

Statistique

Slides 2ème année Apprentissage

Jean-Yves Tournet⁽¹⁾

(1) University of Toulouse, ENSEEIHT-IRIT-TéSA
jyt@n7.fr, <http://perso.tesa.prd.fr/jyt/>

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

- ▶ **Chapitre 1 : Estimation**
 - ▶ **Modèle statistique, qualités d'un estimateur, exemples**
 - ▶ Inégalité de Cramér Rao
 - ▶ Maximum de vraisemblance
 - ▶ Méthode des moments
- ▶ **Chapitre 2 : Tests Statistiques**
 - ▶ Généralités, exemple
 - ▶ Courbes COR
 - ▶ Théorème de Neyman Pearson
 - ▶ Test du χ^2
 - ▶ Test de Kolmogorov

Modèle statistique

Notations

- ▶ Observations

$$x_1, \dots, x_n$$

- ▶ Échantillon

$$X_1, \dots, X_n$$

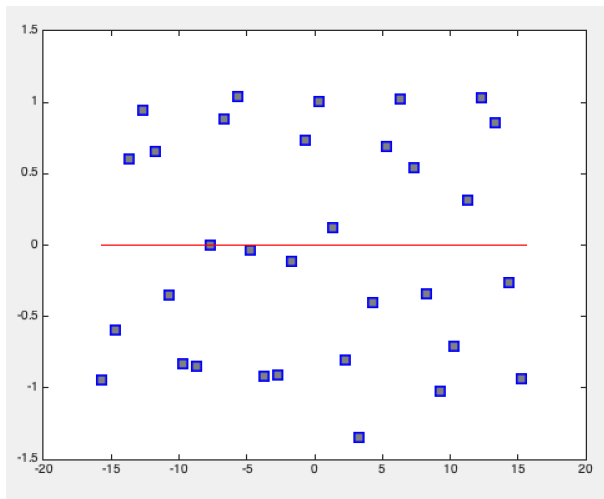
n va iid associées aux observations

- ▶ Estimateur

$$\hat{\theta}(X_1, \dots, X_n) \text{ ou } \hat{\theta}_n \text{ ou } \hat{\theta}$$

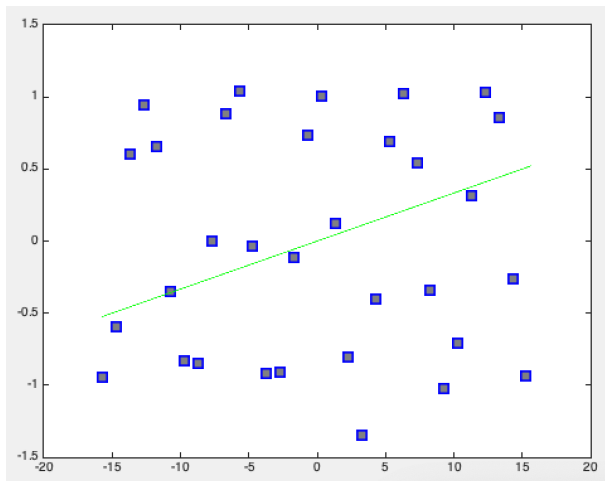
Exemple

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



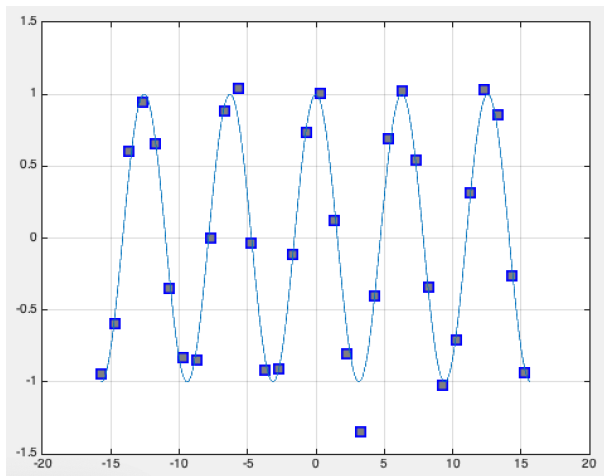
Exemple

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



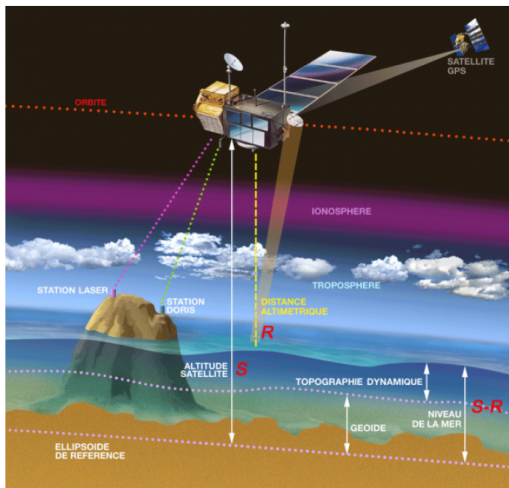
Exemple

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



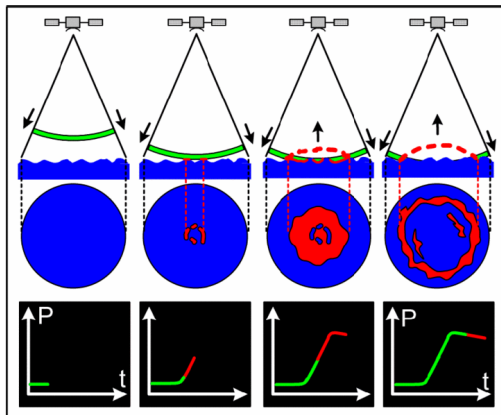
Application réelle

Altimétrie



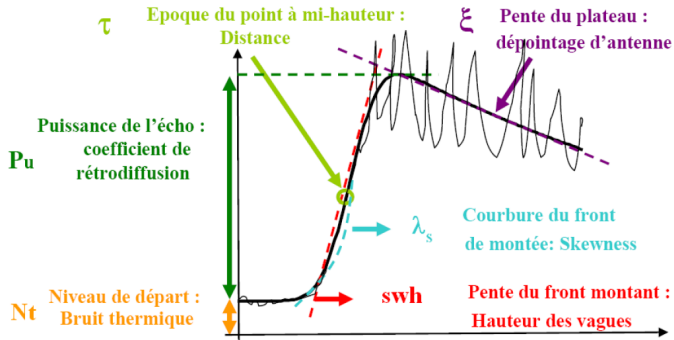
Application réelle

Formation de l'écho altimétrique



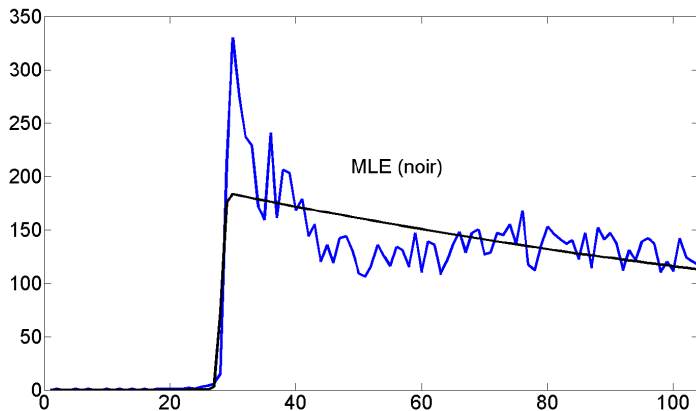
Application réelle

Modèle de Brown



Application réelle

Modèle de Brown



Qualités d'un estimateur

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

- ▶ **Biais** (erreur systématique) : $b_n(\theta) = E(\hat{\theta}_n) - \theta$
- ▶ **Variance**

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

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

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

CS de **convergence** : $\hat{\theta}_n$ est un estimateur convergent si

$$\lim_{n \rightarrow +\infty} b_n(\theta) = \lim_{n \rightarrow +\infty} v_n(\theta) = 0$$

Qualités d'un estimateur

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

- Biais

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

- Matrice de covariance

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

Exemples

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

- Moyenne empirique

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

- Autre estimateur

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

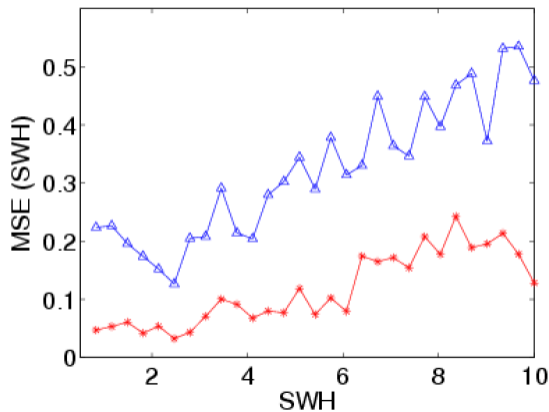
Exemple 2 : $X_i \sim \mathcal{N}(m, \sigma^2)$, $\theta = (m, \sigma^2)^T$

Étude de l'estimateur

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

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
- ▶ **Chapitre 2 : Tests Statistiques**
 - ▶ Généralités, exemple
 - ▶ Courbes COR
 - ▶ Théorème de Neyman Pearson
 - ▶ Test du χ^2
 - ▶ Test de Kolmogorov

Inégalité de Cramér-Rao

Vraisemblance

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

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

► Définition

$$\text{Var}(\hat{\theta}_n) \geq \frac{[1 + b'_n(\theta)]^2}{-E\left[\frac{\partial^2 \ln L(X_1, \dots, X_n; \theta)}{\partial \theta^2}\right]} = \text{BCR}(\theta)$$

BCR(θ) 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 $\text{Var}(\hat{\theta}_n) = \text{BCR}(\theta)$ (Il est unique !)

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

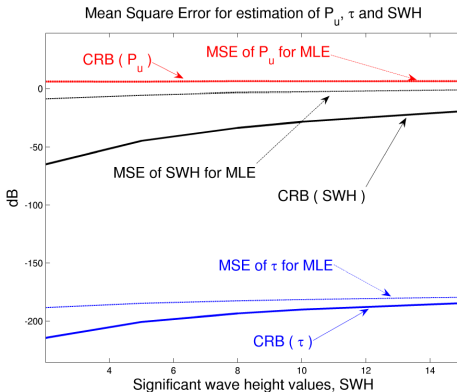
- **Cas où $(X_1, \dots, X_n)$ est un échantillon**

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

- **Cas multivarié** : non traité dans ce cours

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
- ▶ **Chapitre 2 : Tests Statistiques**
 - ▶ Généralités, exemple
 - ▶ Courbes COR
 - ▶ Théorème de Neyman Pearson
 - ▶ Test du χ^2
 - ▶ Test de Kolmogorov

Estimateur du maximum de vraisemblance

Définition

$$\hat{\theta}_{\text{MV}} = \arg \max_{\theta} L(X_1, \dots, X_n; \theta)$$

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

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

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

et on vérifie qu'on a bien un maximum en faisant **un tableau de variations**.

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

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

Contre-exemple : $X_1, \dots, X_n$ échantillon de loi uniforme sur l'intervalle $]0, \theta[$ et θ paramètre à estimer.

Estimateur du maximum de vraisemblance

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

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

pour $i = 1, \dots, p$

Exemples

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

Propriétés

- ▶ Estimateur **asymptotiquement non biaisé**

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

- ▶ Estimateur **convergent**
- ▶ Estimateur **asymptotiquement efficace**

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

- ▶ **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

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

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.

Plan du cours

Résumé

- ▶ **Chapitre 1 : Estimation**
 - ▶ Modèle statistique, qualités d'un estimateur, exemples
 - ▶ Inégalité de Cramér Rao
 - ▶ Maximum de vraisemblance
 - ▶ **Méthode des moments**
- ▶ **Chapitre 2 : Tests Statistiques**
 - ▶ Généralités, exemple
 - ▶ Courbes COR
 - ▶ Théorème de Neyman Pearson
 - ▶ Test du χ^2
 - ▶ Test de Kolmogorov

Méthode des moments

Définition

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

$$\theta = h(m_1, \dots, m_q)$$

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

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

Exemples

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

Méthode des moments

Propriétés

- ▶ Estimateur convergent
- ▶ Normalité Asymptotique

Conclusion

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

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**

Plan du cours

Résumé

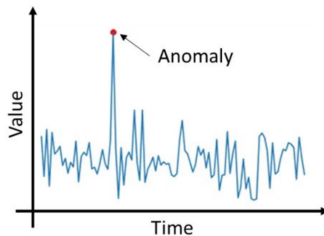
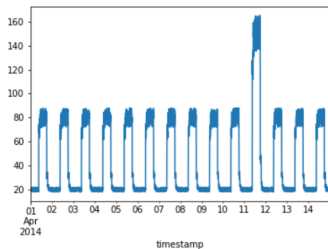
- ▶ **Chapitre 1 : Estimation**
 - ▶ Modèle statistique, qualités d'un estimateur, exemples
 - ▶ Inégalité de Cramér Rao
 - ▶ Maximum de vraisemblance
 - ▶ Méthode des moments
- ▶ **Chapitre 2 : Tests Statistiques**
 - ▶ Généralités, exemple
 - ▶ Courbes COR
 - ▶ Théorème de Neyman Pearson
 - ▶ Test 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
 - ▶ Théorème de Neyman Pearson
 - ▶ Test du χ^2
 - ▶ Test de Kolmogorov

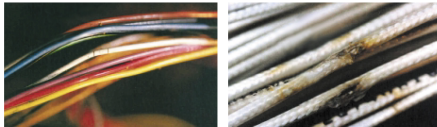
Motivations



Motivations

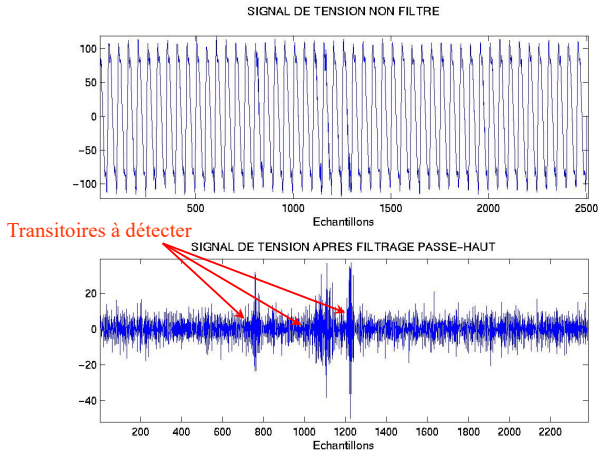


Debris from TWA Flight 800 was pieced together following the fatal crash of the Boeing 747 in 1996. Government investigators concluded that the likely trigger was a short circuit from damaged wiring—wiring “not atypical for an airplane of its age.”



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** $H_0, H_1, \dots$ à partir de n observations $x_1, \dots, x_n$. On se limitera dans ce cours à deux hypothèses H_0 et H_1 . Effectuer un test, c'est déterminer une **statistique de test** $T(X_1, \dots, X_n)$ et un **ensemble** Δ tel que

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

Vocabulaire

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

Définitions

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

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

- ▶ **Risque de seconde espèce** = probabilité de non-détection

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

- ▶ **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

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

- ▶ Problèmes

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

Plan du cours

Résumé

- ▶ **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**
 - ▶ Théorème de Neyman Pearson
 - ▶ Test du χ^2
 - ▶ Test de Kolmogorov

Caractéristiques opérationnelles du récepteur (COR)

Définition

$$PD = h(PFA)$$

Exemple

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

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

► **Probabilité de fausse alarme**

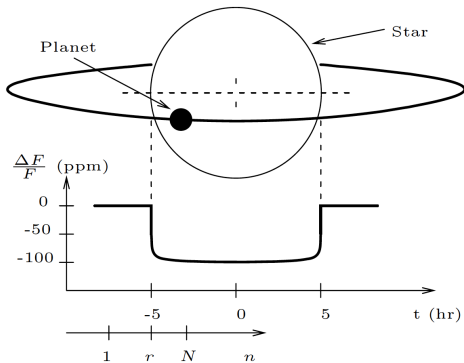
$$\alpha = 1 - F\left(\frac{S_\alpha - m_0}{\frac{\sigma}{\sqrt{n}}}\right) \Leftrightarrow S_\alpha = m_0 + \frac{\sigma}{\sqrt{n}} F^{-1}(1 - \alpha)$$

► **Probabilité de détection**

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

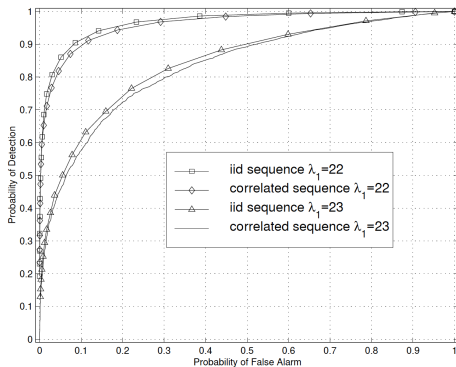
Exemple d'application

Détection d'exoplanètes



Exemple de représentation graphique pour les courbes COR

Données indépendantes ou corrélées ?



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
 - ▶ **Théorème de Neyman Pearson**
 - ▶ Test du χ^2
 - ▶ Test de Kolmogorov

Théorème de Neyman-Pearson

Test paramétrique à hypothèses simples

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \text{ et } H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1 \quad (2)$$

Théorème pour variables aléatoires X_i continues

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

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

Exemple

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

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

Théorème de Neyman-Pearson

Test paramétrique à hypothèses simples

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \text{ et } H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1 \quad (3)$$

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

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

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

Remarque

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

Résumé

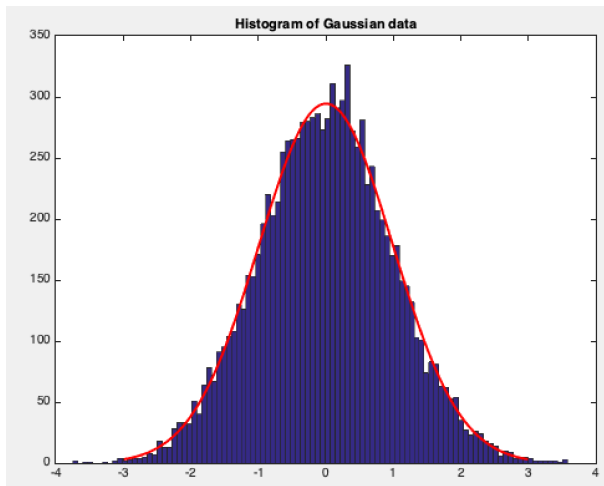
▶ Chapitre 1 : Estimation

- ▶ Modèle statistique, qualités d'un estimateur, exemples
- ▶ Inégalité de Cramér Rao
- ▶ Maximum de vraisemblance
- ▶ Méthode des moments
- ▶ Estimation Bayésienne
- ▶ Intervalles de confiance

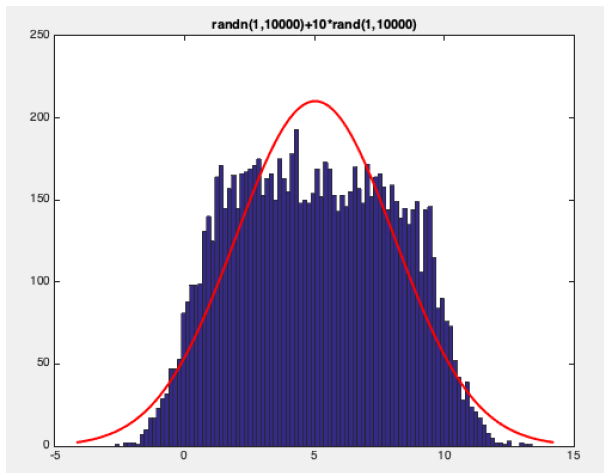
▶ Chapitre 2 : Tests Statistiques

- ▶ Généralités, exemple
- ▶ Courbes COR
- ▶ Théorème de Neyman Pearson
- ▶ Test du χ^2
- ▶ Test de Kolmogorov

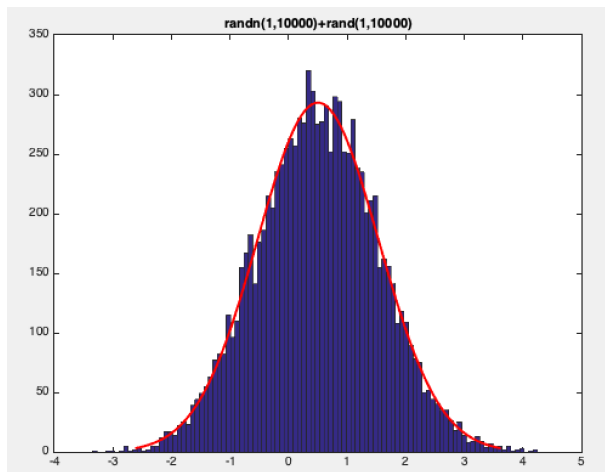
Motivations



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 $(x_1, \dots, x_n)$ 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 , $k = 1, \dots, K$
- ▶ p_k : probabilité qu'une observation x_i appartienne à la classe C_k sachant $X_i \sim L_0$

$$P[X_i \in C_k | X_i \sim L_0]$$

- ▶ n : nombre total d'observations

Loi asymptotique de la statistique du test sous H_0

$$\phi_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \chi_{K-1}^2$$

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 $p_k > 0, \forall k = 1, \dots, K$  Classes **équiprobables**

- **Correction**

Lorsque les paramètres de la loi L_0 sont **inconnus**

$$\phi_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \chi_{K-1-n_p}^2$$

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 {Thé} et {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
 - ▶ Théorème de Neyman Pearson
 - ▶ 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 $(x_1, \dots, x_n)$ 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

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

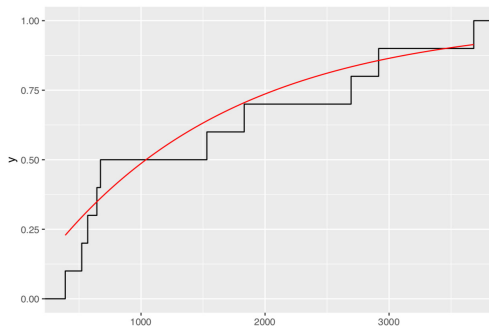
Remarque

- ▶ 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 $\hat{F}_n(x)$ est la fonction de répartition empirique de $(x_1, \dots, x_n)$



D_n est l'écart maximum entre les deux courbes.

Calcul de D_n

À l'aide de l'échantillon ordonné

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

avec

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

Remarques

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

Statistique du test

Loi de D_n sous H_0

- ▶ Indépendante de L_0
- ▶ Loi asymptotique

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

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

$$D_n^+ = \sup_{t \in \mathbb{R}} [\hat{F}_n(t) - F_0(t)] \geq S_{n,\alpha}$$

- Pour tester $H_0 : F = F_0$ contre $H_1 : F \leq F_0$, le test de Kolmogorov rejette H_0 si

$$D_n^- = \sup_{t \in \mathbb{R}} [F_0(t) - \hat{F}_n(t)] \geq S_{n,\alpha}$$

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
$\text{Max}(E_i^+, E_i^-)$	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	$4e - 3$
$\text{Max}(E_i^+, E_i^-)$	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
- ▶ Principe et mise en oeuvre d'un **test du χ^2**
- ▶ Principe et mise en oeuvre d'un **test de Kolmogorov**