

Statistique

Slides 1ère année MFEE

Jean-Yves Tourneret⁽¹⁾

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

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

- ▶ **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

Modèle statistique

Notations

- ▶ Observations

$$x_1, \dots, x_n$$

- ▶ Échantillon

$$X_1, \dots, X_n$$

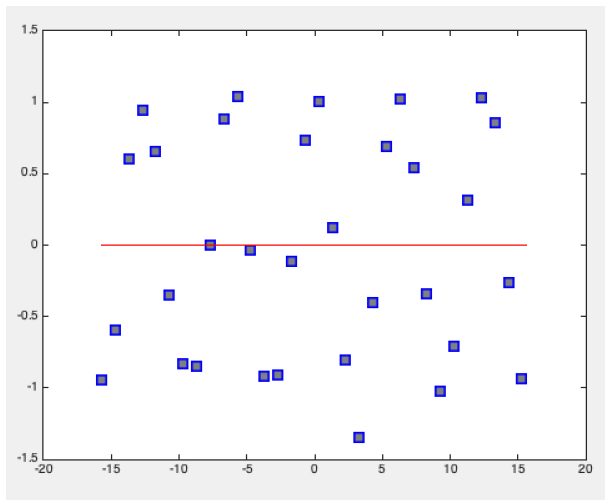
n va iid associées aux observations

- ▶ Estimateur

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

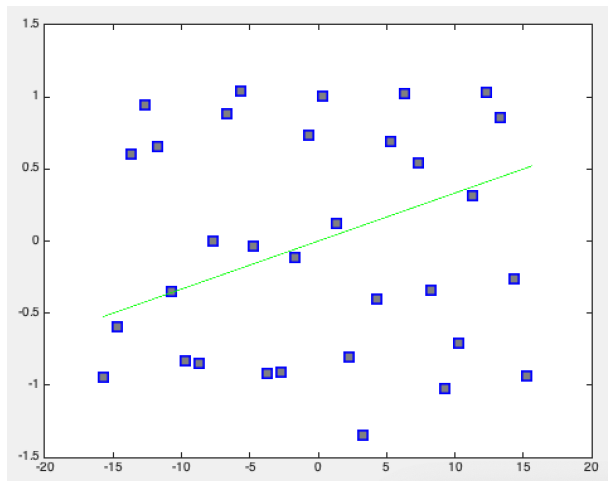
Exemple

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



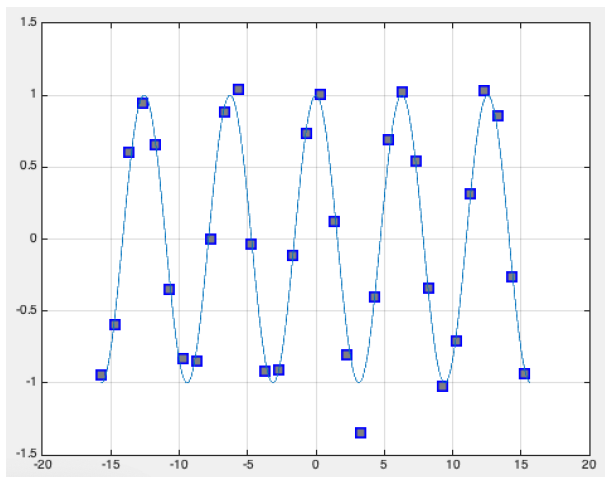
Exemple

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



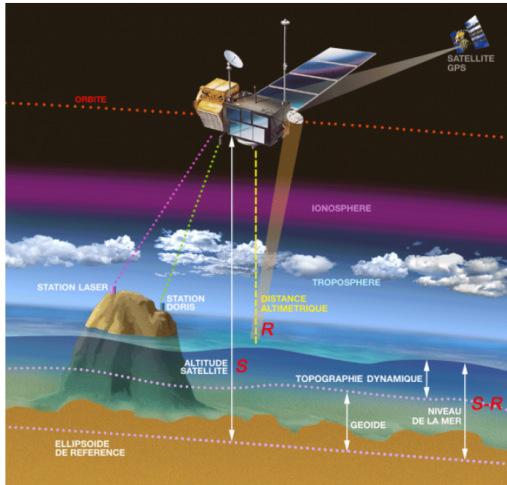
Exemple

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



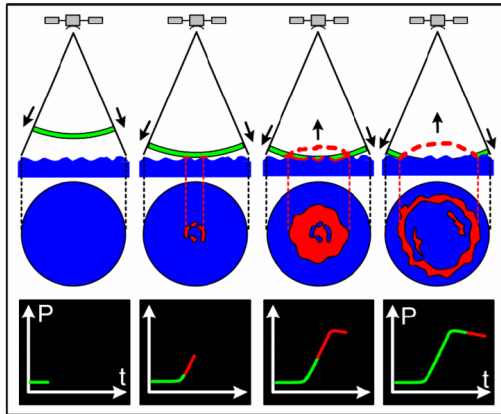
Application réelle

Altimétrie



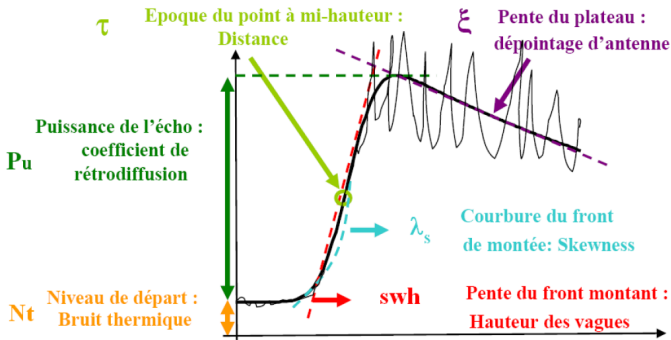
Application réelle

Formation de l'écho altimétrique



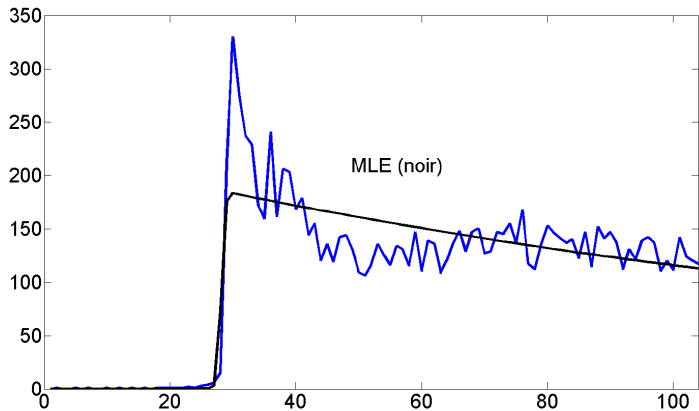
Application réelle

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

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

- ▶ Matrice de covariance

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

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

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

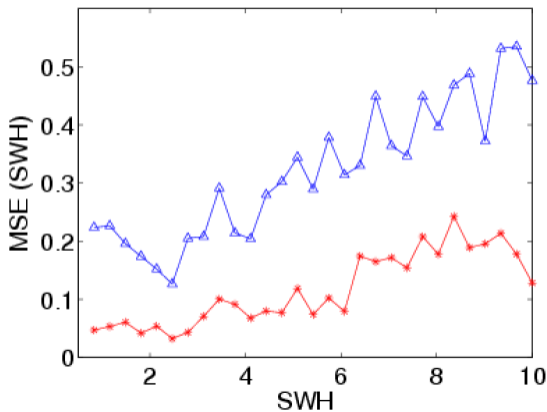
Exemple 2 : $X_i \sim \mathcal{N}(m, \sigma^2)$, $\theta = (m, \sigma^2)^T$

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
 - ▶ 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

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

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 $\text{Var}(\widehat{\theta}_n) = \text{BCR}(\theta)$ (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

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

Cas multivarié

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

► **Définition**

$$\text{Cov}(\hat{\theta}) \geq I_n^{-1}(\theta)$$

avec

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

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

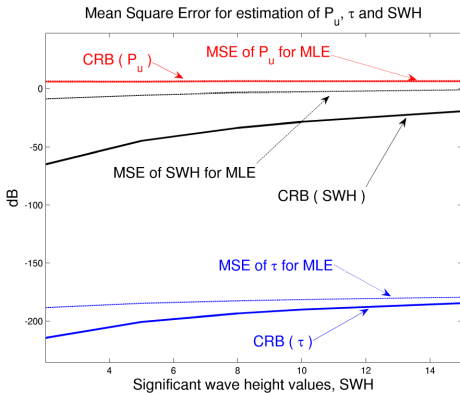
$$\text{Var}(\hat{\theta}_i) \geq [I_n^{-1}(\theta)]_{ii}$$

► **Exemple**

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

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

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

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** ou si ce n'est pas possible en étudiant

$$\frac{\partial^2 \ln L(X_1, \dots, X_n; \hat{\theta}_{\text{MV}})}{\partial \theta^2} < 0$$

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

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

Estimateur du maximum de vraisemblance

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

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

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**

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

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

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

- ▶ **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

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

Soient X_1, \dots, 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 $\hat{\theta}_{MV}$). 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 $\hat{\theta}_{MV}$ non convergent (avec plusieurs maxima locaux de $l'(\theta) = 0$)

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

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

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.

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 $\theta \in \mathbb{R}^p$ à l'aide de la **vraisemblance** de X_1, \dots, X_n (paramétrée par θ) et d'une **loi a priori** $p(\theta)$. Pour cela, on minimise une fonction de coût $c(\theta, \hat{\theta})$ qui représente l'erreur entre θ et $\hat{\theta}$. Deux estimateurs principaux

- ▶ **Estimateur MAP** : l'**estimateur du maximum a posteriori** (MAP) de θ est défini par

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} p(\theta | X_1, \dots, X_n)$$

où $p(\theta | x_1, \dots, x_n)$ est la **loi a posteriori** de θ .

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

$$\hat{\theta}_{\text{MMSE}} = E(\theta | X_1, \dots, X_n)$$

Propriétés des estimateurs Bayésiens

Estimateur MAP

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

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

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

Estimateur MMSE

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

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

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**

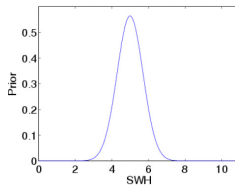
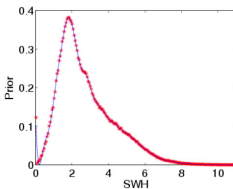
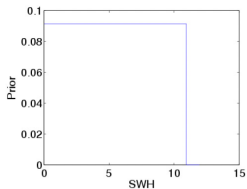
$$\theta | X_1, \dots, X_n \sim \mathcal{N}(m_p, \sigma_p^2)$$

- ▶ **Estimateurs**

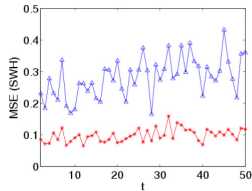
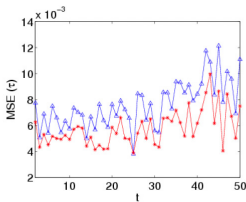
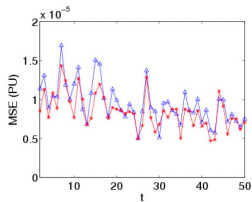
$$\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{MMSE}} = m_p = \bar{X} \left(\frac{n\nu^2}{n\nu^2 + \sigma^2} \right) + \mu \left(\frac{\sigma^2}{\sigma^2 + n\nu^2} \right)$$

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

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

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

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

donc

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

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

- ▶ Variance

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

Que faut-il savoir ?

Estimation statistique

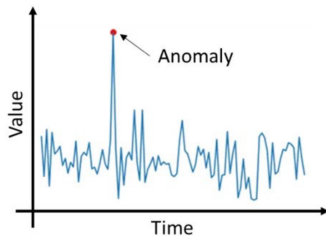
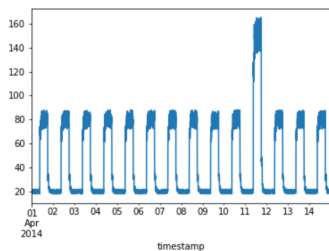
- ▶ Notions de **biais**, **variance** et **convergence** d'un estimateur
- ▶ Calcul d'une **borne de Cramér-Rao** et notion d'**efficacité**
- ▶ Détermination de l'estimateur du **maximum de vraisemblance** (MV)
- ▶ Propriétés de l'estimateur MV
- ▶ Principe et application de la **méthode des moments**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

Plan du cours

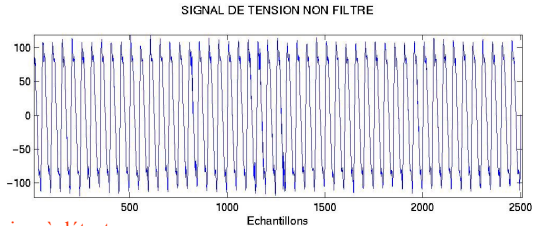
Résumé

- ▶ **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

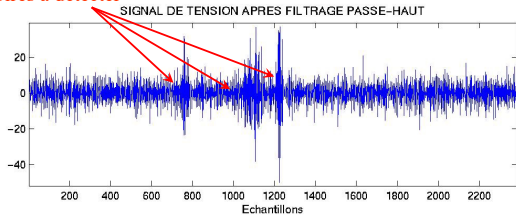
Motivations



Motivations



Transitoires à détecter



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

$$\begin{aligned} H_0 \text{ rejetée si } T(X_1, \dots, X_n) \in \Delta \\ H_0 \text{ acceptée si } T(X_1, \dots, X_n) \notin \Delta. \end{aligned} \tag{1}$$

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, p -valeur**
 - ▶ Théorème de Neyman Pearson
 - ▶ Autres tests paramétriques
 - ▶ Test du χ^2
 - ▶ Test de Kolmogorov

Caractéristiques opérationnelles du récepteur (COR)

Définition

$$PD = h \text{ (PFA)}$$

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**

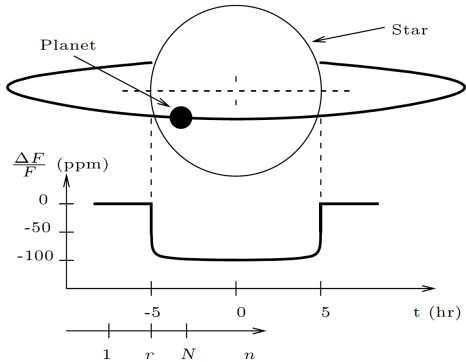
$$\alpha = 1 - F\left(\frac{S_\alpha - m_0}{\frac{\sigma}{\sqrt{n}}}\right) \Leftrightarrow S_\alpha = m_0 + \frac{\sigma}{\sqrt{n}} F^{-1}(1 - \alpha)$$

► **Probabilité de détection**

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

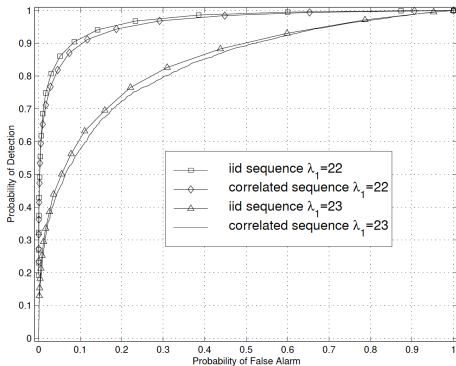
Exemple d'application

Détection d'exoplanètes



Exemple de représentation graphique pour les courbes COR

Données indépendantes ou corrélées ?



p -valeur d'un test

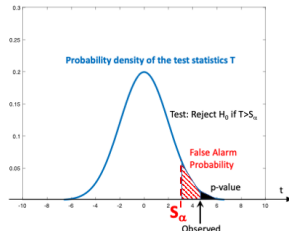
Définition

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

où \mathcal{R}_α est la zone de rejet de H_0 pour α fixé et $\mathbf{x} = (x_1, \dots, x_n)$. **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 : $\alpha^* = 1 - F(T_{\text{obs}})$



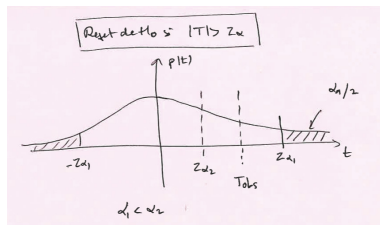
Autre exemple : rejet de H_0 si $|T| > S_\alpha$

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

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

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

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

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
 - ▶ (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

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \text{ et } H_1 : \boldsymbol{\theta} \in \Theta_1 \quad (4)$$

Définition du test

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

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

Remarque

$$L(x_1, \dots, x_n | \hat{\boldsymbol{\theta}}_i^{\text{MV}}) = \sup_{\boldsymbol{\theta} \in \Theta_i} L(x_1, \dots, x_n | \boldsymbol{\theta})$$

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

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

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

Définition du test

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

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

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

Définition du test

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

Remarques

- ▶ Si on pose $U = \sqrt{n} \left(\frac{\bar{X} - m_0}{\sigma} \right)$ et $V = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2$, 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, \dots, X_n) un échantillon gaussien de loi $\mathcal{N}(m, \sigma^2)$ avec une **moyenne m connue** et les hypothèses

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

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

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

Définition du test

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

Remarques

- ▶ La loi de T sous H_0 est une loi du χ_{n-1}^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 deux échantillons Gaussiens sont significativement différents ? (variances connues)

Soient (X_1, \dots, X_n) et (Y_1, \dots, 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

$$H_0 : m_1 = m_2 \text{ et } H_1 : m_1 > m_2 \quad (9)$$

Définition du test

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

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, \dots, X_n) et (Y_1, \dots, 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

$$H_0 : m_1 = m_2 \text{ et } H_1 : m_1 > m_2 \quad (10)$$

Définition du test

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

avec

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

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** (utilisé pour tester l'efficacité d'un médicament pour deux ensembles de patients atteints d'une pathologie)
- ▶ 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, \dots, X_n) et (Y_1, \dots, 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

$$H_0 : m_1 = m_2 \text{ et } H_1 : m_1 > m_2 \quad (11)$$

Définition du test

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

avec

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

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, \dots, X_n) et (Y_1, \dots, 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, \dots, X_n) et (Y_1, \dots, 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

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

Définition du test

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

avec

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

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, \dots, X_n) et (Y_1, \dots, 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

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

Définition du test

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

avec

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

Remarques

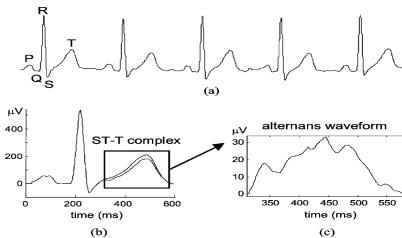
- ▶ Sous l'hypothèse H_0 , T est distribuée suivant une loi de Fisher $\mathcal{F}(n-1, m-1)$
- ▶ Généralisation immédiate à $\sigma_1^2 < \sigma_2^2$ ou à $\sigma_1^2 \neq \sigma_2^2$

Example: T-wave Alternans (TWA) Detection



Context

T-wave alternans (TWA) detection



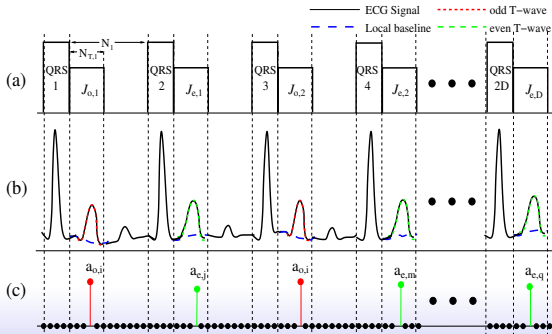
- TWA: a consistent fluctuation in the T waves on an **every-other-beat basis**
- recognized as an index of sudden cardiac death
- a challenging problem: non-visible (microvolt-level) TWA detection

Signal Model



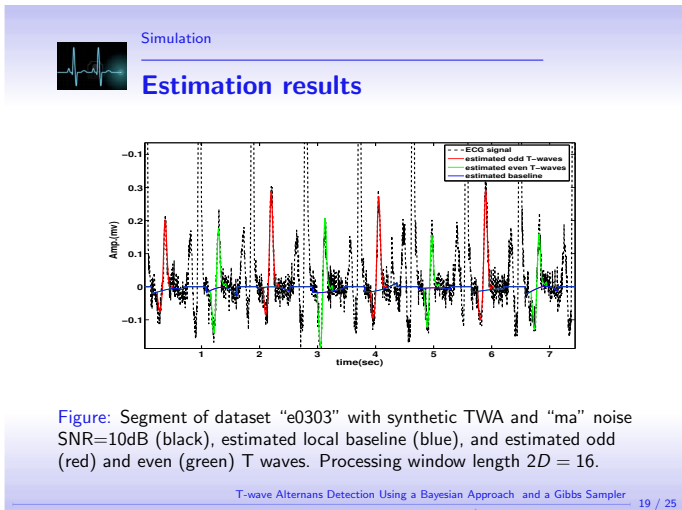
Problem formulation

Signal model for TWA detection



T-wave Alternans Detection Using a Bayesian Approach and a Gibbs Sampler

Estimation Results



Statistical Test



Bayesian TWA detection

Two-sample Student's t -test

- Based on the **assumption of normality** of two samples:

$$\mathcal{H}_0 : \mu_o = \mu_e, \quad \mathcal{H}_1 : \mu_o \neq \mu_e$$

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:

$$t^{(i)} = \frac{\bar{a}_o^{(i)} - \bar{a}_e^{(i)}}{S_{eo}^{(i)} \sqrt{\frac{2}{D}}} \quad (1)$$

$$\bar{a}_o^{(i)} = \frac{1}{D} \sum_{k=1}^D a_{o,k}^{(i)}, \quad \bar{a}_e^{(i)} = \frac{1}{D} \sum_{k=1}^D a_{e,k}^{(i)} \text{ and}$$

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

Detection Results



Simulation

Test decisions

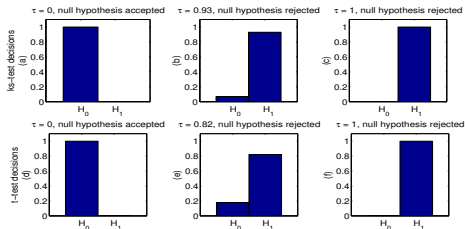
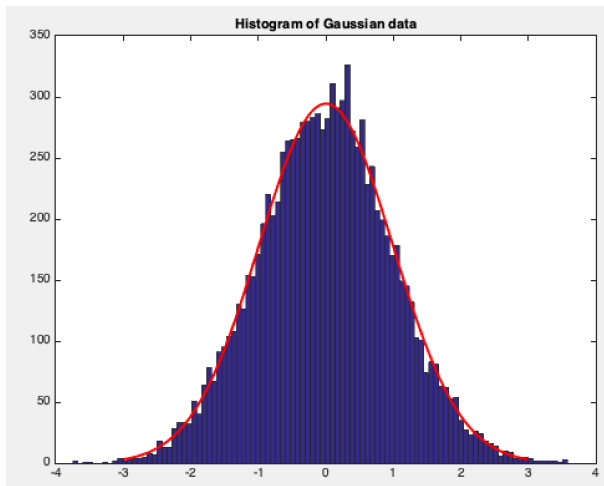


Figure: The KS-test (top) and the t-test (bottom) decisions made for three different 16-beat windows. (1) **no synthetic TWA** and **SNR=10dB**. (2) **synthetic 35 μ V TWA** and **SNR=5dB**. (3) **synthetic 35 μ V TWA** and **SNR=10dB**.

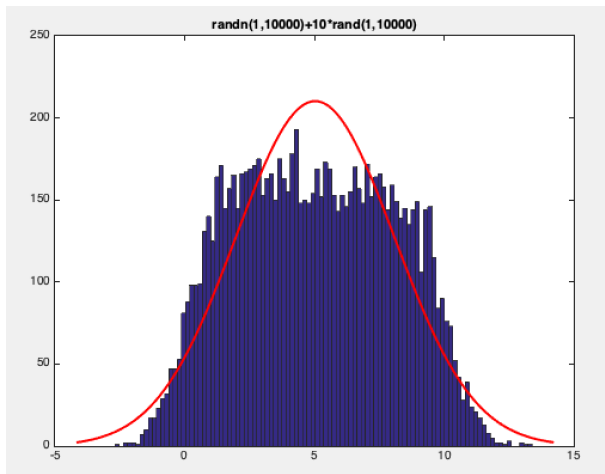
Résumé

- ▶ **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

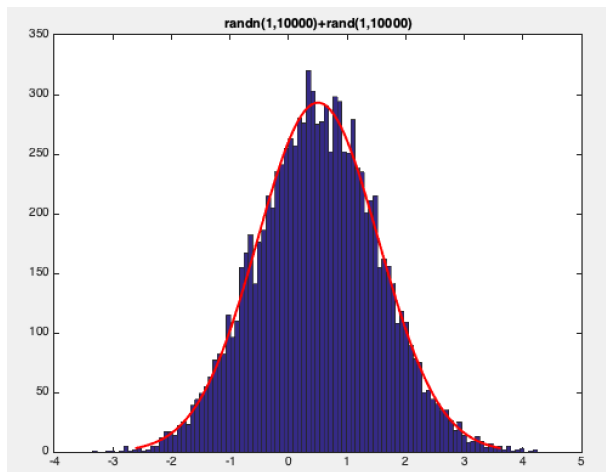
Motivations



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, \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

Énoncé

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)

Énoncé

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, 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, \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

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

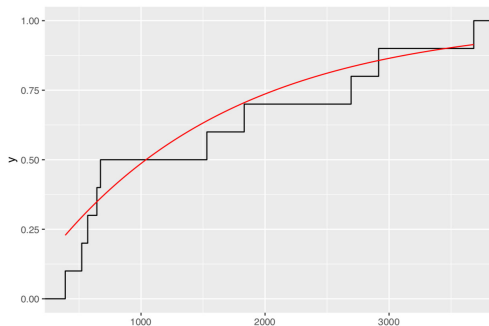
Remarque

- ▶ 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

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

Remarques

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

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 si la loi sous H_1 n'est pas spécifiée

Généralisation aux lois discrètes ou mixtes

D. S. Dimitrova, V. K. Kaishev and S. Tan, "Computing the Kolmogorov-Smirnov distribution when the underlying cdf is purely discrete, mixed or continuous," Journal of Statistical Software, vol. 95, no. 10, oct. 2020.

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
- ▶ 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**