

Université Claude Bernard, Lyon1

Année universitaire 2010/2011

Master Pro 1

STATISTIQUE PARAMETRIQUE

<http://math.univ-lyon1.fr/~gciuperca/>

Gabriela CIUPERCA

Table des matières

1 STATISTIQUE DESCRIPTIVE	5
1.1 Introduction	5
1.1.1 Généralités sur la Statistique	5
1.1.2 Terminologie de base	6
1.2 Statistique descriptive unidimensionnelle	6
1.2.1 Variable quantitative discrète	8
1.2.2 Variable quantitative continue	9
1.2.3 Variable qualitative	9
2 NOTIONS D'ECHANTILLONAGE	10
2.1 Moments empiriques	10
2.2 Fonction de répartition empirique	11
2.3 Rappels	11
3 THEORIE DE L'ESTIMATION	12
3.1 Théorie de l'estimation ponctuelle	12
3.1.1 Propriétés des estimateurs	12
3.1.2 Méthodes d'estimation	15
3.1.3 Familles exponentielles	18
3.2 Estimateur par intervalle	19
4 THEORIE DES TESTS	20
4.1 Tests paramétriques	20
4.1.1 Lemme de Neyman-Pearson	21
4.1.2 Test du rapport de vraisemblance et de Wald	25
4.2 Tests non-paramétriques	27
4.2.1 Théorème de Pearson	27
4.2.2 Test de χ^2 d'ajustement	28
4.2.3 Test de χ^2 d'indépendance	28
4.2.4 Test de Kolmogorov-Smirnov	29
4.2.5 Test de Smirnov, de comparaison de deux échantillons indépendantes	29
4.2.6 Test de la médiane sur des groupes indépendants	29
4.2.7 Test de Spearman	30
4.2.8 Test de Wilcoxon	31
5 REGRESSION LINEAIRE	32
5.1 Généralités sur le Modèle Linéaire	32
5.2 Régression linéaire simple	33
5.2.1 Description des données du modèle	33
5.2.2 Estimation des paramètres du modèle	34
5.2.3 Mesure de l'ajustement	36
5.2.4 Décomposition de la variabilité de Y	36
5.2.5 Evaluation de l'ajustement	37
5.2.6 Tests sur les paramètres	37
5.2.7 Prévision d'une valeur	39
5.3 Régression linéaire multiple	39
5.3.1 Estimation des paramètres	39
5.3.2 Décomposition de la variabilité de Y	41
5.3.3 Mesure de l'ajustement (empirique)	41
5.3.4 Théorème de Gauss-Markov	42

5.3.5	Tests d'hypothèse	42
5.3.6	Sélection des régresseurs	43
6	ANALYSE DE VARIANCE	
6.1	Analyse de variance à un facteur	45
6.1.1	Introduction	45
6.1.2	Terminologie	45
6.1.3	Données	45
6.1.4	Modèles statistiques	46
6.1.5	Estimation des paramètres	47
6.1.6	Tests d'hypothèses	47
6.2	Analyse de variance à deux facteurs	48
6.2.1	Introduction	48
6.2.2	Données	48
6.2.3	Modèle sans interaction (additif) : $r=1$	49
6.2.4	Modèle avec interaction (additif) : $r > 1$	51

Les Statistiques sont une continuation des Probabilités : ces deux disciplines étudient les phénomènes aléatoires :
- en Probabilités les lois des variables aléatoires sont totalement connues et on étudie leurs propriétés ;
- en Statistique la loi est totalement ou partiellement inconnue. Sur la base d'une expérience pratique on essaie de la déduire. Les connaissances de Probabilités jouent un rôle essentiel.

Exemple. Une machine fabrique des objets dont une proportion p (inconnue) est défectueuse. On veut vérifier si la machine est encore en bon état : $p \leq p_0$, pour un p_0 fixé. On prélève au hasard n de ces objets, que l'on vérifie, et à partir de ces observations, on essaie de répondre à la question.

Donc, un problème de statistique typique peut être décrit comme suit : une série d'expériences aléatoires sont réalisées et on mesure les données. Ces données sont des réalisations d'une variable aléatoire. (Pour l'exemple variable aléatoire de Bernoulli).

Problèmes statistiques courants :

- estimer des paramètres (ou la loi). Répondre à la question : est-ce que ces estimateurs ont des bonnes propriétés (si jamais on répète l'expérience on obtient des estimations proches) ;
 - il y a deux éventualités dont une seule est vraie : tests d'hypothèse (exemple : efficacité d'un médicament).
- Les applications de la Statistique sont très nombreuses : prévisions météo (modélisation physique et aléatoire), industrie pharmaceutique : l'efficacité des médicaments, médecine (modélisation de la progression d'une maladie), économétrie,....

Liste des notations

A^t : la matrice A transposée ;

$\mathbb{1}_A$: la fonction indicatrice de A ;

$\xrightarrow[n \rightarrow \infty]{p.s.}$: la convergence presque sûre pour n convergeant vers l'infini ;

$\xrightarrow[n \rightarrow \infty]{IP}$: la convergence en probabilité pour n convergeant vers l'infini ;

$\xrightarrow[n \rightarrow \infty]{\mathcal{L}}$: la convergence en loi pour n convergeant vers l'infini ;

$E[X]$: l'espérance de X ;

$Var[X]$: la variance (matrice de variance-covariance) de X ;

$\mathcal{N}(m, \sigma^2)$: la loi Normale unidimensionnelle d'espérance m et de variance σ^2 ;

$\mathcal{N}_k(m, \Sigma)$: la loi Normale de dimension k , d'espérance m et de matrice de variance-covariance Σ ;

$\chi^2(k)$: la loi de χ^2 à k degrés de liberté ;

$t(k)$: la loi de Student à k degrés de liberté ;

$F(m, n)$: la loi de Fisher à m et n degrés de liberté ;

Chapitre 1

STATISTIQUE DESCRIPTIVE

1.1 Introduction

1.1.1 Généralités sur la Statistique

Définition. On appelle **Statistique** l'ensemble des méthodes (techniques) permettant d'analyser (traiter) des ensembles d'observations (données).

Les méthodes en question relèvent le plus souvent des mathématiques (raison pour laquelle, la Statistique fait partie des Mathématiques appliquées) et font largement appel à l'outil informatique pour leur mise en oeuvre.

Statistique descriptive et statistique inférentielle

De manière approximative, il est possible de classer les méthodes statistiques en deux groupes : celui des méthodes descriptives et celui des méthodes inférentielles.

- La statistique **descriptive**. On regroupe sous ce terme les méthodes dont l'objet principal est la *description* des données étudiées; cette description des données se fait à travers leur *présentation* (la plus synthétique possible), leur *représentation graphique*, et le calcul de *résumés numériques*. Dans cette optique, on ne fait pas appel à des modèles probabilistes.
- La statistique **inférentielle**. Ce terme regroupe les méthodes dont l'objectif principal est de préciser un phénomène sur une population globale, à partir de son observation sur une partie restreinte de cette population, il s'agit donc d'induire (ou encore d'inférer) du particulier au général. Le plus souvent, ce passage ne pourra se faire que moyennant des hypothèses probabilistes.

D'un point de vue méthodologique, on notera que la statistique descriptive précède en général la statistique inférentielle dans une démarche de traitement de données.

1.1.2 Terminologie de base

Population (ou population statistique) : ensemble (au sens mathématique du terme) concerné par une étude statistique.

Individu toute unité de la population.

Echantillon : sous-ensemble de la population sur lequel sont effectivement réalisées les observations.

Taille de l'échantillon n : cardinal du sous-ensemble correspondant.

Enquête (statistique) : opération consistant à observer (ou mesurer, ou questionner,...) l'ensemble des individus d'un échantillon.

Variable : $X : \Omega \rightarrow \begin{cases} \mathbb{R} & \text{si quantitative} \\ E & \text{si qualitative} \end{cases}$

caractéristique (âge, salaire, sexe, ...) définie sur la population et observée sur l'échantillon. Si la variable est à valeurs dans \mathbb{R} (ou un sous-ensemble de \mathbb{R}), elle est dite *quantitative* (âge, salaire, taille,...); sinon elle est dite *qualitative* (sexe, catégorie socio-professionnelle,...).

Données (statistiques) : ensemble des individus observés (échantillon), des variables considérées, et des n observations de ces variables sur ces individus. Elles sont en général présentées sous forme de *tableaux* (individus en lignes et variables en colonnes).

Lorsque n est grand on cherche à synthétiser cette masse d'informations sous une forme exploitable et compréhensible. Une première étape consiste à décrire séparément les résultats obtenus pour chaque variable : c'est la description unidimensionnelle.

1.2 Statistique descriptive unidimensionnelle

Si X est une variable statistique et si ω_i désigne l'individu générique de l'échantillon observé, nous noterons $x_i = X(\omega_i)$ la valeur prise par cette variable sur cet individu. L'échantillon observé sera de dimension n . L'ensemble $\{X(\omega_i); i = 1, \dots, n\}$ constitue ce que l'on appelle la *série statistique brute*. Le but de ce chapitre est d'exposer les outils élémentaires, adaptés au type de variable observée permettant de présenter une série brute de façon synthétique et d'en résumer les principales caractéristiques. La synthèse se fait sous la forme de **tableaux**, de **graphiques**, et de **résumés numériques**. Sont ainsi introduites les notions de médiane, quantile, moyenne, variance, écart-type parallèlement aux représentations graphiques usuelles : diagramme en bâton, histogramme, boîte-à-moustaches, graphiques cumulatives, diagrammes en colonnes, en barre ou en secteurs.

Dans la suite, on distinguera trois cas suivants que la variable étudiée est une :

- variable quantitative discrète (elle ne prend qu'un nombre fini ou dénombrable de valeurs; en général il s'agit d'entiers)
- variable quantitative continue (variable quantitative qui n'est pas discrète)
- variable qualitative (oui/non, femme/homme....).

1.2.1 Variable quantitative discrète

Introduction

Exemples

- 1) Le nombre d'enfants dans une population de 10 familles : 1, 2, 0, 1, 1, 2, 3, 4, 0, 3.
- 2) L'âge (arrondi à l'année près) des 48 salariés d'une entreprise; la série statistique brute est donnée ci-dessous

43 29 57 45 50 29 37 59 46 31 46 24 33 38 49 31
62 60 52 38 38 26 41 52 60 49 52 41 38 26 37 59
57 41 29 33 33 43 46 57 46 33 46 49 57 57 46 43

Présentation des données : Le tableau statistique

Notons x_1, \dots, x_n la suite des observations rangées par ordre croissant, n étant la taille de l'échantillon. Soient z_1, \dots, z_r ces observations rangées par ordre croissant et non répétées (distinctes). Elles s'appellent *modalités*. Dans le tableau statistique la première colonne est l'ensemble de ces r valeurs. Puis on leur fait correspondre dans une seconde colonne leurs effectifs, c'est-à-dire le nombre de réplifications, notés n_1, \dots, n_r . Alors $\sum_{i=1}^r n_i = n$. Dans la troisième colonne on écrit les fréquences : $f_i = n_i/n$. Il peut être utile de compléter le tableau statistique en x rajoutent les fréquences cumulées :

$$F_i = \sum_{j=1}^i f_j$$

On note que $F_r = 1$.

Illustration

Dans le tableau suivant, on a calculé, sur les données présentées dans l'exemple 2, les effectifs, les fréquences et les fréquences cumulées.

z_i	n_i	f_i	F_i
24	1	0.02	0.02
26	2	0.04	0.06
29	3	0.06	0.12
31	2	0.04	0.16
33	4	0.08	0.24
37	2	0.04	0.28
38	4	0.08	0.36
41	3	0.06	0.42
43	3	0.06	0.48
45	1	0.02	0.50
46	6	0.13	0.63
49	3	0.06	0.69
50	1	0.02	0.71
52	3	0.06	0.77
57	5	0.11	0.88
59	2	0.04	0.92
60	2	0.04	0.96
62	1	0.02	1.00

Graphiques usuels

Pour une variable discrète, on rencontre essentiellement deux sortes de représentations graphiques, qui sont en fait complémentaires : le diagramme en bâtons et le diagramme cumulé (en escaliers).

Le diagramme en bâton

Se construit avec les modalités (observations distinctes) en abscisse et les effectifs en ordonnée. Il permet de donner une vision d'ensemble des observations réalisées.

Le diagramme cumulé

Il s'obtient à partir des fréquences cumulées et c'est le graphe d'une fonction appelée *fonction de répartition empirique* et définie ainsi :

$$F_n(x) = \begin{cases} 0 & x < z_1 \\ F_i(x) & z_i \leq x < z_{i+1} \quad (i = 1, \dots, r-1) \\ 1 & x \geq z_r \end{cases}$$

Résumés numériques

La description des données a pour objet le calcul des paramètres ou des valeurs typiques, qui permettent de caractériser de façon simple par un nombre petit de valeurs numériques les données observées. Les valeurs les plus couramment utilisées sont :

- des *paramètres de position* ou de *tendance centrale*. Leur objectif est de fournir un ordre de grandeur de la série étudiée, c'est-à-dire d'en situer le centre, le milieu. Les deux caractéristiques les plus usuelles sont :

- la moyenne
- la médiane

- des *caractéristiques de dispersion* qui permettent de chiffrer la variabilité des valeurs observées autour d'un paramètre de position :

- la variance, l'écart-type
- l'écart moyen absolu
- l'écart moyen à la médiane
- l'intervalle interquartiles

Paramètres de position ces grandeurs donnent un "milieu", une position moyenne autour desquelles les données sont réparties.

Définition La *moyenne empirique* (ou simplement la *moyenne*) est la moyenne arithmétique des observations :

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{j=1}^r n_j z_j$$

La moyenne est fonction de toutes les observations mais elle est sensible aux valeurs extrêmes.

Définition La *médiane* est un paramètre de position tel que la moitié des observations lui sont inférieures (ou

égales) et la moitié supérieures.

Pour calculer la médiane d'un échantillon il faut d'abord ordonner les données en ordre croissant : $\tilde{x}_1 \leq \tilde{x}_2 \leq \dots \leq \tilde{x}_n$, avec \tilde{x}_i les valeurs de x ordonnées.

- si n est impair : $n = 2k + 1$, la médiane est l'observation de rang $\frac{n+1}{2} = k + 1$

$$med(x) = \tilde{x}_{k+1} = \tilde{x}_{\frac{n+1}{2}}$$

Observation : la médiane est une valeur mesurée.

- lorsque n est pair : $n = 2k$, tout nombre compris entre $\tilde{x}_{\frac{n}{2}}$ et $\tilde{x}_{\frac{n}{2}+1}$ répond à la définition et on convient généralement de prendre comme valeur de la médiane la moyenne arithmétique de ces deux observations :

$$med(x) = \frac{\tilde{x}_k + \tilde{x}_{k+1}}{2}$$

La principale propriété de la médiane concerne la place qu'elle occupe par rapport à la moyenne. Dans le cas des distributions symétriques la médiane est égale à la moyenne. Par contre pour les distributions dissymétriques, la moyenne est différente de la médiane. Par exemple si la dissymétrie est à gauche (maximum des fréquences décentrées vers la gauche) alors la médiane est inférieure à la moyenne. La différence entre les deux paramètres est d'autant plus importante, en valeur absolue, que la dissymétrie est plus prononcée.

Paramètres de dispersion Ils servent à préciser la variabilité de la série de données, c'est-à-dire à résumer l'éloignement de l'ensemble des observations par rapport à leur tendance centrale.

a) La variance, l'écart-type

Définition La *variance* d'une série est la moyenne arithmétique des carrés des écarts par rapport à la moyenne :

$$Var(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

L'*écart-type* est la racine carrée de la variance $\sigma = \sqrt{Var(x)}$.

b) Les quartiles, l'écart interquartiles Les quartiles divisent l'échantillon ordonné en ordre croissant en 4 sous-ensembles de même effectif : Q1, Q2, Q3. Donc Q2 est la médiane. Un quart des observations sont inférieures ou égales au premier quartile Q1 et trois quarts des observations lui sont supérieures. Le troisième quartile est supérieur à trois quarts des observations et inférieur à un quart.

La différence Q3-Q1 est *écart interquartiles*.

Les quartiles permettent de construire les diagrammes de type *boxplot* ou *diagramme en boîte à moustaches*. La partie centrale du boxplot est représentée par une boîte de longueur l'écart interquartiles Q3-Q1. On trace à l'intérieur la position de la médiane. La boîte est complétée par des moustaches correspondant aux valeurs :

- partie supérieure : la plus grande valeur inférieure à $Q3 + 1.5(Q3 - Q1)$;
- partie inférieure : la plus petite valeur supérieure à $Q1 - 1.5(Q3 - Q1)$.

On définit les *valeurs extrêmes*, celles qui sortent des "moustaches". On tel graphique permet de repérer les éventuelles valeurs aberrantes et facilite la comparaison de plusieurs distributions. Comparer des diagrammes en boîte est plus aisé que comparer des histogrammes.

1.2.2 Variable quantitative continue

Une variable quantitative est dite continue lorsque les observations qui lui sont associées ne sont pas des valeurs précises mais des intervalles réels. Cela signifie que, dans ce cas, le sous ensemble de \mathbb{R} des valeurs possibles de la variable étudiée a été divisée en r intervalles contigus appelés *classes*.

En général, les deux raisons principales qui peuvent amener à considérer comme continue une variable quantitative sont le grand nombre d'observations distinctes (un traitement en discret sera un peu inconfortable) et le caractère "sensible" d'une variable.

Exemples. l'âge ou le revenu pour un groupe d'individus.

Notons les r classes : $[a_0, a_1[, \dots, [a_{r-1}, a_r]$.

Présentation des données

On présente les données dans un tableau, comme dans le cas discret, en indiquant les classes rangées en ordre croissant.

Les notions d'effectif, de fréquence sont définies de la même façon que dans le cas discret. On indique dans le tableau aussi :

- les centres $c_i = \frac{a_i + a_{i-1}}{2}$ des classes, $i = 1, 2, \dots, r$
- les amplitudes des classes $L_i = a_i - a_{i-1}$;
- les densités des observations dans chaque classe : $h_i = \frac{n_i}{n L_i}$.

Exemple. Pour l'année 1987, la répartition des exploitations agricoles françaises selon SAU (Surface Agricole Utilisée) exprimée en hectares :

SAU	fréq %	F_i	c_i	L_i	h_i
moins de 5	24.0	24	2.5	5	4.8
[5, 10[10.9	34.9	7.5	5	2.18
[10, 20[17.8	52.7	15	10	1.78
[20, 35[20.3	73	27.5	15	1.35
[35, 50[10.2	83.2	42.5	15	0.68
[50, 200]	16.8	100	125	150	0.11

Représentations graphiques

A la place du diagramme en bâtons, on trace un *histogramme* composé de rectangles dont les bases sont les classes et les hauteurs sont les densités des observations. L'aire du rectangle i vaut f_i (la fréquence de la classe correspondante).

Caractéristiques numériques

- La moyenne, la variance et l'écart-type s'obtiennent comme dans le cas discret en prenant à la place des valeurs les centres des classes c_i .
- Les quartiles d'une variable continue peuvent être déterminées de façon directe à partir de la courbe cumulative.

1.2.3 Variable qualitative

Les modalités d'une telle variable ne sont pas numériques, donc on ne peut pas calculer les grandeurs statistiques telles que la moyenne, la variance, ... On peut faire un tableau pour représenter les données en indiquant pour chaque modalité l'effectif et la fréquence.

Exemple. le nombre de sièges occupés par 3 partis politiques : P1, P2, P3.

	effectif	f_i
P1	200	1/3
P2	100	1/6
P3	300	1/2

Les représentations graphiques : diagramme en colonnes et diagramme en secteurs.

Chapitre 2

NOTIONS D'ECHANTILLONAGE

Le schéma théorique de la plupart des problèmes de statistique inférentielle est le suivant : on a un ensemble mesurable (Ω, \mathcal{B}) et cet espace est muni d'une probabilité \mathbb{P}_θ avec $\theta \in \Theta \subseteq \mathbb{R}^p$. Pendant ce cours on va considérer des variables aléatoires X définies sur le champ Borelien de probabilité $\{\Omega, \mathcal{B}, \mathbb{P}_\theta\}$. Sur l'espace (Ω, \mathcal{B}) on se donne n variables aléatoires toutes de même loi \mathbb{P}_θ , (X_1, \dots, X_n) , à valeurs dans un espace mesurable (A, \mathcal{X}) . Les valeurs qu'on mesure pour ces variables aléatoires sont $(x_1, \dots, x_n) = (X_1(\omega), \dots, X_n(\omega))$ pour un certain élément $\omega \in \Omega$.
Définition. On appelle *n-échantillon d'une loi \mathbb{P}_θ* toute famille X_1, \dots, X_n de v.a. indépendantes et de même loi que X . *est une suite de v.a.i.d.*

Puisque les v.a. X_i ont la même loi que X , elles ont les mêmes moments :

$$\mathbb{E}[X_i] = \mathbb{E}[X], \quad \text{Var}(X_i) = \text{Var}(X), \quad \mathbb{E}[X_i^k] = \mathbb{E}[X^k]$$

$$\forall i = 1, \dots, n, k \in \mathbb{N}.$$

2.1 Moments empiriques

On considère le cas des v.a. unidimensionnelles ($\Omega \subseteq \mathbb{R}$). Soit un n-échantillon X_1, \dots, X_n (X_i est une v.a. réelle pour la $i^{\text{ème}}$ expérience).

Définition. On appelle *moment empirique d'ordre p* ($p \in \mathbb{N}$) la v.a.

$$U_p^n = \frac{1}{n} \sum_{i=1}^n X_i^p$$

et on appelle *moment empirique centré d'ordre p* , la v.a.

$$W_p^n = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^p$$

$$\begin{aligned} p=1 : \bar{X}_n = U_1^n \text{ la moyenne empirique} &= \frac{1}{n} \sum_{i=1}^n x_i \\ p=2 : W_2^n = S_n^2 \text{ la variance empirique} &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \end{aligned}$$

Soient $m_p = \mathbb{E}(X^p)$ et $\mu_p = \mathbb{E}(X - m_1)^p$ les moments centrés et ~~non~~ centrés d'ordre p de X (s'ils existent). En utilisant la loi des grands nombres, on obtient le résultat suivant :

Théorème 2.1.1 Si $m_{2p} = \mathbb{E}(X^{2p}) < \infty$, alors

$$U_p^n \xrightarrow[n \rightarrow \infty]{p.s.} m_p, \quad W_p^n \xrightarrow[n \rightarrow \infty]{p.s.} \mu_p$$

Théorème 2.1.2 a) Si $\mathbb{E}(X^p) < \infty$, alors $\mathbb{E}(U_p^n) = \mathbb{E}(X^p) = m_p$. Cas particulier $p=1$, $\mathbb{E}(\bar{X}_n) = m_1 = m$.
 b) Si $\mathbb{E}(X^{2p}) < \infty$, alors $\text{Var}(U_p^n) = \frac{\mathbb{E}(X^{2p}) - \mathbb{E}^2(X^p)}{n}$. Cas particulier $p=1$, $\text{Var}(\bar{X}_n) = \frac{\text{Var}(X)}{n} = \sigma^2/n$.
 c) Si $\text{Var}(X) < \infty$ alors $\mathbb{E}(W_2^n) = \frac{n-1}{n} \text{Var}(X)$

2.2 Fonction de répartition empirique, c'est une approx de $F(x)$

Définition. On appelle *fonction de répartition empirique*, l'application aléatoire : $\mathbb{R} \rightarrow [0,1]$

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq x} = \frac{1}{n} \text{Card}\{X_i; X_i \leq x\}, \quad \forall x \in \mathbb{R}$$

$f: \mathbb{R} \rightarrow [0,1]$ la fonction de répartition de la v.a. $X: \Omega \rightarrow \mathbb{R}$
 $\forall x \in \mathbb{R} \quad f(x) = \mathbb{P}(X \leq x)$

Proposition 2.2.1 $\hat{F}_n(x) \xrightarrow[n \rightarrow \infty]{p.s.} F(x), \quad \forall x \in \mathbb{R}.$

On a un résultat plus fort.

Théorème 2.2.1 (Glivenko-Cantelli) La convergence des fonctions de répartition empirique est p.s. uniforme :

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow[n \rightarrow \infty]{p.s.} 0$$

2.3 Rappels

Rappelons deux résultats de Probabilités utiles à la suite.

Théorème 2.3.1 (Slutsky) Soit la suite de variables aléatoires (X_n) convergeante en loi vers X et la suite (Y_n) convergeante en loi vers C , avec C une constante p.s., alors

$$X_n + Y_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} X + C, \quad Y_n X_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} CX, \quad \frac{X_n}{Y_n} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \frac{X}{C} \text{ (si } C \neq 0)$$

Une application du Théorème 2.3.1 est le résultat suivant qu'on va utiliser dans la théorie de l'estimation, pour énoncer la *delta-méthode*.

Théorème 2.3.2 Soient X_1, \dots, X_n et Y des vecteurs aléatoires de dimension k satisfaisant la condition

$$a_n(X_n - c) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} Y$$

avec $c \in \mathbb{R}^k$ et (a_n) une suite de nombres positifs et $\lim_{n \rightarrow \infty} a_n = \infty$. Soit également la fonction $g : \mathbb{R}^k \rightarrow \mathbb{R}$.

(i) Si g est dérivable en c , alors

$$a_n [g(X_n) - g(c)] \xrightarrow[n \rightarrow \infty]{\mathcal{L}} [\nabla g(c)]^t Y \quad (2.1)$$

avec $\nabla g(c)$ le vecteur, de dimension k , des dérivées partielles de g par rapport à x .

(ii) Supposons que g a des dérivées partielles continues d'ordre $m > 1$ dans un voisinage de c , toutes les dérivées partielles d'ordre j , $1 \leq j \leq m-1$ s'annulent en c , et les dérivées partielles d'ordre m pas toutes nulles en c . Alors

$$a_n^m [g(X_n) - g(c)] \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \frac{1}{m!} \sum_{i_1=1}^k \dots \sum_{i_m=1}^k \frac{\partial^m g}{\partial x_{i_1} \dots \partial x_{i_m}} \Big|_{x=c} Y_{i_1} \dots Y_{i_m} \quad (2.2)$$

avec Y_j la composante j de Y .

appelé également Théorème de la δ -méthode.

Chapitre 3

THEORIE DE L'ESTIMATION

Soit une v.a. X définie sur l'espace de probabilité $(\Omega, \mathcal{B}, P_\theta)$ et supposons que la fonction de répartition F_θ dépend d'un certain nombre de paramètres θ , $\theta \in \Theta \subseteq \mathbb{R}^p$. On suppose que la fonction F_θ est connue, mais pas θ . Soit θ_0 la vraie valeur (inconnue). Le but est de trouver des statistiques (une fonction du n-échantillon (X_1, \dots, X_n)) qui vont approximer le mieux possible, dans un certain sens, θ_0 .

3.1 Théorie de l'estimation ponctuelle

Définition. On appelle *estimateur ponctuel du paramètre* θ_0 (en général on dit θ) toute fonction de l'échantillon, prenant ses valeurs dans Θ : $T_n = T(X_1, \dots, X_n)$ *statistique* $\hat{\theta}_n$. La valeur prise par T pour un n-uplet de données (x_1, \dots, x_n) est l'*estimation* de θ : $T(x_1, \dots, x_n)$.

Exemple 1. On lance une pièce de monnaie et soit la v.a.

$$X = \begin{cases} 0 & \text{si "face"} \\ 1 & \text{si "pile"} \end{cases}$$

alors $X \sim \mathcal{B}(\theta)$, $\theta = p$. On souhaite estimer θ . On lance la pièce 10 fois : $n = 10$. $X_1, \dots, X_{10} \sim \mathcal{B}(\theta)$. Une réalisation de l'échantillon est : 0, 1, 1, 0, 1, 1, 1, 0, 0, 1. Si on prend $\bar{X}_n \in [0, 1]$ comme estimateur, alors $\bar{x}_n = 6/10$. Si on répète 10 fois encore l'expérience : 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, $\bar{x}_n = 5/10$. $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. D'autres estimateurs pour $\theta : 1/2$, $T = X_1$, $T = (X_1 + X_2)/2$.

Exemple 2. $X_1, \dots, X_{10} \sim \mathcal{P}(\lambda)$, λ inconnu. On peut prendre comme estimateur pour λ : \bar{X}_n , mais aussi $\frac{2}{n(n+1)} \sum_{k=1}^n kX_k$.

De ces exemples, c'est claire qu'on doit choisir des estimateurs avec des "bonnes qualités". Par exemple, pour n grand, $\lim T(X_1, \dots, X_n) = \theta_0$ dans un certain sens. Les valeurs de deux estimations ne doivent pas être non plus "trop différentes".

3.1.1 Propriétés des estimateurs

Définition. On dit que l'estimateur $T_n = T(X_1, \dots, X_n)$ est *faiblement* (resp. *fortement*) *consistant* (convergent) si :

$$T_n \xrightarrow[n \rightarrow \infty]{P} \theta_0 : \quad \forall \varepsilon > 0, \lim_{n \rightarrow \infty} IP[|T_n - \theta_0| \geq \varepsilon] = 0$$

respectivement :

$$T_n \xrightarrow[n \rightarrow \infty]{p.s.} \theta_0 : \quad IP[\lim_{n \rightarrow \infty} T_n = \theta_0] = 1$$

Exemple 1. Les moments empiriques sont des estimateurs fortement consistants des moments théoriques. En particulier, \bar{X}_n est estimateur consistant pour $m = \mathbb{E}(X)$.

Exemple 2. $X_i \sim \mathcal{B}(\theta)$, \bar{X}_n est estimateur fortement consistant pour θ , ou encore $\frac{1}{n+2} [\sum_{i=1}^n X_i + 2] \xrightarrow[n \rightarrow \infty]{p.s.} \theta$.

Donc, les estimateurs consistants ne sont pas uniques.

Définition. Pour θ scalaire, on appelle *erreur quadratique* de T_n par rapport à θ_0 , la quantité :

$$d^2(T_n, \theta_0) = \mathbb{E}[T_n - \theta_0]^2$$

Proposition 3.1.1 Si $d^2(T_n, \theta_0) \xrightarrow{n \rightarrow \infty} 0$, alors T_n est un estimateur faiblement consistant de θ . *Preuve avec Biaisymétrie Tchebychev*

Définition. On appelle *biais* de l'estimateur T_n , la quantité : $B(T_n, \theta) = \mathbb{E}(T_n) - \theta$.

L'estimateur est dit sans biais si $B(T_n, \theta) = 0$ et il est dit asymptotiquement sans biais si $B(T_n, \theta) \xrightarrow{n \rightarrow \infty} 0$.

Exemples classiques. 1) U_k^n estimateur sans biais pour $\mathbb{E}(X^k) = m_k$, \bar{X}_n pour $m = \mathbb{E}(X)$.

2) W_2^n estimateur asymptotiquement sans biais pour $\text{Var}(X)$.

3) $S_n^{*2} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{n}{n-1} W_2^n$ est un estimateur sans biais pour μ_2 .

4) Pour x fixé, $\hat{F}_n(x)$ est un estimateur sans biais pour $F(x)$.

$$\mathbb{E}(W_2^n) \stackrel{\text{Thm 2.1.2.c}}{=} \frac{n-1}{n} \text{Var}(x)$$

Proposition 3.1.2 (Fisher, Cochran)

Théorème de Cochran, Fisher

$$\textcircled{1} \quad \sqrt{n} \frac{\bar{X}_n - m}{S_n^*} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1)$$

Dans le cas particulier $X_i \sim \mathcal{N}(m, \sigma^2)$ on a

$$\textcircled{2} \quad \bar{X}_n \sim \mathcal{N}(m, \frac{\sigma^2}{n})$$

$$\textcircled{3} \quad \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - m)^2 \sim \chi^2(n)$$

$$\textcircled{4} \quad \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \sim \chi^2(n-1)$$

$$\textcircled{5} \quad \sqrt{n} \frac{\bar{X}_n - m}{S_n^*} \sim t(n-1)$$

estimateur 6
de centre

et les v.a. \bar{X}_n et W_2^n sont indépendantes.

Preuve. Nous donnons la démonstration que pour les trois dernières affirmations.

Soit le vecteur aléatoire colonne $\mathbf{X} = (X_1, \dots, X_n)$. Nous avons :

$$(n-1)S_n^{*2} = \sum_{i=1}^n X_i^2 + n\bar{X}_n^2 - 2\bar{X}_n \sum_{i=1}^n X_i = \sum_{i=1}^n X_i^2 - n\bar{X}_n^2 \quad (3.1)$$

Supposons d'abord que $m = 0$ et $\sigma = 1$. Alors, $\mathbf{X} \sim \mathcal{N}_n(0, I_n)$. Soit A une matrice $n \times n$ dont la dernière ligne est $(1/\sqrt{n}, \dots, 1/\sqrt{n})$ et telle que $AA^t = A^tA = I_n$ (A est une matrice unitaire). Posons $\mathbf{Y} = A\mathbf{X}$. D'où $\mathbb{E}[\mathbf{Y}] = A\mathbb{E}[\mathbf{X}] = 0_n$, $\text{Var}[\mathbf{Y}] = A\text{Var}[\mathbf{X}]A^t = I_n$. Alors, $\mathbf{Y} = (Y_1, \dots, Y_n) \sim \mathcal{N}_n(0, I_n)$ et $\sum_{i=1}^n Y_i^2 = \sum_{i=1}^n X_i^2$. Puisque la dernière ligne de A est $(1/\sqrt{n}, \dots, 1/\sqrt{n})$ on a $Y_n = \frac{\sum_{i=1}^n X_i}{\sqrt{n}} = \sqrt{n}\bar{X}_n$. Alors $Y_n^2 = n\bar{X}_n^2$.

Alors la relation (3.1) devient $(n-1)S_n^{*2} = \sum_{i=1}^n Y_i^2 - Y_n^2 = \sum_{i=1}^{n-1} Y_i^2 \sim \chi^2(n-1)$. On a également que \bar{X}_n est indépendant de S_n^{*2} .

Si $m \neq 0$ ou $\sigma \neq 1$ soit les variables aléatoires $W_i = \frac{X_i - m}{\sigma} \sim \mathcal{N}(0, 1)$. Nous avons alors $\bar{X}_n = m + \sigma\bar{W}_n$ et $\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \sum_{i=1}^n (W_i - \bar{W}_n)^2 \sim \chi^2(n-1)$.

On a la décomposition :

$$\sqrt{n} \frac{\bar{X}_n - m}{S_n^*} = \sqrt{n} \frac{\bar{X}_n - m}{\sigma} \frac{1}{\sqrt{\frac{S_n^{*2}}{\sigma^2}}}$$

D'autre part $(n-1)S_n^{*2}/\sigma^2 \sim \chi^2(n-1)$ et $\sqrt{n} \frac{\bar{X}_n - m}{\sigma} \sim \mathcal{N}(0, 1)$ et ces deux variables sont indépendantes. D'où $\sqrt{n} \frac{\bar{X}_n - m}{S_n^*} \sim t(n-1)$. ■

Remarquons qu'entre S_n^2 et S_n^{*2} , la loi de χ^2 perd un degré de liberté qui correspond à l'estimation d'un paramètre, l'espérance m .

Proposition 3.1.3 $d^2(T_n, \theta) = \text{Var}(T_n) + B^2(T_n, \theta)$

Si l'estimateur est sans biais, l'erreur quadratique est égale à la variance.

Définition. Un estimateur $T_n(X_1, \dots, X_n)$ est dit *libre* pour le paramètre θ_k si sa loi ne dépend pas de θ_k .

Inégalité de Rao-Cramer. Estimateur efficace.

Soit $X \sim P_\theta$ et f_θ sa fonction de densité si X est continue, sa fonction de fréquence, si X est discrète.

Cas I. $\Theta \subseteq \mathbb{R}$ (scalaire).

Supposons que les conditions suivantes sont satisfaites : l'ensemble $A = \{x; f_\theta(x) > 0\}$ est indépendant de θ , et

$$\forall x \in A, \forall \theta \in \Theta, \exists \frac{\partial}{\partial \theta} \log f_\theta(x) dx < \infty \text{ et } \int_\Omega \frac{\partial}{\partial \theta} f_\theta(x) dx = \frac{\partial}{\partial \theta} \int_\Omega f_\theta(x) dx = 0$$

Définition. L'information de Fisher pour la v.a. $X : I(\theta) = \mathbb{E} \left[\frac{\partial}{\partial \theta} \log f_\theta(X) \right]^2$. densité de fréquence

Théorème 3.1.1 (Inégalité de Rao-Cramer pour une v.a.)

Soit la v.a. X de loi P_θ et soit la v.a. $S(X)$ telle que $\text{Var}[S(X)] < \infty, \forall \theta \in \Theta$. Soit $\psi(\theta) = \mathbb{E}[S(X)]$. Si $0 < I(\theta) < \infty$, alors

$$\forall \theta \in \Theta \quad \text{Var}[S(X)] \geq \frac{[\psi'(\theta)]^2}{I(\theta)} \quad (3.2)$$

Preuve du Théorème 3.1.1

Soit : $l_\theta(x) = \log f_\theta(x)$. On applique l'inégalité de Cauchy aux variables aléatoires : $S(X) - \psi(\theta)$ et $l'_\theta(X) = \frac{\partial l_\theta(X)}{\partial \theta}$.

Donc :

$$\mathbb{E}[(S(X) - \psi(\theta))l'_\theta(X)] \leq \{\mathbb{E}[S(X) - \psi(\theta)]^2\}^{1/2} \{\mathbb{E}[l'_\theta(X)]^2\}^{1/2}$$

Mais $\mathbb{E}[l'_\theta(X)]^2 = I(\theta)$. Alors :

$$\mathbb{E}[S(X) - \psi(\theta)]^2 = \text{Var}[S(X)] \geq \frac{1}{I(\theta)} \mathbb{E}[(S(X) - \psi(\theta))l'_\theta(X)]^2$$

Parce que : $\mathbb{E}[\frac{\partial}{\partial \theta} \log f_\theta(X)] = 0$ on a :

$$\begin{aligned} \mathbb{E}[(S(X) - \psi(\theta))l'_\theta(X)] &= \mathbb{E} \left[S(X) \frac{\partial}{\partial \theta} \log f_\theta(X) \right] = \int_\Omega S(x) \frac{f'_\theta(x)}{f_\theta(x)} f_\theta(x) dx \\ &= \int_\Omega S(x) f'_\theta(x) dx = \frac{\partial}{\partial \theta} \int_\Omega S(x) f_\theta(x) dx = \frac{\partial}{\partial \theta} \mathbb{E}[S(X)] = \psi'(\theta) \end{aligned}$$

Corollaire. Si $S(X)$ est sans biais alors $\text{Var}[S(X)] \geq \frac{1}{I(\theta)}$.

Soit (X_1, \dots, X_n) un n-échantillon de loi P_θ et $I_1(\theta)$ l'information de Fisher pour une v.a. X_i . La densité (la fonction de fréquence) de (X_1, \dots, X_n) est :

$$L_n(\theta; x_1, \dots, x_n) = \prod_{i=1}^n f_\theta(x_i)$$

Alors, l'information de Fisher pour le n-échantillon est :

$$I_n(\theta) = \mathbb{E} \left[\frac{\partial}{\partial \theta} \log L_n(\theta; X_1, \dots, X_n) \right]^2 = \mathbb{E} \left[\frac{\partial}{\partial \theta} \sum_{i=1}^n \log f_\theta(X_i) \right]^2$$

Proposition 3.1.4 $I_n(\theta) = nI_1(\theta)$.

Théorème 3.1.2 (Inégalité de Rao-Cramer pour un estimateur)

Si T_n est un estimateur pour θ et on note $\psi(\theta) = \mathbb{E}[T_n]$ alors l'inégalité de Rao-Cramer devient

$$\forall \theta \in \Theta \quad \text{Var}[T_n] \geq \frac{[\psi'(\theta)]^2}{nI_1(\theta)}$$

et de $\text{Var}[T_n] \geq \frac{1}{nI_1(\theta)}$

Définition. Un estimateur sans biais, pour lequel l'inégalité de Rao-Cramer devient égalité, est dit efficace.

L'inégalité de Rao-Cramer donne une borne inférieure pour la variance des estimateurs sans biais. Donc, si on dispose d'un estimateur dont la variance est égale à cette borne, on sait qu'il est meilleur que tous les autres estimateurs sans biais. Remarquons aussi le rôle joué par l'information de Fisher : plus elle est grande plus la variance du "meilleur" estimateur sans biais est petite.

Cas II. $\theta \in \Theta \subseteq \mathbb{R}^p, p > 1, \theta = (\theta_1, \dots, \theta_p)$.

Définition. On appelle matrice d'information de Fisher :

$$I(\theta) = (I_{ij}(\theta))_{1 \leq i, j \leq p} = \mathbb{E} [\nabla_\theta \log f_\theta(X) \cdot \nabla_\theta \log f_\theta(X)^t]$$

si elle existe et elle est inversible.

Soit $S_\theta(X)$ un vecteur aléatoire de dimension q et de carré intégrable et : $\psi(\theta) = \mathbb{E}[S_\theta(X)] \in \mathbb{R}^q$, $q \geq 1$. L'inégalité de Rao-Cramer est :

$$\forall \theta \in \Theta \quad \text{Var}[S(X)] \geq \nabla_\theta \psi(\theta)^t I^{-1} \nabla_\theta \psi(\theta)$$

Définition. Si $\theta = (\theta_1, \dots, \theta_p) \in \mathbb{R}^p$, un estimateur T_n est *exhaustif* pour le paramètre θ_k , $k \in \{1, \dots, p\}$ si la loi de X sachant $T_n = t$ ne dépend pas de θ_k .

Une statistique exhaustive contient toute l'information sur le paramètre incluse dans l'échantillon. Si T_n est exhaustive pour θ et φ une fonction borélienne strictement monotone, alors $\varphi(T_n)$ est également une statistique exhaustive pour θ .

Théorème 3.1.3 (de factorisation, critère de Neyman)

Un estimateur T_n est exhaustif pour θ s.s.i. existe une fonction borélienne $g : \mathbb{R}^n \rightarrow \mathbb{R}_+$ telle que :

$$L_n(\theta) = h_\theta(T_n)g(X_1, \dots, X_n), \quad \forall \theta \in \Theta \quad (3.3)$$

où h_θ est la densité (fonction de fréquence) de T_n et g ne dépend pas de θ .

Démonstration. Pour des variables discrètes.

Soit $E_n = \{x = (x_1, \dots, x_n) \in \mathbb{R}^n; T_n(x) = t\}$. Nous avons $(T_n = t) = \cup_{x \in E_n} (X = x)$, donc $\mathbb{P}[T_n = t] = \sum_{x \in E_n} h_\theta(t)g(x)$. Par conséquent, si (3.3) est vérifiée,

$$\begin{aligned} \mathbb{P}_\theta[(X_1, \dots, X_n) = (x_1, \dots, x_n) | T_n = t] &= \frac{\mathbb{P}_\theta[(X_1, \dots, X_n) = (x_1, \dots, x_n), T_n = t]}{\mathbb{P}[T_n = t]} \\ &= \begin{cases} 0 & \text{si } x \notin E_n \\ \frac{h_\theta(x)g(x)}{\mathbb{P}[T_n = t]} = \frac{g(x)}{\sum_{y \in E_n} g(y)} & \text{si } x \in E_n \end{cases} \end{aligned}$$

Réciproquement, les fonctions $h_\theta(t) = \mathbb{P}[T_n = t]$ et $g(x_1, \dots, x_n) = \mathbb{P}_\theta[(X_1, \dots, X_n) = (x_1, \dots, x_n) | T_n = t]$ conviennent. ■

Une méthode utile en Statistique est la *delta-méthode*, qui est basée sur le Théorème 2.3.2.

Proposition 3.1.5 Supposons que les conditions du Théorème 2.3.2 sont satisfaites. Soit Y un vecteur aléatoire Gaussien $\mathcal{N}_k(0, \Sigma)$. Alors

$$a_n [g(X_n) - g(c)] \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, [\nabla g(c)]^t \Sigma \nabla g(c))$$

Exemple. Soit (X_n) une suite de variables aléatoires satisfaisant $\sqrt{n}(X_n - c) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1)$. Considérons la fonction $g(x) = x^2$. Si $c \neq 0$, alors, en appliquant la delta-méthode on a $\sqrt{n}(X_n^2 - c^2) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 4c^2)$. Si $c = 0$ la dérivée d'ordre 1 de g en 0 est 0 mais la dérivée seconde est 2. Donc, en appliquant la relation (2.2) on a que $nX_n^2 \xrightarrow[n \rightarrow \infty]{\mathcal{L}} [\mathcal{N}(0, 1)]^2 = \chi^2(1)$.

3.1.2 Méthodes d'estimation

Méthode des moments

Soit X_1, \dots, X_n un n -échantillon de la loi P_θ . Supposons que les moments théoriques d'ordre k existent : $m_k = \mathbb{E}(X^k)$. On cherche un estimateur par résolution du système d'équations en θ obtenu en égalant moment théorique et moment empirique correspondant :

$$m_k = U_k^n, \quad k = 1, 2, \dots, p \quad (3.4)$$

(ou p autres équations). La solution du système (3.4), si elle existe et elle est unique, sera appelée *estimateur par la méthode des moments*.

Méthode du maximum de vraisemblance

On définit la fonction de vraisemblance du n-échantillon par $L_n(\theta) = \prod_{i=1}^n f_\theta(X_i)$. Son interprétation est claire. Par exemple, si la distribution de X est discrète alors $f_\theta(x) = \mathbb{P}_\theta[X = x]$ est la probabilité d'observer le point x et la fonction de vraisemblance : $L_n(\theta) = \mathbb{P}_\theta(X_1) \dots \mathbb{P}_\theta(X_n)$ représente la probabilité d'observer (X_1, \dots, X_n) . Dans le cas continu la fonction de vraisemblance est la densité du vecteur (X_1, \dots, X_n) . Supposons que pour toute valeur (X_1, \dots, X_n) , $L_n(\theta)$ admet un maximum unique. La valeur $\hat{\theta}_n$ pour laquelle ce maximum est atteint :

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} L_n(\theta) \quad (3.5)$$

est appelée *estimation par maximum de vraisemblance*. Si on remplace les valeurs par les v.a. correspondantes on obtient l'estimateur du maximum de vraisemblance (EMV).

Eq : La fonction log étant croissante, il est équivalent de maximiser $\log L_n$ et L_n .

Théorème 3.1.4 Supposons que :

- (i) $f_\theta(x) > 0, \forall x \in \mathbb{R}, \forall \theta \in \Theta, \log f_\theta(x) \in C^4(\Theta \times \mathbb{R})$
- (ii) $-\infty < -I_1(\theta) = \int_{\mathbb{R}} \frac{\partial^2 \log f_\theta(x)}{\partial \theta^2} f_\theta(x) dx < 0, \theta \in \Theta$
- (iii) il existe une fonction $H : \mathbb{R} \rightarrow \mathbb{R}_+$ telle que $\forall \theta \in \Theta$

$$\left| \frac{\partial^3 \log f_\theta(x)}{\partial \theta^3} \right| \leq H(x) \text{ et } \int_{\mathbb{R}} H(x) f_\theta(x) dx = M_\theta < \infty$$

Alors, avec la probabilité 1, pour $n \rightarrow \infty$, l'équation de vraisemblance (3.5) a une solution $\hat{\theta}_n$ consistante et

$$\sqrt{n} \frac{\hat{\theta}_n - \theta^0}{\sigma} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1)$$

$$\text{avec } \sigma^2 = \left[\mathbb{E} \left(\frac{\partial \log f_\theta(X)}{\partial \theta} \Big|_{\theta=\theta^0} \right)^2 \right]^{-1} = \left[I_1(\theta^0) \right]^{-1}$$

Preuve du Théorème 3.1.4

(i) Preuve de l'existence et de la consistance

Fixons θ^0 et considérons la vraisemblance quand le paramètre vaut θ . Maximiser la vraisemblance $L_n(\theta)$ par rapport à θ revient à maximiser $L_n(\theta)/L_n(\theta^0)$. Le maximum de cette fonction, s'il est atteint sur Θ , l'est en un point qui annule la dérivée de : $\log(L_n(\theta)/L_n(\theta^0)) = \sum_{i=1}^n \log(f_\theta(X_i)/f_{\theta^0}(X_i))$. D'après l'inégalité de Jensen :

$$\mathbb{E} \left[\log \frac{f_\theta(X_i)}{f_{\theta^0}(X_i)} \right] < \log \mathbb{E} \left[\frac{f_\theta(X_i)}{f_{\theta^0}(X_i)} \right] = 0$$

L'inégalité est stricte : la fonction log est strictement concave :

$$\mathbb{E} \left[\log \frac{f_\theta(X_i)}{f_{\theta^0}(X_i)} \right] = \log \mathbb{E} \left[\frac{f_\theta(X_i)}{f_{\theta^0}(X_i)} \right] \iff \frac{f_\theta(X_i)}{f_{\theta^0}(X_i)} = c \text{ p.s.}$$

Puisque $f_\theta(\cdot)$ est une densité, la constante c est égale à 1. Mais, $\frac{f_\theta(X_i)}{f_{\theta^0}(X_i)} = 1$, p.s. contredit le fait que P_θ est injective en θ . Lorsque :

$$\mathbb{E} \left| \log \frac{f_\theta(X_i)}{f_{\theta^0}(X_i)} \right| < \infty \quad (3.6)$$

la loi des grandes nombre permet de conclure que, p.s.

$$\frac{1}{n} \sum_{j=1}^n \log \frac{f_\theta(X_j)}{f_{\theta^0}(X_j)} \longrightarrow \mathbb{E} \left[\log \frac{f_\theta(X)}{f_{\theta^0}(X)} \right] < 0 \quad (3.7)$$

Lorsque (3.6) n'est pas vérifiée, alors, on montre que :

$$\mathbb{E} \left[\left(\log \frac{f_\theta(X_i)}{f_{\theta^0}(X_i)} \right)^- \right] = \infty \text{ et } \mathbb{E} \left[\left(\log \frac{f_\theta(X_i)}{f_{\theta^0}(X_i)} \right)^+ \right] < \infty$$

Pour montrer la deuxième partie (pour +), remarquons que :

$$\mathbb{E} \left[\left(\log \frac{f_\theta(X_i)}{f_{\theta^0}(X_i)} \right)^+ \right] = \int_{f_\theta(x) > f_{\theta^0}(x)} \log \frac{f_\theta(x)}{f_{\theta^0}(x)} f_{\theta^0}(x) dx$$

$$\begin{aligned}
&= \int_{f_\theta(x) > f_{\theta^0}(x)} |\log f_\theta(x)| f_{\theta^0}(x) dx + \int_{f_\theta(x) < f_{\theta^0}(x)} |\log f_{\theta^0}(x)| f_\theta(x) dx \\
&\leq \int_{f_\theta(x) > f_{\theta^0}(x)} |\log f_\theta(x)| f_\theta(x) dx + \int_{f_\theta(x) < f_{\theta^0}(x)} |\log f_{\theta^0}(x)| f_{\theta^0}(x) dx \\
&\leq \mathbb{E}_\theta [f_\theta(X)] + \mathbb{E}_{\theta^0} [f_{\theta^0}(X)] < \infty
\end{aligned}$$

Donc si (3.6) n'est pas vérifiée, alors : $\mathbb{E} \left[\left(\log \frac{f_\theta(X_i)}{f_{\theta^0}(X_i)} \right)^- \right] = \infty$.

Donc :

$$\frac{1}{n} \sum_{j=1}^n \log \frac{f_\theta(X_j)}{f_{\theta^0}(X_j)} \rightarrow -\infty, p.s. \quad (3.8)$$

Soit l'ensemble dénombrable : $\Theta_0 = \{\theta = \theta^0 \pm \frac{1}{k}, k \in \mathbb{N}\}$ et pour tout $\theta \in \Theta_0$, considérons N_θ l'ensemble des $\omega \in \Omega$ tels que (3.7) ou (3.8) ont lieu, $\mathbb{P}[N_\theta] = 1$. Puisque Θ_0 est dénombrable, l'intersection des N_θ est de probabilité égale à 1. Soit un ω appartenant à cette intersection. D'après la définition des N_θ , pour tout $\theta \in \Theta_0$, il existe une constante $-\infty \leq l(\theta) < 0$ telle que :

$$\frac{1}{n} \sum_{j=1}^n \log \frac{f_\theta(X_j(\omega))}{f_{\theta^0}(X_j(\omega))} \rightarrow l(\theta) \quad (3.9)$$

Pour un ε fixé, on choisit $k > \varepsilon^{-1}$ et posons $\theta_k = \theta^0 - \frac{1}{k}$ et $\theta'_k = \theta^0 + \frac{1}{k}$. D'après (3.9), il existe un entier $n_0(k, \omega)$ tel que, pour $n \geq n_0(k, \omega)$ on a :

$$\frac{1}{n} \sum_{j=1}^n \log \frac{f_{\theta_k}(X_j(\omega))}{f_{\theta^0}(X_j(\omega))} < 0 \quad \text{et} \quad \frac{1}{n} \sum_{j=1}^n \log \frac{f_{\theta'_k}(X_j(\omega))}{f_{\theta^0}(X_j(\omega))} < 0$$

Considérons la fonction de θ : $\frac{1}{n} \sum_{j=1}^n \log \frac{f_\theta(X_j(\omega))}{f_{\theta^0}(X_j(\omega))}$. Cette fonction est nulle pour $\theta = \theta^0$ et strictement négative pour $\theta = \theta_k$ et $\theta = \theta'_k$. Puisqu'elle est partout dérivable, il existe un point de $]\theta_k, \theta'_k[$ qui annule sa dérivée. Donc on a prouvé que pour tout $n \geq n_0(k, \omega)$, il existe dans l'intervalle $[\theta^0 - \varepsilon, \theta^0 + \varepsilon]$ un point $\hat{\theta}_n(\omega)$ qui est solution des équations de vraisemblance et en plus elle est presque sûrement convergente.

(ii) *Preuve de la Normalité asymptotique* Parce que $\hat{\theta}_n$ maximise $n^{-1} \log L_n(\theta) = \tilde{L}_n(\theta)$, on a : $\tilde{L}'_n(\hat{\theta}_n) = 0$. Alors, par le Théorème des accroissements finis, on a :

$$0 = \tilde{L}'_n(\hat{\theta}_n) = \tilde{L}'_n(\theta^0) + \tilde{L}''_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta^0)$$

avec $\tilde{\theta}_n \in [\hat{\theta}_n, \theta^0]$. Donc :

$$\sqrt{n}(\hat{\theta}_n - \theta^0) = -\sqrt{n} \frac{\tilde{L}'_n(\theta^0)}{\tilde{L}''_n(\tilde{\theta}_n)} \quad (3.10)$$

Mais (voir exercice TD) : $\mathbb{E}[\frac{\partial}{\partial \theta} \log f_\theta(X)] = 0$. Par le TCL :

$$\sqrt{n} \tilde{L}_n(\theta^0) = \sqrt{n} \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f_{\theta^0}(X) - 0 \right] \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, \text{Var}[\frac{\partial}{\partial \theta} \log f_{\theta^0}(X)])$$

En plus :

$$\text{Var}[\frac{\partial}{\partial \theta} \log f_{\theta^0}(X)] = \mathbb{E} \left[\frac{\partial}{\partial \theta} \log f_{\theta^0}(X) \right]^2 = -\mathbb{E} \left[\frac{\partial^2}{\partial \theta^2} \log f_{\theta^0}(X) \right] = I(\theta^0)$$

Pour le dénominateur de (3.10), par la loi des grands nombres, pour tout $\theta \in \Theta$:

$$\tilde{L}''_n(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log f_\theta(X_i) \rightarrow \mathbb{E}[\frac{\partial^2}{\partial \theta^2} \log f_\theta(X)] \quad p.s.$$

Parce que $\hat{\theta}_n \rightarrow \theta^0$ p.s., alors :

$$\tilde{L}''_n(\hat{\theta}_n) \rightarrow \mathbb{E} \left[\frac{\partial^2}{\partial \theta^2} \log f_{\theta^0}(X) \right] = -I(\theta^0)$$

En conclusion :

$$\sqrt{n} \frac{\tilde{L}'_n(\theta^0)}{\tilde{L}''_n(\hat{\theta}_n)} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, I(\theta^0)^{-1})$$

Théorème 3.1.5 Si T_n est un estimateur exhaustif pour θ et l'EMV existe, alors l'EMV est fonction de T_n .

Remarque. Le théorème ne dit pas que l'EMV est exhaustif.

Théorème 3.1.6 Soit la fonction $h : \Theta \subseteq \mathbb{R}^p \rightarrow \Lambda$, avec Λ un intervalle dans \mathbb{R}^m , $1 \leq m \leq p$. Si $\hat{\theta}_n$ est l'EMV de θ alors $h(\hat{\theta}_n)$ est l'EMV de $h(\theta)$.

Méthode des moindres carrés

Si Y est une variable (quantitative ou qualitative) et (X_1, \dots, X_k) sont des variables explicatives quantitatives pour lesquelles on a n mesures, on peut modéliser Y fonction de X_1, \dots, X_k par :

$$Y_i = g(X_{1,i}, \dots, X_{k,i}) + \varepsilon_i \quad i = 1, \dots, n \quad (3.11)$$

appelé *modèle de régression*.

La fonction g_θ dépend d'un ou de plusieurs paramètres inconnus que l'on doit estimer : $\theta \in \Theta \subseteq \mathbb{R}^p$. On considère le cas "simple" : X_1, \dots, X_k variables déterministes. La variable ε est aléatoire (erreur de mesure, erreur de modélisation), donc Y est une v.a. On suppose que ε_i et ε_j sont indépendantes pour $i \neq j$. Pour estimer le paramètre θ , on minimise l'erreur quadratique :

$$EQ(\theta) = \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 = \frac{1}{n} \sum_{i=1}^n [Y_i - g_\theta(X_{1,i}, \dots, X_{k,i})]^2$$

L'estimateur des moindres carrés est : $\hat{\theta}_n = \arg \min_\theta EQ(\theta)$.

Dans certains cas classiques on sait résoudre explicitement ce problème de minimisation. Si la résolution est impossible on fait appel à des algorithmes numériques de minimisation.

Hypothèse sur ε . $\mathbb{E}(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2 > 0$, $Cov(\varepsilon_i, \varepsilon_j) = 0$ pour $i \neq j$.

Dans le cas particulier $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, σ connu, la loi de Y_i est $\mathcal{N}(g(X_{1,i}, \dots, X_{k,i}), \sigma^2)$ et l'estimateur de θ par les moindres carrés coïncide avec l'EMV :

$$L(\theta; Y_1, \dots, Y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - g(X_{1,i}, \dots, X_{k,i}))^2 \right]$$

L maximal s.s.i $\sum_{i=1}^n (Y_i - g(X_{1,i}, \dots, X_{k,i}))^2$ minimal.

3.1.3 Familles exponentielles

Soit X une v.a. de densité (fréquence) $f_\theta(x)$, $\theta \in \Theta \subseteq \mathbb{R}$.

Définition. La densité f_θ est de type exponentiel si elle est de la forme :

$$f_\theta(x) = \exp \{C(\theta)T(x) + D(\theta) + S(x)\} \quad (3.12)$$

où $C, D : \Theta \rightarrow \mathbb{R}$ et T, S mesurables Borel.

eq COST st données à une obs mes

Supposons maintenant que $\theta = (\theta_1, \dots, \theta_p) \in \Theta \subseteq \mathbb{R}^p$, $p > 1$.

Définition. $\{f_\theta(x), \theta \in \Theta\}$ est une famille exponentielle s'il existe p fonctions réelles C_1, \dots, C_p , $D : \Theta \rightarrow \mathbb{R}$ et $T_1, \dots, T_p, S : \mathbb{R} \rightarrow \mathbb{R}$ fonctions mesurables Borel, telles que :

$$f_\theta(x) = \exp \left\{ \sum_{k=1}^p C_k(\theta)T_k(x) + D(\theta) + S(x) \right\}$$

Théorème 3.1.7 ($p=1$) Si $f_\theta(x)$ est une famille exponentielle, alors la v.a. $T(X)$ est aussi une famille exponentielle de densité :

$$g_\theta(t) = \exp \{tC(\theta) + D(\theta) + S^*(t)\}$$

pour une $S^*(t)$ souhaitable.

Il existe un théorème analogue pour $p > 1$.

Théorème 3.1.8 Pour les familles exponentielles

- 1) la statistique $\sum_{i=1}^n T(X_i)$ est de type exponentiel et elle est exhaustive pour le paramètre θ .
- 2) l'EMV est une fonction de $\sum_{i=1}^n T(X_i)$.

Théorème 3.1.9 Si $C(\theta)$ est de classe C^2 et $C'(\theta) \neq 0$, alors $n^{-1} \sum_{i=1}^n T(X_i)$ est un estimateur sans biais de $\psi(\theta) = \mathbb{E}[T(X)]$ et en plus il est efficace.

3.2 Estimateur par intervalle

Donner un résultat sans indication sur sa précision n'a que peu d'intérêt car il n'est pas reproductible. Plutôt que de donner une estimation ponctuelle on propose un intervalle, choisi de manière à contrôler par un niveau de confiance, les chances que le résultat aurait d'être confirmé si on renouvelait l'expérience. Soit (X_1, \dots, X_n) un échantillon de la loi P_θ , $\theta \in \Theta \subseteq \mathbb{R}$ et $\alpha \in (0, 1)$.

Définition. On appelle *estimateur par intervalle* *intervalle de niveau* $(1-\alpha)$ pour θ , un intervalle aléatoire $[A_n, B_n]$ avec A_n et B_n des variables aléatoires fonction de l'échantillon tels que :

$$P[A_n \leq \theta \leq B_n] = \begin{cases} 1 - \alpha & \text{si } P_\theta \text{ est continue} \\ \geq 1 - \alpha & \text{si } P_\theta \text{ est discrète} \end{cases}$$

Si a_n, b_n sont des réalisations pour A_n et B_n alors on obtient un intervalle réel : *intervalle de confiance*.

Définition Soit la constante $\alpha \in (0, 1)$. On appelle *fractile d'ordre* α la valeur u_α telle que : $\alpha = P[W \leq \alpha] = F(u_\alpha)$, avec F la fonction de répartition de la v.a. X . En fait, la fractile est l'inverse de la fonction de répartition : $u_\alpha = F^{-1}(\alpha)$.

Démarche à faire pour trouver l'estimateur par intervalle :

- Le point de départ est un estimateur ponctuel T_n , sans biais, du paramètre et pour lequel on connaît sa loi ;
- On considère éventuellement une v.a. transformée Y_n de T_n , la loi de Y_n ne dépendant plus de θ .
- On considère l'égalité $P[a \leq Y_n \leq b] = 1 - \alpha$ et on déduit a et b fonction de α . $P[a \leq Y_n \leq b] = F(b) - F(a) = 1 - \alpha$. Alors, on prend a et b tels que : $P[Y_n < b] = 1 - \alpha/2$ et $F(a) = \alpha/2$. D'où : $b = u_{1-\alpha/2}$ et $a = u_{\alpha/2}$. Si la loi de Y_n est symétrique, alors : $\frac{\alpha}{2} = P[Y_n \leq a] = P[-Y_n \leq a] = P[Y_n > -a] = 1 - P[Y_n \leq -a]$. D'où $P[Y_n \leq -a] = 1 - \alpha/2$ et $P[Y_n \leq b] = 1 - \alpha/2$. Donc $b = -a = -u_{\alpha/2}$.
- On écrit Y_n fonction de T_n et on obtient les bornes A_n et B_n .

↳ Soit $X_n \sim (\mu, \sigma, P_\theta)$ $\theta \in \Theta \subseteq \mathbb{R}$

$$\begin{cases} A_n = A(X_1, \dots, X_n) \\ B_n = B(X_1, \dots, X_n) \end{cases}$$

Si (x_1, \dots, x_n) est une réalisation de (X_1, \dots, X_n) alors $\begin{cases} a_n = A(x_1, \dots, x_n) \\ b_n = B(x_1, \dots, x_n) \end{cases}$

Alors $[a_n, b_n]$ s'appelle intervalle de confiance.

\Rightarrow Construire l'estimateur par intervalle = trouver A et B .

Construction.

Def: Une hypothèse statistique est un énoncé concernant la loi de la v.a. X
 On teste toujours x hypothèse $\begin{cases} H_0 & \text{hypothèse nulle} \\ H_1 & \text{hypothèse alternative} \end{cases}$
 si $P = P_0$ si H_0 et H_1 sur $\Theta \rightarrow$ test d'hop paramétrique

Chapitre 4

THEORIE DES TESTS

Supposons qu'une machine produit des objets dont certains sont défectueux. Soit θ la probabilité que l'objet soit défectueux. Le fabricant désire avoir $\theta \leq \theta_0$ avec θ_0 donné, faute de quoi il doit réviser ou changer la machine.

Considérons une variable aléatoire X définie sur l'espace Ω et de loi de probabilité \mathbb{P} . Supposons qu'on ne connaît pas \mathbb{P} mais on sait qu'elle peut être seulement une des deux distributions \mathbb{P}_0 ou \mathbb{P}_1 .

def: Une *hypothèse statistique* est un énoncé concernant les caractéristiques (valeurs des paramètres, forme de distribution, ...) d'une ou de plusieurs populations (variables aléatoires).

Le *test statistique* (d'hypothèse) est une démarche qui a pour but de fournir une règle de décision permettant sur la base des résultats de l'échantillon de faire le choix entre deux hypothèses statistiques.

Les hypothèses qui sont envisagées à priori s'appellent, l'*hypothèse nulle* (H_0) et l'*hypothèse alternative* (H_1).

Pour réaliser des test on considère un n -échantillon (X_1, \dots, X_n) et une réalisation (x_1, \dots, x_n) . Sur la base de (X_1, \dots, X_n) on veut décider quelle hypothèse est vraie : $H_0 : \mathbb{P} = \mathbb{P}_0$ ou $H_1 : \mathbb{P} = \mathbb{P}_1$

Pour fournir une règle de décision on utilise une statistique de test. Toute fonction mesurable Borel $\varphi : \Omega^n \rightarrow [0, 1]$ s'appelle fonction *fonction de test*.

La fonction φ est un test de l'hypothèse H_0 contre H_1 avec l'erreur de probabilité α si : $\mathbb{E}[\varphi(X_1, \dots, X_n)] \leq \alpha$ sous H_0 .

Pour décider quelle hypothèse est vraie, on considère une fonction de décision : $\delta : \Omega^n \rightarrow \{H_0, H_1\}$. Si l'hypothèse H_0 est vraie alors $\mathbb{P} = \mathbb{P}_0$. Alors la probabilité que la décision δ fasse une erreur est :

$$\mathbb{P}[\delta(X_1, \dots, X_n) \neq H_0 | H_0] = \mathbb{P}_0[\delta(X_1, \dots, X_n) \neq H_0]$$

et cette probabilité s'appelle *risque de première espèce*.

4.1 Tests paramétriques

On considère que la loi de la v.a. X dépend d'un paramètre θ et on veut faire un test sur ce paramètre : on a à faire à des tests paramétriques.

On teste :

$H_0 : \theta \in \Theta_0$, appelée *hypothèse nulle* (parce qu'elle s'écrit sous la forme $g(\theta) = 0$)

contre :

$H_1 : \theta \in \Theta_1$ l'*hypothèse alternative*

avec $\Theta_0 \cap \Theta_1 = \emptyset$, $\Theta_0 \cup \Theta_1 \subseteq \Theta$.

ie $\Theta_0 = \{\theta_0\}$

Si Θ_0 est formée d'un seul élément on dit que H_0 est une *hypothèse simple*, sinon elle est *composite*.

Pour faire le test on a besoin d'une règle de décision : soit $T_n = T(X_1, \dots, X_n)$ une *statistique de test* et R un sous-ensemble de valeurs possibles de T , appelée *région de rejet* $R = \{(x_1, \dots, x_n) \in \Omega^n; H_1 \text{ acceptée}\}$. Si $T(x_1, \dots, x_n) \in R$ on rejette H_0 et on accepte H_1 . La construction de R est basée sur la connaissance de la loi de T_n sous H_0 .

Définitions : 1) On appelle *risque de première espèce* et on note $\alpha(\theta)$, la probabilité de rejeter H_0 alors qu'elle est vraie :

$$\alpha(\theta) = \mathbb{P}[\delta(X_1, \dots, X_n) = H_1 / \theta \in \Theta_0] = \mathbb{P}[(X_1, \dots, X_n) \in R / \theta \in \Theta_0]$$

On appelle *niveau*, noté α , la valeur la plus élevée du risque de première espèce quand θ parcourt Θ_0 :

$$\alpha = \sup_{\theta \in \Theta_0} \alpha(\theta)$$

Si $H_0 : \theta = \theta_0$ alors $\alpha = \alpha(\theta_0)$.

2) On appelle *risque de deuxième espèce*, noté $\beta(\theta)$, la probabilité d'accepter H_0 alors qu'elle est fautive : $\beta(\theta) = \mathbb{P}[\delta(X_1, \dots, X_n) = H_0 / \theta \in \Theta_1] = \mathbb{P}[(X_1, \dots, X_n) \in R^c / \theta \in \Theta_1]$.

3) On appelle *puissance*, noté $\pi(\theta)$, la probabilité de rejeter H_0 alors qu'elle est fautive. On a $\pi(\theta) = 1 - \beta(\theta)$.

4) *Région de rejet* : $R = \{(x_1, \dots, x_n); H_0 \text{ rejetée}\}$ telle que $\alpha(\theta) = \mathbb{P}[(X_1, \dots, X_n) \in R | H_0 \text{ vraie}]$. Donc R dépend de α .

Alors, la démarche à suivre pour effectuer un test d'hypothèse :

1. Choisir H_0 et H_1 de sorte que la possibilité d'égalité soit dans H_0 ;
2. Fixer α ;
3. Déterminer la région de rejet R ;
4. Regarder si les observations se trouvent ou pas dans R ;
5. Conclure au rejet ou au non rejet de H_0 .

4.1.1 Lemme de Neyman-Pearson

L'idée de base pour la construction de tests est : on fixe un niveau α à une valeur (petite) et on trouve le test de niveau α qui ait une puissance assez grande. Evidemment une idée est d'utiliser, pour construire ce test, un bon estimateur $\hat{\theta}_n$ de θ si on en connaît un. Dans ce cas la région critique portera sur $\hat{\theta}_n$. Pour le même problème de décision, plusieurs tests de même seuil sont souvent possibles. Dans ce cas, le meilleur est celui qui minimise $\beta(\theta)$, donc qui maximise $\pi(\theta)$. On obtient ainsi le *test le uniformément plus puissant (UPP)*.

Le lemme suivant nous donne une manière de trouver le test UPP.

$H_0 : \theta = \theta_0$

$H_1 : \theta = \theta_1$ avec $\theta_1 < \theta_0$ ou $\theta_1 > \theta_0$ ou $\theta_1 \neq \theta_0$.

Lemme 4.1.1 (Lemme de Neyman-Pearson)

Pour un seuil α fixé, soit $L_0 = L(\theta_0)$ et $L_1 = L(\theta_1)$ la densité de (X_1, \dots, X_n) sous H_0 , respectivement H_1 . Alors, il existe une constante $k > 0$ telle que :

$$\delta(X_1, \dots, X_n) = \begin{cases} H_1 & \text{si } L_1(X_1, \dots, X_n) > k L_0(X_1, \dots, X_n) \\ H_0 \text{ ou } H_1 & \text{si } L_1(X_1, \dots, X_n) = k L_0(X_1, \dots, X_n) \\ H_0 & \text{si } L_1(X_1, \dots, X_n) < k L_0(X_1, \dots, X_n) \end{cases}$$

est un test de seuil α et il est le plus puissant (c'est-à-dire $\mathbb{P}_0[\delta(X_1, \dots, X_n) \neq H_0] \leq \alpha$).

Remarque. Ce lemme nous donne aussi la forme de la zone de rejet R :

$$R = \left\{ (x_1, \dots, x_n) / \frac{L_1(x_1, \dots, x_n)}{L_0(x_1, \dots, x_n)} > k \right\}$$

Pour les tests $H_0 : \theta \leq \theta_0$ contre $H_1 : \theta > \theta_1$, en général, il n'est pas possible de trouver le test le plus puissant.

On peut dire que la procédure de test consiste à rejeter l'hypothèse H_0 dans une certaine région R et accepter dans la région complémentaire, on va convenir alors qu'effectuer un test consiste, en chaque point $(X_1, \dots, X_n) \in \bar{X}_n$ à rejeter H_0 avec une certaine probabilité $\Phi(X_1, \dots, X_n)$ et à l'accepter avec la probabilité $1 - \Phi(X_1, \dots, X_n)$. Un test est alors une application $\Phi : \bar{X}_n \rightarrow [0, 1]$ appelée fonction de test. Alors, le niveau du test est : $\alpha = \sup \{ \mathbb{E}[\Phi(X_1, \dots, X_n)] | \theta \in \Theta_0 \}$, la puissance est : $\mathbb{E}[\Phi(X_1, \dots, X_n)] | \theta \in \Theta_1$.

Définition On dit que le test Φ est sans biais si $\mathbb{E}[\Phi(X_1, \dots, X_n)] \geq \alpha, \forall \theta \in \Theta_1$.

Avec ces notations le test de Neyman-Pearson a la fonction de test de la forme :

$$\Phi = \mathbb{1}_{L(\theta_1) > k L(\theta_0)} + \gamma \mathbb{1}_{L(\theta_1) = k L(\theta_0)} \quad (4.1)$$

et le Lemme de Neyman-Pearson peut-être aussi énoncé sous la forme :

Lemme 4.1.2 (Lemme de Neyman-Pearson) (bis)

Soit $\alpha \in (0, 1)$ fixé ; Pour tester H_0 contre H_1 , spécifiées plus haut, il existe $\gamma \in [0, 1]$ et $k \geq 0$ tels que le test (4.1) a les propriétés suivantes :

1. $\mathbb{E}_{\theta_0} \Phi(X_1, \dots, X_n) = \alpha$, avec \mathbb{E}_{θ_0} l'espérance calculée sous l'hypothèse H_0 ;
2. $\mathbb{E}_{\theta_1} \Phi(X_1, \dots, X_n) \geq \alpha$, avec \mathbb{E}_{θ_1} l'espérance calculée sous l'hypothèse H_1 ;
3. Pour toute autre fonction test Φ' telle que $\mathbb{E}_{\theta_0} \Phi'(X_1, \dots, X_n) \leq \alpha$ on a :

$$\mathbb{E}_{\theta_1} \Phi'(X_1, \dots, X_n) \leq \mathbb{E}_{\theta_1} \Phi(X_1, \dots, X_n)$$

Preuve. 1. Remarquons que : $P_{\theta_0}[L(\theta_0) = 0] = 0$ et considérons l'événement aléatoire : $C = \{L(\theta_0) \neq 0\}$. On a :

$$\Psi(z) = P_{\theta_0}[L(\theta_1) > zL(\theta_0)] = P_{\theta_0}\left[\frac{L(\theta_1)}{L(\theta_0)} \mathbb{1}_C > z\right] = 1 - P_{\theta_0}\left[\frac{L(\theta_1)}{L(\theta_0)} \mathbb{1}_C \leq z\right]$$

Puisque $P\left[\frac{L(\theta_1)}{L(\theta_0)} \mathbb{1}_C \leq z\right]$ est une fonction de répartition, alors pour tout z elle est continue à droite et admet une limite à gauche. Donc la fonction Ψ a les mêmes propriétés et en plus elle est décroissante, avec les propriétés :

- $\Psi(z) = 1$ pour $z < 0$;
- $\Psi(0) = P_{\theta_0}[L(\theta_1) > 0]$;
- $\Psi(z) \rightarrow 0$ quand $z \rightarrow \infty$.

Soit la constante : $k = \inf\{z \geq 0 : \Psi(z) < \alpha\}$. Alors on a :

$$\Psi(k) \leq \alpha \leq \Psi(-k) \quad (4.2)$$

On a deux cas :

situation 1 : Ψ est continue au point k . Alors dans la relation (4.2) il y a égalité parce que Ψ décroissante, et le test définit par $\Phi = \mathbb{1}_{L(\theta_1) > kL(\theta_0)}$ qui a la région critique $L(\theta_1) > L(\theta_0)$ est de niveau exactement α .

situation 2 : la fonction Ψ a un saut au point k . Ce saut est d'amplitude : $\Psi(-k) - \Psi(k) = P_{\theta_0}[L(\theta_1) = kL(\theta_0)]$ et on choisit la constante $\gamma : \gamma = \frac{\alpha - \Psi(k)}{\Psi(-k) - \Psi(k)}$. On en déduit :

$\alpha = \Psi(k) + \gamma[\Psi(-k) - \Psi(k)] = P_{\theta_0}[L(\theta_1) > kL(\theta_0)] + \gamma P_{\theta_0}[L(\theta_1) = kL(\theta_0)]$. On obtient un test de niveau α .

3. Soit Φ' un test de niveau au plus α . On a alors :

$$E_{\theta_1}[\Phi' - \Phi] = \int_{\mathcal{X}^n} [\Phi' - \Phi] L(\theta_1) d\mu^{\otimes n} = \int_{\mathcal{A}_1} [\Phi' - \Phi] L(\theta_1) d\mu^{\otimes n} + \int_{\mathcal{A}_2} [\Phi' - \Phi] L(\theta_1) d\mu^{\otimes n} + \int_{\mathcal{A}_3} [\Phi' - \Phi] L(\theta_1) d\mu^{\otimes n}$$

où : $\mathcal{A}_1 = \{L(\theta_1) > kL(\theta_0)\}$, $\mathcal{A}_2 = \{L(\theta_1) = kL(\theta_0)\}$, $\mathcal{A}_3 = \{L(\theta_1) < kL(\theta_0)\}$. Sur \mathcal{A}_1 on a $\Phi = 1$ et donc $\Phi' - \Phi \leq 0$ et

$$\int_{\mathcal{A}_1} [\Phi' - \Phi] L(\theta_1) d\mu^{\otimes n} \leq k \int_{\mathcal{A}_1} [\Phi' - \Phi] L(\theta_0) d\mu^{\otimes n}$$

Sur \mathcal{A}_3 on a $\Phi = 0$, donc $\Phi' - \Phi \geq 0$ et

$$\int_{\mathcal{A}_3} [\Phi' - \Phi] L(\theta_1) d\mu^{\otimes n} = k \int_{\mathcal{A}_3} [\Phi' - \Phi] L(\theta_0) d\mu^{\otimes n}$$

De plus :

$$\int_{\mathcal{A}_2} [\Phi' - \Phi] L(\theta_1) d\mu^{\otimes n} \leq k \int_{\mathcal{A}_2} [\Phi' - \Phi] L(\theta_0) d\mu^{\otimes n}$$

En conclusion : $E_{\theta_1}[\Phi' - \Phi] \leq k E_{\theta_0}[\Phi' - \Phi] k (E_{\theta_0}[\Phi'] - \alpha) \leq 0$.

2. Elle se déduit facilement de la propriété 3 en considérant la fonction test $\Phi'(X_1, \dots, X_n) \equiv \alpha$. Ce test est de niveau α et il a une puissance inférieure ou égale à celle de Φ . Mais la puissance de Φ' est égale à α . ■

Soit L_θ la vraisemblance pour un n-échantillon, pour le paramètre θ .

Définition. On dit que la v.a. X a un *rapport de vraisemblance monotone (MVR)* par rapport à la statistique $T(X_1, \dots, X_n)$ si pour $\theta_1 < \theta_2$, le rapport $L(\theta_2)/L(\theta_1)$ est une fonction non-décroissante de $T(X_1, \dots, X_n)$.

Exemple. Soit $X_1, \dots, X_n \sim \mathcal{U}[0, \theta]$, $\theta > 0$. La distribution jointe des X_1, \dots, X_n est : $L(\theta, x_1, \dots, x_n) = \theta^{-n} \mathbb{1}_{0 \leq \max x_i \leq \theta}$. Si $\theta_1 < \theta_2$ alors considérons le rapport :

$$\frac{L(\theta_2, x_1, \dots, x_n)}{L(\theta_1, x_1, \dots, x_n)} = \left(\frac{\theta_2}{\theta_1}\right)^n \mathbb{1}_{0 \leq \max x_i \leq \theta_2} / \mathbb{1}_{0 \leq \max x_i \leq \theta_1}$$

Soit $R(x_1, \dots, x_n) = \mathbb{1}_{0 \leq \max x_i \leq \theta_2} / \mathbb{1}_{0 \leq \max x_i \leq \theta_1} = 1$ si $\max x_i \in [0, \theta_1]$ et $= \infty$ si $\max x_i \in [\theta_1, \theta_2]$. On définit $R(x_1, \dots, x_n) = \infty$ si $\max x_i > \theta_2$. Donc $L(\theta_2)/L(\theta_1)$ est croissante en $\max x_i$ d'où : la loi uniforme est MLR par rapport à $\max x_i$.

Théorème 4.1.1 Si $L_\theta(X_1, \dots, X_n)$ est MVR par rapport à $T(X_1, \dots, X_n)$, pour tester $H_0 : \theta \leq \theta_0$ contre $H_1 : \theta > \theta_0$, il existe $t_0 \in \mathbb{R}$ tel que :

$$\varphi(X_1, \dots, X_n) = \begin{cases} H_1 & \text{si } T(X_1, \dots, X_n) > t_0 \\ H_0 \text{ ou } H_1 & \text{si } T(X_1, \dots, X_n) = t_0 \\ H_0 & \text{si } T(X_1, \dots, X_n) < t_0 \end{cases}$$

est le test le plus puissant.

Remarque. $H_0 : \theta \geq \theta_0$ contre $H_1 : \theta < \theta_0$ alors :

$$\delta(X_1, \dots, X_n) = \begin{cases} H_1 & \text{si } T(X_1, \dots, X_n) < t_0 \\ H_0 \text{ ou } H_1 & \text{si } T(X_1, \dots, X_n) = t_0 \\ H_0 & \text{si } T(X_1, \dots, X_n) > t_0 \end{cases}$$

A. Tests sur une population

A.1. Tests sur la moyenne d'une loi Normale

On considère un n-échantillon X_1, \dots, X_n avec $X_i \sim \mathcal{N}(m, \sigma^2)$

Cas σ^2 connue

Notons par u_α la *fractile (quartile)* d'ordre α pour la loi Normale : si $Y \sim \mathcal{N}(0, 1)$ alors

$$IP[Y < u_\alpha] = \alpha$$

1) $H_0 : m = m_0$ contre $m \neq m_0$. Statistique de test :

$$Z = \sqrt{n} \frac{\bar{X}_n - m_0}{\sigma} \sim \mathcal{N}(0, 1) \quad \text{sous } H_0 \quad (4.3)$$

Zone de rejet

$$R = \left\{ (x_1, \dots, x_n) / \sqrt{n} \frac{|\bar{x}_n - m_0|}{\sigma} > u_{1-\alpha/2} \right\}$$

Evidemment $u_\alpha = -u_{1-\alpha}$.

2) $H_0 : m \leq m_0$ contre $m > m_0$ ou $H_0 : m = m_0$ contre $m > m_0$. Statistique de test : (4.8). Zone de rejet :

$$R = \left\{ (x_1, \dots, x_n) / \sqrt{n} \frac{\bar{x}_n - m_0}{\sigma} > u_{1-\alpha} \right\} = \left\{ (x_1, \dots, x_n) / \bar{x}_n > m_0 + \frac{\sigma}{\sqrt{n}} u_{1-\alpha} \right\}$$

3) $H_0 : m \geq m_0$ contre $m < m_0$ ou $H_0 : m = m_0$ contre $m < m_0$. Statistique de test : (4.8). Zone de rejet :

$$R = \left\{ (x_1, \dots, x_n) / \sqrt{n} \frac{\bar{x}_n - m_0}{\sigma} < u_\alpha \right\} = \left\{ (x_1, \dots, x_n) / \bar{x}_n < m_0 + \frac{\sigma}{\sqrt{n}} u_\alpha \right\}$$

Cas σ^2 inconnue

Notons par $t_{p,\alpha}$ la *fractile (quartile)* d'ordre α pour la loi Student à p degrés de liberté : si $Y \sim t(p)$ alors $IP[Y < t_{p,\alpha}] = \alpha$. On remplace σ^2 par son estimateur sans biais

$$S^{*2} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

1) $H_0 : m = m_0$ contre $m \neq m_0$. Statistique de test :

$$Z = \sqrt{n} \frac{\bar{X}_n - m_0}{S^*} \sim t(n-1) \quad \text{sous } H_0 \quad (4.4)$$

Zone de rejet :

$$R = \left\{ (x_1, \dots, x_n) / \sqrt{n} \frac{|\bar{x}_n - m_0|}{s^*} > t_{n-1, 1-\alpha/2} \right\}$$

2) $H_0 : m \leq m_0$ contre $m > m_0$ ou $H_0 : m = m_0$ contre $m > m_0$. Statistique de test : (4.4). Zone de rejet :

$$R = \left\{ (x_1, \dots, x_n) / \sqrt{n} \frac{\bar{x}_n - m_0}{s^*} > t_{n-1, 1-\alpha} \right\} = \left\{ (x_1, \dots, x_n) / \bar{x}_n > m_0 + \frac{s^*}{\sqrt{n}} t_{n-1, 1-\alpha} \right\}$$

3) $H_0 : m \geq m_0$ contre $m < m_0$ ou $H_0 : m = m_0$ contre $m < m_0$. Statistique de test : (4.4). Zone de rejet :

$$R = \left\{ (x_1, \dots, x_n) / \sqrt{n} \frac{\bar{x}_n - m_0}{s^*} < t_{n-1, \alpha} \right\} = \left\{ (x_1, \dots, x_n) / \bar{x}_n < m_0 + \frac{s^*}{\sqrt{n}} t_{n-1, \alpha} \right\}$$

A.2. Tests sur la variance d'une loi Normale

On considère un n-échantillon X_1, \dots, X_n avec $X_i \sim \mathcal{N}(m, \sigma^2)$.

Notons par $z_{p,\alpha}$ la *fractile (quartile)* d'ordre α pour la loi χ^2 à p degrés de liberté : si $Y \sim \chi^2(p)$ alors $\mathbb{P}[Y < z_{p,\alpha}] = \alpha$.

1) $H_0 : \sigma = \sigma_0$ contre $\sigma \neq \sigma_0$. Statistique de test :

- m connue

$$Z = \frac{1}{\sigma_0^2} \sum_{i=1}^n (X_i - m)^2 \sim \chi^2(n) \quad \text{sous } H_0 \quad (4.5)$$

- m inconnue

$$Z = \frac{1}{\sigma_0^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \sim \chi^2(n-1) \quad \text{sous } H_0 \quad (4.6)$$

Zone de rejet

- m connue

$$R = \left\{ \frac{1}{\sigma_0^2} \sum_{i=1}^n (x_i - m)^2 > z_{n,1-\alpha} \right\}$$

- m inconnue

$$R = \left\{ \frac{1}{\sigma_0^2} \sum_{i=1}^n (x_i - \bar{x}_n)^2 > z_{n-1,1-\alpha} \right\}$$

2) $H_0 : \sigma \leq \sigma_0$ contre $\sigma > \sigma_0$. Zone de rejet :

- m connue

$$R = \left\{ (x_1, \dots, x_n) / \frac{1}{\sigma_0^2} \sum_{i=1}^n (x_i - m)^2 > z_{n,1-\alpha} \right\}$$

- m inconnue

$$R = \left\{ (x_1, \dots, x_n) / \frac{1}{\sigma_0^2} \sum_{i=1}^n (x_i - \bar{x}_n)^2 > z_{n-1,1-\alpha} \right\}$$

3) $H_0 : \sigma \geq \sigma_0$ contre $\sigma < \sigma_0$. Zone de rejet :

- m connue

$$R = \left\{ (x_1, \dots, x_n) / \frac{1}{\sigma_0^2} \sum_{i=1}^n (x_i - m)^2 < z_{n,\alpha} \right\}$$

- m inconnue

$$R = \left\{ (x_1, \dots, x_n) / \frac{1}{\sigma_0^2} \sum_{i=1}^n (x_i - \bar{x}_n)^2 < z_{n-1,\alpha} \right\}$$

A.3. Tests sur une proportion

On considère un n-échantillon X_1, \dots, X_n avec $X_i \sim B(p)$

1) $H_0 : p = p_0$ contre $p \neq p_0$. Statistique de test :

$$Z = \sqrt{n} \frac{\bar{X}_n - p_0}{\sqrt{p_0(1-p_0)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \quad \text{pour } n \rightarrow \infty \quad (4.7)$$

Zone de rejet : $R = \{(x_1, \dots, x_n) / |z| > u_{1-\alpha/2}\}$.

2) $H_0 : p \leq p_0$ contre $p > p_0$. Zone de rejet : $R = \{(x_1, \dots, x_n) / z > u_{1-\alpha}\}$.

3) $H_0 : p \geq p_0$ contre $p < p_0$. Zone de rejet : $R = \{(x_1, \dots, x_n) / z < -u_\alpha\}$.

B. Tests sur deux populations Normales

Soient deux variables aléatoires $X \sim \mathcal{N}(m_1, \sigma_1^2)$ et $Y \sim \mathcal{N}(m_2, \sigma_2^2)$ pour lesquelles nous considérons deux échantillons X_1, \dots, X_{n_1} , respectivement Y_1, \dots, Y_{n_2} .

1) $H_0 : m_1 = m_2$ contre $H_1 : m_1 \neq m_2$, si σ_1^2, σ_2^2 connues. Statistique de test :

$$Z = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1) \quad \text{sous } H_0 \quad (4.8)$$

Zone de rejet

$$R = \left\{ (x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2}) / |\bar{x}_{n_1} - \bar{y}_{n_2}| > u_{1-\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right\}$$

2) $H_0 : m_1 = m_2$ contre $H_1 : m_1 \neq m_2$, si $\sigma_1^2 = \sigma_2^2$ inconnues. Statistique de test :

$$Z = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2) \quad \text{sous } H_0 \quad (4.9)$$

où

$$S^2 = \left[\sum_{i=1}^{n_1} (X_i - \bar{X}_{n_1})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y}_{n_2})^2 \right] / (n_1 + n_2 - 2)$$

3) $H_0 : m_1 = m_2$ contre $H_1 : m_1 \neq m_2$, si $\sigma_1^2 \neq \sigma_2^2$ inconnues. on a dans ce cas le problème de Fisher-Behrens. La statistique de test (Welch) est :

$$t_{mc} = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{\sqrt{\frac{S_{n_1}^{*2}}{n_1} + \frac{S_{n_2}^{*2}}{n_2}}}$$

mais elle ne suit plus, sous H_0 , une loi $t(n_1 + n_2 - 2)$. On peut approximer cette loi par une loi de Student de degrés de liberté :

$$\nu_{Welch} = \frac{\left[\frac{S_{n_1}^{*2}}{n_1} + \frac{S_{n_2}^{*2}}{n_2} \right]^2}{\frac{(S_{n_1}^{*2}/n_1)^2}{n_1 - 1} + \frac{(S_{n_2}^{*2}/n_2)^2}{n_2 - 1}}$$

qui est une variable aléatoire! On arrondit ce nombre de degrés de liberté à l'entier le plus proche.

4) $H_0 : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_1^2 \neq \sigma_2^2$. Si les moyennes m_1 et m_2 sont inconnues, alors on les estime par \bar{X}_{n_1} et \bar{Y}_{n_2} . Dans ce cas, la statistique de test :

$$Z = \frac{\max(S_{n_1}^{*2}, S_{n_2}^{*2})}{\min(S_{n_1}^{*2}, S_{n_2}^{*2})} \sim F(n_1 - 1, n_2 - 1), \text{ si } \max(S_{n_1}^{*2}, S_{n_2}^{*2}) = S_{n_1}^{*2} \quad (4.10)$$

où

$$S_{n_1}^{*2} = \left[\sum_{i=1}^{n_1} (X_i - \bar{X}_{n_1})^2 \right] / (n_1 - 1), \quad S_{n_2}^{*2} = \left[\sum_{i=1}^{n_2} (Y_i - \bar{Y}_{n_2})^2 \right] / (n_2 - 1)$$

Si les moyennes m_1 et m_2 sont connues, la statistique de test :

$$Z = \frac{\max(S_{n_1}^2, S_{n_2}^2)}{\min(S_{n_1}^2, S_{n_2}^2)} \sim F(n_1, n_2), \text{ si } \max(S_{n_1}^2, S_{n_2}^2) = S_{n_1}^2 \quad (4.11)$$

où

$$S_{n_1}^2 = \left[\sum_{i=1}^{n_1} (X_i - m_1)^2 \right] / n_1, \quad S_{n_2}^2 = \left[\sum_{i=1}^{n_2} (Y_i - m_2)^2 \right] / n_2$$

4.1.2 Test du rapport de vraisemblance et de Wald

On a vu dans la Section précédente que le test le plus puissant (UPP) n'existe pas toujours. Dans cette section nous proposons une solution alternative.

Soit $\theta \in \Theta \subset \mathbb{R}^k$ un vecteur paramètre et X un vecteur aléatoire de densité (fonction de fréquence) f_θ . Considérons le problème du test d'hypothèse $H_0 : X \sim f_\theta, \theta \in \Theta_0$ contre l'hypothèse alternative $H_1 : X \sim f_\theta, \theta \in \Theta_1$.

Définition. Pour tester H_0 contre H_1 , un test de la forme : on rejette H_0 si et seulement si $\lambda(x) < c$, avec c une constante et

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L_n(x_1, \dots, x_n; \theta)}{\sup_{\theta \in \Theta} L_n(x_1, \dots, x_n; \theta)}$$

s'appelle un *test du rapport de vraisemblance*. ($x = (x_1, \dots, x_n)$)

Evidemment $0 \leq \lambda(x) \leq 1$. La constante c est déterminée à partir de la condition :

$$\sup_{\theta \in \Theta_0} P[X = (X_1, \dots, X_n); \lambda(X) < c] = \alpha$$

Remarque. Si H_0 est vraie alors $\lambda(x)$ converge vers 1, pendant que si c'est H_1 qui est vraie, alors $\lambda(x)$ s'éloigne de 1.

Remarque. (Lien avec l'EMV) $\lambda(x) = L_n(\hat{\theta}_0) / L_n(\hat{\theta}_n)$, avec $\hat{\theta}_n$ l'EMV sur Θ et $\hat{\theta}_0$ l'EMV sur Θ_0 .

Théorème 4.1.2 Pour un α fixé, $0 \leq \alpha \leq 1$, les tests du Neyman-Pearson et du rapport de vraisemblance d'une hypothèse simple H_0 contre une hypothèse H_1 simple, sont équivalents.

Exemple 1. Soit $X \sim \mathcal{B}(m, p)$. On teste l'hypothèse $H_0 : p \leq p_0$ contre $H_1 : p > p_0$. Dans ce cas :

$$\lambda(x) = \frac{\sup_{p \leq p_0} C_m^x p^x (1-p)^{m-x}}{\sup_{0 \leq p \leq 1} C_m^x p^x (1-p)^{m-x}}$$

Mais $\sup_{0 \leq p \leq 1} p^x (1-p)^{m-x} = \left(\frac{x}{m}\right)^x \left(1 - \frac{x}{m}\right)^{m-x}$, donc, puisque la fonction $p^x (1-p)^{m-x}$ est croissante et son maximum est atteint dans $p = x/m$, on a

$$\sup_{p \leq p_0} p^x (1-p)^{m-x} = \begin{cases} p_0^x (1-p_0)^{m-x} & \text{si } p_0 < \frac{x}{m} \\ \left(\frac{x}{m}\right)^x \left(1 - \frac{x}{m}\right)^{m-x} & \text{si } \frac{x}{m} \leq p_0 \end{cases}$$

Ce qui implique

$$\lambda(x) = \begin{cases} \frac{p_0^x (1-p_0)^{m-x}}{\left(\frac{x}{m}\right)^x \left(1 - \frac{x}{m}\right)^{m-x}} & \text{si } p_0 < \frac{x}{m} \\ 1 & \text{si } \frac{x}{m} \leq p_0 \end{cases}$$

Notons que $\lambda(x) \leq 1$ pour $mp_0 < x$ et $\lambda(x) = 1$ si $x \leq mp_0$, donc $\lambda(x)$ est une fonction décroissante en x . Alors, $\lambda(x) < c$ si et seulement si $x > c'$, et le test du rapport de vraisemblance rejette H_0 si $x > c'$. D'autre part,

$$\alpha = \sup_{p \leq p_0} P[X > c'] = \sup_{p \leq p_0} \sum_{k=0}^{[c']} C_m^k p^k (1-p)^{m-k} = P_{p_0}[X > c']$$

Parce que X est une v.a. discrète, il est possible que c' n'existe pas. S'il n'existe pas, on choisit l'entier c' tel que :

$$P_{p_0}[X > c'] \leq \alpha \quad \text{et} \quad P_{p_0}[X > c' - 1] > \alpha$$

Exemple 2. Soit $X \sim \mathcal{N}(m, \sigma^2)$ avec m et σ^2 inconnus. On teste l'hypothèse $H_0 : m = m_0$ contre $H_1 : m \neq m_0$. Dans ce cas : $\Theta_0 = \{(m_0, \sigma^2); \sigma^2 > 0\}$ et $\Theta = \{(m, \sigma^2); -\infty < m < \infty, \sigma^2 > 0\}$. Notons $\theta = (m, \sigma^2)$. Alors

$$\sup_{\theta \in \Theta_0} L_n(\mathbf{x}; \theta) = \sup_{\sigma^2 > 0} \left[\frac{1}{(\sigma\sqrt{2\pi})^n} \exp \left\{ -\frac{\sum_{i=1}^n (x_i - m_0)^2}{2\sigma^2} \right\} \right] = L_n(\mathbf{x}; \hat{\sigma}_0^2)$$

avec $\hat{\sigma}_0^2$ l'estimateur du MV de σ^2 , $\hat{\sigma}_0^2 = (1/n) \sum_{i=1}^n (x_i - m_0)^2$. Alors

$$\sup_{\theta \in \Theta_0} L_n(\mathbf{x}; \theta) = \frac{1}{(2\pi/n)^{n/2} \left\{ \sum_{i=1}^n (x_i - m_0)^2 \right\}^{n/2}} e^{-n/2}$$

En tenant compte du fait que l'EMV de θ est (\bar{X}_n, S_n^2) , on a

$$\sup_{\theta \in \Theta} L_n(\mathbf{x}; \theta) = \sup_{m, \sigma^2} \left[\frac{1}{(\sigma\sqrt{2\pi})^n} \exp \left\{ -\frac{\sum_{i=1}^n (x_i - m)^2}{2\sigma^2} \right\} \right] = \frac{1}{(2\pi/n)^{n/2} \left\{ \sum_{i=1}^n (x_i - \bar{x}_n)^2 \right\}^{n/2}} e^{-n/2}$$

Alors

$$\lambda(\mathbf{x}) = \left\{ \frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2}{\sum_{i=1}^n (x_i - m_0)^2} \right\}^{n/2} = \left\{ \frac{1}{1 + n(\bar{x}_n - m_0)^2 / \sum_{i=1}^n (x_i - \bar{x}_n)^2} \right\}^{n/2}$$

Le test du rapport de vraisemblance rejette H_0 si : $\lambda(\mathbf{x}) < c$, et parce que $\lambda(\mathbf{x})$ est décroissante en $n(\bar{x}_n - m_0)^2 / \sum_{i=1}^n (x_i - \bar{x}_n)^2$, on rejette H_0 si

$$\left| \frac{\bar{x}_n - m_0}{\sqrt{\sum_{i=1}^n (x_i - m_0)^2}} \right| > c'$$

ou encore

$$\left| \frac{\sqrt{n}(\bar{x}_n - m_0)}{s_n^*} \right| > c$$

En conclusion, la statistique

$$Z(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - m_0)}{S_n^*} \sim t(n-1)$$

et $c'' = u_{n-1; 1-\alpha/2}$ sa fractile d'ordre $1 - \alpha/2$.

Si on peut trouver une statistique de test, sur la base de $\lambda(\mathbf{x})$ de loi connue sur H_0 , on peut donner la zone de rejet du test du rapport de vraisemblance, pour un α fixé. Par contre, si cette technique ne peut pas être utilisée (pas de loi connue sous H_0), il est difficile, sinon impossible, de trouver le test du rapport de vraisemblance pour un α fixé. Le résultat suivant montre que dans le cas i.i.d. on peut obtenir la loi asymptotique (sous H_0) du rapport de vraisemblance.

Supposons que Θ_0 est déterminé par

$$H_0 : \theta = g(\vartheta) \quad (4.12)$$

avec ϑ un vecteur de dimension $(k-r)$ de paramètres inconnus et g une fonction différentiable de \mathbb{R}^{k-r} à \mathbb{R}^k avec $\partial g(\vartheta)/\partial \vartheta$ de plein rang. Par exemple, si $\Theta = \mathbb{R}^2$ et $\Theta_0 = \{(\theta_1, \theta_2) \in \Theta; \theta_1 = 0\}$, alors $k = 2$, $r = 1$ $g : \mathbb{R} \rightarrow \mathbb{R}^2$, $g(\vartheta) = (g_1(\vartheta), g_2(\vartheta))$, $\vartheta = \theta_2$, $g_1(\vartheta) = 0$ et $g_2(\vartheta) = \vartheta$.

Théorème 4.1.3 *Supposons que H_0 est déterminée par (4.12). Alors, sous H_0 ,*

$$-2 \log \lambda(\mathbf{X}) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \chi^2(r)$$

En conséquence, la zone de rejet asymptotique du rapport de vraisemblance est $\lambda(\mathbf{x}) < e^{-u_{r,1-\alpha}/2}$, avec $u_{r,1-\alpha}$ la fractile d'ordre $1 - \alpha$ de la loi $\chi^2(r)$.

L'hypothèse (4.12) peut être écrite sous la forme :

$$H_0 : R(\theta) = 0 \quad (4.13)$$

avec $R(\theta)$ une fonction continue de \mathbb{R}^k à \mathbb{R}^r , $r \leq k$. Wald (1943) a introduit un test qui rejette H_0 quand la valeur de

$$W_n = [R(\hat{\theta}_n)]^t \left\{ [C(\hat{\theta}_n)]^t [I_n(\hat{\theta}_n)]^{-1} C(\hat{\theta}_n) \right\} R(\hat{\theta}_n)$$

est grande, où $C(\theta) = \partial R(\theta)/\partial \theta$, $I_n(\theta)$ est la matrice de l'information de Fisher de X_1, \dots, X_n et $\hat{\theta}_n$ est l'EMV de θ .

Pour tester $H_0 : \theta = \theta_0$, avec θ_0 connu, $R(\theta) = \theta - \theta_0$ et W_n devient :

$$W_n = (\hat{\theta}_n - \theta_0)^t I_n^{-1}(\hat{\theta}_n) (\hat{\theta}_n - \theta_0)$$

Théorème 4.1.4 (de Wald) *Sous H_0 donnée par (4.13), $W_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \chi^2(r)$*

Alors la zone de rejet de H_0 est $w_n > u_{r,1-\alpha}$ la fractile d'ordre $1 - \alpha$ de la loi $\chi^2(r)$.

4.2 Tests non-paramétriques

4.2.1 Théorème de Pearson

Considérons une variable aléatoire X discrète, avec l'espace d'état : $\Omega = \{v_1, \dots, v_K\}$ et $p_j = IP[X = v_j]$ pour $j = 1, \dots, K$. Pour un n-échantillon (X_1, \dots, X_n) , soient n_1, \dots, n_K les effectifs des valeurs v_1, \dots, v_K : $n_j = \text{Card}\{X_1, \dots, X_n = v_j\}$ pour $j = 1, \dots, K$, avec $n = n_1 + \dots + n_K$.

Théorème 4.2.1 *La variable aléatoire $\sum_{j=1}^K \frac{(n_j - np_j)^2}{np_j}$ converge en loi, pour $n \rightarrow \infty$, vers la loi $\chi^2(K-1)$.*

Preuve du Théorème 4.2.1 La variable aléatoire $Y_{i,j} = \mathbb{1}_{X_i=v_j} \sim \mathcal{B}(p_j)$. Alors, par le TCL :

$$\frac{n_j - np_j}{\sqrt{np_j(1-p_j)}} = \frac{\sum_{i=1}^n Y_{i,j} - n\mathbb{E}[Y_{i,j}]}{\sqrt{n\text{Var}(Y_{i,j})}} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1)$$

Donc

$$\frac{n_j - np_j}{\sqrt{np_j}} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1 - p_j)$$

Par contre $\frac{n_j - np_j}{\sqrt{np_j}}$ et $\frac{n_k - np_k}{\sqrt{np_k}}$ ne sont pas indépendantes pour $j \neq k$. Leur covariance est :

$$\begin{aligned} \text{Cov}\left(\frac{n_j - np_j}{\sqrt{np_j}}, \frac{n_k - np_k}{\sqrt{np_k}}\right) &= \mathbb{E}\left[\frac{n_j - np_j}{\sqrt{np_j}} \cdot \frac{n_k - np_k}{\sqrt{np_k}}\right] \\ &= \frac{1}{n\sqrt{p_j p_k}} (\mathbb{E}[n_j n_k] - np_k \mathbb{E}[n_j] - np_j \mathbb{E}[n_k] + n^2 p_j p_k) = \frac{1}{n\sqrt{p_j p_k}} (\mathbb{E}[n_j n_k] - n^2 p_j p_k) \end{aligned}$$

D'autre part :

$$\begin{aligned} E[n_j n_k] &= E \left[\left(\sum_{l=1}^n \mathbb{1}_{X_l=v_j} \right) \left(\sum_{l'=1}^n \mathbb{1}_{X_{l'}=v_k} \right) \right] = E \left[\sum_{l,l'=1}^n (\mathbb{1}_{X_l=v_j}) (\mathbb{1}_{X_{l'}=v_k}) \right] \\ &= E \left[\sum_{l=l'}^n (\mathbb{1}_{X_l=v_j}) (\mathbb{1}_{X_{l'}=v_k}) + \sum_{l \neq l'}^n (\mathbb{1}_{X_l=v_j}) (\mathbb{1}_{X_{l'}=v_k}) \right] = 0 + E \left[\sum_{l \neq l'}^n (\mathbb{1}_{X_l=v_j}) (\mathbb{1}_{X_{l'}=v_k}) \right] \\ &= n(n-1) E [\mathbb{1}_{X_1=v_j} \mathbb{1}_{X_2=v_k}] = n(n-1) p_j p_k \end{aligned}$$

Alors la covariance calculée plus haut est :

$$\frac{1}{n(n-1)\sqrt{p_j p_k}} [n p_j p_k - n^2 p_j p_k] = -\sqrt{p_j p_k}$$

On a montré jusqu'ici que :

$$\sum_{j=1}^K \frac{(n_j - n p_j)^2}{n p_j} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \sum_{j=1}^K Z_j^2$$

avec les variables aléatoires $Z_j \sim \mathcal{N}(0, 1 - p_j)$ et $E[Z_j^2] = 1 - p_j$, $Cov[Z_j, Z_k] = -\sqrt{p_j p_k}$. On applique ensuite la même technique que pour le Théorème de Cochran (Proposition 3.1.2., voir preuve en TD). ■

4.2.2 Test de χ^2 d'ajustement

Supposons que la variable aléatoire X discrète possède K modalités : v_1, \dots, v_K . Notons par $p_j = P[X = v_j]$ et $p = (p_1, \dots, p_K)$.

Considérons connu le vecteur de probabilités $p^0 = (p_1^0, \dots, p_K^0)$.

On veut tester l'hypothèse : $H_0 : p = p^0$ contre $H_1 : p \neq p^0$.

Pour cela on considère un n échantillon pour la v.a. $X : (X_1, \dots, X_n)$ et (x_1, \dots, x_n) une réalisation. Soient n_1, \dots, n_K les effectifs de chaque valeur possible de X . Les fréquences empiriques sont : $f_k = n_k/n$, pour $k = 1, \dots, K$ et $\hat{p} = (f_1, \dots, f_K)$.

Définition. La "distance" de χ^2 entre les vecteurs de probabilités $p = (p_1, \dots, p_K)$ et $q = (q_1, \dots, q_K)$ est

$$D(p, q) = \sum_{k=1}^K \frac{(p_k - q_k)^2}{q_k}$$

(elle n'est pas une vraie distance, elle n'est pas symétrique).

Considérons la distance $D(\hat{p}, p^0) = \sum_{k=1}^K \frac{(f_k - p_k^0)^2}{p_k^0}$.

Théorème 4.2.2 Si $p_k^0 \neq 0 \forall k = 1, \dots, K$ alors pour $n \rightarrow \infty$:

- sous H_0 , $nD(\hat{p}, p^0) \rightarrow \chi^2(K-1)$ en loi ;

- sous H_1 , $nD(\hat{p}, p^0) \rightarrow \infty$ en probabilité.

Preuve du Théorème 4.2.2

$$\frac{n_k - n p_k^0}{\sqrt{n p_k^0}} = \frac{n_k - n p_k}{\sqrt{n p_k^0}} + \sqrt{n} \frac{p_k - p_k^0}{\sqrt{p_k^0}}$$

Si H_0 est vraie, alors on applique le Théorème de Pearson 4.2.1. Si H_1 est vraie le deuxième terme en valeur absolue converge vers ∞ . ■

Ce théorème permet de construire un *test asymptotique* de l'hypothèse H_0 contre H_1 .

4.2.3 Test de χ^2 d'indépendance

Supposons qu'on a deux variables aléatoires X et Y discrètes ; X possède p modalités : v_1, \dots, v_p et Y possède q modalités : w_1, \dots, w_q .

On veut tester l'hypothèse selon laquelle X et Y sont indépendantes.

$H_0 : X$ et Y indépendantes, contre $H_1 : X$ et Y ne sont pas indépendantes

ou encore :

$H_0 : P[X = v_i, Y = w_j] = P[X = v_i]P[Y = w_j], \forall i = 1, \dots, p, j = 1, \dots, q$

contre

$H_1 : \exists i \in \{1, \dots, p\}, \exists j \in \{1, \dots, q\} \text{ t. q. } P[X = v_i, Y = w_j] \neq P[X = v_i]P[Y = w_j]$

Considérons un échantillon pour X et pour Y . Soient les effectifs :

$n_{ij} = \text{Card}\{x = v_i, y = w_j\}, \quad \forall i = 1, \dots, p, j = 1, \dots, q$
et

$$n = \sum_{i=1}^p \sum_{j=1}^q n_{ij}, \quad n_{.j} = \sum_{i=1}^p n_{ij}, \quad n_{i.} = \sum_{j=1}^q n_{ij}$$

Un estimateur pour $P[X = v_i, Y = w_j]$ est $f_{ij} = n_{ij}/n$, pour $P[X = v_i]$ est $f_{i.} = n_{i.}/n$, pour $P[Y = w_j]$ est $f_{.j} = n_{.j}/n$. On considère la distance de χ^2 entre f_{ij} et $f_{i.}f_{.j}$

$$D = \sum_{i=1}^p \sum_{j=1}^q \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}}$$

$$\left(\frac{f_{11} - f_{1.}f_{.1}}{f_{1.}f_{.1}} \right)^2$$

Théorème 4.2.3 Sous $H_0 : D_n = nD \xrightarrow{\mathcal{L}} \chi^2((p-1)(q-1))$, pour $n \rightarrow \infty$. Sous $H_1 : D_n \xrightarrow{P} \infty$.

Conséquence. La région de rejet de H_0 est : $R = \{d_n > u_{1-\alpha; (p-1)(q-1)}\}$, avec $u_{1-\alpha; (p-1)(q-1)}$ la fractile d'ordre $1 - \alpha$ de la loi $\chi^2((p-1)(q-1))$.

$$\left(\frac{f_{12} - f_{1.}f_{.2}}{f_{1.}f_{.2}} \right)^2$$

4.2.4 Test de Kolmogorov-Smirnov

Soit X une v.a. de fonction de répartition F . On veut tester :

$H_0 : F(x) = F_0(x), \forall x \in \mathbb{R}$, avec $F_0(x)$ une fonction de répartition connue.

Soit la fonction de répartition empirique $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq x}$. On considère la variable aléatoire :

$$K_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|$$

On sait par le théorème de Glivenko-Cantelli que $K_n \xrightarrow{p.s.} 0$, pour $n \rightarrow \infty$. On peut montrer :

Théorème 4.2.4 Sous H_0 , $\sqrt{n}K_n \xrightarrow{\mathcal{L}} K$, avec K une variable aléatoire de loi fixe indépendante de F , définie par :

$$P(K > k) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2j^2 k^2)$$

Les fractiles de cette loi sont tabulées. La région critique de ce test est : $R = \{\sqrt{n}K_n > k_{1-\alpha}\}$ avec $k_{1-\alpha}$ la fractile d'ordre $(1 - \alpha)$ de la loi de $\sqrt{n}K_n$.

4.2.5 Test de Smirnov, de comparaison de deux échantillons indépendantes

Soient X et Y deux variables aléatoires de fonctions de répartition, respectivement, F et G . On veut tester l'hypothèse $H_0 : F(x) = G(x)$ contre $F(x) \neq G(x)$. On dispose de deux échantillons X_1, \dots, X_{n_1} et Y_1, \dots, Y_{n_2} . Soit $\hat{F}_{n_1}(x) = \sum_{i=1}^{n_1} \mathbb{1}_{X_i \leq x}$ et $\hat{G}_{n_2}(x) = \sum_{i=1}^{n_2} \mathbb{1}_{Y_i \leq x}$ les deux fonctions de répartition empiriques des deux échantillons. Alors :

$$\mathbb{P} \left[\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \sup_x |\hat{F}_{n_1}(x) - \hat{G}_{n_2}(x)| < y \right] \rightarrow K(y)$$

avec K la même variable aléatoire que pour le test de Kolmogorov-Smirnov.

4.2.6 Test de la médiane sur des groupes indépendants

(voir cours FG Carpentier)

Soit X_1, \dots, X_{n_1} et Y_1, \dots, Y_{n_2} deux échantillons et $N = n_1 + n_2$.

Hypothèses :

H_0 : Les deux populations parentes ont la même médiane.

H_1 : Les deux populations parentes ont des médianes différentes

Construction de la statistique de test : on détermine la médiane M de la série obtenue en réunissant les deux échantillons. On constitue un tableau de contingence en croisant la variable indépendante et la variable dérivée "position par rapport à M ".

	Gr 1	Gr 2	Ensemble
$\leq M$	N1	N2	N1 + N2
$> M$	N3	N4	N3 + N4
Total	N1 + N3	N2 + N4	N

On fait un test du χ^2 sur le tableau obtenu.

Exemple 31 basketteurs de 14 ans, répartis en deux groupes d'effectifs $n_1 = 12$ et $n_2 = 19$, selon le jugement porté par l'entraîneur (groupe G1 : jugement négatif ; groupe G2 : jugement positif). On a relevé la taille de chaque sujet.

G1 : 152 163 164 173 174 176 177 177 178 178 181 184

G2 : 167 171 172 174 175 176 176 177 179 179 180 182 183 186 188 189 189 193 195

Les deux groupes sont-ils significativement différents du point de vue de la taille ?

Détermination de la médiane :

152 163 164 167 171 172 173 174 174 175 176 176 176 177 177 177 178 178 179 179 180 181 182 183 184 186 188 189 189 193 195 On obtient : Médiane = 177. Tableau de contingence :

	Gr 1	Gr 2	Ensemble
$\leq M$	8	8	16
$> M$	4	11	15
Total	12	19	31

Ici : $D_n = 1.76$. Pour un seuil de 0.05, la fractile = 3.84. On retient H_0 .

4.2.7 Test de Spearman

a) Pour une seule variable

Pour une variable aléatoire X considérons n copies : X_1, X_2, \dots, X_n . Chaque copie X_i a la même loi que X . Avant d'appliquer des techniques statistiques de modélisation, on s'interroge sur l'hypothèse selon laquelle l'ordre dans lequel on effectue les observations n'a pas d'importance, c'est à dire que ces variables sont indépendantes. C'est pourquoi les statistiques d'ordre et de rang des observations jouent un très grand rôle. On testera comme hypothèse nulle :

H_0 : X_1, X_2, \dots, X_n sont indépendantes

On ordonne l'échantillon X_1, X_2, \dots, X_n en ordre croissant, et on note le nouveau échantillon par $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$. Pour une réalisation x_1, x_2, \dots, x_n la réalisation correspondante du échantillon ordonné est $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. A chaque observation X_i on associe son rang R_i dans l'échantillon ordonné.

Exemple $x_1 = 3, x_2 = 1, x_3 = 0, x_4 = 5$ alors $x_{(1)} = 0, x_{(2)} = 1, x_{(3)} = 3, x_{(4)} = 5$. $R_1 = 3, R_2 = 2, R_3 = 1, R_4 = 4$.

Remarque : Si les observations sont distinctes les rangs sont des nombres entiers compris entre 1 et n . Dans le cas des valeurs identiques, on leur assigne un rang égal à la moyenne arithmétique des rangs.

Si l'hypothèse d'indépendance H_0 est vraie alors il n'y a aucune corrélation entre $1, 2, \dots, n$ et R_1, R_2, \dots, R_n .

On construit alors un test basé sur le coefficient de corrélation (de Pearson) entre ces deux ensembles. On obtient ce qui s'appelle *coefficient de Spearman* :

$$r_S = \frac{\sum_{i=1}^n (R_i - \bar{R})(i - \bar{i})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2 \sum_{i=1}^n (i - \bar{i})^2}}$$

où $\bar{R} = \frac{1}{n} \sum_{i=1}^n R_i, \bar{i} = \frac{1}{n} \sum_{i=1}^n i$. Par des calculs élémentaires on peut montrer que r_S s'écrit sous la forme :

$$r_S = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n (R_i - i)^2$$

Dans le cas d'une tendance monotone croissante, $R_i = i$ et $r_S = 1$. Dans le cas d'une tendance monotone décroissante, les classements sont inversés : $R_i = n - i + 1$ et $r_S = -1$. La zone de rejet de l'hypothèse H_0 est :

$$R = \{|r_S| > c\} \quad (4.14)$$

où

- $c = \frac{1}{\sqrt{n-1}} u_{1-\alpha/2}$, avec $u_{1-\alpha/2}$ la fractile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$, pour $n > 30$.

- $c = \frac{t}{\sqrt{n-2+t^2}}$ avec t la fractile d'ordre $1 - \alpha/2$ de la loi Student $t(n-2)$, pour $11 \leq n \leq 30$.

a) Pour deux variables aléatoires

Considérons pour deux variables aléatoires X et Y les échantillons X_1, \dots, X_n , respectivement Y_1, \dots, Y_n . A partir de

ces deux échantillons on veut tester que les deux variables sont indépendantes :

H_0 : X et Y sont indépendantes

On associe aux couples $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ les rangs $(R_1, Q_1), (R_2, Q_2), \dots, (R_n, Q_n)$, R_i et Q_i étant les rangs respectifs de X_i et Y_i dans chacun des deux échantillons. Pour tester l'hypothèse H_0 on utilise la statistique de test :

$$S = \frac{\sum_{i=1}^n (R_i - \bar{R})(Q_i - \bar{Q})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2 \sum_{i=1}^n (Q_i - \bar{Q})^2}}$$

La zone de rejet est la même que dans (4.14).

Remarque. Le test de χ^2 est utilisé surtout pour des variables aléatoires discrètes pendant que le test de Spearman est utilisé pour des lois continues.

4.2.8 Test de Wilcoxon

Considérons pour deux variables aléatoires X et Y les échantillons X_1, \dots, X_n , respectivement Y_1, \dots, Y_m . on considère le cas $n \leq m$. A partir de ces deux échantillons on veut tester que X et Y sont de même loi :

H_0 : X et Y ont la même loi de probabilité

Ce test repose sur l'idée que si l'on mélange les deux séries et qu'on ordonne le tout par valeurs croissantes on doit obtenir un mélange homogène. Pour cela on réordonne les deux suites et on compte le nombre total de couples (X_i, Y_i) où X_i a un rang plus grand que Y_i . Pour tester l'hypothèse H_0 on utilise la statistique de test :

$$W = \sum_{i=1}^n R_i$$

où R_i est le rang de X_i dans l'échantillon global $(X_1, \dots, X_n, Y_1, \dots, Y_m)$ ordonnée de taille $N = m + n$. La zone de rejet est

$$R = \left\{ \left| W - \frac{n(n+m+1)}{2} \right| > u_{1-\alpha/2} \sqrt{\frac{nm(n+m+1)}{12}} \right\}$$

avec $u_{1-\alpha/2}$ la fractile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$, pour $n > 30$

Exemple livre SAPORTA, Page 345

On veut comparer les performances de deux groupes d'élèves à un test d'habileté manuelle. On choisit aléatoirement 8 élèves du premier groupe et 10 du deuxième. Les performances en minutes sont les suivantes :

Groupe 1 : 22 31 14 19 24 28 27 28

Groupe 2 : 25 13 20 11 23 16 21 18 17 26

On réordonne les 18 observations par ordre croissant. Les résultats du premier groupe sont en gras :

Observations : 11 13 14 16 17 18 **19** 20 21 **22** 23 24 25 26 **27 28 28 31**

La somme des rangs des élèves du premier groupe est :

$W = 3 + 7 + 10 + 12 + 15 + 16 + 17 + 18 = 98$

Comme $\frac{98 - 76}{11.25} = 1.96$ on rejette H_0 avec $\alpha = 0.10$.

Adresse internet citée

<http://geai.univ-brest.fr/~carpentier/tdm-index.html>

Chapitre 5

REGRESSION LINEAIRE

5.1 Généralités sur le Modèle Linéaire

Donnons d'abord la forme générale d'un modèle statistique. Soient Y, X_1, \dots, X_p des variables. Dans des nombreux problèmes pratiques on étudie la relation qui peut exister entre Y et X_1, \dots, X_p : $Y = f(X_1, \dots, X_p)$. Mais, assez souvent on met en doute le caractère purement déterministe de cette relation

- soit parce qu'il a des erreurs de mesure
- soit à cause de l'omission volontaire ou non d'éventuelles variables (ce qui est le plus fréquent)

On ajoute un terme d'erreur et on obtient le modèle statistique

$$Y = f(X_1, \dots, X_p) + \varepsilon \quad (5.1)$$

Y est variable expliquée, dépendante, X_1, \dots, X_p variables explicatives, indépendantes.

Définition. Le modèle (5.1) est dit de *régression linéaire* si la fonction f est fonction linéaire de X_1, \dots, X_p

$$f(X_1, \dots, X_p) = a_0 + a_1 X_1 + \dots + a_p X_p \quad (5.2)$$

En ce qui concerne les variables et les paramètres

	Aléatoire	Non aléatoire
Observable	Y	X_1, \dots, X_p
Non observable	ε	a_0, a_1, \dots, a_p

a_0, \dots, a_p paramètres inconnus à estimer. Pour estimer ces paramètres on dispose de n observations des variables Y, X_1, \dots, X_p , notées

Variable	Y	X_1	X_2	\dots	X_p
Observation i	y_i	x_{1i}	x_{2i}	\dots	x_{pi}

Alors, le modèle de régression linéaire peut être écrit

$$Y_i = a_0 + a_1 X_{1i} + \dots + a_p X_{pi} + \varepsilon_i \quad i = 1, \dots, n \quad (5.3)$$

où :

Y_i est une v.a. avec la réalisation y_i

X_{1i} une var (non aléatoire) avec l'observation x_{1i} .

L'étude statistique du modèle linéaire permet

- estimer les paramètres a_0, \dots, a_p par moindres carrés et par intervalle;
- tester l'influence de certaines variables X_j (par test d'hypothèse)
- en déduire le meilleur modèle (par l'étude des résidus)

Notons que les erreurs $\varepsilon_1, \dots, \varepsilon_n$ sont v.a. indépendantes, donc Y_1, \dots, Y_n aussi.

On suppose que les v.a. ε_i suivent une loi Normale : $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \dots, n$.

Le cas le plus simple de régression linéaire est pour $p = 1$: $Y_i = a_0 + a_1 X_{1i} + \varepsilon_i$, $i = 1, \dots, n$, modèle appelé *régression linéaire simple*.

Exemple de régression simple

Pour une ville on mesure la pollution en ozone et la vitesse maximale du vent (m/s) pendant 10 jours. Ecrire un modèle statistique de la pollution fonction de vent :

Y - la concentration de l'ozone (en mg/m^3)

X - vitesse (en m/s)

Obs	1	2	3	4	5	6	7	8	9	10
Y	174	188	176	128	116	88	58	120	92	132
X	1	0.5	1	2	2	2.5	3	2	3	2
\hat{y}_i	171	195	171	122	122	98	74	122	74	122
e_i	3	-7	5	6	-6	-10	-16	-2	18	10
er_i	0.28	-0.65	0.46	0.55	-0.55	-0.92	-1.47	-0.18	1.66	0.92

$$\hat{\sigma} = 10.84$$

- on peut estimer a_0 et a_1
- on peut tester si vraiment il y a un lien linéaire entre la pollution d'ozone et le vent (c'est-à-dire que le modèle linéaire est bon)
- pour un nouveau jour pour lit la vitesse maximale du vent, on peut prévoir la concentration d'ozone (Si par exemple, on prévoit la vitesse du vent par une autre méthode la veille, on peut prévoir pour le lendemain la pollution).

5.2 Régression linéaire simple

5.2.1 Description des données du modèle

La variable à expliquer est Y et la variable indépendante est X . Le modèle statistique est

$$Y = aX + b + \varepsilon \quad (5.4)$$

Pour estimer les paramètres a et b nous disposons de n couples d'observations

Var Obs	1	2	...	i	...	n
Y	y_1	y_2	...	y_i	...	y_n
X	x_1	x_2	...	x_i	...	x_n

Alors, le modèle (5.4) peut être écrit

$$Y_i = aX_i + b + \varepsilon_i \quad i = 1, \dots, n \quad (5.5)$$

On suppose en ce qui concerne les v.a. $\varepsilon_i : \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, avec σ^2 inconnu, pour $i \neq j$, ε_i et ε_j indépendantes, donc $Cov(\varepsilon_i, \varepsilon_j) = 0$

Proposition 5.2.1 1) $\mathbb{E}(Y_i) = aX_i + b$

$$2) Cov(Y_i, Y_j) = \begin{cases} \sigma^2 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}$$

$$3) Y_i \sim \mathcal{N}(aX_i + b, \sigma^2)$$

Preuve

$$1) \mathbb{E}(Y_i) = \mathbb{E}(aX_i + b + \varepsilon_i) = \mathbb{E}(aX_i + b) + \mathbb{E}(\varepsilon_i) = aX_i + b$$

$$2) Cov(Y_i, Y_j) = \mathbb{E}[(Y_i - \mathbb{E}(Y_i))(Y_j - \mathbb{E}(Y_j))] = \mathbb{E}[(Y_i - aX_i - b)(Y_j - aX_j - b)] = \mathbb{E}(\varepsilon_i \varepsilon_j) = \begin{cases} \sigma^2 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}$$

Proposition 5.2.2 $\mathbb{E}(\bar{Y}_n) = a\bar{X}_n + b$ et $Var(\bar{Y}_n) = \sigma^2/n$

Preuve $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$.

$$\mathbb{E}(\bar{Y}_n) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(aX_i + b + \varepsilon_i) = \frac{1}{n} \sum_{i=1}^n (aX_i + b) = \frac{a}{n} \sum_{i=1}^n X_i + b = a\bar{X}_n + b$$

$$\text{Var}(\bar{Y}_n) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n Y_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(Y_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}$$

5.2.2 Estimation des paramètres du modèle

La construction des estimateurs A et B des paramètres réels a et b est basée sur la méthode des moindres carrés.

Définition. Les *estimateurs des moindres carrés* des a et b sont les v.a. A_n et B_n qui minimisent la somme des carrés des termes erreur

$$S(A, B) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n [Y_i - (AX_i + B)]^2$$

Donc, A_n et B_n sont les solutions du système

$$\begin{cases} \frac{\partial S}{\partial A}(A, B) = 0 \\ \frac{\partial S}{\partial B}(A, B) = 0 \end{cases}$$

Résultat.

$$\begin{cases} A_n = \frac{\sum_{i=1}^n (Y_i - \bar{Y}_n)(X_i - \bar{X}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \\ B_n = \bar{Y}_n - A_n \bar{X}_n \end{cases} \quad (5.6)$$

Propriétés

Proposition 5.2.3 Les v.a. A_n et \bar{Y}_n ne sont pas corrélées :

$$\text{Corr}(A_n, \bar{Y}_n) = 0$$

Preuve $\text{Corr}(A_n, \bar{Y}_n) = \frac{\text{Cov}(A_n, \bar{Y}_n)}{\sqrt{\text{Var}(A_n) \cdot \text{Var}(\bar{Y}_n)}}$

Donc il suffit de montrer que $\text{Cov}(A_n, \bar{Y}_n) = 0$.

$$\begin{aligned} A_n &= \frac{\sum_{i=1}^n (Y_i - \bar{Y}_n)(X_i - \bar{X}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} = \frac{\sum_{i=1}^n Y_i(X_i - \bar{X}_n) - \bar{Y} \cdot \sum_{i=1}^n (X_i - \bar{X}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \\ &= \frac{\sum_{i=1}^n Y_i(X_i - \bar{X}_n) - \bar{Y} [n\bar{X}_n - \sum_{i=1}^n X_i]}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} = \frac{\sum_{i=1}^n Y_i(X_i - \bar{X}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \\ &= \sum_{i=1}^n Y_i \left[\frac{X_i - \bar{X}_n}{\sum_{j=1}^n (X_j - \bar{X}_n)^2} \right] = \sum_{i=1}^n c_i Y_i \end{aligned}$$

où $c_i = \frac{X_i - \bar{X}_n}{\sum_{j=1}^n (X_j - \bar{X}_n)^2}$. Propriété pour c_i

$$\sum_{i=1}^n c_i = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)}{\sum_{j=1}^n (X_j - \bar{X}_n)^2} = \frac{\sum_{i=1}^n X_i - n\bar{X}_n}{\sum_{j=1}^n (X_j - \bar{X}_n)^2} = 0$$

Alors

$$\begin{aligned} \text{Cov}(A_n, \bar{Y}_n) &= \text{Cov}\left(\sum_{i=1}^n c_i Y_i, \frac{1}{n} \sum_{j=1}^n Y_j\right) = \sum_{i=1}^n c_i \text{Cov}\left(Y_i, \frac{1}{n} \sum_{j=1}^n Y_j\right) \\ &= \sum_{i=1}^n \frac{c_i}{n} \cdot \sum_{j=1}^n \text{Cov}(Y_i, Y_j) = \sum_{i=1}^n \frac{c_i}{n} \sigma^2 = \frac{\sigma^2}{n} \sum_{i=1}^n c_i = 0 \end{aligned}$$

Proposition 5.2.4 Les v.a. A_n et B_n sont des estimateurs sans biais pour les paramètres a et b .

Preuve

$$\begin{aligned}\mathbb{E}(A_n) &= \mathbb{E} \left[\frac{\sum_{i=1}^n (Y_i - \bar{Y}_n)(X_i - \bar{X}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \right] = \frac{\sum_{i=1}^n (X_i - \bar{X}_n) \mathbb{E}(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(aX_i + b - a\bar{X}_n - b)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} = a \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} = a\end{aligned}$$

$$\mathbb{E}(B_n) = \mathbb{E}(\bar{Y}_n - A_n \bar{X}_n) = \mathbb{E}(\bar{Y}_n) - \mathbb{E}(A_n \bar{X}_n) = a\bar{X}_n + b - \mathbb{E}(A_n)\bar{X}_n = a\bar{X}_n + b - a\bar{X}_n = b$$

En ce qui concerne les variances et les covariances de ces estimateurs on a la proposition suivante :

Proposition 5.2.5

$$\begin{aligned}\text{Var}(A_n) &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}, \quad \text{Var}(B_n) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}_n^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \right] \\ \text{Cov}(A_n, B_n) &= -\frac{\sigma^2 \bar{X}_n}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}\end{aligned}$$

Preuve

$$\begin{aligned}\text{Var}(A_n) &= \text{Var} \left(\sum_{i=1}^n c_i Y_i \right) = \sum_{i=1}^n c_i^2 \text{Var}(Y_i) = \sum_{i=1}^n c_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^n c_i^2 \\ &= \sigma^2 \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\left[\sum_{i=1}^n (X_i - \bar{X}_n)^2 \right]^2} = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \\ \text{Var}(B_n) &= \text{Var}(\bar{Y}_n - A_n \bar{X}_n) = \text{Var}(\bar{Y}_n) + \bar{X}_n^2 \text{Var}(A_n) \\ &= \frac{\sigma^2}{n} + \bar{X}_n^2 \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}_n^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \right] \\ \text{Cov}(A_n, B_n) &= \text{Cov}(A_n, -A_n \bar{X}_n + \bar{Y}_n) = -\text{Cov}(A_n, A_n \bar{X}_n) + \text{Cov}(A_n, \bar{Y}_n) \\ &= -\bar{X}_n \text{Var}(A_n) = -\frac{\sigma^2 \bar{X}_n}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}\end{aligned}$$

En ce qui concerne un estimateur pour la variance σ^2 on a la proposition suivante

Proposition 5.2.6

$$S_n^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - A_n X_i - B_n)^2$$

est un estimateur sans biais pour σ^2 .

Une estimation pour σ :

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{a}_n x_i - \hat{b}_n)^2$$

Exemple. les estimations de a et b sur les données mesurées sont

$$\hat{a}_n = -48.5 \quad \hat{b}_n = 219.5 \quad s_n^2 = 117.5 \quad \hat{\sigma}_n = 10.84$$

$$\text{Var}(\hat{A}_n) = 18.4 \quad \text{Var}(\hat{B}_n) = 78.05 \quad \text{Cov}(\hat{A}_n, \hat{B}_n) =$$

Donc, on peut dire que la pollution d'ozone est liée à la vitesse du vent par la relation linéaire : $Y = -48.5X + 219.5$.

Les lois des estimateurs

On a montré que : $A_n = \sum_{i=1}^n c_i Y_i$, $Y_i \sim \mathcal{N}(aX_i + b, \sigma^2)$, Y_i, Y_j indépendantes pour $i \neq j$ et $\mathbb{E}(A_n) = a$, $\text{Var}(A_n) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}$, d'où

$$A_n \sim \mathcal{N} \left(a, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \right)$$

$B_n = \bar{Y}_n - A_n \bar{X}_n$, \bar{Y}_n et A_n non corrélés, $\mathbb{E}(B_n) = b$, $\text{Var}(B_n) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}_n^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \right]$ $A_n \sim \mathcal{N}$, $\bar{Y}_n \sim \mathcal{N}$, d'où

$$B_n \sim \mathcal{N} \left(b, \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}_n^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \right] \right)$$

Proposition 5.2.7 (sans dém)

- 1) $\frac{\sum_{i=1}^n (Y_i - A_n X_i - B_n)^2}{\sigma^2} = \frac{(n-2)S^2}{\sigma^2} \sim \chi^2(n-2)$
- 2) Les estimateurs A_n et B_n sont indépendantes de S_n^2 .

5.2.3 Mesure de l'ajustement

On dispose de la forme générale des estimateurs. Pour un ensemble de n -couples (x_i, y_i) mesurées, on peut donner une estimation $\hat{a}_n, \hat{b}_n, \hat{\sigma}_n$ (on peut donner une valeur effective) pour a, b, σ . Ainsi la droite de régression la plus proche du nuage de points (x_i, y_i) est définie par l'équation : $y = \hat{a}_n x + \hat{b}_n$; l'estimation de l'observation y_i par le modèle étant :

$$\hat{y}_i = \hat{a}_n x_i + \hat{b}_n \quad (5.7)$$

qui est une réalisation de la v.a. (estimateur) : $\hat{Y}_i = A_n X_i + B_n$.

La différence $e_i = y_i - \hat{y}_i$ s'appelle *résidu*; et en divisant e_i par $\hat{\sigma}_n : \frac{e_i}{\hat{\sigma}_n}$ on a le *résidu réduit*.

Remarque. Il faut faire la différence entre l'erreur $\varepsilon_i = Y_i - aX_i - b$ avec a, b les vraies valeurs mais inconnues (donc ε_i inconnue) et le résidu : $e_i = y_i - \hat{a}_n x_i - \hat{b}_n$ (avec y_i, x_i mesurées) une réalisation de la v.a. ε_i .

Il est souhaitable de donner un indicateur sur la qualité de l'ajustement du modèle $Y_i = aX_i + b + \varepsilon_i$ fournie par l'équation (5.7). Seulement les valeurs des résidus sont insuffisantes :

- d'abord ces différences dépendent de l'unité de mesure;
- elles ne donnent pas une indication sur l'ajustement global.

L'indice le plus couramment employé est le coefficient suivant

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2}{\sum_{i=1}^n (y_i - \bar{y}_n)^2} \in [0, 1]$$

connu sous le nom de *coefficient de détermination*.

Interprétation. Si la valeur de R^2 est proche de 1 on dit que la variable X explique bien la variable Y . Inverse, si R^2 est proche de 0, X n'explique pas bien Y et le modèle de régression linéaire simple considéré n'est pas bon. On va voir plus loin d'où ca vient cette interprétation.

Exemple. $R^2 = 0.93$.

5.2.4 Décomposition de la variabilité de Y

Soit la décomposition (classique) : $y_i - \bar{y}_n = y_i - \hat{y}_i + \hat{y}_i - \bar{y}_n$. Alors

$$\sum_{i=1}^n (y_i - \bar{y}_n)^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}_n)$$

On montre que : $\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}_n) = 0$.

D'où l'équation de la décomposition de la dispersion de Y

$$\sum_{i=1}^n (y_i - \bar{y}_n)^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2$$

Dispersion totale de Y = dispersion due au modèle + dispersion résiduelle

$$ST = SM + SR$$

Remarque. ST ne dépend pas du modèle mais des données mesurées et elles s'appelle totale parce qu'elle donne la mesure de variation des données mesurées par rapport à leur moyenne.

La régression est résumée dans le tableau ci dessous (tableau d'analyse de variance)

Source de variation	Somme des carrés des écarts	Degrés de liberté	Carré moyen
Régression	$SM = \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2$	$1 (= 2-1)$	$SM/1$
Résiduelle	$SR = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n-2$	$SR/(n-2)$
Totale	$ST = \sum_{i=1}^n (y_i - \bar{y}_n)^2$	$n-1$	

En fait, SM donne une mesure de la variabilité (de l'écart) des estimations \hat{y}_i faites par le modèle par rapport à la moyenne \bar{y}_n des données. SR donne une mesure de la variabilité (de l'écart) entre les estimations \hat{y}_i et les vraies valeurs y_i .

Remarque. Le coefficient R^2 est en fait le rapport $R^2 = \frac{SM}{ST} = \frac{ST - SR}{ST} = 1 - \frac{SR}{ST}$.

Maintenant on voit mieux d'où ça vient l'interprétation de R^2 : si R^2 est proche de 1 alors $SR \sim 0$ en fait la différence entre les valeurs mesurées y_i et celles prédites \hat{y}_i est relativement petite. On divise par ST en fait pour avoir un indicateur qui ne tient pas compte de l'unité de mesure.

Remarque.

$$s_n^2 = \hat{\sigma}_n^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{a}_n x_i - \hat{b}_n)^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{SR}{n-2}$$

Exemple. $R^2 = 0.93$. Tableau d'analyse de variance :

Source	S.C.	ddl	C.M.
Modèle	15093	1	15093
Résidu	940	8	117.5
Total	16033	9	

5.2.5 Évaluation de l'ajustement

- Jusqu'à présent on a vu que R^2 nous donne une information sur la qualité de l'ajustement. Mais seulement cette quantité est insuffisante pour l'évaluation du modèle.

On a vu aussi qu'une autre manière simple de détecter les défaillances du modèle consiste à calculer les résidus $e_i = y_i - \hat{y}_i$ et les résidus réduits : $er_i = \frac{e_i}{\hat{\sigma}}$. Puisque les e_i sont des réalisations de la v.a. $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, les er_i sont des réalisations d'une v.a. $\mathcal{N}(0, 1)$.

- un graphique de ces résidus révèle les gros écarts du modèle; une étude systématique des résidus est un élément essentiel de toute analyse de régression.

Si le modèle est correct, les résidus réduits doivent se trouver approximativement entre -2 et 2. Ils ne doivent présenter aucune structure particulière. Si jamais il en présente une, c'est qu'une structure cachée existe dans les données.

5.2.6 Tests sur les paramètres

On va faire des tests sur les paramètres du modèle. On pourrait tester :

- 1) L'hypothèse de lien linéaire effectif entre X_1, \dots, X_n et les variables aléatoires : Y_1, \dots, Y_n . En terme de paramètres, ça signifie qu'on testera l'hypothèse

$$H_0 : a = 0 \quad \text{contre} \quad H_1 : a \neq 0$$

équivalent avec : $H_0 : Y_i = b + \varepsilon_i$, contre $H_1 : Y_i = aX_i + b + \varepsilon_i$.

- 2) L'hypothèse d'un modèle linéaire spécifié : on testera :

$$H_0 : a = a_0 \quad \text{et} \quad b = b_0 \quad \Longleftrightarrow \quad Y_i = a_0 X_i + b_0 + \varepsilon_i$$

contre : $H_1 : a \neq a_0$, ou $b \neq b_0 \quad \Longleftrightarrow \quad Y_i = aX_i + b + \varepsilon_i$.

Test du caractère significatif du modèle

L'hypothèse H_0 à tester est l'hypothèse qu'il n'y a pas de lien linéaire entre X et Y : $H_0 : a = 0$ contre $H_1 : a \neq 0$. En ce qui concerne la statistique utilisée pour tester H_0 , on peut en utiliser deux, qui vont suivre une loi de Student ou une loi de Fisher.

Première méthode : on utilise une v.a. de Student. On sait que

$$Z = \frac{(A_n - a) \sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2}}{S_n} \sim t(n-2)$$

Sous l'hypothèse H_0 cette variable aléatoire devient

$$Z = \frac{A_n \sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2}}{S_n} \sim t(n-2)$$

On calcule la valeur z de la v.a. Z sur les données $(x_i, y_i)_{1 \leq i \leq n}$.

$$z = \frac{\hat{a}_n \sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}{\hat{\sigma}_n} = \frac{\hat{a}_n}{\hat{Var}(A_n)}$$

Deuxième méthode : on utilise une v.a. de Fisher

On sait que

$$\frac{(A_n - a)^2 \sum_{i=1}^n (X_i - \bar{X}_n)^2}{S_n^2} \sim F(1, n-2)$$

Alors, sous l'hypothèse H_0

$$Z = \frac{A_n^2 \sum_{i=1}^n (X_i - \bar{X}_n)^2}{S_n^2} \sim F(1, n-2)$$

On va écrire cette v.a. sous une autre forme (fonction que de Y).

Proposition 5.2.8 Sous l'hypothèse H_0

$$Z = (n-2) \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2}{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}$$

• La statistique utilisée pour tester H_0 est

$$Z = (n-2) \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2}{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2} = \frac{SM/1}{SR/(n-2)} \sim F(1, n-2)$$

• Zone d'acceptation. On fixe le risque α . Par définition de la loi de Fisher : $P(Z \leq f_{1,n-2;1-\alpha}) = 1 - \alpha$ où $f_{1,n-2;1-\alpha}$ est la fractile d'ordre $1 - \alpha$ de la loi de la loi de Fisher. Puisque Z prend que des valeurs positives, la zone d'acceptation est : $ZA_{H_0,\alpha} = [0, f_{1,n-2;1-\alpha}]$.

• On calcule la valeur z de la v.a. Z sur les données (x_i, y_i)

$$z = (n-2) \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2}{\sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

Exemple. On teste l'hypothèse : $H_0 : a = 0$ contre $H_1 : a \neq 0$.

Student : $Z = \frac{A_n \sqrt{\sum_{i=1}^{10} (X_i - \bar{X}_n)^2}}{S_n} \sim t(8)$. L'intervalle de confiance pour $\alpha = 0.05$ est $ZA_{H_0,\alpha} = [-t_{8;0.975}; t_{8;0.975}] = [-2.306; 2.306]$. La valeur de la statistique de test $z = \frac{\hat{a}_n \sqrt{\sum_{i=1}^{10} (x_i - \bar{x}_n)^2}}{\hat{\sigma}_n} = \frac{\hat{a}_n}{\sqrt{Var(A_n)}} = -\frac{48.5}{\sqrt{18.4}} \sim 10$. Donc : $z \notin ZA \Rightarrow H_0$ rejetée, H_1 acceptée. Il y a bien une relation linéaire entre la concentration d'ozone et la vitesse du vent.

Fisher : $Z = 8 \frac{\sum_{i=1}^{10} (\hat{Y}_i - \bar{Y}_n)^2}{\sum_{i=1}^{10} (\hat{Y}_i - Y_i)^2} \sim F(1, 8)$. $ZA_{H_0,\alpha} = [0; f_{1,8;0.95}] = [0; 5.32]$. Valeur de la statistique de test $z = 8 \frac{\sum_{i=1}^{10} (\hat{y}_i - \bar{y}_n)^2}{\sum_{i=1}^{10} (\hat{y}_i - y_i)^2} = 8 \frac{SM}{SR} = \frac{SM}{\hat{\sigma}^2} = \frac{15093}{117.5} \sim 12$. $z \notin ZA \Rightarrow H_0$ rejetée.

Remarque : Par les deux méthodes on devrait obtenir des résultats concordants.

Test d'un modèle linéaire spécifié

On veut tester simultanément les deux paramètres a et b . Puisque les estimateurs A_n et B_n des paramètres a et b ne sont pas indépendants, il serait incorrect de tester successivement a et puis b .

On pose l'hypothèse nulle : $H_0 : a = a_0$ et $b = b_0$ contre l'hypothèse alternative : $H_1 : a \neq a_0$ ou $b \neq b_0$. La construction du test repose sur le théorème suivant, que nous ne démontrerons pas :

Théorème 5.2.1 Sous l'hypothèse H_0 , nous avons

$$Z = \frac{n-2}{2} \frac{\sum_{i=1}^n [(A_n - a_0)X_i + (B_n - b_0)]^2}{\sum_{i=1}^n (Y_i - A_n X_i - B_n)^2} \sim F(2, n-2)$$

Construction de la zone d'acceptation : On fixe un risque α et on calcule (en utilisant les tables de la loi de Fisher) $f_{2,n-2;1-\alpha}$ t.q. $P[Z \leq f_{2,n-2;1-\alpha}] = 1 - \alpha$. La zone d'acceptation est alors $ZA_{H_0,\alpha} = [0; f_{2,n-2;1-\alpha}]$.

Exemple. $H_0 : a = -48, b = 220$, $H_1 : a \neq -48$ ou $b \neq 220$. $Z = 4 \frac{\sum_{i=1}^{10} [(A_n + 48)X_i + (B_n - 220)]^2}{\sum_{i=1}^{10} (Y_i - A_n X_i - B_n)^2} \sim_{H_0} F(2, 8)$. $ZA_{H_0,0.05} = [0; f_{2,8;0.95}] = [0; 4, 46]$.

5.2.7 Prédiction d'une valeur

On est dans la situation suivante : on a n mesures pour les var Y et X : $(y_i, x_i)_{1 \leq i \leq n}$. Entre les var Y et X existe un lien linéaire : $Y_i = aX_i + b + \varepsilon_i$, $i = 1, \dots, n$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. On sait construire des estimateurs A_n et B_n pour les paramètres a et b . Puisqu'on dispose de n données on peut préciser effectivement quelles sont les valeurs de A_n et B_n : \hat{a}_n et \hat{b}_n .

On désire maintenant de prévoir la valeur de Y pour une nouvelle valeur de X : x_{n+1} . On peut fournir deux estimateurs : ponctuel ou par intervalle.

La prédiction la plus naturelle est : $\hat{y}_{n+1} = \hat{a}_n x_{n+1} + \hat{b}_n$ qui est une réalisation de la v.a. $\hat{Y}_{n+1} = A_n X_{n+1} + B_n$, les estimateurs A_n et B_n étant construits à partir des n premières observations. Il faut donner un sens à cette prédiction \hat{Y}_{n+1} : la qualité. Alors, de point de vue statistique, il est plus correct de donner comme prédiction un intervalle, avec un niveau de confiance fixé, \hat{Y}_{n+1} étant le milieu de cet intervalle.

5.3 Régression linéaire multiple

Exemple. Supposons que l'on dispose des données suivantes, pour 3 variables :

Obs	y_i	x_{1i}	x_{2i}
1	10	6	28
2	20	12	40
3	17	10	32
4	12	8	36
5	11	9	34

$$\text{Corr}(Y, X_1) = 0.91 \quad \text{Corr}(Y, X_2) = 0.65$$

On déduit qu'il peut y avoir un lien linéaire entre Y et X_1, X_2

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \varepsilon_i \quad i = 1, \dots, 5 \quad (5.8)$$

avec b_0, b_1, b_2 paramètres inconnus, à estimer.

5.3.1 Estimation des paramètres

Le cadre du problème

Supposons qu'on a un échantillon de n mesures pour $(p+1)$ variables : Y, X_1, \dots, X_p avec $p < n$, Y variable aléatoire, X_i variables non-aléatoires. Comme d'habitude on va noter les valeurs mesurées avec des petites lettres $y_i, x_{1i}, \dots, x_{pi}$ $i = 1, \dots, n$. Ces données mesurées peuvent être représentées sous la forme d'un tableau

Obs	y	x_1	...	x_j	...	x_p
1	y_1	x_{11}	...	x_{j1}	...	x_{p1}
2	y_2	x_{12}	...	x_{j2}	...	x_{p2}
.
i	y_i	x_{1i}	...	x_{ji}	...	x_{pi}
.
n	y_n	x_{1n}	...	x_{jn}	...	x_{pn}

Donc, pour x_{ji} le j c'est pour la variable, le i pour l'observation.

On cherche à construire Y comme fonction linéaire des variables X_1, \dots, X_p .

L'équation modèle pour l'observation i est

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \dots + b_p X_{pi} + \varepsilon_i \quad i = 1, \dots, n \quad (5.9)$$

On a n équations, une pour chaque observation, et elles peuvent être résumées sous la forme matricielle

$$Y = X\beta + \varepsilon \quad (5.10)$$

avec

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1} \quad X = \begin{bmatrix} 1 & X_{11} & \dots & X_{p1} \\ & X_{12} & \dots & X_{p2} \\ & \vdots & \ddots & \vdots \\ 1 & X_{1i} & \dots & X_{pi} \\ & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \dots & X_{pn} \end{bmatrix} \quad \beta = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Pour que le modèle soit complètement spécifié, il faut donner la répartition des erreurs ε_i . On suppose $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \dots, n$. En plus ε_i et ε_j indépendantes, donc $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ pour $i \neq j$.

Les paramètres du modèle sont : b_0, b_1, \dots, b_p et la variance σ^2 . Il faut les estimer en connaissant les n observations.

Conséquences

- 1) $\mathbb{E}(\varepsilon) = 0$ (un vecteur de dimension n de 0)
- 2) $\text{Var}(\varepsilon) = \sigma^2 I_n$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$
- 3) $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$

Preuve 1) $\mathbb{E}(\varepsilon) = E[\varepsilon_1, \dots, \varepsilon_n]^t = 0$

2)

$$\text{Var}(\varepsilon) = \begin{bmatrix} \text{Var}(\varepsilon_1) & \text{Cov}(\varepsilon_1, \varepsilon_2) & \dots & \text{Cov}(\varepsilon_1, \varepsilon_n) \\ \text{Cov}(\varepsilon_2, \varepsilon_1) & \text{Var}(\varepsilon_2) & \dots & \text{Cov}(\varepsilon_2, \varepsilon_n) \\ \dots & \dots & \dots & \dots \\ \text{Cov}(\varepsilon_n, \varepsilon_1) & \text{Cov}(\varepsilon_n, \varepsilon_2) & \dots & \text{Var}(\varepsilon_n) \end{bmatrix} = \sigma^2 I_n$$

3) $\mathbb{E}(Y) = \mathbb{E}(X\beta + \varepsilon) = X\beta + \mathbb{E}(\varepsilon) = X\beta$, $\text{Var}(Y) = \text{Var}(X\beta + \varepsilon) = \text{Var}(\varepsilon) = \sigma^2 I_n$, $X\beta$ déterministe et $\varepsilon \sim \mathcal{N}$ donc $Y = X\beta + \varepsilon \sim \mathcal{N}$. Alors $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$.

Commentaires : 1) La régression linéaire multiple peut être vue comme une extension de la régression simple ($p = 1$).

2) C'est un problème plus difficile : les calculs sont plus difficiles et pratiquement impossible de s'en passer de l'ordinateur.

Estimateurs ponctuels de β et de σ^2

L'estimateur de moindres carrés du vecteur β .

Cet estimateur s'obtient d'après la même procédure que pour la régression simple. C'est le vecteur aléatoire qui minimise la fonction

$$T(\beta) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - b_0 - b_1 X_{1i} - \dots - b_p X_{pi})^2 = \varepsilon^t \varepsilon = (Y - X\beta)^t (Y - X\beta)$$

$$= (Y^t - \beta^t X^t) (Y - X\beta) = Y^t Y - \beta^t X^t Y - Y^t X\beta + \beta^t X^t X\beta$$

or $\beta^t X^t Y$ est un scalaire, donc, égal à son transposé. Donc

$$T(\beta) = Y^t Y - 2Y^t X\beta + \beta^t (X^t X) \beta = Y^t Y - 2\beta^t X^t Y + \beta^t (X^t X) \beta$$

Une condition nécessaire d'existence d'extremum est que la première dérivée de la fonction T par rapport à β soit nulle : $\frac{\partial T}{\partial b_i} = 0 \quad \forall i = 0, 1, \dots, p \Rightarrow -2X^t Y + 2X^t X\beta = 0$,

$(X^t X)\beta = X^t Y \Rightarrow \beta = (X^t X)^{-1} X^t Y$ avec la condition que la matrice $(X^t X)$ soit inversible. Donc, l'estimateur des moindres carrés du vecteur paramètre β est

$$B_n = (X^t X)^{-1} X^t Y$$

$$B_n = \begin{bmatrix} B_0 \\ B_1 \\ \dots \\ B_p \end{bmatrix}, \quad B_i \text{ est l'estimateur des moindres carrés pour } b_i \quad i = 0, \dots, p.$$

Si on a n mesures on obtient une valeur (réalisation) pour la v.a. $B_n : \hat{\beta}_n = (x^t x)^{-1} x^t y$, où y est le vecteur avec les mesures pour Y et x est la matrice avec les mesures pour x_1, \dots, x_p . La valeur prédite de Y par le modèle est $\hat{Y} = X \hat{B}_n$ et $Y - \hat{Y}$ s'appelle résidu.

Propriétés de l'estimateur B_n

1) Estimateur de β sans biais : $\mathbb{E}(B_n) = \beta$.

Preuve. $\mathbb{E}(B_n) = \mathbb{E}[(X^t X)^{-1} X^t Y] = (X^t X)^{-1} X^t \mathbb{E}(Y) = (X^t X)^{-1} X^t (X\beta) = \beta$

2) Variance de B_n . On note par $C = (X^t X)^{-1} X^t$ (une matrice non aléatoire). Donc $B_n = CY$.

$$\text{Var}(B_n) = \text{Var}(CY) = C \text{Var}(Y) C^t = (X^t X)^{-1} X^t \sigma^2 I_n [(X^t X)^{-1} X^t]^t$$

$$= \sigma^2 (X^t X)^{-1} X^t (X^t)^t [(X^t X)^{-1}] = \sigma^2 (X^t X)^{-1} (X^t X) [(X^t X)^t]^{-1} = \sigma^2 (X^t X)^{-1} = \sigma^2 (X^t X)^{-1}$$

Donc,

$$\text{Var}(B) = \begin{bmatrix} \text{Var}(B_0) & \text{Cov}(B_0, B_1) & \dots & \text{Cov}(B_0, B_p) \\ \text{Cov}(B_1, B_0) & \text{Var}(B_1) & \dots & \text{Cov}(B_1, B_p) \\ \dots & \dots & \dots & \dots \\ \text{Cov}(B_p, B_0) & \text{Cov}(B_p, B_1) & \dots & \text{Var}(B_p) \end{bmatrix} = \sigma^2 (X^t X)^{-1}$$

c'est une matrice $(p+1) \times (p+1)$.

3) Chaque élément B_j composant du vecteur B , $j = 0, \dots, p$ est une fonction linéaire des variables Y_1, \dots, Y_n . Cette propriété de linéarité détermine les propriétés statistiques de ces estimateurs. En particulier, puisque les $Y_i \sim \mathcal{N}$, les estimateurs des b_j suivent eux aussi une loi Normale, de variance facilement calculable.

4) Si on note $(X^t X)^{-1} = (c_{ij})_{1 \leq i, j \leq p+1}$, alors

– la variance de l'estimateur B_{i-1} de b_{i-1} est le i -ème élément diagonal de la matrice $\sigma^2 (X^t X)^{-1}$, c'est-à-dire $\sigma^2 c_{ii}$

– $Cov(B_{i-1}, B_{j-1}) = \sigma^2 c_{ij}$ pour $i \neq j$.

Estimateur pour σ^2 . On montre que

$$S_n^2 = \frac{(Y - XB_n)^t (Y - XB_n)}{n - p - 1} = \frac{(Y - \hat{Y})^t (Y - \hat{Y})}{n - p - 1}$$

est un estimateur sans biais de σ^2 . Une estimation de σ^2

$$\hat{\sigma}_n^2 = \frac{(y - x\hat{\beta}_n)^t (y - x\hat{\beta}_n)}{n - p - 1}$$

Propriétés

1) $(n - p - 1) \frac{S_n^2}{\sigma^2} \sim \chi^2(n - p - 1)$

2) B_i et S_n^2 sont indépendantes pour $\forall i = 0, 1, \dots, p$.

5.3.2 Décomposition de la variabilité de Y

Pareil que pour la régression linéaire simple, nous avons

$$\sum_{i=1}^n (y_i - \bar{y}_n)^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2$$

$ST = \sum_{i=1}^n (y_i - \bar{y}_n)^2$ est la somme des carrés totale : représente la variabilité des observations de Y avant de prendre en compte les effets des variables X_1, \dots, X_p .

$SR = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ est la somme des carrés résiduelle (la somme des carrés due aux erreurs) et elle représente la variabilité de Y inexpliquée après que les variables X_1, \dots, X_p ont été utilisées dans l'équation de régression pour prédire Y .

$SM = ST - SR = \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2$ la somme des carrés due au modèle de régression et mesure la variabilité due aux var. indép. X_1, \dots, X_p dans l'équation de régression.

On a le tableau de décomposition (ANOVA) :

Source de variation	ddl	S.C.	Carré moyen
Régression	p	SM	SM/p
Résiduelle	$n - p - 1$	SR	$SR/(n - p - 1)$
Totale	$n - 1$	ST	

5.3.3 Mesure de l'ajustement (empirique)

Est donnée par le coefficient de détermination : $R^2 = \frac{SM}{ST} \in [0, 1]$ qui donne une mesure sommaire, quantitative sur la qualité de la prédiction de Y par les variables X_1, \dots, X_p dans le modèle de régression linéaire. Il représente aussi le carré de la corrélation entre Y et \hat{Y} .

– Si on a une modélisation parfaite : $Y_i = \hat{Y}_i$ alors $SR = 0$, donc $ST + SM$ donc $R^2 = 1$.

– La valeur de R^2 croît si des nouvelles var. indép. sont ajoutées au modèle de régression.

– Similaire à la régression linéaire simple, seulement la valeur de R^2 est insuffisante pour bien caractériser la qualité de l'ajustement.

Exemple. Le tableau d'analyse de variance :

Source de variation	ddl	S.C.	Carré moyen
Modèle (X_1, X_2)	$p = 2$	62.5	31.25
Résidu	2	11.5	5.75
Totale	4	74	

$$R^2 = \frac{62.5}{74} = 0.85$$

5.3.4 Théorème de Gauss-Markov

Considérons un modèle linéaire général : $Y = X\beta + \varepsilon$ avec Y un vecteur aléatoire de dimension $n \times 1$, X une matrice d'ordre $n \times p$ et le vecteur des erreurs ε de dimension $n \times 1$. Soit $B_n = (X^t X)^{-1} X^t Y$ l'estimateur des moindres carrés de β .

Théorème 5.3.1 Soit ψ une application linéaire de \mathbb{R}^p dans \mathbb{R}^q et $Y = X\beta + \varepsilon$ un modèle linéaire où, pour tout i et tout j , $\mathbb{E}[\varepsilon_i] = 0$, $\text{Var}[\varepsilon_i] = \sigma^2 < \infty$ et $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$, pour $i \neq j$. Alors l'estimateur des moindres carrés $\psi(B_n)$ est une estimateur sans biais pour $\psi(\beta)$, uniformément de variance minimale parmi les estimateurs sans biais linéaires en Y .

Preuve. Soient A une matrice de dimension $q \times p$, $U \in \mathbb{R}^q$ et $\psi(U) = AU$. Soit $T(Y)$ un autre estimateur sans biais de $\psi(\beta)$ linéaire en Y , $T(Y) = TY$. On veut montrer que $\text{Var}[TY] \geq \text{Var}[AB_n]$, c'est-à-dire que la matrice $\text{Var}[TY] - \text{Var}[AB_n]$ est positivement définie.

Or $\text{Var}[TY] = T \text{Var}[Y] T^t = \sigma^2 T T^t$. Soit $P = X(X^t X)^{-1} X^t$. On a que $P + (I_n - P)(I_n - P) = I_n$ et donc

$$\begin{aligned} \text{Var}[TY] &= \sigma^2 T T^t = \sigma^2 T I_n T^t = \sigma^2 T [P + (I_n - P)(I_n - P)] T^t = \sigma^2 T P T^t + \sigma^2 T (I_n - P)(I_n - P) T^t \\ &= \sigma^2 (TX)(X^t X)^{-1} (TX)^t + \sigma^2 (T - TP)(T^t - PT^t) \end{aligned}$$

Mais TY et AB_n sont des estimateurs sans biais pour $A\beta$, donc, $\mathbb{E}[TU] = T\mathbb{E}[Y] = TX\beta = A\beta$ pour tout β , ce qui implique $TX = A$. On remplace dans l'équation précédente :

$$\begin{aligned} \text{Var}[TY] &= \sigma^2 A(X^t X)^{-1} A^t + \sigma^2 [T - TX(X^t X)^{-1} X^t][T^t - X(X^t X)^{-1} X^t T^t] \\ &= \text{Var}[AB_n] + \sigma^2 [T - A(X^t X)^{-1} X^t][T^t - X(X^t X)^{-1} A^t] = \text{Var}[AB_n] + \text{Var}[TY - AB_n] \end{aligned}$$

En effet $\text{Var}[TY - AB_n] = \mathbb{E}[(TY - AB_n)(TY - AB_n)^t] = \mathbb{E}[(TY - A(X^t X)^{-1} X^t Y)(TY - A(X^t X)^{-1} X^t Y)^t] = \mathbb{E}[(T - A(X^t X)^{-1} X^t) Y Y^t (T - A(X^t X)^{-1} X^t)^t] = (T - A(X^t X)^{-1} X^t) \mathbb{E}[Y Y^t] (T - A(X^t X)^{-1} X^t)^t = \sigma^2 (T - A(X^t X)^{-1} X^t)(T - A(X^t X)^{-1} X^t)^t$ ■

5.3.5 Tests d'hypothèse

Une fois le modèle de régression multiple fixé et les estimations des paramètres obtenues, on se pose la question sur la contribution des variables X_1, \dots, X_p sur la prédiction de Y .

Un des critères importants dans la sélection d'un modèle est de choisir celui qui, avec moins de variables, fournissait la meilleure description des données étudiées. Dans le cadre de la régression linéaire multiple, p variables peuvent s'avérer superflus et un nombre inférieur q ($q < p$) peut permettre une description aussi bonne.

Il y a 2 types de questions que l'on peut se poser

1. On teste si le groupe entier de variables indépendantes contribue significativement à la prédiction de Y .
2. Test pour ajouter une seule variable, quand les autres variables indépendantes sont déjà dans le modèle

Test de la significativité du modèle de régression entier

On a le modèle complet

$$Y_i = b_0 + b_1 X_{1i} + \dots + b_p X_{pi} + \varepsilon_i$$

Pour ce test l'hypothèse nulle peut se traduire comme :

H_0 : "Toutes les p variables indép. considérées dans le même temps ne produisent pas une variation en Y "

H_0 : " il n'y a pas de régression significative en utilisant les p var indép. dans le modèle

H_0 : $b_1 = b_2 = \dots = b_p = 0$ contre H_1 : $\exists j \in \{1, \dots, p\}$ t.q. $b_j \neq 0$

Sous l'hypothèse H_0 , le modèle réduit est : $Y_i = b_0 + \varepsilon_i$ $i = 1, \dots, n$. Pour faire ce test on utilise la statistique

$$Z = \frac{SM(X_1, \dots, X_p)/p}{SR(X_1, \dots, X_p)/(n-p-1)} = \frac{(ST - SR)/p}{SR/(n-p-1)} \sim F(p, n-p-1)$$

Pour un niveau α fixé, la zone d'acceptation est : $ZA = [0; f_{p, n-p-1, 1-\alpha}]$.

Exemple. $f_{2, 2, 0.95} = 19$ $ZA = [0; 19]$ $z = \frac{31.25}{5.75} = 5.4$

Apport d'une seule variable

Si l'ensemble des variables X_1, \dots, X_p est significatif dans la prévision de Y , on se pose la question d'effacer les variables qui ne servent pas à la prédiction de Y . Sans réduire la généralité, on suppose que l'on teste l'influence de X_p :

H_0 : X_p ne contribue pas de manière significative à la prédiction de Y si X_1, \dots, X_{p-1} sont déjà dans le modèle.

H_1 : X_p contribue de manière significative à la prédiction de Y si X_1, \dots, X_{p-1} sont déjà dans le modèle.

$$H_0 : b_p = 0 | b_j \neq 0, j \in \{1, \dots, p-1\} \quad \text{contre} \quad H_1 : b_p \neq 0 | b_j \neq 0, j \in \{1, \dots, p-1\}$$

Modèle complet : $Y_i = b_0 + b_1 X_{1i} + \dots + b_{p-1} X_{p-1,i} + b_p X_{pi} + \varepsilon_i$.

Modèle réduit : $Y_i = b_0 + b_1 X_{1i} + \dots + b_{p-1} X_{p-1,i} + \varepsilon_i$.

Accepter H_0 signifie que le $p^{\text{ème}}$ facteur n'apporte rien de plus après les $p-1$ variables. Mais ça, ne signifie pas que ce facteur seul n'a pas d'effet sur Y (X_p peut être corrélé avec X_1, \dots, X_{p-1}).

Pour tester l'hypothèse H_0 on utilise la statistique

$$Z = \frac{SM(X_1, \dots, X_p) - SM(X_1, \dots, X_{p-1})}{SR(X_1, \dots, X_p)/(n-p-1)} \sim f(1, n-p-1)$$

où : $SM(X_1, \dots, X_p)$ est la somme des carrés due au modèle dans le modèle complet ; $SM(X_1, \dots, X_{p-1})$ est la somme des carrés due au modèle dans le modèle réduit. Pour un risque α fixé, la zone d'acceptation est : $ZA_{H_0, \alpha} = [0; f_{1, n-p-1; 1-\alpha}]$.

Pour tester H_0 , on peut utiliser aussi une statistique qui suit une loi de Student

$$Z = \frac{B_p}{\sqrt{\text{Var}(B_p)}} \sim t(n-p-1)$$

où B_p est l'estimateur de b_p dans le modèle complet et $\text{Var}(B_p)$ est la variance de cet estimateur. La zone d'acceptation : $ZA = [-t_{n-p-1; 1-\alpha/2}; t_{n-p-1; 1-\alpha/2}]$

5.3.6 Sélection des régresseurs

Plutôt que de chercher à expliquer Y par les p variables explicatives, on peut chercher un ensemble de k ($k \leq p$) variables parmi les p , qui donnent une reconstitution presque aussi satisfaisante de Y .

Les objectifs d'une telle démarche :

- éliminer le nombre de prédictors (régresseurs) ;
- éliminer les variables redondantes qui augmentent de manière non justifiée la variance du modèle.

Les critères du choix

Ils dépendent des usages que l'on fait de la régression :

- reconstitution des y_i ;
- prévision des valeurs futures ;
- estimation précise des paramètres d'un modèle.

Le critère du R^2 est bien adapté au premier objectif. Il n'est pas à l'abri des critiques : il varie de façon monotone avec le nombre de variables : il ne peut qu'augmenter si on rajoute une variable, même peu corrélée avec Y . On ne peut pas donc l'utiliser pour choisir la taille d'un sous-ensemble de régresseurs.

Si l'objectif est de minimiser l'erreur de prévision, le R^2 n'est pas adapté et on préférera des critères tels que le $\hat{\sigma}^2$: plus $\hat{\sigma}^2$ est petit, plus le modèle est meilleur.

Les techniques de sélection

A. Recherche exhaustive

La première idée consiste à faire toutes les régressions possibles :

- à une variable : il y a p régressions ;
- à 2 variables : il y a $p(p-1)/2$ régressions ;
- ;
- à k variables : il y a C_n^k possibilités ;
- ;
- pour en finir avec le modèle complet à p variables.

Or, en total il y a 2^p régressions, y compris le modèle sans régresseurs. Cette procédure est forte longue :

- quand $p = 10$ il y a 1024 modèles possibles ;
- quand $p = 30$ il y a plus d'un milliard.

L'examen de tous les modèles serait d'ailleurs sans intérêt, car nombre d'entre eux sont très voisins.

A k régresseurs fixés, on choisira le modèle qui fournit le R^2 maximum. Si k n'est pas fixé, le modèle avec toutes les variables significatives.

B. Les méthodes pas à pas

Elles procèdent par élimination successive ou ajout successif des variables.

La *méthode descendante* consiste à éliminer la variable la moins significative parmi les p : en général celle qui provoque la diminution la plus faible de R^2 (c'est celle qui a la probabilité d'accepter H_0 la plus proche de 1). On recalcule alors la régression et on recommence jusqu'à l'élimination de $p - 1$ variables ou en fonction du test d'arrêt :

- on part avec le modèle complet [à p variables] :

$$Y_i = b_0 + b_1 X_{1i} + \dots + b_p X_{pi} + \varepsilon_i$$

- on teste : $H_{0j} : b_j = 0 \mid b_1, \dots, b_{j-1}, b_{j+1}, \dots, b_p \neq 0, j \in \{1, \dots, p\}$

S'il existe au moins une hypothèse H_{0j} acceptée, alors on élimine la variable pour laquelle le modèle réduit correspond au R^2 le plus grand : la probabilité d'accepter H_{0j} la plus proche de 1.

Dans le modèle à $p - 1$ variables on teste l'hypothèse si parmi les variables gardées il y a au moins une non significative.

On s'arrête quand on ne peut plus éliminer des variables.

La *méthode ascendante* procède en sens inverse :

- on part de la meilleure régression à une variable (par rapport à R^2) ;
- on cherche parmi les $p - 1$ régressions à 2 variables, incluant la première déjà sélectionnée ;
-
- On s'arrête soit au modèle complet soit quand on ne peut plus introduire de variables significatives.

Chapitre 6

ANALYSE DE VARIANCE

6.1 Analyse de variance à un facteur

6.1.1 Introduction

Exemple. Les 21 candidats à un oral ont été répartis au hasard entre 3 examinateurs. Le premier examinateur a fait passer l'oral à 6 étudiants, le second à 8 étudiants et le troisième à 7 étudiants. Les notes qu'ils ont eu sont :

Examineur	A	B	C
	10,11,11,12,13,15	8,11,11,13,14,15,16,16	10,13,14,14,15,16,16
Effectif	6	8	7
Moyenne	12	13	14

On se demande si la variation des moyennes peut être due au hasard ou si elle tient d'un réel "effet examinateur".

En général, l'analyse de variance (ANOVA) est une technique statistique utilisée pour étudier l'effet des variables qualitatives sur une variable quantitative Y .

6.1.2 Terminologie

- *facteur (variable qualitative)* : une variable qui prend un nombre fini de valeurs, pas nécessairement numériques (une valeur constitue une classe). Pour l'exemple on a le facteur "examineur" qui prend 3 valeurs : A, B, C.
- *niveau (population)* les différentes valeurs prises par un facteur.
- *test de l'effet d'un facteur* tester si les moyennes des populations sont égales.

La variable à modéliser (à prévoir) Y , comme pour la régression linéaire, est une variable qui prend que des valeurs numériques.

Pour l'exemple : Y : notes ; facteur : examinateur ; niveaux : A, B, C.

On utilise un vocabulaire particulier, introduit par les agronomes, qui ont été les premiers à s'intéresser à ce type de problème : la variable qualitative susceptible d'influencer sur la distribution de la variable quantitative étudiée est appelée "facteur" et ses valeurs "populations".

6.1.3 Données

On suppose qu'on a un seul facteur F et on dispose de k échantillons de tailles respectives n_1, \dots, n_k , correspondant chacun à un niveau différent du facteur F . On pose

$$n = \sum_{i=1}^k n_i$$

A chaque expérimentation on mesure la valeur de la variable Y . On peut alors présenter les données à l'aide du tableau suivant :

Niveau (population)	Nb. obs.	Valeurs de Y
1	n_1	$y_{11}, y_{12}, \dots, y_{1n_1}$
2	n_2	$y_{21}, y_{22}, \dots, y_{2n_2}$
...
k	n_k	$y_{k1}, y_{k2}, \dots, y_{kn_k}$

On observe que le nombre d'observations pour chaque population peut ne pas être le même.

Notations : Pour un niveau i :

$$Y_{i.} = \sum_{j=1}^{n_i} Y_{ij}, \quad \bar{Y}_{i.} = \frac{1}{n_i} Y_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

(la moyenne empirique des Y pour la population i)

$$Y_{..} = \sum_{i=1}^k Y_{i.} = \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}, \quad \bar{Y}_{..} = \frac{1}{n} Y_{..} = \frac{1}{n} \sum_{i=1}^k Y_{i.} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}$$

Hypothèse : les k échantillons sont indépendantes et de loi Normale. Plus précisément, on suppose que pour tout couple (i, j) les données y_{ij} sont des réalisations de la v.a. $Y_{ij} \sim \mathcal{N}(m_i, \sigma^2)$ et $Y_{ij}, Y_{i'j'}$ indépendantes pour $i \neq i'$ ou $j \neq j'$.

Autrement dit, pour chaque i , les données y_{i1}, \dots, y_{in_i} sont des réalisations des n_i v.a. Y_{i1}, \dots, Y_{in_i} indépendantes et de même loi $\mathcal{N}(m_i, \sigma^2)$.

L'objet de cet étude sera de savoir si les moyennes m_i sont toutes égales ou non.

6.1.4 Modèles statistiques

Puisque $Y_{ij} \sim \mathcal{N}(m_i, \sigma^2)$ on peut poser :

$$Y_{ij} = m_i + \varepsilon_{ij} \quad i = 1, \dots, k \quad j = 1, \dots, n_i \quad (6.1)$$

avec $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

Paramètres à estimer : m_i la moyenne de la population i , σ^2 la variance.

Le modèle (6.1) peut être écrit sous une forme équivalente :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, \dots, k \quad j = 1, \dots, n_i \quad (6.2)$$

où :

- μ représente une valeur appelée "effet moyen" ;
- α_i représente l'effet du niveau i du facteur F .

Alors, on doit estimer $k + 1$ paramètres : μ et α_i ($i = 1, \dots, k$) plus la variance σ^2 .

Le modèle écrit sous la forme (6.2) a une indétermination, car $(\mu + \alpha_i)$ peut s'obtenir d'une infinité de manières.

On remédie cela, en introduisant une contrainte, qui est en généralement la suivante : $\sum_{i=1}^k n_i \alpha_i = 0$.

En utilisant une notation vectorielle, le modèle (6.1) prend la forme :

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ \dots \\ Y_{1n_1} \\ Y_{21} \\ \dots \\ Y_{2n_2} \\ \dots \\ Y_{k1} \\ \dots \\ Y_{kn_k} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ \dots \\ m_k \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \dots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \dots \\ \varepsilon_{2n_2} \\ \dots \\ \varepsilon_{k1} \\ \dots \\ \varepsilon_{kn_k} \end{bmatrix} \quad (6.3)$$

ou encore

$$Y = X\beta + \varepsilon \quad (6.4)$$

Donc, l'analyse de variance est un modèle linéaire.

6.1.5 Estimation des paramètres

Pour les modèles (6.1) ou (6.3), il faut trouver les valeurs de m_i qui minimise la fonction :

$$T(m_i) = \sum_{i=1}^k \sum_{j=1}^{n_i} \varepsilon_{ij}^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - m_i)^2$$

En faisant des calculs, on obtient que : $\hat{m}_i = \bar{Y}_{i..}$. Sous l'hypothèse de normalité et d'indépendance des échantillons, $\bar{Y}_{i..}$ est un estimateur sans biais de m_i et

$$\hat{m}_i = \bar{Y}_{i..} \sim \mathcal{N}\left(m_i, \frac{\sigma^2}{n_i}\right)$$

Pour le modèle (6.2) les paramètres à estimer sont : μ et les α_i , $i = 1, \dots, k$. On utilise la décomposition : $\varepsilon_{ij} = \bar{\varepsilon}_{..} + (\bar{\varepsilon}_{i.} - \bar{\varepsilon}_{..}) + (\varepsilon_{ij} - \bar{\varepsilon}_{i.})$ et par des calculs élémentaires on obtient :

$$\sum_{i=1}^k \sum_{j=1}^{n_i} \varepsilon_{ij}^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} \bar{\varepsilon}_{..}^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{\varepsilon}_{i.} - \bar{\varepsilon}_{..})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (\varepsilon_{ij} - \bar{\varepsilon}_{i.})^2 \quad (6.5)$$

On écrit les ε fonction des paramètres à estimer :

$$\begin{aligned} \varepsilon_{ij} &= Y_{ij} - \mu - \alpha_i, & \varepsilon_{i.} &= Y_{i.} - n_i \mu - n_i \alpha_i, & \bar{\varepsilon}_{i.} &= \bar{Y}_{i.} - \mu - \alpha_i \\ \varepsilon_{..} &= Y_{..} - \sum_{i=1}^k n_i \mu - \sum_{i=1}^k n_i \alpha_i, & \varepsilon_{..} &= Y_{..} - n \mu, & \bar{\varepsilon}_{..} &= \bar{Y}_{..} - \mu \end{aligned}$$

alors la relation (6.5) devient :

$$\sum_{i=1}^k \sum_{j=1}^{n_i} \varepsilon_{ij}^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{Y}_{..} - \mu)^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..} - \alpha_i)^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 \quad (6.6)$$

Le membre droit de (6.6) est minimisé pour :

$$\hat{\mu} = \bar{Y}_{..}, \quad \hat{\alpha}_i = \bar{Y}_{i.} - \bar{Y}_{..}$$

Il faut vérifier que : $\sum_{i=1}^k n_i \hat{\alpha}_i = 0$:

$$\sum_{i=1}^k n_i \alpha_i = \sum_{i=1}^k n_i \bar{Y}_{i.} - \sum_{i=1}^k n_i \bar{Y}_{..} = \sum_{i=1}^k Y_{i.} - n \bar{Y}_{..} = 0$$

L'estimateur du maximum de vraisemblance modifié pour σ^2 est :

$$S_n^2 = \frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$$

6.1.6 Tests d'hypothèses

Tableau d'analyse de variance

On veut d'abord tester l'hypothèse qu'il n'y a pas k niveaux (populations) différents, mais qu'ils sont tous confondus : les n observations proviennent d'une population unique d'espérance m . Pour le modèle (6.1) ou (6.3), l'hypothèse nulle a la forme :

$H_0 : m_1 = m_2 = \dots = m_k = m$, contre $H_1 : \exists i, j \in \{1, \dots, k\}$ tels que $m_i \neq m_j$.

Ou, équivalent pour le modèle (6.2) : $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$, contre $H_1 : \exists i \in \{1, \dots, k\}$ tel que $\alpha_i \neq 0$.

Sous l'hypothèse H_0 , le modèle a la forme : $M_{réduit} : Y_{ij} = \mu + \varepsilon_{ij}$.

L'estimation pour μ : $\hat{\mu} = \bar{Y}_{..}$ et la prévision de Y_{ij} : $\hat{Y}_{ij} = \hat{\mu}$. Alors, le résidu, sous l'hypothèse H_0 est : $Y_{ij} - \bar{Y}_{..}$.

La variabilité totale est : $\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2$. On peut écrire : $Y_{ij} - \bar{Y}_{..} = (Y_{ij} - \bar{Y}_{i.}) + (\bar{Y}_{i.} - \bar{Y}_{..})$ et par des calculs élémentaires, on obtient :

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2$$

(variabilité totale = variabilité résiduelle + variabilité due au modèle) : ST = SR + SM.

On peut résumer cette décomposition par le tableau d'analyse de variance :

Source de variation	ddl	S.C.	Carré moyen
Modèle	$k - 1$	SM	$SM/(k - 1)$
Résiduelle	$n - k$	SR	$SR/(n - k)$
Totale	$n - 1$	ST	

Test d'égalité des k effets

Pour tester l'hypothèse H_0 on utilise la statistique :

$$Z = \frac{SM/(k-1)}{SR/(n-k)} \sim F(k-1, n-k) \quad (\text{sous } H_0)$$

Pour un risque α fixé, la zone d'acceptation est : $ZA_{H_0, \alpha} = [0, f_{k-1, n-k; 1-\alpha}]$

Exemple. Les modèles attachés :

$$Y_{ij} = m_i + \varepsilon_{ij}, \quad i = 1, 2, 3, \quad j = 1, \dots, n_i, \quad n_1 = 6 \quad n_2 = 8 \quad n_3 = 7 \quad (6.7)$$

ou encore

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, 2, 3, \quad j = 1, \dots, n_i, \quad n_1 = 6 \quad n_2 = 8 \quad n_3 = 7 \quad (6.8)$$

Les estimations des paramètres : $\hat{\mu} = \bar{y}_{..} = 13.04$, $\hat{\alpha}_1 = \bar{y}_{1.} - \bar{y}_{..} = 12 - 13.04 = -1.04$, $\hat{\alpha}_2 = \bar{y}_{2.} - \bar{y}_{..} = 13 - 13.04 = -0.04$, $\hat{\alpha}_3 = \bar{y}_{3.} - \bar{y}_{..} = 14 - 13.04 = 0.96$. On veut tester s'il y a un effet examinateur : les examinateurs n'ont pas le même système de notation :

$H_0 : m_1 = m_2 = m_3 = m$ contre $H_1 : \exists i \neq j$ tel que $m_i \neq m_j$

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ contre $H_1 : \exists i \neq j$ tel que $\alpha_i \neq 0$

SM=12.95, SR=98, $z = \frac{SM/(3-1)}{SR/(21-3)} = 1.19$, $ZA_{H_0, 1-\alpha} = [0, f_{2, 18; 0.096}] = (0, 3.55]$. Donc, H_0 acceptée : les examinateurs ont le même système de notation.

Comparaison de moyennes

Le rejet de l'hypothèse d'égalité des moyennes ne signifie pas que tous les m_i sont différentes entre eux. On cherche souvent à tester l'égalité entre deux moyennes :

$H_0 : m_h = m_j$ contre $H_1 : m_h \neq m_j$ pour $h \neq j$.

On utilise la statistique de test :

$$Z = \frac{\bar{Y}_h - \bar{Y}_j}{\sqrt{\frac{SR}{n-k} \left(\frac{1}{n_h} + \frac{1}{n_j} \right)}}$$

La zone d'acceptation $ZA_{H_0, 1-\alpha} = [-t_{n-k; 1-\alpha/2}, t_{n-k; 1-\alpha/2}]$.

6.2 Analyse de variance à deux facteurs

6.2.1 Introduction

On a vu comment comparer les populations d'un même facteur. Supposons maintenant qu'un expérimentateur souhaite comparer l'influence de trois régimes alimentaires et de deux exploitations sur la production laitière. Les résultats expérimentaux sont dans le tableau suivant.

Expl ↓ R.alim →	A	B	C	Total	Moyenne
1	7	36	2	45	15
2	13	44	18	75	215
Total	20	80	20	120	
Moyenne	10	40	10		20

6.2.2 Données

On suppose qu'on a deux facteurs (variables) F1 et F2. Le nombre de niveaux (valeurs possibles) pour F1 est de p et pour F2 est de q . Pour chaque couple (i, j) de niveaux on a $r (\geq 1)$ observations de la variable dépendante Y . Alors, on peut présenter les données à l'aide du tableau suivant :

F1 / F2	1	i	p
1	y_{111}, \dots, y_{11r}	y_{i11}, \dots, y_{i1r}	y_{p11}, \dots, y_{p1r}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
j	y_{j11}, \dots, y_{j1r}	y_{ij1}, \dots, y_{ijr}	y_{pj1}, \dots, y_{pjr}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
q	y_{q11}, \dots, y_{q1r}	y_{iq1}, \dots, y_{iqr}	y_{pq1}, \dots, y_{pqr}

Dans la cellule (i, j) nous avons les valeurs (observations) y_{ijk} : i donne le niveau (population) du facteur $F1$, j le niveau de $F2$ et k la répétition pour un couple (i, j) . On a pq cellules et dans chaque cellule il y a r observations. Notations :

$$\begin{cases} y_{ij.} = \sum_{k=1}^r y_{ijk} & \bar{y}_{ij.} = \frac{1}{r} y_{ij.} \\ y_{i..} = \sum_{j=1}^q \sum_{k=1}^r y_{ijk} & \bar{y}_{i..} = \frac{1}{qr} y_{i..} \\ y_{.j.} = \sum_{i=1}^p \sum_{k=1}^r y_{ijk} & \bar{y}_{.j.} = \frac{1}{pr} y_{.j.} \\ y_{...} = \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^r y_{ijk} & \bar{y}_{...} = \frac{1}{pqr} y_{...} \end{cases}$$

Les observations y_{ijk} sont des réalisations de la v.a. Y_{ijk} sur laquelle on fait les hypothèses :

$$\begin{cases} Y_{ijk} \sim \mathcal{N}(m_{ij}, \sigma^2) & \forall k = 1, \dots, r \\ Y_{ijk}, Y_{i'j'k'} & \text{indépendantes} \end{cases}$$

En ce qui concerne le nombre r de répétitions on a 2 situations :

- $r > 1$
- $r = 1$. Il n'y a pas de répétition et on va noter Y_{ij} par Y_{ij} .

Alors, les modèles statistiques considérés seront fonction de ces 2 situations. Les problèmes à traiter seront les mêmes que pour un seul facteur :

- écrire un modèle statistique de Y fonction des facteurs ;
- estimer les effets des niveaux des deux facteurs ;
- test d'hypothèse.

6.2.3 Modèle sans interaction (additif) : $r=1$

Le modèle le plus simple est d'additionner les effets du facteur $F1$ avec les effets du facteur $F2$:

$$m_{ij} = \mu + \alpha_i + \beta_j \quad (6.9)$$

où :

- μ est l'effet moyen
- α_i est l'effet dû au niveau i du facteur $F1$;
- β_j est l'effet dû au niveau j du facteur $F2$;

Puisque $Y_{ijk} \sim \mathcal{N}(m_{ij}, \sigma^2)$ on peut considérer un modèle :

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad (6.10)$$

Ce dernier modèle est indéterminé, car on peut obtenir la relation (6.9) par une infinité de manières. On remédie ça, en introduisant des contraintes, par exemple :

$$\sum_{i=1}^p \alpha_i = 0 \quad \sum_{j=1}^q \beta_j = 0$$

Estimation des paramètres

Il faut trouver les valeurs de m_{ij} (ou de μ, α_i, β_j) qui minimisent la fonction :

$$T(m_{ij}) = \sum_{i=1}^p \sum_{j=1}^q \varepsilon_{ij}^2 = \sum_{i=1}^p \sum_{j=1}^q (Y_{ij} - m_{ij})^2 = \sum_{i=1}^p \sum_{j=1}^q (Y_{ij} - \mu - \alpha_i - \beta_j)^2 \quad (6.11)$$

On utilise la même technique que pour l'analyse de variance à un facteur, et on obtient :

$$\hat{\alpha}_i = \bar{Y}_{i.} - \bar{Y}_{..} \quad \hat{\beta}_j = \bar{Y}_{.j} - \bar{Y}_{..} \quad \hat{\mu} = \bar{Y}_{..}$$

La valeur prédite pour Y_{ij} est :

$$\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \bar{Y}_{i.} + \bar{Y}_{.j} - \bar{Y}_{..}$$

Exemple. $F1$ est le régime alimentaire, qui prend 3 valeurs (A, B, C), donc $p = 3$. $F2$ est l'exploitation, qui prend 2 valeurs (1 et 2), donc $q = 2$. Le modèle statistique est :

$$Y_{ij} = \mu + \alpha_i + \beta_j \quad i = 1, 2, 3 \quad j = 1, 2$$

où : α_1 est l'effet de l'exploitation no. 1 sur Y , β_1 est l'effet du régime A sur la production laitière..... Les estimations des paramètres sont : $\hat{\mu} = \bar{y}_{..} = 20$, $\hat{\alpha}_1 = \bar{y}_{1.} - \bar{y}_{..} = 10 - 20 = -10$, $\hat{\alpha}_2 = 20$, $\hat{\alpha}_3 = -10$, $\hat{\beta}_1 = \bar{y}_{.1} - \bar{y}_{..} = 15 - 20 = -5$, $\hat{\beta}_2 = 5$. La prévision de Y_{11} (pour le régime alimentaire A et l'exploitation 1) : $\hat{Y}_{11} = \hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_1 = 20 - 10 - 5 = 5$.

Tableau d'analyse de variance

En partant de l'identité : $Y_{ij} - \bar{Y}_{..} = (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}) + (\bar{Y}_{i.} - \bar{Y}_{..}) + (\bar{Y}_{.j} - \bar{Y}_{..})$. On obtient :

$$\sum_{ij} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{ij} (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2 + q \sum_{i=1}^p (\bar{Y}_{i.} - \bar{Y}_{..})^2 + p \sum_{j=1}^q (\bar{Y}_{.j} - \bar{Y}_{..})^2$$

ou encore $ST = SR + S_{F1} + S_{F2}$. On peut résumer cette décomposition par le tableau d'analyse de variance :

Source de variation	ddl	S.C.	Carré moyen
F1	$p - 1$	S_{F1}	$S_{F1}/(p - 1)$
F2	$q - 1$	S_{F2}	$S_{F2}/(q - 1)$
Résidu	$(p - 1)(q - 1)$	SR	$SR/(p - 1)(q - 1)$
Totale	$pq - 1$	ST	

Test d'hypothèse

On peut tester deux types d'hypothèse : modèle significatif, l'effet de chaque facteur.

Test du modèle. Le modèle n'est pas significatif si aucun des deux facteurs n'influencent Y :

$$H_0 : \alpha_1 = \dots = \alpha_p = \beta_1 = \dots = \beta_q = 0$$

contre :

$$H_1 : \exists i \in \{1, \dots, p\} \text{ ou } \exists j \in \{1, \dots, q\} \text{ t.q. } \alpha_i \neq 0 \text{ ou } \beta_j \neq 0.$$

Le modèle complet est (6.10) et le modèle réduit : $Y_{ij} = \mu + \varepsilon_{ij}$.

Statistique de test :

$$Z = \frac{(S_{F1} + S_{F2})/(p + q - 2)}{SR/(p - 1)(q - 1)} \sim F(p + q - 2, (p - 1)(q - 1)) \quad \text{sous } H_0$$

Test d'un facteur. Supposons que l'on veut tester l'effet de F1.

H_0 F1 n'influe pas Y sachant que F2 est dans le modèle.

$$H_0 : \alpha_1 = \dots = \alpha_p = 0 \text{ contre } H_1 : \exists i \in \{1, \dots, p\} \text{ t.q. } \alpha_i \neq 0.$$

Le modèle complet est (6.10) et le modèle réduit : $Y_{ij} = \mu + \beta_j + \varepsilon_{ij}$. (modèle à un facteur)

L'hypothèse H_0 peut être traduite sous la forme : la moyenne m_{ij} ne dépend pas de i . Statistique de test :

$$Z = \frac{(S_{F1})/(p - 1)}{SR/(p - 1)(q - 1)} \sim F(p - 1, (p - 1)(q - 1)) \quad \text{sous } H_0$$

Exemple. Le tableau d'analyse de variance est :

Source de variation	ddl	S.C.	Carré moyen
F1	2	1200	600
F2	1	150	150
Résidu	2	28	14
Totale	5	1378	

On teste si le modèle est significatif : $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \beta_1 = \beta_2 = 0$:

$$Z = \frac{(S_{F1} + S_{F2})/(3 + 2 - 2)}{SR/2} \sim F(3, 2) \quad \text{sous } H_0$$

$ZA = [0; f_{3,2;0.95}] = [0; 19.2]$, $z = \frac{1350/3}{14} = 32.1 \notin ZA$. Donc H_0 est rejetée et le modèle est significatif. On teste si le facteur régime alimentaire actionne sur la production laitière : $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ sachant que l'exploitation est dans le modèle. L'hypothèse alternative est $H_1 : \exists i \in \{1, 2, 3\}$ t.q. $\alpha_i \neq 0$. Le modèle sous H_0 est $Y_{ij} = \mu + \beta_j + \varepsilon_{ij}$, $i = 1, 2, 3$, $j = 1, 2$. La statistique de test $Z = \frac{S_{F1}/2}{S_{R/2}} \sim F(2, 2)$ sous H_0 . $ZA = [0; f_{2,2;0.95}] = [0; 19.0]$, $z = \frac{600}{14} = 42.86 \notin ZA$. Donc H_0 est rejetée, le régime alimentaire est un facteur influent sur la production laitière.

6.2.4 Modèle avec interaction (additif) : $r > 1$

Dans ce cas, pour chaque couple (i, j) de niveaux on a r ($r > 1$) observations de la variable Y . C'est-à-dire que le tableau de données contient pq cellules et chaque cellule contient r observations. En ce qui concerne l'hypothèse statistique, l'hypothèse que les actions des deux facteurs F1 et F2 s'ajoutent est une hypothèse simplificatrice, qui n'est pas toujours réalisée. Il peut y avoir une interaction des facteurs F1 et F2, c'est-à-dire que pour certains couples (i, j) l'action simultanée de F1 au niveau i et de F2 au niveau j peut être bénéfique sur Y .

C'est ainsi que pour reprendre l'exemple, un certain régime alimentaire peut être particulièrement adapté à une certaine exploitation.

Les mesures sont y_{ijk} qui sont des réalisations de la v.a.

$$Y_{ijk} \sim \mathcal{N}(m_{ij}, \sigma^2) \quad i = 1, \dots, p \quad j = 1, \dots, q \quad k = 1, \dots, r$$

Le modèle statistique considéré est :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \quad \text{avec } \varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2) \quad (6.12)$$

avec γ_{ij} l'effet de l'interaction entre le niveau i du facteur F1 et le niveau j du facteur F2.

Paramètres à estimer : $\mu, \alpha_i, \beta_j, \gamma_{ij}$ et σ^2 . Le modèle (6.12) est indéterminé. On introduit les contraintes :

$$\sum_{i=1}^p \alpha_i = 0, \quad \sum_{j=1}^q \beta_j = 0, \quad \sum_{i=1}^p \gamma_{ij} = 0 \quad \forall j, \quad \sum_{j=1}^q \gamma_{ij} = 0 \quad \forall i$$

Exemple. On suppose que l'on a les mêmes facteurs et que $r = 2$:

Expl ↓ R.alim →	A	B	C
1	7, 8	36, 30	2, 5
2	13, 15	44, 45	18, 20

Estimation des paramètres

Il faut trouver les valeurs de $\mu, \alpha_i, \beta_j, \gamma_{ij}$ qui minimisent la fonction :

$$T(\mu, \alpha_i, \beta_j, \gamma_{ij}) = \sum_{i=1}^p \sum_{j=1}^q \varepsilon_{ijk}^2 = \sum_{i=1}^p \sum_{j=1}^q (Y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2 \quad (6.13)$$

On utilise la même technique que pour l'analyse de variance à un (deux) facteur, et on obtient :

$$\hat{\gamma}_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}, \quad \hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...}, \quad \hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{...}, \quad \hat{\mu} = \bar{Y}_{...}$$

La valeur prédite pour Y_{ij} est :

$$\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij} = \bar{Y}_{ij.}$$

Tableau d'analyse de variance

En utilise l'identité :

$$Y_{ijk} - \bar{Y}_{...} = (Y_{ijk} - \bar{Y}_{ij.}) + (Y_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}) + (\bar{Y}_{i..} - \bar{Y}_{...}) + (\bar{Y}_{.j.} - \bar{Y}_{...})$$

On obtient :

$$\sum_{i,j,k} (Y_{ijk} - \bar{Y}_{...})^2 = \sum_{i,j,k} (Y_{ijk} - \bar{Y}_{ij.})^2 + r \sum_{i,j} (Y_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2 + r q \sum_{i=1}^p (\bar{Y}_{i..} - \bar{Y}_{...})^2 + r p \sum_{j=1}^q (\bar{Y}_{.j.} - \bar{Y}_{...})^2$$

ou encore $ST = SR + S_{F1 \times F2} + S_{F1} + S_{F2} \equiv SR + SM$. On présente usuellement les résultats sous la forme du tableau d'analyse de variance :

Source de variation	ddl	S.C.	Carré moyen
F1	$p - 1$	S_{F1}	$S_{F1}/(p - 1)$
F2	$q - 1$	S_{F2}	$S_{F2}/(q - 1)$
F1*F2	$(p - 1)(q - 1)$	S_{F1*F2}	$S_{F1*F2}/(p - 1)(q - 1)$
Résidu	$pq(r - 1)$	SR	$SR/pq(r - 1)$
Totale	$pqr - 1$	ST	

L'estimateur sans biais de σ^2 :

$$S^2 = \frac{SR}{pq(r - 1)} = \frac{\sum_{i,j,k} (Y_{ijk} - \bar{Y}_{...})^2}{pq(r - 1)}$$

$$SR = \sum_i \sum_j \sum_k (y_{ijk} - \hat{y}_{ijk})^2$$

$$\hat{y}_{ijk} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij}$$

Test d'hypothèse

On peut tester les hypothèses :

- si le modèle est significatif ;
- l'effet d'un facteur sur Y ;
- l'effet de l'interaction entre les deux facteurs sur Y .

Test du modèle. $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_p = 0$ et $\beta_1 = \dots = \beta_q = 0$ et $\gamma_{ij} = 0 \forall i, j$
contre

$H_1 : \exists i \in \{1, \dots, p\}$ ou $\exists j \in \{1, \dots, q\}$ t.q. $\alpha_i \neq 0$ ou $\beta_j \neq 0$ ou $\gamma_{ij} \neq 0$.

Modèle réduit : $Y_{ijk} = \mu + \varepsilon_{ijk}$.

Pour tester l'hypothèse H_0 on utilise la statistique :

$$Z = \frac{(S_{F1} + S_{F2} + S_{F1*F2})/(pq - 1)}{SR/pq(r - 1)} \sim F(pq - 1, pq(r - 1))$$

Test d'un facteur. Sans réduire la généralité on suppose que l'on teste F1.

$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_p = 0$ sachant que F1 et F1*F2 sont dans le modèle.

$H_1 : \exists i \in \{1, \dots, p\}$ t.q. $\alpha_i \neq 0$

Le modèle réduit est : $Y_{ijk} = \mu + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$.

Statistique de test :

$$Z = \frac{S_{F1}/(p - 1)}{SR/pq(r - 1)} \sim F(p - 1, pq(r - 1))$$

Test de l'interaction. $H_0 : \gamma_{ij} = 0, \forall i, j$ sachant que F1 et F2 sont dans le modèle, contre $H_1 : \exists \gamma_{ij} \neq 0$. Le modèle réduit est : $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$.

Statistique de test :

$$Z = \frac{S_{F1*F2}/(p - 1)(q - 1)}{SR/pq(r - 1)} \sim F((p - 1)(q - 1), pq(r - 1))$$

BIBLIOGRAPHIE - TD

- 1) J.P.LECOUTRE- "Statistique et Probabilités"
- 2) A. COMBROUZE - "Probabilités et Statistiques", Vol. 2
- 3) A. ROUGG- "Probabilités et Statistiques"
- 4) P. JAFFARD- "Initiation aux méthodes de la Statistique et du calcul des probabilités"
- 5) G. BAILLARGEON- "Probabilités, Statistique et techniques de régression"
- 6) A. MATTEI- "Inférence et décision statistique"
- 7) C. MOUCHOT- "Exercices pédagogiques et statistique et Econométrie"
- 8) R.A. JOHNSON, G.K. BHATTACHARYYA- "Statistics- Principles and Methods"
- 9) P. DAGNELIE- "Statistique théorique et appliquée", Vol.2.
- 10) A. PHILIPPE, M-C VIANO - Cours de Statistique de base,
www.math.sciences.univ-nantes.fr/~philippe/download/Aphilippe-MCviano-cours-stat-MIM.pdf