Statistique

Slides 1ère année SN

Corinne Mailhes et Jean-Yves Tourneret[1]

(1) University of Toulouse, ENSEEIHT-IRIT-TéSA
Corinne.Mailhes@tesa.prd.fr et jyt@n7.fr

Année $2023 - 2024$

Bibliographie

Quelques références

- B. Lacaze, M. Maubourguet, C. Mailhes et J.-Y. Tourneret, Probabilités et Statistique appliquées, Cépadues, 1997.
- Athanasios Papoulis and S. Unnikrishna Pillai, Probability, Random Variable and Stochastic Processes, McGraw Hill Higher Education, 4th edition, 2002.
- E. L. Lehmann and G. Casella, Theory of Point Estimation, Springer Texts in Statistics, Springer-Verlag, New-York, 2nd edition, 1998.
- H. Van Trees, Detection, Estimation and Modulation Theory, Part I, John Wiley & Sons, New-York 1968.
- S. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice-Hall, Upper Saddle River, New-Jersey, 1993.
- S. Kay, Fundamentals of Statistical Signal Processing: Detection Theory, Prentice-Hall, New-York, 1998.

Plan du cours

Résumé

Modèle statistique

Notations

- Observations

$$x_1, ..., x_n$$

- Échantillon

$$X_1, ..., X_n$$

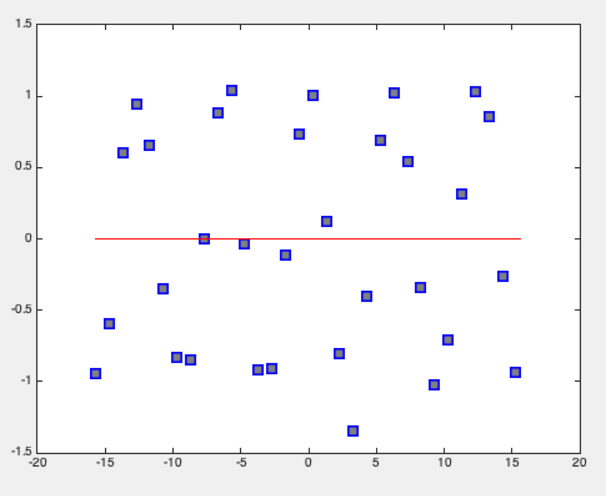
  n va iid associées aux observations

- Estimateur

$$\widehat{\theta}(X_1, ..., X_n) \text{ ou } \widehat{\theta}_n \text{ ou } \widehat{\theta}$$

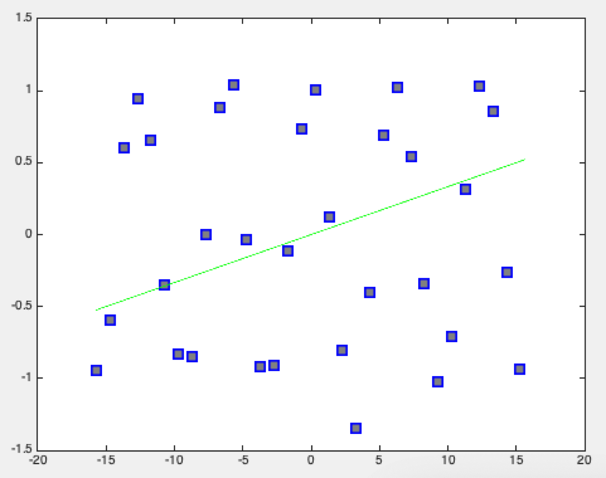
Exemple

Modèle 1 : $x_i = a + e_i$ avec $e_i \sim \mathcal{N}(0, \sigma^2)$



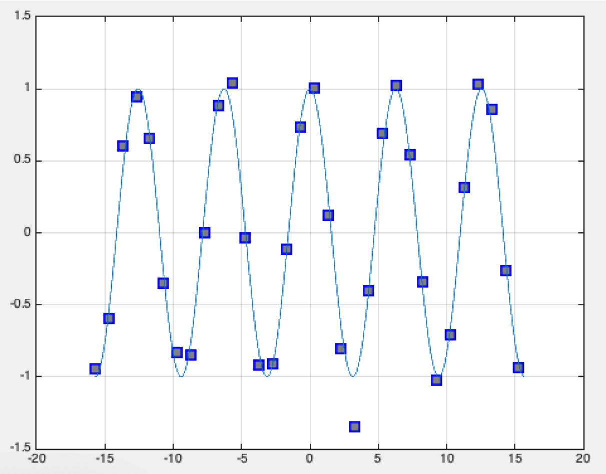
Exemple

Modèle 2 : $x_i = ai + b + e_i$ avec $e_i \sim \mathcal{N}(0, \sigma^2)$



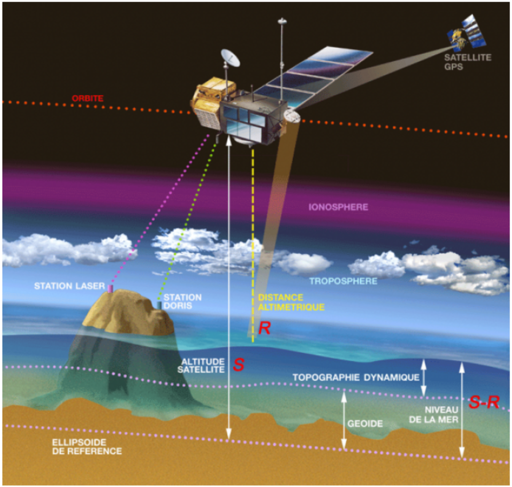
Exemple

Modèle 3 : $x_i = a\cos(i\phi) +$ avec $e_i \sim \mathcal{N}(0, \sigma^2)$



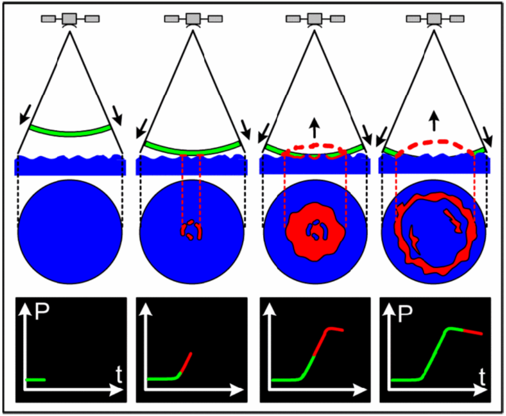
Application réelle

Altimétrie



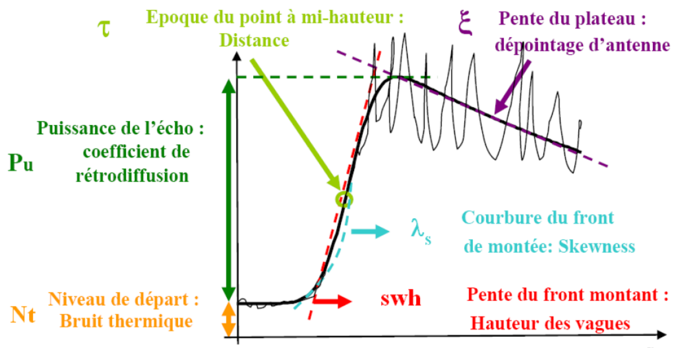
Application réelle

Formation de l'écho altimétrique



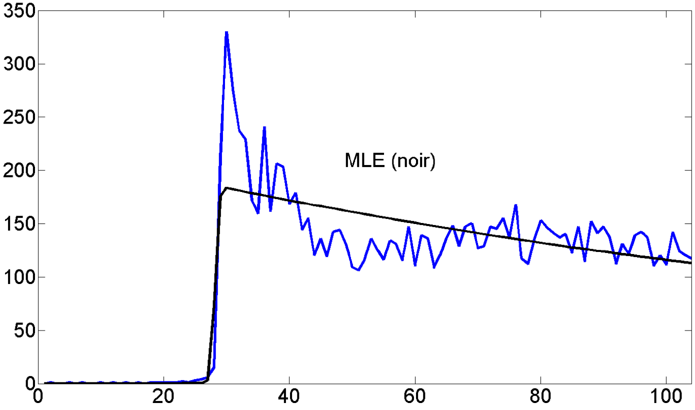
Application réelle

Modèle de Brown



Application réelle

Modèle de Brown



Qualités d'un estimateur

$\theta \in \mathbb{R}$

- Biais (erreur systématique) : $b_n(\theta) = E\left(\widehat{\theta}_n\right) - \theta$
- Variance

$$v_n(\theta) = E\left[\left(\widehat{\theta}_n - E\left(\widehat{\theta}_n\right)\right)^2\right] = E\left[\widehat{\theta}_n^2\right] - E\left(\widehat{\theta}_n\right)^2$$

- Erreur quadratique moyenne (Mean Square Error, MSE) (précision)

$$e_n(\theta) = E\left[\left(\widehat{\theta}_n - \theta\right)^2\right] = v_n(\theta) + b_n^2(\theta)$$

CS de convergence : $\widehat{\theta}_n$ est un estimateur convergent si $\lim\limits_{n\to+\infty} b_n(\theta) = \lim\limits_{n\to+\infty} v_n(\theta) = 0$

Qualités d'un estimateur

$\theta \in \mathbb{R}^p$

- Biais

$$b_n(\boldsymbol{\theta}) = E\left(\widehat{\boldsymbol{\theta}}_n\right) - \boldsymbol{\theta} \in \mathbb{R}^p$$

- Matrice de covariance

$$E\left[\left(\widehat{\boldsymbol{\theta}}_n - E\left(\widehat{\boldsymbol{\theta}}_n\right)\right)\left(\widehat{\boldsymbol{\theta}}_n - E\left(\widehat{\boldsymbol{\theta}}_n\right)\right)^T\right]$$

Exemples

Exemple 1 : $X_i \sim \mathcal{N}(m, \sigma^2)$, $\theta = m$ et σ^2 connue

- Moyenne empirique

$$\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \triangleq \overline{X}_n$$

- Autre estimateur

$$\widetilde{\theta}_n = \frac{2}{n(n+1)} \sum_{i=1}^{n} i X_i$$

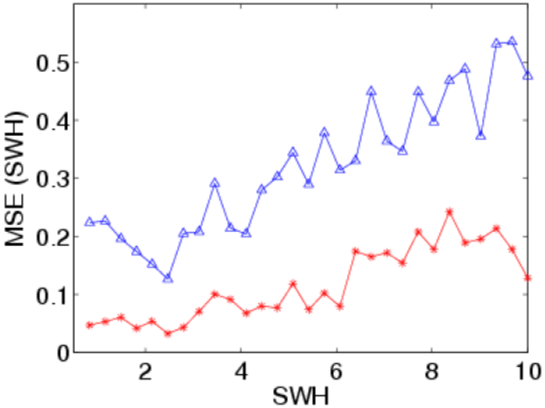
Exemple 2 : $X_i \sim \mathcal{N}(m, \sigma^2)$, $\theta = (m, \sigma^2)^T$

Étude de l'estimateur

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

Mean Square Errors

Comparaison de deux estimateurs



Plan du cours

Résumé

- Chapitre 1 : Estimation
    - Modèle statistique, qualités d'un estimateur, exemples
    - Inégalité de Cramér Rao
    - Maximum de vraisemblance
    - Méthode des moments
    - Estimation Bayésienne
    - Intervalles de confiance
- Chapitre 2 : Tests Statistiques
    - Généralités, exemple
    - Courbes COR, p-valeur
    - Théorème de Neyman Pearson
    - Autres tests paramétriques
    - Test du χ^2
    - Test de Kolmogorov

Inégalité de Cramér-Rao

Vraisemblance

$$L(x_1, ..., x_n; \theta) = \begin{cases} X_i \text{ va discrète : } P[X_1 = x_1, ..., X_n = x_n; \theta] \\ X_i \text{ va continue : } p(x_1, ..., x_n; \theta) \end{cases}$$

Inégalité pour $\boldsymbol{\theta} \in \mathbb{R}$

- Définition

$$\mathrm{Var}\left(\widehat{\theta}_n\right) \geq \frac{\left[1 + b_n'(\theta)\right]^2}{-E\left[\frac{\partial^2 \ln L(X_1,...,X_n;\theta)}{\partial \theta^2}\right]} = \mathrm{BCR}(\theta)$$

  $\mathrm{BCR}(\theta)$ est appelée Borne de Cramér Rao de θ

- Hypothèses
  Log-vraisemblance deux fois dérivable et support de la loi indépendant de θ (contre-exemple : loi $\mathcal{U}[0, \theta]$)

Remarques

Efficacité

Estimateur sans biais tel que $\mathrm{Var}\left(\widehat{\theta}_n\right) = \mathrm{BCR}(\theta)$ (Il est unique !)

Exemple : $X_i \sim \mathcal{N}(m, \sigma^2)$, $\theta = m$ et σ^2 connue

Cas où $(X_1, ..., X_n)$ est un échantillon

$$\mathrm{Var}\left(\widehat{\theta}_n\right) \geq \frac{\left[1 + b_n'(\theta)\right]^2}{-nE\left[\frac{\partial^2 \ln L(X_1;\theta)}{\partial \theta^2}\right]} = \mathrm{BCR}(\theta)$$

Cas multivarié

Inégalité pour un estimateur non biaisé de $\boldsymbol{\theta} \in \mathbb{R}^p$

- **Définition**

$$\operatorname{Cov}\left(\widehat{\boldsymbol{\theta}}\right) \geq I_n^{-1}(\boldsymbol{\theta})$$

avec

$$I_{ij} = E\left[-\frac{\partial^2 \ln L(X_1, ..., X_n; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right], \quad i, j = 1, ..., p$$

et $A \geq B$ signifie $A - B$ matrice semi définie positive

$$x^T(A - B)x \geq 0, \quad \forall x \in \mathbb{R}^p$$

On en déduit

$$\operatorname{Var}\left(\widehat{\theta_i}\right) \geq \left[I_n^{-1}(\boldsymbol{\theta})\right]_{ii}$$

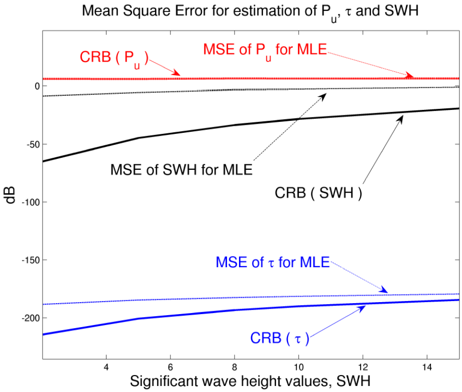
- **Exemple**

$$X_i \sim \mathcal{N}(m, \sigma^2), \boldsymbol{\theta} = (m, \sigma^2)^T$$

.

Exemple

Comparaison des variances d'estimateurs avec les bornes



Plan du cours

Résumé

- Chapitre 1 : Estimation
  - Modèle statistique, qualités d'un estimateur, exemples
  - Inégalité de Cramér Rao
  - Maximum de vraisemblance
  - Méthode des moments
  - Estimation Bayésienne
  - Intervalles de confiance
- Chapitre 2 : Tests Statistiques
  - Généralités, exemple
  - Courbes COR, p-valeur
  - Théorème de Neyman Pearson
  - Autres tests paramétriques
  - Test du χ^2
  - Test de Kolmogorov

Plan du cours

Résumé

- Chapitre 1 : Estimation
  - Modèle statistique, qualités d'un estimateur, exemples
  - Inégalité de Cramér Rao
  - Maximum de vraisemblance
  - Méthode des moments
  - Estimation Bayésienne
  - Intervalles de confiance
- Chapitre 2 : Tests Statistiques
  - Généralités, exemple
  - Courbes COR, p-valeur
  - Théorème de Neyman Pearson
  - Autres tests paramétriques
  - Test du χ^2
  - Test de Kolmogorov

Estimateur du maximum de vraisemblance

Définition

$$\widehat{\boldsymbol{\theta}}_{\mathsf{MV}} = \arg\max_{\theta} L(X_1, ..., X_n; \boldsymbol{\theta})$$

Recherche du maximum pour $\theta \in \mathbb{R}$

Si $L(X_1, ..., X_n; \theta)$ est régulière, on résoud

$$\frac{\partial L(X_1, ..., X_n; \theta)}{\partial \theta} = 0 \text{ ou } \frac{\partial \ln L(X_1, ..., X_n; \theta)}{\partial \theta} = 0$$

et on vérifie qu'on a bien un maximum en faisant un tableau de variations ou si ce n'est pas possible en étudiant

$$\frac{\partial^2 \ln L(X_1, ..., X_n; \widehat{\theta}_{\mathsf{MV}})}{\partial \theta^2} < 0$$

Régularité

Définition

On dit qu'une variable aléatoire X de densité de probabilité $f(x;\theta)$ est régulière si (voir livre de Lehmann, Theory of Point Estimation)

- Le support de la densité f, i.e., $\{x|f(x;\theta) > 0\}$, est indépendant de θ
- $f(x;\theta)$ est au moins trois fois dérivable par rapport à θ
- La vraie valeur de θ appartient à un ensemble compact Θ.

Dans ce cas, la recherche de l'estimateur du maximum de vraisemblance peut se faire en cherchant les racines de

$$\frac{\partial L(X_1, ..., X_n;\theta)}{\partial\theta} = 0 \text{ ou } \frac{\partial \ln L(X_1, ..., X_n;\theta)}{\partial\theta} = 0$$

Estimateur du maximum de vraisemblance

Recherche du maximum pour $\theta \in \mathbb{R}^p$

$$\frac{\partial L(X_1, ..., X_n; \theta)}{\partial \theta_i} = 0 \text{ ou } \frac{\partial \ln L(X_1, ..., X_n; \theta)}{\partial \theta_i} = 0$$

pour $i = 1, ..., p$

Exemples

- Exemple 1 : $X_i \sim \mathcal{P}(\lambda)$, $\theta = \lambda$
- Exemple 2 : $X_i \sim \mathcal{N}(m, \sigma^2)$, $\boldsymbol{\theta} = (m, \sigma^2)^T$

Propriétés

- Estimateur asymptotiquement non biaisé

$$\lim_{n \to +\infty} E\left[\widehat{\boldsymbol{\theta}}_{\text{MV}}\right] - \boldsymbol{\theta} = 0$$

- Estimateur convergent
- Estimateur asymptotiquement efficace

$$\lim_{n \to +\infty} \frac{\operatorname{Var}\left(\widehat{\theta}_i\right)}{\left[I_n^{-1}(\boldsymbol{\theta})\right]_{ii}} = 1$$

- Normalité Asymptotique
- Invariance Fonctionnelle
  Si $\boldsymbol{\mu} = h(\boldsymbol{\theta})$, où h est une fonction bijective d'un ouvert $O \subset \mathbb{R}^p$ dans un ouvert $V \subset \mathbb{R}^p$, alors

$$\widehat{\boldsymbol{\mu}}_{\text{MV}} = h\left(\widehat{\boldsymbol{\theta}}_{\text{MV}}\right)$$

Conclusion

L'estimateur du maximum de vraisemblance possède beaucoup de bonnes propriétés asymptotiques mais peut être difficile à étudier car il est la solution d'un problème d'optimisation.

Remarques sur la convergence

Thèorème (voir, e.g., livre de Lehmann, Theory of Point Estimation)

Soient $X_1, ..., X_n$ des variables aléatoires iid de même densité $f(x_i; \theta)$ avec

- θ appartient à un ouvert $\Theta \in \mathbb{R}$
- le paramètre θ est identifiable, i.e., deux valeurs différentes de θ donnent des densités $f(x_i; \theta)$ différentes
- la log-vraisemblance $l(\theta)$ est dérivable par rapport à θ
- le support de la densité f ne dépend pas de θ

alors l'équation $l'(\theta) = 0$ admet une solution qui converge en probabilité vers θ_0 (pas nécessairement $\hat{\theta}_{\text{MV}}$). Donc s'il y a une unique solution de $l'(\theta) = 0$ et que cette solution maximise la vraisemblance, alors cette solution est l'estimateur du maximum de vraisemblance de θ qui est un estimateur convergent.

Exemple d'estimateur $\widehat{\boldsymbol{\theta}}_{\text{MV}}$ non convergent (avec plusieurs maxima locaux de $l'(\theta) = 0$)

$$f(x_i|\theta) = \frac{1}{2}\mathcal{N}(0,1) + \frac{1}{2}\mathcal{N}\left(\theta, [\exp(-1/\theta^2)]^2\right)$$

Plan du cours

Résumé

Méthode des moments

Définition

Supposons que $X_1, ..., X_n$ ont la même loi de paramètre inconnu $\boldsymbol{\theta} \in \mathbb{R}^p$. En général, le vecteur paramètre à estimer $\boldsymbol{\theta}$ est lié aux premiers moments de la loi des va X_i par une relation notée

$$\boldsymbol{\theta} = h(m_1, ..., m_q)$$

avec $m_k = E[X_i^k]$ et $q \geq p$. Un estimateur des moments de $\boldsymbol{\theta}$ est défini par

$$\widehat{\boldsymbol{\theta}}_{\text{Mo}} = h(\widehat{m}_1, ..., \widehat{m}_q) \text{ avec } \widehat{m}_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

Exemples

- $X_i \sim \mathcal{N}(m, \sigma^2)$, $\boldsymbol{\theta} = (m, \sigma^2)^T$
- $X_i \sim \Gamma(a, b)$, $\boldsymbol{\theta} = (a, b)^T$

Méthode des moments

Propriétés

- Estimateur convergent
- Normalité Asymptotique

Conclusion

L'estimateur des moments possède peu de propriétés mais est généralement simple à déterminer.

Plan du cours

Résumé

- Chapitre 1 : Estimation
  - Modèle statistique, qualités d'un estimateur, exemples
  - Inégalité de Cramér Rao
  - Maximum de vraisemblance
  - Méthode des moments
  - Estimation Bayésienne
  - Intervalles de confiance
- Chapitre 2 : Tests Statistiques
  - Généralités, exemple
  - Courbes COR, p-valeur
  - Théorème de Neyman Pearson
  - Autres tests paramétriques
  - Test du χ^2
  - Test de Kolmogorov

Estimation Bayésienne

Principe

L'estimation Bayésienne consiste à estimer un vecteur paramètre inconnu $\boldsymbol{\theta} \in \mathbb{R}^p$ à l'aide de la vraisemblance de $X_1, ..., X_n$ (paramétrée par $\boldsymbol{\theta}$) et d'une loi *a priori* $p(\boldsymbol{\theta})$. Pour cela, on minimise une fonction de coût $c\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right)$ qui représente l'erreur entre $\boldsymbol{\theta}$ et $\widehat{\boldsymbol{\theta}}$. Deux estimateurs principaux

- Estimateur MMSE : c'est la moyenne de la loi a posteriori

$$\widehat{\boldsymbol{\theta}}_{\text{MMSE}} = E\left(\boldsymbol{\theta} | X_1, ..., X_n\right)$$

- Estimateur MAP : l'estimateur du maximum a posteriori (MAP) de $\boldsymbol{\theta}$ est défini par

$$\widehat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} | X_1, ..., X_n)$$

où $p\left(\boldsymbol{\theta} | x_1, ..., x_n\right)$ est la loi *a posteriori* de $\boldsymbol{\theta}$.

Propriétés des estimateurs Bayésiens

Estimateur MMSE

L'estimateur MMSE minimise l'erreur quadratique moyenne (mean square error, MSE)

$$c\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right) = E\left[\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^T \left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)\right]$$

Estimateur MAP

L'estimateur MAP minimise la fonction de coût $E\left[c\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right)\right]$ avec entre $\boldsymbol{\theta}$ et $\widehat{\boldsymbol{\theta}}$

$$c\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right) = \begin{cases} 1 \text{ si } \left\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right\| > \Delta \\ 0 \text{ si } \left\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right\| < \Delta \end{cases}$$

avec Δ arbitrairement petit (Preuve : voir par exemple livre de H. Van Trees, Detection, Estimation, and Modulation Theory, Part I).

Estimation Bayésienne

Exemple

- Vraisemblance

$$X_i \sim \mathcal{N}(\theta, \sigma^2)$$

- Loi a priori

$$\theta \sim \mathcal{N}(\mu, \nu^2)$$

Solution

- Loi a posteriori

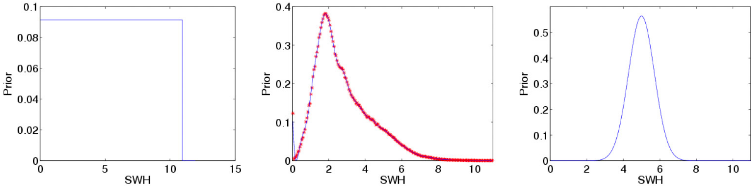
$$\theta | X_1, ..., X_n \sim \mathcal{N}\left(m_p, \sigma_p^2\right)$$

- Estimateurs

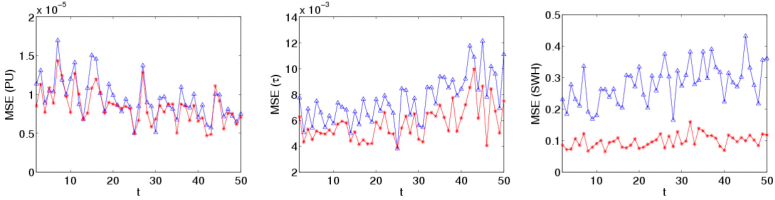
$$\widehat{\theta}_{\text{MAP}} = \widehat{\theta}_{\text{MMSE}} = m_p = \overline{X}\left(\frac{n\nu^2}{n\nu^2 + \sigma^2}\right) + \mu\left(\frac{\sigma^2}{\sigma^2 + n\nu^2}\right)$$

Avec ou sans prior ?

Examples for the Significant Wave Height (SWH)



Dynamic priors



Plan du cours

Résumé

- Chapitre 1 : Estimation
  - Modèle statistique, qualités d'un estimateur, exemples
  - Inégalité de Cramér Rao
  - Maximum de vraisemblance
  - Méthode des moments
  - Estimation Bayésienne
  - Intervalles de confiance
- Chapitre 2 : Tests Statistiques
  - Généralités, exemple
  - Courbes COR, p-valeur
  - Théorème de Neyman Pearson
  - Autres tests paramétriques
  - Test du χ^2
  - Test de Kolmogorov

Intervalle de confiance

Principe

Un intervalle de confiance $[a, b]$ pour le paramètre $\theta \in \mathbb{R}$ est un intervalle tel que $P[a < \theta < b] = \alpha$, où α est le paramètre de confiance (en général $\alpha = 0.99$ ou $\alpha = 0.95$).

Détermination pratique de l'intervalle

- On cherche un estimateur de θ noté $\widehat{\theta}$ (par la méthode des moments, du maximum de vraisemblance, ...)
- On en déduit une statistique $T(X_1, ..., X_n)$ qui dépend de θ de loi connue
- On cherche $c(\theta)$ et $d(\theta)$ tels que

$$P[c(\theta) < T(X_1, ..., X_n) < d(\theta)] = \alpha$$

On en déduit l'intervalle $[a, b]$.

Exemples

- Exemple 1 : $X_i \sim \mathcal{N}(m, \sigma^2)$, m inconnue, σ^2 connue.

$$T = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - m}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

- Exemple 2 : $X_i \sim \mathcal{N}(m, \sigma^2)$, intervalles de confiance pour m et σ^2 inconnue.
  - Moyenne

$$T \sim \mathcal{N}(0,1) \quad \text{et} \quad U = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \sim \chi^2_{n-1}$$

  donc

$$\frac{T}{\sqrt{\frac{U}{n-1}}} \sim t_{n-1}$$

  suit une loi de Student à $n-1$ degrés de liberté.
  - Variance

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2_{n-1}$$

Que faut-il savoir ?

Estimation statistique

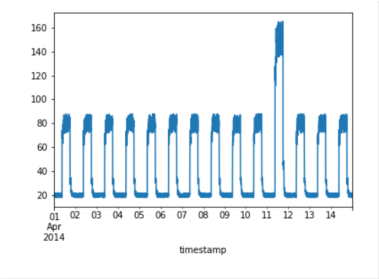
- Notions de biais, variance et convergence d'un estimateur
- Calcul d'une borne de Cramér-Rao et notion d'efficacité
- Détermination de l'estimateur du maximum de vraisemblance (MV)
- Propriétés de l'estimateur MV
- Principe et application de la méthode des moments
- Principe et application de l'estimation Bayésienne
- Détermination des intervalles de confiance

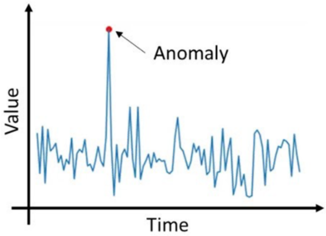
Plan du cours

Résumé

- Chapitre 1 : Estimation
  - Modèle statistique, qualités d'un estimateur, exemples
  - Inégalité de Cramér Rao
  - Maximum de vraisemblance
  - Méthode des moments
  - Estimation Bayésienne
  - Intervalles de confiance
- Chapitre 2 : Tests Statistiques
  - Généralités, exemple
  - Courbes COR, p-valeur
  - Théorème de Neyman Pearson
  - Autres tests paramétriques
  - Test du χ^2
  - Test de Kolmogorov

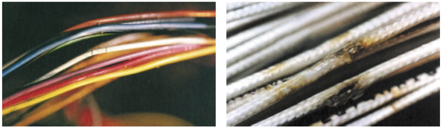
Motivations





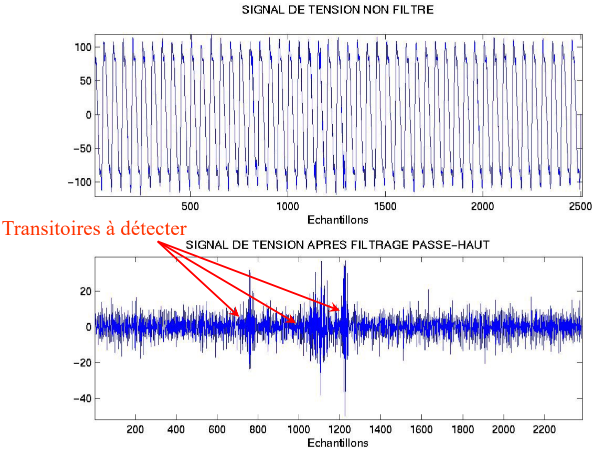
Motivations

Debris from TWA Flight 800 was pieced together following the fatal crash of the Boeing 747 in 1996. Government investigators concluded that the likely trigger was a short circuit from damaged wiring—wiring "not atypical for an airplane of its age."



Wiring taken from U.S. Navy aircraft show [left] cracks in the polyimide insulation that go through to the copper conductor, and [right] faults in PVC-insulated wire that had been hidden under a clamp; the discoloration indicates arcing had occurred.

Motivations



Généralités

Principe

Un test statistique est un mécanisme qui permet de décider entre plusieurs hypothèses $H_0, H_1, ...$ à partir de n observations $x_1, ..., x_n$. On se limitera dans ce cours à deux hypothèses H_0 et H_1. Effectuer un test, c'est déterminer une statistique de test $T(X_1, ..., X_n)$ et un ensemble Δ tel que

$$\begin{aligned} &\mathcal{H}_0 \text{ rejetée si } T(X_1, ..., X_n) \in \Delta \\ &\mathcal{H}_0 \text{ acceptée si } T(X_1, ..., X_n) \notin \Delta. \end{aligned} \tag{1}$$

Vocabulaire

- H_0 est l'hypothèse nulle
- H_1 est l'hypothèse alternative
- $\{(x_1, ..., x_n) | T(x_1, ..., x_n) \in \Delta\}$: région critique

Définitions

- Tests paramétriques et non paramétriques
- Hypothèses simples et hypothèses composites
- Risque de première espèce = probabilité de fausse alarme

$$\alpha = \mathrm{PFA} = P[\text{Rejeter } H_0 | H_0 \text{ vraie}]$$

- Risque de seconde espèce = probabilité de non-détection

$$\beta = \mathrm{PND} = P[\text{Rejeter } H_1 | H_1 \text{ vraie}]$$

- Puissance du test = probabilité de détection : $\pi = 1 - \beta$

Exemple

Changement de moyenne

$$X_i \sim \mathcal{N}(m, \sigma^2), \ \sigma^2 \text{ connue}$$

- Hypothèses

$$H_0 : m = m_0, \ H_1 : m = m_1 > m_0$$

- Exemple de test

$$\text{Rejet de } H_0 \text{ si } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i > S_\alpha$$

- Problèmes
  Déterminer le seuil S_α, le risque β et la puissance du test π.

Plan du cours

Résumé

Caractéristiques opérationnelles du récepteur (COR)

Définition

$$\mathrm{PD} = h\,(\mathrm{PFA})$$

Exemple

$X_i \sim \mathcal{N}(m, \sigma^2),\ \sigma^2$ connue

$$H_0 : m = m_0,\ H_1 : m = m_1 > m_0$$

- Probabilité de fausse alarme

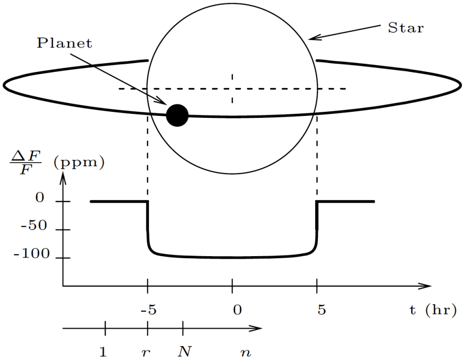
$$\alpha = 1 - F\left(\frac{S_\alpha - m_0}{\frac{\sigma}{\sqrt{n}}}\right) \Leftrightarrow S_\alpha = m_0 + \frac{\sigma}{\sqrt{n}} F^{-1}(1-\alpha)$$

- Probabilité de détection

$$\mathrm{PD} = \pi = 1 - F\left(\frac{S_\alpha - m_1}{\frac{\sigma}{\sqrt{n}}}\right)$$

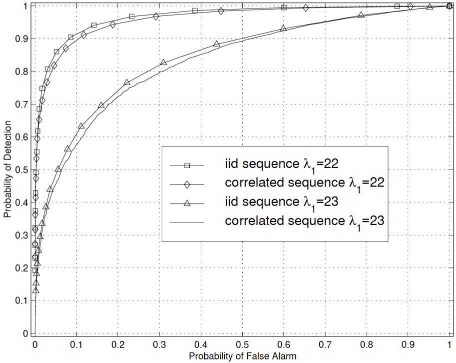
Exemple d'application

Détection d'exoplanètes



Exemple de représentation graphique pour les courbes COR

Données indépendantes ou corrélées ?



p-valeur d'un test

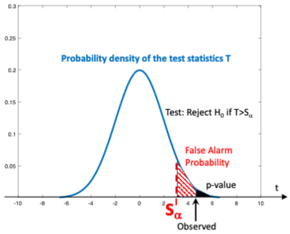
Définition

$$p(\boldsymbol{x}) = \inf\{\alpha \in]0,1[\mid \boldsymbol{x} \in \mathcal{R}_\alpha\}$$

où $\mathcal{R}_\alpha$ est la zone de rejet de H_0 pour α fixé et $\boldsymbol{x} = (x_1, ..., x_n)$. C'est la plus petite valeur de α pour laquelle on rejette H_0.

Calcul de la p-valeur pour le test : Rejet de H_0 si $T > S_\alpha$

- Si $\alpha = 0$, on accepte toujours H_0 donc $S_0 = +\infty$
- Si $\alpha = 1$, on rejette toujours H_0 donc $S_1 = -\infty$
- Plus petite valeur de α pour laquelle on rejette H_0 : $\alpha^* = 1 - F(T_{\text{obs}})$



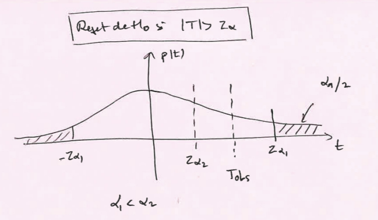
Autre exemple : rejet de H_0 si $|T| > S_\alpha$

Calcul

- Si $\alpha = 0$, on accepte toujours H_0 donc $z_0 = 0$
- Si $\alpha = 1$, on rejette toujours H_0 donc $z_1 = +\infty$
- Plus petite valeur de α pour laquelle on rejette H_0

$$\frac{\alpha^*}{2} = 1 - F(|T_{\text{obs}}|) \Leftrightarrow \alpha^* = 2[1 - F(|T_{\text{obs}})|].$$

Représentation graphique



Plan du cours

Résumé

- Chapitre 1 : Estimation
  - Modèle statistique, qualités d'un estimateur, exemples
  - Inégalité de Cramér Rao
  - Maximum de vraisemblance
  - Méthode des moments
  - Estimation Bayésienne
  - Intervalles de confiance
- Chapitre 2 : Tests Statistiques
  - Généralités, exemple
  - Courbes COR, p-valeur
  - Théorème de Neyman Pearson
  - Autres tests paramétriques
  - Test du χ^2
  - Test de Kolmogorov

Théorème de Neyman-Pearson

Test paramétrique à hypothèses simples

en pratique on utilise le TLC en approximant avec une loi normale

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \text{ et } H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1 \tag{2}$$

Théorème pour variables aléatoires X_i continues

À α fixé, le test qui minimise β (ou maximise π) est défini par

$$\text{Rejet de } H_0 \text{ si } \frac{L(x_1, ..., x_n|H_1)}{L(x_1, ..., x_n|H_0)} = \frac{p(x_1, ..., x_n|\boldsymbol{\theta}_1)}{p(x_1, ..., x_n|\boldsymbol{\theta}_0)} > S_\alpha$$

Exemple

$X_i \sim \mathcal{N}(m, \sigma^2),\ \sigma^2$ connue

$$H_0 : m = m_0,\ H_1 : m = m_1 > m_0$$

Théorème de Neyman-Pearson

Test paramétrique à hypothèses simples

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \text{ et } H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1 \tag{3}$$

Théorème pour variables aléatoires X_i discrètes

Parmi tous les tests de risque de première espèce $\leq \alpha$ fixé, le test de puissance maximale rejette l'hypothèse H_0 si

$$\frac{L(x_1, ..., x_n | H_1)}{L(x_1, ..., x_n | H_0)} = \frac{P[X_1 = x_1, ..., X_n = x_n | \boldsymbol{\theta}_1]}{P(X_1 = x_1, ..., X_n = x_n | \boldsymbol{\theta}_0)} > S_\alpha$$

Exemple : lois de Poisson $X_i \sim \mathcal{P}(\lambda)$

$$\lambda = \lambda_0, \ H_1 : \lambda = \lambda_1 > \lambda_0$$

Test de Neyman-Pearson

Résumé des différentes étapes

- 1) Déterminer la statistique et la région critique du test
- 2) Déterminer la relation entre le seuil S_α et le risque α
- 3) Calculer le risque β et la puissance π du test en fonction de α
- 4) (optionnel) Déterminer les caractéristiques opérationnelles du récepteur
- 5) Application numérique
  - On accepte ou rejette l'hypothèse H_0 en précisant le risque α donné
  - (optionnel) On détermine la p-valeur du test

Remarque

Loi asymptotique : quand n est suffisamment grand, utilisation du théorème de la limite centrale

Résumé

- Chapitre 1 : Estimation
  - Modèle statistique, qualités d'un estimateur, exemples
  - Inégalité de Cramér Rao
  - Maximum de vraisemblance
  - Méthode des moments
  - Estimation Bayésienne
  - Intervalles de confiance
- Chapitre 2 : Tests Statistiques
  - Généralités, exemple
  - Courbes COR, p-valeur
  - Théorème de Neyman Pearson
  - Autres tests paramétriques
  - Test du χ^2
  - Test de Kolmogorov

Test du rapport de vraisemblance généralisé (generalized likelihood ratio)

Test paramétrique à hypothèses composites

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \text{ et } H_1 : \boldsymbol{\theta} \in \Theta_1 \tag{4}$$

Définition du test

$$\text{Rejet de } H_0 \text{ si } \frac{L\left(x_1, ..., x_n | \widehat{\boldsymbol{\theta}}_1^{\text{MV}}\right)}{L\left(x_1, ..., x_n | \widehat{\boldsymbol{\theta}}_0^{\text{MV}}\right)} > S_\alpha$$

où $\widehat{\boldsymbol{\theta}}_0^{\text{MV}}$ et $\widehat{\boldsymbol{\theta}}_1^{\text{MV}}$ sont les estimateurs du maximum de vraisemblance de $\boldsymbol{\theta}$ sous les hypothèses H_0 et H_1.

Remarque

$$L\left(x_1, ..., x_n | \widehat{\boldsymbol{\theta}}_i^{\text{MV}}\right) = \sup_{\boldsymbol{\theta} \in \Theta_i} L\left(x_1, ..., x_n | \boldsymbol{\theta}\right)$$

Est-ce que la moyenne d'un échantillon gaussien augmente ? (σ^2 connue)

Soit $(X_1, ..., X_n)$ un échantillon gaussien de loi $\mathcal{N}(m, \sigma^2)$ avec une variance σ^2 connue. On considère les hypothèses

$$H_0 : m = m_0 \text{ et } H_1 : m = m_1 \text{ avec } m_1 > m_0 \tag{5}$$

Définition du test

$$\text{Rejet de } H_0 \text{ si } T = \sqrt{n}\left(\frac{\overline{X} - m_0}{\sigma}\right) > S_\alpha \quad \text{avec} \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Remarques

- T suit la loi normale $\mathcal{N}(0, 1)$ sous H_0
- Application directe de Neyman-Pearson
- Généralisation immédiate à $m_1 < m_0$ ou à $m_1 \neq m_0$

Est ce que la moyenne d'un échantillon gaussien augmente ? (σ^2 inconnue)

Soit $(X_1, ..., X_n)$ un échantillon gaussien de loi $\mathcal{N}(m, \sigma^2)$ avec une variance σ^2 inconnue et les hypothèses

$$H_0 : m = m_0 \text{ et } H_1 : m = m_1 \text{ avec } m_1 > m_0 \tag{6}$$

Définition du test

$$\text{Rejet de } H_0 \text{ si } T = \sqrt{n}\left(\frac{\overline{X} - m_0}{S_n}\right) > S_\alpha \quad \text{avec} \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Remarques

- Si on pose $U = \sqrt{n}\left(\frac{\overline{X} - m_0}{\sigma}\right)$ et $V = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2$, on a $T = \frac{U}{\sqrt{\frac{V}{n-1}}}$ qui suit une loi de Student à $n-1$ ddl sous H_0
- Généralisation immédiate à $m_1 < m_0$ ou à $m_1 \neq m_0$

Est-ce que la variance d'un échantillon gaussien augmente ? (m connue)

Soit $(X_1, ..., X_n)$ un échantillon gaussien de loi $\mathcal{N}(m, \sigma^2)$ avec une moyenne m connue et les hypothèses

$$H_0 : \sigma^2 = \sigma_0^2 \text{ et } H_1 : \sigma^2 = \sigma_1^2 \text{ avec } \sigma_1^2 > \sigma_0^2 \tag{7}$$

Définition du test

$$\text{Rejet de } H_0 \text{ si } T = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (X_i - m)^2 > S_\alpha$$

Remarques

- La loi de T sous H_0 est une loi du χ_n^2, ce qui permet de déterminer S_α en fonction de α.
- Généralisation immédiate à $\sigma_1^2 < \sigma_0^2$ ou à $\sigma_1^2 \neq \sigma_0^2$

Est-ce que la variance d'un échantillon gaussien augmente ? (m inconnue)

Soit $(X_1, ..., X_n)$ un échantillon gaussien de loi $\mathcal{N}(m, \sigma^2)$ avec une moyenne m inconnue et les hypothèses

$$H_0 : \sigma^2 = \sigma_0^2 \text{ et } H_1 : \sigma^2 = \sigma_1^2 \text{ avec } \sigma_1^2 > \sigma_0^2 \tag{8}$$

Définition du test

$$\text{Rejet de } H_0 \text{ si } T = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (X_i - \overline{X})^2 > S_\alpha \quad \text{avec} \quad \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Remarques

- La loi de T sous H_0 est une loi du χ^2_{n-1}, ce qui permet de déterminer S_α en fonction de α.
- Généralisation immédiate à $\sigma_1^2 < \sigma_0^2$ ou à $\sigma_1^2 \neq \sigma_0^2$

Est-ce que deux échantillons Gaussiens sont significativement différents ? (variances connues)

Soient $(X_1, ..., X_n)$ et $(Y_1, ..., Y_m)$ deux échantillons gaussiens indépendants de lois respectives $\mathcal{N}(m_1, \sigma_1^2)$ et $\mathcal{N}(m_2, \sigma_2^2)$ avec des variances σ_1^2 et σ_2^2 connues, et les hypothèses

$$H_0 : m_1 = m_2 \text{ et } H_1 : m_1 > m_2 \tag{9}$$

Définition du test

$$\text{Rejet de } H_0 \text{ si } T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} > S_\alpha \text{ avec } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \overline{Y} = \frac{1}{m}\sum_{j=1}^{m} Y_j$$

Remarques

- La loi de T sous H_0 est une loi normale $\mathcal{N}(0, 1)$, ce qui permet de déterminer S_α en fonction de α.
- Généralisation immédiate à $m_1 < m_2$ ou à $m_1 \neq m_2$

Est-ce que deux échantillons Gaussiens sont significativement différents ?

Soient $(X_1, ..., X_n)$ et $(Y_1, ..., Y_m)$ deux échantillons gaussiens indépendants de lois respectives $\mathcal{N}(m_1, \sigma_1^2)$ et $\mathcal{N}(m_2, \sigma_2^2)$ avec une même variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$ inconnue, et les hypothèses

$$H_0 : m_1 = m_2 \text{ et } H_1 : m_1 > m_2 \tag{10}$$

Définition du test

$$\text{Rejet de } H_0 \text{ si } T = \frac{\overline{X} - \overline{Y}}{S_{n,m}(\boldsymbol{x}, \boldsymbol{y})\sqrt{\frac{1}{n} + \frac{1}{m}}} > S_\alpha$$

avec

$$S_{n,m}^2(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^n (X_i - \overline{X})^2 + \sum_{j=1}^m (Y_j - \overline{Y})^2}{n + m - 2}$$

Remarques

- La loi de T sous H_0 est une loi de Student à $n + m - 2$ ddl.
- Appelé test de Student ou t-test.
- Généralisation immédiate à $m_1 < m_2$ ou à $m_1 \neq m_2$

Est-ce que deux échantillons Gaussiens sont significativement différents ?

Soient $(X_1, ..., X_n)$ et $(Y_1, ..., Y_m)$ deux échantillons gaussiens indépendants de lois respectives $\mathcal{N}(m_1, \sigma_1^2)$ et $\mathcal{N}(m_2, \sigma_2^2)$ avec des variances σ_1^2 et σ_2^2 inconnues, et les hypothèses

$$H_0 : m_1 = m_2 \text{ et } H_1 : m_1 > m_2 \tag{11}$$

Définition du test

$$\text{Rejet de } H_0 \text{ si } T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_n^2(\boldsymbol{x})}{n} + \frac{S_m^2(\boldsymbol{y})}{m}}} > S_\alpha$$

avec

$$S_n^2(\boldsymbol{x}) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ et } S_m^2(\boldsymbol{y}) = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y})^2$$

Remarques

- Sous l'hypothèse H_0, T converge en loi vers une loi normale $\mathcal{N}(0, 1)$ lorsque n et m tendent vers $+\infty$.
- Généralisation immédiate à $m_1 < m_2$ ou à $m_1 \neq m_2$

Comparaison d'espérances

Remarques générales

- Les tests précédents supposent que les deux échantillons $(X_1, ..., X_n)$ et $(Y_1, ..., Y_m)$ sont indépendants. Si ce n'est pas le cas et que $n = m$, on parle de données appariées. On peut alors considérer les différences $Z_i = X_i - Y_i$ et tester la nullité de l'espérance des Z_i.
- Si l'hypothèse de gaussiannité n'est pas satisfaite, on pourra effectuer un test non paramétrique.

Est-ce que la variance de deux échantillons gaussiens a augmenté ?

Soient $(X_1, ..., X_n)$ et $(Y_1, ..., Y_m)$ deux échantillons gaussiens indépendants de lois respectives $\mathcal{N}(m_1, \sigma_1^2)$ et $\mathcal{N}(m_2, \sigma_2^2)$ avec des moyennes m_1 et m_2 connues, et les hypothèses

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ et } H_1 : \sigma_1^2 > \sigma_2^2 \tag{12}$$

Définition du test

$$\text{Rejet de } H_0 \text{ si } T = \frac{\widetilde{S}_n^2(\boldsymbol{x})}{\widetilde{S}_m^2(\boldsymbol{y})} > S_\alpha$$

avec

$$\widetilde{S}_n^2(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}(X_i - m_1)^2 \ \text{ et } \ \widetilde{S}_m^2(\boldsymbol{y}) = \frac{1}{m}\sum_{j=1}^{m}(Y_j - m_2)^2$$

Remarques

- Sous l'hypothèse H_0, T est distribuée suivant une loi de Fisher $\mathcal{F}(n, m)$
- Appelé F-test.
- Généralisation immédiate à $\sigma_1^2 < \sigma_2^2$ ou à $\sigma_1^2 \neq \sigma_2^2$

Est-ce que la variance de deux échantillons gaussiens a augmenté ?

Soient $(X_1, ..., X_n)$ et $(Y_1, ..., Y_m)$ deux échantillons gaussiens indépendants de lois respectives $\mathcal{N}(m_1, \sigma_1^2)$ et $\mathcal{N}(m_2, \sigma_2^2)$ avec des moyennes m_1 et m_2 inconnues, et les hypothèses

$$H_0 : \sigma_1^2 = \sigma_1^2 \text{ et } H_1 : \sigma_1^2 > \sigma_2^2 \tag{13}$$

Définition du test

$$\text{Rejet de } H_0 \text{ si } T = \frac{S_n^2(\boldsymbol{x})}{S_m^2(\boldsymbol{y})} > S_\alpha$$

avec

$$S_n^2(\boldsymbol{x}) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ et } S_m^2(\boldsymbol{y}) = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y})^2$$

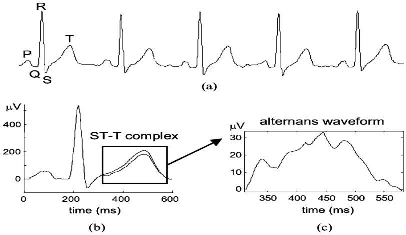
Remarques

- Sous l'hypothèse H_0, T est distribuée suivant une loi de Fisher $\mathcal{F}(n-1, m-1)$
- Généralisation immédiate à $\sigma_1^2 < \sigma_2^2$ ou à $\sigma_1^2 \neq \sigma_2^2$

Example: T-wave Alternans (TWA) Detection

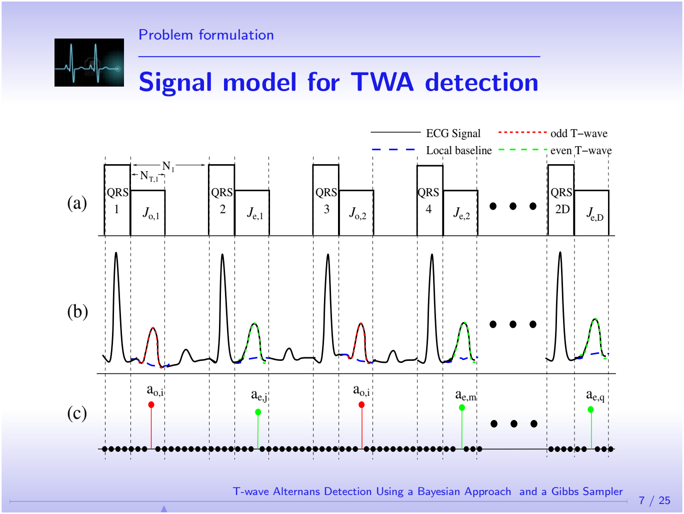
Context

T-wave alternans (TWA) detection



- TWA: a consistent fluctuation in the T waves on an every-other-beat basis
- recognized as an index of sudden cardiac death
- a challenging problem: non-visible (microvolt-level) TWA detection

Signal Model



Estimation Results

Simulation

Estimation results

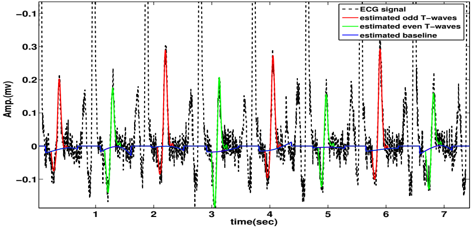


Figure: Segment of dataset "e0303" with synthetic TWA and "ma" noise SNR=10dB (black), estimated local baseline (blue), and estimated odd (red) and even (green) T waves. Processing window length $2D = 16$.

Statistical Test

Bayesian TWA detection

Two-sample Student's t-test

- Based on the assumption of normality of two samples:

$$\mathcal{H}_0 : \mu_\mathrm{o} = \mu_\mathrm{e}, \quad \mathcal{H}_1 : \mu_\mathrm{o} \neq \mu_\mathrm{e}$$

  μ_o and μ_e are the means of the odd and even T-wave amplitude samples.

- The t-test statistic can be computed as follows:

$$t^{(i)} = \frac{\bar{\mathbf{a}}_\mathrm{o}^{(i)} - \bar{\mathbf{a}}_\mathrm{e}^{(i)}}{S_\mathrm{eo}^{(i)} \sqrt{\frac{2}{D}}} \tag{1}$$

$\bar{\mathbf{a}}_\mathrm{o}^{(i)} = \frac{1}{D} \sum_{k=1}^{D} a_{\mathrm{o,k}}^{(i)}$, $\bar{\mathbf{a}}_\mathrm{e}^{(i)} = \frac{1}{D} \sum_{k=1}^{D} a_{\mathrm{e,k}}^{(i)}$ and

$$S_\mathrm{eo}^{(i)} = \sqrt{\frac{1}{2D-2} \left(\sum_{k=1}^{D} \left(a_{\mathrm{o,k}}^{(i)} - \bar{\mathbf{a}}_\mathrm{o}^{(i)} \right)^2 + \sum_{k=1}^{D} \left(a_{\mathrm{e,k}}^{(i)} - \bar{\mathbf{a}}_\mathrm{e}^{(i)} \right)^2 \right)}.$$

Detection Results

Simulation

Test decisions

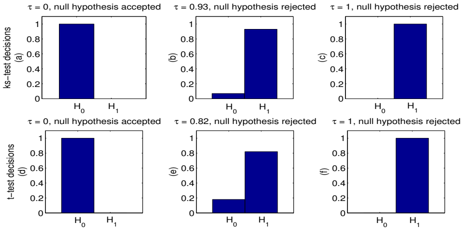
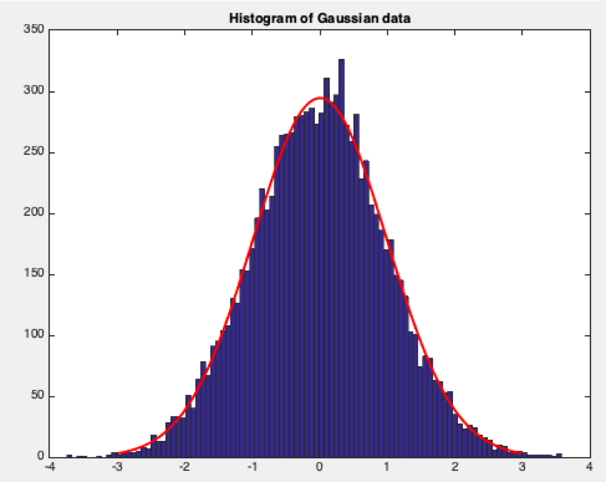


Figure: The KS-test (top) and the t-test (bottom) decisions made for three different 16-beat windows. (1) no synthetic TWA and SNR=10dB. (2) synthetic 35μv TWA and SNR=5dB. (3) synthetic 35μv TWA and SNR=10dB.

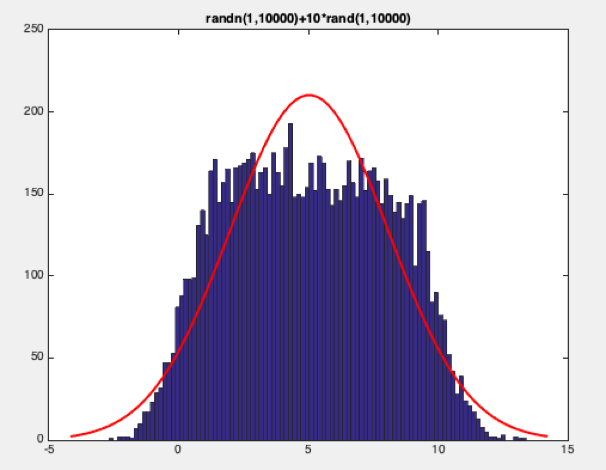
Résumé

- Chapitre 1 : Estimation
  - Modèle statistique, qualités d'un estimateur, exemples
  - Inégalité de Cramér Rao
  - Maximum de vraisemblance
  - Méthode des moments
  - Estimation Bayésienne
  - Intervalles de confiance
- Chapitre 2 : Tests Statistiques
  - Généralités, exemple
  - Courbes COR, p-valeur
  - Théorème de Neyman Pearson
  - Autres tests paramétriques
  - Test du χ^2
  - Test de Kolmogorov

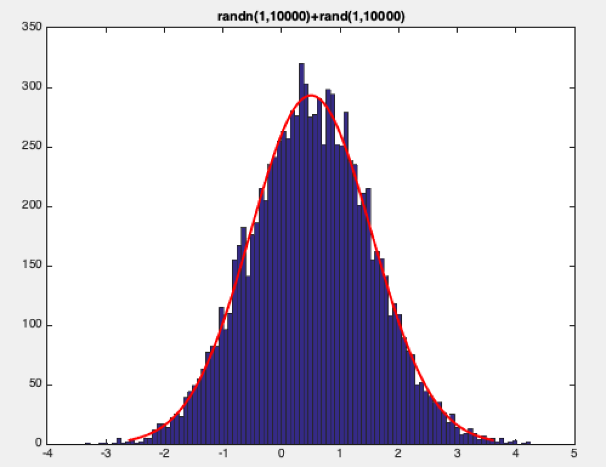
Motivations



Motivations



Motivations



Test du χ^2

Le test du χ^2 est un test non paramétrique d'ajustement (ou d'adéquation) qui permet de tester les deux hypothèses suivantes

$$H_0 : L = L_0, \quad H_1 : L \neq L_0$$

où L_0 est une loi donnée. Le test consiste à déterminer si $(x_1, ..., x_n)$ est de loi L_0 ou non. On se limitera dans ce cours au cas simple où $x_i \in \mathbb{R}$.

Définition du test

$$\text{Rejet de } H_0 \text{ si } \phi_n = \sum_{k=1}^{K} \frac{(Z_k - np_k)^2}{np_k} > S_\alpha$$

Remarque

- L_0 peut être une loi discrète ou continue. Dans le cas discret, on construira les classes en réunissant certaines valeurs de la loi testée.

Test du χ^2

Statistique du test

$$\phi_n = \sum_{k=1}^{K} \frac{(Z_k - np_k)^2}{np_k} > S_\alpha$$

- Z_k : nombre d'observations x_i appartenant à la classe C_k, $k = 1, ..., K$
- p_k : probabilité qu'une observation x_i appartienne à la classe C_k sachant $X_i \sim L_0$
$$P\left[X_i \in C_k | X_i \sim L_0\right]$$
- n : nombre total d'observations

Loi asymptotique de la statistique du test sous H_0

$$\phi_n \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} \chi^2_{K-1}$$

Pour la preuve, voir notes de cours ou livres

Remarques

- Interprétation de ϕ_n
$$\phi_n = \sum_{k=1}^{K} \frac{n}{p_k} \left(\frac{Z_k}{n} - p_k \right)^2$$
Distance entre probabilités théoriques et empiriques
- Nombre d'observations fini
Une heuristique dit que la loi asymptotique de ϕ_n est une bonne approximation pour n fini si 80% des classes vérifient $np_k \geq 5$ et si $p_k > 0, \forall k = 1, ..., K$ ☞ Classes équiprobables
- Correction
Lorsque les paramètres de la loi L_0 sont inconnus
$$\phi_n \underset{n \to \infty}{\overset{\mathcal{L}}{\to}} \chi^2_{K-1-n_p}$$
où n_p est le nombre de paramètres inconnus estimés par la méthode du maximum de vraisemblance
- Constitution des classes dans le cas d'une loi discrète
- Puissance du test : non calculable

Exemple 1

4.13	1.41	-1.16	-0.75	1.96	2.46	0.197	0.24	0.42	2.00
2.08	1.48	1.73	0.82	0.33	-0.76	0.42	4.60	-2.83	0.197
2.59	0.54	4.06	-0.69	4.99	0.67	2.45	5.61	2.13	1.76
5.03	0.85	1.29	0.17	-0.38	2.76	-1.03	1.87	4.48	0.73

Est-il raisonnable de penser que ces observations sont issues d'une population de loi $\mathcal{N}(1,4)$?

Solution

- Classes

$$C_1 :]-\infty, -0.34], C_2 :]-0.34, 1], C_3 :]1, 2.34], C_4 :]2.34, \infty[$$

- Nombres d'observations

$$Z_1 = 7,\ Z_2 = 12,\ Z_3 = 10,\ Z_4 = 11$$

Exemple 1

Solution

- Statistique de test

$$\phi_n = 1.4$$

- Seuils

	χ_2^2	χ_3^2
$S_{0.05}$	5.991	7.815
$S_{0.01}$	9.210	11.345

- Conclusion
  On accepte l'hypothèse H_0 avec les risques $\alpha = 0.01$ et $\alpha = 0.05$.

Exemple 2

Enoncé

On lance un dé 60 fois et on relève les nombres de fois où on a observé les différentes faces

x_i	1	2	3	4	5	6
n_i	15	7	4	11	6	17

On se demande si ce dé est truqué (Hypothèse H_1) ou non (Hypothèse H_0).

1. Déterminer la statistique du test du chi-deux noté ϕ associée à ce problème.
2. Quelle est la loi de cette statistique de test sous l'hypothèse H_0 ?
3. Expliquer comment déterminer le seuil de décision S_α du test du chi-deux à l'aide de la fonction de répartition de la loi déterminée à la question précédente et du risque α de ce test. Pour $\alpha = 0.05$, on trouve $S_{0.05} = 11.07$ et pour $\alpha = 0.01$, on a $S_{0.01} = 15.09$. Que conclut-on ?

Exemple 3 (voir TD)

Enoncé

Un statisticien pose la question suivante à un échantillon de 30 participants : "Préférez-vous boire du thé ou du café ?". Parmi cet échantillon, 10 préfèrent le thé et 20 préfèrent le café. Il désire effectuer un test du chi-deux pour déterminer s'il y a une véritable préférence pour le café dans cet échantillon. Pour cela, il définit l'hypothèse H_0 par "la probabilité de boire du thé est égale à la probabilité de boire du café", i.e., les deux classes $\{\text{Thé}\}$ et $\{\text{Café}\}$ sont équiprobables ($P[\text{Thé}] = P[\text{Café}] = \frac{1}{2}$).

1. Déterminer la statistique du test du chi-deux noté ϕ associée à ce problème.
2. Rappeler la loi de ϕ sous l'hypothèse H_0 (définie par "Il n'y a pas de préférence ni pour le thé, ni pour le café").
3. Expliquer comment déterminer le seuil de décision S_α du test du chi-deux à l'aide de la fonction de répartition de la loi déterminée à la question précédente et du risque α de ce test. Pour $\alpha = 0.05$, on trouve $S_{0.05} = 3.84$. Que conclut-on ?

Résumé

- Chapitre 1 : Estimation
  - Modèle statistique, qualités d'un estimateur, exemples
  - Inégalité de Cramér Rao
  - Maximum de vraisemblance
  - Méthode des moments
  - Estimation Bayésienne
  - Intervalles de confiance
- Chapitre 2 : Tests Statistiques
  - Généralités, exemple
  - Courbes COR, p-valeur
  - Théorème de Neyman Pearson
  - Autres tests paramétriques
  - Test du χ^2
  - Test de Kolmogorov

Test de Kolmogorov

Le test de Kolmogorov est un test non paramétrique d'ajustement (ou d'adéquation) qui permet de tester les deux hypothèses suivantes

$$H_0 : L = L_0, \quad H_1 : L \neq L_0$$

où L_0 est une loi donnée. Le test consiste à déterminer si $(x_1, ..., x_n)$ est de loi L_0 ou non. On se limitera dans ce cours au cas simple où $x_i \in \mathbb{R}$.era dans ce cours au cas simple où $x_i \in \mathbb{R}$.

Définition du test

$$\text{Rejet de } H_0 \text{ si } D_n = \sup_{x \in \mathbb{R}} |\widehat{F}(x) - F_0(x)| > S_\alpha$$

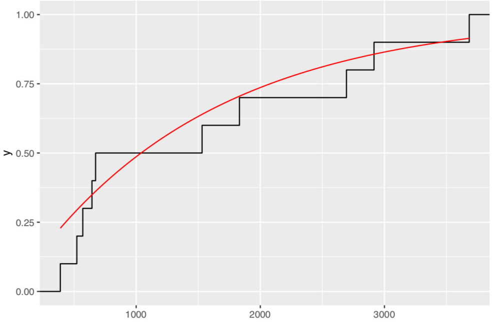
Remarque

- L_0 doit être une loi continue.

Statistique du test de Kolmogorov

Fonctions de répartition

$F_0(x) = P[X \leq x]$ est la fonction de répartition théorique de L_0 et $\widehat{F}_n(x)$ est la fonction de répartition empirique de $(x_1, ..., x_n)$



D_n est l'écart maximum entre les deux courbes.

Calcul de D_n

Á l'aide de l'échantillon ordonné

$$D_n = \max_{i \in \{1,\dots,n\}} \max\{E_i^+, E_i^-\}$$

avec

$$E_i^+ = \left| \widehat{F}_n\left(x_{(i)}^+\right) - F_0\left(x_{(i)}\right) \right|, \; E_i^- = \left| \widehat{F}_n\left(x_{(i)}^-\right) - F_0\left(x_{(i)}\right) \right|$$

Remarques

- $x_{(1)}, ..., x_{(n)}$ est la statistique d'ordre de $x_1, ..., x_n$ telle que $x_{(1)} \leq ... \leq x_{(n)}$
- $\widehat{F}_n\left(x_{(i)}^+\right) = i/n$ et $\widehat{F}_n\left(x_{(i)}^-\right) = (i-1)/n$.

Statistique du test

Loi de D_n sous H_0

- Indépendante de L_0
- Loi asymptotique

$$P[\sqrt{n}D_n < y] \underset{n\to\infty}{\to} \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2y^2) = K(y)$$

Convergence de cette série très rapide (pour $y > 0.56$, les trois premiers termes donnent une approximation avec une erreur inférieure à 10^{-4}).

Détermination du seuil S_α

$$S_{n,\alpha} = \frac{1}{\sqrt{n}} K^{-1}(1-\alpha)$$

Le seuil dépend de α et de n.

Remarques

Puissance du test

Non calculable

Tests unilatéraux

- Pour tester $H_0 : F = F_0$ contre $H_1 : F \geq F_0$, le test de Kolmogorov rejette H_0 si
$$D_n^+ = \sup_{t\in\mathbb{R}} \left[\widehat{F}_n(t) - F_0(t)\right] \geq S_{n,\alpha}$$
- Pour tester $H_0 : F = F_0$ contre $H_1 : F \leq F_0$, le test de Kolmogorov rejette H_0 si
$$D_n^- = \sup_{t\in\mathbb{R}} \left[F_0(t) - \widehat{F}_n(t)\right] \geq S_{n,\alpha}$$

Exemple

Est-il raisonnable de penser que ces observations sont issues d'une population de loi uniforme sur $[0,1]$?

x_i	0.0078	0.063	0.10	0.25	0.32	0.39	0.40	0.48	0.49	0.53
E_i^-	0.0078	0.013	0.00	0.10	0.07	0.14	0.05	0.008	0.04	0.03
E_i^+	0.0422	0.037	0.05	0.05	0.12	0.09	0.10	0.13	0.09	0.08
$\mathrm{Max}(E_i^+, E_i^-)$	0.0422	0.037	0.05	0.1	0.12	0.14	0.10	0.13	0.09	0.08

x_i	0.67	0.68	0.69	0.73	0.79	0.80	0.87	0.88	0.90	0.996
E_i^-	0.17	0.13	0.04	0.03	0.04	0.05	0.07	0.03	0.05	0.046
E_i^+	0.12	0.08	0.09	0.08	0.09	0.00	0.02	0.02	0.00	$4e-3$
$\mathrm{Max}(E_i^+, E_i^-)$	0.17	0.13	0.09	0.08	0.09	0.05	0.07	0.03	0.05	0.046

Résultats

Statistique de test

$$D_n = 0.17$$

Seuils pour $n = 20$

$S_{0.05}$	0.294
$S_{0.01}$	0.352

Conclusion

On accepte l'hypothèse H_0 avec les risques $\alpha = 0.01$ et $\alpha = 0.05$.

Que faut-il savoir ?

Tests statistiques

- Définition et calcul des risques de première et seconde espèce et de la puissance d'un test binaire
- Définition et détermination des courbes COR
- Appliquer le théorème de Neyman-Pearson dans le cas de variables aléatoires discrètes et continues
- Connaître l'existence des tests paramétriques pour tester la valeur de la moyenne ou de la variance d'un échantillon gaussien
- Connaître l'existence des tests paramétriques pour tester l'égalité de moyennes et de variances pour deux échantillons gaussiens indépendants
- Principe et mise en oeuvre d'un test du χ^2
- Principe et mise en oeuvre d'un test de Kolmogorov