$E(\varepsilon) = 0, \quad VAR(\varepsilon) = VAR(Y) = \sigma^2 \qquad (9.51)$$

On dispose d'observations

$$(x_{i1}, x_{i2}, \ldots, x_{in}, y_i) \quad i=1,2,\ldots n>p+1$$

et les principales questions à résoudre sont

- l'estimation des paramètres $\beta_o, \beta_1, \ldots, \beta_p, \sigma^2$;
- l'analyse de la qualité du modèle par le tableau d'analyse de la variance, les tests d'hypothèses sur les paramètres $\beta_1, \ldots, \beta_p$ et l'analyse des résidus.

L'estimation des paramètres $\beta_o \ldots, \beta_p$

L'estimation des paramètres est faite par la méthode des moindres carrés qui a été exposée au chapitre 7. On considère la fonction auxiliaire $S(\beta_o, \ldots, \beta_p)$ définie par:

$$S(\beta_o,\ldots\beta_p) = \sum_{i=i}^{n} \left[y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right]^2 \qquad (9.52)$$

où par convention $x_{io} = 1$ pour tout i. On cherche le minimum de S dans l'espace des paramètres $\beta_o,\beta_1\ldots,\beta_p$

$$\frac{\partial S}{\partial \beta\alpha} = \sum_{i=i}^{n} 2 \left[y_i - \sum_{j=o}^{p} \beta_j x_{ij} \right](-x_{i\alpha}) = 0 \qquad (9.53)$$

$$\alpha = 0,1,2,\ldots,P$$

Après réarrangement le système (9.53) s'écrit:

$$\begin{bmatrix} n & \Sigma x_{i1} & \Sigma x_{i2} & \ldots & \Sigma x_{ip} \\ \Sigma x_{i1} & \Sigma x_{i1}^2 & \Sigma x_{i2}x_{i1} & \ldots & \Sigma x_{ip}x_{i1} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ \Sigma x_{ip} & \Sigma x_{i1}x_{ip} & \Sigma x_{i2}x_{ip} & \ldots & \Sigma x_{ip}^2 \end{bmatrix} \begin{bmatrix} \beta_o \\ \beta_1 \\ . \\ . \\ . \\ \beta_p \end{bmatrix} = \begin{bmatrix} \Sigma y_i \\ \Sigma y_i x_{i1} \\ . \\ . \\ . \\ \Sigma y_i x_{ip} \end{bmatrix} \qquad (9.54)$$

Pour la suite de l'exposé, nous allons utiliser la notation matricielle afin de rendre plus compacte l'écriture des équations. L'opération de transposition sera notée par le symbole '.

Posons

$$\begin{aligned}
&\underset{\sim}{\beta} = (\beta_o,\ldots,\beta_p)' && : \quad \text{vecteur } (p+1)X1 \\
&\underset{\sim}{Y} = (y_1,\ldots,y_n)' && : \quad \text{vecteur } nX1 \\
&\underset{\sim}{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)' && : \quad \text{vecteur } nX1 \\
&X = [x_{ij}] && : \quad \text{matrice } nX(p+1)
\end{aligned} \tag{9.55}$$

Les équations (9.50) et (9.52) deviennent

$$\begin{aligned}
&\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \\
&S = (\underset{\sim}{Y} - X\underset{\sim}{\beta})' (\underset{\sim}{Y} - X\underset{\sim}{\beta}) = \underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon}
\end{aligned} \tag{9.56}$$

et le système d'équations linéaires (9.54) s'écrit:

$$(X'X)\,\underset{\sim}{\beta} = X'\underset{\sim}{Y} \tag{9.57}$$

Si la matrice X'X est inversible on peut exprimer la solution $\hat{\beta}$ de (9.57) par

$$\underset{\sim}{\hat{\beta}} = (X'X)^{-1}\,X'\,\underset{\sim}{Y} = C\,\underset{\sim}{Y} \tag{9.58}$$

où $C = (X'X)^{-1}X'$ est une matrice de valeurs connues.

Les estimations $\underset{\sim}{\hat{\beta}}$ de moindres carrés ont les propriétés suivantes qui ont été démontrées à la section 7.5:

- les estimateurs $\hat{\beta}_j$ sont des combinaisons linéaires des y_i

- les estimateurs $\hat{\beta}_j$ sont sans biais

$$E(\underset{\sim}{\hat{\beta}}) = \underset{\sim}{\beta}$$

- la matrice de variance-covariance de $\underset{\sim}{\hat{\beta}}$ est

$$VAR(\underset{\sim}{\hat{\beta}}) = (X'X)^{-1}\,\sigma^2$$

- les estimations $\hat{\beta}_j$ sont à vraisemblance maximale si $\varepsilon_i \sim N(0,\sigma^2)$ et sont indépendantes.

<u>L'estimation du paramètre σ^2</u>

Nous avons déjà justifié (paragraphe 9.2) une formule générale pour l'estimation du paramètre σ^2. Appliquons cette formule pour le modèle de régression linéaire multiple. Posons

$$\underset{\sim}{x_i} = (x_{io}, x_{i1}, \ldots, x_{ip}) \qquad (9.59)$$

le vecteur 1x(p+1) représentant la i-ième rangée de la matrice X. Alors

$$\hat{y}_i = \underset{\sim}{x_i}\ \underset{\sim}{\hat{\beta}} = \sum_{j=0}^{p} \hat{\beta}_j\, x_{ij} \qquad (9.60)$$

est la prédiction (projection) de la variable y au point $\underset{\sim}{x_i}$.

On propose d'estimer σ^2 par $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)} \sum_{i=i}^{n} (y_i - \hat{y}_i)^2 \qquad (9.61)$$

Sous l'hypothèse de normalité de ε_i on montre que $\hat{\sigma}^2$ constitue une estimation sans biais de σ^2

$$E(\hat{\sigma}^2) = \sigma^2$$

et que

$$(n-p-1)\ \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1} \qquad (9.62)$$

Tableau d'analyse de la variance

Posons

$$SCT = \sum_{i=i}^{n} (y_i - \bar{y})^2 = \text{variabilité totale}$$

$$SCM = \sum_{1=1}^{n} (\hat{y}_i - \bar{y})^2 = \text{variabilité associée au modèle}$$

$$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \text{variabilité associée à l'erreur (résiduelle)}$$

L'équation de décomposition de la variabilité de Y vue pour le modèle de régression linéaire simple (p=1) se généralise au cas du modèle de régression multiple (p>1). On a

$$SCT = SCM + SCE \qquad (9.63)$$

À chacune des sommes de carrés est associée un nombre de degrés de liberté. On montre que

SCT	a	n-1	degrés de liberté
SCM	a	p	degrés de liberté
SCE	a	n-p-1	degrés de liberté

et on résume le tout dans le tableau 9.6

Tableau 9.6: analyse de la variance du modèle:

$$Y_i = \underset{\sim}{x}_i \underset{\sim}{\beta} + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

source de variation	somme de carrés	degrés de liberté	Carrés moyens	F
modèle	SCM	p	$CMM = \frac{SCM}{p}$	$\frac{CMM}{CME}$
erreur	SCE	n-p-1	$CME = \frac{SCE}{n-p-1}$	
Totale	SCT	n-1	---	---

On fait un premier examen de la qualité du modèle à l'aide des quantités R^2 et F.

$$R^2 = \frac{SCM}{SCT} \qquad 0 \leq R^2 \leq 1 \qquad (9.64)$$

est appelé le coefficient de détermination et représente une mesure de l'utilité des variables $X_1, X_2, \ldots, X_p$. En d'autres termes $R^2 = 0$ correspond à l'ajustement du modèle

$$Y = \beta_o + \varepsilon$$

tandis qu'un ajustement parfait ($\hat{y}_i = y_i$ pour tout i, un événement improbable en pratique) donnerait $R^2 = 1$.

La racine carrée R est appelée le coefficient de corrélation multiple entre Y et $X_1, X_2, \ldots, X_p$. Notons que R^2 est un meilleur indicateur que R pour juger de l'adéquation du modèle.

Une deuxième quantité importante est le rapport F

$$F = CMM/CME \qquad (9.65)$$

employé pour mettre à l'épreuve

$$H_N : \quad \beta_1 = \beta_2 = \ldots = \beta_p = 0 \qquad (9.65)$$

contre

$$H_A : \quad \text{non } H_N$$

On a le test suivant de seuil α:

$$\text{on rejette } H_N \text{ si } F > F_{p,n-1-p,\alpha}$$

où $F_{p,n-1-p,\alpha}$ est le 100 $(1-\alpha)$ième percentile d'une variable Fisher.

Nous verrons plus loin d'autres tests d'hypothèses concernant les coefficients $\beta_1, \ldots, \beta_p$.

exemple 9.5 données sur la composition du béton de ciment (chapitre 1)

X_1	X_2	X_3	X_4	Y
7	26	6	60	78.5
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	53	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	19	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

Les variables X_1, X_2, X_3, X_4 représentent des pourcentages de composition de 4 composants chimiques et Y est la chaleur dégagée (cal/gr). On propose le modèle

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Les équations de moindres carrés (9.57) sont:

$$\begin{bmatrix} 13 & 97 & 626 & 153 & 390 \\ 97 & 1139 & 4922 & 769 & 2620 \\ 626 & 4922 & 3350 & 7201 & 15739 \\ 153 & 769 & 7201 & 2293 & 4628 \\ 390 & 2620 & 15739 & 4628 & 15062 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 1240.5 \\ 10032 \\ 62027.8 \\ 13981.5 \\ 34733.3 \end{bmatrix}$$

et la solution est

$$\underset{\sim}{\hat{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{bmatrix} = \begin{bmatrix} 62.40 \\ 1.55 \\ 0.51 \\ 0.10 \\ -0.14 \end{bmatrix}$$

Les sommes de carrés du tableau d'analyse de la variance sont calculées de la manière suivante:

$$SCT = \sum_{i=i}^{n} y_i^2 - \frac{(\Sigma y_i)^2}{n} = 121088.09 - 118372.33$$

$$= 2715.76$$

$$SCM = \sum_{j=1}^{p} \hat{\beta}_j \sum_{i=1}^{n} (x_{ij} - \bar{x}_i)(y_i - \bar{y}) - \frac{(\Sigma y_i)^2}{n}$$

$$= 121040.22 - 118372.33 = 2667.90$$

$$SCE = SCT - SCM = 2715.76 - 2667.90 = 47.86$$

Le tableau d'analyse de la variance est

Tableau 9.7

Source	Somme de Carrés	Degrés de liberté	Carrés moyens	F
Modèle	2667.90	4	666.97	111.48
Erreur	47.86	8	5.98	-
Total	2715.76	12	-	-

$$R^2 = 0.9824$$

On rejette facilement l'hypothèse nulle H_N puisque $F_{4,8,0.005} = 8.81$. À première vue, le modèle semble adéquat (R^2 élevé, F significatif) mais une analyse, effectuée plus loin, démontre l'inadéquation du modèle causée par la présence de variables explicatives colinéaires.

Test d'hypothèses concernant β_j

Sous l'hypothèse d'une distribution normale pour le terme d'erreur on a

$$\frac{\hat{\beta}_j - \beta_j}{ET(\hat{\beta}_j)} \sim T_{n-p-1} \qquad (9.66)$$

où

$$ET(\hat{\beta}_j) = \sqrt{c^{ii}}\,\hat{\sigma} \qquad (9.67)$$

c^{ii} est l'élément (i,i) de la matrice $(X'X)^{-1}$

$$C = [c^{ij}] = (X'X)^{-1} \qquad (9.68)$$

Le résultat (9.66) permet d'effectuer des tests d'hypothèses et de calculer des intervalles de confiance concernant le paramètre β_j.

Le test de l'hypothèse nulle:

$$H_N : \beta_j = 0 \quad \text{contre} \quad H_A : \beta_j \neq 0$$

est de rejeter H_N au seuil α si

$$\frac{\hat{\beta}_j}{ET(\hat{\beta}_j)} > t_{n-p-1,\alpha/2}$$

où $t_{n-p-1,\alpha/2}$ est le 100 $(1-\alpha/2)$-ième percentile d'une distribution de Student avec n-p-1 degrés de liberté.

Exemple 9.6 suite de l'exemple 9.5

Les calculs donnent

i	Coefficient $(\hat{\beta}_j)$	$ET(\hat{\beta}_j)$	$\frac{\hat{\beta}}{ET(\hat{\beta}_j)}$
0	62.40	---	---
1	1.55	0.745	2.08
2	0.51	0.73	0.70
3	0.10	0.71	0.14
4	-0.14	0.70	-0.20

On constate que pour chacun des coefficients $\beta_1, \beta_2, \beta_3, \beta_4$, on ne peut pas rejeter les hypothèses

$$H_j : \beta_j = 0 \qquad j = 1,2,3,4$$

alors que nous avons rejeté (exemple 9.5)

$$H_N : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

Cette situation paradoxale est une indication que le modèle n'est pas satisfaisant et est souvent issu du fait que certaines variables explicatives sont liées par une ou plusieurs dépendances linéaires. Cette situation se rencontre assez fréquemment en analyse de régression multiple. Dans notre exemple 9.6, on peut constater la présence de ce problème car

$$0.95 \leq x_{i1} + x_{i2} + x_{i3} + x_{i4} \leq 0.99 \qquad (9.69)$$

Des critères pour détecter le problème de variables colinéaires et ceux pour analyser la stabilité seront exposés plus loin.

On peut résumer actuellement nos critères pour retenir un modèle:

- R^2 élevé
- test F global significatif
- tests individuels significatifs pour chacun des coefficients β_j
- résidus ne montrant pas d'anomalies:
  variance constante, normalité pas de "tendance anormale" avec chacune des variables explicatives (figure 9.1)
- pas de variables explicatives colinéaires
- absence de points ayant une influcence prépondérante sur l'équation de prédiction (stabilité)

Intervalles de confiance et de prédiction

Sous l'hypothèse de normalité, on a développé les formules suivantes pour les intervalles de confiance et de prédiction:

$$\beta_j : \hat{\beta}_j \pm t_{n-p-1,\alpha/2}\, \hat{\sigma} \sqrt{c^{jj}} \qquad (9.70)$$

où c^{jj} est défini par (9.68)

$$E(Y|\underset{\sim}{x}) : \hat{\underset{\sim}{\beta}}' \underset{\sim}{x} \pm t_{n-p-1,\alpha/2}\, \hat{\sigma} \left(\underset{\sim}{x}'(X'X)^{-1}\underset{\sim}{x} \right)^{1/2} \qquad (9.71)$$

$$Y : \hat{\underset{\sim}{\beta}}\, \underset{\sim}{x} \pm t_{n-p-1,\alpha/2}\, \hat{\sigma} \left(1+\underset{\sim}{x}'(X'X)^{-1}\underset{\sim}{x} \right)^{1/2} \qquad (9.72)$$

Notons la différence entre (9.71) et (9.72): la première est un intervalle de confiance pour la moyenne de Y (une constante) tandis que la deuxième est un intervalle de prédiction pour la valeur de Y (une variable aléatoire).

L'analyse des résidus

Les mêmes techniques graphiques que celles employées en régression linéaire simple s'appliquent dans le cas de régression multiple:

- graphique de r_i versus $\hat{y}_i$
- graphique de r_i versus x_{ij}, $\quad j = 1,2,\ldots p$

9.5 ANALYSE DE STABILITÉ

Qu'arrive-t-il à l'équation obtenue lorsque l'on soumet les données à de "petites" perturbations? Les coefficients de régression sont-ils stables à la suite de ces changements aux données? Si oui, on dit que l'équation est stable, ce qui est une propriété désirable. Une approche pour effectuer une analyse de stabilité consiste à observer l'effet produit sur l'équation en éliminant tour à tour chaque observation. Cette analyse permet de mesurer l'influence de chaque observation sur l'équation. Une observation $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$ est dite INFLUENTE si elle produit un effet démesuré sur les estimations $\hat{\beta}_j, \hat{\sigma}^2$ et R^2. Pour évaluer s'il y a influence ou non, on refait tous les calculs en éliminant l'observation. Nous allons montrer comment identifier les observations influentes à l'aide de certaines mesures diagnostiques basées sur des notions de distance.

La matrice H

On peut exprimer le vecteur de prédiction $\hat{\underset{\sim}{Y}}$ en fonction du vecteur des observations $\underset{\sim}{Y}$. En effet

$$\begin{aligned} \hat{\underset{\sim}{Y}} &= X \hat{\underset{\sim}{\beta}} \\ &= X \left[(X'X)^{-1} X' \underset{\sim}{Y} \right] \\ &= X (X'X)^{-1} X' \underset{\sim}{Y} \\ &= H \underset{\sim}{Y} \end{aligned} \tag{9.73}$$

où

$$H = X(X'X)^{-1} X' = [h_{ij}]_{nxn} \tag{9.74}$$

Cette matrice H est importante pour effectuer l'analyse des points influents. H représente un opérateur de projection opérant sur l'espace engendré par les colonnes de X. Les éléments h_{ij} de H sont définis par

$$h_{ij} = \underset{\sim}{x}'_i \, (X'X)^{-1} \, \underset{\sim}{x}_j \qquad (9.75)$$

où

$\underset{\sim}{x}_i = (x_{i0},\ldots,x_{ip})$ est la i-ième rangée de la matrice X.

On peut montrer que:

$$H' = H$$

$$H^2 = H$$

$$\sum_{i=1}^{n} h_{ii} = \text{rang de la matrice X}$$

$$\sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ij} = 1$$

$$\hat{y}_i = h_{ii}\, y_i + \sum_{j \neq i} h_{ij}\, y_j \qquad (9.76)$$

$$h_{ii} = \frac{1}{n} + (\underset{\sim}{x}_i - \underset{\sim}{\bar{x}})' \, (\chi'\chi)^{-1} \, (\underset{\sim}{x}_i - \underset{\sim}{\bar{x}})$$

où χ est la sous-matrice nxp obtenue de la matrice X en éliminant la première colonne de 1 ($x_{i0} = 1$ tout i) et

$$\underset{\sim}{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$$

$$\underset{\sim}{\bar{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)$$

La quantité h_{ii} est appelée le "levier" de la i-ième observation et mesure l'importance de $\underset{\sim}{x}_i$ dans la détermination de $\hat{y}_i$. On constate que si $h_{ii} \to 1$ alors $\hat{y}_i \to y_i$. On peut interpréter h_{ii} comme une distance entre le point $\underset{\sim}{x}_i$ et le centre $\underset{\sim}{\bar{x}}$. En fait, h_{ii} = constante est l'équation d'un ellipsoïde centrée en $\underset{\sim}{\bar{x}}$ dans l'espace des variables explicatives $x_1, x_2, \ldots, x_p$.

Élimination de la i-ième observation $(x_{i1},\ldots,x_{ip},y_i)$

Posons

$\underset{\sim}{Y}_{(i)}$: vecteur (n-1)x1 formé de toutes les observations sur $\underset{\sim}{Y}$ sauf Y_i

$$\underset{\sim}{Y}_{(i)} = (y_i,\ldots,y_{i-1},y_{i+1},\ldots,y_n)'$$

$X_{(i)}$: sous matrice (n-1) par (p+1) après élimination de la i-ième rangée de X

$$\hat{\underset{\sim}{\beta}}_{(i)} = (X'_{(i)}\, X_{(i)})^{-1}\, X'_{(i)}\, \underset{\sim}{Y}_{(i)} \qquad (9.77)$$

$$\hat{Y}_{(i)} = \underset{\sim}{x}_i\, \hat{\underset{\sim}{\beta}}_{(i)} \qquad (9.78)$$

$$\hat{\underset{\sim}{Y}}_{(i)} = X_{(i)}\, \hat{\underset{\sim}{\beta}}_{(i)}$$

$$\sigma^2_{(i)} = \frac{(\underset{\sim}{Y}_{(i)} - \hat{\underset{\sim}{Y}}_{(i)})'\,(\underset{\sim}{Y}_{(i)} - \hat{\underset{\sim}{Y}}_{(i)})}{m-p-2} \qquad (9.79)$$

$$rse_{(i)} = \frac{Y_i - \hat{Y}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} \qquad (9.80)$$

(résidu studentisé externe)

$$rs_i = \frac{Y_i - \hat{Y}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

(résidu studentisé interne)

$$D_i = \frac{(\underset{\sim}{Y}_{(i)} - \underset{\sim}{\hat{Y}})' \, (\underset{\sim}{Y}_{(i)} - \underset{\sim}{\hat{Y}})}{\hat{\sigma}^2 \, (p+1)} \qquad (9.81)$$

(distance de Cook)

$$= \frac{(\underset{\sim}{\beta}_{(i)} - \underset{\sim}{\hat{\beta}})' \, (X'X) \, (\underset{\sim}{\beta}_{(i)} - \underset{\sim}{\hat{\beta}})}{\hat{\sigma}^2 (p+1)}$$

$$= \frac{1}{p+1} \, (rs_i)^2 \, \frac{h_{ii}}{1-h_{ii}}$$

D_i s'appelle la "distance de Cook" et nous en donnons trois différentes expressions. D_i peut s'interpréter comme une mesure de distance normalisée entre $\underset{\sim}{\hat{Y}}_{(i)}$ et $\underset{\sim}{\hat{Y}}$ ou encore comme une distance entre $\underset{\sim}{\hat{\beta}}_{(i)}$ et $\underset{\sim}{\hat{\beta}}$.

La dernière expression de D_i montre que la distance de Cook dépend

- du résidu studentisé interne rs_i reflétant un manque d'ajustement
- du "levier" h_{ii} reflétant la distance du point $\underset{\sim}{x}_i$ par rapport à $\underset{\sim}{\bar{x}}$

On identifie les observations <u>éventuellement influentes</u> par l'examen des trois quantités rs_i, h_{ii} et D_i selon les critères suivants:

$$|rs_i| \geq 2$$

$$h_{ii} \geq 2(p+1)/n \qquad (9.82)$$

$$\max (D_1, D_2, \ldots, D_n)$$

Après l'identification des observations satisfaisant à l'un des critères (9.82), on refait les calculs en excluant ces observations, et on compare le modèle obtenu avec celui basé sur toutes les données. On pourra alors conclure si les observations sont influentes ou non, c'est-à-dire si le modèle est stable. Il s'agit de porter un jugement comme on en fait souvent en analyse statistique.

Exemple 9.7: Consommation d'essence au USA en 1972

Les variables:

ETAT: état Américain

CONS: consommation annuelle per capita (gal.)

PLIC: proportion de la population ayant un permis de conduire

REV: revenu per capita (000$)

TAXE: coût taxe essence (0.00$/gal.)

LONG: longueur totale des routes principales (000 mi)

On propose le modèle:

$$CONS = \beta_0 + \beta_1 * PLIC + \beta_2 * REV + \beta_3 * TAXE + \beta_4 * LONG + \varepsilon$$

Les données sont présentées au tableau 9.7.

Tableau 9.7: consommation essence USA en 1972 par état

OBS	ETAT	CONS	PLIC	REV	LONG	TAXE
1	AL	554	0.51	3.333	6.594	7.0
2	AP	627	0.55	3.357	4.121	7.5
3	AZ	632	0.60	4.300	3.635	7.0
4	CA	524	0.59	5.002	9.794	7.0
5	CN	456	0.57	5.342	1.333	10.0
6	CO	587	0.63	4.449	4.639	7.0
7	DE	539	0.60	4.983	0.602	8.0
8	FI	574	0.56	4.188	5.975	8.0
9	CA	631	0.58	3.846	9.061	7.5
10	IA	634	0.59	4.318	10.340	7.0
11	ID	648	0.66	3.635	3.274	8.5
12	IL	471	0.52	5.126	14.186	7.5
13	IN	579	0.53	4.391	5.939	8.0
14	KS	649	0.66	4.593	7.834	7.0
15	KY	588	0.49	3.601	4.650	9.0
16	LA	487	0.49	3.528	3.495	8.0
17	MA	414	0.53	4.870	2.351	7.5
18	MD	464	0.51	4.897	2.449	9.0
19	ME	541	0.52	3.571	1.976	9.0
20	MT	524	0.57	4.187	6.930	7.0
21	MN	565	0.61	4.332	8.159	7.0
22	MO	602	0.57	4.206	8.508	7.0
23	MS	577	0.58	3.063	6.524	8.0
24	MT	703	0.59	3.897	6.385	7.0
25	NC	566	0.54	3.721	4.746	9.0
26	ND	713	0.54	3.718	4.725	7.0
27	NE	640	0.68	4.341	6.010	8.5
28	NH	523	0.57	4.092	1.250	9.0
29	NJ	466	0.55	5.126	2.138	8.0
30	NM	698	0.56	3.656	3.985	7.0
31	NV	781	0.67	5.215	2.302	6.0
32	NY	343	0.45	5.319	11.868	8.0
33	OH	498	0.55	4.512	8.507	7.0
34	OK	643	0.63	3.802	7.834	6.6
35	OR	609	0.62	4.296	4.083	7.0
36	PA	463	0.53	4.447	8.577	8.0
37	RI	410	0.54	4.399	0.431	8.0
38	SC	577	0.55	3.448	5.339	8.0
39	SD	865	0.72	4.716	5.915	7.0
40	TN	571	0.52	3.640	6.905	7.0
41	TX	640	0.57	4.045	17.782	5.0
42	UT	591	0.51	3.745	2.611	7.0
43	VA	547	0.52	4.258	4.686	9.0
44	VT	560	0.58	3.865	1.586	9.0
45	WI	507	0.55	4.207	6.580	7.0
46	WN	509	0.57	4.476	3.942	9.0
47	WV	460	0.55	4.574	2.619	8.5
48	WY	968	0.67	4.345	3.905	7.0

Dans un première analyse on obtient

$$CONS = 357.84 + 1335.25*PLIC - 63.36*REV - 33.88*TAXE - 2.62*LONG$$

$$R^2 = 0.67 \qquad F = 21.62$$

Mais le modèle n'est pas satisfaisant car le coefficient de LONG n'est pas significativement différent de zéro. Dans une deuxième analyse nous avons fait l'ajustement avec les variables PLIC, REV et TAXE. Il résulte

$$CONS = 282.17 + 1376.13*PLIC - 64.90*REV - 28.10*TAXE$$

$$R^2 = 0.66 \qquad F = 28.91$$

Ce nouveau modèle est significatif ainsi que chacun des coefficients du modèle. Effectuons une analyse de stabilité de notre modèle en identifiant les observations ayant les plus grandes valeurs de rs_i, h_{ii} et D_i. Un examen du tableau 9.8 nous indique l'état du Wyoming (i=48) est celui qui a la plus grande distance de Cook, soit D_i=0.362, et cela est dû à un grand résidu de 3.79. C'est aussi dans cet état que la consommation est la plus élevée soit 968.

Tableau 9.8: identification des points influents, exemple 9.7

i	Y_i	$\hat{Y}_i$	$r_i = Y_i - \hat{Y}_i$	rs_i	h_{ii}	D_i
1	554	571.0	-16.9	-0.27	0.11	0.002
2	627	610.4	16.6	0.26	0.07	0.001
3	632	632.1	- 0.0	-0.00	0.03	0.000
4	524	572.7	-48.7	-0.75	0.07	0.011
5	456	438.8	17.2	0.29	0.23	0.006
6	587	663.7	-76.7	-1.17	0.06	0.018
7	539	559.6	-20.6	-0.31	0.06	0.002
8	574	556.2	17.8	0.27	0.02	0.000
9	631	619.9	11.0	0.17	0.03	0.000
10	634	617.1	16.9	0.26	0.03	0.001
11	648	715.6	-67.6	-1.10	0.16	0.059
12	471	454.3	16.7	0.26	0.11	0.002
13	579	501.7	77.3	1.17	0.03	0.013
14	649	695.6	-46.6	-0.72	0.08	0.012
15	533	469.8	63.2	1.00	0.17	0.030
16	487	502.7	-15.7	-0.24	0.09	0.001
17	414	484.7	-70.7	-1.09	0.07	0.023
18	464	413.2	50.7	0.80	0.11	0.019
19	541	513.1	27.9	0.44	0.09	0.005
20	524	598.1	-74.1	-1.12	0.03	0.011
21	565	643.7	-78.7	-1.20	0.04	0.014
22	602	596.9	5.1	0.08	0.03	0.000
23	577	656.7	-79.7	-1.27	0.12	0.056
24	703	644.4	58.5	0.89	0.04	0.008
25	566	530.9	35.1	0.54	0.08	0.007
26	713	587.3	125.7	1.93	0.06	0.056
27	640	697.3	-57.3	-0.93	0.16	0.040
28	523	548.1	-25.1	-0.39	0.07	0.003
29	466	481.5	-15.5	-0.24	0.08	0.001
30	698	618.8	79.2	1.21	0.05	0.021
31	781	697.1	83.9	1.38	0.18	0.104
32	343	331.4	11.6	0.20	0.23	0.003
33	498	549.5	-51.5	-0.79	0.05	0.008
34	643	716.9	-73.9	-1.15	0.08	0.027
35	609	659.8	-50.8	-0.77	0.04	0.007
36	463	498.1	-35.0	-0.53	0.04	0.003
37	410	515.0	-10.0	-1.59	0.03	0.020
38	577	590.4	-13.4	-0.21	0.06	0.001
39	865	770.2	94.8	1.56	0.18	0.137
40	571	564.8	6.2	0.10	0.07	0.000
41	640	663.5	-23.5	-0.39	0.20	0.000
42	591	544.2	46.8	0.73	0.08	0.011
43	547	468.5	78.5	1.21	0.07	0.027
44	560	576.6	-16.5	0.26	0.08	0.002
45	507	569.3	-62.2	-0.95	0.04	0.009
46	509	523.1	-14.1	-0.22	0.07	0.001
47	460	503.3	-43.3	-0.66	0.05	0.005
48	968	725.5	242.5	3.79	0.09	0.362

Si on refait l'analyse en excluant l'état du Wyoming on obtient:

$$CONS = 370.22 + 1194.35*PLIC - 64.43*REV$$

$$- 27.06*TAXE$$

$$R^2 = 0.69 \qquad F = 32.09$$

Les résultats sont sensiblement les mêmes avec ou sans l'état du Wyoming. On en conlut que cet état n'est pas un point influent dans la détermination de l'équation. Cela s'explique en partie par le fait que le "levier" de cet état est faible (0.09) et même avec un grand résidu studentisé (3.79), la distance de Cook n'est pas très grande (0.36).

Mais on ne peut considérer aucun des deux derniers modèles comme très satisfaisant car le R^2 est relativement faible. Il faudrait faire une analyse graphique des résidus pour vérifier si le modèle ne souffre pas d'anomalies relativement aux hypothèses de base.

9.6 DÉTECTION DE VARIABLES COLINÉAIRES

Dans l'ajustement d'un modèle de régression multiple, il arrive assez fréquemment qu'une ou plusieurs variables explicatives soient une quasi combinaison linéaire des autres variables explicatives. Cette situation est connue sous le nom de multicolinéarité. Nous allons proposer des diagnostics pour détecter de telles situations. Il existe essentiellement deux méthodes pour pallier à ce problème soit, la construction de modèles par la sélection d'un sous-ensemble de variables ou encore l'utilisation de la régression pseudo-orthogonale. Les techniques de sélection de variables seront exposées à la section 9.7.

Le problème et ses conséquences

Dans le modèle de régression multiple

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

chaque coefficient β_j représente un taux de changement de Y lorsque X_j varie d'une unité, les autres variables étant constantes. (effet marginal).

Cette interprétation est correcte si les variables explicatives sont non-corélées. Dans les études et modèles de régression, les variables explicatives ne sont pas toujours sous le contrôle total de l'expérimentateur comme dans des études lors d'expériences scientifiques ou industrielles. Les variables explicatives sont toujours plus ou moins colinéaires avec comme conséquences possibles:

- l'instabilité des coefficients $\hat{\beta}_j$ causée par l'inversion de la matrice X'X quasi-singulière. Cela peut se manisfester par exemple par des coefficients estimés $\hat{\beta}_j$ dont le signe est contraire à ce que l'on pouvait attendre

- la variance des coefficients $\hat{\beta}_j$ est grande avec comme conséquence que les tests de nullité des coefficients ne sont pas rejetés. On peut, par exemple, se retrouver dans la situation observée avec les données de béton de ciment (cf exemples 9.5 et 9.6): le test F global de la nullité conjointe est rejeté, le R^2 est élevé mais aucune des variables explicatives ne possède un coefficient significativement différent de zéro!

Il est donc nécessaire de proposer des

- diagnostics permettant de détecter une situation de variables colinéaires,

- alternatives pour effectuer une analyse de régression en présence de variables colinéaires.

Détection de variables colinéaires

Il existe plusieurs approches pour détecter si nous sommes en présence d'une situation de variables colinéaires: examen de la matrice R des corrélations, examen de la matrice inverse R^{-1}, examen des valeurs propres de R, décomposition de X en valeurs singulières. Cette dernière approche est supérieure aux trois autres et est disponible dans la procédure REG de SAS pour effectuer de la régression multiple.

Examen de la matrice de corrélation R

La présence d'un ou plusieurs coefficients de corrélation élevés (disons $|r| \geq 0.95$) est une condition généralement suffisante (mais pas nécessaire!) pour engendrer une situation de variables colinéaires.

Examen de la matrice inverse R^{-1}

Les éléments de la diagonale principale de la matrice R^{-1} sont appelés les facteurs inflationnaires de la variance (valeurs d'inflation de la variance). Ils s'écrivent:

$$VIF_j = \frac{1}{1-R_j^2} \qquad (9.83)$$

où R_j^2 est le carré du coefficient de corrélation multiple de la variable X_j avec les autres variables explicatives. La variance de $\hat{\beta}_j$ peut alors devenir très grande si $R_j{}^2$ est près de 1 puisque

$$VAR(\hat{\beta}_j) = \frac{1}{1-R_j{}^2} \quad \frac{1}{\sum_{i=1}^{n} (x_{ij}-\bar{x}_j)^2} \sigma^2 \qquad (9.84)$$

Par exemple

R_j^2	0	0.50	0.90	0.95	0.99
VIF_j	1	1.33	5.26	10.26	100

Examen des valeurs propres de R

Soient $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$ les valeurs propres de la matrice R. La présence de "petites" valeurs propres peut constituer un critère de quasi-singularité. Si toutes les variables sont non-corélées, alors $R=I_p$, la matrice identité d'ordre p. Dans ce cas $\lambda_i=1$ $i=1,2,\ldots p$ et

$$\sum_{i=1}^{p} \frac{1}{\lambda_i} = p$$

On a donc proposé le critère suivant basé sur une comparaison de la matrice R avec la matrice identité I_p:

$$\sum_{i=1}^{p} \frac{1}{\lambda_i} > 5p \qquad (9.85)$$

Si (9.85) est satisfaite, on se déclare en situation de variables colinéaires en ayant donné un sens à "petites valeurs propres".

Décomposition de X en valeurs singulières

Les méthodes précédentes basées sur la matrice R ne permettent pas toujours de détecter facilement une situation de variables colinéaires. Une nouvelle méthode proposée par Belsley, Kuh et Welsh utilise les nombres indices et la décomposition de la variance de chaque estimation de $\hat{\beta}_j$. Cette méthode repose sur la décomposition en valeurs singulières de la matrice X.

On peut montrer que toute matrice X (nxp) peut s'écrire:

$$X = UDV \qquad (9.86)$$

où les matrices U (nxr) et V (rxp) sont orthogonales:

$$U'U = I_r \qquad V'V = I_p$$

et D est une matrice diagonale

$$D = \text{diag}(\mu_1, \mu_2, \ldots, \mu_r)$$

Les nombres μ_i sont appelés les valeurs singulières de X et elles sont liées aux valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_p$ de X'X:

$$D^2 = (\mu_1^2,\ldots,\mu_p^2) = (\lambda_1,\lambda_2,\ldots,\lambda_p) \quad \text{i.e. } \lambda_j = \mu_j{}^2$$

L'équation (9.86) est dite décomposition en valeurs singulières de la matrice X. Les ratios

$$\eta_j = \frac{\mu_{max}}{\mu_j} = \frac{\mu_1}{\mu_j} \tag{9.87}$$

où $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_p$ sont les valeurs singulières indexées en ordre de grandeur décroissant et sont appelés les NOMBRES INDICES

De l'équation (9.86), on a

$$(X'X)^{-1} = VD^{-2}\ V$$

et il s'ensuit que

$$\text{VAR}(\hat{\underset{\sim}{\beta}}) = (X'X)^{-1}\sigma^2 = VD^{-2}V\sigma^2$$

$$\text{VAR}(\hat{\beta}_j) = \sigma^2 \sum_{k=1}^{p} \frac{v_{jk}^2}{\mu_j^2} \tag{9.88}$$

où $V = [v_{jk}]$

Posons

$$\Phi_{jk} = v_{jk}^2/\mu_j^2 \tag{9.89}$$

$$\Phi_j = \sum_{k=1}^{p} \Phi_{jk} \tag{9.90}$$

$$\pi_{jk} = \Phi_{jk}/\Phi_j \tag{9.91}$$

La quantité π_{jk} comprise entre 0 et 1 s'appelle la VARIANCE-PROPORTION DE β_j associée à la valeur singulière μ_j. Le diagnostic de multicolinéarité est basé sur les nombres indices et les nombres variances-proportions tel que présentés dans le tableau 9.9.

Tableau 9.9 étude de la multicolinéarité

valeurs singulières en ordre décroissant	nombres indices	nombres variances-proportions β_1 β_2 ... β_p
$\mu_1=\sqrt{\lambda_1}$	1	π_{11} π_{12} ... π_{1p}
$\mu_2=\sqrt{\lambda_2}$	η_2	π_{21} π_{22} ... π_{2p}
.	.	
.	.	
.	.	
$\mu_p=\sqrt{\lambda_p}$	η_p	$\frac{\pi_{p1}}{1}$ $\frac{\pi_{p2}}{1}$ ... $\frac{\pi_{pp}}{1}$

Le diagnostic de multicolinéarité proposé par Belsley, Kuh et Welsh est

(a) $\eta_k \geq 30$ associé avec

(b) au moins deux nombres variances-proportions supérieurs à 0.50

Ce diagnostic de multicolinéarité est disponible avec l'option COLLIN de la procédure REG de SAS.

Exemple 9.8: données de béton du chapitre 1, suite de l'exemple 9.5

Nous avons p=5 variables explicatives en incluant la variable $X_0=1$. Le tableau de l'analyse de multicolinéarité est

valeurs singulières	nombres indices	nombres variances-proportions				
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
μ_j	η_j					
2.03	1	0.00	0.00	0.00	0.00	0.00
0.74	2.74	0.00	0.01	0.00	0.00	0.00
0.54	3.78	0.00	0.06	0.00	0.00	0.00
0.19	10.46	0.00	0.00	0.00	0.05	0.00
0.008	249.58	0.99	0.93	0.99	0.95	0.99
		1	1	1	1	1

On constate la présence d'un nombre indice de 249.58 auquel est associé 5 nombres variances-proportions plus grands que 0.50. Il s'agit donc d'une situation de multicolinéarité.

Méthodes pour vaincre la multicolinéarité

On a proposé plusieurs méthodes pour vaincre le problème de variables colinéaires:

- élimination de variables par une technique de sélection
- utilisation de la régression pseudo-orthogonale RIDGE qui conserve toutes les variables mais propose des estimations biaisées pour les coefficients de régression
- utilisation de la technique de régression sur les composantes principales formées par des combinaisons linéaires optimales des variables originales.

9.7 Techniques de sélection de variables

La régression est souvent employée comme outil d'exploration afin de déterminer les "meilleures" équations entre une variable Y et une liste de variables explicatives $X_1, X_2, \ldots, X_p$. C'est le cas lorsque l'on cherche à "construire des modèles". C'est aussi une stratégie qui peut être employée lorsque l'on veut éliminer des variables explicatives dans une situation de multicollinéarité.

Critères et coefficients pour juger les équations

Soit p: le nombre total de variables explicatives disponibles:

$$k' = \begin{cases} 1 + p & : \text{ si le modèle contient une constante générale} \\ p & : \text{ autrement} \end{cases}$$

k: le nombre de variables explicatives dans l'équation; peut inclure la variable constante associée à β_0

$$1 \leq k \leq k'$$

La comparaison des équations est faite à l'aide de l'un des 4 critères suivants:

- maximum du coefficient de détermination R_k^2
- maximum du coefficient de détermination ajusté $\bar{R}_k^2$
- minimum du carré moyen résiduel $\hat{\sigma}_k^2$
- coefficient de Mallows C_k voisin de k

Définitions

SCT_k : somme des carrés totaux d'une équation basée sur k coefficients $\beta_0 \ldots, \beta_{k-1}$

SCE_k : somme des carrés due à l'erreur

SCM_k : somme des carrés due au modèle

$$R_k^2 = \frac{SCM_k}{SCT_k} = 1 - \frac{SCE_k}{SCT_k}$$

$$\bar{R}_k^2 = \frac{(n-1)}{(n-k)} (1-R_k^2) \qquad (9.92)$$

$$\hat{\sigma}_k^2 = \frac{SCE_k}{n-k}$$

$$C_k = \frac{SCE_k}{\hat{\sigma}^2_{k'}} + 2k-n$$

$$F_k = \frac{(SCE_k - SCE_{k'})/(k'-k)}{SCE_{K'}/(n-k')}$$

Le coefficient de Mallows C_k est une estimation de l'erreur quadratique moyenne tandis que F_k est la statistique qui permet de tester l'hypothèse nulle

$$H_N : \beta_{k+1} = \ldots = \beta_p = 0$$

pour comparer le modèle global

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

avec le modèle restreint

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + 0.X_{k+1} + \ldots + 0.X_p$$

Relations entre les coefficients

$$F_k = 1 + \frac{C_k - k}{k' - k}$$

$$C_k = (k'-k)\ (F_k - 1) + k$$

$$C_k < k \longleftrightarrow F_k < 1$$

$$C_k < k' \longleftrightarrow F_k < 2$$

$$\bar{R}^2_k < R^2_k$$

$$k_1 < k_2 \rightarrow R_{k1} < R_{k2}$$

Les coefficients R^2_k et C_k sont très utilisés pour comparer les équations. D'autre part on a développé plusieurs stratégies dans la recherche des "bonnes équations".

Examen de toutes les équations

Si l'on dispose de k' variables explicatives, il y a $2^{k'}$ équations candidates; par exemple si k'=20 on a 1,048,576 équations. Actuellement les progiciels comme SAS peuvent assez facilement traiter 25 variables explicatives, soit au delà de 33 000 000 équations dans un temps de quelques minutes.

Sélection ascendante ("forward selection")

On commence avec aucune variable dans le modèle. Pour chaque variable explicative, on calcule un test F pour déterminer si son coefficient de régression est significativement différent de zéro. On retient la variable ayant le plus grand F en autant que celle-ci dépasse une valeur minimale, e.g. le 75-ième percentile d'une distribution de Fisher. On continue l'évaluation avec une nouvelle série de rapports F de la variable résiduelle (résidu de la variable Y avec la première variable choisie) et chacune des variables non choisies lors de la première étape. La variable ayant le plus grand F est retenue. Ainsi de suite. Les variables sont ajoutées une à une jusqu'à ce qu'aucune des variables ne produise un test F significatif. Lorsqu'une variable entre dans le modèle, elle n'en sort jamais.

Elimination descendante ("backward elimination")

On commence avec le modèle contenant toutes les variables. À la première étape on élimine la variable apportant la plus petite contribution i.e. le plus petit rapport F lorsque celui-ci ne dépasse pas un seuil choisi e.g. 75-ième percentile d'une loi de Fisher. La procédure continue jusqu'à ce que toutes les variables possédent un coefficient dont le test F est significatif. Lorsqu'une variable est éliminée à une étape, elle le reste pour toutes les autres.

Progressive ("stepwise")

Cette méthode est une variante de la méthode de sélection avant et diffère par le fait qu'une variable retenue dans le modèle à une étape ne demeure pas nécessairement dans le modèle à une étape ultérieure. Après avoir retenu une variable à une étape, on examine toutes les autres variables retenues à des étapes précédentes. Celles qui ne produisent pas un test F significatif sont éliminées.

Maximum R^2 (option MAXR dans la procédure STEPWISE de SAS)

Cette méthode développée par J. Goodnigh (président de SAS) est supérieure à la méthode progressive et elle est presque aussi bonne que l'examen de toutes les équations. A la première étape, la variable retenue est celle qui possède le plus grand R^2. On détermine ensuite une deuxième variable produisant un accroissement maximal de R^2. On obtient alors provisoirement un modèle à 2 variables. Ce modèle est comparé, au moyen du R^2, à tous les autres modèles à 2 variables. On obtient ainsi le "meilleur" modèle à 2 variables. On continue ainsi pour les modèles à 3 variables, 4 variables etc.

Minimum R^2 (option MINR dans la procédure STEPWISE de SAS)

Méthode analogue à maximum R^2 mais qui procède par élimination descendante en enlevant les variables produisant la plus petite diminution de R^2.

Exemple 9.9: Taux de mortalité sur les voies rapides

Les données proviennent d'une étude menée en 1973 afin d'étudier l'influence des différentes caractéristiques physiques des voies rapides sur le taux de mortalité causée par des accidents de voitures. Le taux de mortalité est mesuré par le nombre de morts par millions de milles parcourus.

Les variables:

Y	= TAUX	= TAUX DE MORTALITÉp0
X_1	= LONG	= LONGUEUR SECTION (MI)
X_2	= TRAF	= TRAFFIC JOURNALIER MOYEN EN 1000
X_3	= VOLCAM	= % VOLUME CAMIONS/VOLUME TOTAL
X_4	= VITES	= VITESSE MAXIMUM PERMISE (M/H)
X_5	= LARG	= LARGEUR SECTION (PI)
X_6	= LARGACT	= LARGEUR ACCOTEMENT (PI)
X_7	= NINTER	= NOMBRE ECHANGEURS PAR MILLE
X_8	= NSIG	= NOMBRE SIGNAUX PAR MILLE
X_9	= NACC	= NOMBRE ACCES PAR MILLE
X_{10}	= NVOIES	= NOMBRE VOIES 2 DIRECTIONS

Tableau 9.10: données exemple 9.9

i	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	4.58	4.99	69	8	55	12	10	1.20	0	4.6	8
2	2.86	16.11	73	8	60	12	10	1.43	0	4.4	4
3	3.02	9.75	49	10	60	12	10	1.54	0	4.7	4
4	2.29	10.65	61	13	65	12	10	0.94	0	3.8	6
5	1.61	20.01	28	12	70	12	10	0.65	0	2.2	4
6	6.87	5.97	30	6	55	12	10	0.34	1.84	24.8	4
7	3.85	8.57	46	8	55	12	8	0.47	0.70	11.0	4
8	6.12	5.24	25	9	55	12	10	0.38	0.38	18.5	4
9	3.29	15.79	43	12	50	12	4	0.95	1.39	7.5	4
10	5.88	8.26	23	7	50	12	5	0.12	1.21	8.2	4
11	4.20	7.03	23	6	60	12	10	0.29	1.85	5.4	4
12	4.61	13.28	20	9	50	12	2	0.15	1.21	11.2	4
13	4.80	5.40	18	14	50	12	8	0	0.56	15.2	2
14	3.85	2.96	21	8	60	12	10	0.34	0	5.4	4
15	2.69	11.75	27	7	55	12	10	0.26	0.60	7.9	4
16	1.99	8.86	22	9	60	12	10	0.68	0	3.2	4
17	2.01	9.78	19	9	60	12	10	0.20	0.10	11.0	4
18	4.22	5.49	9	11	50	12	6	0.18	0.18	8.9	2
19	2.76	8.63	12	8	55	13	6	0.14	0	12.4	2
20	2.55	50.31	12	7	60	12	10	0.05	0.99	7.8	4
21	1.89	40.09	15	13	55	12	8	0.05	0.12	9.6	4
22	2.34	11.81	8	8	60	12	10	0	0	4.3	2
23	2.83	11.39	5	9	50	12	8	0	0.09	11.1	2
24	1.81	22.00	5	15	60	12	7	0	0	6.8	2
25	9.23	3.58	23	6	40	12	2	0.56	2.51	53.0	4
26	8.60	3.23	13	6	45	12	2	0.31	0.93	17.3	2
27	8.21	7.73	7	8	55	12	8	0.13	0.52	27.3	2
28	2.93	14.41	10	10	55	12	6	0	0.07	18.0	2
29	7.48	11.54	12	7	45	12	3	0.09	0.09	30.2	2
30	2.57	11.10	9	8	60	12	7	0	0	10.3	2
31	5.77	22.09	4	8	45	11	3	0	0.14	18.2	2
32	2.90	9.39	5	10	55	13	1	0	0	12.3	2
33	2.97	19.49	4	13	55	12	4	0	0	7.1	2
34	1.84	21.01	5	12	55	10	8	0	0.10	14.0	2
35	3.78	27.16	2	10	55	12	3	0.04	0.04	11.3	2
36	2.76	14.03	3	8	50	12	4	0.07	0	16.3	2
37	4.27	20.63	1	11	55	11	4	0	0	9.6	2
38	3.05	20.06	3	11	60	12	8	0	0	9.0	2
39	4.12	12.91	1	10	55	12	3	0	0	10.4	2

Première analyse: Y régressé sur $X_1, X_2, \ldots, X_{10}$

On obtient

$$\hat{Y} = 11.85 - 0.043*LONG - 0.0006*TRAF - 0.15*VOLCAM - 0.065*VITES - 0.28*LARG + 0.24*NINTER + 0.43*NSIG + 0.08*NACC + 0.06*NVOIES$$

$R^2 = 0.726$ $\quad \hat{\sigma} = 1.41$ $\quad F = 7.81$

Mais l'équation n'est pas satisfaisante pour plusieurs raisons:

- le signe des coefficients révèle des faits troublants: le taux de mortalité diminue avec une augmentation de traffic, camions, vitesse, largeur, échangeurs et une diminution de signalisation
- seul le coefficient de NACC est significativement différent de zéro
- il existe une situation de variables multicollinéaires
- 4 résidus sont supérieurs à 2, soit pour les observations 25,26,27 et 34
- l'observation 25 possède une distance de Cook de 1.12

Deuxième analyse: recherche de bonnes équations selon le maximum de corrélation

Tableau 9.11: trois meilleures équations

k	R^2_k	$\bar{R}^2_k$	$\hat{\sigma}^2_k$	C_k	variables
1	0.3186	0.3002	2.760	37.327	NSIG
1	0.4637	0.4492	2.172	21.925	VITES
1	0.5655	0.5538	1.760	11.118	NACC
2	0.6185	0.5973	1.588	7.4928	NACC VITES
2	0.6326	0.6122	1.530	5.9983	NACC VOLCAM
2	0.6336	0.6133	1.525	5.8906	NACC LONG
3	0.6735	0.6455	1.398	3.6583	NACC LONG VOLCAM
3	0.6769	0.6492	1.384	3.2938	NACC LONG VITES
3	0.6774	0.6497	1.382	3.2468	NACC VOLCAM VITES
4	0.6936	0.6575	1.351	3.5284	NACC LONG VOLCAM NSIG
4	0.7087	0.6744	1.284	1.9198	NACC LONG VITES NSIG
4	0.7125	0.6787	1.267	1.5136	NACC LONG VITES VOLCAM
5	0.7182	0.6755	1.280	2.9139	NACC LONG VITES VOLCAM TRAF
5	0.7198	0.6773	1.273	2.7437	NACC LONG VITES VOLCAM NVOIES
5	0.7279	0.6867	1.236	1.8836	NACC LONG VITES VOLCAM NSIG
6	0.7295	0.6788	1.267	3.709	NACC LONG VITES VOLCAM NSIG TRAF
6	0.7299	0.6793	1.265	3.6673	NACC LONG VITES VOLCAM NSIG LARG
6	0.7305	0.6800	1.262	3.6094	NACC LONG VITES VOLCAM NSIG NINTER
7	0.7318	0.6713	1.297	5.4679	NACC LONG VITES VOLCAM NSIG LARGACT
7	0.7322	0.6718	1.295	5.4236	NACC LONG VITES VOLCAM NSIG NINTER LARGACT
7	0.7324	0.6720	1.294	5.4023	NACC LONG VITES VOLCAM NSIG NINTER LARG
8	0.7352	0.6646	1.323	7.1071	NACC LONG VITES VOLCAM NSIG LARG LARGACT NVOIES
8	0.7354	0.6649	1.322	7.0866	NACC LONG VITES VOLCAM NSIG LARG LARGACT TRAF
8	0.7356	0.6651	1.321	7.0688	NACC LONG VITES VOLCAM NSGIG NINTER LARG LARGACT
9	0.7357	0.6537	1.366	9.0525	NACC LONG VITES VOLCAM NSIG NINTER LARG LARGACT TRAF
9	0.7358	0.6538	1.366	9.045	NACC LONG VITES VOLCAM NSIG LARG LARGACT TRAF NVOIES
9	0.7362	0.6544	1.363	9.0004	NACC LONG VITES VOLCAM NSIG NINTER LARG LARGACT NVOIES
10	0.7362	0.6420	1.412	11	NACC LONG VITES VOLCAM NSIG NINTER LARG LARGACT NVOIES TRAF

Selon le critère du maximum de $\bar{R}^2_k$, l'équation retenue est celle contenant les variables NACC, LONG, VITES, VOLCAM, NSIG. La même équation est retenue selon le critère du minimum de $\hat{\sigma}^2$.

Troisième analyse: recherche de bonnes équations par la méthode séquentielle du MAXR

La méthode du MAXR identifie pour chaque valeur de k entre 2 et 10, l'équation ayant le coefficient de détermination R^2 maximal. Le résultat principal de cette recherche séquentielle est résumé dans le tableau 9.12.

Tableau 9.12: meilleures équations selon la méthode MAXR

étape	équation retenue	R^2	coefficients significatifs
1	1.98 + 0.160*NACC	0.565	oui
2	2.90 + 0.147*NACC - 0.055*LONG	0.633	oui
3	10.21 + 0.098*NACC - 0.098*VITES -0.220*VOLCAM	0.677	oui
4	10.05 + 0.095*NACC - 0.092*VITES - 0.176*VOLCAM - 0.041*LONG	0.712	oui
5	9.33 + 0.084*NACC - 0.087*VITES - 0.137*VOLCAM - 0.041*LONG + 0.478*NSIG	0.728	non
6	9.45 - 0.087*NACC - 0.092*VITES - 0.135*VOLCAM - 0.038*LONG + 0.44*NSIG + 0.27*NINTER	0.730	non
7	11.80 + 0.086*NACC - 0.089*VITES - 0.141*VOLCAM - 0.040*LONG + 0.445*NSIG + 0.269*NINTER - 0.202*LARG	0.732	non
8	11.90 + 0.087*NACC - 0.066*VITES - 0.153*VOLCAM - 0.042*LONG + 0.468*NSIG + 0.339*NINTER - 0.276*LARG - 0.057*LARGACT	0.735	non
9	----	0.736	non
10	----	0.736	non

9.10 Utilisation de SAS: procédures et exemples

Il existe plusieurs procédures SAS pour effectuer l'analyse de régression linéaire et non-linéaire.

Procédure SAS	description sommaire
REG	-régression linéaire multiple -analyse des résidus -identification des points influents -analyse de multicolinéarité
RSQUARE	identification des meilleurs modèles selon plusieurs critères
STEPWISE	construction de modèles selon plusieurs méthodes de sélection de variables
NLIN	ajuste des modèles non-linéaire dans les paramètres
RSREG	ajustement de modèles quadratiques et identification des points stationnaires
GLM	analyse de modèles linéaires dont les modèles de régression et les modèles polynomiaux
COXREGR	analyse de régression d'une variables censurée selon le modèle logistique de Cox
LAV	régression linéaire avec le critère du minimum de valeur absolue
LOGIST	ajustement du modèle de régression logistique de Cox pour une variable à expliquer de type logique (0-1)
RIDGEREG	régression de type RIDGE et produisant des estimations biaisées pour contourner le problème de la multicolinéarité
JACKREG	calcule des estimations robustes pour les coefficients de régression par la méthode du Jacknife
LEAPS	détermination de sous-ensembles de variables explicatives selon la technique du "leaps and bound".

remarque: Les procédures REG RSQUARE STEPWISE NLIN RSREG GLM sont décrites dans le manuel SAS: Statistics, version (1985).

Les autres procédures sont décrites dans le manuel SUGI. Supplemental Library User's Guide, version 5, (1986).

BIBLIOGRAPHIE

Ang A.S, Tangl J.H. (1971)

Probability Concepts in Engineering Planning and Design

volume 1, Basic Principles, John Wiley Inc

Bowker A. H., Lieberman, G.J. (1965)

Méthodes statistiques de 1'ingénieur Dunod, Paris
traduction de Engineering Statistics, Prentice Hall, 1959

Bethea, R.M., Dw-an B.J., Boullion T.L. (1985)

Satistical Methods for Engineers and
Scientists, second edition, Marcel Dekker
Inc

Blake I.F. (1979)

An Introduction to Applied Probability, John Wiley Inc

Benjamin J.R., Cornell A.C. (1970)

Probability, Statistics and Decision for Civil Engineers, McGraw-Hill

Banks J., Heikes R.G. (1984)

Handbook of Tables and Graphs for the Industrial Engineers and
Managers, Reston Publishing Company Inc

Belsley D.A., Kuh E. Welsh R.E. (1980)

Regression Diagnostics: Identifying Influential Data and Sources of
Collinearity, John Wiley Inc

Calot A. (1967)

Cours de calcul des probabilités Dunod, Paris

Draper N.R., Smith H. (1981)

Applied Regression Analysis, 2nd edition John Wiley Inc

Devore J. L. (1982)

Probability and Statistics for Engineering and the Sciences

Brooks/ Cole Publishing Company

Guttman I, Wilks S.S., Hunter J.S. (1982)

Introductory Engineering Statistics, third edition John Wiley inc

Haan C.T. < 1977)

Statistical Methods in Hydrology, Iowa State University Press

Hald A. (1952)

Statistical Theory with Engineering Applications John Wiley Inc

Hines W.W., Montgomery D.C. (1980)

Probability and Statistics in Engineering and Management Science, second edition, John Wiley Inc

Kennedy J.B., Neville A.M. (1986)

Basic Statistical Methods for Engineers and Scientists, third edition

Harper and Row, Publishers, Inc

Meyer S.L. (1975)

Data Analysis for Scientists and Engineers, John Wiley Inc

McCuen, R.H. (1985)

Statistical Methods for Engineers, Prentice-Hall

Miller I., Freund, J.E. (1985)

Probability and Statistics for Engineers, third edition Prentice-Hall

Weisberg S. (1985)

Applied linear Regression, 2nd edition John Wiley Inc

Walpole, R.E., Myers R.H. (1972)

Probability and Statistics for Engineers and Scientists

The Macmillan Company