Mise à niveau
Estimation et Tests statistiques
Master Statistique et econométrie

V. Monbet

Master 1 - 2012

Table des matières

1 Introduction

1.1 Notations

Dans la mesure du possible, nous utiliserons les notations suivantes dans ce cours :

- Variables aléatoires : lettres majuscules (ex : $X_1, \cdots, X_n$)
- Observations : lettres minuscules (ex : $x_1, \cdots, x_n$))
- Paramètres : lettres greques (ex : θ, μ, σ)

Estimation ponctuelle

Dans un premier temps, nous aborderons le problème de l'estimation ponctuelle de paramètres.

Exemple : On suppose que l'on a deux candidats A et B lors d'une élection. On cherche à prédire la proportion de votes pour A à partir d'un échantillon représentatif de taille n sélectionné par un institut de sondage. Chaque individu de l'échantillon donne son intention de vote. On modélise le choix A pour l'individu i par une variable aléatoire

$$X_i = 1 \text{ si } i \text{ vote A}, X_i = 0 \text{ sinon}$$

Les X_i suivent une loi de Bernouilli de paramètre π inconnu. On dispose de n observations $x_1, \cdots, x_n$ de $X_1, \cdots, X_n$. On cherche à inférer π à partir de l'échantillon $x_1, \cdots, x_n$. Une estimation naturelle est

$$\hat{\pi}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

On se pose alors des questions : Cette quantité estime t'elle bien le paramètre pi ? Peut-on lui associer une marge d'erreur ? Que se passe t'il si on dispose d'un échantillon plus grand (c'est à dire si n grandit) ? Existe t'il d'autres quantités qui donnerait une meilleure estimation de p ? En existe t'il une qui est optimale pour un critère bien choisi ?

1.2 Théorie de la décision : tests statistiques

Dans une seconde partie, nous introduirons le concept de test statistique.

Exemple : On teste l'efficacité d'un médicament contre le cholestérol. On dispose pour cela de n individus pour lesquels on a effectué deux mesures du taux de cholestérol, l'une avant et l'autre après le traitement. On note

- X : le taux de cholestérol avant le traitement,
- Y : le taux de cholestérol après le traitement.

On dispose donc des couples d'observations $(x_1, y_1), \cdots, (x_n, y_n)$ et on veut déterminer à l'aide de ces observations si

$$D = Y - X$$

est positif. Etant donnés les caractères aléatoires de D et de l'expérience, on décidera de l'efficacité du traitement si la moyenne observée

$$\bar{d}_n = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i)$$

est plus grande qu'un seuil. Ce seuil est défini en fonction du risque de se tromper qui est fixé par l'expérimentateur et de la taille n de l'échantillon.

2 Estimation

2.1 Généralités

A partir de données d'échantillons représentatifs, on va induire des résultats sur la population-mère (i.e. population dans laquelle les échantillons ont été prélevés). Plus exactement, soit θ un paramètre inconnu intervenant dans la loi de probabilité d'une variable aléatoire X et soient $x_1, \ldots, x_i, \ldots, x_n$ les n valeurs prises par la v.a. X dans un échantillon de taille n prélevé dans la population-mère. On appelle **estimateur** de θ, et l'on note T_n, la fonction qui aux valeurs x_i de l'échantillon fait correspondre la valeur du paramètre θ. On note la valeur numérique de cette estimation par

$$\hat{\theta} = T_n(x_1, \ldots, x_n)$$

Par définition, T_n est une fonction des réalisations d'une v.a., T_n est donc une v.a. dont on peut chercher à déterminer les caractéristiques (loi, ddp, FR, moments, ...).

Exemple : On observe un phénomène de production de pièces manufacturées. Chaque pièce est associée à une mesure (un indicateur de qualité par exemple). Comme on ne peut pas vérifier chaque mesure, on procède à un échantillonnage qui nous fournit donc un échantillon. Supposons que la connaissance de la nature de cet indicateur nous permet de faire l'hypothèse qu'il obéit à une loi de probabilité normale. Le problème est maintenant, au vue de l'échantillon $\{x_i\}$, de proposer une valeur pour la moyenne de cette loi normale. Il faut procéder à une estimation du paramètre vrai μ qui se traduit par la valeur $\hat{\mu}$. Il y a une infinité de manière possible parmi lesquelles on peut citer

- $\hat{\mu} = \frac{1}{n}\sum_i x_i$
- $\hat{\mu} =$médiane$\{x_i\}$
- $\hat{\mu} =$mode$\{x_i\}$
- $\hat{\mu} = x_7$

Quel est le meilleur estimateur de la moyenne ? Existe-t-il ?
Sur ce simple exemple, est résumé le problème fondamental de l'estimation : quelle est la définition mathématique de meilleur ? La réponse est simple, il n'en existe pas. Alors comment comparer les estimateurs. Pour cela, on se sert de plusieurs critères, le plus souvent liés au bon sens :

– le **biais** : On souhaite que l'estimation ne soit pas systématiquement décalée par rapport à la valeur vraie.
  Définition 1 - *Le biais*[1] *d'un estimateur* $\hat{\theta}$ *est défini par*

$$b_\theta(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta$$

  On dit qu'un estimateur est non biaisé ou sans biais si $b_\theta(\hat{\theta}) = 0$.

[1]bias

- la **précision**. Si l'on répète l'estimation sur un autre échantillon, on souhaite obtenir une estimation cohérente, donc peu de variation d'un échantillon à l'autre. On parlera aussi d'**efficacité**. La variance d'un estimateur représente sa précision. Pour tous les estimateurs (ayant même moyenne), il est possible de trouver celui dont la précision sera la meilleure, i.e. dont la variance sera la plus faible. On parle alors d'estimateur à variance minimum. Lorsque l'on compare deux estimateurs, on dira également que T_n est plus efficace que T_n^* si $V(T_n) < V(T_n^*)$.
  Définition 2 - *L'erreur en moyenne quadratique*[2] *(EMQ) de θ est définie par*

$$EMQ = E_\theta((\hat{\theta} - \theta)^2)$$

  On remarque que

$$\text{EMQ} = Var_\theta(\hat{\theta}) + b_\theta(\hat{\theta})^2$$

- la **convergence**. Si l'on peut estimer la valeur du paramètre sur toute la population-mère, la valeur de l'estimation obtenue doit être la valeur vraie du paramètre.
  Définition 3 - *Une suite d'estimateurs $\{\hat{\theta}_n\}$ est dite consistante pour θ si $\{\hat{\theta}_n\}$ converge en probabilité vers θ, c'est à dire si,*

$$\lim_{n\to+\infty} P_\theta[|\hat{\theta}_n - \theta| > \epsilon] = 0$$

  pour tout $\epsilon > 0$ et pour toute valeur de θ.
- la **robustesse**. Dans tout cas concrèt, il existe des sources de perturbations. On souhaite que l'estimation ne soit pas sensible à la présence de valeurs abérantes (outliers en anglais).

Ces différents critères ne sont pas forcément compatibles entre eux, et l'on retrouve des dilemmes classiques, précision vs robustesse, convergence vs complexité.

2.2 Estimation par la méthode des moments

Exemple de construction d'estimateur. Soit X une variable aléatoire telle $E[|X|^r] < \infty$ et $X_1, \cdots, X_n$ de même loi que X. Par la loi des grands nombres,

$$M_n^s = \frac{1}{n}\sum_{i=1}^{n} X_i^s \to E(X^s) = \mu_s \text{ si } X_1, \cdots, X_n \text{ indépendantes}$$

Aussi, si n est assez grand, on s'attend à ce que $M_n^s \simeq \mu_s$.
On appelle estimateur de θ obtenu par la méthode des moments la solution $\hat{\theta}$ du système d'équations

$$\begin{cases} \mu_1(\theta) = M_n^1 \\ \cdots \\ \mu_s(\theta) = M_n^s \end{cases}$$

Exemple - loi de Poisson. $\lambda = E(X)$. D'où $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

[2] mean square error

3 Estimation par intervalles

Un problème évident de l'estimation ponctuelle est de mesurer la qualité de l'estimateur. En effet, $P(\hat{\theta} = \theta)$ est faible (quand elle n'est pas nulle) et il est souvent difficile d'interpréter la précision donnée par la variance de l'estimateur. Ainsi il est souvent intéressant de donner une estimation par interval.

Définition 4 *Soit* $\mathbf{X} = (X_1, \cdots, X_n)$ *des vairables aléatoires dont la loi jointe dépend d'un paramètre* θ *réel. Et soient* $L(\mathbf{X}) < U(\mathbf{X})$ *deux statistiques. Alors l'interval (aléatoire)* $[L(\mathbf{X}), U(\mathbf{X})]$ *est appelé intervalle de confiance à 100p% pour* θ *si*

$$P[L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})] \geq p$$

pour tout θ *avec égalité pour au moins une valeur de* θ.

L'interprétation d'un intervalle de confiance à $100p\%$ est la suivante : si on repète des tirages un grand nombre de fois et qu'on associe un intervalle de confiance à chaque tirage, θ sera dans $100p\%$ de ces intervalles de confiance.

Exemple : intervalle de confiance pour la moyenne.
Soient $X_1, \cdots, X_n$ des variables aléatoires indépendantes et identiquement distribuées suivant une loi de Gauss de moyenne μ et de variance 1. Alors, $\sqrt{n}(\bar{X} - \mu)$ suit une loi de Gauss centrée réduite et

$$P[-1.96 \leq \sqrt{n}(\bar{X} - \mu) \leq 1.96] = 0.95.$$

Or l'évènement $[-1.96 \leq \sqrt{n}(\bar{X} - \mu) \leq 1.96]$ est identique à l'évènement $[\bar{X} - 1.96/\sqrt{n} \leq \mu \leq \bar{X} + 1.96/\sqrt{n}]$ at on a donc

$$P\left[\bar{X} - 1.96/\sqrt{n} \leq \mu \leq \bar{X} + 1.96/\sqrt{n}\right] = 0.95$$

et l'interval $\left[\bar{X} - 1.96/\sqrt{n}, \bar{X} + 1.96/\sqrt{n}\right]$ est un intervalle de confiance à 95% pour μ.
Si on suppose seulement que $X_1, \cdots, X_n$ des variables aléatoires indépendantes et identiquement distribuées de moyenne μ et de variance 1, alors

$$P\left[\bar{X} - 1.96/\sqrt{n} \leq \mu \leq \bar{X} + 1.96/\sqrt{n}\right] \approx 0.95$$

par le TCL.

4 Tests d'hypothèses

4.1 Généralités

Dans cette partie nous énonçons (ou rappelons) un certains nombre de généralités autour des tests d'hypothèse, l'objectif étant d'être capable de bien formuler un test. Considérons un exemple. Un aquaculteur a 114 poissons d'une certaine espèce dans un de ses bassins. On extrait un échantillon de 12 poissons afin de vérifier l'hypothèse selon laquelle la médiane de la longueur des poissons est de 220 mm. On observe les longueurs suivantes :

126	142	156	228	245	246	370	419	433	454	478	503

4.1.1 Hypothèses de test

En premier lieu, nous devons formuler les hypothèses. L'hypothèse que nous voulons vérifier sera appelée *hypothèse nulle* et on la notera H_0. Dans l'exemple, nous poserons

$$H_0 : \theta = 220$$

où θ représente ici la médiane de la longueur des poissons. Nous rassemblerons d'autre part l'ensemble des *hypothèses alternatives* sous H_1 :

$$H_1 : \theta \neq 220$$

Et nous parlerons de tester H_0 contre les alternatives bilatérales H_1 (sous H_1, θ peut être inférieur ou supérieur à 220).
On décide que θ est dans $\Theta_0 = \{220\}$ ou $\Theta_1 = \mathbb{R} - \{220\}$ à l'aide des observations $(x_1, \cdots, x_n)$. Pour cela on cherche une règle de décision qui prend la forme suivante.

- Si $S(x_1, \cdots, x_n) \in$ RC alors on rejette H_0
- Si $S(x_1, \cdots, x_n) \notin$ RC alors on ne rejette pas H_0

où S est une *statistique* (ou fonction) *de test* et RC est une *région critique*. Le plus souvent, S est appelée statistique de test et c'est une fonction d'un estimateur de θ.

4.1.2 Statistique de test

Une fois les hypothèses de test posées, nous devons choisir la **statistique de test** S afin de déterminer la région critique. On va réaliser ici un test naïf appelé *test du signe*. La statistique de test est construite de la façon suivante. Si la médiane est 220, il est également probable pour chaque poisson sélectionné d'être plus ou moins long que 220 mm. Puis on calcule S égale aux nombre d'individus plus long que 220. (Implicitement, on associe à chaque individu un signe - si sa longueur est inférieure à 220 et un signe $+$ sinon ; S est alors la somme des signes $+$). On aura besoin plus loin de connaître la loi de la statistique de test. Il est facile de voir que la loi de la statistique de test est ici une loi binomiale de paramètres $n = 12$ et $\pi = 1/2$.

4.1.3 Région de rejet, Niveau de signification

Région critique. On remarque aisément que l'on va rejetter H_0 si S est trop grande ou trop petite (dominance de signes $+$, ou de signes -) ; la région critique est donc de la forme $\{s$ tels que $s \notin [s_{\text{inf}}, s_{\text{sup}}]\}$. Les bornes s_{inf} et s_{sup} sont déterminées à l'aide de la loi binomiale de telle sorte que

$$P_{H_0}(S < s_{\text{inf}}) = \frac{\alpha}{2} \text{ et } P_{H_0}(S > s_{\text{sup}}) = \frac{\alpha}{2}$$

En s'aidant de la table 1, on trouve que $s_{\text{inf}} = 2$ et $s_{\text{inf}} = 10$ (car $P(x = 0) + P(x = 1) + P(x = 2) < .025$). La règle des tests d'hypothèse est de rejeter H_0 au niveau de signification 0.05 si et seulement si le résultat tombe dans la région de rejet.

La région complémentaire de tous les résultats hors de la région de rejet est appelée **région de non rejet** (ou **d'acceptation**) de l'hypothèse nulle.

En choisissant une région de rejet de probabilité inférieure ou égale au niveau de signification on adopte une attitude dite **conservatrice**.

4.1.4 Les deux espèces d'erreur

Lorsque l'on fait un test d'hypothèse, deux sortes d'erreur sont possibles. On peut rejeter l'hypothèse nulle alors qu'elle est vraie. Ceci se produit si la valeur de la statistique de test tombe dans la région de rejet alors que l'hypothèse H_0 est vraie.

La probabilité de cet évènement est le niveau de signification. On dira aussi que le niveau de signification est la probabilité de rejeter l'hypothèse nulle à tort.

Rejeter l'hypothèse nulle à tort constitue une **erreur de première espèce**.

Si nous ne rejetons pas l'hypothèse nulle alors qu'elle est fausse nous commettons une **erreur de seconde espèce**. C'est le cas si la valeur de la statistique de test tombe dans la région de non rejet (ou d'acceptation) alors que H_0 est fausse (c'est à dire si H_1 est vraie).

Lorsque l'alternative H_1 est de la forme $\theta \neq \theta_0$, notre θ peut prendre une infinité de valeurs ; et la probabilité de rejeter H_0 lorsqu'elle est fausse dépend beaucoup de la vrai valeur de θ (qui est inconnue !).

Lorsque la vraie valeur de θ est dans H_1, la probabilité d'obtenir un résultat dnas la région de rejet est appelée **puissance** du test de H_0 contre H_1.

La puissance d'un test dépend de plusieurs facteurs :

- le niveau de signification du test
- la vraie valeur du paramètre testé
- la taille de l'échantillon
- la nature du test utilisé

De manière générale, plus on tient compte d'informations pertinentes dans un test plus sa puissance est élevée.

Cependant, on ne peut calculer $P_{H_1}(S \notin [s_{\inf}, s_{\sup}])$ que si on choisit une alternative ponctuelle pour H_1 c'est à dire si on fixe la valeur de la médiane sous H_1. En pratique, on dispose généralement pas d'une telle information. On peut alors regarder comment varie la puissance pour différentes valeurs de la médiane. Par exemple, le graphique ci-dessous montre la puissance du test du signe quand π varie. La courbe en noir correspond à un risque de première espèce de 5% et la courbe en rouge à un risque de première espèce de 20%.

r	0	1	2	3	4	5	6
P	0.000	0.003	0.016	0.054	0.121	0.193	0.226
		7	8	9	10	11	12
		0.193	0.121	0.054	0.016	0.003	0.000

TAB. 1 – Probabilités binomiales $P(S = k)$, $n = 12$, $\pi = \frac{1}{2}$

5 Test unilatéral ou bilatéral

Dans l'exemple du nombre de pages dans les livres de la bibliothèque, nous avons posé des hypthèses de tests telles que l'alternative est bilatérale. C'est à dire que si l'on rejette l'hypothèse nulle, la médiane du nombre de page peut-être supérieure ou inférieure à 220.

Dans certains problèmes, il est plus pertinent de considérer une hypothèse alternative unilatérale. On pose alors

$$H_0 : \theta \leq \theta_0 \text{ contre } H_1 : \theta > \theta_0$$

ou

$$H_0 : \theta \geq \theta_0 \text{ contre } H_1 : \theta < \theta_0$$

La définition de la région de rejet du test dépend de la forme de l'hyopthèse alternative (voir TD 1, ex. 2).

Le choix d'un test unilatéral ou bilatéral dépend de la logique de la situation expérimentale et doit être fait avant d'inspecter les données.

5.1 Exemple 1 : test de Student

Un contrôle anti-dopage a été effectué sur 16 sportifs. On a mesuré la variable X de moyenne μ, qui est le taux (dans le sang) d'une certaine substance interdite. Voici les données obtenues :

0.35	0.4	0.65	0.27	0.14	0.59	0.73	0.13
0.24	0.48	0.12	0.70	0.21	0.13	0.74	0.18

La variable X est supposée gausienne et de variance $\sigma^2 = 0.04$. On veut tester, au niveau 5% l'hypothèse selon laquelle le taux moyen dans le sang de la population des sportifs est égal à 0.4.

On pose des hypothèses de test unilatérales :

$$H_0 : \mu = \mu_0 = 0.4 \text{ contre } H_1 : \mu > 0.4$$

La statistique de test est la moyenne empirique. Si on note $X_1, \cdots, X_n$ l'échantillon de variables aléatoires de même loi que X, la moyenne empirique est donnée par

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Intuitivement, on comprend bien qu'on va rejeter H_0 si $\bar{X}_n - \mu_0$ est trop grande en valeur absolue c'est à dire si la moyenne empirique est trop éloignée de la moyenne sous H_0.

Sous H_0, $Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$ suit une loi de Gauss de moyenne 0 et de variance 1. D'autre part, d'après la remarque faite plus haut on comprend qu'on rejette H_0 si $Z > z_0$. Pour construire la région critique, on cherche donc z_0 tel que

$$P(Z > z_0) = \alpha$$

$$P(Z > z_0) = 1 - \Phi(z_0)$$

où on note Φ la fonction de répartition de la loi Gauss de moyenne 0 et de variance 1. Ainsi z_0 est tel que

$$1 - \Phi(z_0) = \alpha$$

ce qui s'écrit encore

$$z_0 = \Phi^{-1}(1 - \alpha)$$

D'après la table de la fonction de répartition inverse de la loi normale, on en déduit que $z_0 = 1.64$ car $\alpha = 0.05$.

Finalement, on rejette donc H_0 si

$$\bar{X}_n - \mu_0 > 1.64 \frac{\sigma}{\sqrt{n}}$$

.

Remarques

– Lorsque le nombre d'observations n est grand (supérieur à 30), d'après le théorème de limite centrale on a que la statistique de test

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

suit approximativement une loi de Gauss quelque soit la loi de la variable X considérée.

5.2 Exemple 2 : test de Fisher

Le test de Fisher permet de tester l'égalité de deux variances dans un cadre gaussien.

- Condition d'utilisation : $X \sim \mathcal{N}\left(m_1, \sigma_1^2\right)$ dans population 1
et $X \sim \mathcal{N}\left(m_2, \sigma_2^2\right)$ dans population 2.

- La statistique de test F est telle que :
H_0 vraie $(\sigma_1^2 = \sigma_2^2) \Rightarrow F \sim Fisher\,(n_1 - 1; n_2 - 1)$ si $s_1^2 \geq s_2^2$
$F \sim Fisher\,(n_2 - 1; n_1 - 1)$ si $s_2^2 \geq s_1^2$

- Une réalisation de F est définie par : $f = s_1^2/s_2^2$ si $s_1^2 \geq s_2^2$
$f = s_2^2/s_1^2$ si $s_2^2 \geq s_1^2$

- Zone de non-rejet de H_0 au risque α

Hypothèse	Zone de non-rejet H_0
(1) $H_0 : \sigma_1^2 = \sigma_2^2 \quad H_1 : \sigma_1^2 \neq \sigma_2^2$	$\left[1; F_{n_2-1;\alpha/2}^{n_1-1}\right]$
(2) $H_0 : \sigma_1^2 = \sigma_2^2 \quad H_1 : \sigma_1^2 > \sigma_2^2$	$\left[1; F_{n_2-1;\alpha}^{n_1-1}\right]$
(3) $H_0 : \sigma_1^2 = \sigma_2^2 \quad H_1 : \sigma_1^2 < \sigma_2^2$	$\left[1; F_{n_1-1;\alpha}^{n_2-1}\right]$ avec $f = s_2^2/s_1^2$

$F_{n_2-1;\alpha}^{n_1-1}$ est lu dans la table de loi Fisher-Snedecor (α) à colonne $(n_1 - 1)$, ligne $(n_2 - 1)$.

- On acceptera H_0 si f appartient à la zone de non-rejet H_0, rejettera H_0 sinon.

Exercice

On veut comparer la précision de deux méthodes de dosage du menthol dans l'essence de menthe poivrée. Pour cela, on dose le menthol dans 16 flacons par ces deux méthodes. Les variances des résultats obtenus sont respectivement 0,013 g^2/L^2 (méthode 1) et 0,024 g^2/L^2 (méthode 2). Peut-on dire, au risque de 5%, que ces deux méthodes n'ont pas la même précision (on fera les hypothèses nécessaires) ?

5.3 Exemple 3 : test du chi2

Les tests basés sur la statistique du χ^2 permettent d'étudier les distributions de variables qualitative. Considérons l'exemple suivant dans lequel on cherche à savoir si les enfants de parents fumeurs ont plus de chance d'être fumeur que les autres. Les données sont reportées dans la table 2

	Père et mère fumeurs	Père fumeur mère non fumeur	Père non fumeur mère fumeur	Père et mère non fumeur	Total
fumeur	13	16	7	29	65
non fumeur	5	24	6	23	58

TAB. 2 –

- Hypothèses de test H_0 : le comportement des parents n'a pas d'effet contre H_1 : le comportement des parents a un d'effet ; ce qui peut se traduire par H_0 : indépendance des variables "enfant" et "parents" contre H_1 : dépendance des 2 variables
- Seuil de test : $\alpha = 0.05$
- Statistique de test

$$D = \sum_{i=1^2} \sum j = 1^4 \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$$

  où N_{ij} est l'effectif observé dans la case (i, j) et E_{ij} est l'effectif espéré (ou théorique) sous l'hpothèse H_0. Si on vérifie l'hpothèse H_0, alors D doit être proche de 0. De plus, on peut montrer que, sous H_0 , D suit une loi du χ^2 à $(nb_{lignes} - 1)(nb_{colonnes} - 1))$

  On remarque que sous H_0 : $\pi_{ij} = \pi_i * \pi_j$. On a de plus $E_{ij} = N\pi_{ij}$. Ainsi $E_{1,1} = 123 * 13/65 * 13/18$

  On a donc que $D = 5.54$.
- Région de rejet. $\chi^2_3(0.95) = 7.81$.
- Conclusion. On accepte donc H_0.

Références

Cadre B., Vial C., Statistique Mathématique Cours et exercices corrigés. Ellipse.

Fourdrinier D., (2002). Statistique inférentielle. Dunod.

Jolion J.M., (2003). Probabilité et Statistique. Cours de l'INSA. *http ://rfv.insa-lyon.fr/ jolion*

Kaufman P., (1994). Statistique : Information, Estimation, Test. Dunod.

Saporta G., (1990). Probablités, analyse des données et statistique. Edition Technip.
Reau J.P., Chauvat G., (1996). Probabilités et statistiques. Excercices et corrigés, Armand Colin, Collection cursus TD, série économie.
Scherrer B., (1988). Biostatitiques. Edition Gaetan Morin.
Schwartz D., (1984). Méthodes statistiques á l'usage des médecins et des biologistes, Flammarion, Médecine-Sciences, Collection Statistique en biologie et médecine.