ÉCOLE NATIONALE DES PONTS ET CHAUSSÉES

Année universitaire 2006 – 2007

COURS DE STATISTIQUE et ANALYSE des DONNÉES

12 juin 2007

Préambule

Le mot **Statistique** vient de l'allemand **Statistik**, qui désigne des données utiles à l'État, *der Staat* (dix-huitième siècle). Ce premier sens renvoie donc à ce que nous appelons aujourd'hui, *au pluriel*, **"des statistiques"**, c'est-à-dire des données chiffrées, généralement de grande ampleur, avec un accent sur le fait que ces données ne sont pas à transmettre brutes à l'utilisateur ("le prince" dans le sens initial) mais doivent être triées, clarifiées et organisées en fonction d'usages bien précis.

On peut se mettre à parler, historiquement de **"Statistique"**(*au singulier*) en tant que science quand, dans le courant du dix-neuvième siècle, on prend en compte le fait que ces données sont entachées d'aléatoire (en particulier en recourant au tirage d'échantillons) et que donc on modélise leur recueil par l'usage des méthodes mathématiques du calcul des probabilités. Ce point de vue connaît un grand essor dans la première moitié du vingtième siècle.

Dans la deuxième moité du vingtième siècle le progrès des moyens de calcul permet de faire subir des traitements simplificateurs à des masses de données de plus en plus grandes, et ce éventuellement préalablement à la mise en place de modèles probabilistes ; c'est alors qu'émerge la terminologie **"Analyse des données"**, qui ne doit pas être vue comme s'opposant à "Statistique" mais complémentaire, les deux points de vue s'étant largement interpénétrés dans leur progression commune jusqu'en ce début du vingt-et-unième siècle.

Dans tout ce cours nous désignerons du vocable **"statisticien"** le praticien dont nous décrirons et justifierons les techniques qu'il emploie face aux données qui lui sont fournies.

Table des matières

Préambule **iii**

A Introduction 1

I L'activité du statisticien **3**
I.1 Le contexte . . . 3
I.2 La démarche . . . 5
I.3 Le modèle . . . 5

B Statistique décisionnelle 7

II Rappels sur le modèle paramétrique **9**
II.1 Un exemple introductif . . . 9
II.1.1 Présentation d'un jeu de données . . . 9
II.1.2 Remarques sur la modélisation . . . 9
II.1.3 Quels questionnements peut-on soulever ? . . . 11
II.2 Définitions et rappels . . . 12
II.2.1 Les modèles paramétriques . . . 12
II.2.2 Rappels sur la LFGN et le TCL . . . 13
II.3 Estimation ponctuelle : les estimateurs . . . 13
II.3.1 Propriétés des estimateurs . . . 14
II.3.2 L'estimateur du maximum de vraisemblance . . . 16
II.4 Résumé des données : les statistiques exhaustives . . . 18
II.5 Précision de l'estimation : l'intervalle de confiance . . . 19
II.6 Procédure de décision : les tests . . . 23
II.7 Utiliser des données extérieures : l'approche bayesienne . . . 29
II.8 Résumé des procédures de test . . . 31
II.8.1 Le modèle de Bernoulli . . . 31

III Modèle linéaire gaussien **33**
III.1 Généralités . . . 33
III.1.1 Plans d'expériences . . . 33
III.1.2 Le modèle général . . . 33
III.1.3 Exemples . . . 34

III.2 Lois associées aux échantillons gaussiens 37
III.2.1 Théorème de Cochran 37
III.2.2 Statistiques fondamentales 38
III.3 Le modèle gaussien 39
III.3.1 Un exemple de données réelles à loi gaussienne 39
III.3.2 Étude du modèle 40
III.3.3 Estimation 41
III.3.4 Intervalle de confiance et tests pour la moyenne 41
III.3.5 Intervalle de confiance pour la moyenne 41
III.3.6 Intervalles de confiance et tests pour la variance 43
III.3.7 Analyse des données réelles 44
III.3.8 Approche bayesienne 44
III.4 Comparaison de moyennes de deux échantillons gaussiens 45
III.5 Analyse de la variance à un facteur 46
III.5.1 Comparaison de plusieurs échantillons gaussiens 46
III.5.2 Cadre général du modèle linéaire gaussien 48
III.5.3 Répartition de la variation 48
III.5.4 Test pour l'égalité des moyennes de k échantillons gaussiens 49
III.6 Régression linéaire multiple 52
III.6.1 Définition du modèle 52
III.6.2 Estimation 53
III.6.3 Test de l'utilité des régresseurs 55
III.6.4 Analyse des résidus et des observations 57
III.7 Résumé 60
III.7.1 Le modèle gaussien à variance connue 60
III.7.2 Le modèle gaussien à variance inconnue 60
III.7.3 Comparaison de moyennes de deux échantillons gaussiens 61
III.7.4 Analyse de la variance à un facteur 61
III.7.5 Régression multiple 62

IV Modèles discrets **63**
IV.1 Problématique des modèles discrets 63
IV.2 Tests du χ^2 64
IV.2.1 Test d'adéquation à une loi discrète 64
IV.2.2 Test d'adéquation à une famille de lois discrètes 67
IV.3 La régression logistique 68
IV.3.1 Exemple introductif 69
IV.3.2 Une solution raisonnable : la régression logistique 71
IV.3.3 Comment tester l'influence de la variable explicative ? 73
IV.4 Comparaison de proportions 75
IV.4.1 Présentation de la problématique sur un exemple 75
IV.4.2 Échantillons non appariés 75
IV.4.3 Échantillons appariés 79
IV.5 Résumé des tests 80
IV.5.1 Test d'adéquation à une loi discrète : le test du χ^2 80
IV.5.2 Test d'indépendance entre deux variables qualitatives 81

IV.5.3 Régression logistique 81
IV.5.4 Comparaison de deux proportions sur échantillons non appariés . . . 82
IV.5.5 Comparaison de deux proportions sur échantillons appariés 83

V Tests non paramétriques 85
V.1 Pourquoi la statistique non paramétrique ? 85
V.1.1 Se libérer du modèle paramétrique 85
V.1.2 Quand faut-il utiliser du non paramétrique ? 85
V.1.3 Exemples 86
V.2 Les outils statistiques du non paramétrique 88
V.3 Problèmes à un seul échantillon 90
V.3.1 Ajustement à une loi : le test de Kolmogorov 90
V.3.2 Un exemple 92
V.3.3 Test de normalité 93
V.4 Problèmes à deux échantillons 93
V.4.1 Que signifie échantillon apparié, non apparié ? 93
V.4.2 Deux échantillons appariés : le test de Wilcoxon 94
V.4.3 Deux échantillons non appariés 97
V.4.4 Tests paramétriques ou tests non paramétriques ? 102
V.5 Conclusions 102
V.6 Annexe 102
V.7 Résumé 103
V.7.1 Test de Kolmogorov 103
V.7.2 Test de Wilcoxon 104
V.7.3 Test de Kolmogorov-Smirnov 104
V.7.4 Test de Mann et Whitney 105

VI Modélisation statistique des valeurs extrêmes 107
VI.1 Introduction 107
VI.2 Modèles de valeurs extrêmes 110
VI.2.1 La loi généralisée des extrêmes (la loi du maximum par blocs) 111
VI.2.2 La loi des dépassements (modèle POT) 114
VI.3 Inférence 117
VI.3.1 Inférence du modèle GEV 117
VI.3.2 Inférence du modèle POT 122
VI.4 Traitement décisionnel de la construction d'une digue 123
VI.5 Conclusions et perspectives 125
VI.6 Sur la période de retour et la loi de Poisson 125
VI.6.1 Période de retour 125
VI.6.2 La loi de Poisson 126

VII Séries chronologiques 131
VII.1 Introduction 131
VII.1.1 Motivations et objectifs 131
VII.1.2 Exemples de séries temporelles 132
VII.1.3 Repères historiques 134

VII.1.4 Principaux modèles statistiques pour l'étude des séries temporelles . 135
VII.2 Processus univariés à temps discret 137
VII.2.1 Processus stochastique du second ordre 137
VII.2.2 Processus stationnaire 138
VII.2.3 Autocovariance et autocorrélations 139
VII.2.4 Estimation des moments pour les processus stationnaires 141
VII.2.5 Tests de blancheur 142
VII.2.6 Densité spectrale 143
VII.3 Décomposition saisonnière 145
VII.3.1 Principe de la décomposition saisonnière 145
VII.3.2 Décomposition saisonnière à l'aide de la régression linéaire 145
VII.4 Prévision et processus des innovations 148
VII.5 Étude des processus AR 149
VII.5.1 Définition 149
VII.5.2 Processus AR canonique et écriture $MA(\infty)$ 150
VII.5.3 Autocorrélations simples d'un processus AR 151
VII.5.4 Autocorrélations partielle d'un processus AR 152
VII.5.5 Exemples 152
VII.6 Processus MA 153
VII.6.1 Processus MA canonique et écriture $AR(\infty)$ 154
VII.6.2 Autocorrélations simples d'un processus MA 154
VII.6.3 Autocorrélations partielles d'un processus MA 155
VII.7 Processus ARMA 156
VII.7.1 Processus ARMA canonique et écritures $MA(\infty)$ et $AR(\infty)$ 157
VII.7.2 Autocorrélations d'un processus ARMA 158
VII.7.3 Densité spectrale d'un processus ARMA 159
VII.7.4 Estimation des processus ARMA 159
VII.7.5 Choix de modèle 162
VII.8 Pratique des modèles SARIMA 163
VII.8.1 Méthodologie 163
VII.8.2 Exemple 166

VIII Apprentissage Statistique 173
VIII.1 Introduction 173
VIII.2 Description formelle et exemples 173
VIII.2.1 Problématique 173
VIII.2.2 Exemples 174
VIII.2.3 Lien entre classement binaire et régression aux moindres carrés . . . 176
VIII.2.4 Consistance universelle 177
VIII.3 Les algorithmes d'apprentissage et leur consistance 177
VIII.3.1 Algorithmes par moyennage locale 177
VIII.3.2 Algorithmes par minimisation du risque empirique 180
VIII.4 Au delà de la consistance universelle 186

C Analyse exploratoire 193

IX Analyse des données **195**
IX.1 Introduction . . . 195
IX.1.1 Objectif . . . 195
IX.1.2 Notations . . . 196
IX.1.3 Exemple . . . 196
IX.2 L'Analyse en Composantes Principales . . . 201
IX.2.1 Problématique . . . 201
IX.2.2 Choix de la métrique . . . 201
IX.2.3 Moindre déformation du nuage . . . 202
IX.2.4 Principales relations entre variables . . . 207
IX.2.5 Dualité : imbrication des deux objectifs . . . 208
IX.2.6 Nombre d'axes (ou de composantes) à analyser . . . 209
IX.2.7 Eléments supplémentaires . . . 209
IX.2.8 Aides à l'interprétation . . . 210
IX.3 Classification automatique . . . 211
IX.3.1 Problématique . . . 211
IX.3.2 Mesure de dissimilarité . . . 211
IX.3.3 Inerties . . . 212
IX.3.4 Algorithmes de partitionnement : les "centres mobiles" . . . 213
IX.3.5 Classification ascendante hiérarchique (CAH) . . . 214
IX.3.6 Méthodes mixtes . . . 216
IX.3.7 Caractérisation des classes . . . 216
IX.4 Complémentarité des deux techniques . . . 217
IX.5 Limites . . . 219

D Annexe 221

X Rappels et compléments de probabilités **223**
X.1 Définitions et rappels généraux . . . 223
X.1.1 Probabilité et loi . . . 223
X.1.2 Variables aléatoires . . . 227
X.1.3 Espérance-Variance . . . 233
X.1.4 Conditionnement, indépendance . . . 236
X.2 Lois de variables aléatoires remarquables . . . 242
X.2.1 Variables aléatoires discrètes . . . 242
X.2.2 Variables aléatoires absolument continues . . . 244
X.3 Théorèmes limites . . . 249
X.3.1 Convergences . . . 249
X.3.2 Théorème central limite . . . 250
X.3.3 Convergences en loi remarquables . . . 251

XI Tables statistiques **253**
XI.1 Quantiles de la loi $\mathcal{N}(0,1)$ 253
XI.2 Fonction de répartition de la loi $\mathcal{N}(0,1)$ 254
XI.3 Quantiles de la loi du χ^2 255
XI.4 Quantiles de la loi de Student 256
XI.5 Quantiles de la loi de Fisher-Snedecor 257

Bibliographie et logiciels **259**

Première partie

Introduction

Chapitre I

Comment caractériser l'activité du statisticien ?

I.1 Le contexte

a. Le statisticien n'invente pas son domaine d'investigation, mais se trouve confronté à des données, qui peuvent être de plusieurs natures :

- données brutes telles que des enregistrements de qualité de pièces produites, de phénomènes météorologiques, d'observations démographiques, de cours de bourse ...
- données issues de protocoles expérimentaux, tels que des expériences biologiques, agronomiques ...

Dans tous les cas, de manière plus ou moins explicite selon la connaissance des conditions de recueil, et même si ces données sont très vastes, on peut considérer qu'il ne s'agit que d'une **connaissance imparfaite d'une réalité sous-jacente** (en ceci la statistique ne diffère pas d'autres activités scientifiques expérimentales).

b. Le statisticien n'invente pas sa problématique, mais il a un interlocuteur, qui lui exprime, de manière plus ou moins confuse, des attentes face à ces données :

- **Clarifier les données** pour en extraire des résumés pertinents.
- **Modéliser la part d'aléa qui intervient dans le phénomène** qui a conduit à ces données ; *insistons sur le fait qu'il ne s'agit pas ici de modéliser la "structure" (physique, biologique, économique ...) de ce phénomène.* **Dans la plupart des cas cette modélisation prendra la forme d'hypothèses sur une loi de probabilité régissant le phénomène qui a conduit aux observations** : le modèle prend en compte une plus ou moins vaste famille de lois de probabilité et le statisticien doit, au vu des données, **dégager une sous-famille** dans laquelle il conseillera de considérer que se situe la loi du phénomène ; le type de sous-famille à mettre en évidence dépend des besoins exprimés par l'interlocuteur, et le "conseil" du statisticien ne pourra jamais être considéré comme prémuni contre tout risque d'erreur.
- **Prendre une décision**, celle-ci pouvant être opérationnelle (interruption d'une production défectueuse, acceptation d'un médicament, publication d'un sondage ...) ou de nature scientifique, typiquement le choix d'une certaine modélisation (le mot étant pris ici dans un sens très large, et pas seulement statistique) dont on fera ultérieurement usage si se représente un contexte analogue.

– Effectuer une certaine **prévision de l'avenir** si se poursuit le phénomène enregistré (ceci étant un cas particulier du point précédent dans le cas de séries chronologiques, mis en évidence en raison de son caractère important pratiquement).

Bien sûr les distinctions faites ci-dessus ne sont pas aussi tranchées dans la pratique ; donnons quelques exemples :

– Le statisticien peut avoir participé à l'élaboration du protocole de recueil des données : une meilleure éducation à la statistique de la population, en particulier chez les chercheurs, les ingénieurs, les financiers ... doit en particulier conduire à inciter les utilisateurs à associer le plus en amont possible le statisticien à l'élaboration des conditions de recueil des données sur lesquelles il doit travailler.
– Le statisticien doit respecter les besoins de son interlocuteur, mais il peut l'aider à les formaliser et le renseigner sur la possibilité de les satisfaire ou non au vu des données disponibles ; en particulier les demandes de son interlocuteur ont souvent une présentation qualitative (qui sera marquée par des guillements dans les exemples donnés ci-dessous) ; le statisticien doit lui rendre sensible la nécessité de quantifier le contexte de travail pour le rendre sujet à un traitement scientifique.
– Le choix des résumés "intéressants" des données est souvent lié à la fois au questionnement de l'interlocuteur et au choix du modèle.
– Enfin, évidemment, le mot *interlocuteur* est à prendre dans un sens très large, celui-ci pouvant être effectivement une personne physique mais aussi un concept abstrait tel qu'une théorie scientifique soumise au feu de l'expérimentation et dont v le statisticien doit assimiler les problématiques propres.

Il n'en reste pas moins que cette présentation du travail du statisticien permet (même si les personnalités à casquettes multiples sont souvent les plus productives) de distinguer son métier de ceux :

– du **mathématicien** ; mais la caractérisation des modèles et des qualités des outils utilisés pour résoudre les questions dégagées passe par l'intermédiaire des mathématiques et de l'idéalisation qu'elles permettent : on verra par exemple l'importance des théorèmes asymptotiques du calcul des probabilités qui, en renseignant sur "ce qui se passe" quand la taille d'un échantillon tend vers l'infini, justifient les techniques employées pour des échantillons finis (ils le sont par essence même) mais "de taille assez grande" ;
– de l'**informaticien** ; mais la coopération est indispensable pour élaborer les outils de calcul dégageant les caractéristiques pertinentes des grandes masses de données (cas de beaucoup d'observations appartenant chacune à un espace de taille assez importante) ; il y a lieu ici de se méfier de la mise sur le marché d'outils purement informatiques prétendant au traitement automatique de données en faisant l'économie d'hypothèses modélisatrices des phénomènes sous-jacents, ce que recouvre parfois le vocable nouveau et très en vogue de *fouille des données* (en anglais *data mining*) ;
– du **chercheur impliqué dans un domaine d'application donné** ; mais chaque branche scientifique a suscité des problématiques statistiques qui lui sont particulièrement utiles (et qui ont souvent migré ensuite vers d'autres branches) et a donc développé des corps de spécialistes dont la qualité tient à leur double compétence : économètres, biomètres, démographes, fiabilistes industriels ...

I.2 La démarche

Les considérations ci-dessus justifient le plan des chapitres qui suivent. On procédera à la considération des points suivants :
a. formalisation d'un modèle,
b. réflexion sur les questionnements envisageables dans ce modèle,
c. élaboration d'outils associés aux questionnements dégagés et exploration des qualités qu'on peut en attendre.

Ces différents points se retrouveront à la fois dans la pratique de la résolution d'un problème concret de statistique, une fois son contexte bien analysé, et dans les exercices qui pourront être proposés à des étudiants dans le cadre de ce cours.

Dans le chapitre II, nous nous efforcerons aussi de dégager l'environnement de telles études, ce qui peut comporter :
a. une réflexion sur l'origine de ces données (*nous pourrons par exemple fournir ici une liste non limitative de conditions de recueil envisageables ; d'autres peuvent être imaginées en cours*),
b. une réflexion sur les résumés de ces données qui paraissent pertinents,
c. une interrogation sur la possibilité de prendre en compte des informations extérieures aux données.

I.3 Le modèle

Comme dans toute activité scientifique, l'élément central dans cette démarche est celui de la **formalisation d'un modèle**. Nous devons insister ici sur cinq points :

a. La première étape d'une modélisation en statistique consiste en le **choix de l'ensemble de toutes les réalisations possibles du phénomène étudié** (ensemble souvent noté Ω) ; l'observation effective dont dispose le statisticien est donc vue comme un élément ω de ce Ω, ensemble qui peut être "très gros" et dont le choix fait intervenir à la fois des considérations "réalistes" et des considérations mathématiques : par exemple pour 1000 observations prenant des valeurs entières entre 0 et 120 (âges d'êtres humains en années) on poura prendre $\Omega = \{0, \ldots, 120\}^{1000}$, mais il pourra être utile pour le traitement mathématique de "plonger" cet espace dans $\mathbb{R}^{1000}$.

b. Une fois adopté un tel modèle, les résumés des données auxquels on s'intéresse se formalisent comme des applications de Ω, dans un autre espace Ω'. De telles applications sont aussi appelées des **statistiques**. Le choix des résumés jugés pertinents peut être fondé sur des considérations propres au ω observé (on est alors, comme ce sera le cas dans le cadre de l'*analyse des données*, jadis appelée *statistique descriptive*) ou sur des considérations portant sur le phénomène qui a donné naissance à ce ω (on est alors, comme ce sera le cas en II.1.1 et III.3.1, dans le cadre de la *statistique inférentielle*, qu'on aborde au point **c** qui suit).

c. Dans le cadre de la statistique inférentielle, le caractère fortuit de l'observation effective, parmi toutes les réalisations auxquelles aurait pu donner naissance le phénomène observé (et dont l'ensemble a été modélisé par Ω), se traduit mathématiquement en complétant le modèle par le **choix d'une famille de probabilités** ; en ce sens, adopter un modèle statistique consiste à admettre (au moins jusqu'à révision de ce point de vue) que la "vraie" probabilité qui régit le phénomène appartient à cette famille, classiquement notée $\mathcal{P} = \{P_\theta ; \theta \in \Theta\}$;

on parlera donc du **modèle statistique** $(\Omega, \mathcal{P})$. Il est souvent possible de donner un sens concret, dans le modèle, à ce **paramètre** θ (*sinon, il est toujours possible mathématiquement de prendre pour ensemble* Θ *l'ensemble de celles des probabilités sur* Ω *que l'on considère susceptibles de régir le phénomène et d'écrire que* $P_\theta = \theta$*).*

Pour des raisons d'ordre mathématique (voir X.1.1), il peut être nécessaire de préciser la tribu $\mathcal{F}$ de parties de Ω (appelées **évènements** dans ce modèle) sur laquelle sont définies les probabilités P_θ (mais nous omettrons la référence à $\mathcal{F}$ quand cela ne prêtera pas à confusion). Dans le cadre d'un tel modèle, les statistiques (voir **b.** ci-dessus) ont la structure mathématique de **variables aléatoires** (v.a., voir X.1.1) de $(\Omega, \mathcal{F})$ dans un autre espace $(\Omega', \mathcal{F}')$.

d. La succession des points évoqués en I.2 ci-dessus ne doit pas être considérée comme rigide. Par exemple la formalisation du problème permet de guider des choix de résumés, ou bien les questionnements ont été présentés par l'interlocuteur du statisticien préalablement au choix du modèle (mais il importe de les interpréter dans ce cadre).

e. Un modèle est certes un "carcan" dans lequel on s'enferme pour pouvoir mener une étude rigoureuse (en particulier mathématique) afin d'affiner, au vu des données, notre connaissance de là où se situent, dans le cadre de ce modèle, les caractéristiques essentielles des observations ou bien la probabilité qui régit le phénomène (autrement dit, en général, où se situe la vraie valeur du paramètre). **Ceci ne doit pas bannir tout esprit critique sur le choix du modèle**, et, en avançant dans l'étude, on peut mettre en évidence des anomalies qui conduisent à reprendre le travail à partir d'un modèle différent.

Deuxième partie

Statistique décisionnelle

Chapitre II

Rappels sur le modèle paramétrique

II.1 Un exemple introductif

II.1.1 Présentation d'un jeu de données

L'interlocuteur du statisticien est ici un industriel, responsable d'une machine qui produit des pièces classées soit "bonnes" (ce qui est noté 0), soit "défectueuses" (ce qui est noté 1). Il a observé un "lot" de n pièces ; voici la suite des observations, pour $n = 100$:

00010 10100 00000 00000 00000 01000 00100 00100 01000 00010
10100 01000 00000 10100 00001 00101 00100 11010 00101 00000

On notera $x = (x_1, \ldots, x_{100})$ la suite de 0 et de 1 observée. Le modèle le plus simple que l'on puisse proposer ici consiste à supposer que chaque $x_i \in \mathcal{X} = \{0, 1\}$ est la réalisation d'une variable aléatoire (v.a.) X_i de loi de Bernoulli (voir X.2.1), ces v.a. étant supposées **indépendantes et identiquement distribuées (i.i.d.)** *(on a de la chance : les initiales sont les mêmes en français et en anglais !)*. L'ensemble des lois pour X_i est l'ensemble des lois de Bernoulli : $\mathcal{P} = \{\mathcal{B}(p), p \in [0, 1]\}$. Bien sûr, quand on fera l'inférence à partir des données, la vraie valeur du paramètre sera inconnue. L'espace des observations est $\mathcal{X}^n = \{0, 1\}^n$ (avec $n = 100$), et on a, pour $x \in \mathcal{X}^n$,

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = p^y (1 - p)^{n-y},$$

où $y = \sum_{i=1}^{n} x_i$ est le nombre de pièces "défectueuses".

Il est important de justifier le modèle et de s'assurer qu'il traduit bien les conditions d'obtention des observations.

II.1.2 Remarques sur la modélisation

Pour comprendre comment on peut justifier un tel modèle, remarquons que, comme dans la plupart des problèmes statistiques, l'information sur les conditions de recueil des observations débouche immédiatement sur le questionnement suivant : **Y a-t-il lieu de considérer qu'est intervenue une part d'aléatoire ? si oui, où se situe cette intervention du hasard ?**

Très grossièrement, on peut considérer que l'aléatoire peut en général avoir deux types de provenance :

A. Variabilité intrinsèque du phénomène étudié. On peut imaginer, ici, que de "petites variations" dans la fabrication (par exemple dans la position de la pièce dans la machine qui la façonne) influent sur la qualité ; ces variations ne sont ni maîtrisables ni descriptibles en détail, et seule la notion de "probabilité de sortie d'une pièce défectueuse" peut en rendre compte ; les paramètres de réglage de la machine pouvant évoluer au cours du fonctionnement, il peut être nécessaire de ne pas considérer cette probabilité comme constante ; une information, même imparfaite, sur cette probabilité peut déboucher sur des décisions de réglage de la machine.

B. Échantillonnage. Ici on ne se situe plus pendant le phénomène, mais après celui-ci : dans notre exemple, la masse de la production durant une certaine période est considérée dans son ensemble et on s'intéresse à la proportion de pièces défectueuses dans cette production ; un échantillon est tiré "au hasard" dans cette population et on observe les qualités des pièces ainsi extraites ; en déduire une information, même imparfaite, sur la proportion de pièces défectueuses dans la production totale peut déboucher sur des décisions consistant à refuser de mettre la production en vente telle quelle ou à arrêter la production ; nous reviendrons en II.1.3 ci-dessous sur de telles considérations de "décision".

Plus précisément, dans notre exemple, le modèle i.i.d. proposé peut se justifier dans les conditions expérimentales suivantes, qui correspondent chacune à l'un des deux types d'aléatoire que nous venons d'évoquer ; **le lecteur remarquera que l'interprétation du paramètre p de la loi de Bernoulli est différente dans les deux cas, mais que celui-ci n'est jamais connu du statisticien.**

A. *On a pris chronologiquement 100 pièces produites pendant un certain laps de temps* ; **on admet** que la production a été stable durant cette période, cette stabilité étant caractérisée par la constance de la probabilité p pour chaque pièce produite d'être défectueuses ; on admet aussi que les petites variations aléatoires pouvant influer sur la qualité des pièces ne se répercutent pas d'une pièce à celles qui suivent ; le fait d'introduire des "trous" entre les interventions sur la chaîne pour extraire les pièces à observer peut favoriser la satisfaction de l'exigence d'indépendance (voir X.1.4), mais, en allongeant le temps global de recueil, il peut détériorer l'exigence de conservation de la loi. Le paramètre p a ici un caractère théorique qui explique qu'il ne puisse pas être directement observable.

B. *On dispose d'une masse de N pièces produites*, dans laquelle p est la proportion de pièces défectueuses et **on admet** que, dans la procédure de tirage de l'échantillon, le "mélange" des pièces a été bien effectué : ou bien chaque pièce tirée est remise dans la masse (qui est à nouveau "bien mélangée"), ou bien elle n'est pas remise mais alors N est suffisamment grand devant 100 pour qu'on puisse approximativement considérer que les 100 observations sont indépendantes (en effet, en toute rigueur, chaque fois qu'on tire une pièce, la proportion de pièces défectueuses dans la population restante est influencée par la qualité de la pièce tirée, mais cette influence est très faible si N est très grand). Le paramètre p ici a une situation concrète et serait atteignable si on voulait faire un recensement complet de la production considérée ; mais le fait d'extraire un échantillon est justement dû à la volonté de ne pas faire un tel recensement (trop long, ou trop coûteux, ou parfois même irréalisable si il se trouve que toute observation détériore la pièce). Il est clair que, dans cette situation **B**, la masse de pièces produites ayant été brassée, on a perdu toute possibilité de s'interroger sur la stabilité de la production au cours de la période de recueil.

Ces considérations montrent que l'hypothèse d'i.i.d., pourtant fort simplificatrice mathé-

matiquement, est souvent sujette à caution dans les applications. Par exemple elle ne serait sans doute pas utilisable dans les situations expérimentales (réalistes) suivantes :

C. 100 pièces extraites sans replacement d'une masse de 500 pièces ;

D. 100 pièces extraites "au hasard" à des instants répartis sur un long laps de temps ;

E. succession de m "sous-échantillons" de taille k chacun (avec $k.m = 100$), extraits chacun d'une certaine masse de pièces produites dans un laps de temps où on peut considérer la production comme "stable" (par exemple $m = 2$, $k = 50$, ou bien $m = k = 10$).

II.1.3 Quels questionnements peut-on soulever ?

On considère le modèle des v.a. i.i.d. introduit dans II.1.1.

a. Le statisticien peut avoir à proposer à son interlocuteur une valeur pour p, qu'on appellera l'**estimation ponctuelle** de p. L'estimation ponctuelle sera présentée au paragraphe II.3 Mais une telle information paraît souvent un peu sommaire pour être vraiment opérationnelle pour l'industriel ; il sait en effet que, partant d'une réalité "floue" car aléatoire, il ne peut pas espérer une réponse exacte ; il aura donc souvent besoin d'avoir un ordre de grandeur de l'imprécision de la réponse.

b. Cet ordre de grandeur est fourni par **un intervalle de confiance** (familièrement une "fourchette") sur p : $[p^-, p^+]$. Il faut alors considérer (sans que, ici encore, ce puisse être une certitude) que cet intervalle, calculé à partir des observations, contient la valeur inconnue de p. La construction des intervalles de confiance sera traitée au paragraphe II.5. Cet intervalle de confiance peut servir à étayer une *décision* (par exemple arrêter la production si l'intervalle de confiance a une intersection non vide avec l'intervalle $[p_0, 1]$ des valeurs de p jugées inadmissibles).

c. Les procédures de **décisions** dépendent des questions posées. Le statisticien peut avoir à expliciter en quoi consiste le renseignement le plus utile pour l'interlocuteur (ici l'industriel) en vue d'une décision que ce dernier a à prendre (mettre en vente ou non la production, réparer ou non la machine ...) ; par exemple, l'industriel, pour des raisons de qualité, devrait arrêter la production s'il apparaissait que la probabilité p de produire des pièces défectueuses était montée au dessus d'un certain seuil p_0 (par exemple 0.2) ; mais un arrêt coûte cher, et il ne le fera que si le statisticien arrive à le convaincre de la nécessité d'y procéder ; tout ce qui intéresse l'industriel, au vu des données, c'est donc de savoir s'il doit considérer que $p \leq p_0$ (et continuer à produire) ou que $p > p_0$ (et se résoudre à arrêter). Une technique pour répondre à cette question sera appelée **test de l'hypothèse nulle** $H_0 = \{p \leq p_0\}$ **contre l'hypothèse alternative (ou contre-hypothèse)** $H_1 = \{p > p_0\}$. Il s'agit donc d'associer à la partition (H_0, H_1), donnée dans l'espace des paramètres $\Theta = [0, 1]$, une partition $(\mathcal{X}_0, \mathcal{X}_1)$, **fabriquée dans l'espace des observations** $\mathcal{X}^n = \{0, 1\}^n$, de telle sorte que l'on conclue en faveur de H_0 si $x \in \mathcal{X}_0$ et en faveur de H_1 si $x \in \mathcal{X}_1$. L'ensemble $\mathcal{X}_1$, où on rejette H_0, est dite **région critique du test** (ou zone de rejet). Les procédures de test seront traitées au paragraphe II.6.

Remarque II.1. De même, dans le cadre du modèle E (avec 2 tranches de taille 50) on peut, par exemple :

- chercher à proposer une estimation pour les paramètres p_a et pour p_b, correspondant chacun à un des deux échantillons ;
- chercher à élaborer deux intervalles de confiance, respectivement pour p_a et pour p_b.

– s'interroger sur l'apparition d'une dégradation entre la première période et la seconde ; tout ce qui intéresse l'industriel, au vu des données, c'est de savoir s'il doit considérer que $p_b \leq p_a$ (et être serein car il n'y aurait pas de dégradation) ou que $p_b > p_a$ (et prendre des mesures car il y aurait dégradation) ;

$\Diamond$

II.2 Définitions et rappels

II.2.1 Les modèles paramétriques

On dit que la suite de variables aléatoires (v.a.) $X = (X_1, \ldots, X_n)$ forme un **échantillon** (ou n-échantillon si on veut insister sur la taille de l'échantillon) si les v.a. $X_1, \ldots, X_n$ sont **indépendantes et identiquement distribuées** (**i.i.d.**) On note P la loi de X_i, et par convention la loi de $X = (X_1, \ldots, X_n)$ est notée $\mathrm{P}^{\otimes n}$. Il s'agit de la **loi produit**. On suppose que la loi inconnue P appartient à une famille de probabilités $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$, où θ est un paramètre d-dimensionnel et Θ un sous-ensemble de $\mathbb{R}^d$. On dit que le modèle est **paramétrique**. Par exemple les familles $\mathcal{P} = \{\mathcal{B}(p); p \in [0,1]\}$, où $\mathcal{B}(p)$ est la loi de Bernoulli (cf. paragraphe II.1.1), et $\mathcal{P} = \left\{\mathcal{N}(\mu, \sigma^2); \mu \in \mathbb{R}, \sigma > 0\right\}$, où $\mathcal{N}(\mu, \sigma^2)$ est la loi gaussienne de moyenne μ et de variance σ^2, correspondent à des modèles paramétriques. Dans le deuxième cas le paramètre $\theta = (\mu, \sigma)$ est bi-dimensionnel.

Nous étudierons dans ce chapitre des modèles paramétriques **dominés** , c'est-à-dire qu'il existe une mesure σ-finie, ν, par rapport à laquelle toute les probabilités de $\mathcal{P}$ sont absolument continues (i.e. possèdent une densité). On note $p(\cdot; \theta)$ la densité de P_θ par rapport à ν. En particulier la densité de l'échantillon $X = (X_1, \ldots, X_n)$, où X_i, à valeur dans $\mathcal{X}$, est de loi P_θ, est donnée par la densité produit :

$$p_n(x; \theta) = p(x_1; \theta) \cdots p(x_n; \theta), \quad \text{où} \quad x = (x_1, \ldots, x_n) \in \mathcal{X}^n.$$

Cette définition très générale recouvre les deux cas particuliers que nous considérerons essentiellement, à savoir :

– Les lois discrètes où la densité de l'échantillon est la probabilité d'observer $(x_1, \ldots, x_n)$ (ici la mesure ν est la mesure de comptage sur $\mathcal{X}$). Plus particulièrement, pour le modèle de Bernoulli, $\mathcal{P} = \{\mathcal{B}(p); p \in [0,1]\}$, on a

$$p_n(x; p) = p^y (1-p)^{n-y}, \quad \text{où} \quad y = \sum_{i=1}^n x_i \quad \text{et} \quad x = (x_1, \ldots, x_n) \in \{0,1\}^n.$$

– Les lois à densité où la densité de l'échantillon correspond à la densité usuelle (ici la mesure ν est la mesure de Lebesgue sur $\mathcal{X} = \mathbb{R}$ ou $\mathcal{X} = \mathbb{R}^n$). Plus particulièrement, pour le modèle gaussien $\mathcal{P} = \left\{\mathcal{N}(\mu, \sigma^2); \mu \in \mathbb{R}, \sigma > 0\right\}$, on a

$$p_n(x; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \, \mathrm{e}^{-\sum_{i=1}^n (x_i - \mu)^2 / 2\sigma^2}, \quad \text{où} \quad x = (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

La fonction réelle $\theta \mapsto p_n(x; \theta)$ à x fixé est appelée la **vraisemblance** (*"likelihood"* en anglais) de l'échantillon ou de la réalisation $x = (x_1, \ldots, x_n)$. (On utilise parfois la notation $L_n(x; \theta)$ pour la vraisemblance.) Elle est définie sur Θ. Cette fonction est d'une grande importance pour les modèles paramétriques, comme nous le verrons par la suite.

Une **statistique** S est une fonction de l'échantillon $X = (X_1, \ldots, X_n)$ réelle ou vectorielle. Cette fonction est indépendante de $\mathrm{P} \in \mathcal{P}$.

Exemple II.2. Les fonctions $S(X) = \sum_{i=1}^n X_i$, $S(X) = X_1$, $S(X) = (X_1 - X_2, X_3)$ sont des statistiques de l'échantillon $(X_1, \ldots, X_n)$. $\Diamond$

II.2.2 Rappels sur la LFGN et le TCL

Nous rappelons les deux théorèmes fondamentaux des probabilités. Le premier indique que pour un échantillon $X_1, \ldots, X_n$, la moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ converge vers la moyenne $\mu = \mathbb{E}[X_1]$, pourvu que cette dernière existe. Le second précise quelle est la répartition, i.e. la loi, asymptotique de la moyenne empirique autour de la moyenne (voir X.3.1 pour les différentes notions de convergence de suites de v.a.).

Théorème II.3 (Loi forte des grands nombres (LFGN))**.** *Soit $(X_n, n \in \mathbb{N}^*)$ une suite de v.a. réelles ou vectorielles* ***indépendantes, et identiquement distribuées*** *(i.i.d.) et* ***intégrables*** *($\mathbb{E}[|X_n|] < \infty$). Alors la moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$ converge* ***presque sûrement*** *(p.s.) vers la moyenne $\mu = \mathbb{E}[X_1]$. Ainsi on a*

$$\boxed{\bar{X}_n = \frac{1}{n}\sum_{k=1}^n X_k \xrightarrow[n\to\infty]{\textbf{p.s.}} \mu.}$$

Théorème II.4 (Théorème central limite (TCL))**.** *Soit $(X_n, n \in \mathbb{N}^*)$ une suite de v.a. réelles* ***indépendantes et identiquement distribuées*** *(i.i.d.). On suppose qu'elles sont de* ***carré intégrable*** *($\mathbb{E}[X_n^2] < \infty$). On pose $\mu = \mathbb{E}[X_n]$, $\sigma^2 = \mathrm{Var}(X_n)$ et la moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$. La suite de v.a. $\sqrt{n}(\bar{X}_n - \mu) = \frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n}}$ converge en loi vers la loi gaussienne $\mathcal{N}(0, \sigma^2)$:*

$$\boxed{\sqrt{n}(\bar{X}_n - \mu) = \frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n}} \xrightarrow[n\to\infty]{\textbf{en loi}} \mathcal{N}(0, \sigma^2).}$$

Ce théorème peut se généraliser au cas de v.a. vectorielles, voir le paragraphe X.3.2.

Nous utiliserons également le résultat suivant qui est une conséquence du théorème de Slutsky.

Théorème II.5. *Soit $(Y_n, n \geq 1)$ une suite de v.a. réelles qui converge en loi vers la loi gaussienne $\mathcal{N}(0, \sigma^2)$, où $\sigma^2 > 0$. Soit $(Z_n, n \geq 1)$ une suite de v.a. positives qui converge p.s. ou en loi vers la constante σ^2. Alors la suite $(Y_n/\sqrt{Z_n}, n \geq 1)$ converge en loi vers la loi gaussienne centrée réduite $\mathcal{N}(0, 1)$.*

II.3 Estimation ponctuelle : les estimateurs

On considère un modèle paramétrique $\mathcal{P} = \{\mathrm{P}_\theta; \theta \in \Theta\}$, et un échantillon $(X_1, \ldots, X_n)$, où X_i est de loi $\mathrm{P} \in \mathcal{P}$ inconnue. On notera $\mathbb{E}_\theta[f(X_1, \ldots, X_n)]$ l'espérance de $f(X_1, \ldots, X_n)$ et $\mathbb{P}_\theta(A)$ la probabilité de l'évènement A, quand $\mathrm{P} = \mathrm{P}_\theta$.

Nous avons vu dans l'exemple introductif (paragraphe II.1.3) que le premier objectif est d'estimer le paramètre inconnu. Plus généralement on peut chercher à estimer une fonction de ce paramètre $g(\theta)$.

Définition II.6. *Un **estimateur**, Z, de $g(\theta)$ est une statistique construite à partir de l'échantillon $(X_1, \ldots, X_n)$, à valeurs dans $g(\Theta)$.*

Exemple II.7. Pour le modèle de Bernoulli (cf. paragraphe II.1.1), $1/2$, X_1 et $\bar{X}_n = \frac{1}{n}\sum_{i=1} X_i$ sont des estimateurs de p. Évidemment ils ont des propriétés différentes. ◊

Dans le paragraphe qui suit nous donnons des propriétés que l'on peut attendre d'un estimateur, et qui permettent de mesurer d'une certaine manière sa qualité d'approximation de la vraie valeur inconnue du paramètre. Le paragraphe II.3.2 sera consacrée à la construction de l'estimateur du maximum de vraisemblance qui possède en général de très bonnes qualités.

II.3.1 Propriétés des estimateurs

Estimateur sans biais

Le biais est l'écart entre la moyenne de l'estimateur et ce que l'on cherche à estimer. Le biais d'un estimateur, Z, intégrable (i.e. $\mathbb{E}_\theta[|Z|] < \infty$ pour tout $\theta \in \Theta$) de $g(\theta)$ est donné par $\mathbb{E}_\theta[Z] - g(\theta)$. L'estimateur est dit **sans biais** si $\mathbb{E}_\theta[Z] = g(\theta)$ pour tout $\theta \in \Theta$.

Exemple II.8. Pour le modèle de Bernoulli, X_1 et $\bar{X}_n = \frac{1}{n}\sum_{i=1} X_i$ sont des estimateurs sans biais de p. En effet, comme les v.a. X_i sont des variables de Bernoulli de paramètre p, on a $\mathbb{E}_p[X_1] = p$ et $\mathbb{E}_p[\bar{X}_n] = p$. En revanche $1/2$ est un estimateur biaisé de p. ◊

L'absence de biais est une propriété intéressante, mais elle n'est pas conservée par transformation non affine. Ainsi dans l'exemple précédent $\bar{X}_n$ est un estimateur sans biais de p, mais $\bar{X}_n(1-\bar{X}_n)$ n'est pas un estimateur sans biais de la variance $\sigma^2 = p(1-p)$. En effet, on a $\mathbb{E}_p[\bar{X}_n(1-\bar{X}_n)] = p(1-p) - n^{-1}p(1-p)$.

Risque

On peut également mesurer la précision d'un estimateur, Z, de $g(\theta)$ en regardant son écart moyen avec $g(\theta)$. Une quantité traditionnellement utilisée est le **risque quadratique**[1] défini, si $g(\theta)$ est réel, par $R(Z;\theta) = \mathbb{E}_\theta[(Z - g(\theta))^2]$. Bien sûr cette quantité n'a de sens que si Z est de carré intégrable (i.e. $\mathbb{E}_\theta[Z^2] < \infty$ pour tout $\theta \in \Theta$).

On dit que Z_1 est un estimateur **préférable** à Z_2 si son risque quadratique est plus faible : pour tout $\theta \in \Theta$, on a

$$R(Z_1, \theta) \leq R(Z_2, \theta).$$

Nous verrons au paragraphe II.4 une méthode générale pour améliorer les estimateurs, au sens du risque quadratique.

Remarque II.9. Sauf dans des cas triviaux il n'existe pas d'estimateur préférable à tous les autres. En effet supposons que Z soit un estimateur de θ préférable à tous les autres estimateurs. Entre autres Z est préférable à $Z_{\theta_0} = \theta_0$. Donc on a

$$R(Z;\theta_0) \leq R(Z_{\theta_0};\theta_0) = 0.$$

[1] On peut considérer d'autres risques du type $\mathbb{E}_\theta[\psi(Z - g(\theta))]$, où ψ est une fonction convexe.

On en déduit donc que $Z = \theta$ $\mathbb{P}_\theta$-p.s ! Cela implique que les densités $p(\cdot;\theta)$ ont des supports disjoints pour tout θ. Ainsi dès que l'on dispose d'une seule observation, on peut en déduire le vrai paramètre p.s. Cela correspond aux cas triviaux de l'estimation ponctuelle. ◊

Exemple II.10. Pour le modèle de Bernoulli, l'estimateur $\bar{X}_n = \frac{1}{n}\sum_{i=1} X_i$ est préférable à l'estimateur X_1 pour l'estimation de p. En effet, on a

$$R(X_1; p) = \mathbb{E}_p[(X_1 - p)^2] = \mathrm{Var}_p(X_1) = p(1-p)$$

et

$$R(\bar{X}_n; p) = \mathbb{E}_p[(\bar{X}_n - p)^2] = \mathrm{Var}_p(\bar{X}_n) = p(1-p)/n.$$

◊

Estimateurs convergents

En général, on désire que l'estimation soit d'autant plus précise que la taille de l'échantillon est grande. Contrairement aux deux notions précédentes, on regarde maintenant le comportement asymptotique d'une famille d'estimateurs construits à partir d'échantillons de plus en plus grands.

Soit $(Z_n, n \geq 1)$ une suite d'estimateurs de $g(\theta)$, où Z_n est construit à partir du n-échantillon $(X_1, \ldots, X_n)$. On dit que la suite $(Z_n, n \geq 1)$, ou plus succintement Z_n, est un estimateur **fortement convergent** de $g(\theta)$, si pour tout $\theta \in \Theta$, on a $\mathbb{P}_\theta$-p.s.

$$\lim_{n\to\infty} Z_n = g(\theta).$$

Si la convergence précédente à lieu seulement en probabilité, on parle d'estimateur faiblement convergent. En fait, dans ce cours nous considérerons essentiellement que des estimateurs fortement convergent, et par souci de simplicité on dira convergent pour fortement convergent. (De nombreux auteurs utilisent le mot anglais "*consistant*" à la place de convergent.)

En général les résultats de convergence découlent de résultats du type loi forte des grands nombres. Remarquons de plus que cette propriété est conservée quand on considère une fonction de la suite d'estimateur (qui converge alors vers la fonction de la quantité estimée), pourvu que la fonction soit continue.

Exemple II.11. Pour le modèle de Bernoulli, la moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{i=1} X_i$ est un estimateur convergent de p. Ceci découle de la loi forte des grands nombres (voir théorème II.3). Par continuité de la fonction $u \mapsto u(1-u)$, on a également que l'estimateur $\bar{X}_n(1-\bar{X}_n)$ est un estimateur convergent de la variance $p(1-p)$. ◊

Normalité asymptotique

Enfin, pour des estimateurs convergents, il est intéressant de préciser la vitesse de convergence. Cela nous permettra de construire des intervalles de confiance asymptotiques (voir le paragraphe II.5).

Soit $(Z_n, n \geq 1)$ une suite d'estimateurs de $g(\theta)$, où Z_n est construit à partir du n-échantillon $(X_1, \ldots, X_n)$. On dit que la suite $(Z_n, n \geq 1)$, ou plus succintement Z_n, est un estimateur **asymptotiquement normal** de $g(\theta)$, si pour tout $\theta \in \Theta$, on a sous $\mathbb{P}_\theta$

$$\sqrt{n}(Z_n - g(\theta)) \xrightarrow[n\to\infty]{\text{en loi}} \mathcal{N}(0, \Sigma(\theta)),$$

où $\Sigma(\theta)$ est la variance de la loi limite, appelée **variance asymptotique** (ou matrice de covariance asymptotique si $g(\theta)$ est multidimensionnel).

Les résultats de normalité asymptotique découlent souvent de résultats du type théorème central limite. (Cette propriété est également conservée quand on considère une fonction de la suite d'estimateur pourvu que la fonction soit dérivable.)

Exemple II.12. Pour le modèle de Bernoulli, l'estimateur $\bar{X}_n = \frac{1}{n}\sum_{i=1} X_i$ est un estimateur asymptotiquement normal de p de variance asymptotique $\sigma(p)^2 = p(1-p)$. Ceci découle du TCL (voir théorème II.4) et du fait que $\mathrm{Var}_p(X_1) = p(1-p)$. ◊

II.3.2 L'estimateur du maximum de vraisemblance

Exemple II.13. Au cours d'une enquête, on recherche un suspect dont la taille est d'environ 1m80. Afin d'orienter rapidement les recherches doit-on interroger plutôt un suspect masculin ou féminin ?

On peut au vue des données (cf National Center for Health Statistics (1988-1994)) dire que la taille d'un homme (aux U.S.A.) a pour loi une loi gaussienne de moyenne μ_H =1m76 et d'écart type σ_H =0m073, alors que la taille d'une femme a pour loi une loi gaussienne de moyenne μ_F =1m62 et d'écart type σ_F =0m069. On note $p(x;\mu,\sigma^2)$ la densité de la loi gaussienne $\mathcal{N}(\mu,\sigma^2)$.

Si la taille du suspect est 1m80, alors on se concentre sur la recherche d'un suspect masculin. Le choix est en fait guidé par l'allure de la densité. Et plus particulièrement ayant observé la taille x_0 =1m80, il est raisonnable de supposer que le suspect est un homme car $p(x_0;\mu_H,\sigma_H^2) \geq p(x_0;\mu_F,\sigma_F^2)$. On a choisit le paramètre $\theta \in \{(\mu_H,\sigma_H^2),(\mu_F,\sigma_F^2)\}$ qui maximise la vraisemblance $\theta \mapsto p(x_0;\theta)$.

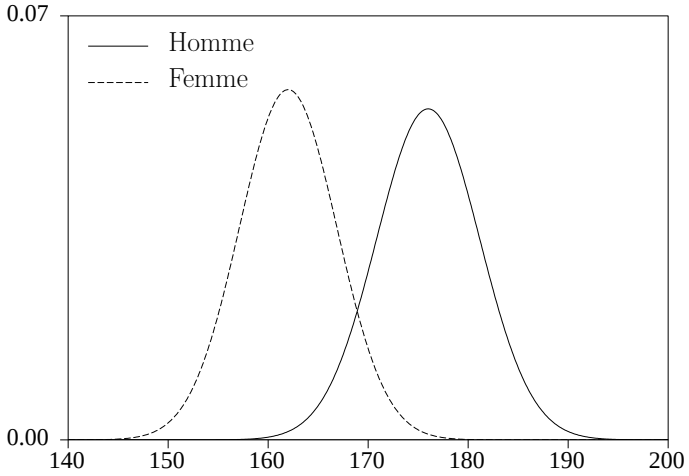


FIG. II.1 – Densités gaussiennes pour la taille d'un homme et d'une femme.

◊

Définition II.14. *Supposons que pour tout $x = (x_1, \ldots, x_n)$, il existe* ***une et une seule valeur*** *de $\theta \in \Theta$ telle que la vraisemblance soit maximale. On note cette valeur $\hat{\theta}_n(x_1, \ldots, x_n)$. La variable aléatoire $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$, à valeurs dans Θ, est appelée* ***estimateur du maximum de vraisemblance*** *(EMV) de θ.*

On définit également la **log-vraisemblance** du modèle comme le logarithme de la vraisemblance : c'est la fonction

$$\boxed{\theta \mapsto \ell_n(x;\theta) = \log(p_n(x;\theta)) = \sum_{i=1}^{n} \log(p(x_i;\theta)).}$$

Comme la fonction log est croissante, maximiser la vraisemblance ou la log-vraisemblance revient au même. Maximiser la log-vraisemblance donne lieu souvent à des calculs plus simples.

Remarque II.15. Sous des hypothèses de régularité assez générale, les EMV sont convergents et asymptotiquement normaux. (Dans la plupart des exemples du cours, on pourra vérifier directement ces deux propriétés.) En revanche les EMV sont en général biaisés. Enfin, quitte à reparamétriser la famille de loi par $g(\theta)$, quand g est une bijection, il est clair que si $\hat{\theta}_n$ est l'EMV de θ, alors $g(\hat{\theta}_n)$ est l'EMV de $g(\theta)$.

Ainsi, dans les modèles paramétriques "*réguliers*", si $\hat{\theta}_n$ est l'EMV de θ pour n observations indépendantes de même loi, associé au vrai paramètre θ_0 inconnu, alors on peut montrer que si g est une fonction régulière, $g(\hat{\theta}_n)$ est un estimateur de $g(\theta_0)$ convergent et asymptotiquement normal. Supposons que le paramètre, θ, est vectoriel de dimension $d \geq 1$. La variance asymptotique est donnée par $(\nabla g(\theta_0))' V(\theta_0) (\nabla g(\theta_0))$, où $I = V^{-1}$ est, au signe près, la matrice hessienne de la log-vraisemblance évaluée en θ_0 :

$$I = \left(-\mathbb{E}_{\theta_0} \left[\frac{\partial^2 \ell_n(X;\theta)}{\partial\theta_i \partial\theta_j} \right], 1 \leq i,j \leq d \right).$$

La matrice I est aussi appelée matrice d'information de Fisher. ◊

Exemple II.16. Pour le modèle de Bernoulli, la vraisemblance est définie par

$$p_n(x;p) = \mathbb{P}_p(X_1 = x_1, \ldots, X_n = x_n) = p^y (1-p)^{n-y},$$

où $x = (x_1, \ldots, x_n) \in \{0,1\}^n$ et $y = \sum_{i=1}^n x_i$. La log-vraisemblance est définie par

$$\ell_n(x;p) = y \log(p) + (n-y) \log(1-p).$$

On a $\partial \ell_n(x;p)/\partial p = y/p - (n-y)/(1-p)$. La dérivée de la log-vraisemblance est décroissante sur $[0,1]$ et s'annule en $p = y/n$. On en déduit donc que la log-vraisemblance atteint son maximum pour $p = \sum_{i=1}^n x_i/n$. L'estimateur du maximum de vraisemblance est donc la moyenne empirique : $\hat{p}_n = \bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. Nous avons déjà vu que cet estimateur est sans biais, convergent et asymptotiquement normal. Pour les données du paragraphe II.1.1, on a $n = 100$, $\sum_{i=1}^n x_i = 22$ et l'EMV de p est $\hat{p}_n = 0.22$. ◊

II.4 Résumé des données : les statistiques exhaustives

Exemple II.17. Pour l'exemple proposé dans le paragraphe II.1.1, il semble naturel de supposer que l'ordre dans lequel sont observées les pièces bonnes ou défectueuses n'apporte aucune information pertinente sur le paramètre inconnu, p. Et donc, on peut résumer la suite observées $(x_1, \dots, x_n)$ par le nombre total $y_n = \sum_{i=1}^n x_i$ de pièces défectueuse. Pour étayer cette intuition, on calcule la loi de l'échantillon, $(X_1, \dots, X_n)$, conditionnellement à $Y_n = \sum_{i=1}^n X_i$. On a pour $y \in \{0, \dots, n\}$ et $x_1, \dots, x_n \in \{0, 1\}$ tel que $\sum_{i=1}^n x_i = y$,

$$\mathbb{P}_p(X_1 = x_1, \dots, X_n = x_n | Y_n = y) = \frac{\mathbb{P}_p(X_1 = x_1, \dots, X_n = x_n)}{\mathbb{P}_p(Y_n = y)} = \frac{p^y(1-p)^{n-y}}{C_n^y p^y (1-p)^{n-y}} = \frac{1}{C_n^y}.$$

Si $\sum_{i=1}^n x_i \neq y$, on a $\mathbb{P}_p(X_1 = x_1, \dots, X_n = x_n | Y_n = y) = 0$. En conclusion, on en déduit que la loi de $(X_1, \dots, X_n)$ sachant $\sum_{i=1}^n X_i = y$ est la loi uniforme sur $\{(x_1, \dots, x_n) \in \{0,1\}^n; \sum_{i=1}^n x_i = y\}$. En particulier cette loi ne dépend pas du paramètre p. Ainsi toute l'information sur p de l'échantillon $(X_1, \dots, X_n)$ est contenue dans la statistique Y_n. Pour toute étude sur p, on peut résumer l'échantillon à Y_n. ◊

Définition II.18. *Une statistique S est **exhaustive** si la loi conditionnelle de l'échantillon $(X_1, \dots, X_n)$ sachant S est indépendante du paramètre θ.*

Le théorème (admis) suivant permet d'exhiber aisément des statistiques exhaustives.

Théorème II.19 (de Halmos-Savage ou théorème de factorisation)**.** *La statistique $S = S(X_1, \dots, X_n)$ est exhaustive si et seulement si la densité $p_n(x_1, \dots, x_n; \theta)$ de la loi de l'échantillon $(X_1, \dots, X_n)$ se factorise de la façon suivante : il existe des fonctions Ψ et l telles que pour tout $x = (x_1, \dots, x_n) \in \mathcal{X}^n$, $\theta \in \Theta$,*

$$\boxed{p_n(x; \theta) = \Psi(S(x), \theta) l(x).}$$

En particulier, s'il existe une statistique exhaustive donnée par le théorème de factorisation, alors l'estimateur du maximum de vraisemblance est une fonction de la statistique exhaustive.

Soit $Z = Z(X_1, \dots, X_n)$ un estimateur de $g(\theta)$. La loi de $(X_1, \dots, X_n)$ sachant une statistique exhaustive S est indépendante de θ. En particulier l'espérance conditionnelle, voir p. 238, $\mathbb{E}_\theta[Z|S] = \mathbb{E}_\theta[Z(X_1, \dots, X_n)|S]$ est une fonction de S qui ne dépend plus de θ. C'est donc un estimateur, et on le note $\mathbb{E}[Z|S]$, en supprimant l'indice θ. Le théorème suivant assure que cette procédure permet d'améliorer les estimateurs.

Théorème II.20 (de Rao-Blackwell)**.** *Soit Z un estimateur de $g(\theta)$ que l'on suppose de carré intégrable. Si S est une statistique exhaustive, alors l'estimateur $Z_S = \mathbb{E}[Z|S]$ est un estimateur préférable à Z.*

Nous retiendrons donc de ce théorème qu'il est utile de construire des estimateurs à l'aide de la statistique exhaustive.

Démonstration. On admet que si Z est de carré intégrable alors Z_S est aussi de carré intégrable. On a

$$R(Z; \theta) = \mathbb{E}_\theta[Z^2] - 2g(\theta)\mathbb{E}_\theta[Z] + g(\theta)^2,$$

et

$$R(Z_S;\theta) = \mathbb{E}_\theta[Z_S^2] - 2g(\theta)\mathbb{E}_\theta[Z_S] + g(\theta)^2.$$

Les propriétés de l'espérance conditionnelle impliquent que $\mathbb{E}_\theta[Z_S] = \mathbb{E}_\theta[Z]$ et que $\mathbb{E}_\theta[ZZ_S] = \mathbb{E}_\theta[Z_S^2]$, voir p. 238. On en déduit donc que

$$R(Z;\theta) - R(Z_S;\theta) = \mathbb{E}_\theta[Z^2] - \mathbb{E}_\theta[Z_S^2] = \mathbb{E}_\theta[(Z - Z_S)^2].$$

En particulier, $R(Z;\theta) \geq R(Z_S;\theta)$ pour tout $\theta \in \Theta$. L'estimateur Z_S est donc préférable à Z. □

Exemple II.21. Dans le modèle de Bernoulli, on a vu que X_1 était également un estimateur sans biais du paramètre p. On désire améliorer cet estimateur avec le théorème de Rao-Blackwell. On a vu dans l'exemple II.17 que $Y_n = n\bar{X}_n$ est une statistique exhaustive. Pour calculer $\mathbb{E}[X_1|Y_n]$, on rappelle que la loi conditionnelle de $(X_1, \ldots, X_n)$ sachant Y_n est la loi uniforme sur $\{(x_1, \ldots, x_n) \in \{0,1\}^n; \sum_{i=1}^n x_i = Y_n\}$. Un calcul de dénombrement élémentaire, permet de vérifier que la loi conditionnelle marginale de X_1 est la loi de Bernoulli de paramètre $Y_n/n = \bar{X}_n$. En particulier, on a $\mathbb{E}[X_1|Y_n] = \bar{X}_n$. Sur cet exemple précis, la procédure d'amélioration permet de retrouver l'EMV.

◊

II.5 Précision de l'estimation : l'intervalle de confiance

Reprenons l'exemple du paragraphe II.1.1. Il s'agit d'associer à l'observation $(x_1, \ldots, x_n)$ un intervalle dans lequel on considère que se trouve p. Ici encore les qualités de la moyenne empirique, $\bar{x}_n = n^{-1}\sum_{i=1}^n x_i$, en tant qu'estimation ponctuelle de p, incitent à prendre un intervalle $[I_n^-, I_n^+]$ auquel appartient $\bar{x}_n$. Comme toujours en contexte aléatoire, on ne peut espérer affirmer avec certitude que p sera dans cet intervalle (à moins de prendre $[0,1]$, ce qui est sans intérêt !). On choisit donc un **niveau de confiance**, proche de 1, noté $1-\alpha$, tel que la probabilité qu'il contienne le paramètre soit au moins égale au niveau de confiance : pour tout $p \in [0,1]$,

$$\mathbb{P}_p(p \in [I_n^-, I_n^+]) \geq 1 - \alpha.$$

On se trouve ici face à un dilemme : pour garantir le niveau de confiance, l'intervalle ne doit pas être trop court mais, pour être intéressant, il ne doit pas être trop long. On doit donc chercher des intervalles aussi courts que possible, au niveau de confiance $1-\alpha$ imposé, et ce uniformément en p, d'où la difficulté du problème.

On définit l'intervalle de confiance pour l'estimation d'un paramètre réel ou d'une fonction réelle d'un paramètre.

Définition II.22. *Soit $x = (x_1, \ldots, x_n) \mapsto I(x)$ une application de l'ensemble des observations à valeurs dans les intervalles.*

*On dit que I est un intervalle de confiance par **excès** pour $g(\theta)$ de niveau $1-\alpha$ si pour tout $\theta \in \Theta$, on a*

$$\mathbb{P}_\theta\left(g(\theta) \in I(X_1, \ldots, X_n)\right) \geq 1 - \alpha.$$

*On dit que l'intervalle aléatoire I est un **intervalle de confiance** de niveau $1-\alpha$ pour $g(\theta)$ si l'inégalité ci-dessus est une égalité (pour tout $\theta \in \Theta$).*

Si la taille de l'échantillon est faible, une étude détaillée du modèle permet d'exhiber les intervalles de confiance, voir l'exemple ci-dessous sur le modèle de Bernoulli. (Il existe des techniques spécifiques pour établir des intervalles de confiance, mais nous ne les détaillons pas ici.) En revanche, dans le cas où la taille de l'échantillon est grande, on peut utiliser des intervalles de confiance asymptotiques construits à partir des comportements asymptotiques des estimateurs.

Définition II.23. *Soit $x = (x_1, \dots, x_n) \mapsto I_n(x)$, où $n \in \mathbb{N}^*$, une suite d'applications à valeurs intervalles. On dit que la suite $(I_n, n \geq 1)$ est un **intervalle de confiance asymptotique** pour $g(\theta)$ de niveau asymptotique $1-\alpha$, si pour tout $\theta \in \Theta$, on a*

$$\lim_{n\to\infty} \mathbb{P}_\theta(g(\theta) \in I_n(X_1, \dots, X_n)) = 1 - \alpha.$$

Nous aurons besoin de la notion de quantile d'une loi.

Définition II.24. *Soit Y une variable aléatoire réelle de fonction de répartition $F(x) = \mathbb{P}(Y \leq x)$. Le **quantile** (on parle aussi de fractile), $q_r \in \mathbb{R}$, d'ordre $r \in]0,1[$ de la loi de Y est défini par*

$$\boxed{q_r = \inf\{x; F(x) \geq r\},}$$

et on a $F(q_r) \geq r$. Si la fonction de répartition de Y est continue, alors $F(q_r) = r$. Si la fonction de répartition est de plus strictement croissante au point q_r, alors q_r est l'unique solution de l'équation $F(x) = r$.

Exemple II.25. Suite du paragraphe II.1.1 sur le modèle de Bernoulli. Rappelons que la loi de $Y_n = n\bar{X}_n = \sum_{i=1}^n X_i$ est la loi binomiale de paramètre (n, p).

Pour construire des intervalles de confiance sur p de niveau $1-\alpha$, on peut par exemple définir $\tilde{I}_n^-(p)$ (resp. $\tilde{I}_n^+(p)$) comme étant le quantile d'ordre $\alpha/2$ (resp. $1-\alpha/2$) de la loi binomiale $\mathcal{B}(n,p)$. Ainsi, pour tout $p \in [0,1]$, on a

$$\mathbb{P}_p(Y_n \in [\tilde{I}_n^-(p), \tilde{I}_n^+(p)]) \geq 1 - \alpha.$$

Les fonctions $p \mapsto \tilde{I}_n^-(p)$ et $p \mapsto \tilde{I}_n^+(p)$ sont croissantes. En effet, soit $U_1, \dots, U_n$ des v.a. indépendantes de loi uniforme sur $[0,1]$. Pour $p \in [0,1]$, la v.a. $\mathbf{1}_{\{U_i \leq p\}}$ est une v.a. de Bernoulli de paramètre p. Ainsi $\sum_{i=1}^n \mathbf{1}_{\{U_i \leq p\}}$ est une variable aléatoire de loi binomiale de paramètre (n,p). Ainsi si $p \geq p'$, on a $\mathbf{1}_{\sum_{i=1}^n \{U_i \leq p\}} \geq \sum_{i=1}^n \mathbf{1}_{\{U_i \leq p'\}}$, et on en déduit

$$F_{\mathcal{B}(n,p)}(x) = \mathbb{P}(\sum_{i=1}^n \mathbf{1}_{\{U_i \leq p\}} \leq x) \leq \mathbb{P}(\sum_{i=1}^n \mathbf{1}_{\{U_i \leq p'\}} \leq x) = F_{\mathcal{B}(n,p')}(x). \tag{II.1}$$

Ainsi la fonction de répartition de la loi binomiale de paramètre (n,p), $F_{\mathcal{B}(n,p)}$, minore celle de la loi binomiale de paramètre (n,p'), $F_{\mathcal{B}(n,p')}$. On déduit alors que pour tout $r \in [0,1]$, le quantile d'ordre r de la loi binomiale de paramètre (n,p) est supérieur au quantile d'ordre r de la loi binomiale de paramètre (n,p'). (On dit que $\mathcal{B}(n,p)$ majore $\mathcal{B}(n,p')$ pour l'ordre stochastique, voir aussi p. 228.) On a ainsi vérifié que les fonctions $\tilde{I}_n^-$ et $\tilde{I}_n^+$ sont croissantes en p (voir le graphique II.2).

Il est maintenant facile de construire un intervalle de confiance (par excès) pour p de niveau α. En effet, on déduit de la croissance des fonctions $\tilde{I}_n^-$ et $\tilde{I}_n^+$, que

$$\bar{X}_n = \frac{1}{n} Y_n \in [\frac{1}{n} \tilde{I}_n^-(p), \frac{1}{n} \tilde{I}_n^+(p)] \Longleftrightarrow p \in [I_n^-(\bar{X}_n), I_n^+(\bar{X}_n)],$$

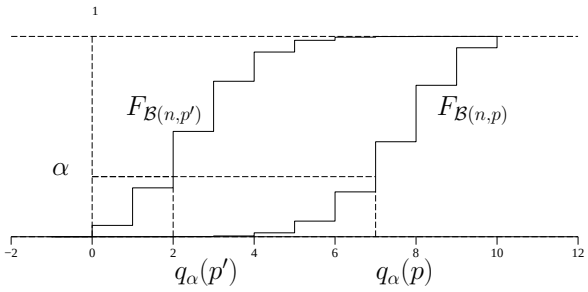


FIG. II.2 – Fonctions répartitions de $\mathcal{B}(n,p)$ et $\mathcal{B}(n,p')$, avec $n=10$ et $p \geq p'$, ainsi que les quantiles d'ordre α correspondants : $q_\alpha(p)$ et $q_\alpha(p')$.

où I_n^- et I_n^+ sont les inverses des fonctions $n^{-1}\tilde{I}_n^+$ et $n^{-1}\tilde{I}_n^-$ (au sens où $I_n^-(z) \leq p \Leftrightarrow z \leq n^{-1}\tilde{I}_n^+(p)$ et $I_n^+(z) \geq p \Leftrightarrow z \geq n^{-1}\tilde{I}_n^-(p)$). Donc pour tout $p \in [0,1]$, on a

$$\mathbb{P}_p(p \in [I_n^-(\bar{X}_n), I_n^+(\bar{X}_n)]) \geq 1-\alpha.$$

L'intervalle aléatoire $[I_n^-(\bar{X}_n), I_n^+(\bar{X}_n)]$ est donc un intervalle de confiance par excès de niveau $1-\alpha$. (Le fait que l'intervalle de confiance soit par excès est dû au fait que $\bar{X}_n$ est une variable discrète.) Le calcul numérique des courbes I^+ et I^- repose simplement sur le calcul des quantiles de la loi binomiale. Les logiciels scientifiques proposent des algorithmes pour le calcul des quantiles. Cela permet de construire des abaques qui ont l'avantage d'apporter une information visuelle (voir le graphique II.3).

En conclusion, si l'on observe les données $(x_1, \ldots, x_n)$ qui proviennent de la réalisation d'un échantillon de Bernoulli, on estime le paramètre inconnu p par $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ et on fournit l'intervalle de confiance (par excès) de niveau $1-\alpha$:

$$\boxed{I_1 = [I_n^-(\bar{x}_n), I_n^+(\bar{x}_n)].} \tag{II.2}$$

Bien que les intervalles ci-dessus soient numériquement calculables pour tout n, il est intéressant d'utiliser une approche asymptotique si n est grand. En particulier, cela permet d'avoir des ordres de grandeurs et de se forger une intuition. Pour cela, on utilise le fait que $\bar{X}_n$ est un estimateur asymptotiquement normal de p de variance asymptotique $\sigma^2 = p(1-p)$ (voir le paragraphe II.3). Remarquons que $\bar{X}_n(1-\bar{X}_n)$ est un estimateur convergent de σ^2 (c'est en fait l'EMV de $p(1-p)$). On déduit du théorème II.5, que la suite $(Z_n, n \geq 1)$, où

$$Z_n = \sqrt{n}\frac{\bar{X}_n - p}{\sqrt{\bar{X}_n(1-\bar{X}_n)}},$$

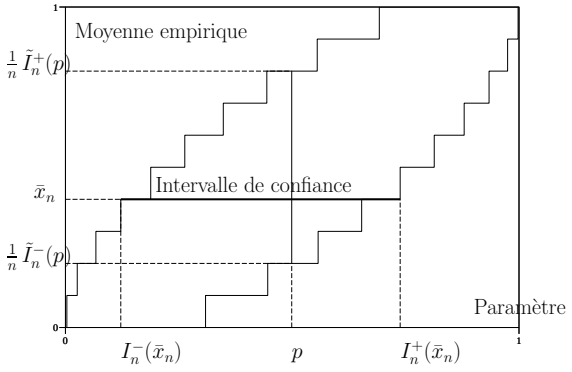


FIG. II.3 – Construction des intervalles de confiance, $[I_n^-(\bar{x}_n), I_n^+(\bar{x}_n)]$, de niveau de confiance 95% (par excès) pour la loi de Bernoulli avec $n = 10$.

converge en loi vers la loi gaussienne centrée réduite $\mathcal{N}(0,1)$. En particulier, on a pour tout $a \in \mathbb{R}$,

$$\mathbb{P}_p(Z_n \in [-a,a]) \xrightarrow[n\to\infty]{} \mathbb{P}(G \in [-a,a]),$$

où G est de loi $\mathcal{N}(0,1)$. Remarquons que

$$Z_n \in [-a,a] \Longleftrightarrow p \in \Big[\bar{X}_n \pm \frac{a\sqrt{\bar{X}_n(1-\bar{X}_n)}}{\sqrt{n}}\Big].$$

On déduit donc de ce qui précède que si $\phi_{1-\alpha/2}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$ (voir les tables du chapitre XI), alors $\Big[\bar{X}_n \pm \frac{\phi_{1-\alpha/2}\sqrt{\bar{X}_n(1-\bar{X}_n)}}{\sqrt{n}}\Big]$ est un intervalle de confiance asymptotique de niveau $1-\alpha$.

En conclusion, on fournit l'intervalle de confiance asymptotique de niveau $1-\alpha$ pour p :

$$\boxed{I_2 = \Big[\bar{x}_n \pm \frac{\phi_{1-\alpha/2}\sqrt{\bar{x}_n(1-\bar{x}_n)}}{\sqrt{n}}\Big].} \tag{II.3}$$

En fait cet intervalle de confiance est de bonne qualité pour $np(1-p)$ grand (par exemple $np(1-p) \geq 5$), voir le graphique II.4. La mauvaise qualité de l'intervalle de confiance, pour $p \simeq 0$ et $p \simeq 1$, est due à la mauvaise estimation de $\sigma^2 = p(1-p)$ par $\bar{X}_n(1-\bar{X}_n)$. Remarquons que la convergence de $\sqrt{n}(\bar{X}_n - p)/\sqrt{p(1-p)}$ vers la loi gaussienne centrée, permet d'assurer que

$$\lim_{n\to\infty} \mathbb{P}_p\left(\bar{X}_n \in [p \pm \frac{\phi_{1-\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}]\right) = \mathbb{P}(G \in [-\phi_{1-\alpha/2}, a_{1-\alpha/2}]) = 1-\alpha.$$

En calcul élémentaire permet d'obtenir que

$$\bar{X}_n \in [p \pm \frac{\phi_{1-\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}] \Leftrightarrow p \in [J^-(\bar{X}_n), J^+(\bar{X}_n)],$$

où, pour $r \in [0,1]$,

$$J^{\pm}(r) = \frac{r + (2n)^{-1}\phi^2_{1-\alpha/2}}{1 + \phi^2_{1-\alpha/2}/n} \pm \frac{1}{\sqrt{n}}\,\phi_{1-\alpha/2}\frac{\sqrt{r(1-r) + \phi^2_{1-\alpha/2}/4n}}{1 + \phi^2_{1-\alpha/2}/n}.$$

On peut donc fournir l'intervalle de confiance asymptotique suivant de niveau $1-\alpha$ pour p :

$$\boxed{I_3 = [J^-(\bar{x}_n), J^+(\bar{x}_n)].} \tag{II.4}$$

Le comportement des intervalles de confiance I_1 I_2 et I_3 est présenté au graphique II.4. On remarque que l'approximation gaussienne est médiocre pour n petit. Et que l'intervalle de confiance asymptotique I_3 est meilleur, en particulier pour les valeurs extrêmes de p, que l'intervalle de confiance asymptotique I_2.

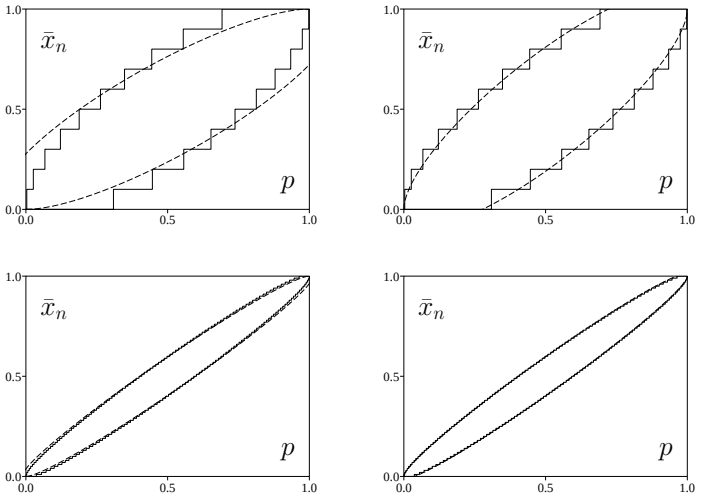


FIG. II.4 – Intervalles de confiance de niveau 95% pour p dans un modèle de Bernoulli, avec $n = 10$ (en haut) et $n = 100$ (en bas) observations. La moyenne empirique est lue en ordonnée. En coupant au niveau de la moyenne empirique on obtient un intervalle de confiance par excès (formule (II.2)) délimités par les courbes en trait plein, et un intervalle de confiance asymptotique (formule (II.3) à gauche, formule (II.4) à droite) délimités par les courbes en pointillées.

Si on fait l'application numérique du paragraphe II.1.1, on obtient $n = 100$, $\bar{x}_n = 0.22$, et pour $\alpha = 5\%$, il vient

$$I_1 \simeq [0.152, 0.314], \quad I_2 \simeq [0.139, 0.301], \quad I_3 \simeq [0.150, 0.311].$$

$\Diamond$

II.6 Procédure de décision : les tests

Exemple II.26. Comme nous l'avons vu au paragraphe II.1.3, il est naturel pour l'industriel, plutôt que d'estimer la vrai valeur du paramètre p, de répondre à la question suivante : est-ce

que $p \in [0, p_0]$ (i.e. la production continue), ou bien est-ce que $p \in]p_0, 1]$ (i.e. il faut arrêter la production pour régler ou changer la machine). Ici p_0 est un seuil que l'industriel s'est fixé pour le taux maximal de pièces "défectueuses" qu'il peut se permettre avant de devoir interrompre la production pour régler ou changer la machine. $\Diamond$

Les **tests d'hypothèses** sont des règles de décisions qui à partir des observations permettent de considérer que le paramètre θ inconnu appartient à $H_0 = \{p \leq p_0\}$ (**hypothèse nulle**) ou $H_1 = \{p > p_0\}$ (**hypothèse alternative**), (H_0, H_1) formant une partition de l'ensemble des paramètres Θ. Plus précisément, un test est défini par sa **région critique** (appelé aussi **zone de rejet**), notée ici W, qui est un sous ensemble de l'ensemble des observations possibles : si on observe $x = (x_1, \ldots, x_n)$, alors

- soit x **n'est pas** dans la région critique, on **accepte alors** H_0 et on rejette H_1 (i.e. on dit que $\theta \in H_0$).
- soit x **est** dans la région critique, on **rejette alors** H_0 et on accepte H_1 (i.e. on dit que $\theta \in H_1$).

Bien sûr, les données étant aléatoires, la décision peut être erronée. Il convient de distinguer deux types d'erreurs. L'**erreur de 1ère espèce** consiste à rejeter H_0 à tort : elle est définie par

$$\mathbb{P}_\theta(W), \quad \theta \in H_0,$$

où par définition $\mathbb{P}_\theta(W) = \mathbb{P}_\theta((X_1, \ldots, X_n) \in W)$. Dans l'exemple ci-dessus, l'erreur de 1ère espèce revient à arrêter la production alors que le taux de pièces "défectueuses" est encore acceptable.

L'**erreur de 2ème espèce** consiste à accepter H_0 à tort (i.e. rejet H_1 à tort) : elle est définie par

$$1 - \mathbb{P}_\theta(W), \quad \theta \in H_1.$$

Dans l'exemple ci-dessus, l'erreur de 2ème espèce consiste à continuer la production alors que le taux de défaillance est supérieur au seuil p_0 fixé par l'industriel.

On définit la **puissance** d'un test de région critique W comme la probabilité de rejeter H_0 à raison : c'est la fonction définie sur H_1 par $\theta \mapsto \rho_W(\theta) = \mathbb{P}_\theta(W)$. (La quantité $1 - \rho_W(\theta)$ représente l'erreur de 2ème espèce.)

Généralement les erreurs de 1ère et 2ème espèce n'ont pas la même importance. Par convention, on choisit H_0 de sorte que **l'on désire minimiser en priorité l'erreur de 1ère espèce**. Ceci traduit le fait que rejeter H_0 à tort est plus préjudiciable que rejeter H_1 à tort. Cette raison entraîne une dissymétrie entre l'hypothèse nulle et l'hypothèse alternative. **Le statisticien doit bien comprendre le contexte, pour choisir correctement l'hypothèse nulle et l'hypothèse alternative !** (En général, le but d'un test est de rejeter H_0. Quand on ne peut pas rejeter H_0, car l'erreur de 1ère espèce est trop importante, on peut utiliser un autre test ou augmenter le nombre de mesures.)

Le **niveau** du test de région critique W est défini par

$$\boxed{\alpha_W = \sup_{\theta \in H_0} \mathbb{P}_\theta(W).}$$

Il s'agit du maximum de l'erreur de 1ère espèce quand θ parcours H_0. On cherche donc en priorité des tests de niveau faible. Ceci se traduit par le **principe de Neyman** qui consiste

à ne retenir que les tests dont le niveau est inférieur à un **seuil α fixé a priori**. La valeur de α est faible ; les valeurs typiques sont 10%, 5%, 1%,... Ensuite, parmi les tests de niveau inférieur à α, on cherche à minimiser l'erreur de $2^{\text{ème}}$ espèce. On recherche donc **les tests de niveau inférieur à α et de puissance maximale**.

Remarque II.27. Il est clair que le choix de H_0 et de H_1 dépendent de la problématique. Dans le cas du modèle du paragraphe II.1.1, repris dans l'exemple II.26, le choix de $H_0 = \{p \leq p_0\}$ et $H_1 = \{p > p_0\}$ implique que l'on arrête la production (i.e. on rejette H_0) si l'on est vraiment convaincu (avec une marge d'erreur au plus égale au seuil du test) que le seuil, p_0, de taux de pièces "défectueuses" est dépassé. Ce choix correspond au cas où l'arrêt de la production à un coût très important.

Si en revanche on choisit $H_0 = \{p \geq p_0\}$ et $H_1 = \{p < p_0\}$, cela signifie que l'on accepte de continuer la production que si l'on est vraiment convaincu (toujours avec une marge d'erreur au plus égale au seuil du test) que le seuil, p_0, de taux de pièces "défectueuses" n'est pas encore atteint. C'est par exemple le point de vue d'un acheteur ou d'un organisme de contrôle de la qualité (dans ce cas la qualité de la production est un enjeu qui prime sur le coût d'arrêt de la production).

En particulier, le choix de H_0 et H_1 dépend directement des préoccupations de l'interlocuteur ! ◊

Une bonne idée pour construire un test concernant $g(\theta)$ consiste à utiliser l'EMV de $g(\theta)$, comme l'illustre l'exemple qui suit.

Exemple II.28. Nous poursuivons l'étude de l'exemple II.26 concernant le modèle de Bernoulli.

On suppose que pour l'industriel, il est très coûteux d'arrêter la production pour réparer ou changer la machine. Dans ce cas, on ne veut absolument pas commettre (traduire par "on veut minimiser") l'erreur suivante : arrêter la production (i.e. rejeter $p \leq p_0$ et accepter $p > p_0$) si $p \leq p_0$. On en déduit donc que l'erreur de $1^{\text{ère}}$ espèce consiste à rejeter $\{p \leq p_0\}$ à tort. Donc l'hypothèse nulle est $H_0 = \{p \leq p_0\}$ et l'hypothèse alternative $H_1 = \{p > p_0\}$.

On considère l'EMV de p, $\bar{X}_n$. On a vu que $\bar{X}_n$ est un estimateur convergent de p. En particulier, sous H_0, $\bar{X}_n$ prend des valeurs proches de p dans $[0, p_0]$, et sous H_1 des valeurs proches de p dans $]p_0, 1,]$. Il est naturel de choisir la région critique suivante construite à partir de l'EMV :

$$W_n = \{(x_1, \ldots, x_n);\ \bar{x}_n > a\}$$

où n est la taille de l'échantillon, et a une constante à déterminer.

Il faut maintenant calculer le niveau du test : $\alpha_{W_n} = \sup_{p \leq p_0} \mathbb{P}_p(\bar{X}_n > a)$. Intuitivement la moyenne empirique a "tendance" à augmenter avec p. On s'attend donc à ce que le supremum soit atteint en p_0. C'est effectivement le cas, comme le démontre l'inégalité (II.1) (rappelons que sous $\mathbb{P}_p$, la loi de $n\bar{X}_n$ est la loi binomiale de paramètre (n, p)). On en déduit donc que

$$\alpha_{W_n} = \mathbb{P}_{p_0}(\bar{X}_n > a) = \mathbb{P}(Z > na) = 1 - \mathbb{P}(Z \leq na),$$

où Z est de loi binomiale de paramètre (n, p_0). En particulier si on se fixe un seuil α, alors il faut choisir a supérieur ou égal à $q_{1-\alpha}/n$, où $q_{1-\alpha}$ est le quantile d'ordre $1 - \alpha$ de la loi binomiale de paramètre (n, p_0). Parmi tous les tests construits pour $a \geq q_{1-\alpha}/n$, on recherche maintenant celui de puissance maximale. La puissance est donnée par la fonction définie sur

$]p_0, 1]$ par $p \mapsto \mathbb{P}_p(\bar{X}_n > a)$. Il est clair que la puissance est maximale pour a le plus petit possible, à savoir $a = q_{1-\alpha}/n$. On en déduit donc que le test de seuil α retenu est

$$W_n = \{(x_1, \ldots, x_n);\ \bar{x}_n > q_{1-\alpha}/n\} = \{(x_1, \ldots, x_n);\ \sum_{i=1}^{n} x_i > q_{1-\alpha}\}. \tag{II.5}$$

Si on fait l'application numérique du paragraphe II.1.1 avec $p_0 = 0.2$, on obtient $n = 100$, $\sum_{i=1}^{n} x_i = 22$, $\bar{x}_n \geq p_0$ et pour $\alpha = 5\%$, on a $q_{1-\alpha} = 27$. Il vient $W_n = \{\bar{x}_n > 0.27\}$ et $\alpha_{W_n} = \mathbb{P}_{p_0}(W_n) = 0.03$ (le fait que α_{W_n} soit strictement inférieur à α est dû au fait que $\bar{X}_n$ est une v.a. discrète). La valeur observée n'est pas dans la région critique. On accepte donc H_0 au niveau de 5% (en fait au niveau de 3%).

Remarquons que plus p est proche de 0 (i.e. plus on est loin de H_1), plus l'erreur de 1ère espèce est faible. De même, plus p est proche de 1 (i.e. plus on est loin de H_0), plus l'erreur de 2ème espèce est faible et la puissance élevée. Enfin quand $p \in H_1$ se rapproche de p_0 (i.e. de H_0), on remarque que l'erreur de 2ème espèce croît vers $1 - \alpha_{W_n}$. L'erreur de 2ème espèce est loin d'être négligeable. (On pourra vérifier que les tests présentés dans les différents chapitres possèdent ce comportement).

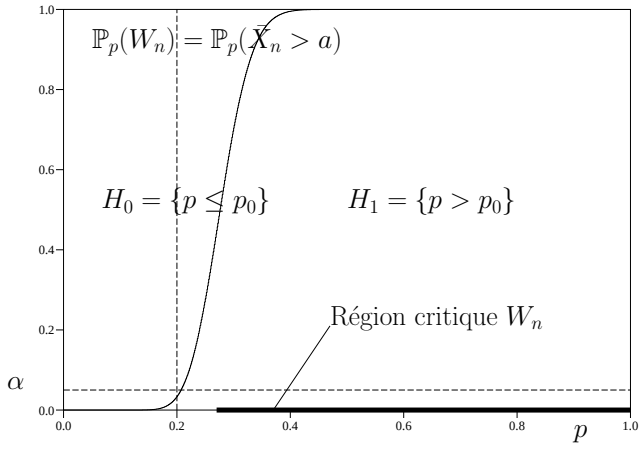


FIG. II.5 – Fonction $p \mapsto \mathbb{P}_p(W_n)$ (erreur de 1ère sur H_0 et puissance sur H_1) et région critique du test (II.5), avec $\alpha = 5\%$, $p_0 = 0.2$ et $a = 0.27$.

Il est intéressant de remarquer qui si l'on adopte le point de vue d'un acheteur ou d'un organisme de contrôle de la production, alors il est préférable de minimiser l'erreur suivante : continuer la production (i.e. rejeter $p \geq p_0$ et accepter $p < p_0$) si $p \geq p_0$. Dans ce cas on choisit $H_0 = \{p \geq p_0\}$ et $H_1 = \{p < p_0\}$. Des calculs similaires à ceux qui précédent permettent d'établir que le test de seuil α et de puissance maximale construit à l'aide de l'EMV de p a pour région critique

$$W'_n = \{(x_1, \ldots, x_n);\ \sum_{i=1}^{n} x_i < q_\alpha\},$$

où q_α est le quantile d'ordre α de la loi binomiale de paramètre (n, p_0). En particulier on remarquera que le complémentaire de $W'_n \cup W_n$ dans $[0,1]$ représente la zone où l'on accepte H_0 que l'on ait choisi $H_0 = \{p \leq p_0\}$ ou $H_0 = \{p \geq p_0\}$. $\Diamond$

Si l'on peut écrire la région critique, construite pour un échantillon de taille n, sous la forme $W = \{(x_1, \ldots, x_n); S(x_1, \ldots, x_n) \geq a\}$, abrégé par $W = \{S \geq a\}$ (ou sous la forme $W = \{(x_1, \ldots, x_n); S(x_1, \ldots, x_n) \leq a\} = \{S \leq a\}$ abrégé par $W = \{S \leq a\}$), où $S = S(X_1, \ldots, X_n)$ est une statistique, alors on dit que S est une **statistique de test** et que W est la région critique associée à la statistique de test S.

Dans ce cas, pour une observation $x^{\text{obs}} = (x_1^{\text{obs}}, \ldots, x_n^{\text{obs}})$, on définit la p**-valeur** du test de région critique $W = \{S \geq a\}$ par

$$p\text{-valeur} = \sup_{\theta \in H_0} \mathbb{P}_\theta(S \geq S^{\text{obs}}),$$

où $S^{\text{obs}} = S(x^{\text{obs}})$. Bien sûr la p-valeur du test de région critique $W = \{S \leq a\}$ est donnée par $\sup_{\theta \in H_0} \mathbb{P}_\theta(S \leq S^{\text{obs}})$. La p-valeur traduit la probabilité, sous H_0, d'observer pire que les observations dont on dispose. Si la p-valeur est très faible, cela signifie que ce que l'on observe est très rare (sous H_0). En particulier cela remet en cause l'hypothèse H_0. Suivant la valeur de la p-valeur et le contexte qui permet d'apprécier le niveau de risque retenu, on pourra alors rejeter H_0. En revanche si la p-valeur n'est pas faible, cela signifie qu'il est raisonnable d'observer cette valeur sous H_0. On acceptera alors H_0. La p-valeur est un indicateur de la confiance que l'on accorde à H_0, on peut également la voir comme l'inverse d'une "distance" à H_0. Il est plus précis que la réaction binaire du rejet ou acceptation dépendant d'un niveau parfois arbitraire ! Remarquons que si H_0 est simple, alors la p-valeur est sous H_0 la réalisation d'une variable aléatoire uniforme sur $[0,1]$.

Enfin la construction des tests et des régions critiques est souvent facilitée quand on regarde le comportement asymptotique des statistiques de test lorsque la taille de l'échantillon tend vers l'infini.

Si $(W_n, n \geq 1)$ est une suite de régions critiques où W_n correspond aux échantillons de taille n, on dit que le test W_n est de niveau asymptotique α, si

$$\boxed{\limsup_{n \to \infty} \sup_{\theta \in H_0} \mathbb{P}_\theta(W_n) = \alpha.}$$

Le test est dit **convergent** si pour tout $\theta \in H_1$, on a

$$\boxed{\lim_{n \to \infty} \mathbb{P}_\theta(W_n) = 1.}$$

Un test convergent assure que quand $n \to \infty$, si on est sous H_1, alors les observations sont p.s. dans la région critique.

Exemple II.29. Suite de l'exemple II.28. On peut vérifier que les tests W_n et W'_n sont des tests convergents de niveau asymptotique α. Plutôt que de faire cette démonstration académique, on construit naturellement un test asymptotique à partir de la normalité asymptotique de $\bar{X}_n$, l'EMV de p.

On peut structurer la modélisation et la réponse du statisticien de la manière suivante.

1. Choix d'un modèle (issu de la discussion entre le statisticien et son interlocuteur) : le modèle paramétrique de Bernoulli : $(X_n, n \geq 1)$ suite de v.a. i.i.d. de loi de Bernoulli $p \in [0,1]$.
2. Choix des hypothèses (issues des préoccupations de l'interlocuteur du statisticien) : $H_0 = \{p \leq p_0\}$, $H_1 = \{p > p_0\}$, avec $p_0 \in]0,1[$.
3. Choix de la statistique de test
$$\zeta_n = \sqrt{n}\frac{\bar{X}_n - p_0}{\sqrt{p_0(1-p_0)}}.$$
4. Comportement asymptotique sous H_0. On déduit de la loi forte des grands nombres que pour $p < p_0$, $\bar{X}_n - p_0$ converge p.s. vers $p - p_0 < 0$. On en déduit donc que pour $p < p_0$, on a p.s.
$$\lim_{n\to\infty} \zeta_n = -\infty.$$
Pour $p = p_0$, on déduit du TCL la convergence en loi de $(\zeta_n, n \geq 1)$ vers $\mathcal{N}(0,1)$.
5. Comportement asymptotique sous H_1. On déduit de la loi forte des grands nombres que $(\bar{X}_n - p_0, n \geq 1)$ converge p.s. vers $p - p_0 > 0$. On en déduit donc que sous H_0, p.s.
$$\lim_{n\to\infty} \zeta_n = +\infty.$$
6. Région critique du test. Vu le comportement asymptotique de la statistique de test sous H_0 et H_1, on considère la région critique
$$W_n = \{\zeta_n \geq a\}. \tag{II.6}$$
7. Le test est de niveau asymptotique α, si $a = \phi_{1-\alpha}$, où $\phi_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $\mathcal{N}(0,1)$. En effet, grâce à (II.1), on déduit que pour $p \in H_0$, on a
$$\mathbb{P}_p(\zeta_n \geq a) \leq \mathbb{P}_{p_0}(\zeta_n \geq a).$$
On déduit du TCL que $\lim_{n\to\infty} \mathbb{P}_{p_0}(\zeta_n \geq a) = \mathbb{P}(G \geq a)$, où G est de loi $\mathcal{N}(0,1)$. En particulier, on a
$$\lim_{n\to\infty} \sup_{p\in H_0} \mathbb{P}_p(\zeta_n \geq \phi_{1-\alpha}) = \lim_{n\to\infty} \mathbb{P}_{p_0}(\zeta_n \geq \phi_{1-\alpha}) = \alpha.$$
8. Le test est convergent. (Cela découle du fait que la statistique de test converge vers $+\infty$ sous H_1.)
9. La p-valeur du test est donnée par
$$p\text{-valeur} = \mathbb{P}(G \geq \zeta_n^{\text{obs}}),$$
où ζ_n^{obs} est la valeur de la statistique de test ζ_n évaluée en les observations (i.e. on remplace $\bar{X}_n$ dans la définition de ζ_n par $\bar{x}_n^{\text{obs}} = \frac{1}{n}\sum_{i=1}^n x_i^{\text{obs}}$).

Le statisticien peut alors fournir la p-valeur à son interlocuteur, et l'aider à conclure. Si on fait l'application numérique du paragraphe II.1.1, on obtient pour $p_0 = 0.2$, $n = 100$, $\bar{x}_n = 0.22$, $\zeta_n = 0.5$, et la p-valeur asymptotique de 0.31. (La p-valeur associée au test (II.5) est de 0.26.) En particulier on ne rejette pas H_0 au seuil de 5% ni au seuil de 10%. On voit ici que le calcul de la p-valeur donne une information plus complète que la simple acceptation de H_0 au niveau de 5%.

$\Diamond$

Comme dans l'exemple précédent, nous ferons ressortir pour chaque test les 9 points suivants :

1. Description du modèle (guidé par le contexte).
2. Les hypothèses (guidées par le contexte).
3. La statistique de test (utilisant l'EMV ou les statistiques exhaustives dans les modèles paramétriques).
4. Loi ou comportement (en général asymptotique) sous H_0.
5. Loi ou comportement (en général asymptotique) sous H_1.
6. Région critique du test.
7. Niveau (exact, par excès ou asymptotique) du test.
8. Puissance ou convergence du test.
9. Calcul de la p-valeur du test.

À l'issu de ce travail, il faut conclure : si la p-valeur est faible, on rejette H_0, sinon on accepte H_0. La notion de p-valeur "faible" dépend du contexte, elle n'est pas la même suivant qu'il s'agit d'un risque nucléaire, de la mise sur le marché d'un médicament ou d'une nouvelle lessive. Enfin, si on rejette H_0, il faut aussi se poser des questions sur la pertinence du modèle choisi.

Exercice II.30. Reprendre l'exemple II.29 avec $H_0 = \{p \geq p_0\}$ et $H_1 = \{p < p_0\}$. (Bien sûr, comme $\hat{p}_n \in H_0$, il est inutile de faire le test, car il n'est pas raisonnable de rejeter H_0.) On vérifiera que la p-valeur asymptotique est égale à 0.69. ⧫

Exercice II.31. On reprend le modèle décrit dans l'exemple II.29 ainsi que les notations. En considérant la statistique de test $|\zeta_n|$, donner un test asymptotique pour $H_0 = \{p = p_0\}$ et $H_1 = \{p \neq p_0\}$. ⧫

Exemple II.32. On considère le problème **E** posé au paragraphe II.1.1 : deux sous-échantillons chacun i.i.d., l'un de paramètre p_a, l'autre de paramètre p_b. Si on veut tester l'hypothèse $H_0 = \{p_b \leq p_a\}$, des considérations de bon sens analogues aux précédentes conduisent à un rejet si $\bar{x}_n^{(a)}$, proportion observée sur le premier sous-échantillon, est significativement plus élevée que $\bar{x}_n^b$, proportion observée sur le second sous-échantillon. Les calculs, exacts ou approchés, en fonction du choix de α seront présentés au chapitre IV. $\Diamond$

II.7 Utiliser des données extérieures : l'approche bayesienne

Suite du paragraphe II.1.1 sur le modèle de Bernoulli.

Une critique qui est souvent faite à l'ensemble des traitements des observations issues du paragraphe II.1.1, est que toutes les valeurs de p appartenant à $[0,1]$ y soient traitées de manière indifférente ; or il en est rarement ainsi dans la pratique. Certes on peut décider de restreindre l'espace du paramètre, Θ, à un sous-intervalle de $[0,1]$, mais cela ne suffit pas : souvent on a "de bonnes raisons" de penser que certaines zones de valeurs, pour p sont "plus crédibles" que d'autres. Ceci conduit naturellement à *munir l'ensemble* $[0,1]$ *d'une probabilité qui traduit cette crédibilité*; notons la Q. Elle est appelée **la probabilité a priori** (ou en bref **l'a priori**). Cette démarche est dite **bayésienne**.

L'espace produit $\Theta \times \{0,1\}^n$ est donc muni d'une probabilité, soit R, relativement à laquelle Q apparaît comme une **probabilité marginale** (projection sur l'ensemble Θ des paramètres) et $\mathbb{P}_p$ apparaît comme une **probabilité conditionnelle** : loi de l'observation, conditionnellement au paramètre. On peut donc donner, inversement, dans ce cadre, un sens à la loi du paramètre conditionnellement à l'observation. Cette probabilité est dite **probabilité a posteriori** (ou en bref **l'a posteriori**) . C'est cette démarche que, au dix-huitième siècle, T. Bayes avait qualifiée de *recherche de la probabilité des causes*. Elle peut être vue comme une "révision", à la lumière des observations, de la crédibilité qui avait été choisie initialement sur Θ. Dans cette logique, c'est sur elle que vont reposer alors toutes les méthodes statistiques.

La critique que l'on peut faire à cette démarche est le caractère relativement arbitraire du choix de la probabilité a priori : le statisticien devrait en principe la faire exprimer par son interlocuteur, mais ceci est souvent difficile à expliciter. De plus les calculs auxquels on est conduit sont souvent difficiles (à telle enseigne que c'est pour les résoudre qu'ont été effectuées certaines des avancées les plus notables du calcul statistique sur ordinateur ces dernières années). On est donc souvent amené à faire jouer, pour le choix de la probabilité a priori, des considérations de nature mathématique visant à faciliter les calculs.

Dans l'exemple considéré, on choisit volontiers pour loi a priori une loi Beta (voir p. 249), de paramètre $(a,b) \in \mathbb{R}_+^*$. En particulier, si $a = b = 1$, il s'agit de la loi uniforme sur $[0,1]$, qui traduit une information a priori dite par antiphrase "non informative". S'il en est ainsi, la loi du couple $(p,(x_1,\ldots,x_n))$ admet pour densité, par rapport à la mesure produit $\lambda_{[0,1]} \times \delta_{\{0,1\}^n}$ (où $\lambda_{[0,1]}$ est la restriction à l'intervalle $[0,1]$ de la mesure de Lebesgue sur $\mathbb{R}$ (voir X.1.2) et $\delta_{\{0,1\}^n}$ est la mesure de comptage sur $\{0,1\}^n$) l'application :

$$(p,(x_1,\ldots,x_n)) \mapsto \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)} p^y (1-p)^{n-y},$$

où $B(a,b) = \int_0^1 p^{a-1}(1-p)^{b-1}dp$ et $y = \sum_{i=1}^n x_i = n\bar{x}_n$. Un calcul élémentaire permet d'en déduire que la loi a posteriori est la loi Beta de paramètre $(a+y, b+n-y)$.

On prend alors pour **estimation** ponctuelle de p (estimation dite **bayésienne**) l'espérance mathématique de la loi a posteriori, c'est-à-dire après un calcul élémentaire

$$\hat{p}_n = \mathbb{E}[p|X_1,\ldots,X_n] = \frac{a + \sum_{i=1}^n X_i}{a+b+n}.$$

Si on observe $y = \sum_{i=1}^n x_i$, l'estimation de p est donc $\dfrac{a+y}{a+b+n}$. Dans le cas particulier de l'a priori uniforme, on trouve pour estimation $\frac{1+y}{2+n}$. Dans tous les cas, à a et b fixés, cette estimation, quand n augmente, s'éloigne de $\frac{a}{a+b}$ (espérance de la probabilité a priori) pour tendre vers $\bar{x}_n$. On retrouve donc asymptotiquement l'EMV de p. Ceci exprime que, plus la

taille de l'échantillon est élevée, plus forte est l'influence des observations par rapport à l'idée a priori que l'on se faisait sur le paramètre.

La recherche d'un **intervalle de confiance** au niveau de confiance $1-\alpha$ se résout ici en fournissant un intervalle qui, pour la loi a posteriori, a une probabilité égale à $1-\alpha$; ceci laissant une certaine faculté de choix, on le prendra en général symétrique en probabilité, c'est-à-dire laissant au dessus et au dessous des intervalles de probabilité a posteriori égale à $\frac{\alpha}{2}$. Autrement dit l'intervalle de confiance a pour extrémités les quantiles a posteriori d'ordres $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$.

Le **problème de test** est abordé ici de manière tout à fait différente de ce que l'on avait fait en II.6. Puisque l'interlocuteur s'interroge sur l'appartenance de p à l'intervalle $[0,p_0]$, et qu'il avait été "capable" de donner une valeur a priori (avant toute expérimentation) à une probabilité de cet intervalle, on va lui fournir comme réponse à ses préoccupations la valeur de la probabilité a posteriori de $[0,p_0]$. À lui d'en tirer profit ! Il peut en particulier, s'il doit absolument trancher entre l'hypothèse nulle H_0 et la contre-hypothèse H_1, opter pour celle des deux qui a la plus forte probabilité a posteriori. Cette attitude semble nécessiter chez l'interlocuteur du statisticien une meilleure familiarité avec les notions probabilistes que l'attitude "classique" vue ci-dessus ; mais cette dernière, si elle semble moins exigeante, est souvent source de malentendus.

II.8 Résumé des procédures de test

Nous résumons les tests vus pour les exemples présentés au paragraphe II.1.1 et III.3.1.

II.8.1 Le modèle de Bernoulli

1. Modèle : $(X_k, 1\leq k\leq n)$ suite de v.a. i.i.d. de loi de Bernoulli : $\mathcal{P}=\{\mathcal{B}(p), \in [0,1]\}$.
2. $H_0=\{p\leq p_0\}$, $H_1=\{p>p_0\}$, avec $p_0\in]0,1[$.
3. Statistique de test : $\bar{X}_n$.
4. Comportement sous H_0 : $\bar{X}_n$ proche de $[0,p_0]$.
5. Comportement sous H_1 : $\bar{X}_n$ proche de $]p_0,1]$.
6. Région critique : $W_n=\{\bar{X}_n>a\}$.
7. Niveau α par excès : $a=q_{1-\alpha}/n$, où $q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de $\mathcal{B}(n,p)$.
8. Puissance du test $\mathbb{P}_p(\bar{X}_n>a)$, $p\in]0,1]$.
9. p-valeur : $\mathbb{P}_{p_0}(\bar{X}_n\geq \bar{x}_n^{\text{obs}})$.
10. Variante : test asymptotique (n grand), statistique de test : $\zeta_n=\sqrt{n}\dfrac{\bar{X}_n-p_0}{\sqrt{p_0(1-p_0)}}$. Sous H_0, ζ_n converge p.s. vers $-\infty$ ou en loi vers $\mathcal{N}(0,1)$. Sous H_1, ζ_n converge p.s. vers $+\infty$. Région critique $W_n=\{\zeta_n\geq a\}$. Niveau asympotique α : $a=\phi_{1-\alpha}$, où $\phi_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de $\mathcal{N}(0,1)$. Le test est convergent. p-valeur : $\mathbb{P}(G\geq \zeta_n^{\text{obs}})$ où G de loi $\mathcal{N}(0,1)$.

Chapitre III

Modèle linéaire gaussien

III.1 Généralités

III.1.1 Plans d'expériences

Le statisticien planifie une expérience statistique en fonction d'un objectif qui est souvent l'étude de l'effet de certains **facteurs** de variabilité d'un phénomène. Ces facteurs sont présents sous plusieurs **modalités**.

La technique de bon sens lorsque plusieurs facteurs sont à étudier est de ne modifier qu'un facteur à la fois. Par exemple, si on dispose de 3 facteurs présents chacun sous p modalités, cette technique conduirait à fixer 2 facteurs puis étudier dans chacun des cas l'effet du troisième facteur, soit $3p^2$ expériences. Dans beaucoup de cas le coût, l'efficacité, ou les possibilités effectives d'expérimentation, recommandent de minimiser le nombre d'expériences tout en conservant un cadre expérimental rigoureux. En répondant à ces critères, la méthode des plans d'expérience initiée au début du $XX^{\text{ème}}$ siècle par Ronald A.Fisher s'est imposée dans le cadre industriel pour tester des médicaments, des variétés de plantes, des procédés de fabrication, etc...

L'objectif de la construction de plans d'expérience est de mettre en place un dispositif expérimental permettant d'aboutir à une interprétation statistique des résultats notamment à l'aide de tests d'hypothèses. Pour cela il faut construire un modèle statistique qui distinguera parmi les facteurs de variabilité les **facteurs contrôlés** et les **facteurs aléatoires**.

- Les facteurs contrôlés sont :

  quantitatifs : on parle alors de **facteurs explicatifs**.

  qualitatifs : ils sont distingués dans ce cas par leur position dans le dispositif expérimental.

- Les facteurs aléatoires, ou **erreur expérimentale**, représentent la part de variabilité non associée à un facteur contrôlé de l'expérience : erreurs de mesures, variabilité due à un tirage aléatoire, etc...

III.1.2 Le modèle général

Ce type d'expérience statistique peut être décrit avec le modèle général suivant :

$$X = f(\theta) + \varepsilon,$$

où

- $X = (X_i)_{i=1,\dots,n}$ désigne les observations effectuées.
- $\theta = (\theta_1, \dots, \theta_p)$ est un vecteur de paramètres inconnu caractérisant les facteurs contrôlés que l'on souhaite étudier à l'aide de ces observations.
- $\varepsilon = (\varepsilon_i)_{i=1,\dots,n}$ sont des variables aléatoires indépendantes et centrées, représentant l'erreur expérimentale. Le modèle est gaussien si ε est un vecteur gaussien centré (voir X.2.2).
- $f(\cdot)$ est une application connue qui fixe le modèle. Ce modèle est linéaire si $f(\theta)$ est une application $\theta \mapsto R\theta$ où R est une matrice. Le modèle s'écrit alors matriciellement :

$$X = R\theta + \varepsilon.$$

Dans la suite nous considérerons des modèles linéaires gaussiens. Ces deux hypothèses (linéarité et caractère gaussien de l'erreur) doivent être validées. Pour les vérifier on peut, soit utiliser la connaissance *a priori* que l'on a du modèle, soit construire des tests.

Dans certains cas, lorsqu'il y a plusieurs observations, le caractère gaussien peut être une conséquence du théorème de la limite centrale (voir X.3.2). Enfin, dans de nombreux cas, on peut rendre le modèle gaussien et linéaire en modifiant le modèle, ou en effectuant des transformations sur les observations.

III.1.3 Exemples

Dans ce paragraphe nous proposons des exemples illustrant la problématique précédente. Dans les sections suivantes, nous donnerons les éléments permettant de résoudre ce type de problèmes.

Exemple III.1. Le tableau ci-dessous représente des mesures de hauteurs d'arbres en mètres effectuées dans 3 forêts distinctes. On rassemble dans un même tableau les mesures effectuées dans les 3 forêts dans le but de les comparer.

Le facteur étudié est ici l'influence de la forêt sur la hauteur de ces arbres. La variabilité de la hauteur due ici au tirage d'un échantillon aléatoire dans chaque forêt se décompose donc naturellement en une partie controlée, le facteur (forêt), et une partie aléatoire, la variabilité intrinsèque à la pousse des arbres due au terrain, à la lumière, à la présence ou non d'un autre arbre à proximité...

On peut supposer que les hauteurs des différents arbres sont indépendantes (ce qui exige que l'on ne mesure pas des arbres trop rapprochés les uns des autres), et que, pour la forêt numéro k, la mesure d'un arbre suit une loi gaussienne de moyenne m_k et de variance σ_k^2 ; on peut alors comparer les 3 échantillons 2 à 2. Mais si la variabilité des hauteurs des arbres peut être considérée comme identique d'une forêt à l'autre ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$) on observe trois échantillons gaussiens de même variance σ^2 et de moyennes différentes qui représentent l'effet de chaque forêt (les modalités du facteur "forêt") sur la pousse des arbres. L'hypothèse d'égalité des variances est appelée **homoscédasticité**. Avec ces hypothèses on peut alors écrire :

$$X_{i,j} = m_i + \varepsilon_{i,j} \quad \text{pour la } j\text{-ième mesure de la forêt } i, \quad j = 1, \dots, n_i, \quad i = 1, 2, 3,$$

où $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Ceci s'écrit avec une notation matricielle :

$$X = R\theta + \varepsilon,$$

Foret 1 $n_1 = 13$ arbres	Foret 2 $n_2 = 14$	Foret 3 $n_3 = 10$
23.4	22.5	18.9
24.4	22.9	21.1
24.6	23.7	21.2
24.9	24.0	22.1
25.0	24.4	22.5
26.2	24.5	23.5
26.3	25.3	24.5
26.8	26.0	24.6
26.8	26.2	26.2
26.9	26.4	26.7
27.0	26.7	
27.6	26.9	
27.7	27.4	
	28.5	

TAB. III.1 – Hauteurs d'arbres dans 3 forêts

où ε est un vecteur aléatoire gaussien, et

$$X = (X_{1,1}, \ldots, X_{1,n_1}, X_{2,1}, \ldots, X_{2,n_2}, X_{3,1}, \ldots, X_{3,n_3})^t,$$

$$R = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ - & - & - \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ - & - & - \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix}, \quad \theta = \begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix}$$

Ce problème est un problème **d'analyse de la variance à un facteur**. Il sera étudié en III.5. Pour répondre à la question "existe-t-il un effet forêt", on construira un test statistique dont l'hypothèse nulle est :

$$H_0 = \{m_1 = m_2 = m_3\}.$$

$\Diamond$

Exemple III.2. Le tableau suivant donne le nombre de jours de pluie et la hauteur de pluie en mm, observés pendant toute l'année à Paris de 1956 à 1995.

Années	1956	1957	1958	1959	1960	1961	1962	1963	1964	1965
Jours	154	161	193	131	198	152	159	159	146	196
Hauteur	545	536	783	453	739	541	528	559	521	880
Années	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975
Jours	192	161	176	173	199	141	170	156	198	164
Hauteur	834	592	634	618	631	508	740	576	668	658
Années	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985
Jours	135	179	171	172	170	197	173	177	177	163
Hauteur	417	717	743	729	690	746	700	623	745	501
Années	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Jours	176	180	167	140	149	140	154	155	192	162
Hauteur	611	707	734	573	501	472	645	663	699	670

TAB. III.2 – Jour et quantité de pluie par années

Une représentation sur un graphique (fig. III.1) des données avec en abscisse le nombre de jours de pluie et en ordonnée la hauteur de pluie permet de constater que l'ensemble des points forme un nuage allongé et que la quantité de pluie augmente lorsque le nombre de jours de pluie augmente.

Le facteur hauteur de pluie est alors un facteur à expliquer par le facteur explicatif contrôlé nombre de jours de pluie.

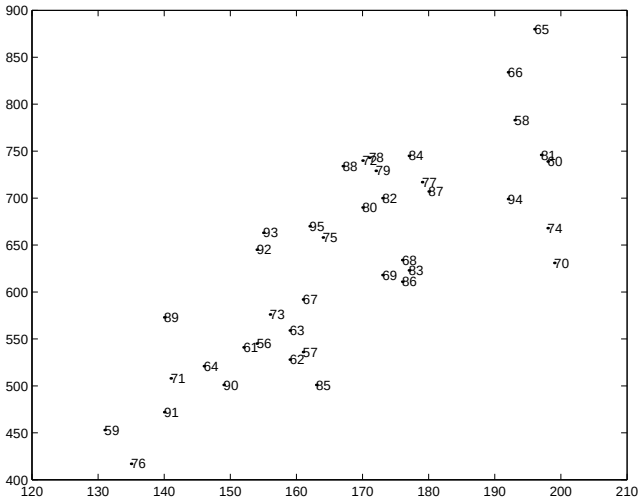


FIG. III.1 – Représentation des données

La question que l'on se pose est de savoir si ces deux quantités sont liées par une relation affine, de calculer les paramètres de cette relation et d'avoir une indication sur le caractère prédictif de ce modèle (autrement dit, peut-on déduire de façon satisfaisante la hauteur de pluie à partir du nombre de jours de pluie ?).

Le modèle statistique que l'on propose est le suivant :

$$X_i = \beta + \alpha\, r_i + \varepsilon_i$$

où :
- $X = (X_i)_{i=1,\dots,n}$ désigne la hauteur de pluie.
- $(r_i)_{i=1,\dots,n}$ désigne le nombre de jours de pluie
- la droite d'équation
$$x = \alpha\, r + \beta$$
est appelée droite de régression ; α et β sont à estimer à partir des observations.
- $\varepsilon = (\varepsilon_i)_{i=1,\dots,n}$ représente les écarts aléatoires entre les observations et la droite. On supposera que c'est une suite de variables aléatoires indépendantes de loi $\mathcal{N}(0, \sigma^2)$.

Le modèle peut alors s'écrire :

$$X = R\theta + \varepsilon$$

en notant :

$$R = \begin{pmatrix} 1 & r_1 \\ 1 & r_2 \\ \vdots & \vdots \\ 1 & r_n \end{pmatrix}, \quad \text{et} \quad \theta = \begin{pmatrix} \beta \\ \alpha \end{pmatrix}$$

C'est un modèle de **régression linéaire simple** qui sera étudié en III.6.

$\Diamond$

III.2 Lois associées aux échantillons gaussiens

Rappelons pour commencer les définitions des lois associées aux échantillons gaussiens qui, comme la loi de Student, nous serons utiles dans la suite (voir X.2.2).

Définition III.3. *Si $(X_1, \dots, X_n)$ est un échantillon de loi normale $\mathcal{N}(0,1)$, alors la loi de la v.a. $\sum_{i=1}^n X_i^2$ est la* **loi du chi-deux** *à n degrés de liberté, notée $\chi^2(n)$.*

Définition III.4. *Si $X \sim \mathcal{N}(0,1)$, $Y \sim \chi^2(n)$ et que X et Y sont indépendantes, alors $\frac{X}{\sqrt{Y/n}} \sim t(n)$,* **loi de Student** *à n degrés de liberté.*

Définition III.5. *Si $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$ et que X et Y sont indépendantes, alors $\frac{X/n}{Y/m} \sim \mathcal{F}(n,m)$,* **loi de Fisher** *(ou de Fisher-Snedecor) à n et m degrés de liberté.*

Enfin, on utilise souvent la convention pratique suivante : si une v.a. X a pour loi F, on note aF la loi de aX. Ainsi, on notera $\sigma^2\chi^2(n)$ la loi de $\sum_{i=1}^n X_i^2$ dans le cas où $(X_1, \dots, X_n)$ forment un n-échantillon de la loi $\mathcal{N}(0, \sigma^2)$.

III.2.1 Théorème de Cochran

C'est l'outil fondamental pour l'étude des échantillons gaussiens et du modèle linéaire gaussien (la notation $||\cdot||$ désigne la norme euclidienne dans $\mathbb{R}^n$).

Théorème III.6. *Soit $X = (X_1, \dots, X_n)$ un n-échantillon de $\mathcal{N}(0,1)$, et $E_1, \dots, E_p$ une décomposition de $\mathbb{R}^n$ en p sous-espaces deux-à-deux orthogonaux, avec $dim(E_j) = d_j$, $j = 1, \dots, p$. Alors on a :*

(i) Les composantes de X dans toute base orthonormale de $\mathbb{R}^n$ forment encore un n-échantillon de $\mathcal{N}(0,1)$.

(ii) Les vecteurs aléatoires $X_{E_1}, \dots, X_{E_p}$, qui sont les projections de X sur $E_1, \dots, E_p$, sont indépendants.

(iii) Les variables aléatoires $||X_{E_1}||, \dots, ||X_{E_p}||$ sont indépendantes, et $||X_{E_j}||^2 \sim \chi^2(d_j)$, $j = 1, \dots, p$.

Une formulation équivalente consiste à dire (par exemple avec $p = 2$), que si P_1 et P_2 sont deux projecteurs orthogonaux de $\mathbb{R}^n$ sur deux sous-espaces orthogonaux E_1 et E_2 de dimensions d_1 et d_2, alors $P_1X = X_{E_1}$ et $P_2X = X_{E_2}$ sont indépendants, et $||P_1X||^2$ et $||P_2X||^2$ sont indépendants et ont pour lois respectivement $\chi^2(d_1)$ et $\chi^2(d_2)$.

III.2.2 Statistiques fondamentales

Plaçons-nous donc dans le cas où $(X_1, \dots, X_n)$ est un n-échantillon de la loi $\mathcal{N}(\mu, \sigma^2)$. Les statistiques utiles pour les problèmes de test ou d'intervalle de confiance sur les paramètres μ et σ^2 sont fonction de la moyenne empirique, que nous notons

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

et de la variance empirique, dont nous choisissons ici la "version sans biais" (voir III.3.1) :

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{n}{n-1} \left[\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \left(\bar{X}\right)^2 \right].$$

Utilisons le théorème III.6 dans le cas où $p = 2$ et où on projette X sur le sous-espace E de dimension 1 engendré par le vecteur (normé) de $\mathbb{R}^n$, $e_1 = \frac{1}{\sqrt{n}} \mathbf{1}_n$ (où on note $\mathbf{1}_n$ le vecteur de dimension n ayant toute ses coordonnées égales à 1). On obtient $X_E = \sqrt{n}\bar{X} \frac{1}{\sqrt{n}} \mathbf{1}_n$. La norme de la projection de X sur l'orthogonal de E (de dimension $n-1$) est

$$||X - X_E||^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

qui suit la loi $\sigma^2 \chi^2(n-1)$ (c'est le point (iii) du théorème de Cochran à ceci près qu'il faut tenir compte de la variance σ^2). On en déduit les résultats suivants, utiles pour le statisticien :

Proposition III.7. *Soit $X = (X_1, \dots, X_n)$ un n-échantillon de $\mathcal{N}(\mu, \sigma^2)$. Alors on a :*

(i) Les v.a. $\bar{X}$ et S^2 sont indépendantes.

(ii) $(n-1)S^2 \sim \sigma^2 \chi^2(n-1)$.

(iii) $\dfrac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t(n-1)$.

Remarquons que la v.a. $\sum_{i=1}^{n}(X_i-\mu)^2$ suit elle-même la loi $\sigma^2\chi^2(n)$ mais, si μ est inconnu, son calcul n'est pas accessible. Le point (ii) exprime intuitivement le fait que l'on perd un degré de liberté en raison du remplacement de μ, inconnu, par son estimateur $\bar{X}$. De même la v.a. $\sqrt{n}(\bar{X}-\mu)/\sigma \sim \mathcal{N}(0,1)$, autrement dit le point (iii) signifie que la loi de Student remplace la loi normale comme loi de la moyenne empirique normalisée dans le cas où σ est inconnu et doit être remplacé par son estimateur S.

III.3 Le modèle gaussien

Nous illustrons dans un premier temps les concepts du modèle paramètrique sur le modèle gaussien. Ce modèle est très (trop ?) couramment utilisé pour analyser des données continues. Cet usage fréquent est dû à la simplicité des calculs et à la généralité du TCL (sous des hypothèses très faibles, la somme de nombreux petits bruits suit asymptotiquement une loi gaussienne).

III.3.1 Un exemple de données réelles à loi gaussienne

On a enregistré le taux d'alcool dans le sang (en dg/l) de n sujets : voici le tableau des observations, avec $n = 30$ (extrait de l'ouvrage de D. Schwartz, *Méthodes statistiques à l'usage des médecins et des biologistes*, Flammarion).

27	,	26	,	26	,	29	,	10	,	28	,	26	,	23	,	14	,	37
16	,	18	,	26	,	27	,	24	,	19	,	11	,	19	,	16	,	18
27	,	10	,	37	,	24	,	18	,	26	,	23	,	26	,	19	,	37

On notera $(x_1, \ldots, x_{30})$ cette suite de résultats observée. Les valeurs s'échelonnant entre 10 et 37, la précision étant l'unité, il serait maladroit de modéliser ceci comme les réalisations de v.a. discrètes : le nombre de valeurs distinctes envisageables devrait être grand, de l'ordre de la quarantaine, car rien n'interdit de penser qu'auraient pu être observées des valeurs en dehors de l'intervalle ici présent. Il est plus raisonnable de considérer qu'il y a, sous-jacent à ces observations, un phénomène à valeurs réelles, dont les observations recueillies sont une discrétisation, l'arrondi se faisant à la précision du décigramme par litre.

Les modèles les plus simples que l'on puisse envisager ici sont des modèles d'échantillonnage : on admet que l'on a observé les réalisations de n v.a. X_i indépendantes et identiquement distribuées.

Pour voir si un tel modèle est approprié, il faut d'abord se demander comment a été constitué cet échantillon.

Le problème essentiel est, comme dans le premier paragraphe, celui de la source de variabilité (cause de l'aléatoire). Celle-ci a en fait ici plusieurs origines simultanées : variation d'individu à individu et, pour chaque individu, imprécision de l'appareil de mesure et effet de l'erreur d'arrondi. On peut espérer que la première cause est dominante, mais alors il reste à s'interroger sur les conditions de choix des individus sur lesquels a été effectuée la prise de sang. Voici quelques situations possibles :

• Expérience scientifique contrôlée, par exemple faisant intervenir 30 sujets en bonne santé, ayant bu tous la même quantité d'alcool, dans des conditions d'alimentation identiques, et testés dans un temps déterminé après l'absorption : on jauge alors la variabilité des réactions individuelles des organismes.

• Contrôle systématique après un bon repas ou au sortir d'un boîte de nuit : on jauge alors la variabilité de consommation de gens placés dans un même contexte social, mêlé à l'effet individuel de cette consommation sur le sang.
• Rapport de police : on jauge alors la variabilité du taux parmi des analyses sur des conducteurs que la police a jugé utile de citer, par exemple après un premier filtrage à l'alcootest.

Revenant à l'ouvrage d'où ont été extraites ces données, nous lisons : *30 sujets en état d'ébriété*. Nous savons donc que nous nous limitons à une catégorie dont il faut donner la définition exacte (non fournie dans le texte) : celle-ci est-elle **externe** à l'enregistrement (par exemple détermination sur le comportement ou sur l'alcootest) ou **interne** à l'enregistrement (on aurait procédé à des observations sur une plus grande masse de sujets et retenu uniquement ceux pour lesquels le taux était supérieur ou égal à 10 dg/l, procédure dite de **censure** des données) ?

Il est assez évident que, quelles que soient les conditions de recueil, elles ont dû assurer l'**indépendance** des n v.a. X_i dont les observations résultent. **Le problème de l'identité de leurs lois et du choix de la famille à laquelle serait supposée appartenir cette loi commune est plus délicat**.

En tout état de cause, si l'on pose (comme nous allons le faire) un modèle i.i.d., il faudra retenir que la loi commune traduit une variabilité intrinsèque et de mesure dans l'ensemble des individus satisfaisant aux critères retenus pour la population que les sujets testés sont censés représenter, et celle-là seule (ici une certaine notion de l'*état d'ébriété*).

Nous l'avons dit, les praticiens utilisent souvent dans un tel contexte une modélisation avec pour loi commune une **loi normale**, de moyenne μ et variance σ^2 (non nulle) inconnues, $\mathcal{N}(\mu, \sigma^2)$. Le paramètre est donc bi-dimensionnel $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$. La probabilité $\mathcal{N}(\mu, \sigma^2)$ a pour support $\mathbb{R}$ tout entier, alors qu'ici (comme presque toujours dans la pratique) les données sont fondamentalement bornées ; cet usage suppose donc que, pour la zone de valeurs de μ et σ envisageables, la probabilité du complémentaire de l'intervalle des valeurs effectivement atteignables par les taux d'alcool soit négligeable. Cette condition de validité serait incontestablement mise en défaut dans le cas de valeurs censurées évoqué ci-dessus.

III.3.2 Étude du modèle

On considère donc un échantillon $(X_1, \ldots, X_n)$ de v.a. indépendantes et de même loi gaussienne : $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), \theta = (\mu, \sigma^2) \in \mathbb{R} \times]0, \infty[\}$. La densité de la loi $\mathcal{N}(\mu, \sigma^2)$ est

$$p(x_1; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \, \mathrm{e}^{-(x_1-\mu)^2/2\sigma^2} \, .$$

La vraisemblance du modèle est pour $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$,

$$\begin{aligned} p_n(x; \mu, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \, \mathrm{e}^{-\sum_{i=1}^n (x_i-\mu)^2/2\sigma^2} \\ &= (2\pi\sigma^2)^{-n/2} \, \mathrm{e}^{-n\frac{(\bar{x}_n-\mu)^2+v_n}{2\sigma^2}}, \end{aligned}$$

où $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ et $v_n = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2$. On déduit du théorème de factorisation II.19 que

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i, \quad \text{et} \quad V_n = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X}_n)^2.$$

sont des statistiques exhaustives (i.e. elles contiennent toute l'information sur le paramètre (μ, σ)). Traditionnellement, on considère

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

au lieu de V_n (car S_n^2 est un estimateur sans biais de σ^2), cf la proposition III.7. Bien sûr la statistique $(\bar{X}_n, S_n^2)$ est une statistique exhaustive ; sa loi est donnée dans la proposition III.7.

III.3.3 Estimation

Pour calculer l'estimateur du maximum de vraisemblance de (μ, σ^2), on considère la log-vraisemblance

$$\ell_n(x; \mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - n\frac{(\bar{x}_n - \mu)^2 + v_n}{2\sigma^2}.$$

En calculant les dérivées partielles, il vient

$$\frac{\partial}{\partial \mu}\ell_n(x; \mu, \sigma^2) = n\frac{\bar{x}_n - \mu}{\sigma^2},$$

et

$$\frac{\partial}{\partial \sigma^2}\ell_n(x; \mu, \sigma^2) = -\frac{n}{2\sigma^2} + n\frac{(\bar{x}_n - \mu)^2 + v_n}{2\sigma^4}.$$

En particulier, les dérivées de la log-vraisemblance s'annulent pour $\mu = \bar{x}_n$ et $\sigma^2 = v_n$. Ensuite, on vérifie sans difficulté que la log-vraisemblance atteint son maximum pour $(\mu, \sigma^2) = (\bar{x}_n, v_n)$. On en déduit donc que l'EMV de $\theta = (\mu, \sigma^2)$ est $(\bar{X}_n, V_n)$. On déduit de la proposition III.7 que $\mathbb{E}_\theta[\bar{X}_n] = \mu$ et que $\mathbb{E}_\theta[S_n^2] = \sigma^2$. (En revanche V_n est un estimateur biaisé de σ^2, d'où le choix traditionnel de S_n^2). Ainsi l'estimateur $\hat{\theta}_n = (\bar{X}_n, S_n^2)$ est un estimateur sans biais de θ.

On peut démontrer qu'il n'existe pas d'estimateur sans biais de μ préférable (au sens du risque quadratique) à $\bar{X}_n$ autre que lui même. Et que parmi les estimateurs de la forme aS_n^2, $a > 0$, de σ^2, $\frac{n-1}{n+1}S_n^2 = \frac{1}{n+1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ est l'estimateur de risque quadratique le plus faible.

Par la loi forte des grands nombre $\bar{X}_n$ et S_n^2 sont des estimateurs convergents. Ainsi $\hat{\theta}_n$ est un estimateur convergent de θ. (On peut également vérifier qu'il est asymptotiquement normal, mais cela ne nous sera pas utile par la suite).

III.3.4 Intervalle de confiance et tests pour la moyenne

III.3.5 Intervalle de confiance pour la moyenne

On déduit de la proposition III.7, que la loi de $\sqrt{n}\frac{\bar{X}_n - \mu}{S_n}$ est la loi $t(n-1)$. La loi de Student est symétrique, ainsi si $t_{n-1,1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi $t(n-1)$,

alors $-t_{n-1,1-\alpha/2}$ est le quantile d'ordre $\alpha/2$. En particulier, une v.a. de loi $t(n-1)$ appartient à $[-t_{n-1,1-\alpha/2}, t_{n-1,1-\alpha/2}]$ avec probabilité $1-\alpha$. Comme

$$\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \in [-t_{n-1,1-\alpha/2}, t_{n-1,1-\alpha/2}] \Longleftrightarrow \mu \in [\bar{X}_n \pm t_{n-1,1-\alpha/2}\frac{S_n}{\sqrt{n}}],$$

on en déduit que $[\bar{X}_n \pm t_{n-1,1-\alpha/2}\frac{S_n}{\sqrt{n}}]$ est un intervalle de confiance de niveau $1-\alpha$ pour μ.

On remarque que la longueur de l'intervalle de confiance $[\bar{x}_n \pm t_{n-1,1-\alpha/2}\frac{s_n}{\sqrt{n}}]$, où $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x}_n)^2$ tend bien vers 0 quand la taille de l'échantillon tend vers l'infini (à $\bar{x}_n$ et s_n fixé). Il est aussi d'autant plus long que s_n est plus élevé (ceci est naturel : la fluctuation des données contrarie la confiance que l'on a en elles, confiance qui se traduirait par un intervalle de confiance assez court).

Exercice III.8. Si la variance est connue et égale à σ_0^2, c'est-à-dire si l'on considère le modèle $\mathcal{P} = \{\mathcal{N}(\mu, \sigma_0^2), \mu \in \mathbb{R}\}$, vérifier que l'intervalle de confiance de μ de niveau $1-\alpha$ est alors $[\bar{X}_n \pm \phi_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}]$, où $\phi_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$. ♦

Tests pour la moyenne

On considère les hypothèses $H_0 = \{\mu = \mu_0\}$ et $H_1 = \{\mu \neq \mu_0\}$, où μ_0 est donné. (On parle d'hypothèse bilatérale, par opposition à l'exercice III.10, où parle d'hypothèse unilatérale). Il est naturel de comparer la moyenne empirique avec moyenne proposée, μ_0. Toutefois, sous H_0, la loi de $\bar{X}_n - \mu_0$ est la loi $\mathcal{N}(0, \sigma^2/n)$, qui dépend du paramètre inconnu σ^2. On considère donc la statistique de test

$$\zeta_n = \sqrt{n}\frac{\bar{X}_n - \mu_0}{S_n}.$$

La loi de la statistique de test sous H_0 est la loi de Student de paramètre $n-1$. La loi de ζ_n sous H_1 est la loi de Student décentrée, mais nous ne l'expliciterons pas ici. On remarque que sous H_1, $\bar{X}_n - \mu_0$ converge p.s. vers $\mu - \mu_0 \neq 0$ quand $n \to \infty$. On a toujours que S_n converge p.s. vers σ^2. On en déduit donc que sous H_1, p.s.

$$\lim_{n\to\infty} |\zeta_n| = +\infty.$$

Il est donc naturel de considérer la région critique

$$W_n = \{(x_1, \ldots, x_n);\ |\zeta_n(x)| \geq a\}, \tag{III.1}$$

où $\zeta_n(x) = \sqrt{n}\frac{\bar{x}_n - \mu_0}{s_n}$, avec $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ et $s_n = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x}_n)^2$. D'après le comportement de la statistique de test sous H_1, on en déduit que le test W_n est convergent. (En fait la statistique telle qu'elle a été définie au paragraphe II.6 serait plutôt $|\zeta_n|$.)

Comme sous H_0, la loi de ζ_n est la loi de Student de paramètre $n-1$, on en déduit que le niveau du test W_n est

$$\sup_{\theta \in H_0} \mathbb{P}_\theta(W_n) = \mathbb{P}(|Z| \geq a),$$

où Z est de loi $t(n-1)$. Pour obtenir un test de niveau α, on choisit $a = t_{n-1,1-\alpha/2}$, le quantile d'ordre $1-\alpha/2$ de loi de Student de paramètre $n-1$.

La p-valeur du test est donnée par

$$p\text{-valeur} = \mathbb{P}(|Z| \geq |\zeta_n^{\text{obs}}|), \tag{III.2}$$

où ζ_n^{obs} est la statistique de test évaluée en les observations.

Remarque III.9. On peut étudier la réponse du test en fonction de n, $\bar{x}_n$ et s_n.
- à n et s_n fixés, si $\bar{x}_n$ s'éloigne de μ_0, alors $|\zeta_n|$ augmente et on a tendance à rejeter le test.
- à n et $\bar{x}_n$ fixés, si s_n diminue, alors $|\zeta_n|$ augmente et on a tendance à rejeter le test. Cela traduit le fait que si s_n est petit alors la variabilité des données est petite et $\bar{x}_n$ donne une estimation précise du vrai paramètre μ. Des petits écarts entre $\bar{x}_n$ et μ_0 deviennent significatifs.
- à $\bar{x}_n$ et s_n fixés, si n augmente, alors $|\zeta_n|$ augmente et on a tendance à rejeter le test. En effet, plus la taille de l'échantillon est grande est plus $\bar{x}_n$ donne une estimation précise du vrai paramètre μ.

$\Diamond$

Exercice III.10. Écrire le test pour les hypothèses unilatérales $H_0 = \{\mu \leq \mu_0\}$ et $H_1 = \{\mu > \mu_0\}$. ⧫

Exercice III.11. Tester les hypothèses $H_0 = \{\mu = \mu_0\}$ et $H_1 = \{\mu \neq \mu_0\}$, où μ_0 est donné dans le modèle gaussien à variance connue : $\mathcal{P} = \{\mathcal{N}(\mu, \sigma_0^2), \mu \in \mathbb{R}\}$. ⧫

III.3.6 Intervalles de confiance et tests pour la variance

Le raisonnement est identique dans le cas de la variance : la construction d'intervalles de confiance ou de tests se fait à partir de la connaissance de la loi, sous l'hypothèse nulle ou à la frontière de celle-ci, de l'estimateur du paramètre d'intérêt.

Intervalles de confiance pour la variance

L'estimateur (sans biais) de σ^2 est la variance empirique sans biais S^2, et le point (ii) de la proposition III.7 permet d'écrire par exemple que, si $\chi^2_{n,1-\alpha}$ est le quantile d'ordre $(1-\alpha)$ de la loi $\chi^2(n)$,

$$\mathbb{P}\left(\chi^2_{n-1,\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1,1-\alpha/2}\right) = 1-\alpha,$$

d'où l'on déduit un intervalle de confiance pour la variance (bilatéral dans cet exemple) de niveau de confiance $(1-\alpha)$:

$$\left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\,;\,\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right].$$

Tests pour la variance

On peut aussi, en suivant la démarche introduite au chapitre II, construire des tests pour des hypothèses relatives au paramètre σ^2. Considérons par exemple le test de

$$H_0 = \{\sigma^2 \leq \sigma_0^2\} \quad \text{contre} \quad H_1 = \{\sigma^2 > \sigma_0^2\}.$$

à la frontière de H_0, i.e. lorsque la valeur du paramètre est σ_0^2, la statistique

$$Z = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1).$$

Cette statistique aura "tendance à croître" avec σ sous l'hypothèse alternative (et de plus $S^2 \to \sigma^2$ p.s. en vertu de la loi forte des grands nombres), d'où le choix d'une région de rejet de la forme $]c, +\infty[$, où c est calibré (le plus petit possible) de sorte que $\mathbb{P}_{\sigma_0}(Z > c) = \alpha$. Ceci amène donc à choisir pour c le quantile d'ordre $(1-\alpha)$ de la loi $\chi^2(n-1)$, autrement dit à conclure

$$\text{Rejet de } H_0 \text{ si} \quad \frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{n-1,1-\alpha}.$$

Le lecteur pourra construire les tests relatifs aux situations suivantes :

$$H_0 = \{\sigma^2 \geq \sigma_0^2\} \quad \text{contre} \quad H_1 = \{\sigma^2 < \sigma_0^2\},$$
$$H_0 = \{\sigma^2 = \sigma_0^2\} \quad \text{contre} \quad H_1 = \{\sigma^2 \neq \sigma_0^2\}.$$

III.3.7 Analyse des données réelles

On choisit le modèle $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\}$. On obtient l'estimation de (μ, σ^2) à l'aide de $(\bar{x}_n, s_n^2)$:

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i = 22.9 \quad \text{et} \quad s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x}_n)^2 = 53.0.$$

L'intervalle de confiance de niveau 95% de μ est donné par

$$[\bar{x}_n \pm t_{n-1,1-\alpha/2}\frac{s_n}{\sqrt{n}}] = [20.2, 25.6].$$

La p-valeur associée au test de région critique (III.1), définie par (III.2), est pour $\mu_0 = 20$,

$$p\text{-valeur} = \mathbb{P}(|Z| \geq \zeta_n^{\text{obs}}) = 0.037, \quad \text{où} \quad \zeta_n = \sqrt{n}\frac{\bar{x}_n - \mu_0}{s_n} = 2.18.$$

En particulier on rejette $H_0 = \{\mu = \mu_0\}$ au niveau de 5%.

III.3.8 Approche bayesienne

Comme pour le cas discret (voir le paragraphe II.7), relevons que le fait d'avoir traité indifféremment toutes les valeurs possibles de $\mu \in \mathbb{R}$ et, s'il y a lieu, de $\sigma \in]0, +\infty[$ est peu conforme à la réalité expérimentale. On peut donc, ici aussi, envisager de munir l'espace des valeurs du paramètre d'une probabilité a priori.

Nous nous limitons ici au cas où la variance σ_0 est connue (modèle $\mathcal{P} = \{\mathcal{N}(\mu, \sigma_0^2), \mu \in \mathbb{R}\}$) et où on adopte pour loi a priori sur μ une loi normale $\mathcal{N}(\nu, \tau^2)$ (avec, bien sûr, le problème pratique du choix de ces paramètres ν et τ^2 de la loi a priori, qui sont qualifiés d'**hyperparamètres**). On peut alors calculer la loi a posteriori, une fois observée une suite $(x_1, \ldots, x_n)$ de moyenne empirique $\bar{x}_n$; c'est la loi $\mathcal{N}(\frac{\sigma_0^2\nu + n\tau^2\bar{x}_n}{\sigma_0^2 + n\tau^2}, \frac{\sigma_0^2\tau^2}{\sigma_0^2 + n\tau^2})$.

Toutes les techniques statistiques en découlent selon les principes vus au paragraphe II.7. En particulier l'estimation bayésienne de μ est l'espérance de la loi a posteriori, c'est-à-dire $\dfrac{\sigma_0^2\nu + n\tau^2\bar{x}_n}{\sigma_0^2 + n\tau^2}$. On constate que, quand n augmente, l'estimation s'éloigne de l'espérance de la loi a priori, ν, pour converger vers la moyenne empirique $\bar{x}_n$. Ceci s'interprète aisément : plus on a de données, plus leur poids est prédominant par rapport à l'idée que l'on se faisait a priori de la valeur de μ.

III.4 Comparaison de moyennes de deux échantillons gaussiens

Nous considérons ici le problème de la comparaison de paramètres relatifs à deux populations, au vu de deux échantillons. Nous supposons disposer d'observations réelles, gaussiennes, **de même variance inconnue**, et de moyennes éventuellement différentes, sur lesquelles vont porter les hypothèses nulles à tester. Cette situation peut être vue comme un cas simple de l'analyse de la variance traitée en III.5, et on peut se reporter en III.1 et à la section suivante pour un exemple de problématique.

L'observation consiste en deux échantillons **indépendants entre eux** (non appariés)

$$\begin{array}{lll}(X_{1,1},\ldots,X_{1,n_1}) & \text{i.i.d.} & \mathcal{N}(\mu_1,\sigma^2)\\ (X_{2,1},\ldots,X_{2,n_2}) & \text{i.i.d.} & \mathcal{N}(\mu_2,\sigma^2),\end{array}$$

et on souhaite tester, par exemple,

$$H_0 = \{\mu_1 = \mu_2\} \quad \text{contre} \quad H_1 = \{\mu_1 \neq \mu_2\}.$$

Comme dans le cas de la comparaison de proportions fondée sur des échantillons non appariés (voir IV.4.2), il est pratique de faire porter le test sur l'hypothèse nulle équivalente

$$H_0 = \{\mu_1 - \mu_2 = 0\}.$$

Puisque l'on a fait l'hypothèse de modélisation que les deux populations sont de même variance σ^2 (hypothèse simplificatrice, mais qui sera également celle de l'analyse de la variance), la loi de la statistique d'intérêt, différence des moyennes empiriques des deux populations, est donc

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \sigma^2\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)\right),$$

où $\bar{X}_i = \sum_{j=1}^{n_i} X_{i,j}/n_i$, $i = 1, 2$. Notons alors S_1^2 et S_2^2 les variances empiriques sans biais des deux populations, autrement dit $(n_i - 1)S_i^2 = \sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_i)^2$, $i = 1, 2$. En appliquant le résultat de la proposition III.7 (ii) à chaque échantillon on obtient leurs lois : $(n_i - 1)S_i^2 \sim \sigma^2\chi^2(n_i - 1)$, $i = 1, 2$. Les statistiques S_1 et S_2 étant indépendantes, on en déduit (voir X.2.2, p. 246) que

$$U^2 = (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 \sim \sigma^2\chi^2(n_1 + n_2 - 2),$$

autrement dit $U^2/\sigma^2 \sim \chi^2(n_1 + n_2 - 2)$ et donc $U^2/(n_1 + n_2 - 2)$ est un estimateur sans biais de σ^2. La loi de $(\bar{X}_1 - \bar{X}_2)$ centrée réduite est

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim \mathcal{N}(0, 1),$$

et elle est indépendante de U^2 (puisque $\bar{X}_i$ est indépendante de S_i^2, $i = 1, 2$ par le théorème de Cochran). Le rapport de ces deux v.a. convenablement normalisé, voir la définition III.4, suit donc une loi de Student. Or dans ce rapport le paramètre "de nuisance" σ^2 s'élimine :

$$\frac{Z\sqrt{n_1+n_2-2}}{U/\sigma} = \frac{(\bar{X}_1 - \bar{X}_2 - (\mu_1-\mu_2))\sqrt{n_1+n_2-2}}{U\sqrt{1/n_1+1/n_2}} \sim t(n_1+n_2-2).$$

Cette statistique est utilisable pour construire des intervalles de confiance pour $\mu_1 - \mu_2$, selon la procédure décrite au chapitre I. Elle sert également à construire selon la procédure habituelle des tests relatifs à des hypothèses portant sur $\mu_1 - \mu_2$. Par exemple, le test de l'égalité des moyennes utilise la statistique de test

$$T = \frac{(\bar{X}_1 - \bar{X}_2)\sqrt{n_1+n_2-2}}{U\sqrt{1/n_1+1/n_2}}.$$

Sous H_0, T suit la loi de Student à n_1+n_2-2 degré de liberté, $t(n_1+n_2-2)$. Cette statistique "a tendance à s'éloigner de 0" lorsque $|\mu_1 - \mu_2|$ croît. Le test de niveau α rejette donc H_0 lorsque $|T| > t_{n_1+n_2-2,1-\alpha/2}$. Si l'on observe la valeur t pour la v.a. T, la p-valeur pour ce test est donnée par $\mathbb{P}(|T| > t)$, où $T \sim t(n_1+n_2-2)$.

Exemple III.12. Reprenons l'exemple III.1, et supposons que l'on ne dispose que des données relatives aux forêts 1 et 2. Le test de l'égalité des moyennes de ces forêts donne :

$$\bar{X}_1 = 25.97, \quad S_1 = 1.36, \quad \bar{X}_2 = 25.38, \quad S_2 = 1.77, \quad \text{et} \quad t = 0.9542.$$

Cela correspond à la p-valeur $\alpha_p = 0.3491$. On ne peut pas rejeter H_0 (égalité des deux moyennes) à tous les niveaux usuels (1%, 5%, 10%).

$\Diamond$

III.5 Analyse de la variance à un facteur

III.5.1 Comparaison de plusieurs échantillons gaussiens

La problématique qui conduit au modèle linéaire gaussien, et plus précisément à l'analyse de variance a déjà été introduite sur l'exemple III.1 des forêts. Voici un autre contexte donnant lieu à la même modélisation.

Une situation classique, par exemple en agronomie, consiste en l'évaluation du rendement de plusieurs variétés de blé et en la comparaison de ces rendements. Si l'on souhaite comparer k variétés de blé $(V_1, \ldots, V_k)$, et que l'on dispose de n parcelles "identiques" sur lesquelles on peut planter les différentes variétés, on peut réaliser l'expérience suivante :

1. regrouper au hasard les n parcelles en k groupes, d'effectifs $n_1, n_2, \ldots, n_k$, de sorte que $\sum_{i=1}^k n_i = n$, les n_i, $i = 1, \ldots, k$ n'étant pas nécessairement égaux ;
2. sur les n_i parcelles du i-ème groupe, planter la variété V_i et observer les rendements $X_{i,1}, \ldots, X_{i,n_i}$.

Il est alors raisonnable de supposer que le rendement observé sur une parcelle sur laquelle a été plantée la variété V_i est la somme d'un effet moyen dû à la variété de blé, et d'un ensemble de facteurs imprévisibles, les "perturbations", que l'on modélise par les réalisations

d'une v.a. centrée. Dans cet exemple, les perturbations aléatoires représentent les variations entre parcelles dues au sol, au climat, à la manière de semer... On admet fréquemment que ces perturbations sont convenablement modélisées par une loi gaussienne (car résultant de l'addition de nombreux effets), et que l'amplitude de ces perturbations est la même pour toutes les expériences (hypothèse dite *d'homoscédasticité*).

Ces considérations conduisent à poser

$$X_{i,j} = m_i + \varepsilon_{i,j}, \quad i = 1, \dots, k, \quad j = 1, \dots, n_i,$$

où m_i est le rendement moyen de la variété V_i, qui est inconnu, et les $\varepsilon_{i,j}$ sont des v.a. indépendantes de loi $\mathcal{N}(0, \sigma^2)$, où $\sigma^2 > 0$ est inconnu également. Autrement dit, les $(\varepsilon_{i,j})$ forment un n-échantillon de la loi $\mathcal{N}(0, \sigma^2)$.

Notation vectorielle

Rangeons les $\varepsilon_{i,j}$ en colonne par groupes, et notons ε le vecteur de $\mathbb{R}^n$ obtenu. Ce vecteur aléatoire

$$\varepsilon = \big(\varepsilon_{1,1}, \dots, \varepsilon_{1,n_1}, \varepsilon_{2,1}, \dots, \varepsilon_{2,n_2}, \dots\dots, \varepsilon_{k,1}, \dots, \varepsilon_{k,n_k}\big)^t$$

est un vecteur gaussien de loi $\mathcal{N}(0, \sigma^2 I_n)$ (où I_n est la matrice identité à n lignes et n colonnes, voir X.2.2, p.245). De même, on note X le vecteur des n observations rangées dans le même ordre :

$$X = \big(X_{1,1}, \dots, X_{1,n_1}, X_{2,1}, \dots, X_{2,n_2}, \dots\dots, X_{k,1}, \dots, X_{k,n_k}\big)^t,$$

et μ le vecteur des effets moyens associés aux observations :

$$\mu = \big(\underbrace{m_1, \dots, m_1}_{\times n_1}, \underbrace{m_2, \dots, m_2}_{\times n_2}, \dots\dots, \underbrace{m_k, \dots, m_k}_{\times n_k}\big)^t,$$

(on rappelle que ^{t} désigne la transposition, de sorte que X et μ sont des matrices à n lignes et 1 colonne). Le modèle s'écrit alors vectoriellement sous la forme "observation = moyenne + bruit" :

$$X = \mu + \varepsilon, \quad \mu \in \mathbb{R}^n, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

autrement dit $X \sim \mathcal{N}(\mu, \sigma^2 I_n)$, et le vecteur μ se trouve ici appartenir à un sous-espace vectoriel de $\mathbb{R}^n$ de dimension $k < n$, si bien que le paramètre inconnu du modèle, (μ, σ^2), consiste en $k + 1$ paramètres réels (avec $\sigma^2 > 0$).

Test

Plus que l'estimation de ces paramètres inconnus, le problème qui se pose souvent à l'agronome est celui du test de l'hypothèse nulle H_0 : "le facteur variété de blé n'a pas d'effet", aussi appelé test d'homogénéité des moyennes :

$$H_0 : m_1 = m_2 = \cdots = m_k \quad \text{contre} \quad H_1 : \text{"c'est faux"}.$$

L'agronome choisit cette hypothèse nulle car les données lui font penser qu'il y a au contraire une différence entre les variétés de blé. Il envisage donc de rejeter H_0 afin de s'efforcer d'en déduire ensuite la (les) variété(s) significativement la (les) meilleure(s), et ce rejet de H_0 a des conséquences importantes puisqu'il lui faudra ensuite, par exemple, distribuer à ses clients la variété jugée meilleure et retirer les autres. Le niveau α du test borne ainsi la probabilité que cette décision coûteuse soit prise à mauvais escient.

III.5.2 Cadre général du modèle linéaire gaussien

L'introduction générale et l'exemple précédent permettent de dégager le cadre formel ci-dessous. On effectue n observations $X = (X_1, \ldots, X_n)$, et chaque observation est l'addition d'un "effet moyen" et d'un "bruit". Si on considère le vecteur des observations $X \in \mathbb{R}^n$, le modèle s'écrit

$$X = \mu + \varepsilon,$$

et on fait les hypothèses (de modèle) suivantes :

- $M1$ l'effet moyen μ est inconnu et non observable, mais $\mu \in E$, sous espace vectoriel de $\mathbb{R}^n$, fixé et de dimension k ;
- $M2$ le vecteur aléatoire ε (non observable) a pour loi $\mathcal{N}(0, \sigma^2 I_n)$ et le paramètre $\sigma^2 > 0$ est inconnu.

Estimation

Ayant observé X, le point de E le plus proche de X est sa projection sur E, $X_E = \mu + \varepsilon_E$, qui est l'estimateur intuitif de μ. La projection sur l'orthogonal de E, $X - X_E = \varepsilon - \varepsilon_E$ ne contient pas d'information sur μ (elle est centrée) : c'est un indicateur de la dispersion des observations, qu'il est naturel d'utiliser pour estimer σ^2. On précise ceci dans le résultat suivant, conséquence directe du théorème III.6.

Proposition III.13. *On observe $X = \mu + \varepsilon$ avec les hypothèses $M1$ et $M2$. Alors on a :*

(i) X_E *est un estimateur sans biais de* μ.

(ii) $||X - X_E||^2/(n-k)$ *est un estimateur sans biais de* σ^2.

(iii) X_E *et* $X - X_E$ *sont indépendants.*

(iv) $||X_E - \mu||^2 \sim \sigma^2 \chi^2(k)$ *et* $||X - X_E||^2 \sim \sigma^2 \chi^2(n-k)$.

On peut montrer également que, pour tout vecteur $u \in \mathbb{R}^n$, le produit scalaire $\langle u, X_E \rangle$ est l'estimateur de $\langle u, \mu \rangle$ sans biais de variance minimum.

III.5.3 Répartition de la variation

On peut avoir une idée intuitive des quantités (ou résumés) issues des données qui sont pertinentes au regard de la décision à prendre (existe-t-il une différence entre les variétés de blé ?). Il est ainsi naturel de calculer les **moyennes empiriques par groupe** et de regarder si celles-ci sont "assez différentes" pour justifier le rejet de H_0. Cette différence est à apprécier relativement à la dispersion des données dans chaque groupe, c'est-à-dire (avec les hypothèses faites sur le modèle) relativement au paramètre σ^2 ou plutôt à son estimation.

Il est d'usage de noter les différentes moyennes empiriques intervenant en remplaçant par un point l'indice de sommation sur lequel se fait la moyenne. Les moyennes empiriques de chaque sous-échantillon (groupe) sont donc

$$X_{i,\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{i,j}, \quad i = 1, \ldots, k$$

et la moyenne empirique de toutes les observations (la "moyenne générale") est

$$X_{\cdot,\cdot} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{i,j}.$$

La **variation totale** des observations est

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - X_{\cdot,\cdot})^2,$$

où SST est l'acronyme utilisé par les logiciels de statistique anglo-saxons pour *Total Sum of Squares*. Cette variation totale est la somme de la **variation interclasses**, écarts entre les moyennes des groupes et la moyenne générale pondérés par les effectifs des groupes,

$$SSM = \sum_{i=1}^{k} n_i (X_{i,\cdot} - X_{\cdot,\cdot})^2,$$

et de la **variation intraclasses**, donc à l'intérieur des groupes,

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - X_{i,\cdot})^2.$$

Il est intuitivement clair que la variation interclasses sera d'autant plus grande que les moyennes empiriques des groupes seront éloignées les unes des autres. Cette variation interclasses est pour cette raison souvent appelée **variation expliquée par le modèle**, ce qui justifie l'acronyme SSM (*Sum of Squares for the Model*) souvent utilisé. De même, la variation intraclasses mesure sur chaque groupe les écarts entre les individus et la moyenne de leur groupe d'appartenance. Sous l'hypothèse d'homoscedasticité, il s'agit donc, à $n_i - 1$ près, de l'estimateur de la variance σ^2 calculé sur le groupe i, puis sommé sur tous les groupes. Remarquons que dans le cas $k = 2$ on retrouve la statistique U^2 utilisée en III.4 pour le cas de deux échantillons. Cette variation intraclasses est pour cette raison souvent appelée **variation résiduelle**, ou encore chez les anglo-saxons SSE pour *Sum of Squares of the Error*. Par rapport à ce que l'on a dit plus haut, on voit que plus la quantité SSM sera grande, plus on aura tendance à rejeter H_0, et ceci est à moduler avec la quantité SSE qui est liée à la dispersion. Nous verrons que ces quantités apparaissent dans la statistique du test d'homogénéité.

III.5.4 Test pour l'égalité des moyennes de k échantillons gaussiens

Le test de H_0 : "le facteur variété de blé n'a pas d'effet", que nous avions déjà exprimé comme le test de l'égalité des k moyennes des groupes, peut encore s'exprimer comme

$$H_0 = \{\mu \in H\} \quad \text{contre} \quad H_1 = \{\mu \in E \setminus H\},$$

où H est le sous-espace vectoriel de E auquel appartient μ sous l'hypothèse nulle, c'est-à-dire lorsque μ n'est constitué que d'une seule valeur moyenne commune à tous les groupes ;

autrement dit, si on note $\mathbf{1}_n$ le vecteur dont toutes les composantes valent 1, $H = \{\lambda \mathbf{1}_n, \lambda \in \mathbb{R}\}$ qui est de dimension 1.

Sous H_0, la projection de X sur H est $X_H = \mu + \varepsilon_H$, et donc la projection sur l'orthogonal de H dans E est $X_E - X_H = \varepsilon_E - \varepsilon_H$, projection du vecteur gaussien centré ε, d'où $||X_E - X_H||^2 \sim \sigma^2 \chi^2(k-1)$. Si au contraire on se place dans l'hypothèse alternative, il se produit un *décentrage* puisque $X_E - X_H = (\mu - \mu_H) + (\varepsilon_E - \varepsilon_H)$. Il est donc naturel d'utiliser $||X_E - X_H||^2$ comme statistique de test puisqu'elle a tendance à croître avec $||\mu - \mu_H||^2$.

Malheureusement la loi de $||X_E - X_H||^2$ n'est connue sous H_0 qu'à σ^2 près. On procède donc comme pour le test de Student, en remplaçant σ^2 par son estimateur

$$\frac{||X - X_E||^2}{n-k}$$

qui est sans biais quelle que soit la localisation de μ dans E. Techniquement, ceci revient à utiliser la loi d'un rapport de v.a. convenablement normalisé et dans lequel le paramètre σ disparaît : les v.a. $||X - X_E||^2/\sigma^2$ et $||X_E - X_H||^2/\sigma^2$ sont indépendantes et suivent chacune une loi du chi-deux centrée (sous H_0 seulement pour $||X_E - X_H||^2/\sigma^2$), donc (voir la définition III.5)

$$F = \frac{||X_E - X_H||^2/k - 1}{||X - X_E||^2/n - k} \sim \mathcal{F}(k-1, n-k) \quad \text{sous } H_0.$$

Sous l'hypothèse alternative $||X_E - X_H||^2/\sigma^2$ suit une loi du chi-deux décentrée, de décentrage $||\mu - \mu_H||^2/\sigma^2$. Le numérateur $||X_E - X_H||^2/k - 1$, comme F, ont tendance à croître quand le paramètre de décentrage augmente (voir X.2.2, p. 246) : F suit une loi de Fisher décentrée. On vérifie que le décentrage, et donc cette statistique de test, tend vers $+\infty$ lorsque $n \to \infty$. Le test de niveau α conduit ainsi à

$$\text{rejeter } H_0 \text{ dès que } F > \mathcal{F}_{k-1,n-k,1-\alpha},$$

où $\mathcal{F}_{k-1,n-k,1-\alpha}$ est le quantile d'ordre $(1-\alpha)$ de la loi de Fisher $\mathcal{F}(k-1, n-k)$.

Pour appliquer ce test, calculons la projection de X sur E. Considérons la base de E donnée par $(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_k})$, où $\mathbf{1}_{n_i}$ est le vecteur de $\mathbb{R}^n$ ayant n_i coordonnées égales à 1 et toutes les autres nulles, celles égales à 1 correspondant au groupe i dans la notation vectorielle du modèle d'analyse de la variance, autrement dit :

$$\begin{aligned}
\mathbf{1}_{n_1} &= (\underbrace{1, \ldots, 1}_{\times n_1}, 0, \ldots\ldots, 0)^t, \\
\mathbf{1}_{n_2} &= (\underbrace{0, \ldots, 0}_{\times n_1}, \underbrace{1, \ldots, 1}_{\times n_2}, 0, \ldots\ldots, 0)^t, \\
&\ldots \\
\mathbf{1}_{n_k} &= (0, \ldots\ldots, 0, \underbrace{1, \ldots, 1}_{\times n_k})^t,
\end{aligned}$$

de sorte que $\mu = \sum_{i=1}^k m_i \mathbf{1}_{n_i}$, et $X_E = \sum_{i=1}^k X_{i,\cdot} \mathbf{1}_{n_i}$. On déduit du point (i) de la proposition III.13 que $X_{i,\cdot}$ est un estimateur sans biais de m_i pour $i = 1, \ldots, k$ (remarquons que l'on pouvait déduire ce résultat directement de l'étude faite au chapitre II sur le modèle gaussien

en raisonnant sous-échantillon par sous-échantillon). Il s'ensuit que la norme de la projection sur l'orthogonal de E est la résiduelle :

$$||X - X_E||^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - X_{i,\cdot})^2 = SSE.$$

D'autre part, $X_H = X_{\cdot,\cdot}\mathbf{1}_n$, d'où

$$||X_E - X_H||^2 = \sum_{i=1}^{k} n_i (X_{i,\cdot} - X_{\cdot,\cdot})^2 = SSM.$$

Remarquons enfin que $SST = ||X - X_H||^2$, et que $SST = SSM + SSE$ est une conséquence du théorème de Pythagore.

Table d'analyse de la variance

Les résultats du test de l'appartenance de μ à H, quand le modèle stipule son appartenance à E, sont habituellement résumés dans une table d'analyse de la variance, dont le format est à peu de choses près toujours le même quel que soit le logiciel de statistique utilisé. Dans le cas du test d'homogénéité, cette table contient les éléments suivants :

Variabilité	SS	DF	MS	Fisher	p-valeur
Interclasse	$\|\|X_E - X_H\|\|^2$	$k-1$	$MSM = \dfrac{\|\|X_E - X_H\|\|^2}{k-1}$	$f = \dfrac{MSM}{MSE}$	$\mathbb{P}(F > f)$
Intraclasse	$\|\|X - X_E\|\|^2$	$n-k$	$MSE = \dfrac{\|\|X - X_E\|\|^2}{n-k}$		
Totale	$\|\|X - X_H\|\|^2$	$n-1$			

Les acronymes utilisés ici sont encore une fois ceux des logiciels de statistique anglo-saxons : SS pour *Sum of Squares*, DF pour *Degrees of Freedom* et MS pour *Mean Squares* (MS *of the Model* pour le carré moyen associé à la variation interclasses et ME *of the Error* pour celui associé à la variation intraclasses). La p-valeur est $\mathbb{P}(F > f)$, calculée lorsque F suit la loi sous l'hypothèse nulle $\mathcal{F}(k-1, n-k)$.

Exemple III.14. Si on reprend l'exemple III.1, on obtient la table suivante pour l'analyse de la variance des hauteurs d'arbre dans les 3 forêts :

`Variabilité`	`SS`	`DF`	`MS`	`Fisher`	p-`valeur`
`Interclasse`	49.25	2	24.62	7.18	0.0025
`Intraclasse`	116.566	34	3.43		
`Totale`	165.82	36			

On rejette donc H_0 au niveau usuel de 5%, ainsi que pour tout niveau $\alpha \geq 0.0025$. L'effet du facteur "forêt" est significatif. ◊

Généralisation

Cette technique se généralise à tous les modèles du type

$$X = \mu + \varepsilon, \quad \mu \in \mathbb{R}^n, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$

où il est stipulé que $\mu \in E$ (sous-espace vectoriel de $\mathbb{R}^n$, de dimension $k < n$) et où l'hypothèse nulle est $\mu \in H$ (sous-espace vectoriel de E, de dimension $h < k$) ; il suffit de remplacer, dans la table d'analyse de la variance ci-dessus, $k - 1$ par $k - h$ et alors F suit, sous l'hypothèse nulle, la loi $\mathcal{F}(k - h, n - k)$.

III.6 Régression linéaire multiple

Rappel de la problématique

La problématique a été introduite sur un exemple en III.1. Reprenons-la avec une autre situation. Il s'agit ici de modéliser un phénomène aléatoire observé par une combinaison linéaire ou affine de *variables explicatives*, dont les valeurs sont déterministes et connues pour chaque expérience (ou observation) réalisée. Par exemple, si l'on souhaite "expliquer" la durée d'une certaine maladie (en jours) après l'admission de patients à l'hôpital, on peut penser que cette durée est liée à certaines variables quantitatives (i.e., à valeur numériques). On relèvera par exemple les nombres de bactéries de certains types présentes dans l'organisme du patient à son arrivée, ainsi que des indicateurs de son état général (poids, température,...).

Si l'on dispose de n observations de ces variables explicatives ainsi que de la variable à expliquer (l'observation de la variable à expliquer est donc faite a posteriori dans cet exemple, lorsque les n patients ont quitté l'hopital) on peut étudier la pertinence de cette modélisation linéaire. Il est possible de tester la significativité du modèle, et celle de certaines variables explicatives. Il est possible aussi d'estimer les liens entre variables explicatives et variable à expliquer et éventuellement de faire ensuite de la *prédiction*, c'est à dire ici d'estimer la durée d'hospitalisation d'un nouveau patient à partir de la connaissance des valeurs des variables explicatives dans son cas.

III.6.1 Définition du modèle

On observe un phénomène aléatoire X et l'on suppose ce phénomène influencé par p variables explicatives ou *régresseurs*, $R^1, \ldots, R^p$. Parfois, X est aussi appelée la *variable dépendante*, et $R^1, \ldots, R^p$ les *variables indépendantes* (où encore, en Économétrie, X est appelée la *variable exogène* et les R^j les *variables endogènes*).

On réalise n observations, autrement dit $X = (X_1, \ldots, X_n)$, et on note $R_i^1, \ldots, R_i^p$ les conditions expérimentales pour la i-ème observation X_i, c'est à dire les valeurs (déterministes) des p régresseurs lors de l'expérience i. On fait comme on l'a dit l'hypothèse d'une relation linéaire ou affine entre les régresseurs et la variable à expliquer X et, comme en analyse de la variance, on suppose observer la somme de l'effet de ces régresseurs et d'un ensemble de perturbations non observables, que l'on résume par un "bruit" gaussien centré. Ce modèle s'écrit ainsi

$$X_i = \sum_{j=1}^{p} \alpha_j R_i^j + \varepsilon_i, \quad \text{ou bien} \quad X_i = \beta + \sum_{j=1}^{p} \alpha_j R_i^j + \varepsilon_i, \quad i = 1, \ldots, n,$$

où $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ est un n-échantillon de la loi $\mathcal{N}(0, \sigma^2)$ (l'hypothèse d'homoscédasticité est présente ici aussi, puisque σ^2 ne dépend pas de i). Les paramètres inconnus à estimer sont $(\beta, \alpha_1, \ldots, \alpha_p, \sigma^2)$ dans le cas affine (on retire β dans le cas linéaire sans constante).

Notation vectorielle

Considérons par exemple le cas affine, et posons

$$M = \begin{bmatrix} 1 & R_1^1 & \cdots & R_1^p \\ \vdots & \vdots & \ldots & \vdots \\ 1 & R_n^1 & \cdots & R_n^p \end{bmatrix} = \left[\mathbf{1}_n\, R^1 \cdots R^p\right],$$

la matrice $n \times (p+1)$ des régresseurs (la colonne de 1, $\mathbf{1}_n$, étant considérée comme un régresseur particulier lorsqu'elle figure dans le modèle). Posons aussi $\theta \in \mathbb{R}^{p+1}$ le paramètre du modèle, où $\theta = (\beta, \alpha_1, \ldots, \alpha_p)^t$. Le modèle s'écrit vectoriellement :

$$X = M\theta + \varepsilon, \quad \text{avec } M\theta \in E \text{ et } \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

où $E = \{Mu, u \in \mathbb{R}^{p+1}\}$ est le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de M. Ce modèle s'inscrit ainsi dans le cadre général du modèle linéaire gaussien décrit en III.5.2, avec adoption des hypothèses $M1$ et $M2$ qui y ont été faites.

On suppose que la dimension de E est $p+1$, c'est à dire que les p régresseurs et $\mathbf{1}_n$ sont linéairement indépendants, ou ce qui revient au même que $\text{rang}(M) = p+1$, ou encore que la matrice symétrique $M^T M$ est elle-même de rang $p+1$. Cette hypothèse n'est pas une réelle perte de généralité puisque, si elle n'est pas vérifiée, cela signifie que l'un des régresseurs est combinaison linéaire des autres ; il n'apporte alors pas d'explication supplémentaire et il suffit de le retirer.

Exemple III.15. La régression simple. C'est la situation où l'on dispose d'un seul régresseur ($p = 1$) que nous notons simplement R. Le modèle s'écrit

$$X_i = \beta + \alpha R_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

ce qui revient à dire que $X_i \sim \mathcal{N}(\beta + \alpha R_i, \sigma^2)$. On visualise ce modèle dans l'espace des variables (R, X) par le fait que les observations "tombent" dans un "tunnel gaussien" d'amplitude σ le long de la droite d'équation $x = \beta + \alpha r$. L'exemple III.2 des données de pluie est de ce type.

$\Diamond$

III.6.2 Estimation

On applique dans ce cadre les résultats de la proposition III.13. La projection de X sur E est l'estimateur sans biais de $M\theta$. Il s'écrit $X_E = M\hat{\theta}$, où $\hat{\theta} \in \mathbb{R}^{p+1}$ est l'estimateur sans biais de θ. Il est tel que $X - M\hat{\theta}$ est orthogonal à tout vecteur de E, autrement dit pour tout vecteur $u \in \mathbb{R}^{p+1}$, $\langle Mu, X - M\hat{\theta}\rangle = 0$, ce qui donne

$$\boxed{\hat{\theta} = (M^t M)^{-1} M^t X.}$$

Remarquons que, si l'on note P le projecteur sur E (donc tel que $X_E = PX$), celui-ci s'écrit $P = M(M^tM)^{-1}M^t$. La résiduelle est $||X - X_E||^2 = \langle X, X - X_E\rangle = X^t(I-P)X$, soit

$$||X - X_E||^2 = X^t\left[I - M(M^tM)^{-1}M^t\right]X.$$

D'après le point (iv) de la proposition III.13, $||X - X_E||^2 \sim \sigma^2\chi^2\big(n-(p+1)\big)$, et l'on estime (sans biais) la variance par

$$\boxed{\hat{\sigma}^2 = \frac{||X - X_E||^2}{n-(p+1)}.}$$

Remarque : dans le cas de la régression sans constante, il suffit de retirer la colonne $\mathbf{1}_n$ de M et de remplacer $p+1$ par p.

Variances des estimateurs

On déduit immédiatement de l'expression de $\hat{\theta}$ que sa matrice de variances-covariances (voir X.1.3) est

$$\mathrm{Var}(\hat{\theta}) = \sigma^2(M^tM)^{-1}.$$

Exemple III.16. La régression simple (suite de l'exemple III.15).

Il est facile de mener les calculs "à la main" dans le cas de la régression simple. La matrice des régresseurs est $M = [\mathbf{1}_n\, R]$, d'où

$$(M^tM)^{-1} = \frac{1}{n\sum_{i=1}^n (R_i - \bar{R})^2}\begin{bmatrix} \sum_{i=1}^n R_i^2 & -\sum_{i=1}^n R_i \\ -\sum_{i=1}^n R_i & n \end{bmatrix},$$

et le calcul de $\hat{\theta}$ donne

$$\hat{\alpha} = \frac{\mathrm{Cov}(R,X)}{\mathrm{Var}(R)}, \qquad \hat{\beta} = \bar{X} - \hat{\alpha}\bar{R},$$

où $\bar{R} = \sum_{i=1}^n R_i/n$ est la moyenne empirique de R, et

$$\mathrm{Var}(R) = \frac{1}{n}\sum_{i=1}^n (R_i - \bar{R})^2, \quad \mathrm{Cov}(R,X) = \frac{1}{n}\sum_{i=1}^n (R_i - \bar{R})(X_i - \bar{X}) = \frac{1}{n}\sum_{i=1}^n R_iX_i - \bar{R}\bar{X},$$

sont les variances et covariances empiriques (qui ont le sens de mesures descriptives ici puisque R n'est pas aléatoire). On peut remarquer que ces estimateurs coïncident avec les **estimateurs des moindres carrés** de la droite de régression de X sur R, c'est à dire la pente et la constante de la droite d'équation $X = b + aR$ qui minimisent les carrés des écarts $\sum_{i=1}^n (X_i - b - aR_i)^2$.

On déduit immédiatement de l'expression de $\mathrm{Var}(\hat{\theta})$ l'expression des variances de $\hat{\alpha}$ et $\hat{\beta}$, ainsi que la covariance entre les deux estimateurs (ils ne sont pas indépendants). Comme ils sont des estimateurs sans biais des paramètres qu'ils estiment, et suivent des lois gaussiennes (car combinaisons linéaires de X), on a finalement :

$$\hat{\alpha} \sim \mathcal{N}\left(\alpha\,,\, \frac{\sigma^2}{\sum_{i=1}^n (R_i - \bar{R})^2}\right), \qquad \hat{\beta} \sim \mathcal{N}\left(\beta\,,\, \frac{\sigma^2\sum_{i=1}^n R_i^2}{n\sum_{i=1}^n (R_i - \bar{R})^2}\right).$$

Le projeté est $X_E = \hat{\beta}\mathbf{1}_n + \hat{\alpha}R$, et on peut écrire directement la résiduelle

$$SSE = ||X - X_E||^2 = \sum_{i=1}^{n}(X_i - \hat{\beta} - \hat{\alpha}R_i)^2,$$

écarts entre les valeurs observées et les valeurs *ajustées* par le modèle. Elle suit la loi $\sigma^2\chi^2(n-2)$, et l'estimateur sans biais de la variance est $||X - X_E||^2/(n-2)$ qui est indépendant de $\hat{\theta}$. La connaissance des lois des estimateurs de (β, α), qui dépendent de σ^2, ainsi que de la loi de l'estimateur de σ^2 et cette propriété d'indépendance permet de construire des intervalles de confiance ou des tests pour β et α analogues aux intervalles de confiance et tests de Student construits en III.3.6.

$\Diamond$

III.6.3 Test de l'utilité des régresseurs

Dans le modèle $X = M\theta + \varepsilon$ avec p régresseurs et la constante (cas affine), on souhaite souvent tester l'utilité d'une partie des régresseurs, autrement dit une hypothèse nulle de la forme

$$H_0 = \{\text{“}R^{q+1}, \ldots, R^p \text{ sont inutiles”}\} \quad \text{contre} \quad H_1 = \{\text{“c'est faux”}\},$$

où $1 \leq q < p$, et où on a éventuellement effectué une permutation de l'ordre des régresseurs. La contre-hypothèse se comprend comme H_1 : “l'un des $R^j, q+1 \leq j \leq p$ au moins est utile”. L'hypothèse nulle, si elle n'est pas rejetée, permet alors de simplifier le modèle de régression, en ne conservant qu'une partie des variables qui étaient a priori explicatives. Les hypothèses peuvent se reformuler comme

$$H_0 = \{\alpha_j = 0, j = q+1, \ldots, p\} \quad \text{contre} \quad H_1 = \{\text{il existe au moins un } \alpha_j \neq 0\},$$

autrement dit comme l'appartenance, sous H_0, de l'effet moyen à un sous-espace vectoriel de E de dimension plus petite, ce qui nous ramène à la méthode employée pour le test d'homogénéité en analyse de la variance (voir le passage “Généralisation” en III.5.4, p. 49). En effet, le sous-modèle associé à H_0 s'écrit $X_i = \beta + \sum_{j=1}^{q}\alpha_j R_i^j + \varepsilon_i$, $i = 1, \ldots, n$, ou vectoriellement (en indiçant par 0 les quantités qui diffèrent sous l'hypothèse nulle)

$$X = M_0\theta_0 + \varepsilon, \quad M_0 = [\mathbf{1}_n\, R^1 \cdots R^q], \quad \theta_0 = (\beta, \alpha_1, \ldots, \alpha_q)^t,$$

et donc $M_0\theta_0 \in H = \{M_0 w : w \in \mathbb{R}^{q+1}\}$, où H est de dimension $q+1$. On teste ainsi

$$H_0 = \{M\theta \in H\} \quad \text{contre} \quad H_1 = \{M\theta \in E \setminus H\}.$$

Sous H_0, on estime l'effet moyen par la projection de X sur H c'est à dire $X_H = M_0\hat{\theta}_0$ avec $\hat{\theta}_0 = (M_0^t M_0)^{-1}M_0^t X$. On procède ensuite comme pour le test d'homogénéité : sous H_0, $||X_E - X_H||^2 \sim \sigma^2\chi^2(p-q)$ mais σ est inconnu. On prend le rapport avec la résiduelle normalisée qui, elle, suit toujours la loi $\chi^2(n-p-1)$, pour construire la statistique de test

$$F = \frac{||X_E - X_H||^2/(p-q)}{||X - X_E||^2/(n-p-1)} \sim \mathcal{F}(p-q, n-p-1) \quad \text{sous } H_0.$$

La loi du χ^2 du numérateur (normalisé convenablement) se décentre sous l'hypothèse alternative, d'où le test au niveau γ qui conduit à

$$\text{rejeter } H_0 \text{ dès que } F > \mathcal{F}_{p-q,n-p-1,1-\gamma}.$$

Table d'analyse de la variance pour le modèle de régression

Lorsqu'ils traitent un modèle de régression, la plupart des logiciels de statistique calculent les estimateurs des paramètres et effectuent des tests individuels de nullité de ces paramètres (p tests de Student de $H_0 : \alpha_j = 0$, $j = 1, \ldots, p$, fondés sur les lois que nous avons donné plus haut). Ils fournissent également une *table d'analyse de variance associée au modèle de régression*. Il s'agit du résultat du test de Fisher pour l'hypothèse nulle "pas de modèle de régression", autrement dit "aucun régresseur n'est significatif". C'est la réalisation du test ci-dessus pour $H = \{\lambda \mathbf{1}_n, \lambda \in \mathbb{R}\}$.

Coefficient de détermination

Lorsque il y a une constante dans la régression, on appelle coefficient de détermination, ou R^2, le nombre

$$R^2 = \frac{||X_E - \bar{X}\mathbf{1}_n||^2}{||X - \bar{X}\mathbf{1}_n||^2} \in [0, 1].$$

C'est un indicateur de la "qualité" de la régression : plus le R^2 est proche de 1, meilleure est l'adéquation du modèle aux données (on parle aussi de pourcentage de la variabilité expliquée par le modèle de régression).

Remarquons que pour le test de Fisher associé à l'hypothèse nulle "aucun régresseur n'est significatif", le sous-espace vectoriel H est celui engendré par $\mathbf{1}_n$ ce qui entraîne que $X_H = \bar{X}\mathbf{1}_n$. Dans ce cas il existe un lien simple entre le R^2 et la statistique du test de Fisher :

$$F = \frac{(n - (p + 1))}{p} \frac{R^2}{1 - R^2}.$$

Exemple III.17. La régression simple, (suite et fin de l'exemple III.16).

Nous terminons l'étude détaillée de la régression simple avec le test de non effet du seul régresseur présent dans le modèle :

$$H_0 = \{\alpha = 0\} \quad \text{contre} \quad H_1 = \{\alpha \neq 0\}.$$

Remarquons que, ici, il est possible de construire ce test de deux manières : à partir de la loi de $\hat{\alpha}$ en utilisant la loi de Student (qui provient, rappelons-le, de l'obligation d'estimer σ^2 par la résiduelle), ou bien à partir du test de Fisher. On vérifie que les statistiques de ces deux tests sont liées par la relation $F = T^2$, et ils donnent la même p-valeur. Nous allons utiliser ici la seconde méthode.

Sous H_0, le modèle est simplement $X = \beta \mathbf{1}_n + \varepsilon$ (il s'agit donc d'un n-échantillon de $\mathcal{N}(\beta, \sigma^2)$), et $X_H = \bar{X}\mathbf{1}_n$. Nous avons déjà précisé l'expression de la résiduelle dans ce cas. La "somme des carrés du modèle" est

$$SSM = ||X_E - X_H||^2 = \sum_{i=1}^{n} (\hat{\beta} + \hat{\alpha} R_i - \bar{X})^2,$$

et la statistique de test

$$F = \frac{||X_E - X_H||^2}{||X - X_E||^2/(n-2)} \sim \mathcal{F}(1, n-2) \quad \text{sous } H_0.$$

On rejette donc H_0 au niveau γ si $F > \mathcal{F}_{1,n-2,1-\gamma}$. Enfin, si on a observé la valeur f de la statistique F, la p-valeur de ce test est $\mathbb{P}(F > f)$, où $F \sim \mathcal{F}(1, n-2)$.

Dans le cas de la régression simple, le coefficient de détermination $R^2 = SSM/SST$ est aussi le carré du coefficient de corrélation entre X et R.

$\Diamond$

Exemple III.18. Si on reprend l'exemple III.2, on obtient les résultats suivants :

– Les estimations des paramètres valent : $\hat{\alpha} = 4.55$ et $\hat{\beta} = -128.07$. Sur le graphique (fig. III.2) on a représenté la **droite de régression**.

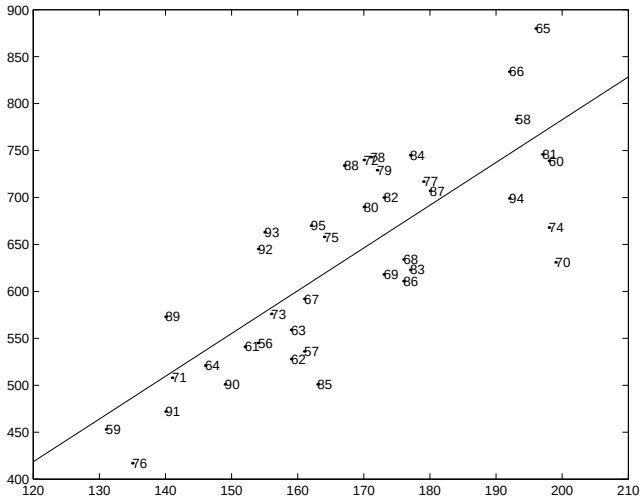


FIG. III.2 – Droite de régression sur le nuage de points

– Les intervalles de confiance de Student sont : $I_{0.05}(\alpha) = [3.40; 5.70]$ et $I_{0.05}(\beta) = [-322; 66]$

– Le calcul du R^2 et du test de $H_0 = \{\alpha = 0\}$ donnent :

R^2	Fisher	p-valeur
0.6294	64.52	$< 10^{-4}$

donc on rejette clairement H_0.

$\Diamond$

III.6.4 Analyse des résidus et des observations

Il peut arriver que le modèle soit influencé par des observations atypiques ou suspectes. Cela peut entrainer des interprétations et des conclusions erronées. Un moyen raisonnable pour se prémunir de ce type de problèmes est de réaliser une analyse des résidus et des observations.

Un résidu e_i correspond à l'écart entre l'observation X_i et son estimation $\hat{X}_i$ (avec $\hat{X} = M\hat{\theta} = M(M^tM)^{-1}M^tX$) : $e_i = X_i - \hat{X}_i$. Cependant, la valeur de ces résidus simples dépendent de l'échelle des valeurs prises par la variable à expliquer. Pour y remédier, il suffit de réduire chaque résidu par l'écart-type des résidus : (il n'y a pas besoin de les centrer, car

leur moyenne est nulle par construction du critère des moindres carrés). Ce nouveau résidu (nommé résidu réduit ou standardisé) vaut : $e_{si} = e_i/\hat{\sigma}$.

Cependant ces résidus réduits nous renseignent seulement au niveau individuel. En d'autres termes, ils ne nous permettent pas de les comparer entre eux. Pour cela, il faudrait obtenir un autre résidu éliminant l'effet individuel. La réponse est donnée par une mesure nommée levier ("leverage" en anglais) qui permet d'identifier des observations influentes dans le sens des variables candidates à l'explication, alors l'effet de levier peut avoir une forte influence sur les résultats. Pour une observation particulière (R_i, X_i), l'impact de celle-ci sur la droite des moindres carrés peut être déterminé grâce à la formule de prédiction de X_i par R_i : $\hat{X} = M\hat{\theta} = M(M^tM)^{-1}M^tX$. On note $H = M(M^tM)^{-1}M^t$ la matrice d'estimation (appelée aussi matrice "hat"). On pose $H = (H_{i,j}; 1 \leq i, j \leq n)$. De cette expression, on peut constater que chaque valeur $\hat{X}_i$ correspond à une combinaison linéaire pondérée des observations $X = (X_1, ..., X_n)$, où chaque H_{ij} est le poids attribué à chaque X_i.

Exemple III.19. La régression simple (suite de l'exemple III.16). On a

$$H_{ij} = 1/n + \frac{(R_i - \bar{R})(R_j - \bar{R})}{\sum_{j=1}^{n}(R_j - \bar{R})^2}. \tag{III.3}$$

En particulier, on voit que plus un point R_i est loin du centre de gravité des variables explicatives (la moyenne $\bar{R}$), plus le poids correspondant à X_j est élevé, et influence donc la détermination de $\hat{X}_i$. Par contre, les estimations des X's seront d'autant moins influencées que leur R associé sera proche de $\bar{R}$. Il est clair également que chaque X_j a un impact sur $\hat{X}_i$, c'est-à-dire qu'une valeur extrême de R_j influencera tous les $\hat{X}_i$. ◊

La droite des moindres carrés est très sensible aux points extrêmes. On dit qu'elle n'est pas robuste. Il suffit de le vérifier en dimension un, pour s'en convaincre.

En pratique, il est plutôt intéressant de considérer H_{ii} (éléments de la diagonale de la matrice H) qui traduit l'influence de X_i sur $\hat{X}_i$.

On se place à partir de maintenant dans le cas de la régression simple, cf les exemples III.15, III.16 et III.19. On a

$$H_{ij} = \frac{1}{n} + \frac{(R_i - \bar{R})^2}{\sum_{j=1}^{n}(R_j - \bar{R})^2}.$$

Quand le levier H_{ii} est élevé, $\hat{X}_i$ est plus sensible à des changements en X_i que lorsque H_{ii} est relativement petit (R_i est proche de $\bar{R}$). En effet, ceci se concrétise en récrivant le coefficient de la pente $\hat{\alpha}$ sous la forme suivante :

$$\hat{\alpha} = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(X_i - \bar{X})}{\sum_{i=1}^{n}(R_i - \bar{R})^2} = \sum_{i=1}^{n} \frac{(R_i - \bar{R})^2}{\sum_{j=1}^{n}(R_j - \bar{R})^2} \frac{(X_i - \bar{X})}{(R_i - \bar{R})} = \sum_{i=1}^{n}(H_{ii} - \frac{1}{n})\frac{(X_i - \bar{X})}{(R_i - \bar{R})}.$$

Par conséquent $\hat{\alpha}$ revient à une moyenne pondérée des pentes individuelles : $\dfrac{(X_i - \bar{X})}{(R_i - \bar{R})}$ par le levier $H_{ii} - \dfrac{1}{n}$.

Ce levier va nous permettre de construire un résidu dans lequel l'effet individuel a été retiré. Comme nous l'avons vu précédemment, nous pouvons écrire matriciellement $\hat{X} = HX$. Par conséquent, le vecteur des résidus e_i peut etre réécrit comme suit : $X - HX = (I - H)X$ où I est la matrice identité. On peut montrer que la variance des e_i (on a $\mathbb{E}(e_i) = 0$, par construction des moindres carrés) : $\text{Var}(e_i) = \hat{\sigma}^2(1 - H_{ii})$. Par conséquent, le nouveau résidu réduit estimé, nommé résidu approximé (ou résidu studentisé), vaut :

$$\tilde{e}_{si} = \frac{e_i}{\hat{\sigma}\sqrt{1 - H_{ii}}}.$$

Une observation i sera mal reconstituée par le modèle, si le résidu approximé est trop grand en valeur absolue. En identifiant la loi des $\tilde{e}_{si}$ à une loi Normale centrée réduite, on choisit couramment un seuil critique de 2 écarts-types, mais on peut prendre aussi 2,5 ou 3 écarts-types.

Bien que plus correct que les deux précédents résidus, le résidu approximé ne tient pas compte du fait que chaque e_i n'est pas indépendant de $\hat{\sigma}^2$. D'ailleurs, les résidus approximés sont parfois appelés : résidus studentisés internes. Pour pallier cet inconvénient, un autre résidu a été introduit et est nommé résidu studentisé ou Rstudent (parfois résidu studentisé externe, car l'effet de la dépendance discutée plus haut est gommé). Pour cela, il suffit d'éliminer le résidu e_i correspondant à l'observation i, mais pas n'importe comment. En effet, on calcule tout d'abord une nouvelle régression sans l'observation i, alors on obtient une nouvelle variance corrigée des résidus (indépendante de l'observation i) :

$$\hat{\sigma}^2_{(i)} = ((n-2)\hat{\sigma}^2 e_i^2/(1 - H_{ii}))/(n-3).$$

En fait, cette formule permet de ne pas recalculer autant de régressions qu'il y a d'observations dans notre échantillon, et de tenir compte sur les n couples estimés des coefficients $(\hat{\beta}, \hat{\alpha})$, de l'élimination individuelle d'un couple d'observations (X_i, R_i). Enfin, on divise par $(n-3)$, car on a enlevé une observation. Le Rstudent, pour chaque individu i, prend la forme suivante :

$$e_{ti} = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1 - H_{ii}}}.$$

L'avantage de cette nouvelle mesure est l'on peut effectuer un test statistique sur chaque RStudent. En effet, si l'analyste veut vérifier que l'observation i est un point non influent, alors sous cette hypothèse, la réalisation d'un Rstudent doit être issue d'une loi de Student à $(n-3)$ degrés de liberté : H_0 : i n'est pas atypique contre H_1 : i est atypique.

On choisit alors un risque d'erreur α de rejeter H_0, alors qu'elle est vraie. Par exemple, pour $\alpha = 0.05$, si $t_{n-3}(0.975)$ est le quantile d'ordre $1-\alpha/2$ de la loi de student de paramètre $n-3$, alors la règle de décision est la suivante : si $|e_{ti}| > |t_{n-3}(0.975)|$ alors à un seuil 5%, l'observation i peut être considérer comme influençant fortement la droite des moindres carrés. Cela permet aussi de construire un intervalle, hors duquel, les observations sont atypiques au modèle : $[-t_{n-3}(0.975), t_{n-3}(0.975)]$.

Un autre indicateur nommé DFfits permet aussi de mesurer l'influence globale de chaque observation (X_i, R_i) sur le modèle car ils sont plus riches que le levier qui ne permet que de détecter l'influence des X_i. Il est défini par

$$\text{DFfits}_i = \frac{\hat{X}_i - \hat{X}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{H_{ii}}},$$

où $\hat{X}_{i(i)}$ correspond à l'estimation de la réponse X_i avec les coefficients de la régression calculés sans le point i. Celle-ci peut être réécrite de la façon suivante :

$$\text{DFfits}_i = \sqrt{\frac{h_i}{(1-h_i)}} e_{ti}.$$

Plusieurs règles empiriques sont également proposées pour déterminer les observations influentes : si un DFfits est élevé alors l'observation i sera jugé influençant la pente de la droite des moindres carrés.

Ces indicateurs doivent être utilisés avec prudence. Ils représentent seulement des aides à la décision.

III.7 Résumé

III.7.1 Le modèle gaussien à variance connue

1. Modèle : $(X_k, 1 \leq k \leq n)$ suite de v.a. i.i.d. de loi gaussienne à variance, σ_0^2, connue : $\mathcal{P} = \{\mathcal{N}(\mu, \sigma_0^2), \mu \in \mathbb{R}\}$.
2. $H_0 = \{\mu = \mu_0\}$, $H_1 = \{\mu \neq \mu_0\}$, avec $\mu_0 \in \mathbb{R}$.
3. Statistique de test : $\zeta_n = \sqrt{n}\dfrac{\bar{X}_n - \mu_0}{\sigma_0}$.
4. Loi sous H_0 : $\mathcal{N}(0,1)$.
5. Loi sous H_1 : gaussienne réduite décentrée.
6. Région critique : $W_n = \{|\zeta_n| \geq a\}$.
7. Niveau exact α : $a = \phi_{1-\alpha/2}$, où $\phi_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de $\mathcal{N}(0,1)$.
8. Test convergent.
9. p-valeur : $\mathbb{P}(|G| \geq |\zeta_n^{\text{obs}}|)$ où G de loi $\mathcal{N}(0,1)$.
10. Variante : $H_0 = \{\mu \leq \mu_0\}$, $H_1 = \{\mu > \mu_0\}$. Même statistique de test. Région critique : $W_n = \{\zeta_n \geq a\}$. Niveau exact α : $a = \phi_{1-\alpha}$. Test convergent. p-valeur : $\mathbb{P}(G \geq \zeta_n^{\text{obs}})$.

III.7.2 Le modèle gaussien à variance inconnue

1. Modèle : $(X_k, 1 \leq k \leq n)$ suite de v.a. i.i.d. de loi gaussienne : $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\}$.
2. $H_0 = \{\mu = \mu_0\}$, $H_1 = \{\mu \neq \mu_0\}$, avec $\mu_0 \in \mathbb{R}$.
3. Statistique de test : $\zeta_n = \sqrt{n}\dfrac{\bar{X}_n - \mu_0}{S_n}$.
4. Loi sous H_0 : Student de paramètre $n-1$.
5. Comportement asymptotique sous H_1 : ζ_n converge p.s. vers $-\infty$ ou $+\infty$.
6. Région critique : $W_n = \{|\zeta_n| \geq a\}$.
7. Niveau exact α : $a = t_{n-1,1-\alpha/2}$, où $t_{n-1,1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student de paramètre $n-1$.
8. Test convergent.

9. p-valeur : $\mathbb{P}(|T| \geq \zeta_n^{\text{obs}})$ où T de loi de Student de paramètre $n-1$.
10. Variante : $H_0 = \{\mu \leq \mu_0\}$, $H_1 = \{\mu > \mu_0\}$. Région critique : $W_n = \{\zeta_n \geq a\}$. Niveau exact α : $a = t_{n-1,1-\alpha}$. Test convergent. p-valeur : $\mathbb{P}(T \geq \zeta_n^{\text{obs}})$.

III.7.3 Comparaison de moyennes de deux échantillons gaussiens

1. Modèle : deux suites indépendantes de v.a. indépendantes $(X_{1,1}, \ldots, X_{1,n_1})$, de même loi $\mathcal{N}(\mu_1, \sigma^2)$, et $(X_{2,1}, \ldots, X_{2,n_2})$, de même loi $\mathcal{N}(\mu_2, \sigma^2)$, avec $\mu_1, \mu_2 \in \mathbb{R}$ et $\sigma > 0$.
2. $H_0 = \{\mu_1 = \mu_2\}$, $H_1 = \{\mu_1 \neq \mu_2\}$.
3. Statistique de test : $T_{n_1,n_2} = \dfrac{(\bar{X}_1 - \bar{X}_2)\sqrt{n_1+n_2-2}}{U_{n_1,n_2}\sqrt{1/n_1 + 1/n_2}}$, avec $U^2_{n_1,n_2} = \sum_{j=1}^{n_1}(X_{1,j} - \bar{X}_1)^2 + \sum_{j=1}^{n_2}(X_{2,j} - \bar{X}_2)^2$ et $\bar{X}_i = \dfrac{1}{n_i}\displaystyle\sum_{j=1}^{n_i} X_{i,j}$ pour $i \in \{1,2\}$.
4. Loi sous H_0 : Student de paramètre $n_1 + n_2 - 2$.
5. Comportement asymptotique sous H_1 quand $\min(n_1, n_2)$ tend vers l'infini : T_{n_1,n_2} converge p.s. vers $-\infty$ ou $+\infty$.
6. Région critique : $W_{n_1,n_2} = \{|T_{n_1,n_2}| \geq a\}$.
7. Niveau exact α : $a = t_{n_1+n_2-2,1-\alpha/2}$, où $t_{n,1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student de paramètre n.
8. Test convergent.
9. p-valeur : $\mathbb{P}(|T| \geq |T^{\text{obs}}_{n_1,n_2}|)$ où T de loi de Student de paramètre $n_1 + n_2 - 2$.

III.7.4 Analyse de la variance à un facteur

1. Modèle : pour $i = 1 \ldots k$, $j = 1 \ldots n_i$,

$$X_{ij} = m_i + \varepsilon_{i,j},$$

et les $n = \sum_{i=1}^k n_i$ v.a. $\varepsilon_{i,j}$ sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. Les niveaux du facteur d'intérêt $m_1, \ldots, m_k \in \mathbb{R}$, et la variance $\sigma^2 \in]0, +\infty[$ sont inconnues.

2. $H_0 = \{m_1 = \ldots = m_k\}$ (le facteur n'a pas d'influence),
$H_1 = \{\exists i \neq j; m_i \neq m_j\}$, (au moins un des niveaux du facteur a de l'influence).
3. Statistique de test : $F = \dfrac{||X_E - X_H||^2/k - 1}{||X - X_E||^2/n - k}$, où

$$||X_E - X_H||^2 = \sum_{i=1}^{k} n_i (X_{i,\cdot} - X_{\cdot,\cdot})^2 = SSM,$$

$$||X - X_E||^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (X_{i,j} - X_{i,\cdot})^2 = SSE.$$

4. Comportement sous H_0 : F suit une loi de Fischer : $\mathcal{F}(k-1, n-k)$.
5. Comportement sous H_1 : $F \to \infty$ quand $n_i \to \infty$ pour $i = 1 \ldots k$.

6. Région critique : $W_n = \{F > a\}$.
7. Niveau α : $a = \mathcal{F}_{k-1,n-k,1-\alpha}$, où $\mathcal{F}_{k-1,n-k,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de Fisher $\mathcal{F}(k-1, n-k)$.
8. Le test est convergent.
9. p-valeur : $\mathbb{P}(F \geq f^{\text{obs}})$.
10. Le test est résumé par la table d'analyse de la variance :

Variabilité	SS	DF	MS	Fisher	p-valeur
Interclasse	$\|\|X_E - X_H\|\|^2$	$k-1$	$MSM = \frac{\|\|X_E - X_H\|\|^2}{k-1}$	$f = \frac{MSM}{MSE}$	$\mathbb{P}(F > f^{\text{obs}})$
Intraclasse	$\|\|X - X_E\|\|^2$	$n-k$	$MSE = \frac{\|\|X - X_E\|\|^2}{n-k}$		
Totale	$\|\|X - X_H\|\|^2$	$n-1$			

III.7.5 Régression multiple

1. Modèle : pour $i = 1 \dots n$
$$X_i = \beta + \sum_{j=1}^{p} \alpha_j R_i^j + \varepsilon_i.$$
Les v.a. ε_i, $i = 1 \dots n$ sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. Les coefficients de la régression $\beta, \alpha_1, \dots, \alpha_p$ et la variance σ^2 sont inconnues.
2. $H_0 = \{\alpha_{q+1} = \dots = \alpha_p = 0\}$ (les $p-q$ régresseurs $R^{q+1}, \dots, R^p$ sont inutiles),
$H_1 = \{\exists j \in \{q+1, \dots, p\}, \alpha_j \neq 0\}$ (un au moins des $p-q$ régresseurs $R^{q+1}, \dots, R^p$ est utile).
3. Statistique de test :
$$F = \frac{||X_E - X_H||^2/(p-q)}{||X - X_E||^2/(n-p-1)},$$
où X_E est la projection orthogonale de X sur l'espace vectoriel, E, engendré par $\mathbf{1}, R^1, \dots, R^p$, et X_H est la projection orthogonale de X sur l'espace vectoriel, H, engendré par $\mathbf{1}, R^1, \dots, R^q$.
4. Comportement sous H_0 : F suit une loi de Fischer : $\mathcal{F}(p-q, n-p-1)$.
5. Comportement sous H_1 : $F \to \infty$ quand $n \to \infty$.
6. Région critique : $W_n = \{F > a\}$.
7. Niveau α : $a = \mathcal{F}_{p-q,n-p-1,1-\alpha}$, où $\mathcal{F}_{p-q,n-p-1,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de Fisher $\mathcal{F}(p-q, n-p-1)$.
8. Le test est convergent.
9. p-valeur : $\mathbb{P}(F \geq f^{\text{obs}})$.

Chapitre IV

Modèles discrets

IV.1 Problématique des modèles discrets

- Est-ce que ce dé est pipé ?
- Est-ce que les études des enfants dépendent de la catégorie socioprofessionnelle du père ?
- Est-ce que l'âge influence la présence de maladie cardiaque ?
- Est-ce que le nouveau traitement médical est plus efficace que le précédent ?

Pour répondre à ces quatre problèmes bien réels, il est non seulement nécessaire de disposer de données, mais aussi de modèles statistiques adéquats et de tests statistiques associés. Ici, les données sont discrètes. Elles sont également nommées qualitatives ou encore catégorielles.

Dans le premier problème, il y a six modalités (les six faces numérotées, mais il pourrait très bien s'agir de couleurs). Si le dé n'est pas pipé, la probabilité d'apparition de chaque face doit être "proche" de 1/6, ce sera notre hypothèse nulle. L'expérience consistera à lancer ce dé un certain nombre de fois, et à évaluer la proportion observée d'apparition de chaque face. On comparera alors la distribution observée (les six proportions observées) à la distribution théorique (les probabilités égales à 1/6). Pour cela, nous utiliserons un test d'adéquation à une loi discrète, où la statistique du χ^2 associée mesure la distance entre ces deux distributions.

Dans le deuxième exemple, les catégories socioprofessionnelles (CSP) et les types d'étude correspondent à des noms. Les données sont généralement présentées sous forme d'un tableau à double entrées croisant ces deux variables. Les marges horizontale et verticale représentent respectivement les distributions des proportions observées des CSP et des types d'étude, alors qu'à l'intérieur du tableau les cases sont relatives à la distribution croisée observée. Pour tester l'indépendance de ces deux variables (ce sera notre hypothèse nulle), on peut utiliser la statistique du χ^2 d'indépendance qui mesure la distance entre la distribution croisée observée et le produit des deux distributions marginales.

Dans le troisième exemple, la maladie cardiaque est soit présente, soit absente (type booléen), par contre l'âge est exprimé numériquement en année. Pour chaque âge, on considère qu'il y a une probabilité d'avoir une maladie cardiaque. L'objectif est alors de rechercher si cette probabilité croît ou décroît significativement avec l'âge. Notre hypothèse nulle correspondra à l'absence d'une liaison. Cette analyse suit la même démarche vue au chapitre III dans le cadre du modèle linéaire gaussien, la seule différence est que la variable à expliquer est booléenne, à la place d'être numérique. La méthode est nommée régression logistique.

Dans le dernier problème, l'effet du nouveau traitement peut s'exprimer en terme de succès ou d'échec. Pour réaliser le test, une solution raisonnable est de mesurer la proportion de succès (ou d'échecs) dans deux échantillons (appariés ou non appariés comme nous le verrons par la suite). L'hypothèse nulle sera relative à l'absence d'effet du traitement, c'est-à-dire à l'égalité statistique des deux proportions. Ce test est nommé : comparaison de deux proportions.

IV.2 Tests du χ^2

IV.2.1 Test d'adéquation à une loi discrète

Le problème

On observe n v.a. $(X_i)_{1\leq i\leq n}$, indépendantes et de même loi, à valeurs dans un espace fini $A = \{a_1, \ldots, a_k\}$. Cette loi, inconnue, est caractérisée par la suite $\underline{p} = (p_1, \ldots p_k)$ (avec $\sum_{j=1}^k p_j = 1$), où pour tout $j = 1, \ldots, k$, la quantité p_j désigne la probabilité d'observer a_j (indépendante de i en raison de l'identique distribution des X_i) ; soit $p_j = \mathbb{P}(X_i = a_j)$. La loi jointe du n-uplet $\underline{X} = (X_i)_{1\leq i\leq n}$ est : pour tout $(x_1, \cdots, x_n) \in A^n$,

$$\mathbb{P}_{\underline{p}}(X_i = x_i,\ \ 1 \leq i \leq n) = \prod_{i=1}^n \mathbb{P}_{\underline{p}}(X_i = x_i) = \prod_{j=1}^k p_j^{\mathrm{card}(\{i\,;\, x_i = a_j\})}.$$

Remarque IV.1. Il en est ainsi, par exemple, si on procède à un sondage dans une population divisée en k catégories, les tirages des n individus pouvant être considérés comme indépendants, et, à chaque fois, la probabilité d'être dans une catégorie donnée étant égale à la proportion (inconnue) d'individus de cette catégorie dans la population totale. C'est bien le cas si on effectue des tirages "avec remises" et "brassage" de la population, mais un tel "modèle d'urne", quoique traditionnel, n'est pas très réaliste. Cependant, on peut considérer qu'on est approximativement dans le modèle proposé si on fait porter le tirage sur des individus distincts (tirage "sans remise") mais dans un contexte où la taille totale de la population est très grande par rapport à celle de l'échantillon.

$\Diamond$

On avance l'hypothèse que le paramètre est $\underline{p}^0 = (p_1^0, \ldots, p_k^0)$, où $p_j^0 > 0$, pour tout $j = 1, \ldots, k$. **Le but est de tester, à un niveau donné α, cette hypothèse nulle simple**, $H_0 = \{\underline{p} = \underline{p}^0\}$, contre l'hypothèse alternative $H_1 = \{\underline{p} \neq \underline{p}^0\}$.

Intuitions

Pour tout $j = 1, \ldots, k$ on note $N_j = \mathrm{card}(\{i : X_i = a_j\}) = \sum_{i=1}^n \mathbf{1}_{\{X_i = a_j\}}$ la variable aléatoire de comptage du nombres de fois où l'état a_j est visité par les v.a. X_i, $i = 1, \ldots, n$. La v.a. N_j suit une loi binomiale (voir X.2.1, p. 242) de paramètres (n, p_j). On rappelle que $\mathbb{E}[N_j] = np_j$, que la v.a. $\hat{P}_j = \frac{N_j}{n}$ est un **estimateur convergent sans biais** de p_j (voir le chapitre II).

Il y a donc lieu de penser que, s'il est vrai que $\underline{p} = \underline{p}^0$, la suite des effectifs observés $n_j = \mathrm{card}(\{i : x_i = a_j\})$ sera telle que la suite des fréquences observées, $\underline{\hat{p}} = (\hat{p}_1, \ldots, \hat{p}_k) =$

$(\frac{n_1}{n}, \ldots, \frac{n_k}{n})$, sera "proche" (en raison de la loi forte des grands nombres citée précèdemment) de la suite mise en test $\underline{p}^0 = (p_1^0, \ldots, p_k^0)$.

Avec cette notation, il vient que $\mathbb{P}_{\underline{p}}(X_i = x_i,\ 1 \leq i \leq n) = \prod_{j=1}^k p_j^{n_j}$, ce qui met en évidence, par la méthode de Halmos-Savage (voir le théorème II.19), que la v.a. k-dimensionnelle $\underline{N} = (N_j)_{1 \leq j \leq k}$ est **exhaustive**, ce qui justifie que nous fassions porter notre test sur cette suite des effectifs observés, ou, ce qui revient au même, sur la suite des fréquences observées $\underline{\hat{P}} = (\hat{P}_j)_{1 \leq j \leq k}$. La loi de $\underline{N}$ est la loi multinomiale de paramètres n et $\underline{p} = (p_1, \ldots, p_k)$, notée $\mathcal{M}(n, \underline{p})$ (voir X.2.1, p. 243). On peut vérifier que $\underline{\hat{P}}$ est l'estimation par maximum de vraisemblance de $\underline{p}$.

On souhaite donc pouvoir caractériser une "distance" entre la **suite des fréquences observées** $\underline{\hat{p}}$ et la **suite des fréquences théoriques** $\underline{p}^0$, de manière à rejeter l'hypothèse nulle si cette distance est supérieure à une certaine valeur frontière. Pour réaliser ce programme, il faut que :

- **la loi, sous l'hypothèse nulle, de cette distance soit (au moins approximativement) connue** de sorte que la frontière sera le quantile d'ordre $1-\alpha$ de cette loi (le rejet à tort de l'hypothèse nulle sera bien alors de probabilité approximativement égale à α),
- **si l'hypothèse nulle n'est pas satisfaite**, cette distance ait tendance à prendre des valeurs d'autant plus grandes que la vraie valeur du paramètre $\underline{p}$ est plus "éloignée" de $\underline{p}^0$ (ce qui, là aussi, conduit à souhaiter disposer d'une distance entre $\underline{p}$ et $\underline{p}^0$, gouvernant la loi de la distance entre la v.a. $\underline{\hat{P}}$ et $\underline{p}^0$).

Outils

On définit la **distance du χ^2** (ou **distance du chi-deux**) , entre deux probabilités sur un ensemble fini à k éléments, $\underline{p} = (p_j)_{1 \leq j \leq k}$ et $\underline{q} = (q_j)_{1 \leq j \leq k}$, par :

$$D(\underline{p}, \underline{q}) = \sum_{j=1}^k \frac{(p_j - q_j)^2}{q_j} \, .$$

Remarquons que, faute de symétrie entre $\underline{p}$ et $\underline{q}$, cet objet n'est pas une "distance" au sens mathématique traditionnel du terme (on parle parfois de "pseudo-distance" du χ^2).

On démontre (*nous l'admettrons*) que, **si l'hypothèse nulle est satisfaite**, la loi de la v.a. $n.D(\underline{\hat{P}}, \underline{p}^0)$ tend, quand n tend vers l'infini, vers la loi du chi-deux à $k-1$ degrés de liberté (voir X.2.2, p. 246). Ceci conduit, pour n "assez grand" (notion qui sera précisée empiriquement dans la suite), à fonder sur $n.D(\underline{\hat{P}}, \underline{p}^0)$ le test, au niveau α, de l'hypothèse $H_0 = \{\underline{p} = \underline{p}^0\}$, le rejet ayant lieu si

$$n \sum_{j=1}^k \frac{(\hat{p}_j - p_j^0)^2}{p_j^0} \geq \chi^2_{k-1,1-\alpha} \, ,$$

où $\chi^2_{k-1,1-\alpha}$ désigne le quantile d'ordre $1-\alpha$ de la loi du chi-deux à $k-1$ degrés de liberté, disponible dans des tables ou via les ordinateurs. C'est ce que l'on appelle le **test du χ^2**.

Critère pratique. *On considère souvent que l'approximation fournie par la loi du χ^2 à $k-1$ degrés de liberté pour la loi de $n.D(\underline{\hat{P}}, \underline{p}^0)$ est valide si tous les produits $np_j^0(1-p_j^0)$ sont supérieurs ou égaux à 5.*

Intéressons nous maintenant à la **puissance** de ce test, c'est-à-dire considérons les situations où $\underline{p} \neq \underline{p}^0$. On démontre (*nous l'admettrons*) que, si la loi commune des v.a. X_i est caractérisée par la valeur $\underline{p}$ du paramètre, alors la loi de $n.D(\underline{\hat{P}}, \underline{p}^0)$ est bien approchée, quand n tend vers l'infini, par la loi dite du χ^2 **décentré** à $k-1$ degrés de liberté, $\chi^2_{k-1,\delta}$ (voir X.2.2, p. 246), avec pour coefficient d'excentricité $\delta = n.D(p, p_0)$.

Il se produit alors une circonstance heureuse concernant la famille des lois $\chi^2_{k-1,\delta}$: elle est, à nombre de degrés de liberté fixé (ici $k-1$) **stochastiquement croissante** avec le coefficient d'excentricité δ, c'est-à-dire que, pour tout $t > 0$, la probabilité qu'une v.a. suivant la loi $\chi^2_{k-1,\delta}$ dépasse t est fonction croissante de δ.

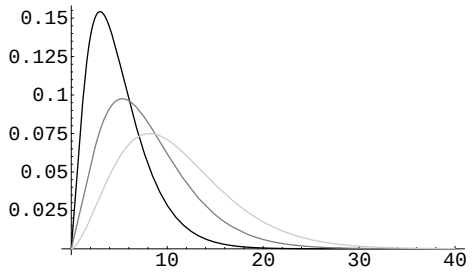


FIG. IV.1 – Densité du chi-2 à 5 degrés de liberté avec décentrage de 0, 3 et 6.

Afin d'illustrer davantage le phénomène d'excentricité engendré par δ nous pouvons rappeler que $\mathbb{E}[\chi^2_{k,\delta}] = k + \delta$ et $\mathrm{Var}(\chi^2_{k,\delta}) = 2(k + 2\delta)$.

Généralisation à l'adéquation d'une loi quelconque

Soient n v.a. $(X'_i)_{1\leq j\leq n}$ indépendantes, de même loi inconnue Q, et à valeurs dans un espace mesurable $(E', \mathcal{E}')$. On veut tester l'hypothèse que cette loi coincide avec une loi proposée Q^0.

On suppose que l'ensemble E' est infini, ou bien fini mais à cardinal trop élevé pour qu'on puisse lui appliquer raisonnablement le test du χ^2. Un artifice parfois pratiqué est de considérer une partition finie de E', soit $(A_1, \ldots, A_k)$, et de se contenter de tester si les valeurs des $\mathbb{P}_Q(X'_1 \in A_j)$ sont égales aux $\mathbb{P}_{Q^0}(X'_1 \in A_j)$. Il suffit donc de relever, pour chaque observation x'_i, à quelle partie A_j elle appartient, ce qui revient à considérer des v.a. $X_i = j\mathbf{1}_{\{X'_i \in A_j\}}$ à valeurs dans l'ensemble $\{1, \ldots, k\}$. On se trouve ramené au problème précédent, mais au prix d'une certaine tricherie sur le problème posé : on ne sait plus distinguer entre deux lois différentes, mais identiques sur la partition choisie. En d'autres termes, on teste en fait l'hypothèse nulle selon laquelle la loi inconnue Q appartient à l'ensemble des lois qui affectent à chacune des parties A_j la même probabilité que Q^0 ; l'hypothèse alternative est alors le complémentaire de cet ensemble dans l'ensemble de toutes les probabilités sur $(E', \mathcal{E}')$.

Si malgré cet inconvénient on décide de procéder ainsi, il reste à choisir la partition. Si son effectif k est fixé, on constate que l'approximation par la loi asymptotique de χ^2 sera d'autant meilleure que la suite des $\mathbb{P}_{Q^0}(A_j)$ sera plus proche de la suite constante dont tous les éléments valent $\frac{1}{k}$; en effet c'est ainsi que la plus forte des valeurs $n\mathbb{P}_{Q^0}(A_j)(1-\mathbb{P}_{Q^0}(A_j))$ sera la plus faible possible ; or ces valeurs sont les variances des effectifs N_j et, pour chaque

j, plus cette variance est élevée, plus l'estimation de p_j par la fréquence observée n_j/n (qui nous a servi à justifier heuristiquement le test) est mauvaise.

Quant au choix de k, il sera l'objet d'un compromis entre le désir d'élever k (pour ne pas trop dénaturer le problème initial) et le désir d'élever $n\frac{1}{k}(1-\frac{1}{k})$ (à rendre au moins égal à 5) pour valider au mieux l'approximation par la loi du chi-deux.

Dans le cas particulier d'une loi Q^0 sur $\mathbb{R}$ admettant une densité et donc (voir X.1.2) à fonction de répartition, soit F^0, continue, on prendra généralement une partition $(A_1, \ldots, A_k)$ en intervalles (bornés ou non) tous de probabilité $\frac{1}{k}$, donc délimités par les points s_j (où $1 \leq j \leq k-1$) vérifiant $F^0(s_j) = \frac{j}{k}$. Mais cette méthode reste médiocre et les méthodes de type non-paramétrique, qui seront vues au chapitre V, sont en général meilleures.

IV.2.2 Test d'adéquation à une famille de lois discrètes

Présentation générale

Le modèle est ici le même qu'en IV.2.1 : on observe n v.a. X_i, indépendantes et de même loi, à valeurs dans un espace fini, soit $A = \{a_1, \ldots, a_k\}$. Cette loi, inconnue, est caractérisée par la suite $\underline{p} = (p_1, \ldots p_k)$, où, pour tout j (avec $1 \leq j \leq k$), p_j désigne la probabilité d'observer a_j.

Ici l'hypothèse à tester n'est plus réduite à une valeur bien déterminée $\underline{p}^0$, mais elle exprime que le paramètre appartient à une famille $(\underline{p}_\theta,\ \theta \in \Theta)$, où l'on note $\underline{p}_\theta = (p_{1,\theta}, \ldots, p_{k,\theta})$ un vecteur de poids de probabilité indexé par un paramètre θ. Attention : Θ n'est pas ici l'ensemble des paramètres du modèle tout entier mais paramétrise seulement l'hypothèse nulle.

Une idée naturelle est de reprendre la méthode du test d'adéquation vue en IV.2.1 en y remplaçant $\underline{p}^0$ par $\underline{p}_{\hat{\theta}}$, où $\hat{\theta}$ est une estimation de θ. C'est ce que l'on appelle un **test du χ^2 adaptatif**. On démontre alors que **si l'ensemble Θ des valeurs possibles pour θ est une partie ouverte d'intérieur non vide de** $\mathbb{R}^h$ (avec $h < k-1$) la loi de $nD(\underline{\hat{P}}, \underline{p}_{\hat{\theta}})$ tend, sous l'hypothèse nulle, vers la loi du χ^2 à $k-h-1$ degrés de liberté, **sous des conditions de régularité que nous ne préciserons pas ici**, mais qui sont satisfaites si $\hat{\theta}$ est une estimation par maximum de vraisemblance. Donc **on procède comme dans le test du χ^2 d'adéquation, en remplaçant seulement le nombre de degrés de liberté $k-1$ par $k-h-1$.**

Exemple : test du χ^2 d'indépendance

Les v.a. i.i.d. X_i sont ici de la forme (Y_i, Z_i), où les "premières composantes" Y_i sont à valeurs dans $A = \{a_1, \ldots, a_k\}$, et les "secondes composantes" Z_i sont à valeurs dans $B = \{b_1, \ldots, b_m\}$.

On note, pour tout $j = 1, \ldots, k$, et tout $\ell = 1, \ldots, m$, $p_{j,\ell} = \mathbb{P}((Y_i, Z_i) = (a_j, b_\ell))$. Le paramètre est donc $\underline{p} = (p_{j,\ell})_{1 \leq j \leq k, 1 \leq \ell \leq m}$.

On veut tester l'hypothèse que les 2 composantes sont indépendantes, autrement dit que la loi commune des couples (Y_i, Z_i) est une loi produit, c'est-à-dire encore que tous les $p_{j,\ell}$ sont de la forme :

$$\forall (j,\ell) \in A \times B,\ \ p_{j,\ell} = \mathbb{P}(Y_i = a_j, Z_i = b_\ell) = \mathbb{P}(Y_i = a_j)\mathbb{P}(Z_i = b_\ell) = q_j r_\ell,$$

où nécessairement, pour tout j, $q_j = \sum_{\ell=1}^{m} p_{j,\ell}$ et, pour tout ℓ, $r_\ell = \sum_{j=1}^{k} p_{j,\ell}$. Les q_j caractérisent la loi commune des v.a. Y_i et les r_ℓ caractérisent la loi commune des v.a. Z_i ; ces lois sont appelées aussi première et seconde lois marginales des X_i (voir X.1.1, p. 225).

Ainsi, **sous l'hypothèse nulle**, le paramètre, caractérisé d'une part par les k valeurs q_j (de somme égale à 1) et d'autre part par les m valeurs r_ℓ (aussi de somme égale à 1), appartient à un espace de dimension $h = k + m - 2$. On supposera que les q_j et les r_ℓ sont tous non nuls, ce qui assure que, sous l'hypothèse nulle, l'ensemble de paramétrage est une partie ouverte de $\mathbb{R}^{k+m-2}$

Étant observé un échantillon de taille n, soit $(y_i, z_i)_{1\leq i\leq n}$, notons, pour tout couple (j, ℓ), $n_{j,\ell}$ l'effectif des observations égales à (a_j, b_ℓ) et $\hat{p}_{j,\ell}$ leur fréquence ($\hat{p}_{j,\ell} = \frac{n_{j,\ell}}{n}$). On estime alors chaque q_j de la première marge par la fréquence marginale correspondante $\hat{q}_j = \frac{1}{n}\sum_{\ell=1}^{m} n_{j,\ell}$ et de même, pour la seconde marge, chaque r_ℓ par la fréquence marginale correspondante $\hat{r}_\ell = \frac{1}{n}\sum_{j=1}^{k} n_{j,\ell}$.

Alors, **si l'hypothèse nulle est satisfaite**, on estime, pour tout couple (j, ℓ), $p_{j,\ell}$, par le produit des fréquences marginales $\hat{q}_j\hat{r}_\ell$ (pour mimer la formule d'indépendance citée plus haut).

Nous admettons que les conditions de validité de la méthode sont satisfaites, $\hat{q}_j$ et $\hat{r}_\ell$ étant respectivement des estimateurs par maximum de vraisemblance de q_j et r_ℓ. Le test, au seuil α, consiste donc à rejeter l'hypothèse d'indépendance si :

$$n\sum_{j=1}^{k}\sum_{\ell=1}^{m}\frac{(\hat{p}_{j,\ell} - \hat{q}_j\hat{r}_\ell)^2}{\hat{q}_j\hat{r}_\ell} \geq \chi^2_{(k-1)(m-1),1-\alpha},$$

autrement dit

$$n\sum_{j=1}^{k}\sum_{\ell=1}^{m}\frac{(\frac{n_{j,\ell}}{n} - \frac{n'_j.n''_\ell}{n^2})^2}{\frac{n'_j.n''_\ell}{n^2}} \geq \chi^2_{(k-1)(m-1),1-\alpha},$$

où :

- $n_{j,\ell}$ est le nombre d'observations égales à (a_j, b_ℓ),
- $n'_j = \sum_{\ell=1}^{m} n_{j,\ell}$ est le nombre d'observations dont la première composante est égale à a_j,
- $n''_\ell = \sum_{j=1}^{k} n_{j,\ell}$ est le nombre d'observations dont la seconde composante est égale à b_ℓ,
- $\chi^2_{(k-1)(m-1),1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi du χ^2 à $(k-1)(m-1)$ degrés de liberté (en effet $km - (k+m-2) - 1 = (k-1)(m-1)$).

IV.3 La régression logistique

Au chapitre III le paragraphe III.6, portant sur la *régression linéaire* était consacré à un modèle dans lequel une variable aléatoire absolument continue, de loi gaussienne, était "expliquée" par une ou plusieurs variables. Nous allons maintenant nous intéresser à des situations où la variable à expliquer est discrète, et même, plus précisément ici (quoique ce soit généralisable), dichotomique (aussi dite booléenne : elle ne peut prendre que deux valeurs, notées conventionnellement 0 et 1).

i	Age	CHD	i	Age	CHD	i	Age	CHD	i	Age	CHD
1	20	0	26	35	0	51	44	1	76	55	1
2	23	0	27	35	0	52	44	1	77	56	1
3	24	0	28	36	0	53	45	0	78	56	1
4	25	0	29	36	1	54	45	1	79	56	1
5	25	1	30	36	0	55	46	0	80	57	0
6	26	0	31	37	0	56	46	1	81	57	0
7	26	0	32	37	1	57	47	0	82	57	1
8	28	0	33	37	0	58	47	0	83	57	1
9	28	0	34	38	0	59	47	1	84	57	1
10	29	0	35	38	0	60	48	0	85	57	1
11	30	0	36	39	0	61	48	1	86	58	0
12	30	0	37	39	1	62	48	1	87	58	1
13	30	0	38	40	0	63	49	0	88	58	1
14	30	0	39	40	1	64	49	0	89	59	1
15	30	0	40	41	0	65	49	1	90	59	1
16	30	1	41	41	0	66	50	0	91	60	0
17	32	0	42	42	0	67	50	1	92	60	1
18	32	0	43	42	0	68	51	0	93	61	1
19	33	0	44	42	0	69	52	0	94	62	1
20	33	0	45	42	1	70	52	1	95	62	1
21	34	0	46	43	0	71	53	1	96	63	1
22	34	0	47	43	0	72	53	1	97	64	0
23	34	1	48	43	1	73	54	1	98	64	1
24	34	0	49	44	0	74	55	0	99	65	1
25	34	0	50	44	0	75	55	1	100	69	1

TAB. IV.1 – Les données CHD

IV.3.1 Exemple introductif

On dispose d'un échantillon de 100 personnes sur lequel on a mesuré deux variables (voir tableau IV.1), l'âge du patient, X (noté AGE dans le tableau et les graphiques ci-dessous) et la présence (1) ou l'absence (0) d'une maladie cardiaque, Y (notée CHD dans le tableau et les graphiques ci-dessous).

L'objectif de l'étude est de savoir si l'âge a un effet sur la présence de la maladie cardiaque. Sur le graphique suivant, on observe deux bandes parallèles de points, où chacun représente l'âge de l'individu avec la présence ($Y = 1$) ou l'absence ($Y = 0$) de la maladie. On peut notamment constater qu'il y a plus de points rassemblés vers les jeunes pour $Y = 0$, alors qu'un regroupement se matérialise plutôt vers les plus âgés pour $Y = 1$. Cependant, cette constatation n'est pas suffisante pour en déduire une relation significative entre la prédisposition à une maladie cardiaque et l'âge.

Considérons tout d'abord ce à quoi conduirait l'application (évidemment maladroite) à cette situation d'une régression linéaire simple usuelle, c'est-à-dire posons une relation affine du type :

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

où ϵ suivrait la loi normale $\mathcal{N}(0, \sigma^2)$. On supposerait de plus que les v.a. à expliquer Y_i (où $1 \leq i \leq 100$) sont indépendantes. Notons que les observations explicatives x_i n'intervenant qu'à titre de conditionnements dans le modèle, il n'est pas besoin de faire d'hypothèse de nature probabiliste les concernant ; en fait elles peuvent être selon les protocoles de recueil des données aussi bien aléatoires (cas de sujets admis dans l'étude sans considération préalable de leur âge) que déterministes (cas d'un panel de sujets composé afin d'équilibrer les âges).

On pourrait être tenté d'estimer ce modèle à l'aide de la méthode des moindres carrés comme dans le chapitre III ; alors on obtiendrait :

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = -0.5380 + 0.0218X.$$

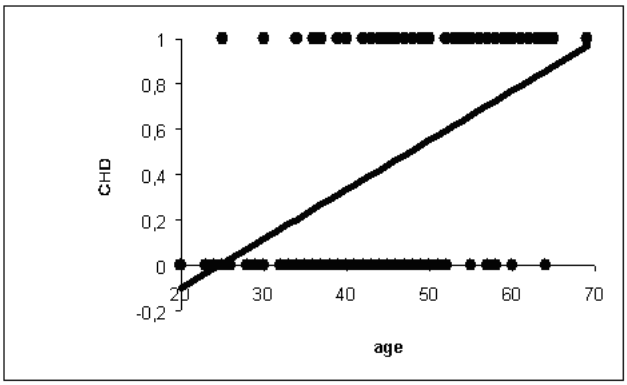


FIG. IV.2 – Régression linéaire de CHD avec l'âge

Cette régression signifierait que le fait d'avoir une année de plus produirait un accroissement de 0,02 en terme de prédisposition à avoir la maladie cardiaque (prédisposition assimilée à la probabilité d'avoir cette maladie cardiaque). On poserait par conséquent, en termes de probabilité conditionnele (voir X.1.4, p. 238),

$$p(x) = \mathbb{P}(Y = 1 | X = x).$$

On voit alors, sur le graphique précédent, les inconvénients de ce modèle linéaire en probabilité. Un accroissement constant de l'âge produit également un accroissement constant de $p(x)$. De plus, $p(x)$ peut très bien alors sortir de l'intervalle $[0, 1]$, ce qui est logiquement impossible pour une probabilité. En effet, dans notre exemple, nous voyons que la probabilité estimée $\hat{p}(x)$ serait négative pour les patients âgés de 25 ans et moins ; par exemple, pour une personne ayant 23 ans, l'estimation de la probabilité d'avoir une maladie cardiaque serait :

$$\hat{p}(23) = -0.5380 + 0.0218 \times 23 = -0.036.$$

De même, si l'on voulait prévoir la probabilité d'avoir une maladie cardiaque pour une personne de plus de 70 ans, alors les valeurs estimées seraient supérieures à l'unité.

De plus les résidus pour un âge fixé ne peuvent prendre que deux valeurs possibles : si $Y = 0$, alors $\epsilon(x) = -p(x)$ et si $Y = 1$, alors $\epsilon(x) = 1 - p(x)$, ce qui est contraire à l'hypothèse de normalité (loi continue) des résidus, introduite dans le cadre du modèle linéaire gaussien. Par conséquent, la loi de ϵ n'est pas continue, mais discrète : si $Y = 1$, alors $\epsilon = 1 - p(x)$ avec une probabilité $p(x)$ et si $Y = 0$, alors $\epsilon = -p(x)$ avec une probabilité $1 - p(x)$. Comme ϵ a une distribution d'espérance nulle, la variance de ϵ est égale à $p(x)(1 - p(x))$. Ce résultat est à nouveau en contradiction avec l'hypothèse de variance constante des résidus du modèle linéaire gaussien usuel, car ici la variance dépend de l'âge.

En fait, à la lumière de ces constatations, la distribution conditionnelle de Y suit une loi de Bernoulli avec une probabilité d'être malade fournie par l'espérance conditionnelle (voir X.1.4, p. 238) :

$$p(x) = \mathbb{E}[Y|X = x].$$

On peut donc conclure, comme on pouvait s'en douter, que la régression linéaire simple usuelle n'est pas une bonne solution pour résoudre notre problème.

IV.3.2 Une solution raisonnable : la régression logistique

Le modèle

Soit Y une v.a. (dite "à expliquer") dont la distribution conditionnelle est une loi de Bernoulli dont le paramètre, noté $p(x)$, dépend de la valeur x d'une variable réelle X dite "explicative" : $\mathcal{B}(p(x))$. On rappelle que son espérance mathématique est égale à $p(x)$ et sa variance égale à $p(x)(1 - p(x))$. Avec une notation du type "probabilité conditionnelle" (voir X.1.4) et avec la convention $0^0 = 1$, on pourra noter ceci :

$$\mathbb{P}(Y = y|X = x) = p(x)^y(1 - p(x))^{1-y}.$$

Pour préciser le modèle, il faut maintenant fixer une classe de fonctions à laquelle est supposée appartenir l'application $x \mapsto p(x)$ (ce rôle était joué, dans la tentative de modèle linéaire ci-dessus, par les fonctions affines).

Une classe $\mathcal{L}$ de telles fonctions "raisonnable" pour modéliser $p(x)$ doit satisfaire les conditions suivantes :

- valeurs dans l'intervalle $[0, 1]$,
- monotonie en x,
- stabilité par changement d'origine et d'échelle sur la variable explicative (comme en régression linéaire) : si la fonction p appartient à $\mathcal{L}$, il en est de même des fonctions $x \mapsto p(b_0 + b_1 x)$.

La **régression logistique** consiste en l'emploi des fonctions du type

$$p(x) = g(\beta_0 + \beta_1 x)$$

avec

$$g(t) = \frac{\exp(t)}{1 + \exp(t)}.$$

On voit que ces fonctions ont toutes une même "forme", et que donc, comme en régression linéaire simple, tout ce qu'il y a à estimer consiste en un *paramètre de position* (β_0) et un *paramètre d'échelle* (β_1) ; on notera ψ le couple (β_0, β_1).

Ce modèle possède de plus les propriétés suivantes :

- *Variation* : Si $\beta_1 = 0$, la loi de la variable à expliquer ne dépend pas de la variable explicative (la fonction p est constante) ; sinon, p est strictement croissante (resp. décroissante) si $\beta_1 > 0$ (resp. $\beta_1 < 0$).
- *Limites* : Si $\beta_1 \neq 0$, et si x parcourait tout $\mathbb{R}$ (ce qui est en général impossible dans un modèle réaliste), alors $p(x)$ parcourrait tout l'intervalle $]0,1[$; si $\beta_1 > 0$ (resp. $\beta_1 < 0$), on a $\lim_{x\to-\infty} p(x) = 0$ et $\lim_{x\to+\infty} p(x) = 1$ (resp. $\lim_{x\to-\infty} p(x) = 1$ et $\lim_{x\to+\infty} p(x) = 0$).
- *Symétrie par rapport au point* $(-\frac{\beta_0}{\beta_1}, 1/2)$: si $x + x' = -2\frac{\beta_0}{\beta_1}$, alors $p(x) + p(x') = 1$.

La terminologie "logistique" vient de l'anglais LOGIT (*LOGarithm of Inverse Transformation*) qui désigne la fonction inverse de g, c'est-à-dire h définie par $h(u) = \ln(\frac{u}{1-u})$.

Les estimateurs

Procédons maintenant à l'*estimation par maximum de vraisemblance* dans ce modèle du paramètre $\psi = (\beta_0, \beta_1)$; on note désormais $p_\psi(x) = g(\beta_0 + \beta_1 x) = \dfrac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$.

Les v.a. Y_i étant indépendantes et de lois de Bernoulli, calculons la vraisemblance (relativement à la mesure de comptage sur $\{0,1\}^n$, comme en II.2), associée à la suite d'observations $\underline{y} = (y_1, \ldots, y_n)$, pour la suite de valeurs explicatives $\underline{x} = (x_1, \ldots, x_n)$:

$$L(\underline{y}|\underline{x};\psi) = \prod_{i=1}^{n} (p_\psi(x_i))^{y_i} (1 - p_\psi(x_i))^{1-y_i}.$$

Alors l'estimateur du maximum de vraisemblance est la valeur (ici unique) de ψ en laquelle prend son maximum la fonction $L(\underline{y}|\underline{x};\cdot)$ ou, ce qui est plus pratique pour la résolution, $\ell(\underline{y}|\underline{x};\cdot)$, où la log-vraisemblance ℓ est définie ici par :

$$\ell(\underline{y}|\underline{x};\cdot) = \sum_{i=1}^{n} y_i \ln(p_\psi(x_i)) + (1 - y_i) \ln(1 - p_\psi(x_i))$$

(avec la convention $0 \ln(0) = 0$).

C'est la solution du système de deux équations à deux inconnues :

$$\frac{\partial \ell(\underline{y}|\underline{x};\beta_0,\beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} \left[y_i - \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right] = 0,$$

$$\frac{\partial \ell(\underline{y}|\underline{x};\beta_0,\beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} x_i \left[y_i - \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right] = 0.$$

Ces expressions sont non linéaires en β_0 et β_1 (contrairement à la situation correspondante en régression linéaire), ce qui demande de faire appel à des méthodes spécifiques de résolution. Il s'agit généralement de méthodes de résolution numérique itérative (Newton-Raphson, Score de Fisher et de Rao, moindres carrés repondérés itératifs). Nous ne détaillerons pas ces

méthodes dans le présent cours ; les logiciels statistiques standards offrent ce type d'algorithme.

Quand la taille de l'échantillon tend vers l'infini, ces estimateurs sont convergents et asymptotiquement normaux. Leurs variances asymptotiques sont théoriquement connues en fonction de ψ. Des approximations de ces variances sont donc obtenues, pour n assez grand, en remplaçant dans les formules donnant ces variances asymptotiques, ψ par son estimation.

Retour à l'exemple

Dans notre exemple, l'estimateur du maximum de vraisemblance donne :

$$\hat{\beta}_0 = -5.3095 \ , \ \hat{\beta}_1 = 0.1109 \text{ et } \ell(\underline{y}; \hat{\beta}_0, \hat{\beta}_1) = -53.677.$$

Le modèle estimé est :

$$\hat{p}(x) = \frac{\exp(-5.3095 + 0.1109x)}{1 + \exp(-5.3095 + 0.1109x)}.$$

Les estimations des variances sont $\mathrm{Var}(\hat{\beta}_0) = 1.2856$ et $\mathrm{Var}(\hat{\beta}_1) = 0.0005$.

Cela nous permet de tracer la courbe logistique estimée de la présence de maladie cardiaque en fonction de l'âge des patients. On voit nettement sur le graphique ci-dessous comment, dans ce modèle, l'estimation de la probabilité d'avoir une maladie cardiaque augmente avec l'âge. Elle est, par exemple, de 0.073 à 25 ans, 0.194 à 35 ans, 0.421 à 45 ans, 0.688 à 55 ans et 0.870 à 65 ans.

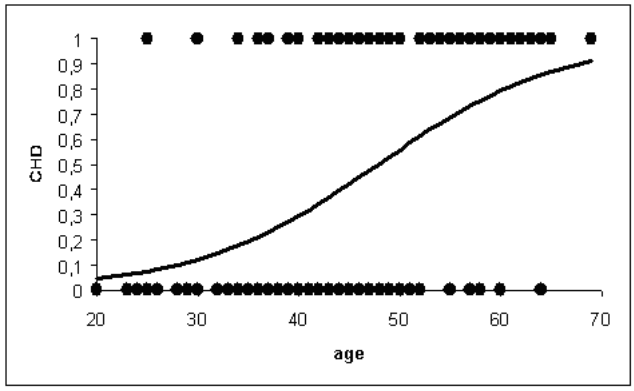


FIG. IV.3 – Régression logistique de CHD avec l'âge

Cependant cela ne nous permet pas d'affirmer que la présence d'une maladie cardiaque augmente statistiquement (significativement) avec l'âge. Pour répondre à cette question (et préciser le sens de "significativement") il convient d'avoir recours à des tests statistiques, comme dans le cadre du modèle linéaire gaussien.

IV.3.3 Comment tester l'influence de la variable explicative ?

Dans l'exemple introductif, il s'agit de savoir si statistiquement l'âge influe sur la présence de maladie cardiaque. La démarche classique des tests, utilisée dans le cadre du modèle linéaire gaussien au chapitre III, est la même pour le modèle logit. L'hypothèse nulle a la

forme suivante : $H_0 = \{\beta_1 = 0\}$ (pas d'influence de la variable X sur Y) et l'hypothèse alternative est : $H_1 = \{\beta_1 \neq 0\}$ (influence effective de la variable X sur Y).

Plusieurs tests ont été proposés pour ce test, tous fondés sur une même idée très générale (qui valait déjà pour le modèle linéaire gaussien) et qui est la suivante : estimons par maximum de vraisemblance le paramètre ψ d'une part dans le modèle général (ici le modèle logistique), d'où l'estimation $\hat{\psi} = (\hat{\beta}_0, \hat{\beta}_1)$, et d'autre part dans le modèle restreint à l'hypothèse nulle, d'où l'estimation $\hat{\psi}_{H_0} = (\tilde{\beta}_0, 0)$; alors ces deux estimations ont tendance à être "proches" si l'hypothèse nulle est satisfaite et "éloignées" si elle ne l'est pas. (On aurait pu aussi utiliser un test construit sur la normalité asymptotique de $\hat{\beta}_1$ sous H_0.)

Les différentes techniques de test diffèrent par le choix de l'outil statistique caractérisant cette proximité. Nous détaillons ici seulement le **test du rapport de vraisemblances**, qui consiste à considérer la différence des log-vraisemblances (ou plutôt, pour des raisons de calcul précisées ensuite, son double), c'est-à-dire :

$$\Lambda(\underline{y}|\underline{x}) = 2(\ell(\underline{y}|\underline{x}; \hat{\psi}) - \ell(\underline{y}|\underline{x}; \hat{\psi}_{H_0})).$$

Remarquons que $\Lambda(\underline{y}|\underline{x})$ est nécessairement positif ou nul, car $\ell(\underline{y}|\underline{x}, \hat{\psi})$ est le maximum de $\ell(\underline{y}|\underline{x}; \cdot)$ pris sur un ensemble qui contient celui utilisé pour la recherche de $\hat{\psi}_{H_0}$.

Il est clair alors que l'on aura d'autant plus tendance à rejeter l'hypothèse nulle que $\Lambda(\underline{y})$ sera plus grand. Ce test peut être mis en œuvre car (nous l'admettrons), sous l'hypothèse nulle, la v.a. Λ suit asymptotiquement (quand la taille de l'échantillon, n, tend vers l'infini) une loi du χ^2 à 1 degré de liberté (autrement dit la loi du carré d'une v.a. de loi normale centrée réduite).

Il reste à expliciter $\Lambda(\underline{y}|\underline{x})$. Notons n_1 (resp. n_0) le nombre d'observations de la variable à expliquer égales à 1 (resp. 0) ; autrement dit soit $n_1 = \sum_{i=1}^n y_i$ et $n_0 = \sum_{i=1}^n (1 - y_i) = n - n_1$. Alors, rappelant que sous H_0, p est constant estimé par n_1/n, et que sous H_1, $\hat{p}(x) = \dfrac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}$, il vient :

$$\Lambda(\underline{y}|\underline{x}) = 2[\sum_{i=1}^{n} (y_i \ln(\hat{p}(x_i)) + (1 - y_i) \ln(1 - \hat{p}(x_i)) - (n_0 \ln(n_0) + n_1 \ln(n_1) - n \ln(n))].$$

Pour notre exemple, nous obtenons $n_0 = 57$, $n_1 = 43$ et

$$\Lambda(\underline{y}|\underline{x}) = 2[53,677 - (57 \ln(57) + 43 \ln(43) - 100 \ln(100))] = 29.31$$

alors la p-valeur associée à ces données est égale à $1 - F_{\chi^2(1)}(29.31)$, où $F_{\chi^2(1)}$ désigne la fonction de répartition de la loi du χ^2 à un degré de liberté. On constate que

$$1 - F_{\chi^2(1)}(29.31) < 10^{-4},$$

de sorte que l'hypothèse nulle (non influence de l'âge sur la maladie cardiaque) serait rejetée (*à condition qu'on accepte le modèle de régression logistique comme modèle général*) même à un niveau de signification aussi sévère que 10^{-4}.

IV.4 Comparaison de proportions

IV.4.1 Présentation de la problématique sur un exemple

Une situation classique, en statistique médicale, consiste en l'appréciation des effets comparés de 2 traitements (ou l'appréciation des effets d'un traitement face à l'absence d'administration de traitement) ; nous nous plaçons ici dans la situation la plus élémentaire, où les effets des traitements sont simplement résumés en "succès" ou "échec". Des équivalents peuvent être bien sûr proposés pour des applications en sciences humaines, en agronomie, industrielles etc...

On se trouve ici typiquement en présence de deux échantillons de suites de 0 (échec) et de 1 (succès), et nous supposerons que les v.a. observées, que nous noterons $X_{1,i}$ (où $1 \leq i \leq n_1$) et $X_{2,j}$ (où $1 \leq j \leq n_2$) sont indépendantes. Pour aller plus loin dans l'élaboration du modèle, il faut distinguer deux types de protocoles expérimentaux, selon que les échantillons sont dits **appariés** ou **non appariés**.

Dans le cas d'échantillons non appariés, on a pris indépendamment deux groupes de sujets et on a appliqué le premier traitement à l'un et le second traitement à l'autre ; le modèle associé sera très simple, mais on s'interdit alors toute prise en compte d'influences sur les sujets autres que la seule administration des traitements, et on risque ainsi de passer à côté de facteurs importants.

Dans le cas d'échantillons appariés, nécessairement de même taille n, le protocole a consisté en la sélection préalable de n "situations" comparables, dans chacune desquelles on a administré chacun des deux traitements ; un tel cas serait, si ceci est techniquement réalisable, celui où un même individu a été observé successivement sous le premier et sous le second traitement, en admettant qu'il y ait eu un intervalle de temps suffisant pour que l'effet du premier se soit estompé lors de l'administation du second, ceci étant en particulier le cas si en fait le "premier traitement" consistait en l'absence de traitement ; ou, plus simplement, on peut considérer que des sujets sont réunis par couples de manière qu'ils aient dans chaque couple des caractéristiques communes susceptibles d'agir sur la réceptivité aux traitements (par ex. l'âge, certains éléments du dossier médical ...) et que dans chaque couple les deux sujets reçoivent des traitements différents.

IV.4.2 Échantillons non appariés

Le modèle naturel ici est de considérer que les v.a. $X_{1,i}$, pour $1 \leq i \leq n_1$, sont i.i.d., la loi commune étant de Bernoulli avec paramètre p_1, et les v.a. $X_{2,j}$, pour $1 \leq j \leq n_2$, sont i.i.d., la loi commune étant de Bernoulli avec paramètre p_2. Le paramètre du modèle est donc $\theta = (p_1, p_2)$, qui appartient à $\Theta = [0,1]^2$ (Θ peut aussi, si on est suffisamment renseigné, être modélisé comme une partie stricte de $[0,1]^2$).

On sait (voir l'exemple II.17) que les proportions de 1 observées dans chacun des échantillons fournissent un résumé exhaustif des données. Nous travaillerons donc systématiquement sur ces proportions $\bar{x}_1$ et $\bar{x}_2$ (ou, ce qui revient au même, sur les effectifs $y_1 = n_1.\bar{x}_1$ et $y_2 = n_2.\bar{x}_2$).

Si on considère séparément les deux échantillons, on se trouve dans la situation déjà étudiée au II.1.1 pour ce qui est d'estimations, tests, intervalles de confiance sur p_1 et p_2 . Nous allons ici nous intéresser aux tests des **hypothèses nulles** suivantes :

- $H_0 = \{p_1 = p_2\}$: **identité des effets des deux traitements** ; autrement dit on cherchera à répondre à la question : *les proportions de succès observées différent-elles significativement, à un niveau de signification choisi au préalable ?*
- $H_1 = \{p_1 > p_2\}$: **le traitement 1 est plus efficace que le traitement 2** (en d'autres termes il a une plus forte probabilité de succès) ; autrement dit on cherchera à répondre à la question : *la proportion de succès observées avec le traitement 2 est-elle significativement plus forte que celle du traitement 1, ce qui conduirait à réfuter l'hypothèse que le premier était meilleur, à un niveau de signification choisi au préalable* ; c'est ainsi en particulier que l'on procède si le traitement 1 était déjà en usage et que l'on se demande si cela "vaut vraiment la peine" de mettre en usage le nouveau traitement 2. (En fait, on pourra vérifier que les tests mis en place pour cette hypothèse alternative correspond également à l'hypothèse nulle $H_0 = \{p_1 \leq p_2\}$: le traitement 1 n'est pas meilleur que le traitement 2.)

Nous serons amenés, comme au paragraphe II.6, à distinguer selon que les échantillons sont "petits" (et nous nous heurterons à des problèmes de calculs de nature combinatoire) ou "grands" (et nous aurons recours à l'approximation normale).

Petits échantillons non appariés : test bilatéral

Considérons tout d'abord le test de l'hypothèse nulle $H_0 = \{p_1 = p_2\}$, situation de test dite **bilatérale** car on inclut dans la contre-hypothèse, H_1, aussi bien le cas $\{p_1 < p_2\}$ que le cas $\{p_1 > p_2\}$.

Le bon sens est que la réfutation de l'hypothèse se fera si les proportions observées $\bar{x}_1$ et $\bar{x}_2$ sont suffisamment éloignées, c'est-à-dire intuitivement si, dans le cas où $p_1 = p_2$, la probabilité d'obtenir un éloignement des deux proportions "au moins aussi important" que celui enregistré est inférieure ou égale au seuil de signification choisi, α ; la difficulté pour donner un sens à cette intuition est double :

- l'hypothèse nulle n'est pas "simple", c'est-à-dire ne se réduit pas à une seule loi, mais se compose de tous les couples (p, p) où $p \in [0, 1]$ (éventuellement p appartient à un intervalle plus petit délimité au préalable lors de la constitution du modèle) ; en particulier la loi de $\bar{X}_1 - \bar{X}_2$ dépend de cette valeur commune inconnue p ;
- il faut donner un sens à la notion d'éloignement des deux proportions "au moins aussi important" que celui enregistré ; ne disposant pas de la loi de la différence $\bar{x}_1 - \bar{x}_2$, il nous est difficile de donner à ceci un sens symétrique, mais ce n'est pas trop gênant : nous nous contenterons de préciser la notion d'évènement consistant en un éloignement "au moins aussi important que celui observé **et dans le même sens**", et procèderons au rejet de l'hypothèse si les probabilités de cet évènement sont, sous l'hypothèse à tester, inférieures ou égales à $\frac{\alpha}{2}$.

La première difficulté se résout grâce à la remarque suivante : introduisons la variable aléatoire $Y = Y_1 + Y_2$ (on rappelle que $Y_1 = n_1.\bar{X}_1$ et $Y_2 = n_2.\bar{X}_2$). Le nombre y est le total de succès observés dans la réunion des deux échantillons. On constate alors que, **sous l'hypothèse nulle, la loi du couple (Y_1, Y_2), conditionnellement à Y, ne dépend plus de p $(= p_1 = p_2)$** ; un calcul élémentaire montre en effet que, pour $y_1 + y_2 = y$, on a :

$$\mathbb{P}_{p,p}((Y_1, Y_2) = (y_1, y_2) | Y = y) = \frac{y!(n_1 + n_2 - y)!n_1!n_2!}{y_1!(n_1 - y_1)!y_2!(n_2 - y_2)!(n_1 + n_2)!} ,$$

cette probabilité conditionnelle étant évidemment nulle si $y_1 + y_2 \neq y$.
(**Attention :** il ne s'agit pas d'exhaustivité de Y dans le modèle statistique considéré, mais d'une exhaustivité dans un modèle qui serait restreint à l'hypothèse nulle.)

Cette remarque conduit à une technique dite de **test conditionnel**. On considère la partition de l'ensemble Ω de tous les couples (y_1, y_2) observables en les "tranches" $\Omega_y = \{(y_1, y_2); y_1 + y_2 = y\}$ (y parcourant l'ensemble des valeurs pouvant être prises par la v.a. $Y_1 + Y_2$, donc comprises entre 0 et $n_1 + n_2$) . On fabrique séparément sur chaque Ω_y une région de rejet C_y dont, si l'hypothèse nulle est satisfaite, la probabilité, conditionnellement à l'évènement $\{Y = y\}$, est inférieure ou égale à α (le singulier "la probabilité" se justifiant ici en vertu de la remarque précédente). La région de rejet globale du test est alors $C = \bigcup_{y=0}^{n_1+n_2} C_y$ et on a bien :

$$\forall p \in]0,1[, \ \ \mathbb{P}_{p,p}(C) = \sum_{y=0}^{n_1+n_2} \mathbb{P}_{p,p}(Y = y)\mathbb{P}_{p,p}(C_y|Y = y) \leq \sum_{y=0}^{n_1+n_2} \mathbb{P}_{p,p}(Y = y)\alpha = \alpha \ .$$

Repérer si une observation d'effectifs (y_1, y_2) conduit au rejet, c'est-à-dire appartient à C_y, où $y = y_1 + y_2$, est alors élémentaire :

- si $\bar{x}_1 = \bar{x}_2$ (autrement dit $\frac{y_1}{n_1} = \frac{y_2}{n_2}$), il n'y a évidemment pas rejet ;
- si $\bar{x}_1 \neq \bar{x}_2$, on repère quel est le sens de cette différence ; soit par exemple $\bar{x}_1 < \bar{x}_2$, autrement dit $\frac{y_1}{n_1} < \frac{y_2}{n_2}$) ; on fait la liste de tous les couples "pires" (au sens large) que (y_1, y_2), à nombre total de succès, y, fixé, autrement dit tous les couples de la forme $(y_1', y_2') = (y_1 - j, y_2 + j)$, où $j \geq 0$, avec bien sûr $y_1 - j \geq 0$ et $y_2 + j \leq n_2$; les probabilités conditionnelles de ces couples, sous l'hypothèse nulle, sont connues : elles valent $\dfrac{y!(n_1 + n_2 - y)!n_1!n_2!}{y_1'!(n_1 - y_1')!y_2'!(n_2 - y_2')!(n_1 + n_2)!}$ (il s'agit des poids de probabilité d'une loi hypergéométrique $\mathcal{H}(n_1 + n_2, y_1 + y_2, \frac{n_1}{n_1+n_2})$, voir X.2.1, p. 243) ; il y a rejet si et seulement si la somme de ces probabilités (qui est la p-valeur de ce test) est inférieure ou égale à $\frac{\alpha}{2}$.

Petits échantillons non appariés : test unilatéral

Considérons maintenant le test de l'hypothèse nulle $H_0 = \{p_1 \geq p_2\}$ (situation de test dite **unilatérale** car la contre-hypothèse est $H_1 = \{p_1 < p_2\}$).

Il n'y aura évidemment pas rejet si les proportions observées "vont dans le sens" de l'hypothèse nulle, c'est-à-dire si $\bar{x}_1 \geq \bar{x}_2$. En revanche, il y aura rejet si on a une inégalité $\bar{x}_1 < \bar{x}_2$ significativement nette, c'est-à-dire, suivant la même ligne que dans le cas du test bilatéral, si, quand on est à la situation "frontière" $p_1 = p_2$, la probabilité, conditionnellement au nombre total y de succès observés, d'obtenir une situation "pire" que celle observée, est inférieure ou égale à α ; cette inégalité serait encore satisfaite, a fortiori, si on s'enfonçait dans la contre-hypothèse, c'est-à-dire si on considérait des couples (p_1, p_2) tels que $p_1 < p_2$.

La technique est donc la même que dans le cas bilatéral, à ceci près que la comparaison de la somme de probabilités calculée se fait avec α et non avec $\frac{\alpha}{2}$.

Grands échantillons non appariés

On considère que chacun des deux échantillons est assez grand pour qu'on applique à la v.a. "proportion de succès", dans chacun de ces échantillons, l'approximation normale.

Pratiquement, aux degrés de précision couramment exigés par les praticiens, cela suppose que $n_1 p_1(1-p_1) \geq 5$ et $n_2 p_2(1-p_2) \geq 5$, ou bien que l'on dispose de renseignements préalables sur des ordres de grandeur de p_1 et p_2 qui permettent d'affirmer qu'il en est bien ainsi, ou bien on fait une vérification approchée de ces conditions en y remplaçant p_1 et p_2 par leurs estimations respectives $\bar{x}_1$ et $\bar{x}_2$.

On déduit du TCL que les suites indépendantes $\sqrt{n_1}(\bar{X}_1 - p_1)$ et $\sqrt{n_2}(\bar{X}_2 - p_2)$ convergent respectivement en loi vers la loi $\mathcal{N}(0, p_1(1-p_1))$ et $\mathcal{N}(0, p_2(1-p_2))$. Si p_1 et p_2 sont égaux de valeur commune p, il est alors facile de vérifier que si $(n_1(k), n_2(k))_{k\in\mathbb{N}}$ est une suite qui diverge vers l'infini, telle que $\lim_{k\to\infty} n_1(k)/n_2(k)$ existe, et cette limite est finie et non nulle, alors la suite de v.a. $\sqrt{\dfrac{n_1(k)n_2(k)}{n_1(k)+n_2(k)}}(\bar{X}_1 - \bar{X}_2)$ converge en loi vers $\mathcal{N}(0, p(1-p))$. De plus, on a la convergence p.s. de $\bar{X} = (n_1\bar{X}_1 + n_2\bar{X}_2)/(n_1+n_2)$ vers p. On en déduit que la statistique de test

$$\zeta_{n_1,n_2} = \sqrt{\frac{n_1(k)n_2(k)}{n_1(k)+n_2(k)}}\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\bar{X}(1-\bar{X})}}$$

converge en loi vers la loi $\mathcal{N}(0,1)$.

Considérons tout d'abord le test bilatéral, dont l'hypothèse nulle est $H_0 = \{p_1 = p_2\}$ et l'hypothèse alternative $H_1 = \{p_1 \neq p_2\}$.

Le rejet se fera alors si la distance $|\bar{x}_1 - \bar{x}_2|$ entre les proportions observées est assez élevée. En effet, sous H_1, la statistique de test diverge p.s. vers $+\infty$ ou $-\infty$, car $\bar{X}_1 - \bar{X}_2$ converge vers $p_1 - p_2 \neq 0$ et $\bar{X}(1-\bar{X})$ a une limite finie. On en déduit la région critique de la forme $]-\infty, -a] \cup [a, \infty[$ pour $a > 0$. Le test est de niveau asympotique α pour $a = \phi_{1-\alpha/2}$, le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$.

On rejette l'hypothèse si et seulement si $\left|\zeta^{\text{obs}}_{n_1,n_2}\right| = \sqrt{\dfrac{n_1 n_2}{n_1+n_2}}\dfrac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\bar{x}(1-\bar{x})}} > \phi_{1-\alpha/2}$, où $\bar{x} = (n_1\bar{x}_1 + n_2\bar{x}_2)/(n_1+n_2)$. La p-valeur est donnée par $\mathbb{P}(|G| \geq \left|\zeta^{\text{obs}}_{n_1,n_2}\right|)$, où G est de loi $\mathcal{N}(0,1)$. En d'autres termes, Φ désignant la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0,1)$, il y a rejet si la p-valeur $2(1 - \Phi(\left|\zeta^{\text{obs}}_{n_1,n_2}\right|)$ est plus petite que α ; on rappelle que la p-valeur d'une observation, pour une technique de test donnée, est le plus petit choix du niveau de signification pour lequel l'observation effectuée conduise au rejet.

Passons au cas du test unilatéral, dont l'hypothèse nulle est $H_0 = \{p_1 \geq p_2\}$ et l'hypothèse alternative $H_1 = \{p_1 < p_2\}$.

Ici on rejette l'hypothèse nulle si $\bar{x}_2$ est, au niveau de signification fixé, significativement plus grand que $\bar{x}_1$. On est donc conduit à la technique suivante, qui assure (aux approximations faites près) une probabilité de rejet de α si on est sur la frontière entre hypothèse nulle et hypothèse alternative, c'est-à-dire si $p_1 = p_2$, et une probabilité de rejet inférieure à α si on est au sein de l'hypothèse alternative, c'est-à-dire si $p_1 < p_2$: on rejette l'hypothèse si et seulement si $\zeta^{\text{obs}}_{n_1,n_2} = \sqrt{\dfrac{n_1 n_2}{n_1+n_2}}\dfrac{\bar{x}_2 - \bar{x}_1}{\sqrt{\bar{x}(1-\bar{x})}} > \phi_{1-\alpha}$.

En d'autres termes il y a rejet si la p-valeur $1 - \Phi(\zeta^{\text{obs}}_{n_1,n_2})$ est plus petite que α.

IV.4.3 Échantillons appariés

Rappelons qu'il s'agit de prendre en compte le fait que des facteurs variant de couple à couple peuvent influer sur l'efficacité de chacun des traitements, c'est-à-dire qu'il y a en fait autant de valeurs du paramètre que de variables aléatoires. Nous modélisons donc n v.a. $X_{1,i}$ et n v.a. $X_{2,i}$, supposées comme précédemment toutes indépendantes, la loi de $X_{1,i}$ (resp. $X_{2,i}$) étant de Bernoulli avec paramètre $p_{1,i}$ (resp. $p_{2,i}$) : $X_{1,i}$ (resp. $X_{2,i}$) est le résultat obtenu par application du traitement 1 (resp. 2) dans la paire i (ces v.a. valant 1 en cas d'efficacité, 0 dans le cas contraire).

Petits échantillons appariés

Nous présentons d'abord le test bilatéral de l'hypothèse de non influence d'un changement de traitement, l'expérimentation se faisant sur des couples de sujets. L'hypothèse nulle est donc composée des n égalités $p_{1,i} = p_{2,i}$ (où $1 \leq i \leq n$). L'hypothèse alternative est très vaste, puisqu'elle comporte tous les couples de suites $((p_{1,i})_{1\leq i\leq n}, (p_{2,i})_{1\leq i\leq n})$ ne vérifiant pas ces n égalités (mais dans la pratique le test ne détectera bien, notion que l'on pourrait traduire par la convergence du test, pour procéder à un rejet, que les situations où il existe un suffisamment grand nombre d'inégalités bien marquées). La difficulté technique, dans la fabrication du test, vient du fait que l'hypothèse nulle elle-même est "complexe", de dimension n car le paramètre s'y compose de n couples (p_i, p_i).

Il est intuitif que l'on aura tendance à rejeter l'hypothèse nulle si on observe un suffisamment grand nombre de couples où les 2 traitements ne donnent pas le même résultat, avec parmi eux une tendance suffisamment nette en faveur de l'un des deux traitements. Cette intuition nous conduit à négliger, dans la fabrication du test, les couples à résultats identiques $(0,0)$ ou $(1,1)$ et à recourir, comme dans le cas des échantillons non appariés, à une technique de test conditionnel, le conditionnement se faisant cette fois relativement à la v.a. $Z = Z_{0,1} + Z_{1,0}$, où $Z_{0,1}$ (resp. $Z_{1,0}$) désigne le nombre de couples pour lesquels a été observé $(0,1)$ (resp. $(1,0)$) ; en d'autres termes, Z est la v.a. : "nombre de couples pour lesquels les deux résultats diffèrent".

Un calcul élémentaire établit alors que, conditionnellement à l'évènement $\{Z = z\}$, et si l'hypothèse nulle est satisfaite, la loi de $Z_{0,1}$ est la loi binomiale de paramètres z et $\frac{1}{2}$ (ce qui équivaut au fait que la loi de $Z_{1,0}$ est aussi la loi binomiale de paramètres z et $\frac{1}{2}$). C'est naturel, puisque ceci traduit une situation d'équilibre. Le rejet de l'hypothèse nulle, au seuil α, s'effectuera alors si la valeur observée $z_{0,1}$ est en dehors de l'intervalle des quantiles d'ordre $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ de la loi $\mathcal{B}(z, \frac{1}{2})$ ou, ce qui revient au même, si $\inf(z_{0,1}, z_{1,0})$ est inférieur au quantile d'ordre $\frac{\alpha}{2}$ de la loi $\mathcal{B}(z, \frac{1}{2})$.

Ce test est appelé **test des signes**, car souvent on marque du signe + les couples $(0,1)$, du signe - les couples $(1,0)$ et du signe 0 les couples $(0,0)$ ou $(1,1)$, et on est donc amené à considérer la proportion de l'un des signes + ou - sur l'ensemble des signes autres que 0.

On peut aussi, de manière équivalente, procéder en utilisant la p-valeur du test qui est ici le double de la valeur en $\inf(z_{0,1}, z_{1,0})$ de la fonction de répartition de $\mathcal{B}(z, \frac{1}{2})$: il y a rejet si elle est inférieure ou égale à α.

Si on considère maintenant pour hypothèse nulle la famille des inégalités $p_{1,i} \geq p_{2,i}$, le rejet se fera évidemment si le nombre de couples $(1,0)$ est significativement faible devant le nombre de couples $(0,1)$; le principe du test est donc le même que dans

le cas bilatéral qu'on vient d'étudier, mais cette fois le rejet se fait quand $z_{1,0}$ (et non plus $\inf(z_{0,1}, z_{1,0})$) est inférieur au quantile d'ordre α (et non plus $\frac{\alpha}{2}$) de la loi $\mathcal{B}(z, \frac{1}{2})$.

Grands échantillons appariés

Si la valeur observée de z est assez grande (en pratique, comme d'habitude, si $z\frac{1}{2}(1-\frac{1}{2}) \geq 5$, c'est-à-dire $z \geq 20$), on approxime la loi $\mathcal{B}(z, \frac{1}{2})$ par la loi normale $\mathcal{N}(\frac{z}{2}, \frac{z}{4})$; autrement dit, pour z grand, on approxime la loi, conditionnellement à $Z = Z_{0,1} + Z_{1,0}$, de $\frac{Z_{0,1}-Z_{1,0}}{\sqrt{Z_{0,1}+Z_{1,0}}}$ par la loi $\mathcal{N}(0,1)$ (en effet $Z_{0,1} - Z_{1,0} = 2Z_{0,1} - Z$).

Le test en résulte immédiatement, qui consiste, **dans le cas bilatéral**, à rejeter l'hypothèse nulle si $\frac{|z_{0,1} - z_{1,0}|}{\sqrt{z_{0,1} + z_{1,0}}} \geq \phi_{1-\alpha/2}$, où $\phi_{1-\alpha/2}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de $\mathcal{N}(0,1)$.

Nous laissons au lecteur le soin d'effectuer l'adaptation au cas unilatéral.

IV.5 Résumé des tests

IV.5.1 Test d'adéquation à une loi discrète : le test du χ^2

L'objectif est de déterminer si les données discrètes observées proviennent d'une loi donnée ou non.

1. Description du modèle : $(X_j, 1 \leq j \leq n)$ est une suite de v.a. i.i.d. à valeurs dans $A = \{a_1, \ldots, a_k\}$. Une loi P_p sur A est décrite par le paramètre $p = (p_1, \ldots, p_k)$, où $p_i = \mathbb{P}_p(X_1 = a_i)$.
2. Les hypothèses : $H_0 = \{p = p^0\}$ et $H_1 = \{p \neq p^0\}$, où p_0 est donné.
3. La statistique de test :
$$\zeta_n = n \sum_{i=1}^{k} \frac{(\hat{p}_i - p_i^0)^2}{p_i^0},$$
où $\hat{p}_i$ est le nombre d'occurrence de a_i divisé par n.
4. Sous H_0, $(\zeta_n, n \geq 1)$ converge en loi vers $\chi^2(k-1)$.
5. Sous H_1, $(\zeta_n, n \geq 1)$ diverge vers $+\infty$.
6. Région de critique du test asymptotique : $[a, +\infty[$.
7. Niveau asymptotique du test égal à α : a est le quantile d'ordre $1-\alpha$ de la loi $\chi^2(k-1)$.
8. Le test est convergent.
9. La p-valeur asymptotique est donnée par
$$p\text{-valeur} = \mathbb{P}(Z \geq \zeta_n^{\mathrm{obs}}),$$
où Z est de loi $\chi^2(k-1)$, et ζ_n^{obs} est la statistique de test calculée avec les observations.

Le test asymptotique est considéré valide si $np_i^0(1-p_i^0) \geq 5$ pour tout i.

IV.5.2 Test d'indépendance entre deux variables qualitatives

L'objectif est de vérifier si deux variables catégorielles sont indépendantes ou non.

1. Description du modèle : $((Y_i, Z_i), 1 \leq i \leq n)$ est une suite de v.a. i.i.d. respectivement à valeurs dans $A = \{a_1, \ldots, a_k\}$ et $B = \{b_1, \ldots, b_m\}$. Une loi commune P_p des couples (Y_i, Z_i) sur (A, B) est décrite par le paramètre $p = (p_{j,h})_{1 \leq j \leq k, 1 \leq l \leq m}$ où $p_{j,l} = \mathbb{P}_p((Y_i, Z_i) = (a_j, b_l))$.
2. Les hypothèses : $H_0 = \{p_{j,l} = q_j r_l\}_{1 \leq j \leq k, 1 \leq l \leq m}$ et $H_1 = \{\exists j, l; p_{j,l} \neq q_j r_l\}$, où $q_j = \sum_{l=1}^{m} p_{j,l}$ et $r_l = \sum_{j=1}^{k} p_{j,l}$.
3. La statistique de test :
$$\zeta_n = n \sum_{j=1}^{k} \sum_{l=1}^{m} \frac{(\hat{p}_{j,l} - \hat{q}_j \hat{r}_l)^2}{\hat{q}_j \hat{r}_l},$$
où $\hat{p}_{j,l}$, $\hat{q}_j$ et $\hat{r}_l$ sont respectivement les nombres d'occurrence de (a_j, b_l), de a_j et de b_l divisé par n.
4. Sous H_0, $(\zeta_n, n \geq 1)$ converge en loi vers $\chi^2((k-1)(m-1))$.
5. Sous H_1, $(\zeta_n, n \geq 1)$ diverge vers $+\infty$.
6. Région de critique du test asymptotique : $[a, +\infty[$.
7. Niveau asymptotique du test égal à α : a est le quantile d'ordre $1-\alpha$ de la loi $\chi^2((k-1)(m-1))$.
8. Le test est convergent.
9. La p-valeur asymptotique est donnée par
$$p\text{-valeur} = \mathbb{P}(Z \geq \zeta_n^{\mathrm{obs}}),$$
où Z est de loi $\chi^2((k-1)(m-1))$, et ζ_n^{obs} est la statistique de test calculée avec les observations.

Le test asymptotique est considéré valide si $n\hat{q}_j\hat{r}_l(1 - \hat{q}_j\hat{r}_l) \geq 5$ pour tout (j, l).

IV.5.3 Régression logistique

L'objectif est de savoir si une variable explique significativement ou non une variable réponse booléenne.

1. Description du modèle : $((X_i, Y_i), 1 \leq i \leq n)$ est une suite de v.a. i.i.d. respectivement à valeurs dans $\mathbb{R}$ et dans $\{0;1\}$. La liason entre les deux variables est décrite par le modèle logit : $\mathbb{P}(Y_i = 1 | X_i = x) = p_\psi(x) = \dfrac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$, où $\psi = (\beta_0, \beta_1) \in \mathbb{R}^2$.
2. Les hypothèses : $H_0 = \{\beta_1 = 0\}$ et $H_1 = \{\beta_1 \neq 0\}$.
3. La statistique de test :
$$\Lambda_n(Y|X) = 2(\ell_n(Y|X; \hat{\psi}) - \ell_n(Y|X; \hat{\psi}_{H_0}))$$
où $Y = (Y_1, \ldots, Y_n)$, $X = (X_1, \ldots, X_n)$, $\hat{\psi} = (\hat{\beta}_0, \hat{\beta}_1)$ et $\hat{\psi}_{H_0} = (\tilde{\beta}_0, 0)$ sont respectivement les paramètres estimés sous le modèle complet et sous le modèle restreint à l'hypothèse nulle, et $\ell_n(Y|X; \psi) = \sum_{i=1}^{n} Y_i \ln p_\psi(X_i) - (1 - Y_i) \ln(1 - p_\psi(X_i))$.

4. Sous H_0, $\Lambda_n(Y|X)$ converge en loi vers $\chi^2(1)$.
5. Sous H_1, $\Lambda_n(Y|X)$ diverge vers $+\infty$.
6. Région de critique du test asymptotique : $[a, +\infty[$.
7. Niveau asymptotique du test égal à α : a est le quantile d'ordre $1-\alpha$ de la loi $\chi^2(1)$.
8. Le test est convergent.
9. La p-valeur asymptotique est donnée par
$$p\text{-valeur} = \mathbb{P}(Z \geq \Lambda_n(y^{\text{obs}}|x^{\text{obs}}),$$
où Z est de loi $\chi^2(1)$, et $\Lambda_n(y^{\text{obs}}|x^{\text{obs}})$ est la statistique de test calculée avec les observations.

IV.5.4 Comparaison de deux proportions sur échantillons non appariés

1. Description du modèle : $(X_{1,k}, 1 \leq k \leq n_1)$ et $(X_{2,j}, 1 \leq j \leq n_2)$ sont deux suites indépendantes de v.a. i.i.d., de loi de Bernoulli de paramètres respectifs p_1 et p_2.
2. Les hypothèses : $H_0 = \{p_1 = p_2\}$ et $H_1 = \{p_1 \neq p_2\}$.
3. La statistique de test : Y_1 conditionnellement à $Y = y$ où $Y_1 = \sum_{k=1}^{n_1} X_{1,k}$, $Y_2 = \sum_{j=1}^{n_2} X_{2,j}$ et $Y = Y_1 + Y_2$.
4. Sous H_0, la loi de Y_1 conditionnellement à $Y = y$ est la loi hypergéométrique $\mathcal{H}(n_1 + n_2, y, p)$ où $p = n_1/(n_1 + n_2)$.
5. Sous H_1, Y_1 conditionnellement à $Y = y$ a tendance à prendre de très petites valeurs ou de très grandes valeurs.
6. Région de critique du test : $\{y - n_2, \ldots, y_- - 1\} \cup \{y_+, \ldots, n_1\}$.
7. Niveau du test par excès égal à α : y_-, resp. y_+, est le quantile d'ordre $\alpha/2$, resp. $1-\alpha/2$, de la loi $\mathcal{H}(n_1 + n_2, y, p)$.
8. Puissance du test : non crucial.
9. p-valeur : si $y_1/n_1 < y_2/n_2$ (correspondant à $H_0 = \{p_1 \geq p_2\}$ et $H_1 = \{p_1 < p_2\}$), elle est donnée par
$$\sum_{r=\max(y-n_2,0)}^{y_1} \frac{C_{n_1}^r C_{n_2}^{y-r}}{C_y^r},$$
si $y_1/n_1 > y_2/n_2$ (correspondant à $H_0 = \{p_1 \leq p_2\}$ et $H_1 = \{p_1 > p_2\}$), elle est donnée par
$$\sum_{r=y_1}^{\min(n_1,y)} \frac{C_{n_1}^r C_{n_2}^{y-r}}{C_y^r}.$$
10. Variante pour grands échantillons : statistique de test
$$\zeta_{n_1,n_2} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\bar{X}(1-\bar{X})}},$$
où $\bar{X}_1 = Y_1/n_1$, $\bar{X}_2 = Y_2/n_2$ et $\bar{X} = (Y_1 + Y_2)/(n_1 + n_2)$; sous H_0, ζ_{n_1,n_2} converge en loi ($n_1 \to \infty$, $n_2 \to \infty$) vers la loi $\mathcal{N}(0,1)$; sous H_1, ζ_{n_1,n_2} diverge vers $+\infty$ ou $-\infty$; région critique $]-\infty, -a] \cup [a, +\infty[$; niveau asymptotique α pour $a = \phi_{1-\alpha/2}$, le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$; le test est convergent ; p-valeur égale à $\mathbb{P}(|G| \geq |\zeta_{n_1,n_2}^{\text{obs}}|)$, où G de loi $\mathcal{N}(0,1)$.

IV.5.5 Comparaison de deux proportions sur échantillons appariés

1. Description du modèle : $((X_{1,i}, X_{2,i}), 1 \leq i \leq n)$ suite de v.a. indépendantes, $X_{1,i}$ de loi de Bernoulli de paramètre $p_{1,i}$, indépendante de $X_{2,i}$, de loi de Bernoulli de paramètre $p_{2,i}$.
2. Les hypothèses : $H_0 = \{p_{1,i} = p_{2,i}; 1 \leq i \leq n\}$ et $H_1 = \{p_{1,i}$ significativement plus grand (ou plus petit) que $p_{2,i}$ pour un nombre significatif d'indices $i\}$.
3. La statistique de test : $Z_{0,1}$ conditionnellement à $Z = z$, où $Z_{0,1} = \sum_{i=1}^{n} \mathbf{1}_{\{X_{1,i}=0, X_{2,i}=1\}}$, $Z_{1,0} = \sum_{i=1}^{n} \mathbf{1}_{\{X_{1,i}=1, X_{2,i}=0\}}$ et $Z = Z_{0,1} + Z_{1,0}$ (Z représente le nombre de couples pour lesquels les résultats différent).
4. Sous H_0, la loi de $Z_{0,1}$ conditionnellement à $Z = z$ est la loi $\mathcal{B}(z, 1/2)$.
5. Sous H_1, conditionnellement à $Z = z$, $Z_{0,1}$ a tendance à prendre de très petites valeurs ou de très grandes valeurs.
6. Région de critique du test : $\{0, \ldots, z_- - 1\} \cup \{z_+, \ldots, n\}$.
7. Niveau du test par excès égal à α : z_- (resp. z_+) est le quantile d'ordre $\alpha/2$ (resp. $1 - \alpha/2$) de la loi $\mathcal{B}(z, 1/2)$.
8. Puissance du test : non crucial.
9. La p-valeur est $2\mathbb{P}(W \leq \min(z_{0,1}^{\text{obs}}, z - z_{0,1}^{\text{obs}}))$, où W est de loi $\mathcal{B}(z, 1/2)$ et $z_{0,1}^{\text{obs}}$ est la statistique de test calculée avec les observations.
10. Variante pour grands échantillons, sous des hypothèses raisonnables sur les $p_{j,i}$ sous H_0 et H_1 : statistique de test
$$\zeta_n = \frac{Z_{1,0} - Z_{0,1}}{\sqrt{Z_{1,0} + Z_{0,1}}};$$
sous H_0, ζ_n converge en loi vers la loi $\mathcal{N}(0,1)$; sous H_1, ζ_n diverge vers $+\infty$ ou $-\infty$; région critique $]-\infty, -a] \cup [a, +\infty[$; niveau asymptotique α pour $a = \phi_{1-\alpha/2}$, le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$; le test est convergent ; p-valeur égale à $\mathbb{P}(|G| \geq |\zeta_{n_1,n_2}^{\text{obs}}|)$, où G de loi $\mathcal{N}(0,1)$.

Le test asymptotique est considéré valide si $z > 20$.

Chapitre V

Tests non paramétriques

V.1 Pourquoi la statistique non paramétrique ?

V.1.1 Se libérer du modèle paramétrique

La statistique *non paramétrique* cherche à faire le moins d'hypothèses possible sur la forme éventuelle du modèle qui a donné naissance aux données observées. Elle se distingue donc de la statistique dite *paramétrique* déclinée jusqu'ici, qui était caractérisée, dans le contexte général des modèles statistiques par une famille paramétrée de probabilités $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$, où l'ensemble de paramètres Θ était contenu dans $\mathbb{R}^k$, et on dit souvent dans un tel cas qu'il s'agit d'un modèle à k paramètres (en sous-entendant l'adjectif "réels"). En statistique non paramétrique, cette inclusion dans $\mathbb{R}^k$ n'est plus stipulée. La loi de probabilité des observations appartient à un ensemble plus vaste, par exemple l'ensemble des lois absolument continues sur $\mathbb{R}$. Les buts restent néanmoins les mêmes : construire des tests. Pour cela, elle utilise des outils spécifiques : essentiellement les rangs, les signes et les statistiques d'ordre des observations. Les calculs impliqués sont souvent bien plus simples que dans le cas paramétrique.

Dans ce cours, **nous insisterons sur les tests non paramétriques**. La grande différence avec les tests paramétriques précédents tient à la taille des hypothèses H_0 et H_1. Dans le modèle normal linéaire, par exemple, ces hypothèses sont décrites par un (petit) nombre de paramètres réels comme les moyennes dans les cas simples. Dans le cas non paramétrique, par contre, la description des hypothèses nulles H_0 du type "ces deux échantillons sont tirés d'une même loi" ne peut pas se ramener à un nombre fini de paramètres réels. Quant aux hypothèses alternatives H_1, elles sont souvent tellement vastes qu'elles échappent à toute tentative de description. Nous essaierons néanmoins de préciser dans certains cas la forme possible d'hypothèses alternatives afin de justifier la construction des tests.

V.1.2 Quand faut-il utiliser du non paramétrique ?

Le non paramétrique (expression familière pour "modèle non paramétrique") permet d'aborder des problèmes complexes qu'il est difficile de modéliser correctement en utilisant un modèle où les relations doivent être explicitées à l'aide de paramètres réels : notations données par des examinateurs ayant des barèmes différents, petits échantillons quand on n'a aucun moyen de justifier l'hypothèse gaussienne, données quantitatives avec variance hétérogènes.

Il existe 2 situations : la première situation est celle où la question posée entraine natu-

rellement le choix d'un modèle non paramétrique (par exemple les données sont-elles gaussiennes ?) et une seconde situation dans laquelle on peut hésiter entre un modèle paramétrique et un modèle non paramétrique (par exemple ces 2 séries de données ont-elles la même loi de probabilité ?)

Dans ce dernier cas la difficulté est qu'il n'existe pas systématiquement un test non paramétrique adapté à une situation donnée. Pour mettre en évidence cette difficulté, précisons que le modèle linéaire gaussien (voir chapitre 2) permet de traiter un nombre arbitraire de facteurs explicatifs et de régresser sur un nombre quelconque de variables (la limite informatique du nombre de variables sur un gros système est de l'ordre de 1 000). A contrario, les méthodes non paramétriques d'analyse de la variance ne permettent que de traiter deux facteurs au plus (avec des contraintes d'équilibre) et celles de régression non paramétrique ne peuvent être mises en œuvre que sur une seule variable ! On observe aussi que les tests paramétriques sont souvent plus puissants que les tests non paramétriques équivalents mais les tests non paramétriques sont plus simples à mettre en œuvre. La stratégie à adopter est donc un compromis entre un modèle avec de nombreux paramètres mais un test plus puissant, et un modèle avec une variabilité plus grande mais un test plus simple.

V.1.3 Exemples

Exemple V.1. Revenons sur l'exemple III.3.1 concernant la mesure de taux d'alcool dans le sang. (en dg/l) de n sujets : voici le tableau des observations, avec $n = 30$ (extrait de l'ouvrage de D. Schwartz, *Méthodes statistiques à l'usage des médecins et des biologistes*, Flammarion).

27	26	26	29	10	28	26	23	14	37	16	18	26	27	24
19	11	19	16	18	27	10	37	24	18	26	23	26	19	37

TAB. V.1 – Taux d'alcool dans le sang. (en dg/l) de $n = 30$ sujets

Nous avons étudié ces données en supposant qu'elles étaient gaussiennes : grâce aux tests non paramétriques et en particulier au test de Kolmogorov on peut valider cette hypothèse. Mais construire un test pour justifier du caractère gaussien des données ne peut se faire que dans le cadre non-paramétrique car l'hypothèse alternative est naturellement l'ensemble de toutes les autres lois de probabilité. ◊

Exemple V.2. Des sociologues s'intéressent à la maladie et à l'anxiété qui en découle. Des sociétés où il existe une explication orale de la maladie ("c'est parce qu'on a mangé du poison qu'on est malade", "c'est parce qu'on a reçu un sort", etc...) sont-elles différentes de celles où aucune explication orale n'existe ? Des "juges" différents ont donné une "note d'anxiété" à chacune de ces sociétés ; elles sont consignées dans le tableau ci-après, où les effectifs des deux échantillons sont respectivement $m = 16$ (sociétés sans explications orales) et $n = 23$ (sociétés avec explications orales).

Dans cet exemple les notes n'ont pas une valeur intrinsèque de mesure et il y a de plus beaucoup d'ex-aequo rendant difficile l'assimilation de cette loi à une loi gaussienne. En revanche le classement des sociétés les unes par rapport aux autres est significatif. Or nous verrons que par leur construction les tests non-paramétriques sont particulièrement adaptés à la comparaison de deux échantillons par l'intermédiaire de leurs statistiques d'ordre.

Sociétés sans expl. orales	Notes d'anxiété	Sociétés avec expl. orales	Notes d'anxiété
Lapp	13	Marquesans	17
Chamorro	12	Dobuans	16
Samoans	12	Baiga	15
Arapesh	10	Kwoma	15
Balinese	10	Thonga	15
Hopi	10	Alorese	14
Tanala	10	Chagga	14
Paiute	9	Navaho	14
Chenchu	8	Dahomeans	13
Teton	8	Lesu	13
Flathead	7	Masai	13
Papago	7	Lepeha	12
Wenda	7	Maori	12
Warrau	7	Pukapukans	12
Wogeo	7	Trobianders	12
Ontong	6	Kwakiull	11
		Manus	11
		Chiricahua	10
		Comanche	10
		Siriono	10
		Bena	8
		Slave	8
		Kurtatchi	6

TAB. V.2 – Tableau extrait de Siegel, 1956 : Angoisse dans les sociétés primitives.

◊

Exemple V.3. On cherche à étudier l'effet d'une certaine drogue sur la pression artérielle. Huit personnes souffrant d'hypertension sont tirées au hasard dans une population d'hypertendus. On mesure la pression artérielle de chacune d'elle, immédiatement avant puis 2 heures après l'administration de cette drogue. Les résultats sont les suivants :

N° du patient	1	2	3	4	5	6	7	8
avant traitement X	192	169	187	160	167	176	185	179
après traitement Y	196	165	166	157	170	145	168	177

Dans cet exemple, on a 2 petits échantillons appariés (c.f. paragraphe IV.4) : la statistique non-paramétrique permet de construire des tests simples (par exemple le test de Wilcoxon) fondés sur le signe des différences entre les échantillons et le rang des différences, permettant de tester que les 2 distributions suivent la même loi de probabilité.

◊

V.2 Les outils statistiques du non paramétrique

Les problèmes de test, en statistique non-paramétrique, conduisent souvent à ce que, sous l'hypothèse nulle, les observations soient i.i.d. alors que sous l'hypothèse alternative, l'une au moins des hypothèses d'indépendance ou d'identité des distributions est mise en défaut. Ainsi, sous H_0, il est possible de contruire des statistiques de test dont la loi est calculable. Les tests intéressants seront ceux basés sur les statistiques dites libres :

Définition V.4. *Une statistique est dite libre, pour des observations i.i.d., si sa loi ne dépend que de la taille de l'échantillon (et non de la loi commune des observations composant cet échantillon).*

Pour de petits échantillons, on peut alors construire des tables et en déduire la construction de tests. Pour les grandes valeurs, on montre en général que la statistique étudiée converge en loi (souvent vers une loi gaussienne ou la loi d'une fonction simple d'un vecteur gaussien). Ces statistiques sont souvent calculées à partir **des signes et des rangs** d'un échantillon.

Considérons 3 observations x_1, x_2 et x_3 vérifiant : $x_3 < x_1 < x_2$; il est naturel de dire que, en croissant, le premier est "3", le second est "1" et le troisième est "2" ; en d'autres termes, le rang de "1" est $r_1 = 2$, celui de "2" est $r_2 = 3$ et celui de "3" est $r_3 = 1$. Ceci nous conduit à la notion générale de suite des rangs $(r_1, \dots r_n)$ associée à une suite d'observations réelles toutes distinctes $(x_1, \dots, x_n)$; c'est la permutation (application bijective de $\{1, \dots, n\}$ dans lui-même) dont la réciproque, soit $(s_1, \dots, s_n)$, vérifie : $x_{s_1} < x_{s_2} < \dots < x_{s_n}$.

Cette définition est bien établie si les x_i sont tous distincts. S'il n'en est pas ainsi, il faut remplacer les inégalités strictes dans la définition précédente par des inégalités larges, et plusieurs permutations y satisfont en cas d'ex-aequo. Nous ne considèrerons que des situations où les observations i.i.d. sont de loi commune à fonction de répartition continue de sorte que les variables aléatoires de rangs, $R_1, \dots, R_n$ seront uniquement définies sauf sur une partie de probabilité nulle de $\mathbb{R}^n$; c'est pourquoi nous n'aurons aucun problème pour le calcul de la loi de statistiques dans la définition desquelles interviendront les rangs des observations. Dans ce contexte, on a :

Proposition V.5. *Soit $(X_1, \dots, X_n)$ est un n-échantillon d'une loi de fonction de répartition F continue, de médiane*[1] *μ. Alors les v.a.*[2] *(signe $(X_1 - \mu), \dots,$ signe$(X_n - \mu))$ sont i.i.d. avec*

$$\mathbb{P}(\textit{signe } (X_i - \mu) = +) = \mathbb{P}(\textit{signe } (X_i - \mu) = -) = \frac{1}{2}.$$

En outre les rangs $R_1, \dots, R_n$ des X_i $(1 \leq i \leq n)$ dans l'ordre croissant vérifient que pour toute permutation σ de $\{1, \dots, n\}$,

$$\mathbb{P}(R_1 = \sigma(1), ..., \ R_n = \sigma(n)) = \tfrac{1}{n!}$$

Démonstration. Le premier résultat est une simple conséquence de l'indépendance et de la définition de la médiane.

[1] La médiane d'une v.a. X est le nombre μ tel que $\mathbb{P}(X \leq \mu) = \mathbb{P}(X \geq \mu)$, ou l'un de ces nombres μ s'il n'y a pas unicité, ce qui se produit s'il existe un intervalle non réduit à un point sur lequel la fonction de répartition de X prend la valeur 1/2.

[2] signe (t) prend ses valeurs dans l'ensemble $\{-, +\}$, en convenant que signe $(0) = +$; cette convention n'est pas gênante quand F est continue car alors $\mathbb{P}(X = 0) = 0$.

Soit τ une permutation de $1, ..., n$, alors $X_{\tau(1)}, ..., X_{\tau_{(n)}}$ est encore un $n-$échantillon de la loi F. Donc pour une permutation σ, on a

$$\mathbb{P}\left(R_1 = \sigma(1), ..., R_n = \sigma(n)\right) = \mathbb{P}(R_{\tau(1)} = \sigma(1), ..., R_{\tau(n)} = \sigma(n)).$$

En prenant $\tau = \sigma$ on trouve que cette probabilité est égale à : $\mathbb{P}(R_1 = 1, ..., R_n = n)$. En conclusion, tous les classements ont la même probabilité, comme il y en a $n!$, cette probabilité vaut $\frac{1}{n!}$. □

De ce théorème on déduit le résultat fondamental suivant : **Toute fonction g des rangs $g(R_1, ..., R_n)$ a une loi qui ne dépend pas de la fonction de répartition F mais seulement de n.** Ainsi $g(R_1, ..., R_n)$ est une statistique libre. Remarquons aussi que **les rangs sont invariants par transformation strictement croissante**. Le résultat suivant très important permet de ramener le cas où F est une fonction de répartition quelconque à celui de la loi uniforme sur $[0, 1]$.

Proposition V.6. *Soit $(X_1, \ldots, X_n)$ un n-échantillon suivant la loi de fonction de répartition F continue. Alors $(F(X_1), \ldots, F(X_n))$ est un n-échantillon suivant la loi uniforme sur $[0, 1]$ et les rangs des $F(X_i)$ sont p.s. les mêmes que ceux des X_i. Enfin p.s., $\forall t \in \mathbb{R}$, $\sum_{i=1}^n \mathbf{1}_{\{X_i \leq t\}} = \sum_{i=1}^n \mathbf{1}_{\{F(X_i) \leq F(t)\}}$.*

Démonstration. Nous allons effectuer la démonstration dans le cas où F est strictement croissante de $\mathbb{R}$ dans $]0, 1[$, ce qui assure que cette fonction continue admet un inverse strictement croissant et continu F^{-1}. Pour $u \in]0, 1[$,

$$\mathbb{P}(F(X_i) \leq u) = \mathbb{P}(X_i \leq F^{-1}(u)) = F(F^{-1}(u)) = u,$$

ce qui assure que les variables aléatoires $F(X_i)$ sont i.i.d. suivant la loi uniforme sur $[0, 1]$. Les deux autres assertions découlent facilement de la croissance stricte de F.

Dans le cas général où F n'est pas strictement croissante, la preuve est donnée en annexe en fin de chapitre. □

Définition V.7. *Soit $X = (X_1, X_2, ..., X_n)$ un n-échantillon suivant la loi de fonction de répartition continue F. Le vecteur aléatoire $\left(X_{(1)}, X_{(2)}, ..., X_{(n)}\right)$ obtenu en ordonnant les valeurs des X_i par ordre croissant est appelé statistique d'ordre de X. La variable $X_{(k)}$ est la $k^{ième}$ statistique d'ordre.*

Proposition V.8. *La fonction de répartition de la $k^{ième}$ statistique d'ordre est*

$$\mathbb{P}\left(X_{(k)} \leq x\right) = \sum_{j=k}^{n} C_n^j (F(x))^j \left(1 - F(x)\right)^{n-j}.$$

Démonstration. Par définition $\{X_{(k)} \leq x\}$ signifie "il y a au moins k valeurs dans X qui sont inférieures à x". Soit j le nombre de valeurs effectivement inférieures à x ; il y a C_n^j façons de choisir les X_i correspondants et la probabilité que ces X_i soient inférieurs à x est bien : $F(x)^j \left(1 - F(x)\right)^{n-j}$. En combinant tous ces éléments, on trouve le résultat. □

Un avantage incontesté du non paramétrique est de permettre de travailler sur des **données ordinales**, c'est-à-dire qui expriment l'ordre et ainsi de s'affranchir de la quantification absolue, car les échelles de quantifications et les procédures d'attributions de mesures sont souvent peu robustes. Par exemple, dans des tests de dégustation, on demande simplement aux juges de classer les divers produits. Enfin, des notes, attribuées par différents notateurs n'ayant pas les mêmes barèmes, à la fois en niveau moyen et en dispersion, posent d'importants problèmes statistiques pour les ramener à un barème commun. Il est très souvent préférable de considérer que la note en elle-même n'a aucune valeur, mais qu'elle exprime seulement les préférences du notateur : la note est considérée comme une donnée ordinale et le non paramétrique permet de s'affranchir de nombreuses difficultés provenant de l'échelle utilisée pour la notation.

V.3 Problèmes à un seul échantillon

V.3.1 Ajustement à une loi : le test de Kolmogorov

C'est un test d'ajustement à une loi, comme le test du χ^2 (voir X.1.3), mais qui s'applique à une variable quantitative. On veut tester l'hypothèse selon laquelle les données observées sont tirées d'une loi dont la fonction de répartition est F_0. Dans toute cette section, on considère que la vraie fonction de répartion inconnue F et F_0 **sont continues**. Le test est basé sur la différence entre la fonction de répartition F_0 de cette loi théorique et la fonction de répartition empirique $\hat{F}$ dont on rappelle la définition (voir X.1.3, p. 236) :

Définition V.9. *On définit la fonction de répartition empirique du n-échantillon* $\underline{X} = (X_1, ..., X_n)$*, par la fonction en escalier suivante :*

$$\hat{F}_{\underline{X}}(t) = \frac{\text{Card}\left(\{1 \leq i \leq n : X_i \leq t\}\right)}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq t\}}.$$

Remarque V.10. Notons que $\hat{F}_{\underline{X}}$ est continue à droite. ◊

Le test de Kolmogorov permet de tester l'hypothèse H_0 : "Les observations sont un échantillon de la loi F_0" contre sa négation. La statistique $D_{\underline{X}}$ de ce test est alors basée sur la distance maximale entre F_0 et $\hat{F}$, c'est à dire :

$$D_{\underline{X}} = \sup_{t \in \mathbb{R}} \left| F_0(t) - \hat{F}_{\underline{X}}(t) \right|.$$

Il s'agit d'un choix de distance raisonnable, car d'après le théorème de Glivenko-Cantelli (voir X.3.3, p. 251), sous H_0, $D_{\underline{X}}$ converge p.s. vers 0 lorsque n tend vers l'infini. La zone de rejet est alors de la forme : $\left\{D_{\underline{X}} > a\right\}$. Notons que comme $\hat{F}_{\underline{X}}$ est constante et égale à i/n sur l'intervalle $[X_{(i)}, X_{(i+1)}[$ tandis que F_0 est croissante sur cet intervalle,

$$\sup_{t \in [X_{(i)}, X_{(i+1)}[} \left| F_0(t) - \hat{F}_{\underline{X}}(t) \right| = \max(|F_0(X_{(i)}) - \frac{i}{n}|, |F_0(X_{(i+1)}) - \frac{i}{n}|).$$

On en déduit l'expression suivante très utile en pratique

$$D_{\underline{X}} = \max_{1 \leq i \leq n} \max(|F_0(X_{(i)}) - \frac{i-1}{n}|, |F_0(X_{(i)}) - \frac{i}{n}|).$$

La légitimité du choix de $D_{\underline{X}}$ comme statistique de test repose sur la proposition suivante :

Proposition V.11. *Sous H_0, $D_{\underline{X}}$ est une statistique libre.*

Démonstration. D'après la proposition V.6, presque sûrement

$$D_{\underline{X}} = \sup_{t\in\mathbb{R}} \left| F_0(t) - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{U_i \leq F_0(t)\}} \right|$$

où les variables $U_i = F_0(X_i)$ sont i.i.d. suivant la loi uniforme sur $[0,1]$. Il suffit ensuite de faire le changement de variable $u = F_0(t)$ pour conclure. □

La loi de $D_{\underline{X}}$ sous H_0 a été tabulée, ce qui donne des valeurs seuils a_α à ne pas dépasser pour que H_0 soit acceptable au niveau α. Les moyens actuels de calcul informatique permettent également d'approcher la loi de $D_{\underline{X}}$ à l'aide de simulations. Pour n grand, il existe une approximation décrite par la proposition suivante :

Proposition V.12. *Sous H_0, en posant $\zeta_n = \sqrt{n} D_{\underline{X}}$, on dispose du résultat asymptotique suivant : la suite $(\zeta_n, n \geq 1)$ converge en loi et pour tout $y > 0$, on a*

$$\mathbb{P}(\zeta_n \leq y) \xrightarrow[n\to\infty]{} \sum_{k=-\infty}^{\infty} (-1)^k \exp\left(-2k^2y^2\right).$$

Démonstration. Comme pour $t \in \mathbb{R}$, $\hat{F}_{\underline{X}}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq t\}}$ où les variables $\mathbf{1}_{\{X_i\leq t\}}$ sont i.i.d. suivant la loi de Bernoulli $\mathcal{B}(F_0(t))$, le TCL entraîne que $\sqrt{n}(F_0(t) - \hat{F}_{\underline{X}}(t))$ converge en loi vers Y_t de loi normale centrée $\mathcal{N}(0, F_0(t)(1-F_0(t)))$. Plus généralement, le théorème de la limite centrale multidimensionnel (voir X.3.2) assure que $\sqrt{n}(F_0(t_1) - \hat{F}_{\underline{X}}(t_1), \ldots, F_0(t_k) - \hat{F}_{\underline{X}}(t_k))$ converge en loi vers un vecteur gaussien centré $(Y_{t_1}, \ldots, Y_{t_k})$ de covariance donnée par $\mathrm{Cov}(Y_{t_i}, Y_{t_j}) = F_0(\min(t_i, t_j)) - F_0(t_i)F_0(t_j)$. En fait on montre que le processus $\sqrt{n}(F_0(t) - \hat{F}_{\underline{X}}(t))_{t\in\mathbb{R}}$ converge en loi vers "un processus gaussien centré" tel que $\mathrm{Cov}(Y_s, Y_t) = F_0(\min(s,t)) - F_0(s)F_0(t)$ et on montre que pour tout $y > 0$,

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}} |Y_t| \leq y\right) = \sum_{-\infty}^{+\infty} (-1)^k \exp\left(-2k^2y^2\right).$$

□

Proposition V.13. *Sous H_1, $\zeta_n = \sqrt{n} D_{\underline{X}}$ tend p.s. vers $+\infty$ avec n.*

Le test est donc nécessairement unilatéral à droite (rejet des valeurs trop grandes).

Démonstration. Sous H_1 la fonction de répartition commune des X_i, notée F est différente de F_0. Soit $t_1 \in \mathbb{R}$ tel que $F_0(t_1) \neq F(t_1)$. D'après la loi forte des grands nombres $\hat{F}_{\underline{X}}(t_1) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i\leq t_1\}}$ converge p.s. vers $\mathbb{E}\left[\mathbf{1}_{\{X_i\leq t_1\}}\right] = F(t_1)$. Donc $\sqrt{n}|F_0(t_1) - \hat{F}_{\underline{X}}(t_1)|$ tend p.s. vers $+\infty$ de même pour $\sqrt{n} D_{\underline{X}}$. □

Remarque V.14. Si F_0 est **non continue** (par exemple lorsqu'il s'agit d'une loi discrète), le test de Kolmogorov sous sa forme classique n'est pas valide (la proposition V.12 n'est valable que si F_0 est continue) : on peut montrer que $D_{\underline{X}}$ est alors plus "concentrée" à proximité de zéro que quand F est continue. ◊

Remarque V.15. On peut aussi envisager des contre-hypothèses plus fines, du type unilatéral : "la loi des données a une fontion de répartition F telle que $F \prec F_0$ au sens où $\forall t \in \mathbb{R}$, $F(t) \leq F_0(t)$ et $\exists t_0 \in \mathbb{R}$, $F(t_0) < F_0(t_0)$". Dans ce cas, la statistique de test s'écrit sans la valeur absolue (et sa loi est différente). $\Diamond$

V.3.2 Un exemple

On dispose des 10 données suivantes :

$$\underline{x} = (2.2\ ,\ 3.3\ ,\ 5.3\ ,\ 1.8\ ,\ 4.3\ ,\ 6.3\ ,\ 4.5\ ,\ 3.8\ ,\ 6.4\ ,\ 5.5)$$

La question naïve "ces observations proviennent-elles d'une loi normale de moyenne 4 et de variance 4 ? "va être formalisée sous l'énoncé : "tester, au niveau de signification 0.05, l'hypothèse nulle selon laquelle ces observations, supposées indépendantes et identiquement distribuées, ont pour loi commune la loi normale de moyenne 4 et variance 4 ".

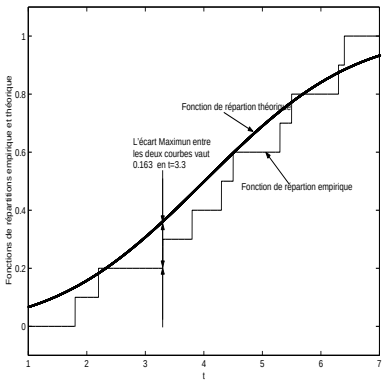


FIG. V.1 – Le test de Kolmogorov s'appuie sur la distance entre fonction de répartition empirique et théorique.

On calcule la fonction empirique dessinée sur la figure V.1. Elle montre que $D_{\underline{x}} = 0.163$, écart maximal obtenu en $t = 3.3$. Cette valeur est-elle plausible, au niveau 0.05, sous l'hypothèse H_0 ? Les praticiens ont l'habitude de faire la transformation de l'axe des abscisses $u = F(t)$. Cette transformation permet de travailler dans le carré $[0,1] \times [0,1]$ (cf figure V.2) où $D_{\underline{X}}$ mesure alors l'écart de la fonction de répartition empirique par rapport à la première bissectrice.

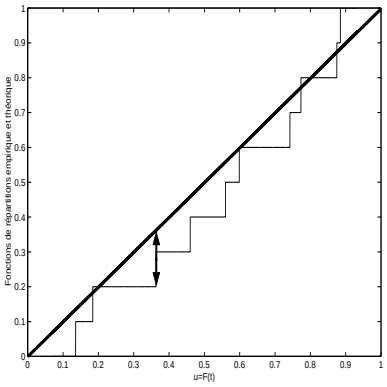


FIG. V.2 – Présentation usuelle de la distance de Kolmogorov.

En utilisant une table ou bien en approchant les quantiles de la loi de $D_{\underline{X}}$ sous H_0 par simulation d'un grand nombre de réalisations suivant cette loi, on remarque que la valeur observée $D_{\underline{x}} = 0.163$ est inférieure au quantile d'ordre 0.95 de la loi de $D_{\underline{X}}$: 0.410. (La p-valeur est de 0.963.)

L'hypothèse de référence H_0 est acceptée.

V.3.3 Test de normalité

Revenons à l'exemple V.1 des mesures de taux d'alcoolémie. On peut de la même manière tester H_0 : "Les données suivent une loi gaussienne de moyenne 23 et de variance 49" contre l'alternative : "c'est faux". On trouve $D_{\underline{x}} = 0.132$ donc on ne rejette pas H_0 pour les niveaux habituellement utilisés (quantile asymptotique d'ordre 0.95 égal à 0.242, et p-valeur asymptotique égale à 0.637). Dans ce problème on pourrait tester H_0 : "Les données suivent une loi gaussienne" contre l'alternative : "c'est faux", à l'aide du test de normalité de Lilliefors : ce test utilise la statistique de Kolmogorov déterminée par la distance entre la loi empirique et la loi gaussienne dont l'espérance est la moyenne empirique et la variance, la variance empirique. Les quantiles sont différents des quantiles du test de Kolmogorov et peuvent être calculés par simulation. Il existe de nombreux tests de normalité (test de Pearson construit avec une approche de discrétisation et un test du χ^2, test de Shapiro-Wilk, ...).

V.4 Problèmes à deux échantillons

V.4.1 Que signifie échantillon apparié, non apparié ?

La distinction entre échantillon apparié et échantillon non-apparié a déjà été faite dans IV.4 ; reprenons la avec un autre exemple. Deux correcteurs doivent corriger 400 copies et on souhaite savoir s'ils ont ou non le même comportement (on dira aussi : "le même barème") ; deux modes de répartition des copies entre les correcteurs sont considérés.

Dans un premier cas, chacune des 400 copies est corrigée par les deux correcteurs, "en aveugle", c'est-à-dire que chaque correcteur ignore la note mise par son collègue. On enregistre ensuite pour chaque copie ses deux notes, ce qui justifie la terminologie : les échantillons de notes sont **appariés** ; en d'autres termes, il y a deux facteurs identifiés dans l'élaboration de

la note : la copie et le correcteur. Pour donner un sens à la problématique qui nous intéresse ici (comparaison des deux correcteurs), il faut donc modéliser la situation en introduisant pour chaque copie une notion de "valeur intrinsèque", de sorte que ce sont les variations de chaque correcteur autour de cette valeur qui caractérisent ces correcteurs et dont il faut comparer les lois (et non les lois des notes elles-mêmes, qui varient de copie à copie pour chaque correcteur). Dans cette situation, la comparaison des barèmes peut se faire par un test de Wilcoxon.

Dans un second cas, les deux correcteurs se partagent le paquet et chaque copie n'est corrigée qu'une fois ; les deux paquets se composent respectivement de m et $400 - m$ copies. Le seul facteur identifiable est donc le correcteur et il n'y a plus de notion d'appariement qui ait un sens : on dit donc que les échantillons de notes sont **non appariés**. On suppose implicitement que, globalement, ces deux paquets sont "de même valeur" (un exemple célèbre de non respect de cette clause s'est présenté en France lors d'un examen national d'allemand où la répartition des copies était faite par ordre alphabétique ; le rapport sur cet examen s'étonnait ensuite que les zones de la fin de l'alphabet, plus fréquentes comme première lettre du nom de famille dans les régions frontalières de l'Allemagne que dans le reste de la France, recueillent significativement de meilleurs notes !). Ici, c'est sur l'ensemble des notes de l'un et l'autre paquets que l'on fondera la comparaison des deux correcteurs. Dans cette situation, la comparaison des barèmes peut par exemple se faire par un test de Mann et Whitney.

V.4.2 Deux échantillons appariés : le test de Wilcoxon

Principe

Dans l'exemple V.3 de l'étude d'une drogue sur la pression artérielle, on désire savoir si l'administration de cette drogue diminue la pression. On peut reformuler cette question de la façon suivante : la loi de la pression artérielle des patients a-t-elle été décalée vers les valeurs inférieures après l'administration de la drogue ?

Pour y répondre, on se place dans le modèle non paramétrique de décalage suivant : on suppose que les variables $(X_1, \dots, X_n)$ (pressions avant administration de la drogue) sont i.i.d. de fonction de répartition F continue et indépendantes des variables $(Y_1, \dots, Y_n)$ (pressions après administration) i.i.d. de fonction de répartition $F_\mu(t) = F(t - \mu)$ où $\mu \in \mathbb{R}$. Dans ce modèle, les variables Y_i ont même loi que les variables $X_i + \mu$. En effet,

$$\mathbb{P}(Y_1 \leq t) = F(t - \mu) = \mathbb{P}(X_1 \leq t - \mu) = \mathbb{P}(X_1 + \mu \leq t).$$

On souhaite tester $H_0 = \{\mu = 0\}$ contre $H_1 = \{\mu < 0\}$ (cela revient au même de choisir $H_0 = \{\mu \geq 0\}$ contre $H_1 = \{\mu < 0\}$).

Pour répondre à cette question, on présente le test de Wilcoxon[3] construit à partir de la statistique de Wilcoxon. On ordonne les variables $Z_i = X_i - Y_i$ suivant l'ordre croissant des valeurs absolues pour obtenir la suite $Z_{(1)}, \dots, Z_{(n)}$ avec $|Z_{(1)}| \leq |Z_{(2)}| \leq \dots \leq |Z_{(n)}|$. On calcule ensuite

$$T^+ = \sum_{k=1}^{n} k \mathbf{1}_{\{Z_{(k)} > 0\}}.$$

[3]Ce test est également appelé test des *signes et rangs*.

L'expression de T^+ mêle donc les rangs des valeurs absolues des différences $Z_i = X_i - Y_i$ et leur signes. Sous H_0, ces différences ont une loi symétrique autour de 0 : en effet, comme X_i et Y_i sont indépendantes et de même loi, $X_i - Y_i$ a même loi que $Y_i - X_i$. En outre comme F est continue, $\mathbb{P}(Z_i = 0) = 0$.

La proposition suivante (assez intuitive), permet alors de déterminer la loi de T^+ sous H_0 puisqu'elle assure que les variables aléatoires $(\mathbf{1}_{\{Z_{(k)}>0\}},\ 1 \leq k \leq n)$ sont alors i.i.d. suivant la loi de Bernoulli $\mathcal{B}(\frac{1}{2})$:

Proposition V.16. *Si Z suit une loi symétrique autour de zéro telle que $\mathbb{P}(Z = 0) = 0$, alors sa valeur absolue et son signe sont indépendants.*

Démonstration. La symétrie de la loi de Z, jointe au fait que $\mathbb{P}(Z = 0) = 0$, exprime que si B est une partie borélienne de $\mathbb{R}^+$ et qu'on note $-B$ sa symétrique ($-B = \{x \leq 0 : -x \in B\}$), on a :

$$\mathbb{P}(|Z| \in B) = \mathbb{P}(Z \in -B \cup B) = 2\,\mathbb{P}(Z \in B).$$

Il en résulte pour le choix $B = \mathbb{R}_+$ que :

$$\mathbb{P}(\text{signe}(Z) = +) = \mathbb{P}(\text{signe}(Z) = -) = \frac{1}{2}$$

donc :

$$\mathbb{P}\left(\text{signe}(Z) = +, |Z| \in B\right) = \mathbb{P}(Z \in B) = \frac{1}{2}\mathbb{P}(|Z| \in B) = \mathbb{P}(\text{signe}(Z) = +)\,\mathbb{P}(|Z| \in B).$$

De même, on a $\mathbb{P}\left(\text{signe}(Z) = -, |Z| \in B\right) = \mathbb{P}(\text{signe}(Z) = -)\,\mathbb{P}(|Z| \in B)$. La définition de l'indépendance entre signe(Z) et $|Z|$ est ainsi vérifiée (voir X.1.4, p. 239). □

On déduit de la proposition précédente que, sous H_0, les variables aléatoires $\mathbf{1}_{\{Z_{(k)}>0\}}$ sont i.i.d. de loi de Bernoulli $\mathcal{B}(\frac{1}{2})$. On a ainsi

$$\mathbb{E}[T^+] = \frac{n\,(n+1)}{4}$$

$$\text{Var}\left(T^+\right) = \frac{1}{4}\sum k^2 = \frac{n\,(n+1)\,(2n+1)}{24}.$$

Et même si les différents termes de T^+ n'ont pas même variance à cause du coefficient k, on peut néanmoins montrer la normalité asymptotique de T^+. On considère la statistique de test

$$\zeta_n = \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}.$$

Proposition V.17. *Sous H_0, la suite $(\zeta_n, n \geq 1)$ converge en loi vers la loi normale centrée réduite $\mathcal{N}(0,1)$.*

La preuve de cette proposition est donnée en annexe en fin de chapitre.

On étudie maintenant le comportement de la statistique de test sous H_1. Comme $|Z_{(1)}| \leq |Z_{(2)}| \leq \ldots \leq |Z_{(n)}|$, pour $1 \leq j \leq k \leq n$, $Z_{(k)} + Z_{(j)}$ est positif si et seulement si $Z_{(k)}$ l'est. Donc $k\mathbf{1}_{\{Z_{(k)}>0\}} = \sum_{j=1}^{k} \mathbf{1}_{\{Z_{(j)}+Z_{(k)}>0\}}$ et

$$T^+ = \sum_{1\leq j\leq k\leq n} \mathbf{1}_{\{Z_{(j)}+Z_{(k)}>0\}}.$$

Mais les doubletons $\{Z_{(j)}, Z_{(k)}\}$ avec $j \leq k$ sont en bijection avec les doubletons $\{Z_i, Z_j\}$ avec $i \leq j$. D'où

$$T^+ = \sum_{1\leq i\leq j\leq n} \mathbf{1}_{\{Z_i+Z_j>0\}} = \sum_{1\leq i\leq j\leq n} \mathbf{1}_{\{X_i-(Y_i-\mu)+X_j-(Y_j-\mu)>2\mu\}},$$

où les variables aléatoires $Y_i - \mu$ ont même loi que les X_i. En utilisant cette expression, on peut démontrer que lorsque n tend vers l'infini, ζ_n tend p.s. vers $-\infty$ ou $+\infty$ suivant que $\mu > 0$ ou $\mu < 0$.

Ainsi pour l'exemple de l'étude de l'effet d'une drogue ($H_1 = \{\mu < 0\}$), la zone critique est de la forme $[a, +\infty[$.

Remarque V.18. Si on choisit comme hypothèse alternative :
- $H_1 = \{\mu > 0\}$, la zone critique est de la forme $]-\infty, -a]$,
- $H_1 = \{\mu \neq 0\}$ alors la zone critique est de la forme $]-\infty, a] \cup [a, +\infty[$.

$\Diamond$

Pour les petits échantillons ($n < 20$), on consultera les tables pour trouver les valeurs seuils de T^+ ou on écrira quelques lignes de programmation afin de simuler un grand nombre de réalisations de T^+ sous l'hypothèse nulle. Pour les grands échantillons ($n \geq 20$), on utilisera l'approximation gaussienne. Ce résultat reste approximativement valable quand les données comportent des ex-aequo. Voyons plus précisément comment traiter ces cas en pratique.

Traitement des ex-aequo

C'est un des problèmes permanents de la statistique non paramétrique. En théorie, on sait pourtant que si on travaille avec des lois F continues, il ne devrait pas apparaître d'ex-aequo (probabilité d'occurrence nulle !). En pratique, on en trouve souvent et surtout quand on travaille sur des notes. Par exemple, si un ou plusieurs examinateurs notent 200 individus de 1 à 20, on est assuré de trouver des ex-aequo en grand nombre. Ce problème est donc incontournable en pratique. Nous donnons ci-dessous les trois principales réponses possibles et qui s'appliquent à l'ensemble des tests de ce chapitre.

- **Randomisation** : on départage tous les ex-aequo par un tirage au sort auxiliaire : on jette une pièce en l'air... Cette méthode est la plus séduisante sur le plan théorique, cependant elle a l'inconvénient d'introduire un hasard exogène qui peut "brouiller" les résultats pour les petits échantillons.
- **Suppression** : dans un test de signe, si on a deux données appariées égales, on supprime la paire correspondante.

– **Rang moyen** : dans les tests de rangs, quand plusieurs valeurs sont égales, on leur donne la moyenne des rangs qu'elles auraient si elles étaient différentes. **C'est la méthode la plus employée dans les tests de rangs, et c'est celle que nous conseillons**, et que nous utiliserons dans les exemples à venir. Elle n'est pas parfaitement rigoureuse sur le plan théorique, mais pour la plupart des tests, il existe des corrections pour tenir compte des égalisations.

Exemple V.19. Revenons sur l'exemple V.3 concernant l'étude de l'effet d'une certaine drogue.

$(X-Y)$	-4	+4	+21	+3	-3	+31	+17	+2
rang($\lvert X-Y\rvert$)	4.5	4.5	7	2.5	2.5	8	6	1
rang des dif. ≥ 0		4.5	7	2.5		8	6	1

On en déduit $t^+ = 29$. Des simulations ou une table de Wilcoxon donne la p-valeur sous H_0 : $\mathbb{P}(T^+ \geq 29) = 0.074$. Au niveau de signification de 0.05 (comme pour tout niveau inférieur à 0.074), $t^+ = 29$ tombe dans la région d'acceptation de H_0 ; la drogue n'a pas d'effet significatif sur la tension.

$\Diamond$

V.4.3 Deux échantillons non appariés

On dispose de deux échantillons $X_1, ..., X_m$ et $Y_1, ..., Y_n$, issus de variables aléatoires toutes indépendantes, de fonctions de répartition continues, notées respectivement F pour le premier, et G pour le second. On veut tester l'hypothèse H_0 : "$F = G$" Nous allons pour cela décrire deux tests. Dans le test de Kolmogorov-Smirnov, l'alternative est "$F \neq G$". Pour le test de Mann et Whitney, l'alternative, un peu plus restrictive, sera précisée ultérieurement.

Test de Kolmogorov-Smirnov pour deux échantillons

On décrit d'abord le test de Kolmogorov-Smirnov qui généralise le test de Kolmogorov (voir ci-dessus V.3.1) au cas de deux échantillons. Ce test très utilisé vise donc à répondre à la question "ces deux échantillons sont ils issus d'une même loi ?".

Comme dans le cas à un seul échantillon, ce test de Kolmogorov-Smirnov est basé sur la distance entre deux fonctions de répartition ; il s'agit ici des fonctions de répartition empirique $\hat{F}_{\underline{X}}$ associée à $\underline{X}$ et $\hat{G}_{\underline{Y}}$ associée à $\underline{Y}$:

$$D_{\underline{X},\underline{Y}} = \sup_{t\in\mathbb{R}} \left| \hat{F}_{\underline{X}}(t) - \hat{G}_{\underline{Y}}(t) \right|$$

De même, la zone de rejet de niveau α est prise de la forme $\{D_{\underline{X},\underline{Y}} > a\}$, où a est encore indépendant de F sous H_0 grâce au résultat suivant.

Proposition V.20. *Sous H_0, la statistique $D_{\underline{X},\underline{Y}}$ est libre. On notera sa loi $\mathcal{D}(m,n)$.*

Démonstration. D'après la proposition V.6,

$$D_{\underline{X},\underline{Y}} = \sup_{t\in\mathbb{R}} \left| \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{\{U_i \leq F(t)\}} - \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{\{V_j \leq F(t)\}} \right|$$

où $U_1 = F(X_1), ..., U_m = F(X_m)$ et $V_1 = F(Y_1), ..., V_n = F(Y_n)$ sont deux échantillons indépendants de la loi uniforme sur $[0, 1]$. On conclut en effectuant le changement de variable $u = F(t)$. □

Remarque V.21. En pratique, pour calculer $D_{\underline{x},\underline{y}}$ on trie globalement les x_i et les y_j par ordre croissant pour obtenir $\underline{z} = (z_1, \ldots, z_{n+m})$. En notant $s_k = 1/m$ si z_k provient de $\underline{x}$ et $s_k = -1/n$ si z_k provient de $\underline{y}$, on a

$$D_{\underline{x},\underline{y}} = \max_{1 \le k \le n+m} \left| \sum_{l=1}^{k} s_l \right|.$$

◊

La loi de Kolmogorov-Smirnov $\mathcal{D}(m, n)$ est tabulée pour des petites valeurs de m et n, elle est aussi accessible facilement par simulation. Pour le cas $m = n$, et si $D_{n,n}$ désigne une v.a. de loi $\mathcal{D}(n, n)$, on dispose aussi de la formule (pour tout entier strictement positif k) :

$$\mathbb{P}(nD_{n,n} > k) = 2 \sum_{j=1}^{[n/k]} (-1)^{j+1} \frac{(n!)^2}{(n-jk)!(n+jk)!}.$$

Pour les échantillons de grande taille, lorsque m et n tendent vers l'infini, on a le comportement asymptotique suivant.

Proposition V.22. *Si on pose* $\zeta_{m,n} = \sqrt{\frac{mn}{m+n}} D_{\underline{X},\underline{Y}}$, *lorsque* $\min(m, n) \to \infty$, *on a*

- *sous* H_0, $\forall y > 0$, $\mathbb{P}(\zeta_{m,n} \le y) \to \sum_{-\infty}^{+\infty} (-1)^k \exp(-2k^2y^2)$,
- *sous* H_1, *p.s.* $\zeta_{m,n}$ *tend vers* $+\infty$.

L'expression asymptotique de la fonction de répartition sous H_0 ne peut être utilisée en pratique que pour des valeurs assez élevées de m et n (de l'ordre de 40 au moins) car la convergence annoncée est lente.

Ce test de Kolmogorov-Smirnov est intéressant dans le cas très général où l'on ne connaît aucun lien entre les deux échantillons. On propose maintenant un autre test, très simple à mettre en œuvre.

Test de Mann et Whitney

Le test de Mann et Whitney est basé sur l'enchevêtrement des observations des deux échantillons. On note $\mathbf{1}_{\{Y_j > X_i\}}$ la v.a. qui vaut 1 si $Y_j > X_i$ et 0 sinon. La statistique $U_{\underline{Y},\underline{X}}$ du test s'écrit comme la somme des $\mathbf{1}_{\{Y_j > X_i\}}$.

$$U_{\underline{Y},\underline{X}} = \sum_{\substack{i=1..m \\ j=1..n}} \mathbf{1}_{\{Y_j > X_i\}}$$

Dans ces conditions, on a :

Proposition V.23. *Sous H_0, la statistique $U_{\underline{Y},\underline{X}}$ est libre. On notera sa loi $\mathcal{U}(m,n)$. En outre, $\mathbb{P}(Y_j > X_i) = \frac{1}{2}$.*

Démonstration. Si H_0 est vraie, on dispose d'un $(m+n)$-échantillon de la loi de fonction de répartition F : $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$. D'après la proposition V.6, $(F(X_1), \ldots, F(X_m), F(Y_1), \ldots, F(Y_n))$ est un $(m+n)$-échantillon de la loi uniforme et p.s.,

$$\sum_{\substack{i=1..m \\ j=1..n}} \mathbf{1}_{\{Y_j > X_i\}} = \sum_{\substack{i=1..m \\ j=1..n}} \mathbf{1}_{\{F(Y_j) > F(X_i)\}},$$

ce qui assure que la statitisque $U_{\underline{Y},\underline{X}}$ est libre. Comme $\mathbb{P}(Y_j > X_i) = \mathbb{P}(F(Y_j) > F(X_i))$ et que $F(Y_j)$ et $F(X_i)$ sont des v.a. indépendantes de loi uniforme sur $[0,1]$, on en déduit que

$$\mathbb{P}(Y_j > X_i) = \mathbb{P}(F(Y_j) > F(X_i)) = \int_{[0,1]^2} \mathbf{1}_{\{u>v\}}\, dudv = \frac{1}{2}.$$

□

Si $\mathbb{P}(Y_j > X_i) < \frac{1}{2}$, $U_{\underline{Y},\underline{X}}$ va avoir tendance à être plus petit que sous H_0 tandis que si $\mathbb{P}(Y_j > X_i) > \frac{1}{2}$, $U_{\underline{Y},\underline{X}}$ va avoir tendance à être plus grand que sous H_0. Cela se voit au niveau de l'espérance car $\mathbb{E}[U_{\underline{Y},\underline{X}}] = mn\mathbb{P}(Y_j > X_i)$.

On choisit comme hypothèse alternative $H_1 = \{\mathbb{P}(Y_j > X_i) \neq \frac{1}{2}\}$. La région de rejet de $H_0 = \{F = G\}$ contre H_1 au niveau α est de la forme :

$$\text{“}U_{\underline{Y},\underline{X}} \text{ n'appartient pas à } [u^-_{m,n,\alpha/2}, u^+_{m,n,\alpha/2}]\text{”}.$$

Les valeurs de $u^-_{m,n,\alpha/2}$ et $u^+_{m,n,\alpha/2}$ (quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi $\mathcal{U}(m,n)$) ont été calculées et tabulées pour les petites valeurs à partir de la loi uniforme.

L'espérance mathématique et la variance d'une v.a. $U_{m,n}$ de loi $\mathcal{U}(m,n)$ valent respectivement $\mathbb{E}[U_{m,n}] = \frac{mn}{2}$ et $\mathrm{Var}(U_{m,n}) = \frac{mn(m+n+1)}{12}$. Pour m et n supérieurs à 10, l'approximation de la loi $\mathcal{U}(m,n)$ par la loi normale d'espérance mathématique $\dfrac{mn}{2}$ et variance $\dfrac{mn(m+n+1)}{12}$ est justifiée par la proposition suivante :

Proposition V.24. *Lorsque* $\min(m,n)$ *tend vers l'infini, les v.a. définies par*

$$\zeta_{m,n} = \frac{U_{m,n} - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}},$$

où $U_{m,n}$ est une v.a. de loi $\mathcal{U}(m,n)$, convergent en loi vers la loi normale centrée réduite $\mathcal{N}(0,1)$.

On peut montrer que sous $H_1 = \{\mathbb{P}(Y_j > X_i) \neq \frac{1}{2}\}$, la statistique de test $\zeta_{m,n}$ tend p.s. vers $+\infty$ ou $-\infty$ lorsque m et n tendent vers l'infini de telle sorte que le rapport $\frac{m}{n}$ ait une limite dans $]0,\infty[$. La région critique est donc de la forme $]-\infty,-a] \cup [a,+\infty[$, et le test de Mann et Whithney est convergent. Pour un test de niveau asymptotique α, on choisit $a = \phi_{1-\alpha/2}$, le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$.

Remarque V.25. On peut expliciter des sous ensembles remarquables de l'hypothèse alternative du test de Mann et Whitney.

- Soit W une v.a. de fonction de répartion F_W et Z une v.a. de fonction répartition F_Z. On dit que la loi de W majore strictement pour l'ordre stochastique la loi Z, si $F_W(t) \leq F_Z(t)$ pour tout $t \in \mathbb{R}$ avec inégalité stricte pour au moins une valeur de t. On le note alors $Z \prec W$. Si de plus Z et W sont indépendantes, on a $\mathbb{P}(W > Z) > \frac{1}{2}$. Pour le test de Mann et Whitney, on peut donc considérer comme hypothèse alternative $H_1 = \{X_i \prec Y_j\} \cup \{Y_j \prec X_i\}$ sans changer la région critique.
- On peut se placer dans un modèle de décalage en supposant que la fonction de répartition commune des Y_j est de la forme $F_\mu(t) = F(t - \mu)$ où $\mu \in \mathbb{R}$, ce qui signifie que Y_j a même loi que $X_i + \mu$. Alors pour $\mu > 0$, $X_i \prec Y_j$ tandis que pour $\mu < 0$, $Y_j \prec X_i$. On peut alors tester l'hypothèse $H_0 = \{\mu = 0\}$ contre l'hypothèse $H_1 = \{\mu \neq 0\}$ sans changer la région critique.

Enfin si on prend comme hypothèse alternative $H_1 = \{\mathbb{P}(Y_j > X_i) < \frac{1}{2}\}$ ou $H_1 = \{Y_j \prec X_i\}$ ou $H_1 = \{\mu < 0\}$, la région critique est de la forme $]-\infty, a]$ tandis que si on prend $H_1 = \{\mathbb{P}(Y_j > X_i) > \frac{1}{2}\}$ ou $H_1 = \{X_i \prec Y_j\}$ ou $H_1 = \{\mu > 0\}$, la région critique est de la forme $[a, +\infty[$.

$\Diamond$

Dès que m et n sont grands, le calcul de U sous la forme donnée ci-dessus devient fastidieux. Dans ce cas, on utilise la méthode suivante :

- on classe globalement les X_i et Y_j, d'où une suite $\underline{Z} = (Z_1, \ldots, Z_{n+m})$,
- pour tout j (où $1 \leq j \leq n$) on repère, et on note S_j, le rang de Y_j dans la suite $\underline{Z}$,
- on calcule $R_{\underline{Y}} = \sum_{j=1}^{n} S_j$. On vérifie élémentairement que les v.a. $U_{\underline{Y},\underline{X}}$ et $R_{\underline{Y}}$ sont liées par :

$$U_{\underline{Y},\underline{X}} = R_{\underline{Y}} - \frac{n(n+1)}{2}.$$

Le traitement des ex-aequo se fait par les mêmes méthodes que pour les tests à un échantillon.

Remarque V.26. On peut évidemment échanger le rôles des deux échantillons. On considère alors les statistiques $U_{\underline{X},\underline{Y}}$, calculée à partir du nombre de couples (i, j) tels que $X_i > Y_j$ et $R_{\underline{X}}$, calculée à partir de la somme des rangs des X_i dans la suite $\underline{Z}$; elles vérifient $U_{\underline{X},\underline{Y}} = R_{\underline{X}} - \frac{m(m+1)}{2}$. D'autre part on a : $R_{\underline{X}} + R_{\underline{Y}} = \frac{(n+m)(n+m+1)}{2}$.

$\Diamond$

Exemple V.27. Revenons sur l'exemple V.2 de l'angoisse dans les sociétés primitives. Pour tester l'hypothèse H_0 : "les deux types de société sont différentes" contre sa négation, nous allons utiliser les deux tests précédents. Le tableau des rangs est le suivant :

Sociétés sans expl. orales	Notes d'anxiété	Rangs	Sociétés avec expl. orales	Notes d'anxiété	Rangs
Lapp	13	20.5	Marquesans	17	39
Chamorro	12	24.5	Dobuans	16	38
Samoans	12	24.5	Baiga	15	36
Arapesh	10	16	Kwoma	15	36
Balinese	10	16	Thonga	15	36
Hopi	10	16	Alorese	14	33
Tanala	10	16	Chagga	14	33
Paiute	9	12	Navaho	14	33
Chenchu	8	9.5	Dahomeans	13	29.5
Teton	8	9.5	Lesu	13	29.5
Flathead	7	5	Masai	13	29.5
Papago	7	5	Lepeha	12	24.5
Wenda	7	5	Maori	12	24.5
Warrau	7	5	Pukapukans	12	24.5
Wogeo	7	5	Trobianders	12	24.5
Ontong	6	1.5	Kwakiull	11	20.5
			Manus	11	20.5
			Chiricahua	10	16
			Comanche	10	16
			Siriono	10	16
			Bena	8	9.5
			Slave	8	9.5
			Kurtatchi	6	1.5
		$R_{\underline{x}} = 200$ $m = 16$			$R_{\underline{y}} = 580$ $n = 23$

Application du test de Kolmogorov-Smirnov. On calcule les fonctions de répartition empiriques des deux échantillons, puis on trouve $D_{\underline{x},\underline{y}} = 0.5516$. Or d'après les tables de ce test au niveau 0.01 (ou d'après les résultats fournis par les logiciels), on a : $\mathbb{P}(D_{\underline{x},\underline{y}} > 0.5045) = 0.01$. On peut donc rejeter, au seuil 0.01, l'hypothèse nulle et affirmer la différence entre les deux sociétés.

A-t-on le même résultat en utilisant le test de Mann et Whitney ?

Application du test de Mann et Whitney. On trouve : $U_{\underline{y},\underline{x}} = 580 - 12 \times 23 = 580 - 276 = 304$ de même que $U_{\underline{x},\underline{y}} = \sum$ des rangs des $X - \frac{m(m+1)}{2} = 200 - 8 \times 17 = 64$.

L'approximation gaussienne a pour moyenne $16 \times 23 = 184$ et pour variance 35.02^2. La v.a. centrée et réduite $\dfrac{U_{\underline{y},\underline{x}} - 184}{35.02}$ prend pour valeur 3.43 ce qui est très grand pour une loi normale centrée réduite, dont la fonction de répartition vaut $1 - 3.10^{-4}$ en 3.43. Par conséquent on met en évidence une différence, significative à tout niveau $\alpha \geq 3.10^{-4}$, entre les deux types de sociétés, de la même manière qu'avec le test de Kolmogorov. ◊

V.4.4 Tests paramétriques ou tests non paramétriques ?

Les tests de Kolmogorov-Smirnov ou de Mann et Whitney sont le pendant non paramétrique du test paramétrique de Student. Ces trois tests travaillent tous sur deux échantillons mutuellement indépendants. Dans le cas où toutes les hypothèses du test de Student sont vérifiées (même variance, normalité...) on démontre qu'il est optimal. Dans ce cas précis, l'efficacité du test de Mann et Whitney par rapport au test de Student est de 96% ; c'est à dire qu'il a une précision comparable. Dès que l'on a le moindre doute sur la validité des hypothèses du test de Student, il vaut mieux faire un test non paramétrique.

Un des avantages du test de Student est de fournir un estimateur de la différence entre les deux populations ainsi qu'un intervalle de confiance.

V.5 Conclusions

Ces notes de cours sont loin d'être exhaustives et prétendent simplement montrer quelques méthodes. L'affranchissement des hypothèses contraignantes des tests paramétriques évite beaucoup d'hypothèses que l'on sait rarement justifier mais se paye par l'absence de méthodes générales. Chaque problème donne lieu à une statistique particulière, à des tables différentes. Notons qu'avec les ordinateurs personnels, il est désormais très facile de reconstruire les tables en approchant **par simulation** les distributions des statistiques de test sous l'hypothèse nulle et d'en évaluer les valeurs critiques selon le niveau désiré.

Pour donner des recommandations générales d'emploi de la statistique non paramétrique, on peut dire que, s'il existe un test non paramétrique qui répond au problème que l'on se pose et que l'on connaît ce test, ce n'est pas une mauvaise stratégie que de l'utiliser. Cependant, ce serait une erreur d'éliminer sciemment un facteur ou un régresseur pertinent dans un problème pour rentrer dans le moule non paramétrique.

V.6 Annexe

Preuve de la proposition V.6. On introduit le pseudo-inverse $F^{-1}(u) = \inf\{x; F(x) \geq u\}$. Avec cette définition, on vérifie facilement que pour F continue, on a

$$\forall (x,u) \in \mathbb{R}\times]0,1[,\ F(x) < u \Leftrightarrow x < F^{-1}(u) \ \text{ et } \ F(F^{-1}(u)) = u. \tag{V.1}$$

Soit $u \in]0,1[$. L'ensemble $\{t \in \mathbb{R} : F(t) = u\}$ est fermé non vide par continuité de F et c'est un intervalle par croissance de F. On le note $[\underline{t}_u, \bar{t}_u]$. Comme pour tout $x \in \mathbb{R}$, $\mathbb{P}(X_i = x) = 0$, on a alors

$$\mathbb{P}(F(X_i) = u) = \mathbb{P}(X_i = \underline{t}_u) + \mathbb{P}(X_i \in]\underline{t}_u, \bar{t}_u]) = 0 + F(\bar{t}_u) - F(\underline{t}_u) = 0. \tag{V.2}$$

Cette propriété implique que les variables $F(X_i)$ sont presque sûrement distinctes ; leurs rangs sont alors identiques à ceux des X_i. On en déduit également en utilisant les propriétés (V.1) que

$$\begin{aligned}\mathbb{P}(F(X_i) \leq u) &= \mathbb{P}(F(X_i) < u) + \mathbb{P}(F(X_i) = u) = \mathbb{P}(X_i < F^{-1}(u)) + 0 \\ &= \mathbb{P}(X_i \leq F^{-1}(u)) - \mathbb{P}(X_i = F^{-1}(u)) = F(F^{-1}(u)) - 0 = u.\end{aligned}$$

Ainsi les variables $F(X_i)$ sont uniformément réparties sur $[0,1]$.
On note $\mathcal{U} = \{u \in]0,1[: \bar{t}_u - \underline{t}_u > 0\}$, ensemble qui est au plus dénombrable puisque pour $n \in \mathbb{N}^*$, $\{u \in [F(-n), F(n)] : \bar{t}_u - \underline{t}_u > 1/n\}$ comporte au plus $2n^2+2$ éléments. On a donc en utilisant (V.2),

$$\mathbb{P}\left(X_i \in \bigcup_{u\in\mathcal{U}} [\underline{t}_u, \bar{t}_u]\right) = \sum_{u\in\mathcal{U}} \mathbb{P}(F(X_i) = u) = 0.$$

Comme pour x en dehors de l'union $\bigcup_{u\in\mathcal{U}}[\underline{t}_u, \bar{t}_u]$ des intervalles de constance de F, $\forall t \in \mathbb{R}$, $x \le t \Leftrightarrow F(x) \le F(t)$, on conclut que p.s. $\forall t \in \mathbb{R}$, $\sum_{i=1}^n \mathbf{1}_{\{X_i \le t\}} = \sum_{i=1}^n \mathbf{1}_{\{F(X_i)\le F(t)\}}$. □

Preuve de la proposition V.17. Il suffit d'après le paragraphe X.3.1 de vérifier la convergence de la fonction caractéristique $\Phi_{\zeta_n}(u) = \mathbb{E}[e^{iu\zeta_n}]$ vers la fonction caractéristique de la loi $\mathcal{N}(0,1)$: $\Phi_{\mathcal{N}(0,1)}(u) = e^{-u^2/2}$, où

$$\zeta_n = \frac{1}{2\sqrt{v_n}} \sum_{k=1}^n k(\mathbf{1}_{\{Z_{(k)}>0\}} - \mathbf{1}_{\{Z_{(k)}\le 0\}}) \quad \text{et} \quad v_n = \frac{n(n+1)(2n+1)}{24}.$$

Comme les variables aléatoires $Z_{(k)}$ sont i.i.d. de loi $\mathcal{B}(1/2)$, on a

$$\Phi_{\zeta_n}(u) = \prod_{k=1}^n \mathbb{E}\left[e^{i\frac{uk}{2\sqrt{v_n}}(\mathbf{1}_{\{Z_{(k)}>0\}} - \mathbf{1}_{\{Z_{(k)}\le 0\}})}\right] = \prod_{k=1}^n \cos\left(\frac{uk}{2\sqrt{v_n}}\right).$$

Comme en 0, $\ln(\cos(x)) + \frac{x^2}{2} = O(x^4)$, il existe une constante $C > 0$ telle que

$$\left|\ln(\Phi_{\zeta_n}(u)) + \sum_{k=1}^n \frac{u^2k^2}{8v_n}\right| \le C \sum_{k=1}^n \frac{u^4k^4}{16v_n^2}.$$

En remarquant que $\frac{1}{8v_n}\sum_{k=1}^n k^2 = \frac{1}{2}$ et que $\frac{1}{v_n^2}\sum_{k=1}^n k^4 \le \frac{n^5}{v_n^2}$ tend vers 0 avec n, on conclut que $\ln(\Phi_{\zeta_n}(u))$ converge vers $-u^2/2$. □

V.7 Résumé

V.7.1 Test de Kolmogorov

1. Modèle non paramétrique : $\underline{X} = (X_i, 1 \le i \le n)$ i.i.d. de fonction de répartition F continue.
2. Hypothèses : $H_0 = \{F = F_0\}$ et $H_1 = \{F \ne F_0\}$
3. Statistique de Kolmogorov

$$D_{\underline{X}} = \max_{1\le i\le n} \max\left(|F_0(X_{(i)}) - \frac{i-1}{n}|, |F_0(X_{(i)}) - \frac{i}{n}|\right)$$

où $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ est le réordonnement croissant des X_i.
Statistique de test : $\zeta_n = \sqrt{n} D_{\underline{X}}$.
4. Sous H_0, lorsque n tend vers l'infini, ζ_n converge en loi vers la loi de fonction de répartition $\mathbf{1}_{\{y>0\}} \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2y^2)$.

5. Sous H_1, ζ_n tend p.s. vers $+\infty$.
6. Région critique : $[a, +\infty[$, avec $a > 0$.
7. Test convergent pour $n \to +\infty$.
8. Pour un niveau asymptotique α, a est donné par $\sum_{k=-\infty}^{+\infty}(-1)^k \exp(-2k^2a^2) = 1 - \alpha$.
9. La p-valeur asymptotique du test est donnée par $1 - \sum_{k=-\infty}^{+\infty}(-1)^k \exp\left(-2k^2{\zeta_n^{\text{obs}}}^2\right)$.

V.7.2 Test de Wilcoxon

1. Modèle non paramétrique de décalage : $\underline{X} = (X_i,\ 1 \leq i \leq n)$ i.i.d. de fonction de répartition $F(t)$ continue indépendants de $\underline{Y} = (Y_i,\ 1 \leq i \leq n)$ i.i.d. de fonction de répartition $F_\mu(t) = F(t - \mu)$ où $\mu \in \mathbb{R}$ (Y_i a même loi que $X_i + \mu$).
2. Hypothèses : $H_0 = \{\mu = 0\}$ et $H_1 = \{\mu \neq 0\}$.
3. Statistique de Wilcoxon : On classe les variables $Z_i = X_i - Y_i$ suivant l'ordre croissant des valeurs absolues et on note $Z_{(1)}, \ldots, Z_{(n)}$ les variables ainsi obtenues ($|Z_{(1)}| \leq |Z_{(2)}| \leq \ldots \leq |Z_{(n)}|$) ; la statistique de Wilcoxon est $T^+ = \sum_{k=1}^{n} k\mathbf{1}_{\{Z_{(k)}>0\}}$.

   Statistique de test :
   $$\zeta_n = \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}.$$
4. Sous H_0, lorsque n tend vers l'infini, ζ_n converge en loi vers la loi $\mathcal{N}(0,1)$.
5. Sous H_1, ζ_n tend p.s. vers $-\infty$ si $\mu > 0$ ou vers $+\infty$ si $\mu < 0$.
6. Région critique : $]-\infty, -a] \cup [a, +\infty[$, avec $a > 0$.
7. Test convergent pour $n \to +\infty$.
8. Pour un niveau asymptotique α (on recommande $n > 20$), a est donné par le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite.
9. La p-valeur asymptotique du test est donnée par $\mathbb{P}\left(|G| \geq |\zeta_n^{\text{obs}}|\right)$ où G de loi $\mathcal{N}(0,1)$.
10. Variantes : H_0 inchangé,
    - $H_1 = \{\mu > 0\}$, région critique $]-\infty, -a]$ où a est le quantile d'ordre $1 - \alpha$ de la loi normale centrée réduite,
    - $H_1 = \{\mu < 0\}$, région critique $[a, +\infty[$ où a est le quantile d'ordre $1 - \alpha$ de la loi normale centrée réduite.

V.7.3 Test de Kolmogorov-Smirnov

1. Modèle non paramétrique : $\underline{X} = (X_i,\ 1 \leq i \leq m)$ i.i.d. de fonction de répartition F continue indépendants de $\underline{Y} = (Y_j,\ 1 \leq j \leq n)$ i.i.d. de fonction de répartition G continue
2. Hypothèses : $H_0 = \{G = F\}$ et $H_1 = \{G \neq F\}$.

3. Statistique de Kolmogorov-Smirnov :

$$D_{\underline{X},\underline{Y}} = \max_{1\leq k\leq n+m} \left| \sum_{l=1}^{k} S_l \right|$$

où pour $1 \leq l \leq m+n$, $S_l = 1/m$ si dans le réordonnement croissant des X_i et des Y_j, le l-ième élément provient de $\underline{X}$ et $S_l = -1/n$ sinon.
Statistique de test :

$$\zeta_{m,n} = \sqrt{\frac{mn}{m+n}} D_{\underline{X},\underline{Y}}.$$

4. Sous H_0, lorsque $\min(m,n)$ tend vers l'infini, $\zeta_{m,n}$ converge en loi vers la loi de fonction de répartition $\mathbf{1}_{\{y>0\}} \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2y^2)$.
5. Sous H_1, $\zeta_{m,n}$ tend p.s. vers $+\infty$.
6. Région critique : $[a, +\infty[$, avec $a > 0$.
7. Test convergent pour $\min(m,n) \to +\infty$.
8. Pour un niveau asymptotique α, a est donné par $\sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2a^2) = 1-\alpha$.
9. La p-valeur asymptotique du test est donnée par $1 - \sum_{k=-\infty}^{+\infty} (-1)^k \exp\left(-2k^2 {\zeta_{m,n}^{\text{obs}}}^2\right)$.

V.7.4 Test de Mann et Whitney

1. Modèle non paramétrique : $\underline{X} = (X_i,\ 1 \leq i \leq m)$ i.i.d. de fonction de répartition F continue indépendants de $\underline{Y} = (Y_j,\ 1 \leq j \leq n)$ i.i.d. de fonction de répartition G continue.
2. Hypothèses : $H_0 = \{F = G\}$ et $H_1 = \{\mathbb{P}(Y_j > X_i) \neq \frac{1}{2}\}$.
3. Statistique de Mann et Whitney :

$$U_{\underline{Y},\underline{X}} = \sum_{j=1}^{n} S_j - \frac{n(n+1)}{2},$$

où S_j désigne le rang de Y_j dans le classement des X_i et Y_k par ordre croissant.
Statistique de test :

$$\zeta_{m,n} = \frac{U_{\underline{Y},\underline{X}} - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}}.$$

4. Sous H_0, lorsque $\min(m,n)$ tend vers l'infini, $\zeta_{m,n}$ converge en loi vers la loi $\mathcal{N}(0,1)$.
5. Sous H_1, $\zeta_{m,n}$ tend p.s. vers $-\infty$ si $\mathbb{P}(Y_j > X_i) < \frac{1}{2}$, ou vers $+\infty$ si $\mathbb{P}(Y_j > X_i) > \frac{1}{2}$.
6. Région critique : $]-\infty, -a] \cup [a, +\infty[$, avec $a > 0$.
7. Test convergent pour $m, n \to +\infty$ avec $\frac{m}{n}$ admettant une limite dans $]0, +\infty[$.
8. Pour un niveau asymptotique α, a est donné par le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$.
9. La p-valeur asymptotique du test est donnée par $\mathbb{P}\left(|G| \geq |\zeta_{m,n}^{\text{obs}}|\right)$ où G de loi $\mathcal{N}(0,1)$.
10. Variantes : H_0 inchangé

- $H_1 = \{\mathbb{P}(Y_1 > X_1) < \frac{1}{2}\}$, région critique $]-\infty, -a]$ où a est le quantile d'ordre $1-\alpha$ de la loi $\mathcal{N}(0,1)$.
- $H_1 = \{\mathbb{P}(Y_1 > X_1) > \frac{1}{2}\}$, région critique $[a, +\infty[$ où a est le quantile d'ordre $1-\alpha$ de la loi $\mathcal{N}(0,1)$.

Chapitre VI

Modélisation statistique des valeurs extrêmes

VI.1 Introduction

Comme le relate le journal La Petite Gironde du 7 mars 1930, les débits d'une rivière comme la Garonne (voir aussi figure VI.1) sont très variables : *A Cadillac, la plus grosse*

FIG. VI.1 – La Garonne en colère

inondation du siècle : Jeudi matin, les eaux ont atteint 11 m . 77, à 10 heures, soit 30

centimètres de plus qu'en 1875. Le sol de la halle est recouvert par les eaux. La place du Châteaux, sur laquelle notre dépositaire avait installé un magasin de vente en plein vent ressemble à s'y méprendre au Marché neuf ou à la place Mériadeck, tant les inondés ont transporté de meubles, ustensiles et objets de toutes sortes. Le courant est d'une telle violence que, seule, la vedette des Ponts et Chaussées traverse le fleuve dont la berge gauche se situe maintenant à dix-sept kilomètres à l'ouest. La route de Saint Macaire est submergée au point que les eaux ont pénétré dans la chapelle des aliénés. Dans cet établissement aux quartiers très élevés, tous les pensionnaires sont en lieu sûr et largement ravitaillés. La décrue commencée à midi aujourd'hui semble devoir être lente . (...)

Faut- il construire une digue de protection contre les crues ou considérer que le risque de débordement est acceptable ? Si l'ingénieur décide de se protéger par un ouvrage de génie civil, jusqu'à quelle hauteur h construire la digue ? En termes d'ingénierie hydraulique, h est souvent dénommé *crue de projet* (on conviendra que $h = 0$ représente l'alternative *ne pas protéger*) et sera exprimé dans la même unité que les débits de la rivière (puisqu'en une section donnée de la rivière, il y a une correspondance biunivoque entre la hauteur et le débit).

Décider de construire ou de ne pas construire un ouvrage destiné à protéger un site contre le débordement d'une rivière repose sur la connaissance de la probabilité d'apparition d'une crue dommageable. Pour l'évaluer, on dispose généralement de quelques enregistrements des débits passés de la rivière. Dans le tableau VI.1 par exemple, on a enregistré le maximum de débit journalier qui s'est écoulé au cours de chacune des années de 1913 à 1977. Les données x sont formées de la collection (jour de la mesure j, x_j= débit enregistré en m^3/s).

année	Max	année	Max	année	Max	année	Max	année	Max
1913	4579	1926	3200	1939	2800	1952	6721	1965	4968
1914	4774	1927	6332	1940	5553	1953	2700	1966	5163
1915	4968	1928	4968	1941	5163	1954	3000	1967	2600
1916	4774	1929	1950	1942	3100	1955	5747	1968	2530
1917	3400	1930	7500	1943	3600	1956	2300	1969	4073
1918	6137	1931	3700	1944	4579	1957	3200	1970	3120
1919	4189	1932	3600	1945	3200	1958	2900	1971	4696
1920	4579	1933	2500	1946	950	1959	4968	1972	5377
1921	2800	1934	3700	1947	1850	1960	3400	1973	3956
1922	4384	1935	6137	1948	2000	1961	4774	1974	4228
1923	5747	1936	4189	1949	1900	1962	2300	1975	3200
1924	3200	1937	5747	1950	2600	1963	2700	1976	4209
1925	3100	1938	3200	1951	2900	1964	3300	1977	4482

TAB. VI.1 – Débits annuels maximaux (en m^3/s) de la Garonne à Mas d'Agenais sur la période 1913-1977

Mais d'autres types de données auraient pu être collectées. La figure VI.2 présente ainsi la série des débits de crue de la Garonne à Mas d'Agenais qui ont dépassé le seuil de $2500m^3/s$ sur la période 1913-1977.

Avec cette autre campagne de collecte d'informations, plusieurs débits de pointe peuvent apparaître une même année, tandis que certaines années ne feront pas partie des valeurs

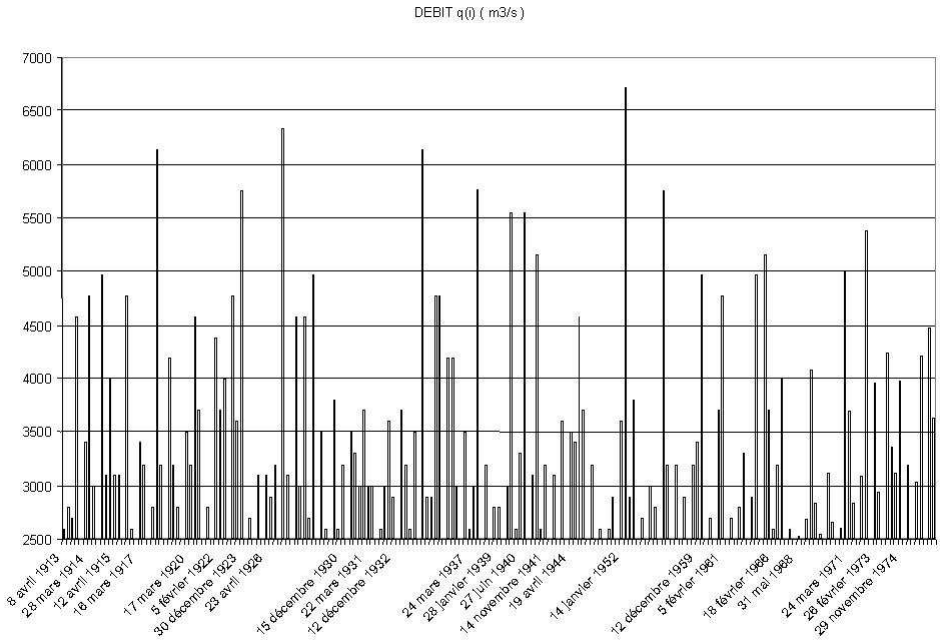


FIG. VI.2 – 151 dépassement des pointes de crue au delà de $2500m^3/s$ de la Garonne à Mas d'Agenais durant la période 1913-1977

extrêmes.

Au vu de ces données, l'ingénieur doit prendre la décision h. Imaginons qu'il y ait n mesures de débit x_j dépassant $2500m3/s$. Une recommandation courante de l'ingénierie est la décision :

$$h = k \times \max_{i=1...n}(x_1, ..., x_i, ...x_n)$$

où k est un *coefficient de sécurité* valant par exemple 3. Est-ce que cette stratégie définit toujours un pari intéressant ? Que se passe-t-il quand l'échantillon de données s'accroît avec le temps ? Que faire lorsque l'on ne dispose pas d'une série de débits, au site où l'on envisage de construire l'ouvrage (site non jaugé) ? Peut-on justifier la décision de façon plus formelle, et montrer par exemple que cette façon de procéder réalise un arbitrage rationnel entre les investissements de génie civil consentis et les dommages évités ? Ici entre en scène le statisticien. S'appuyer sur un modèle probabiliste permet d'analyser le bon fonctionnement et/ou l'intégrité d'une protection menacée par un événement externe dommageable. Cet événement dommageable apparaît comme une variable aléatoire réelle Z (e.g. débit du fleuve) dont la fonction de répartition $G(z|\theta)$ est caractérisée par un vecteur de paramètres inconnus θ. Le plus souvent, les dommages à assumer en cas de défaillance de la protection sont une fonction croissante de Z : le coût des dommages augmente avec les surfaces inondées.

On propose alors de choisir une valeur de projet h grâce à un quantile, z_p, fixant le niveau de protection qui a la probabilité p (faible !) d'être dépassé :

$$p = \mathbb{P}(Z > z_p) = 1 - G(z_p|\theta)$$

La probabilité de défaillance p est fixée par le décideur, d'après une norme nationale ou internationale. Le travail de l'analyste est

- de choisir un modèle $G(z|\theta)$ réaliste pour représenter les valeurs extrêmes, d'en discuter les propriétés et d'en connaître les limites,
- d'estimer la valeur de projet via une inférence statistique sur θ,
- d'étudier la sensibilité du résultat fournit aux hypothèses de modélisation et de fournir une fourchettes d'incertitudes sur la recommandation.

Le but de ce paragraphe est de mettre la théorie des valeurs extrêmes, et plus spécifiquement les modèles GEV *(Generalized Extreme Values)* et POT *(Peak Over Threshold)* à la portée des élèves. Une excellente introduction à la théorie des valeurs extrêmes est disponible dans [3]. Pour des applications en ingénierie hydraulique, le manuel [6], se lit facilement.

VI.2 Modèles de valeurs extrêmes

La première idée d'un modélisateur pour choisir un modèle $G(z|\theta)$ réaliste est de s'appuyer sur une vérité mathématique. Si les hypothèses sur lesquelles il appuie sa réflexion se rapprochent de circonstances idéalisées particulières où a été démontré un théorème du calcul des probabilités, il pourra être tenté d'utiliser comme modèle la structure particulière des lois du hasard issue de ce théorème.

Mais cette idée doit être examinée avec soin. D'abord la construction d'un modèle est toujours en soi une succession de vérités mathématiques à l'intérieur d'un corps d'hypothèses (c'est son support rationnel). C'est dans la transposition concrète des hypothèses que réside l'interprétation et celle-ci ne peut prendre la forme d'une déduction mathématique absolue. Le prototype de théorème justificatif concerne la loi normale. On ne peut mieux faire que rappeler ici la boutade de H. Poincaré : *Tout le monde croit à la loi normale : les physiciens parce qu'ils pensent que les mathématiciens l'ont démontrée et les mathématiciens parce qu'ils croient qu'elle a été vérifiée par les physiciens.*

En fait, l'hypothèse d'effet additif d'un grand nombre de *causes* justifiant mathématiquement la loi normale, n'est qu'une aide à l'interprétation *physique qualitative* de certaines variables particulières. Le modèle résultant n'est pas justifié de façon absolue mais s'il est validé par des données, il peut être privilégié (vis-à-vis d'autres modèles équivalents) pour autant que l'additivité ait réellement un sens.

Les valeurs extrêmes d'un signal quelconque peuvent être définies de deux façons différentes. La première considère le maximum des observations régulièrement espacées sur une période fixe ou bloc (e.g. le maximum annuel des observations journalières). Pourvu que la taille du bloc soit assez grande, les maxima peuvent être considérés comme des tirages indépendants dans une *loi généralisée des valeurs extrêmes* ou modèle GEV. La seconde manière considère que les observations qui excèdent un seuil fixé constituent un *processus ponctuel de Poisson* et, pourvu que ce seuil soit assez haut, les dépassements du seuil fixé ont une *distribution de Pareto généralisée* : c'est le modèle POT. Les deux modèles présentent trois paramètres formant un vecteur tridimensionnel, que nous noterons θ, prenant ses valeurs dans Θ qui représente l'ensemble des états de la nature (i.e. l'ensemble des valeurs possibles de θ). L'avantage du modèle GEV est qu'il donne directement la *valeur de projet,*

i.e. l'événement extrême associé à une période de retour fixée (décennale, centennale, etc.). Or, on peut montrer que cette valeur de projet peut aussi être déterminée à partir d'un modèle POT. L'avantage du modèle POT est qu'en choisissant judicieusement le seuil on peut augmenter les données et donc réduire l'incertitude sur la valeur de projet.

Toutefois, insistons sur le fait que pour l'estimation des quantiles élevés, il n'existe pas de recette miracle, mais plutôt de nombreuses techniques qui ensemble permettent d'en avoir un ordre de grandeur.

VI.2.1 La loi généralisée des extrêmes (la loi du maximum par blocs)

Considérons un ensemble de variables aléatoires indépendantes $X_1, X_2...X_n$ ayant en commun la même fonction de répartition F (hypothèse *iid* pour *indépendantes et identiquement distribuées*) et considérons le maximum $M_n = \max(X_1, X_2...X_n)$. Dans les applications, les X_i sont souvent enregistrés à intervalle de temps régulier : par exemple les pluies moyennes de la journée ou les débits journaliers d'une rivière si bien que M_n correspondra au *record* sur une période de temps n. Ainsi, si n est le nombre de jours d'un mois donné, M_n pourra désigner la pluie de la journée la plus humide du mois en question.

On vérifie facilement que la fonction de répartition de M_n est F^n. Passer à la limite quand n tend vers l'infini n'a pas de sens, car tous les points x du domaine de définition plus petits que le supremum de ce domaine sont tels que $F(x) < 1$, et, formulé ainsi, la limite de $F^n(x)$ est 0 (sauf pour le supremum où elle vaut 1).

Pour éviter cette difficulté, on se donne le droit de renormaliser M_n avec deux suites μ_n (translation) et σ_n (échelle) en $Y_n = \frac{M_n - \mu_n}{\sigma_n}$. La question devient : existe-t-il de telles suites de constantes qui permettent de stabiliser la répartition de Y_n quand n tend vers l'infini vers une fonction de répartition G non dégénérée ? (On dit que la fonction de répartition d'une variable aléatoire X est dégénérée ou triviale si p.s. X est égal à une constante.)

La loi du maximum renormalisé

Si X est une variable aléatoire, et $\alpha > 0$, $\beta \in \mathbb{R}$ deux constantes, la transformation $X - \beta$ est appelé *translation* et $\frac{X}{\alpha}$ *changement d'échelle*. On dit qu'une fonction de répartition est *max-stable* si pour tout $n \geq 1$, la loi de $M_n = \max_{1 \leq i \leq n} X_i$, où les v.a. $(X_i, i \geq 1)$ sont iid de fonction de répartition G, est à une translation et un changement d'échelle constante. Autrement dit, G est max-stable si pour tout $n \geq 1$, s'il existe des constantes $\alpha_n > 0$ (changement d'échelle) et β_n (translation) telles que :

$$G^n(\alpha_n z + \beta_n) = G(z)$$

Théorème VI.1 (Loi du maximum renormalisé)**.** *Soit $(X_i, i = 1..n)$ une suite de variables aléatoires* iid*. Supposons qu'il existe deux suites de translations/changements d'échelle μ_n et σ_n tels que $Y_n = \frac{M_n - \mu_n}{\sigma_n}$ converge en loi vers une distribution non triviale de fonction de répartition G.*

Alors G est max-stable et est de la forme GEV (generalized extreme value) à trois paramètres $\theta = (\mu \in \mathbb{R}, \sigma > 0, \xi \in \mathbb{R})$:

$$G(y) = \exp\left(-\left(1 + \xi\left(\frac{y-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}\right), \quad \textit{avec } y \textit{ tel que } \left(1 + \xi\left(\frac{y-\mu}{\sigma}\right)\right) > 0, \qquad \text{(VI.1)}$$

et par continuité quand $\xi = 0$

$$G(y) = \exp\left(-\exp-\left(\frac{y-\mu}{\sigma}\right)\right), \quad y \in \mathbb{R}.$$

Trois cas remarquables apparaissent :

- Pour $\xi < 0$, on parle de loi de Weibull. Le support de G, i.e. $\{x; G(x) < 1\}$, est borné à droite. Autrement dit x_M, le quantile d'ordre 1 de G, est fini.
- Pour $\xi = 0$, on parle de loi de Gumbel. G tend vers 1 à l'infini à vitesse exponentielle.
- Pour $\xi > 0$, on parle de loi de Fréchet. G tend vers 1 à l'infini à vitesse polynomiale.

La condition suffisante de ce théorème est facile à montrer avec un peu d'algèbre à partir de l'équation (VI.1) . La condition nécessaire, montrer que si une telle limite existe, alors G est de la forme (VI.1), fait appel à la théorie des fonctions à variation lente qui dépasse le cadre de ce cours. Pour la démonstration on pourra consulter [4], page 322.

La loi GEV a l'avantage d'être explicite, de ne dépendre que de 3 paramètres : dans la pratique c'est une loi qu'on pourra utiliser pour modéliser la pluie maximale, la crue maximale annuelle d'une rivière (voir par exemple le tableau VI.1, bien que les débits successifs d'une rivière ne soient pas indépendants), la vitesse maximale du vent durant l'année en un lieu donné ou la plus grande intensité de secousses sismiques d'une région durant une année. Dans ces modèles, les paramètres μ (translation) et σ (échelle) sont spécifiques des dimensions auxquelles on étudie le phénomène (la *taille des données élémentaires* dont on prendrait le maximum) tandis que le paramètre ξ règle le comportement des queues de distributions. Intuitivement, plus ξ est grand, plus la distribution est concentrée à l'infini. La structure initiale de la loi du phénomène élémentaire F oriente donc le signe et la valeur de ξ.

Exemple VI.2. On peut facilement démontrer la loi du maximum renormalisé pour des queues de distribution typiques. A cet effet, distinguons trois cas :

-i) Supposons que la fonction de répartition F ait un support borné à droite, i.e. $x_M = \inf\{x; F(x) = 1\} < \infty$, et un comportement polynomial en ce point. Plus précisément, on suppose que pour un certain $\xi < 0$ et $\alpha > 0$:

$$1 - F(x) \underset{x \to x_M^-}{\sim} (\frac{x_M - x}{\alpha})^{-1/\xi}.$$

Ceci est par exemple le cas pour les lois uniformes (pour lesquelles $\xi = -1$). Dés lors,la fonction de répartition de $\dfrac{M_n - x_M}{n^\xi}$ est (pour $x < 0$)

$$\mathbb{P}\left(\frac{M_n - x_M}{n^\xi} \leq x\right) = \mathbb{P}(M_n \leq n^\xi x + x_M) = F(n^\xi x + x_M)^n,$$

et on a

$$F(n^\xi x + x_M)^n = \mathrm{e}^{n \log\left(1 - \frac{1}{n}(-\frac{x}{\alpha})^{-1/\xi} + o(1/n)\right)} \xrightarrow[n\to\infty]{} \mathrm{e}^{-(-\frac{x}{\alpha})^{-1/\xi}}.$$

Le maximum renormalisé $\dfrac{M_n - x_M}{n^\xi}$ (échelle grossissante) converge effectivement vers une distribution de Weibull.

-ii) Supposons que F ait un comportement polynomial à l'infini (queue dite "lourde"). Plus précisément, on suppose que pour un certain $\xi > 0$ et $\alpha > 0$:

$$1 - F(x) \underset{x\to+\infty}{\sim} (\frac{x}{\alpha})^{-1/\xi}.$$

Ceci est le cas pour les lois de Cauchy. La fonction de répartition de $\frac{M_n}{n^\xi}$ est (pour $x < 0$) $F(n^\xi x)^n$. Et on a

$$F(n^\xi x)^n = \mathrm{e}^{n \log\left(1-\frac{1}{n}(\frac{x}{\alpha})^{-1/\xi}+o(1/n)\right)} \xrightarrow[n\to\infty]{} \mathrm{e}^{-(\frac{x}{\alpha})^{-1/\xi}}.$$

Donc le maximum renormalisé $\frac{M_n}{n^\xi}$ (échelle rétrécissante) converge effectivement vers une distribution de Fréchet.

-iii) Supposons enfin que F ait un comportement exponentiel à l'infini. Plus précisément, soit $\sigma > 0$ et on suppose que

$$1 - F(x) \underset{x\to+\infty}{\sim} \mathrm{e}^{-\frac{x-\mu}{\sigma}}.$$

On a donc que $F(x + \sigma\log(n))^n \to_{n\to\infty} \exp(-\exp(-\frac{x-\mu}{\sigma}))$ et le maximum translaté $M_n - \sigma\log(n)$ converge effectivement vers une distribution de Gumbel.

On peut également montrer que le maximum renormalisé d'un échantillon de v.a. de loi gaussienne converge vers une loi de Gumbel. En revanche, pour les lois discrètes prenant un nombre fini de valeurs $(x_1 < ... < x_d)$ avec probabilité positive (comme par exemple la loi Bernoulli), le maximum converge en loi vers la distribution triviale de la v.a. constante égale à x_d. ◊

La valeur de projet du modèle GEV

La valeur de projet ou niveau de retour, z_p, (quantile d'ordre $1-p$) associé à la période de retour $T(z_p) = 1/p$ est défini à l'annexe A. Il est obtenu en posant

$$1 - p \equiv G\left(z_p|\sigma, \xi, \mu\right).$$

Après quelques manipulations élémentaires, on trouve, avec $x_p = -\log(1-p)$,

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left(1 - x_p^{-\xi}\right) & \text{si } \xi \neq 0, \\ \mu - \sigma\log x_p & \text{si } \xi = 0. \end{cases}$$

Remarquons que x_p peut être interprété comme l'inverse de la période de retour pour les petites valeurs de p. En effet quand p est petit

$$x_p \approx p = \frac{1}{1 - G(z_p)} = \frac{1}{T(z_p)}$$

Ainsi, le maximum annuel d'une année quelconque a la probabilité p de dépasser la hauteur z_p et, dans un repère cartésien, les couples $(\log x_p, z_p)$ dessinent une droite si $\xi = 0$ (Gumbel), une courbe concave si $\xi < 0$ (Weibull) ou convexe si $\xi > 0$ (Fréchet).

Si l'axe des abscisses est en coordonnée logarithmique, on arrive aux mêmes conclusions avec les couples (x_p, z_p).

VI.2.2 La loi des dépassements (modèle POT)

L'autre modèle caractéristique des extrêmes est celui des dépassements encore appelé *POT* pour *Peaks over Threshold* . Considérons un ensemble de variables aléatoires indépendantes $(X_n, n \geq 1)$ de même loi de fonction de répartition F. Appelons u un niveau seuil et étudions la loi des dépassements au-delà de ce seuil. Le théorème de Pickands stipule que lorsque u croît vers l'infini, on sait caractériser à la fois l'intensité et la fréquence des dépassements.

Définition VI.3. *La fonction de répartition de la loi de Pareto généralisée de paramètre* $(\sigma > 0, \xi \in \mathbb{R})$ *est, si* $\xi \neq 0$,

$$1 - \left(1 + \xi \frac{y}{\sigma}\right)^{-\frac{1}{\xi}}, \text{ pour } 0 \leq y \text{ si } \xi > 0 \text{ ou } 0 \leq y \leq \sigma/(-\xi) \text{ si } \xi < 0,$$

et (par continuité) si $\xi = 0$

$$1 - \exp(-y/\sigma) \text{ pour } y > 0.$$

Théorème VI.4 (Loi des dépassements)**.** *Soient* $(X_n, n \geq 1)$ *une suite de variables aléatoires* iid *de fonction de répartition* F *vérifiant la loi du maximum renormalisée. On note* $x_M = \inf\{x; F(x) = 1\} \in]-\infty, \infty]$ *le maximum du support de* F*. On se donne un seuil de dépassement* u *qui croît vers* x_M *et tel que* $\lim_{u \to x_M^-, n \to \infty} n(1 - F(u)) = \lambda \in]0, +\infty[$*. Dés lors, asymptotiquement quand* $u \to x_M^-$ *et* $n \to \infty$,

- *le nombre de dépassements de l'échantillon de taille* n, $K = \text{Card}\,\{i \in \{1, \ldots, n\}; X_i > u\}$*, suit une loi de Poisson de paramètre* λ.
- *Conditionnellement aux nombres de dépassements, les intensités des dépassements forment des variables aléatoires indépendantes de loi de* Pareto généralisée *: pour* $u + y < x_M$,

$$\mathbb{P}(X \leq u + y | X > u) \simeq 1 - \left(1 + \xi \frac{y}{\sigma(u)}\right)^{-\frac{1}{\xi}}, \tag{VI.2}$$

où ξ *est l'indice de la loi limite du maximum renormalisé.*

Remarque VI.5. Pour $\xi \neq 0$, il est possible de choisir l'échelle $\sigma(u) > 0$ telle que :

$$\sigma(u) \underset{u \to x_M^-}{\sim} \begin{cases} \xi(u - x_M) & \text{si } \xi < 0 \text{ (et } x_M < +\infty), \\ \xi u & \text{si } \xi > 0 \text{ (et } x_M = +\infty). \end{cases}$$

$\Diamond$

Rappelons que la distribution de Poisson de paramètre λ s'écrit :

$$\mathbb{P}(K = k) = \exp(-\lambda) \frac{\lambda^k}{k!}, \quad k \in \mathbb{N},$$

où λ est la valeur moyenne du nombre de dépassements du seuil u. Le paragraphe VI.6.2 sur la loi de Poisson permet de comprendre pourquoi c'est ici un cas limite de tirage binomial de paramètre $1 - F(u)$. On utilisera ainsi ce résultat théorique comme modèle pour décrire les températures d'une saison supérieures à un seuil ou les débits d'une rivière dépassant un niveau de référence (voir l'exemple la figure VI.2). Il y a une liaison étroite entre la GEV du paragraphe précédent et ce modèle de dépassement (POT). La loi du maximum sur une période

de temps donnée d'un modèle POT est la loi GEV. La loi conditionnelle du dépassement d'un seuil quand on sait que l'observation issue d'un modèle GEV dépasse ce seuil est la loi de Pareto généralisée. On peut fortement justifier les hypothèses du modèle utilisé : pour peu que l'on travaille avec un seuil suffisamment élevé et que l'hypothèse d'indépendance soit acceptable pour les crues de ce niveau, les conditions asymptotiques s'appliquent et entraînent la validité progressive de la représentation mathématique (VI.2). D'un autre coté, il a été simplifié pour les besoins du calcul (tout en restant réaliste pour certains cas) en posant $\xi = 0$ auquel cas l'équation (VI.2) devient par continuité la loi exponentielle :

$$\mathbb{P}(X \leq u + y \,|X > u) \simeq 1 - \left(\exp - \left(\frac{y}{\sigma(u)}\right)\right)$$

Le modèle POT pour lequel le nombre de dépassements suit une loi de Poisson avec l'intensité du dépassement exponentielle ($\xi = 0$) est encore appelé modèle de renouvellement-dépassement.

Exemple VI.6. On peut aussi vérifier la loi des dépassements pour les trois distributions typiques de l'exemple VI.2. En effet, la fonction de répartition des dépassements s'écrit :

$$\mathbb{P}(X \leq u + y|X > u) = 1 - \frac{1 - F(u + y)}{1 - F(u)}.$$

-i) Loi limite de Weibull ($\xi < 0$). On considère le cas $1 - F(x) \underset{x \to x_M^-}{\sim} (\frac{x_M - x}{\alpha})^{-1/\xi}$. On prend comme échelle $\sigma(u) = \xi(u - x_M)$ et on vérifie que

$$\mathbb{P}(X - u \leq \sigma(u)y|X - u > 0) \xrightarrow[u \to x_M^-]{} 1 - (1 + \xi y)^{-1/\xi},$$

qui est bien la fonction de répartition de la loi de Pareto généralisée avec $\xi < 0$.

-ii) Loi limite de Fréchet ($\xi > 0$). On considère le cas $1 - F(x) \underset{x \to +\infty}{\sim} (\frac{x}{\alpha})^{-1/\xi}$. On prend comme échelle $\sigma(u) = \xi u$ et on vérifie que

$$\mathbb{P}(X - u \leq \sigma(u)y|X - u > 0) \xrightarrow[u \to +\infty]{} 1 - (1 + \xi y)^{-1/\xi},$$

qui est bien la fonction de répartition de la loi de Pareto généralisée avec $\xi > 0$.

-iii) Loi limite de Gumbel ($\xi = 0$). On considère le cas $1 - F(x) \underset{x \to +\infty}{\sim} \mathrm{e}^{-\frac{x-\mu}{\sigma}}$. On prend comme échelle $\sigma(u) = \sigma$ et on vérifie que

$$\mathbb{P}(X - u \leq \sigma y|X - u > 0) \xrightarrow{u \to +\infty} 1 - \mathrm{e}^{-y},$$

qui est bien la fonction de répartition de la loi de Pareto généralisée avec $\xi = 0$.

$\Diamond$

Du modèle GEV au modèle POT

Donnons une idée heuristique de comment passer du modèle GEV au modèle POT. Soit une suite de variables *iid* à temps discret de fonction de répartition F . Sous l'hypothèse *iid*,

les *extrêmes* sont les observations élémentaires qui dépassent un seuil $u > 0$ fixé (cf. figure VI.3). On s'intéresse alors à la probabilité qu'une variable aléatoire élémentaire quelconque dépasse un certain niveau $y > 0$ quand on sait qu'elle dépasse le seuil fixé :

$$\mathbb{P}\left(X > y + u | X > u\right) = \frac{1 - F\left(y + u\right)}{1 - F\left(u\right)} \tag{VI.3}$$

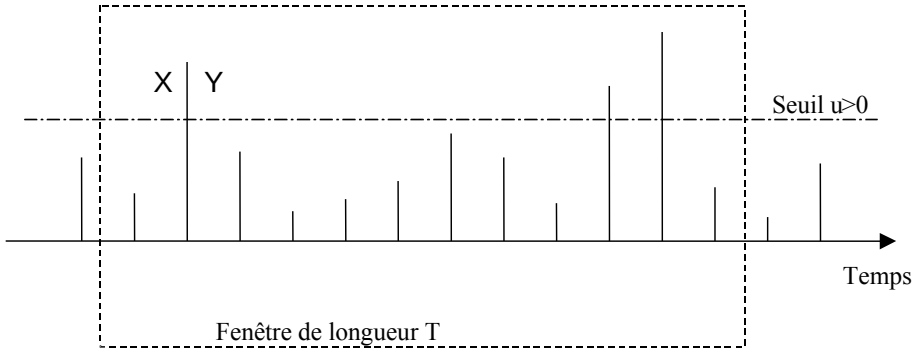


FIG. VI.3 – Au dépassement Y du seuil $u > 0$ correspond l'intensité $X = Y + u$

On sait que la distribution du maximum des observations élémentaires tend asymptotiquement vers la distribution GEV. A u fixé, il existe donc pour un n suffisamment grand, deux constantes μ_n et σ_n,telles que la loi du maximum de n variables aléatoires *iid* de loi F réalise l'approximation :

$$\mathbb{P}(M_n \leq u) = \mathbb{P}\left(\frac{M_n - \mu_n}{\sigma_n} \leq \frac{u - \mu_n}{\sigma_n}\right) \simeq G\left(\frac{u - \mu_n}{\sigma_n}\right).$$

Comme $\mathbb{P}(M_n \leq u) = F(u)^n$, on en déduit que

$$-n \log F\left(u\right) \simeq \left(1 + \xi\left(\frac{u - \mu_n}{\sigma_n}\right)\right)^{-1/\xi}$$

Si u est suffisamment proche de x_M, on utilise un développement au premier ordre du logarithme autour de 0 :

$$-\log F\left(u\right) = \log(1 - (1 - F(u)) \simeq 1 - F\left(u\right)$$

pour obtenir

$$1 - F(u) \simeq \frac{1}{n}\left(1 + \xi\left(\frac{u-\mu_n}{\sigma_n}\right)\right)^{-1/\xi}$$

Si cette relation tient pour un seuil $u > 0$, elle tiendra aussi pour tout niveau qui le dépasse, par exemple le niveau $y+u$. Dès lors, en substituant dans (VI.3) on trouve que la distribution de Pareto généralisée est candidate à la loi des dépassements quand u est suffisamment élevé, puisque :

$$\mathbb{P}(X > y + u | X > u) \simeq \left(1 + \frac{\xi y}{\sigma(u)}\right)^{-1/\xi}$$

où $\sigma(u) = \sigma_n + \xi(u - \mu_n) > 0$.

Du modèle POT au modèle GEV

On considère M, le maximum d'un grand nombre de variables aléatoires iid. Pour un seuil u (grand), $M - u$, qui représente les dépassements au dessus du seuil u, se comporte asymptotiquement comme le maximum de K variables aléatoires $(Y_i, i \in \{1, \ldots, K\})$, où les variables aléatoires $(Y_i, i \geq 1)$ sont indépendantes de loi de Pareto généralisée de paramètre $(\sigma(u), \xi)$, et indépendantes de K de loi de Poisson de paramètre $\lambda > 0$. On a donc

$$\mathbb{P}(M - u \leq y) \simeq \mathbb{P}(\max_{1 \leq i \leq K} Y_i \leq y).$$

La fonction de répartition du maximum est obtenue en sommant la répartition conjointe sur toutes les valeurs possibles de K :

$$\begin{aligned}\mathbb{P}\left(\max_{1 \leq i \leq K} Y_i \leq y\right) &= \sum_{k=0}^{\infty} \mathbb{P}\left(\max_{1 \leq i \leq K} Y_i \leq y | K = k\right) \mathbb{P}(K = k) \\ &= \exp\left[-\lambda\left(1 + \frac{\xi}{\sigma(u)} y\right)^{-1/\xi}\right],\end{aligned}$$

où l'on a tenu compte de la loi de Poisson et de la loi de Pareto généralisée. Il vient donc

$$\mathbb{P}(M \leq x) \simeq \exp\left[-\lambda\left(1 + \frac{\xi}{\sigma(u)}(x - u)\right)^{-1/\xi}\right].$$

Il est facile de montrer grâce à un simple reparamétrage que cette distribution limite est une GEV.

VI.3 Inférence

VI.3.1 Inférence du modèle GEV

Le modèle GEV n'est pas régulier : le support de la distribution dépend de la valeur des paramètres : $\mu - \sigma/\xi$ est une borne supérieure de la distribution si $\xi < 0$ et une borne inférieure si $\xi > 0$. Du fait de cette violation des conditions de régularité, les propriétés des estimateurs du maximum de vraisemblance (existence, convergence, normalité asymptotique) ne sont pas automatiquement garanties. Smith [9] a étudié ce problème de théorie en détail et formule les conclusions suivantes :

- si $\xi > -0.5$, alors les estimateurs du maximum de vraisemblance possèdent les propriétés asymptotiques ordinaires,
- si $-1 < \xi < -0.5$, alors on peut généralement calculer les estimateurs du maximum de vraisemblance, mais ils ne possèdent pas les propriétés asymptotiques classiques,
- si $\xi < -1$, il peut même être impossible de calculer les estimateurs du maximum de vraisemblance.

Les cas ennuyeux, $\xi < -0.5$, correspondent en pratique à des distributions avec une queue très courte bornée à droite qui ne sont que rarement rencontrées dans la pratique, ce qui fait que ces limites théoriques sont moins strictes qu'elles ne le paraissent de prime abord. Par contre la normalité des estimateurs est une propriété asymptotique, qui peut n'être obtenue que pour un nombre très important de données, condition irréaliste quand on travaille avec des événements rares. De ce fait la théorie proposera souvent des intervalles de confiance symétriques (alors que l'on s'attend à plus d'incertitudes à droite qu'à gauche) et trop optimistes (le théorème de Cramer-Rao donne la borne inférieure théoriquement atteignable en situation asymptotique).

Le modèle de GEV par maximum de vraisemblance

Rappelons que la vraisemblance d'un modèle aléatoire est la densité de sa loi de probabilité. La vraisemblance du modèle généralisé des extrêmes, pour $j = 1...m$ années d'enregistrements $y = (y_1, ...y_j, ...y_m)$ supposées indépendantes s'écrit donc :

$$l(y; \mu, \sigma, \xi) = \left(\frac{1}{\sigma}\right)^m \prod_{j=1}^{m} \exp\left(-\left(1 + \xi\left(\frac{y_j - \mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}\right)\left(1 + \xi\left(\frac{y_j - \mu}{\sigma}\right)\right)^{-\frac{1}{\xi}-1}.$$

On travaille généralement avec la log-vraisemblance, souvent plus maniable, $L(y; \mu, \sigma, \xi) = \log(l(y; \mu, \sigma, \xi))$, notée abusivement par la suite $L(\mu, \sigma, \xi)$:

$$L(\mu, \sigma, \xi) = -m\log(\sigma) - \left(\frac{1}{\xi} + 1\right)\sum_{j=1}^{m}\log\left(1 + \xi\left(\frac{y_j - \mu}{\sigma}\right)\right) - \sum_{j=1}^{m}\left(1 + \xi\left(\frac{y_j - \mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}.$$

Un algorithme d'optimisation numérique est indispensable pour trouver le maximum de $L(\mu, \sigma, \xi)$ sous les m contraintes $1 + \xi\left(\frac{y_j - \mu}{\sigma}\right) > 0$.

Le cas $\xi = 0$ requiert un traitement séparé. En posant $\rho = 1/\sigma$ et $\log(\lambda) = \mu/\sigma$, la vraisemblance se présente sous la forme

$$L(\mu, \sigma, \xi) = -m\log(\lambda\rho) - \rho\left(\sum_{j=1}^{n} y_j\right) - \lambda\left(\sum_{j=1}^{n}\exp(-\rho y_j)\right)$$

Exemple VI.7. Prenons comme exemple illustratif le niveau journalier de la mer à Port Pirie (Australie). Cet exemple est tiré de [3]. Les données couvrent la période 1923-1987 et peuvent être obtenues sur le site :

`http ://www.maths.bris.ac.uk/~masgc/ismev/summary.html`.

La figure VI.4 montre le profil du maximum annuel et le graphe des niveaux de retour. La variabilité du signal semble stationnaire et il est donc raisonnable de postuler que les maxima sont iid.

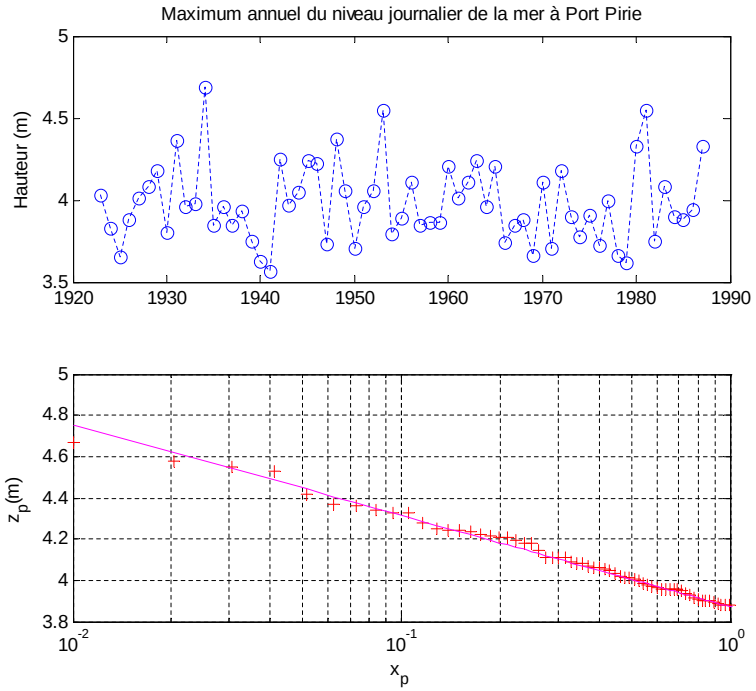


FIG. VI.4 – Chronique des maxima annuels et graphe des niveaux de retour

Appelons $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ l'estimateur du maximum de vraisemblance trouvé par évaluation numérique. On trouve ici $(\hat{\mu} \simeq 3.9, \hat{\sigma} \simeq 0.2, \hat{\xi} \simeq -0.05)$. La théorie fournit aussi les intervalles de crédibilité après calcul de la matrice de variance-covariance V (inverse de la matrice d'information de Fisher), voir la remarque II.15.

$$V = \begin{pmatrix} -\frac{\partial^2 L}{\partial \mu^2} & -\frac{\partial^2 L}{\partial \mu \partial \sigma} & -\frac{\partial^2 L}{\partial \mu \partial \xi} \\ -\frac{\partial^2 L}{\partial \mu \partial \sigma} & -\frac{\partial^2 L}{\partial \sigma^2} & -\frac{\partial^2 L}{\partial \sigma \partial \xi} \\ -\frac{\partial^2 L}{\partial \mu \partial \xi} & -\frac{\partial^2 L}{\partial \sigma \partial \xi} & -\frac{\partial^2 L}{\partial \xi^2} \end{pmatrix}^{-1}$$

soit

$$V \simeq \begin{pmatrix} 0.00078 & 0.000197 & -0.00107 \\ 0.000197 & 0.00041 & -0.00078 \\ -0.00107 & -0.00078 & 0.00965 \end{pmatrix}$$

Prenant la racine carrée de la diagonale, on obtient que les écart-types pour $\hat{\mu}, \hat{\sigma}$ et $\hat{\xi}$ sont respectivement sont $0.028, 0.020$ et 0.098. L'approximation normale fournit les intervalles de confiance correspondants pour μ, σ, ξ (pour un niveau de 95% approximativement $\pm$ 2 écart-types autour de l'estimation). On constate en particulier que celui correspondant à ξ est $[-0.24, 0.14]$, ce qui contient la valeur 0 et n'exclut pas le modèle plus simple de Gumbel. Le choix d'une distribution à support borné pour représenter les données ne va donc pas de soi.

$\Diamond$

Inférence des quantiles de la GEV par maximum de vraisemblance

Rappelons que le quantile associé à la période de retour $T = \frac{1}{p}$ est , en posant $x_p = -\log(1-p)$,

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left(1 - x_p^{-\xi}\right) & \text{si } \xi \neq 0, \\ \mu - \sigma \log x_p & \text{si } \xi = 0. \end{cases}$$

Avec $T = 100$ ans ($p = 0.01$), la hauteur centennale est obtenue en injectant ($\hat{\mu} = 3.9, \hat{\sigma} = 0.2, \hat{\xi} = -0.05$) dans ces formules, $\hat{z}_{0.01} = 4.69$. On peut aussi y associer un intervalle de confiance, car la variance de z_p peut se calculer par :

$$\mathrm{Var}(z_p) \simeq (\nabla z_p)' V (\nabla z_p),$$

avec

$$(\nabla z_p)' = \left(\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \xi}\right) = \left(1, \frac{1 - x_p^{-\xi}}{-\xi}, \frac{\left(1 - x_p^{-\xi}\right) - \xi x_p^{-\xi} \log x_p}{\xi^2/\sigma}\right)$$

(voir la remarque II.15), que l'on évalue avec $\hat{\mu} \simeq 3.9, \hat{\sigma} \simeq 0.2, \hat{\xi} \simeq -0.05$. Numériquement, l'intervalle de confiance à 95% ainsi calculé est $[4.38, 5]$.

On peut reprendre ces mêmes calculs avec un modèle de Gumbel à deux paramètres (puisque $\xi = 0$). On trouve $\hat{\mu} \simeq 3.87, \hat{\sigma} \simeq 0.195$ avec des écart-types associés 0.03 et 0.019. L'estimation de la hauteur centennale est légèrement plus forte que précédemment, $\hat{z}_{0.01} \simeq 4.77$, mais l'intervalle de confiance beaucoup plus étroit, ce qui représente le fait qu'une grosse part de l'incertitude est portée par ξ, qui traduit le comportement des queues de distribution.

Vérification du modèle

En rangeant les données y_j , $j = 1..m$ par ordre croissant on obtient un échantillon ordonné $y_{(i)}$, $i = 1..m$. La distribution empirique évaluée en $y_{(i)}$ peut être évaluée par un estimateur non-paramétrique :

$$\hat{G}_e(y_{(i)}) = \frac{i}{m+1}$$

(on notera la présence de $m+1$ et non m au dénominateur pour autoriser la possibilité de dépasser la plus grande donnée enregistrée).

En remplaçant les valeurs inconnues par leur estimation dans la distribution théorique, il vient :

$$\hat{G}_{mv}(z_{(i)}) = \exp\left(-\left(1 + \hat{\xi}\left(\frac{y_{(i)} - \hat{\mu}}{\hat{\sigma}}\right)\right)^{-\frac{1}{\hat{\xi}}}\right)$$

Si le modèle GEV marche bien G est proche de $\hat{G}$ et le graphe $(\hat{G}_{mv}(y_{(i)}), \hat{G}_e(y_{(i)}),\ i = 1..m.\ )$ des probabilités (*probability plot*) ne doit pas s'éloigner de la première diagonale. Il en va de même si on regarde le graphe des quantiles (*qq plot*) , c'est-à-dire le graphe $(\left(\hat{G}_{mv}^{-1}\left(\frac{i}{m+1}\right), y_{(i)}\right),\ i = 1..m)$.

Le modèle de Gumbel ($\xi = 0$) par ajustement linéaire

La technique du *qq plot* précédent suggère une heuristique d'estimation des paramètres d'un modèle de Gumbel fondée sur la répartition empirique : en effet, en papier de Gumbel, c'est à dire après avoir effectué la transformation $y \longmapsto \log(-\log(y))$, les quantiles empiriques $\hat{G}_{mv}^{-1}\left(\frac{i}{m+1}\right)$ et les données associées s'alignent sur une droite de pente $\frac{1}{\sigma}$ et d'ordonnée à l'origine $\frac{\mu}{\sigma}$.

Sur l'exemple des niveaux de la mer à Port-Pirie, on obtient les estimateurs $\hat{\mu} = 3.9, \hat{\sigma} = 0.2, \hat{z}_{0.01} = 4.8$. Malheureusement cette technique simple ne fournit pas d'intervalle de confiance des résultats qu'elle produit (il faut avoir recours à des méthodes plus élaborées de type bootstrap), mais ce calcul peut être utile pour procurer des valeurs initiales intéressantes en entrée d'un algorithme de recherche du maximum de vraisemblance de la GEV.

Autres estimateurs que ceux du maximum de vraisemblance

Dans la littérature des extrêmes, d'autres estimateurs que ceux du maximum de vraisemblance ont été proposés pour évaluer les paramètres inconnus (μ, σ, ξ). En particulier, l'estimation du coefficient ξ qui gouverne la forme de la queue de distribution est cruciale. Ces estimateurs sont tous basés sur la statistique d'ordre $Y_{(1)} < Y_{(2)} < ...Y_{(k)} < ...Y_{(n)}$, et on montre leur convergence en considérant les écarts de $k(n)$ valeurs ordonnées consécutives. Pour montrer la convergence en probabilité de ces estimateurs vers ξ quand $k(n)$ tend vers l'infini avec n, on se place dans les circonstances où $\frac{k(n)}{n} \underset{n\longrightarrow\infty}{\longrightarrow} 0$ pour que la proportion du nombre de termes considérés dans l'estimateur n'explose pas de façon asymptotique.

Estimateur de Pickands L'estimateur de Pickands ([8]) s'exprime comme :

$$\hat{\xi}_k = \frac{1}{\log(2)} \log\left(\frac{Y_{(n-k)} - Y_{(n-2k)}}{Y_{(n-2k)} - Y_{(n-4k)}}\right).$$

Il est valable quelque soit le signe de ξ . La preuve de la convergence est donnée à la page 332 de [4]. En pratique on est à n fixé et on trace le graphique de cet estimateur en fonction du nombre k d'observations considérées, mais le comportement est très volatil au départ, ce qui nuit à la lisibilité du graphique. De plus, l'estimateur est mécaniquement très sensible à la taille de l'échantillon sur lequel on travaille, ce qui le rend peu robuste.

Estimateur de Hill L'estimateur de Hill ([5]) ne fonctionne que pour le domaine d'attraction de Fréchet ($\xi > 0$). Il est donné par

$$\hat{\xi}_k = \frac{1}{k} \sum_{j=1}^{k} \log\left(Y_{(n-j+1)}\right) - \log(Y_{(n-k)}).$$

Il s'interprète comme la pente d'un *qqplot* sur une zone de fortes valeurs (recherche de la pente à l'infini). Ses conditions d'applications sont sujettes à de nombreuses discussions dans la communauté statistique. Pour la preuve de la convergence, on consultera [4], page 334.

VI.3.2 Inférence du modèle POT

Le modèle POT souffre des mêmes difficultés d'estimation que le modèle GEV.

Choix du seuil

Quand on a affaire à un jeu réel de données, comme des débits journaliers, il faut choisir le seuil au delà duquel la modélisation POT est réaliste.

Dans le cas des débits par exemple, il faut discuter la notion d'indépendance : deux événements de crues peuvent être considérés comme indépendants s'il sont séparés par un intervalle de temps suffisamment long (vis à vis des temps de réaction de la rivière pour concentrer les pluies d'orage) et une seule valeur (le maximum de débit) doit être retenue sur un intervalle de crue. Le seuil ne doit donc pas être choisi trop bas (de plus l'approximation asymptotique ne tiendrait pas). Le choisir trop haut réduit drastiquement le nombre de données. L'idée est donc de choisir le seuil le plus bas qui rencontre ces deux exigences.

Du point de vue de la théorie statistique, si Y suit une loi de Pareto généralisée de paramètres σ et $\xi < 1$ son espérance est

$$\mathbb{E}(Y) = \frac{\sigma}{1-\xi}$$

et on a

$$\mathbb{E}(Y - u \,|Y > u\,) = \frac{\xi u + \sigma}{1-\xi}.$$

Si dans un modèle POT , $X - u_0$ dénote le dépassement au delà d'un niveau u_0, on a de même

$$\mathbb{E}(X - u_0 \,|X > u_0\,) \simeq \frac{\sigma_{u_0}}{1-\xi}$$

En passant du niveau u_0 au niveau u, on change le paramètre d'échelle de σ_{u_0} à $\sigma(u) = \sigma_{u_0} + \xi u$, par conséquent l'espérance des dépassements est une fonction linéaire du seuil :

$$\mathbb{E}(X - u \,|X > u) \simeq \frac{\xi u + \sigma_{u_0}}{1-\xi}$$

En considérant l'échantillon ordonné, $x_{(1)} \leq \cdots \leq x_{(n)}$, et en appelant n_u le nombre de points qui dépassent le seuil u, la courbe $(u, \frac{1}{n_u}\sum_{i=1}^{n_u}(x_{(i)} - u))$ doit être approximativement linéaire dans la zone des seuils ou le modèle de Pareto est valide. La figure VI.5 donne un exemple type de courbe pour laquelle on voit que la courbe peut être considérée linéaire uniquement qu'à partir de la seconde graduation en u. Pour les fortes valeurs de u, il ne reste que peut de données donc une grande variabilité de l'estimation de l'espérance d'estimation.

Estimation des paramètres du modèle POT

L'estimation peut se faire de façon séparée pour l'estimation du paramètre de la loi de Poisson réglant l'occurrence des événements (le maximum de vraisemblance ou la méthode des moments fournissent le même résultat) et les paramètres σ et ξ de la Pareto qui dirigent la valeur observée. Pour cette dernière la log-vraisemblance s'écrit :

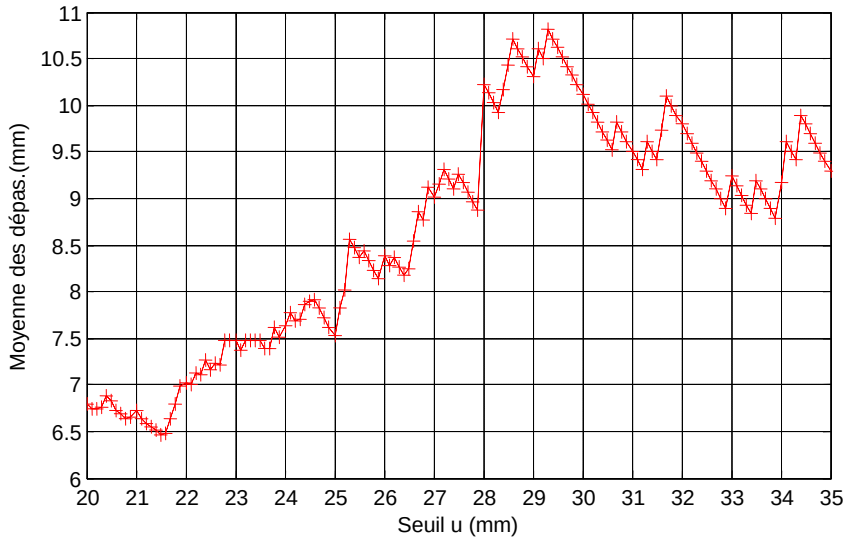


FIG. VI.5 – Exemple de choix du seuil : prendre $u = 22$ unités au moins

$$L(\sigma, \xi) = -k \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{k} \log(1 + \xi \frac{y_i}{\sigma})$$

On doit alors avoir recours à des méthodes numériques pour en rechercher le maximum .

Dans le cas Gumbel ($\xi = 0$), les choses sont plus simple car il s'agit d'estimer le paramètre d'une loi exponentielle. On prend alors $\hat{\sigma} = \frac{1}{k} \sum_{i=1}^{k} y_i$.

VI.4 Traitement décisionnel de la construction d'une digue

Posons le problème décisionnel complet pour la construction d'une digue. La notion de période de retour d'événements décrite à l'annexe 1 de ce chapitre est essentielle pour comprendre la pratique traditionnelle d'ingénierie. On suppose ici que les crues de la rivière au delà d'un seuil u sont régies par un modèle POT avec $\xi = 0$. On considère que toutes les valeurs sont exprimés à partir du seuil u , pris comme niveau de référence. Si on construit une digue de hauteur h on investit $C(h)$, coût d'amortissement annuel de l'ouvrage de protection élevé jusqu'à la hauteur h. Si une crue X excède la valeur h, on subit un dommage $D(X, h)$, nul quand $X < h$. Le nombre moyen de crue dépassant le seuil u est donné par le paramètre de la loi de Poisson du nombre de dépassement, $\lambda > 0$, est supposé connu. Le dommage moyen est donc $\int_{x=h}^{+\infty} \lambda k D(x, h) \sigma^{-1} e^{-\sigma^{-1} x} dx$ où k est un facteur lié à l'actualisation, supposé

fixe et connu dans cet exemple. La dépense moyenne totale vaut alors :

$$W(h,\sigma,\lambda) = C(h) + \int_{x=h}^{+\infty} \lambda k D(x,h)\, \sigma^{-1} e^{-x/\sigma} dx.$$

Plaçons-nous dans le cadre d'une démarche économique classique, fondée sur l'hypothèse que λ et σ^{-1} sont connus. En supposant C et D approximativement linéaires, respectivement $C(h) \simeq C_0 h$ et $D(x,h) \simeq D_0(x-h)$, au voisinage de la hauteur de digue idéale, l'équation précédente donne une hauteur de digue idéale $h^*\left(\lambda, \sigma^{-1}\right)$ telle que :

$$\left.\frac{\partial W(h,\sigma,\lambda)}{\partial h}\right|_{h=h^*} = 0. \tag{VI.4}$$

Comme

$$\frac{\partial W(h,\sigma,\lambda)}{\partial h} = \frac{\partial C(h)}{\partial h} + \int_{x=h}^{+\infty} \lambda k \frac{\partial D(x,h)}{\partial h} \sigma^{-1} e^{-x/\sigma} dx - kD(h,h)\,\lambda \sigma^{-1} e^{-h/\sigma},$$

il vient $0 = C_0 - kD_0\lambda(1 - G(h^*))$ avec $G(y) = \left(1 - \int_{x=0}^{y} \sigma^{-1} e^{-x/\sigma} dx\right)$. On en déduit que

$$h^*(\lambda,\sigma) = \sigma \log(\lambda k D_0 / C_0). \tag{VI.5}$$

La période de retour $T(h^*)$ définie par (VI.6) est égale à

$$T(h^*) = \frac{1}{\lambda(1 - G(h^*))} = \frac{kD_0}{C_0}.$$

En d'autres termes, si on est sans incertitude sur les caractéristiques des aléas hydrologiques et avec ce modèle hydrologique de renouvellement-dépassement, la digue idéale est telle que la période de retour de la crue maximale annuelle de projet est égale au rapport de la valeur marginale du coût d'investissement sur la valeur marginale du coût des dommages. Ce court développement mathématique justifie donc une pratique courante d'ingénierie hydraulique qui opère de la façon suivante :

- Sélectionner une période de retour de la crue de projet vingtennale, centennale ou millennale selon l'importance des enjeux économiques associés à l'ouvrage. L'équation (VI.5) montre le lien $T(D,C)$ entre T et les composantes des fonctions de coût.
- Estimer λ et σ à partir des données statistiques (n crues $x_i, i = 1..n$, se sont produites sur r années). Les estimateurs classiques sont, en notant $S(n)$ le débordement cumulé sur n années :

$$\hat{\lambda} = \frac{n}{r}, \quad \hat{\sigma} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{S(n)}{n}.$$

- Estimer la hauteur de la digue idéale par :

$$h^*\left(\hat{\lambda}, \hat{\sigma}\right) = \frac{S(n) \log\left(\frac{n}{r} T(D,C)\right)}{n}$$

En supposant $T(D,C) = 100$ ans, on trouve avec les données de l'exemple de la Garonne : $\hat{\sigma}^{-1} = 0.98 \times 10^{-3}$, $\hat{\lambda} = 2.385$ d'où une hauteur de digue équivalente à $\hat{\sigma} \log\left(\hat{\lambda} T\right) = 5586\ m^3/s$ au-dessus du seuil de $2500 m^3/s$ pris comme référence pour les dépassements, soit finalement une digue d'une hauteur contenant un débit de 8086 m^3/s pour se prémunir contre la crue centennale.

Pour des fonctions de coûts plus compliquées ou si on travaille avec un modèle plus général à trois paramètres, la résolution de (VI.4) ne se fait plus de façon analytique et on a recours à la simulation numérique pour le calcul de l'intégrale et la recherche de l'optimum.

VI.5 Conclusions et perspectives

Les modèles GEV et POT sont relativement faciles à mettre en œuvre. Sur une même chronique complète de données, le modèle POT exploite généralement mieux l'information. On peut d'ailleurs comparer la valeur de projet ainsi obtenue avec celle déduite d'un modèle GEV sur la même période. Un simple graphique "seuil vs moyenne des dépassements" permet d'orienter le choix du seuil qui reste malgré tout une opération délicate. Dans le doute, remonter un peu le seuil est certainement une bonne idée.

Enfin, les modèles GEV et POT sont fondés sur l'hypothèse que le processus stochastique à temps discret sous-jacent est constitué de populations indépendantes et de même loi. C'est une hypothèse forte et critiquable dans bon nombre de situations réelles où les effets saisonniers sont difficilement contestables (sans même considérer le changement climatique). Ainsi, la lame d'eau journalière dépend de la carte du temps et, en situation cyclonique, les jours pluvieux se suivent. Tant que le processus stochastique sous-jacent est stationnaire, les modèles GEV et POT sont relativement peu sensibles à la dépendance des populations élémentaires. Pour les processus non stationnaires, des modélisations plus élaborées sont nécessaires. En situation stationnaire, une analyse de sensibilité (hauteur du seuil, prédiction de records historiques enregistrés sur site) constitue souvent le meilleur test de la pertinence du modèle.

La rareté de données (par la définition même de l'occurrence de situations d'extrêmes) rend généralement l'extrapolation des quantiles élevés périlleuse, d'autant plus que les intervalles de confiance (calculé en situation asymptotique d'abondance de données) sous-estiment généralement l'incertitude des valeurs de projet. Le recours au paradigme Bayésien ([1],[2]) peut être une façon de proposer des estimations avec des fourchettes d'incertitudes plus réalistes et d'incorporer l'expertise en plus de l'information apportée par les données ([7]).

VI.6 Sur la période de retour et la loi de Poisson

VI.6.1 Période de retour

Période de retour du modèle GEV

La période de retour $T(y)$ d'un phénomène récurrent associé à une variable aléatoire X de fonction de répartition F est la durée moyenne qui sépare deux événements du phénomène de valeur dépassant un seuil y . Soit une séquence $\{X_1, X_2, X_3, ...X_n, X_{n+1}\}$ de réalisations *iid* (indépendantes et identiquement distribuées) de même loi F. Pour la calculer, fixons un

seuil y et appelons Z_y la variable aléatoire entière *longueur de l'intervalle de temps séparant deux réalisations du phénomène* X dépassant le seuil y. L'événement $\{Z_y = n\}$, pour $n \geq 1$, coïncide avec l'événement réalisé par $\{X_2 < y; X_3 < y; ...X_n < y; X_{n+1} > y | X_1 > y\}$; la probabilité d'un tel événement est donc :

$$\mathbb{P}\left(Z_y = n\right) = \left(F\left(y\right)\right)^{n-1}\left(1 - F\left(y\right)\right), \quad n \geq 1.$$

La variable Z_y suit la loi géométrique de paramètre $1 - F(y)$. En particulier son espérance est $\mathbb{E}[Z_y] = 1/(1 - F(y))$. Par définition la période de retour associée à la valeur y, $T(y)$, est la valeur moyenne de Z_y :

$$T(y) = 1/(1 - F(y)). \tag{VI.6}$$

Période de retour du modèle POT

Reprenons maintenant le modèle de dépassement-renouvellement, pour lequel cette formule n'est pas directement applicable puisqu'il peut se produire plus d'une crue dommageable par année. Supposons d'abord qu'il se produise exactement r crues sur une année : la probabilité pour que la plus grande des crues soit inférieure à x vaut $(G\left(x\right))^r$. Si le nombre de crue suit une de loi de Poisson de paramètre θ, et que ce nombre est indépendant de la hauteur de la crue, la fonction de répartition, F, de la crue maximale annuelle est :

$$\begin{aligned} F(x) &= \sum_{r=0}^{+\infty} \frac{\theta^r \, \mathrm{e}^{-\theta}}{r!} G(x)^r \\ &= \mathrm{e}^{-\theta(1-G(x))} \, . \end{aligned}$$

La période de retour correspondante est

$$T(y) = \frac{1}{1 - F(y)} = \frac{1}{1 - \mathrm{e}^{-\theta(1-G(y))}} \, .$$

Quand T est grand (i.e. y grand et $G(y)$ proche de 1), on a

$$T(y) \simeq \frac{1}{\theta(1 - G(y))} \, .$$

VI.6.2 La loi de Poisson

Le modèle de Poisson permet de représenter l'occurrence d'événements rares. Elle s'obtient par passage à la limite de la loi binomiale. Soit X_n une variable aléatoire suivant une loi binomiale de paramètres (n, p_n), avec $n \geq 1$ et $p_n \in]0, [1$. La variable X_n représente le nombre d'occurrences d'un évènement de probabilité p_n lors de n observations. On a pour $0 \leq k \leq n$,

$$\mathbb{P}(X_n = k) = \frac{n!}{k!(n-k)!} p_n^k (1 - p_n)^{n-k}.$$

La loi de Poisson est la loi limite de X_n, quand on dispose de beaucoup d'observations (n grand), mais que l'on considère un évènement très rare (p_n petit), et que l'on dispose de quelques occurrences en moyenne ($np_n \approx 1$). Un calcul élémentaire assure que si $n \to \infty$,

$p_n \to 0$ et $np_n \to \theta > 0$, alors $\mathbb{P}(X_n = k) \to \dfrac{\theta^k}{k!}\,\mathrm{e}^{-\theta}$. On retrouve ainsi comme limite la probabilité pour qu'une variable de loi de Poisson de paramètre θ soit égale à k.

La loi de Poisson est très employée dans le cas d'événements rares, par exemple pour le nombre de bactéries que l'on va pouvoir décompter lorsqu'on effectue un prélèvement dans un milieu où elles se répartissent de façon homogène. Elle a l'avantage de ne dépendre que d'un seul paramètre λ qui fixe l'espérance d'une variable aléatoire de Poisson. C'est en même temps un manque de souplesse, puisque ce paramètre fixe aussi la variance, ce qui restreint la portée du modèle à des phénomènes où moyenne et variance seront égales.

Bibliographie

[1] J. O. Berger and D. Rios Insua. Recent developments in bayesian inference with applications in hydrology. In E. Parent, P. Hubert, B. Bobée, and J. Miquel, editors, *Bayesian Methods in Hydrological Sciences*, pages 43–62. UNESCO Publishing, 1998.

[2] J. Bernier, E. Parent, and J-J. Boreux. *Statistique de l'Environnement. Traitement Bayésien de l'Incertitude*. Lavoisier, Paris, 2000.

[3] S. Coles. *An Introduction to Statistical Modelling of Extremes Values*. Springer-Verlag, Londres, 2001.

[4] J.F. Delmas and B. Jourdain. *Modèles Aléatoires*. Springer, 2006.

[5] B.M. Hill. A simple general approach about the tail of a distribution. *Ann. Stat.*, 3 :1163–1174, 1975.

[6] Jacques Miquel. *Guide Pratique D'estimation Des Probabilités de Crues*. Eyrolles, Paris, 1984.

[7] E. Parent and J. Bernier. Bayesian POT modeling for historical data. *Journal of Hydrology*, 274 :95–108, 2003.

[8] J. Pickands. Statistical inference using extreme order statistics. *Ann. Stat.*, 3 :11–130, 1975.

[9] R. L. Smith. Maximum likelihood estimation in a class of non regular cases. *Biometrika*, 72 :67–90, 1985.

Chapitre VII

Séries chronologiques

VII.1 Introduction

VII.1.1 Motivations et objectifs

Une **série temporelle** , ou série chronologique, est un ensemble d'observations qui se distinguent par le rôle important que joue l'ordre dans lequel elles ont été recueillies. Les objectifs du cours sont les suivants :

1. Comprendre les problématiques posées par le lien temporel.
2. Etre capable de mettre en oeuvre des techniques de base (statistiques) sur des séries temporelles.

L'importance de ce domaine est illustrée par les nombreux domaines d'application :
- Economie : prévision d'indices économiques...
- Finance : évolution des cours de la bourse...
- Démographie : analyse de l'évolution d'une population...
- Météorologie : analyse de données climatiques...
- Médecine : analyse d'electrocardiogrammes...
- Géophysique : analyse de données sismiques...
- Théorie du signal : transmission de signaux bruités...
- Traitement d'images : analyse d'images satellites, médicales...
- Energie : prévision de la consommation d'éléctricité...

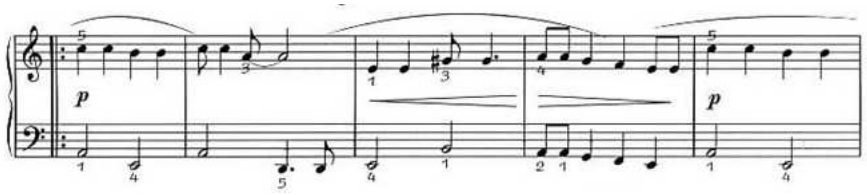
Les buts poursuivis sont multiples :
- Description : détermination de composantes...
- Filtrage : transformation de la série dans le but d'éliminer certaines caractéristiques ou des valeurs aberrantes...
- Modélisation : recherche de causalité...
- Prévision.

Il existe évidemment des interactions entre ces différents objectifs. Afin de mener l'étude pratique d'une série temporelle, on ne doit pas négliger la phase descriptive, pour envisager éventuellement un filtrage de la série. La modélisation, qui s'avère souvent plus facile sur une série filtrée, fournit les outils pour effectuer des prévisions.

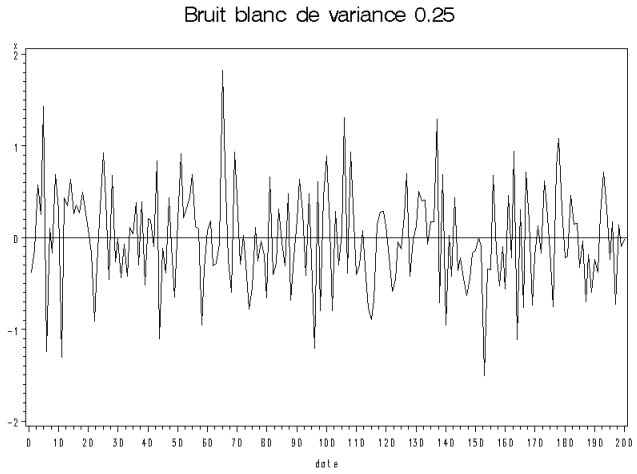
VII.1.2 Exemples de séries temporelles

Dans ce document, les exemples, les traitements et donc les graphiques sont obtenus à l'aide du logiciel SAS.

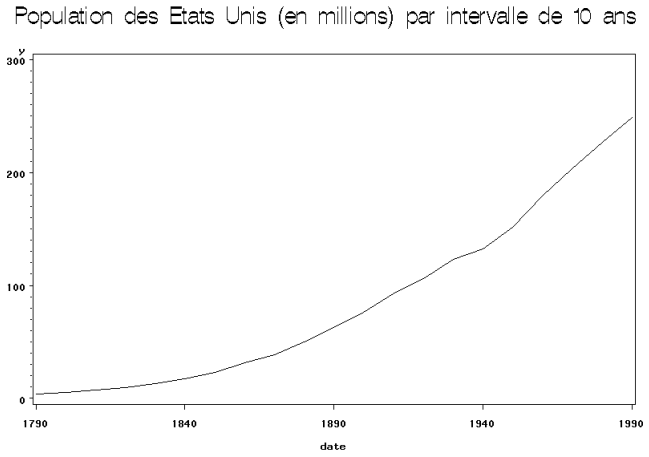
1. La partition de musique est un exemple de série temporelle, même s'il est rare de la modéliser mathématiquement (si on excepte la musique sérielle : Cf. Arnold Schönberg) : modifier l'emplacement d'une note a des conséquences musicales évidentes.



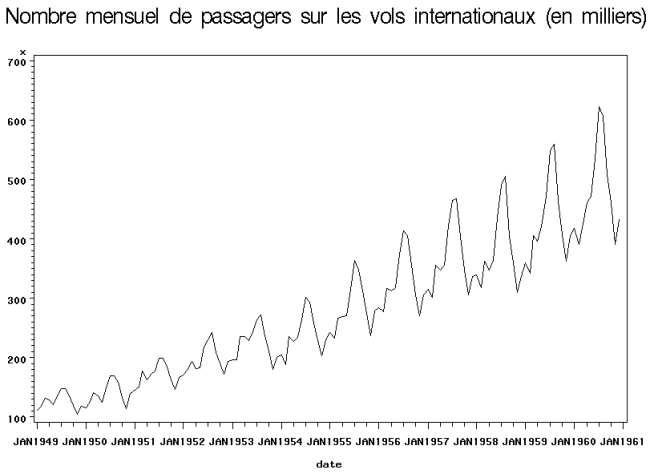
2. Un bruit blanc fort est constitué de variables aléatoires indépendantes et identiquement distribuées (*i.i.d.*), d'espérance nulle. Voici un exemple simulé à partir d'une loi normale $\mathcal{N}(0, 0.25)$.



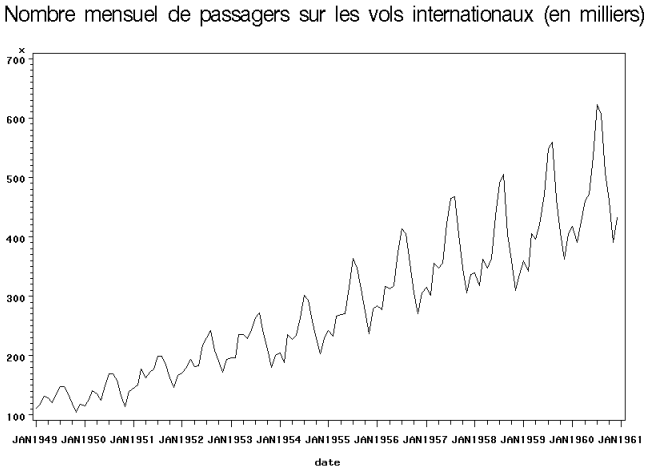
3. La série suivante représente la population des Etats Unis de 1790 à 1990. On peut observer une tendance de type exponentiel puis linéaire, ainsi qu'un point d'inflexion dont il faudrait déterminer la cause si on veut modéliser le plus correctement possible cette série.



4. Le graphique ci-dessous représente le nombre mensuel de passagers aériens de janvier 1949 à décembre 1960. C'est une des séries les plus utilisées comme exemple d'application et elle figurera abondamment dans ce document sous la dénomination "Nombre de passagers aériens".



On observe une tendance de type exponentiel ainsi qu'une saisonnalité de période 12 (saisonnalité annuelle) qui s'accentue avec le temps. Si on effectue une transformation logarithmique de cette série, on se ramène à une tendance linéaire et à une saisonnalité non volatile :

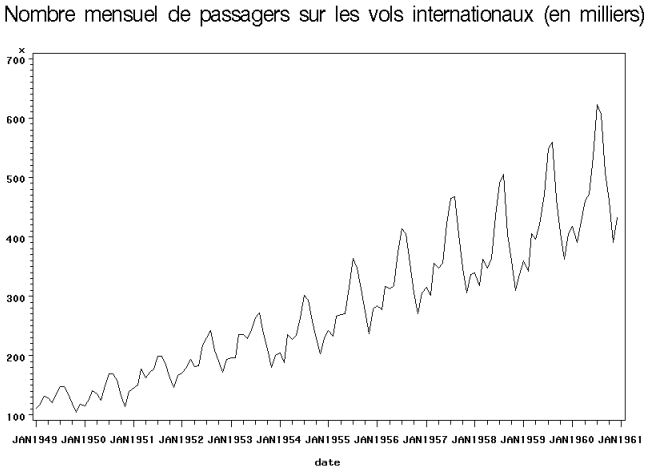


VII.1.3 Repères historiques

On peut distinguer trois phases dans l'histoire de l'analyse des séries temporelles :

1. Les séries temporelles apparaissent avec l'utilisation de la représentation graphique, en astronomie.

   Le plus ancien diagramme connu figure ci-dessous ; il représente l'inclinaison des planètes en fonction du temps (illustration d'un commentaire du *Songe de Scipion* de Cicéron, extrait des *Saturnales* de Macrobius, en 395 après J.C).



2. A partir des 18ème et 19ème siècles, on passe de la visualisation graphique aux premières techniques temporelles (déterministes). Citons deux voies très importantes :
   - les travaux de Schuster (1898, 1906) à partir de ceux de Fourier (1807) et Stokes (1879), sur l'analyse fréquentielle d'une série temporelle (un signal est approché par une somme de sinusoïdales) ;
   - les travaux de Pearson (1916) sur la décomposition d'une série temporelle en termes de composantes tendantielle, cyclique, saisonnière et accidentelle.

3. A partir du 20ème siècle, l'aléatoire est pris en compte, notamment à l'aide des travaux de Yule (1927). Ces travaux sont issus de l'observation du mouvement oscillatoire d'un pendule bombardé de petits pois lancés par un enfant !

   Il y a ensuite de nombreux contributeurs aux méthodes aléatoires : Cramer (1937, 1951), Wold (1938), Kolmogorov (1939)...

VII.1.4 Principaux modèles statistiques pour l'étude des séries temporelles

On présente ici les principales familles de modèles utilisés pour traiter les séries temporelles.

Modèles autorégressifs ("Auto-Regressive")

Ils ont été introduits par Yule en 1927. On prend en compte une dépendance linéaire du processus à son propre passé :

$$AR(p): \quad X_t = \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + \varepsilon_t,$$

où $p \in \mathbb{N}^*$ est l'ordre du processus, $\varphi_1, \ldots, \varphi_p$ sont des constantes réelles et $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc (cf. définition VII.3).

Modèles moyennes mobiles ("Moving Average")

Ils ont également été introduits en 1927, par Slutsky. Un processus moyenne mobile est la somme d'un bruit blanc et des ses retards :

$$MA(q): \quad X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

où $q \in \mathbb{N}^*$ est fixé et $\theta_1, \ldots, \theta_q$ sont des constantes réelles.

Modèles ARMA ("Auto-Regressive Moving Average")

Développés par Box & Jenkins en 1970, les modèles ARMA sont une combinaison des modèles autorégressif et moyenne mobile :

$$ARMA(p,q): \quad X_t - \varphi_1 X_{t-1} - ... - \varphi_p X_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q}.$$

Les modèles ARIMA (AutoRegressive Integrated Moving Average) et SARIMA (Seasonnal AutoRegressive Integrated Moving Average) (un processus SARIMA est un processus ARMA intégré avec une composante saisonnière) ont ensuite été développés afin de pouvoir modéliser un grand nombre de phénomènes réels qui présentent des tendances et/ou des saisonnalités. On applique en fait des modèles ARMA à des séries dites différenciées ; par exemple, pour un ARIMA d'ordre 1, on suppose que $X_t - X_{t-1}$ est un ARMA (on note en général $\nabla X_t = X_t - X_{t-1}$ ou encore $(I - B) X_t$ en utilisant les opérateurs identité I et rétrograde ("backward") $BX_t = X_{t-1}$).

Modèles ARCH ("Auto-Regressive Conditional Heteroskedasticity")

En 1982, Engle a proposé des modèles autorégressifs prenant en compte une "volatilité stochastique" :

$$ARCH(p): \quad X_t = \varepsilon_t \sqrt{h_t} \text{ avec } h_t = \alpha_0 + \alpha_1 X_{t-1}^2 + ... + \alpha_p X_{t-p}^2,$$

où $p \geq 1$ est fixé et $\alpha_0, \ldots, \alpha_p$ sont des constantes positives.

Modèles à mémoire longue

Les modèles envisagés la plupart du temps sont des modèles dits à mémoire courte : deux instants éloignés du processus n'ont que très peu d'interaction entre eux. Il existe d'autres modélisations pour les processus à mémoire longue, tels que les processus FARIMA (fractional ARIMA) introduits en 1980 par Granger.

Modèles multivariés ("Vector Auto-Regressive")

On est parfois contraints de modéliser un phénomène multiple, ou plusieurs séries ayant de fortes relations entre elles. Les modèles autorégressifs vectoriels sont un exemple d'une modélisation multivariée :

$$VAR: \quad X_t = AX_{t-1} + E_t \quad \text{où} \quad X_t = \left(X_t^1, X_t^2\right),$$

où A est une matrice carrée constante et $(E_t)_{t\in\mathbb{Z}}$ est un bruit blanc multidimensionnel.

Modèles non-paramétriques

Tous les modèles envisagés précédemment sont paramétriques : l'estimation d'un ou plusieurs paramètres suffit pour déterminer les relations temporelles d'un processus. On peut cependant considérer que la fonction de lien n'est pas paramétrée ; on cherche alors à déterminer une fonction f dans des classes de fonctions adaptées traduisant la relation temporelle, par exemple :

$$X_t = f\left(X_{t-1}, ..., X_{t-p}\right) + \varepsilon_t.$$

La fonction f peut être estimée à l'aide de la méthode des noyaux, des séries de Fourier, des ondelettes...

Modèles semi-paramétriques

Les modèles non-paramétriques souffrent du "fléau de la dimension". On utilise alors une modélisation non-paramétrique sur une ou plusieurs combinaison de variables, par exemple

$$X_t = f\left(\theta_1 X_{t-1} + ... + \theta_p X_{t-p}\right) + \varepsilon_t,$$

où $p \geq 1$ est fixé et $\theta_1, \ldots, \theta_p$ sont des constantes.

VII.2 Processus univariés à temps discret

Cette section pose les bases de la modélisation probabiliste des séries temporelles. Il définit les notions de processus stochastique et de stationnarité, ainsi que des outils d'analyse comme les autocorrélogrammes et le périodogramme.

Soit $(x_t)_{t\in T}$ une famille d'observations d'un phénomène qui peut être physique, économique, biologique... Chaque observation $x_t \in \mathbb{R}^d$ a été enregistrée à un temps spécifique $t \in T$ et on appelle **série temporelle** cet ensemble d'observations.

Si T est dénombrable (en général $T \subset \mathbb{Z}$), on parle de série temporelle **à temps discret** . Si T n'est pas dénombrable (en général un intervalle de $\mathbb{R}$), on parle de série temporelle **à temps continu** .

On considère des séries temporelles à valeur dans $\mathbb{R}^d$. Si $d = 1$, on parle de série **univariée** . Si $d > 1$, on parle de série **multivariée** .

On désignera dans la suite par série temporelle une **série temporelle univariée à temps discret**.

On considère en Statistique que l'observation x est la réalisation d'une variable aléatoire X. De manière analogue, une série temporelle $(x_t)_{t\in T}$ est considérée comme la réalisation d'un processus stochastique (d'une suite de variables aléatoires) $(X_t)_{t\in T}$.

VII.2.1 Processus stochastique du second ordre

Définition VII.1. *Soit $X := (X_t)_{t\in T}$ un processus stochastique (i.e. une suite de variables aléatoires). Le processus X est dit du second ordre si pour tout $t \in T$, X_t est une variable aléatoire de carré intégrable* i.e. $\mathbb{E}(|X_t|^2) < +\infty$.

Voici deux exemples de processus du second ordre qui sont fondamentaux dans la suite.

Définition VII.2. *$(\varepsilon_t)_{t\in\mathbb{Z}}$ est un **bruit blanc fort** si :*
- *$(\varepsilon_t)_{t\in\mathbb{Z}}$ est une suite de variables aléatoires réelles indépendantes et identiquement distribuées (i.i.d.),*
- *$\forall t \in \mathbb{Z} : \mathbb{E}(\varepsilon_t) = 0$ et $\mathbb{E}(\varepsilon_t^2) = \sigma^2$.*

Définition VII.3. *$(\varepsilon_t)_{t\in\mathbb{Z}}$ est un **bruit blanc faible** si :*
- *$(\varepsilon_t)_{t\in\mathbb{Z}}$ est une suite de variables aléatoires réelles identiquement distribuées,*
- *$\forall (t,t') \in \mathbb{Z}^2, t \neq t' : \mathbb{C}ov(\varepsilon_t, \varepsilon_{t'}) = 0$,*
- *$\forall t \in \mathbb{Z} : \mathbb{E}(\varepsilon_t) = 0$ et $\mathbb{E}(\varepsilon_t^2) = \sigma^2$.*

On rappelle maintenant un résultat important pour l'étude de processus du second ordre.

Proposition VII.4. *Soient $(X_n)_{n\in\mathbb{Z}}$ et $(Y_n)_{n\in\mathbb{Z}}$ deux processus tels que*

$$\lim_{n\to+\infty} \mathbb{E}\left[\Big(\sum_{i=n}^{+\infty} X_i\Big)^2 + \Big(\sum_{i=-\infty}^{-n} X_i\Big)^2\right] = 0 \text{ et } \lim_{n\to+\infty} \mathbb{E}\left[\Big(\sum_{i=n}^{\infty} Y_i\Big)^2 + \Big(\sum_{i=-\infty}^{-n} Y_i\Big)^2\right] = 0.$$

Alors, on a

$$\mathbb{C}ov\left(\sum_{i=-\infty}^{+\infty} X_i, \sum_{j=-\infty}^{+\infty} Y_j\right) = \sum_{i=-\infty}^{+\infty}\sum_{j=-\infty}^{+\infty} \mathbb{C}ov(X_i, Y_j)$$

VII.2.2 Processus stationnaire

Dans de très nombreux cas, on ne peut pas renouveler la suite de mesures dans des conditions identiques (par exemple le taux de chômage mensuel). Alors pour que le modèle déduit à partir d'une suite d'observations ait un sens, il faut que toute portion de la trajectoire observée fournisse des informations sur la loi du processus et que des portions différentes, mais de même longueur, fournissent les mêmes indications. D'où la notion de stationnarité.

Définition VII.5. *Un processus $(X_t)_{t\in T}$ est* ***fortement stationnaire*** *ou strictement stationnaire si, pour tous $k \geq 1$, $(t_1, ..., t_k) \in T^k$, h tel que $(t_1 + h, \ldots, t_k + h) \in T^k$, les vecteurs $(X_{t_1}, ..., X_{t_k})$ et $(X_{t_1+h}, ..., X_{t_k+h})$ ont même loi.*

Cette propriété très forte est très difficile à vérifier, d'où la notion de stationnarité faible.

Définition VII.6. *Un processus $(X_t)_{t\in T}$ du second ordre est* ***faiblement stationnaire*** *ou stationnaire à l'ordre 2, si son espérance $\mathbb{E}(X_t)$ et ses autocovariances $\mathbb{C}ov(X_s, X_t)$ sont invariantes par translation dans le temps :*

- $\forall t \in T, \quad \mathbb{E}(X_t) = \mathbb{E}(X_0)$,
- $\forall (s,t) \in T^2, \forall h \,/ (s+h, t+h) \in T^2 , \quad \mathbb{C}ov(X_s, X_t) = \mathbb{C}ov(X_{s+h}, X_{t+h})$.

Dans la suite, on notera $\mu_X = \mathbb{E}(X_0)$ et $\gamma_X(h) = \mathbb{C}ov(X_0, X_h)$.

Remarque VII.7.

- La stationnarité faible est bien plus facile à étudier et à vérifier que la stationnarité stricte.
- Un processus stationnaire n'est pas obligatoirement borné ou "sympathique" (par exemple, un bruit blanc).
- Pour un processus du second ordre, la stationnarité stricte implique la stationnarité faible. La réciproque est fausse (elle n'est vraie que pour les processus gaussiens).

◊

Exemple VII.8.

1. Un bruit blanc fort est fortement stationnaire.
2. Un bruit blanc faible est faiblement stationnaire.
3. Un processus présentant une tendance et/ou une saisonnalité n'est pas faiblement stationnaire.

◊

On désignera dans la suite par processus stationnaire, un processus faiblement stationnaire (donc du second ordre). De plus, on supposera que le processus est à temps discret et plus précisément que $T = \mathbb{Z}$.

Proposition VII.9. *Soit $(X_t)_{t\in\mathbb{Z}}$ un processus stationnaire (au sens faible). Soit une suite $(a_i)_{i\in\mathbb{Z}}$ telle que $\sum_{i\in\mathbb{Z}} |a_i| < +\infty$. Le processus $(Y_t)_{t\in\mathbb{Z}}$ où $Y_t = \sum_{i\in\mathbb{Z}} a_i X_{t-i}$, est stationnaire et*

- $\mu_Y = \mu_X \sum_{i\in\mathbb{Z}} a_i$ *où* $\mu_X = \mathbb{E}(X)$,
- $\gamma_Y(h) = \sum_{i\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} a_i a_j \gamma_X(h + j - i)$.

Cela implique notamment qu'une somme de v.a.r d'un processus stationnaire est stationnaire.

Notons que la somme de deux processus stationnaires n'est pas forcément stationnaire.

Il existe des non-stationnarités particulières dont :

- $(X_t)_{t\in\mathbb{Z}}$ est un processus **non-stationnaire TS** (Trend Stationary) s'il peut s'écrire sous la forme :$X_t = f(t) + Y_t$, où f est une fonction déterministe et $(Y_t)_{t\in\mathbb{Z}}$ un processus stationnaire.
- $(X_t)_{t\in\mathbb{Z}}$ est un processus **non-stationnaire DS** (Difference Stationary) s'il est stationnaire après d différenciations : $\nabla^d X_t = (I-B)^d X_t$ où $BX_t = X_{t-1}$. Si $d=2$, cela signifie que le processus défini pour tout $t \in \mathbb{Z}$ par $\nabla^2 X_t = X_t - 2X_{t-1} + X_{t-2}$ est stationnaire.

Ces processus peuvent néanmoins être rendus stationnaires par une suite de transformations usuelles (désaisonnalisation, différenciation, transformation non linéaire...). Une méthodologie classique est celle de Box-Jenkins que nous étudierons par la suite et en TP.

VII.2.3 Autocovariance et autocorrélations

L'autocovariance (resp. l'autocorrélation) d'un processus stochastique est la covariance (resp. la corrélation) de ce processus avec une version décalé dans le temps de lui même. Ces fonctions sont bien définies pour un processus stationnaire. Dans la suite on considère un processus stationnaire $(X_t)_{t\in\mathbb{Z}}$.

Fonction d'autocovariance

Définition VII.10. *On appelle* ***fonction d'autocovariance*** *du processus X la fonction γ suivante :*

$$\forall h \in \mathbb{Z} : \gamma(h) = \mathbb{C}ov(X_t, X_{t-h})$$

Proposition VII.11. *La fonction d'autocovariance vérifie les propriétés suivantes :*

- $\gamma(0) \geq 0$,
- $|\gamma(h)| \leq \gamma(0)$,
- *γ est une fonction symétrique : pour tout $h \in \mathbb{N}$, $\gamma(-h) = \gamma(h)$,*
- *γ est une fonction semi-définie positive :*

$$\forall n \in \mathbb{N}^*, \forall (a_i)_{i\in\{1,\dots,n\}} \in \mathbb{R}^n : \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma(i-j) \geq 0.$$

Comme la fonction γ est symétrique, on calculera la fonction d'autocovariance pour $h \in \mathbb{N}$.

Réciproquement, si une fonction γ vérifie :

- $\gamma(-h) = \gamma(h)$, $h \in \mathbb{N}^*$,
- $\forall n \in \mathbb{N}^*, \forall (a_i)_{i\in\{1,\dots,n\}} \in \mathbb{R}^n : \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma(i-j) \geq 0$,

alors c'est une fonction d'autocovariance.

Autocorrélogramme simple

Définition VII.12. *On appelle* ***autocorrélogramme simple*** *du processus X la fonction ρ suivante :*

$$\forall h \in \mathbb{Z} : \rho(h) = \mathbb{C}orr(X_t, X_{t-h}) = \frac{\gamma(h)}{\gamma(0)}$$

L'autocorrélogramme simple est donc la fonction d'autocovariance renormalisée. Il vérifie des propriétés similaires.

Proposition VII.13. *On a :*
- $\rho(0) = 1$,
- $|\rho(h)| \leq 1$,
- ρ *est une fonction symétrique :* $\forall h \in \mathbb{N} : \rho(-h) = \rho(h)$,
- ρ *est une fonction définie positive :*

$$\forall n \geq 0, \forall (a_i)_{i\in\{1,...,n\}} : \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \rho(i-j) \geq 0.$$

Définition VII.14. *On appelle* ***matrice d'autocorrélation*** *de* $(X_t, ..., X_{t-h+1})$ *avec* $h \in \mathbb{N}^*$ *la matrice suivante :*

$$R_h = \begin{bmatrix} 1 & \rho(1) & \rho(2) & ... & \rho(h-1) \\ \rho(1) & 1 & \rho(1) & ... & ... \\ ... & ... & ... & ... & \rho(1) \\ \rho(h-1) & ... & ... & \rho(1) & 1 \end{bmatrix} \tag{VII.1}$$

Proposition VII.15. *La fonction* ρ *est une fonction définie positive si et seulement si* $\forall h \in \mathbb{N}^* : \det R_h \geq 0$.

La seconde condition fixe de nombreuses contraintes aux corrélations, par exemple :

$$\det R_3 \geq 0 \Leftrightarrow [1-\rho(2)]\left[1+\rho(2)-2\rho^2(1)\right] \geq 0.$$

Autocorrélogramme partiel

Une seconde façon de mesurer l'influence entre le processus et son décalé dans le temps est de calculer la corrélation entre deux instants en enlevant une partie de l'information contenue entre ces deux instants. Plus précisement, on calcule la corrélation entre le terme $X_t - \mathbb{EL}(X_t|X_{t-1}, ..., X_{t-h+1})$ et le terme $X_{t-h} - \mathbb{EL}(X_{t-h}|X_{t-1}, ..., X_{t-h+1})$, où la quantité $\mathbb{EL}(X_t|X_{t-1}, ..., X_{t-h+1})$ (resp. $\mathbb{EL}(X_{t-h}|X_{t-1}, ..., X_{t-h+1})$) désigne la régression linéaire de X_t (resp. X_{t-h}) sur X_{t-1}, ...,X_{t-h+1}.

On rappelle que la régression linéaire d'une variable aléatoire Z sur n variables aléatoires $Y_1, \ldots, Y_n$ est la variable aléatoire, notée $\mathbb{EL}(Z|Y_1, ..., Y_n)$, est la combinaison linéaire des $Y_1, \ldots, Y_n$ qui minimise $\bar{Z} \mapsto \mathbb{V}ar(Z - \bar{Z})$. Plus précisément, on a

$$\mathbb{EL}(Z|Y_1, ..., Y_n) = \sum_{i=1}^{n} a_i Y_i \quad \text{avec} \quad (a_1, \ldots, a_n) = \underset{(x_1,...,x_n)\in\mathbb{R}^n}{\operatorname{argmin}} \mathbb{V}ar\big(Z - \sum_{i=1}^{n} x_i Y_i\big).$$

Définition VII.16. *On appelle* ***autocorrélogramme partiel*** *du processus* X *la fonction* r *suivante :* $r(1) = \rho(1)$ *et pour tout* $h \in \mathbb{N}^*$,

$$\begin{aligned} r(h) &= \mathbb{C}orr(X_t, X_{t-h}|X_{t-1}, ..., X_{t-h+1}) \\ &= \frac{\mathbb{C}ov\left(X_t - \mathbb{EL}(X_t|X_{t-1}, ..., X_{t-h+1}), X_{t-h} - \mathbb{EL}(X_{t-h}|X_{t-1}, ..., X_{t-h+1})\right)}{\mathbb{V}ar\left(X_t - \mathbb{EL}(X_t|X_{t-1}, ..., X_{t-h+1})\right)}. \end{aligned}$$

Théorème VII.17. *Soit $(X_t)_{t\in\mathbb{Z}}$ un processus stationnaire. On considère la régression linéaire de X_t sur $X_{t-1},...,X_{t-h}$:*

$$\mathbb{EL}\left(X_t|X_{t-1},...,X_{t-h}\right)=\sum_{i=1}^{h}a_i\left(h\right)X_{t-i},$$

et $\varepsilon_t=X_t-\mathbb{EL}\left(X_t|X_{t-1},...,X_{t-h}\right)$. Alors on a

- *$\mathbb{E}\left(\varepsilon_t\right)=0$ et il existe $\sigma>0$ tel que pour tout $t\in\mathbb{Z}$, $\mathbb{E}\left(\varepsilon_t^2\right)=\sigma^2$*
- *$\forall i\in\{1,...,h\}:\mathbb{E}\left(\varepsilon_t X_{t-i}\right)=0$*

Et on a $a_h\left(h\right)=r\left(h\right)$ ainsi que

$$\begin{pmatrix}\rho\left(1\right)\\ \rho\left(2\right)\\ ...\\ \rho\left(h\right)\end{pmatrix}=R_h\begin{pmatrix}a_1\left(h\right)\\ a_2\left(h\right)\\ ...\\ a_h\left(h\right)\end{pmatrix}.$$

D'après cette proposition, si on a une estimation de $\left(\rho\left(1\right),...,\rho\left(h\right)\right)$ et donc de la matrice d'autocorrélation R_h définie par (VII.1), alors on est capable d'estimer $\left(a_1\left(h\right),...,a_h\left(h\right)\right)$ (en inversant la matrice R_h). On obtient aussi, d'après le théorème précédent, une estimation de l'autocorrélogramme partiel $r(h)$.

Une façon moins coûteuse d'inverser la matrice R_h pour déterminer les autocorrélations partielles est d'utiliser l'algorithme de Durbin-Levinson.

Algorithme VII.18. Algorithme de **Durbin-Levinson** :

$a_1\left(1\right)=\rho\left(1\right)$

$\forall h\geq 2,\forall i\in\{1,...,h-1\}:a_i\left(h\right)=a_i\left(h-1\right)-a_h\left(h\right)a_{h-i}(h-1),$

$$\forall h\geq 2:a_h\left(h\right)=\frac{\rho\left(h\right)-\sum_{i=1}^{h-1}\rho\left(h-i\right)a_i\left(h-1\right)}{1-\sum_{i=1}^{h-1}\rho\left(i\right)a_i\left(h-1\right)}.$$ ◊

VII.2.4 Estimation des moments pour les processus stationnaires

Soit $(X_t)_{t\in\mathbb{Z}}$ un processus stationnaire.

Espérance

L'estimateur naturel (sans biais) de $\mathbb{E}\left(X\right)=\mu$ à partir de $(X_1,...,X_T)$ est la moyenne empirique $\overline{X}_T$:$\overline{X}_T=\frac{1}{T}\sum_{i=1}^{T}X_i$. C'est un estimateur convergent vers μ lorsque T tend vers l'infini.

Autocovariance et autocorrélation

Si on dispose de $(X_1,...,X_T)$ alors on peut considérer les deux estimateurs suivants de la fonction d'autocorrélation $\gamma(h)$: pour tout $h\in\{1,\ldots,T-1\}$

$$\widehat{\gamma}\left(h\right)=\frac{1}{T}\sum_{t=h+1}^{T}\left(X_t-\overline{X}_T\right)\left(X_{t-h}-\overline{X}_T\right),$$

et

$$\bar{\gamma}(h) = \frac{1}{T-h} \sum_{t=h+1}^{T} \left(X_t - \overline{X}_T\right)\left(X_{t-h} - \overline{X}_T\right).$$

Ils sont tous les deux convergents et asymptotiquement sans biais mais le biais de $\bar{\gamma}(h)$ est plus petit que celui de $\widehat{\gamma}(h)$. Par contre le premier est défini positif tandis que le second ne l'est pas en général.

On en déduit l'estimateur suivant de l'autocorrélogramme simple $\rho(h)$:

$$\forall h \in \{1, ..., T-1\} : \widehat{\rho}(h) = \frac{\sum_{t=h+1}^{T} \left(X_t - \overline{X}_T\right)\left(X_{t-h} - \overline{X}_T\right)}{\sum_{t=1}^{T} \left(X_t - \overline{X}_T\right)^2}$$

Remarque VII.19. On peut noter que :
- Appliquer ces estimateurs avec $h = T-1$ pose problème.
- On peut faire les calculs même lorsque le processus n'est pas stationnaire !
- Les estimations des autocorrélations partielles se déduisent des estimations des autocorrélations simples grâce à l'algorithme de Durbin-Levinson.

$\Diamond$

VII.2.5 Tests de blancheur

Il existe différents tests de blancheur nécessaires notamment pour valider les modélisations SARIMA (Seasonal ARIMA), notamment le **test de Portmanteau** .

On considère le test suivant (de Portmanteau ou de Ljung-Box) :

$$\begin{cases} H_0 : (X_t)_{t\in\mathbb{Z}} \text{ est un bruit blanc} \\ H_1 : (X_t)_{t\in\mathbb{Z}} \text{ n'est pas un bruit blanc} \end{cases}$$

Si on dispose de $(X_1, ..., X_T)$, on considère la statistique suivante (calculée sur les k premières estimations des autocorrélations) :

$$\bar{Q}_k = T\sum_{h=1}^{k} \widehat{\rho}^2(h)$$

Une trop grande valeur de $\bar{Q}_k$ indique que les autocorrélations sont trop importantes pour être celles d'un bruit blanc. On peut également considérer la statistique (qui converge plus vite) :

$$Q_k = T(T+2)\sum_{h=1}^{k} \frac{\widehat{\rho}^2(h)}{T-h}$$

Sous H_1, Q_k diverge vers l'infini quand T tend vers l'infini. Asymptotiquement, sous H_0, Q_k suit une loi du χ^2 à k degrés de liberté. On rejette donc l'hypothèse H_0 au niveau α si :

$$Q_k > \chi^2_k(1-\alpha)$$

où $\chi^2_k(1-\alpha)$ désigne le quantile d'ordre $1-\alpha$ d'une loi du χ^2 à k degrés de liberté. En général on préfère travailler sur la p-valeur :

$$p\text{-valeur} = \mathbb{P}\left(\chi^2_k \geq Q_k\right)$$

On rejette l'hypothèse H_0 au niveau α si : p-valeur $< \alpha$.

Exemple VII.20. Le logiciel SAS permet d'effectuer le test de blancheur. La sortie suivante correspond au test de blancheur appliquée à un bruit blanc simulé :

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations				
6	7.09	6	0.2135	-0.044	0.042	-0.152	0.035	0.065	0.041
12	11.45	12	0.4907	-0.074	-0.105	0.002	-0.006	0.057	0.032
18	13.96	18	0.7320	-0.027	-0.032	0.032	0.000	-0.093	0.003
24	20.56	24	0.6643	0.022	-0.114	-0.059	0.040	-0.093	-0.044

◊

Dans cet exemple, on a choisit $k = 6, 12, 18$ et 24. On peut lire par exemple :

$$q_{18} = 13.96 \text{ où } q_{18} \text{ est la réalisation de } Q_{18},$$
$$p\text{-valeur} = 0.7320.$$

On ne rejette donc pas (fort heureusement) l'hypothèse de blancheur de cette série.

VII.2.6 Densité spectrale

Définition VII.21. *Soit $(X_t)_{t\in\mathbb{Z}}$ un processus stationnaire de fonction d'autocovariance γ.*

On appelle ***densité spectrale*** *de X la transformée de Fourier discrète, f, de la suite des autocovariances (lorsqu'elle existe) :*

$$f(\omega) = \frac{1}{2\pi}\sum_{h=-\infty}^{+\infty}\gamma(h)\,e^{-ih\omega}$$

Cette densité spectrale existe lorsque :

$$\sum_{h=-\infty}^{+\infty}|\gamma(h)| < +\infty$$

Si elle existe elle est continue, positive, paire et 2π-périodique (ce qui permet de l'étudier uniquement sur $[0,\pi]$).

Théorème VII.22 (Théorème spectral)**.** *Si f est la densité spectrale de X alors :*

$$\gamma(h) = \int_{-\pi}^{+\pi} f(\omega)\,e^{ih\omega}d\omega.$$

Exemple VII.23. Soit $(\varepsilon_t)_{t\in\mathbb{Z}}$ un bruit blanc faible de variance σ^2. On a : $\gamma(h) = \sigma^2$ si $h = 0$ et $\gamma(h) = 0$ sinon. On en déduit donc :

$$f(\omega) = \frac{\sigma^2}{2\pi}$$

Réciproquement, si la densité spectrale est constante, alors le processus correspondant est un bruit blanc faible. ◊

Définition VII.24. *Soit $(X_t)_{t\in\mathbb{Z}}$ un processus stationnaire.*
Si on dispose de $(X_1, ..., X_T)$, on appelle ***périodogramme*** *la fonction I_T suivante :*

$$I_T(\omega) = \frac{1}{T}\left|\sum_{t=1}^{T} X_t e^{-it\omega}\right|^2$$

Si le processus $(X_t)_{t\in\mathbb{Z}}$ admet une densité spectrale, alors $\frac{1}{2\pi} I_T(\omega)$ est un estimateur sans biais de la densité spectrale (mais non-convergent !). On peut résoudre le problème de convergence en lissant le périodogramme et en considérant l'estimateur suivant :

$$\hat{f}(\omega) = \frac{1}{2\pi} \sum_{|j|\leq m_T} W_T(j) I_T\left(g(T,\omega) + \frac{2\pi j}{T}\right),$$

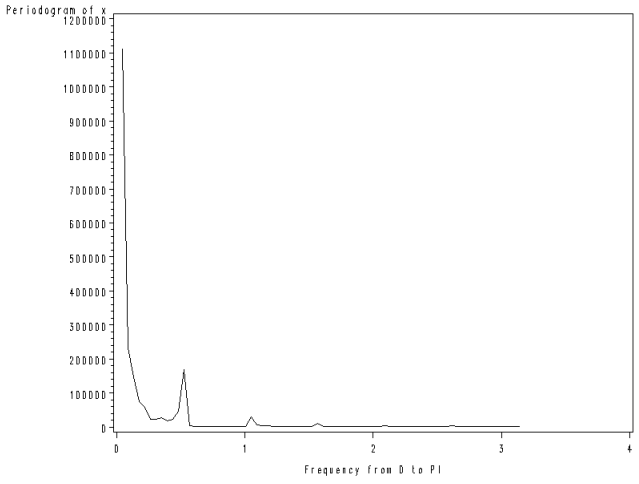
où

- $g(T,\omega)$ est le multiple de $2\pi/T$ le plus proche de ω,
- $m_T \to +\infty$ et $\frac{m_T}{T} \to 0$ quand $T \to +\infty$,
- $\forall j \in \mathbb{Z}, W_T(j) \geq 0, W_T(-j) = W_t(j)$,
- $\sum_{|j|\leq m_T} W_T(j) = 1$ et $\sum_{|j|\leq m_T} W_T^2(j) \to 0$.

Par exemple les poids constants $W_T(j) = 1/(2m_T + 1)$ conviennent, mais il existe d'autres choix avec de meilleures propriétés.

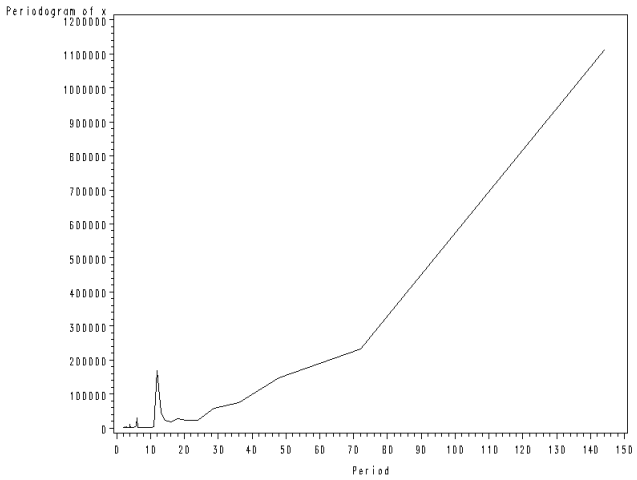
Une grande valeur du périodogramme suggère que la série a une composante saisonnière à la fréquence correspondante.

Exemple VII.25. Le graphique ci-dessous représente le périodogramme lissé du nombre de passagers aériens, dans le domaine fréquentiel :



Le graphique ci-dessous représente le périodogramme du nombre de passagers aériens, dans le domaine temporel[1] :

[1]On passe du domaine fréquentiel au domaine temporel à l'aide de la relation : $T = \frac{2\pi}{\varpi}$.



◊

VII.3 Décomposition saisonnière

Cette partie présente la décomposition saisonnière, à l'aide de la régression et des moyennes mobiles.

VII.3.1 Principe de la décomposition saisonnière

Il est courant de décomposer un processus $(X_t)_{t\in\{1,\dots,T\}}$ en différents termes :
- Une tendance : T_t
- Une composante saisonnière : S_t (de période p)
- Un résidu aléatoire : ε_t
- Et éventuellement une composante cyclique (de long terme) : C_t

Il existe différentes modèles possibles dont :
- Le modèle additif : $X_t = T_t + S_t + \varepsilon_t$, où $\mathbb{E}(\varepsilon_t) = 0$ et $\mathbb{V}(\varepsilon_t) = \sigma^2$.
- Le modèle multiplicatif : $X_t = T_t S_t (1 + \varepsilon_t)$, où $\mathbb{E}(\varepsilon_t) = 0$ et $\mathbb{V}(\varepsilon_t) = \sigma^2$.

VII.3.2 Décomposition saisonnière à l'aide de la régression linéaire

Principe

On suppose que les composantes tendantielles et saisonnières sont des combinaisons linéaires de fonctions connues dans le temps :

$$T_t = \sum_{i=1}^{n} \alpha_i T_t^i, \qquad S_t = \sum_{j=1}^{p} \beta_j S_t^j,$$

où :

$$S_t^j = \begin{cases} 1 & \text{si } t = j\ [p], \\ 0 & \text{sinon.} \end{cases}$$

Exemple VII.26. On peut considérer les tendances suivantes :

1. Tendance linéaire : $T_t = \alpha_0 + \alpha_1 t$
2. Tendance quadratique : $T_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2$
3. Tendance exponentielle : $T_t = c\alpha^t$
4. Tendance de Gomperz : $T_t = \exp\left(c_1\alpha^t + c_2\right)$

Les deux premiers cas peuvent se résoudre à l'aide de la régression linéaire. Les deux suivants se résolvent par optimisation (en minimisant par exemple l'erreur quadratique). $\Diamond$

Estimation dans le cas d'un modèle additif

Le modèle considéré est le suivant :

$$\forall t \in \{1, ..., T\} : X_t = \sum_{i=1}^{n} \alpha_i T_t^i + \sum_{j=1}^{p} \beta_j S_t^j + \varepsilon_t,$$

où $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc. On cherche à minimiser l'erreur quadratique

$$\sum_{t=1}^{T} (X_t - \sum_{i=1}^{n} \alpha_i T_t^i - \sum_{j=1}^{p} \beta_j S_t^j)^2$$

en les paramètre $\alpha_1, \ldots, \alpha_n$ et $\beta_1, \ldots, \beta_p$, où n et p sont fixés. On utilise les notations suivantes :

$$D = \begin{bmatrix} T_1^1 & ... & T_1^n \\ ... & ... & ... \\ T_T^1 & ... & T_T^n \end{bmatrix} \qquad S = \begin{bmatrix} S_1^1 & ... & S_1^p \\ ... & ... & ... \\ S_T^1 & ... & S_T^p \end{bmatrix}.$$

Avec ces notations, le système s'écrit sous la forme vectorielle suivante :

$$X = D\alpha + S\beta + \varepsilon = Yb + \varepsilon,$$

où :

$$X = \begin{pmatrix} X_1 \\ ... \\ X_T \end{pmatrix}; \varepsilon = \begin{pmatrix} \varepsilon_1 \\ ... \\ \varepsilon_T \end{pmatrix}; \alpha = \begin{pmatrix} \alpha_1 \\ ... \\ \alpha_n \end{pmatrix}; \beta = \begin{pmatrix} \beta_1 \\ ... \\ \beta_p \end{pmatrix}; Y = [D, S]; b = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

Un estimateur de b dans $X = Yb + \varepsilon$ par les moindres carrés ordinaires (MCO) est :

$$\widehat{b} = \left(Y'Y\right)^{-1} Y'X.$$

Ainsi on obtient les estimateurs de α et β :

$$\begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix} = \begin{bmatrix} D'D & D'S \\ S'D & S'S \end{bmatrix}^{-1} \begin{pmatrix} D'X \\ S'X \end{pmatrix}$$

De plus, ce sont des estimateurs sans biais

$$\mathbb{E}\begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix},$$

et de variance

$$\mathbb{V}\begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix} = s^2 \begin{bmatrix} D'D & D'S \\ S'D & S'S \end{bmatrix}^{-1}$$

où $s^2 = \dfrac{1}{T-n-p}\sum_{t=1}^{T}\varepsilon_t^2$ est un estimateur convergent (lorsque T tend vers l'infini) et sans biais de la variance de ε_t.

Résultats

Une fois les estimations effectuées, on peut obtenir la série ajustée $\widehat{X}_t$ et la série corrigée des variations saisonnières X_t^{CVS} :

- $\widehat{X}_t = \sum_{i=1}^{n}\widehat{\alpha}_i T_t^i + \sum_{j=1}^{p}\widehat{\beta}_j S_t^j$
- $X_t^{CVS} = X_t - \widehat{X}_t$

Prévision

Si on désire prévoir X_{T+h}, on peut supposer que le modèle est toujours valable à cet horizon, ce qui donne :

$$X_{T+h} = \sum_{i=1}^{n}\alpha_i T_{T+h}^i + \sum_{j=1}^{p}\beta_j S_{T+h}^j + \varepsilon_{T+h}.$$

La meilleure prévision au sens de l'erreur quadratique moyenne est l'espérance de X_{T+h} notée $X_T(h)$ avec

$$X_T(h) = \mathbb{E}(X_{T+h}) = \sum_{i=1}^{n}\alpha_i T_{T+h}^i + \sum_{j=1}^{p}\beta_j S_{T+h}^j,$$

dont l'estimateur $\widehat{X}_T(h)$ est

$$\widehat{X}_T(h) = \widehat{X}_{T+h} = \sum_{i=1}^{n}\widehat{\alpha}_i T_{T+h}^i + \sum_{j=1}^{p}\widehat{\beta}_j S_{T+h}^j.$$

La notation $\widehat{X}_T(h)$ est utilisé pour signifier que l'on prédit à l'horizon h connaissant les données jusqu'à T. On peut également estimer des intervalles de confiance (asymptotiques).

Exemple

Considérons le nombre de passagers aériens. La croissance du nombre de passagers aérien est exponentielle. On effectue donc une transformation logarithmique sur les données. Et on considère le modèle suivant :

$$\forall t \in \{1, ..., 144\} : X_t = at + b + \sum_{j=1}^{12} c_j S_t^j + \varepsilon_t$$

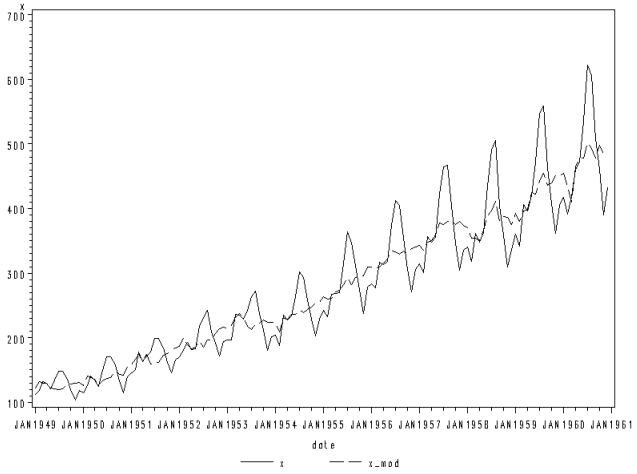
où on a tenu compte d'une saisonalité annuelle :

$$S_t^j = \begin{cases} 1 & \text{si } t = j\,[12], \\ 0 & \text{sinon.} \end{cases}$$

Ce modèle est l'analogue du premier modèle historique de Buys-Ballot (1847).

Comme $\sum_{j=1}^{12} S_t^j = 1$, le modèle indéterminé. Pour résoudre ce problème on peut soit poser $c_{12} = 0$ soit contraindre $\sum_{j=1}^{12} c_j = 0$.

Le graphique suivant représente les résultats obtenus : série brute et série corrigée des variations saisonnières (on a effectué une transformation logarithmique aux données avant traitement, puis une transformation exponentielle aprés traitement).



Remarque VII.27. Il est aussi possible d'utiliser des moyennes mobiles (ou glissantes) pour effectuer les décompositions saisonnières. Une moyenne mobile est une combinaison linéaire d'opérateurs retard :

$$M = \sum_{i=-m_1}^{m_2} \theta_i B^{-i}$$

où $(m_1, m_2) \in \mathbb{N}^2$, $(\theta_{-m_1}, \theta_{m_2}) \in \mathbb{R}^{*2}$ et B est l'opérateur retard : $BX_t = X_{t-1}$ et plus généralement $B^i X_t = X_{t-i}$. Un algorithme classique de desaisonnalisation utilisant les moyennes mobiles est l'algorithme $X11$ mis au point par le "Bureau of Census". ◊

VII.4 Prévision et processus des innovations

On considère un processus stationnaire $(X_t)_{t\in\mathbb{Z}}$. Soit t fixé. La prévision linéaire optimale de X_t sachant son passé est définie comme la régression linéaire de X_t sur l'espace fermé engendré par les combinaisons linéaires des $(X_i)_{i\leq t-1}$. On la note $\widehat{X}_t$,

$$\widehat{X}_t = c_0 + \sum_{i=-\infty}^{t-1} a_i X_i, \tag{VII.2}$$

où $(c_0,(a_i)_{1\leq i\leq t-1}) = \text{argmin}_{c'_0,(a'_i)_{1\leq i\leq t-1}} \mathbb{E}[(X_t - c'_0 + \sum_{i=-\infty}^{t-1} a'_i X_i)^2]$. Les erreurs de prévision successives $\varepsilon_t = X_t - \widehat{X}_t$ forment un processus appelé **processus des innovations**.

Proposition VII.28. *Le processus des innovations, ou plus simplement innovations, est un bruit blanc.*

Remarque VII.29. On utilise souvent des algorithmes itératifs afin de déterminer les prévisions basées sur $(X_1,...,X_{T+1})$, à partir de celles basées sur $(X_1,...,X_T)$. L'algorithme de Durbin-Levinson et l'algorithme des innovations sont couramment utilisés. ◊

VII.5 Étude des processus AR

Cette partie ainsi que les suivantes présentent les modèles proposés par Box et Jenkins. Ils sont utilisés pour traiter de nombreuses séries temporelles.

Sauf mention contraire, on considère dans ce chapitre des processus centrés. Si le processus $(X_t)_{t\in\mathbb{Z}}$ n'est pas centré, on obtient les mêmes résultats mais sur le processus $(Y_t)_{t\in\mathbb{Z}}$ tel que $Y_t = X_t - \mu_X$.

VII.5.1 Définition

Définition VII.30. *Soit $(\varepsilon_t)_{t\in\mathbb{Z}}$ un bruit blanc (faible) de variance σ^2 et $p \geq 1$.*

*On dit qu'un processus $(X_t)_{t\in\mathbb{Z}}$ est un processus autorégressif ou encore **processus AR** (AutoRegressive) d'ordre p, noté $AR(p)$, si :*

- *$(X_t)_{t\in\mathbb{Z}}$ est stationnaire,*
- *$\forall t \in \mathbb{Z} : X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$ où $(\varphi_1,...,\varphi_p) \in \mathbb{R}^p$ sont des constantes et $\varphi_p \neq 0$.*

On utilise généralement la notation suivante :

$$\Phi(B) X_t = \varepsilon_t, \tag{VII.3}$$

où $\Phi(B) = I - \sum_{i=1}^{p} \varphi_i B^i$.

Exemple VII.31. Soit $(\varepsilon_t)_{t\in\mathbb{Z}}$ un bruit blanc faible de variance σ^2. Soit le processus $(X_t)_{t\in\mathbb{Z}}$ suivant :

$$X_t = X_{t-1} + \varepsilon_t.$$

On a, pour $h \in \mathbb{N}^*$: $X_t - X_{t-h} = \varepsilon_t + ... + \varepsilon_{t-h+1}$, d'où :

$$\mathbb{E}\left[(X_t - X_{t-h})^2\right] = h\sigma^2$$

Si le processus $(X_t)_{t\in\mathbb{Z}}$ était stationnaire, on aurait :

$$\begin{aligned}
\mathbb{E}\left[(X_t - X_{t-h})^2\right] &= \mathbb{E}\left(X_t^2\right) + \mathbb{E}\left(X_{t-h}^2\right) - 2\mathbb{E}\left(X_t X_{t-h}\right) \\
&= \mathbb{E}\left(X_t^2\right) + \mathbb{E}\left(X_{t-h}^2\right) + \mathbb{E}\left(X_t^2\right) + \mathbb{E}\left(X_{t-h}^2\right) - \mathbb{E}\left[(X_t + X_{t-h})^2\right] \\
&= 4\mathbb{V}(X_t) - \mathbb{E}\left[(X_t + X_{t-h})^2\right] \\
&\leq 4\mathbb{V}(X_t).
\end{aligned}$$

Or il est impossible d'avoir, pour tout $h \in \mathbb{N}^* : h\sigma^2 \leq 4\mathbb{V}(X_t)$. Le processus $(X_t)_{t\in\mathbb{Z}}$ n'est donc pas stationnaire. $\Diamond$

Dans l'exemple précédent, on a $\Phi(x) = 1 - x$, et Φ possède une racine (ici égale à 1) de module égal à 1. Plus généralement, on peut montrer que si $\Phi(B)$ a une racine de module égal à 1 (sur le cercle unité) alors le processus $(X_t)_{t\in\mathbb{Z}}$ n'est pas stationnaire, et il ne peut donc pas s'agir d'un processus AR.

VII.5.2 Processus AR canonique et écriture $MA(\infty)$

Définition VII.32. *Soit $(X_t)_{t\in\mathbb{Z}}$ un processus $AR(p)$ associé à Φ par VII.3. Si les racines de Φ sont toutes à l'extérieur du cercle unité alors on a sa représentation* ***canonique*** *.*

Proposition VII.33. *Si X est sous sa forme canonique, alors le bruit blanc associé est le processus des innovations : $\varepsilon_t = X_t - \widehat{X}_t$ où $\widehat{X}_t$ est donné par (VII.2).*

Si le polynôme $\Phi(B)$, que l'on suppose inversible, a des racines à l'intérieur du cercle unité, alors le processus $(\varepsilon_t)_{t\in\mathbb{Z}}$ n'est pas le processus des innovations. Pour se ramener à une écriture canonique (c'est à dire une écriture où les résidus sont l'innovation) on transforme l'opérateur $\Phi(B)$.

Soient $z_i = \dfrac{1}{\lambda_i}$, $i \in \{1, ..., p\}$, les racines de $\Phi(z)$.

Supposons que les $(z_i)_{i\in\{1,...,r\}}$ soient toutes de module inférieur à 1 (à l'extérieur du cercle unité) et que les $(z_i)_{i\in\{r+1,...,p\}}$ soient toutes de module supérieur à 1 (à l'intérieur du cercle unité). On peut écrire :

$$\Phi(B) = \prod_{i=1}^{p} (I - \lambda_i B)$$

On construit alors un nouveau polynôme $\Phi^*(B)$ qui a toutes ses racines à l'extérieur du cercle unité :

$$\Phi^*(B) = \prod_{i=1}^{r} \left(I - \frac{1}{\lambda_i} B\right) \prod_{i=r+1}^{p} (I - \lambda_i B),$$

et on obtient l'écriture canonique du processus.

Proposition VII.34. *On définit le processus $(\eta_t)_{t\in\mathbb{Z}}$ par $\eta_t = \Phi^*(B)X_t$. Alors $(\eta_t)_{t\in\mathbb{Z}}$ est un bruit blanc faible et c'est le processus des innovations du processus AR $(X_t)_{t\in\mathbb{Z}}$.*

On préfère travailler avec la représentation canonique car la représentation inverse ne fera apparaître que les instants passés du processus. Dans la suite, on considère des processus AR canoniques.

Proposition VII.35. *Soit $(\varepsilon_t)_{t\in\mathbb{Z}}$ un bruit blanc faible de variance σ^2.*

Soit $(X_t)_{t\in\mathbb{Z}}$ un processus $AR(p)$ canonique : $\Phi(B) X_t = \varepsilon_t$. Il admet alors une ***écriture*** *$MA(\infty)$:*

$$X_t = \Phi^{-1}(B)\, \varepsilon_t = \varepsilon_t + \sum_{i=1}^{+\infty} \psi_i \varepsilon_{t-i},$$

où $(\psi_i)_{i\in\mathbb{N}}$ est une suite réelle.

VII.5.3 Autocorrélations simples d'un processus AR

Soit $(\varepsilon_t)_{t\in\mathbb{Z}}$ un bruit blanc faible de variance σ^2 et $(X_t)_{t\in\mathbb{Z}}$ un processus $AR(p)$ canonique : $X_t = \sum_{i=1}^{p}\varphi_i X_{t-i} + \varepsilon_t$ pour $t\in\mathbb{Z}$ et $\varphi = (\varphi_1,\ldots,\varphi_p)\in\mathbb{R}^p$.

Lemma VII.36. *On a* $\gamma(0) = \dfrac{\sigma^2}{1-\sum_{i=1}^{p}\varphi_i\rho(i)}$ *et* $\rho(h) = \sum_{i=1}^{p}\varphi_i\rho(h-i)$ *pour tout* $h\in\mathbb{N}^*$.

Démonstration. On a :

$$\begin{aligned}
\gamma(0) &= \mathbb{C}ov(X_t, X_t)\\
&= \mathbb{C}ov\left(X_t, \sum_{i=1}^{p}\varphi_i X_{t-i} + \varepsilon_t\right)\\
&= \sum_{i=1}^{p}\varphi_i \mathbb{C}ov(X_t, X_{t-i}) + \mathbb{C}ov(X_t, \varepsilon_t)\\
&= \sum_{i=1}^{p}\varphi_i\gamma(i) + \mathbb{C}ov\left(\sum_{i=1}^{p}\varphi_i X_{t-i} + \varepsilon_t, \varepsilon_t\right)\\
&= \sum_{i=1}^{p}\varphi_i\gamma(i) + \mathbb{C}ov(\varepsilon_t, \varepsilon_t) \text{ car } \varepsilon_t \perp \mathcal{H}_{-\infty}^{t-1}(X)\\
&= \sum_{i=1}^{p}\varphi_i\gamma(0)\rho(i) + \sigma^2.
\end{aligned}$$

On a de même :

$$\begin{aligned}
\forall h\in\mathbb{N}^* : \gamma(h) &= \mathbb{C}ov(X_t, X_{t-h})\\
&= \mathbb{C}ov\left(\sum_{i=1}^{p}\varphi_i X_{t-i} + \varepsilon_t, X_{t-h}\right)\\
&= \sum_{i=1}^{p}\varphi_i \mathbb{C}ov(X_{t-i}, X_{t-h}) + \mathbb{C}ov(\varepsilon_t, X_{t-h})\\
&= \sum_{i=1}^{p}\varphi_i\gamma(h-i) \text{ car } \varepsilon_t \perp \mathcal{H}_{-\infty}^{t-1}(X).
\end{aligned}$$

□

On obtient finalement le système suivant, appelé **<u>équations de Yule-Walker</u>** :

$$\begin{cases}
\gamma(0) = \dfrac{\sigma^2}{1-\sum_{i=1}^{p}\varphi_i\rho(i)}\\
\begin{pmatrix}\rho(1)\\ \rho(2)\\ \ldots\\ \rho(p)\end{pmatrix} = \begin{bmatrix}1 & \rho(1) & \ldots & \rho(p-1)\\ \rho(1) & 1 & \ldots & \ldots\\ \ldots & \ldots & \ldots & \rho(1)\ldots\\ \rho(p-1) & \ldots & \rho(1) & 1\end{bmatrix}\begin{pmatrix}\varphi_1\\ \varphi_2\\ \ldots\\ \varphi_p\end{pmatrix} = R_p\varphi.
\end{cases}$$

Les autocorrélations simples sont solution d'une équation de récurrence linéaire simple d'ordre p.

Remarque VII.37. Si les racines de $\Phi(z)$: $z_i = \dfrac{1}{\lambda_i}$, $i \in \{1, ..., p\}$, sont réelles et distinctes alors on obtient :

$$\rho(h) = \sum_{i=1}^{p} c_i \lambda_i^h$$

Comme les racines sont de modules strictement plus grand que 1, les autocorrélations simples décroissent exponentiellement vites des vers 0.

Dans le cas général, on obtient une décroissance des autocorrélations simples vers 0, de type exponentiel ou sinusoïdal amorti. ◊

VII.5.4 Autocorrélations partielle d'un processus AR

Si $(X_t)_{t\in\mathbb{Z}}$ est un processus $AR(p)$ alors ses autocorrélations partielles s'annulent à partir du rang $p+1$: $r(p) \neq$ et pour $h \geq p+1$ on a $r(h) = 0$.

Réciproquement, il s'agit d'une condition nécessaire et suffisante pour qu'un processus $(X_t)_{t\in\mathbb{Z}}$ soit un $AR(p)$.

De plus, si le processus est sous sa forme canonique, on a $r(p) = \varphi_p$. On ne peut rien dire a priori sur $r(h)$ pour $h \in \{1, ..., p-1\}$ (on sait simplement que $r(1) = \rho(1)$, ce qui est vrai pour tous les processus stationnaires).

VII.5.5 Exemples

On présente dans les figures VII.1, VII.2 et VII.3 les autocorrélogrammes de trois processus $AR(1)$ pour lesquels on a simulé 5000 valeurs.

Mean of Working Series -0.02172
Standard Deviation 1.048352
Number of Observations 5000

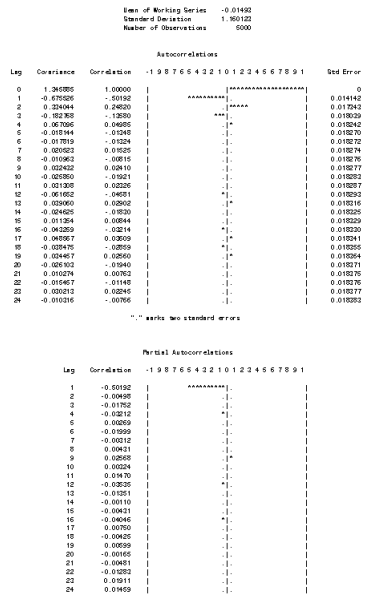
Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	1.099044	1.00000	|********************|	0
1	0.318711	0.28999	|******|	0.014142
2	0.074232	0.06753	|^ |	0.015285
3	-0.010826	-.00985	| |	0.015345
4	-0.020759	-.01892	| |	0.015346
5	-0.000106	-.00028	| |	0.015351
6	-0.018714	-.01703	| |	0.015352
7	0.011538	0.01050	| |	0.015356
8	0.022691	0.02065	| |	0.015357
9	0.027411	0.02494	| |	0.015363
10	0.0020121	0.00274	| |	0.015371
11	-0.015314	-.01448	| |	0.015371
12	-0.046167	-.04200	\^|	0.015372
13	-0.012284	-.01118	| |	0.015396
14	-0.016770	-.01526	| |	0.015398
15	-0.030021	-.02732	\^|	0.015401
16	-0.027478	-.02500	\^|	0.015411
17	0.010286	0.00946	| |	0.015419
18	-0.0012580	-.00114	| |	0.015420
19	0.0069716	0.00634	| |	0.015420
20	-0.010231	-.00935	| |	0.015421
21	-0.0087720	-.00798	| |	0.015422
22	0.0044624	0.00407	| |	0.015422
23	0.021897	0.01992	| |	0.015423
24	-0.007717	-.00707	| |	0.015428

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.28999	|******|
2	-0.01797	| |
3	-0.02657	\^|
4	-0.00825	| |
5	0.00135	| |
6	-0.01621	| |
7	0.00394	| |
8	0.01269	| |
9	0.01452	| |
10	-0.01019	| |
11	-0.01349	| |
12	-0.03564	\^|
13	0.01272	| |
14	-0.01507	| |
15	-0.02205	| |
16	-0.01200	| |
17	0.00240	| |
18	-0.01203	| |
19	0.00852	| |
20	-0.01318	| |
21	-0.00119	| |
22	0.00795	| |
23	0.01978	| |
24	-0.02200	| |

FIG. VII.1 – $X_t = 0.3X_{t-1} + \varepsilon_t$



Mean of Working Series -0.01492
Standard Deviation 1.160123
Number of Observations 5000

Autocorrelations

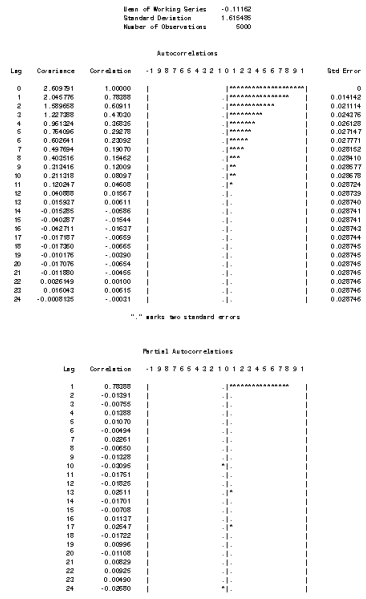
Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	1.345826	1.00000	|********************|	0
1	-0.675526	-.50192	|**********| |	0.014142
2	0.324044	0.24520	| |*****| |	0.017343
3	-0.182768	-.13580	| |***| |	0.018029
4	0.067095	0.04985	| |* | |	0.018242
5	-0.018144	-.01348	| |. | |	0.018270
6	-0.017819	-.01324	| |. | |	0.018272
7	0.000423	0.00315	| |. | |	0.018274
8	-0.010963	-.00815	| |. | |	0.018276
9	0.022432	0.02410	| |. | |	0.018277
10	-0.025850	-.01921	| |. | |	0.018283
11	0.021306	0.02326	| |. | |	0.018287
12	-0.061652	-.04581	| |*| |	0.018292
13	0.039060	0.02902	| |* | |	0.018316
14	-0.024625	-.01830	| |. | |	0.018325
15	0.011354	0.00844	| |. | |	0.018329
16	-0.043259	-.03214	| |*| |	0.018330
17	0.045957	0.03009	| |* | |	0.018341
18	-0.025475	-.02859	| |*| |	0.018355
19	0.034457	0.02560	| |* | |	0.018364
20	-0.026163	-.01940	| |. | |	0.018371
21	0.010274	0.00762	| |. | |	0.018375
22	-0.016457	-.01148	| |. | |	0.018376
23	0.009213	0.02845	| |. | |	0.018377
24	-0.010216	-.00766	| |. | |	0.018383

"." marks two standard errors

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.50192	| **********|. |
2	-0.00498	| |. |
3	-0.01752	| |. |
4	-0.03212	| |*| |
5	0.00269	| |. |
6	-0.01999	| |. |
7	-0.00312	| |. |
8	0.00431	| |. |
9	0.02648	| |* |
10	0.00224	| |. |
11	0.01470	| |. |
12	-0.03535	| |*| |
13	-0.01351	| |. |
14	-0.00110	| |. |
15	-0.00431	| |. |
16	-0.04046	| |*| |
17	0.00750	| |. |
18	-0.00425	| |. |
19	0.00699	| |. |
20	-0.00165	| |. |
21	-0.00481	| |. |
22	-0.01252	| |. |
23	0.01911	| |. |
24	0.01469	| |. |

FIG. VII.2 – $X_t = -0.5X_{t-1} + \varepsilon_t$



Mean of Working Series -0.11162
Standard Deviation 1.615485
Number of Observations 5000

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	2.609791	1.00000	|********************|	0
1	2.045776	0.78388	| |****************| |	0.014142
2	1.589658	0.60911	| |************| |	0.021114
3	1.227288	0.47030	| |*********| |	0.024076
4	0.961284	0.36835	| |*******| |	0.026128
5	0.764096	0.29278	| |******| |	0.027147
6	0.602941	0.23098	| |*****| |	0.027771
7	0.497694	0.19070	| |****| |	0.028152
8	0.402516	0.15462	| |***| |	0.028410
9	0.312416	0.12009	| |**| |	0.028577
10	0.211218	0.08097	| |**| |	0.028678
11	0.122917	0.04708	| |*| |	0.028724
12	0.048888	0.01867	| |. | |	0.028739
13	0.015327	0.00611	| |. | |	0.028740
14	-0.015285	-.00595	| |. | |	0.028741
15	-0.040287	-.01544	| |. | |	0.028741
16	-0.042711	-.01637	| |. | |	0.028742
17	-0.017857	-.00659	| |. | |	0.028744
18	-0.017360	-.00665	| |. | |	0.028745
19	-0.010176	-.00290	| |. | |	0.028745
20	-0.017076	-.00654	| |. | |	0.028745
21	-0.011880	-.00455	| |. | |	0.028745
22	0.002649	0.00100	| |. | |	0.028746
23	0.016043	0.00615	| |. | |	0.028746
24	-0.0008125	-.00031	| |. | |	0.028746

"." marks two standard errors

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.78388	| |****************| |
2	-0.01251	| |. | |
3	-0.00755	| |. | |
4	0.01288	| |. | |
5	0.01070	| |. | |
6	-0.00494	| |. | |
7	0.02261	| |. | |
8	-0.00650	| |. | |
9	-0.01228	| |. | |
10	-0.03095	| |*| |
11	-0.01751	| |. | |
12	-0.01525	| |. | |
13	0.02511	| |* | |
14	-0.01701	| |. | |
15	-0.00708	| |. | |
16	0.01137	| |. | |
17	0.02547	| |* | |
18	-0.01722	| |. | |
19	0.00996	| |. | |
20	-0.01108	| |. | |
21	0.00829	| |. | |
22	0.00925	| |. | |
23	0.00490	| |. | |
24	-0.02550	| |*| |

FIG. VII.3 – $X_t = 0.8X_{t-1} + \varepsilon_t$

VII.6 Processus MA

Définition VII.38. *Soit $(\varepsilon_t)_{t\in\mathbb{Z}}$ un bruit blanc faible de variance σ^2.*

On dit qu'un processus $(X_t)_{t\in\mathbb{Z}}$ est un processus ***processus MA*** *(Moving Average) ou*

encore à moyenne mobile d'ordre q, noté $MA(q)$, si :

$$\forall t \in \mathbb{Z} : X_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

où $(\theta_1, ..., \theta_q) \in \mathbb{R}^q$ et $\theta_q \neq 0$.

On utilise généralement la notation suivante : $X_t = \Theta(B)\,\varepsilon_t$, où $\Theta(B) = I + \sum_{i=1}^{q} \theta_i B^i$.

Un processus MA est toujours stationnaire.

VII.6.1 Processus MA canonique et écriture $AR(\infty)$

Si les racines de Θ sont toutes à l'extérieur du cercle unité alors on a sa représentation **canonique**. Le bruit blanc associé est alors l'innovation.

On préfère travailler avec la représentation canonique car la représentation inverse ne fera apparaître que les instants passés du bruit blanc. On peut passer d'une représentation non canonique à une représentation canonique en inversant les racines qui sont à l'intérieur du cercle unité (cf. la transformation similaire effectuée sur les processus AR). Dans la suite, on considère des processus MA canoniques.

Proposition VII.39. *Soit $(X_t)_{t\in\mathbb{Z}}$ un processus $MA(q)$. Si X est sous forme canonique, alors il admet alors une **écriture** $AR(\infty)$:*

$$\varepsilon_t = \Theta^{-1}(B)\, X_t = X_t + \sum_{i=1}^{+\infty} \pi_i X_{t-i} \quad \textit{c'est-à-dire} \quad X_t = -\sum_{i=1}^{+\infty} \pi_i X_{t-i} + \varepsilon_t,$$

où $(\pi_i)_{i\in\mathbb{N}}$ est une suite réelle.

VII.6.2 Autocorrélations simples d'un processus MA

Soit $(X_t)_{t\in\mathbb{Z}}$ un processus $MA(q)$ canonique : pour tout $t \in \mathbb{Z}$, $X_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$.

Comme $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc, on a :

$$\gamma(0) = \mathbb{V}\left(\varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}\right) = \sigma^2 \left(1 + \sum_{i=1}^{q} \theta_i^2\right).$$

On a de même :

$$\begin{aligned}
\forall h \in \mathbb{N}^* : \gamma(h) &= \mathbb{C}ov\,(X_t, X_{t-h}) \\
&= \mathbb{C}ov\left(\varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}, \varepsilon_{t-h} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-h-j}\right) \\
&= \begin{cases} \theta_h + \sum\limits_{i=h+1}^{q} \theta_i \theta_{i-h} & \text{si } h \in \{1, ..., q\}, \\ 0 & \text{si } h > q. \end{cases}
\end{aligned}$$

Proposition VII.40. *Si $(X_t)_{t\in\mathbb{Z}}$ est un processus $MA(q)$ alors ses autocorrélations simples s'annulent à partir du rang $q+1$:*

$$\begin{cases} \rho(q) \neq 0, \\ \forall h \geq q+1 : \rho(h) = 0. \end{cases}$$

Il s'agit d'une condition nécessaire et suffisante pour qu'un processus $(X_t)_{t\in\mathbb{Z}}$ soit un $MA(q)$.

VII.6.3 Autocorrélations partielles d'un processus MA

Les autocorrélations partielles sont solution d'une équation de récurrence linéaire simple d'ordre q. Elles décroissent vers 0 de manière exponentielle ou sinusoïdale amortie.

Exemples

On présente dans les figures VII.4, VII.5 et VII.6 les autocorrélogrammes de trois processus $MA(1)$ pour lesquels on a simulé 5000 valeurs.

Mean of Working Series -0.01587
Standard Deviation 1.048805
Number of Observations 4999

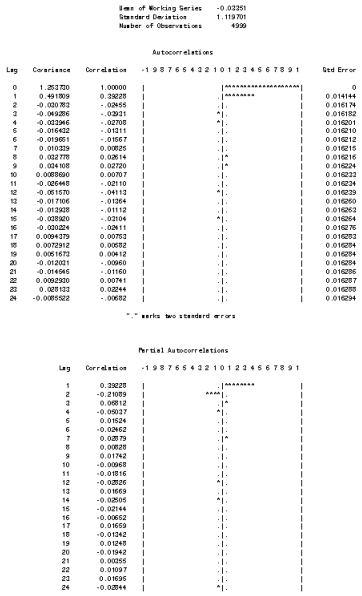
Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	1.099991	1.00000	||********************|	0
1	-0.302542	-.27595	******|. |	0.014144
2	-0.0021597	-.00257	.|. |	0.015152
3	-0.002041	-.00204	.|. |	0.015183
4	-0.010257	-.00935	.|. |	0.015183
5	0.012511	0.01228	.|. |	0.015189
6	-0.026945	-.02460	.|. |	0.015191
7	0.014535	0.01248	.|. |	0.015199
8	0.0026197	0.00229	.|. |	0.015201
9	0.020580	0.01871	.|. |	0.015201
10	-0.0066347	-.00612	.|. |	0.015206
11	0.0077299	0.00703	.|. |	0.015206
12	-0.045764	-.04159	*|. |	0.015207
13	0.019942	0.01813	.|. |	0.015220
14	-0.0016530	-.00151	.|. |	0.015234
15	-0.012416	-.01129	.|. |	0.015234
16	-0.026827	-.02439	.|. |	0.015236
17	0.021028	0.02321	.|* |	0.015244
18	-0.016464	-.01497	.|. |	0.015254
19	0.017008	0.01546	.|. |	0.015257
20	-0.012521	-.01139	.|. |	0.015260
21	-0.0056171	-.00502	.|. |	0.015262
22	-0.0065691	-.00615	.|. |	0.015262
23	0.025963	0.02633	.|* |	0.015262
24	-0.0064120	-.00583	.|. |	0.015271

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.27595	******|. |
2	-0.08562	**|. |
3	-0.04852	*|. |
4	-0.02182	*|. |
5	-0.00174	.|. |
6	-0.02666	*|. |
7	-0.00120	.|. |
8	0.00610	.|. |
9	0.02287	.|. |
10	0.00786	.|. |
11	0.01170	.|. |
12	-0.03572	*|. |
13	-0.00286	.|. |
14	-0.00109	.|. |
15	-0.01363	.|. |
16	-0.02612	*|. |
17	0.01074	.|. |
18	-0.01044	.|. |
19	0.01016	.|. |
20	-0.00469	.|. |
21	-0.00779	.|. |
22	-0.01068	.|. |
23	0.02564	.|* |
24	0.00715	.|. |

FIG. VII.4 – $X_t = \varepsilon_t - 0.3\varepsilon_{t-1}$



Mean of Working Series	-0.02351
Standard Deviation	1.119701
Number of Observations	4999

Autocorrelations

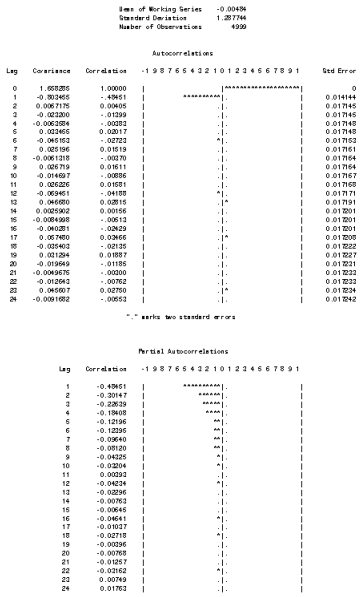
Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	1.253720	1.00000	||********************|	0
1	0.491809	0.39228	|******** |	0.014144
2	-0.020783	-.02455	|. |	0.016174
3	-0.069286	-.05931	*|. |	0.016182
4	-0.022945	-.02708	*|. |	0.016201
5	-0.016422	-.01211	|. |	0.016210
6	-0.019951	-.01557	|. |	0.016212
7	0.010239	0.00825	|. |	0.016215
8	0.022778	0.02614	|. |	0.016216
9	0.024106	0.02720	|. |	0.016224
10	0.008590	0.00707	|. |	0.016232
11	-0.026448	-.02110	|. |	0.016234
12	-0.051570	-.04113	*|. |	0.016239
13	-0.017106	-.01264	|. |	0.016260
14	-0.013928	-.01112	|. |	0.016262
15	-0.028920	-.02104	*|. |	0.016264
16	-0.020024	-.02411	|. |	0.016276
17	0.0094379	0.00752	|. |	0.016283
18	0.0072912	0.00582	|. |	0.016284
19	0.0051673	0.00412	|. |	0.016284
20	-0.012021	-.00960	|. |	0.016284
21	-0.014945	-.01160	|. |	0.016286
22	0.0092920	0.00741	|. |	0.016287
23	0.025103	0.02244	|. |	0.016288
24	-0.0085522	-.00682	|. |	0.016294

"." marks two standard errors

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.39228	|******** |
2	-0.21089	****|. |
3	0.06512	|* |
4	-0.09317	*|. |
5	0.01524	|. |
6	-0.02462	|. |
7	0.02779	|* |
8	0.00628	|. |
9	0.01742	|. |
10	-0.00968	|. |
11	-0.01816	|. |
12	-0.02526	*|. |
13	0.01669	|. |
14	-0.02505	*|. |
15	-0.02144	|. |
16	-0.00652	|. |
17	0.01659	|. |
18	-0.01242	|. |
19	0.01248	|. |
20	-0.01942	|. |
21	0.00255	|. |
22	0.01937	|. |
23	0.01695	|. |
24	-0.02744	*|. |

FIG. VII.5 – $X_t = \varepsilon_t + 0.5\varepsilon_{t-1}$



Mean of Working Series	-0.00436
Standard Deviation	1.257744
Number of Observations	4999

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	1.582285	1.00000	||********************|	0
1	-0.803465	-.49451	|*********|. |	0.014144
2	0.0067175	0.00405	|. |	0.017145
3	-0.023300	-.01399	|. |	0.017145
4	-0.0063524	-.00382	|. |	0.017148
5	0.032466	0.02017	|. |	0.017148
6	-0.045153	-.02723	*|. |	0.017153
7	0.025196	0.01519	|. |	0.017161
8	-0.0061315	-.00370	|. |	0.017164
9	0.026719	0.01611	|. |	0.017164
10	-0.014697	-.00886	|. |	0.017167
11	0.026226	0.01581	|. |	0.017169
12	-0.069461	-.04188	*|. |	0.017171
13	0.046820	0.02815	|* |	0.017191
14	0.0025902	0.00156	|. |	0.017201
15	-0.0084998	-.00513	|. |	0.017201
16	-0.060281	-.02629	|. |	0.017201
17	0.057450	0.02466	|* |	0.017208
18	-0.025403	-.02135	|. |	0.017222
19	0.021294	0.01837	|. |	0.017227
20	-0.019549	-.01185	|. |	0.017231
21	-0.0049675	-.00200	|. |	0.017233
22	-0.012643	-.00762	|. |	0.017233
23	0.045607	0.02750	|* |	0.017234
24	-0.0091682	-.00553	|. |	0.017242

"." marks two standard errors

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.49451	*********|. |
2	-0.20147	****|. |
3	-0.22629	****|. |
4	-0.19408	****|. |
5	-0.12196	*|. |
6	-0.12295	*|. |
7	-0.09640	*|. |
8	-0.08120	*|. |
9	-0.04225	*|. |
10	-0.02894	*|. |
11	0.00232	|. |
12	-0.04234	*|. |
13	-0.02296	|. |
14	-0.00763	|. |
15	-0.00645	|. |
16	-0.04641	*|. |
17	-0.01937	|. |
18	-0.02718	*|. |
19	-0.00296	|. |
20	-0.00768	|. |
21	-0.01257	|. |
22	-0.03162	*|. |
23	0.00749	|. |
24	0.01763	|. |

FIG. VII.6 – $X_t = \varepsilon_t - 0.8\varepsilon_{t-1}$

VII.7 Processus ARMA

Soit $(\varepsilon_t)_{t\in\mathbb{Z}}$ un bruit blanc faible de variance σ^2.

Définition VII.41. *On dit qu'un processus $(X_t)_{t\in\mathbb{Z}}$ est un* ***processus ARMA*** *(AutoRegressive Moving Average) d'ordre (p,q), noté $ARMA(p,q)$, si :*

- *$(X_t)_{t\in\mathbb{Z}}$ est stationnaire.*
- *il existe $\varphi=(\varphi_1,\ldots,\varphi_p)\in\mathbb{R}^p$ avec $\varphi_p\neq 0$ et $\theta=(\theta_1,\ldots,\theta_q)\in\mathbb{R}^q$ avec $\theta_q\neq 0$ tel que pour tout $t\in\mathbb{Z}$,*

$$X_t-\sum_{i=1}^{p}\varphi_i X_{t-i}=\varepsilon_t+\sum_{i=1}^{q}\theta_i\varepsilon_{t-i}.$$

On utilise généralement la notation suivante :

$$\Phi(B)X_t=\Theta(B)\varepsilon_t$$

où $\Phi(B)=I-\sum_{i=1}^{p}\varphi_i B^i$ et $\Theta(B)=I+\sum_{i=1}^{q}\theta_i B^i$.

Un procesus $AR(p)$ est un processus $ARMA(p,0)$; un procesus $MA(q)$ est un processus $ARMA(0,q)$.

VII.7.1 Processus ARMA canonique et écritures $MA(\infty)$ et $AR(\infty)$

Définition VII.42. *Soit $(X_t)_{t\in\mathbb{Z}}$ un processus $ARMA(p,q)$:*

$$\Phi(B)X_t=\Theta(B)\varepsilon_t$$

La représentation est :

- ***minimale*** *si Φ et Θ n'ont pas de facteurs communs.*
- ***causale*** *si Φ a toutes ses racines à l'extérieur du cercle unité.*
- ***inversible***[2] *si Θ a toutes ses racines à l'extérieur du cercle unité.*
- ***canonique*** *si la représentation est causale et inversible.*

Proposition VII.43. *Le bruit blanc associé à un processus ARMA sous sa forme canonique est le processus d'innovation.*

Proposition VII.44. *Soit $(X_t)_{t\in\mathbb{Z}}$ un processus $ARMA(p,q)$ canonique : $\Phi(B)X_t=\Theta(B)\varepsilon_t$. Si de plus il est minimal alors :*

1. *Il admet une* ***écriture*** *$MA(\infty)$:*

$$\begin{aligned}X_t&=\Phi^{-1}(B)\Theta(B)\varepsilon_t\\&=\varepsilon_t+\sum_{i=1}^{+\infty}\psi_i\varepsilon_{t-i}\end{aligned}$$

où $(\psi_i)_{i\in\mathbb{N}}$ est une suite réelle.
En posant $\psi_i=0$ pour $i<0$, $\theta_0=1$ et $\theta_i=0$ pour $i>q$, on a :

$$\forall i\in\mathbb{N}:\psi_i-\sum_{j=1}^{p}\varphi_j\psi_{i-j}=\theta_i$$

[2] Le terme "inversible" est un point de terminologie ici. Pour un processus ARMA, Φ est toujours inversible (au sens stationnaire) et Θ est inversible (au sens stationnaire) si aucune de ses racines n'est sur le cercle unité.

2. Il admet une ***écriture*** $AR(\infty)$ *:*

$$\varepsilon_t = \Theta^{-1}(B)\,\Phi(B)\,X_t = X_t + \sum_{i=1}^{+\infty}\pi_i X_{t-i} \Leftrightarrow X_t = -\sum_{i=1}^{+\infty}\pi_i X_{t-i} + \varepsilon_t$$

où $(\pi_i)_{i\in\mathbb{N}}$ *est une suite réelle.*
En posant $\pi_i = 0$ *pour* $i < 0$, $\varphi_0 = -1$ *et* $\varphi_i = 0$ *pour* $i > p$, *on a :*

$$\forall i \in \mathbb{N} : \pi_i + \sum_{j=1}^{q}\theta_j \pi_{i-j} = -\varphi_i$$

VII.7.2 Autocorrélations d'un processus ARMA

On peut déterminer les autocorrélations simples à l'aide deux calculs différents
– On obtient à l'aide de la représentation $MA(\infty)$:

$$\begin{aligned}\gamma(h) &= \mathbb{C}ov(X_t, X_{t-h}) \\ &= \mathbb{C}ov\left(\varepsilon_t + \sum_{i=1}^{+\infty}\psi_i\varepsilon_{t-i}, \varepsilon_{t-h} + \sum_{i=1}^{+\infty}\psi_i\varepsilon_{t-h-i}\right) \\ &= \sigma^2\sum_{i=0}^{+\infty}\psi_i\psi_{i+h} \text{ où } \psi_0 = 1\end{aligned}$$

– On a :

$$X_t - \varphi_1 X_{t-1} - ... - \varphi_p X_{t-p} = \varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q}$$

D'où :

$$\begin{aligned}\gamma(h) - \varphi_1\gamma(h-1) - ... - \varphi_p\gamma(h-p) &= \mathbb{C}ov(\varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q}, X_{t-h}) \\ &= \mathbb{C}ov\left(\varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q}, \sum_{i=0}^{+\infty}\psi_i\varepsilon_{t-h-i}\right) \\ &= \begin{cases} \sigma^2\sum_{i=0}^{+\infty}\theta_{h+i}\psi_i & \text{si } h \in \{0, ..., q\} \\ 0 & \text{si } h > q \end{cases}\end{aligned}$$

Les autocorrélations simples décroissent vers 0.
– Si $p > q$ la décroissance est de type exponentiel ou sinusoïdal amorti.
– Si $q \geq p$, les $q-p-1$ premières valeurs ont un comportement quelconque et les suivantes décroissent.
Il existe des propriétés similaires sur les autocorrélations partielles.

Méthode du coin

Il n'existe pas de caractérisation simple des processus ARMA, basée sur les autocorrélations simples ou partielles. La **méthode du coin** repose sur des propriétés des matrices d'autocorrélation.

Soient les matrices suivantes et leurs déterminants respectifs :

$$\Omega_{i,j} = \begin{bmatrix} \rho(i) & \rho(i-1) & ... & \rho(i-j+1) \\ \rho(i-1) & \rho(i) & ... & ... \\ ... & ... & ... & \rho(i-1) \\ \rho(i-j+1) & ... & \rho(i-1) & \rho(i) \end{bmatrix}$$

$$\Delta_{i,j} = \det(\Omega_{i,j})$$

Soit $(X_t)_{t\in\mathbb{Z}}$ un processus $ARMA(p,q)$ canonique minimal :

$$\Phi(B) X_t = \Theta(B) \varepsilon_t$$

On a alors :
– $\forall (i,j) / i > q, j > p : \Delta_{i,j} = 0$
– $\forall (i,j) / i \leq q : \Delta_{i,p} \neq 0$
– $\forall (i,j) / j \leq p : \Delta_{q,j} \neq 0$

On peut visualiser ce résultat sous forme matricielle en représentant la matrice $M = (\Delta_{i,j})_{(i,j)\in\{1,...,k\}^2}$ (pour k assez grand) et faire ainsi apparaître un coin :

$$M = \left[\begin{array}{cccccc} \Delta_{1,1} & ... & \Delta_{1,p} & \Delta_{1,p+1} & ... & \Delta_{1,k} \\ ... & & ... & ... & & ... \\ \Delta_{q,1} & ... & \Delta_{q,p} & \Delta_{q,p+1} & ... & \Delta_{q,k} \\ \Delta_{q+1,1} & ... & \Delta_{q+1,p} & \multicolumn{3}{c}{\multirow{3}{*}{0}} \\ ... & & ... & & & \\ \Delta_{k,1} & ... & \Delta_{k,p} & & & \end{array}\right]$$

VII.7.3 Densité spectrale d'un processus ARMA

Proposition VII.45. *Soit $(X_t)_{t\in\mathbb{Z}}$ un processus $ARMA(p,q)$ (pas forcément sous sa forme canonique), $\Phi(B) X_t = \Theta(B) \varepsilon_t$. Sa densité spectrale vaut :*

$$f(\omega) = \frac{\sigma^2}{2\pi} \frac{\left|\Theta\left(e^{-i\omega}\right)\right|^2}{\left|\Phi\left(e^{-i\omega}\right)\right|^2}$$

VII.7.4 Estimation des processus ARMA

Soit $(\varepsilon_t)_{t\in\mathbb{Z}}$ un bruit blanc faible de variance σ^2 et $(X_t)_{t\in\mathbb{Z}}$ un processus $ARMA(p,q)$ canonique minimal : $\Phi(B) X_t = \Theta(B) \varepsilon_t$.

Principe

Le but est d'estimer les coefficients des polynômes Φ et Θ, ainsi que σ^2.

Une première approche consiste à déterminer ces valeurs à partir des autocorrélations. En général, ces estimateurs ne sont pas efficaces[3]. C'est pourquoi, on utilise des estimations préliminaires comme première étape dans des méthodes itératives, du type maximum de vraisemblance ou moindres carrés.

Estimation préliminaire

Cas des processus AR

[3]Un estimateur efficace est un estimateur sans biais atteignant la borne FDCR (donc de variance minimale).

Dans le cas d'un processus $AR(p)$ canonique, on peut utiliser les équations de Yule-Walker :

$$\begin{cases} \gamma(0) = \dfrac{\sigma^2}{1 - \sum_{i=1}^{p} \varphi_i \rho(i)} \\ \begin{pmatrix} \rho(1) \\ \rho(2) \\ ... \\ \rho(p) \end{pmatrix} = \begin{bmatrix} 1 & \rho(1) & ... & \rho(p-1) \\ \rho(1) & 1 & ... & ... \\ ... & ... & ... & \rho(1) ... \\ \rho(p-1) & ... & \rho(1) & 1 \end{bmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ ... \\ \varphi_p \end{pmatrix} = R_p \varphi. \end{cases}$$

On en déduit :

$$\begin{cases} \widehat{\sigma}^2 = \widehat{\gamma}(0) \left(1 - \sum_{i=1}^{p} \widehat{\varphi}_i \widehat{\rho}(i)\right) \\ \begin{pmatrix} \widehat{\varphi}_1 \\ \widehat{\varphi}_2 \\ ... \\ \widehat{\varphi}_p \end{pmatrix} = \begin{bmatrix} 1 & \widehat{\rho}(1) & ... & \widehat{\rho}(p-1) \\ \widehat{\rho}(1) & 1 & ... & ... \\ ... & ... & ... & \widehat{\rho}(1) ... \\ \widehat{\rho}(p-1) & ... & \widehat{\rho}(1) & 1 \end{bmatrix}^{-1} \begin{pmatrix} \widehat{\rho}(1) \\ \widehat{\rho}(2) \\ ... \\ \widehat{\rho}(p) \end{pmatrix} \end{cases}$$

En notant $\varphi = (\varphi_1, ..., \varphi_p)$ et $\widehat{\varphi} = (\widehat{\varphi}_1, ..., \widehat{\varphi}_p)$, on a les résultats suivants :

$$\begin{array}{l} \sqrt{n}(\widehat{\varphi} - \varphi) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2 \Sigma_p^{-1}\right) \\ \widehat{\sigma}^2 \xrightarrow{\mathbb{P}} \sigma^2 \end{array}$$

où :

$$\Sigma_p = \begin{bmatrix} \gamma(0) & \gamma(1) & ... & \gamma(p-1) \\ \gamma(1) & \gamma(0) & ... & ... \\ ... & ... & ... & \gamma(1) ... \\ \gamma(p-1) & ... & \gamma(1) & \gamma(0) \end{bmatrix} = \gamma(0) R_h.$$

Cas des processus MA et ARMA

Si on a une partie moyenne mobile, on utilise plutôt l'algorithme des innovations. On considère ici le cas d'un processus $ARMA(p,q)$ canonique minimal. Le cas d'un processus $MA(q)$ n'est qu'un cas particulier.

Considérons l'écriture $MA(\infty)$ du processus :

$$X_t = \varepsilon_t + \sum_{i=1}^{+\infty} \psi_i \varepsilon_{t-i}$$

On peut estimer les coefficients $(\psi_i)_{i \in \{1,...,n\}}$ par $\left(\widehat{\psi}_1(n), ..., \widehat{\psi}_{p+q}(n)\right)$ à l'aide de l'algorithme des innovations. On se fixe n a priori ; on le prend suffisamment grand pour avoir assez d'équations nécessaires à la résolution du problème, sachant que $\psi_i \underset{i \to +\infty}{\to} 0$.

En posant $\psi_i = 0$ pour $i < 0$, $\theta_0 = 1$ et $\theta_i = 0$ pour $i > q$, on a :

$$\forall i \in \mathbb{N} : \psi_i - \sum_{j=1}^{p} \varphi_j \psi_{i-j} = \theta_i$$

On utilise ces relations pour obtenir une première estimation de $(\varphi_1, ..., \varphi_p)$ et $(\theta_1, ..., \theta_q)$ à partir de $\left(\widehat{\psi}_1(n), ..., \widehat{\psi}_{p+q}(n)\right)$.

Estimation par la méthode du maximum de vraisemblance

En général, on ajoute une hypothèse sur la loi des résidus. On suppose ici que les résidus sont distribués selon une loi normale $\mathcal{N}\left(0, \sigma^2\right)$, hypothèse qu'il faudrait vérifier à l'aide d'un test (Kolmogorov par exemple).

Si l'hypothèse de normalité n'est pas vérifiée, on peut tout de même considérer que la vraisemblance normale est un critère d'ajustement qui peut convenir.

Si on dispose d'un échantillon $(X_1, ..., X_T)$, la vraisemblance s'écrit :

$$V\left(x_1, ..., x_T; \varphi_1, ..., \varphi_p, \theta_1, ..., \theta_q, \sigma^2\right) = \frac{1}{(2\pi)^{\frac{T}{2}}} \frac{1}{\sqrt{\det \Sigma_T}} \exp\left(-\frac{1}{2} x' \Sigma_T^{-1} x\right)$$

où $x = (x_1, ..., x_T)$ et Σ_T est la matrice de variance-covariance de $(X_1, ..., X_T)$.

Remarquons que $(X_1, ..., X_T)$ est un vecteur gaussien, en tant que transformation linéaire du vecteur gaussien $(\varepsilon_1, ..., \varepsilon_T)$.

La maximisation de cette vraisemblance n'est pas simple et nécessite l'utilisation d'algorithmes d'optimisation.

Notons pour tout $i \in \{2, ..., T\}$, $\widehat{X}_i = \mathbb{EL}(X_i | X_{i-1}, ..., X_1)$.

On peut utiliser l'algorithme des innovations pour calculer les erreurs de prévision au pas 1, $\varepsilon_i = X_i - \widehat{X}_i$, ainsi que leur variance v_i. On évite ainsi le calcul direct de Σ_T^{-1} et de $\det \Sigma_T$. On a :

$$\begin{pmatrix} X_1 \\ X_2 \\ ... \\ X_T \end{pmatrix} = C_T \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_T \end{pmatrix}$$

avec la matrice C_T (estimée par l'algorithme des innovations) définie par

$$C_T = \begin{bmatrix} 1 & 0 & ... & ... & 0 \\ \theta_1(1) & 1 & 0 & ... & 0 \\ ... & ... & ... & ... & ... \\ \theta_{T-1}(T-1) & \theta_{T-2}(T-1) & ... & \theta_1(T-1) & 1 \end{bmatrix}$$

Les $(\varepsilon_i)_{i \in \{1,...,T\}} = \left(X_i - \widehat{X}_i\right)_{i \in \{1,...,T\}}$ ne sont pas corrélés, la matrice de covariance de $(\varepsilon_i)_{i \in \{1,...,T\}}$ vaut :

$$V_T = \begin{bmatrix} v_0 & 0 & ... & 0 \\ 0 & v_1 & ... & ... \\ ... & ... & ... & 0 \\ 0 & ... & 0 & v_{T-1} \end{bmatrix}$$

On a $\Sigma_T = C_T V_T C_T'$ et $\det \Sigma_T = (\det C_T)^2 \det V_T = v_0 ... v_{T-1}$, et donc

$$x' \Sigma_T^{-1} x = \sum_{i=1}^{T} \frac{(x_i - \widehat{x}_i)^2}{v_{i-1}},$$

où les prévisions $\widehat{x}_i$ sont données par l'algorithme des innovations.

La vraisemblance s'écrit donc :

$$V\left(x_1,...,x_T;\varphi_1,...,\varphi_p,\theta_1,...,\theta_q,\sigma^2\right)=\frac{1}{(2\pi)^{\frac{T}{2}}}\frac{1}{\sqrt{v_0...v_{T-1}}}\exp\left(-\frac{1}{2}\sum_{i=1}^{T}\frac{\left(x_i-\widehat{x}_i\right)^2}{v_{i-1}}\right)$$

L'algorithme des innovations nous indique que :

$$\widehat{X}_{i+1}=\begin{cases}\sum\limits_{j=1}^{i}\theta_j\left(i\right)\left(X_{i+1-j}-\widehat{X}_{i+1-j}\right) & \text{si } 1\leq i\leq\max\left(p,q\right)\\ \varphi_1X_i+...+\varphi_pX_{i+1-p}+\sum\limits_{j=1}^{q}\theta_j\left(i\right)\left(X_{i+1-j}-\widehat{X}_{i+1-j}\right) & \text{si } i>\max\left(p,q\right).\end{cases}$$

On définit $r_i=v_i/\sigma^2$. On peut réécrire la vraisemblance comme suit :

$$V\left(x_1,...,x_T;\varphi_1,...,\varphi_p,\theta_1,...,\theta_q,\sigma^2\right)=\frac{1}{(2\pi\sigma^2)^{\frac{T}{2}}}\frac{1}{\sqrt{r_0...r_{T-1}}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{T}\frac{\left(x_i-\widehat{x}_i\right)^2}{r_{i-1}}\right)$$

On utilise la log-vraisemblance. L'estimateur du maximum de vraisemblance vérifie :

- $\left(\widehat{\varphi}_1,...,\widehat{\varphi}_p,\widehat{\theta}_1,...,\widehat{\theta}_q\right)=\underset{(\varphi_1,...,\varphi_p,\theta_1,...,\theta_q)}{\arg\min}\left\{\log\left[\frac{1}{T}S\left(\varphi_1,...,\varphi_p,\theta_1,...,\theta_q\right)\right]+\frac{1}{T}\sum_{i=1}^{T}\log r_{i-1}\right\}$
- $\widehat{\sigma}^2=\frac{1}{T}S\left(\widehat{\varphi}_1,...,\widehat{\varphi}_p,\widehat{\theta}_1,...,\widehat{\theta}_q\right)$ où $S\left(\widehat{\varphi}_1,...,\widehat{\varphi}_p,\widehat{\theta}_1,...,\widehat{\theta}_q\right)=\sum_{i=1}^{T}\frac{\left(x_i-\widehat{x}_i\right)^2}{r_{i-1}}$

La maximisation s'effectue de manière itérative, et l'estimation préliminaire fournit des valeurs initiales. Les estimateurs obtenus sont efficaces.

Estimation par la méthode des moindres carrés

On cherche cette fois à minimiser la quantité suivante :

$$S\left(\varphi_1,...,\varphi_p,\theta_1,...,\theta_q\right)=\sum_{i=1}^{T}\frac{\left(x_i-\widehat{x}_i\right)^2}{r_{i-1}}$$

L'estimateur des moindres carrés vérifie :

- $\left(\widehat{\varphi}_1,...,\widehat{\varphi}_p,\widehat{\theta}_1,...,\widehat{\theta}_q\right)=\underset{(\varphi_1,...,\varphi_p,\theta_1,...,\theta_q)}{\arg\min}S\left(\varphi_1,...,\varphi_p,\theta_1,...,\theta_q\right)$
- $\widehat{\sigma}^2=\frac{1}{T-p-q}S\left(\widehat{\varphi}_1,...,\widehat{\varphi}_p,\widehat{\theta}_1,...,\widehat{\theta}_q\right)$

Les estimateurs obtenus sont efficaces.

Remarque VII.46. Il existe d'autres méthodes d'estimation qu'on ne présente pas ici, notamment la méthode du maximum de vraisemblance conditionnel et la méthode des moindres carrés conditionnels. ◊

VII.7.5 Choix de modèle

Les critères de choix de modèles sont un compromis entre l'ajustement de la série modélisée (on cherche à minimiser la variance résiduelle non expliquée : $\widehat{\sigma}^2$), et la règle de parcimonie (on cherche à minimiser le nombre de paramètre $p+q$), ces deux critères étant antagonistes. Ces critères sont établis en calculant une distance, appelée distance de Kullback, entre la loi inconnue et l'ensemble des lois proposées par le modèle.

– Critère d'Akaïke. La fonction que l'on désire minimiser est

$$AIC(p,q) = \log\left(\widehat{\sigma}^2\right) + 2\frac{p+q}{T}$$

– Critère de Schwarz. La fonction que l'on désire minimiser est

$$BIC(p,q) = \log\left(\widehat{\sigma}^2\right) + (p+q)\frac{\log(T)}{T}$$

VII.8 Pratique des modèles SARIMA

Cette partie présente la mise en oeuvre pratique des modèles SARIMA.

VII.8.1 Méthodologie

Synthèse

La démarche adoptée par Box et Jenkins est la suivante :

1. Stationnarisation
2. Identification a priori de modèles potentiels
3. Estimation des modèles potentiels
4. Vérification des modèles potentiels
5. Choix définitif d'un modèle
6. Prévision à l'aide du modèle choisi
7. Analyse a posteriori

Stationnarisation

Méthodes

La plupart des séries temporelles présentent une tendance et/ou une saisonnalité, et ne sont donc pas modélisables par un processus stationnaire. Afin de se ramener à un processus ARMA, il faut stationnariser la série[4] et différentes méthodes sont envisageables :

– Décomposition saisonnière
Cette méthode permet d'éliminer tendance et saisonnalité, sources évidentes de non-stationnarité ; il se peut néanmoins que la série résultant de la décomposition ne soit toujours pas stationnaire.

– Différenciation
C'est la méthode employée par les modèles ARIMA et SARIMA. On ne modélise pas la série brute mais la série différenciée, en "tendance" à l'aide de $\nabla^d = (I-B)^d$, ou la série différenciée en "saisonnalité" à l'aide de $\nabla_s^D = (I-B^s)^D$. De manière générale, ∇^d permet de stationnariser des séries possédant une tendance polynomiale de degré d, et ∇_s^D des séries possédant une composante saisonnière de période s.
Si les processus stationnaires peuvent être approchés par des modèles ARMA, il n'est pas certain que la différenciation permette de stationnariser tous les processus.

[4] En toute rigueur, on devrait parler de stationnariser un processus et non une série temporelle.

– Méthode de Box-Cox
Elle permet une stationnarisation en "variance" (ou encore de stationnariser des séries présentant une tendance exponentielle). On utilise la transformation suivante : $\frac{X_t^\lambda - 1}{\lambda}$ avec $\lambda \in \mathbb{R}$. Il existe des méthodes alternatives si X_t n'est pas une série positive. Remarquons que :

$$\frac{X_t^\lambda - 1}{\lambda} \xrightarrow{\lambda \to 0} \log(X_t)$$

Il existe des tests de stationnarité, mais aucun n'est "universel". Citons tout de même le test de Dickey-Fuller.

Pratique de la différenciation dans les modèles SARIMA

On utilise très souvent une méthode empirique basée sur l'autocorrélogramme simple.
– On effectue une différenciation en "tendance" si :
  – Les autocorrélations $\widehat{\rho}(h)$[5] sont proches de 1 pour un grand nombre de retards.
  – Les premières autocorrélations $\widehat{\rho}(h)$ sont proches les unes des autres (même si elles ne sont pas forcément proches de 1).
  On parle souvent de décroissance lente des autocorrélations simples.
– On effectue une différenciation en "saisonnalité" si des comportements similaires sont observés de manière périodique. Par exemple, si $\widehat{\rho}(12)$, $\widehat{\rho}(24)$, ... sont proches de 1, on utilise une différenciation en "saisonnalité" avec $s = 12$.

Remarques

1. Quelle que soit la méthode, on procède de manière itérative : on effectue une première différenciation ; si celle-ci n'est pas suffisante, on en effectue une seconde...
2. En pratique, on a souvent $d \leq 2$ et $D \leq 2$.

On travaille dorénavant sur une suite stationnarisée.

Identification a priori de modèles potentiels (Ordre de grandeur de p et q

Une fois la stationnarisation effectuée, on peut se consacrer aux choix potentiels des polynômes AR et MA.

Il existe différentes méthodes pour identifier un modèle $ARMA(p,q)$:
– Méthode de Box et Jenkins
Il s'agit d'une méthode heuristique pour majorer p et q.
  – Pour un processus $AR(p)$:
  On peut montrer que :

$$\forall h > p : \sqrt{n}\widehat{r}(h) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

  On peut définir un intervalle de confiance à 95% et on recherche à partir de quelle valeur, 95% des $\widehat{r}(h)$ sont dans l'intervalle $\left[-\frac{1.96}{\sqrt{n}}, \frac{1.96}{\sqrt{n}}\right]$.

[5] Il s'agit ici d'un abus de langage car si la sortie nommée "autocorrélations simples" nous permet de douter de la stationnarité du processus sous-jacent, nous ne devrions pas parler alors d'autocorrélations simples.

– Pour un processus $MA(q)$:
On peut montrer que :

$$\forall h > q : \sqrt{n}\widehat{\rho}(h) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, 1 + 2\sum_{k=1}^{q}\rho^2(k)\right)$$

On peut définir un intervalle de confiance à 95% et on recherche à partir de quelle valeur, 95% des $\widehat{\rho}(h)$ sont dans l'intervalle suivant :
$$\left[-\frac{1.96}{\sqrt{n}}\left(1 + 2\sum_{k=1}^{q}\widehat{\rho}^2(k)\right)^{\frac{1}{2}}, \frac{1.96}{\sqrt{n}}\left(1 + 2\sum_{k=1}^{q}\widehat{\rho}^2(k)\right)^{\frac{1}{2}}\right]$$

– Méthode du coin
On utilise la méthode du coin avec les estimations de $\widehat{\rho}(h)$. Cette méthode ne permet pas toujours d'aboutir, surtout si on a des effets saisonniers.

– Méthode empirique
En pratique (surtout pour les modèles SARIMA), on essaye d'identifier les autocorrélations simples et partielles "significatives" pour caler ensuite des polynômes AR et MA qui reflettent ces liens temporels.
Afin d'obtenir des modèles potentiels, l'idéal est de regarder l'autocorrélogramme partiel afin d'émettre une hypothèse sur la partie autorégressive (simple et saisonnière), la tester puis regarder l'autocorrélogramme simple (et partiel) du résidu afin d'identifier complètement un modèle. Cette démarche par étape permet en général d'obtenir plusieurs modèles potentiels.

Estimation des modèles potentiels

On estime les modèles potentiels à l'aide des méthodes classiques : maximum de vraisemblance ou moindres carrés.

Vérification des modèles potentiels

Afin de vérifier la validité des modèles estimés, on doit vérifier :

– Significativité des paramètres
Par exemple, pour le coefficient AR d'ordre p, on effectue le test suivant :

$$\begin{cases} H_0 : \text{le processus est un } ARMA(p, q) \\ H_1 : \text{le processus est un } ARMA(p-1, q) \end{cases}$$

On utilise pour cela la statistique de test suivante :

$$t = \frac{|\widehat{\varphi}_p|}{\mathbb{V}(\widehat{\varphi}_p)},$$

où $\mathbb{V}(\widehat{\varphi}_p)$ est la variance (que nous ne précisons pas ici) de $\widehat{\varphi}_p$.
Le test de Student permet de rejeter H_0 au niveau 5% si $|t|$ est supérieur à 1.96.
Il existe des résultats similaire pour les coeffcients MA.

– Blancheur du résidu
On vérifie que le résidu est bien un bruit blanc, à l'aide du test de Portmanteau par exemple.

Remarque VII.47. Les logiciels fournissent généralement la valeur de la statistique de test, ainsi que la p-valeur. On rejette H_0 si la p-valeur est inférieure au niveau du test α. ◊

Choix définitif d'un modèle

Ce choix s'opère entre les modèles potentiels retenus. Il y a plusieurs critères possibles :

- Des critères d'information basés sur l'information de Kullback (par exemple, les critères d'Akaike et de Schwartz).
- Des critères basés sur le pouvoir prédictif.

Une fois ce choix effectué, le modèle retenu est utilisé à des fins de prévision.

Prévision à l'aide du modèle choisi

La fonction de prévision s'obtient assez facilement à partir des écritures autorégressive ou moyenne mobile.

Analyse a posteriori

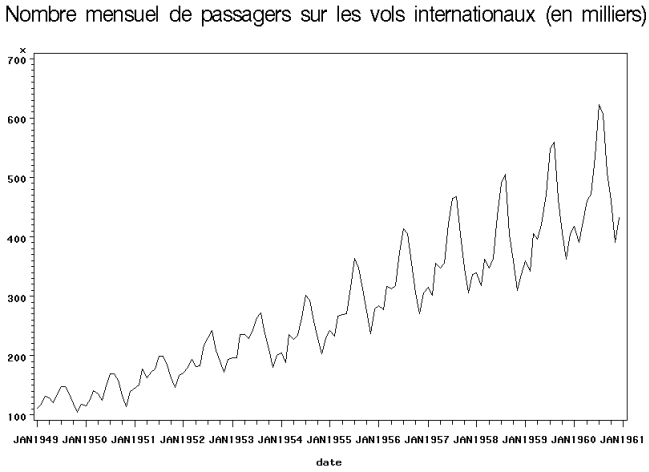
L'analyse a posteriori permet de voir les écarts entre les prévisions et les réalisations, en tronquant la série d'un certain nombre de points ; le modèle doit être correctement estimé sur la série tronquée, et les écarts entre prévisions et réalisations doivent être faibles. On utilise des critères d'erreur comme l'erreur quadratique moyenne (Root Mean Square Error : RMSE) ou l'erreur relative absolue moyenne (Mean Average Percentage Error : MAPE) :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{X}_i\right)^2}$$

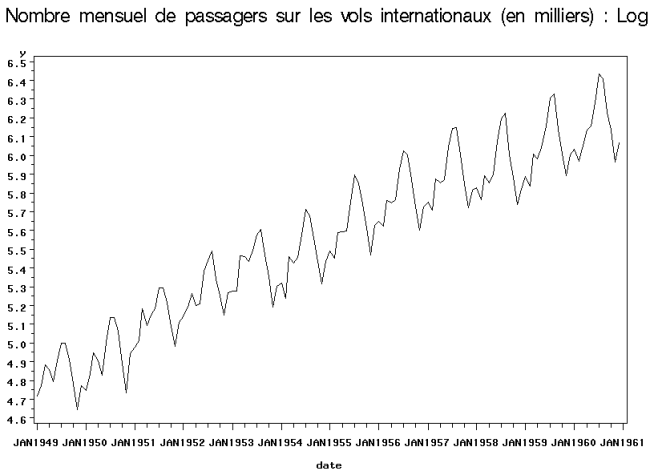
$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{X_i - \widehat{X}_i}{X_i}\right|$$

VII.8.2 Exemple

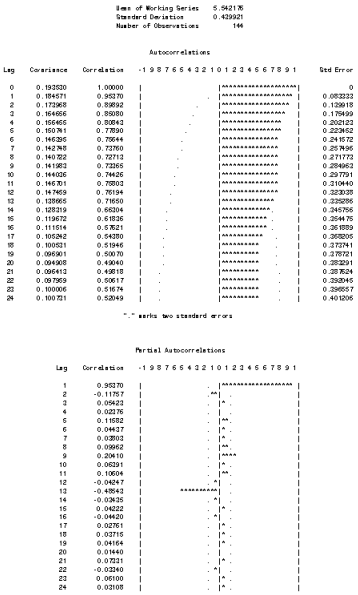
On considère le nombre de passagers aériens qu'on cherche à modéliser à l'aide la méthode de Box et Jenkins.



Le graphique de la série brute, figure VII.8.2, montre une série avec une tendance (de type parabolique ou exponentielle), ainsi qu'une saisonnalité (de période 12). On constate également un accroissement de la variabilité, ce qui explique la transformation logarithmique opérée par la suite.



On voit que la série transformée, figure VII.8.2, présente une tendance (quasiment linéaire) et conserve une saisonnalité (de période 12) ; c'est cette dernière série qui est modélisée par la suite (pour revenir à la série de base, il suffit d'effectuer une transformation exponentielle).



Mean of Working Series	5.542176
Standard Deviation	0.429921
Number of Observations	144

Autocorrelations

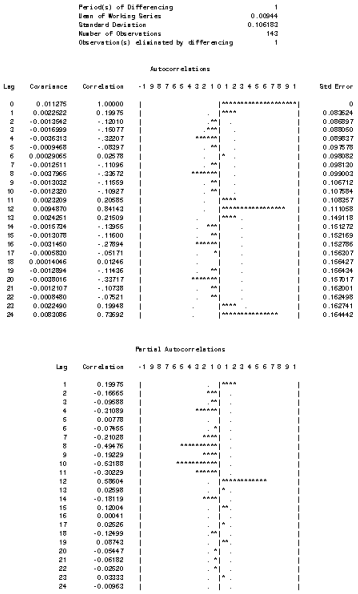
Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.193530	1.00000	|********************|	0
1	0.184571	0.95370	|******************** |	0.083333
2	0.172968	0.89392	|******************** |	0.139618
3	0.164656	0.85098	|*****************. |	0.175499
4	0.155455	0.80343	|****************. |	0.202123
5	0.150761	0.77920	|****************. |	0.223462
6	0.146326	0.75644	|***************. |	0.241572
7	0.142765	0.73803	|***************. |	0.257696
8	0.140722	0.72712	|***************. |	0.271772
9	0.141983	0.73365	|***************. |	0.284963
10	0.144035	0.74426	|***************. |	0.297791
11	0.146701	0.75802	|***************. |	0.310440
12	0.147459	0.76194	|***************. |	0.323028
13	0.138665	0.71650	|**************. |	0.335286
14	0.128319	0.66304	|*************. |	0.345756
15	0.119672	0.61836	|************. |	0.354475
16	0.111514	0.57621	|************. |	0.361889
17	0.105242	0.54380	|***********. |	0.368205
18	0.100531	0.51946	|**********. |	0.373741
19	0.096901	0.50070	|**********. |	0.378721
20	0.094908	0.49040	|**********. |	0.383291
21	0.096412	0.49818	|**********. |	0.387624
22	0.097959	0.50617	|**********. |	0.392046
23	0.100006	0.51674	|**********. |	0.396557
24	0.100731	0.52049	|**********. |	0.401206

"." marks two standard errors

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.95370	|******************** |
2	-0.11757	| .** . |
3	0.09422	| . * . |
4	0.02276	| . . |
5	0.11652	| . ** . |
6	0.04627	| . * . |
7	0.03503	| . * . |
8	0.09962	| . ** . |
9	0.20410	| . **** . |
10	0.06291	| . * . |
11	0.10604	| . ** . |
12	-0.04247	| . * . |
13	-0.48543	| ********* . |
14	-0.03435	| . * . |
15	0.04222	| . * . |
16	-0.04420	| . * . |
17	0.02761	| . * . |
18	0.03715	| . * . |
19	0.04164	| . * . |
20	0.01440	| . . |
21	0.07231	| . * . |
22	-0.02340	| . * . |
23	0.06100	| . * . |
24	0.03108	| . * . |

L'autocorrélogramme simple, figure VII.8.2, de la série montre que les autocorrélations simples décroissent lentement vers 0, ce qui indique un problème de non-stationnarité. On effectue donc une différenciation $(I - B)$. Remarquons qu'il est inutile de commenter l'autocorrélogramme partiel pour l'instant.



Period(s) of Differencing	1
Mean of Working Series	0.00944
Standard Deviation	0.106183
Number of Observations	143
Observation(s) eliminated by differencing	1

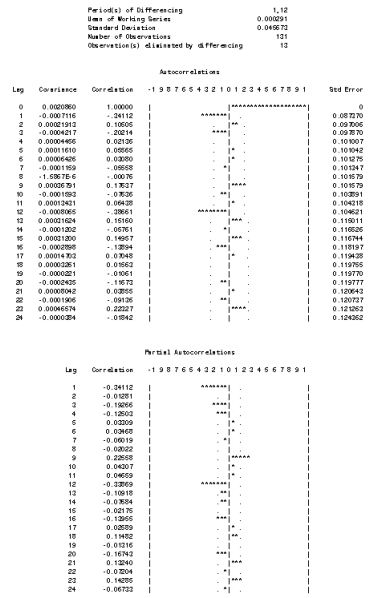
Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.011275	1.00000	|********************|	0
1	0.002252	0.19975	| . **** . |	0.083624
2	-0.001342	-.12010	| . **| . |	0.088397
3	-0.001699	-.15077	| . **| . |	0.089060
4	-0.003613	-.32057	| . *****| . |	0.089837
5	-0.000946	-.08397	| . *| . |	0.093578
6	0.000906	0.08078	| . *| . |	0.093692
7	-0.001251	-.11096	| . *| . |	0.093820
8	-0.003766	-.33572	| . ******| . |	0.099002
9	-0.001032	-.11659	| . *| . |	0.106712
10	-0.001220	-.10927	| . *| . |	0.107534
11	0.002209	0.20685	| . **** . |	0.108257
12	0.009370	0.84143	| . ******************|	0.111958
13	0.002451	0.21909	| . **** . |	0.149118
14	-0.001724	-.13955	| . **| . |	0.151272
15	-0.001378	-.11600	| . *| . |	0.152169
16	-0.002146	-.27894	| . *****| . |	0.152786
17	-0.000520	-.05171	| . *| . |	0.156317
18	0.00014046	0.01246	| . *| . |	0.156427
19	-0.001284	-.11426	| . *| . |	0.156424
20	-0.003916	-.33717	| . ******| . |	0.157017
21	-0.001207	-.10738	| . *| . |	0.162001
22	-0.000486	-.0521	| . *| . |	0.162498
23	0.002490	0.19948	| . **** . |	0.162741
24	0.008086	0.73692	| . **************|	0.164442

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.19975	| . **** . |
2	-0.16665	| . **| . |
3	-0.09688	| . *| . |
4	-0.31089	| . *****| . |
5	0.00778	| . . |
6	-0.07455	| . *| . |
7	-0.21028	| . ***| . |
8	-0.49476	| *********| . |
9	-0.19229	| . ***| . |
10	-0.53188	| **********| . |
11	-0.30229	| . *****| . |
12	0.53604	| . **********| . |
13	0.02598	| . * . |
14	-0.18119	| . ***| . |
15	0.12004	| . **. |
16	0.00041	| . . |
17	0.02526	| . * . |
18	-0.12699	| . **| . |
19	0.08742	| . *\. |
20	-0.05447	| . *| . |
21	-0.06182	| . *| . |
22	-0.02520	| . *| . |
23	0.03332	| . *\. |
24	-0.00963	| . . |

L'autocorrélogramme simple, figure VII.8.2, de la série ainsi différenciée montre encore que les corrélations multiples de 12 décroissent lentement vers 0. On applique cette fois la différenciation $(I - B^{12})$.



Period(s) of Differencing	1,12
Mean of Working Series	0.000291
Standard Deviation	0.045672
Number of Observations	131
Observation(s) eliminated by differencing	13

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.0020860	1.00000	|********************|	0
1	-0.0007116	-.34112	*******| .	0.087370
2	0.00021913	0.10505	. |*| .	0.093506
3	-0.0004217	-.20214	****| .	0.093870
4	0.00004456	0.02135	. | .	0.101007
5	0.00011610	0.05565	. |* .	0.101042
6	0.00006426	0.03080	. |* .	0.101275
7	-0.0001159	-.05558	. *| .	0.101347
8	-1.53E-6	-.00076	. | .	0.101579
9	0.00026791	0.12837	. |****	0.101579
10	-0.0001952	-.09356	. *| .	0.102891
11	0.00013421	0.06431	. |* .	0.104218
12	-0.0008065	-.38661	*******| .	0.104621
13	0.00021624	0.10366	. |*** .	0.115011
14	-0.0001202	-.05761	. *| .	0.116526
15	0.00031200	0.14957	. |*** .	0.116744
16	-0.0002788	-.13364	. ***| .	0.118197
17	0.00014703	0.07048	. |* .	0.119438
18	0.00003361	0.01612	. | .	0.119755
19	-0.0000221	-.01061	. | .	0.119770
20	-0.0002435	-.11673	. **| .	0.119777
21	0.00009042	0.04335	. |* .	0.120643
22	-0.0001906	-.09135	. **| .	0.120727
23	0.00046574	0.22327	. |**** .	0.121262
24	-0.0000784	-.01542	. | .	0.124262

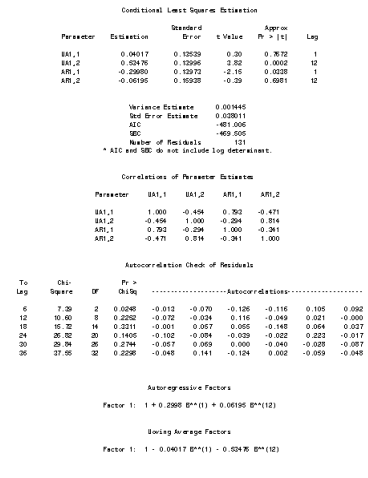
Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.34112	*******| .
2	-0.01281	. | .
3	-0.19266	. ***| .
4	-0.12503	. **| .
5	0.02209	. |* .
6	0.03468	. |* .
7	-0.06019	. *| .
8	-0.02022	. | .
9	0.22688	. |*****
10	0.04207	. |* .
11	0.04669	. |* .
12	-0.33969	*******| .
13	-0.10918	. *| .
14	-0.07894	. *| .
15	-0.02175	. | .
16	-0.12965	. **| .
17	0.02839	. |* .
18	0.11462	. |** .
19	-0.01216	. | .
20	-0.16743	. ***| .
21	0.13240	. |*** .
22	-0.07204	. *| .
23	0.14286	. |*** .
24	-0.06733	. *| .

L'autocorrélogramme simple, figure VII.8.2, de la série doublement différenciée ne semble pas poser de problème de stationnarité. La série sur laquelle on travaille est donc :

$$Y_t = (I - B)\left(I - B^{12}\right)\log\left(X_t\right)$$

On constate que certaines autocorrélations simples et partielles de cette série sont significativement différentes de 0 ; voici trois modèles qui sont testés, et les résultats obtenus pour chacun de ces modèles, voir figures VII.8.2, VII.8.2 et VII.8.2.



Conditional Least Squares Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.04017	0.12539	0.32	0.7632	1
MA1,2	0.53476	0.12996	3.82	0.0002	12
AR1,1	-0.29690	0.12972	-2.15	0.0338	1
AR1,2	-0.06195	0.15928	-0.29	0.6981	12

Variance Estimate	0.001445
Std Error Estimate	0.038011
AIC	-491.006
SBC	-469.505
Number of Residuals	131

* AIC and SBC do not include log determinant.

Correlations of Parameter Estimates

Parameter	MA1,1	MA1,2	AR1,1	AR1,2
MA1,1	1.000	-0.494	0.792	-0.471
MA1,2	-0.494	1.000	-0.294	0.814
AR1,1	0.792	-0.294	1.000	-0.341
AR1,2	-0.471	0.814	-0.341	1.000

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations				
6	7.39	2	0.0248	-0.012	-0.070	-0.126	-0.116	0.105	0.092
12	10.60	8	0.2252	-0.072	-0.024	0.116	-0.049	0.021	-0.000
18	15.32	14	0.3211	-0.001	0.057	0.055	-0.148	0.064	0.027
24	26.92	20	0.1405	-0.102	-0.084	-0.039	-0.022	0.223	-0.017
30	29.94	26	0.2744	-0.057	0.069	0.000	-0.040	-0.028	-0.087
36	37.55	32	0.2298	-0.048	0.141	-0.124	0.002	-0.059	-0.048

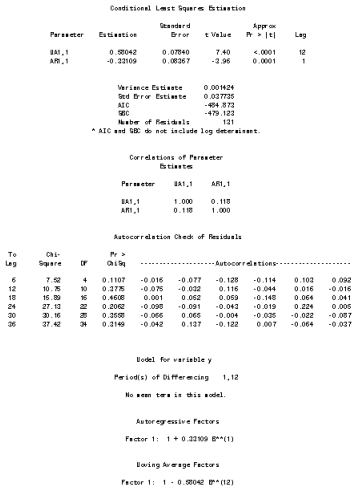
Autoregressive Factors

Factor 1: 1 + 0.2968 B**(1) + 0.06195 B**(12)

Moving Average Factors

Factor 1: 1 - 0.04017 B**(1) - 0.53476 B**(12)

FIG. VII.7 – Modèle 1 : $\left(I - \varphi_1 B - \varphi_{12} B^{12}\right) Y_t = \left(I + \theta_1 B + \theta_{12} B^{12}\right) \varepsilon_t$



Conditional Least Squares Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.58042	0.07840	7.40	<.0001	12
AR1,1	-0.33109	0.08267	-2.96	0.0001	1

Variance Estimate	0.001424
Std Error Estimate	0.037735
AIC	-494.572
SBC	-479.123
Number of Residuals	131

^ AIC and SBC do not include log determinant.

Correlations of Parameter Estimates

Parameter	MA1,1	AR1,1
MA1,1	1.000	0.118
AR1,1	0.118	1.000

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	7.52	4	0.1107	-0.015	-0.077	-0.128	-0.114	0.102	0.092
12	10.75	10	0.3775	-0.075	-0.032	0.116	-0.044	0.016	-0.016
18	15.89	16	0.4608	0.001	0.052	0.059	-0.148	0.064	0.041
24	27.13	22	0.2062	-0.098	-0.091	-0.042	-0.019	0.224	0.005
30	30.16	28	0.3588	-0.066	0.065	-0.004	-0.035	-0.022	-0.057
36	37.42	34	0.3149	-0.042	0.137	-0.122	0.007	-0.064	-0.037

Model for variable y

Period(s) of Differencing 1,12

No mean term in this model.

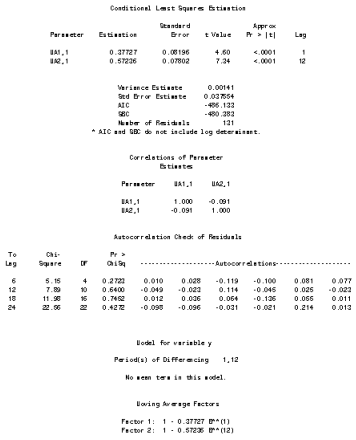
Autoregressive Factors

Factor 1: 1 + 0.33109 B**(1)

Moving Average Factors

Factor 1: 1 - 0.58042 B**(12)

FIG. VII.8 – Modèle 2 :$(I - \varphi_1 B)\, Y_t = \left(I + \theta_{12} B^{12}\right) \varepsilon_t$



Conditional Least Squares Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.37727	0.08196	4.60	<.0001	1
MA2,1	0.57235	0.07502	7.34	<.0001	12

Variance Estimate	0.00141
Std Error Estimate	0.037554
AIC	-495.123
SBC	-480.383
Number of Residuals	131

^ AIC and SBC do not include log determinant.

Correlations of Parameter Estimates

Parameter	MA1,1	MA2,1
MA1,1	1.000	-0.091
MA2,1	-0.091	1.000

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	5.15	4	0.2723	0.010	0.028	-0.119	-0.100	0.081	0.077
12	7.89	10	0.6400	-0.049	-0.023	0.114	-0.045	0.025	-0.023
18	11.98	16	0.7462	0.012	0.035	0.064	-0.135	0.055	0.011
24	22.66	22	0.4212	-0.098	-0.096	-0.031	-0.021	0.214	0.012

Model for variable y

Period(s) of Differencing 1,12

No mean term in this model.

Moving Average Factors

Factor 1: 1 - 0.37727 B**(1)
Factor 2: 1 - 0.57235 B**(12)

FIG. VII.9 – Modèle 3 : $Y_t = (I + \theta_1 B)\left(I + \theta_{12} B^{12}\right) \varepsilon_t$

Afin de lire quel est le modèle testé sous SAS, il faut regarder la fin du listing dans lequel apparaissent les polynômes moyenne mobile et autorégressif.

On constate que seuls les modèles 2 et 3 conviennent. En effet les coefficients estimés dans le modèle 1 ne sont pas tous significatifs ($MA1,1 = -\theta_1$ et $AR1,2 = -\varphi_{12}$). On peut également remarquer que le test de Portmanteau n'est pas validé sur les 6 premières autocorrélations, ce qui n'est pas forcément trop grave car les résultats sont toujours plus incertains sur six autocorrélations. Ces lectures de tests sont effectuées avec un niveau de test de 5%.

Le modèle 3 présente un AIC plus faible que le modèle 2 ; on choisit donc le modèle 3 pour effectuer la prévision. On pourrait également prendre comme critères le BIC ou encore l'écart-type du résidu.

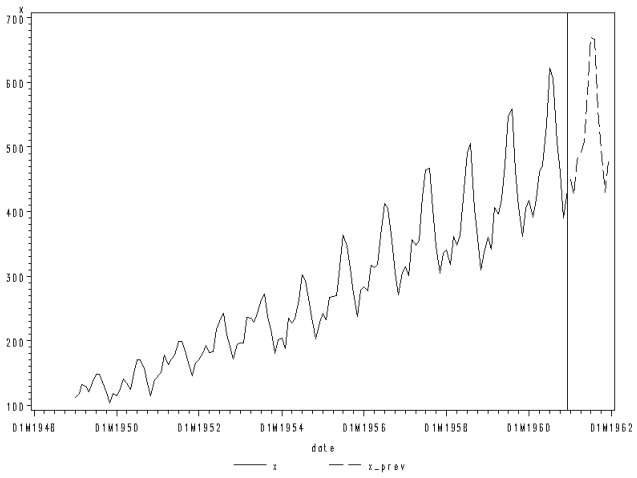
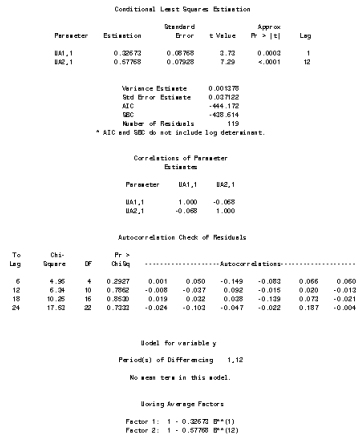


FIG. VII.10 – Prévision par le modèle 3

La prévision effectuée, cf. figure VII.8.2, par le modèle 3 semble raisonnable vu le passé de la série.

On effectue une analyse a posteriori, cf. figure VII.8.2 et VII.8.2, en :

- tronquant la série de 12 points ;
- estimant le modèle 3 sur la série tronquée (on constate que le modèle est correctement estimé) ;
- prévoyant les douze points manquants à l'aide du modèle SARIMA ainsi estimé.



Conditional Least Squares Estimation

Parameter	Estimation	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.32672	0.08768	3.72	0.0003	1
MA2,1	0.57768	0.07928	7.29	<.0001	12

Variance Estimate	0.001378
Std Error Estimate	0.037122
AIC	-444.172
SBC	-438.614
Number of Residuals	119

* AIC and SBC do not include log determinant.

Correlations of Parameter Estimates

Parameter	MA1,1	MA2,1
MA1,1	1.000	-0.068
MA2,1	-0.068	1.000

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations				
6	4.95	4	0.2927	0.001	0.050	-0.149	-0.083	0.066	0.060
12	6.34	10	0.7862	-0.008	-0.027	0.092	-0.015	0.020	-0.012
18	10.25	16	0.8530	0.019	0.032	0.028	-0.129	0.073	-0.021
24	17.52	22	0.7222	-0.024	-0.102	-0.047	-0.022	0.187	-0.004

Model for variable y

Period(s) of Differencing 1,12

No mean term in this model.

Moving Average Factors

Factor 1: 1 - 0.32672 B**(1)
Factor 2: 1 - 0.57768 B**(12)

FIG. VII.11 – Analyse a posteriori

On obtient les résultats suivants :

$$RMSE = 18,5$$
$$MAPE = 2,9$$

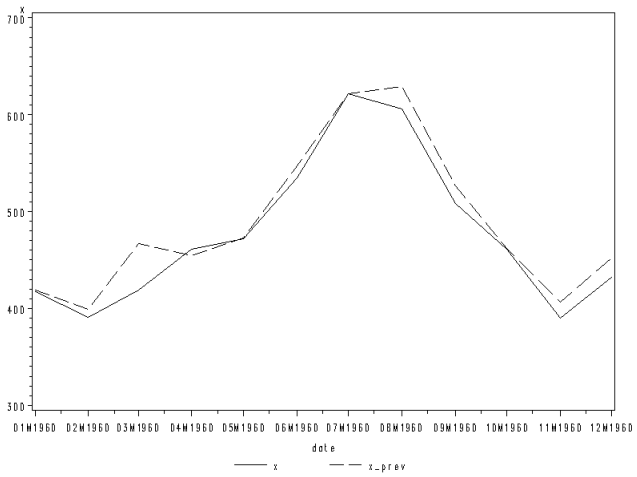


FIG. VII.12 – L'interprétation des critères d'erreur dépend de la série et de la qualité de prévision exigée. Dans le cas présent, un MAPE de 2.9% semble satisfaisant a priori.

Chapitre VIII

Apprentissage Statistique

VIII.1 Introduction

Nous avons vu en introduction de ce cours que la statistique comprend deux pans :
- la statistique décisionnelle (ou inférencielle) qui utilise les bases de données pour prédire la valeur de variables non observées
- la statistique descriptive qui a pour but de décrire les liens existant entre les différentes variables observées.

En statistique décisionnelle, les deux types de modèles considérés sont les modèles paramétriques (qui basent leur prédiction sur un nombre de paramètres fini indépendant de la taille de la base de données) et les modèles non paramétriques.

L'apprentissage statistique est la branche non paramétique de la statistique décisionnelle qui s'intéresse aux bases de données composées de n couples, souvent appelés couples entrée-sortie, supposés *indépendants et identiquement distribués*. Le but d'un algorithme d'apprentissage statistique est de proposer pour toute nouvelle entrée une prédiction de la sortie associée à cette entrée.

Les procédures d'apprentissage statistique sont utiles lorsqu'une modélisation paramètrique de la loi générant les données n'est pas accessible ou lorsque la complexité du modèle est telle qu'elle empêche son utilisation pour la prédiction.

Ces méthodes sont devenues incontournables dans de nombreuses applications pratiques (classement et analyse d'images, reconnaissance d'objets, classement de documents textuels (par exemple : pourriel vs non pourriel), diagnostic médical, analyse de séquences génétiques ou de protéines, prédiction du rendement d'actifs financiers, interface cerveau-machine, ...).

VIII.2 Description formelle et exemples

VIII.2.1 Problématique

Nous observons une base de données composée de n couples $Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)$ que nous supposons être des réalisations indépendantes d'une même loi $\mathbb{P}$ inconnue. Les $X_1, \ldots, X_n$ appartiennent à un espace $\mathcal{X}$ et s'appellent les entrées. Typiquement, $\mathcal{X} = \mathbb{R}^d$ pour un grand entier d. Les $Y_1, \ldots, Y_n$ appartiennent à un espace $\mathcal{Y}$, et s'appellent les sorties. Typiquement, $\mathcal{Y}$ est fini ou $\mathcal{Y}$ est un sous-ensemble de $\mathbb{R}$.

But de l'apprentissage statistique : prédire la sortie Y associée à toute nouvelle entrée X, où il est sous-entendu que la paire (X, Y) est une nouvelle réalisation de la loi $\mathbb{P}$, cette réalisation étant indépendante des réalisations précédemment observées.

Une *fonction de prédiction* est une fonction (mesurable) de $\mathcal{X}$ dans $\mathcal{Y}$. Dans ce chapitre, nous supposons que toutes les quantités que nous manipulons sont mesurables. L'ensemble de toutes les fonctions de prédiction est noté $\mathcal{F}(\mathcal{X}, \mathcal{Y})$. La base de données $Z_1, \ldots, Z_n$ est appelée *ensemble d'apprentissage*, et sera parfois notée Z_1^n. Un *algorithme d'apprentissage* est une fonction qui à tout ensemble d'apprentissage renvoie une fonction de prédiction, i.e. une fonction de l'union $\cup_{n\geq 1}\mathcal{Z}^n$ dans l'ensemble $\mathcal{F}(\mathcal{X}, \mathcal{Y})$, où $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. C'est un estimateur de "la meilleure" fonction de prédiction, où le terme "meilleure" sera précisé ultérieurement.

Soit $\ell(y, y')$ la perte encourue lorsque la sortie réelle est y et la sortie prédite est y'. La fonction $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ est appelée *fonction de perte*.

Exemple du classement : $\ell(y, y') = \mathbf{1}_{y \neq y'}$ (i.e. $\ell(y, y') = 1$ si $y \neq y'$ et $\ell(y, y') = 0$ sinon). Un problème d'apprentissage pour lequel cette fonction de perte est utilisée est appelé problème de *classement* (ou plus couramment par anglicisme *classification*). L'ensemble $\mathcal{Y}$ considéré en classement est le plus souvent fini, voire même de cardinal deux en *classement binaire*.

Exemple de la régression L_p *:* $\mathcal{Y} = \mathbb{R}$ et $\ell(y, y') = |y - y'|^p$ où $p \geq 1$ est un réel fixé. Dans ce cas, on parle de régression L_p. La tâche d'apprentissage lorsque $p = 2$ est aussi appelée *régression aux moindres carrés*.

La qualité d'une fonction de prédiction $g : \mathcal{X} \to \mathcal{Y}$ est mesurée par son *risque* (ou *erreur de généralisation*) :

$$R(g) = \mathbb{E}\big[\ell\big(Y, g(X)\big)\big]. \tag{VIII.1}$$

Le risque est donc l'espérance par rapport à loi $\mathbb{P}$ de la perte encourue sur la donnée (X, Y) par la fonction de prédiction g. La qualité d'un algorithme d'apprentissage $\hat{g}_n$, construit à partir de Z_1^n, peut être mesurée par son risque moyen $\mathbb{E}R[\hat{g}_n]$, où il est sous-entendu que l'espérance est prise par rapport à la loi de l'ensemble d'apprentissage.

La "meilleure" fonction de prédiction est la (ou plus rigoureusement une) fonction de $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ minimisant R. Une telle fonction n'existe pas nécessairement mais existe pour les fonctions de pertes usuelles (notamment celles que nous considérerons par la suite). Cette "meilleure" fonction sera appelée *fonction cible* ou *fonction oracle*.

Exemple du classement : $\ell(y, y') = \mathbf{1}_{y \neq y'}$. La fonction qui à une entrée x renvoie la sortie la plus probable (au sens de la distribution conditionnelle de Y sachant $X = x$: $\mathcal{L}(Y|X = x)$) est "la" fonction cible en classement.

Exemple de la régression aux moindres carrés : $\mathcal{Y} = \mathbb{R}$ et $\ell(y, y') = |y - y'|^2$. La fonction qui à une entrée x renvoie la sortie moyenne $\mathbb{E}(Y|X = x)$ est "la" fonction cible en régression aux moindres carrés.

VIII.2.2 Exemples

Dans ce paragraphe, nous proposons des exemples illustrant la problématique précédente.

Exemple VIII.1. La reconnaissance de caractères manuscrits est un des problèmes sur lequel les méthodes d'apprentissage ont permis des avancées fulgurantes. Le contexte est

le suivant : nous disposons d'une image numérisée d'un caractère manuscrit. Cette image est essentiellement un tableau de nombre réels indiquant l'intensité lumineuse en chacun des pixels. Nous souhaitons trouver la fonction qui à ce tableau de réels renvoie le caractère présent dans l'image. A l'heure actuelle, les meilleures méthodes pour trouver une telle fonction sont de nature statistique : elles reposent donc sur

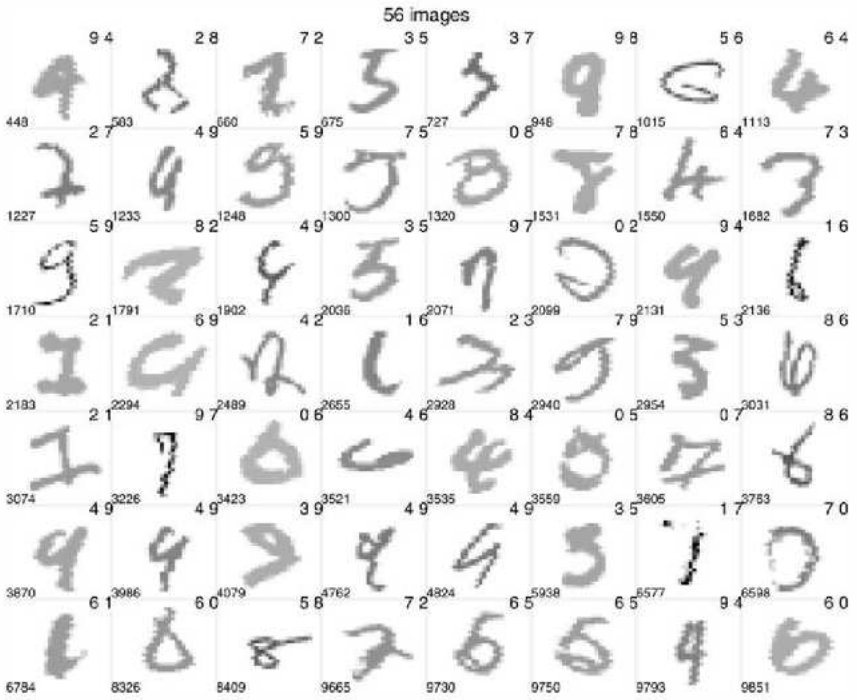


FIG. VIII.1 – Reconnaissance de chiffres manuscrits. Les 56 erreurs sur les 10 000 caractères de la base de test (MNIST/VSV2) d'un des meilleurs algorithmes de reconnaissance de caractères manuscrits [2, 7]. Le premier nombre en haut à droite indique la valeur prédite et le second indique la vraie valeur. Le nombre en bas à gauche est le numéro de l'image (de 1 à 10 000).

1. la constitution d'une base d'images de caractères où les images sont étiquetées par le caractère qu'elle contient. Un X_i correspond donc à une de ces images et un Y_i désigne le caractère que X_i contient.
2. l'utilisation de cette base pour proposer une estimation non paramétrique de la fonction cible.

Le taux d'erreur sur la reconnaissance de chiffres manuscrits (problème intéressant notamment les centres de tri postaux) sont de l'ordre de 0.5% pour les meilleurs algorithmes (voir figure VIII.1). ◊

Exemple VIII.2. Ces dernières années ont vu un essor de la recherche sur la manière d'interfacer le cerveau humain avec un ordinateur. Un des enjeux fondamentaux de ce domaine est de permettre aux personnes ayant perdu l'usage de leurs mains de communiquer avec un ordinateur.

Une des méthodes d'interface cerveau-machine consiste à placer des électrodes à la surface de leur cerveau (méthode non invasive) ou à l'intérieur (méthode invasive). Le but est de déduire de l'observation des signaux électriques les pensées et/ou volontés du sujet. A l'heure actuelle, ces méthodes reposent sur

- la constitution d'une base d'apprentissage : il est demandé à un (ou des) sujet(s) sous observation à penser à quelque chose de précis (par exemple, bouger la souris vers le haut, le bas, la gauche ou la droite ; autre exemple : à une des lettres de l'alphabet). La nature de cette pensée est une sortie Y_i associée à l'entrée X_i qui est le signal électrique observé pendant la seconde suivant la requête.
- l'apprentissage de la fonction cible à partir de cette base de signaux électriques étiquetés.

$\Diamond$

Exemple VIII.3. Pour faire face aux fluctuations incontrôlées du marché, les banques proposent aujourd'hui des produits dont les fluctuations sont indépendantes de la tendance (baissière ou haussière) du marché. Ces placements, dits de gestion alternative, reposent sur l'achat des actions qui vont croître le plus (ou du moins baisser le moins) et la vente pour le même montant des actions qui vont baisser le plus (ou croître le moins). La difficulté pour mettre en place ce type de produit est d'identifier ces actions "sur-performantes" et "sous-performantes".

Une méthode utilisée par les banques est

- de recenser un ensemble de paramètres caractérisant l'action. Ces paramètres proviennent autant des analystes techniques (qui étudie les courbes à travers des paramètres tels que la moyenne mobile, les seuils de résistance, ...) que des analystes financiers (qui étudie les paramètres de la société : chiffre d'affaires, bénéfices, indices de rentabilité, ...).
- de constituer une base de données où une entrée X_i est un vecteur des paramètres décrits ci-dessus et la sortie Y_i associée évalue la sur/sous-performance de l'action sur la période suivant l'observation de ces paramètres (typiquement de l'ordre d'une semaine)
- l'apprentissage de la fonction cible qui, à une date donnée, pour chaque action du marché, associe aux paramètres observés l'indice de sur/sous-performance de l'action.

$\Diamond$

VIII.2.3 Lien entre classement binaire et régression aux moindres carrés

Dans cette section, nous considérons le problème de prédiction binaire, c'est-à-dire où la sortie ne peut prendre que deux valeurs. C'est en particulier la problématique des logiciels de lutte contre les pourriels (ou, par anglicisme, spam). Sans perte de généralité, nous pouvons considérer : $\mathcal{Y} = \{0; 1\}$. Le théorème suivant précise le lien entre classement binaire et régression aux moindres carrés dans le contexte de la prédiction binaire.

Considérons $\mathcal{Y} = \{0; 1\}$. Soit η^* la fonction cible en régression aux moindres carrés définie par $\eta^*(x) = \mathbb{E}(Y|X = x) = \mathbb{P}(Y = 1|X = x)$. Soit g^* la fonction cible en classement définie par

$$g^*(x) = \mathbf{1}_{\eta^*(x)\geq 1/2} = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1|X = x) \geq 1/2 \\ 0 & \text{sinon} \end{cases}$$

Pour toute fonction de régression $\eta : \mathcal{X} \to \mathbb{R}$, on définit la fonction de classement $g_\eta \triangleq \mathbf{1}_{\eta\geq 1/2}$.

Théorème VIII.4. *Nous avons*

$$R_{cla}(g_\eta) - R_{cla}(g^*) \leq 2\sqrt{R_{reg}(\eta) - R_{reg}(\eta^*)},$$

où R_{cla} et R_{reg} désignent respectivement les risques en classement et en régression aux moindres carrés : précisément $R_{cla}(g_\eta) = \mathbb{P}[Y \neq g_\eta(X)]$ et $R_{reg}(\eta) = \mathbb{E}\big[\big(Y - \eta(X)\big)\big]$.

Autrement dit, si η est une "bonne" fonction de régression, alors sa version seuillée g_η est une "bonne" fonction de classement.

VIII.2.4 Consistance universelle

Définition VIII.5. *Un algorithme d'apprentissage est dit* consistant par rapport à la loi *de (X, Y) si et seulement si*

$$\mathbb{E}R(\hat{g}_n) \underset{n\to+\infty}{\longrightarrow} R(g^*).$$

Un algorithme d'apprentissage est dit universellement consistant *si et seulement si il est consistant par rapport à toute loi possible du couple (X, Y).*

Théorème VIII.6. *Si un algorithme $\hat{\eta}$ est universellement consistant pour la régression aux moindres carrés à sorties dans $[0; 1]$ $\big($i.e. $\mathcal{Y} = [0; 1]$, $\ell(y, y') = (y - y')^2\big)$, alors l'algorithme $\hat{g} = \mathbf{1}_{\hat{\eta}\geq 1/2}$ est universellement consistant pour le problème de classement binaire à sorties dans $\{0; 1\}$ $\big($i.e. $\mathcal{Y} = \{0; 1\}$, $\ell(y, y') = \mathbf{1}_{y\neq y'}\big)$.*

Démonstration. Le point de départ consiste à remarquer que si $\hat{\eta}$ est universellement consistant pour la régression aux moindres carrés à sorties dans $[0; 1]$, alors $\hat{\eta}$ est en particulier consistant par rapport à toute distribution telle que les sorties sont dans $\{0; 1\}$ avec probabilité un (i.e. presque sûrement). Le résultat annoncé est une conséquence directe du théorème VIII.4 et du théorème de Jensen. □

VIII.3 Les algorithmes d'apprentissage et leur consistance

Dans ce paragraphe[1], nous considérons (sauf mention contraire) le problème de régression aux moindres carrés défini par $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = [-B; B]$ pour $B > 0$ et $\ell(y, y') = (y - y')^2$. Une fonction cible est alors $\eta^* : x \mapsto \mathbb{E}(Y|X = x)$.

VIII.3.1 Algorithmes par moyennage locale

Idée principale : Pour estimer $\mathbb{E}(Y\|X = x)$, moyenner les Y_i des X_i proches de x.

Cette idée nous amène à nous intéresser aux algorithmes d'apprentissage de la forme

$$\hat{\eta} : x \mapsto \textstyle\sum_{i=1}^n W_i(x) Y_i,$$

où les poids réels $W_i(x)$ vont être des fonctions bien choisies de $x, n, X_1, \ldots, X_n$.

[1] Ce paragraphe s'inspire largement des six premiers chapitres du livre [4].

Exemple 1 : Algorithme par partition : soit $\{A_1, A_2, \ldots\}$ une partition finie ou dénombrable de $\mathcal{X}$ (i.e. $\cup_j A_j = \mathcal{X}$ et $A_j \cap A_k = \emptyset$ pour $j \neq k$). Soit $A(x)$ l'élément de la partition qui contient x. L'algorithme par partition considère les poids

$$W_i(x) = \frac{\mathbf{1}_{X_i \in A(x)}}{\sum_{l=1}^n \mathbf{1}_{X_l \in A(x)}}, \tag{VIII.2}$$

où nous utilisons la convention $\frac{0}{0} = 0$.

Exemple 2 : Algorithme par noyau (ou *estimateur de Nadaraya-Watson*) : Soient $K : \mathbb{R}^d \to \mathbb{R}_+$ et $h > 0$ un paramètre (dit de largeur du noyau). L'algorithme par noyau considère les poids

$$W_i(x) = \frac{K(\frac{x - X_i}{h})}{\sum_{l=1}^n K(\frac{x - X_l}{h})}, \tag{VIII.3}$$

où nous utilisons toujours la convention $\frac{0}{0} = 0$. Les deux noyaux les plus courants sont le noyau fenêtre $K(x) = \mathbf{1}_{\|x\| \leq 1}$ et le noyau gaussien $e^{-\|x\|^2}$, avec $\|\cdot\|$ la norme euclidienne.

Exemple 3 : L'*algorithme des k-plus proches voisins* (k-p.p.v.) considère les poids

$$W_i(x) = \begin{cases} \frac{1}{k} & \text{si } X_i \text{ fait partie des } k\text{-p.p.v. de } x \text{ dans } X_1, \ldots, X_n \\ 0 & \text{sinon} \end{cases}. \tag{VIII.4}$$

L'algorithme des k-plus proches voisins

Pour définir parfaitement l'algorithme des k-p.p.v., il faut préciser ce qui se passe lorsque le point x à classer est à égale distance de plusieurs points de l'ensemble d'apprentissage. La règle la plus simple consiste à traiter ces problèmes d'égalité de distances par tirage au sort : pour une distance d et $x \in \mathbb{R}^d$, si l'ensemble $E = \big\{i \in \{1, \ldots, n\} : \|X_i - x\| = d\big\}$ est de cardinal $|E| \geq 2$, alors les points $(X_i)_{i \in E}$ sont ordonnés en tirant une permutation aléatoire (suivant la loi uniforme sur l'ensemble des $|E|!$ permutations de E).

Pour cette gestion des égalités éventuelles de distances, nous avons

Théorème VIII.7. *Si $k_n \underset{n \to +\infty}{\longrightarrow} +\infty$ et $k_n/n \underset{n \to +\infty}{\longrightarrow} 0$, l'algorithme des k_n-p.p.v. est universellement consistant.*

- L'algorithme des k-p.p.v. nécessite de conserver en mémoire tous les points de la base d'apprentissage, d'où un stockage coûteux (en $\mathrm{O}(n)$).
- Une implémentation naïve de l'algorithme repose sur le calcul des distances entre le point pour lequel la prédiction est recherchée et les points de l'ensemble d'apprentissage, d'où un algorithme en $\mathrm{O}(n)$. Des méthodes basées sur la construction d'un arbre associé à l'ensemble d'apprentissage ont un coût moyen en $\mathrm{O}(\log n)$. Cela nécessite néanmoins la construction de l'arbre, ce qui a en général un coût en $\mathrm{O}(n \log n)$ (voir par exemple 'kd-tree' sur `en.wikipedia.org`). Une implémentation de cette méthode est accessible en ligne sous le nom d'¡¡approximate nearest neighbour¿¿ (`www.cs.umd.edu/~mount/ANN/`) qui permet également de rechercher les plus proches voisins de manière approchée (ce qui mène à un gain de temps considérable). Une astuce pour améliorer la recherche approchée des plus proches voisins est de construire plusieurs arbres de recherche (si possible les plus différents possibles).

– Le choix du paramètre k peut être effectué par une méthode dite de validation croisée que nous décrivons ci-dessous. Cette méthode, couramment utilisée pour trouver les 1 ou 2 paramètres de réglage d'un algorithme d'apprentissage, repose sur l'estimation du risque moyen $\mathbb{E}R(\hat{g})$ d'un algorithme par

$$\frac{1}{n}\sum_{j=1}^{p}\sum_{(x,y)\in B_j}\ell\big[y,\hat{g}(\cup_{k\neq j}B_k)(x)\big] \tag{VIII.5}$$

où p est l'ordre de la validation croisée et $B_1,\ldots,B_p$ est une partition équilibrée (i.e. $n/p-1<|B_j|<n/p+1$) de l'ensemble d'apprentissage. Moins formellement, il faut couper l'ensemble d'apprentissage en p parties, entraîner l'algorithme sur $p-1$ de ces parties, regarder la perte qu'encourt cet algorithme sur les données de la p-ème partie, et faire cela p fois (correspondant aux p parties pouvant être laissées de côté lors de l'apprentissage). Pour $p=n$, l'estimateur du risque est appelé *erreur ¡¡laisser-un-de-côté¿¿¿*.
Pour les k-p.p.v., chaque k définit un algorithme. Choisir k par validation croisée d'ordre p signifie choisir le k qui minimise l'erreur de validation croisée donnée par (VIII.5). En pratique, on prend k d'ordre 5 à 10 suivant les contraintes de temps de calcul.

Algorithme par noyau

Le théorème suivant donne un résultat de consistance universelle pour l'algorithme par noyau.

Théorème VIII.8. *On note par $\mathcal{B}(0,u)$ la boule euclidienne de $\mathbb{R}^d$ de centre 0 et de rayon $u>0$. Si il existe $0<r\leq R$ et $b>0$ tels que*

$$\forall u\in\mathbb{R}^d \quad b\mathbf{1}_{\mathcal{B}(0,r)}\leq K(u)\leq\mathbf{1}_{\mathcal{B}(0,R)}$$

et si $h_n\underset{n\to+\infty}{\longrightarrow}0$ et $nh_n^d\underset{n\to+\infty}{\longrightarrow}+\infty$, alors l'algorithme par noyau défini par

$$\hat{g}(x)=\sum_{i=1}^{n}\left(\frac{K(\frac{x-X_i}{h})}{\sum_{l=1}^{n}K(\frac{x-X_l}{h})}\right)Y_i$$

(avec la convention $\frac{0}{0}=0$) est universellement consistant.

Remarque VIII.9. C'est un algorithme très employé par les praticiens, notamment avec le noyau gaussien. Il est néanmoins à utiliser avec précaution pour des dimensions de l'espace d'entrées supérieures à 10. ◊

Algorithme par partition

Tout d'abord, rappelons que cet l'algorithme repose sur une partition de $\mathbb{R}^d$ $A_1,A_2,\ldots$ finie ou dénombrable et que pour tout $x\in\mathbb{R}^d$, nous notons $A(x)$ l'élément de la partition qui contient le point x. En pratique, cette partition est prise d'autant plus fine que la taille n de l'nesemble d'apprentissage est grande. Le théorème suivant indique une bonne manière de choisir la partition en fonction de n.

Théorème VIII.10. *On note encore par $\mathcal{B}(0,u)$ la boule euclidienne de $\mathbb{R}^d$ de centre 0 et de rayon $u>0$. Le diamètre de A_j est noté* $\mathrm{Diam}(A_j) = \sup_{x_1,x_2\in A_j} \|x_1-x_2\|$. *Si pour tout $R>0$*

$$\begin{cases} \max\limits_{j:A_j\cap\mathcal{B}(0,R)\neq\emptyset} \mathrm{Diam}(A_j) \underset{n\to+\infty}{\longrightarrow} 0 \\ \frac{|\{j:A_j\cap\mathcal{B}(0,R)\neq\emptyset\}|}{n} \underset{n\to+\infty}{\longrightarrow} 0 \end{cases}$$

alors l'algorithme par partition définie par pour tout $x\in\mathbb{R}^d$

$$\hat{g}(x) = \sum_{i=1}^n \left(\frac{\mathbf{1}_{X_i\in A(x)}}{\sum_{l=1}^n \mathbf{1}_{X_l\in A(x)}} \right) Y_i$$

(avec la convention $\frac{0}{0}=0$) est universellement consistant.

Remarque VIII.11. Pour une grille (régulière) de $\mathbb{R}^d$ de pas H_n, les hypothèses du théorème sont simplement $H_n \underset{n\to+\infty}{\longrightarrow} 0$ et $nH_n^d \underset{n\to+\infty}{\longrightarrow} +\infty$. En pratique, contrairement à l'algorithme par noyau gaussien, il y a un effet escalier (bien souvent non désiré) au bord des éléments de la partition, puisque la sortie prédite pour une entrée x change brusquement lorsque x change d'élément de la partition. ◊

Remarque VIII.12. Dans les arbres de décision, les éléments de la partition sont définis par les réponses à un ensemble de questions, ces questions étant éventuellement choisies en fonction de l'ensemble d'apprentissage observé. Chaque noeud d'un arbre de décision correspond à un ensemble de l'espace $\mathbb{R}^d$ des entrées. Les contraintes sur les parties de $\mathbb{R}^d$ associées aux noeuds sont les suivantes :

- la racine est associée à $\mathbb{R}^d$
- si $A\subset\mathbb{R}^d$ est l'ensemble associé à un noeud et si $A^{(1)},\dots,A^{(k)}$ sont les parties associées aux fils de ce noeud, alors

$$A^{(1)}\cup\cdots\cup A^{(k)} = A \qquad \text{et} \qquad \forall 1\le i<j\le k \quad A^{(i)}\cap A^{(j)}=\emptyset,$$

  autrement dit $A^{(1)},\dots,A^{(k)}$ est une partition de A.

Dans un arbre de décision binaire, tout noeud possède zéro ou deux fils. Donc tout noeud peut être vu comme une question sur l'entrée x dont la réponse est ¡¡oui¿¿ ou ¡¡non¿¿. Les questions typiques sur x sont : la i-ème composante de x est elle plus grande qu'un certain seuil (cf. figure VIII.3) ? De quel côté x est-il par rapport à un certain hyperplan de $\mathbb{R}^d$? x appartient-il à un certain hyperrectangle ?

L'arbre de décision est une variante de l'algorithme par partition qui est très utilisée notamment en raison de sa simplicité d'interprétation, de la rapidité de l'algorithme et de sa capacité à être mis à jour de manière dynamique (les branches de l'arbre peuvent être développées au fur et à mesure que de nouvelles données viennent compléter l'ensemble d'apprentissage). ◊

VIII.3.2 Algorithmes par minimisation du risque empirique

Principe de la minimisation du risque empirique

Rappelons tout d'abord que le risque d'une fonction de prédiction $g:\mathcal{X}\to\mathcal{Y}$ est défini par

$$R(g) = \mathbb{E}\ell[Y,g(X)].$$

Le but d'un algorithme d'apprentissage est de trouver une fonction de prédiction dont le risque est aussi faible que possible (autrement dit aussi proche que possible du risque des fonctions cibles).

La distribution $\mathbb{P}$ générant les données étant inconnue, le risque R et les fonctions cibles sont inconnus. Néanmoins, le risque $R(g)$ peut être estimé par son équivalent empirique

$$r(g) = \tfrac{1}{n} \sum_{i=1}^{n} \ell[Y_i, g(X_i)].$$

Si nous supposons $\mathbb{E}\big\{\ell[Y, g(X)]\big\}^2 < +\infty$, alors la L.F.G.N. et le T.C.L. permettent d'affirmer

$$\begin{aligned} r(g) &\underset{n\to+\infty}{\xrightarrow{\text{p.s.}}} R(g) \\ \sqrt{n}\big[r(g) - R(g)\big] &\underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}\big(0, \operatorname{Var} \ell[Y, g(X)]\big). \end{aligned}$$

Pour toute fonction de prédiction g, la variable aléatoire $r(g)$ effectue donc des déviations en $\mathrm{O}(1/\sqrt{n})$ autour de sa moyenne $R(g)$.

Puisque nous cherchons une fonction qui minimise le risque R et puisque ce risque est approché par le risque empirique r, il est naturel de considérer l'algorithme d'apprentissage, dit de *minimisation du risque empirique*, défini par

$$\hat{g}_{\mathrm{MRE}} \in \operatorname*{argmin}_{g\in\mathcal{G}} r(g), \tag{VIII.6}$$

où $\mathcal{G}$ est un sous-ensemble de $\mathcal{F}(\mathcal{X}, \mathcal{Y})$.

Prendre $\mathcal{G} = \mathcal{F}(\mathcal{X}, \mathcal{Y})$ n'est pas une bonne idée. Tout d'abord, cela entraîne un problème de choix puisqu'en général, pour tout ensemble d'apprentissage, il existe une infinité de fonctions de prédiction minimisant le risque empirique (voir figure VIII.4). Par ailleurs et surtout, si on prend l'algorithme du plus proche voisin comme minimiseur du risque empirique (en régression aux moindres carrés ou en classement), alors on peut montrer que cet algorithme est ¡¡loin¿¿ d'être universellement consistant.

Prendre $\mathcal{G} = \mathcal{F}(\mathcal{X}, \mathcal{Y})$ mène en général à un *surapprentissage* dans la mesure où l'algorithme résultant a un risque empirique qui peut être très inférieure à son risque réel (même lorsque la taille de l'ensemble d'apprentissage tend vers l'infini).

En pratique, il faut prendre $\mathcal{G}$ suffisamment grand pour pouvoir raisonnablement approcher toute fonction tout en ne le prenant pas trop grand pour éviter que l'algorithme ¡¡surapprenne¿¿. La ¡¡grandeur¿¿ de l'ensemble $\mathcal{G}$ est appelée *capacité* ou *complexité*. Un autre point de vue consiste à rajouter à $r(g)$ une pénalisation, quand par exemple, la fonction g est trop irrégulière. Ces deux approches sont en fait proches l'une de l'autre. L'approche par pénalisation sera adoptée pour les S.V.M. (Section VIII.3.2).

Soit $\tilde{g}$ une fonction minimisant le risque sur $\mathcal{G}$:

$$\tilde{g} \in \operatorname*{argmin}_{g\in\mathcal{G}} R(g). \tag{VIII.7}$$

On suppose le minimum atteint pour simplifier l'exposé. D'après l'inégalité

$$R(\hat{g}_{\mathrm{MRE}}) \geq R(\tilde{g}) \geq R(g^*),$$

L'excès de risque de $\hat{g}_{\mathrm{MRE}}$ se décompose en deux termes positifs, appelés *erreur d'estimation* et *erreur d'approximation* (ou *biais*) :

$$R(\hat{g}_{\mathrm{MRE}}) - R(g^*) = \underbrace{R(\hat{g}_{\mathrm{MRE}}) - R(\tilde{g})}_{\text{erreur d'estimation}} + \underbrace{R(\tilde{g}) - R(g^*)}_{\text{erreur d'approximation}} \quad ,$$

Plus $\mathcal{G}$ est grand, plus l'erreur d'approximation est faible mais plus l'erreur d'estimation est en général grande. Il y a donc un compromis à trouver dans le choix de $\mathcal{G}$. Ce compromis est souvent appelé *dilemme ¡¡biais-variance¿¿* , où le terme variance provient du lien entre l'erreur d'estimation et la variabilité de l'ensemble d'apprentissage que nous avons supposé dans notre formalisme être une réalisation de variables aléatoires i.i.d..

Réseaux de neurones

Les réseaux de neurones sont nés de la volonté de modéliser le fonctionnement du cerveau. Un neurone biologique reçoit des stimuli de ses voisins, et produit un signal lorsque son seuil d'activation est dépassé. La modélisation de Mc Culloch et Pitts (1943) considère que le neurone fait une combinaison linéaire de ses entrées, d'où la fonction d'activation

$$g(x) = \mathbf{1}_{\sum_{j=1}^{d} a_j x^{(j)} + a_0 > 0},$$

où les x_j sont les stimuli envoyés par les neurones voisins et $-a_0$ est le seuil d'activation.

Définition VIII.13. *Une sigmoïde σ est une fonction croissante telle que*

$$\begin{cases} \sigma(x) & \underset{x\to-\infty}{\longrightarrow} & 0 \\ \sigma(x) & \underset{x\to+\infty}{\longrightarrow} & 1 \end{cases}$$

Voici des exemples de sigmoïdes.

1. $\sigma(x) = \mathbf{1}_{x\geq 0}$
2. $\sigma(x) = \frac{1}{1+\exp(-x)}$
3. $\sigma(x) = \frac{1}{2} + \frac{1}{\pi}\arctan x$

Définition VIII.14.

– *Un* neurone *(artificiel) est une fonction définie sur $\mathbb{R}^d$ par*

$$g(x) = \sigma\big(\textstyle\sum_{j=1}^{d} a_j x^{(j)} + a_0\big) = \sigma(a \cdot \tilde{x})$$

où $a = (a_0, \cdots, a_d)^t$, $\tilde{x} = (1, x^{(1)}, \cdots, x^{(d)})^t$ et σ est une sigmoïde.

– *Un* réseau de neurones à une couche cachée *est une fonction $f : \mathbb{R}^d \to \mathbb{R}$ définie par*

$$f(x) = \textstyle\sum_{j=1}^{k} c_j \sigma(a_j \cdot \tilde{x}) + c_0$$

où $\tilde{x} = (1, x^{(1)}, \cdots, x^{(d)})^t$. La sigmoïde σ, le nombre de neurones k et les paramètres $a_1, \ldots, a_k \in \mathbb{R}^{d+1}$, $c_0, c_1, \ldots, c_k \in \mathbb{R}$ caractérisent le réseau.

Le théorème suivant donne un résultat de consistance universelle pour l'algorithme de minimisation du risque empirique sur un ensemble de réseaux de neurones à une couche cachée.

Théorème VIII.15. *Soient (k_n) une suite d'entiers et (β_n) une suite de réels. Soit $\mathcal{F}_n$ l'ensemble des réseaux de neurones à une couche cachée tels que $k \leq k_n$ et $\sum_{i=0}^{k} |c_i| \leq \beta_n$. L'algorithme $\hat{f}$ qui produit pour l'ensemble d'apprentissage Z_1^n la fonction de prédiction de $\mathcal{F}_n$ minimisant l'erreur quadratique empirique, i.e.*

$$\hat{f} \in \underset{f \in \mathcal{F}_n}{\operatorname{argmin}} \sum_{i=1}^{n} [Y_i - f(X_i)]^2,$$

est universellement consistant si $k_n \to +\infty$, $\beta_n \to +\infty$ et $\frac{k_n \beta_n^4 \log(k_n \beta_n^2)}{n} \underset{n \to +\infty}{\longrightarrow} 0$.

La minimisation de l'erreur quadratique empirique est un problème d'optimisation non convexe en raison de la sigmoïde. Par conséquent, il n'existe pas en général d'algorithme permettant d'obtenir systématique le minimum global. Il faut donc avoir recours à des heuristiques pour trouver les meilleurs minima locaux (et éventuellement) un minimum global.

L'algorithme des réseaux de neurones est une méthode d'apprentissage puissante ayant notamment donné d'excellents résultats sur les problèmes de reconnaissance de visages ([3]) et de reconnaissance de chiffres manuscrits ([9]). Des conseils sur la manière d'implémenter cet algorithme sont donnés dans [6, 5, 9].

Machines à Vecteurs Supports ou Séparateurs à Vastes Marges (S.V.M.)

Pour simplifier l'exposé, nous présentons cet algorithme dans le cadre du classement binaire avec $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{-1; +1\}$ et $\ell(y, y') = \mathbf{1}_{y \neq y'}$. Une fonction cible est alors

$$g^* : x \mapsto \begin{cases} +1 & \text{si } P(Y = +1 | X = x) \geq 1/2 \\ -1 & \text{sinon} \end{cases}$$

Soit $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ une fonction symétrique. Soit $\mathcal{H}$ l'espace vectoriel engendré par les fonctions

$$k(x, \cdot) : x' \mapsto k(x, x').$$

On suppose que

$$\langle \textstyle\sum_{1 \leq j \leq J} \alpha_j k(x_j, \cdot), \sum_{1 \leq k \leq K} \alpha'_k k(x'_k, \cdot) \rangle_{\mathcal{H}} = \sum_{j,k} \alpha_j \alpha'_k k(x_j, x'_k) \tag{VIII.8}$$

définit un produit scalaire sur $\mathcal{H}$, où $J, K \in \mathbb{N}$, α, α' sont des vecteurs quelconques de $\mathbb{R}^J$, et $x_1, \ldots, x_J$ et $x'_1, \ldots, x'_K$ sont respectivement des J-uplet et K-uplet d'éléments de $\mathcal{X}$.

C'est en particulier le cas pour les trois exemples suivants

- le noyau linéaire $k(x, x') = \langle x, x' \rangle_{\mathbb{R}^d}$,
- les noyaux polynômiaux $k(x, x') = \left(1 + \langle x, x' \rangle_{\mathbb{R}^d}\right)^p$ pour $p \in \mathbb{N}^*$,
- les noyaux gaussiens $k(x, x') = e^{-\|x - x'\|^2 / (2\sigma^2)}$ pour $\sigma > 0$.

Plus généralement, (VIII.8) définit un produit scalaire dès que k est semi-définie positive (i.e. pour tout $J \in \mathbb{N}$, $\alpha \in \mathbb{R}^J$ et tout J-uplet $x_1, \ldots, x_J$ de $\mathcal{X}$, $\sum_{1 \leq j,k \leq J} \alpha_j \alpha_k k(x_j, x_k) \geq 0$).

L'espace $\mathcal{H}$ muni du produit scalaire $\langle\ ,\ \rangle_{\mathcal{H}}$ est dit à *noyau reproduisant* car pour tout $f \in \mathcal{H}$, nous avons $\langle f, k(x, \cdot) \rangle_{\mathcal{H}} = f(x)$. La fonction k est appelée fonction *noyau*.

Définition VIII.16. *Soit "sgn" la fonction signe : $sgn(x) = \mathbf{1}_{x \geq 0} - \mathbf{1}_{x < 0}$. Notons u_+ la partie positive d'un réel u : $u_+ = \max(u, 0)$. Une machine à vecteur support de noyau k et*

de paramètre $C > 0$ est un algorithme d'apprentissage qui produit la fonction de prédiction $x \mapsto sgn\big\{\hat{h}(x) + \hat{b}\big\}$, où $(\hat{h}, \hat{b})$ est une solution du problème

$$\min_{b\in\mathbb{R}, h\in\mathcal{H}} \; C\sum_{i=1}^n \big(1 - Y_i[h(X_i) + b]\big)_+ + \tfrac{1}{2}\|h\|_{\mathcal{H}}^2 \qquad (\mathcal{P})$$

Si on impose $b = 0$ (et on ne minimise donc que sur $\mathcal{H}$), on parle de S.V.M. sans terme constant.

Remarque VIII.17. La machine à vecteur support de noyau linéaire $k(x, x') = \langle x, x'\rangle$ et de paramètre $C > 0$ est un algorithme d'apprentissage qui produit la fonction de prédiction $x \mapsto \mathrm{sgn}\big\{\langle w, x\rangle_{\mathbb{R}^d} + b\big\}$, où $w \in \mathbb{R}^d$ et $b \in \mathbb{R}$ minimisent

$$C\sum_{i=1}^n \big(1 - Y_i[\langle w, X_i\rangle_{\mathbb{R}^d} + b]\big)_+ + \tfrac{1}{2}\|w\|_{\mathbb{R}^d}^2.$$

$\Diamond$

Le théorème suivant donne un résultat de consistance universelle pour les S.V.M. à noyau gaussien.

Théorème VIII.18. *Soit $\mathcal{X}$ un compact de $\mathbb{R}^d$ et soit $\sigma > 0$. La S.V.M. (avec ou sans terme constant) de noyau gaussien $k : (x, x') \mapsto e^{-\|x-x'\|^2/(2\sigma^2)}$ et de paramètre $C = n^{\beta-1}$ avec $0 < \beta < 1/d$ est universellement consistante.*

Concentrons-nous à présent sur la résolution du problème d'optimisation $(\mathcal{P})$ et sur l'interprétation de la solution proposée. Des techniques d'optimisation classiques montrent que le problème d'optimisation $(\mathcal{P})$ sur un e.v. de dimension infinie se ramène (par passage à son problème dual) à un problème d'optimisation sur un e.v. de dimension n qui est quadratique et avec des contraintes linéaires simples.

Le théorème suivant explique comment calculer la solution de $(\mathcal{P})$ en pratique.

Théorème VIII.19. *Il existe (h, b) solution de $(\mathcal{P})$. De plus, h est unique et s'écrit*

$$h = \sum_{j=1}^n \alpha_j Y_j k(X_j, \cdot), \qquad \text{(VIII.9)}$$

où α est un vecteur de $\mathbb{R}^n$ qui maximise

$$\sum_{i=1}^n \alpha_i - \tfrac{1}{2}\sum_{1\le i,j\le n} \alpha_i\alpha_j Y_i Y_j k(X_i, X_j) \qquad (\mathcal{D})$$

sous les contraintes : $\sum_{i=1}^n \alpha_i Y_i = 0$ et $\forall i, 0 \le \alpha_i \le C$. Par ailleurs, si il existe une composante de α telle que $0 < \alpha_i < C$, alors $Y_i - h(X_i)$ est constant sur l'ensemble des i tel que $0 < \alpha_i < C$. Soit b la valeur commune. Le couple (h, b) est solution de $(\mathcal{P})$.

Idées de la preuve. Pour l'unicité du h solution de $(\mathcal{P})$, il faut montrer tout d'abord que h appartient nécessairement l'espace vectoriel engendrée par $k(X_1, \cdot), \ldots, k(X_n, \cdot)$, puis utiliser la stricte convexité de $h \mapsto \|h\|_{\mathcal{H}}^2$. (Attention, l'unicité de h n'implique pas en général celle des α_i.)

Pour obtenir $(\mathcal{D})$, la première étape consiste à remarquer que $(\mathcal{P})$ est égal à

$$\min_{\substack{b\in\mathbb{R}, h\in\mathcal{H}\\ \xi_i\ge 0\\ \xi_i\ge 1-Y_i[h(X_i)+b]}} \quad C\sum_{i=1}^n \xi_i + \tfrac{1}{2}\|h\|_{\mathcal{H}}^2,$$

ce qui ramène le problème convexe $(\mathcal{P})$ à un problème quadratique avec contraintes linéaires. La suite de la preuve consiste à écrire le dual de ce problème et à utiliser les conditions de Karush-Kuhn-Tucker. $\square$

Remarque VIII.20. La fonction de prédiction produite par une S.V.M. de noyau k peut donc s'écrire $x \mapsto \mathrm{sgn}\big\{\sum_{i=1}^n \alpha_i' k(X_i, x) + b\big\}$ où $\alpha_1', \ldots, \alpha_n'$ et b sont des paramètres judicieusement choisis. Une S.V.M. de noyau k réalise donc une (éventuellement partielle) séparation linéaire des données $X_1, \ldots, X_n$ représentées respectivement par $k(X_1, \cdot), \ldots, k(X_n, \cdot)$ dans l'espace vectoriel $\mathcal{H}$. Les α_i utilisés dans (VIII.9) et obtenus par résolution de $(\mathcal{D})$ sont intrinsèquement liés à la position de $k(X_i, \cdot)$ par rapport à l'hyperplan $h(x) + b = 0$ (voir figure VIII.5). Connaître α_i et Y_i permet de situer $k(X_i, \cdot)$ par rapport à la frontière de décision (d'équation $h(x) + b = 0$) sans avoir besoin de calculer $h(X_i) + b$. Cela permet une interprétation rapide des résultats d'une S.V.M.. Notons en particulier que

1. tout point mal classé X_i vérifie nécessairement $\alpha_i = C$ (réciproque fausse),
2. tout point X_i tel que $0 < \alpha_i < C$ se trouve sur les courbes d'équations respectives $h(x) + b = -1$ et $h(x) + b = +1$,
3. tout point X_i tel que $\alpha_i = 0$ n'influe pas sur la fonction de prédiction produite par une S.V.M. au sens où en retirant ces points de l'ensemble d'apprentissage et en recalculant la solution de $(\mathcal{D})$ pour ce nouvel ensemble d'apprentissage, nous retrouvons la même fonction de prédiction : $x \mapsto \mathrm{sgn}[h(x) + b]$.

$\Diamond$

Dans le contexte du classement binaire, l'algorithme des réseaux de neurones décrits dans le Théorème VIII.15 peut être également utilisé (par exemple en seuillant à 0 la fonction obtenue par minimisation du risque empirique quadratique : $\hat{g}(x) = +1$ si $\hat{f}(x) \geq 0$ et $\hat{g}(x) = -1$ sinon). L'inconvénient de cet algorithme d'apprentissage est qu'il doit résoudre un problème d'optimisation non convexe. Rien n'empêche en pratique l'algorithme de minimisation de tomber dans un minimum local. En conséquence, suivant l'implémentation et les heuristiques utilisées pour minimiser le risque empirique, des résultats très inégaux peuvent être observés.

L'algorithme des S.V.M. est en général privilégié par les praticiens car il possède les mêmes propriétés théoriques que l'algorithme des réseaux de neurones sans en avoir les inconvénients pratiques : les S.V.M. résolvent un problème d'optimisation convexe dont la solution est le plus souvent unique et facile à obtenir. De nombreux logiciels implémentant les S.V.M. dans différents langages informatiques sont accessibles en ligne.

Choix du paramètre C et du noyau k. Pour $\mathcal{X} = \mathbb{R}^d$, les noyaux les plus populaires sont

- le noyau linéaire $(x, x') \mapsto \langle x, x' \rangle_{\mathbb{R}^d}$,
- les noyaux polynômiaux $(x, x') \mapsto \big(1 + \langle x, x' \rangle_{\mathbb{R}^d}\big)^p$ pour $p \in \mathbb{N}^*$,
- les noyaux gaussiens $(x, x') \mapsto e^{-\|x - x'\|^2/(2\sigma^2)}$ pour $\sigma > 0$.

Attention, seul le noyau gaussien peut mener à un algorithme universellement consistant. Il doit donc être en général préféré en pratique.

Pour choisir C, la méthode la plus employée est de faire une validation croisée. Pour limiter le temps de calcul, il faut prendre tout d'abord une grille géométrique grossière (par exemple, $\{10^5/n,\ 10^4/n,\ 10^3/n,\ 10^2/n,\ 10/n\}$), puis prendre une grille géométrique plus fine autour du C ayant donné la plus faible erreur de validation croisée (cf. (VIII.5)), à condition que ce C ne soit pas sur les bords de la grille grossière.

Pour choisir la largeur σ du noyau gaussien, le principe est le même que pour choisir C sauf que l'intervalle de recherche est $[\sigma_{\min}; \sigma_{\max}]$, où $\sigma_{\min}$ est la médiane de l'ensemble des distances d'un point de l'ensemble d'apprentissage à son plus proche voisin et $\sigma_{\max}$ est le diamètre de $\{X_1, \ldots, X_n\}$ pour la norme euclidienne sur $\mathbb{R}^d$, i.e. $\sigma_{\max} = \max_{i,j} \|X_i - X_j\|_{\mathbb{R}^d}$.

VIII.4 Au delà de la consistance universelle

Les résultats de consistance universelle sont des résultats asymptotiques. Ils disent juste que si nous avons suffisamment de données (et un ordinateur suffisamment puissant pour les traiter) alors tout algorithme universellement consistant produira une prédiction très proche de la meilleure prédiction possible.

Les résultats de consistance universelle ne disent pas le nombre de données nécessaires pour avoir une garantie du type $\mathbb{E}R(\hat{g}) \leq R(g^*) + \epsilon$ pour $\epsilon > 0$ fixé. Pour que ce nombre existe, il faudrait avoir un résultat de consistance universelle uniforme, i.e.

$$\lim_{n\to+\infty} \sup_{\mathbb{P}} \left\{\mathbb{E}R(\hat{g}) - R(g^*)\right\} = 0,$$

la consistance universelle n'affirmant que

$$\sup_{\mathbb{P}} \lim_{n\to+\infty} \left\{\mathbb{E}R(\hat{g}) - R(g^*)\right\} = 0.$$

En général, ce nombre n'existe pas d'après le théorème suivant.

Théorème VIII.21. *Si $|\mathcal{X}| = +\infty$, il n'existe pas d'algorithme d'apprentissage uniformément universellement consistant ni en régression aux moindres carrés ($\ell(y, y') = (y - y')^2$) ni en classement ($\ell(y, y') = \mathbf{1}_{y\neq y'}$).*

L'absence d'algorithme universellement uniformément consistant nous amène à définir un ¡¡bon¿¿ algorithme d'apprentissage comme étant un algorithme universellement consistant et ayant une propriété de convergence uniforme sur une classe de probabilités paraissant pertinente pour le problème à traiter. Plus précisément, si $\mathcal{P}$ est un ensemble de probabilités sur $\mathcal{Z}$ dans laquelle nous pensons que $\mathbb{P}$ est, nous souhaitons que le bon algorithme satisfasse

$$\lim_{n\to+\infty} \sup_{\mathbb{P}\in\mathcal{P}} \left\{\mathbb{E}R(\hat{g}) - R(g^*)\right\} = 0$$

et également avoir une suite $\sup_{\mathbb{P}\in\mathcal{P}} \left\{\mathbb{E}R(\hat{g}) - R(g^*)\right\}$ décroissant le plus vite possible vers 0 pour que peu de données soient nécessaires à l'algorithme pour prédire efficacement. L'ensemble $\mathcal{P}$ doit être pensé comme une modélisation de notre a priori, et il en résulte un a priori implicite sur la fonction cible. L'obtention d'algorithmes incorporant un a priori et étant efficace lorsque l'a priori est correct est au coeur de la recherche actuelle en apprentissage statistique.

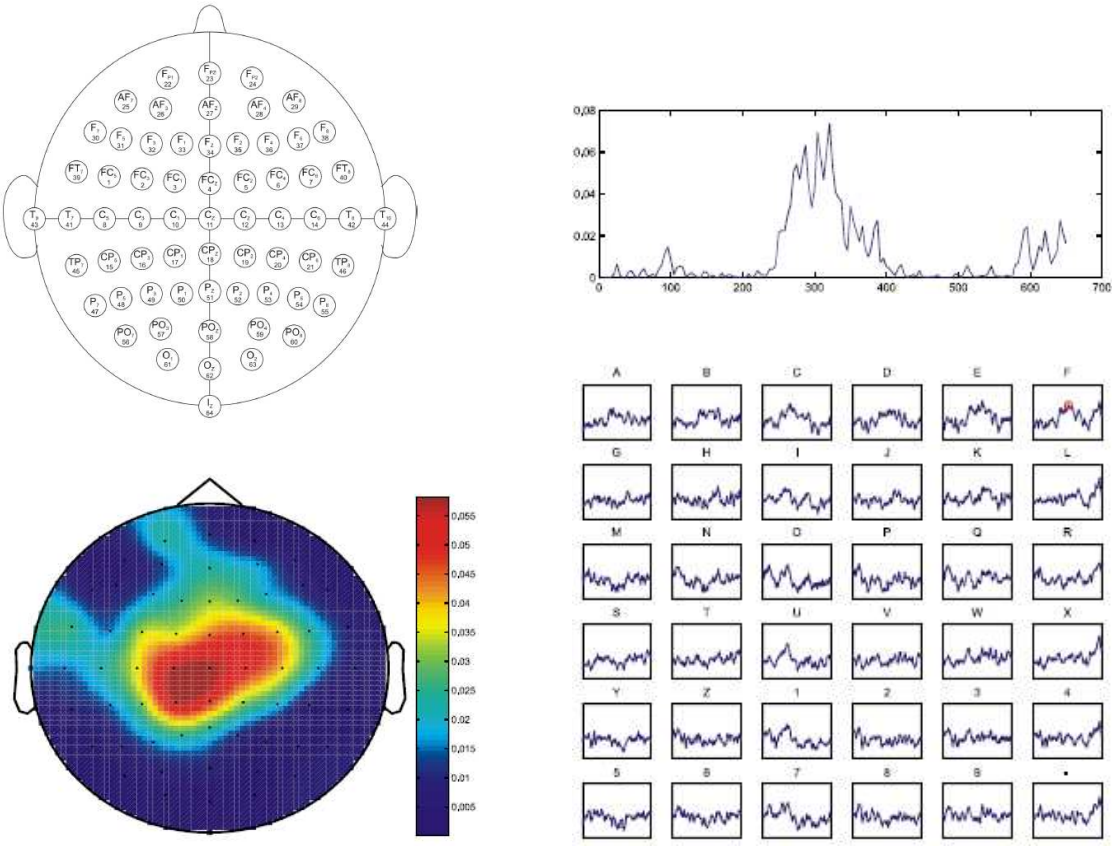


FIG. VIII.2 – *En haut à gauche :* exemple d'emplacement d'électrodes sur la surface du cerveau du sujet vu de haut. *En haut à droite :* exemple de signal électrique observé à une électrode durant les 700 ms suivant le stimulus. L'activité observée environ 300 ms après le stimulus est caractéristique de la réaction du sujet au stimulus. *En bas à gauche :* topographie (vue de haut) de la variance du signal due à la présence ou non de stimuli. Cette topographie montre que le signal est relativement peu localisé en un point précis de la surface du cerveau. *En bas à droite :* signaux électriques moyens observés à une électrode pour différents stimuli [1, 8].

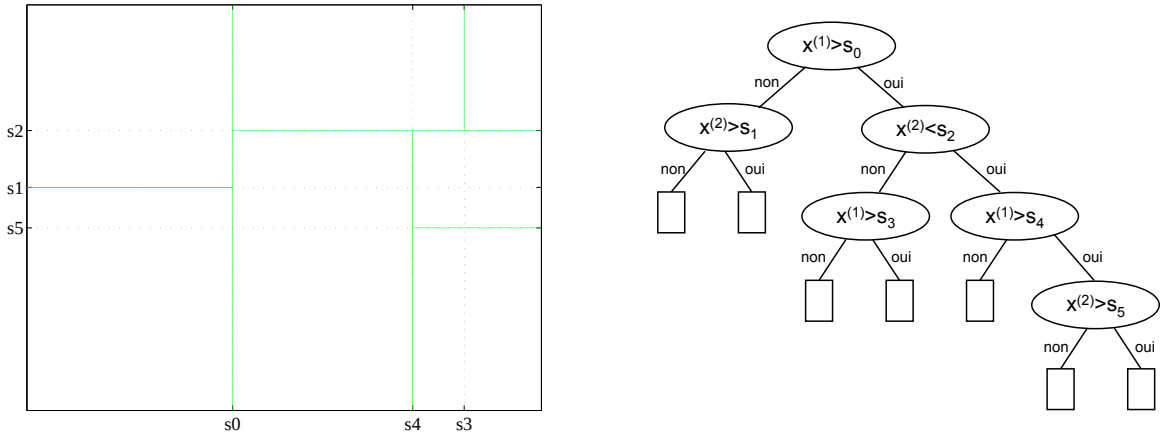


FIG. VIII.3 – *A gauche :* Exemple de partition provenant d'un arbre de décision. *A droite :* Arbre de décision correspond à cette partition.

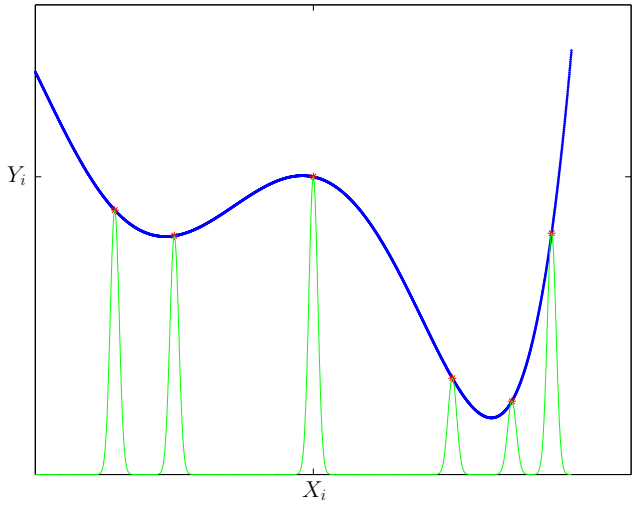


FIG. VIII.4 – Surapprentissage en régression aux moindres carrés : les couples entrées-sorties sont représentés par des croix. Les deux courbes minimisent le risque empirique $\frac{1}{n}\sum_{i=1}^{n}[Y_i - g(X_i)]^2$ (puisqu'elles ont toutes les deux un risque empirique nulle). La courbe fine semble apprendre par coeur la valeur des sorties associées aux entrées de l'ensemble d'apprentissage. On dit qu'elle "surapprend". Au contraire, la courbe épaisse explique plus simplement les données.

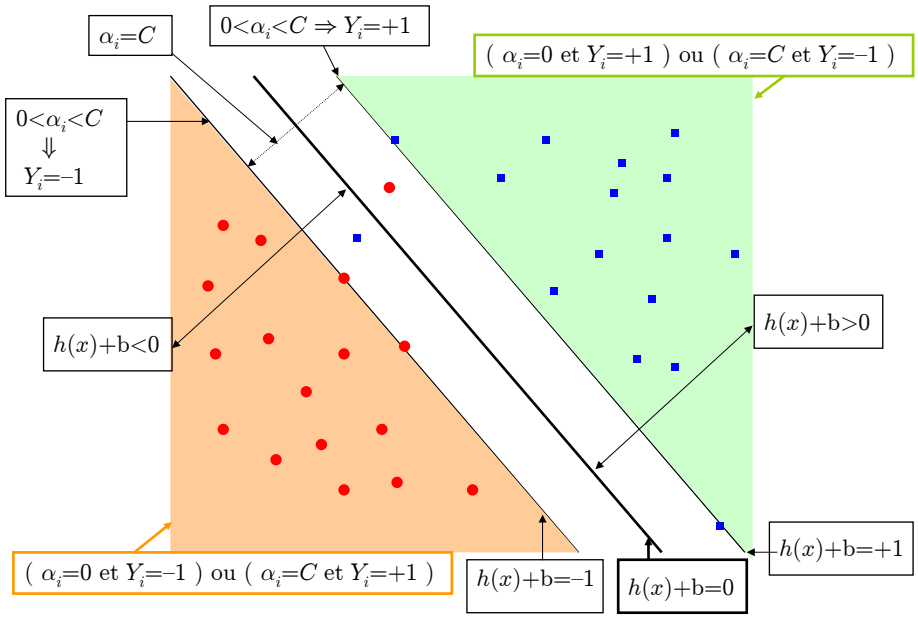


FIG. VIII.5 – Interprétation géométrique des S.V.M.. Le plan du dessin représente l'espace vectoriel (de fonctions de $\mathcal{X}$ dans $\mathbb{R}$) engendré par les fonctions $k(X_1,\cdot),\ldots,k(X_n,\cdot)$. Les coefficients α_i sont ceux obtenues par résolution de $(\mathcal{D})$ et le couple (h,b) est la solution présentée dans le Théorème VIII.19 : $h = \sum_{j=1}^{n} \alpha_j Y_j k(X_j,\cdot)$. Les points ronds et carrés représentent les points de l'ensemble d'apprentissage (par le biais des fonctions $k(X_i,\cdot)$ qui leur sont associées).

Bibliographie

[1] B. Blankertz, K.-R. Mller, G. Curio, T. Vaughan, G. Schalk, J. Wolpaw, A. Schlgl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schrder, and N. Birbaumer. Bci competition 2003 : Progress and perspectives in detection and discrimination of eeg single trials. *IEEE Transactions on Biomedical Engineering*, 51(6) :1044–1051, 2004.

[2] D. DeCoste and B. Schölkopf. Training invariant support vector machines. *Machine Learning*, 46 :161–190, 2002.

[3] C. Garcia and M. Delakis. Convolutional face finder : A neural architecture for fast and robust face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11) :1408–1423, 2004. article non accessible en ligne, me le demander.

[4] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2004.

[5] Y. LeCun, 2005. `http://www.cs.nyu.edu/~yann/2005f-G22-2565-001/diglib/lecture09-optim.djvu`, requires the djvu viewer `http://djvu.org/download/`.

[6] Y. LeCun, L. Bottou, G. Orr, and K. Muller. Efficient backprop, `http://yann.lecun.com/exdb/publis/pdf/lecun-98b.pdf`. In G. Orr and M. K., editors, *Neural Networks : Tricks of the trade*. Springer, 1998.

[7] Y. LeCun and C. Cortes. MNIST page. `http://yann.lecun.com/exdb/mnist/`.

[8] G. Schalk, D. McFarland, T. Hinterberger, N. Birbaumer, and J. Wolpaw. BCI2000 : a general-purpose brain-computer interface system. *IEEE Transactions on Biomedical Engineering*, 51(6) :1034–1043, 2004.

[9] P. Simard, D. Steinkraus, and J. Platt. Best practice for convolutional neural networks applied to visual document analysis, `http://research.microsoft.com/~patrice/PDF/fugu9.pdf`. *International Conference on Document Analysis and Recogntion (ICDAR), IEEE Computer Society*, pages 958–962, 2003.

Troisième partie

Analyse exploratoire

Chapitre IX

Analyse des données

IX.1 Introduction

IX.1.1 Objectif

Dans toute étude appliquée, la démarche première du statisticien est de décrire et d'explorer les données dont il dispose, avant d'en tirer de quelconques lois ou modèles prédictifs. Or la statistique traite généralement du grand nombre et, les outils informatiques aidant, les bases de données deviennent de plus en plus volumineuses, tant en largeur (quantité d'informations recueillies) qu'en hauteur (nombre d'unités sur lesquelles ces informations sont recueillies).

Cette phase d'exploration descriptive des données n'est en conséquence pas aisée : si le statisticien est déjà outillé pour analyser la distribution, sur la population d'étude, d'une variable quelconque, ou la relation entre deux variables quelles qu'elles soient - et qu'il peut donc développer séquentiellement cette démarche à l'ensemble des informations présentes dans ses données -, ces outils basiques ne permettent pas d'appréhender ce vaste ensemble informatif dans sa globalité. Il ne s'agit naturellement pas d'en donner alors une vision exhaustive, mais bien de répondre à l'une des principales missions du statisticien : extraire d'une masse de données ce qu'il faut en retenir, en la synthétisant ou en simplifiant les structures - le tout avec un souci de neutralité objective qui en garantit la crédibilité.

Les techniques d'analyse de données répondent à ce besoin, et en particulier les deux présentées ici :

- l'Analyse en Composantes Principales (ACP), aînée des méthodes d'analyse factorielle qui s'appuient sur la réduction de rang découlant des travaux de décomposition matricielle d'Eckart et Young (1936), détermine les principales relations linéaires dans un ensemble de variables numériques ;
- les méthodes de classification automatique permettent de résoudre le problème de l'appréhension des "individus", en les regroupant au sein de classes homogènes, sur la base d'informations communes.

Dans chaque cas, il s'agit bien de réduire un ensemble complexe et de grande dimension à ses principaux éléments, de façon à en mieux comprendre les structures sous-jacentes.

IX.1.2 Notations

On dispose de p variables ou caractères $X^1, \ldots, X^j, \ldots, X^p$, que l'on observe sur n unités statistiques - ou individus : on note X_i^j la valeur de la variable X^j observée sur le i-ième individu. Cet ensemble de données dit actif peut donc être mis sous la forme d'un tableau X à n lignes et p colonnes, et de terme courant X_i^j.

Dans la suite - et c'est très généralement le cas en analyse des données, contrairement aux autres domaines de la statistique - on confondra la notion de variable avec le vecteur de dimension n qui la définit sur notre population active : X^j ; de même, chaque individu sera assimilé au vecteur de dimension p qui compile ses valeurs sur les variables actives : X_i.

Chaque individu est affecté d'un poids m_i (tel que $m_i > 0$ et $\sum_{i=1}^{n} m_i = 1$). Ces poids peuvent résulter d'un plan de sondage, ou bien traduire une "taille" nécessaire à la problématique (nombre d'habitants d'une ville, chiffre d'affaire d'une entreprise, etc.).

IX.1.3 Exemple

Nous choisirons ici un exemple emprunté à Michel Tenenhaus[1], décrivant les caractéristiques techniques de 24 modèles de voitures de l'année 1989.

Cet exemple comporte volontairement un nombre réduit d'individus (voitures) et de variables (caractéristiques), pour en faciliter la compréhension.

Pour comprendre ce qu'apportent les méthodes d'analyse de données, menons au préalable une brève analyse descriptive de ce tableau, tant du point de vue des individus que des variables.

[1] Tenenhaus M. (1994), Méthodes statistiques en gestion, Dunod

Modèle	Cylindre (cm3)	Puissance (ch)	Vitesse (km/h)	Poids (kg)	Longueur (cm)	Largeur (cm)
Honda Civic	1396	90	174	850	369	166
Renault 19	1721	92	180	965	415	169
Fiat Tipo	1580	83	170	970	395	170
Peugeot 405	1769	90	180	1080	440	169
Renault 21	2068	88	180	1135	446	170
Citroën BX	1769	90	182	1060	424	168
Bmw 530i	2986	188	226	1510	472	175
Rover 827i	2675	177	222	1365	469	175
Renault 25	2548	182	226	1350	471	180
Opel Omega	1998	122	190	1255	473	177
Peugeot 405 Break	1905	125	194	1120	439	171
Ford Sierra	1993	115	185	1190	451	172
Bmw 325iX	2494	171	208	1300	432	164
Audi 90 Quattro	1994	160	214	1220	439	169
Ford Scorpio	2933	150	200	1345	466	176
Renault Espace	1995	120	177	1265	436	177
Nissan Vanette	1952	87	144	1430	436	169
VW Caravelle	2109	112	149	1320	457	184
Ford Fiesta	1117	50	135	810	371	162
Fiat Uno	1116	58	145	780	364	155
Peugeot 205	1580	80	159	880	370	156
Peugeot 205 Rallye	1294	103	189	805	370	157
Seat Ibiza SX I	1461	100	181	925	363	161
Citroën AX Sport	1294	95	184	730	350	160

Etude descriptive des individus

Vu le faible nombre de variables (qui seront souvent plus nombreuses, dans les problématiques relevant de l'analyse des données), on peut représenter chaque individu par un graphique en étoile (cf. Fig.IX.1, Fig.IX.2), chaque valeur étant reportée sur la branche correspondante (qui va de la valeur minimale à la valeur maximale observée).

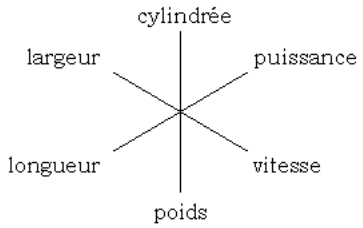


FIG. IX.1 – Graphique en étoile

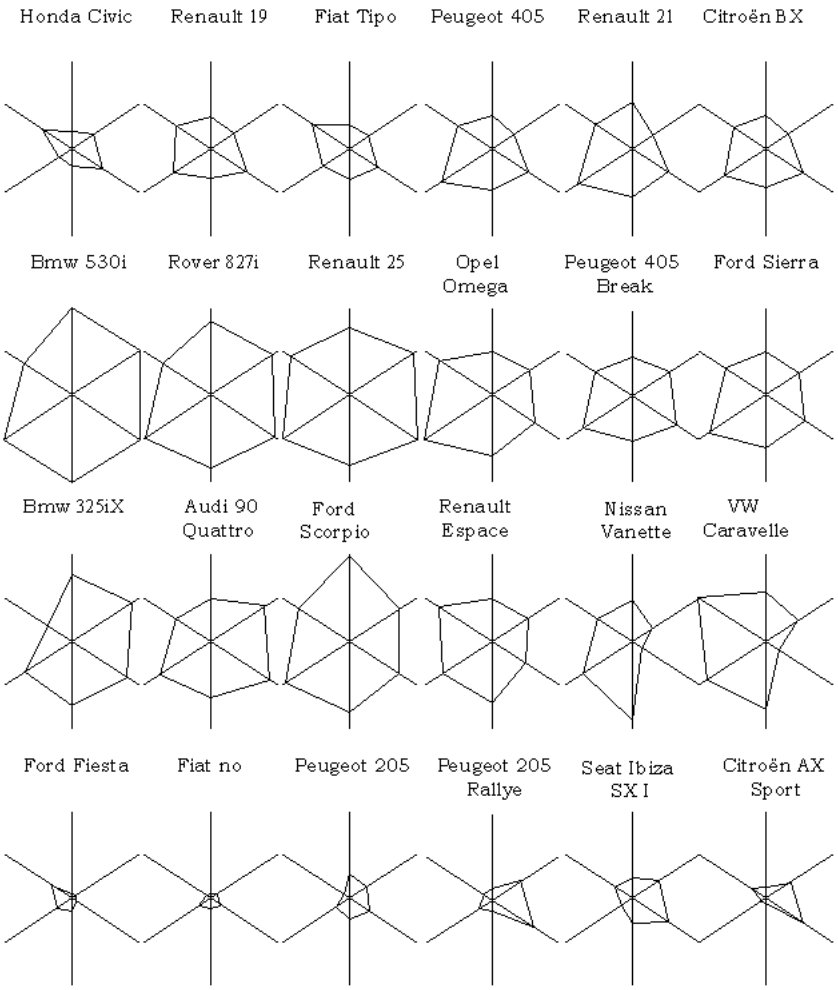


FIG. IX.2 – Graphiques en étoile des différents modèles

On constate que les étoiles sont plus ou moins grosses, mais généralement harmonieuses : toutes les caractéristiques évoluent globalement dans le même sens.

On peut cependant distinguer certains modèles, dont l'étoile a une forme particulière :

- les petites voitures sportives (en bas à droite), qui sont particulièrement rapides par rapport à leur gabarit ;
- les "vans" (Nissan Vanette, VW Caravelle), qui sont particulièrement lents pour le leur.

Etude descriptive des variables

Classiquement, on peut se livrer à une analyse de la distribution de chaque variable :

	Cylindrée (cm3)	Puissance (ch)	Vitesse (km/h)	Poids (kg)	Longueur (cm)	Largeur (cm)
Minimum	1116	50	135	730	350	155
1er quartile	1550	90	173	914	371	164
Médiane	1929	102	182	1128	436	169
3ème quartile	2078	131	196	1305	453	175
Maximum	2986	188	226	1510	473	184
Moyenne	1906	114	183	1111	422	169
Ecart-type	517	38	25	225	40	7

Ce tableau est cependant relativement peu parlant pour le non-spécialiste, à cause notamment de l'hétérogénéité des variables. Plus visuellement, pour mieux appréhender la forme de ces distributions, on peut tracer des histogrammes (cf. Fig.IX.3).

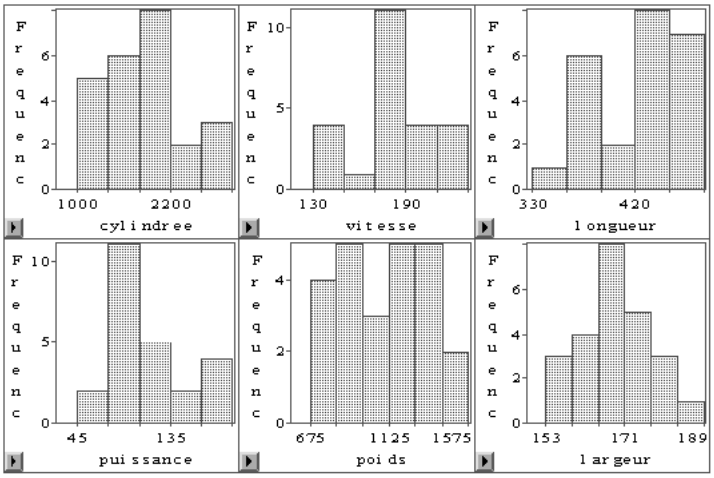


FIG. IX.3 – Histogramme des différentes variables

Plus intéressant est sans doute l'étude de la relation entre ces différentes variables. Des outils d'analyse connus peuvent ainsi être mis en œuvre pour appréhender les distributions bivariées, comme le "scatter plot" (cf. Fig.IX.4), ou encore le calcul des coefficients de corrélation linéaire (Bravais-Pearson), après avoir vérifié que la forme linéaire des nuages les légitimait. La corrélation linéaire de deux variables d'intérêt $Y = (Y_i, 1 \leq i \leq n)$ et $Z = (Z_i, 1 \leq i \leq n)$ est donnée par

$$\mathrm{Corr}\,(Y,Z) = \frac{\mathrm{Cov}\,(Y,Z)}{\sigma_Y \sigma_Z} = \frac{\sum_{i=1}^n m_i \left(Y_i - \bar{Y}\right)\left(Z_i - \bar{Z}\right)}{\sqrt{\sum_{i=1}^n m_i \left(Y_i - \bar{Y}\right)^2}\sqrt{\sum_{i=1}^n m_i \left(Z_i - \bar{Z}\right)^2}},$$

où $\bar{Y} = \sum_{i=1}^n m_i Y_i$ et $\bar{Z} = \sum_{i=1}^n m_i Z_i$.

Voici le tableau des corrélations linéaires :

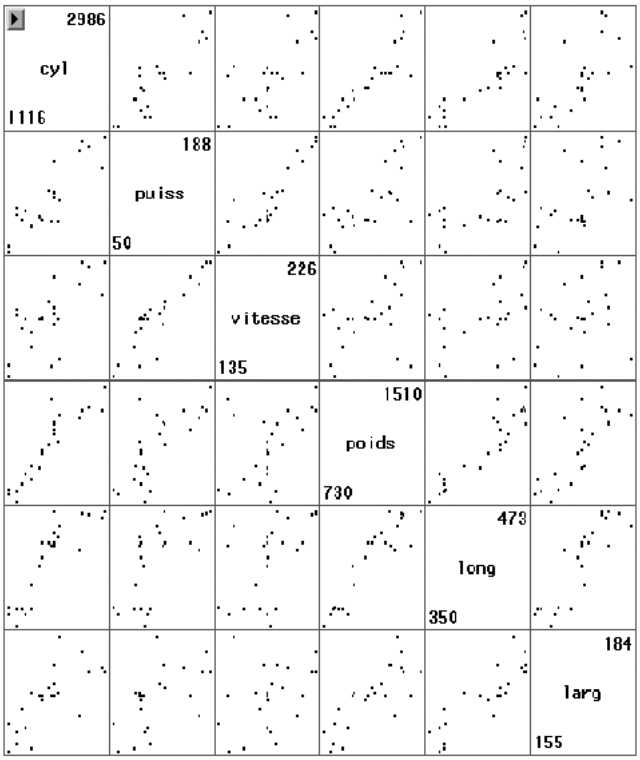


FIG. IX.4 – Scatter plot des différentes variables

	Cylindrée	Puissance	Vitesse	Poids	Longueur	Largeur
Cylindrée	1	0,86	0,69	0,90	0,86	0,71
Puissance	0,86	1	0,89	0,75	0,69	0,55
Vitesse	0,69	0,89	1	0,49	0,53	0,36
Poids	0,90	0,75	0,49	1	0,92	0,79
Longueur	0,86	0,69	0,53	0,92	1	0,86
Largeur	0,71	0,55	0,36	0,79	0,86	1

Toutes les corrélations calculées sont positives, ce qui indique que les 6 variables sont globalement corrélées - l'harmonie des étoiles le laissait déjà présager...

Ce tableau peut être assimilé à un tableau de proximités entre variables :

– cylindrée, poids et longueur sont particulièrement proches ;
– vitesse et puissance sont proches ;
– longueur et largeur le sont aussi.

Manquent cependant des outils de synthèse, qui permettraient de dégager la structure essentielle de ces données : nous allons en développer deux, parmi les plus puissants.

IX.2 L'Analyse en Composantes Principales

IX.2.1 Problématique

On se place ici dans la situation où les p variables d'intérêt $X^1, \ldots, X^j, \ldots, X^p$, sont numériques. Pour appréhender l'information contenue dans le tableau numérique X, on peut tenter de visualiser l'un ou l'autre des nuages de points qui en découlent : les p variables, dans $\mathbb{R}^n$; ou, plus classiquement[2], les n individus, dans $\mathbb{R}^p$. Mais très souvent, le nombre d'individus n est grand (plusieurs centaines, voire plusieurs milliers), et le nombre de variables p peut atteindre quelques dizaines. Quoiqu'il en soit, même avec des outils de visualisation performants, X ne peut être appréhendé de façon simple dans sa globalité, ni les relations entre les variables.

La problématique est alors double :
- Comment visualiser la forme du nuage des individus ?
- Comment synthétiser les relations entre variables ?

L'ACP permet justement de répondre à ce type de besoin.

IX.2.2 Choix de la métrique

La méthode d'Analyse en Composantes Principales requiert un espace vectoriel muni d'un produit scalaire. Dans ce chapitre, nous considérerons l'espace euclidien $\mathbb{R}^p$ muni de son produit scalaire canonique. La métrique associée est donnée par

$$\|X_i - X_{i'}\|^2 = \sum_{j=1}^{p} \left(X_i^j - X_{i'}^j\right)^2.$$

Définition IX.1. *Soient* $\bar{X}^j = \sum_{i=1}^n m_i X_i^j$ *et* $\sigma_j^2 = \sum_{i=1}^n m_i \left(X_i^j - \bar{X}^j\right)^2$ *la moyenne et la variance de la variable d'intérêt* X^j*. La* **représentation réduite** *de l'individu* i *est donnée par* $\tilde{X}_1^j, \ldots, \tilde{X}_p^j$*, où pour tout* $1 \leq j \leq p$*,*

$$\tilde{X}_i^j = \frac{X_i^j - \bar{X}^j}{\sigma_j}.$$

Une **ACP normée** *est une ACP menée sur la représentation réduite.*

Les différentes variables X^j pouvant être hétérogènes, et correspondre à des échelles de mesure disparates, la représentation réduite est utilisée pour éviter que le choix de ces unités ait une influence dans le calcul des distances. Cette représentation rend les variables centrées et de variance 1, et égale donc leur contribution au calcul des distances[3]

$$\|\tilde{X}_i - \tilde{X}_{i'}\|^2 = \sum_{j=1}^{p} \left(\frac{X_i^j}{\sigma_j} - \frac{X_{i'}^j}{\sigma_j}\right)^2 = \sum_{j=1}^{p} \frac{\left(X_i^j - X_{i'}^j\right)^2}{\sigma_j^2}$$

Pour simplifier la présentation, on considérera dans la suite que les variables ont été déjà réduites.

[2] C'est en effet l'extension de la démarche naturellement adoptée lors des analyses bivariées, où la forme des nuages-individus dans une base de variables est la traduction géométrique des liaisons entre ces variables.

[3] La métrique associée pour la représentation initiale est dite "inverse des variances".

IX.2.3 Moindre déformation du nuage

Pour visualiser le nuage des individus (et donc en connaître la forme, pour savoir comment sont liées nos p variables), il est nécessaire de réduire la dimension de l'espace qui le porte. Les méthodes d'analyse factorielle (dont l'ACP fait partie) réduisent cette dimension par projection orthogonale sur des sous-espaces affines.

Inerties d'un nuage de points

Définition IX.2. *Soit G le barycentre d'un nuage de points $X_1, \dots, X_n$ pondérés par les poids $m_1, \dots, m_n$. L'***inertie** *du nuage $X_1, \dots, X_n$ est donnée par*

$$I = \sum_{i=1}^{n} m_i \|X_i - G\|^2.$$

*L'***inertie** *J_H du nuage* **autour du sous-espace affine** *H est donnée par*

$$J_H = \sum_{i=1}^{n} m_i \|X_i - X_i^*\|^2,$$

où X_i^* le projeté orthogonal de X_i sur H.

L'inertie J_H autour de H mesure la déformation du nuage lorsque celui-ci est projeté orthogonalement sur H. Pour que la représentation des données par leur projection sur un sous-espace affine ait un sens, il faut qu'elle modifie peu la forme du nuage de points, donc qu'elle minimise l'inertie J_H.

Théorème IX.3. *Soit $\mathcal{H}$ une famille de sous-espaces affines invariante par translation (autrement dit, si $H \in \mathcal{H}$, alors tout sous-espace affine parallèle à H appartient à $\mathcal{H}$). Soit H_{opt} un sous-espace affine minimisant la déformation du nuage projeté orthogonalement sur un élément de $\mathcal{H}$, au sens où $J_{H_{opt}} = \min_{H \in \mathcal{H}} J_H$. Alors :*

- *Le centre de gravité*[4] *G du nuage de points appartient à H_{opt}.*
- *Le sous-espace affine H_{opt} maximise l'inertie du nuage projeté :*

$$I_{H_{opt}} = \max_{H \in \mathcal{H}} I_H,$$

  *où $I_H = \sum_{i=1}^{n} m_i \|X_i^* - G^*\|^2$ est l'***inertie du nuage projeté** *orthogonalement sur H.*
- *Pour $H \in \mathcal{H}$, la moyenne pondérée des carrés des distances entre les points du nuage projeté : $\sum_{i \neq i'} m_i m_{i'} \|X_i^* - X_{i'}^*\|^2$, est maximale pour $H = H_{opt}$.*

Démonstration. Soit $H \in \mathcal{H}$. Soit G^* le projeté orthogonal du barycentre G sur H, et H_G le sous-espace parallèle à H passant par G. Par hypothèse d'invariance par translation de $\mathcal{H}$, on a $H_G \in \mathcal{H}$. La projection orthogonale sur H_G n'est autre que celle sur H, translatée du

[4] Dans la représentation réduite, le centre de gravité est au centre du repère : $G(\bar{X}^1, \dots, \bar{X}^p) = O(0, \dots, 0)$.

vecteur $G - G^*$. En conséquence, on a

$$\begin{aligned} J_{H_G} &= \sum_{i=1}^{n} m_i \|X_i - (X_i^* + G - G^*)\|^2 \\ &= \sum_{i=1}^{n} m_i \|(X_i - X_i^*) - (G - G^*)\|^2 \\ &= \sum_{i=1}^{n} m_i \left(\|X_i - X_i^*\|^2 - 2\langle X_i - X_i^*, G - G^* \rangle + \|G - G^*\|^2 \right) \\ &= \sum_{i=1}^{n} m_i \left(\|X_i - X_i^*\|^2 - \|G - G^*\|^2 \right), \end{aligned}$$

d'où $J_{H_G} = J_H - \|G - G^*\|^2$. De cette égalité, constituant le théorème de Huyghens, découle la première assertion du théorème[5].

Pour la deuxième asssertion, considérons un sous-espace affine H passant par G. Par le théorème de Pythagore, on a

$$\begin{aligned} J_H &= \sum_{i=1}^{n} m_i \|X_i - X_i^*\|^2 \\ &= \sum_{i=1}^{n} m_i \left(\|X_i - G\|^2 - \|X_i^* - G\|^2 \right) \\ &= I - I_H, \end{aligned}$$

où I est l'inertie du nuage, et I_H l'inertie du nuage projeté orthogonalement sur H. L'inertie I étant indépendante du sous-espace affine considéré, on obtient qu'elle est maximale pour $H = H_{\text{opt}}$.

Pour la dernière assertion, il suffit de remarquer la relation suivante entre la moyenne pondérée des carrés des distances entre les points du nuage projeté, et l'inertie du nuage projeté :

$$\begin{aligned} \sum_{i \neq i'} m_i m_{i'} \|X_i^* - X_{i'}^*\|^2 &= \sum_{i=1}^{n} \sum_{i'=1}^{n} m_i m_{i'} \left\| (X_i^* - G^*) - (X_{i'}^* - G^*) \right\|^2 \\ &= \sum_{i'=1}^{n} m_{i'} \sum_{i=1}^{n} m_i \|X_i^* - G^*\|^2 + \sum_{i=1}^{n} m_i \sum_{i'=1}^{n} m_{i'} \|X_{i'}^* - G^*\|^2 \\ &\qquad - 2 \left\langle \sum_{i=1}^{n} m_i (X_i^* - G^*), \sum_{i'=1}^{n} m_{i'} (X_{i'}^* - G^*) \right\rangle \\ &= I_H + I_H + 0 \\ &= 2 I_H. \end{aligned}$$

□

[5]Puisque nous travaillons dans la représentation réduite, nous aurions pu omettre G. Par souci d'homogénéité des formules, nous l'avons conservé dans le calcul précédent.

En résumé, la moindre déformation d'un nuage de points par projection orthogonale sur un sous-espace affine est obtenue, de manière équivalente, par minimisation de l'inertie par rapport au sous-espace affine, par maximisation de l'inertie du nuage projeté, ou par maximisation de la somme des distances entre les points projetés[6].

Résolution séquentielle et décomposition de l'inertie

Cette section montre que la recherche d'un sous-espace affine de dimension fixée maximisant l'inertie du nuage projeté peut être menée de manière séquentielle et que l'inertie se décompose en la somme des inerties du nuage projeté sur des droites orthogonales, dites directions principales de l'ACP.

Puisque l'inertie du nuage projeté sur un sous-espace affine est invariante par translation de ce sous-espace, nous assimilerons dans cette section le sous-espace affine avec son sous-espace vectoriel associé.

Désignons par $\mathcal{H}_k$ l'ensemble des sous-espaces vectoriels de dimension k avec par convention $\mathcal{H}_0 = \{\{0\}\}$. Soit H_k le sous-espace vectoriel de $\mathcal{H}_k$ portant l'inertie maximale, au sens où $I_{H_k} = \max_{H \in \mathcal{H}_k} I_H$.

L'orthogonal d'un sous-espace vectoriel H de $\mathbb{R}^p$ est défini par $H^\perp = \{u \in \mathbb{R}^p : u \perp H\}$.

Théorème IX.4. *Soit u_{k+1} le vecteur de $H_k^\perp$ maximisant l'inertie du nuage projeté sur la droite (G, u_{k+1}). On a*

$$H_{k+1} = H_k \oplus u_{k+1},$$

d'où

$$H_k = u_1 \oplus \cdots \oplus u_k.$$

Ce théorème est basé sur la propriété suivante des inerties.

Proposition IX.5. *Si E et F sont deux sous-espaces vectoriels orthogonaux, i.e. pour tout $u \in E$ et $v \in F$ on a $\langle u, v \rangle = 0$, alors :*

$$I_{E \oplus F} = I_E + I_F.$$

Preuve de la proposition. Soit O l'origine de notre espace $\mathbb{R}^p$. Soient M un point de $\mathbb{R}^p$ et M_E, M_F et $M_{E\oplus F}$ les projetés respectifs de M sur les sous-espaces affines (O, E), (O, F) et $(O, E \oplus F)$. Le résultat découle de la relation

$$\overrightarrow{OM_{E\oplus F}} = \overrightarrow{OM_E} + \overrightarrow{OM_F}$$

et du théorème de Pythagore. □

Preuve du théorème IX.4. Soit $E_{k+1} \subset \mathbb{R}^p$ un sous-espace vectoriel de dimension $k+1$. Comme $\dim E_{k+1} = k+1$ et $\dim H_k^\perp = p - k$, on a : $\dim(E_{k+1} \cap H_k^\perp) \geq 1$. Il existe donc un vecteur non nul u appartenant à $E_{k+1} \cap H_k^\perp$. Soit F le supplémentaire orthogonal de u dans E_{k+1} : $F \perp u$ et $E_{k+1} = F \oplus u$. Par la Proposition IX.5, on a

$$I_{E_{k+1}} = I_F + I_u$$

[6] à une différence négligeable près : pour les deux derniers critères, l'ensemble des sous-espace affines solutions est invariant par translation, tandis que pour le premier, le sous-espace affine passe nécessairement par le centre de gravité G.

et

$$I_{H_k \oplus u} = I_{H_k} + I_u. \tag{IX.1}$$

Comme H_k est par définition le sous-espace de dimension k portant l'inertie maximale, on a $I_{H_k} \geq I_F$, donc $I_{H_k \oplus u} \geq I_{E_{k+1}}$. Le résultat découle alors de l'égalité (IX.1) et du fait que $u \in H_k^{\perp}$. □

En conséquence, toute solution au problème de réduction de dimension s'obtient de proche en proche, par somme directe de droites portant des inerties maximales.

Soit Γ la matrice de variance-covariance associée au nuage de points (dans la représentation réduite, les moyennes $\bar{X}^j$ sont nulles) :

$$\Gamma = \sum_{i=1}^{n} m_i X_i (X_i)^t,$$

autrement dit $\Gamma_{j,j'} = \sum_{i=1}^{n} m_i X_i^j X_i^{j'}$ est la covariance entre les variables d'intérêt X^j et $X^{j'}$.

Théorème IX.6. *Les assertions suivantes caractérisent la résolution séquentielle du problème de réduction de dimension par moindre déformation.*

- *Les vecteurs portant l'inertie maximale à chaque pas de la décomposition sont les vecteurs propres de la matrice de variance-covariance du nuage. Plus précisément, le vecteur u_k, défini dans le Théorème IX.4, est un vecteur propre de Γ associée à la k-ième plus grande valeur propre.*
- *La k-ième plus grande valeur propre λ_k de Γ vaut l'inertie du nuage projeté sur le k-ième axe propre u_k :*

$$I_{u_k} = \lambda_k.$$

- *l'inertie sur H_k est la somme des inerties sur les k axes propres principaux :*

$$I_{H_k} = \sum_{l=1}^{k} \lambda_l.$$

Démonstration. Cherchons d'abord le vecteur unitaire, i.e. de norme 1, u maximisant l'inertie du nuage projeté sur u. Considérons la projection du nuage sur la direction donnée par le vecteur unitaire u. Le projeté X_i^* de l'individu i s'écrit

$$X_i^* = \langle u, X_i \rangle u$$

et l'inertie du nuage projeté (nous nous plaçons toujours dans le cadre de la représentation réduite) est

$$I_u = \sum_{i=1}^{n} m_i \| \langle u, X_i \rangle u \|^2 = \sum_{i=1}^{n} m_i \langle u, X_i \rangle^2 = \sum_{i=1}^{n} m_i u^t X_i (X_i)^t u = u^t \Gamma u.$$

La matrice Γ est symétrique, semi-définie positive ; elle est diagonalisable, a toutes ses valeurs propres réelles, et il existe une base orthonormale de vecteurs propres de $\mathbb{R}^p$. Notons $\lambda_1 \geq$

... $\geq \lambda_p$ les valeurs propres triées par ordre décroissant, et $u'_1, ..., u'_p$ les vecteurs propres unitaires associés. Alors

$$I_u = \sum_{j=1}^{p} \lambda_j \langle u, u'_j \rangle^2 \leq \lambda_1 \sum_{j=1}^{p} \langle u, u'_j \rangle^2 = \lambda_1 ||u||^2 = \lambda_1.$$

Il suffit alors de choisir $u = u'_1$ pour maximiser I_u.

La meilleure droite de projection du nuage est celle de vecteur directeur u_1, associé à la plus grande valeur propre λ_1 de la matrice Γ.

Pour les directions suivantes, on répète le procédé, en cherchant le vecteur directeur u_2 orthogonal à u_1 portant l'inertie maximale. Pour tout vecteur u orthogonal à u'_1, on a

$$I_u = \sum_{j=2}^{p} \lambda_j \langle u, u'_j \rangle^2 \leq \lambda_2.$$

Donc le maximum est atteint pour $u = u'_2$, et ainsi de suite.

Au passage, on a également prouvé la deuxième assertion du théorème : $I_{u_k} = \lambda_k$. La troisième assertion découle alors de la Proposition IX.5 et du Théorème IX.4. □

L'inertie I du nuage de points est donc égale à la trace de matrice de variance-covariance, ce qui implique $I = p$, en ACP normée. (En ACP non normée, elle vaut la somme des variances.) On définit la part d'inertie expliquée sur le l-ième axe propre : $\tau_l = \lambda_l / I$. L'inertie portée par un sous-espace de dimension k est donc au mieux $\sum_{l=1}^{k} \tau_l$ pour cent de l'inertie totale I.

Retour à l'exemple

Sur notre exemple concernant les 24 modèles de voitures, on peut chercher à visualiser les proximités (en termes de distance normée sur les 6 caractéristiques techniques) entre modèles sur le premier plan factoriel (u_1 horizontalement, u_2 verticalement) (cf. Fig.IX.5).

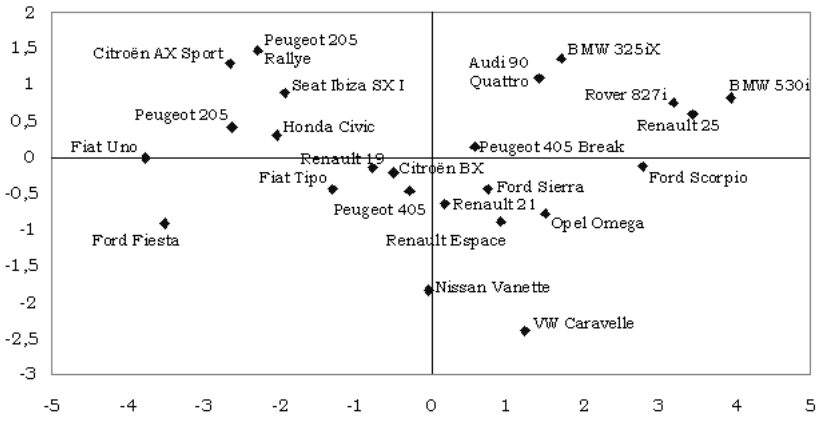


FIG. IX.5 – Projection des individus sur le premier plan factoriel

Dans cet exemple, l'inertie $I = 6$ (nombre de variables) se décompose sur les premiers axes ainsi : $I_1 = 4,66$ (donc $\tau_1 = 78\%$), $I_2 = 0,92$ (donc $\tau_2 = 15\%$). On visualise donc de façon simplifiée, mais optimale ($\tau_{1\oplus 2} = I_{u_1 \oplus u_2}/I$ =93% de l'inertie représentée sur ce plan), les proximités entre modèles, selon la distance choisie.

Les vecteurs directeurs de ces deux premiers axes s'expriment ainsi, dans l'ancienne base :

Vecteur propre	Cylindrée	Puissance	Vitesse	Poids	Longueur	Largeur
u_1	0,44	0,41	0,34	0,43	0,43	0,38
u_2	0,03	0,42	0,66	-0,26	-0,30	-0,48

Reste à interpréter véritablement ces axes, et à comprendre quels sont les principales relations linéaires entre les caractéristiques techniques...

IX.2.4 Principales relations entre variables

Les composantes principales

La diagonalisation vue précédemment permet de définir p nouvelles variables[7] appelées composantes principales :

$$C^\alpha = \sum_{j=1}^{p} u_\alpha^j X^j = X u_\alpha \in \mathbb{R}^n,$$

ou encore $C_i^\alpha = \langle X_i, u_\alpha \rangle$. Elles sont donc combinaisons linéaires des variables d'intérêt X^j initiales. Elles sont centrées puisque les X^j le sont, et on a :

$$\mathrm{Cov}\left(C^\alpha, C^\beta\right) = \sum_{j=1}^{p}\sum_{j'=1}^{p} u_\alpha^j u_\beta^{j'} \mathrm{Cov}\left(X^j, X^{j'}\right) = u_\alpha^t \Gamma u_\beta = \lambda_\beta u_\alpha^t u_\beta.$$

Donc $\mathrm{Cov}\left(C^\alpha, C^\beta\right) = \begin{cases} 0 \text{ si } \alpha \neq \beta, \\ \lambda_\alpha \text{ si } \alpha = \beta. \end{cases}$

On peut calculer la covariance entre les composantes principales et les variables initiales :

$$\mathrm{Cov}\left(C^\alpha, X^j\right) = \sum_{j'=1}^{p} u_\alpha^{j'} \mathrm{Cov}\left(X^{j'}, X^j\right) = \sum_{j'=1}^{p} u_\alpha^{j'} \Gamma_{j',j} = \lambda_\alpha u_\alpha^j.$$

Il s'ensuit que

$$\mathrm{Corr}\left(C^\alpha, X^j\right) = \frac{\mathrm{Cov}\left(C^\alpha, X^j\right)}{\sqrt{\mathrm{Var}(C_\alpha)\,\mathrm{Var}(X_j)}} = \sqrt{\lambda_\alpha} u_\alpha^j.$$

Donc $\sum_{j=1}^{p} \mathrm{Corr}^2\left(C^\alpha, X^j\right) = \lambda_\alpha$.

Autrement dit, la première composante principale C^1 est la combinaison linéaire unitaire des X^j de variance maximale ; de même, parmi les combinaisons linéaires des variables initiales non corrélées avec C^1, C^2 est celle de variance maximale ; etc.

[7] De même que précédemment, on confondra sous le vocable variable la forme linéaire, et sa réalisation sur nos n individus, soit encore le vecteur de $\mathbb{R}^n$ associé.

Ces variables, dans l'ordre présenté, sont donc celles qui résument le mieux les corrélations linéaires entre les X^j.

Pour visualiser les corrélations entre les composantes principales et les X^j, on établit des représentations planes où, en prenant par exemple $\left(C^\alpha, C^\beta\right)$ comme base orthogonale de ce plan, chaque X^j est figuré par un vecteur de coordonnées $\left(\mathrm{Corr}\left(C^\alpha, X^j\right), \mathrm{Corr}\left(C^\beta, X^j\right)\right)$, à l'intérieur du cercle unité[8] , dit des corrélations.

Retour à l'exemple

On voit, dans cet exemple (cf. Fig.IX.6), que C^1 est très corrélée avec tous les X^j, et peut être considérée comme un résumé de la "taille" de la voiture : C^1 est minimale pour les voitures de faible cylindrée, de faible poids, de faibles dimensions, de faible puissance... et maximale pour les voitures grosses, grandes, puissantes, rapides.

Quant à C^2, non corrélée à C^1, elle est maximale pour les voitures de petit gabarit mais rapides et puissantes ; minimale pour celles qui sont imposantes, mais lentes.

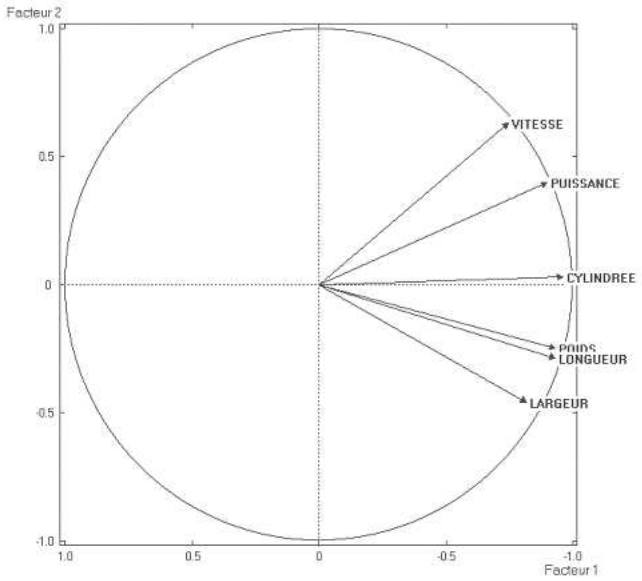


FIG. IX.6 – Cercle des corrélations pour le premier plan factoriel

IX.2.5 Dualité : imbrication des deux objectifs

Ecrivons : $C_i^\alpha = \langle X_i, u_\alpha \rangle$. Autrement dit, la valeur d'un individu sur C^α vaut la coordonnée de cet individu en projection sur l'axe (O, u_α).

[8] Ce vecteur est dans le cercle unité car, dans $\mathbb{R}^n$ muni du produit scalaire $\langle x, y \rangle = \sum_{i=1}^n m_i x_i y_i$, c'est le vecteur projeté orthogonal du vecteur unitaire X^j sur le plan engendré par les vecteurs orthonormés $C^\alpha / \mathrm{Var}(C^\alpha)$ et $C^\beta / \mathrm{Var}(C^\beta)$.

Les deux objectifs de départ sont donc équivalents, la recherche d'axes d'inertie maximale pour le nuage des individus revenant à construire des combinaisons linéaires des variables de variance maximale.

Pour ce qui concerne notre exemple, les deux représentations planes précédentes doivent être considérées simultanément : plus une voiture se projette à droite plus sa valeur sur C^1 est élevée, donc plus elle est grosse et puissante ; et plus un modèle se projette haut plus sa vitesse est remarquable pour ses dimensions.

IX.2.6 Nombre d'axes (ou de composantes) à analyser

Combien d'axes analyser ? Il existe plusieurs critères de décision.

Le premier (Kaiser) veut qu'on ne s'intéresse en général qu'aux axes dont les valeurs propres sont supérieures à la moyenne (qui vaut 1 en ACP normée).

Un second (dit du coude, ou de Cattell) utilise le résultat suivant : lorsque des variables sont peu corrélées, les valeurs propres de la matrice d'inertie décroissent régulièrement - et l'ACP présente alors peu d'intérêt. A l'inverse, lorsqu'il existe une structure sur les données, on observe des ruptures dans la décroissance des valeurs propres (cf. Fig.IX.7). On cherchera donc à ne retenir que les axes correspondant aux valeurs qui précèdent la décroissance régulière. Analytiquement, cela revient à chercher un point d'inflexion dans la décroissance des valeurs propres, et de ne pas aller au-delà dans l'analyse.

Ainsi, dans notre exemple, on ne s'intéressera qu'aux 2 (voire 3) premiers axes.

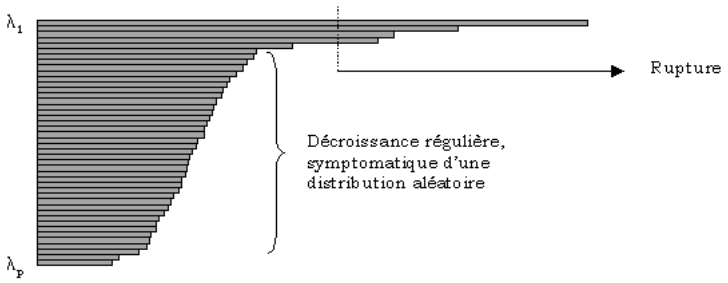


FIG. IX.7 – Représentation des valeurs propres

En tout état de cause, le critère décisif est sans doute celui de l'interprétabilité : inutile de retenir un axe auquel on ne sait donner d'interprétation.

IX.2.7 Eléments supplémentaires

Toute l'analyse précédente s'est fondée sur un tableau de données croisant individus et variables numériques, que l'on appelle individus et variables actives, correspondant à une problématique clairement énoncée. On peut souhaiter, après cette première analyse, y intégrer d'autres informations : il s'agit d'information auxiliaire, ou d'éléments supplémentaires, qui peuvent revêtir plusieurs formes.

- Individus : après avoir transformé leurs valeurs sur les variables actives par centrage et, le cas échéant, réduction (avec les moyenne et écart-type calculés sur les individus actifs), on les projette sur les axes (O, u_α).
- Variables numériques : on calcule leurs corrélations avec les composantes principales, et on les projette à l'intérieur du cercle des corrélations.
- Variables nominales : on représente chaque modalité par le barycentre des individus qui la prennent, dans l'espace des individus.

IX.2.8 Aides à l'interprétation

Si, pour les variables numériques (actives comme supplémentaires), la visualisation des vecteurs à l'intérieur du cercle des corrélations donne toute l'information nécessaire à l'analyse, il peut être utile de définir, pour chaque individu, les aides suivantes :

- La contribution à l'inertie du nuage, qui croît avec le poids et l'excentricité de l'individu :

$$CTR\,(X_i) = \frac{m_i \|X_i - G\|^2}{I}$$

- La contribution à l'inertie portée par un axe (O, u_α) :

$$CTR_\alpha\,(X_i) = \frac{m_i\,(C_i^\alpha)^2}{\lambda_\alpha}$$

  Par construction : $\sum_{i=1}^n CTR\,(X_i) = 1$, et $\sum_{i=1}^n CTR_\alpha\,(X_i) = 1$. La valeur de ces contributions dépend donc fortement du nombre d'individus actifs : une contribution de 5% sera considérée comme forte si l'on manipule les données de milliers d'individus, nettement moins si l'on n'en a qu'une vingtaine (de façon générale, on considèrera que l'individu i a une contribution importante si elle dépasse son poids m_i).
- La qualité de projection sur l'axe (O, u_α) est donnée par le carré du cosinus de l'angle :

$$CO2_\alpha\,(X_i) = \frac{(C_i^\alpha)^2}{\|X_i - G\|^2}.$$

  Par orthogonalité des u_α, la qualité de projection d'un individu sur un sous-espace principal est additive : $CO2_{\alpha+\beta}\,(X_i) = CO2_\alpha\,(X_i) + CO2_\beta\,(X_i)$. D'autre part, on remarque que $\sum_{\alpha=1}^p CO2_\alpha\,(X_i) = 1$; de même que précédemment, cette qualité dépend fortement du nombre initial de variables : on pourra être exigeant si l'on n'en manipule qu'une poignée, on le sera moins s'il y en a davantage.

Pour un axe donné, l'examen parallèle des CTR et des $CO2$ des individus qui s'y projettent peut donner lieu à quatre cas de figure, dont un pose problème ($CO2$ faible-CTR forte), qui apparaît lorsqu'un individu a un poids m_i trop fort par rapport aux autres :

	CTR faible	CTR forte
$CO2$ faible	Elément peu contributif quasi indépendant de l'axe	Elément très contributif mais peu illustratif de l'axe
$CO2$ forte	Elément peu contributif mais bien illustratif de l'axe	Elément particulièrement caractéristique de l'axe

IX.3 Classification automatique

IX.3.1 Problématique

L'objectif est à présent d'opérer des regroupements homogènes au sein de notre population de n individus, sur la base de l'observation des p descripteurs $X^1, \ldots, X^j, \ldots, X^p$, à présent quelconques.

Malheureusement, le nombre de partitions augmente très rapidement avec le nombre d'individus n. Par exemple, pour $n = 4$ (A,B,C et D), il existe 15 partitions possibles :

ABCD	ABC D	ABD C	ACD B	BCD A
AB CD	AC BD	AD BC	AB C D	AC B D
AD B C	BC A D	BD A C	CD A B	A B C D

Plus généralement, les quantités cherchées sont données par les nombres de Bell, dont voici quelques valeurs, qui montrent bien que leur croissance est explosive :

n	4	6	10
P_n	15	203	115 975

Or, en analyse des données, on a généralement à traiter des ensembles de plusieurs milliers d'individus ! On conçoit bien que, même avec les outils de calcul modernes, il serait impossible de comparer toutes les partitions possibles afin de choisir la meilleure, selon un critère qu'on se serait fixé...

C'est la raison pour laquelle ont été développées des méthodes algorithmiques variées répondant à cette problématique. Nous en présentons ici quelques unes[9] parmi les plus utilisées, et indiquons comment les faire fonctionner au mieux.

IX.3.2 Mesure de dissimilarité

La première étape, pour un tel problème, est de définir une mesure de dissimilarité $d : E \times E \to \mathbb{R}^+$, où E est l'ensemble des individus. Diverses propriétés sont souhaitables pour une telle mesure, et notamment, $\forall (i, j, k) \in E \times E \times E$:

1. $d(i, i) = 0$,
2. $d(i, j) = d(j, i)$,
3. $d(i, j) = 0 \Rightarrow i = j$,
4. $d(i, j) \leq d(i, k) + d(k, j)$,
5. $d(i, j) \leq \max(d(i, k), d(k, j))$.

Selon les propriétés vérifiées par d, la terminologie diffère. On parle ainsi de :

[9] Les méthodes dérivées des réseaux de neurones et développées par Kohonen, souvent appréciées pour l'aspect visuel de leurs résultats, ne seront ainsi pas abordées ici.

Indice de dissimilarité	(1)	(2)			
Indice de distance	(1)	(2)	(3)		
Ecart	(1)	(2)		(4)	
Distance	(1)	(2)	(3)	(4)	
Ecart ultramétrique	(1)	(2)		(4)	(5)
Distance ultramétrique	(1)	(2)	(3)	(4)	(5)

De nombreuses mesures sont ainsi proposées dans la littérature, dont les plus anciennes (comme l'indice de distance de Jaccard, 1908) se fondaient sur le nombre de caractéristiques (qualitatives) que i et j partagent ou non.

Le choix d'une mesure dépend essentiellement de la nature des descripteurs ; le plus souvent, le choix se porte sur une distance euclidienne appropriée. Ainsi, si le tableau X est numérique, on pourra utiliser la distance canonique $M = I$ si l'ensemble des variables est homogène, la distance inverse des variances $M = \text{diag}\left(1/\sigma_j^2\right)$ si l'ensemble est hétérogène, ou la distance de Mahalanobis ($M = V^{-1}$, où V est la matrice de variance-covariance) si l'on veut sphériciser le nuage.

Dans la suite, on utilisera des distances classiques qu'on notera $\|.\|$.

IX.3.3 Inerties

Au centre des méthodes les plus utilisées - car ayant de bonnes propriétés "généralistes", et cohérentes avec les méthodes d'analyse factorielle notamment -, on retrouve la notion d'inertie : si l'on a déjà défini dans le cadre de l'ACP l'inertie totale I du nuage de points, et G son centre d'inertie, on va ici définir de nouvelles notions.

Si l'ensemble de nos points est regroupé en K classes (i.e. K sous-nuages N_k), de poids $\left(M_k = \sum_{i \in N_k} m_i\right)_{1 \leq k \leq K}$, de centre d'inertie $(G_k)_{1 \leq k \leq K}$, et d'inertie $(I_k)_{1 \leq k \leq K}$, on définit :

- l'inertie intraclasse : $I_W = \sum_{k=1}^{K} I_k$,
- l'inertie interclasse : $I_B = \sum_{k=1}^{K} M_k \|G_k - G\|^2$.

Or on a (théorème de Huyghens) :

$$\begin{aligned} I &= \sum_{i=1}^{n} m_i \|X_i - G\|^2 \\ &= \sum_{k=1}^{K} \sum_{i \in N_k} m_i \left[\|X_i - G_k\|^2 + \|G_k - G\|^2 + 2 \langle X_i - G_k, G_k - G \rangle \right] \\ &= \sum_{k=1}^{K} I_k + \sum_{k=1}^{K} M_k \|G_k - G\|^2 \\ &= I_W + I_B. \end{aligned}$$

Un critère usuel de classification sera alors, à K fixé, de minimiser l'inertie intraclasse (i.e. rendre les classes les plus homogènes possible) - ce qui revient donc à maximiser l'inertie interclasse (i.e. séparer le plus possible les classes).

La qualité d'une classification pourra alors être évaluée par le ratio I_B/I, interprétable comme une part d'inertie des n points expliquée par leur synthèse en K barycentres.

IX.3.4 Algorithmes de partitionnement : les "centres mobiles"

Plusieurs algorithmes de classification sont d'inspiration géométrique : ils sont connus sous le nom de "méthodes de partitionnement", et leur principe est de partir d'une partition arbitraire, améliorée itérativement jusqu'à convergence. Tous nécessitent de choisir un nombre de classes a priori.

La plus connue de ces méthodes est celle des centres mobiles, due principalement à Forgy (1965). Son principe est le suivant : si l'on veut K classes, on choisit K points dans l'espace des individus ; on affecte ensuite chacun des n individus à celui de ces K points qui lui est le plus proche (au sens de la distance d choisie au départ) ; les K points de départ sont remplacés par les K (ou moins, si un point n'a attiré personne...) barycentres des individus affectés à chacun ; puis on réitère l'affectation, jusqu'à convergence.

Proposition IX.7. *Cet algorithme fait diminuer à chaque itération la variance intraclasse.*

Démonstration. Soient N_k^t les sous-nuages constitués et C_k^t les points auxquels ils se rattachent (barycentres des N_k^{t-1}), à la t-ième itération. Considérons le critère suivant :

$$v(t) = \sum_{k=1}^{K} \sum_{i \in N_k^t} m_i \left\| X_i - C_k^t \right\|^2 .$$

L'inertie intraclasse s'écrit :

$$I_W(t) = \sum_{k=1}^{K} \sum_{i \in N_k^t} m_i \left\| X_i - C_k^{t+1} \right\|^2 .$$

D'après le théorème de Huyghens, on a :

$$v(t) = I_W(t) + \sum_{k=1}^{K} M_k \left\| C_k^t - C_k^{t+1} \right\|^2 .$$

Par minimisation des distances pour chaque individu, on a également : $I_W(t) \geq v(t+1)$. On obtient ainsi $v(t+1) \leq I_W(t) \leq v(t)$, et donc $I_W(t+1) \leq I_W(t)$.

□

Cet algorithme a été légèrement sophistiqué dans deux autres méthodes : les "k-means" (les barycentres ne sont pas recalculés à la fin des affectations, mais après chacune : l'ordre d'apparition des individus n'est donc pas neutre), et les "nuées dynamiques" (où ce n'est plus un seul point qui représente une classe).

Toutes ces méthodes ont pour avantage de converger rapidement vers un minimum local de l'inertie intraclasse. En revanche, leurs deux défauts principaux sont de devoir fixer K, et surtout de converger vers un résultat qui dépend des K points choisis initialement, souvent de façon arbitraire.

IX.3.5 Classification ascendante hiérarchique (CAH)

Ce type d'algorithme produit des suites de partitions emboîtées, à n, $n-1$, ..., 1 classes, par regroupements successifs[10]. La partition en k classes est ainsi obtenue en agrégeant les deux classes les plus proches (éventuellement réduites à un seul individu), au sens d'une nouvelle distance ultramétrique ∇ à définir, parmi celles en $k+1$ classes.

Une fois cet algorithme terminé, on ne récupère donc pas directement une partition, mais une hiérarchie de partitions, indicée par la distance ∇ entre éléments agrégés, et que l'on peut présenter sous formes de dendogrammes, ou mobiles de Calder. En analysant les indices correspondant aux dernières itérations, on décide de retenir la partition qui semble la meilleure - généralement celle qui précède une valeur de l'indice ∇ brutalement plus élevée. Comme en analyse factorielle, l'analyse visuelle de l'historique des indices a pour but de détecter une structure dans les données, autrement dit de choisir un K "naturel".

Ce principe nécessite donc que l'on définisse au préalable une distance ∇ entre groupes d'individus : c'est ce qu'on appelle le choix d'une stratégie d'agrégation. De nombreuses ont été proposées dans la littérature, plus ou moins adaptées au type de données manipulées.

L'une des plus simples est ainsi la stratégie du minimum, où on définit la distance entre deux groupes A et B par : $\nabla(A,B) = \min_{i\in A, j\in B} d(i,j)$. Elle a cependant l'inconvénient de construire, par chaînage, des classes filiformes, qui peuvent ne pas bien répondre à la recherche affichée d'homogénéité intraclasse.

Une autre stratégie est très fréquemment adoptée en espace euclidien : il s'agit de la stratégie moment-partition, encore appelée méthode de Ward. Elle est définie par :

$$\nabla(A,B) = \frac{m_A m_B}{m_{A+} m_B} \|G_A - G_B\|^2$$

où m_A et m_B sont les poids de A et B, et G_A et G_B leurs barycentres respectifs.

Or si l'on agrège A et B, la perte d'inertie interclasse due à cette agrégation vaut :

$$\Delta I_B = m_A \|G_A - G\|^2 + m_B \|G_B - G\|^2 - m_{A\cup B} \|G_{A\cup B} - G\|^2 .$$

Par ailleurs, on sait que, d'après Huyghens :

$$\begin{aligned} &m_A \|G_A - G\|^2 + m_B \|G_B - G\|^2 \\ &= m_A \|G_A - G_{A\cup B}\|^2 + m_B \|G_B - G_{A\cup B}\|^2 + (m_A + m_B) \|G_{A\cup B} - G\|^2 . \end{aligned}$$

Donc on a

$$\begin{aligned} \Delta I_B &= m_A \|G_A - G_{A\cup B}\|^2 + m_B \|G_B - G_{A\cup B}\|^2 \\ &= m_A \left\| G_A - \frac{m_A G_A + m_B G_B}{m_{A+} m_B} \right\|^2 + m_B \left\| G_B - \frac{m_A G_A + m_B G_B}{m_{A+} m_B} \right\|^2 \\ &= \frac{m_A}{(m_{A+} m_B)^2} \|m_B G_A - m_B G_B\|^2 + \frac{m_B}{(m_{A+} m_B)^2} \|m_A G_B - m_A G_A\|^2 \\ &= \frac{m_A m_B}{m_{A+} m_B} \|G_A - G_B\|^2 . \end{aligned}$$

[10] La classification descendante hiérarchique, opérant par dichotomies successives, n'est que rarement utilisée, car ses propriétés sont moins bonnes, et sa programmation plus hardue.

Itération	Classe formée				∇
	Nom	Regroupant :		Effectif	
1	A	Citroën EX	Peugeot 405	2	0,004
2	B	Ford Sierra	P. 405 break	2	0,009
3	C	A	Renault 19	3	0,011
4	D	Rover 827i	Renault 25	2	0,012
5	E	Cit. AX Sport	P. 205 Rallye	2	0,013
6	F	Honda Civic	Seat Ibiza SX I	2	0,015
7	G	Ren. Espace	Opel Omega	2	0,023
8	H	Fiat Uno	Ford Fiesta	2	0,024
9	I	C	Renault 21	4	0,026
10	J	D	Bmw 530i	3	0,032
11	K	A. 90 Quattro	Bmw 325iX	2	0,035
12	L	G	B	4	0,037
13	M	F	E	4	0,040
14	N	I	Fiat Tipo	5	0,048
15	O	H	Peugeot 205	3	0,059
16	P	J	Ford Scorpio	4	0,062
17	Q	Niss. Vanette	VW Caravelle	2	0,106
18	R	Q	L	6	0,200
19	S	O	M	7	0,225
20	T	P	K	6	0,226
21	U	R	N	11	0,293
22	V	U	T	17	1,429
23	W	V	S	24	3,072

FIG. IX.8 – Méthode de Ward appliquée à l'exemple

Autrement dit, avec la méthode de Ward, on agrège à chaque itération les classes dont l'agrégation fait perdre le moins d'inertie interclasse : il s'agit donc d'une optimisation pas-à-pas, qui a l'avantage de ne pas dépendre d'un choix initial arbitraire, mais l'inconvénient - outre sa gourmandise en temps de calcul - d'être vraisemblablement assez éloigné d'un optimum si le nombre de pas $n - K$ est trop élevé.

Si l'on applique à nos 24 modèles de voitures l'algorithme qui vient d'être défini, avec la même distance que pour l'ACP - c'est-à-dire la distance "inverse des variances" - et la méthode de Ward, on obtient les regroupements successifs indiqués Fig.IX.8.

On peut ainsi constater que la perte d'inertie interclasse (représentée par le ∇) augmente lentement lors des premières agrégations ; le premier "saut" remarquable est situé au niveau de la 17ème itération, et plus encore de la 18ème. Les pertes d'inertie forment alors un plateau, pour augmenter fortement lors des deux dernières itérations.

Deux possibilités s'offrent alors :

– si le besoin réside en une classification assez fruste, on peut pousser l'algorithme jusqu'à la 21ème itération, et récupérer ainsi 3 classes. On conserve alors 1.429+3.072=4.501 d'inertie interclasse, ce qui représente 75% de l'inertie totale ;

– si l'on souhaite une analyse un peu plus fine, on pourra par exemple s'arrêter après la 17ème itération, et récupérer alors 7 classes. L'inertie interclasse vaut alors 5.445, ce qui représente 91% de l'inertie totale ; autrement dit, résumer un ensemble de 24 modèles en 7 types ne fait perdre que 9% de l'information initiale[11].

[11] En analyse de données, le rôle pivot que joue l'inertie la rend pratiquement assimilable à un niveau d'information, en langage courant.

Les classes obtenues pour ces deux possibilités sont indiquées Fig.IX.9.

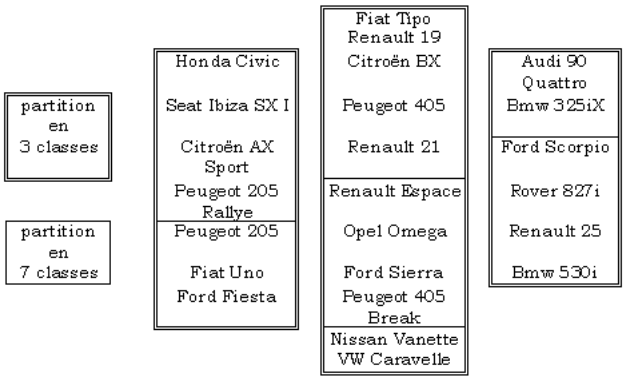


FIG. IX.9 – Partitions en 3 et 7 classes

IX.3.6 Méthodes mixtes

Les méthodes mixtes, en cela qu'elles combinent efficacement les avantages des deux types d'algorithmes vus précédemment et permettent d'en annuler les inconvénients, sont finalement les plus efficaces.

Ces méthodes ont pour cœur algorithmique la CAH, suivie mais aussi parfois précédée d'une méthode de type centres mobiles :

- Si le nombre d'individus n est élevé (plusieurs milliers), on lance plusieurs fois un algorithme de type centres mobiles, en faisant varier le choix des K points de départ : le produit des partitions obtenues permet d'obtenir les groupes - éventuellement réduits à un singleton - d'individus toujours classés ensemble : c'est ce que l'on appelle des formes fortes. Leur nombre peut être assez réduit (quelques dizaines, au plus quelques centaines), et leur définition assez robuste si l'espace de départ est bien balayé, et que le nombre d'essais est suffisant.
- On lance une CAH sur ces formes fortes, avec la méthode de Ward : on chemine ainsi vers une partition en K classes, en minimisant à chaque pas la perte d'inertie interclasse.
- On lance les centres mobiles sur la partition issue de la CAH, avec pour points de départ les barycentres des K classes : on est ainsi assuré d'aboutir à un optimum local, en autorisant quelques réaffectations individuelles - ce que n'autorisait pas la CAH.

IX.3.7 Caractérisation des classes

Une fois la classification aboutie, il faut l'interpréter : pour cela, chaque classe est décrite grâce aux variables actives - celles sur lesquelles on a souhaité différencier les classes ; mais aussi avec toute autre variable supplémentaire (quel qu'en soit le type) dont on connaît les valeurs sur notre population de n individus.

Généralement, on commence par calculer la distance entre le barycentre de chaque classe et le barycentre global, afin de connaître l'excentricité globale de chaque groupe.

Ensuite, on compare, pour chaque variable, sa distribution sur chaque classe et sur l'ensemble de la population. Des tests probabilistes ont été développés (Lebart L., Morineau A., Piron M., Statistique exploratoire multidimensionnelle, Dunod, 2000, 3ème édition) dans cette optique, qui diffèrent selon la nature des variables :

- Lorsque les variables sont numériques, on compare leurs moyennes sur les différents groupes (en tenant compte des effectifs et des variances respectifs).
- Lorsque les variables sont nominales, on compare, pour chaque modalité, sa proportion dans la classe à sa proportion dans la population, afin de déterminer les modalités significativement sur- (ou sous-) représentées. Il est aussi possible de mesurer le pouvoir caractéristique de la variable elle-même (et non plus de chaque modalité) sur la classe.

Globalement, il est alors possible de quantifier la force discriminante de chaque variable sur la classification dans son ensemble, afin de déterminer quelles sont les caractéristiques expliquant le plus les différenciations construites. En fin de course, il est alors fréquent de "nommer" chacune des classes obtenues par un qualificatif résumant la caractérisation.

Pour illustrer ce qui caractérise une classe, on peut aussi rechercher l'individu le plus typique (ou central) de la classe, ou bien encore un noyau d'individus la représentant bien.

Sur notre exemple, les tableaux ci-dessous montrent le pouvoir discriminant des différentes caractéristiques techniques ainsi que la caractérisation de chacune des 7 classes.

Caractéristique	Valeur moyenne (rappel)	Pouvoir discriminant F
Puissance	114 ch	48,8
Poids	1111 kg	36,3
Longueur	422 cm	33,6
Vitesse	183 km/h	32,9
Cylindrée	1906 cm3	29,4
Largeur	169 cm	12,7

TAB. IX.1 – Pouvoir discrimminant des caractéristiques techniques.

IX.4 Complémentarité des deux techniques

L'ACP est un outil puissant pour analyser les corrélations entre plusieurs variables ou, ce qui revient finalement au même, dégager les axes d'allongement maximal d'un nuage de points de grande dimension. Elle permet d'obtenir des facteurs[12] orthonormés et hiérarchisés, ainsi que de nouvelles variables non corrélées à forte variance (i.e. à fort pouvoir informatif).

Elle possède cependant quelques limites, qu'une classification automatique menée sur les mêmes variables (ou sur une sélection réduite de facteurs) corrige avantageusement :

- Du point de vue de l'interprétation : si l'ACP permet de visualiser la structure des corrélations, ces visualisations ne sont réalisables que sur des plans, ce qui limite la perception précise des phénomènes si les variables actives sont nombreuses et leurs

[12] Cette façon de procéder, très courante, a l'avantage de gommer les résidus non structurels représentés par les derniers axes factoriels, et rend généralement le travail de classification plus robuste, car l'hétérogénéité de la population est réduite par cet artifice.

Classe	Excentricité	Caractéristiques significatives moyennées	Résumé générique
Honda Civic Seat Ibiza SX I Citroën AX Sport Peugeot 205 Rallye	6,07	Longueur = 363 cm Poids = 827 kg	Petites sportives
Fiat Uno Peugeot 205 Ford Fiesta	11,11	Largeur = 158 cm Vitesse = 146 km/h	Petites
Fiat Tipo Renault 19 Peugeot 405 Renault 21 Citroën BX	0,63	Puissance = 89 ch Poids = 1 042 kg	Berlines moyennes
Renault Espace Opel Omega Ford Sierra Peugeot 405 Break	1,26	Largeur = 174 cm Longueur = 450 cm	Volumineuses
Nissan Vanette VW Caravelle	5,19	Vitesse = 146 km/h Poids = 1 375 kg	Vans
Audi 90 Quattro Bmw 325iX	4,22	Puissance = 166 ch Vitesse = 211 km/h	Routières
Ford Scorpio Rover 827i Renault 25 Bmw 530i	11,51	Cylindrée = 2 786 cm3 Puissance = 174 ch	Grandes routières

TAB. IX.2 – Caractérisation de chacune des 7 classes.

relations complexes. En outre, la contrainte d'orthogonalité des axes rend souvent l'interprétation délicate au-delà du premier plan factoriel, car les informations délivrées sont alors résiduelles, "sachant" les proximités déjà observées sur les axes précédents. Les méthodes de classification, en revanche, prennent en compte l'ensemble des dimensions actives du nuage, et corrigent donc les distorsions des projections factorielles. Par là même, elles compensent l'abstraction que l'on peut reprocher à l'analyse factorielle - qui se concentre sur le lien entre les variables et la recherche de composantes synthétiques - en appréhendant les individus tels qu'ils sont en réalité, et non tels qu'ils apparaissent en projection.
D'autre part, l'interprétation d'une ACP se heurte au problème de l'espace numérique, dont le continuum n'est pas altéré par l'obtention de la nouvelle base par diagonalisation ; tandis que la classification se traduit par une partition de cet espace en quelques zones, ce qui peut être plus aisé à décrire[13].

[13]D'ailleurs, au-delà de l'interprétation des composantes principales, il est toujours utile d'observer la forme du nuage d'individus projeté sur le plan factoriel : on détermine ainsi des zones de répartition plus ou moins denses (qui sont à l'origine des corrélations), qui peuvent être mises en relation avec les groupes obtenus par

– Du point de vue des représentations :
En tant qu'aides à l'interprétation, les représentations peuvent intégrer la remarque précédente, en faisant figurer sur les projections factorielles du nuage des individus la partition obtenue par classification : il est ainsi possible de faire figurer ces classes par leur barycentre, et/ou par leur "enveloppe", voire en remplaçant chaque point projeté par le code de sa classe d'appartenance, si le nombre de points est assez restreint.
C'est d'ailleurs bien souvent la seule façon - avec les variables supplémentaires qualitatives, dont les barycentres des modalités sont projetés - d'interpréter la projection factorielle des individus, tant leur nombre rend parfois impossible toute analyse directe.

On peut ainsi, pour reprendre notre exemple, faire figurer sur le premier plan factoriel agrémenté de la signification des axes les enveloppes correspondant aux deux classifications possibles, en sept ou trois classes, pour aboutir à un schéma (cf. Fig.IX.10) résumant au mieux l'information contenue dans le tableau initial.

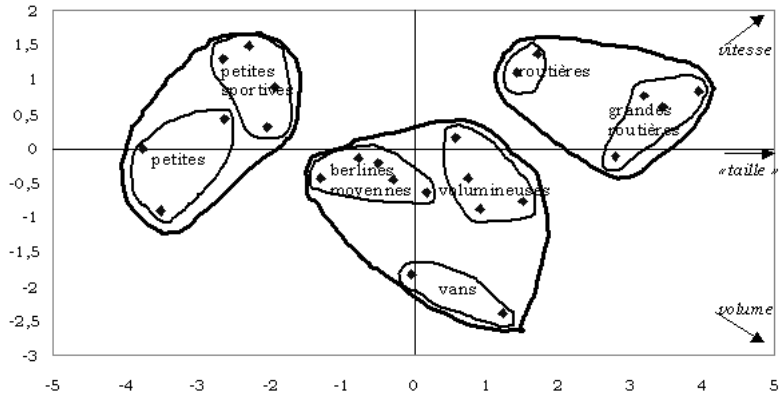


FIG. IX.10 – ACP et classification

IX.5 Limites

Il convient de citer en conclusion les principales limites des méthodes développées.

Une des principales faiblesses de ces techniques est la forte sensibilité aux points extrêmes :

– En ce qui concerne l'ACP, ce manque de robustesse est notamment lié au rôle central qu'y joue la corrélation de Bravais-Pearson : les points extrêmes, en perturbant les moyennes et corrélations, polluent alors fortement l'analyse[14] - on peut cependant envisager de les déplacer en point supplémentaire.
– Cette sensibilité naturelle aux valeurs extrêmes perturbe également les méthodes de classification :

méthode de classification.

[14]Et ce d'autant que dès qu'un axe est perturbé, tous les suivants sont affectés, d'après la contrainte d'orthogonalité des axes.

- Dans la méthodes des centres mobiles[15], le calcul des barycentres est directement affecté, ce qui modifie l'ensemble des affectations de la zone concernée
- Lors d'une CAH "standard", un point extrême, par son éloignement, est bien souvent isolé, et ne s'agrège aux autres que très tardivement. Certes, les premiers regroupements ne sont pas perturbés par sa présence ; mais les classes finalement obtenues le sont, si l'arrêt de la procédure est postérieur au premier rattachement de ce point extrême.

Une solution efficace contre cette faiblesse est alors de modifier dès le départ la topologie de l'espace, de façon à atténuer les distances entre le(s) point(s) extrême(s) et les autres.

D'autre part, l'ACP est inadaptée aux phénomènes non linéaires qui plus est en grande dimension. Pour ce genre de problème, d'autres méthodes ont été développées, comme l'ACPN (Analyse en Composantes Principales par Noyau), (voir à se sujet *Kernel principal component analysis*, B. Schölkopf, A. Smola, and K.-R. Müller, 199).

[15]La méthode parente des "nuées dynamiques" peut cependant évacuer ce problème, car on n'y calcule pas nécessairement de barycentre, mais on utilise un noyau représentatif.

Quatrième partie

Annexe

Chapitre X

Rappels et compléments de probabilités

Ce chapitre est conçu comme une succession de NOTICES sur des notions et résultats essentiels de Calcul des Probabilités, utiles pour le reste du cours.

X.1 Définitions et rappels généraux

*Le lecteur est supposé connaître les définitions de base du calcul des probabilités : nous rappelons ici la terminologie usuelle en langue française, en insistant sur le fait (essentiel pour leur usage en Statistique) qu'il s'agit d'***outils de modélisation***, ce qui implique que leur usage suppose à chaque fois des* **choix***, effectués au vu de la réalité concrète dont on veut rendre compte, choix éventuellement révisables au cours d'une étude et que l'honnêteté scientifique et statistique devrait contraindre à toujours bien expliciter.*

X.1.1 Probabilité et loi

Espace mesurable, éventualités, évènements

On appelle **espace mesurable** tout couple $(\Omega, \mathcal{F})$, où $\mathcal{F}$ est une tribu (autrement dit une σ-algèbre) de parties de Ω ; Ω est parfois appelé **univers** ou **référentiel** ; les éléments de Ω sont appelés **éventualités**, ceux de $\mathcal{F}$ sont appelés **évènements** ; le fait que $\mathcal{F}$ est une tribu signifie que le complémentaire de tout évènement est un évènement et que, pour toute famille finie ou infinie dénombrable d'évènements, leur union et leur intersection sont encore des évènements. Dans la pratique les éventualités ω correspondent à la description la plus fine possible, dans la modélisation choisie, de ce à quoi l'on s'intéresse dans le phénomène étudié. Les évènements sont des sous-ensembles d'éventualités dont la réalisation est susceptible d'intéresser les usagers de cette modélisation (un évènement est "réalisé" si l'éventualité qui survient lui appartient).

Mesure de probabilité

On appelle **mesure de probabilité** (ou **loi de probabilité**, ou encore en bref **probabilité**), définie sur l'espace mesurable $(\Omega, \mathcal{F})$, toute application, soit $\mathbb{P}$, de $\mathcal{F}$ dans $[0, 1]$,

vérifiant $\mathbb{P}(\Omega) = 1$ et $\sigma - additive$ (c'est-à-dire telle que, pour toute famille finie ou infinie dénombrable d'évènements disjoints, soit $(A_i)_{i\in I}$, on a $\mathbb{P}(\bigcup_{i\in I} A_i) = \sum_{i\in I} \mathbb{P}(A_i)$).

Dans la pratique de la modélisation, à partir des hypothèses faites sur le phénomène étudié, on choisit les probabilités de certains évènements remarquables (autrement dit la valeur de $\mathbb{P}(A)$ pour certains éléments A de $\mathcal{F}$) ; le *calcul des probabilités* consiste à en déduire les probabilités de certains autres évènements à la signification concrète jugée intéressante.

L'exemple le plus élémentaire de cette démarche est celui d'un ensemble Ω fini muni de la tribu de toutes ses parties ; $\mathbb{P}$ est alors entièrement déterminée par les probabilités des singletons (appelés en calcul des probabilités **évènements élémentaires**) $\mathbb{P}(\{\omega\})$ (souvent noté en bref $\mathbb{P}(\omega)$ et, pour tout évènement A, il vient

$$\mathbb{P}(A) = \sum_{\omega\in A} \mathbb{P}(\omega) \ .$$

Un **cas particulier** de cette situation est celui où on considère fondée l'hypothèse dite d'**équiprobabilité** : cela signifie que tout évènement élémentaire $\{\omega\}$ vérifie $\mathbb{P}(\omega) = 1/\mathrm{card}(\Omega)$ (où $\mathrm{card}(\Omega)$ désigne le nombre d'éléments de Ω) ; alors on en déduit que, pour tout évènement A, $\mathbb{P}(A) = \mathrm{card}(A)/\mathrm{card}(\Omega)$.

Variable aléatoire (en bref v.a.)

Etant donnés deux espaces mesurables $(\Omega, \mathcal{F})$ et $(\Omega', \mathcal{F}')$, une **variable aléatoire** (ou, en termes généraux de théorie de la mesure, une **application mesurable**) du premier dans le second est une application, soit X, de Ω dans Ω', telle que l'image réciproque de tout évènement A' appartenant à $\mathcal{F}'$ (c'est-à-dire $\{\omega : X(\omega) \in A'\}$) appartienne à $\mathcal{F}$.

En calcul des probabilités, il est fréquent d'adopter pour cette image réciproque la notation $\{X \in A'\}$, qui est en soi monstrueuse (l'être mathématique X ne peut être un élément de l'être mathématique A') et qui se lit : X *prend ses valeurs dans* A'. Concrètement, l'usage d'une variable aléatoire X traduit un "affaiblissement" (on pourrait dire aussi un "filtrage") de l'information apportée par les éventualités de Ω ; la *condition de mesurabilité* que l'on a imposée ($\{X \in A'\} \in \mathcal{F}$) assure que tous les évènements auxquels on peut vouloir s'intéresser sur cette information affaiblie peuvent s'analyser en termes d'évènements relatifs à l'information complète apportée par l'espace $(\Omega, \mathcal{F})$.

Une variante naturelle de la notation $\{X \in A'\}$ est celle qui intervient si A' est un évènement élémentaire $\{\omega'\}$; on note alors $\{X = \omega'\}$ pour $\{\omega : X(\omega) = \omega'\}$. Remarquons que, si Ω' est fini ou infini dénombrable, il suffit, pour démontrer la mesurabilité de X, de s'assurer que, pour tout ω', $\{X = \omega'\}$ appartient bien à $\mathcal{F}$; ceci n'est pas vrai dans le cas général.

Loi d'une variable aléatoire (autrement dit probabilité image)

Reprenant les notations de la notice précédente, la **loi de la v.a.** X (relativement à la mesure de probabilité $\mathbb{P}$) est la mesure de probabilité P_X définie sur l'espace d'arrivée de X, $(\Omega', \mathcal{F}')$, par :

$$P_X(A') = \mathbb{P}(\{X \in A'\})$$

(on notera en bref $\mathbb{P}(X \in A')$ pour $\mathbb{P}(\{X \in A'\})$). Concrètement, A' et $\{X \in A'\}$ représentent le même "évènement" (au sens intuitif du terme), même si, mathématiquement, il s'agit de

sous-ensembles définis dans des espaces différents. Il importe donc qu'ils aient bien numériquement la même probabilité, que l'on porte son attention sur l'espace $(\Omega', \mathcal{F}')$ (intérêt pour l'information filtrée par X) ou sur l'espace $(\Omega, \mathcal{F})$ (information complète). P_X est aussi appelée **probabilité image** (ou **mesure image**) de $\mathbb{P}$ par X (terminologie de la théorie de la mesure et de l'intégration) ; on la note alors aussi $X(\mathbb{P})$.

Q étant une mesure de probabilité sur Ω', la propriété Q *est la loi de la v.a.* X (autrement dit l'égalité $P_X = Q$) se dit aussi : X **suit la loi** Q ; on note ceci : $X \sim Q$.

Dans la pratique, il arrive fréquemment que l'on considère simultanément une famille de variables aléatoires, $(X_i)_{i\in I}$, définies sur un même espace mesurable. Il sera alors nécessaire de bien préciser les espaces $(\Omega_i, \mathcal{F}_i)$ d'arrivée de chacune de ces v.a. mais souvent leur espace de départ commun, ainsi que la mesure de probabilité dont il est muni, pourront être seulement évoqués, par exemple comme lieu où se formulent des hypothèses d'indépendance (voir ci-dessous X.1.4), sans recevoir de notation particulière ; s'il en est ainsi, on trouvera des écritures du type $\mathbb{P}(X \in B)$ (avec le graphisme conventionnel $\mathbb{P}$) pour traduire l'expression : *probabilité que* X *prenne ses valeurs dans* B.

Espace mesurable produit, loi marginale

Il est fréquent que la réalisation d'un phénomène observé se décompose sous forme d'une famille d'observations plus élémentaires ; en d'autres termes on doit modéliser l'éventualité ω comme une famille $(\omega_i)_{i\in I}$ (souvent $I = \{1, \ldots, n\}$ et $\omega = (\omega_1, \ldots, \omega_n)$). On est donc amené à associer à chaque observation élémentaire un espace mesurable $(\Omega_i, \mathcal{F}_i)$ et les éventualités "complètes" sont des éléments de l'espace produit $\Omega_I = \prod_{i\in I} \Omega_i$; celui-ci peut être muni d'une tribu, dite *tribu produit*, $\mathcal{F}_I = \prod_{i\in I} \mathcal{F}_i$ qui est la plus petite tribu contenant tous les pavés finis, c'est-à-dire contenant toutes les parties de Ω_I de la forme $\prod_{i\in I} A_i$, où pour tout i, $A_i \in \mathcal{F}_i$ et $A_i = \Omega_i$ sauf pour un nombre fini d'indice.

Attention : En notation traditionnelle de théorie des ensembles, $\prod_{i\in I} \mathcal{F}_i$ désignerait plutôt l'ensemble des familles $(A_i)_{i\in I}$ où, pour tout i, $A_i \in \mathcal{F}_i$; c'est pourquoi certains ouvrages notent $\bigotimes_{i\in I} \mathcal{F}_i$ ce que nous avons noté $\prod_{i\in I} \mathcal{F}_i$ et disent qu'il s'agit du **produit tensoriel** des tribus $\mathcal{F}_i$.

Dans le cas de deux espaces (où donc $I = \{1, 2\}$), on note la tribu produit $\mathcal{F}_1 \times \mathcal{F}_2$, ou $\mathcal{F}_1 \otimes \mathcal{F}_2$ si on adopte la notation du type "produit tensoriel".

Concrètement, cette définition exprime que l'on définit bien un évènement dans Ω_I en spécifiant simultanément la réalisation d'évènements relatifs à chacune des composantes ω_i. Si on veut spécifier seulement la réalisation d'événements relatifs à certaines composantes (disons les ω_j, où $j \in J$, avec $J \subset I$), on considère dans Ω_I des parties de la forme $\prod_{i\in I} B_i$ où, si $i \in J$, $B_i = A_i$ et, si $i \notin J$, $B_i = \Omega_i$. En particulier, dans Ω_I, un évènement relatif à une unique composante j s'exprime $\prod_{i\in I} B_i$ où $B_j = A_j$ et, si $i \neq j$, $B_i = \Omega_i$. ; en d'autres termes, c'est l'image réciproque de A_j par l'application "projection" de Ω_I sur Ω_j, soit π_j ; la définition de $\mathcal{F}_I$ peut alors s'interpréter en disant que c'est la plus petite tribu relativement à laquelle toutes les projections π_j (où $j \in I$) soient mesurables.

Il est essentiel de retenir que **la donnée d'une probabilité "individuelle"** $\mathbb{P}_i$ **sur chacun des espaces** $(\Omega_i, \mathcal{F}_i)$ **ne suffit pas à déterminer une probabilité sur l'espace mesurable produit** $(\Omega_I, \mathcal{F}_I)$; tout ce que l'on sait c'est que toute modélisation cohérente doit faire intervenir sur $(\Omega_I, \mathcal{F}_I)$ une probabilité $\mathbb{P}$ dont, pour tout i, l'image par la projection π_i est $\mathbb{P}_i$; en d'autres termes, $\mathbb{P}_i = \mathbb{P}_{\pi_i}$, puisque c'est la loi, relativement à $\mathbb{P}$, de la variable

aléatoire projection, π_i. On dit aussi que $\mathbb{P}_i$ est la **loi marginale** (ou en bref **marge**) de $\mathbb{P}$ pour la composante d'indice i .

Dans le cas particulier où tous les espaces mesurables $(\Omega_i, \mathcal{F}_i)$ sont identiques (soit ici $(\Omega, \mathcal{F})$ cet espace commun) et où $I = \{1, \ldots, n\}$, on emploie la notation de type "puissance", c'est-à-dire $(\Omega^n, \mathcal{F}^n)$ (ou bien, avec les notations du type "produit tensoriel", $(\Omega^n, \mathcal{F}^{\otimes n})$.

La construction ci-dessus (espace produit) intervient généralement dans des contextes où on considère une **famille de v.a.**, $X_I = (X_i)_{i \in I}$, implicitement toutes définies sur un même espace. Chaque X_i est à valeurs dans un espace $(\Omega_i, \mathcal{F}_i)$ et donc X_I est à valeurs dans l'espace $(\Omega_I, \mathcal{F}_I)$.

Absolue continuité d'une probabilité par rapport à une mesure σ-finie. Densité

Considérons sur $(\Omega, \mathcal{F})$ à la fois une probabilité $\mathbb{P}$ et une mesure μ (application σ-additive de $\mathcal{F}$ dans $[0, +\infty]$) qui est supposée de plus σ-**finie**, c'est-à-dire telle que ou bien $\mu(\Omega)$ soit fini, ou bien il existe une partition de Ω en une suite $(F_1, \ldots, F_n, \ldots)$ d'éléments de $\mathcal{F}$ telle que, pour tout n, $\mu(F_n)$ soit fini. $\mathbb{P}$ est dite **absolument continue par rapport à** μ si tout évènement A vérifiant $\mu(A) = 0$ vérifie aussi $\mathbb{P}(A) = 0$. Ceci équivaut (théorème de Radon-Nikodym) à l'existence d'une application mesurable, soit f, de $(\Omega, \mathcal{F})$ dans $([0, +\infty[, \mathcal{B})$, où $\mathcal{B}$ est la tribu borélienne (voir si besoin X.1.2, p. 227 ci-dessous), appelée **densité** de $\mathbb{P}$ par rapport à μ, telle que

$$\forall A \in \mathcal{F} \quad \mathbb{P}(A) = \int_A f(\omega) d\mu(\omega) \; .$$

On retiendra qu'une densité n'est définie qu'à une égalité μ presque partout près, c'est-à-dire que, si f est une densité de $\mathbb{P}$ par rapport à μ, il en est de même de toute application mesurable, soit g, de $(\Omega, \mathcal{F})$ dans $([0, +\infty[, \mathcal{B})$ vérifiant : $\mu(\{\omega; f(\omega) \neq g(\omega)\}) = 0$. En d'autres termes, la "bonne notion" pour la densité serait plutôt celle d'un élément de $L^1_+(\mu)$, ensemble des classes d'équivalence pour la relation d'égalité μ presque-partout sur l'ensemble $\mathcal{L}^1_+(\mu)$ des fonctions réelles positives ou nulles μ-intégrables.

Un cas élémentaire d'absolue continuité est celui où Ω est fini ou infini dénombrable, muni, comme il est usuel alors, de la tribu de toutes ses parties. Alors on peut considèrer sur Ω sa **mesure de comptage** (ou mesure de dénombrement) δ_Ω : pour tout A, $\delta_\Omega(A)$ est le nombre d'éléments de A ; toute probabilité $\mathbb{P}$ est alors absolument continue par rapport δ_Ω, et la densité (unique ici) est l'application : $\omega \mapsto \mathbb{P}(\{\omega\})$.

Par extension, si X est une v.a. à valeurs dans un espace mesurable $(\Omega', \mathcal{F}')$ et μ' est une mesure σ-finie sur cet espace, on dit que X est absolument continue par rapport à μ' si sa loi P_X l'est ; la densité de P_X par rapport à μ' est alors aussi appelée densité de X.

Absolue continuité entre deux mesures σ-finies

Les définitions d'absolue continuité et de densité d'une probabilité par rapport à une mesure σ-finie s'étendent sans difficulté à celles d'absolue continuité et de densité d'une mesure σ-finie par rapport à une autre mesure σ-finie. Etant donné 3 mesures σ-finies μ, ν et θ, la propriété de transitivité suivante est immédiate : si ν est absolument continue par rapport à μ, avec f pour densité, et θ est absolument continue par rapport à ν, avec g pour densité, alors θ est absolument continue par rapport à μ, avec le produit gf pour densité.

On emploie parfois une notation de type "différentielle", bien adaptée pour rendre compte de cette propriété : notant $\frac{d\nu(x)}{d\mu(x)}$ "la" valeur en x de la densité de ν par rapport à μ, il vient : $\frac{d\theta(x)}{d\mu(x)} = \frac{d\theta(x)}{d\nu(x)}\frac{d\nu(x)}{d\mu(x)}$.

Deux mesures σ-finies sont dites **équivalentes** si chacune est absolument continue par rapport à l'autre. Si ν est absolument continue par rapport à μ, avec f pour densité, il faut et il suffit, pour qu'inversement μ soit absolument continue par rapport à ν, que f soit μ-presque partout strictement positive, et alors $1/f$ définit une densité de μ par rapport à ν, ce qui se note aussi : $\frac{d\mu(x)}{d\nu(x)} = \frac{1}{\frac{d\nu(x)}{d\mu(x)}}$.

Calcul de la densité de la loi d'une v.a.

Soit donné, sur un espace mesurable $(\Omega, \mathcal{F})$, une mesure σ-finie μ et une probabilité $\mathbb{P}$ absolument continue par rapport à μ ; soit f une densité de $\mathbb{P}$ par rapport à μ (voir p. 226). Soit X une v.a. définie sur $(\Omega, \mathcal{F})$, à valeurs dans un espace $(\Omega', \mathcal{F}')$. Il n'est pas toujours facile de calculer la densité de la loi de X, soit P_X (aussi appelée probabilité image de $\mathbb{P}$ par X), relativement à une mesure σ-finie μ' sur $(\Omega', \mathcal{F}')$.

Un cas particulier où ce calcul est élémentaire est celui où la densité f se factorise à travers X et où on prend pour mesure sur Ω' la mesure image de μ par X, $X(\mu)$, qui est définie par :

$$\forall B \in \mathcal{F}' \quad (X(\mu))(B) = \mu(X \in B) \ .$$

Cela signifie qu'il existe g, application de Ω' dans $\mathbb{R}_+$ telle que : $\forall \omega \ f(\omega) = g(X(\omega))$. Il résulte alors immédiatement de la définition de la densité que g est une densité de P_X (loi de X) par rapport à $X(\mu)$.

Il en est en particulier ainsi dans la situation suivante, que nous allons décrire dans le vocabulaire des applications mesurables plutôt que dans celui des variables aléatoires : soit ϕ une application bijective et bimesurable (c'est-à-dire mesurable ainsi que sa réciproque ϕ^{-1}) de Ω dans Ω' ; alors les densités f de $\mathbb{P}$ par rapport à μ et g de $\phi(\mathbb{P})$ par rapport à $\phi(\mu)$ sont liées par les relations :

$$f(\omega) = g(\phi(\omega)) \quad , \quad g(\omega') = f(\phi^{-1}(\omega')) \ .$$

Un défaut de cette propriété est que, si μ a été introduite naturellement lors d'une modélisation, son image $\phi(\mu)$ peut n'être pas très maniable (voir en X.1.2 des situations où cette difficulté peut être contournée).

X.1.2 Variables aléatoires

Tribu borélienne

L'ensemble $\mathbb{R}$ des nombres réels est systématiquement muni de la tribu dite **borélienne**, notée $\mathcal{B}$, qui est la plus petite tribu contenant les demi-droites infinies à gauche et fermées à droite ($]-\infty, x]$) ; elle contient alors aussi nécessairement tous les intervalles, bornés ou non, quelle que soit leur nature (ouvert ou fermé) en leurs extrémités ; plus généralement, elle

contient toutes les parties ouvertes et fermées pour la topologie usuelle de la droite réelle. Les éléments de $\mathcal{B}$ sont appelés **parties boréliennes** (ou en bref **boréliens**).

Si A est une partie borélienne de $\mathbb{R}$, on appelle encore tribu borélienne sur A, et on note $\mathcal{B}_A$ (ou simplement $\mathcal{B}$ s'il n'y a pas de risque de confusion sur A) la restriction à A de la tribu borélienne de $\mathbb{R}$ (nous l'avons utilisé ci-dessus, pour $A = [0, +\infty[$, en X.1.1).

De nombreuses modélisations conduisent à considérer des éventualités appartenant à une partie borélienne, soit A, de $\mathbb{R}$, en particulier un intervalle ou un ensemble fini ou infini dénombrable : voir par exemple une durée exprimée en nombre entiers d'unités de temps (minutes, jours, coups dans un jeu ...) qui s'exprime dans $\mathbb{N}$; on peut indifféremment, sans que cela prête à confusion, considérer que la probabilité $\mathbb{P}$ qui régit ce phénomène est définie sur l'espace mesurable $(A, \mathcal{B}_A)$ ou sur $(\mathbb{R}, \mathcal{B})$, avec bien sûr dans ce dernier cas $\mathbb{P}(A) = 1$ (et donc $\mathbb{P}(\bar{A}) = 0$, où $\bar{A}$ désigne le complémentaire de A dans $\mathbb{R}$) ; on dit que $\mathbb{P}$ est **concentrée** sur A, ou **portée** par A.

Fonction de répartition d'une probabilité

Une probabilité P sur $(\mathbb{R}, \mathcal{B})$ est caractérisée par sa **fonction de répartition**, c'est-à-dire l'application, soit F_P, de $\mathbb{R}$ dans $[0, 1]$, définie par : $F_P(x) = P(]-\infty, x])$. Cette fonction est croissante, c.à.d.l.à.g. (c'est-à-dire continue à droite et admettant une limite à gauche en tout point), sa limite en $-\infty$ est 0 et sa limite en $+\infty$ est 1. Elle ne peut-être discontinue qu'au plus en une infinité dénombrable de points et, en tout point de discontinuité x, le saut de F_P vaut $P(\{x\})$.

Variable aléatoire réelle (v.a.r.), fonction de répartition d'une v.a.r., ordre stochastique

Une application de Ω, muni de la tribu $\mathcal{F}$, dans $\mathbb{R}$, soit X, est une **v.a.r.** (variable aléatoire réelle) si, pour tout $B \in \mathcal{B}$, l'image réciproque de B par X (dont on rappelle qu'on la note $\{X \in B\}$) appartient à $\mathcal{F}$; il suffit en fait qu'il en soit ainsi pour tout B qui est une demi-droite infinie à gauche et fermée à droite ($]-\infty, x]$) et on note alors naturellement $\{X \leq x\}$ pour $\{X \in]-\infty, x]\}$ (on aura évidemment des notations analogues du type $\{X > x\}$, $\{x \leq X \leq x'\}$ ...).

Si l'espace mesurable $(\Omega, \mathcal{F})$ est muni d'une probabilité P, on appelle **fonction de répartition** de X (relativement à P, s'il y a lieu de le préciser) la fonction de répartition (voir X.1.2) de sa loi P_X (voir X.1.1), c'est-à-dire l'application F_X de $\mathbb{R}$ dans $[0, 1]$ définie par :

$$F_X(x) = F_{P_X}(x) = P(X \leq x) = P_X(]-\infty, x]) \ .$$

La fonction de répartition sert à définir sur les variables aléatoires **l'ordre stochastique** : étant donné, sur un même espace mesurable $(\Omega, \mathcal{F})$ muni d'une probabilité P, deux v.a. X et Y, on dit que X est stochastiquement plus grande que Y (noté $Y \preceq X$) si, quel que soit $t \in \mathbb{R}$, la probabilité que X dépasse t est supérieure à celle que Y dépasse t, autrement dit si

$$\forall t \in \mathbb{R} \ \ F_X(t) \leq F_Y(t).$$

Cette propriété s'exprime aussi couramment par la phrase : X **a tendance à prendre de plus grandes valeurs que** Y.

Cette expression *tendance à prendre de plus grandes valeurs* s'emploie aussi quand on considère une seule variable aléatoire X, mais deux probabilités, soit P_1 et P_2, sur l'espace mesurable $(\Omega, \mathcal{F})$; si on note $F_{P_i,X}$ (où i vaut 1 ou 2) la fonction de répartition de X relativement à la probabilité P_i, on dira que X **a tendance à prendre de plus grandes valeurs sous** P_1 **que sous** P_2 si

$$\forall t \in \mathbb{R} \;\; F_{P_1,X}(t) \leq F_{P_2,X}(t).$$

Variable aléatoire réelle discrète

Une v.a.r. X est dite **discrète** si elle ne prend ses valeurs que dans une partie finie ou une infinie dénombrable de $\mathbb{R}$, soit B. Alors, pour tout $t \in \mathbb{R}$, $F_X(t) = \sum_{x \in B; x \leq t} \mathbb{P}(X = x)$, la notation $\sum$ désignant ici soit la sommation usuelle d'un nombre fini de termes, soit la somme d'une série à termes positifs . Ce mode de calcul justifie une autre appellation, un peu démodée, de la fonction de répartition, à savoir **fonction de cumul** (ou **des probabilités cumulées**).

En particulier le calcul de $P_X(B)$ ne fait intervenir que des sommes finies si B est fini (notons $B = \{x_1, \ldots, x_k\}$, avec la suite $(x_1, \ldots, x_k)$ strictement croissante) ou infini dénombrable, à condition dans ce dernier cas que les éléments de B puissent être ordonnés en croissant (notons $B = \{x_1, \ldots, x_k, \ldots\}$). Il en est ainsi si X est à valeurs entières positives ou nulles, non bornées, ce qui est fréquent dans la modélisation de phénomènes de durée. F_X est alors constante sur chacun des intervalles, fermés à gauche, ouverts à droite, $[x_{i-1}, x_i[$, ainsi que sur $]-\infty, x_1[$ (où elle vaut 0) et, s'il y a lieu, sur $[\sup(B), +\infty[$ (où elle vaut 1). En chaque x_i, le saut de F_X vaut $p_i = \mathbb{P}(X = x_i)$.

Tribu borélienne puissance

Etant donné un ensemble I, l'ensemble puissance $\mathbb{R}^I$ (ensemble des familles $(x_i)_{i \in I}$ où, pour tout i, $x_i \in \mathbb{R}$) est muni de la tribu puissance (voir X.1.1) $\mathcal{B}^I$ (ou $\mathcal{B}^{\otimes I}$) qui est aussi la plus petite tribu contenant toutes les parties de la forme $\prod_{i \in I}]-\infty, x_i[$, où $x_i \in \mathbb{R} \cup \{+\infty\}$ et $x_i = +\infty$ sauf pour un nombre fini d'indice $i \in I$.

Variable aléatoire réelle multidimensionnelle ou vecteur aléatoire

La loi d'une famille finie $X = (X_i)_{i \in I}$ de v.a.r. réelles soit P_X, est entièrement caractérisée par sa **fonction de répartition** (I-dimensionnelle), c'est-à-dire l'application F_X de $\mathbb{R}^I$ dans $[0, 1]$ définie par :

$$F_X((x_i)_{i \in I}) = \mathbb{P}(\forall i \;\; X_i \leq x_i) = \mathbb{P}(\bigcap_{i \in I} \{X_i \leq x_i\}) = P_X(\prod_{i \in I}]-\infty, x_i])$$

Cette notion nous sert essentiellement dans le cas de variables n-dimensionnelles (cas où $I = \{1, \cdots, n\}$). On dit alors qu'on a un **vecteur aléatoire** et on adopte souvent la notation matricielle :

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} .$$

Mesure de Lebesgue, densité par rapport à la mesure de Lebesgue

Sauf mention contraire l'absolue continuité (en bref **a.c.**) pour une probabilité P (ou plus généralement une mesure σ-finie) sur $\mathbb{R}$ (ou sur un intervalle de $\mathbb{R}$) s'entend toujours relativement à la **mesure de Lebesgue**, c'est-à-dire la mesure, notée λ, qui à tout intervalle borné d'extrémités inférieure a et supérieure b associe sa longueur $b-a$, et donc à tout intervalle non borné (autrement dit toute droite ou demi-droite) associe la valeur $+\infty$.

Si la probabilité P est a.c., sa fonction de répartition (voir p. 228) F_P est continue (autrement dit, pour tout x, $P(\{x\})=0$) et est dérivable sauf éventuellement aux points d'un ensemble N vérifiant $\lambda(N)=0$; on dit qu'elle est presque partout dérivable, et pour tout choix "réaliste" de P dans une modélisation, N est un ensemble fini, ou (plus rarement) infini dénombrable.

On obtient alors une **densité** de P (sous-entendu : par rapport à λ), soit f_P, **par dérivation** : on prend, pour tout $x \notin N$, $f_P(x)=F'_P(x)$ et, s'il y a lieu, on prolonge arbitrairement à N. On remarque que, sur tout intervalle ouvert $]a,b[$ tel que $P(]a,b[)=0$ (s'il en existe), F_P est constante et on prend f_P nulle. On dispose alors, en tout point x en lequel F_P est dérivable, d'une interprétation "à la physicien" de la densité : la probabilité d'un "petit intervalle" de longueur Δx centré en x vaut approximativement $f_P(x)\Delta x$.

L'intégration par rapport à la mesure de Lebesgue est l'intégration "usuelle" et on a donc, pour tout intervalle, borné ou non, d'extrémités inférieure a et supérieure b :

$$P([a,b])=P([a,b[=P(]a,b])=P(]a,b[)=F_P(b)-F_P(a)=\int_a^b f_P(x)dx \ .$$

En particulier on reconstitue la fonction de répartition à partir de la densité :

$$F_P(x)=\int_{-\infty}^{x} f_P(t)dt \ .$$

On remarque que si, comme on le fait dans certains ouvrages, on identifie la probabilité P avec sa fonction de répartition F_P, on obtient une grande similitude entre la notation "différentielle" des densités (voir X.1.1), $\frac{dP(x)}{d\lambda(x)}$, et la notation différentielle des déérivées $\frac{dF_P(x)}{dx}$.

Une v.a.r. X est dite absolument continue (ou plus succintement continue) s'il en est ainsi de sa loi P_X, et on appelle densité de X toute densité de P_X (par rapport à λ) ; soit f_X une densité de X ; alors, $\mathbb{P}$ désignant ici la mesure de probabilité adoptée sur l'espace mesurable sur lequel est définie la v.a. X, il vient :

$\mathbb{P}(a \le X \le b)=\mathbb{P}(a \le X < b)=\mathbb{P}(a < X \le b)=\mathbb{P}(a < X < b)=\int_a^b f_X(x)dx$

et

$$F_X(x)=\int_{-\infty}^{x} f_X(t)dt \ .$$

Image de la mesure de Lebesgue

Une situation techniquement importante est la suivante : soit ϕ un **C^1-difféomorphisme** (c'est-à-dire une application bijective, continuement différentiable ainsi que sa réciproque ϕ^{-1}) d'un intervalle ouvert U dans un intervalle ouvert V de $\mathbb{R}$. Alors la mesure image de λ_U

(restriction de la mesure de Lebesgue à U) par ϕ est absolument continue par rapport à la mesure de Lebesgue λ_V et admet pour densité l'application $\frac{d\phi(\lambda)}{d\lambda}$ définie par :

$$\frac{d\phi(\lambda)}{d\lambda}(y) = |(\phi^{-1})'(y)| = \frac{1}{|\phi'(\phi^{-1}(y))|} \ .$$

Il en résulte (voir X.1.1) que, si P est une probabilité sur $\mathbb{R}$, concentrée sur U (voir p. 227) et admettant f comme densité par rapport à λ, son image $\phi(P)$ admet par rapport à λ la densité g définie par :

$$g(y) = \frac{d\phi(P)}{d\lambda}(y) = \frac{d\phi(P)}{d\phi(\lambda)}(y)\frac{d\phi(\lambda)}{d\lambda}(y)$$

c'est-à-dire

$$g(y) = f(\phi^{-1}(y))|(\phi^{-1})'(y)| = \frac{f(\phi^{-1}(y))}{|\phi'(\phi^{-1}(y))|} \ .$$

Inversement, on a :

$$f(x) = g(\phi(x))|\phi'(x)| \ .$$

On fera le lien avec les calculs classiques par "changement de variable" en intégration : soit B un évènement contenu dans V ; alors

$$(\phi(P))(B) = \int_V \mathbf{1}_B(y)g(y)dy$$

d'où, par le changement de variable $y = \phi(x)$,

$$(\phi(P))(B) = \int_U \mathbf{1}_B(\phi(x))g(\phi(x))|\phi'(x)| \ dx = \int_{\phi^{-1}(B)} g(\phi(x))|\phi'(x)| \ dx \ ;$$

or par définition de la probabilité image, $\phi(P))(B) = P(\phi^{-1}(B)))$ et donc on a aussi :

$$(\phi(P))(B) = \int_{\phi^{-1}(B)} f(x)dx \ ;$$

on a donc pour tout B

$$\int_{\phi^{-1}(B)} g(\phi(x))|\phi'(x)| \ dx = \int_{\phi^{-1}(B)} f(x)dx \ ,$$

ce qui est bien cohérent avec l'égalité

$$f(x) = g(\phi(x))|\phi'(x)| \ .$$

Mesure de Lebesgue multi-dimensionnelle

Sauf mention contraire l'absolue continuité (en bref **a.c.**) pour une probabilité P sur $\mathbb{R}^n$ s'entend toujours relativement à la **mesure de Lebesgue n-dimensionnelle**, c'est-à-dire la mesure, notée λ^n (où simplement λ s'il n'y a pas de risque de confusion), qui à tout pavé $\prod_{i=1}^n [a_i, b_i[$ (la nature des extrémités étant en fait indifférente) associe son volume $\prod_{i=1}^n (b_i - a_i)$ (et donc à tout pavé non borné associe la valeur $+\infty$).

Si P est a.c., sa fonction de répartition (voir p. 229) F_P est continue et est presque partout différentiable par rapport aux n variables. On obtient alors une **densité** de P (en sous-entendant : p.r.p. λ^n), soit f_P, **par dérivation** : on prend, pour tout $(x_1, \ldots, x_n)$ pour lequel cela a un sens, $f_P(x_1, \ldots, x_n) = \frac{\partial^n F_P}{\partial x_1 \ldots \partial x_n}(x_1, \ldots, x_n)$ et, s'il y a lieu, on prolonge arbitrairement à l'ensemble, de mesure nulle pour μ, pour lequel cette dérivée n'est pas définie. On dispose alors, en tout point $(x_1, \ldots, x_n)$ en lequel F_P est dérivable par rapport aux n variables, d'une interprétation "à la physicien" de la densité : la probabilité d'un "petit pavé" de volume $\Delta x_1 \ldots \Delta x_n$ centré en $(x_1, \ldots, x_n)$ vaut approximativement $f_P(x_1, \ldots, x_n)\Delta x_1 \ldots \Delta x_n$.

L'intégration par rapport à la mesure de Lebesgue n-dimensionnelle est l'intégration "usuelle" des fonctions de n variables et on a donc, pour tout pavé $\prod_{i=1}^n [a_i, b_i[$:

$$P(\prod_{i=1}^{n} [a_i, b_i[) = \int_{a_1}^{b_1} \ldots \int_{a_n}^{a_n} f_P(x_1, \ldots, x_n) dx_1 \ldots dx_n \ .$$

En particulier on reconstitue la fonction de répartition à partir de la densité :

$$F_P(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f_P(t_1, \ldots, t_n) dt_1 \ldots dt_n \ .$$

Une v.a.r. n-dimensionnelle $X = (X_1, \ldots, X_n)$ est dite absolument continue s'il en est ainsi de sa loi P_X, et on appelle densité de X toute densité de P_X (par rapport à λ) ; soit f_X une densité de X ; il vient :

$$P(\forall i \ \ a_i \leq X_i \leq b_i) = \int_{a_1}^{b_1} \ldots \int_{a_n}^{a_n} f_X(x_1, \ldots, x_n) dx_1, \ldots, dx_n$$

(les inégalités larges étant remplaçables par des inégalités strictes) et

$$F_X(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f_X(t_1, \ldots, t_n) dt_1 \ldots dt_n \ .$$

Image de la mesure de Lebesgue multidimensionnelle

Nous généralisons au cas multidimensionnel les résultats donnés en X.1.2 dans le cas unidimensionnel : soit ϕ un C^1**-difféomorphisme** d'une partie ouverte U de $\mathbb{R}^n$ dans une partie ouverte V de $\mathbb{R}^n$. Alors la mesure image de λ_U^n (restriction de la mesure de Lebesgue n-dimensionnelle à U) par ϕ est absolument continue par rapport à la mesure de Lebesgue λ_V^n et admet pour densité l'application $\frac{d\phi(\lambda)}{d\lambda}$ définie par :

$$\frac{d\phi(\lambda)}{d\lambda}(y) = |\det(J(\phi^{-1})(y))| = \frac{1}{|\det(J(\phi))(\phi^{-1}(y))|} \ ,$$

où det signifie *déterminant* et où $J(\phi)(x)$ est le jacobien de ϕ en x, c'est-à-dire, si on note, pour $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, $\phi(x) = (\phi_1(x), \ldots, \phi_n(x))$, la matrice, à n lignes et n colonnes, dont le terme géneral est $\dfrac{\partial \phi_i}{\partial x_j}(x)$ (et de même pour $J(\phi^{-1})(y)$).

Il en résulte (voir X.1.1) que, si P est une probabilité sur $\mathbb{R}^n$, concentrée sur U (voir p. 227) et admettant f comme densité par rapport à λ^n, son image $\phi(P)$ admet par rapport à λ^n la densité g définie par :

$$g(y) = \frac{d\phi(P)}{d\lambda^n}(y) = \frac{d\phi(P)}{d\phi(\lambda^n)}(y) \frac{d\phi(\lambda^n)}{d\lambda^n}(y)$$

c'est-à-dire

$$g(y) = f(\phi^{-1}(y))|\det(J(\phi^{-1})(y))| \ .$$

Inversement, on a :

$$f(x) = g(\phi(x))|\det(J(\phi))(x)| \ .$$

Comme dans le cas réel, voir p. 230, on peut faire le lien avec les calculs classiques par "changement de variable" en intégration à plusieurs variables.

X.1.3 Espérance-Variance

Notion d'espérance mathématique, v.a.r. centrée

Soit X une v.a.r. définie sur l'espace mesurable $(\Omega, \mathcal{F})$, de loi P_X ; on appelle **espérance mathématique** (ou en bref **espérance**) de X par rapport à P, et on note $\mathbb{E}(X)$ (ou $\mathbb{E}_P(X)$ s'il est utile de préciser la probabilité P) l'intégrale (si elle a un sens) de X par rapport à P, autrement dit (théorème dit "de transfert") l'intégrale de l'identité par rapport à P_X ; ceci s'exprime avec différentes notations (avec ou sans lettre muette) :

$$\mathbb{E}(X) = \int_\Omega X \ dP = \int_\Omega X(\omega) \ dP(\omega) = \int_{\mathbb{R}} x \ dP_X(x) \ .$$

$\mathbb{E}(X)$ est définie (valant éventuellement $+\infty$ ou $-\infty$) comme différence des espérances des deux v.a. positives X^+ $(= \sup(X, 0))$ et X^- $(= \sup(-X, 0))$, **sauf si** $\mathbb{E}(X^+) = \mathbb{E}(X^-) = +\infty$. On a alors : $\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$.

X est dite **intégrable** (par rapport à P) si $-\infty < \mathbb{E}(X) < +\infty$, ce qui équivaut à $\mathbb{E}(X^+) < +\infty$ et $\mathbb{E}(X^-) < +\infty$, autrement dit $\mathbb{E}(|X|) < +\infty$. L'ensemble des v.a. intégrables par rapport à P est noté $\mathcal{L}^1(P)$.

$\mathbb{E}$ est, sur $\mathcal{L}^1(P)$, un opérateur linéaire $(\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y))$, positif (si $X \geq 0$, alors $\mathbb{E}(X) \geq 0$) et unitaire $(\mathbb{E}(1) = 1)$.

L'espérance est un *indicateur de valeur centrale* de la loi de X (usage justifié en particulier par la propriété $\mathbb{E}(X - a) = \mathbb{E}(X) - a$). Si X est intégrable, la v.a. $X - \mathbb{E}(X)$ est dite déduite de X par **centrage** (une v.a.r. **centrée** est une v.a.r. d'espérance nulle).

Calcul de l'espérance mathématique d'une v.a.r.

Si la loi P_X est absolument continue par rapport à une mesure σ-finie μ (voir X.1.1) et de densité f_X relativement à cette mesure, on a :

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x f_X(x) d\mu(x) \ ,$$

qui a toujours un sens (éventuellement $+\infty$ ou $-\infty$), sauf si à la fois $\int_0^{+\infty} x f_X(x) d\mu(x) = +\infty$ et $\int_{-\infty}^0 x f_X(x) d\mu(x) = -\infty$; alors :

$$\mathbb{E}(X) = \int_0^{+\infty} x f_X(x) d\mu(x) + \int_{-\infty}^0 x f_X(x) d\mu(x) \ .$$

Ceci conduit à deux modes de calcul :

– **Calcul intégral usuel si X est absolument continue p.r.p. la mesure de Lebesgue.**
On a dans ce cas :

$$\mathbb{E}(X) = \int_0^{+\infty} x f_X(x) dx + \int_{-\infty}^0 x f_X(x) dx \ ,$$

à mener soit par calcul de primitives (à condition que f_X soit continue ou continue par morceaux) soit par techniques de Monte-Carlo.

– **Calcul de sommes finies ou de sommes de séries à termes positifs si X est discrète**
Dans ce cas X est (voir X.1.1) absolument continue par rapport à la mesure de comptage sur l'ensemble B (contenu dans $\mathbb{R}$) des valeurs que prend X, l'application $x \mapsto P_X(\{x\})$ s'interprétant comme une densité par rapport à cette mesure ; si on note $B^+ = \{x \in B; x > 0\}$ et $B^- = \{x \in B; x < 0\}$, on a dans ce cas :

$$\mathbb{E}(X) = \sum_{x \in B^+} x P_X(\{x\}) + \sum_{x \in B^-} x P_X(\{x\}) \ ,$$

sauf si on a à la fois $\sum_{x\in B^+} x P_X(\{x\}) = +\infty$ et $\sum_{x\in B^-} x P_X(\{x\}) = -\infty$, ceci ne pouvant se produire que si les ensembles B^+ et B^- sont tous deux infinis dénombrables. Dans le cas (très fréquent dans les modélisations usuelles ; voir X.1.2) où B est fini (notons alors, en croissant, $B = \{x_1, \cdots, x_k\}$) ou bien infini dénombrable et tel que ses élements puissent être ordonnés en croissant (notons de même $B = \{x_1, \cdots, x_k, \cdots\}$), $\mathbb{E}(X)$ a nécessairement un sens et, notant $p_i = P(X = x_i) = P_X(\{x_i\})$, il vient :
- si B est fini, $\mathbb{E}(X) = \sum_{i=1}^k x_i p_i$,
- si B est infini dénombrable, $\mathbb{E}(X) = \sum_{i=1}^{+\infty} x_i p_i = x_1 + \sum_{i=1}^{+\infty} (x_i - x_1) p_i$ (cette dernière somme étant celle d'une série à termes positifs).

Variance d'une v.a. réelle

Si X est intégrable, on définit la **variance** de X (relativement à P) comme le nombre positif ou nul (éventuellement $+\infty$), noté ici $\mathrm{Var}(X)$ (ou $\mathrm{Var}_P(X)$ s'il est utile de préciser la probabilité P) défini par :

$$\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \ .$$

La racine carrée de la variance est appelée **écart-type** de X (relativement à P), noté ici $\sigma(X)$ (ou $\sigma_P(X)$).

Les propriétés et modes de calcul de la variance se déduisent de ceux des espérances.

L'ensemble des v.a.r. de variance finie (autrement dit de second moment, $\mathbb{E}(X^2)$, fini) est noté $\mathcal{L}^2(P)$ (aussi dit : ensemble des v.a. de carré intégrable).

La variance et l'écart-type sont des *indicateurs de dispersion* de la loi de X (usage justifié en particulier par les propriétés $\sigma(X - a) = \sigma(X)$ et $\sigma(bX) = |b|\sigma(X)$). Si X est de variance finie, la v.a. $(X - \mathbb{E}(X))/\sigma(X)$ est dite déduite de X par **centrage** et **réduction** (une v.a.r. **réduite** est une v.a.r. d'écart-type égal à 1).

Une v.a.r. de variance nulle est p.s. constante : sa loi est la **loi de Dirac** en un point x_0, c'est-à-dire portée par ce seul point.

Covariance de deux v.a. réelles

Soit X_1 et X_2 deux v.a.r. définies sur l'espace mesurable $(\Omega, \mathcal{F})$. Si X_1^2 et X_2^2 sont intégrables, il en est ainsi (en vertu de l'inégalité de Hölder) du produit X_1X_2 et on définit la **covariance** de X_1 et X_2 (relativement à P) comme le nombre réel, noté ici $\mathrm{Cov}(X_1, X_2)$ (ou $\mathrm{Cov}_P(X_1, X_2)$ si nécessaire) défini par :

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}((X_1 - \mathbb{E}(X_1)(X_2 - \mathbb{E}(X_2)) = \mathbb{E}(X_1X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2) \ .$$

Etant donné n v.a. $(X_1, \ldots, X_n)$, on a de manière évidente l'égalité :

$$\mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}(X_i, X_j) \ .$$

Si X_1 et X_2 ne sont pas presque sûrement constantes (et donc de variance non nulle), on appelle **corrélation** de X_1 et X_2 (relativement à P) le nombre, appartenant à l'intervalle $[-1, +1]$ (encore une conséquence de l'inégalité de Holder), noté ici $\mathrm{Corr}(X_1, X_2)$ (ou bien $\mathrm{Corr}_P(X_1, X_2)$ si nécessaire) défini par :

$$\mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)\,\mathrm{Var}(X_2)}} \ .$$

La corrélation est un *indicateur de dépendance* entre X_1 et X_2 (ce mot de "dépendance" étant pris ici dans un sens "intuitif"). En particulier la valeur absolue de la corrélation est maximale si il y a une relation affine entre X_1 et X_2, à condition qu'aucune des ces deux v.a. ne soit constante : si $X_2 = aX_1 + b$ avec $a > 0$, il vient $\mathrm{Corr}(X_1, X_2) = +1$ et, si $X_2 = aX_2 + b$ avec $a < 0$, il vient $\mathrm{Corr}(X_1, X_2) = -1$.

Espérance mathématique d'un vecteur aléatoire

Soit $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ un vecteur aléatoire (voir X.1.2). On appelle espérance mathématique de X, le vecteur $\mathbb{E}(X) = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_n) \end{pmatrix}$.

Il résulte de la linéarité et de l'unitarité de l'opérateur $\mathbb{E}$ pour les v.a.r. que, si A est une matrice à m lignes et n colonnes et B un vecteur de $\mathbb{R}^m$, alors $\mathbb{E}(AX + B) = A\mathbb{E}(X) + B$.

Variance d'un vecteur aléatoire

On appelle **matrice de variances et covariances** de X, ou en bref **variance** de X, la matrice à n lignes et n colonnes $\mathrm{Var}(X)$ de terme général $v_{i,j}$ tel que :
- pour tout i, $v_{i,i} = \mathrm{Var}(X_i)$,
- pour tout (i, j) tel que $i \neq j$, $v_{i,j} = \mathrm{Cov}(X_i, X_j)$.

Autrement dit, l'espérance d'une matrice aléatoire étant définie de manière évidente dans la ligne de l'espérance d'un vecteur aléatoire, on a $\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))(X - \mathbb{E}(X))^t)$, où le symbole ^{t} désigne la transposition des matrices.

$\mathrm{Var}(X)$ est une matrice symétrique ; elle est positive, c'est-à-dire que pour tout vecteur u (élément de $\mathbb{R}^n$, représenté par une matrice unicolonne à n lignes) elle vérifie $u^t \mathrm{Var}(X)u \geq 0$ (où u^t est donc une matrice uniligne à n colonnes).

$\mathrm{Var}(X)$ n'est pas nécessairement de rang n (autrement dit elle n'est pas nécessairement définie positive). Si elle est de rang $k < n$, le support de la loi de X est contenu dans l'hyperplan affine, de dimension k, passant par $\mathbb{E}(X)$ et parallèle à l'hyperplan vectoriel $\mathrm{Im}(\mathrm{Var}(X))$ (ensemble des vecteurs de la forme $\mathrm{Var}(X)u$ où u parcourt $\mathbb{R}^n$).

Si A est une matrice à m lignes et n colonnes, la variance du vecteur aléatoire m-dimensionnel AX est : $\mathrm{Var}(AX) = A\,\mathrm{Var}(X)A^t$; en particulier, si $m = 1$, $Y = AX$ est une v.a. réelle admettant pour variance le nombre positif ou nul $\mathrm{Var}(Y) = A\,\mathrm{Var}(X)A^t$ (ce qui explique la positivité de $\mathrm{Var}(X)$ citée plus haut).

Emploi du terme "empirique"

A toute suite finie $(x_1, \ldots, x_n)$ d'éléments de $\mathbb{R}$ on associe la **probabilité empirique**, sur $(\mathbb{R}, \mathcal{B})$ définie par

$$B \mapsto \frac{\mathrm{Card}(\{i : x_i \in B\})}{n}$$

(autrement dit, on affecte la masse $\frac{1}{n}$ à chaque x_i, mais attention aux ex-aequo).

Tout objet relatif à la probabilité empirique sera qualifié d'empirique ; en voici des exemples usuels :

- fonction de répartition empirique : $t \mapsto \dfrac{1}{n}\mathrm{card}(\{i : x_i \leq t\}) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \mathbf{1}_{\{x_i \leq t\}}$,
- espérance mathématique empirique : $\bar{x}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$,
- variance empirique : $v_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2 = \dfrac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$.

Inégalité de Jensen

Pour toute fonction convexe $\varphi : \mathbb{R} \to \mathbb{R}$ et toute variable aléatoire X intégrable à valeurs dans $\mathbb{R}$:

$$\varphi(\mathbb{E}X) \leq \mathbb{E}\varphi(X). \tag{X.1}$$

X.1.4 Conditionnement, indépendance

Conditionnement par un évènement

Etant donné, sur un espace mesurable $(\Omega, \mathcal{F})$, une mesure de probabilité $\mathbb{P}$ et un évènement de probabilité non nulle B, la **mesure de probabilité déduite de $\mathbb{P}$ par conditionnement par B**, notée P^B, est définie, sur le même espace $(\Omega, \mathcal{F})$, par :

$$\mathbb{P}^B(A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}\,.$$

Elle est concentrée sur B (c'est-à-dire que $\mathbb{P}^B(B) = 1$). $\mathbb{P}^B(A)$ se lit, si on n'a pas besoin de faire référence à $\mathbb{P}$, **probabilité de A conditionnée par B**, ou **probabilité de A**

conditionnellement à B, ou encore **probabilité de** A **sachant** B, ou enfin (avec un pléonasme) **probabilité conditionnelle de** A **sachant** B .

On emploie aussi la notation $\mathbb{P}(A|B)$ (ou parfois $\mathbb{P}_B(A)$) là où nous venons de noter $\mathbb{P}^B(A)$.

Concrètement, le remplacement de $\mathbb{P}$ par $\mathbb{P}^B$ se présente quand un certain contexte d'informations, relatif à une situation aléatoire, a conduit à la modélisation par $(\Omega, \mathcal{F}, \mathbb{P})$, puis que l'*on fournit l'information supplémentaire* que B est réalisé. Ceci conduit à affecter la valeur 1 à la probabilité de B (autrement dit 0 à son complémentaire) et à maintenir "l'équilibre interne" entre les valeurs des probabilités des événements inclus dans B ; c'est bien ce que fait P^B.

Conditionnement par une v.a. discrète

Dans la pratique de la modélisation intervient souvent la démarche inverse de celle décrite ci-dessus : l'analyse de la situation concrète justifie de porter son attention sur une partition finie ou infinie dénombrable de Ω, soit $(B_i)_{i\in I}$ et de choisir :

- d'une part les probabilités $\mathbb{P}(B_i)$,
- d'autre part, pour tout i tel que $\mathbb{P}(B_i) \neq 0$, la probabilité conditionnelle sachant B_i, $\mathbb{P}^{B_i}$.

Alors, si on note J $(\subset I)$ l'ensemble des j tels que $\mathbb{P}(B_j) \neq 0$, on obtient $\mathbb{P}$ par la "formule des probabilités conditionnelles"

$$\forall A \in \mathcal{F} \;\; \mathbb{P}(A) = \sum_{j\in J} \mathbb{P}^{B_j}(A)\mathbb{P}(B_j) \; .$$

Remarquons que la donnée de la famille $(\mathbb{P}(B_i))_{i\in I}$ peut s'interpréter comme la donnée de la loi, P_X, d'une variable aléatoire X, à valeurs dans un espace $(\Omega', \mathcal{F}')$ et pour laquelle l'ensemble des valeurs prises est indexé par I ; notons les x_i, où $i \in I$, de telle sorte que, pour tout i, $B_i = \{X = x_i\}$ (notation introduite en X.1.1) ; toute application de Ω' dans l'ensemble des probabilités sur $(\Omega, \mathcal{F})$ qui, à tout x_i, où $i \in J$, associe $\mathbb{P}^{[X=x_i]}$ est appelée **probabilité conditionnelle relativement à** X ; notons la $\mathbb{P}^X$ et remarquons que le fait que, pour tout ω' non égal à l'un des x_i (où $i \in J$), $\mathbb{P}^{[X=\omega']}$ soit arbitraire (et en particulier ne puisse pas s'interpréter comme une probabilité conditionnelle usuelle) n'est pas gênant car l'ensemble de ces ω' est de probabilité nulle (en effet $\sum_{j\in J} \mathbb{P}(X = x_j) = \sum_{j\in J} \mathbb{P}_X(\{x_j\}) = 1$). Dans ce contexte, la "formule des probabilités conditionnelles" s'écrit :

$$\forall A \in \mathcal{F} \;\; \mathbb{P}(A) = \int_{\Omega'} \mathbb{P}^{[X=\omega']}(A) \; dP_X(\omega') \; ,$$

autrement dit :

$$\forall A \in \mathcal{F} \;\; \mathbb{P}(A) = \mathbb{E}_{P_X}(P^X(A)) \; .$$

Rappelons par ailleurs que, pour presque tout ω', la probabilité $\mathbb{P}^{[X=\omega']}$ est portée par l'évènement $\{X = \omega'\}$.

On emploie aussi la notation $\mathbb{P}(A|X = \omega')$ là où nous avons noté $\mathbb{P}^{[X=\omega']}(A)$.

Conditionnement par une variable aléatoire : cas général

Cette notice est assez difficile ; le lecteur peut se contenter de rechercher le calcul des lois conditionnelles figurant en X.1.4.

Soit donné, sur un espace mesurable $(\Omega, \mathcal{F})$, une mesure de probabilité $\mathbb{P}$ et soit X une v.a. à valeurs dans $(\Omega', \mathcal{F}')$. Alors on appelle **probabilité conditionnelle relativement à** X toute application de Ω' dans l'ensemble des probabilités sur $(\Omega, \mathcal{F})$ qui à tout ω' associe une probabilité, notée $\mathbb{P}^{[X=\omega']}$, qui vérifie les deux propriétés obtenues dans le cas des v.a. discrètes :

• **concentration :** pour presque tout ω', la probabilité $\mathbb{P}^{[X=\omega']}$ est portée par l'évènement $\{X = \omega'\}$,

• **reconstitution :**

$$\forall A \in \mathcal{F} \;\; \mathbb{P}(A) = \int_{\Omega'} \mathbb{P}^{[X=\omega']}(A) \; dP_X(\omega') \; .$$

Comme dans le cas discret, on emploie aussi la notation $\mathbb{P}(A|X = \omega')$ là où nous venons de noter $\mathbb{P}^{[X=\omega']}(A)$.

On en déduit que pour toute v.a. Y, $\mathbb{P}$-intégrable,

$$\int_{\Omega} Y(\omega) d\mathbb{P}(\omega) = \int_{\Omega'} (\int_{\Omega} Y(\omega) d\mathbb{P}^{[X=\omega']}(\omega)) dP_X(\omega') \; ,$$

autrement écrit de manière condensée :

$$\mathbb{E}(Y) = \mathbb{E}_{P_X}(\mathbb{E}^X(Y)) \; ;$$

$\mathbb{E}^X(Y)$ est appelé **espérance conditionnelle de** Y **relativement à** X. On emploie aussi la notation $\mathbb{E}(Y|X = \omega')$ pour ce qui, dans la logique de la notation que nous venons d'introduire, s'écrit $\mathbb{E}^{[X=\omega']}(Y)$.

Insistons sur le fait qu'il s'agit d'une définition descriptive et non constructive : en général, $\mathbb{P}^{[X=\omega']}$ ne peut pas être obtenu de manière élémentaire car $\mathbb{P}(X = \omega')$ peut fort bien valoir 0 pour tout ω' (c'est par exemple le cas si X est à valeurs réelles et de fonction de répartition continue). Comme toute définition descriptive, celle-ci pose des problèmes d'existence ; en fait, dans tous les cas qui nous seront nécessaires, nous disposerons d'un outil constructif en termes de densités (voir X.1.4 ci-dessous) .

Un cas particulier utile est celui de la loi d'un couple de v.a. (X_1, X_2), à valeurs dans un espace produit $\Omega_1 \times \Omega_2$, que l'on veut conditionner par X_1 ; la loi conditionnelle de (X_1, X_2) est alors caractérisée par celle de X_2, soit $P_{X_2}^{X_1}$, qui vérifie

$$P_{X_2}(A_2) = \int_{\Omega_1} P_{X_2}^{[X_1=x_1]}(A_2) dP_{X_1}(x_1)$$

et bien sûr, sur l'espace produit, $P^{[X_1=x_1]}$ est portée par le "fil" $\{x_1\} \times \Omega_2$.

Espérance conditionnelle

Nous donnons ici une autre interprétation pour l'espérance conditionnelle d'une v.a.r., Z, de carré intégrable par rapport à une v.a. X, notée $\mathbb{E}[Z|X]$. La v.a.r. $\mathbb{E}[Z|X]$ s'interprète comme la projection orthogonale, avec le produit scalaire défini sur l'espace des v.a.r. de

carré intégrable par $(Y,Z)=\mathbb{E}[YZ]$, sur l'ensemble, $\mathcal{H}_X$, des fonctions réelles mesurables de X de carré intégrables. En particulier, la v.a.r. $\mathbb{E}[Z|X]$ est une fonction réelle mesurable de X, et pour tout $\varphi(X)\in\mathcal{H}_X$, on a pour toute fonction réelle φ, telle que $\varphi(X)$ est de carré intégrable,

$$\mathbb{E}[Z\varphi(X)]=(Z,\varphi(X))=(\mathbb{E}[Z|X],\varphi(X))=\mathbb{E}[\mathbb{E}[Z|X]\varphi(X)],$$

En particulier, comme $\mathbf{1}\in\mathcal{H}_X$ et $\mathbb{E}[Z|X]\in\mathcal{H}_X$, on a

$$\mathbb{E}[\mathbb{E}[Z|X]]=\mathbb{E}[Z]\quad\text{et}\quad\mathbb{E}[Z\mathbb{E}[Z|X]]=\mathbb{E}[\mathbb{E}[Z|X]^2].$$

Remarquons enfin que $\mathbb{E}[\varphi(X)|X]=\varphi(X)$, que l'on peut comprendre en remarquant que si X est connu, alors $\varphi(X)$ est aussi connu.

Indépendance de 2 évènements

Deux évènements A et B sont dits **indépendants** (relativement à une probabilité $\mathbb{P}$) si $\mathbb{P}(A\cap B)=\mathbb{P}(A)\mathbb{P}(B)$.

Si $\mathbb{P}(B)\neq 0$, cette définition équivaut à $\mathbb{P}^B(A)=\mathbb{P}(A)$, ce qui justifie la terminologie "indépendants" ; en effet alors une information sur la réalisation de B n'a aucune influence sur la valeur de la probabilité affectée à A.

Les évènements de probabilité 0 (dits **presque impossibles**) ou de probabilité 1 (dits **presque certains**) sont indépendants de tout autre évènement.

Indépendance de 2 variables aléatoires

Deux v.a., définies sur un même espace $(\Omega,\mathcal{F})$, sont dites **indépendantes** (relativement à une probabilité $\mathbb{P}$ sur cet espace) si tout évènement exprimé en termes de réalisation de X_1 est indépendant de tout évènement exprimé en termes de réalisation de X_2 ; autrement dit, X_1 étant à valeurs dans $(\Omega_1,\mathcal{F}_1)$ et X_2 étant à valeurs dans $(\Omega_2,\mathcal{F}_2)$, on a, pour tout couple (B_1,B_2), où $B_1\in\mathcal{F}_1$ et $B_2\in\mathcal{F}_2$, l'égalité

$$\mathbb{P}(\{X_1\in B_1\}\cap\{X_2\in B_2\})=\mathbb{P}(X_1\in B_1)\mathbb{P}(X_2\in B_2)\ .$$

Cette propriété peut se traduire aussi, à l'aide des lois de X_1, X_2 et de celle du couple (X_1,X_2) (qui est définie sur l'espace produit $(\Omega_1\times\Omega_2,\mathcal{F}_1\times\mathcal{F}_2)$ (voir X.1.1), par l'égalité : $P_{(X_1,X_2)}(B_1\times B_2)=P_{X_1}(B_1)P_{X_2}(B_2)$. On dit que la loi du couple (X_1,X_2) est **le produit** des lois de X_1 et X_2. Cette notion de produit de probabilités s'étend immédiatement au produit des mesures σ-finies : si μ_1 (resp. μ_2) est une mesure σ-finie sur Ω_1 (resp. Ω_2) il existe sur $\Omega_1\times\Omega_2$ une unique mesure notée $\mu_1\times\mu_2$ vérifiant $(\mu_1\times\mu_2)(B_1\times B_2)=\mu_1(B_1)\mu_2(B_2)$ (ici encore, si on adopte pour les tribus la notation tensorielle, on note $\mu_1\otimes\mu_2$ là où nous avons noté $\mu_1\times\mu_2$).

Si X_1 et X_2 sont indépendantes, il en est de même de tout couple de v.a. $\phi(X_1)$ et $\psi(X_2)$.

Si X_1 et X_2 sont indépendantes et intégrables, on a $\mathbb{E}(X_1X_2)=\mathbb{E}(X_1)\mathbb{E}(X_2)$ (mais cette égalité ne caractérise pas l'indépendance).

Si X_1 et X_2 sont indépendantes et de carrés intégrables, on a (voir X.1.3) $\mathrm{Cov}(X_1,X_2)=0$, autrement dit, si de plus $\mathrm{Var}(X_1)>0$ et $\mathrm{Var}(X_2)>0$, $\mathrm{Corr}(X_1,X_2)=0$; ceci complète l'interprétation déjà donnée de la corrélation comme *indicateur de dépendance* entre les 2 v.a. (on dit aussi qu'elles sont **non corrélées**) : mais cette propriété ne caractérise pas

l'indépendance. Il en résulte que $\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$ (et cette égalité non plus ne caractérise pas l'indépendance).

Si X_1 est absolument continue par rapport à une mesure μ_1 sur $(\Omega_1, \mathcal{F}_1)$, relativement à laquelle elle admet une densité f_1 et X_2 est a.c. par rapport à une mesure μ_2 sur $(\Omega_2, \mathcal{F}_2)$, relativement à laquelle elle admet une densité f_2 et si de plus X_1 et X_2 sont indépendantes, alors le couple (X_1, X_2) de v.a. indépendantes est aussi absolument continu par rapport à la mesure produit $\mu_1 \times \mu_2$, relativement à laquelle il admet pour densité l'application : $(\omega_1, \omega_2) \mapsto f_1(\omega_1) f_2(\omega_2)$.

Réciproquement, si la densité f de (X_1, X_2) par rapport à la mesure produit $\mu_1 \times \mu_2$ s'écrit sous la forme $f(\omega_1, \omega_2) = g_1(\omega_1) g_2(\omega_2)$, les v.a. X_1 et X_2 sont indépendantes et leurs densités, par rapport respectivement à μ_1 et μ_2, sont proportionnelles à $g_1(\omega_1)$ et à $g_2(\omega_2)$. Cette propriété s'étend aux mesures σ-finies : si une mesure μ sur $\Omega_1 \times \Omega_2$ admet par rapport à $\mu_1 \times \mu_2$ une densité f de la forme $f(\omega_1, \omega_2) = g_1(\omega_1) g_2(\omega_2)$, elle est elle-même une mesure produit, soit $\nu_1 \times \nu_2$, et les densités de ν_1 (resp. ν_2) par rapport à μ_1 (resp. μ_2) sont proportionnelles respectivement à $g_1$et g_2.

S'il s'agit de v.a. réelles, une condition nécessaire et suffisante d'indépendance s'exprime à l'aide des fonctions de répartition (voir X.1.2) : $F_{(X_1,X_2)}(x_1, x_2) = F_{X_1}(x_1) F_{X_2}(x_2)$.

Indépendance de n évènements

n (où $n \geq 2$) évènements $A_1, \ldots, A_n$ sont dits **indépendants** (on précise parfois **indépendants dans leur ensemble**) si, pour toute sous-famille $(A_i)_{i \in I}$, où $I \subseteq \{1, \ldots, n\}$ et $\mathrm{card}(I) \geq 2$, on a : $\mathbb{P}(\bigcap_{i \in I} A_i) = \prod_{i \in I} \mathbb{P}(A_i)$.

Alors, si J et K sont deux parties disjointes de $\{1, \ldots, n\}$, tout évènement exprimé à partir de $(A_j)_{j \in J}$ est indépendant de tout évènement exprimé à partir de $(A_k)_{k \in K}$.

Indépendance de n variables aléatoires

Soit $n \geq 2$. n variables aléatoires X_i (où $1 \leq i \leq n$), définies sur un même espace $(\Omega, \mathcal{F})$, sont dites **indépendantes** (relativement à une probabilité $\mathbb{P}$ sur cet espace) si toute suite d'évènements $(\{X_i \in B_i\})_{1 \leq i \leq n}$ est composée d'évènements indépendants, autrement dit si, pour toute famille $(B_i)_{1 \leq i \leq n}$ où, pour tout i, $B_i \in \mathcal{F}_i$, on a : $\mathbb{P}(\bigcap_{i=1}^n \{X_i \in B_i\}) = \prod_{i=1}^n \mathbb{P}(X_i \in B_i)$.

En d'autres termes, si on note P_i la loi de X_i, l'indépendance des n variables aléatoires équivaut au fait que la loi P_X de la variable aléatoire $X = (X_i)_{1 \leq i \leq n}$, à valeurs dans l'espace produit (voir X.1.1) $(\Omega_I, \mathcal{F}_I) = (\prod_{i=1}^n \Omega_i, \prod_{i=1}^n \mathcal{F}_i)$, vérifie les égalités : $P_X(\prod_{i=1}^n B_i) = \prod_{i=1}^n P_i(B_i)$. On dit que P_X est le **produit des probabilités** P_i et on la note donc aussi $\prod_{i=1}^n P_i$ (ou $\bigotimes_{i=1}^n P_i$). Dans le cas où les n v.a. indépendantes sont à valeurs dans un même espace $(\Omega, \mathcal{F})$ et de même loi Q, on parlera de **probabilité puissance** Q^n (ou $Q^{\otimes n}$) sur l'espace puissance $(\Omega^n, \mathcal{F}^n)$.

Si, pour tout i, X_i admet la densité f_i par rapport à une mesure σ-finie μ_i sur Ω_i, l'indépendance des n variables aléatoires équivaut au fait que la loi P_X de la v.a. $(X_i)_{1 \leq i \leq n}$ admet pour densité par rapport à la mesure produit $\prod_{i=1}^n \mu_i$, l'application

$$(\omega_1, \ldots, \omega_1) \mapsto \prod_{i=1}^n f_i(\omega_i) \ .$$

Si $X_1, \cdots, X_n$ sont indépendantes et J et K sont deux parties disjointes de $\{1, \ldots, n\}$, il y a indépendance pour tout couple $(\phi((X_j)_{j\in J}), \psi((X_k)_{k\in K}))$.

Les autres propriétés des familles de n v.a. (ou v.a.r.) indépendantes reproduisent celles des couples ; nous ne les détaillons pas ici .

Calcul de la densité de la loi d'une v.a. : cas d'un espace produit

On sait (voir X.1.1) que la densité de la loi d'une variable aléatoire est souvent délicate à calculer : un cas où on dispose d'une méthode simple est celui d'un couple (X_1, X_2) de v.a., à valeurs dans un espace produit $(\Omega_1 \times \Omega_1, \mathcal{F}_1 \times \mathcal{F}_2)$.

Soit sur cet espace une mesure produit $\mu = \mu_1 \times \mu_2$ (c'est-à-dire, comme pour les probabilités produits, que, pour tout pavé mesurable $A_1 \times A_2$, on a $\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$). On suppose la loi $P_{(X_1,X_2)}$ de (X_1, X_2) absolument continue par rapport à μ, et de densité f. Alors il résulte du théorème de Fubini que la loi de X_1 admet pour densité par rapport à μ_1 l'application f_1 définie par : $f_1(x_1) = \int_{\Omega_2} f(x_1, x_2) d\mu_2(x_2)$ (on est bien dans le cas général annoncé, en prenant pour X_1 la projection de $\Omega_1 \times \Omega_1$ sur Ω_1).

Cette situation intervient en particulier dans des situations où $\Omega_1 = \mathbb{R}^{k_1}$, $\Omega_2 = \mathbb{R}^{k_2}$ et μ_1 et μ_2 sont les mesures de Lebesgue ; alors, si on note $x_1 = (t_1, \ldots, t_{k_1})$ et $x_2 = (t_{k_1+1}, \ldots, t_{k_1+k_2})$, la densité de la loi de X_1 s'écrit :

$$f_1(t_1, \ldots, t_{k_1}) = \int_{\mathbb{R}^{k_2}} f(t_1, \ldots, t_{k_1}, t_{k_1+1}, \ldots, t_{k_1+k_2}) dt_{k_1+1} \ldots dt_{k_1+k_2}.$$

Calcul de la densité d'une loi conditionnelle

Les deux cas de calcul de densité dégagés en X.1.1 et p. 241 nous donnent des possibilités de calcul de densité de probabilité conditionnelle relativement à la v.a. X ; on rappelle que, pour tout ω' appartenant à l'espace d'arrivée de X et tout évènement A dans l'espace de départ de X, on note $\mathbb{P}^{[X=\omega']}(A)$ (ou $\mathbb{P}(A|X = \omega')$) "la" probabilité de A sachant que X prend la valeur ω' (voir p. 238 pour une définition rigoureuxe de cette notion).

a. Si la densité f se factorise à travers X, c'est-à-dire que $f(\omega) = g(X(\omega))$, alors, pour presque tout ω', l'application qui à tout $\omega \in \{X = \omega'\}$ associe le quotient $f(\omega)/g(\omega')$ est une densité de $\mathbb{P}^{[X=\omega']}$ par rapport à une mesure concentrée sur $\{X = \omega'\}$ (dont nous ne détaillerons pas ici la construction).

b. Soit un couple (X_1, X_2) de v.a. , à valeurs dans un espace produit $(\Omega_1 \times \Omega_1, \mathcal{F}_1 \times \mathcal{F}_2)$. La loi $P_{(X_1,X_2)}$ de (X_1, X_2) admet f pour densité par rapport à $\mu_1 \times \mu_2$. Alors la loi de X_2 conditionnelle relativement à X_1, $P_{X_2}^{X_1}$, est telle que, pour presque tout x_1, $P_{X_2}^{[X_1=x_1]}$ admet pour densité par rapport à μ_2 l'application $f_2^{[X_1=x_1]}$ définie par : $f_2^{[X_1=x_1]}(x_2) = \frac{f(x_1,x_2)}{f_1(x_1)}$.

N.B. Le "presque tout x_1" qui figure dans cette propriété est destiné à couvrir les cas où il existe des valeurs x_1 telles que $f_1(x_1) = 0$; on n'a pas à en tenir compte s'il n'y pas de telles valeurs x_1.

On remarque que **dans le cas où X_1 et X_2 sont indépendantes**, et où donc (voir X.1.4) on peut prendre $f(x_1, x_2) = f_1(x_1).f_2(x_2)$, on obtient que $f_2^{[X_1=x_1]}(x_2) = f_2(x_2)$, ce qui est très satisfaisant, car cela exprime que le conditionnement de X_2 par X_1 n'a pas d'effet sur la densité de X_2.

X.2 Lois de variables aléatoires remarquables

X.2.1 Variables aléatoires discrètes

L'expression *variable aléatoire discrète* a été introduite en X.1.2 pour les v.a. réelles. On l'emploie pour toute v.a. X prenant ses valeurs dans un ensemble fini ou infini dénombrable, soit Ω'. Sa loi P_X est donc entièrement déterminée par ses valeurs pour les évènements élémentaires, $P_X(\{x\})$ $(= \mathbb{P}(X = x))$, où $x \in \Omega'$. L'abus de notation consistant à noter $P_X(x)$ au lieu de $P_X(\{x\})$ est fréquent.

Nous pouvons (voir X.1.1) considérer l'application $x \mapsto P_X(x)$ comme une densité de X ; il ne s'agit pas ici de la densité "usuelle" (c'est-à-dire relativement à la mesure de Lebesgue, que nous utiliserons pour les lois étudiées en X.2.2 ci-dessous) mais de la densité relativement à la mesure de comptage sur Ω.

Loi de Dirac

La loi d'une variable aléatoire presque-sûrement constante a (autrement dit une probabilité concentrée sur un évènement élémentaire $\{a\}$) est appelée loi de Dirac en ce point et notée $\mathcal{D}(a)$, ou $\delta(a)$: $(\mathcal{D}(a))(\{a\}) = 1$.

Si $a \in \mathbb{R}$ et $X \sim \mathcal{D}(a)$, on a $\mathbb{E}(X) = a$ et $\mathrm{Var}(X) = 0$ (les v.a. réelles p.s. constantes sont les seules de variance nulle)

Loi de Bernoulli

Nous la noterons $\mathcal{B}(p)$, où $p \in [0, 1]$; c'est la loi de probabilité sur $\{0, 1\}$ qui affecte la valeur p à $\{1\}$ et donc la valeur $1 - p$ à $\{0\}$: $(\mathcal{B}(p))(\{1\}) = p$ et $(\mathcal{B}(p))(\{0\}) = 1 - p$.

Si $X \sim \mathcal{B}(p)$, on a $\mathbb{E}(X) = p$ et $\mathrm{Var}(X) = p(1 - p)$.

Dans le cas particulier où $p = 0$ (resp. $p = 1$), il s'agit de la loi de Dirac en 0 (resp. 1) $\mathcal{D}(0)$ (resp. $\mathcal{D}(1)$).

Loi Binomiale

Nous la noterons $\mathcal{B}(n, p)$, où n est un entier strictement positif et $p \in [0, 1]$; c'est la loi de la somme de n v.a. indépendantes suivant toutes la loi de Bernoulli $\mathcal{B}(p)$. En d'autres termes, c'est la loi du nombre de "succès" dans une succession de n expériences indépendantes résultant toutes en un succès, avec probabilité p, ou un échec.

Dans la terminologie du contrôle de qualité c'est la loi du nombre de "bonnes pièces" quand on observe successivement n pièces d'une production supposée constante dans ses performances (la probabilité de production d'une "bonne pièce" est toujours p et celle d'une "pièce défectueuse" $1 - p$) et de telle sorte que la qualité de toute nouvelle pièce observée ne soit pas affectée par celles des pièces observées antérieurement ; on parle aussi dans ce cas de "tirage avec remise" car ceci est équivalent à la situation où on tire successivement n pièces d'un lot comportant une proportion p de "bonnes pièces", en remettant à chaque fois la pièce observée dans le lot et en "brassant" celui-ci de manière à se retrouver pour chaque tirage dans la situation initiale (ceci est transposable évidemment en termes, historiquement classiques, d'urnes contenant 2 types de boules).

On remarque que $\mathcal{B}(1, p) = \mathcal{B}(p)$.

$\mathcal{B}(n,p)$ a pour support l'ensemble d'entiers $\{0,\dots,n\}$, et, pour tout k dans cet ensemble, si $X \sim \mathcal{B}(n,p)$, on a

$$\mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k},$$

où $\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$ (coefficient binomial, aussi noté C_n^k). On a $\mathbb{E}(X) = np$ et $\mathrm{Var}(X) = np(1-p)$.

Si $p=0$ (resp. $p=1$), il s'agit de la **loi de Dirac** en 0 (resp. n), $\mathcal{D}(0)$ (resp. $\mathcal{D}(1)$).

La somme de deux variables aléatoires indépendantes binomiales, de lois respectives $\mathcal{B}(n_1,p)$ et $\mathcal{B}(n_2,p)$ (même valeur de p), suit la loi $\mathcal{B}(n_1+n_2,p)$.

Loi Multinomiale

Nous la noterons $\mathcal{M}(n,\underline{p})$, où n est un entier strictement positif et $\underline{p} = (p_1,\dots,p_m)$ est une probabilité sur un ensemble à m éléments, autrement dit est une suite de m nombres positifs ou nuls de somme égale à 1.

Si on effectue n expériences indépendantes, à valeurs dans un ensemble à m éléments, noté par exemple $\{a_1,\dots,a_m\}$, de telle sorte qu'à chaque expérience la probabilité d'obtention de a_j (où $1 \le j \le m$) soit p_j, $\mathcal{M}(n,\underline{p})$ est la loi de la suite des effectifs de résultats égaux respectivement à $a_1,\dots,a_m$.

$\mathcal{M}(n,\underline{p})$ a pour support l'ensemble des suites de longueur m d'entiers positifs ou nuls de somme égale à n. Pour toute telle suite $\underline{x} = (x_1,\dots,x_m)$, on a si $X = (X_1,\dots,X_m) \sim \mathcal{M}(n,p_1,\dots,p_m)$,

$$\mathbb{P}(X=(x_1,\dots,x_m)) = \frac{n!}{\prod_{j=1}^m x_j!} \prod_{j=1}^m p_j^{x_j},$$

où $\frac{n!}{\prod_{j=1}^m x_j!}$ est appelé coefficient multinomial. Pour tout j tel que $1 \le j \le m$, $X_j \sim \mathcal{B}(n,p_j)$.

Loi Hypergéométrique

Nous la noterons $\mathcal{H}(N,n,p)$, où N est un entier strictement positif, n est un entier strictement positif inférieur à N et $p \in [0,1]$, de sorte que $m = pN$ soit un entier strictment positif.

Dans la terminologie du contrôle de qualité (transposable évidemment en termes d'urnes contenant 2 types de boules) c'est la loi du nombre de "bonnes pièces" quand on tire avec équiprobabilité un sous-ensemble de n pièces dans un lot de N pièces dans lequel la proportion de "bonnes pièces" est p et celle de "pièces défectueuses" est $1-p$; on parle aussi dans ce cas de "tirage sans remise" car ceci est équivalent à la situation où on tire successivement n pièces distinctes, de telle sorte que à chaque tirage toutes les pièces subsistantes aient la même probabilité d'être tirées.

Si on note $m = pN$, cette probabilité est concentrée sur l'ensemble des entiers k vérifiant $\sup(0, n+m-N) \le k \le \inf(m,n)$ et , pour tout tel k, on a si $X \sim \mathcal{H}(N,n,p)$,

$$\mathbb{P}(X=k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

(autrement noté $\mathbb{P}(X=k)=\dfrac{C_m^k C_{N-m}^{n-k}}{C_N^k}$)On a aussi $\mathbb{E}(X)=np$ (comme pour $\mathcal{B}(n,p)$) et $\mathrm{Var}(X)=\dfrac{N-n}{N-1}np(1-p)$ (plus petit que pour $\mathcal{B}(n,p)$).

Si $N\to\infty$, on a pour tout $k\in\mathbb{N}$, $\mathbb{P}(X=k)\to\mathbb{P}(S=k)$, où $S\sim\mathcal{B}(n,p)$. Concrètement, pour le statisticien, ceci exprime qu'un tirage avec remise approxime d'autant mieux un tirage sans remise que l'effectif de la population est plus grand devant la taille n de l'échantillon tiré.

Loi de Poisson

Nous la noterons $\mathcal{P}(\lambda)$, où $\lambda\in\mathbb{R}_+^*$ (autrement dit λ est strictement positif).
Son support est l'ensemble des entiers naturels $\mathbb{N}$ et si $X\sim\mathcal{P}(\lambda)$, on a pour $k\in\mathbb{N}$,

$$\mathbb{P}(X=k)=\mathrm{e}^{-\lambda}\,\frac{\lambda^k}{k!},$$

ainsi que $\mathbb{E}(X)=\lambda$ et $\mathrm{Var}(X)=\lambda$.

La somme de n v.a. indépendantes suivant des lois de Poisson de paramètres $\lambda_1\ldots,\lambda_n$ suit la loi de Poisson de paramètre $\sum_{i=1}^n\lambda_i$.

X.2.2 Variables aléatoires absolument continues

Dans toute cette partie, l'absolue continuité (a.c.) s'entend par rapport à la mesure de Lebesgue sur $\mathbb{R}$ ou $\mathbb{R}^n$ (voir X.1.2). Donc ici, si $n=1$ (ce qui sera le cas sauf pour la loi de Gauss multi-dimensionnelle), si on considère une v.a. réelle X, elle admet une densité (définie à une égalité presque sûre près), c'est-à-dire une fonction f de $\mathbb{R}$ dans $\mathbb{R}_+$ liée à sa fonction répartition F par les relations $F(x)=\mathbb{P}(X\le x)=\int_{-\infty}^x f(t)dt$ et, pour presque tout x, $f(x)=F'(x)$; il en résulte que, pour tout couple (a,b) vérifiant $a\le b$, on a :

$$\mathbb{P}(a\le X\le b)=F(b)-F(a)=\int_a^b f(t)dt$$

(les inégalités larges pouvant être remplacées par des inégalités strictes) et en particulier, pour tout x, $\mathbb{P}(X=x)=0$.

Loi normale (ou de Gauss) unidimensionnelle

Nous la noterons $\mathcal{N}(\mu,\sigma^2)$, où $\mu\in\mathbb{R}$ et $\sigma\in\mathbb{R}_+$; on distingue 2 cas dans sa définition :
– si $\sigma=0$, c'est la loi de Dirac en μ, $\mathcal{D}(\mu)$ (et il n'y a donc pas absolue continuité),
– si $\sigma\neq 0$, elle admet pour densité l'application définie sur $\mathbb{R}$ par :

$$x\mapsto\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2}).$$

Une caractérisation unifiée (que σ soit nul ou non) peut-être donnée par l'utilisation de la fonction caractéristique, application de $\mathbb{R}$ dans $\mathcal{C}$ définie sur $\mathbb{R}$ par

$$t\ \mapsto\ \mathbb{E}[\mathrm{e}^{itX}]=\exp(it\mu-\frac{\sigma^2t^2}{2})\ ,$$

où $X \sim \mathcal{N}(\mu, \sigma^2)$. On a aussi $\mathbb{E}(X) = \mu$ et $\mathrm{Var}(X) = \sigma^2$.

Un rôle "central" (voir ci-dessous X.3.2) est joué par la **loi normale centrée réduite** $\mathcal{N}(0,1)$, que nous pourrons noter aussi plus simplement $\mathcal{N}$. Il est clair que la loi $\mathcal{N}(\mu, \sigma^2)$ est caractérisée par le fait qu'elle est l'image de $\mathcal{N}$ par l'application de $\mathbb{R}$ dans $\mathbb{R}$, affine croissante : $x \mapsto \mu + \sigma x$. Dans de nombreux ouvrages, la fonction de répartition de $\mathcal{N}$ est notée Φ et sa densité est notée ϕ : $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$; le graphe de ϕ est appelé *courbe en cloche de Gauss*.

La somme de n v.a. indépendantes suivant des lois normales de paramètres $(\mu_1, \sigma_1^2) \ldots (\mu_n, \sigma_n^2)$ suit la loi normale de paramètre $(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$.

Loi normale (ou de Gauss) multidimensionnelle

Soit n un entier strictement positif ; on va considérer ici des probabilités sur $\mathbb{R}^n$, notées $\mathcal{N}(\mu, \Sigma)$, où $\mu \in \mathbb{R}^n$ et Σ est une matrice $n \times n$ symétrique positive, mais non nécessairement définie positive ; on rappelle (voir X.1.3) que cela signifie que, pour tout $u \in \mathbb{R}^n$, $u^t \Sigma u \geq 0$, mais que $u \neq 0$ n'implique pas nécessairement que $u^t \Sigma u > 0$ (la notation ^{t} désigne la transposition des matrices). Une v.a. X à valeurs dans $\mathbb{R}^n$ et suivant une loi de Gauss sera appelée **vecteur gaussien**.

Cas particulier

Soit $\mu = 0_n$ (vecteur nul de $\mathbb{R}^n$) et $\Sigma = I_n$ (matrice identité). Alors $\mathcal{N}(0_n, I_n) = (\mathcal{N}(0,1))^n$, c'est-à-dire (voir X.1.4) la puissance n-ième de la loi normale unidimensionnelle centrée réduite ; autrement dit $\mathcal{N}(0_n, I_n)$ est la la loi d'un n-uple de v.a. réelles indépendantes et de même loi normale d'espérance 0 et variance 1.

$\mathcal{N}(0_n, I_n)$ est donc absolument continue et admet pour densité par rapport à la mesure de Lebesgue n-dimensionnelle l'application

$$(x_1 \ldots, x_n) \mapsto \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp(-\frac{x_i^2}{2}),$$

autrement écrite

$$x \mapsto \frac{1}{(2\pi)^{n/2}} \exp(-\frac{1}{2}(x^t.x)) \ .$$

Cas général

Σ étant symétrique positive, il existe A, matrice $n \times n$ telle que $AA^t = \Sigma$. Alors $\mathcal{N}(\mu, \Sigma)$ est la loi de probabilité image de $\mathcal{N}(0_n, I_n)$ par l'application affine de $\mathbb{R}^n$ dans $\mathbb{R}^n$: $x \mapsto \mu + Ax$.

On remarque que $\mathcal{N}(\mu, \Sigma)$ est concentrée sur l'image de l'application $x \mapsto \mu + Ax$, c'est-à-dire le sous-espace affine passant par μ et parallèle au sous-espace vectoriel $\mathrm{Im}(A)$. Ce sous-espace est de dimension égale au rang de A, qui est aussi le rang de Σ. $\mathcal{N}(\mu, \Sigma)$ n'est donc absolument continue par rapport à la mesure de Lebesgue n-dimensionnelle λ^n que si Σ est de rang n, autrement dit inversible (autrement dit encore ici définie positive). Elle admet alors pour densité l'application, de $\mathbb{R}^n$ dans $\mathbb{R}_+^*$:

$$x \mapsto \frac{1}{(2\pi)^{n/2} \sqrt{\det(\Sigma)}} \exp(-\frac{1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu)) \ ,$$

où $\det(\Sigma)$ désigne le déterminant de la matrice Σ (non nul car celle-ci est ici inversible).

Si un vecteur aléatoire X suit la loi $\mathcal{N}(\mu, \Sigma)$, on a $\mathbb{E}(X) = \mu$ et $\mathrm{Var}(X) = \Sigma$ (on rappelle (voir X.1.3) que $\mathrm{Var}(X)$ désigne la matrice de variances et covariances du vecteur aléatoire X).

Ces notations usuelles présentent une petite incohérence avec le cas unidimensionnel : c'est la matrice notée ci-dessus A qui généralise au cas n-dimensionnel l'écart-type noté par la minuscule σ et c'est la majuscule Σ qui correspond à σ^2.

Une caractérisation unifiée (que Σ soit ou non inversible) peut-être donnée par l'utilisation de la fonction caractéristique définie sur $\mathbb{R}^n$:

$$u \;\mapsto\; \mathbb{E}[\mathrm{e}^{i\,u^t X}] = \exp\left(i\,u^t\mu - \frac{1}{2}\,u^t\Sigma u\right)\,.$$

L'image de $\mathcal{N}(\mu, \Sigma)$ par une application affine, $x \mapsto C + Dx$, est $\mathcal{N}(C + D\mu\ ,\ D\Sigma D^T)$.

Soit X un vecteur aléatoire n-dimensionnel, de loi $\mathcal{N}(\mu, \Sigma)$; décomposons le sous la forme $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, où X_1 est n_1-dimensionnel et X_2 est n_2-dimensionnel (avec $n_1 + n_2 = n$). Soient $M = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ et $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ (avec $\Sigma_{21} = \Sigma_{12}^t$) les décompositions correspondantes de M et de Σ. Alors X_1 suit la loi $\mathcal{N}(\mu_1, \Sigma_{11})$ et, si Σ_{11} est inversible, la loi de X_2 conditionnellement à X_1 (voir X.1.4 pour la notion et pour le calcul) est :

$$x_1 \mapsto \mathcal{N}(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)\ ,\ \Sigma_{22} - \Sigma_{21}.\Sigma_{11}^{-1}.\Sigma_{12})\ .$$

Loi du χ^2 (ou chi-deux ou khi-deux)

Nous noterons $\chi^2(n)$, et appellerons **loi du χ^2 à n degrés de liberté** la loi de la **somme des carrés de n v.a. réelles indépendantes toutes de loi normale centrée réduite.**

Elle est concentrée sur $\mathbb{R}_+$, absolument continue par rapport à la mesure de Lebesgue, et de densité :

$$x \mapsto \frac{1}{2^{n/2}\Gamma(n/2)} e^{-\frac{x}{2}} x^{\frac{n}{2}-1} \mathbf{1}_{\mathbb{R}_+}(x),$$

où Γ est la fonction eulérienne Gamma : $\Gamma(a) = \int_0^{+\infty} e^{-t} t^{a-1} dt$.

Si $X \sim \chi^2(n)$, on a $\mathbb{E}(X) = n$ et $\mathrm{Var}(X) = 2n$.

Il résulte de la définition que, si X_1 et X_2 sont indépendantes, $X_1 \sim \chi^2(n_1)$ et $X_2 \sim \chi^2(n_2)$, alors $X_1 + X_2 \sim \chi^2(n_1 + n_2)$.

Loi du χ^2 décentré

La loi de la **somme des carrés de n v.a. réelles indépendantes $X_1, \ldots, X_n$, la loi de X_i étant $\mathcal{N}(\mu_i, 1)$** (elles sont donc toutes réduites, mais pas de même espérance) ne dépend des paramètres μ_i qu'au travers du paramètre d'excentricité $\delta = \sum_{i=1}^n \mu_i^2$. Nous noterons cette loi $\chi^2(n, \delta)$, et l'appellerons **loi du χ^2 décentré à n degrés de liberté, et de paramètre d'excentricité δ.**

On retrouve $\chi^2(n) = \chi^2(n, 0)$.

La loi $\chi^2(n, \delta)$ est concentrée sur $\mathbb{R}_+$, absolument continue par rapport à la mesure de Lebesgue, mais nous ne donnerons pas l'expression de sa densité qui est trop compliquée pour être utilisable.

Si $X \sim \chi^2(n,\delta)$, on a $\mathbb{E}(X) = n + \delta$ et $\mathrm{Var}(X) = 2(n + 2\delta)$.

À n fixé, les lois $\chi^2(n,\delta)$ forment une famille stochastiquement croissante en δ : on rappelle (voir X.1.2) que cela signifie que, pour tout $x \geq 0$, l'application $\delta \mapsto (\chi^2(n,\delta))([x,+\infty[)$ est croissante : plus δ est grand, plus une v.a. qui suit la loi $\chi^2(n,\delta)$ a tendance à prendre de fortes valeurs.

Il résulte de la définition que, si X_1 et X_2 sont indépendantes, $X_1 \sim \chi^2(n_1,\delta_1)$ et $X_2 \sim \chi^2(n_2,\delta_2)$, alors $X_1 + X_2 \sim \chi^2(n_1 + n_2, \delta_1 + \delta_2)$.

Loi de Student

On appelle **loi de Student à n degrés de liberté** et on note $\mathcal{T}(n)$ la loi du **quotient $\frac{\sqrt{n}X}{\sqrt{Y}}$, où X et Y sont deux variables aléatoires indépendantes, suivant respectivement les lois $\mathcal{N}(0,1)$ et $\chi^2(n)$** (dans cette notation, le $\mathcal{T}$ est traditionnel, et non pas l'initiale $\mathcal{S}$, pour des raisons historiques : Student lui-même notait t, et nous pourrons aussi employer cette notation t dans ce cours).

Elle est absolument continue par rapport à la mesure de Lebesgue, et de densité :

$$x \mapsto \frac{1}{\sqrt{n}B(\frac{1}{2},\frac{n}{2})}(1 + \frac{x^2}{n})^{-(n+1)/2} ,$$

où B est la fonction eulérienne Beta : $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$.

Si $X \sim \mathcal{T}(n)$, on a $\mathbb{E}(X) = 0$ (à condition que $n > 1$) et $\mathrm{Var}(X) = \frac{n}{n-2}$ (à condition que $n > 2$).

Pour $n = 1$, on obtient la **loi de Cauchy** dont l'espérance mathématique n'est pas définie : si X suit une loi de Cauchy, on a $\mathbb{E}(X^+) = \mathbb{E}(X^-) = +\infty$ (les notations X^+ et X^- ont été définies en X.1.3).

Loi de Fisher (ou de Fisher-Snedecor)

On appelle **loi de Fisher à n et m degrés de liberté** et on note $\mathcal{F}(n,m)$ la loi du quotient $\frac{m}{n}\frac{X}{Y}$, où X et Y sont deux variables aléatoires indépendantes, suivant respectivement les lois $\chi^2(n)$ et $\chi^2(m)$.

Elle est concentrée sur $\mathbb{R}_+$, absolument continue par rapport à la mesure de Lebesgue, et de densité :

$$x \mapsto \frac{1}{B(n/2,m/2)} n^{n/2} m^{m/2} \frac{x^{n/2\,-\,1}}{(m+nx)^{(n+m)/2}} \mathbf{1}_{\mathbb{R}_+}(x),$$

où B est la fonction eulérienne Beta : $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$.

Si $X \sim \mathcal{F}(n,m)$, on a $\mathbb{E}(X) = \dfrac{m}{m-2}$ (à condition que $m > 2$) ainsi que $\mathrm{Var}(X) = \dfrac{2m^2(n+m-2)}{n(m-4)(m-2)^2}$ (à condition que $m > 4$).

Loi uniforme

Etant donné un intervalle de longueur non nulle $[a,b]$, on appelle **loi uniforme sur** $[a,b]$, et on note $\mathcal{U}([a,b])$ la probabilité concentrée sur $[a,b]$ et de densité :

$$x \mapsto \frac{1}{b-a}\mathbf{1}_{[a,b]}(x).$$

On note en bref $\mathcal{U}$ pour $\mathcal{U}([0,1])$.

La fonction de répartition de $\mathcal{U}([a,b])$ s'explicite élémentairement : si $a \leq x \leq b$, elle vaut $F_{\mathcal{U}([a,b])}(x) = \dfrac{x-a}{b-a}$.

Si $X \sim \mathcal{U}([a,b])$, on a $\mathbb{E}(X) = \dfrac{a+b}{2}$ et $\mathrm{Var}(X) = \dfrac{(b-a)^2}{12}$.

Loi exponentielle

Etant donné $\theta > 0$, on appelle **loi exponentielle** de paramètre θ, et on note $\mathcal{E}(\theta)$ (ou $\mathrm{Exp}(\theta)$) la probabilité concentrée sur $\mathbb{R}_+$ et de densité :

$$x \mapsto \theta e^{-\theta x}\mathbf{1}_{\mathbb{R}_+}(x).$$

Si $\theta = 1$, on note $\mathcal{E}$ pour $\mathcal{E}(1)$ (ou Exp pour Exp(1)) et on parle de loi exponentielle ("tout court").

La fonction de répartition de $\mathcal{E}(\theta)$ s'explicite élémentairement : si $x \geq 0$, elle vaut

$$F_{\mathcal{E}(\theta)}(x) = 1 - e^{-\theta x}\ .$$

Si $X \sim \mathcal{E}(\theta)$, on a $\mathbb{E}(X) = \frac{1}{\theta}$ et $\mathrm{Var}(X) = \frac{1}{\theta^2}$.

Une propriété remarquable de la loi exponentielle est qu'elle modélise les durées de vie "sans vieillissement" : si la loi de X est $\mathcal{E}(\theta)$, il en est de même, pour tout $t > 0$, de la loi de $X - t$ conditionnellement à l'évènement (de probabilité non nulle) $\{X \geq t\}$.

Loi Gamma

Etant donné $a > 0$ et $\theta > 0$, on appelle **loi Gamma** de paramètre (a,θ), et on note $\mathcal{G}(a,\theta)$ (ou $\gamma(a,\theta)$) la loi concentrée sur $\mathbb{R}_+$ et de densité :

$$x \mapsto \frac{\theta^a}{\Gamma(a)} e^{-\theta x} x^{a-1}\mathbf{1}_{\mathbb{R}_+}(x)\ ,$$

où Γ est la fonction eulérienne Gamma : $\Gamma(a) = \int_0^{+\infty} e^{-t}t^{a-1}dt$.

On retrouve en particulier la loi exponentielle : $\mathcal{G}(1,\theta) = \mathcal{E}(\theta)$.

Si $\theta = 1$, on note $\mathcal{G}(a)$ pour $\mathcal{G}(a,1)$ et on parle de loi Gamma de paramètre (au singulier) a (ou en bref "loi Gamma a").

Si $X \sim \mathcal{G}(a,\theta)$, on $\mathbb{E}(X) = \dfrac{a}{\theta}$ et $\mathrm{Var}(X) = \dfrac{a}{\theta^2}$.

Si les v.a. X_1 et X_2 sont indépendantes et suivent respectivement les lois $\mathcal{G}(a_1,\theta)$ et $\mathcal{G}(a_2,\theta)$, la v.a. $X_1 + X_2$ suit la loi $\mathcal{G}(a_1+a_2,\theta)$; en particulier, si a est un entier n, $\mathcal{G}(n,\theta)$ est la loi de la somme de n v.a. indépendantes toutes de loi exponentielle de paramètre θ.

Il y a un lien entre les lois Gamma et du chi-deux : $\mathcal{G}(\frac{n}{2},\frac{1}{2}) = \chi^2(n)$.

Loi Beta

Etant donné $a > 0$ et $b > 0$, on appelle **loi Beta de seconde espèce** (mais, n'en présentant pas ici d'autre, nous dirons en bref **loi Beta**) de paramètre (a, b), et on note $\mathcal{B}^{(2)}(a, b)$ (ou $\beta(a, b)$) la **loi du quotient $\frac{X}{X+Y}$, où X et Y sont indépendantes et suivent des lois Gamma, respectivement $\mathcal{G}(a)$ et$\mathcal{G}(b)$.**

Elle est concentrée sur $[0, 1]$ et de densité :

$$x \mapsto \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}\mathbf{1}_{[0,1]} ,$$

où B est la fonction eulérienne Beta : $B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$.

Si $a = b = 1$, on retrouve la loi uniforme sur $[0, 1]$: $\mathcal{B}^{(2)}(1, 1) = \mathcal{U}$.

Si $X \sim \mathcal{B}^{(2)}(a, b)$, on a $\mathbb{E}(X) = \dfrac{a}{a+b}$ et $\mathrm{Var}(X) = \dfrac{ab}{(a+b)^2(a+b+1)}$.

Pour mémoire : X et Y étant prises comme dans la définition précédente, la loi de $\frac{X}{Y}$ est dite loi Beta de première espèce de paramètre (a, b).

Loi de Weibull

Etant donné $\alpha > 0$ et $\theta > 0$, on appelle **loi de Weibull** de paramètre de forme α et de paramètre d'échelle θ et on note $\mathcal{W}(\alpha, \theta)$ la probabilité concentrée sur $\mathbb{R}_+$ et de densité :

$$x \mapsto \alpha\theta x^{\alpha-1} \exp(-\theta x^\alpha)\mathbf{1}_{\mathbb{R}_+}(x).$$

Si $\alpha = 1$, on retrouve la loi exponentielle de paramètre θ ; si $X \sim \mathcal{W}(\alpha, \theta)$, X^α suit la loi exponentielle de paramètre θ.

Si $X \sim \mathcal{W}(\alpha, \theta)$, on a $\mathbb{E}(X) = \dfrac{\Gamma(1+\frac{1}{\alpha})}{\theta^{\frac{1}{\alpha}}}$ et $\mathrm{Var}(X) = \dfrac{\Gamma(1+\frac{2}{\alpha}) - \Gamma^2(1+\frac{2}{\alpha})}{\theta^{\frac{2}{\alpha}}}$.

Les lois de Weibull sont utilisées en fiabilité industrielle pour modéliser des lois de durée de vie d'appareils en raison de la propriété suivante : étant fixé $t > 0$, la probabilité que l'appareil vive encore au moins pendant la durée t sachant qu'il a déjà atteint l'âge x est :

- fonction croissante de x si $\theta < 1$,
- fonction constante de x si $\theta = 1$ (cas de la loi exponentielle),
- fonction décroissante de x si $\theta > 1$.

Elles sont donc utilisées pour modéliser des durées de vie d'appareils en phase de jeunesse si $\theta < 1$ (l'élimination des défauts de jeunesse améliore la probabilité de vivre au moins un certain temps), en phase de maturité si $\theta = 1$ (pas de phénomène sensible d'amélioration ni de vieillissement), en phase de vieillissement si $\theta > 1$.

X.3 Théorèmes limites

X.3.1 Convergences

Il existe plusieurs modes de convergence de variables aléatoires : presque-sûre, en probabilité, en loi, dans L^p (quadratique si $p = 2$). Le statisticien a besoin essentiellement de la convergence en loi, parfois de la convergence p.s., et nous ne traiterons pas ici des convergences dans L^p.

Rappelons qu'on dit qu'une suite $(X_n)_{n\in\mathbb{N}}$ de v.a. à valeurs dans $\mathbb{R}^d$ **converge presque sûrement** (p.s.) vers la v.a. Y aussi à valeurs dans $\mathbb{R}^d$, si $\mathbb{P}(\lim_{n\to\infty} X_n = Y) = 1$.

Rappelons qu'on dit qu'une suite $(X_n)_{n\in\mathbb{N}}$ de v.a. à valeurs dans $\mathbb{R}^d$ **converge en loi** (ou **converge faiblement**) vers la v.a. Y, aussi à valeurs dans $\mathbb{R}^d$ si, P_n désignant le loi de X_n et Q celle de Y, on a, pour toute fonction, f, continue bornée de $\mathbb{R}^d$ dans $\mathbb{R}$, $\lim_{n\to+\infty} \mathbb{E}_{P_n}(f) = \mathbb{E}_Q(f)$. En particulier la convergence p.s. implique la convergence en loi.

La convergence en loi ne faisant intervenir que les lois des v.a., on dit aussi (un peu improprement car il ne s'agit pas d'objets de même nature) que la suite $(X_n)_{n\in\mathbb{N}}$ converge en loi vers Q, ou, plus correctement, que **la suite des probabilités** $(P_n)_{n\in\mathbb{N}}$ **converge étroitement** vers Q.

Un résultat souvent utile est le suivant : si X_n tend en loi vers Y et si ϕ est une application continue de $\mathbb{R}^n$ dans $\mathbb{R}^m$, alors $\phi(X_n)$ tend en loi vers $\phi(Y)$; en d'autres termes, si P_n tend étroitement vers Q et si ϕ est une application continue de $\mathbb{R}^n$ dans $\mathbb{R}^m$, alors $\phi(P_n)$ (image de P_n par ϕ, autrement dit loi de ϕ relativement à P_n) tend étroitement vers $\phi(Q)$.

Une caractérisation pratique de cette convergence, pour des **v.a. réelles**, est la suivante, à l'aide de fonctions de répartition : P_n tend étroitement vers Q si et seulement si, pour tout point de continuité x de F_Q, on a $\lim_{n\to+\infty} F_{P_n}(x) = F_Q(x)$.

Le cas de $\mathbb{R}^d$, avec $d > 1$, se ramène à celui de $\mathbb{R}$, par le résultat suivant : P_n tend étroitement vers Q si et seulement si, pour toute application linéaire v de $\mathbb{R}^d$ dans $\mathbb{R}$, on a $v(P_n)$ qui tend étroitement vers $v(Q)$.

Une autre caractérisation pratique de cette convergence, utilise la fonction caractéristique : P_n tend étroitement vers Q si et seulement si, pour $u \in \mathbb{R}^d$, on a la convergence de $\Phi_{P_n}(u) = \mathbb{E}[\mathrm{e}^{iu^tX_n}]$ vers $\Phi_P(u) = \mathbb{E}[\mathrm{e}^{iu^tY}]$.

X.3.2 Théorème central limite

Théorème central limite (ou "de la limite centrale") unidimensionnel

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. réelles indépendantes et de même loi, de second ordre, c'est-à-dire admettant une espérance mathématique μ et une variance finie $\sigma^2 > 0$.

Alors, quand n tend vers $+\infty$, $\frac{\sqrt{n}}{\sigma}(\frac{\sum_{i=1}^n X_i}{n} - \mu)$ tend en loi vers la loi normale centrée réduite $\mathcal{N}(0,1)$.

En d'autres termes, $\sqrt{n}(\frac{\sum_{i=1}^n X_i}{n} - \mu)$ tend en loi, quand n tend vers $+\infty$, vers la loi normale centrée $\mathcal{N}(0,\sigma^2)$.

Conséquence élémentaire (et historiquement première) : **convergence des lois binomiales vers la loi normale** . Si Y_n suit une loi binomiale $\mathcal{B}(n,p)$, $\dfrac{Y_n - np}{\sqrt{np(1-p)}}$ tend en loi, quand n tend vers $+\infty$, vers la loi normale centrée réduite $\mathcal{N}$.

Théorème central limite multidimensionnel

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. à valeurs dans $\mathbb{R}^m$, indépendantes et de même loi, de second ordre. Notons μ leur espérance mathématique et Σ leur matrice de variances et covariances.

Alors, quand n tend vers $+\infty$, $\sqrt{n}(\dfrac{\sum_{i=1}^n X_i}{n} - \mu)$ tend en loi vers la loi normale n-dimensionnelle centrée $\mathcal{N}(0,\Sigma)$ (voir X.2.2).

X.3.3 Convergences en loi remarquables

Convergence des lois de Student

La suite des lois de Student à n degrés de liberté converge étroitement, quand n tend vers l'infini, vers la loi normale centrée réduite.

Convergence des lois binomiales vers la loi de Poisson

Si Y_n suit la loi de Bernouilli $\mathcal{B}(n, p_n)$ et $\lim_{n\to\infty} np_n = \lambda > 0$ alors Y_n tend en loi vers la loi de Poisson $\mathcal{P}(\lambda)$ (voir X.2.1).

Ce théorème fournit une meilleure approximation que le théorème central limite pour les lois binomiales à n "grand" et p "petit".

Convergence des lois du chi-deux vers la loi normale

Si, pour tout n, X_n suit la loi $\chi^2(n)$, la suite des lois des v.a.r. $\sqrt{2X_n} - \sqrt{2n-1}$ converge en loi, quand le nombre de degrés de liberté n tend vers l'infini, vers la loi normale centrée réduite.

Théorème de Glivenko-Cantelli

A toute suite finie $(x_1, \ldots, x_n) \in \mathbb{R}^n$ on associe la **probabilité empirique** $\hat{\mu}_{x_1,\ldots,x_n} = \frac{1}{n}\sum_{i=1}^n \mathcal{D}(x_i)$, où $\mathcal{D}(x_i)$ est la mesure de Dirac en x_i (voir X.1.3 et X.2.1). On note $\hat{F}_{x_1,\ldots,x_1}$ la fonction de répartition empirique de la suite $(x_1, \ldots, x_n)$, c'est-à-dire la fonction de répartition de la probabilité empirique :

$$\forall t \in \mathbb{R} \quad \hat{F}_{x_1,\ldots,x_n}(t) = \frac{1}{n}\mathrm{Card}(\{i : x_i \leq t\}).$$

Soit $(X_n)_{n\in\mathbb{N}}$ une suite infinie de v.a. i.i.d., de fonction de répartition commune F. Alors, pour tout $n \in \mathbb{N}$ et tout $t \in \mathbb{R}$, $\hat{F}_{X_1,\ldots,X_n}(t)$ est une v.a. à valeurs dans $[0, 1]$. Le théorème de Glivenko-Cantelli énonce la convergence vers 0 (en loi, et même presque sûrement) de la suite des v.a. ;$S_n = \sup_{t\in\mathbb{R}} |\hat{F}_{X_1,\ldots,X_n}(t) - F(t)|$.

Chapitre XI

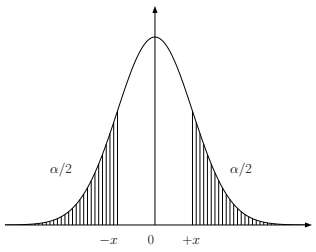
Tables statistiques

Les tables qui suivent ont été générées à l'aide du logiciel Scilab.

XI.1 Quantiles de la loi $\mathcal{N}(0,1)$

Soit X une v.a. de loi $\mathcal{N}(0,1)$, on pose

$$2\int_x^\infty \mathrm{e}^{-y^2/2}\ \frac{dy}{\sqrt{2\pi}} = \mathbb{P}(|X| \geq x) = \alpha.$$



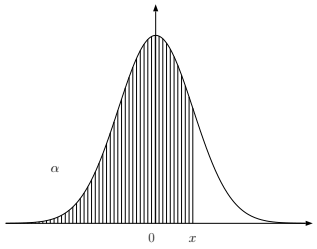
La table donne les valeurs de x en fonction de α. Par exemple $\mathbb{P}(|X| \geq 0.6280) \simeq 0.53$.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	∞	2.5758	2.3263	2.1701	2.0537	1.9600	1.8808	1.8119	1.7507	1.6954
0.1	1.6449	1.5982	1.5548	1.5141	1.4758	1.4395	1.4051	1.3722	1.3408	1.3106
0.2	1.2816	1.2536	1.2265	1.2004	1.1750	1.1503	1.1264	1.1031	1.0803	1.0581
0.3	1.0364	1.0152	0.9945	0.9741	0.9542	0.9346	0.9154	0.8965	0.8779	0.8596
0.4	0.8416	0.8239	0.8064	0.7892	0.7722	0.7554	0.7388	0.7225	0.7063	0.6903
0.5	0.6745	0.6588	0.6433	0.6280	0.6128	0.5978	0.5828	0.5681	0.5534	0.5388
0.6	0.5244	0.5101	0.4959	0.4817	0.4677	0.4538	0.4399	0.4261	0.4125	0.3989
0.7	0.3853	0.3719	0.3585	0.3451	0.3319	0.3186	0.3055	0.2924	0.2793	0.2663
0.8	0.2533	0.2404	0.2275	0.2147	0.2019	0.1891	0.1764	0.1637	0.1510	0.1383
0.9	0.1257	0.1130	0.1004	0.0878	0.0753	0.0627	0.0502	0.0376	0.0251	0.0125

XI.2 Fonction de répartition de la loi $\mathcal{N}(0,1)$

Soit X une v.a. de loi $\mathcal{N}(0,1)$, on pose

$$\int_{-\infty}^{x} \mathrm{e}^{-y^2/2} \ \frac{dy}{\sqrt{2\pi}} = \mathbb{P}(X \leq x) = \alpha.$$



La table donne les valeurs de α en fonction de x. Par exemple $\mathbb{P}(X \leq 1.96) \simeq 0.97500$.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861

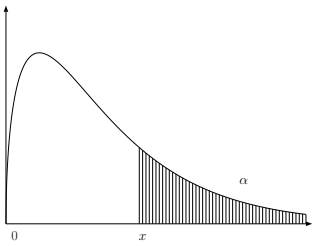
La table suivante donne les valeurs de $1-\alpha$ pour les grandes valeurs de x.

x	2	3	4	5	6	7	8	9	10
$1-\alpha$	2.28e-02	1.35e-03	3.17e-05	2.87e-07	9.87e-10	1.28e-12	6.22e-16	1.13e-19	7.62e-24

XI.3 Quantiles de la loi du χ^2

Soit X_n une v.a. de loi $\chi^2(n)$, on pose

$$\int_x^\infty \frac{1}{2^{n/2}\Gamma(n/2)}\, y^{\frac{n}{2}-1}\, \mathrm{e}^{-y/2}\, dy = \mathbb{P}(X_n \geq x) = \alpha.$$



La table donne les valeurs de x en fonction de n et α. Par exemple $\mathbb{P}(X_8 \geq 20.09) \simeq 0.01$.

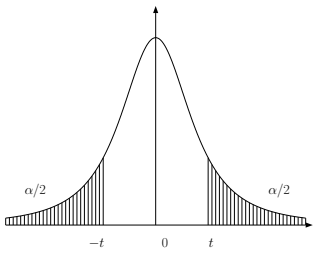
$n\backslash\alpha$	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.001
1	0.0002	0.0010	0.0039	0.0158	2.71	3.84	5.02	6.63	10.83
2	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	13.82
3	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	16.27
4	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	18.47
5	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	20.52
6	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	22.46
7	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	24.32
8	1.65	2.18	2.73	3.49	13.36	15.51	17.53	**20.09**	26.12
9	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	27.88
10	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	29.59
11	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	31.26
12	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	32.91
13	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	34.53
14	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	36.12
15	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	37.70
16	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	39.25
17	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	40.79
18	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	42.31
19	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	43.82
20	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	45.31
21	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	46.80
22	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	48.27
23	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	49.73
24	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	51.18
25	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	52.62
26	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	54.05
27	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	55.48
28	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	56.89
29	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	58.30
30	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	59.70

Lorsque $n > 30$, on peut utiliser l'approximation $\sqrt{2\chi^2(n)} - \sqrt{2n-1} \overset{\text{Loi}}{\simeq} \mathcal{N}(0,1)$.

XI.4 Quantiles de la loi de Student

Soit X_n une variable aléatoire de loi de Student de paramètre n. On pose

$$2\int_t^\infty \frac{\Gamma((n+1)/2)}{\sqrt{\pi n}\,\Gamma(n/2)} \frac{1}{\left(1+\frac{y^2}{n}\right)^{(n+1)/2}}\, dy = \mathbb{P}(|X_n| \geq t) = \alpha.$$



La table donne les valeurs de t en fonction de n et α.
Par exemple $\mathbb{P}(|X_{20}| \geq 2.086) \simeq 0.05$.

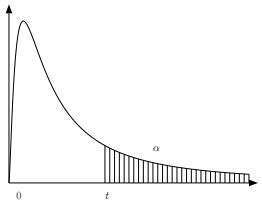
$n\backslash\alpha$	0.900	0.800	0.700	0.600	0.500	0.400	0.300	0.200	0.100	0.050	0.020	0.010	0.001
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
80	0.126	0.254	0.387	0.526	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.416
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

XI.5 Quantiles de la loi de Fisher-Snedecor

Soit $X_{n,m}$ une v.a. de loi de Fisher-Snedecor de paramètre (n, m). On pose

$$\mathbb{P}(X_{n,m} \geq t) = \alpha.$$

La table donne les valeurs de t en fonction de n, m et $\alpha \in \{0.05; 0.01\}$. Par exemple $\mathbb{P}(X_{4,20} \geq 4.43) \simeq 0,01$.



	$n=1$		$n=2$		$n=3$		$n=4$		$n=5$	
m	α =0.05	α =0.01	α =0.05	α =0.01	α =0.05	α =0.01	α =0.05	α =0.01	α =0.05	α =0.01
1	161.45	4052.18	199.50	4999.50	215.71	5403.35	224.58	5624.58	230.16	5763.65
2	18.51	98.50	19.00	99.00	19.16	99.17	19.25	99.25	19.30	99.30
3	10.13	34.12	9.55	30.82	9.28	29.46	9.12	28.71	9.01	28.24
4	7.71	21.20	6.94	18.00	6.59	16.69	6.39	15.98	6.26	15.52
5	6.61	16.26	5.79	13.27	5.41	12.06	5.19	11.39	5.05	10.97
6	5.99	13.75	5.14	10.92	4.76	9.78	4.53	9.15	4.39	8.75
7	5.59	12.25	4.74	9.55	4.35	8.45	4.12	7.85	3.97	7.46
8	5.32	11.26	4.46	8.65	4.07	7.59	3.84	7.01	3.69	6.63
9	5.12	10.56	4.26	8.02	3.86	6.99	3.63	6.42	3.48	6.06
10	4.96	10.04	4.10	7.56	3.71	6.55	3.48	5.99	3.33	5.64
11	4.84	9.65	3.98	7.21	3.59	6.22	3.36	5.67	3.20	5.32
12	4.75	9.33	3.89	6.93	3.49	5.95	3.26	5.41	3.11	5.06
13	4.67	9.07	3.81	6.70	3.41	5.74	3.18	5.21	3.03	4.86
14	4.60	8.86	3.74	6.51	3.34	5.56	3.11	5.04	2.96	4.69
15	4.54	8.68	3.68	6.36	3.29	5.42	3.06	4.89	2.90	4.56
16	4.49	8.53	3.63	6.23	3.24	5.29	3.01	4.77	2.85	4.44
17	4.45	8.40	3.59	6.11	3.20	5.18	2.96	4.67	2.81	4.34
18	4.41	8.29	3.55	6.01	3.16	5.09	2.93	4.58	2.77	4.25
19	4.38	8.18	3.52	5.93	3.13	5.01	2.90	4.50	2.74	4.17
20	4.35	8.10	3.49	5.85	3.10	4.94	2.87	4.43	2.71	4.10
21	4.32	8.02	3.47	5.78	3.07	4.87	2.84	4.37	2.68	4.04
22	4.30	7.95	3.44	5.72	3.05	4.82	2.82	4.31	2.66	3.99
23	4.28	7.88	3.42	5.66	3.03	4.76	2.80	4.26	2.64	3.94
24	4.26	7.82	3.40	5.61	3.01	4.72	2.78	4.22	2.62	3.90
25	4.24	7.77	3.39	5.57	2.99	4.68	2.76	4.18	2.60	3.85
26	4.23	7.72	3.37	5.53	2.98	4.64	2.74	4.14	2.59	3.82
27	4.21	7.68	3.35	5.49	2.96	4.60	2.73	4.11	2.57	3.78
28	4.20	7.64	3.34	5.45	2.95	4.57	2.71	4.07	2.56	3.75
29	4.18	7.60	3.33	5.42	2.93	4.54	2.70	4.04	2.55	3.73
30	4.17	7.56	3.32	5.39	2.92	4.51	2.69	4.02	2.53	3.70
40	4.08	7.31	3.23	5.18	2.84	4.31	2.61	3.83	2.45	3.51
80	3.96	6.96	3.11	4.88	2.72	4.04	2.49	3.56	2.33	3.26
120	3.92	6.85	3.07	4.79	2.68	3.95	2.45	3.48	2.29	3.17
∞	3.84	6.63	3.00	4.61	2.60	3.78	2.37	3.32	2.21	3.02

	$n=6$		$n=8$		$n=12$		$n=24$		$n=\infty$	
m	α =0.05	α =0.01	α =0.05	α =0.01	α =0.05	α =0.01	α =0.05	α =0.01	α =0.05	α =0.01
1	233.99	5858.99	238.88	5981.07	243.91	6106.32	249.05	6234.63	254.31	6365.86
2	19.33	99.33	19.37	99.37	19.41	99.42	19.45	99.46	19.50	99.50
3	8.94	27.91	8.85	27.49	8.74	27.05	8.64	26.60	8.53	26.13
4	6.16	15.21	6.04	14.80	5.91	14.37	5.77	13.93	5.63	13.46
5	4.95	10.67	4.82	10.29	4.68	9.89	4.53	9.47	4.36	9.02
6	4.28	8.47	4.15	8.10	4.00	7.72	3.84	7.31	3.67	6.88
7	3.87	7.19	3.73	6.84	3.57	6.47	3.41	6.07	3.23	5.65
8	3.58	6.37	3.44	6.03	3.28	5.67	3.12	5.28	2.93	4.86
9	3.37	5.80	3.23	5.47	3.07	5.11	2.90	4.73	2.71	4.31
10	3.22	5.39	3.07	5.06	2.91	4.71	2.74	4.33	2.54	3.91
11	3.09	5.07	2.95	4.74	2.79	4.40	2.61	4.02	2.40	3.60
12	3.00	4.82	2.85	4.50	2.69	4.16	2.51	3.78	2.30	3.36
13	2.92	4.62	2.77	4.30	2.60	3.96	2.42	3.59	2.21	3.17
14	2.85	4.46	2.70	4.14	2.53	3.80	2.35	3.43	2.13	3.00
15	2.79	4.32	2.64	4.00	2.48	3.67	2.29	3.29	2.07	2.87
16	2.74	4.20	2.59	3.89	2.42	3.55	2.24	3.18	2.01	2.75
17	2.70	4.10	2.55	3.79	2.38	3.46	2.19	3.08	1.96	2.65
18	2.66	4.01	2.51	3.71	2.34	3.37	2.15	3.00	1.92	2.57
19	2.63	3.94	2.48	3.63	2.31	3.30	2.11	2.92	1.88	2.49
20	2.60	3.87	2.45	3.56	2.28	3.23	2.08	2.86	1.84	2.42
21	2.57	3.81	2.42	3.51	2.25	3.17	2.05	2.80	1.81	2.36
22	2.55	3.76	2.40	3.45	2.23	3.12	2.03	2.75	1.78	2.31
23	2.53	3.71	2.37	3.41	2.20	3.07	2.01	2.70	1.76	2.26
24	2.51	3.67	2.36	3.36	2.18	3.03	1.98	2.66	1.73	2.21
25	2.49	3.63	2.34	3.32	2.16	2.99	1.96	2.62	1.71	2.17
26	2.47	3.59	2.32	3.29	2.15	2.96	1.95	2.58	1.69	2.13
27	2.46	3.56	2.31	3.26	2.13	2.93	1.93	2.55	1.67	2.10
28	2.45	3.53	2.29	3.23	2.12	2.90	1.91	2.52	1.65	2.06
29	2.43	3.50	2.28	3.20	2.10	2.87	1.90	2.49	1.64	2.03
30	2.42	3.47	2.27	3.17	2.09	2.84	1.89	2.47	1.62	2.01
40	2.34	3.29	2.18	2.99	2.00	2.66	1.79	2.29	1.51	1.80
80	2.21	3.04	2.06	2.74	1.88	2.42	1.65	2.03	1.32	1.49
120	2.18	2.96	2.02	2.66	1.83	2.34	1.61	1.95	1.25	1.38
∞	2.10	2.80	1.94	2.51	1.75	2.18	1.52	1.79	1.00	1.00

Bibliographie et logiciels

Quelques éléments de bibliographie commentée

1. Livres classiques de référence.

Dacunha-Castelle D., Duflo M. *Probabilités et Statistique*, Tomes I et II, Masson (2ème édition) 1994.

Dacunha-Castelle D., Duflo M. *Exercices de Probabilités et Statistique*, Tomes I et II, Masson (2ème édition) 1994.

Ouvrages riches couvrant un large champ des probabilités et de la statistique mathématique. Rédaction assez condensée, de nombreuses propriétés étant introduites à partir des exercices. Niveau mathématique élevé (Maîtrise "forte" ou 3ème cycle).

Saporta G., *Probabilités, Statistique et Analyse des Données*, 2ème édition, Technip, 1990.

Ouvrage conçu à partir d'un enseignement du Conservatoire National des Arts et Métiers. Très bonne lisibilité à partir de connaissances mathématiques de niveau premier cycle. Confronte plusieurs approches des données.

Monfort A. *Cours de Statistique Mathématique*, Economica, 1988.

Ouvrage classique orienté vers les propriétés mathématiques sous-jacentes aux modèles de la statistique.

Lebart L., Morineau A., Piron M. A. *Statistique exploratoire multidimensionnelle* (2ème édition), Dunod, 1997.

Ouvrage de référence en Analyse des Données multidimensionnelles, dont il couvre les principales branches. Très bonne lisibilité à partir de connaissances mathématiques de niveau premier cycle.

2. Livres orientés vers des modèles ou des domaines d'application.

Bernier J., Parent E., Boreux J.J. *Statistiques pour l'environnement. Traitement bayésien des incertitudes*, Editions Tec et Doc, Lavoisier, 2000.

Accès à la statistique bayésienne, à partir d'un domaine d'application où elle est particulièrement pratiquée.

Bouyer J., *Epidémiologie : principes et méthodes quantitatives*, Editions INSERM (1993).

Très clair pour débuter sur des méthodes usitées en statistique médicale.

Conover W.J. *Practical nonparametric statistics* (2nd edition), J. Wiley and sons, 1980.

Présentation complète des bases de la statistique non paramétrique ; exercices corrigés ; nombreuses tables et références.

Gourieroux C., *Econométrie des variables qualitatives*, Economica (1989).

Ouvrage très théorique, mais accompagné de nombreux exemples.

Robert C.P. *L'analyse statistique bayésienne*, Economica,1992.

Ouvrage très complet sur les méthodes bayésiennes. Existe aussi en version anglaise.

Deux logiciels

Les logiciels indiqués ici sont d'un accès gratuit et conçus dans un but pédagogique. Leur pratique peut fournir un complément utile aux travaux pratiques qui sont proposés dans le cadre de ce cours à l'aide du logiciel SCILAB.

SMEL (Statistique Médicale En Ligne)

Il s'agit d'un outil français d'apprentissage à la statistique à la fois par simulation interactive et par emploi de données réelles. Il est structuré en 4 chapitres : Probabilités, Statistique descriptive, Tests, Estimation.

Accessible sur : http ://www. math-info.univ-paris5.fr/smel

R

Logiciel gratuit proche du logiciel S, qui est l'un des logiciels les plus couramment utilisés par les statisticiens professionnels. Les lignes suivantes sont extraites de sa présentation dans une lettre intitulée **R FAQ** (Frequently Asked Questions about R), qui est disponible et régulièrement mise à jour sur http ://www.ci.tuwien.ac.at/-hornik/R/ :

R *is a system for statistical computation and graphics. It consists of a language plus a run-time environment with graphics, a debugger, acccess to certain system functions, and the ability to run programs stored in script files .. it is very similar in appearance to* S. R *has a home page at* http///www.r-project.org .

Voici quelques autres références utiles sur R :

http ://www.univ-st-etienne.fr/infsci/linfo/l8/tp1/tp1.html (introduction en français)

http ://www.agr.kuleuven.ac.be/vakken/StatisticsByR/ (présentation et références en anglais)

http ://www.ma.hw.ac.uk/ stan/R/ (présentation et références en anglais)

http ://rweb.stat.umn.edu/Rweb/ (interface web de R)

http ://www.stat.Berkeley.EDU/users/terry/zarray/Html/Rintro.html (applications aux microarrays)

Index

a posteriori, 30
a priori, 30
absolue continuité, 226
AIC, 163
BIC, 163
Akaïke (Critère d'), 163
algorithme d'apprentissage, 174
algorithme des k-plus proches voisins, 178
algorithme par noyau, 178, 179
algorithme par partition, 178, 179
ACP, 201
analyse de la variance, 35, 61
apprentissage, 173
AR (Processus), 149
    canonique, 150
$AR(\infty)$ (Ecriture)
    d'un processus MA, 154
$AR(\infty)$ (Ecriture)
    d'un processus ARMA, 158
ARMA (Processus), 157
    canonique, 157
    causal, 157
    Densité spectrale d'un, 159
    inversible, 157
    minimal, 157
Autocorrélation (Matrice d'), 140
Autocorrélogramme
    partiel, 140
    simple, 139
Autocovariance (Fonction d'), 139

biais, 14, 182
biais-variance, 182
Blancheur (Test de), 142
borélien, 227
Box-Cox (Méthode de), 164
Bruit blanc
    faible, 137
    fort, 137

capacité, 181
caractère prédictif, 36
classement, 174
classement binaire, 174, 176
classification, 174, 211
classification ascendante hiérarchique, 214
classification binaire, 176
coefficient de détermination, 56
Coin (Méthode du), 158
complexité, 181
composantes principales, 207
consistance, 177
contribution à l'inertie, 210
corrélation, 235
covariance, 235

décentrage, 50
Décomposition saisonnière, 145
    à l'aide de la régression linéaire, 145
densité, 227, 230
Densité spectrale, 143
densité (calcul de), 230, 232, 241
densité multi-dimensionnelle, 231
distance du χ^2, 65
Durbin-Levinson (Algorithme de), 141

échantillon, 12
échantillons appariés, 93
ensemble d'apprentissage, 174
erreur d'approximation, 182
erreur d'estimation, 182
erreur de généralisation, 174
erreur ¡¡laisser-un-de-côté¿¿, 179
erreur de $1^{\text{ère}}$ espèce, 24
erreur de $2^{\text{ème}}$ espèce, 24
espérance, 233, 235
    conditionnelle, 238

empirique, 236
espace mesurable, 223, 225
estimateur, 14
amélioré, 18
asymptotiquement normal, 15
consistant, 15
convergent, 15
des moindres carrés, 54
du maximum de vraisemblance, 17
préférable, 14
sans biais, 14
estimateur de Nadaraya-Watson, 178
estimation, 11, 13
évènements, 223
indépendants, 239, 240
ex-aequo, 96

facteur, 33
aléatoire, 33
contrôlé, 33
modalités, 33
qualitatif, 33
quantitatif, 33
fonction cible, 174
fonction de prédiction, 174
fonction oracle, 174
fonction de répartition, 228
empirique, 90
fractile, 20

homoscédasticité, 34
hypothèse
alternative, 24
bilatérale, 42
nulle, 24
unilatérale, 42

i.i.d., 9, 12
inégalité de Jensen, 236
inertie, 202
Innovation, 149
intervalle de confiance, 11, 19

Jensen, 236

log-vraisemblance, 17
loi, 224
beta, 249
binomiale, 242, 251
de Bernoulli, 242
de Dirac, 242
de Fisher, 247
de Fisher décentrée, 50
de Pareto généralisée, 114
de Poisson, 126, 244, 251
de Student, 247, 251
de Weibull, 249
du χ^2, 246, 251
du χ^2 décentrée, 50, 246
exponentielle, 248
Gamma, 248
gaussienne, 244, 245
GEV, 111
hypergéométrique, 243
multinomiale, 243
normale, 244, 245
uniforme, 248
loi conditionnelle, 241
loi forte des grands nombres, 13
loi image, 224
loi marginale, 225
loi produit, 12, 225

MA (Processus), 153
MA (Processus)
canonique, 154
$MA(\infty)$ (Ecriture)
d'un processus AR, 150
d'un processus ARMA, 157
machines à vecteurs supports, 183
matrice de covariance, 235
mesure de Lebesgue, 230, 231
mesure de probabilité, 223
mesure image, 230
minimisation du risque empirique, 181
modèle statistique, 6
modèle dominé, 12
modèle paramétrique, 12
moyenne empirique, 13

niveau d'un test, 24, 27
noyau, 183

ordre stochastique, 20, 100, 228

Périodogramme, 144
paramètre, 6
plans d'expérience, 33
Portmanteau (Test de), 142
prédiction, 52
principe de Neyman, 24
probabilité conditionnelle, 236
probabilité empirique, 236
probabilité puissance, 240
Processus
    faiblement stationnaire, 138
    fortement stationnaire, 138
    non stationnaire DS, 139
    non stationnaire TS, 139
p-valeur, 27, 46

qualité de projection, 210
quantile, 20

R^2, 56
régresseurs, 52
régression, 52, 62
    droite de, 37
    linéaire simple, 37, 53, 54, 58
    logistique, 71
régression (aux moindres carrés), 174, 176
rang, 88
région critique, 24
réseaux de neurones, 182
risque, 174

Série temporelle, 131, 137
    à temps continu, 137
    à temps discret, 137
    multivariée, 137
    univariée, 137
Schwarz (Critère de), 163
SSE, 49
SSM, 49
statistique, 13
    d'ordre, 89
    de test, 27
    exhaustive, 18
    libre, 88
support vector machine, 183
surapprentissage, 181

table d'analyse de la variance, 51
TCL, 13
test, 11, 24
    convergent, 27
    d'homogénéité des moyennes, 47
    des signes et rangs, 94
    de comparaison, 75, 79
    de Kolmogorov, 90
    de Kolmogorov-Smirnov, 97
    de Mann et Whitney, 98
    de Wilcoxon, 94
    du χ^2, 65
    du χ^2 adaptatif, 67
    du χ^2 d'indépendance, 67
    niveau, 24, 27
    puissance, 24
théorème
    central limite, 13
    de Cochran, 37
    de Glivenko-Cantelli, 251
tribu, 227
tribu borélienne, 229

validation croisée, 179
variable
    à expliquer, 52
    aléatoire, 224
    dépendante, 52
    discrète, 229
    endogène, 52
    exogène, 52
    explicative, 52
    indépendante, 52, 239, 240
variance, 234, 235
    asymptotique, 16
    empirique, 236
variation
    interclasses, 49
    intraclasses, 49
    totale, 49
vecteur aléatoire, 229
vraisemblance, 12

Yule Walker (Equations de), 151

zone de rejet, 24