Initiation aux statistiques

Notions de base d'analyse statistique

Paul FRAUX

Version du 5 février 2025

Avant-Propos

Ce polycopié présente les fondamentaux des outils statistiques, utile dans tout cursus ingénieurs. Il approfondi le cours, et parfois va plus loin (parties marquées d'un H.P.).
Ce cours vise à faire de l'analyse chiffrée pour décrire un phénomène qui dépend du hasard, comprendre et modéliser des phénomènes complexes, aider à la prise de déscision et faire des prévisions.
Pour cela, les objectifs de ce cours sont :

- savoir calculer les statistiques usuelles,
- connaître les principaux concepts de l'estimation,
- savoir utiliser l'estimateur de maximum de vraisemblance pour un modèle régulier,
- savoir parler des paramètres des tests (risque, puissance),
- savoir mettre en oeuvre les test courants,
- savoir utiliser les regressions multilinéaires, et utiliser des statistiques pour valider ce modèle,
- se sensibiliser à la notion de modélisation parcimonieuse.

On commencera par rappeler les outils de Probabilités nécessaire, avant de faire une introduction aux statistiques usuelles. Puis nous présenterons les concepts de l'estimation en insistant sur l'estimateur de maximum de vraisemblance. Nous verrons ensuite les notions de test, avant de revenir sur les modèles gaussion avec les regressions multilinéaires et les tests associés. Nous finirons par une brève initiations au méthodes de recherche de modèles parcimonieux.
Ce polycopié est nouveau, et contient presque sûrement des erreurs, incohérence et fautes d'orthographes. De plus, certaines notations sont encore implicites et mériterai d'être rappellé en début de polycopié. Merci de nous les signaler pour contribuer à son amélioration.

Notations

Ensembles classiques :

- Pour n un entier naturel, on note $[\![1, n]\!] := \{1, \cdots n\}$ l'ensemble des entiers compris entre 1 et n.
- $\mathbb{N}$ est l'ensemble des entiers naturel et $\mathbb{N}^*$ l'ensemble des entiers supérieurs à 1.
- $\mathbb{R}$ est l'ensemble des nombres réels.
- $\mathcal{P}(E)$ est l'ensemble contenant tous les sous-ensembles d'un ensemble E.
- E^n est l'ensemble des n tuples d'un ensemble E et $E^{\mathbb{N}}$ l'ensemble des suites à valeurs dans E.
- E^F est l'ensemble des fonctions de E dans F.

Fonctions utiles :

- $C(E, F)$ est l'ensemble des fonctions continues, entre deux ensembles topologiques, de E vers F. En particulier $C(\mathbb{R}, \mathbb{R})$ est l'ensemble des fonctions réelles continues.
- $C^k(\mathbb{R}, E)$ est l'ensemble des fonctions k fois dérivable, de $k^{\text{ème}}$ dérivée continues à valeur dans E un espace vectoriel normé complet.
- Pour A un sous-ensemble d'un ensemble E, on note $\mathbb{1}_A$ la fonction

$$\mathbb{1}_A : \begin{cases} E & \to & \mathbb{R} & \\ x & \mapsto & 1 & \text{si } x \in A, \\ & & 0 & \text{si } x \notin A. \end{cases}$$

Concepts probabilistes :

- Ω est un univers, $\mathcal{T}$ l'ensemble des évènements et $\mathbb{P}$ une mesure de probabilité.
- A^c ou $\overline{A}$ est l'événement complémentaire de A, *i.e.* celui où A ne se produit pas.
- $\mathbb{P}_A$ ou $\mathbb{P}(\cdot|A)$ est la probabilité conditionnellement à l'événement A.
- X, Y, Z sont des variables ou des vecteurs aléatoires réels.
- $\mathbb{E}(X)$ est l'espérance de X une variable aléatoire réel, $\text{Var}(X)$ sa variance et σ son écart-type.
- Σ est sa matrice de covariance et ρ la matrice de corrélation d'un vecteur aléatoire réel.
- F_X ou F est la fonction de répartition de X, f_x ou f est sa densité s'il en a une.
- M_X est la fonction génératrice des moments de X.
- φ_X est la fonction caractéristique de X.
- iid est l'abréviation pour indépendante et identiquement distribuée. Se dit pour une suite de variables aléatoire indépendante dans leur ensemble et de même loi de probabilité.

Concepts statistiques :

- $(\Omega, \mathcal{T}, (\mathbb{P})_{\mathbb{P}\in\mathcal{P}}, \Theta, \mathcal{X}, \mathcal{G}, X)$ est une expérience statistique.
- $\overline{x}$ la moyenne empirique.
- S_y^2 la variance empirique d'un échantillon
- $Cov_{x,\hat{y}}$ la covariance empirique de deux échantillons.
- Plus généralement, x, y, z (en minuscule) sont des réalisations des variables aléatoires X, Y ou Z.
- T_n est une statistique
- Le biais d'un estimateur de $f(\theta)$ est la quantité $\mathbb{E}_\theta(T_n) - f(\theta)$
- $I(\theta)$ est l'information de Fischer au point θ. Sc est la fonction score associé.

Algèbre linéaire :

- $M_{n,p}$ est l'ensemble des matrices de n lignes et p colonnes.
- S_n est l'ensemble des matrices symétriques, S_n^+ des matrices symétriques semi-définies positives et S_n^{++} des matrices symétriques définies positives.
- T_n est l'ensemble des matrices triangulaires inférieures.
- $A^t \in M_{p,n}$ est la matrice transposée de $A \in M_{n,p}$.

Table des matières

1 Premières statistiques **8**

1.1 Les expériences statistiques et les statistiques . . . 9
1.2 Moyenne et variance empirique . . . 10
1.3 L'exemple fonctionnel fondamental : la fonction de répartition . . . 14
1.4 Exercices . . . 18

2 Quelques lois usuelles en statistiques **20**

2.1 Loi normale . . . 20
2.2 Loi du χ^2 . . . 22
2.3 Loi de Student . . . 24
2.4 Loi de Fisher-Snedecor . . . 26
2.5 Loi de Kolmogorov . . . 28
2.6 Exercices . . . 28

3 Estimateur du maximum de vraisemblance **29**

3.1 Estimateur et premiers critères d'évaluations . . . 30
3.1.1 Comment estimer une performance . . . 30
3.1.2 Biais d'un estimateur . . . 31
3.1.3 Convergence d'une suite d'estimateurs . . . 32
3.2 Modèles réguliers . . . 33
3.2.1 Régularité d'une expérience . . . 33
3.2.2 Efficacité d'un estimateur . . . 35
3.3 Estimation du maximum de vraisemblance . . . 38
3.3.1 Définition et méthode de calcul pour les modèles réguliers . . . 38
3.3.2 Propriétés asymptotiques . . . 39
3.4 Amélioration d'estimateur (H.P.) . . . 41
3.5 Exercices . . . 45

4 Tests **47**

4.1 La problématique des tests . . . 47
4.1.1 Hypothèses statistiques . . . 47
4.1.2 Test statistique . . . 48
4.1.415 Erreurs et incertitude . . . 49
4.1.5 Courbe ROC et évaluation AUC de tests . . . 51
4.2 Tests paramétriques . . . 53
4.2.1 Test paramétrique avec deux hypothèses simples (Neyman-Pearson) . . . 53

4.2.2 Test paramétrique avec hypothèse composite . . . 55
4.2.3 Un autre test composite : le test de proportion . . . 58
4.3 Exercices . . . 60

5 Les modèles gaussiens 61
5.1 Rappels sur les vecteurs gaussiens . . . 61
5.2 Normes de vecteurs gaussiens et théorème de Cochran (H.P.) . . . 63
5.3 Test d'égalités de paramètres dans un modèle gaussien . . . 66
5.3.1 Test d'égalité des variances . . . 66
5.3.2 Test d'égalité des moyennes . . . 67
5.4 Estimation par régressions linéaires avec plusieurs variables, et introductions au test d'hypothèses linéaires . . . 67
5.4.1 Problème des moindres carrés . . . 67
5.4.2 Validation de la régression . . . 69
5.4.3 Test d'hypothèse linéaire avec bruit gaussien . . . 71
5.5 Exercices . . . 72

6 Tests de χ^2 et test de Kolmogorov-Smirnov 75
6.1 Test du χ^2 et tests d'indépendance . . . 75
6.1.1 Ajustement à une loi connue . . . 75
6.1.2 Comment choisir les classes $C_1, ..., C_l$? . . . 78
6.1.3 Que faire si les paramètres définissant F_0 ne sont pas connus ? . . . 79
6.1.4 Test d'égalité de loi (d'homogénéité) et d'indépendance de variable qualitative . . . 79
6.2 Test d'adéquation à une loi de Kolmogorov . . . 81
6.3 Test d'égalité de loi de Kolmogorov-Smirnov . . . 82
6.4 Exercices . . . 84

7 A propos de la parcimonie 85

8 Solutions et pistes de corrections des exercices 90
8.1 Premières statistiques . . . 90
8.2 Quelques lois usuelles en statistiques . . . 91
8.3 Estimateur du maximum de vraisemblance . . . 91
8.4 Tests . . . 96
8.5 Les modèles gaussiens . . . 98
8.6 Tests de χ^2 et test de Kolmogorov-Smirnov . . . 104

A Inverse généralisé de fonction de répartition 106

B Propriétés probabilistes supplémentaires (H.P.) 108
B.1 Inégalité de Jensen pour les fonctions convexes . . . 108
B.2 Inégalité d'Hölder . . . 109

C Compléments d'algèbre linéaire 111
C.1 Factorisation de Choleski . . . 111

D Espérance conditionnelle de variables aléatoires (H.P.) 113

Bibliographie **116**

Chapitre 1

Premières statistiques

Pourquoi s'embêter à apprendre les statistiques ? À quoi peuvent-elles me servir ? Nous allons prendre un exemple, un jeu de pile ou face, et voir le type de questions qui trouvent des réponses en statistiques.
L'on s'apprête à jouer à pile ou face avec une pièce de monnaie. Nous soupçonnons que la pièce n'est pas parfaitement équilibrée, c'est-à-dire que la probabilité p de tomber sur face n'est pas $\frac{1}{2}$.
La première question que l'on pourrait se poser est la suivante : quelle est cette valeur p, et peut-on l'*estimer* ? On cherche alors à prédire cette valeur pour modifier notre stratégie en conséquent, à l'aide d'un certain nombre de réalisations $(x_1, ..., x_n)$ de l'expérience "tirer une pièce". La stratégie de résolution de cette question consiste à construire, à partir des données (*via* une fonction), une "estimation" $\hat{\theta}$ de p. Bien sûr, cette estimation ne peut pas dépendre de l'estimande p, mais l'on cherchera à s'assurer que $\hat{\theta}$ soit probablement proche de p. Un tel problème est dit "*estimation ponctuelle*", et consiste à donner une valeur précise.
Savoir qu'un estimateur est probablement proche de la vraie valeur est parfois d'intérêt limité. On pourra préférer donner une région de confiance, par exemple un intervalle. Pour la pièce, cela consisterait à donner un bon encadrement de la probabilité de tirer face. Un intervalle de confiance est la construction de deux fonctions des données $\overline{\theta_n}$ et $\underline{\theta_n}$ telles que l'estimande appartiennent à l'intervalle aléatoire $[\underline{\theta_n}(x_1, ..., x_n), \overline{\theta_n}(x_1, ..., x_n)] = [\underline{\theta_n}, \overline{\theta_n}]$ avec forte probabilité. La question devient alors de trouver un équilibre entre la taille de l'intervalle (région) que l'on veut petit(e) pour plus de précisions, et la probabilité que l'estimande appartienne à la région, que l'on souhaite grande. Répondre à ce type de question correspond à trouver une "*région ou intervalle de confiance*".
Un dernier problème que l'on va chercher à résoudre est celui de *tester* une hypothèse. Par exemple, comment trouver une règle de décision pour refuser l'hypothèse que votre adversaire triche en sa faveur ($p < \frac{1}{2}$) ? Il s'agit de donner une règle de décision qui permette de minimiser le risque de rejeter à tort l'hypothèse, sans pour autant toujours l'accepter. Cette partie des statistiques est ce que l'on appelle les *tests*.

Ces questions, différentes en essence, ont des réponses intimement liées, comme nous allons pouvoir le voir dans ce cours.

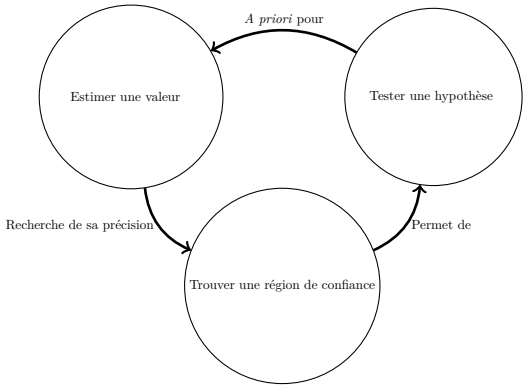


FIGURE 1.1 – Trois questions, intimement reliées

Certaines questions de statistique ne seront pas traitées dans ce document, comme la question de la *mise à jour de la connaissance et de la prise de décision* avec des méthodes Bayesiennes, et Minimax. Le lecteur intéressé pourra se laisser tenter par les lectures suivantes :

- [Rob01] pour une introduction aux méthodes bayésiennes ;
- [Dud89] Pour les bases théoriques à la constructions des espérances conditionnelles, nécessaire à la méthode bayesienne ;
- [GvdV17] pour un développement des méthodes Bayésienne pour des statistiques non paramétriques, et une étude de leurs performances.

1.1 Les expériences statistiques et les statistiques

Cette section est là pour donner un cadre formel dans lequel nous discuterons par la suite, et peut être omise en première lecture.
Partons de l'exemple précédent. Nous disposons d'un espace probabilisable $(\Omega, \mathcal{T})$, c'est-à-dire un univers (les résultats possibles des n lancer de pièces, donc $\Omega = \{0,1\}^n$) et d'une tribu (l'ensemble de tous les événements descriptibles, donc ici $\mathcal{T} = \mathcal{P}(\Omega)$).
Sur cette espace, l'on considère une *famille de loi* $\mathcal{P}$, où chaque loi est susceptible de régir le phénomène (ici toutes les lois de variables aléatoires issues de la répétition indépendante d'une expérience de Bernoulli).
On peut alors choisir pour cette famille de loi une *paramétrisation*, qui joue le rôle d'un système de coordonné des lois. C'est-à-dire qu'il s'agit avant tout d'un choix de convenance, qui aide à la description. Une *paramétrisation* est la donnée d'un ensemble Θ, que l'on demande souvent inclus dans $\mathbb{R}^n$, et d'une application surjective de Θ dans la famille des lois $\mathcal{P}$. On dira alors que la paramétrisation est *identifiable* si cette application est injective (avec le lancer de pièce, une paramétrisation identifiable est la probabilité de tirer face avec la pièce, et $\Theta = [0,1]$). Cette hypothèse d'identifiabilité est surtout une hypothèse de convenance, et n'est absolument pas nécessaire, comme l'on pourrait le voir en s'intéressant à la théorie du mélange.
Maintenant, il est naturel que le statisticien n'ai pas accès à tout l'univers, mais seulement à une partie des données. C'est ce qui motive la définition suivante, inspiré des variables aléatoires :

Définition 1 :

On appelle expérience statistique *la donnée de*

$$(\Omega, \mathcal{T}, \Theta, (\mathbb{P}_\theta)_{\theta\in\Theta}, \mathcal{X}, \mathcal{G}, X)$$

où $(\Omega, \mathcal{T})$ est un espace probabilisable, $\mathcal{P}$ est une famille de probabilité sur Ω, Θ en est une paramétrisation, $\mathcal{X}$ un espace d'observation et $\mathcal{G}$ une tribu sur cet espace. Enfin, X est une variable aléatoire, c'est-à-dire une application mesurable de $(\Omega, \mathcal{T})$ dans $(\mathcal{X}, \mathcal{G})$.
Si les expériences sont construites à partir de répétition indépendante d'une expérience de base, on parlera d'expérience produit*. Dans ce cas, l'univers doit être une puissance de l'espace de base (Ω^n), et les tribus, lois, et observation doivent être des tribus et lois produits ($\mathcal{T} = \sigma(\times_{i=1}^n \mathcal{T}_0)$, $\mathcal{P} = (\mathbb{P}_n = \mathbb{P}^{\otimes n})$, $\mathcal{X} = {\mathcal{X}_0}^n$).*

Remarque : Bien souvent, on se contentera de donner la famille de loi $\mathcal{P}$, le reste étant implicite.

Notation : On notera les concepts probabilistes d'intérêt sous la loi $\mathbb{P}_\theta$ avec un θ en indice. Par exemple $\mathbb{E}_\theta$ est l'espérance et V_θ la variance suivant $\mathbb{P}_\theta$.

Le statisticien cherche bien souvent à extraire un maximum d'information de ses données, c'est pourquoi il se permet un panel large d'outils :

Définition 2 :

On appelle Statistique *toute application mesurable depuis l'espace des observations*

Exemple : La moyenne empirique, la variance empirique et la fonction de répartition empirique sont trois exemples de statistiques.
Notation : Dans toute la suite de ce document, on fixe $(\Omega, \mathcal{T}, \Theta, (\mathbb{P}_\theta)_{\theta\in\Theta}, \mathcal{X}, \mathcal{G}, X)$ une expérience statistique.

1.2 Moyenne et variance empirique

La première statistique que l'on voit le plus souvent est celle de la moyenne empirique :

Définition 3 :

Soit $(X_i)_{i\in[\![1,n]\!]}$ des variables aléatoires. On appelle moyenne empirique la variable aléatoire

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

On la note aussi parfois $\overline{X}_n$. On note $\overline{x}$ une réalisation de la moyenne empirique.

On l'utilisera dans le cadre de variables aléatoire indépendantes et identiquement distribuées. Elle est naturelle : il s'agit de remplacer une intégrale par sa version discrète, en espérant approcher suffisamment la quantité finale.

Proposition 4 :

Soit $(X_i)_{i\in[\![1,n]\!]}$ des variables aléatoires indépendantes et identiquement distribuées ayant un moment d'ordre 2. On note $m = \mathbb{E}(X_1)$ leur moyenne et $\sigma^2 = V(X_1)$ leur variance commune.
Alors :

$$\mathbb{E}(\overline{X}) = m, \qquad V(\overline{X}) = \frac{\sigma^2}{n},$$

$$\overline{X} \xrightarrow{L^2} m, \qquad \sqrt{n}\frac{\overline{X} - m}{\sigma} \xrightarrow{Loi} \mathcal{N}(0,1)$$

Démonstration.
Les deux premiers points sont des conséquences immédiates des propriétés de l'espérance et de la variance. Les deux derniers points sont les conséquences de la loi des grands nombres et du théorème central limite, dont les hypothèses sont clairement vérifiées. □

Mais une fois que l'on a une idée de la moyenne de la loi que l'on cherche à apprendre, une bonne idée est de regarder l'étalement de cette loi autour de cette moyenne. Pour cela, nous utiliserons la variance empirique :

Définition 5 :

Soit $(X_i)_{i\in[\![1,n]\!]}$ des variables aléatoires, on appelle variance empirique des (X_i) la quantité

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

Proposition 6 :

Soit $(X_i)_{i\in[\![1,n]\!]}$ des variables aléatoires indépendantes et identiquement distribuées ayant un moment d'ordre 2, alors

$$\mathbb{E}(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2) = \frac{n-1}{n}\sigma^2.$$

Démonstration.
Commençons par remarquer que pour $m \in \mathbb{R}$,

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

En effet,

$$\begin{aligned}
\sum_{i=1}^{n}(X_i-\overline{X})^2 &= \sum_{i=1}^{n}\left(X_i-m-(\overline{X}-m)\right)^2,\\
&= \sum_{i=1}^{n}(X_i-m)^2+(\overline{X}-m)^2-2(\overline{X}-m)(X_i-m),\\
&= \sum_{i=1}^{n}(X_i-m)^2+\sum_{i=1}^{n}(\overline{X}-m)^2-\sum_{i=1}^{n}2(\overline{X}-m)(X_i-m),\\
&= \sum_{i=1}^{n}(X_i-m)^2+n(\overline{X}-m)^2-2(\overline{X}-m)\sum_{i=1}^{n}(X_i-m),\\
&= \sum_{i=1}^{n}(X_i-m)^2+n(\overline{X}-m)^2-2n(\overline{X}-m)^2,\\
&= \sum_{i=1}^{n}(X_i-m)^2-n(\overline{X}-m)^2.
\end{aligned}$$

En prenant alors $m=\mathbb{E}(X_1)$ et en intégrant l'égalité précédente, nous obtenons :

$$\begin{aligned}
\mathbb{E}(\sum_{i=1}^{n}(X_i-\overline{X})^2) &= \mathbb{E}\left(\sum_{i=1}^{n}(X_i-m)^2-n(\overline{X}-m)^2\right),\\
&= nV(X_1)-nV(\overline{X}),\\
&= (n-1)V(X_1).
\end{aligned}$$

Ce qui nous donne bien le résultat annoncé. □

On remarque alors qu'en moyenne il y a un décalage entre l'écart-type et la quantité que nous avons considérée (on parle de *biais*). C'est pourquoi nous allons préférer la quantité suivante :

Définition 7 :

Soit $(X_i)_{i\in[\![1,n]\!]}$ des variables aléatoires.

On appelle écart-type empirique débiaisé la quantité :

$$S_X^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2.$$

Proposition 8 :

Soit $(X_i)_{i\in[\![1,n]\!]}$ des variables aléatoires indépendantes et identiquement distribuées ayant un moment d'ordre 4, alors en notant μ le moment d'ordre 4 commun et σ^2 la variance commune, l'on a :

$$V(S_X^2)=\frac{1}{n}\left(\mu-\frac{n-3}{n-1}\sigma^4\right).$$

Démonstration. Comme S_X^2 est invariant si l'on modifie X_i par une constante, on peut supposer $\mathbb{E}(X_1) = 0$.
Calculons :

$$\begin{aligned}
\mathbb{E}\left[(\sum_i X_i^2 - n\overline{X})^2\right] &= \mathbb{E}\left[\sum_i X_i^4 + \sum_{i\neq j} X_i^2 X_j^2 + n^2\overline{X}^4 - 2n\overline{X}^2\sum_i X_i^2\right], \\
&= n\mu + n(n-1)\sigma^4 + n^2\mathbb{E}\left[\overline{X}^4\right] - 2n\mathbb{E}\left[\sum_{i,k}\frac{1}{n^2}X_i^2X_k^2 + \sum_i\sum_{j\neq k} X_i^2X_jX_k\right], \\
&= n\mu + n(n-1)\sigma^4 + n^2\mathbb{E}[\overline{X}^4] - \frac{2}{n}\left(n\mu + n(n-1)\sigma^4 + 0\right), \\
&= (n-2)\mu + (n-2)(n-1)\sigma^4 + n^2\mathbb{E}[\overline{X}^4].
\end{aligned}$$

où nous avons utilisé l'indépendance pour simplifier l'espérance de produit et le fait que l'on a supposé les X_i de moyenne nulle.
Maintenant,

$$\begin{aligned}
n^4\mathbb{E}\left[\overline{X}^4\right] &= \mathbb{E}\left[(\sum_i X_i)^4\right], \\
&= \mathbb{E}\left[\sum_i X_i^4 + C_3^4\sum_{i=1}\sum_{j\neq i} X_i^3X_j + \frac{C_2^4}{2}\sum_i\sum_{j\neq i} X_i^2X_j^2 + C_2^4\sum_{i,j,k \text{ distincts}} X_i^2X_jX_k + \sum_{i,j,k,l \text{ distincts}} X_iX_jX_kX_l\right] \\
&= n\mu + 0 + 3n(n-1)\sigma^4 + 0 + 0
\end{aligned}$$

Donc nous avons

$$\mathbb{E}\left[(\sum_i X_i^2 - n\overline{X})^2\right] = \frac{(n-1)^2}{n}\mu + \frac{(n-1)}{n}(n^2 - 2n + 3)\sigma^4$$

Donc en écrivant $V(S_n^2) = \mathbb{E}[(S_n^2)^2] - \mathbb{E}(S_n^2)^2$, nous pouvons conclure que

$$V(S_n^2) = \frac{1}{n}\mu + \frac{n^2 - 2n + 3 - n(n-1)}{n(n-1)}\sigma^4 = \frac{1}{n}\mu + \frac{-n+3}{n(n-1)}\sigma^4$$

Qui est bien le résultat annoncé. □

Un résultat important, que nous reverrons dans la partie 5.3 est le suivant, et il permettra de calibrer de nombreux tests :

Théoreme 9 :

Soit (X_i) une suite iid de variables, de loi $\mathcal{N}(m, \sigma^2)$. Alors :

$$\overline{X} \sim \mathcal{N}(m, \frac{\sigma^2}{n}), \qquad \overline{X} \text{ et } S_X^2 \text{ sont indépendants}$$

$$\frac{(n-1)S_X^2}{\sigma^2} = \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\sigma^2} \sim \chi^2_{n-1} \textit{ loi du chi-2 à (n-1) degré de libertés}$$

1.3 L'exemple fonctionnel fondamental : la fonction de répartition

Dans les exemples précédents, nous n'avons pas conservé toute l'information disponible avec l'expérience aléatoire. Si nous voulons conserver toute l'information, quoi de mieux que la fonction de répartition ? En effet, elle contient toute l'information de la loi selon laquelle ont été tirés nos observations. C'est dans cette philosophie d'approche de la fonction de répartition théorique que nous posons la définition suivante :

Définition 10 :

Soient $(X_i)_{i\in[\![1,n]\!]}$ des variables aléatoires, et (x_i) une réalisation. On définit la fonction de répartition empirique de ces réalisations comme la fonction :

$$F_n : t \mapsto \frac{Card\{k \in [\![1,n]\!], x_k \leqslant t\}}{n} = \frac{1}{n}\sum_{k=1}^{n} 1_{]-\infty,t]}(x_k).$$

Remarque : La fonction de répartition est un exemple de fonction aléatoire.
La fonction de répartition n'est pas continue. Elle a n points de discontinuités.

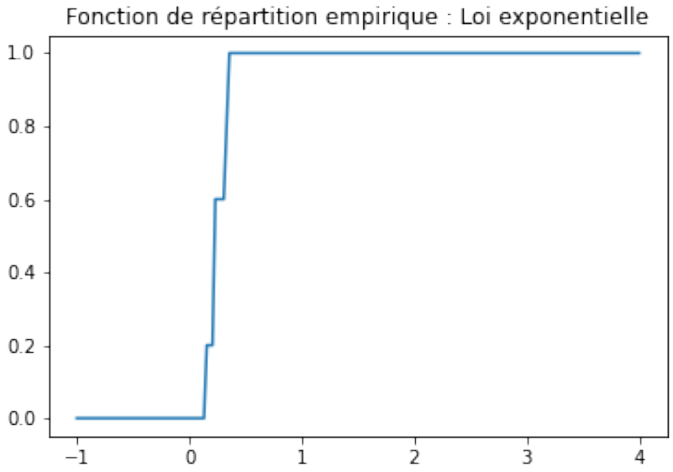


FIGURE 1.2 – Un tirage de fonction de répartition

Pour comprendre la fonction de répartition, commençons par essayer de comprendre les variables aléatoires $(F_n(t))$ pour $t \in \mathbb{R}$ fixé.

Proposition 11 :

Soit $(X_i)_{i\in[\![1,n]\!]}$ des variables aléatoires indépendantes et identiquement distribuées, F la fonction caractéristique de la loi commune, F_n la fonction caractéristique empirique correspondante et $t \in \mathbb{R}$. Alors

$$nF_n(t) \sim \mathcal{B}in(n, F(t)).$$

Démonstration.
Avec la deuxième écriture, il est évident que $nF_n(t)$ est une somme de variables de Bernoulli indépendantes. Leur probabilité de succès est alors $\mathbb{P}(X_i \in]-\infty, t]) = F(t)$ □

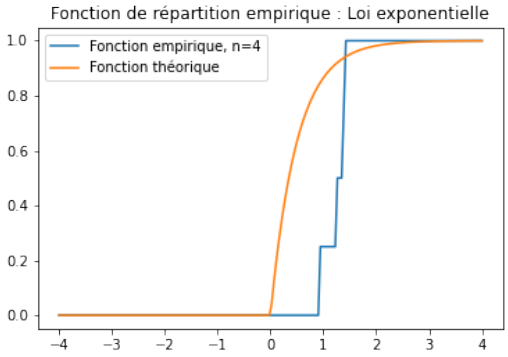
Mais alors une conséquence immédiate est le résultat suivant, justifiant notre définition de la fonction de répartition empiriques :

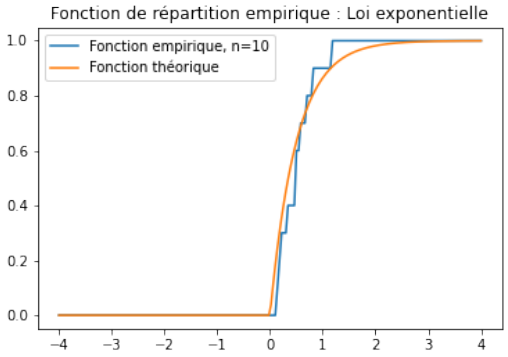
Théoreme 12 : (Bernoulli)
Soit $(X_i)_{i\in[\![1,n]\!]}$ des variables aléatoires indépendantes et identiquement distribuées, F la fonction caractéristique de la loi commune, F_n la fonction caractéristique empirique correspondante et $t \in \mathbb{R}$. Alors :

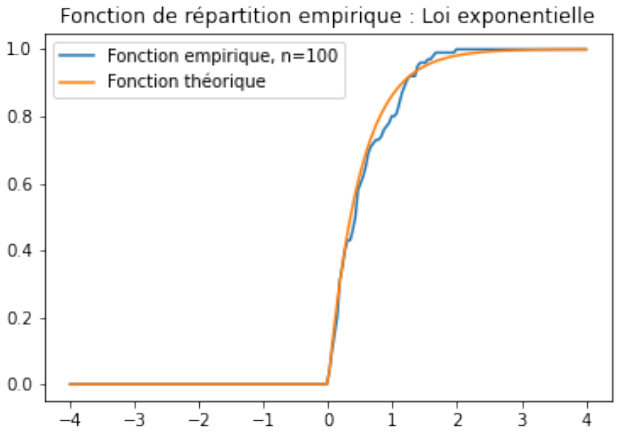
$$\lim_{n\to+\infty} F_n(t) \overset{p.s.}{=} F(t)$$

Démonstration. Il s'agit d'une application immédiate de la loi forte des grands nombres : en effet, pour $X_n \sim \mathcal{B}in(n,p)$, l'on a que $\frac{X_n}{n} \overset{p.s.}{\to} p$. □

Ce théorème signifie qu'en tout point t, la fonction de répartition empirique tend vers la fonction de répartition théorique. Nous pouvons l'observer via des simulations numériques :







Nous avons donc une convergence simple de fonction. La question naturelle qui suit est alors : peut-on faire mieux, et obtenir une convergence uniforme ? La réponse est positive !

Théoreme 13 : (Glivenko - Cantelli)

Avec les hypothèses et notations précédentes, on a presque surement que

$$F_n \overset{CVU}{\to} F.$$

De plus, la loi de $\sup_{t\in\mathbb{R}} |F_n(t) - F(t)|$ ne dépend que de n, et pas de la loi commune des variables aléatoires.

Démonstration. Il s'agit en fait d'une généralisation du deuxième théorème de Dini, auquel nous allons nous ramener.
Pour cela, on définit l'inverse généralisé de F comme la fonction
$F^{\leftarrow} : s \mapsto \inf\{x,\ F(x) \geqslant s\}$ (voir l'annexe A page 106 pour la preuve des propriétés de cette fonction).
On se donne alors une suite de variable aléatoire indépendante (U_i) de loi uniforme sur $[0,1]$, et alors la fonction de répartition des variables $F^{\leftarrow} \circ U_i$ est la fonction F. En particulier, nous avons les égalités en loi suivante :

$$\begin{aligned}
\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| &= \sup_{t\in\mathbb{R}} \left| \frac{1}{n} \sum_{k=1}^{n} 1_{\{F^{\leftarrow}\circ U_k \leqslant t\}} - F(t) \right|, \\
&= \sup_{t\in\mathbb{R}} \left| \frac{1}{n} \sum_{k=1}^{n} 1_{\{U_k \leqslant F(t)\}} - F(t) \right|, \\
&\text{c'est à dire en changeant de variable que :} \\
&= \sup_{s\in F(\mathbb{R})} \left| \frac{1}{n} \sum_{k=1}^{n} 1_{\{U_k \leqslant s\}} - s \right|, \\
&= \sup_{s\in[0,1]} \left| \frac{1}{n} \sum_{k=1}^{n} 1_{\{U_k \leqslant s\}} - s \right|.
\end{aligned}$$

Remarquons dès à présent que nous venons de montrer le deuxième point : la loi du plus grand écart entre fonction de répartition empirique et théorique est indépendante de la loi commune.

Grâce à la loi forte des grands nombres, pour tout $s \in [0,1]$, nous avons convergence presque sûre de $|\frac{1}{n}\sum_{k=1}^{n} 1_{\{U\leqslant s\}}$ vers s. Notons A_s l'ensemble des expériences pour lesquelles il y a convergence, et l'on aura $\mathbb{P}(A_s) = 1$. Nous aimerions en déduire une convergence presque sûre en tout $s \in [0,1]$, mais cet ensemble est indénombrable. Nous allons donc donner un argument supplémentaire pour dépasser cet obstacle.
Tout d'abord, $[0,1] \cap \mathbb{Q}$ est dénombrable, donc

$$\mathbb{P}\Big(\bigcap_{s\in[0,1]\cap\mathbb{Q}} \{\omega, \omega \in A_s\}\Big) = 1$$

Soit $s \in [0,1]$ et $(v_l)_{l\in\mathbb{N}}$ une suite croissante et $(w_l)_{l\in\mathbb{N}}$ une suite décroissante de limite commune s. Soit alors $l \in \mathbb{N}$, alors pour tout $n \in \mathbb{N}$, par croissance des ensembles correspondants, pour tout $\omega \in A := \cap_{s\in\mathbb{Q}} A_s$, l'on a

l'inégalité numérique suivante :

$$\frac{1}{n}\sum_{k=1}^{n} 1_{\{U_k(\omega)\leqslant v_l\}} \leqslant \frac{1}{n}\sum_{k=1}^{n} 1_{\{U_k(\omega)\leqslant s\}} \leqslant \frac{1}{n}\sum_{k=1}^{n} 1_{\{U_k(\omega)\leqslant w_l\}}.$$

En passant alors à la limite respectivement inférieure et supérieure dans les deux inégalités quand n tend vers l'infini, l'on obtient

$$v_l \leqslant \liminf_{n\to+\infty} \frac{1}{n}\sum_{k=1}^{n} 1_{\{U_k(\omega)\leqslant s\}} \leqslant \limsup_{n\to+\infty} \frac{1}{n}\sum_{k=1}^{n} 1_{\{U_k(\omega)\leqslant s\}} \leqslant w_l.$$

Maintenant, comme les termes centraux sont indépendants de l, nous obtenons par encadrement que

$$\liminf_{n\to+\infty} \frac{1}{n}\sum_{k=1}^{n} 1_{\{U_k(\omega)\leqslant s\}} = \limsup_{n\to+\infty} \frac{1}{n}\sum_{k=1}^{n} 1_{\{U_k(\omega)\leqslant s\}} = s$$

Donc pour tout $\omega \in A$, nous avons bien pour tout $s \in [0,1]$ la convergence simple suivante $\lim_{n\to+\infty} \frac{1}{n}\sum_{k=1}^{n} 1_{\{U_k(\omega)\leqslant s\}} = s$.

Nous pouvons alors conclure grâce au deuxième théorème de Dini :

Théoreme 14 :

La convergence simple d'une suite (f_n) de fonctions réelles d'une variable réelle définies et croissantes (non nécessairement continues) sur un segment $[a,b]$ de $\mathbb{R}$ vers une fonction continue f sur $[a,b]$ implique la convergence uniforme.

Démonstration. Notons d'abord que la fonction f est croissante comme limite simple de fonctions croissante.
Soit $\epsilon > 0$, et on fixe k un entier tel que $k \geqslant \frac{f(a)-f(b)}{\epsilon}$. D'après le théorème des valeurs intermédiaires appliqué à f, il existe (a_i) une subdivision $(a = a_0 < a_1 < \cdots < a_{k-1} < a_k = b$ telle que $\forall i \in [\![0, k-1]\!], f(a_i) - f(a_{i+1}) \leqslant \epsilon$.
On se donne à présent un N tel que

$$\forall n \geqslant N, \forall i \in [\![0,k]\!], |f_n(a_i) - f(a_i)| \leqslant \epsilon$$

Soit à présent $n \geqslant N$ et on se donne un $x \in [a,b[$, et l'on note i l'indice tel que $a_i \leqslant x < a_{i+1}$.
Maintenant, comme f_n et f sont croissantes, $f_n(a_i) \leqslant f_n(x) \leqslant f_n(a_{i+1})$ et $f(a_i) \leqslant f(x) \leqslant f(a_{i+1})$.
Finalement,

$$\begin{aligned} |f_n(x) - f(x)| &\leqslant |f_n(x) - f_n(a_i)| + |f_n(a_i) - f(a_i)| + |f(a_i) - f(x)|, \\ &\leqslant |f_n(a_{i+1}) - f_n(a_i)| + |f_n(a_i) - f(a_i)| + \epsilon, \\ &\leqslant 3\epsilon + \epsilon + \epsilon. \end{aligned}$$

Ce qui nous donne bien la convergence uniforme. □

□

Comme toujours, les mathématiciens sont avides (de connaissance) et à peine la convergence uniforme est établie qu'il demande de connaître la vitesse de convergence. Deux solutions s'offrent à eux : soit tabuler les lois des différents supremum en fonction de n, soit trouver un résultat sur la dispersion. Les deux ont été faits, et ainsi l'on a le résultat :

Théoreme 15 : (*Dvoretzky, Kiefer, Wolfowitz et Massart*) *admis*

Soit $(X_i)_{i\in[\![1,n]\!]}$ des variables aléatoires indépendantes et identiquement distribuées, F la fonction caractéristique de la loi commune, F_n la fonction caractéristique empirique correspondante et $t \in \mathbb{R}$. Alors :

$$\forall\ \epsilon > 0, \mathbb{P}(\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| > \epsilon) \leqslant 2e^{-2n\epsilon^2}$$

On trouve ainsi une décroissance exponentielle des queux de la distribution, qui permet de nombreux tests, comme nous le verrons dans la partie 6.2.

Exemple d'utilisation : Si l'on note D_n ce supremum, alors $\mathbb{P}(D_n > 0.04) \leqslant 0.1$ dès que $n \geqslant 996$. Et dès $n = 1992$, cette estimation chute à 1%.

1.4 Exercices

Une petite série d'exercices, dont on retrouvera la correction page 90.

Exercice 1 :

Montrer que la fonction de répartition empirique vérifie les propriétés d'une fonction de répartition : limites en $\pm\infty$, continuité à droite et limite à gauche.

Exercice 2 :

On suppose que X admette un moment d'ordre 4. Montrer les convergences L^2 et presque sûre de la variance empirique d'un échantillon de réalisation de variables indépendante de même loi que X, et donner la limite. Que dire de la vitesse de convergence ?

Exercice 3 :

Julie habite en face du métro. Elle part de chez elle 25 min avant le début de ses cours à l'INP.
Son temps d'attente du métro suit une loi normale d'espérance 5 et d'écart type 3 et la durée de son trajet ensuite suit une loi normale d'espérance 15 et d'écart type 4. Une fois au pied de l'ENSEEIHT, elle met 1 min à rejoindre sa salle de cours.
On suppose que le temps d'attente du métro et le temps de trajet sont indépendants. Pour répondre aux questions, l'on pourra se servir de la table 2.2.

- Donner la loi du temps total de trajet de Julie.
- Ce matin Julie a un cours de Statistiques, quelle est la probabilité qu'elle arrive à l'heure ? qu'elle ait plus de 5 minutes de retard ?
- Sur un trimestre, elle effectue 80 trajets pour venir à la faculté. On note $X_1, X_2, ..., X_{80}$ les 80 variables aléatoires représentants les temps de parcours de Julie pour ces trajets. Quel est le nom et la loi de la variable aléatoire Y représentant son temps moyen de parcours pour aller de chez elle à l'INP ?
- Quelle est la probabilité que, sur un semestre, elle passe plus de 27h30min en trajet pour venir à la ENSEEIHT (temps d'attente et de transport) ?

Exercice 4 :

On considère une suite (X_n) de variables aléatoires réelles de Loi de Cauchy (c'est-à-dire de densité $x \mapsto \frac{1}{\pi(1+x^2)}$). Calculer la loi de la moyenne empirique de ces variables. Que remarque-t-on ?

Exercice 5 :

Écrivez un algorithme sous Matlab pour retrouver les courbes des fonctions de répartition présentées.

Chapitre 2

Quelques lois usuelles en statistiques

2.1 Loi normale

Les lois normales sont centrales en probabilité et en statistiques. Nous renvoyons à la section 5.1 pour tous les rappels sur cette loi, et nous concentrerons ici sur l'intuition et la manipulation concrète des lois normales unidimensionnelles. Une loi normale est décrite par deux paramètres : la moyenne m, qui décrit la localisation de la variable, et la variance σ^2 qui décrit l'étalement autour de cette moyenne (où un σ^2 petit correspond à un étalement moindre).

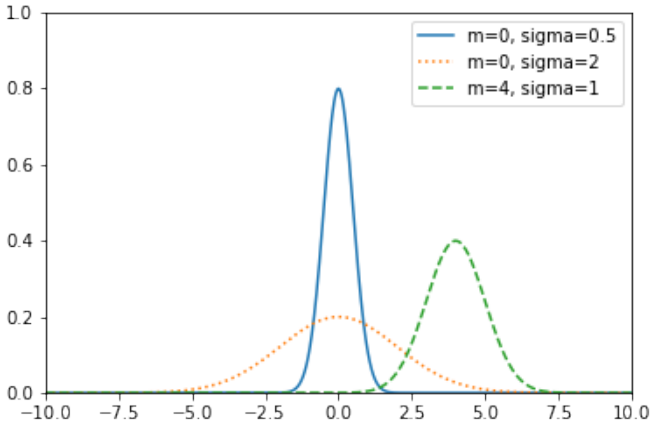


FIGURE 2.1 – Densité de lois normales

Lorsqu'on cherche à manipuler ou calculer des probabilités d'évènements "simple" (*i.e.* intersection fini d'évènements du type $X \in]-\infty, a|$ ou de leurs complémentaires) pour des cas concrets, on pourra utiliser la table 2.2. Pour cela, il s'agit de commencer par se ramener à une loi normale centrée et réduite avant d'aller chercher dans la table la valeur de la probabilité.

	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7703	0,7734	0,7764	0,7793	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8906	0,8925	0,8943	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

FIGURE 2.2 – Table de la fonction de répartition d'une loi Normale centrée réduite

Exemple : Une usine produit des bobines de fil électriques. Soit X la variable aléatoire qui à toute bobine extraite de la production associe sa longueur de fil en mètres. On suppose que X suit la loi normale d'espérance 50 et d'écart-type 0,2.

Calculer la probabilité que la longueur du fil de la bobine soit inférieure à 50,19 .

On pose $Z = \frac{X-50}{0.2} \sim \mathcal{N}(0,1)$. On trouve alors

$$\mathbb{P}(X \leqslant 50.19) = \mathbb{P}(Z \leqslant 0.95) = 0.8289$$

où l'on a lu la probabilité dans la 6^e colonne de la 10^e ligne ($\mathbb{P}(Z \leqslant 0.95) = F_Z(0.95) = F_Z(0.9 + 0.05)$). Trouver à présent un écart $a > 0$ tel que 95% des bobines aient entre $50 - a$ et $50 + a$ mètres de fils.

Une fois la loi est centré, il y aura autant de chance de dépasser la longueur maximale que de ne pas atteindre la longueur minimale :

$$\mathbb{P}(-b \leqslant Z \leqslant b) = 2\mathbb{P}(Z \leqslant b) - 1.$$

Comme l'on veut $0.95 = \mathbb{P}(50 - a \leqslant X \leqslant 50 + a) = \mathbb{P}(-\frac{a}{0.2} \leqslant Z \leqslant \frac{a}{0.2}) = 2\mathbb{P}(Z \leqslant \frac{a}{0.2}) - 1 = 2F_Z(\frac{a}{0.2}) - 1$, on doit chercher l'antécédent pour la fonction de répartition de 0.975. On trouve 1.96 comme valeur, il faut donc prendre $a = 1.96 * 0.2 = 0.392m$.

2.2 Loi du χ^2

Il y a deux manières de définir une loi du χ^2

Définition 16 : (et proposition)

Soit Y une variable aléatoire réelle, $k \in \mathbb{N}^*$ un entier et $X = (X_1, ..., X_k)^t$ un vecteur gaussien standard. Les deux propositions suivantes sont équivalentes. Lorsqu'elles sont vérifiées, on dit que Y suit une loi du χ^2_k.

- *Y suit une loi $\gamma(\frac{k}{2}, \frac{1}{2})$, loi absolument continue par rapport à la mesure de Lebesgue de densité*

$$g_{\frac{k}{2},\frac{1}{2}} : x \mapsto \frac{x^{\frac{1}{2}(k-2)} e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} \mathbb{1}_{\mathbb{R}^+}(x)$$

- *Y suit la même loi que la variable $\sum_{i=1}^{k} X_i^2$, somme de k variables gaussiennes indépendantes centrées et réduites.*

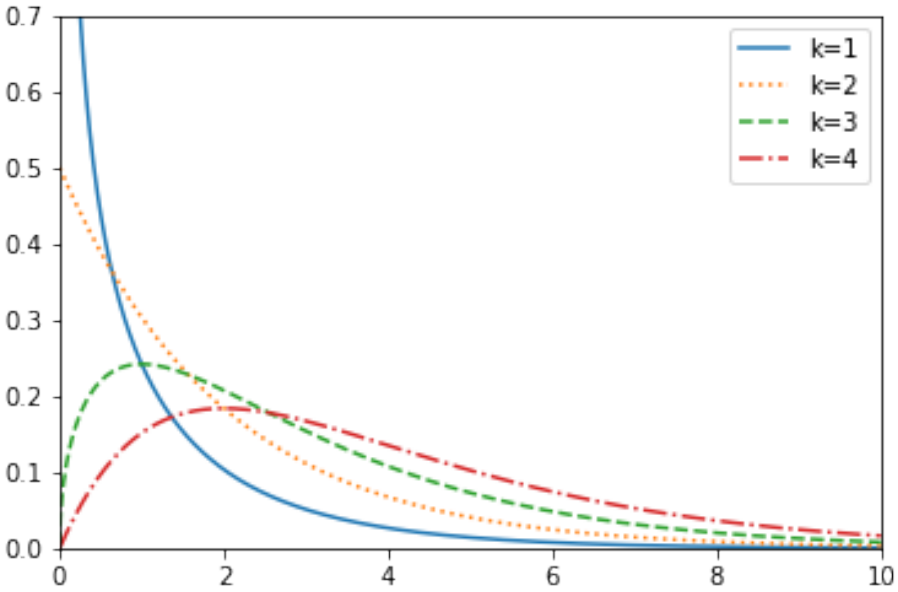


FIGURE 2.3 – Densité de lois du χ^2 pour divers paramètres

Démonstration. Rappelons que pour deux lois gamma indépendantes $X \sim \gamma(p, \lambda)$ et $Y \sim \gamma(q, \lambda)$, alors $X+Y \sim \gamma(p+q, \lambda)$.
En effet, notons $g_{a,\lambda}(t) = \frac{\lambda^a}{\Gamma(a)} t^{a-1} e^{-\lambda t}$ la densité d'une loi $\gamma(a, \lambda)$. La densité de la somme de deux variables indépendantes est le produit de convolution, ce qui nous donne :

$$\begin{aligned} f_{X+Y}(t) = g_{p,\lambda} \star g_{q,\lambda}(t) &= \int_{\mathbb{R}} g_{p,\lambda}(u) g_{q,\lambda}(t-u) du \\ &= \int_0^t \frac{\lambda^p}{\Gamma(p)} u^{a-1} e^{-\lambda u} \frac{\lambda^q}{\Gamma(q)} (t-u)^{q-1} e^{-\lambda(t-u)} du \\ &= \frac{\lambda^{p+q}}{\Gamma(p)\Gamma(q)} e^{-\lambda t} \int_0^t u^{p-1} (t-u)^{q-1} du \\ &= \frac{\lambda^{p+q}}{\Gamma(p)\Gamma(q)} e^{-\lambda t} t^{p+q-1} \int_0^1 u^{p-1} (1-u)^{q-1} du \\ &= g_{p+q,\lambda}(t) \left[\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \int_0^1 u^{p-1} (1-u)^{q-1} du \right] \end{aligned}$$

Comme f_{X+Y} et $g_{p+q,\lambda}$ sont des densités, la constante entre crochet vaut 1.

Revenons à notre proposition. Grâce au rappel, nous savons qu'il suffit de montrer que le carré d'une gaussienne unidimensionnelle centré réduite suit une loi $\gamma(\frac{1}{2}, \frac{1}{2})$. Il s'agit d'un simple changement de variable, dont les détails sont laissés en exercices. □

Comme pour la loi normale, la loi du χ^2 est tabulé :

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value									
	0.995	**0.99**	**0.975**	**0.95**	**0.90**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

FIGURE 2.4 – Table des valeurs critiques pour des lois de χ^2

2.3 Loi de Student

De manière similaire, l'on peut prendre plusieurs routes pour définir les lois de Student. La démonstration de l'équivalence des définitions est similaire à celle de χ^2 et est laissée au lecteur.

Définition 17 :

Soit Z une variable aléatoire réelle et pour $n \in \mathbb{N}$, $X \sim \mathcal{N}(0,1)$ et $Y \sim \chi^2_n$. Les deux propositions suivantes sont équivalentes. Lorsqu'elles sont vérifiées, on dit que Z suit une loi de Student à n degré de liberté (abrégé en d.d.l.).

- *La loi de Z et absolument continue par rapport à la mesure de Lebesgue, de densité*

$$g_Z : x \mapsto \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

- *Z suit la même loi que la variable $\frac{X}{\sqrt{\frac{Y}{n}}}$.*

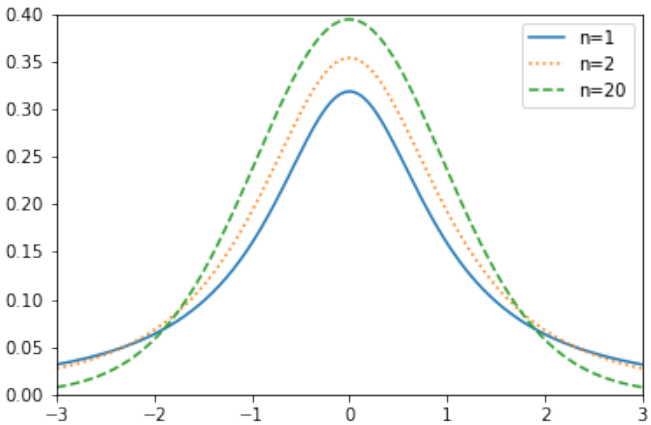


FIGURE 2.5 – Densité de lois de Student pour divers degrés de libertés

L'on peut aisément montrer que

Proposition 18 :
Soit Z une variable suivant une loi de Student à n degré de liberté. Si $n > 1$, alors $\mathbb{E}(Z) = 0$ et Si $n > 2$, alors $V(Z) = \frac{n}{n-2}$

Démonstration. Les conditions sur n sont là pour assurer l'existence des moments.
Pour calculer l'espérance, nous pouvons prendre la première définition et remarquer que la densité est paire.
Pour calculer la variance, l'on pourra remarquer que si z_k suivent des lois de Student à k degré de liberté

$$g_{Z_n}(t) + \frac{1}{n}t^2 g_{z_n}(t) = (1+\frac{t^2}{n})g_{z_n}(t) = \frac{\sqrt{n-1}}{\sqrt{n}}\frac{\Gamma(\frac{n+1}{2})\Gamma(\frac{n-2}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{n-1}{2})}g_{z_{n-2}}\left(\frac{t\sqrt{n-1}}{\sqrt{n}}\right),$$

avant d'effectuer un changement de variable et d'utiliser la relation fonctionnelle de Γ pour conclure. □

$1-\alpha$	75 %	80 %	85 %	90 %	95 %	97,5 %	99 %	99,5 %	99,75 %	99,9 %	99,95 %
k											
1	1,000	1,376	1,963	3,078	6,314	12,71	31,82	63,66	127,3	318,3	636,6
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	14,09	22,33	31,60
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	7,453	10,21	12,92
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	4,773	5,893	6,869
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,104	3,485	3,767

FIGURE 2.6 – Tabulation des quantiles des lois de Stendent

2.4 Loi de Fisher-Snedecor

L'on suit encore une fois le même parcours pour présenter les lois de Fisher-Snedecor.

Définition 19 :

Soit Z une variable aléatoire réelle et pour $(n,m) \in \mathbb{N}^2$, $X \sim \chi^2_n$ et $Y \sim \chi^2_m$. Les deux propositions suivantes sont équivalentes. Lorsqu'elles sont vérifiées, on dit que Z suit une loi de Fisher-Snedecor à n et m degré de liberté (abrégé en d.d.l.).

- *La loi de Z et absolument continue par rapport à la mesure de Lebesgue, de densité*

$$g_Z : x \mapsto \frac{\Gamma(\frac{n+m}{2}) n^{\frac{n}{2}} m^{\frac{m}{2}}}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \frac{x^{\frac{n}{2}-1}}{(nz+m)^{\frac{n+m}{2}}} \mathbb{1}_{\mathbb{R}^+}(x)$$

- *Z suit la même loi que la variable $\frac{X/n}{Y/p}$.*

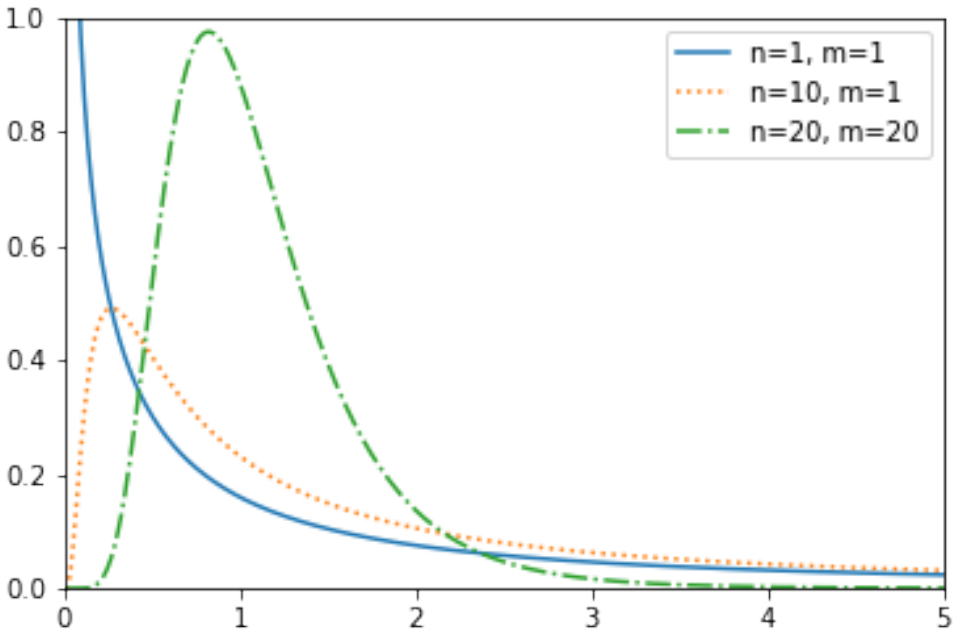


FIGURE 2.7 – Densité de lois de Fisher-Snedecor pour divers degrés de libertés

Des calculs intégraux permettent de montrer que :

Proposition 20 :

Soit Z une variable aléatoire de loi de Fisher-Snedecor à n et m degrés de libertés.
Alors si $m > 2$, $\mathbb{Z} = \frac{m}{m-2}$ et si $m > 4$, $V(Z) = 2\frac{m^2(m+n-2)}{n(m-2)^2(m-4)}$

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	>25
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.08	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
>120	3.84	2.99	2.60	2.37	2.21	2.10	1.94	1.75	1.52	1.00

FIGURE 2.8 – Quantile d'ordre 95% d'une loi de Fisher-Snedecor

2.5 Loi de Kolmogorov

Nous parlons rapidement de cette loi par souci de complétude.

Il s'agit d'une famille de loi très pratique pour les tests adéquation de loi. Pour $n \in \mathbb{N}^*$, la loi de Kolmogorov est celle de la variable aléatoire

$$\sup_{t\in[0,1]} |\frac{Card(U_i \leqslant t)}{n} - t|$$

où (U_i) est une famille de variable aléatoire indépendante de loi uniforme sur $[0, 1]$.

Le plus important à retenir est que cette loi est tabulé :

$n \backslash^{\alpha}$	0.001	0.01	0.02	0.05	0.1	0.15	0.2
1		0.99500	0.99000	0.97500	0.95000	0.92500	0.90000
2	0.97764	0.92930	0.90000	0.84189	0.77639	0.72614	0.68377
3	0.92063	0.82900	0.78456	0.70760	0.63604	0.59582	0.56481
4	0.85046	0.73421	0.68887	0.62394	0.56522	0.52476	0.49265
5	0.78137	0.66855	0.62718	0.56327	0.50945	0.47439	0.44697
6	0.72479	0.61660	0.57741	0.51926	0.46799	0.43526	0.41035
7	0.67930	0.57580	0.53844	0.48343	0.43607	0.40497	0.38145
8	0.64098	0.54180	0.50654	0.45427	0.40962	0.38062	0.35828
9	0.60846	0.51330	0.47960	0.43001	0.38746	0.36006	0.33907
10	0.58042	0.48895	0.45662	0.40925	0.36866	0.34250	0.32257
11	0.55588	0.46770	0.43670	0.39122	0.35242	0.32734	0.30826
12	0.53422	0.44905	0.41918	0.37543	0.33815	0.31408	0.29573
13	0.51490	0.43246	0.40362	0.36143	0.32548	0.30233	0.28466
14	0.49753	0.41760	0.38970	0.34890	0.31417	0.29181	0.27477
15	0.48182	0.40420	0.37713	0.33760	0.30397	0.28233	0.26585
16	0.46750	0.39200	0.36571	0.32733	0.29471	0.27372	0.25774
17	0.45440	0.38085	0.35528	0.31796	0.28627	0.26587	0.25035
18	0.44234	0.37063	0.34569	0.30936	0.27851	0.25867	0.24356
19	0.43119	0.36116	0.33685	0.30142	0.27135	0.25202	0.23731
20	0.42085	0.35240	0.32866	0.29407	0.26473	0.24587	0.23152
25	0.37843	0.31656	0.30349	0.26404	0.23767	0.22074	0.20786
30	0.34672	0.28988	0.27704	0.24170	0.21756	0.20207	0.19029
35	0.32187	0.26898	0.25649	0.22424	0.20184	0.18748	0.17655
40	0.30169	0.25188	0.23993	0.21017	0.18939	0.17610	0.16601
45	0.28482	0.23780	0.22621	0.19842	0.17881	0.16626	0.15673
50	0.27051	0.22585	0.21460	0.18845	0.16982	0.15790	0.14886
OVER 50	$\frac{1.94947}{\sqrt{n}}$	$\frac{1.62762}{\sqrt{n}}$	$\frac{1.51743}{\sqrt{n}}$	$\frac{1.35810}{\sqrt{n}}$	$\frac{1.22385}{\sqrt{n}}$	$\frac{1.13795}{\sqrt{n}}$	$\frac{1.07275}{\sqrt{n}}$

FIGURE 2.9 – Quantile de niveau $1 - \alpha$ des lois de Kolmogorov

2.6 Exercices

Une série d'exercice dont on trouvera des corrections page 91

Exercice 6 :

Écrire un programme testant l'adéquation d'observation avec une des lois de ce chapitre. Vous pourrez pour cela comparer les fréquences d'appartenance à des intervalles bien choisis à la probabilité théorique. L'on pourra approximer l'intégrale de la densité par la méthode des carrés à gauche (ou toute autre méthode numérique d'intégration).

Chapitre 3

Estimateur du maximum de vraisemblance

Dans le chapitre 1, nous avons introduit formellement les expériences statistiques, ainsi que les statistiques. Pour rappel, une statistique est une application mesurable T_n de $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ dans $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$:

$$T_n : \begin{cases} \mathbb{R}^n & \to & \mathbb{R} \\ (x_1, ..., x_n) & \mapsto & T_n(x_1, ..., x_n) \end{cases}$$

Dans ce chapitre, nous allons introduire la notion d'estimateur, ainsi que les propriétés générales que l'on peut attendre d'un bon estimateur.

Définition 21 :

Soit $F(\theta)$ une quantité qui dépend de la loi $\mathbb{P}_\theta$ On appelle estimateur de $F(\theta)$ une statistique T_n(ou une suite de statistique $(T_n)_{n\in\mathbb{N}}$).
Comme l'on cherche souvent à retrouver cette valeur, il est naturel de demander que l'estimateur permette en moyenne de retrouver $F(\theta)$. Ceci se traduirait par le fait que

$$\mathbb{E}(T_n(X_1, .., X_n)) = F(\theta) \qquad (ou\ \mathbb{E}[T_n(X_1, .., X_n)]_{n\in\mathbb{N}} = F(\theta)\ ).$$

Exemple 1 : Prenons $(\mathbb{P}_\theta)_{\theta\in\mathbb{R}_+}$ la famille des lois uniformes $\mathcal{U}([0, \theta])$. Des exemples d'estimateurs de θ sont :

- deux fois la moyenne empirique, c'est-à-dire $T_n(x_1, ..., x_n) = \dfrac{2}{n}\sum_{i=1}^{n} x_i$.
  Il s'agit d'un bon candidat, puisque par linéarité, $\mathbb{E}_\theta(\overline{X}) = \frac{\theta}{2}$.
- la plus grande des valeurs, c'est-à-dire $T_n(x_1, ..., x_n) = \sup_{k\in[\![1,n]\!]} x_k$
  Il s'agit d'un bon candidat, car l'on peut prouver que pour des lois uniformes, $\mathbb{E}_\theta(\sup_{k\in[\![1,n]\!]} X_k) = \frac{n}{n+1}\theta$.
- la statistique $T_n(x_1, ..., x_n) = \sup_{k\in[\![1,n]\!]} x_k + \inf_{k\in[\![1,n]\!]} x_k$
  Il s'agit d'un bon candidat, car l'on peut prouver que pour des lois uniformes, $\mathbb{E}_\theta(\sup_{k\in[\![1,n]\!]} X_k + \inf_{k\in[\![1,n]\!]} X_k) = \theta$.
- la statistique $T_n(x_1, ..., x_n) = \dfrac{n+1}{n} \sup_{k\in[\![1,n]\!]} x_k$.

Exemple 2 : Si (T_n) est une suite de statistique, l'estimateur prendre la limite des $T_n(x_1, ...x_n)$ lorsqu'elle existe, et une valeur arbitraire sinon, est un estimateur courant. Nous pourrons définir des qualités similaires à une simple statistique, que l'on appellera asymptotique.

3.1 Estimateur et premiers critères d'évaluations

3.1.1 Comment estimer une performance

Afin d'évaluer la performance d'un estimateur, il est important de choisir un coût pour l'écart entre la 'vrai' valeur que l'on cherche à estimer et ce que l'on estime. Il y a de nombreux choix possibles de tels coûts. Le plus simple à calculer, manipuler et le plus répandu est le coût en moyenne quadratique. Parmi les autres, l'on pourra citer les coûts L^p. Plus adapté aux tests et à l'estimation non paramétrique, on peut également citer la distance en variation, l'entropie relative et la distance de Hellinger entre la loi de la statistique et la loi théorique. On pourra ainsi s'intéresser à :

- [CS04] pour les définitions des mesures d'informations, distance en variation et entropie relatives, et leurs propriétés,
- [Vaa98] pour l'utilisation de la distance de Hellinger en analyse paramétrique,
- [Bir13] pour l'utilisation de boules pour la distance de Hellinger en estimation.

Mais revenons à un cadre plus simple, et supposons par la suite que nous cherchons à minimiser l'erreur quadratique, c'est-à-dire la moyenne du carré de la distance entre l'estimateur et la quantité désirée, c'est-à-dire que l'on veut pour le "bon" θ minimiser la quantité

$$\mathbb{E}_\theta \left[\left(T_n(X_1, ..., X_n) - F(\theta) \right)^2 \right].$$

Nous allons chercher à comprendre cette quantité. Une première remarque que nous pouvons faire est que nous pouvons appliquer une relation de Pythagore (avec comme produit scalaire entre variables aléatoire la covariance) :

Proposition 22 :
Soit T_n un estimateur de $F(\theta)$, alors on a que

$$\mathbb{E}_\theta \left[\left(T_n(X_1, ..., X_n) - F(\theta) \right)^2 \right] = V_\theta(T_n) + \left(\mathbb{E}_\theta[T_n] - F(\theta) \right)^2.$$

Démonstration. On obtient directement le résultat avec le théorème de Pythagore. Une preuve alternative est de dire que :

$$\begin{aligned}
\mathbb{E}_\theta \left[(T_n(X_1, ..., X_n) - F(\theta))^2 \right] &= \mathbb{E}_\theta \left[(T_n(X_1, ..., X_n) - \mathbb{E}_\theta[T_n] + \mathbb{E}_\theta[T_n] - F(\theta))^2 \right] \\
&= \mathbb{E}_\theta \Big[(T_n(X_1, ..., X_n) - \mathbb{E}_\theta[T_n])^2 + 2 (T_n(X_1, ..., X_n) - \mathbb{E}_\theta[T_n]) (\mathbb{E}_\theta[T_n] - F(\theta)) \\
&\qquad + (\mathbb{E}_\theta[T_n] - F(\theta))^2 \Big] \\
&= V_\theta(T_n) + 0 + (\mathbb{E}_\theta[T_n] - F(\theta))^2
\end{aligned}$$

□

Intuition et vocabulaire : Cette proposition nous dit que l'erreur quadratique est composée elle-même de deux termes d'erreur. On parle de décomposition en biais-variance. L'erreur quadratique est augmentée par la dispersion suivant la loi $\mathbb{P}_\theta$ de la statistique (terme de variance), et par l'importance de l'erreur en moyenne de la statistique (deuxième terme). On appelle le terme $\mathbb{E}[T_n] - F(\theta)$ le *biais* de la statistique. Si ce terme est nul, l'on dira que la statistique est *sans biais*.

Une des questions essentielles que nous nous poserons sera la dépendance du risque quadratique vis-à-vis de la taille de l'échantillon. Elle revient à se demander à quelle vitesse nous pouvons approcher la bonne valeur, mais également comment obtenir une telle vitesse d'approche.

3.1.2 Biais d'un estimateur

Lorsqu'on regarde la décomposition précédente, nous avons parlé de décomposition en biais-variance. Nous connaissons déjà la variance avec nos connaissances probabilistes, le statisticien ajoute à son arsenal, dans cette courte section, la notion de biais.

Définition 23 :
On dit qu'un estimateur T_n d'un paramètre $f(\theta)$ est sans biais *si* $\mathbb{E}[T_n(X_1, ..., X_n)] = f(\theta)$.
On dit qu'une suite d'estimateur (T_n) est asymptotiquement sans biais *si la limite des espérances existe et* $\lim_{n\to+\infty} \mathbb{E}[T_n(X_1, ..., X_n)] = f(\theta)$.

Exemple : Soit (X_n) une suite iid de variables, de loi envisagée uniforme sur $[0, \theta]$. On cherche à estimer θ.

- Montrons que deux fois la moyenne empirique est sans biais.
  En effet, $\mathbb{E}(\frac{2}{n}\sum_{i=1}^n X_i) = \frac{2}{n}\sum_{i=1}^n \mathbb{E}(X_i) = \frac{2}{n}\sum_{i=1}^n \frac{\theta}{2} = \theta$.

- Montrons que la suite d'estimateur qui prend le plus grand des résultats est biaisé, mais asymptotiquement sans biais.
  D'abord, pour $0 \leqslant t \leqslant \theta$, on a que $\mathbb{P}_\theta(sup(X_i) \leqslant t) = \prod_{i=1}^n \mathbb{P}_\theta(X_i \leqslant t) = \left(\frac{t}{\theta}\right)^n$.
  Ensuite, pour Y une variable aléatoire, on peut montrer avec le théorème de Tonelli que

$$\mathbb{E}(Y) = \int_0^{+\infty} \mathbb{P}(Y > t)dt$$

  Donc finalement,

$$\begin{aligned}
\mathbb{E}_\theta(\sup X_i) &= \int_0^\theta \mathbb{P}_\theta(\sup X_i > t), \\
&= \int_0^\theta (1 - \left(\frac{t}{\theta}\right)^n)dt \\
&= \theta \int_0^1 (1 - u^n)du \\
&= \theta \frac{n}{n+1}
\end{aligned}$$

  Donc en particulier, l'estimateur est biaisé, mais asymptotiquement sans biais.

Proposition 24 :
S'il existe un estimateur sans biais de variance minimale parmi tous les estimateurs sans biais, alors il est unique à vérifier cette propriété.

Démonstration. Soit T et T' deux estimateurs vérifiant la propriété (en particulier, ils ont la même variance). Considérons alors l'estimateur $U = \frac{T+T'}{2}$.
Il est clair par linéarité de l'intégrale que U est sans biais. Maintenant,

$$\begin{aligned} V(U) &= \frac{1}{4}V(T) + \frac{1}{4}V(T') + 2Cov(\frac{T}{2}, \frac{T'}{2}) \\ &\leqslant \frac{V(T) + V(T') + 2\sqrt{V(T)V(T')}}{4} \text{ par l'inégalité de Cauchy-Schwarz} \\ &= \frac{(\sqrt{V(T)} + \sqrt{V(T')})^2}{4} \\ &= V(T) \text{ par égalité des variances de T et T'} \end{aligned}$$

En particulier, comme T est de variance minimale, il y a égalité dans l'inégalité précédente. Par le cas d'égalité dans l'inégalité de Cauchy-Schwarz, T et T' sont positivement affinement lié. Comme ils sont de même moyenne et variance, ils sont donc égaux. □

Définition 25 :
On dit qu'un estimateur T sans biais est plus efficace *qu'un autre estimateur T' sans biais si sa variance est plus petite.*

3.1.3 Convergence d'une suite d'estimateurs

Une fois que l'on s'est assuré de se rapprocher suffisamment du résultat en moyenne dans notre estimation, l'on peut chercher de se rapprocher tout court. Il s'agit de la notion de convergence d'estimateur.

Définition 26 :
On dit qu'une suite d'estimateur T_n de $f(\theta)$ est convergente *(ou consistante, consistancy en anglais) si elle converge en probabilité vers ce $f(\theta)$, c'est-à-dire*

$$\forall \theta, \forall \epsilon > 0, \mathbb{P}_\theta(|T_n(X_1, ..., X_n) - f(\theta)| > \epsilon) \to 0$$

(H.P.) On dira qu'une suite d'estimateur est fortement consistante *si la convergence est presque sûre :*

$$\forall \theta, \mathbb{P}_\theta \left(\lim_{n \to +\infty} T_n(X_1, ..., X_n) = f(\theta) \right) = 1$$

Remarque : Comme la limite est une constante, demander seulement la convergence en loi vers l'estimée n'affaiblie pas la condition. Il s'agit donc de la plus faible des notions de convergence possible pour des expériences statistiques.

Proposition 27 :

Soit T_n une suite d'estimateur de $f(\theta)$ et ϕ une fonction continue en ce point.
Alors si T_n est convergente, alors $\phi(T_n)$ est une suite d'estimateur convergents vers $\phi(f(\theta))$.
Si T_n est fortement consistante, alors il en est de même pour $\phi(T_n)$

Démonstration. Il s'agit juste d'appliquer la définition de la continuité. □

Corollaire 3.1.1. *On suppose que la famille de loi indépendante identiquement distribuée admet une espérance pour toute loi.*
Si l'on peut écrire $\theta = f(\mathbb{E}(X))$ avec f une fonction continue défini au voisinage de Θ, alors l'estimateur

$$f\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

est convergent et fortement consistant.

Démonstration. D'après la loi des grands nombres, on a bien que $\frac{1}{n}\sum_{i=1}^{n} X_i \overset{p.s.}{\to} \mathbb{E}(X)$.
Donc la suite d'estimateurs est fortement consistant, et comme la convergence presque sûre implique la convergence en probabilité, elle est également convergente. □

Proposition 28 :

Une suite d'estimateurs (T_n), asymptotiquement sans biais et telle que

$$\lim_{n\to+\infty} V(T_n) = 0,$$

est convergente.

Démonstration. En utilisant la décomposition biais-variance, l'on obtient que

$$\mathbb{E}\left[(T_n - f(\theta))^2\right] = V(T_n) + (f(\theta) - \mathbb{E}(T_n))^2 \underset{n\to+\infty}{\to} 0$$

Comme la convergence en moyenne quadratique implique la convergence en probabilité, la suite d'estimateur est bien convergente. □

3.2 Modèles réguliers

Afin de faciliter nos calculs et obtenir des résultats plus fins, nous allons faire une série d'hypothèses simplificatrice sur la famille des lois. On parle alors d'hypothèse de régularité.

3.2.1 Régularité d'une expérience

Pour pouvoir parler de la régularité, nous allons avoir besoin de la vraisemblance, connu en probabilité sous le nom de densité. Le résultat que nous utiliserons est le suivant :

Définition 29 :

On dit qu'une famille de loi est dominée par une mesure ν si toutes les lois sont absolument continues par rapport à cette mesure. En particulier, il existe une fonction $(x, \theta) \mapsto f(x, \theta)$ intégrable en la première variable telle que pour toutes fonctions g mesurable bornée,

$$\mathbb{E}_\theta(g(X)) = \int_R g(x) f(x, \theta) d\nu(x).$$

On appelle alors le choix d'une telle fonction dans la classe d'équivalence d'égalité presque partout une vraisemblance des lois.

Remarque : En particulier, la connaissance de la vraisemblance permet de calculer la probabilité sous $\mathbb{P}_\theta$ de n'importe quel événement en prenant pour g une indicatrice.

Remarque fondamentalle : Dans le cas d'une suite de variables iid, on peut choisir comme mesure dominante une mesure produit. Dans ce cas, la vraisemblance d'un vecteur aléatoire s'écrira comme un produit.

Ainsi, considérons des lois absolument continues par rapport à Lebesgue (lois uniformes, normales, exponentielles, gammas, etc). Un vecteur aléatoire de variables indépendantes suivant ces lois auront alors une vraisemblance par rapport à Lebesgue égale au produit : Si $X_i \sim \mathcal{E}(\lambda_i)$, alors la vraisemblance du vecteur $(X_1, ..., X_n)$ est :

$$f : (x_1, ..., x_n, \lambda_1, ..., \lambda_n) \mapsto \prod_{i=1}^{n} \lambda_i e^{-\lambda_i x_i} \mathbb{1}_{\mathbb{R}_+}(x_i)$$

De la même manière, pour des lois discrète, nous obtiendrons une vraisemblance par rapport à la mesure de comptage égale à :

$$f : (x_1, ..., x_n, \theta) \mapsto \prod_{i=1}^{n} \mathbb{P}_\theta(X_i = x_i)$$

Nous avons maintenant tous les outils pour définir un modèle (ou expérience statistique) régulier(e).

Définition 30 :

Soit une expérience statistique dominée par une mesure μ, et $f(x, \theta)$ une vraisemblance associée.
On considère la suite de propriétés suivante par rapport à la mesure μ :

$$\{x \in \mathbb{R}, f(x, \theta) > 0\} \text{ est indépendant de } \theta. \tag{H1}$$

$$\Theta \text{ est un interval ouvert.} \tag{H2}$$

$$\partial_\theta f(x, \theta) \text{ et } \partial^2_\theta f(x, \theta) \text{ existent et sont intégrables par rapport à la loi } \mathbb{P}_\theta. \tag{H3}$$

Il est possible de deriver deux fois la vraisemblance sous le signe intégrale - hypothèses de dominations. (H4)

$$\text{La fonction } Sc_\theta(x) \mapsto \partial_\theta \ln(f(x, \theta)) \text{ (pour score) est de carré intégrable.} \tag{H5}$$

Lorsque les cinq propriétés H1-5 *sont vérifiées, on dit que le modèle est régulier.*

Remarques sur les hypothèses :

- La première hypothèse correspond à pouvoir, quitte à modifier le support de la mesure μ, mettre la densité sous la forme d'une exponentielle. On retrouve alors parmi les modèles vérifiant cette hypothèse tous les modèles exponentiels (dont on trouvera une introduction dans [PB76]).
- La deuxième hypothèse assure un espace de paramètre satisfaisant pour une analyse fonctionnelle.
- La troisième et quatrième hypothèse permettent de calculer facilement les variations des quantités d'intérêts statistiques par rapport au paramètre. En particulier, elles simplifieront la recherche d'optimum
- La dernière hypothèse est là pour permettre l'étude qui va suivre de l'information de Fisher.

Exemple de modèles réguliers :

- La famille des lois normales avec comme paramètres moyenne et variance.
- La famille des lois Gamma.
- La famille des lois de Poissons.

3.2.2 Efficacité d'un estimateur

Un des résultats principal des modèles réguliers est une minoration du risque quadratique, minoration qui décrit l'impossibilité d'avoir une statistique parfaite.
L'idée de cette minoration est de définir une notion de quantité d'information que l'on peut extraire d'une loi.
Cette quantité d'information sera directement reliée à la variation du score d'une probabilité (qui mesure la variation de la densité au voisinage d'une loi). Plus précisément :

Définition 31 :
L'on considère un modèle régulier, et l'on note $(x,\theta) \mapsto f(x,\theta)$ la vraisemblance.
On note $\mathcal{X}$ l'ensemble $\{x \in \mathbb{R} \mid f(x,\theta) > 0\}$ (cet ensemble est bien défini grâce à l'hypothèse (H1)*, et on l'appelle* support *du modèle).*
On appelle fonction score du modèle l'application :

$$Sc : \begin{cases} \mathcal{X} \times \Theta & \to & \mathbb{R} \\ (x,\theta) & \mapsto & \partial_\theta \ln(f(x,\theta)), \end{cases}$$

qui est bien défini grâce à l'hypothèse (H3)

Remarque : La présence du logarithme dans la définition du score permet, dans le cas indépendant, d'écrire la fonction score d'un vecteur comme somme de score de chacune des composantes.
Dans le cas d'un vecteur, l'on notera de même $Sc_n(\mathbf{x},\theta) = \partial_\theta \ln(f(\mathbf{x},\theta))$ la fonction score correspondante.
Dans le cas particulier de variables indépendantes, on pourra vérifier que si l'on note Sc le score de la première composante, l'on a que $Sc_n(\mathbf{x},\theta) = \sum_{i=1}^n Sc(x_i,\theta)$.

Point d'attention : Il faut éviter de confondre la fonction score (ici noté Sc) avec la variance empirique S !

À partir de ce score, il est alors possible de définir l'information de Fisher, qui est une mesure de l'information qu'apporte la variable aléatoire sur la loi (et donc sur la valeur de θ).

Définition 32 :

Pour une expérience statistique dans un modèle régulier de vraisemblance $f(x,\theta)$, on définit l'information de Fisher *de la variable aléatoire réelle X (sous-entendu sous la loi $\mathbb{P}_\theta$) comme la quantité :*

$$\mathcal{I}_X(\theta) := -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln(f(X,\theta))\right].$$

De même, on définit l'information de Fisher *du vecteur aléatoire réel $\mathbf{X}$ comme la quantité :*

$$\mathcal{I}_{\mathbf{X}}(\theta) := -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln(f(\mathbf{X},\theta))\right].$$

Intuition : L'information de Fisher donne une estimation moyenne de la courbure de la Log-vraisemblance. Si l'information est grande, la Log-vraisemblance varie beaucoup. L'estimation du maximum de vraisemblance (voir la partie suivante), obtenue en prenant le paramètre pour lequel la chance d'obtenir le résultat tirée est maximale, est meilleur : elle est plus sensible à un petit déplacement de paramètre ; on pourra donc obtenir plus d'information sur le vrai paramètre à partir de l'échantillon.

Remarque : Dans le cas où les composantes d'un vecteur aléatoire $\mathbf{X} = (X_i)_{i\in[\![1,n]\!]}$ sont indépendantes dans leur ensemble et de même loi, on aura la relation sur l'information de Fisher :$\mathcal{I}_{\mathbf{X}}(\theta) = n\mathcal{I}_{X_1}(\theta)$.

On peut directement relier l'information de Fisher avec la norme quadratique de la fonction score :

Proposition 33 :

Pour une expérience statistique avec un modèle régulier,

$$I_X(\theta) = \mathbb{E}_\theta\left[Sc(x,\theta)^2\right].$$

Démonstration. Commençons d'abord par remarque que grâce à (H4), l'on a que :

$$\int \frac{\partial^2}{\partial\theta^2}\left(f(x,\theta)\right)d\nu(x) = \frac{\partial^2}{\partial\theta^2}\left(\int f(x,\theta)d\nu(x)\right) = \frac{\partial^2}{\partial\theta^2}1 = 0.$$

Maintenant, grâce au théorème de transfert pour les mesures à densité, l'on peut calculer :

$$\begin{aligned}\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln(f(X,\theta))\right] &= \int \frac{\partial^2}{\partial\theta^2}\ln(f(X,\theta))\ f(x,\theta)d\nu(x)\\ &= \int \frac{\partial}{\partial\theta}\frac{\partial_\theta f}{f}(x,\theta)\ f(x,\theta)d\nu(x)\\ &= \int\left(\frac{\partial^2}{\partial\theta^2}f(x,\theta) - \frac{\partial_\theta f\times\partial_\theta f}{f^2}(x,\theta)\ f(x,\theta)\right)d\nu(x)\end{aligned}$$

Enfin, comme la fonction score est de carré intégrable, on peut utiliser la linéarité de l'intégrale et trouver que

$$\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln(f(X,\theta))\right] = \int\frac{\partial^2}{\partial\theta^2}f(x,\theta)d\nu(x) - \int Sc(x,\theta)^2\ f(x,\theta)d\nu(x) = 0 - \mathbb{E}_\theta\left[Sc^2\right]$$

qui est bien le résultat voulu. □

Un des résultats majeur pour minorer le risque quadratique utilise cette information de Fisher :

Théoreme 34 : *Cramer-Rao*

On considère un modèle régulier (hypothèse (H1) à (H5)). Soit h une fonction dérivable sur Θ, et soit une statistique bornée T_n qui est un estimateur sans biais de $h(\theta)$. On a la minoration de la variance suivante :

$$V_\theta(T_n) \geqslant \frac{(h'(\theta))^2}{I_{\mathbf{X}}(\theta)}.$$

En particulier, pour un estimateur de θ sans biais,

$$V_\theta(T_n) \geqslant \frac{1}{I_{\mathbf{X}}(\theta)}.$$

Démonstration. On se donne une statistique T_n qui est un estimateur borné sans biais de $h(\theta)$.
Alors

$$\begin{aligned}
|h'(\theta)| &= \left|\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[T_n(\mathbf{X})\right]\right| \\
&= \left|\frac{\partial}{\partial\theta}\int T_n(\mathbf{x})f(\mathbf{x},\theta)d\nu(x)\right|, \\
&\qquad \text{et par dérivation sous l'intégrale,} \\
&= \left|\int T_n(\mathbf{x})\frac{\partial}{\partial\theta}\left(f(\mathbf{x},\theta)\right)d\nu(x)\right| \\
&= \left|\int T_n(\mathbf{x})f(\mathbf{x},\theta)\times\frac{\partial f(\mathbf{x},\theta)}{f(\mathbf{x},\theta)}d\nu(x)\right|, \\
&\qquad \text{et comme } \int \mathbb{E}_\theta(T_n)\partial_\theta f(\mathbf{x},\theta)d\nu(x) = \mathbb{E}_\theta(T_n)\partial_\theta\int f(\mathbf{x},\theta)d\nu(x) = \mathbb{E}_\theta(T_n)\partial_\theta(1) = 0 \\
&= \left|\int T_n(\mathbf{x})\times\frac{\partial f(\mathbf{x},\theta)}{f(\mathbf{x},\theta)}f(\mathbf{x},\theta)d\nu(x) - \int \mathbb{E}_\theta(T_n)\partial_\theta f(\mathbf{x},\theta)d\nu(x)\right|, \\
&= \left|\int (T_n(\mathbf{x})-\mathbb{E}_\theta(T_n))\times\frac{\partial f(\mathbf{x},\theta)}{f(\mathbf{x},\theta)}f(\mathbf{x},\theta)d\nu(x)\right|, \\
&= \mathbb{E}_\theta\left[(T_n(\mathbf{X})-\mathbb{E}_\theta(T_n))\times\frac{\partial f(\mathbf{X},\theta)}{f(\mathbf{X},\theta)}\right] \\
&\qquad \text{et donc d'après l'inégalité de Cauchy-Schwarz,} \\
|h'(\theta)| &\leqslant \sqrt{\mathbb{E}_\theta\left[(T_n(\mathbf{X})-\mathbb{E}_\theta(T_n))^2\right]}\sqrt{\mathbb{E}_\theta\left[Sc(\mathbf{X},\theta)^2\right]} \\
&= \sqrt{V_\theta\left[T_n(\mathbf{X})\right]}\sqrt{I_X(\theta)}.
\end{aligned}$$

En prenant alors le carré de cette inégalité, l'on retrouve bien l'inégalité de Cramer-Rao. □

Remarquons que dans la preuve, l'hypothèse sur le caractère borné de T_n n'a servi que pour la majoration dans la dérivation sous l'intégrale. Des hypothèses plus larges sont donc possibles — voir par exemple le Théorème 5.10 en page 120 de [LC98].

Rappelons la définition suivante,

Définition :
Un estimateur T sans biais est dit plus efficace *qu'un autre estimateur T' sans biais si sa variance est plus petite.*

Le théorème de Cramer-Rao donne une borne inférieur sur la variance d'un estimateur sans biais, et donc une barrière minimale pour la relation d'efficacité entre statistiques. C'est ceci qui nous pousse à poser la définition suivante :

Définition 35 :
On dit qu'une statistique T_n est efficace en θ si elle est sans biais et atteint la borne de Cramer-Rao en θ. En langage propositionnel :

$$\mathbb{E}_\theta(T_n) = h(\theta) \qquad et \qquad V_\theta(T_n) = \frac{(h'(\theta)^2}{I_X(\theta)}$$

3.3 Estimation du maximum de vraisemblance

Le premier estimateur à considérer lors d'une recherche d'approximation est l'estimateur du maximum de vraisemblance. Il s'agit de chercher le paramètre θ pour lequel la probabilité $\mathbb{P}_\theta$ met le plus grand poids sur le résultat obtenu.

3.3.1 Définition et méthode de calcul pour les modèles réguliers

Définition 36 :
On considère un modèle régulier de vraisemblance $f(\mathbf{x}, \theta)$. On appelle estimateur du maximum de vraisemblance la statistique $\hat{\theta}$ qui renvoie un des paramètres θ pour lequel la vraisemblance de la réalisation est maximale.

$$f(\mathbf{x}, \hat{\theta}) = \max_{\theta \in \Theta} f(\mathbf{x}, \theta)$$

Méthode dans un modèle régulier : Si l'on a un modèle explicite, que l'on suppose régulier (en fait les hypothèses (H1), (H2) et (H3) suffisent), l'on peut pour trouver l'estimateur du maximum de vraisemblance appliquer la méthode suivante :

1. Calculer la vraisemblance, et en déduire la Log-vraisemblance
2. Calculer et chercher les points d'annulations de la fonction score :

$$\partial_\theta \ln\left(f(\mathbf{x}, \theta)\right)$$

3. Chercher le maximum de la Log-vraisemblance parmi les points d'annulation de la fonction score et le bord de Θ si l'hypothèse (H2) n'est pas vérifiée.

   Pour vérifier qu'un point est bien un maximum, l'on pourra sur un intervalle utiliser un argument de convexité (*i.e.* calculer la dérivée seconde et étudier son signe), ou encore tracer un tableau de variation pour éliminer les optimums locaux des candidats possibles.

Remarque : Pour pouvoir calculer l'Estimateur du Maximum de Vraisemblance (E.M.V., ou MLE en anglais), il faut que le modèle soit spécifié (gaussien, exponentiel, etc.)
Attention cependant, il n'y a pas forcément unicité de la solution de l'équation de vraisemblance, et donc de l'estimateur.

Une hypothèse essentielle à une étude plus fine est l'identifiabilité, c'est-à-dire que la connaissance de la loi permette de retrouver le paramètre.

Définition 37 :
On dit qu'un modèle est identifiable *si l'on a que :*

$$\theta_1 \neq \theta_2 \iff \mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}$$

3.3.2 Propriétés asymptotiques

C'est bien beau d'avoir un estimateur, mais encore faut-il qu'il converge. Nous allons dans cette partie chercher à montrer un résultat de convergence et donnerons un résultat de vitesse de convergence.

Proposition 38 :
On considère un modèle identifiable d'une suite de variable aléatoire indépendante de même loi, avec l'expérience statistique qui est dominée, vérifie (H1) et (H2), et que la Log-vraisemblance est continue. On suppose de plus que l'adhérence de Θ est compacte, et on rappelle que le vrai paramètre recherché θ_0 est dans Θ un ouvert.
On suppose de plus qu'il existe une suite de statistique $\hat{\theta}_n$ qui calcule un maximum de vraisemblance avec probabilité 1 sous la loi $\mathbb{P}_{\theta_0}$, c'est-à-dire que :

$$\mathbb{P}_{\theta_0}\left(\forall\theta, f(X_1,...X_n,\hat{\theta}_n(X_1,...X_n)) \geqslant f(X_1,...X_n,\theta)\right) = 1.$$

Enfin, on suppose que pour ce θ_0, les variables aléatoires $(\ln(f(X_1,\theta)))_\theta$ sont toutes intégrables.

Alors sous la loi $\mathbb{P}_{\theta_0}$, $\hat{\theta}_n$ converge presque sûrement vers θ_0 :

$$\mathbb{P}_{\theta_0}\left(\lim_{n\to+\infty}\hat{\theta}_n = \theta_0\right) = 1$$

Démonstration. (inspiré de [MK62])
On peut commencer par remarquer que grâce à l'inégalité de Jensen (voir page 108), l'on a l'inégalité suivante pour tout $\theta \in \Theta$:

$$\mathbb{E}_{\theta_0}\left[\ln\left(\frac{f(\mathbf{X},\theta)}{f(\mathbf{X},\theta_0)}\right)\right] \leqslant \ln\left(\mathbb{E}_{\theta_0}\left[\frac{f(\mathbf{X},\theta)}{f(\mathbf{X},\theta_0)}\right]\right)$$

Or, le membre de droite se calcule sans peine. En effet,

$$\mathbb{E}_{\theta_0}\left[\frac{f(\mathbf{X},\theta)}{f(\mathbf{X},\theta_0)}\right] = \int\left(\frac{f(\mathbf{x},\theta)}{f(\mathbf{x},\theta_0)}\right) f(\mathbf{x},\theta_0)d\nu((x) = \int f(\mathbf{x},\theta)d\nu((x) = 1$$

Donc l'on a

$$\mathbb{E}_{\theta_0}\left[\ln\left(\frac{f(\mathbf{X},\theta)}{f(\mathbf{X},\theta_0)}\right)\right] \leqslant 0$$

Ce qui se réécrit

$$\mathbb{E}_{\theta_0}\left[\ln\left(f(\mathbf{X},\theta)\right)\right] \leqslant \mathbb{E}_{\theta_0}\left[\ln\left(f(\mathbf{X},\theta_0)\right)\right]$$

Maintenant, par l'hypothèse d'un modèle issu d'une suite de variable aléatoire indépendante, la Log-vraisemblance vérifie que pour tout paramètre θ,

$$\frac{1}{n}\ln\left(f(\mathbf{X},\theta)\right) = \frac{1}{n}\sum_{i=1}^{n}\ln\left(f(X_i,\theta)\right).$$

L'on utilise encore une fois le fait que $\left(\ln\left(f(X_i,\theta)\right)\right)_{i\in\mathbb{N}}$ est une suite de variable indépendante (par transfert de l'indépendance), intégrable et de même loi. En effet, d'après la loi des grands nombres, l'on obtiendra pour tous θ la convergence presque sûre sous $\mathbb{P}_{\theta_0}$ que

$$\frac{1}{n}\sum_{i=1}^{n}\ln\left(f(X_i,\theta)\right) \overset{p.s.\mathbb{P}_{\theta_0}}{\to} \mathbb{E}_{\theta_0}\left[\ln\left(f(X_1,\theta)\right)\right].$$

En particulier, l'on obtiendra que

$$\mathbb{P}_{\theta_0}\left(\lim_{n\to+\infty}\frac{1}{n}\ln\left(f(\mathbf{X},\theta)\right) \leqslant \lim_{n\to+\infty}\frac{1}{n}\ln\left(f(\mathbf{X},\theta_0)\right)\right) = 1$$

De plus, comme $\hat{\theta}_n$ est un maximum de vraisemblance, l'on a que

$$\mathbb{P}_{\theta_0}\left(\liminf_{n\to+\infty}\frac{1}{n}\ln\left(f(\mathbf{X},\hat{\theta}_n)\right) \geqslant \lim_{n\to+\infty}\frac{1}{n}\ln\left(f(\mathbf{X},\theta_0)\right)\right) = 1$$

En particulier, si pour chaque événement l'on considère une valeur d'adhérence,

$$\mathbb{P}_{\theta_0}\left(\forall\theta^*, (\exists\phi \text{ extractrice}, \lim_n \hat{\theta}_{\phi(n)} = \theta^*) \Rightarrow \liminf_{n\to+\infty}\frac{1}{n}\ln\left(f(\mathbf{X},\hat{\theta}_{\phi(n)})\right) = \lim_{n\to+\infty}\frac{1}{n}\ln\left(f(\mathbf{X},\theta_0)\right)\right) = 1$$

ou encore en utilisant l'indépendance

$$\mathbb{P}_{\theta_0}\left(\forall\theta^*, (\exists\phi \text{ extractrice}, \lim_n \hat{\theta}_{\phi(n)} = \theta^*) \Rightarrow \liminf_{n\to+\infty}\ln\left(f(X_1,\hat{\theta}_{\phi(n)})\right) = \ln\left(f(X_1,\theta_0)\right)\right) = 1$$

Ce qui par continuité de la log-vraisemblance donne :

$$\mathbb{P}_{\theta_0}\left(\forall\theta^*, (\exists\phi \text{ extractrice}, \lim_n \hat{\theta}_{\phi(n)} = \theta^*) \Rightarrow \ln\left(f(X_1,\theta^*)\right) = \ln\left(f(X_1,\theta_0)\right)\right) = 1$$

Ce qui donne par identifiabilité que

$$\mathbb{P}_{\theta_0}\left(\forall\theta^*, (\exists\phi \text{ extractrice}, \lim_n \hat{\theta}_{\phi(n)} = \theta^*) \Rightarrow \theta^* = \theta_0\right) = 1$$

Ce qui donne par compacité de Θ que

$$\mathbb{P}_{\theta_0}\left(\lim_{n\to+\infty}\hat{\theta}_n = \theta_0\right) = 1$$

□

Remarque : L'hypothèse de compacité, tout comme celle sur l'existence d'une suite $\hat{\theta}_n$ d'estimateur de maximum de vraisemblance, peut être supprimée, quitte à affaiblir la conclusion (voir ainsi [LC98, Théorème 3.7, page 447]).

Nous avons donc la convergence presque sûre de l'estimateur du maximum de vraisemblance vers la valeur du paramètre de la loi. Comme toujours, la question qui suit est alors celle de la vitesse de convergence. Nous allons citer le résultat principal dans cette direction, sans donner de démonstration :

Théoreme 39 : *Normalité asymptotique de l'EMV-admis*

On se donne un modèle qui vérifie les hypothèses de régularité H1 à H5, mais également d'information positive $+\infty > I(\theta) > 0$. On suppose de plus que la log-vraisemblance est trois fois dérivable et qu'il existe M intégrable sous la vraie loi $\mathbb{P}_\theta$ (*i.e.* $\mathbb{E}_\theta(M(X)) < +\infty$) avec

$$\exists c, \forall x, \forall \theta_0 - c < \theta < \theta + c, \left| \frac{\partial^3}{\partial \theta^3} \ln f(\mathbf{x}, \theta) \right| \leqslant M(x).$$

L'on se donne également $(\hat{\theta_n})$ une suite de racine de

$$\partial_\theta \ln(f(\mathbf{x}, \theta)) = 0.$$

Si cette suite est un estimateur convergent de θ, alors, sous la loi $\mathbb{P}_{\theta_0}$ on a la convergence en loi

$$\sqrt{n}\left(\hat{\theta_n} - \theta_0\right) \overset{Loi}{\to} \mathcal{N}(0, \frac{1}{I(\theta_0)}).$$

Remarque : Ce résultat nous dit que l'estimateur du maximum de vraisemblance est asymptotiquement efficace, et renforce ainsi ses bons points.

3.4 Amélioration d'estimateur (H.P.)

Maintenant que nous sommes armés des outils permettant la description de bons estimateurs, nous pouvons nous demander comment améliorer ceux-ci. Nous allons voir dans cette section une introduction à une méthode d'amélioration théorique.

Statistiques exhaustives

Il faut tout d'abord se rendre compte que certaines statistiques sont plus complètes que d'autres, et contienne plus d'informations que d'autres sur les paramètres d'une loi de probabilité. C'est cette notion que nous allons définir la notion de statistique exhaustive (également appelé suffisante). Nous aurons besoin pour cela des espérances conditionnelles, décrite dans l'annexe D.

Définition 40 :

Considérons un modèle paramétré. Nous dirons qu'une statistique S est exhaustive si la loi conditionnelle de l'échantillon sachant la statistique est libre du paramètre :

$$\mathbb{P}(X \in A | S, \theta) = \mathbb{P}(X \in A | S)$$

Remarque : D'un point de vue purement pratique, si S est une statistique exhaustive, le rapport des vraisemblances peut s'écrire comme une fonction uniquement de S.

Dans des domaines dominés, il est possible de se former une intuition, en reformulant cette définition :

Proposition 41 : *(admise)*

Soit un modèle dominé et paramétré, et nous notons f_θ la densité de la loi $\mathbb{P}_\theta$.

S est une statistique exhaustive si et seulement si il existe h et g deux applications mesurables telles que

$$\forall (x,\theta) \in \mathcal{X} \times \Theta, f_\theta(x) = h(x)g(S(x),\theta).$$

Exemple : Considérons (X_i) la répétition de loi de Poisson $\mathcal{P}(\lambda)$ indépendantes. Alors, nous pouvons réécrire la vraisemblance comme :

$$f(k_1,...,k_n,\lambda) = \mathbb{P}_\lambda(X_1 = k_1,...,X_n = k_n) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{k_i}}{k_i!} = e^{-n\lambda}\lambda^{\sum_{i=1}^n k_i} \cdot \prod_{i=1}^{n}\frac{1}{k_i!}.$$

En particulier, $S = \sum_{i=1}^n k_i$ est une statistique exhaustive

Cette définition permet d'énoncer le théorème suivant :

Théoreme 42 : Rao-Blackwell-admis

Soit un modèle paramétrique et S une statistique exhaustive. Alors, pour tout estimateur T sans biais, l'estimateur

$$E = \mathbb{E}(T|S)$$

est (sans biais et) de variance inférieure à T.

Exemple d'application : Considérons la répétition de lois de Poisson $\mathcal{P}(\lambda)$ indépendantes, et la statistique exhaustive $S = \sum_{i=1}^n X_i$. Considérons l'estimateur de $e^{-\lambda}$(douteux) $T_1 = \delta_{0=X_1}$ à valeur dans $\{0,1\}$, et appliquons-lui le théorème. D'abord, T_1 est bien sans biais car $\mathbb{E}(T_1) = \mathbb{P}(X_1 = 0) = e^{-\lambda}$. L'estimateur suivant est donc plus efficace

$$T_2 := \mathbb{E}(T_1|S).$$

Calculons le :

$$\begin{aligned} T_2(k_1,...,k_n) &= \mathbb{E}(T_1|S)(k_1,...,k_n) \\ &= 1\mathbb{P}(X_1 = 0|\sum_{i=1}^n X_i = \sum_{i=1}^n k_i) + 0\mathbb{P}(X_1 \neq 0|\sum_{i=1}^n X_i = \sum_{i=1}^n k_i) \\ &= \frac{\mathbb{P}(X_1 = 0, \sum_{i=1}^n X_i = \sum_{i=1}^n k_i)}{\mathbb{P}(\sum_{i=1}^n X_i = \sum_{i=1}^n k_i)} \\ &= \frac{\mathbb{P}(X_1 = 0)\mathbb{P}(\sum_{i=2}^n X_i = \sum_{i=1}^n k_i)}{\mathbb{P}(\sum_{i=1}^n X_i = \sum_{i=1}^n k_i)} \end{aligned}$$

Or, comme une somme de loi de Poisson suit une loi de poisson de paramètre, la somme des paramètres, nous

pouvons directement dire (en notant $S := \sum_{i=1}^{n} k_i$)) :

$$T_2(k_1, ..., k_n) = e^{-\lambda} \frac{e^{-(n-1)\lambda} ((n-1)\lambda)^S}{S!} \times \left(\frac{e^{-n\lambda}(n\lambda)^S}{S!} \right)^{-1}$$
$$= (n-1)^S \frac{1}{n^S} = (1 - \frac{1}{n})^S$$

Nous pourrions alors vérifier que la variance a effectivement diminué, mais nous nous contenterons de remarquer que cet estimateur est devenu convergent (à l'aide de la convergence presque sûre de $\frac{S}{n}$ vers λ).
La question qui suit est alors, comme toujours, celle de l'optimalité du résultat précédent. Pourrions-nous trouver un estimateur plus efficace que celui-ci ? La réponse est positive, si l'on est plus attentif dans le choix de notre statistique exhaustive. Introduisons la notion de statistique complète pour savoir que demander pour atteindre l'optimalité :

Définition 43 :
On dit qu'une statistique S à valeurs dans $\mathbb{R}^q$ est complète si, pour toute fonction mesurable g telle que $g \circ S$ soit intégrable, alors

$$(\forall \theta \in \Theta, \mathbb{E}_\theta[g(S(X))] = 0) \Rightarrow (\forall \theta \in \Theta, g(S(X)) = 0 \text{ presque surement pour } \mathbb{P}_\theta)$$

Lorsque nous arriverons à trouver une telle statistique, alors nous saurons améliorer un estimateur en un estimateur optimal :

Théoreme 44 : Lehmann-Scheffé-admis
Soit un modèle paramétrique et S une statistique exhaustive **complète**. Alors, pour tout estimateur T sans biais, l'estimateur

$$E = \mathbb{E}(T|S)$$

est (sans biais et) de variance inférieure à tout estimateur sans biais.

Exemple d'application : Toujours avec le modèle de loi de Poisson indépendantes, vérifions que la statistique $S = \sum X_i$ est complète. Pour $\lambda > 0$ et g telle que $g \circ S \in L^1$, calculons :

$$\mathbb{E}\left[g\left(\frac{1}{n} \sum X_i \right) \right] = \sum_{k_1,...,k_n} g\left(\frac{1}{n} \sum k_i \right) e^{-n\lambda} \frac{\lambda^{\sum k_i}}{k_1!...k_n!}$$
$$= \sum_{k=0}^{+\infty} \sum_{\substack{k_1,...,k_n \\ \sum k_i = k}} g\left(\frac{1}{n} \sum k_i \right) e^{-n\lambda} \frac{\lambda^{\sum k_i}}{k_1!...k_n!}$$
$$= e^{-n\lambda} \sum_{k=0}^{+\infty} g\left(\frac{k}{n} \right) \lambda^k \sum_{\substack{k_1,...,k_n \\ \sum k_i = k}} \frac{1}{k_1!...k_n!}$$

Donc si cette quantité est nulle pour tout λ, la série entière suivante est nulle

$$\sum_{k=0}^{+\infty} \left(g\left(\frac{k}{n}\right) \sum_{\substack{k_1,\dots,k_n \\ \sum k_i = k}} \frac{1}{k_1!...k_n!} \right) \lambda^k = 0$$

Donc $\forall k \in \mathbb{N}, g\left(\frac{k}{n}\right) = 0$.
En particulier, $g \circ S = 0$, et la statistique est bien complète. Ainsi, la statistique $(1 - \frac{1}{n})^S$ est optimale, et donc on s'attend à ce qu'elle soit meilleure que $e^{-\frac{1}{n}\Sigma X_i}$ (ce qui n'est pas étonnant puisque cette dernière statistique est biaisée).

3.5 Exercices

Une petite série d'exercices, dont on retrouvera la correction page 91.

Exercice 7 :

On considère l'expérience statistique issue de la répétition indépendante de n tirage $(X_1, ..., X_n)$ de variable aléatoire suivant des lois de Poisson de paramètre $\lambda > 0$.

- Montrer que la moyenne empirique $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur sans biais de λ.
- Montrer que la moyenne empirique converge presque sûrement et en moyenne quadratique vers λ. En déduire le caractère convergent de l'estimateur.
- Montrer que la moyenne empirique est asymptotiquement normale, c'est-à-dire que $\sqrt{n}(\overline{X_n} - \lambda)$ converge en loi vers une loi normale, dont on précisera les paramètres.
- Montrer que la variance empirique $S_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$ est un estimateur sans biais, et montrer que $\sqrt{n}\,(S_n - \lambda)$ converge en loi vers une loi normale dont on précisera les paramètres.
  On pourra utiliser que pour une loi de Poisson, le moment centré d'ordre 4 vérifie $\mathbb{E}[(X-\lambda)^4] = \lambda + 3\lambda^2$*, mais également le lemme de Slusky (voir page 57) qui permet de dire que si* Y_n *converge en loi vers* Y *et que* X_n *converge en probabilité vers* X*, alors* X_nY_n *(respectivement* $X_n + Y_n$*) converge en loi vers* XY *(respectivement* $X+Y$*).*
- Quel estimateur privilégier pour avoir de meilleurs résultats asymptotiques ?

Exercice 8 :

On considère l'expérience statistique issue de la répétition indépendante de n tirage $(X_1, ..., X_n)$ de variable aléatoire suivant des lois exponentielles de paramètre $\lambda > 0$.

- Calculer l'estimateur du maximum de vraisemblance
- Montrer que la moyenne empirique converge presque sûrement et en moyenne quadratique vers $\frac{1}{\lambda}$. En déduire le caractère convergent de l'estimateur.

Exercice 9 :

On considère l'expérience statistique issue de la répétition indépendante de n tirage $(X_1, ..., X_n)$ de variable aléatoire suivant des lois de Bernoulli de paramètre $0 < p < 1$.

- Calculer l'estimateur du maximum de vraisemblance de p.
- L'estimateur est-il sans biais ? convergent ? efficace ?

Exercice 10 :

Examen 2004

On considère l'expérience statistique issue de la répétition indépendante de n tirage $(X_1, ..., X_n)$ de variable aléatoire suivant la même loi de densité de paramètre $\theta > 0$ et $\mu \in \mathbb{R}$:

$$\forall x > 0, \qquad f(x) = \frac{1}{\theta x \sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\theta^2}}$$

On peut vérifier que $Y_i = \ln(X_i)$ suivent des lois normales $\mathcal{N}(\mu, \theta^2)$ et l'on rappelle que la fonction génératrice des moments d'une loi normale vérifie

$$M_Y(t) = \mathbb{E}(e^{tY}) = e^{\mu t + \frac{\theta^2 t^2}{2}}$$

- Montrer que l'espérance et la variance de X s'écrivent

$$\mathbb{E}(X) = e^{\mu + \frac{\theta^2}{2}} \qquad \text{et} \qquad V(X) = e^{2\mu+\theta^2}\left(e^{\theta^2} - 1\right).$$

- Déterminer, en supposant θ connu, l'estimateur du maximum de vraisemblance de μ , noté $\hat{\mu}_{MV}$, construit à partir d'observation de $X_1, ..., X_n$. L'on prendra le soin d'établir un tableau de variation associé à la fonction à maximiser

- Montrer que $\hat{\mu}_{MV}$ est un estimateur sans biais, convergent de μ.

- Montrer que $\hat{\mu}_{MV}$ est un estimateur efficace de μ.

- Nous ne supposons plus θ connu. À l'aide des valeurs de l'espérance et de la variance de X, proposer un estimateur de (μ, θ) qui ne dépend que de

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad \overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Exercice 11 :

On considère l'expérience statistique issue de la répétition indépendante de n tirage de variable aléatoire, avec $\Theta =]-1, 1[$ et de loi $\mathbb{P}_\theta$ la loi sur $\{0, 1, -1\}$ avec des probabilités respectives $1-\theta$, $\frac{\theta}{2}$ et $\frac{\theta}{2}$. Calculer l'estimateur du maximum de vraisemblance en fonction du nombre de résultats nul n_1.

Exercice 12 :

On considère un modèle régulier $(g(\theta, x))_{\theta\in\Theta}$ dominé par une mesure ν, d'information de Fisher $I(\theta) > 0$. Rappelons que pour tout $x \in \mathbb{R}$, la fonction $\theta \mapsto f(\theta, x)$ est de classe D^2. Considérons $\Phi : \theta \mapsto \Phi(\theta)$ un C^1 difféomorphisme.

- Montrer que pour tout $x \in \mathbb{R}$, la fonction $\eta \mapsto h(\eta, x) = g(\Phi^{-1}(\eta), x)$ est de classe C^1 et calculer sa dérivée.

- En déduire l'information de Fisher du modèle $(h(\eta, x))_{\eta\in\Phi(\Theta)}$ dominé par la mesure ν en passant par la fonction score.

- Donner la borne de Cramer-Rao associé à un estimateur non biaisé de $\Phi(\theta)$.

Chapitre 4

Tests

Les tests servent au mathématicien, à l'ingénieur, au biologiste, à l'économiste et à toutes les autres professions utilisant les mathématiques que je ne puis citer ici pour des raisons de manque de place[1]. Ils leur permettent, une fois qu'ils se sont décidé sur un risque à prendre, de donner une réponse positive ou négative à une question scientifique.

4.1 La problématique des tests

La problématique des tests est vaste, prenons un exemple jouet avec nous pour ne pas nous perdre, et suivront au cours de cette section la démarche à suivre, tout en décrivant le vocabulaire :
Les moteurs d'appareils électroménagers d'une marque donnée ont une durée de vie que l'on peut modéliser par une variable aléatoire réelle de loi gaussienne $\mathcal{N}(3000, (150)^2)$. À la suite d'une modification dans la fabrication des moteurs, le fabricant affirme que les nouveaux moteurs ont une durée de vie supérieure à celle des anciens moteurs.

4.1.1 Hypothèses statistiques

Les tests permettent de savoir si l'on doit abandonner une idée actuelle en faveur d'une alternative raisonnable. Pour cela, il est nécessaire avant même de concevoir le test de dégager les hypothèses.

Définition 45 :
On appelle hypothèse statistique *toute proposition sur la nature de la loi d'une expérience statistique.*

Exemple d'hypothèses :

- Le paramètre de la loi vaut une valeur m_0.
- Le paramètre de la loi est dans une certaine région de l'espace des paramètres.
- La loi appartient à une famille donnée de lois.
- La loi est issue d'un tirage indépendant, etc.

Pour construire un test, on commencera toujours par formuler deux hypothèses statistiques :

- Une hypothèse conservatrice H_0 aussi appelée hypothèse nulle,
- Une hypothèse alternative H_1.

1. mais dont j'ai une liste fort élégante, qui ne rentre malheureusement pas dans cette marge

On doit choisir ces hypothèses de manière à ce qu'elles ne se réalisent jamais simultanément, mais également à ce qu'elles couvrent l'ensemble des éventualités envisageables.

△ Attention, si pour l'instant la distinction entre H_0 et H_1 est purement sémantique, la procédure de construction des tests conduit à briser la symétrie. Ces deux hypothèses ne jouent pas le même rôle, ce qui se lira dans la conclusion possible d'un test.

Pour tester les dires du fabricant des nouveaux moteurs, on se donne deux hypothèses statistiques :

- H_0 : la durée de vie des moteurs suivent des lois normales indépendantes de moyenne inchangée $m = m_0 = 3000$
- contre H_1 : la durée de vie des moteurs suivent des lois normales indépendantes de moyenne supérieures $m > m_0 = 3000$

4.1.2 Test statistique

Pour construire un test, une fois l'hypothèse H_0 formulée, il faut choisir une statistique de travail et une règle de décision. Le test statistique est alors une procédure qui consiste à calculer une statistique puis à appliquer une règle de décision préalablement fixée qui nous conduit à refuser ou à ne pas pouvoir pour l'instant refuser H_0 en prenant certains risques de se tromper.

En traitement du signal, les tests statistiques apparaissent dans les problèmes de détection : problèmes d'alarme, détection de la présence d'un signal noyé dans du bruit, etc.

Définition 46 :

On appelle test *sur une expérience statistique la donnée d'une statistique* $T : X \to \{0, 1\}$.
Lorsque T *vaut 1, l'on dira que l'on rejette l'hypothèse nulle* H_0.
Le plus souvent, un test sera décrit par la donnée d'une statistique $S : X \to \mathbb{R}^n$ *et d'une* région d'acceptation $A \subset \mathbb{R}^n$ *en posant* $T = 1_{S \in A}$.
Dans ce cas, on appellera zone critique la frontière (topologique) de la région d'acceptation.

On peut penser le test comme le Booléen qui répond a la question : doit-on raisonnablement rejeter H_0 *? Si le test est bien construit, l'on obtiendra de bons résultats.*

Introduisons un peu de vocabulaire pour décrire les catégories de test possibles.

Définition 47 :

On dit qu'un test est paramétrique *lorsque les hypothèses* H_0 *et* H_1 *portent sur la valeur d'un paramètre.*
Lorsque ce n'est pas le cas, on parle de test non paramétrique.

Pour les tests paramétriques, on distingue les hypothèses simples *qui fixent de façon unique la valeur d'un paramètre et les* hypothèses composites *selon lesquelles les paramètres appartiennent à une région contenant au moins deux points.*

Pour les deux hypothèses statistiques que nous nous sommes fixés dans le cadre paramétrique des durées de vie des moteurs suivant des lois normales de même écart-type :

- $H_0 : m = m_0 = 3000$ est une hypothèse simple
- $H_1 : m > m_0 = 3000$ est une hypothèse composite.

Nous choisissons (pour des raisons que nous verrons plus tard) de prendre comme statistique de test la moyenne empirique $\overline{x}$, et comme règle de décision

$$\begin{cases} \text{si } \overline{x} > \lambda_c & \text{on rejette } H_0 \\ \text{si } \overline{x} \leqslant \lambda_c & \text{on ne rejette pas} H_0 \end{cases}$$

Avec cette règle, on a que $]-\infty, \lambda_c[$ est la région d'acceptation, que $]\lambda_c, +\infty[$ est la zone de rejet, et $\{\lambda_c\}$ est la zone critique

4.1.415 Erreurs et incertitude

Dans un monde réel, à information partielle, il est illusoire d'espérer obtenir la "bonne" réponse à chaque fois. Ceci est en particulier vrai pour les tests. L'avantage des tests est qu'ils peuvent être conçus pour connaitre la fréquence d'erreur que l'on aurait en suivant leur procédure un grand nombre de fois sur des cas différents. Définissons le vocabulaire utilisé pour décrire ces erreurs :

Définition 48 :

*L'*erreur de première espèce *consiste à rejeter H_0 à tort lorsque les données sont des tirages d'une loi qui vérifie l'hypothèse nulle.*
Pour $\mathbb{P}$ vérifiant H_0, on appelle risque de première espèce *(sous-entendu vis-à-vis de $\mathbb{P}$) la probabilité sous cette loi de rejeter H_0.*

*L'*erreur de seconde espèce *consiste à ne pas rejeter H_0 lorsque les données sont des tirages d'une loi qui vérifie l'hypothèse alternative.*
Pour $\mathbb{P}$ vérifiant H_1, on appelle risque de seconde espèce *(sous-entendu vis-à-vis de $\mathbb{P}$) la probabilité sous cette loi de ne pas rejeter H_0.*

De manière plus générale, on appelle puissance *du test pour la loi $\mathbb{P}$ (vérifiant H_0 ou H_1) la probabilité sous ce test de rejeter H_0.*

On appelle niveau *du test le suppremum du risque de première espèce sur toutes les lois vérifiant H_0.*

Remarque : En traitement du signal, on donne un deuxième nom du risque de première espèce : la probabilité de fausse alarme.

Si l'on fait un tableau pour résumer : l'on trouve que α est le risque de première espèce, que β est le risque de

	Etat réel	(vraie probabilité)
	La loi vérifie H_0	La loi vérifie H_1
Test ne rejette pas H_0	$1-\alpha$	β
Test rejette H_0	α	$1-\beta$

seconde espèce et que $1-\beta$ est la puissance du test.

Remarque : Être dans la zone d'acceptation ne signifie pas que H_0 est vraie, mais seulement que les observations dont on dispose ne sont pas incompatibles avec cette hypothèse et que l'on n'a pas de raisons suffisantes de lui préférer H_1 au vu des résultats expérimentaux. C'est pour cela que nous préférons l'appellation "ne pas pouvoir rejeter H_0" à "accepter H_0".

Principe d'incertitude :

Il n'est en général pas possible de minimiser à la fois le niveau de test et le risque de seconde espèce.
En effet, une telle optimisation n'est possible que s'il est possible de faire un test qui ne se trompe jamais, et donc que si aucune des lois vérifiant H_0 prend avec probabilité positive des valeurs qu'une des lois vérifiant H_1 peut prendre avec probabilité positive (c'est-à-dire si les lois ont des supports disjoints).
Pour se forger une intuition du phénomène, considérons le problème suivant. Nous voulons tester un échantillon suivant une loi supposée normale d'écart-type connu pour donner sa valeur entre deux hypothèses. Nous choisissons alors une statistique de test et une région de rejet, disons $\overline{x}$ et une zone de la forme $\overline{x} > \mu_a$. Utilisons alors le graphique suivant :

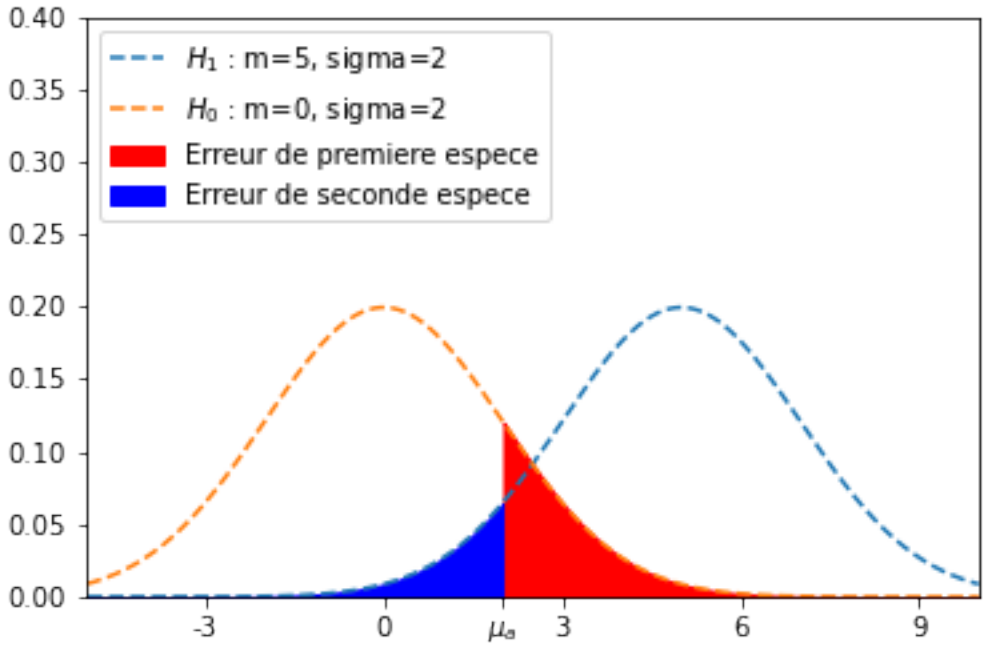


FIGURE 4.1 – Illustrations des diverses zones dans un test sur des moyennes de loi normales

Alors, sous H_0, la probabilité de dépasser μ_a, et donc de se tromper sur la réalité avec une erreur de première espèce, est égale à l'aire en rouge. De même, l'aire en bleue vaut le risque de seconde espèce.
Ce schéma souligne l'intuition : si l'on cherche à diminuer l'aire en bleu, alors l'aire en rouge augmentera.

Cette image montre uniquement l'idée générale. Nous allons dans la partie suivante chercher une méthode visuelle pour mesurer l'efficacité des tests.

4.1.5 Courbe ROC et évaluation AUC de tests

Connaitre la puissance d'un test en fonctions de la précision demandé est un élément central. C'est pour cette raison que les ingénieurs électriques et radar, en charge de la détection en terrain hostile, ont développé un outil graphique pour le décrire : la courbe ROC (Receiver Operating Characteristic). Cet outil a depuis trouvé des applications en médecine, en radiologie, en prévision de catastrophes naturelles et en météorologie.

Prenons par exemple le graphique suivant, extrait du papier de médecine de 2001 intitulé *Head-to-head comparison of N-terminal pro-brain natriuretic peptide, brain natriuretic peptide and N-terminal pro-atrial natriuretic peptide in diagnosing left ventricular dysfunction*, et écrit par ***Angelika Hammerer-Lercher, Elke Neubauer, Silvana Muller, Otmar Pachinger, Bernd Puschendorf, Johannes Mair***.

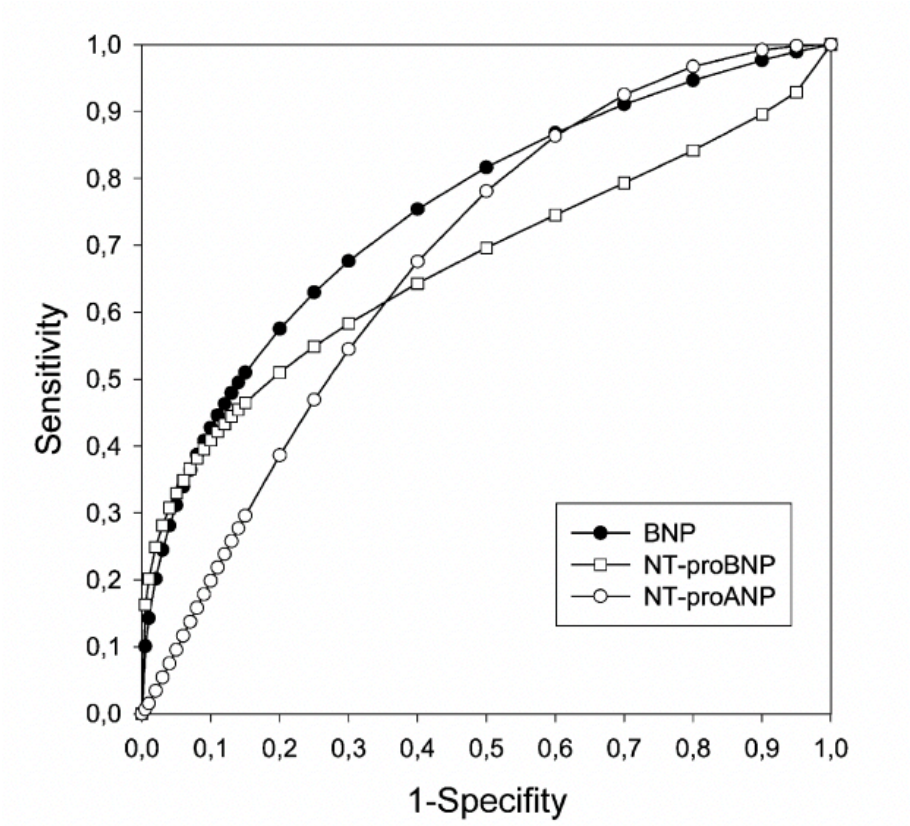


FIGURE 4.2 – Courbe ROC comparant trois tests pour la détection de l'anomalie LVEF.

Essayons de comprendre ce qu'il représente, et comment se servir de ce genre de courbe pour comparer des tests. Remarquons d'abord qu'il y a trois courbes, une pour chacun des trois tests de détections existants. En abscisse, nous retrouvons la specificity (ou erreur de première espèce) et en ordonné la sensitivity (ou puissance) du test. Ainsi, chaque courbe donne la puissance du test en fonction du niveau demandé.

Définition 49 :

On appelle courbe ROC d'un test (donné par une statistique et un ensemble de région de rejet des divers niveaux $\{I_\alpha\}$) la courbe représentative de la puissance du test en fonction du niveau du test.

Ce graphique permet de détecter immédiatement les tests à éviter à tout prix :

Proposition 50 :
Si un premier test à une courbe ROC en dessous d'un autre, alors pour tout niveau α, le deuxième test est plus puissant.
Si la courbe ROC d'un test est sous la bissectrice $y = x$, alors passer au complémentaire des zones de rejet permet d'obtenir un test plus puissant.
Un test dont la courbe ROC est la bissectrice est aussi efficace qu'un test ignorant les données et tirant un résultat au hasard.

Démonstration.
Le premier point est immédiat avec la définition de la courbe.

Pour le deuxième point, notons I_α la zone de rejet de niveau α et $\pi_1(\alpha)$ la puissance du test.
Alors en notant π_2 la puissance du test obtenu en passant au complémentaire, l'on trouve : $\mathbb{P}(X \notin I_\alpha \ |H_1) = 1 - \alpha$ donc $\pi_2(1-\alpha) = \mathbb{P}(X \notin I_\alpha \ |H_1) = 1 - \pi_1(\alpha)$.
Nous pouvons réécrire cela comme $\pi_2(\alpha) = 1 - \pi_1(1-\alpha)$, ou encore que la nouvelle courbe ROC est toujours au dessus de la bissectrice (car il s'agit de la symétrique de l'ancienne autour du 0.5,0.5), donc plus puissante.

Enfin, pour un niveau α donné, un test suivant une Bernoulli indépendante des données de paramètre α est bien de niveau α et de puissance α. Sur la bissectrice, l'on n'est pas mieux que le hasard. □

Une des (nombreuses) manière de résumer la qualité d'un test est de calculer l'aire sous la courbe.

Définition 51 :
On appelle aire sous la courbe (sous-entendu d'un test), ou AUC, l'aire sous la courbe ROC.

La pertinence de cette évaluation est justifiée par le fait suivant.
Un test qui se trompe toujours sous l'alternative a une AUC de 0, s'il n'est pas meilleur que le hasard, il aura une AUC de 0.5 et s'il ne se trompe jamais sous l'alternative son AUC vaudra 1.

Ce critère permet alors de comparer la pertinence relative de deux tests. Ainsi, dans les résultats du papier de médecine, l'on peut lire :
ROC analysis showed that among the evaluated natriuretic peptides, BNP had the best diagnostic performance to detect patients with [...] LVEF. The area under curve AUC was greater than those of the NT-natriuretic peptides.[...] The current results concur with our previous findings, demonstrating that BNP was a superior marker compared with ANP, NT-ANP, and cGMP for the identification of patients with asymptomatic LVD. .

Les auteurs utilisent ainsi la notion d'aire sous la courbe (Area Under Curve, ou AUC) pour comparer au mieux les trois tests qui les intéressent.

Méthode de construction empirique de courbe ROC :

Bien souvent, l'ingénieur est plus intéressé par la construction de telles courbes ROC pour des données réelles. Pour ce faire, il suivra la méthodologie suivante :

- Pour un grand nombre de α répartit sur $[0,1]$, déterminer les zones de rejet (par exemple déterminer μ_α pour des zones de la forme $]-\infty, \mu_\alpha]$). Pour cela, soit nous connaissons les lois sous H_0 et nous faisons un calcul théorique ; soit nous n'avons accès qu'à un grand nombre de tirages suivant des lois de H_0 et utilisons alors les quantiles empiriques.
- Effectuer un grand nombre de tirages vérifiant H_1 et utiliser les fréquences empiriques pour estimer la puissance du test de niveau α.
- Tracer alors la fonction en escalier (ou affine par morceau) avec des sauts sur les diverses valeurs de α évalués.
- (Pas systématique) Tracer les intervalles de confiance associés à la construction, obtenu pour des très grands ensembles de données par des intervalles de confiance de loi normales.

4.2 Tests paramétriques

Maintenant que nous avons tout le vocabulaire nécessaire pour décrire les tests d'hypothèses, nous allons présenter un ensemble de test portant sur la valeur d'un paramètre.

4.2.1 Test paramétrique avec deux hypothèses simples (Neyman-Pearson)

On se place encore une fois dans le cadre d'un modèle dominé (voir la définition page 34).
Il existe donc une mesure ν et une fonction $f : \Omega \times \Theta \to \mathbb{R}_+$ telle que toute loi $\mathbb{P}_\theta$ soit absolument continue par rapport à ν et de densité $f(\cdot, \theta)$.
On veut alors tester une hypothèse nulle simple contre une hypothèse alternative simple :

$$H_0 : \theta = \theta_0 \qquad \text{contre} \qquad H_1 : \theta = \theta_1.$$

Théoreme 52 :

Pour α fixé, il existe un test de la statistique "rapport de vraisemblance", éventuellement randomisé de niveau exactement α. Il consiste à comparer le rapport de vraisemblance à un seuil τ_α :

Si $\frac{f(\mathbf{X},\theta_1)}{f(\mathbf{X},\theta_0)} > \tau_\alpha$, alors on rejette H_0
Si $\frac{f(\mathbf{X},\theta_1)}{f(\mathbf{X},\theta_0)} < \tau_\alpha$, alors on ne rejette pas H_0
Enfin, si $\frac{f(\mathbf{X},\theta_1)}{f(\mathbf{X},\theta_0)} = \tau_\alpha$, on choisit de rejeter ou de ne pas rejeter suivant le résultat d'un tirage d'une loi uniforme.

De plus, ce test est le plus puissant sous l'hypothèse alternative parmi tous les tests de niveau inférieur ou égal à α.
On appelle ce test *le test de Neyman-Pearson*, ou *test du rapport des vraisemblances*.

Démonstration. On note G la fonction de répartition du rapport de vraisemblance sous $\mathbb{P}_{\theta_0}$, c'est-à-dire

$$G : t \mapsto \mathbb{P}_{\theta_0}\left(\frac{f(\mathbf{X},\theta_1)}{f(\mathbf{X},\theta_0)} \leqslant t\right)$$

Et on note $G^{\leftarrow}$ la fonction quantile associé (voir annexe A) :

$$G^{\leftarrow}(p) = \inf\{x : G(x) \geqslant p\}.$$

Posons alors comme seuil $\tau_\alpha := G^{\leftarrow}(1-\alpha)$. On a alors

$$\mathbb{P}_{\theta_0}\left(\frac{f(\mathbf{X},\theta_1)}{f(\mathbf{X},\theta_0)} > \tau_\alpha\right) = 1 - G(\tau_\alpha) \leqslant \alpha$$

Car l'inverse généralisé vérifie $G(G^{\leftarrow}(1-\alpha)) \geqslant 1-\alpha$.
Maintenant, si $G(\tau_\alpha) > 1-\alpha$, on sait que

$$\mathbb{P}_{\theta_0}\left(\frac{f(\mathbf{X},\theta_1)}{f(\mathbf{X},\theta_0)} = \tau_\alpha\right) = G(\tau_\alpha) - G(\tau_\alpha^-)$$

On considère alors U une variable aléatoire uniforme sur[0,1] indépendante des mesures.
Si l'on prend alors le test qui rejette H_0 si $\frac{f(\mathbf{X},\theta_1)}{f(\mathbf{X},\theta_0)} > \tau_\alpha$ ou si $U \leqslant \frac{G(\tau_\alpha)-(1-\alpha)}{G(\tau_\alpha)-G(\tau_\alpha^-)}$ alors ce test (randomisé) est de niveau exactement α. On le note T pour la suite.

Montrons à présent que ce test est le plus puissant parmi les tests de niveau inférieur à *alpha*. Prenons la convention qu'un test est une application qui vaut 1 si l'on rejette H_0 et 0 sinon, et considérons T' au autre test de niveau inférieur à α. Alors, avec ν la mesure dominante, l'on a

$$\begin{aligned}
\mathbb{P}_{\theta_1}(T=1) - \mathbb{P}_{\theta_1}(T=1) &= \mathbb{E}_{\theta_1}(T-T') \\
&= \int (T-T') f(x,\theta_1) d\nu(x) \\
&= \int (T-T') \frac{f(\mathbf{X},\theta_1)}{f(\mathbf{X},\theta_0)} f(x,\theta_0) d\nu(x) + \int (T-T') 1_{f(x,\theta_0)=0} f(x,\theta_1) d\nu(x) \\
&\text{or sur l'événement } f(x,\theta_0) = 0, \text{le rapport de vraisemblance est infini, et} \\
&\text{donc T=1 et donc } T - T' \geqslant 0 \\
&\geqslant \mathbb{E}_{\theta_1}\left[(T-T')\frac{f(\mathbf{X},\theta_1)}{f(\mathbf{X},\theta_0)}\right] \\
&\geqslant \mathbb{E}_{\theta_0}\left[(T-T')\left(\frac{f(\mathbf{X},\theta_1)}{f(\mathbf{X},\theta_0)} - \tau_\alpha\right)\right] + \tau_\alpha \mathbb{E}_{\theta_0}\left[(T-T')\right] \\
&\geqslant \mathbb{E}_{\theta_0}\left[(T-T')\left(\frac{f(\mathbf{X},\theta_1)}{f(\mathbf{X},\theta_0)} - \tau_\alpha\right)\right] \\
&(\text{car } \mathbb{E}_{\theta_0}[(T_i)] = \mathbb{P}_{\theta_0}(T=1) \text{ est le niveau du test i}) \\
&\geqslant 0 \text{ car } (T-T')\left(\frac{f(\mathbf{X},\theta_1)}{f(\mathbf{X},\theta_0)} - \tau_\alpha\right) \geqslant 0
\end{aligned}$$

Qui nous donne bien que la puissance sous l'alternative est maximale. □

Méthodologie : Ce résultat théorique permet de construire concrètement un test uniformément le plus puissant parmi les tests de niveau donné. Pour cela, on part de l'inégalité du théorème puis, par équivalences successives, on aboutit à une statistique de test et à la règle de décision correspondante. En général, il n'y a pas besoin de connaître la valeur de τ_α, on préférera la calculer lorsqu'il faut donner la règle de décision.

4.2.2 Test paramétrique avec hypothèse composite

On veut à présent tester des hypothèses plus complexes. En effet, il faut bien souvent envisager plus que deux lois potentielles. On pourra par exemple se rappeler l'exemple décrit dans la section précédente des durées de vie des moteurs, qui correspond à :

$$\text{Une hypothèse simple contre une hypothèse composite : } \begin{cases} H_0 : \theta = \theta_0 \\ \quad \text{contre} \quad . \\ H_1 : \theta > \theta_0 \end{cases}$$

Plus généralement, on voudrait considérer les tests paramétriques les plus complexes possibles :

$$\text{Une hypothèse composite contre une hypothèse composite : } \begin{cases} H_0 : \theta \in \Theta_0 \\ \quad \text{contre} \quad , \text{ avec } \Theta_0 \cap \Theta_1 = \varnothing. \\ H_1 : \theta \in \Theta_1 \end{cases}$$

En voyant l'efficacité du test de Neyman-Pearson, il est naturel de chercher à le généraliser. Une première idée serait de prendre comme statistique de test le rapport

$$\frac{\sup\limits_{\theta \in \Theta_0} f(\mathbf{X}, \theta)}{\sup\limits_{\theta \in \Theta_1} f(\mathbf{X}, \theta)}.$$

Comme on peut rencontrer des valeurs trop petites au dénominateur, on peut alors modifier légèrement cette statistique. On pose alors :

Définition 53 :

On pose $\Theta = \Theta_0 \cup \Theta_1$. On appelle test du rapport des maximums de vraisemblance le test basé sur la statistique de test :

$$\Lambda_n = \frac{\sup\limits_{\theta \in \Theta_0} f(\mathbf{X}, \theta)}{\sup\limits_{\theta \in \Theta} f(\mathbf{X}, \theta)}.$$

On prendra alors une zone de rejet de la forme $\Lambda_n < \lambda_\alpha$

Remarque : La statistique Λ_n vérifie $0 \leqslant \Lambda_n \leqslant 1$.

Comme nous avons généralisé les hypothèses, le résultat précédent est perdu. En revanche, on peut prouver le résultat suivant dans le cas d'une hypothèse simple contre une hypothèse composite :

Théoreme 54 :

On considère un modèle identifiable, régulier (voir page 34), paramétrisé par $\mathbb{R}$, d'information de Fisher strictement positive ($I_n(f)(\theta) > 0$) et de dérivée seconde de la Log-vraisemblance uniformément continue en θ. On suppose aussi que le modèle est issu du tirage de variables aléatoire indépendante et de même loi.
On souhaite tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$.
Si l'on considère la statistique du test du rapport des maximums de vraisemblance (avec l'EMV l'estimateur de maximum de vraisemblance) :

$$\Lambda_n := \begin{cases} \frac{f(\mathbf{X},\theta_0)}{f(\mathbf{X},\hat{\theta})} \\ \text{où} \\ \hat{\theta_n}\text{est l'EMV,} \end{cases}$$

alors on a convergence en loi du logarithme de la statistique de test vers une loi du χ^2 à un degré de liberté :

$$-2\log(\Lambda_n) \overset{Loi}{\to} \chi^2_1.$$

Démonstration.
Notons (X_i) la suite des réalisations et $\hat{\theta}_n$ l'estimateur du maximum de vraisemblance des n premières réalisations.
D'après le théorème de Taylor Lagrange appliqué à la log-vraisemblance en θ_0, il existe des variables aléatoires $\theta^*_{i,n}$ telles que

$$\forall n \in \mathbb{N}, \forall i \leqslant n, \qquad \log\left(f(X_i,\theta_0)\right) = \log\Big(f(X_i,\hat{\theta}_n(\mathbf{X}))\Big) + (\theta_0 - \hat{\theta}_n)\left(\partial_\theta \log f\right)(X_i,\hat{\theta}_n(\mathbf{X})) + \frac{1}{2}(\theta_0 - \hat{\theta}_n)^2\left(\partial^2_\theta \log f\right)(X_i,\theta^*_{i,n})$$

De plus, par uniforme continuité de $\partial^2_\theta \log f$ en θ_0, il existe une fonction $c : \mathbb{R}_+ \to \mathbb{R}_+$ croissante, indépendante de x, et de limite nulle en 0 telle que

$$\forall x, |\partial^2_\theta \log f(x,\theta) - \partial^2_\theta \log f(x,\theta_0)| \leqslant c(|\theta - \theta_0|)$$

Avec ces deux résultats, nous pouvons calculer :

$$\begin{aligned} -2\log(\Lambda_n) &= -2\sum_{i=1}^{n}\Big[\log f(X_i,\theta_0) - \log f(X_i,\hat{\theta}_n)\Big] \\ &= -2\sum_{i=1}^{n}(\theta_0 - \hat{\theta}_n)\left(\partial_\theta \log f\right)(X_i,\hat{\theta}_n(\mathbf{X})) - \sum_{i=1}^{n}(\theta_0 - \hat{\theta}_n)^2\left(\partial^2_\theta \log f\right)(X_i,\theta^*_{i,n}) \\ &= 0 - \sum_{i=1}^{n}(\theta_0 - \hat{\theta}_n)^2\left(\partial^2_\theta \log f\right)(X_i,\theta^*_{i,n}) \text{ car } \hat{\theta}_n \text{ est l'EMV.} \\ &= \frac{1}{I(\theta_0)}\left[\sqrt{I(\theta_0)}(\theta_0 - \hat{\theta}_n)\sqrt{n}\right]^2\left[\frac{1}{n}\sum_{i=1}^{n}\left(\partial^2_\theta \log f\right)(X_i,\theta^*_{i,n})\right] \\ &\quad \text{en faisant artificiellement apparaitre les termes manquants} \\ &= \left[\sqrt{I(\theta_0)}(\theta_0 - \hat{\theta}_n)\sqrt{n}\right]^2\frac{1}{I(\theta_0)}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\partial^2_\theta \log f\right)(X_i,\theta_0) + O(c(|\theta_0 - \theta^*_{i,n}|))\right] \end{aligned}$$

Or par croissance de c, $c(|\theta_0-\theta^*_{i,n}|) \leqslant c(|\theta_0-\hat{\theta}_n|)$. Comme $\hat{\theta}_n$ converge presque sûrement vers θ_0 sous l'hypothèse H_0 (voir 3.3.2), $c(|\theta_0 - \hat{\theta}_n|)$ converge presque surement vers 0.
De plus, d'après la loi des grands nombres, encore sous H_0,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\partial^2_\theta \log f\right)(X_i,\theta_0) \overset{p.s.}{\to} \mathbb{E}_{\theta_0}\left[\left(\partial^2_\theta \log f\right)(X,\theta_0)\right] = I(\theta_0)$$

Encore une fois, avec les résultats de la section 3.3.2,

$$\sqrt{I(\theta_0)}(\theta_0 - \hat{\theta}_n)\sqrt{n} \overset{Loi}{\to} \mathcal{N}(0,1)$$

Par composition avec une fonction continue, $\left[\sqrt{I(\theta_0)}(\theta_0 - \hat{\theta}_n)\sqrt{n}\right]^2$ converge alors en loi vers une χ^2 à un degré de liberté.
Nous obtenons donc le produit d'une variable aléatoire qui converge presque surement vers 1 et d'une variable aléatoire qui converge en loi vers une χ^2 à un degré de liberté. Nous pouvons alors conclure avec le lemme de Scheffé, corollaire du lemme de Slutsky :

Lemme : *Slutsky et Scheffé* Soit $X_n \overset{\mathbb{P}}{\to} c \in \mathbb{R}$ et $Y_n \overset{\mathbb{L}\times\supset}{\to} Y$ deux suites de variable aléatoire, l'une convergente en loi vers une constante et l'autre convergente en loi.
Alors le couple (X_n, Y_n) converge en loi vers (X, Y) et en particulier le produit des deux suites converge en loi vers le produit des limites :

$$X_n Y_n \overset{\mathbb{L}\times\supset}{\to} cY$$

Démonstration. Quitte à remplacer X_n par $X_n - c$, on peut supposer que X_n converge en probabilité (et donc en loi comme la limite est constante) vers 0. Montrons alors la convergence en loi du couple (X_n, Y_n) vers $(0, Y)$ en utilisant la caractérisation de Paul Lévy par les fonctions caractéristiques :

$$|\varphi_{(X_n,Y_n)}(s,t) - \varphi_{(0,Y)}(s,t)| \leqslant |\varphi_{(X_n,Y_n)}(s,t) - \varphi_{(0,Y_n)}(s,t)| + |\varphi_{(0,Y_n)}(s,t) - \varphi_{(0,Y)}(s,t)|$$

Par convergence en loi, $|\varphi_{(0,Y_n)}(s,t) - \varphi_{(0,Y)}(s,t)| = |\varphi_{Y_n}(t) - \varphi_Y(t)| \underset{t\to 0}{\to} 0$.
Maintenant, $|\varphi_{(X_n,Y_n)}(s,t) - \varphi_{(0,Y_n)}(s,t)| = |\mathbb{E}(e^{isX_n+itY_n} - e^{itY_n})| \leqslant \mathbb{E}(|e^{isX_n} - 1|)$,
Comme l'exponentielle est continue en 0, pour $\epsilon > 0$, il existe $\delta > 0$ tel que si $|x| \leqslant \delta$, alors $|e^{itx} - 1| \leqslant \epsilon$.
Donc

$$\begin{aligned}\mathbb{E}(|e^{isX_n} - 1|) &= \mathbb{E}(|e^{isX_n} - 1|\mathbb{1}_{|X_n|\leqslant\delta}) + \mathbb{E}(|e^{isX_n} - 1|\mathbb{1}_{|X_n|>\delta}) \\ &\leqslant \epsilon + 2\mathbb{P}(|X_n| > \delta)\end{aligned}$$

Et comme la convergence en loi implique la convergence en probabilité, on en déduit la convergence simple des fonctions caractéristiques, et donc la convergence en loi du couple. Enfin, le produit est une fonction continue, donc on a bien la convergence en loi annoncée. □

□

L'on pourra donc avec le rapport des vraisemblances et la fonction quantile d'un χ^2 développer un test asymptotique. Nous allons voir un autre test utilisant cette notion de convergence pour assurer une cohérence.

4.2.3 Un autre test composite : le test de proportion

On considère une expérience aléatoire où l'issue se ramène à « succès ou échec », c'est-à-dire une expérience qui peut se modéliser avec une expérience de Bernoulli. Par exemple,

- On a obtenu pile lors d'un lancer ;
- Le composant électronique testé fonctionne ;
- L'électeur sondé vote oui au référendum.

On note p la probabilité de succès lors de l'expérience. On veut tester une des trois options suivantes :

$$\begin{cases} H_0 & : & p = p_0 \\ \text{contre} & & \\ H_1 & : & p > p_0 \quad \text{test unilatéral à droite} \\ & ou & p < p_0 \quad \text{test unilatéral à gauche} \\ & ou & p \neq p_0 \quad \text{test bilateral} \end{cases}$$

On répète n fois, dans des conditions identiques et indépendantes, l'expérience aléatoire. On note X_n le nombre de succès obtenus. On sait que $X_n \sim \mathcal{B}in(n,p)$. On sait déjà, d'après la loi des grands nombres, que

$$\frac{X_n}{n} \overset{p.s.}{\to} p.$$

On peut utiliser la vitesse de convergence donnée par le théorème central limite :

$$\frac{X_n - p}{\sqrt{np(1-p)}} \overset{Loi}{\to} \mathcal{N}(0,1)$$

En effet, si cela ne nous donne malheureusement pas encore une statistique, sous H_0 l'on aura

$$Z_n := \frac{X_n - p_0}{\sqrt{np_0(1-p_0)}} \overset{Loi}{\to} \mathcal{N}(0,1)$$

On pourra alors pour notre test comparer Z_n à une valeur critique, fonction du risque de 1^{re} espèce choisi. Pour que cette approximation soit valide, il faut prendre n suffisamment grand ($n \geqslant 30$) et éviter les proportions p trop faibles ou trop importantes ($np_0 > 5$ ou $n(1-p_0) > 5$).

Pour décrire cette valeur critique, nous pouvons utiliser la fonction quantile de la loi normale (que l'on pourra trouver dans des tables numériques) q_α. Cette fonction est l'inverse de la fonction de répartition d'une loi normale centré réduite, c'est-à-dire que pour $Z \sim \mathcal{N}(0,1)$,

$$\mathbb{P}(Z < q_\alpha) = \alpha.$$

En fonction du type d'hypothèse que l'on souhaite tester, on choisira à priori une région d'acceptation différente :

- Pour le test unilatéral à droite, on s'attend à ce que sous l'hypothèse alternative, la statistique soit plus grande. On choisit donc de rejeter H_0 si $Z_n > q_{1-\alpha}$.
- Pour le test unilatéral à gauche, on s'attend à ce que sous l'hypothèse alternative, la statistique soit plus petite. On choisit donc de rejeter H_0 si $Z_n < q_\alpha$.
- Enfin, pour le test bilatéral, on n'a pas de raison de préférer une direction, et l'on prend donc une zone d'acceptation symétrique. On choisit donc de rejeter H_0 si $q_{\frac{\alpha}{2}} > Z_n$ ou si $Z_n > q_{1-\frac{\alpha}{2}}$.

Une généralisation : le test d'égalité de proportion

Il arrive souvent que ce qui nous intéresse n'est pas la valeur exacte de la proportion, mais la comparaison de la proportion entre deux échantillons. Quelques exemples de cela :

- Lorsqu'on veut étudier l'évolution d'une proportion ;
- Lorsqu'on veut savoir si un phénomène à un impact sur la variable d'intérêt (est-ce qu'un médicament a un effet supérieur à du placebo ?) ;

Nous avons alors deux séries $(X_i) \sim B(p_1)$ et $(Y_i) \sim B(p_2)$, supposé indépendants, et nous voulons tester l'hypothèse

$$\begin{cases} H_0 : & p_1 = p_2 \\ \text{contre} & \\ H_1 : & p_1 > p_2 \quad \text{test orienté de comparaison des proportions.} \\ \quad ou & p_1 \neq p_2 \quad \text{test bilateral} \end{cases}$$

Nous allons pouvoir réutiliser le théorème centrale limite. Notons $\hat{p_1}^{(n)}$ la fréquence de réussite des X_i dans les n premiers termes, et $\hat{p_2}^{(n)}$ celle des Y_i. Pour un nombre de réalisations n_i grand $n_i > 5$ et des proportions théoriques ni trop grandes ni trop faibles ($n_i p_i > 5$ et $n_i(1-p_i) > 5$), nous pouvons utiliser l'approximation gaussienne, et approcher chaque fréquence par une loi normale de moyenne p_i et d'écart-type $\sqrt{\frac{p_i(1-p_i)}{n_i}}$.
Alors en utilisant l'indépendance, nous pourrons approcher la différence des deux proportions par une loi normale de moyenne $p_1 - p_2$ et d'écart type $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$.
Donc sous H_0, si l'on note p le paramètre commun, nous pouvons choisir la statistique :

$$\frac{\hat{p_1}^{(n_1)} - \hat{p_2}^{(n_2)}}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \underset{n_1 \& n_2 \to +\infty}{\overset{Loi}{\to}} \mathcal{N}(0,1)$$

Comme nous ne connaissons pas sous l'hypothèse H_0 la valeur de p (puisqu'il s'agit d'une hypothèse composite), nous allons devoir l'approcher par le meilleur estimateur que l'on puisse trouver :

$$\hat{p} := \frac{n_1 \hat{p_1}^{(n_1)} + n_2 \hat{p_2}^{(n_2)}}{n_1 + n_2}$$

Nous prendrons donc la statistique de test

$$\frac{\hat{p_1}^{(n_1)} - \hat{p_2}^{(n_2)}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{où} \quad \hat{p} := \frac{n_1 \hat{p_1}^{(n_1)} + n_2 \hat{p_2}^{(n_2)}}{n_1 + n_2}.$$

Et comme dans le point précédent, nous construirons alors des zones de rejet asymptotique à l'aide de la loi normale en fonction du type d'hypothèse.

La question des tests est loin d'être épuisé. Pour aller plus loin, l'on pourra voir :
Dans le chapitre 5 deux autres tests, propres au modèle Gaussiens.
Dans le chapitre 6 un ensemble de tests non paramétriques, qui proposent d'étudier des questions plus générales lorsqu'on ne s'intéresse pas forcément à une loi en particulier.

4.3 Exercices

Une série d'exercices dont on trouvera des corrections page 96

Exercice 13 :

Nous considérons des variables aléatoires indépendantes $X_1, ..., X_n$ de même loi $\mathcal{N}(0, \frac{1}{\lambda^2})$.
Nous voulons choisir entre deux hypothèses :

$$H_0 : \lambda = 1 \qquad \text{contre} \qquad H_1 : \lambda = \lambda_1 < 1.$$

- Construire le test de Neyman-Pearson.
- On désire calculer l'erreur de seconde espèce pour le test de niveau $\alpha = 5\%$ issus de $n = 30$ réalisation, et une alternative $\lambda_1 = 0.9$. Donner la formule générale de cette erreur et sa valeur numérique.

Exercice 14 :

Soit $a \geqslant 0$ connu, et soient $X_1, ..., X_n$ des variables aléatoires indépendantes. La loi de X_j est $\mathcal{N}(m, j^{2a})$. On souhaite choisir entre $H_0 : m = 0$ et $H_1 : m = 2$ à partir des observations des X_j.
Construire le test de Neymann-Pearson associé pour un risque $\alpha = 0,05$. Que dire de la zone de rejet pour $a > \frac{1}{2}$ et n grand ?

Exercice 15 :

Lors d'une enquête, menée en 2014 sur 490 français, 387 estimaient que la France était en déclin. À la même question posée en août 2019 à 400 personnes, 242 estimaient que la France était en déclin.
Peut-on en conclure, à l'aide d'un test d'hypothèse avec un niveau d'erreur de 5%, que la proportion de Français pessimistes a baissé entre 2014 et 2019 ?

Chapitre 5

Les modèles gaussiens

Les vecteurs Gaussien ont une place centrale en statistiques pour deux raisons : ils apparaissent souvent spontanément dans des phénomènes limites, comme le prouve le théorème central limite, et ils sont également utilisés, car ils permettent souvent de faire des calculs exacts des lois des statistiques, voir parfois les réduisent en un problème d'algèbre linéaire.

5.1 Rappels sur les vecteurs gaussiens

Rappelons qu'une loi normale de moyenne m et de variance σ^2, noté $\mathcal{N}(m, \sigma^2)$, est la loi d'une variable aléatoire réelle X absolument continue par rapport à la mesure de Lebesgue, de densité :

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

Nous avions montré que la fonction caractéristique de $\mathcal{N}(m, \sigma^2)$ était

$$t \mapsto \varphi_X(t) := \mathbb{E}(e^{itx}) = e^{itm - \frac{t^2\sigma^2}{2}}.$$

Nous avions également défini dans le cours de probabilité un vecteur gaussien :

Définition 55 :

On dit qu'un vecteur aléatoire $X = (X_1, ...X_n)^t$ *est un* vecteur gaussien *si pour toute collection de réels* $(\lambda_1, ...\lambda_n)$*, la variable aléatoire* $\sum_{i=1}^n \lambda_i X_i$ *suit une loi normale.*

Proposition 56 :

Pour $X_1, ...X_n$ une suite de variable aléatoire gaussienne indépendante, le vecteur aléatoire $X = (X_1, ...X_n)^t$ est gaussien.
Si en plus les X_i sont centrés et réduit, on parlera de *vecteurs gaussiens standard*.

Démonstration. La preuve se fait directement en calculant pour toute combinaison linéaire la fonction caractéristique, et en remarquant qu'il s'agit de la fonction caractéristique d'une loi normale. □

Rappelons que nous avions défini la matrice de covariance d'un ensemble de variable $(X_i)_{i\in[\![1,n]\!]}$ comme la matrice carrée K dont les coefficients sont

$$K_{i,j} = \mathbb{E}(X_iX_j) - \mathbb{E}(X_i)\mathbb{E}(X_j).$$

Cette matrice joue un rôle central pour les vecteurs gaussiens.
Nous avons déjà vu que pour un vecteur aléatoire et un vecteur fixé $\lambda \in \mathbb{R}^n$,

$$Var(<\lambda, X>) = Var(\sum_i \lambda_i X_i) = \lambda^t K \lambda$$

De ce fait, et de la stabilité de l'ensemble des vecteurs gaussien par modification linéaire découle immédiatement la proposition suivante :

Proposition 57 :
Si $X = (X_1, ...X_n)^t$ est un vecteur gaussien de matrice de covariance K et $A \in M_{p,n}$ est une matrice, alors le vecteur AX est gaussien, de matrice de covariance

$$AKA^t.$$

Démonstration. Le caractère Gaussien est immédiat, puisque pour tous $\lambda \in \mathbb{R}^n$, l'on a

$$<\lambda, AX> = <A^t\lambda, X>.$$

Pour la matrice de covariance, il s'agit d'utiliser directement la bilinéarité de la covariance :

$$Cov\left[(AX)_i, (AX)_j\right] = \sum_q \sum_p a_{i,q} Cov\left[X_q, X_p\right] a_{j,p} = AKA^t_{i,j}$$

□

Nous avions déjà vu qu'une matrice de covariance est symétrique, semi-définie positive. Nous allons pouvoir utiliser la proposition d'algèbre linéaire suivante, vu dans l'annexe C.1, pour caractériser la loi d'un vecteur Gaussien :

Proposition 58 :
Soit $K \in M_{n,n}$ une matrice symétrique semi-définie positive, alors il existe (au moins) une matrice $L \in M_{n,n}$ telle que $K = L^tL$.

Ceci permet directement de dire que tout vecteur gaussien est la transformation linéaire d'un vecteur gaussien standard.

Corollaire 5.1.1. *Soit $(Y_1, ..., Y_n)^t$ un vecteur gaussien, alors il existe $L \in M_{n,n}$ et $(X_1, ..., X_n)^t$ vecteur gaussien standard tels que Y et LX sont de même loi.*

Démonstration. L'on peut remarquer que pour Y un vecteur Gaussien, sa fonction caractéristique ne dépend que de sa moyenne m et de sa matrice de covariance K :

$$\forall \lambda \in \mathbb{R}^n, \quad \varphi_Y(\lambda) = \mathbb{E}[e^{i<\lambda, X>}] = e^{i<\lambda, m> - \frac{\lambda^t K \lambda}{2}}$$

□

Dans les faits, nous utiliserons plutôt la caractérisation suivante :

Théoreme 59 :

Soit $m \in R^d$, et Λ une matrice $d \times d$ symétrique *définie positive*. On considère X un vecteur aléatoire, alors on a l'équivalence entre les points

- X est un vecteur gaussien de moyenne m et de matrice de covariance Λ ;
- X admet une densité par rapport à Lebesgue de la forme

$$\frac{1}{(\sqrt{2\pi})^d\sqrt{det(\Lambda)}}\exp(-\frac{(x-m)^t\Lambda(x-m)}{2})$$

Lorsque c'est le cas, la variable aléatoire X_j est alors de loi $\mathcal{N}(m_j, \Lambda_{jj})$.

Démonstration. La formule est trivialement vraie pour les vecteurs gaussiens standards. Le cas général découle de la formule de changement de variables pour la mesure de Lebesgue et de la proposition précédente. □

Définition 60 :

Soit $(X_1, ...X_n)^t$ un vecteur gaussien centré. On appelle espace gaussien *engendré par les X_i l'espace vectoriel engendré par la famille des X_i :*

$$Vect(X_i) = \{\sum_{i=1}^{n} \lambda_i X_i | \lambda \in \mathbb{R}^n\}$$

5.2 Normes de vecteurs gaussiens et théorème de Cochran (H.P.)

Une des propositions étonnante des espaces gaussiens est que l'indépendance s'y lit très facilement, et se réduit même à un problème géométrique :

Proposition 61 :

Soit Y, Z deux variables aléatoires centré appartenant à un même espace gaussien.
Alors Y et Z sont indépendantes si et seulement si elles ont orthogonales pour le produit scalaire sur les vecteurs centré qu'est la covariance (*i.e.* si elles sont décorrélées) :

$$\mathbb{E}[< Y.Z >] = Cov(Y, Z) = 0.$$

Démonstration. On peut, sans perte de généralité, supposer que le vecteur gaussien $(X_1, ..., X_n)$ générant l'espace gaussien est un vecteur standard (centré de variance I_n).
Comme le sens direct est immédiat, nous allons prouver le sens réciproque. On écrit pour cela $Y = \sum_{i=1}^n \lambda_i X_i$ et $Z = \sum_{i=1}^n \tilde{\lambda}_i X_i$.

Le caractère décorrélé des variables Y et Z se lit dans les paramètres comme $\mathbb{E}[YZ] = \sum_{i=1}^{n} \lambda_i \tilde{\lambda}_i = 0$.
Pour montrer l'indépendance, nous allons montrer que la fonction caractéristique du couple (Y, Z) est le produit de leurs fonctions caractéristiques respectives. Soit $\mu \in \mathbb{R}$ et $\tilde{\mu} \in \mathbb{R}$. Alors,

$$\begin{aligned}
\mathbb{E}[e^{i\mu Y} e^{i\tilde{\mu} Z}] &= \mathbb{E}[e^{i\mu \sum_i \lambda_i X_i} e^{i\tilde{\mu} \sum_i \tilde{\lambda}_i X_i}] \\
&= \mathbb{E}[\prod_i e^{i(\mu\lambda_i + \tilde{\mu}\tilde{\lambda}_i) X_i}] \\
&= \prod_i \mathbb{E}[e^{i(\mu\lambda_i + \tilde{\mu}\tilde{\lambda}_i) X_i}] \quad \text{par indépendance des } X_i \\
&= \prod_i e^{-\frac{(\mu\lambda_i + \tilde{\mu}\tilde{\lambda}_i)^2}{2}} \\
&= exp(-\frac{1}{2} \sum_i (\mu\lambda_i + \tilde{\mu}\tilde{\lambda}_i)^2) \\
&= exp(-\frac{1}{2} \sum_i \mu^2\lambda_i^2 + \tilde{\mu}^2\tilde{\lambda}_i^2) \quad \text{Par orthogonalité des coefficients - hypothèse de décorrelation} \\
&= \mathbb{E}[e^{i\mu \sum_i \lambda_i X_i}] \mathbb{E}[e^{i\tilde{\mu} \sum_i \tilde{\lambda}_i X_i}] \\
&= \mathbb{E}[e^{i\mu Y}] \mathbb{E}[e^{i\tilde{\mu} Z}]
\end{aligned}$$

Ce qui nous donne bien l'indépendance. □

Une conséquence immédiate de la preuve est le résultat suivant :

Corollaire 5.2.1. *Considérons deux sous espaces vectoriels E et E' d'un espace gaussien. Alors toutes les variables aléatoires de E sont indépendantes de toutes les variables de E' si et seulement si E et E' sont orthogonales.*

Une deuxième propriété fondamentale des vecteurs gaussien standard est leur invariance par transformation orthogonale, qui permet d'obtenir le résultat suivant en utilisant la théorie spectrale (voir annexe C.1) :

Proposition 62 :
Soit $X = (X_1, ..., X_n)^t$ un vecteur gaussien centré de matrice de covariance K et soit $M \in S_n^+$ une matrice symétrique semi-définie positive. On considère également (Z_i) des variables aléatoires indépendantes de loi χ_1^2.
Alors si l'on note $K = LL^t$ une décomposition de Cholesky (L triangulaire supérieure) et (λ_i) les valeurs propres avec multiplicités de la matrice L^tML, la variable aléatoire X^tMX est distribué comme $\sum_i \lambda_i Z$.

Démonstration. Sans perte de généralité, on peut supposer $X = LY$ avec $Y \sim \mathcal{N}(0, I_n)$ un vecteur gaussien standard.
D'après le théorème spectral, il existe O une matrice orthogonale telle que

$$L^tML = O^t \text{diag}(\lambda_1, ..., \lambda_n) O.$$

Or OY est encore un vecteur standard, puisque $O^tO = I_n$. Donc

$$X^tMX = Y^tL^tMLY = (OY)^t\text{diag}(\lambda_1, ..., \lambda_n)(OY)$$

Ce qui nous donne directement le résultat. □

Nous avons à présent tous les ingrédients pour montrer le résultat suivant. Il s'agit d'un résultat fondamental pour la construction de nombreux tests.

Théoreme 63 : *Théorème de Cochran*

Soit $X \sim \mathcal{N}(0, I_n)$ un vecteur gaussien standard. L'on considère une décomposition orthogonale $\mathbb{R}^n = \oplus_{i=1}^k E_i$ où les E_i sont des sous-espaces vectoriels deux à deux orthogonaux de $\mathbb{R}^n$. On note π_{E_i} la projection orthogonale sur l'espace E_i.
Alors la famille $(\pi_{E_i}X)_{1\leqslant i\leqslant k}$ est une famille indépendante, et pour chaque j, on connait la loi de la norme :

$$||\pi_{E_j}X||_2^2 \sim \chi^2_{dim(E_j)}$$

Démonstration. Remarquons d'abord que la loi de la norme quadratique est une conséquence immédiate de la proposition 62. En effet, la matrice de covariance de $\pi_{E_j}X$ est $\pi_{E_j}\pi_{E_j}^t = \pi_{E_j}$. Or les valeurs propres d'une matrice de projection sont 1 avec comme multiplicité la dimension de l'espace stable $dim(E_j)$ et 0.

Pour montrer l'indépendance, nous pouvons utiliser la proposition 61et son corollaire pour montrer que π_{E_j} est indépendant de $(\pi_{E_i})_{i<j}$. En effet, les π_{E_i} sont bien dans un espace gaussien comme image par une application linaire d'un espace gaussien, et sont orthogonaux puisque $\pi_{E_j}X \in E_j \perp \oplus_{i=1}^{j-1} E_i$ (et donc en particulier $\forall Y \in Vect(\pi_{E_i})_{i<j}, \mathbb{E}(<\pi_{E_j}X, Y>) = 0$). □

Ceci nous permet de montrer directement le résultat suivant d'estimation d'une loi gaussienne en dimension 1 que nous avions admis :

Théoreme 64 :
Soit (X_i) une suite iid de variables, de loi $\mathcal{N}(m, \sigma^2)$. Alors :

$$\overline{X} \sim \mathcal{N}(m, \frac{\sigma^2}{n}), \qquad \overline{X} \text{ et } S_X^2 \text{ sont indépendants}$$

$$\frac{(n-1)S_X^2}{\sigma^2} = \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\sigma^2} \sim \chi^2_{n-1} \textit{ loi du chi-2 à (n-1) degré de libertés}$$

Démonstration. Sans perte de généralité, l'on peut supposer que $m = 0$ et que $\sigma = 1$. L'on remarque alors que

$$\overline{X}_n = \frac{1}{n}\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} \times \begin{pmatrix}1 & \cdots & 1\end{pmatrix} \times X.$$

C'est-à-dire que $\overline{X}_n$ est la projection orthogonale de X sur l'espace engendré par $(1, \ldots, 1)^t$. Notons π le

projecteur. D'après le théorème de Cochran, πX et $X - \pi X$ sont indépendants, et la norme quadratique de $X - \pi X$ suit une loi χ^2_{n-1}.
Ceci nous donne les deux derniers points du théorème, la loi de $\overline{X}_n$ étant immédiate. □

Pour se convaincre de l'utilité théorique de ce théorème, nous allons voir comment nous en servir pour construire des tests d'hypothèse sur des variables gaussiennes dans la section qui suit.

5.3 Test d'égalités de paramètres dans un modèle gaussien

On considère $(X_1, ..., X_{n_1})$ un échantillon de taille n_1 d'une loi normale $\mathcal{N}(m_1, \sigma_1^2)$ et $(Y_1, ..., Y_{n_2})$ un échantillon de taille n_2 d'une loi normale $\mathcal{N}(m_2, \sigma_2^2)$. On suppose de plus que (X_i, Y_j) sont indépendants dans leur ensemble. Les paramètres $m_1, m_{2,\,1,\,2}$ sont inconnus, mais nous souhaitons tester la "ressemblance" de la loi de X avec celle de Y. Pour cela, nous commencerons par tester l'égalité des variances.

5.3.1 Test d'égalité des variances

L'on veut donc tester

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad \text{contre} \qquad H_1 : \sigma_1^2 \neq \sigma_2^2.$$

Comme toujours, pour construire un test, il faut choisir une statistique de test. Le corollaire du théorème de Cochran nous donne une idée. En effet, si l'on pose

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \overline{X})^2 \qquad \text{avec } \overline{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i$$

et

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_j - \overline{X})^2 \qquad \text{avec } \overline{X} = \frac{1}{n_2} \sum_{j=1}^{n_2} Y_j,$$

alors l'on aura pour $k \in \{1, 2\}$ que

$$\frac{1}{\sigma_k} S_k^2 \sim \chi^2_{n_k - 1}.$$

L'on choisit donc comme statistique de test, puisque ces deux variables sont par hypothèse indépendante, la variable

$$Stat = \frac{S_1^2}{S_2^2}.$$

En effet, $\frac{\sigma_2^2}{\sigma_1^2} Stat$ a la même loi que le rapport de deux lois du χ^2 avec respectivement $n_1 - 1$ et $n_2 - 1$ degrés de liberté. Rappelons qu'il s'agit d'une loi tabulée, appelé loi de Fisher-Snedecor, et noté $\mathcal{F}_{n_1-1, n_2-1}$. Sous H_0, le rapport des variances disparait car égal à 1.
Notons $\mathcal{F}_{n,p;\alpha}$ le quantile d'ordre α de la loi $\mathcal{F}_{n,p}$. Pour construire le test, il ne nous reste plus qu'à choisir une zone de rejet.
Nous rejetterons H_0 si

$$Stat > \mathcal{F}_{n_1-1, n_2-1; 1-\alpha} \quad \text{ou bien si} \quad Stat < \mathcal{F}_{n_1-1, n_2-1; \alpha}.$$

Nous admettrons que ce test est le test le plus puissant parmi tous les tests de niveau α (test uniformément plus puissant U.P.P. ou U.M.P en anglais).
Une critique principale de ce test est sa sensibilité à la non-normalité (l'on dit que le test n'est pas robuste). C'est pour cela que les statisticiens ont développé des alternatives comme le test de Bartlett ou le test de Levene, que nous ne ferons que citer ici.

5.3.2 Test d'égalité des moyennes

On considère à présent que l'on sait que les variances des deux échantillons sont égales (par exemple en effectuant le test précédent), et l'on veut alors tester lorsque $\sigma_1^2 = \sigma_2^2 = \sigma^2$ l'hypothèse

$$H_0 : m_1 = m_2 \qquad \text{contre} \qquad H_1 : m_1 \neq m_2.$$

Pour choisir la statistique de test, le théorème de Cochran vient encore à la rescousse. En effet, l'on a alors

$$\frac{\overline{X} - \overline{Y}}{\sigma} \sim \mathcal{N}(m_1 - m_2, \frac{1}{n_1} + \frac{1}{n_2})$$

$$Z = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2} \text{ comme somme de deux variables indépendante de loi du } \chi^2,$$

et comme les moyennes empiriques sont indépendantes des variances empiriques, l'on a la loi de la statistique :

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cdot \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}} = \frac{\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\sqrt{\frac{Z}{n_1+n_2-2}}} \sim \text{ Student à } (n_1 + n_2 - 1) \text{ d.d.l.}$$

L'on choisit donc ce rapport comme statistique de test et construit les zones de rejet directement avec la fonction de répartition de la loi de Student. Nous admettrons que si l'on prend la zone de rejet symétrique, ce test est le test le plus puissant parmi tous les tests de niveau α.

5.4 Estimation par régressions linéaires avec plusieurs variables, et introductions au test d'hypothèses linéaires

Donnons-nous des variables aléatoires $(Y_k)_{k\in\{1,\dots,l\}}$ qui admettent comme espérance mathématique une certaine fonction $h(\theta)$ connue. Estimer θ par la méthode des moindres carrés, c'est prendre comme estimateur $\hat{\theta}_n$ un minimiseur de l'erreur quadratique :

$$\hat{\theta}_n = arg\min_{\theta} \sum_{k=1}^{l} (y_k - h_k(\theta))^2.$$

Le problème le plus simple imaginable est celui de la régression gaussienne avec design fixe et bruit homoschédastique.

5.4.1 Problème des moindres carrés

Ce problème est relativement simple. Il cherche à expliquer Y par J régresseurs ou variables explicatives (X_j), à un bruit près, avec une relation linéaire de la forme :

$$Y = a_0 + < \mathbf{X}, \theta > +\epsilon$$

le design fixe fait référence à $< \mathbf{X}, \theta >$, et l'hypothèse de bruit homoschédastique est l'hypothèse que le bruit est une gaussienne centré de variance fixé σ^2.

On observe alors les valeurs de Y et des X_j pendants n répétition indépendantes de l'expérience (toutes ces variables sont mesurées sur n individus, et l'on demande bien entendu que $n > J$) et l'on récolte ainsi un ensemble de $n \times (J+1)$ valeurs observées, donc connues :

$$\begin{array}{cccc} y_1 & x_{1,1} & \cdots & x_{1,J} \\ y_2 & x_{2,1} & \cdots & x_{2,J} \\ y_3 & x_{3,1} & \cdots & x_{3,J} \\ \vdots & \vdots & & \vdots \\ y_n & x_{n,1} & \cdots & x_{n,J} \end{array}$$

Le but est de retrouver $\theta := \begin{pmatrix} a_1 \\ \vdots \\ a_J \end{pmatrix}$ tel que

$$\forall k \in [\![1, n]\!], Y_k = a_0 + x_{k,1}a_1 + x_{k,2}a_2 + ... + x_{k,J}a_J + \epsilon_k$$

avec les ϵ_k des variables aléatoires réelles indépendantes.
En posant

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,J} \\ 1 & x_{2,1} & \cdots & x_{2,J} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,J} \end{pmatrix}, \quad \theta = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_J \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

les hypothèses de modélisation se réécrivent de manière matricielle comme

$$\mathbf{Y} = \mathbf{X} \times \theta + \mathbf{U}.$$

Le problème des moindres carrés devient alors de trouver

$$\hat{\theta}_n = arg\min_{\theta} \sum_{k=1}^{n} (y_k - a_0 - a_1 x_{i,1} - ... - a_J x_{i,J})^2 = arg\min_{\theta} ||Y - X\theta||^2_{\mathbb{R}^n}$$

C'est-à-dire que l'on cherche le meilleur hyperplan permettant d'expliquer Y par les X_i.
Nous allons prouver que :

Proposition 65 :

Supposons que X est de rang plein (*i.e.* X est de rang J + 1, ou encore que les vecteurs colonnes ne sont pas colinéaires).
L'estimateur $\hat{\theta}$ des moindres carrés vérifie :

$$\hat{\theta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

et c'est un estimateur sans biais si $\mathbb{E}_\theta(U) = 0$.

Démonstration.
Comme $\mathbb{R}^n$ est un espace euclidien, nous avons la relation de Pythagore :

$$||Y - X\theta||^2_{\mathbb{R}^n} = ||Y - \hat{Y}||^2_{\mathbb{R}^n} + ||\hat{Y} - X\theta||^2_{\mathbb{R}^n}$$

avec $\hat{Y}$ la projection orthogonale de Y sur l'espace vectoriel engendré par les colonnes de X.
Comme X est de plein rang, la matrice X^tX est définie positive, donc inversible. Notons H la matrice (appelé *hat matrix*, ou *matrice d'influence*) la matrice :

$$H = X(X^tX)^{-1}X^t$$

et remarquons que cette matrice laisse invariant les colonnes de X (puisque $HX = X$) et tout vecteur V orthogonal aux colonnes de X sont dans le noyau de H (car $X^tV = 0$). Donc H est le projecteur orthogonal sur le sous-espace engendré par les colonnes de X.

En particulier,

$$\hat{Y} = HY = X(X^tX)^{-1}X^tY$$

Donc puisque X est de plein rang, il existe un unique minimiseur θ (car il doit vérifier $X\theta = \hat{Y}$) :

$$\hat{\theta} = (X^tX)^{-1}X^tY.$$

Enfin, pour le biais, comme X^tX est inversible, nous avons que

$$\begin{aligned}\hat{\theta} - \theta_0 = \hat{\theta} - \left((X^tX)^{-1}X^tX\right)\theta_0) &= (X^tX)^{-1}X^tY - (X^tX)^{-1}X^t(Y-U) \\ &= (X^tX)^{-1}X^t\,(U)\end{aligned}$$

Donc par linéarité,

$$\mathbb{E}(\hat{\theta} - \theta_0) = (X^tX)^{-1}X^t\,(\mathbb{E}(U)) = 0$$

□

Remarque : Il est également possible de trouver ce minimum par une étude de différentielle de la quantité à minimiser.

Cas particulier J=1 : Lorsqu'on utilise une seule variable explicative, nous pouvons écrire pour le minimiseur $(\hat{a}, \hat{b})$ au sens quadratique de l'équation $y = ax + b + \epsilon$ avec bruit centré que :

$$\hat{a} = \frac{\sum_{i=1}^n (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

$$\hat{b} = \overline{y} - \hat{a}\overline{x}.$$

Remarque : Lorsque les variables aléatoires d'erreur U_k sont gaussiennes, l'estimateur de maximum de vraisemblance des a_j coïncide avec l'estimateur des moindres carrés de ces coefficients.

5.4.2 Validation de la régression

Lorsqu'on veut s'interesser à la qualité d'une régression linéaire, l'on pourra calculer le coefficient de corrélation multiple.

Définition 66 :

On appelle coefficient de détermination *la quantité*

$$R^2 = \frac{Cov_{y,\hat{y}}}{S_y^2} = \frac{\sum_{i=1}^n (y_i - \overline{y})(\hat{y}_i - \overline{\hat{y}})}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

On appelle coefficient de corrélation multiple *le coefficient R de corrélation linéaire entre la série* $(y_1, y_2, ..., y_n)$ *et la série ajustée* $(\hat{y_1}, \hat{y_2}, ..., \hat{y_n}) = X\hat{\theta}$ *sa racine carré.*

Interprétation : Le coefficient R^2 s'interprète comme la proportion de variabilité expliquée par la régression. Géométriquement, R représente la valeur du cosinus de l'angle formé dans $\mathbb{R}^n$ entre $y - \overline{y}$ et $\hat{y} - \overline{y}$ où $\overline{y}$ désigne le vecteur de $\mathbb{R}^n$ dont toutes les coordonnées sont égales à $\overline{y}$. Le coefficient R^2 est donc compris entre 0 et 1. Plus il est proche de 1 et meilleure est la régression.

Nous allons voir une deuxième expression du coefficient de détermination qui justifie cette interprétation et facilite parfois son calcul :

Proposition 67 :

Le coefficient de détermination R^2 se réécrit comme :

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

Démonstration. Comme les numérateurs sont les mêmes, considérons uniquement les numérateurs. On peut écrire :

$$\begin{aligned}\sum_{i=1}^n (y_i - \overline{y})(\hat{y}_i - \overline{\hat{y}}) &= \sum_{i=1}^n (y_i - \overline{y})(\hat{y}_i - y_i + y_i - \overline{\hat{y}}) \\ &= \sum_{i=1}^n (y_i - \overline{y})(\hat{y}_i - y_i) + \sum_{i=1}^n (y_i - \overline{y})(y_i - \overline{\hat{y}}) \\ &= \left[\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - y_i) + \sum_{i=1}^n (\hat{y}_i - \overline{y})(\hat{y}_i - y_i)\right] + \left[\sum_{i=1}^n (y_i - \overline{y})(y_i - \overline{y}) + (\overline{y} - \overline{\hat{y}}) \sum_{i=1}^n (y_i - \overline{y})\right]\end{aligned}$$

Le vecteur $\hat{y}$ est la projection du vecteur y sur l'espace engendré par les colonnes de X, et le vecteur $\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ appartient à cet espace. Par définition de la projection orthogonale, $\hat{y} - y$ est orthogonale à l'espace engendré par les colonnes de X, et donc en particulier au vecteur $\hat{y} - \overline{y} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$. C'est-à-dire que

$$\sum_{i=1}^n (\hat{y}_i - \overline{y})(\hat{y}_i - y_i) = 0$$

Enfin, par définition de la moyenne empirique,

$$(\overline{y} - \overline{\hat{y}}) \sum_{i=1}^n (y_i - \overline{y}) = 0$$

Donc nous avons bien

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

□

Réecriture de la variance d'un échantillon :

Avec le même argument que dans la preuve $\hat{y} - y$ et $\hat{y} - \overline{y}$ sont orthogonaux. En particulier, l'on peut utiliser le théorème de Pythagore pour écrire :

$$||y - \overline{y}||^2 = ||\hat{y} - y||^2 + ||\hat{y} - \overline{y}||^2$$

ce qui se développe en :

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

Définition 68 :

On appelle somme des carrés d'écarts totale *d'un échantillon, noté* ***ST*** *ou SST en anglais (pour (total sum of squares), la variance (biaisé) de l'échantillon :*

$$\sum_{i=1}^{n}(y_i - \overline{y})^2$$

On appelle somme des carrés des erreurs *(sous entendu de la régression linéaire) d'un échantillon, noté* ***SE*** *ou SSE en anglais (sum of squared errors), l'écart quadratique de la régression linéaire à l'échantillon :*

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

On appelle somme des carrés de la régression *d'un échantillon, noté* ***SR*** *ou SSR en aglais (regression sum of squares) les carrés de la régression de l'échantillon recentré :*

$$\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$$

L'expression précédente se réécrit alors

$$ST{=}SE + SR$$

Remarque : Avec ces notations, $R^2 = \frac{SR}{ST}$ Cette réécriture va nous permettre de formuler un test sur le caractère significatif de la régression.

5.4.3 Test d'hypothèse linéaire avec bruit gaussien

Dans cette partie, nous allons supposer que le bruit (la partie aléatoire du modèle) est gaussien, centré et de variance σ^2, c'est-à-dire que $U \sim \mathcal{N}(0, \sigma^2 I_n)$ ou encore que $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ et sont indépendants.
Résumons la situation de la section précédente

Type de variation	Variable correspondante	Somme des carrés	Dimension de l'espace de la variable/ Degrés de libertés	Carrés moyens
De la régression	$\hat{y} - \overline{y}$	SR	J	$\frac{SR}{J}$
Erreur dans la régression	$y - \hat{y}$	SE	n-J-1	$\frac{SR}{n-J-1}$
Totale	$y - \overline{y}$	ST	n-1	

Nous voulons alors tester l'hypothèse

$$H_0 : (a_1 = a_2 = \cdot \cdot \cdot = a_J = 0),$$

c'est-à-dire que la régression linéaire n'est pas significative.

Or, sous H_0, nous avons que $y = U \sim \mathcal{N}(0, \sigma^2 I_n)$ est une dilatation d'une gaussienne standard. Comme $\hat{y}$ est la projection orthogonale de y, nous allons pouvoir appliquer le théorème de Cochran :

Proposition 69 :

Sous l'hypothèse H_0, $\frac{SR}{J}$ et $\frac{SE}{N-J-1}$ sont indépendantes et de loi respective χ^2_J et χ^2_{n-J-1}.
En particulier,

$$\frac{\frac{SR}{J}}{\frac{SE}{n-J-1}}$$

suit une loi de Fisher-Snedecor à J et n − J − 1 degrés de liberté.
Le test qui rejette H0 si ce quotient dépasse le quantile d'ordre $1-\alpha$ de la loi de Fisher-Snedecor à J et n − J − 1 degrés de liberté est un test de risque α.

Démonstration. Il s'agit d'une application directe du théorème de Cochran. □

5.5 Exercices

Une série d'exercices, on trouvera des pistes de correction à la page 98.

Exercice 16 :

Le but de cet exercice est de donner dans un exemple le cheminement de la démonstration du corollaire du théorème de Cochran.
On considère un vecteur aléatoire à valeurs dans $\mathbb{R}^3$, de composantes X_1, X_2 et X_3 dans la base canonique. On suppose que X_1, X_2 et X_3 sont des variables aléatoires réelles indépendantes de loi gaussienne $\mathcal{N}(0,1)$.

- Quelle est la densité du triplet (X_1, X_2, X_3) ?
  Soit M la matrice de changement de base orthonormée suivante :

  $$M = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} \end{pmatrix}$$

  Vous pourrez admettre par la suite que $MM^t = I_3$.
- Donner la loi de $Y = M^t X$.
- En déduire la loi des composantes Y_1, Y_2 et Y_3 de Y.
- Montrer qu'avec $\overline{X} = \frac{1}{3}(X_1 + X_2 + X_3)$, l'on a que

  $$\sum_{i=1}^{3} X_i^2 = \sum_{i=1}^{3} Y_i^2$$

  $$Y_2^2 + Y_3^2 = \sum_{i=1}^{3} X_i^2 - 3\overline{X}^2$$

  En déduire que les variables $\overline{X}$ et $S_X^2 = \frac{1}{2}\sum_{i=1}^{3}(X_i - \overline{X})^2$ sont indépendantes.
- Donner la loi de $\overline{X}$ et de $2S_X^2$.

Exercice 17 :

On considère un système électronique, alimenté par une source continu de tension x connue. On observe à la sortie une tension Y. Nous supposerons que Y peut s'écrire de la forme

$$Y = ux + v + A$$

Avec $A \sim \mathcal{N}(0, \sigma^2)$. On désire estimer les valeurs de u et v. Pour cela, on mesure les valeurs y_j de Y pour diverses valeurs x_j de X. Ces valeurs sont "entachées" des erreurs A_j.

- On effectue ces mesures à Yaoundé (Cameroun) par $30°C$. En supposant que les A_j soient issus de tirages indépendants, déduire du tableau suivant les estimations $\hat{U}$ et $\hat{V}$ de u et v.

x_j :	5	7	9	11	13	15	17	19	21
y_j :	6	8	9.6	9.1	10	10.4	13.7	14.1	14.6

- On refait à présent ces mesures au Pôle Nord par $-30°C$. Avec les mêmes hypothèses, déduire les estimations $\hat{U}$ et $\hat{V}$ de u et v. Quel remarque peut-on faire ?

x_j :	2	4	6	8	10	12	14	16
y_j :	8	8	10.5	10.1	12.8	12.8	13.1	14.2

- Donner un estimateur de la variance σ du bruit. Le calculer pour les deux séries de mesure
- Montrer avec un test adapté que l'on peut supposer que les deux séries ont la même variance du bruit (ce qui revient à tester que les mesures ont la même précision) avec un risque de 5%. L'on pourra s'aider des tables du chapitre 2.
- Étudier avec des tests d'égalité de moyenne si les valeurs de u sont les mêmes pour les deux séries de mesures. *Vous pourrez admettre que* $\hat{U} \sim \mathcal{N}(u, \frac{\sigma^2}{n})$
- Étudier avec des tests d'égalité de moyenne si les valeurs de v sont les mêmes pour les deux séries de mesures. *Vous pourrez admettre que* $\hat{V} \sim \mathcal{N}(v, \frac{\sigma^2}{n}(1 + \frac{\overline{x}_i^2}{S_{x_i}^2})$

Exercice 18 : (Régression non paramétrique - base de Fourrier-(HP)* - Exercice et figure tiré des TD du Pr. Guyader)**

Nous nous intéressons à un problème plus difficile : prédire une fonction entière. Nous observons alors les variables aléatoires $Y_i = f(\frac{i}{n}) + \epsilon_i$ pour $i \in [\![1, n]\!]$, somme des valeurs d'une fonction avec un bruit. Nous supposons donc que les ϵ_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$ avec σ connu, et que $f : [0, 1] \to \mathbb{R}$ est une fonction inconnue qui est le paramètre d'intérêt.

Réécriture du modèle

- Quel est la difficulté particulière de ce modèle statistique ?
  Pour résoudre ce problème, on propose de projeter f sur une base de fonction bien choisie. Supposons pour la fin de l'exercice que

$$\forall t \in [0, 1], f(t) = a_0 + \sum_{k=1}^{K} a_k \cos(2k\pi t) + b_k \sin(2k\pi t)$$

Les inconnues deviennent alors $(a_i)_{0\leqslant i\leqslant K}$ et $(b_i)_{1\leqslant i\leqslant K}$.

- Sous cette hypothèse, écrire le modèle comme un modèle linéaire gaussien $Y = X\beta + \epsilon$ et préciser X, β et le nombre de variables explicatives p.

- On suppose à présent $2K+1 \leqslant n$. Vérifier que le modèle est identifiable (que la matrice est de rang plein), et calculer l'estimateur des moindres carrés $\hat{\beta}$ de β.
  En déduire un estimateur $\hat{\mu}$ de $\mu = [f(\frac{i}{n})]_{i\in[\![1,n]\!]}$, et proposer un estimateur $\hat{f}$ de la fonction f.
  On pourra se rappeler les formules de trigonométrie, en notant j une racine carré de -1 :

$$\cos(A)\cos(B) = \frac{1}{2}Re(e^{j(A+B)} + e^{j(A-B)})$$

$$\cos(A)\sin(B) = \frac{1}{2}Im(e^{j(A+B)} - e^{j(A-B)})$$

$$\sin(A)\sin(B) = \frac{1}{2}Re(e^{j(A-B)} - e^{j(A+B)})$$

Overfitting et choix du modèle :

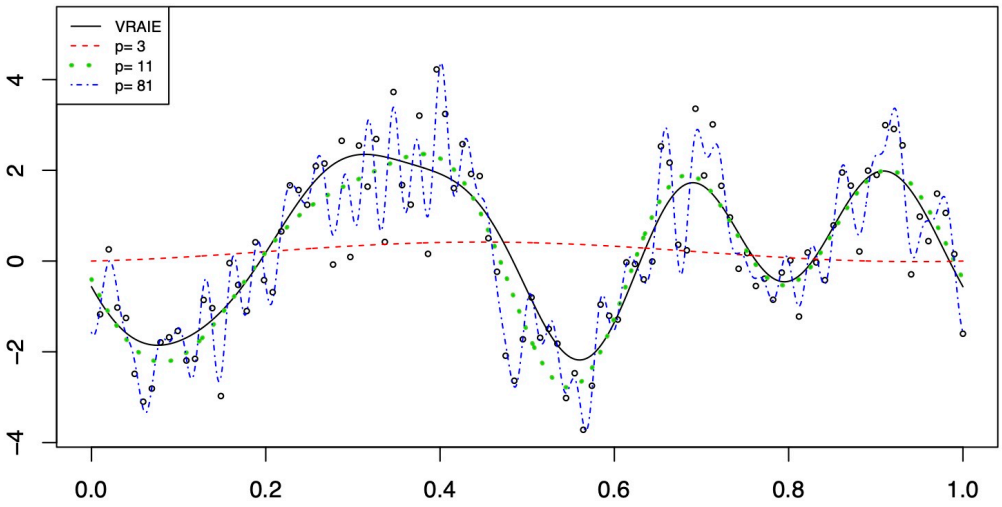
- Calculé la somme des carrés de l'erreur quadratique renormalisé :

$$r_n = \frac{\mathbb{E}(||Y - X\hat{\beta}||^2)}{n}$$

  Pour K fixé, que se passe-t-il quand n tend vers l'infini ?

- On suppose que p=n, donner alors la valeur de r_n. Que peut-on dire des valeurs de $\hat{f}$ aux points $\frac{i}{n}$ pour $i \in [\![1, n]\!]$?

- Nous prenons un polynôme trigonométrique avec coefficients non nuls que pour $k \leqslant 11$ dans la décomposition ci-dessus. Nous faisons n = 101 observations et choisissions diverses valeurs de K. Nous obtenons la figure ci-dessous Quels phénomènes observe-t-on ? Quelle règle proposez-vous au vu des points précédents et de cette observation pour un bon ajustement ?



Chapitre 6

Tests de χ^2 et test de Kolmogorov-Smirnov

Jusqu'à présent, nous avons travaillé dans des expériences statistiques où des informations *a priori* permettaient de réduire l'étude à un ensemble de loi paramétré par un nombre fini de valeurs réelles. Nous allons voir dans cette partie un ensemble de test qui peuvent être mis en œuvre dans un cadre plus général.
On se donne deux séries de n variables aléatoires $(X_1, ..., X_n)$ et $(Y_1, ..., Y_n)$ et l'on suppose que chaque série est constituée de variables indépendantes et identiquement distribuées de loi respective X et Y. On souhaite vérifier si la loi de X est égale à une certaine loi F_0 donnée (donc parfaitement connue), ou si les deux séries suivent la même loi ($X \overset{loi}{=} Y$) ou encore si les séries sont indépendantes.

6.1 Test du χ^2 et tests d'indépendance

Le principe du test du χ^2 est toujours le même. Il s'agit de choisir judicieusement un découpage de $\mathbb{R}$ en diverses classes $C_1, ..., C_l$ et d'utiliser une statistique dépendante des fréquences d'apparition des diverses classes.

6.1.1 Ajustement à une loi connue

Dans cette section, l'on souhaite vérifier que la loi F de X est égale à une certaine loi F_0 donnée (donc parfaitement connue). C'est-à-dire l'on souhaite tester

$$H_0 : F = F_0 \qquad \text{contre} \qquad H_1 : F \neq F_0.$$

Ce genre de test s'appellent des *test d'ajustement*.

Donnons-nous $l \in \mathbb{N}^*$ et une partition de $\mathbb{R}$ en classe $(C_1, ..., C_l)$. L'on se donne Z qui suit la loi F_0 et l'on pose

$$p_k := \mathbb{P}(Z \in C_k).$$

Posons également les variables aléatoires $U_i^{(k)} := \mathbb{1}_{C_k}(X_i)$ et $W^{(k)} := \sum_{i=1}^n U_i^{(k)}$.
Sous H_0, $U_i^{(k)}$ suit une loi de Bernoulli de paramètre p_k et $W^{(k)}$ suit une loi Binomiale de paramètre $\mathcal{B}in(n, p_k)$ (elle représente le nombre de réalisations présente dans la classe C_k).
En particulier, l'on connait son espérance et sa variance : $\mathbb{E}(W^{(k)}) = np_k$ et $V(W^{(k)}) = np_k(1 - p_k)$.
L'on sait alors avec le théorème central limite que

$$\frac{W^{(k)} - np_k}{\sqrt{np_k(1 - p_k)}} \overset{Loi}{\to} \mathcal{N}(0, 1)$$

L'on peut donc raisonnablement supposer que la variable

$$T = \sum_{k=1}^{l} \frac{(W^{(k)} - np_k)^2}{np_k}$$

soit, à des constantes c_k près, une variable aléatoire suivant une loi du χ^2. Comme chacune des variables ne sont pas indépendantes, il faut néanmoins faire attention (en particulier les constantes ne seront pas $\sqrt{p_k(1-p_k)}$). Mais l'on va pouvoir montrer le théorème suivant :

Théoreme 70 :

Sous l'hypothèse H_0, la suite de variables aléatoire suivante converge en loi vers une loi du χ^2 à $l-1$ degré de liberté :

$$T_n = \sum_{k=1}^{l} \frac{(W^{(k)} - np_k)^2}{np_k} \overset{Loi}{\to} \chi^2_{l-1}$$

Démonstration. Nous allons utiliser une généralisation du théorème 54 à la statistique de test du rapport de vraisemblance pour un paramètre dans $\mathbb{R}^n$, qui dit que dans ce cas, le logarithme du maximum du rapport de vraisemblance converge en loi vers une χ^2 à n degré de liberté.
La vraisemblance d'un tirage sous la loi F_0 (avec $p_l = 1 - \sum_{k<l} p_k$) est alors

$$f_{p_1,...,p_l}(n_1,...,n_l) = C \prod_{k=1}^{l} p_k^{n_k}$$

On peut alors facilement montrer que l'estimateur de maximum de vraisemblance des valeurs p_i est $\hat{p_i} = \frac{n_i}{n}$.
La statistique de rapport des maximums de vraisemblance est alors

$$-2\ln(\Lambda_n) := \frac{f_{p_1,...,p_l}(n_1,...,n_l)}{f_{\hat{p_1},...,\hat{p_{l-1}},}(n_1,...,n_l)} = n^n \prod_{k=1}^{l} \left(\frac{p_k}{n_k}\right)^{n_k}$$

D'où comme $-2\ln(\Lambda_n) = \sum_{k=1}^{l} 2n_i \ln(\frac{n_k}{np_k})$, l'on aura avec nous $l-1$ degrés de liberté (puisque $\sum p_k = 1$) que :

$$\sum_{k=1}^{l} 2n_k \ln(\frac{n_k}{np_k}) \overset{loi}{\to} \chi^2_{l-1}.$$

Mais le membre de gauche n'est pas exactement notre statistique. Posons $\Delta_i = \frac{n_i - np_i}{np_i}$ et réécrivons notre expression :

$$\begin{aligned}
-2\ln(\Lambda_n) &= 2\sum_{k=1}^{l} n_k \ln(\frac{n_k}{np_k}), \\
&= 2\sum_{k=1}^{l} (n_k - np_k + np_k)\ln(1+\Delta_k), \\
&= 2\sum_{k=1}^{l} [(n_k - np_k) + np_k)]\,(\Delta_k - \frac{1}{2}\Delta_k^2 + O(\Delta^3), \\
&= 2\sum_{k=1}^{l} (n_k - np_k)\Delta_k + np_k\Delta_k - \frac{np_k}{2}\Delta_k^2 + O(n\Delta^3),
\end{aligned}$$

Or maintenant, comme $\sum_{k=1}^{l} np_k\Delta_k = \sum_{k=1}^{l} n_i - np_i = n - n = 0$ et comme $(n_k - np_k) = np_k\Delta_k$, l'on a que :

$$-2\ln(\Lambda_n) = \sum_{k=1}^{l} np_k\Delta_k^2 + O(n\Delta^3)$$

Enfin, avec la loi forte des grands nombres, $\sqrt{n}\Delta_k$ converge en loi vers une loi normale, donc avec le lemme de Slusky et une composition avec une application continue,

$$\frac{n}{\sqrt{n}^3} \cdot (\sqrt{n}\Delta_k)^3 = \frac{}{\sqrt{n}} \cdot (\sqrt{n}\Delta_k)^3$$

converge en loi vers 0, donc en probabilité vers 0. L'on a bien alors que la statistique T_n à la même loi asymptotique que la statistique de rapport des maximums de vraisemblance.

□

Remarque et preuve alternative : L'on peut remarquer que T_n est en fait la norme du vecteur $(\frac{n_k - np_k}{\sqrt{np_k}})_{k\in\{1,\dots,l\}}$. Or l'on peut montrer que ce vecteur est de matrice de covariance :

$$Cov(\frac{n_k - np_k}{\sqrt{np_k}}) = Id_l - \begin{pmatrix} \sqrt{p_1} \\ \vdots \\ \sqrt{p_l} \end{pmatrix} \times \begin{pmatrix} \sqrt{p_1} & \cdots & \sqrt{p_l} \end{pmatrix} := \Gamma(F_0)$$

Qui est la matrice de projection sur l'orthogonal de l'espace engendré par $\begin{pmatrix} \sqrt{p_1} \\ \vdots \\ \sqrt{p_l} \end{pmatrix}$.
Comme avec le théorème central limite, nous avons la convergence en loi vectorielle

$$\frac{n_k - np_k}{\sqrt{np_k}} \to \mathcal{N}(0, \Gamma(F_0))$$

le théorème de Cochran permet de conclure sur la loi asymptotique de T_n.

Cette loi asymptotique permet encore une fois de créer un test, portant sur notre hypothèse H_0.

Définition 71 :

Soit F_0 une loi, (C_i) une partition de $\mathbb{R}$ en classes, et T défini comme ci-dessus.
Si l'on se donne α un risque de première espèce, et s_α le quantile d'ordre $1-\alpha$ d'une loi du χ^2_{l-1},
Alors le test

On rejette H_0 si $T > s_\alpha$ *On ne rejette pas H_0 si $T \leqslant s_\alpha$*

est un test de niveau asymptotique α.

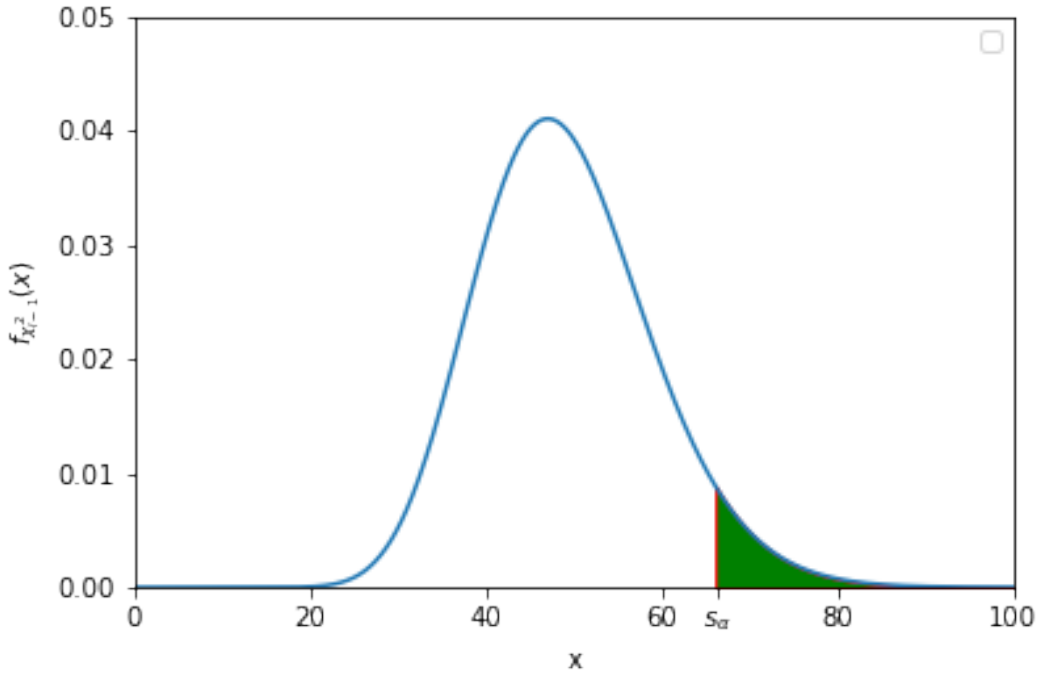


FIGURE 6.1 – Illustration du test du χ^2 avec l = 50 classes

Rappel et Remarque : La densité d'une loi du χ^2_{l-1} à $l-1$ degré de liberté a pour densité

$$f_{\chi^2_{l-1}}(x) = x \mapsto \frac{x^{\frac{1}{2}(k-2)}e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}\mathbb{1}_{\mathbb{R}^+}(x).$$

Lors de calculs explicite, l'on préfèrera se rapporter à une table des valeurs de la fonction de répartition plutôt qu'un calcul d'intégral numérique pour trouver le quantile d'ordre $1-\alpha$.

6.1.2 Comment choisir les classes $C_1, ..., C_l$?

Pour éviter de diviser par des nombres trop petits et que l'un des termes de la somme devienne artificiellement trop important, il est d'usage de demander que le nombre moyen d'observations par classe np_k soit supérieur à 5 pour toutes les classes, *i.e.*

$$\min_{k\in\{1,...,l\}} np_k > 5.$$

Pour faciliter cela, étant donné que n est fixé et que $\sum_{k=1}^{l} p_k = 1$, on a intérêt à ce que les classes soient les plus équiprobables possibles. Pour des variables aléatoires réelles continues, l'on détermine des classes C_k de manière à avoir $p_k = \frac{1}{l}$. Dans le cas des variables aléatoires réelles discrètes, ce n'est pas toujours possible d'avoir de telles classes et l'on cherche alors des classes C_k de manière à avoir $p_k \approx \frac{1}{l}$. Enfin, pour des variables aléatoires qui renvoient un caractère plutôt qu'un nombre (on parle de variable qualitative, comme par exemple la classe socio-professionnelle d'un individu, la marque d'une voiture...), l'on vérifiera la condition $np_k > 5$, et si cela n'est pas le cas, l'on cherchera à regrouper intelligemment des caractères.
Pour des raisons de simplicité, lorsque le test est effectué à la main, on choisit un nombre de classes faible : $l = 3, 4$ ou 5. Dans le cas d'une loi F_0 ayant une symétrique par rapport à sa moyenne, on pourra profiter de la symétrie pour avoir l'équiprobabilité en prenant l pair (par exemple l=4).

6.1.3 Que faire si les paramètres définissant F_0 ne sont pas connus ?

Il arrive bien souvent que seule la forme de la loi F_0 soit connue via la modélisation (loi de Poisson, normale…), mais que l'on ignore certains paramètres. L'on pourra alors estimer à l'aide de l'échantillon ces paramètres.

Proposition 72 : (*admis*)
Soit $m \in \mathbb{N}^*$ un entier et $F_0(\theta_1, \cdots, \theta_m)$ une famille de loi. Si l'on note $(\hat{\theta_i})$ l'estimateur de maximum de vraisemblance des paramètres de la loi, $\hat{p_k} = \mathbb{P}_{F_0(\hat{\theta_1}, \cdots, \hat{\theta_m})}(X \in C_k)$ et n_k le nombre de résultats présent dans la classe C_k, alors

$$T_n := \sum_{k=1}^{l} \frac{(n_k - n\hat{p_k})^2}{n\hat{p_k}} \overset{Loi}{\to} \chi^2_{l-m-1}$$

Démonstration. Voir [MK62] points 30.11 à 30.19.
Une autre preuve possible est d'estimer plus finement le maximum de vraisemblance avec l'information de Fisher et le gradient de la Log-vraisemblance, avant d'étudier le comportement asymptotique de couple $(T_n, \hat{\theta})$ puis de conclure avec le théorème de Cochran. □

Ainsi, on réduira le nombre de degrés de liberté de la loi du χ^2 utilisée pour le test d'ajustement selon le nombre de paramètres estimés m.

6.1.4 Test d'égalité de loi (d'homogénéité) et d'indépendance de variable qualitative

Une première application de la section précédente est la solution de la question d'égalité des lois. L'on considère $(x_1, ..., x_n)$ et $(y_1, ..., y_n)$ deux échantillons d'observation, et l'on se demande si ces échantillons ont été tirés avec la même loi. L'on cherche donc à tester

$$H_0 : F_X = F_Y \qquad \text{contre} \qquad H_1 : F_X \neq F_Y.$$

Notons alors, pour $C_1, ..., C_l$ des classes d'une partition de $\mathbb{R}$, les valeurs

$$\forall k \in \{1, ..., l\} \qquad p_{k,X} = \mathbb{P}(X \in C_k) \quad \text{et} \quad p_{k,Y} = \mathbb{P}(Y \in C_k).$$
$$n_{k,x} = Card(x_i \in C_k) \quad \text{et} \quad n_{k,y} = Card(x_i \in C_k).$$

Sous l'hypothèse nulle H_0, $p_{k,X} = p_{k,Y} := p_k$, on peut alors montrer que l'estimateur du maximum de vraisemblance sera :

$$\hat{p_k} = \frac{n_{k,x} + n_{k,y}}{2n}$$

Nous pouvons alors considérer le test du χ^2 avec $2l$ classes et l paramètres estimés :

Proposition 73 :

On pose

$$T_n = \sum_{k=1}^{l} \frac{(n_{k,x} - n_{k,y})^2}{n_{k,x} + n_{k,y}}$$

Si l'on se donne α un risque de première espèce, et s_α le quantile d'ordre $1-\alpha$ d'une loi du χ^2_{l-1},
Alors le test

$$\text{On rejette } H_0 \text{ si } \ T > s_\alpha \qquad \text{On ne rejette pas } H_0 \text{ si } \ T \leqslant s_\alpha$$

est un test de niveau asymptotique α.

Démonstration. Pour montrer cela, il suffit de montrer que $T_n \overset{Loi}{\to} \chi^2_{l-1}$.
Mais ceci est immédiat, puisque

$$\sum_{k=1}^{l} \frac{(n_{k,x} - n_{k,y})^2}{n_{k,x} + n_{k,y}} = \sum_{k=1}^{l} \frac{(n_{k,x} - n\frac{n_{k,x}+n_{k,y}}{2n})^2}{n\frac{n_{k,x}+n_{k,y}}{2n}} + \sum_{k=1}^{l} \frac{(n_{k,y} - n\frac{n_{k,x}+n_{k,y}}{2n})^2}{n\frac{n_{k,x}+n_{k,y}}{2n}}$$

qui est exactement la statistique de test du χ^2 avec $2l$ classes (observation de x ou de y et classe d'appartenance) et l paramètre estimé (probabilité des diverses classes). □

L'on peut également appliquer la section précédente à la question de l'indépendance. L'on considère $(x_1, ..., x_n)$ et $(y_1, ..., y_n)$ deux échantillons d'observation, et l'on se demande si ces échantillons ont été tirés avec des lois indépendantes. On note Ω_1 (de cardinal l_1) l'ensemble des classes pour X et Ω_2 (de cardinale l_2) l'ensemble des classes pour Y. L'hypothèse d'indépendance peut se réécrire comme le fait que la loi jointe est le produit des lois marginales. Il faut donc trouver les lois marginales avec un estimateur du maximum de vraisemblance, et il y aura $(l_1 - 1)$ et $(l_2 - 1)$ possibilité respectivement pour chaque loi marginales, soit $l_1 + l_2 - 2$ paramètres à déterminer.
Remarquons que sur $\Omega_1 \times \Omega_2$ il y a $l_1 \times l_2 - 1$ lois possibles.

L'on cherche à tester

$$H_0 : F_X \text{ et } F_Y \text{ sont indépendants} \qquad \text{contre} \qquad H_1 : F_X \text{ et } F_Y \text{ ne sont pas indépendants.}$$

Notons alors, pour $C_{1,x}, ..., C_{l_1,x}, C_{1,y}, ..., C_{l_2,y}$ les classes respectives des deux échantillons. Alors pour toute paire d'indice (i,k), on pose

$$p_{i,k} = \mathbb{P}(X \in C_{i,x}, Y \in C_{k,y}) \quad p_{i,\cdot} = \mathbb{P}(X \in C_{i,x}) \qquad p_{\cdot,k} = \mathbb{P}(Y \in C_{k,y}),$$

$$n_{i,k} = Card((x_m, y_n) \in C_{i,x} \times C_{k,y}) \quad n_{i,\cdot} = Card(x_m \in C_{i,x}) = \sum_{k=1}^{l_2} n_{i,k} \qquad n_{i,\cdot} = Card(x_m \in C_{i,x}) = \sum_{i=1}^{l_1} n_{i,k}.$$

Sous l'hypothèse nulle H_0, $p_{i,\cdot} \cdot p_{\cdot,k} = p_{i,k}$ et l'on peut alors montrer que l'estimateur du maximum de vraisemblance sera bien les valeurs

$$\hat{p_{i,k}} = \frac{n_{i,\cdot}}{n} \times \frac{n_{\cdot,k}}{n}$$

Nous pouvons alors considérer le test du χ^2 avec $l_1 \times l_2 - 1$ classes et $l_1 + l_2 - 2$ paramètres estimés :

Proposition 74 :

On pose

$$T_n = \sum_{i=1}^{l_1} \sum_{k=1}^{l_2} \frac{(n_{i,k} - \frac{n_{i,.} n_{.,k}}{N})^2}{\frac{n_{i,.} n_{.,k}}{N}}$$

Si l'on se donne α un risque de première espèce, et s_α le quantile d'ordre $1-\alpha$ d'une loi du $\chi^2_{(l_1-1)\times(l_2-1)}$, Alors le test

$$\text{On rejette } H_0 \text{ si } T > s_\alpha \qquad \text{On ne rejette pas } H_0 \text{ si } T \leqslant s_\alpha$$

est un test de niveau asymptotique α.

Démonstration. Pour montrer cela, il suffit de montrer que $T_n \overset{Loi}{\to} \chi^2_{(l_1-1)\times(l_2-1)}$.
Mais ceci est immédiat, puisque la statistique T_n est exactement la statistique de test du χ^2 avec $l_1 \times l_2 - 1$ classes et $l_1 + l_2 - 2$ paramètres estimés (probabilités des diverses classes pour les marginales). Soit bien un total de $l_1 \times l_2 - 1 - (l_1 + l_2 - 2) = (l_1 - 1) \times (l_2 - 1)$ degré de liberté. □

Attention à ne pas trop utiliser ce test, qui n'est performant que pour des échantillons de grande taille

6.2 Test d'adéquation à une loi de Kolmogorov

Pour de petits nombres d'observation et pour des lois F_0 continues, l'on pourra préférer utiliser le test de Kolmogorov, qui utilise toute l'information des observations à l'exception de l'ordre dans lesquelles elles ont été obtenues. Pour cela, ce test utilise la fonction de répartition empirique.
Rappel : La fonction de répartition empirique est la variable aléatoire à image fonctionnelle suivante :

$$\hat{F_n} : t \mapsto \frac{Card\{k \in [\![1, n]\!], X_k \leqslant t\}}{n} = \frac{1}{n} \sum_{k=1}^{n} 1_{]-\infty,t]}(X_k).$$

Prenons F la fonction de répartition de nos variables aléatoires indépendantes (X_i), l'on montre en exercice que si F est continue, alors la statistique

$$D_n = \sup_{t \in \mathbb{R}} |\hat{F_n}(t) - F(t)|$$

est égale en loi à une variable aléatoire qui ne dépend que de la loi uniforme et de n. Il est alors possible d'étudier cette deuxième variable aléatoire, de la tabuler ou de donner des comportements asymptotiques. Par exemple, l'on peut montrer que :

Proposition 75 : (admis)

La statistique D_n vérifie que pour tout $\mu > 0$,

$$\mathbb{P}(D_n > \mu) \sim 2 \sum_{k=0}^{\infty} (-1)^{k-1} \exp(-2k^2\mu^2 n)$$

Remarque : Plutôt que de chercher dans les tables, il est possible lors de la recherche de $s_\alpha(n)$, quantile d'ordre $1-\alpha$ de la distribution de D_n, d'utiliser que pour $n > 80$, $s_{0.05}(n) \sim \frac{1.3581}{\sqrt{n}}$ et $s_{0.01}(n) \sim \frac{1.6276}{\sqrt{n}}$.

Comme la loi de D_n peut-être tabulé, nous pouvons alors construire un test d'ajustement (l'hypothèse nulle est $H_0 : F_X = F$) à partir des quantiles d'ordre $1-\alpha$ de la distribution de D_n noté $s_\alpha(n)$. On prend comme règle de décision

$$\begin{cases} \text{Si } d_n(x_1,...,x_n) > s_\alpha(n) & \text{On rejette } H_0 \\ \text{Si } d_n(x_1,...,x_n) \leqslant s_\alpha(n) & \text{On ne peut pas rejetter } H_0. \end{cases}$$

Remarque : Pour calculer la statistique d_n, il suffit de remarquer que F est croissant et que $\hat{F}_n$ ne change qu'en les points x_i. Ce point va nous permettre de simplifier la recherche du supremum. Notons $(x_1^*,...,x_n^*)$ les observations rangées par ordre croissant (on parle de la statistique de l'ordre), et supposons pour simplifier qu'elles soient toutes distinctes, c'est-à-dire que $\forall i \in \{1,...,n-1\}, x_i^* < x_{i+1}^*$ (ce qui est presque-sûr puisque X est supposé continu). On a alors la réécriture

$$d_n = \sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| = \sup_{i\in\{1,...,n\}} \left(\max[|\frac{i-1}{n} - F(x_i^*)|, |\frac{i}{n} - F(x_i^*)|] \right).$$

Dit autrement, il suffit de regarder l'écart au niveau des observations, comme illustré dans la figure 6.2 :

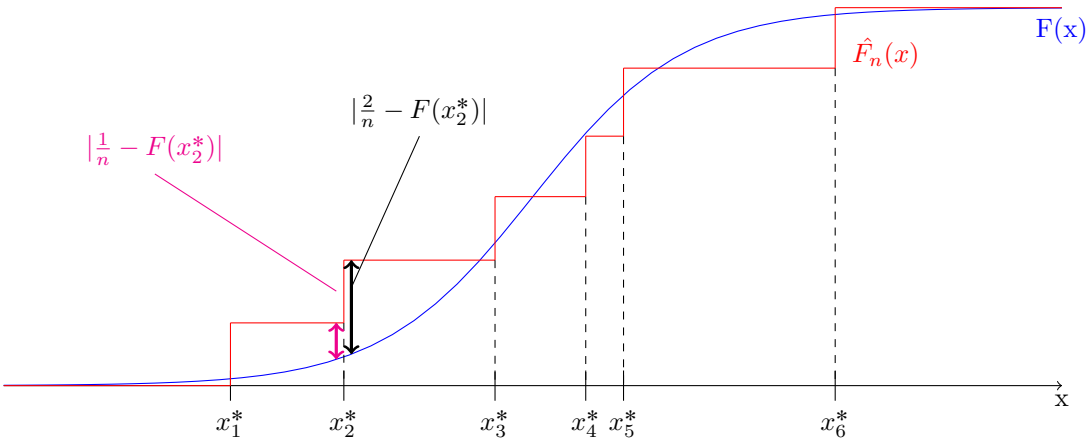


FIGURE 6.2 – Illustration de la méthode de calcul de la statistique D_n

Nous allons voir dans la section suivante comment ce test peut être légèrement modifié pour vérifier l'homogénéité de deux séries d'observation.

6.3 Test d'égalité de loi de Kolmogorov-Smirnov

Maintenant, supposons que l'on ne connaisse plus la loi de la variable d'intérêt, mais cherchons plutôt à tester l'égalité des lois de deux séries d'observations. Nous pourrons suivre la procédure suivante :
Considérons deux échantillons de tailles n et m respectivement :
$(X_1,...,X_n)$ des variables aléatoires indépendantes de loi F_1
$(Y_1,...,Y_n)$ des variables aléatoires indépendantes de loi F_2.
Nous cherchons à tester l'hypothèse

$$H_0 : F_1 = F_2 \text{ et il s'agit d'une loi continue.}$$

contre l'alternative générale.

Soient $\hat{F_1}$ et $\hat{F_2}$ les fonctions de répartition empirique des deux échantillons. Inspiré par le test de Kolmogorov, nous posons comme statistique du test :

$$\Delta := \sup_{t \in \mathbb{R}} |\hat{F_1}(t) - \hat{F_2}(t)|.$$

Sous H_0, la variable aléatoire Δ suit la loi de Kolmogorov-Smirnov, dont on utilise la table afin de déterminer s_α, le quantile d'ordre $1 - \alpha$, qui vérifie :

$$\alpha = \mathbb{P}(\Delta > s_\alpha | H_0)$$

On rejette alors H_0 si la valeur observée δ de la statistique satisfait $\delta > s_\alpha$ et on ne rejette pas H_0 lorsque $\delta \leqslant s_\alpha$.

6.4 Exercices

Une série d'exercice dont on trouvera des corrections page 104

Exercice 19 :

Soit (X_i) une suite de variables aléatoires indépendante et de même loi que la variable X. L'on note $F : t \mapsto \mathbb{P}(X \leqslant t)$ la fonction de répartition de la loi et $\hat{F_n}$ la fonction de répartition empirique de l'échantillon.
Montrer que la loi de la variable

$$\sup_{t\in\mathbb{R}} |\hat{F_n}(t) - F(t)|$$

ne dépend pas de la loi de X lorsque F est continu.
L'on pourra se servir de l'inverse généralisé de la fonction de répartition

Exercice 20 :

Au cours d'une année, les sites de productions d'une entreprise d'électronique de pointe doivent renouveler certains des outils d'usinages nécessaires à la production d'une entreprise. L'hypothèse de modélisation la plus simple est de dire que le nombre d'outils à remplacer dans un site de production suit une loi de Poisson, de paramètre λ inconnu. Au cours de l'année passée, l'entreprise a compté le nombre d'outils remplacé dans ses divers sites. Les résultats sont présentés dans le tableau suivant :

Nombre d'outils renouvelés	0	1	2	3	4	5	6	7	8	9
Nombre de sites	1	2	2	3	3	5	5	2	3	2

Nous admettrons que l'estimateur de maximum de vraisemblance pour le paramètre des lois de Poisson est la moyenne empirique.
L'entreprise souhaite vérifier que les données (x_i) sont bien issues d'un échantillon de loi de Poisson de paramètre inconnus.

L'on donne à l'aide de la fonction Malab, dont la documentation décrit "$POISSCDF(X, Lambda)$ compute the Poisson cumulative distribution function with parameter $Lambda$ at the value X". poisscdf([0 1 2 3 4 5 6 7 8 9 10 11], 3) = [0.0498, 0.1991, 0.4232, 0.6472, 0.8153, 0.9161, 0.9665, 0.9881, 0.9962, 0.9989, 0.9997, 0.9999]
poisscdf([0 1 2 3 4 5 6 7 8 9 10 11], 4)=[0.0183, 0.0916, 0.2381, 0.4335, 0.6288, 0.7851, 0.8893, 0.9489, 0.9786, 0.9919, 0.9972, 0.9991]
poisscdf([0 1 2 3 4 5 6 7 8 9 10 11], 5)=[0.0067, 0.0404, 0.1247, 0.265 , 0.4405, 0.616 , 0.7622, 0.8666, 0.9319, 0.9682, 0.9863, 0.9945] poisscdf([0 1 2 3 4 5 6 7 8 9 10 11], 6)=[0.0025, 0.0174, 0.062 , 0.1512, 0.2851, 0.4457, 0.6063, 0.744 , 0.8472, 0.9161, 0.9574, 0.9799]

- L'entreprise souhaite effectuer un test du χ^2 avec un risque $\alpha = 0.05$. Calculez la statistique d'un tel test en répartissant les données dans trois classes judicieusement choisies. Combien de degré de liberté aura-t-on ? En s'aidant de la table à la page 23, donnez la région d'acceptation. Conclure.

- Effectuer à present un test de Kolmogorov. Est-ce que votre conclusion évolue ?

Chapitre 7

A propos de la parcimonie

Lorsqu'on modélise un phénomène de la réalité, l'on cherche souvent à répondre à l'un des objectifs suivants :

- **Description**, on recherche de façon exploratoire les liaisons entre une variable d'intérêt et d'autres variables potentiellement explicatives.
- **Explication**, on souhaite confirmer ou affiner une connaissance a priori du phénomène par l'estimation des paramètres et des tests appropriés.
- **Prédiction**, on souhaite exploiter le modèle pour prévoir des valeurs de la variable d'intérêt à partir de valeurs de variables explicatives. L'accent est alors mis sur la qualité des estimateurs et des variables explicatives, judicieusement sélectionnées.

Mais de manière bien concrète, si le modèle que l'on utilise pour prédire un évènement met trop de temps à délivrer le résultat par des moyens de calculs à notre disposition, alors le modèle est inutile, voire néfaste (des unités de temps, de calculs et d'énergie aurait pu être investie ailleurs).
La *parcimonie* est justement cet équilibre désirable entre précision du modèle et coûts du modèle (temps de calcul, nombre de variables nécessaires ...).

Pour des méthodes d'estimations parcimonieuses que nous ne regarderons pas ici, l'on pourra s'intéresser à :
- [Liu09] pour l'aspect pratique de l'estimation de quantité sous forme d'espérance avec la méthode de Monte-Carlo ;
- [LPW06] pour une deuxième méthode pratique d'estimation avec des chaines de Markov, et l'étude de leurs propriétés ;

Algorithmes de sélection de paramètres pour la régression linéaire

Rappelons que la régression linéaire, étudié dans 5.4, consiste à trouver la "meilleure" (au sens de l'erreur quadratique) estimation affine des observations $(y_i)_{i\in[\![1,n]\!]}$ à partir des réalisations des paramètres explicatifs $(x_{i,j})_{(i,j)\in[\![1,n]\!]\times[\![1,J]\!]}$, c'est-à-dire à trouver des paramètres $(a_j)_{i\in[\![0,J]\!]}$ réalisant le minimum de l'erreur quadratique :

$$\begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} = arg\min_{b_j} \left(\sum_{i=1}^{n} (y_i - b_0 - x_{i,1}b_1 - x_{i,2}b_2 - \cdots - x_{i,J}b_J)^2 \right)$$

Mais une fois fait ce calcul d'optimisation, nous pouvons nous demander si toutes les variables d'explications était vraiment nécessaires. Si nous sommes un peu plus souples dans les erreurs acceptées, pourrions-nous drastiquement diminuer la dimension du modèle (et donc les temps de calculs qui suivront) ? Pire, ne sommes-nous pas en train de faire un sur-apprentissage, valable uniquement pour les données obtenues, mais peu généralisable ?

Pour savoir ce que nous acceptons de sacrifier comme précision au profit de la diminution du nombre de paramètres, il va falloir nous fixer un critère. Pour que celui-ci fonctionne, il faut que ce critère soit grand quand le nombre de dimensions est grand ou quand l'erreur est grande. Nous verrons ici deux critères répondant à ces attentes.

Définition 76 :

Soit $(y_i)_{i\in[\![1,n]\!]}$ une série d'observations, et $(x_{i,j})_{(i,j)\in[\![1,n]\!]\times[\![1,J]\!]}$ la série d'observations correspondantes des J variables explicatives.
Pour $\Omega_r \subset [\![1,J]\!]$ un sous-ensemble des explications de taille $r \leqslant J$, on appelle :

- C_p de Mallows *de Ω_p la quantité :*

$$C_p = (n-r-1)\frac{SE(\Omega_p)}{SE} - (n-2(r+1))$$

- critère d'Akaike *de Ω_p la quantité :*

$$AIC(\Omega_r) = n\ln(\frac{SE(\Omega_p)}{n}) + 2(r+1)$$

où SE est l'erreur quadratique du système entier

$$SE = \min_{(b_j)\in\mathbb{R}^J}\left(\sum_{i=1}^{n}(y_i - b_0 - x_{i,1}b_1 - x_{i,2}b_2 - \cdots - x_{i,J}b_J)^2\right) = \min_{b_j}\left(\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{J}x_{i,j}b_j)^2\right)$$

et $SE(\Omega_r)$ est la somme des carrés d'erreur obtenue en ne retenant que les r variables de Ω_r parmi les J initialement considérées :

$$SE(\omega_r) = \min_{(b_j)\in\mathbb{R}^r}\left(\sum_{i=1}^{n}(y_i - b_0 - x_{i,j_1}b_1 - x_{i,j_2}b_2 - \cdots - x_{i,j_r}b_J)^2\right) = \min_{b_j}\left(\sum_{i=1}^{n}(y_i - b_0 - \sum_{j\in\Omega_r}x_{i,j}b_j)^2\right)$$

Remarque : Dans le cadre d'un modèle avec bruit gaussien centré et de variance connu égale à σId, les deux critères sont équivalant dans les résultats que donnera l'application des algorithmes de sélections. En revanche, ce n'est plus le cas lorsque l'on sort de ce cadre.
L'intuition derrière la construction des algorithmes de sélection est que plus la valeur du critère est petite et plus le modèle sélectionné est parcimonieux tout en restant efficace.

Nous voudrions donc trouver un modèle ne contenat que des variables d'importance. Mais dans le cas général et évidemment le plus courant en pratique, les variables ne sont pas pré-ordonnées par importance. Il existe des algorithmes de sélection global (l'algorithme de Furnival et Wilson dit de "leaps and bound" par exemple, implémenté dans la plupart des librairies de statistique), mais ceux-ci ne sont souvent utilisables que pour un nombre de paramètres inférieur à 15.
Lorsque J est grand, il n'est pas raisonnable de penser explorer les 2^J ensembles de variables explicatives afin de

sélectionner le meilleur au sens de l'un des critères ci-dessus. Des algorithmes d'exploration pas-à-pas (algorithme de descente) existent. Ils se regroupent en trois types :

- **Sélection** *(forward)* L'algorithme commence avec aucune variable. À chaque étape, on cherche à ajouter la variable qui diminue le plus le critère de sélection. L'algorithme s'arrête lorsque toutes les variables sont présentes ou lorsque le critère ne s'améliore plus.

- **Élimination** *(backward)* L'algorithme démarre avec toutes les variables. La variable dont la suppression conduit à la plus petite valeur du critère est alors retirée. La procédure s'arrête lorsque le critère ne décroit plus.

- **Mixte** *(stepwise)* Algorithme mélangeant les deux précédents. Il ajoute une étape d'élimination après chaque étape de sélection, de façon à éliminer d'éventuelles variables devenues inutiles après l'introduction d'une nouvelle variable.

L'analyse en composante principale pour diminuer la dimension (H.P.)

Un autre cas de figure potentiel est si nous ne sommes pas assurés par la modélisation que nos variables explicatives soit bien indépendante et que nous ayons alors fait le meilleur choix possible dans leur description. Il est alors possible de faire une analyse en composante principale pour décrire les directions dans l'espace des variables explicative ayant la plus grande importance dans la variation de l'échantillon. Nous ne décrirons ici que le principe et la méthode numérique de calcul.

L'analyse en composante principale (ou ACP) consiste à trouver une ou plusieurs nouvelles variables explicatives, combinaison linéaire des anciennes, expliquant la plus grande partie de la variance *via* une projection orthogonale. Considérons donc un ensemble de variables $(X_i)_{i\in[\![1,n]\!]}$ et donnons-nous K réalisations de l'échantillon, que nous écrivons de forme matricielle comme :

$$\tilde{M} := \begin{pmatrix} x_{1,1} & \cdots & x_{n,1} \\ \vdots & & \vdots \\ x_{1,K} & \cdots & x_{n,K} \end{pmatrix}$$

Et nous ne conserverons que la version recentrée de l'échantillon :

$$M := \begin{pmatrix} x_{1,1} - \overline{x}_1 & \cdots & x_{n,1} - \overline{x}_n \\ \vdots & & \vdots \\ x_{1,K} - \overline{x}_1 & \cdots & x_{n,K} - \overline{x}_n \end{pmatrix}$$

Si l'on se donne une direction u, les réalisations des projection des variables explicatives (X_i) sur la droite engendrée par u est le vecteur Mu. Comme celui-ci est de moyenne empirique nulle, sa variance empirique est alors

$$(Mu)^t Mu = ||Mu||_2^2$$

Définition 77 :

On appelle première composante principale la direction u_1 maximisant après projection la variance empirique :

$$u_1 := Arg \max_{u\in\mathbb{R}^n, ||u||_2=1} \left(||Mu||_2^2\right)$$

On définit par récurrence une p^{ieme} composante principale comme la direction u_p maximisant après projection la variance empirique restante :

$$u_p := Arg \max_{\substack{u\in\mathbb{R}^n, ||u||_2=1 \\ \forall i\in[\![1,p]\!], <u_i,u>=0}} \left(||Mu||_2^2\right)$$

De manière équivalente, les composantes principales sont les vecteurs propres de la matrice M^tM si l'on a classé les valeurs propres par ordre croissant.

Démonstration. L'équivalence découle du fait que M^tM est une matrice symétrique, et admet donc une diagonalisation en vecteurs propres orthogonaux avec des valeurs propres réelles (voir le théorème spectral dans n'importe quel livre d'analyse matricielle). Notons v_i ces vecteurs propres, de valeurs propres $\mu_1 \geqslant \mu_2 ... \geqslant \mu_n$. Si l'on écrit $u = \sum_{i=p}^n \lambda_i v_i$ avec $\sum_{i=p}^n \lambda_i^2 = 1$, alors on a que :

$$\begin{aligned} ||Mu||_2^2 &= < Mu, Mu > = \sum_{p\leqslant i,j\leqslant n} \lambda_i\lambda_j\mu_i\mu_j < u_i, u_j > \\ &\leqslant \mu_p^2 \sum_{p\leqslant i\leqslant n} \lambda_i^2 \\ &= \mu_p^2 = ||Mv_p||^2 \end{aligned}$$

□

Finalement, la question de l'ACP se ramène à un problème de diagonalisation de la matrice de covariance. Pour faciliter le traitement numérique, il est également possible d'utiliser la décomposition en valeur singulière de la matrice rectangulaire M.
Bien souvent, les données sont corrigées pour être de variance empirique 1 en divisant dans la matrice M les diverses colonnes par leur écart-type empirique (on parle alors d'ACP normée), ce qui ramène le calcul à une diagonalisation de la matrice de corrélation.

Une fois la décomposition en valeur singulière effectuée, la question du nombre de variables à conserver se pose souvent. Un certain nombre de règles empirique existent, basé sur les valeurs propres associés aux facteurs :

- Conserver les facteurs pour lesquels la valeur propre est supérieure à 1 (seuil de Kaiser, souvent considéré comme trop permissif).
- Conserver les facteurs u_p pour lesquels la valeur propre est supérieure à $\sum_{i=p}^n \frac{1}{i}$ (test des bâtons brisés à 5% correspondant à une distribution uniforme de la dispersion sur les axes).
- Conserver les facteurs u_p pour lesquels la valeur propre est supérieure à $1 - 2\sqrt{\frac{p-1}{n-1}}$ (test unilatéral de conformité à 5% de loi asymptotiquement normale).
- Tracer les valeurs propres en fonction de leur numéro d'apparition, et ne garder que les valeurs propres avant la cassure nette de pente.

- Tirer aléatoirement un échantillon de variables aléatoires indépendantes de loi normale centré-réduite, et comparer son ACP à celle des variables renormalisé centrée et réduite

Dans tous les cas, il est pertinent de tracer cette courbe des valeurs propres en fonction de leur numéro d'apparition (appelé éboulis de valeur propre) et le critère correspondant.
Par exemple, si l'on cherche à extraire les groupes de questions pertinentes du test de rentrée de 2021 du parcours 3EA de Toulouse-INP, l'on trouve le graphique suivant :

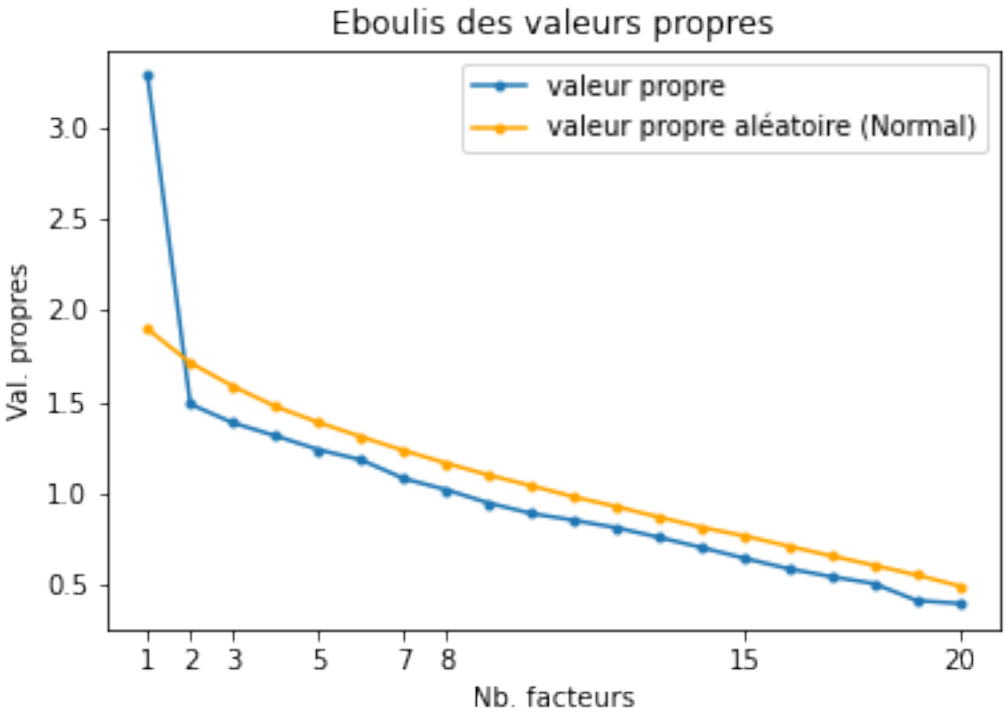


FIGURE 7.1 – Éboulis de valeurs propres sur une ACP

Il faut ensuite savoir combien de recombinaison garder, suivant les critères, l'on gardera que le premier facteur (qui se trouve correspondre à la note totale), les 8 premiers facteurs (test de Kaiser et des bâtons brisés) ou encore tous les facteurs (test de conformité). Au vu du test aléatoire, il semble pertinent de ne conserver que la première composante. Ainsi, il n'y a pas de groupe de question expliquant mieux la dispersion entre individus que la note totale.

Chapitre 8

Solutions et pistes de corrections des exercices

8.1 Premières statistiques

1) Montrer que la fonction de répartition empirique vérifie les propriétés d'une fonction de répartition : limites en $\pm\infty$, continuité à droite et limite à gauche.
Il s'agit de la fonction de répartition d'une variable aléatoire uniforme sur $\{x_1, ..., x_n\}$

2) On suppose que X admette un moment d'ordre 4. Montrer les convergences L^2 et presque sûre de la variance empirique d'un échantillon de réalisation de variables indépendante de même loi que X, et donner la limite. Que dire de la vitesse de convergence ?
Nous notons $m = \mathbb{E}(X)$ et $\sigma^2 = V(X)$.
La loi forte des grands nombres appliquée deux fois nous assure que $S_X^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1} = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^n (X_i - m)^2\right) + \frac{n}{n-1}(\overline{X} - m)^2 \overset{L^2;p.s.}{\to} \sigma^2 + 0$.
Le théorème central limite nous assure que la convergence sera en $\frac{1}{\sqrt{n}}$.

3) Julie habite en face du métro. Elle part de chez elle 25 min avant le début de ses cours à l'INP.
Son temps d'attente du métro suit une loi normale d'espérance 5 et d'écart type 3 et la durée de son trajet ensuite suit une loi normale d'espérance 15 et d'écart type 4. Une fois au pied de l'ENSEEIHT, elle met 1 min à rejoindre sa salle de cours.
On suppose que le temps d'attente du métro et le temps de trajet sont indépendants. Pour répondre aux questions, l'on pourra se servir de la table 2.2.

- Donner la loi du temps total de trajet de Julie.
  Le temps total de trajet X de Julie suit une loi normale, d'espérance 21 et de variance 25.

- Ce matin Julie a un cours de Statistiques, quelle est la probabilité qu'elle arrive à l'heure ? qu'elle ait plus de 5 minutes de retard ?
  On sait que $Z = \frac{X-21}{5}$ suit une loi centrée réduite. En regardant dans la table, l'on trouve $\mathbb{P}(X \leqslant 25) = \mathbb{P}(Z \leqslant \frac{4}{5}) \approx 0.7881$ et $\mathbb{P}(X \geqslant 30) = \mathbb{P}(Z \geqslant \frac{9}{5}) \approx 1 - 0.9641$.

- Sur un trimestre, elle effectue 80 trajets pour venir à la faculté. On note $X_1, X_2, ..., X_{80}$ les 80 variables aléatoires représentants les temps de parcours de Julie pour ces trajets. Quel est le nom et la loi de la variable

aléatoire Y représentant son temps moyen de parcours pour aller de chez elle à l'INP ?
Il s'agit de la moyenne empirique, de loi $\mathcal{N}(21, \frac{5}{\sqrt{80}})$.

- Quelle est la probabilité que, sur un semestre, elle passe plus de 27h30min en trajet pour venir à la ENSEEIHT (temps d'attente et de transport) ?
  Il s'agit de calculer $\mathbb{P}(Z_2 \geqslant \frac{27.5\times 60 - 80\times 21}{5\times\sqrt{80}}) \approx 0.7454$

4) On considère une suite (X_n) de variables aléatoires réelles de Loi de Cauchy (c'est-à-dire de densité $x \mapsto \frac{1}{\pi(1+x^2)}$). Calculer la loi de la moyenne empirique de ces variables.
On va utiliser le fait que, via un calcul laissé au lecteur, la fonction caractéristique d'une loi de Cauchy est

$$t \mapsto e^{-|t|}$$

Alors la fonction caractéristique de la moyenne empirique vérifie pour $t \in \mathbb{R}$:

$$\varphi_{\frac{1}{n}\sum_{i=1}^n X_i}(t) = \varphi_{\sum_{i=1}^n X_i}(\frac{t}{n}) = \prod_{i=1}^n \varphi_{X_i}(\frac{t}{n}) = \left(e^{-|\frac{t}{n}|}\right)^n = e^{-|t|}.$$

On reconnait alors la fonction caractéristique d'une loi de Cauchy !
Que remarque-t-on ?
On remarque qu'en particulier que la moyenne empirique ne converge pas presque surement vers une constante. Ceci n'est pas trop étonnant, puisque la loi de Cauchy n'a pas d'espérance.

5) Écrivez un algorithme sous Matlab pour retrouver les courbes des fonctions de répartition présentées.
Utilisez les méthodes de générations de variables pseudos-aléatoire vu en BE de probabilités pour générer les variables aléatoires. Il s'agit ensuite de tracer les bonnes fonctions.

8.2 Quelques lois usuelles en statistiques

La correction des exercices de la page 28

6) Écrire un programme testant l'adéquation d'observation avec une des lois de ce chapitre. Vous pourrez pour cela comparer les fréquences d'appartenance à des intervalles bien choisis à la probabilité théorique. L'on pourra approximer l'intégrale de la densité par la méthode des carrés à gauche (ou toute autre méthode numérique d'intégration). Si l'on n'a pas trouvé de test, l'on pourra s'inspirer du test de Kolmogorov après lecture du chapitre correspondant.

8.3 Estimateur du maximum de vraisemblance

La correction des exercices de la page 45

7) On considère l'expérience statistique issue de la répétition indépendante de n tirage $(X_1, ..., X_n)$ de variable aléatoire suivant des lois de Poisson de paramètre $\lambda > 0$.

- Montrer que la moyenne empirique $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$ est un estimateur sans biais de λ.
  Par linéarité de l'intégrale,

  $$\mathbb{E}_\lambda(\overline{X_n}) = \mathbb{E}_\lambda(X_1) = \lambda$$

  Donc l'estimateur est sans biais.

- Montrer que la moyenne empirique converge presque sûrement et en moyenne quadratique vers λ. En déduire le caractère convergent de l'estimateur.

  La convergence presque-sûre est la conséquence de la loi forte des grands nombres.
  La convergence L^2 est directe. En effet,

  $$\mathbb{E}_\lambda\left[(\overline{X_n}-\lambda)^2\right] = V(\overline{X_n}) = \frac{V_\lambda(X_1)}{n^2} \to 0$$

  En particulier, l'estimateur moyenne empirique converge en probabilité, et est donc convergent.

- Montrer que la moyenne empirique est asymptotiquement normale, c'est-à-dire que $\sqrt{n}(\overline{X_n}-\lambda)$ converge en loi vers une loi normale, dont on précisera les paramètres.
  Il s'agit d'une application directe du théorème central limite

- Montrer que la variance empirique $S_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X_n})^2$ est un estimateur sans biais, et montrer que $\sqrt{n}\,(S_n-\lambda)$ converge en loi vers une loi normale dont on précisera les paramètres.
  On pourra utiliser que pour une loi de Poisson, le moment centré d'ordre 4 vérifie $\mathbb{E}[(X-\lambda)^4] = \lambda + 3\lambda^2$, *mais également le lemme de Slusky (voir page 57) qui permet de dire que si* Y_n *converge en loi vers* Y *et que* X_n *converge en probabilité vers* X, *alors* X_nY_n *(respectivement la somme) converge en loi vers* XY *(respectivement la somme des limites.*
  Nous avons déjà vu le calcul pour le biais de la variance empirique dans la section 1.2.
  Maintenant, par le théorème central limite, avec la première indication pour calculer la variance de l'estimateur, on trouve :

  $$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n (X_i-\lambda)^2 - \lambda\right) \overset{Loi}{\to} \mathcal{N}(0, \lambda+2\lambda^2) \qquad \sqrt{n}\left(\overline{X_n}-\lambda\right) \overset{Loi}{\to} \mathcal{N}(0,\lambda)$$

  Alors,

  $$\begin{aligned}\sqrt{n}\,(S_n-\lambda) &= \sqrt{n}\left(\frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X_n})^2 - \lambda\right)\\ &= \sqrt{n}\left(\frac{1}{n-1}[\sum_{i=1}^n (X_i-\lambda)^2 - n(\overline{X_n}-\lambda)^2] - \lambda\right)\\ &= \frac{n}{n-1}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n (X_i-\lambda)^2 - \lambda\right) + \frac{n}{n-1}\sqrt{n}\left(\overline{X_n}-\lambda\right)^2 - \frac{\sqrt{n}}{n-1}\lambda\end{aligned}$$

  Maintenant, comme $\overline{X_n}-\lambda$ converge presque surement vers 0 (et donc en particulier converge en probabilité vers 0) et que $\sqrt{n}(\overline{X_n}-\lambda)$ converge en loi, $\frac{n}{n-1}\sqrt{n}\left(\overline{X_n}-\lambda\right)^2$ converge en loi vers 0. Comme la limite est constante, la convergence est même en probabilité.
  Donc, comme d'après le lemme de Slusky $\frac{n}{n-1}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n (X_i-\lambda)^2 - \lambda\right) \overset{Loi}{\to} \mathcal{N}(0,\lambda+2\lambda^2)$, nous obtenons en appliquant encore une fois le lemme que

  $$\sqrt{n}\,(S_n-\lambda) \overset{Loi}{\to} \mathcal{N}(0,\lambda+2\lambda^2)$$

- Quel estimateur privilégier pour avoir de meilleurs résultats asymptotiques ?
  Comme S_n^2 a une plus grande variance asymptotique, l'on privilégiera la moyenne empirique $\overline{X_n}$ afin de diminuer l'écart autour du paramètre d'intérêt.

8) On considère l'expérience statistique issue de la répétition indépendante de n tirage $(X_1, ..., X_n)$ de variable aléatoire suivant des lois exponentielles de paramètre $\lambda > 0$.

- Calculer l'estimateur du maximum de vraisemblance.
  La Log-vraisemblance est la fonction
  $$\ln(f) : (x_1, ..., x_n) \mapsto \sum_{i=1}^{n} \ln(\lambda) - \lambda x_i = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$
  Cette application est bien dérivable en λ, de dérivé
  $$\partial_\theta \ln \circ f(\mathbf{X}, \theta) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i.$$
  Comme celle-ci s'annule pour $\theta = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\overline{X_n}}$, et que la Log-vraisemblance est concave (sa dérivée seconde existe et est négative), l'on a bien que l'estimateur du maximum de vraisemblance est
  $$\frac{1}{\overline{X_n}}$$
  .
- Montrer que la moyenne empirique converge presque sûrement et en moyenne quadratique vers λ. En déduire le caractère convergent de l'estimateur du maximum de vraisemblance.
  Par la loi forte des grands nombres, la moyenne empirique converge presque surement et dans L^2 vers l'espérance, qui vaut $\frac{1}{\lambda}$.
  L'EMV converge presque surement (par les opérations sur les limites réelles) vers λ donc converge en probabilité.

9) On considère l'expérience statistique issue de la répétition indépendante de n tirage $(X_1, ..., X_n)$ de variable aléatoire suivant des lois de Bernoulli de paramètre $0 < p < 1$.

- Calculer l'estimateur du maximum de vraisemblance de p.
  La Log-vraisemblance est la fonction
  $$\ln(f) : (x_1, ..., x_n) \mapsto \ln(\mathbb{P}_p(\forall i, \quad X_i = x_i)) = \ln(p^{\sum x_i}(1-p)^{n - \sum x_i})$$
  Cette application est bien dérivable en p, de dérivé
  $$\partial_\theta \ln \circ f(\mathbf{X}, p) = \frac{1}{p} \sum x_i - \frac{1}{1-p}(n - \sum_{i=1}^{n} x_i).$$
  Cette expression s'annule uniquement pour $p = \overline{X}$. Comme la dérivée seconde de la log-vraisemblance est
  $$-\frac{1}{p^2} \sum x_i - \frac{1}{(1-p)^2}(n - \sum_{i=1}^{n} x_i) < 0,$$
  on a bien trouvé un maximum de la Log-vraisemblance.
- L'estimateur est-il sans biais ? convergent ? efficace ?
  Il est sans biais car $\mathbb{E}_p(\overline{X}) = p$.
  Il est convergent grâce à la loi des grands nombres.
  Pour montrer qu'il est efficace, commençons par calculer l'information de Fisher :
  $$I(p) = -\mathbb{E}_p(\partial_\theta^2 \ln \circ f(\mathbf{X}, p)) = -\mathbb{E}_p\left[-\frac{1}{p^2} \sum x_i - \frac{1}{(1-p)^2}(n - \sum_{i=1}^{n} x_i)\right] = \frac{np}{p^2} + \frac{n(1-p)}{1-p}$$

Donc $I(p) = \frac{n}{p(1-p)}$. En particulier, la borne de Cramer-Rao pour un estimateur de p est $\frac{p(1-p)}{n}$. Comme $V(\overline{\mathbf{X}}) = p(1-p)\frac{n}{n^2}$, l'estimateur est donc efficace.

10) Examen 2004

On considère l'expérience statistique issue de la répétition indépendante de n tirage $(X_1, ..., X_n)$ de variable aléatoire suivant la même loi de densité de paramètre $\theta > 0$ et $\mu \in \mathbb{R}$:

$$\forall x > 0, \qquad f(x) = \frac{1}{\theta x\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\theta^2}}$$

On peut vérifier que $Y_i = \ln(X_i)$ suivent des lois normales $\mathcal{N}(\mu, \theta^2)$ et l'on rappelle que la fonction génératrice des moments d'une loi normale vérifie

$$M_Y(t) = \mathbb{E}(e^{tY}) = e^{\mu t + \frac{\theta^2 t^2}{2}}$$

- Montrer que l'espérance et la variance de X s'écrivent

$$\mathbb{E}(X) = e^{\mu + \frac{\theta^2}{2}} \qquad \text{et} \qquad V(X) = e^{2\mu+\theta^2}\left(e^{\theta^2} - 1\right).$$

L'on vérifie que

$$\mathbb{E}(X) = \mathbb{E}(e^{ln(X)}) = e^{\mu * 1 + \frac{\theta^2 * 1^2}{2}}$$

et que

$$V(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = e^{2\mu+\theta^2}\left(e^{\theta^2} - 1\right)$$

- Déterminer, en supposant θ connu, l'estimateur du maximum de vraisemblance de μ , noté $\hat{\mu}_{MV}$, construit à partir d'observation de $X_1, ..., X_n$. L'on prendra le soin d'établir un tableau de variation associé à la fonction à maximiser

La Log-vraisemblance de l'échantillon s'écrit

$$\ln f(x_1, ..., x_n, \mu, \theta) = -n(\ln(\theta) - \ln(\sqrt{2\pi})) - \sum_{i=1}^{n} ln(x_i) - \sum_{i=1}^{n} \frac{(ln(x_i) - \mu)^2}{2\theta^2}$$

Sa dérivée suivant μ s'écrit

$$\partial_\mu \ln f(x_1, ..., x_n, \mu, \theta) = \sum_{i=1}^{n} \frac{ln(x_i) - \mu}{\theta^2}$$

Qui est bien décroissante en μ et s'annule pour $\mu = \frac{1}{n}\sum_{i=1}^{n} \ln(x_i)$.
En particulier, la log-vraisemblance est concave et admet pour maximum $\hat{\mu}_{MV} = \frac{1}{n}\sum_{i=1}^{n} \ln(x_i)$

- Montrer que $\hat{\mu}_{MV}$ est un estimateur sans biais, convergent de μ.

Avec le rappel, $\mathbb{E}(\ln(X)) = \mathbb{E}(Y) = \mu$ donc l'estimateur n'a pas de biais. Avec la loi des grands nombres appliquée aux lois normales, l'estimateur converge presque surement vers μ et en particulier converge en probabilité.

- Montrer que $\hat{\mu}_{MV}$ est un estimateur efficace de μ.

Commençons par calculer la borne de Cramer-Rao.
La dérivée seconde de la log-vraisemblance est

$$\partial_\mu^2 \ln(f(x_1, ..., x_n, \mu, \theta) = -\sum_{i=1}^{n} \frac{\mu)}{\theta^2} = -\frac{n}{\theta^2}$$

En particulier, l'information de Fisher est

$$I(\mu) = \mathbb{E}(-\partial_\mu^2 \ln(f(x_1, ..., x_n, \mu, \theta))) = \frac{n}{\theta^2}$$

Comme la variance de $\hat{\mu}_{MV}$ est celle d'une somme de loi normale, elle vaut $V(\hat{\mu}_{MV}) = \frac{V(\ln(X_i))}{n} = \frac{\theta^2}{n} = \frac{1}{I(\mu)}$.
L'estimateur est donc efficace

- Nous ne supposons plus θ connu. À l'aide des valeurs de l'espérance et de la variance de X, proposer un estimateur de (μ, θ) qui ne dépend que de

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad \overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

.Notons $m_1 = \mathbb{E}(X)$ et $m_2 = \mathbb{E}(X^2)$.
Alors avec la première question, nous avons le système

$$\begin{cases} \mu + \frac{\theta^2}{2} &= \ln(m_1), \\ 2\mu + 2\theta^2 &= \ln(m_2), \end{cases} \iff \begin{cases} \theta^2 &= \ln(m_2) - 2\ln(m_1), \\ \mu &= 2\ln(m_1) - \frac{1}{2}\ln(m_2). \end{cases}$$

Nous pouvons alors poser comme estimateur

$$\begin{cases} \hat{\mu} &= 2\ln(\overline{X^1}) - \frac{1}{2}\ln(\overline{X^2}), \\ \hat{\theta^2} &= \ln(\overline{X^2}) - 2\ln(\overline{X}). \end{cases}$$

11) On considère l'expérience statistique issue de la répétition indépendante de n tirage de variable aléatoire, avec $\Theta =]-1, 1[$ et de loi $\mathbb{P}_\theta$ la loi sur $\{0, 1, -1\}$ avec des probabilités respectives $1-\theta$, $\frac{\theta}{2}$ et $\frac{\theta}{2}$.
Calculer l'estimateur du maximum de vraisemblance en fonction du nombre de résultats nuls n_1.
La statistique est dominée par la mesure de comptage sur $\{0, 1, -1\}$.
La vraisemblance est l'application

$$(\mathbf{x}, \theta) \mapsto \prod_{x_i=0}(1-\theta) \times \prod_{x_i=1}\frac{\theta}{2} \times \prod_{x_i=-1}\frac{\theta}{2}$$

La Log-vraisemblance est alors

$$(\mathbf{x}, \theta) \mapsto \sum_{x_i=0}\log(1-\theta) + \sum_{x_i\neq 0}\log(\frac{\theta}{2})$$

dont la dérivée s'annule en $\hat{\theta}$ tel que

$$(n-n_1)\frac{2}{\hat{\theta}} - \frac{n_1}{1-\hat{\theta}} = 0$$

C'est à dire pour $\hat{\theta} = \frac{2(n-n_1}{3n_1}$.
Ce résultat doit-être ramené dans [-1,1] s'il sort de l'interval.

12) On considère un modèle régulier $(g(\theta, x))_{\theta\in\Theta}$ dominé par une mesure ν, d'information de Fisher $I(\theta) > 0$. Rappellons que pour tout $x \in \mathbb{R}$, la fonction $\theta \mapsto f(\theta, x)$ est de classe D^2. Considérons $\Phi : \theta \mapsto \Phi(\theta)$ un C^1 difféomorphisme.

- Montrer que pour tout $x \in \mathbb{R}$, la fonction $\eta \mapsto h(\eta, x) = g(\Phi^{-1}(\eta), x)$ est de classe C^1 et calculer sa dérivée.
Il s'agit d'une composition de fonction de classe C^1. Nous avons donc directement le caractère C^1 et l'on peut calculer la dérivée :

$$\partial_\eta h(\eta, x) = (\Phi^{-1})'(\eta)\partial_\theta g(\Phi^{-1}(\eta), x)$$

- En déduire l'information de Fisher du modèle $(h(\eta,x))_{\eta\in\Phi(\Theta)}$ dominé par la mesure ν en passant par la fonction score.
  On peut directement calculer :

$$\begin{aligned} I(\eta) &= \mathbb{E}_\eta\left[\partial_\eta \ln h(X,\eta)\right] \\ &= \int \frac{\partial_\eta h(x,\eta)^2}{h(x,\eta)} \mathbb{1}_{h(x,\eta)>0} d\nu(x) \\ &= \left((\Phi^{-1})'(\eta)\right)^2 \int \frac{\partial_\theta g(x,\Phi^{-1}(\eta))^2}{g(x,\Phi^{-1}(\eta))} \mathbb{1}_{g(x,\Phi^{-1}(\eta))>0} d\nu(x) \\ &= \left((\Phi^{-1})'(\eta)\right)^2 I(\Phi^{-1}(\eta)) \\ &= \frac{1}{\Phi'(\Phi^{-1}(\eta))^2} I(\Phi^{-1}(\eta)) \end{aligned}$$

- Donner la borne de Cramer-Rao associé à un estimateur non biaisé de $\Phi(\theta)$.
  On déduit du point précédent que la borne associée à $\Phi(\theta)$ est nécessairement $\frac{\Phi'(\theta)^2}{I(\theta)}$.

8.4 Tests

La correction des exercices de la page 60.

13) Nous considérons des variables aléatoires indépendantes $X_1,...,X_n$ de même loi $\mathcal{N}(0,\frac{1}{\lambda^2})$.
Nous voulons choisir entre deux hypothèse :

$$H_0 : \lambda = 1 \qquad \text{contre} \qquad H_1 : \lambda = \lambda_1 < 1.$$

- Construire le test de Neyman-Pearson.
  Raisonnons par équivalence (sur la partie du test non randomisé) pour trouver une statistique de test à partir du rapport des vraisemblances :

$$\begin{aligned} \frac{f(x_1,...,x_n,1)}{f(x_1,...,x_n,\lambda_1)} > C_\alpha &\iff \frac{1}{\lambda_1^n} e^{-\frac{((1-\lambda_1^2)}{2}\sum_{i=1}^n x_i^2} > C_\alpha \\ \text{(En prenant le logarithme, fonction croissante)} &\iff -n\ln(\lambda_1) - \frac{(1-\lambda_1^2)}{2}\sum_{i=1}^n x_i^2 > \ln(C_\alpha) \\ \text{(Comme } -\frac{(1-\lambda_1^2)}{2} < 0) &\iff \sum_{i=1}^n x_i^2 < \frac{2}{1-\lambda_1^2}(-n\ln(\lambda_1) - \ln(C_\alpha)) \end{aligned}$$

  Nous prenons alors comme statistique

$$Y = \sum_{i=1}^n x_i^2$$

  Et comme le test de Neyman-Pearson rejette H_0 lorsque le rapport des vraisemblances est supérieur à C_α, nous rejetterons lorsque $Y < \mu_\alpha$.

$$\begin{cases} si \quad Y < \mu_alpha & \text{on rejette } H_0 \\ si \quad Y > \mu_alpha & \text{on ne peut pas rejetter } H_0 \end{cases}$$

  où μ_alpha est le quantile d'ordre α d'une loi χ^2_n

- On désire calculer l'erreur de seconde espèce pour le test de niveau $\alpha = 5\%$ issus de $n = 30$ réalisation, et une alternative $\lambda_1 = 0.9$. Donner la formule générale de cette erreur et sa valeur numérique.
  Pour $n = 30$ et $\alpha = 5\%$, on trouve avec la table 2.4 $\mu_\alpha = 18.493$.
  Or, sous H_1, $Y \times \lambda_1^2 = \sum_{i=1}^n \frac{x_i^2}{1/\lambda_1^2} \sim \chi_n^2$. Donc

$$\mathbb{P}(\text{Accepter } H_0 \ | H_1 \quad \text{vrai}) = \mathbb{P}(Y > \mu_\alpha | H_1 \quad \text{vrai}) = \mathbb{P}(Y\lambda_1^2 > \mu_\alpha \lambda_1^2 | H_1 \quad \text{vrai})$$

  De nouveau avec la table, l'on trouve que pour $Z \sim \chi_{30}^2$

$$\mathbb{P}(Z > \mu_\alpha \times \lambda_1^2) = \mathbb{P}(Z > 14.979) = 0.01$$

  Donc nous aurons une erreur de seconde espèce de 1%.

14) Soit $a \geqslant 0$ connu, et soient $X_1, ..., X_n$ des variables aléatoires indépendantes. La loi de X_j est $\mathcal{N}(m, j^{2a})$. On souhaite choisir entre $H_0 : m = 0$ et $H_1 : m = 2$ à partir des observations des X_j.
Construire le test de Neymann-Pearson associé pour un risque $\alpha = 0.05$.
Que dire de la zone de rejet pour $a > \frac{1}{2}$ et n grand ?
La zone de rejet est de la forme

$$\tau_\alpha < \frac{f(\mathbf{X}, m = 0)}{f(\mathbf{X}, m = 2)}.$$

Nous pouvons réécrire cela comme :

$$\begin{aligned}
\tau_\alpha < \frac{f(\mathbf{X}, m = 0)}{f(\mathbf{X}, m = 2)} &\iff \tau_\alpha < \prod_{j=1}^n e^{-\frac{x_j^2}{2j^{2a}} + \frac{(x_j-2)^2}{2j^{2a}}} \\
&\iff \ln(\tau_\alpha) < \sum_{j=1}^n -\frac{x_j^2}{2j^{2a}} + \frac{(x_j-2)^2}{2j^{2a}} \\
&\iff \ln(\tau_\alpha) < \sum_{j=1}^n \frac{-4x_j + 4}{2j^{2a}} \\
&\iff C_\alpha > \sum_{j=1}^n \frac{x_j}{j^{2a}}
\end{aligned}$$

Or la variable $\sum_{j=1}^n \frac{x_j}{j^{2a}} \sim \mathcal{N}(m\sum \frac{1}{j^{2a}}; \sum \frac{1}{j^{2a}})$. Donc $C_\alpha = 1.65 \times \frac{1}{\sqrt{\sum \frac{1}{j^{2a}}}}$. En particulier, pour $a > \frac{1}{2}$, nous avons $C_\alpha \underset{n\to+\infty}{\to} \frac{1.65}{\sqrt{\zeta(2a)}}$. et donc la zone de rejet "converge" vers une zone limite.

15) Lors d'une enquête, menée en 2019 sur 490 français, 387 estimaient que le niveau de vie en France était en déclin. À la même question posée en août 2023 à 400 personnes, 242 estimaient déjà que cela était le cas.
Peut-on en conclure, à l'aide d'un test d'hypothèse avec un niveau d'erreur de 5%, que la proportion de Français pessimistes a diminué entre 2019 et 2023 ?
Nous allons faire un test d'égalité de proportion. Notons p_1 la proportion de pessimiste en 2019 et p_2 celle de 2023. Nous allons tester

$$\begin{cases} H_0 & : \ p_1 = p_2 \\ \text{contre } H_1 & : \ p_1 > p_2 \end{cases}$$

Les effectifs $n_1 = 490$ et $n_2 = 400$ sont bien supérieurs à 30, et les hypothèses sur les proportions sont également vérifiées. Nous pouvons donc utiliser l'approximation gaussienne.
Nous avons $\hat{p_1} = \frac{387}{490} \approx 0.7898$ et $\hat{p_2} = \frac{242}{400} \approx 0.608$. L'estimateur de la proportion est

$$\hat{p} = \frac{387 + 242}{490 + 400} \approx 0.7067.$$

et la statistique d'intérêt est alors :

$$\frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \approx 5.824$$

Au vu de l'hypothèse, la zone de rejet est de la forme $[q(0.05); +\infty[$. En regardant dans les tables de la loi normale centrée réduite (page 21), nous trouvons qu'il faut prendre $q(0.05) = 1.65$. Comme la statistique est supérieure à cette valeur, nous rejetons l'hypothèse H_0, et pouvons en conclure qu'avec un niveau d'erreur de 5%, la proportion de pessimiste a diminué.

8.5 Les modèles gaussiens

La correction des exercices de la page 72.

16) On considère un vecteur aléatoire à valeurs dans $\mathbb{R}^3$, de composantes X_1, X_2 et X_3 dans la base canonique. On suppose que X_1, X_2 et X_3 sont des variables aléatoires réelles indépendantes de loi gaussienne $\mathcal{N}(0,1)$.

- Quelle est la densité du triplet (X_1, X_2, X_3) ?
  Par indépendance, il s'agit du produit des densités des composantes.
  On reconnait alors la densité d'un vecteur Gaussien, de moyenne $0_{\mathbb{R}^3}$ et de matrice de covariance I_3.
  Soit M la matrice de changement de base orthonormée suivante :

$$M = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} \end{pmatrix}$$

  Vous pourrez admettre par la suite que $MM^t = I_3$.

- Donner la loi de $Y = M^t X$.
  Il s'agit d'une transformation linéaire d'un vecteur gaussien, donc $Y \sim \mathcal{N}(M^t 0_{\mathbb{R}^3}, M^t I_3 (M^t)^t) = \mathcal{N}(0_{\mathbb{R}^3}, I_3)$.

- En déduire la loi des composantes Y_1, Y_2 et Y_3 de Y.
  Il s'agit de calculer des lois marginales. L'on trouvera $Y_i \sim \mathcal{N}(0,1)$.

- Montrer qu'avec $\overline{X} = \frac{1}{3}(X_1 + X_2 + X_3)$, l'on a que

$$\sum_{i=1}^{3} X_i^2 = \sum_{i=1}^{3} Y_i^2$$

$$Y_2^2 + Y_3^2 = \sum_{i=1}^{3} X_i^2 - 3\overline{X}^2$$

  En déduire que les variables $\overline{X}$ et $S_X^2 = \frac{1}{2}\sum_{i=1}^{3}(X_i - \overline{X})^2$ sont indépendantes.
  Le premier point consiste à remarquer que M étant orthogonale, l'application linéaire canonique associée préserve la norme.
  Le deuxième découle directement du fait que $Y_1 = \frac{3}{\sqrt{3}}\overline{X}$.
  Enfin, comme $S_X^2 = \frac{1}{2}\sum_{i=1}^{3}(X_i - \overline{X})^2 = \frac{1}{2}(Y_2^2 + Y_3^3)$. Comme Y_1 et (Y_2, Y_3) sont indépendantes, nous en déduisons le dernier point.

- Donner la loi de $\overline{X}$ et de $2S_X^2$.
  Avec le point précédent, $\overline{X} \sim \mathcal{N}(0, \frac{1}{3})$ et $2S_X^2 \sim \chi_2^2$. Nous avons bien redémontré le corollaire du théorème de Cochran dans ce cas particulier.

17) On considère un système électronique, alimenté par une source continu de tension x connue. On observe à la sortie une tension Y. Nous supposerons que Y peut s'écrire de la forme

$$Y = ux + v + A$$

Avec $A \sim \mathcal{N}(0, \sigma^2)$. On désire estimer les valeurs de u et v. Pour cela, on mesure les valeurs y_j de Y pour diverses valeurs x_j de X. Ces valeurs sont "entachées" des erreurs A_j.

- On effectue ces mesures à Yaoundé (Cameroun) par $30°C$. En supposant que les A_j soient issus de tirages indépendants, déduire du tableau suivant les estimations $\hat{U}$ et $\hat{V}$ de u et v.

x_j :	5	7	9	11	13	15	17	19	21
y_j :	6	8	9.6	9.1	10	10.4	13.7	14.1	14.6

On sait que l'estimateur pour une variable explicative sont :

$$\hat{u_1} = \frac{Cov(x,y)}{\tilde{S_x}^2} \qquad \hat{v_1} = \overline{y} - \hat{u_1}\overline{x}$$

Donc ici, nous trouvons

$$\hat{u_1} = 0.52 \qquad \hat{v_1} = 3.87$$

- On refait à présent ces mesures au Pôle Nord par $-30°C$. Avec les mêmes hypothèses, déduire les estimations $\hat{U}$ et $\hat{V}$ de u et v. Quel remarque peut-on faire ?

x_j :	2	4	6	8	10	12	14	16
y_j :	8	8	10.5	10.1	12.8	12.8	13.1	14.2

On trouve :

$$\hat{u_2} = 0.47 \qquad \hat{v_2} = 6.98$$

si l'estimateur de U reste proche, celui de V change beaucoup. La valeur moyenne semble beaucoup impactée par la température, mais sa réponse au signal l'est peu. Nous allons essayer de le montrer

- Donner un estimateur de la variance σ du bruit. Le calculer pour les deux séries de mesure.
  On pourra démontrer ou utiliser le fait que pour une estimation $\hat{Y}$ de Y par la méthode des moindres carrés avec bruit gaussien par J paramètres X_i sur n mesures, $Y - \hat{Y}$ suit une loi normale de dimension (n-J-1)

  En reprenant les notations de la démonstration de la preuve des moindres carré, $Y - \hat{Y} = (X(X^tX)^{-1}X^t - I_n)A$. Or la matrice $X(X^tX)^{-1}X^t - I_n$ est la matrice de projection sur l'orthogonal de l'espace des colonnes de X. Donc le théorème de Cochran montre l'indication. On peut alors estimer la variance du bruit avec la variance empirique de ce vecteur.
  Pour la première série, on trouve

$$\hat{\sigma_1}^2 := \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{u_1}x_i - \hat{v_1})^2 = 0.69$$

$$\hat{\sigma_2}^2 = 0.539$$

- Montrer avec un test adapté que l'on peut supposer que les deux séries ont la même variance du bruit (ce qui revient à tester que les mesures ont la même précision) avec un risque de 5%. L'on pourra s'aider des tables du chapitre 2.
  On sait déjà (encore avec le théorème de Cochran) que $(n_i-2)\frac{\hat{\sigma_i^2}}{\sigma_i^2}$ suit une loi du χ^2 à n_i-2 degré de liberté.
  On veut tester $H_0 : \sigma_1^2 = \sigma_2^2$ contre l'alternative.
  Sous H_0, on peut donc faire un test d'égalité de variance, puisque alors $\frac{(n_2-2)\sigma_2^2}{(n_1-2)\sigma_1^2}$ suit une loi de Fisher-Snedecor à $n_2-2=6$ et $n_1-2=7$ degrés de libertés.
  On trouve ici $\frac{(n_1-2)\sigma_1^2}{(n_2-2)\sigma_2^2} = 1.49$.
  Avec la table 2.8, l'on trouve le quantile $q_{1-\alpha} = 3.87$.
  On ne peut donc pas rejeter l'hypothèse, à un risque de 5% que

$$\sigma_1^2 = \sigma_2^2$$

- Étudier avec des tests d'égalité de moyenne si les valeurs de u sont les mêmes pour les deux séries de mesures.
  Vous pourrez admettre que $\hat{U} \sim \mathcal{N}(u, \frac{\sigma^2}{n})$
  Nous voulons tester $H_0 : u_1 = u_2$ contre l'alternative. Nous savons que $\hat{u_i} \sim \mathcal{N}(u_i, \frac{\sigma}{n_i})$.
  En particulier,

$$\hat{u_1} - \hat{u_2} \sim \mathcal{N}(u_1 - u_2, \sigma(\frac{1}{n_1} + \frac{1}{n_2}))$$

$$\frac{(n_1-2)\hat{\sigma}_1 + (n_2-2)\hat{\sigma}_2}{\sigma^2} \sim \chi^2_{n_1+n_2-4}$$

  Donc sous H_0, l'on a que

$$\frac{\hat{u_1} - \hat{u_2}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}(0,1)$$

  Et donc que

$$S = \frac{\hat{u_1} - \hat{u_2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{\frac{(n_1-2)\hat{\sigma}_1+(n_2-2)\hat{\sigma}_2}{n_1+n_2-4}}} \sim t_{n_1+n_2-4} \text{ loi de Student à} n_1+n_2-4 \text{ degrés de libertés.}$$

  (*Loi du rapport d'une loi normale par la racine d'une loi d'une variable du* χ^2_n *divisé par son nombre de degrés de liberté*).
  Pour l'échantillon, $S = 0.0308$. le quantile d'ordre 0.975 (car nous cherchons une région de rejet pour un test bilatéral d'une loi symétrique) pour une loi de Student à 13 degrés de liberté est 1,1604.
  Comme $-1.1604 \leqslant S \leqslant 1.1604$, nous ne pouvons pas rejeter H_0 avec un risque de 5%.

- Étudier avec des tests d'égalité de moyenne si les valeurs de v sont les mêmes pour les deux séries de mesures.
  Vous pourrez admettre que $\hat{V} \sim \mathcal{N}\left(v, \frac{\sigma^2}{n}(1 + \frac{\overline{x}_i^2}{S^2_{x_i}})\right)$
  Comme pour la question précédente, nous voulons tester $H_0 : v_1 = v_2$ contre l'alternative. Nous savons que $\hat{v_i} \sim \mathcal{N}\left(v_i, \frac{\sigma^2}{n_i}(1 + \frac{\overline{x}^2}{S^2_X})\right)$.
  Avec le même raisonnement qu'à la question précédente, l'on trouve que

$$S' = \frac{\hat{v_1} - \hat{v_2}}{\sqrt{\frac{1}{n_1}(1 + \frac{\overline{x}_1^2}{S^2_{X_1}}) + \frac{1}{n_2}(1 + \frac{\overline{x}_2^2}{S^2_{x_2}})}\sqrt{\frac{(n_1-2)\hat{\sigma}_1+(n_2-2)\hat{\sigma}_2}{n_1+n_2-4}}} \sim t_{n_1+n_2-4}$$

  Pour l'échantillon, $S' = -3.31$. Comme $S' \notin [-1.1604; 1.1604]$, nous rejetons H_0 avec un risque de 5%.

18) **(Régression non paramétrique - base de Fourrier - D'après le TD du Pr. Guyader)**
Nous nous intéressons à un problème plus difficile : considérons que nous observons les variables aléatoires $Y_i = f(\frac{i}{n}) + \epsilon_i$ pour $i \in [\![1, n]\!]$, somme des valeurs d'une fonction avec un bruit. Nous supposons donc que les ϵ_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$ avec σ connu, et que $f : [0, 1] \to \mathbb{R}$ est une fonction inconnue qui est le paramètre d'intérêt.
Réécriture du modèle

- Quel est la difficulté particulière de ce modèle statistique ?
  Il ne s'agit pas d'un problème linéaire.

  Pour résoudre ce problème, on propose de projeter f sur une base de fonction bien choisie. Supposons pour la fin de l'exercice que

$$\forall t \in [0, 1], f(t) = a_0 + \sum_{k=1}^{K} a_k \cos(2k\pi t) + b_k \sin(2k\pi t)$$

  Les inconnues deviennent alors $(a_i)_{0 \leqslant i \leqslant K}$ **et** $(b_i)_{1 \leqslant i \leqslant K}$**.**

- Sous cette hypothèse, écrire le modèle comme un modèle linéaire gaussien $Y = X\beta + \epsilon$ et préciser X, β et le nombre de variables explicatives p.
  Nous avons alors $p = 2K + 1$,

$$\beta = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_K \\ b_1 \\ \vdots \\ b_n \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & \cos(2\pi\frac{1}{n}) & \cdots & \cos(2k\pi\frac{1}{n}) & \cdots & \cos(2K\pi\frac{1}{n}) & \sin(2\pi\frac{1}{n}) & \cdots & \sin(2k\pi\frac{1}{n}) & \cdots & \sin(2K\pi\frac{1}{n}) \\ \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots & & \vdots \\ 1 & \cos(2\pi\frac{i}{n}) & \cdots & \cos(2k\pi\frac{i}{n}) & \cdots & \cos(2K\pi\frac{i}{n}) & \sin(2\pi\frac{i}{n}) & \cdots & \sin(2k\pi\frac{i}{n}) & \cdots & \sin(2K\pi\frac{i}{n}) \\ \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 \end{pmatrix}$$

- On suppose à présent $2K + 1 \leqslant n$. Vérifier que le modèle est identifiable (que la matrice est de rang plein), et calculer l'estimateur des moindres carrés $\hat{\beta}$ de β.
  En déduire un estimateur $\hat{\mu}$ de $\mu = [f(\frac{i}{n})]_{i \in [\![1,n]\!]}$, et proposer un estimateur $\hat{f}$ de la fonction f.
  On pourra se rappeler les formules de trigonométrie, en notant j une racine carré de -1 *:*

$$\cos(A)\cos(B) = \frac{1}{2} Re(e^{j(A+B)} + e^{j(A-B)})$$

$$\cos(A)\sin(B) = \frac{1}{2} Im(e^{j(A+B)} - e^{j(A-B)})$$

$$\sin(A)\sin(B) = \frac{1}{2} Re(e^{j(A-B)} - e^{j(A+B)})$$

S'il existe une combinaison linéaire des colonnes nulle $\lambda_0 \sum_{k=1}^{K} \lambda_k X_{k+1} + \mu_k X_{K+1+k} = 0$, alors le polynôme trigonométrique

$$g : t \mapsto \lambda_0 + \sum_{k=1}^{K} \lambda_k \cos(2k\pi t) + \mu_k \sin(2k\pi t)$$

admet $2K+1$ zéro sur $[0,1]$ et est donc la fonction nulle et les coefficients sont nuls. En particulier, le système est identifiable.

Calculons la matrice $X^t X$ à l'aide de l'indication :

$$X^t X = \begin{pmatrix} n & \cdots & \sum_{i=1}^n \cos(2k\pi \frac{i}{n}) & \cdots & \cdots & \sum_{i=1}^n \sin(2k\pi \frac{i}{n}) & \cdots \\ \vdots & & \vdots & & & \vdots & \\ \sum_{i=1}^n \cos(2\tilde{k}\pi \frac{i}{n}) & \cdots & \sum_{i=1}^n \cos(2k\pi \frac{i}{n}) \cos(2\tilde{k}\pi \frac{i}{n}) & \cdots & \cdots & \sum_{i=1}^n \sin(2k\pi \frac{i}{n}) \cos(2\tilde{k}\pi \frac{i}{n}) & \cdots \\ \vdots & & \vdots & & & \vdots & \\ \sum_{i=1}^n \sin(2\tilde{k}\pi \frac{i}{n}) & \cdots & \sum_{i=1}^n \cos(2k\pi \frac{i}{n}) \sin(2\tilde{k}\pi \frac{i}{n}) & \cdots & \cdots & \sum_{i=1}^n \sin(2k\pi \frac{i}{n}) \sin(2\tilde{k}\pi \frac{i}{n}) & \cdots \end{pmatrix}$$

Avec l'indication, pour $k \neq \tilde{k}$,

$$\begin{aligned} \sum_{i=1}^n \cos(2k\pi \frac{i}{n}) \sin(2\tilde{k}\pi \frac{i}{n}) &= \sum_{i=1}^n \frac{1}{2} Im(e^{2j\pi(\tilde{k}+k)\frac{i}{n}} - e^{2j\pi(\tilde{k}-k)\frac{i}{n}}) \\ &= \frac{1}{2} Im(\frac{1-e^{2j\pi(\tilde{k}+k)}}{1-e^{2j\pi\frac{(\tilde{k}+k)}{n}}} - \frac{1-e^{2j\pi(\tilde{k}-k)}}{1-e^{2j\pi\frac{(\tilde{k}-k)}{n}}} \\ &= 0 \end{aligned}$$

De la même manière, l'on prouve pour $k \neq \tilde{k}$

$$\sum_{i=1}^n \cos(2k\pi \frac{i}{n}) \cos(2\tilde{k}\pi \frac{i}{n}) = 0$$

$$\sum_{i=1}^n \sin(2k\pi \frac{i}{n}) \sin(2\tilde{k}\pi \frac{i}{n}) = 0$$

et pour $k \neq 0$,

$$\sum_{i=1}^n \cos(2k\pi \frac{i}{n}) = 0$$

$$\sum_{i=1}^n \sin(2k\pi \frac{i}{n}) = 0$$

Enfin, pour $k = \tilde{k}$, l'on prouve que

$$\sum_{i=1}^n \cos(2k\pi \frac{i}{n}) \cos(2\tilde{k}\pi \frac{i}{n}) = \frac{n}{2}$$

$$\sum_{i=1}^n \sin(2k\pi \frac{i}{n}) \sin(2\tilde{k}\pi \frac{i}{n}) = \frac{n}{2}$$

Donc finalement $X^t X$ est diagonale de la forme :

$$X^t X = \begin{pmatrix} n & 0 & \cdots & 0 \\ 0 & \frac{n}{2} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & \frac{n}{2} \end{pmatrix}$$

Finalement,

$$\hat{\beta} = diag(\frac{2}{n}, \frac{1}{n}, ..., \frac{1}{n}) * X^t Y$$

$$\hat{\mu} = X * diag(\frac{2}{n}, \frac{1}{n}, ..., \frac{1}{n}) * X^t Y$$

$$\hat{f} = \hat{\beta}_1 + \sum_{k=1}^{K} \hat{\beta}_{k+1} \cos(2k\pi t) + \hat{\beta}_{k+K+1} \sin(2k\pi t)$$

Overfitting et choix du modèle :

- Calculé la somme des carrés de l'erreur quadratique renormalisé :

$$r_n = \frac{\mathbb{E}(||Y - X\hat{\beta}||^2)}{n}$$

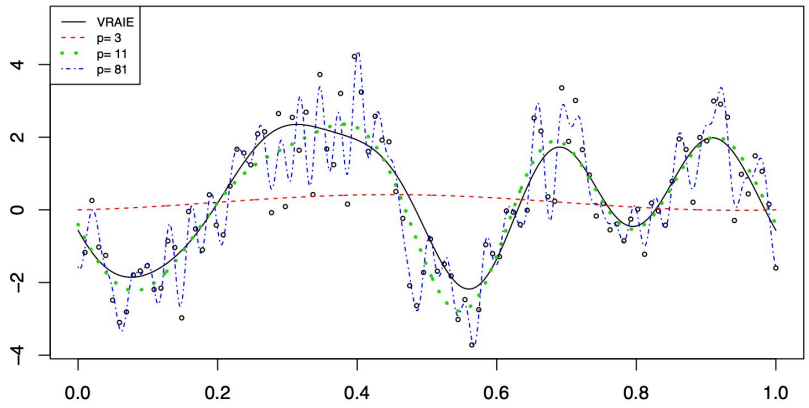
Pour K fixé, que se passe-t-il quand n tend vers l'infini ?
Par définition, $Y - X\hat{\beta} = Y - \hat{Y}$ est la projection de Y sur l'espace des colones de X. En particulier, d'après le théorème de Cochran, $||\frac{Y - X\hat{\beta}}{\sigma}||^2 \sim \chi^2_{n-p}$.
Nous pouvons alors en déduire l'erreur quadratique renormalisé :

$$r_n = \frac{\mathbb{E}(||Y - X\hat{\beta}||^2)}{n} = \frac{(n-p)\sigma}{n}$$

Pour K fixé, cette erreur moyenne tend vers σ^2 : le bruit prédomine.

- On suppose que p=n, donner alors la valeur de r_n. Que peut-on dire des valeurs de $\hat{f}$ aux points $\frac{i}{n}$ pour $i \in [\![1, n]\!]$?
Nous avons alors que si $p = n$, alors $r_n = 0$.
Nous avons trop de paramètre , et les valeurs de $\hat{f}$ valent Y_i.

- Nous prenons un polynôme trigonométrique avec coefficients non nuls que pour $k \leqslant 11$ dans la décomposition ci-dessus. Nous faisons n = 101 observations et choisissions diverses valeurs de K. Nous obtenons la figure ci-dessous Quels phénomènes observe-t-on ? Quelle règle proposez-vous au vu des points précédents et de cette observation pour un bon ajustement ?



Nous voyons une mauvaise estimation si p=3 (manque de paramètres dans ce cas) et si $p \approx n$ (overfitting). En revanche, pour $p \approx \frac{n}{10}$, l'estimation est bonne.
Nous pouvons alors proposer la règle suivante : le nombre d'observations doit être supérieur à 10 fois le nombre de variables explicatives envisagé.

8.6 Tests de χ^2 et test de Kolmogorov-Smirnov

La correction des exercices de la page 84.

19) Soit (X_i) une suite de variables aléatoires indépendante et de même loi que la variable X. L'on note $F : t \mapsto \mathbb{P}(X \leqslant t)$ la fonction de répartition de la loi et $\hat{F}_n$ la fonction de répartition empirique de l'échantillon.
Montrer que la loi de la variable

$$\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)|$$

ne dépend pas de la loi de X lorsque F est continu.
L'on pourra se servir de l'inverse généralisé de la fonction de répartition
Nous allons suivre l'indication. Notons $F^{\leftarrow}$ l'inverse généralisé de la fonction de répartition F.
Rappelons que U une variable suivant une loi uniforme sur $[0,1]$, $F^{\leftarrow} \circ U$ à la même loi que X (voir annexe A). Nous pouvons donc, sans perte de généralité sur les lois, supposer qu'il existe (U_i) une suite de variables aléatoires indépendante et suivant une loi uniforme sur $[0,1]$ telle que $X_k = F^{\leftarrow} \circ U_i$.
Or

$$\begin{aligned} F_n(t) &= \frac{Card\{k \in [\![1,n]\!], X_k \leqslant t\}}{n} \\ &= \frac{Card\{k \in [\![1,n]\!], U_i \leqslant F^{\leftarrow}(t)\}}{n} \end{aligned}$$

et

$$F(t) = \mathbb{P}(X \leqslant t) = \mathbb{P}(U \leqslant F^{\leftarrow}(t)).$$

Ce qui nous donne, en posant $s = F(t)$, que

$$\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| = \sup_{s\in F(\mathbb{R})} |\hat{F_{n,U}}(s) - s| = \sup_{s\in]0,1[} |\hat{F_{n,U}}(s) - s|.$$

Car si F est continue, $F(\mathbb{R}) =]0,1[$ et donc la loi de la variable aléatoire $\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)|$ ne dépend pas de la loi de X (puisqu'elle est égale en loi à une variable aléatoire ne dépendant que de la loi uniforme).

20) Au cours d'une année, les sites de productions d'une entreprise d'électronique de pointe doivent renouveler certains des outils d'usinages nécessaires à la production d'une entreprise. L'hypothèse de modélisation la plus simple est de dire que le nombre d'outils à remplacer dans un site de production suit une loi de Poisson, de paramètre λ inconnu.
Au cours de l'année passée, l'entreprise a compté le nombre d'outils remplacé dans ses divers sites. Les résultats sont présentés dans le tableau suivant :

Nombre d'outils renouvelés	0	1	2	3	4	5	6	7	8	9
Nombre de sites	1	2	2	3	3	5	5	2	3	2

Nous admettrons que l'estimateur de maximum de vraisemblance pour le paramètre des lois de Poisson est la moyenne empirique.
L'entreprise souhaite vérifier que les données (x_i) sont bien issues d'un échantillon de loi de Poisson de paramètre inconnus.

L'on donne à l'aide de la fonction Malab, dont la documentation décrit "$POISSCDF(X, Lambda)$ compute the Poisson cumulative distribution function with parameter $Lambda$ at the value X". poisscdf([0 1 2 3 4 5 6 7 8 9 10 11], 3) = [0.0498, 0.1991, 0.4232, 0.6472, 0.8153, 0.9161, 0.9665, 0.9881, 0.9962, 0.9989, 0.9997, 0.9999]
poisscdf([0 1 2 3 4 5 6 7 8 9 10 11], 4)=[0.0183, 0.0916, 0.2381, 0.4335, 0.6288, 0.7851, 0.8893, 0.9489, 0.9786, 0.9919, 0.9972, 0.9991]
poisscdf([0 1 2 3 4 5 6 7 8 9 10 11], 5)=[0.0067, 0.0404, 0.1247, 0.265 , 0.4405, 0.616 , 0.7622, 0.8666, 0.9319, 0.9682, 0.9863, 0.9945] poisscdf([0 1 2 3 4 5 6 7 8 9 10 11], 6)=[0.0025, 0.0174, 0.062 , 0.1512, 0.2851, 0.4457, 0.6063, 0.744 , 0.8472, 0.9161, 0.9574, 0.9799]

- L'entreprise souhaite effectuer un test du χ^2 avec un risque $\alpha = 0.05$. Calculez la statistique d'un tel test en répartissant les données dans trois classes judicieusement choisies. Combien de degré de liberté aura-t-on ? En s'aidant de la table à la page 23, donnez la région d'acceptation. Conclure.
  On trouve $\hat{\lambda} \approx 5$.
  Au vu de la fonction de répartition correspondante, ont choisi les classes : $C_1 = \{0, 1, 2, 3\}$, $C_1 = \{4, 5\}$ et C_3 les autres valeurs. En particulier, $\mathbb{P}(Y \in C_1) = 0.256$, $\mathbb{P}(Y \in C_2) = 0.36$ et $\mathbb{P}(Y \in C_3) = 0.384$.
  L'on compte 8 valeurs dans C_1, 8 dans C_2 et 12 dans C_3. Comme il y a 3 classe et un paramètre estimé, il faut comparer à un test du χ^2 avec 1 degré de liberté.
  Après calcul, la statistique de test vaut $\sum_{l=1}^{3} \frac{(n_i - np_i)^2}{np_i} = 0.67$
  Il faut comparer cela à la valeur critique de 3.841, qui nous donne comme région d'acceptation $[0, 3.841]$.
  Nous ne pouvons donc pas rejeter l'hypothèse que ces données suivent bien une loi de Poisson.

- Effectuer à présent un test de Kolmogorov. Est-ce que votre conclusion évolue ?
  Le tableau des fréquences empiriques est :

Nombre d'outils renouvelés	0	1	2	3	4
Fréquence cumulée de sites	0.0357	0.1071	0.1785	0.2857	0.3928
Fréquence cumulée théorique	0.0067	0.0404	0.1247	0.265	0.4405

5	6	7	8	9
0.5714	0.75	0.8214	0.9285	1
0.616	0.7622	0.8666	0.9319	0.9682

  L'on trouve donc que

$$\sup |\hat{F}_n(t) - F(t)| = 0.047$$

  Il faut comparer cette valeur à 0.26404 (voir table 2.9 page 28), et donc encore une fois, nous ne pouvons pas rejeter l'hypothèse que ces données suivent bien une loi de Poisson.

Annexe A

Inverse généralisé de fonction de répartition

Soit F une fonction de répartition, donc croissante et continue à droite. On cherche ici à inverser la fonction du mieux possible. Malheureusement, la fonction F n'est pas nécessairement injective. Pour certaines valeurs, nous aurons donc plusieurs choix d'inverses possibles. Mais tous ces choix ne se valent pas, et il y a un choix préférentiel : l'inverse généralisé. Il s'agit de la fonction $F^{\leftarrow} : s \mapsto \inf\{x,\ F(x) \geqslant s\}$

Proposition 78 :

$F^{\leftarrow}$ est croissante, continue à gauche et vérifie pour tout u que :

$$F(F^{\leftarrow}(y)) \geqslant y$$

$$\{t, F(t) \geqslant u\} = \{t, t \geqslant F^{\leftarrow}(u)\}$$

Démonstration.

- Soit $u < v$, comme $\{x,\ F(x) \geqslant v\} \subset \{x,\ F(x) \geqslant u\}$, on obtient directement que $F^{\leftarrow}$ est croissante.
- Soit (s_n) une suite croissante et convergente vers $s \in \mathbb{R}$. La fonction $F^{\leftarrow}$ étant croissante, $F^{\leftarrow}(s_n)$ converge vers un l. Montrons que $l = F^{\leftarrow}(s)$.
  Par croissance de $F^{\leftarrow}$, on a que $l \leqslant F^{\leftarrow}(s)$. Soit $m < F^{\leftarrow}(s)$, alors il existe $x_O \geqslant m$ tel que $F(x_0) < s$.
  Par convergence, il existe un rang n tel qu'à partir de ce rang, $s_n > F(x_0)$, et donc par croissance de F que $F^{\leftarrow}(s_n) \geqslant x_0 \geqslant m$.
  En particulier, à la limite, $l \geqslant m$. Puis en faisant tendre m vers $F^{\leftarrow}(s)$, l'on en déduit que $l \geqslant F^{\leftarrow}(s)$.
- Il existe une suite (x_n) qui converge par valeur supérieure vers $F^{\leftarrow}(y)$ tels que $F(x_n) \geqslant y$. Comme F est continu à droite, on en déduit que $F(F^{\leftarrow}(y)) \geqslant y$
- Procédons par double inclusion : Soit t tel que $F(t) \geqslant u$. Alors par définition, $F^{\leftarrow}(u) = inf(x,\ F(x) \geqslant u) \geqslant t$.
  Soit maintenant t tel que $F(t) < u$. Comme F est continu à droite, il existe $t_0 > t$ tel que $F(t_0) < u$, et donc nous en déduisons par croissance de F que $F^{\leftarrow}(u) \geqslant t_0 > t$.

□

Dans le cadre particulier des statistiques, on appelle également l'inverse généralisé d'une fonction de répartition la fonction quantile de la variable aléatoire.

Exemple : Pour X une variable aléatoire, la médiane est la valeur t telle que la probabilité que X lui soit inférieur est de $\frac{1}{2}$. C'est donc $F^{\leftarrow}(\frac{1}{2})$.

La fonction quantile n'a pas qu'un intérêt théorique. En effet, le lemme suivant nous dit qu'elle permet de simuler très facilement n'importe quelle variable aléatoire.

Lemme : *de transformation des quantiles*
Soit $U : (\Omega, \mathcal{T}, \mathbb{P}) \to ([0,1], \mathcal{B}[0,1])$ une variable aléatoire de loi uniforme sur [0, 1]. Soit $F : \mathbb{R} \to [0,1]$ une fonction de répartition. Alors la variable aléatoire $X := F^{\leftarrow} \circ U : (\Omega, \mathcal{T}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ est une variable aléatoire de fonction de répartition F.

Démonstration.
Soit U et F comme dans l'énoncé, et soit $t \in \mathbb{R}$, alors avec la proposition précédente, et comme nous connaissons la fonction de répartition d'une loi uniforme,

$$\mathbb{P}(F^{\leftarrow} \circ U \leqslant t) = \mathbb{P}(U \leqslant F(t)) = F(t)$$

□

Annexe B

Propriétés probabilistes supplémentaires (H.P.)

Nous allons dans cette partie voir certaines des inégalités classiques de la théorie des probabilités, qui trouvent leur utilisation autant dans les démonstrations théoriques que dans la vérification de l'efficacité d'algorithmes.

B.1 Inégalité de Jensen pour les fonctions convexes

Cette propriété est essentielle si l'on cherche à majorer des espérances, mais aussi lorsque l'on cherche des inégalités de concentrations ou des majorations de probabilités pour des tests.

Commençons par rappeler des résultats de bases sur les fonctions convexes :

Définition 79 :

On appelle fonction convexe toute application $f : I \subset \mathbb{R} \to \mathbb{R}$, où I est un intervalle, vérifiant :

$$\forall (x, y) \in V^2, \forall \lambda \in [0, 1] : f\left[\lambda x + (1 - \lambda)y\right] \leqslant \lambda f(x) + (1 - \lambda) f(y)$$

L'on peut aisément montrer les résultats suivant :

Proposition 80 :

Soit f une fonction convexe. On a que

- $$\forall a < b < c, \quad \frac{f(a) - f(b)}{a - b} \leqslant \frac{f(a) - f(c)}{a - c} \leqslant \frac{f(b) - f(c)}{b - c}.$$
- f est continu.
- f est dérivable sauf en un nombre dénombrable de points.
- Si f est dérivable, sa dérivée est croissante.
- Si f est deux fois dérivable, sa dérivée seconde est positive.

Ces résultats sont laissés en exercice.

Théoreme 81 :

Soit $(\Omega, \mathcal{T}, \mathbb{P})$ un espace probabilisé.

Soit f une fonction convexe sur un intervalle réel $[a, b]$ et X une variable aléatoire à valeurs dans I, dont l'espérance $\mathbb{E}(f(X))$ existe, alors

$$f(\mathbb{E}(X)) \leqslant \mathbb{E}(f(X))$$

Démonstration.

Soit $z \in [a, b]$. Comme f est une fonction convexe, pour z fixé, il existe $l \in \mathbb{R}$ tel que

$$\forall x \in [a, b], f(x) \geqslant f(z) + (z - x) \times l$$

En effet, il suffit de prendre $l = \limsup_{x \to z^-} \frac{f(z)-f(x)}{z-x}$ et la propriété découle directement de la croissance des taux d'accroissement (premier point de la propriété précédente).

Remarque : si f est dérivable en z, il s'agit alors de prendre la dérivée de f en z, et d'utiliser qu'une fonction convexe est au-dessus de ses tangentes.

Donc pour la variable aléatoire X, l'on a l'inégalité :

$$f(X) \geqslant f(z) + (z - X) \times l.$$

Par la croissance et linéarité de l'espérance, il en vient que

$$\mathbb{E}[f(X)] \geqslant f(z) + (z - \mathbb{E}[X]) \times l.$$

Si l'on prend maintenant au début du raisonnement que $z = \mathbb{E}[X]$, on obtient bien que

$$\mathbb{E}[f(X)] \geqslant f(\mathbb{E}[X]).$$

□

B.2 Inégalité d'Hölder

Une autre application théorique des fonctions convexes est l'inégalité d'Hölder dans le cadre probabiliste :

Proposition 82 :

Soit $(\Omega, \mathcal{T}, \mathbb{P})$ un espace probabilisé.

Soit $(p, q, r) \in]0, +\infty]$ tels que $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$.

Alors pour toutes variables X et Y tels que X^p et Y^q soient intégrables, $|XY|^r$ admet une espérance finie, et de plus :

$$\mathbb{E}\left[|XY|^r\right]^{\frac{1}{r}} \leqslant \mathbb{E}\left[|X|^p\right]^{\frac{1}{p}} \times \mathbb{E}\left[|Y|^q\right]^{\frac{1}{q}}$$

Démonstration.

La fonction $-\ln$ est une fonction convexe, donc pour $u, v \in \mathbb{R}^+$, l'on a que

$$-\ln(\frac{r}{p}u^{\frac{p}{r}} + \frac{r}{q}v^{\frac{q}{r}}) \leqslant -\frac{r}{p}\ln(u^{\frac{p}{r}}) - \frac{r}{q}\ln(v^{\frac{q}{r}}) = -\ln(u) - \ln(v).$$

C'est-à-dire que $uv \leqslant \frac{r}{p}u^{\frac{p}{r}} + \frac{r}{q}v^{\frac{q}{r}}$.

Supposons tout d'abord que $\mathbb{E}\left[|X|^p\right] = \mathbb{E}\left[|Y|^q\right] = 1$ On applique alors l'inégalité précédente à $|X|^r$ et $|Y|^r$ pour obtenir que

$$|XY|^r \leqslant \frac{r}{p}|X|^p + \frac{r}{q}|Y|^q.$$

Mais alors en intégrant, l'on obtient bien que

$$\mathbb{E}(|XY|^r) \leqslant \frac{r}{p}\mathbb{E}(|X|^p) + \frac{r}{q}\mathbb{E}(|Y|^q) = \frac{r}{p} + \frac{r}{q} = 1 = \mathbb{E}\left[|X|^p\right] \times \mathbb{E}\left[|Y|^q\right]$$

Maintenant, pour le cadre général, il suffit de se rendre compte que si $X \neq 0$ et $Y \neq 0$,

$$\mathbb{E}\left[\left|\frac{X}{\mathbb{E}\left[|X|^p\right]^{\frac{1}{p}}}\right|^p\right] = 1 = \mathbb{E}\left[\left|\frac{Y}{\mathbb{E}\left[|Y|^q\right]^{\frac{1}{q}}}\right|^q\right].$$

□

Annexe C

Compléments d'algèbre linéaire

Dans cette partie, nous allons donner des résultats d'algèbres linéaires utilisés dans ce polycopié. L'on notera $M_{n,p}$ les matrices de n lignes et p colonnes.

C.1 Factorisation de Choleski

Nous allons commencer par montrer le théorème suivant :

Théoreme 83 : Factorisation de Cholesky

Soit $K \in S_n^+$ une matrice symétrique semi-définie positive.
Alors il existe une matrice de permutation $P \in M_{n,n}$ et une matrice triangulaire inférieur $L \in T_n$ à coefficients diagonaux positifs tels que :

$$PKP^t = L \times L^t$$

Si K est défini positive, alors on peut prendre $P = I_n$ et alors L est unique.

Nous ne traiterons que le cas défini positif, le cas général pouvant s'en déduire en utilisant que pour une matrice symétrique, $Ker(A) = Im(A)^\perp$ pour le produit scalaire (hermitien) canonique de $\mathbb{R}^n$ ($\mathbb{C}^n$). Nous ne montrerons pas l'unicité de la décomposition, qui sera laissé au lecteur.

Démonstration. Nous allons faire une démonstration en construisant un algorithme, et nous aurons pour cela besoin de la proposition matricielle suivante :

Proposition 84 :
Soit $K \in S_n^{++}$ une matrice symétrique définie positive écrite par bloc de la forme :

$$K := \begin{pmatrix} A & B^t \\ B & W \end{pmatrix}$$

Alors A est définie positive, donc inversible, et le complément de Schur de A dans la matrice K définie comme la matrice

$$W - BA^{-1}B^t$$

est défini positif.

Démonstration. Le caractère défini positif de A est immédiat avec la caractérisation

$$A \text{ est défini positif } \iff \forall X \in \mathbb{K}^n, \vec{X}^t A \vec{X} \geqslant 0$$

Maintenant, supposons par l'absurde que le complément de Schur de A dans K n'est pas défini positif, il existe donc $\vec{u}$ non nul tel que

$$\vec{u}^t (W - BA^{-1}B^t)\vec{u} \leqslant 0$$

On pose alors le vecteur non nul

$$\vec{v} = \begin{pmatrix} -A^{-1}B^t\vec{u} \\ \vec{u} \end{pmatrix}$$

qui vérifie bien

$$\vec{v}^t W \vec{v} = \vec{u}^t (W - BA^{-1}B^t)\vec{u} \leqslant 0$$

Ce qui contredit le fait que K est défini positif. □

Construisons à présent par récurrence forte sur n l'existence d'une décomposition de Cholesky pour toute matrice définie positive.

Initialisation : Pour $n = 1$, l'existence d'une telle décomposition est triviale.

Hérédité : Supposons l'existence d'une telle décomposition (sans matrice de permutation) pour toute matrice définie positive de taille $1 \leqslant k \leqslant n$, et prenons $K \in M_{n+1,n+1}$ une matrice définie positive, et choisissons $1 \leqslant k \leqslant n$. Nous pouvons alors écrire K de la forme

$$K = \begin{pmatrix} A & B^t \\ B & W \end{pmatrix}$$

avec $A \in M_{k,k}$.

D'après l'hypothèse de récurrence forte et le fait que $C = W - BA^{-1}B^t$ est définie positive, les deux matrices A et C admettent chacune une décomposition de Cholesky. Il existe donc $(L_1, L_2) \in T_n^2$ deux matrices triangulaires inférieures à coefficients diagonaux positifs tels que $A = L_1 L_1^t$ et $W - BA^{-1}B^t = L2L_2^t$. Nous trouvons alors une décomposition de Cholesky en écrivant K comme :

$$K = \begin{pmatrix} L_1 & 0 \\ B(L_1^t)^{-1} & L_2 \end{pmatrix} \times \begin{pmatrix} L_1^t & L_1^{-1}B^t \\ 0 & L_2^t \end{pmatrix}$$

□

Remarque : Si l'on suit l'algorithme, la décomposition de Cholesky est facile à calculer. On peut montrer que la complexité de l'algorithme est de l'ordre de n^3.

On préfère parfois trouver une racine carrée de la matrice K, c'est-à-dire une matrice A, qui commute avec K, telle que

$$K = A^2$$

. De plus, l'on veut parfois trouver une matrice de plus petite dimension "proche" de la matrice originelle pour accélérer les calculs en statistique de grande dimension. Pour le lecteur intéressé, nous renvoyons alors à la décomposition spectrale et la décomposition en valeur singulière. L'on pourra par exemple lire :

Le livre [HJ90] pour tous les résultats théoriques de l'analyse matricielle,

Le livre [Eld07] pour des méthodes algorithmiques de régression le long des valeurs principale, pour des régressions mal posées.

Annexe D

Espérance conditionnelle de variables aléatoires (H.P.)

Nous allons dans cette annexe donner la définition et quelques propriétés utiles à l'intuition des espérances conditionnelle. L'idée principale qu'il faut retenir de l'espérance conditionnelle d'une variable Y sachant X est que l'on cherche à trouver la meilleure approximation de Y à partir de la valeur de X (ou des valeurs de ses coordonnées). Soit $(\Omega, \mathcal{T}, \mathbb{P})$ un espace probabilisé, et $X : (\Omega, \mathcal{T}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}^n))$ une variable ou un vecteur aléatoire, c'est-à-dire une application mesurable. La première question à se poser est celle de l'information maximale que l'on peut extraire de X, les événements accessibles à partir des valeurs de X ou encore comme la meilleure reconstruction du signal original Y, après émission, en fonction de la déformation bruitée obtenue à la réception.

Définition 85 :

On appelle tribu engendrée *par X la tribu*

$$\mathcal{T}_X = X^{-1}(\mathcal{B}(\mathbb{R}^n)) = \{X^{-1}(A) | A \in \mathcal{B}(\mathbb{R}^n)\}$$

L'on peut montrer (théorème de Doob) que $Y : \Omega \to \mathbb{R}^n$ est $\mathcal{T}_X$ mesurable si et seulement si il existe une application h, mesurable pour la tribu borélienne, telle que $Y = h(X)$.

Rappelons que $L^2(\Omega, F, \mu)$ est l'ensemble des classes (pour l'égalité μ presque partout) d'applications réelles F mesurables de carré intégrable.
Nous avons naturellement l'inclusion suivante :

Proposition 86 :

$L^2(\Omega, \mathcal{T}_X, \mu)$ est un sous-espace vectoriel **fermé** de $L^2(\Omega, \mathcal{T}, \mu)$ pour toute mesure μ.

Démonstration. L'inclusion est immédiate.
L'on prend $Z_i \in L^2(\Omega, \mathcal{T}_X, \mu)$ tel que Z_i converge en moyenne quadratique dans $L^2(\Omega, \mathcal{T}, \mu)$ vers Z. On se donne aussi h_i la suite des applications mesurables telles que $Z_i = h_i(X)$.
Quitte à extraire, l'on peut supposer que h_i converge vers la limite inférieur ponctuelle. Mais alors $h_i(X)$ converge en moyenne quadratique vers Z et donc l'on peut trouver une extractrice ϕ telle que $h_{\phi(i)}(X)$ converge presque partout vers Z. Comme $h_{\phi(i)}(X)$ converge vers $h(X)$, par unicité de la limite presque sûre l'on aura bien que $Z = h(X)$ est $\mathcal{T}_X$ mesurable. □

L'espace des fonctions de carré intégrable est intéressant puisqu'il s'agit d'un espace de Hilbert avec le produit scalaire :

$$(X, Y) \mapsto < X, Y >= \mathbb{E}(XY)$$

Mais la projection orthogonale dans les espaces fonctionnels peut ne pas être continue si l'espace d'arrivée n'est pas fermé. Heureusement, ce n'est pas le cas ici, ce qui permet la définition suivante :

Définition 87 :
Soit X et Y deux variables aléatoires quelconques de $(\Omega, \mathcal{T}, \mathbb{P})$.
L'on appelle espérance conditionnelle de Y sachant X noté $\mathbb{E}^X(Y)$ *ou* $\mathbb{E}(Y|X)$ *la projection orthogonale de Y sur* $L^2(\Omega, \mathcal{T}_X, \mathbb{P})$

Avec les propriétés de la projection orthogonale, l'on peut directement énoncer des propriétés de l'espérance conditionnelles :

Proposition 88 :
Soit X et Y deux variables aléatoires de carré intégrable. Alors :

- L'espérance conditionnelle de Y sachant X est la meilleure manière de minimiser l'erreur quadratique moyenne que l'on puisse avoir de Y par une variable de la forme h(X), i.e.

$$\forall h \text{ borel-mesurable, } \mathbb{E}\left[(Y - \mathbb{E}^X(Y))^2\right] = ||Y - \mathbb{E}^X(Y)||^2 \leqslant ||Y - h(X)||^2 = \mathbb{E}\left[(Y - h(X))^2\right].$$

- L'espérance conditionnelle est caractérisée par le comportement de Y vis-à-vis des valeurs de X, c'est-à-dire, parmi les applications $\mathcal{T}_X$ mesurables, par la propriété

$$\forall h \text{ borel-mesurable bornée, } \mathbb{E}(h(X)Y) =< h(X), \mathbb{E}^X(Y) >= \mathbb{E}(h(X)\mathbb{E}^X(Y)).$$

- Si Y est indépendante de X, alors $\mathbb{E}^X(Y) = \mathbb{E}(Y)$.
- Pour toute fonction $g : \mathbb{R}^n \to \mathbb{R}^n$ mesurable bornée, $\mathbb{E}^X(g(X)Y) = g(X)\mathbb{E}^X(Y)$
- L'espérance conditionnelle préserve la moyenne (Propriété de l'espérance totale) : $\mathbb{E}\left[\mathbb{E}^X(Y)\right] = \mathbb{E}(Y)$
- L'espérance conditionnelle est une fonction croissante en Y.

Démonstration.

- Il s'agit de l'inégalité de Pythagore
- Il s'agit de dire que si π est une projection orthogonale sur un sous-espace F de H, alors $(Id - \pi)(x)$ est toujours orthogonale à F, c'est-à-dire $\forall (x, z) \in H^2, < \pi(z), x - \pi(x) >= 0$. Ceci implique alors que :

  $$\mathbb{E}(h(X)Y) =< h(X), Y >=< h(X), \mathbb{E}^X(Y) >= \mathbb{E}(h(X)\mathbb{E}^X(Y)).$$

  Le caractère suffisant de cette égalité pour les fonctions bornée vient de la densité de cet ensemble.
- L'on applique le point précédent en utilisant l'indépendance pour dire que pour toute fonction h Borel-mesurable et bornée, $h(X)$ et Y sont indépendantes, donc :

  $$\mathbb{E}(h(X)\mathbb{E}^X(Y)) = \mathbb{E}(h(X)Y) = \mathbb{E}(h(x))\mathbb{E}(Y) = \mathbb{E}\left[h(X)\mathbb{E}(Y)\right]$$

  Comme $\mathbb{E}(Y)$ est une constante, elle est bien $\mathcal{T}_X$ mesurable.
- Pour g mesurable bornée, l'on se donne h mesurable bornée et l'on vérifie que

  $$\mathbb{E}(h(X)g(X)\mathbb{E}^X(Y)) = \mathbb{E}(h(X)g(X)Y) = \mathbb{E}(h(X)\mathbb{E}^X(g(X)Y))$$

  L'on peut alors conclure avec le deuxième point comme $g(X)\mathbb{E}^X(Y)$ est $\mathcal{T}_X$ mesurable.
- L'on applique la deuxième propriété à la fonction constante $h = 1$.
- L'on commence par remarquer que dans le point 2 l'on pourrait remplacer l'hypothèse d'application h borel-mesurable borné par h borel-mesurable borné et positive.
  La conclusion découle alors de la croissance de l'intégrale.

□

L'espérance conditionnelle de Y sachant X est la meilleure approximation possible de Y avec juste la connaissance de X

Le deuxième point de la propriété est central. Il permet de généraliser l'espérance conditionnelle à des variables aléatoires seulement continues. Nous admettrons ce point général, et donnerons une expression particulière dans le cas d'une probabilité absolument continue vis-à-vis d'une probabilité issue d'un couple de variable indépendante.

Proposition 89 :
Soit (X, Y) deux variables aléatoires, et l'on suppose que la loi du couple vu comme une mesure est à densité par rapport à une mesure produit, de densité $(x, y) \mapsto f(x, y)d\nu(x)d\mu(y)$.
Alors si l'on note $k(x)d\nu(x) := \int_{\mathbb{R}} f(x, y)d\mu(y)$ la loi marginale de X, l'on aura :

$$\mathbb{E}^X(Y) = \int_{\mathbb{R}} y\frac{f(X, y))}{k(x)}\mathbb{1}_{k(x)\neq 0}d\mu(y).$$

Démonstration. Supposons X et Y de carré intégrable.
Posons $g_Y(X) = \int_{\mathbb{R}} y\frac{f(X,y))}{k}\mathbb{1}_{k\neq 0}d\mu(y)$ et donnons-nous h une application borel-mesurable bornée.
L'on calcule :

$$\begin{aligned}\mathbb{E}(h(X)g_Y(X)) &= \int_R h(x)g_Y(x)k(x)d\nu(x)\\ &= \int_R h(x)\int_{\mathbb{R}} y\frac{f(x,y))}{h(x)}\mathbb{1}_{h(x)\neq 0}d\mu(y)k(x)d\nu(x)\\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} h(x)yf(x,y)d\nu(x)d\mu(y)\\ &= \mathbb{E}(h(X)Y)\end{aligned}$$

Ce qui montre bien que $g_Y(X) = \mathbb{E}^X(Y)$ avec le deuxième point de la propriété précédente. □

Exemple : Lorsque X et Y sont des variables aléatoires discrètes, l'on aura

$$\mathbb{E}^X(Y)(x_i) = \sum_j y_j\mathbb{P}(Y = y_j|X = x_i)$$

Et l'on retrouve bien l'espérance par rapport à la loi conditionnelle à $\{X = x_i\}$.
En fait, plus généralement, si A est un événement de probabilité non nulle et que $X = \mathbb{1}_A$, on trouvera que l'espérance conditionnelle par rapport à X est l'espérance par rapport à la loi conditionnelle de A.

$$\mathbb{E}^X(Y) = \mathbb{E}_{\mathbb{P}(\cdot|A)}(Y)$$

Bibliographie

[Bir13] Lucien Birgé : *Robust tests for model selection.* Inst. Math. Stat., 2013.

[CS04] Imre Csiszár et Paul Shields : *Information Theory and Statistics : A Tutorial.* Now Foundations and Trends, 2004.

[Dud89] R. M. Dudley : *Real Analysis and Probability.* Cambridge Studies in Advanced Mathematics, 1989.

[Eld07] Lars Eldén : *Matrix Methods in Data Mining and Pattern Recognition.* Society for Industrial and Applied Mathematics, 2007.

[GvdV17] Subhashis Ghosal et Aad van der Vaart : *Fundamentals of Nonparametric Bayesian Inference.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017.

[HJ90] Roger A. Horn et Charles R. Johnson : *Matrix Analysis.* Cambridge University Press, 1990.

[LC98] Erich L Lehmann et George Casella : *Theory of point estimation (Second edition).* Springer Science & Business Media, 1998.

[Liu09] Jun Liu : *Monte Carlo Strategies in Scientic Computing.* Springer series in statistics, 02 2009.

[LPW06] David A. Levin, Yuval Peres et Elizabeth L. Wilmer : *Markov chains and mixing times.* American Mathematical Society, 2006.

[MK62] A. Stuart M. Kendall : *The Advanced Theory of Statistics. Volume 2, Inference and Relationship*, volume 2. Journal of the Royal Statistical Society Series C, 1962.

[PB76] K. Doksum P. Bickel : *Mathematical statistics : Basic ideas and selected topics.* Holden-Day, 1976.

[Rob01] Christian Robert : *The Bayesian Choice.* Springer series in statistics, 01 2001.

[Vaa98] A. W. van der Vaart : *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.