TP pour les leçons de statistique

1 Tests d'ajustement pour des lois discrètes

Leçon : *principes de tests statistiques*

Le but de cet exercice consiste à comparer deux tests d'ajustement pour des lois discrètes. Ces tests sont le test χ^2 et le le test du maximum de vraisemblance.

On va considérer le modèle statistique suivant. En se basant sur un n-échantillon $\mathbf{X} = (X_1, \dots, X_n)$ de loi discrète $P = (p_1, \dots, p_d)$, où $p_k = P(X_1 = k),\ k = 1, \dots, d$, on veut tester *l'hypothèse nulle simple*

$$H_0 : P = P_0 = (p_1^0, \dots, p_d^0),$$

contre *l'alternative composite*

$$H_1 : P \neq P_0.$$

Soient

$$\widehat{p}_k = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\Big(X_i = k\Big), \quad k = 1, \dots, , d.$$

Le test χ^2. Ce test accepte l'hypothèse H_0 si

$$T^\chi(\mathbf{X}) \leq h_\alpha^\chi,$$

où

$$T^\chi(\mathbf{X}) = n \sum_{k=1}^{d} \frac{(\widehat{p}_k - p_k^0)^2}{p_k^0}$$

et le seuil h_α^χ est défini comme la valeur minimale de x t.q.

$$P_0\Big(T_\chi(\mathbf{X}^n) > x\Big) \leq \alpha$$

La motivation de ce test se base sur
Théorème. *Lorsque* $n \to \infty$

$$T^{\chi}(\mathbf{X}) \xrightarrow{P_0} \chi^2_{d-1},$$

où χ^2_{d-1} *est la loi* χ^2 *à* $d-1$ *degrés de liberté.*

Le test du maximum de vraisemblance. Ce test accepte l'hypothèse H_0 si

$$T^{MV}(\mathbf{X}) \leq h^{MV}_{\alpha}$$

où

$$T^{MV}(\mathbf{X}) = -2n \sum_{k=1}^{d} \widehat{p}_k \log \frac{p^0_k}{\widehat{p}_k}$$

et le seuil h^{MV}_{α} est défini comme la valeur minimale de x t.q.

$$P_0 \left(T^{MV}(\mathbf{X}) > x \right) \leq \alpha$$

Le résultat suivant permet d'obtenir h^{MV}_{α} pour les valeurs n suffisamment grandes
Théorème. *Lorsque* $n \to \infty$

$$T^{MV}(\mathbf{X}) \xrightarrow{P_0} \chi^2_{d-1},$$

où χ^2_{d-1} *est la loi* χ^2 *avec* $d-1$ *degrés de liberté.*

- Programmer une fonction de MATLAB qui génère une matrice de variables aléatoires indépendantes de loi discrète

$$P(X_i = k) = p_k, \quad k = 1, \ldots, d.$$

- Programmer les tests χ^2 et du maximum de vraisemblance.
  - Générer une $m \times n$-matrice dont les composantes sont des variables aléatoires de loi uniforme $P_0 = (1/4, 1/4, 1/4, 1/4)$
  - Pour chaque colonne de cette matrice calculer les statistiques T^{MV} et T^{χ}. Donc on obtient deux vecteurs ligne de taille m

$$\begin{aligned} T^{MV} &= (T^{MV}_1, \ldots, T^{MV}_m) \\ T^{\chi} &= (T^{\chi}_1, \ldots, T^{\chi}_m) \end{aligned}$$

– Calculer les quantiles empiriques $\bar{h}_\alpha^{MV}$, $\bar{h}_\alpha^{\chi}$ d'ordre $\alpha = 0.001$ pour T^{MV} et T^{χ}. C'est-à-dire trouver, par exemple, $\bar{h}_\alpha^{\chi}$ t.q.

$$\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\Big(T_i^{\chi} \geq \bar{h}_\alpha^{\chi}\Big) = \alpha.$$

- Tester

$$H_0 : \quad P = (0.25, 0.25, 0.25, 0.25)$$

contre l'alternative

$$H_1 : \quad P = (0.528, 0.3262, 0.1417, 0.0041).$$

  – Générer une $m \times n$-matrice dont les composantes sont des variables aléatoires de loi discrète $P = (0.528, 0.3262, 0.1417, 0.0041)$.
  – Pour chaque colonne de cette matrice calculer les statistiques T^{MV} et T^{χ}.
  – Calculer l'erreur de deuxième espèce de ces tests. Par exemple, l'erreur de deuxième espèce du test χ^2 ce calcule comme

$$\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\Big(T_i^{\chi} \leq \bar{h}_\alpha^{\chi}\Big)$$

- Lancer votre programme pour $n = 10, 20, 30, 40, 50$ ($m = 30000$) et tracer le log de l'erreur de deuxième espèce en fonction de n. Illustrer vos résultats comme sur la figure 1.

1.1 Une application

Le problème courant en cryptographie consiste à décider si la suite de bits

$$B^n = (b_1, b_2, \ldots, b_n)$$

provient d'un bon générateur des nombres aléatoires ou pas. Soit

$$\Sigma^n = (\sigma_1, \sigma_2, \ldots, \sigma_n)$$

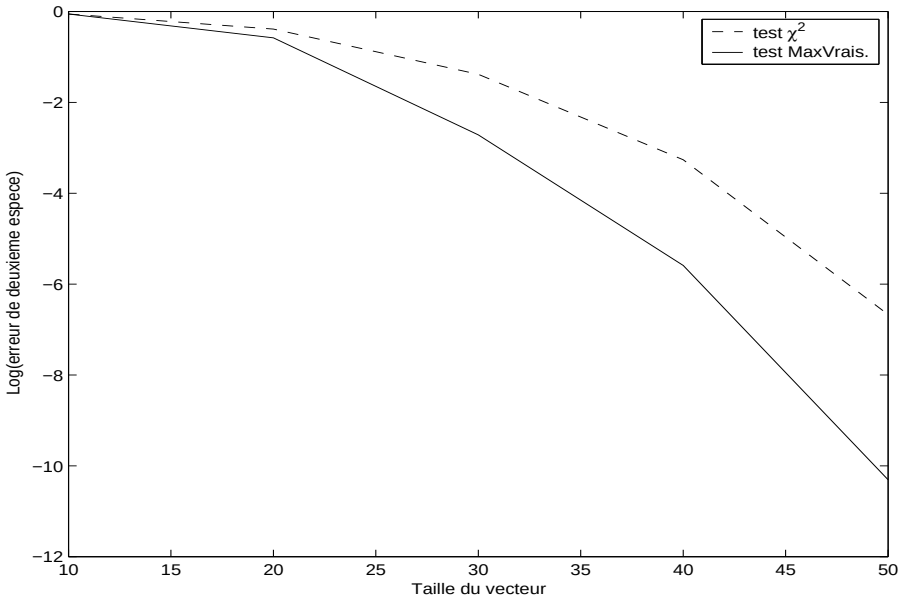


Figure 1:

un vecteur de composantes $\sigma_i \in \{0, 1\}$ qui sont les v.a. indépendantes avec $\mathbf{P}(\sigma_i = 0) = 0.5$. Par $\mathcal{L}\{X\}$ on va noter la loi de probabilité de X. Alors, il nous faut tester l'hypothèse simple

$$\begin{aligned} &H_0 : \ \mathcal{L}\{B^n\} = \mathcal{L}\{\Sigma^n\} \\ &\text{contre} \\ &H_1 : \ \mathcal{L}\{B^n\} \neq \mathcal{L}\{\Sigma^n\}. \end{aligned} \tag{1}$$

Malheureusement, ce problème est trop difficile à aborder. C'est pourquoi on cherche très souvent à transformer (1) en problème de test plus simple. Il y a une approche appelée *test par blocs* qui permet de le faire sans beaucoup de difficultés techniques. L'idée est banale:

- on partage B^n en blocs de petit taille s. Par exemple, pour $s = 2$ on a

$$B^n = \Big((b_1, b_2), (b_3, b_4) \ldots, (b_{n-1}, b_n) \Big)$$

- avec chaque petit bloc $(b_l, b_{l+1}, \ldots, b_{l+s})$ on associe un nombre entier X_l par

$$X_l = 1 + \sum_{i=0}^{s-1} b_{l+i} 2^i$$

Cela nous permet de transformer B^n en vecteur d'entiers

$$\mathbf{X} = (X_1, \ldots, X_{n/s}).$$

- on teste l'hypothèse simple

$$H_0 : \quad X_i \text{ sont i.i.d. avec } \mathbf{P}\Big(X_i = k\Big) = \frac{1}{2^s}, \quad k = 1, \ldots, 2^s$$

contre l'alternative composite

$$H : \quad X_i \text{ sont i.i.d. et il existe } k \text{ t. q. } \mathbf{P}\Big(X_i = k\Big) \neq \frac{1}{2^s},$$

Donc on arrive au problème de test d'ajustement pour la loi discrète uniforme. La pratique courante consiste à utiliser le test de χ^2, mais les simulations précédentes montrent que le test du maximum de vraisemblance peut être plus efficace que χ^2 sur des certaines alternatives.

2 Test de Kolmogorov

Leçon : *principes de tests statistiques*
fonction de répartition empirique

Il s'agit d'un test non paramétrique d'ajustement à une distribution entièrement spécifiée de fonction de répartition $F(x)$. Le problème consiste à tester

$$H_0 : \ F(x) = F_0(x) \quad \text{contre} \quad H_1 : F(x) \neq F_0(x).$$

en se basant sur un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n)$.

Soit

$$\bar{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}$$

la fonction de répartition empirique et

$$D_n(\mathbf{X}) = \sqrt{n} \max_x |\bar{F}_n(x) - F_0(x)|.$$

La région critique de test de Kolmogorov est donnée par

$$\{X : D_n(\mathbf{X}) \geq h_\alpha\}$$

où h_α est défini par

$$P_0\Big(D_n(\mathbf{X}\Big) > h_\alpha) = \alpha.$$

L'idée de ce test se repose sur

Théorème (Kolmogorov) *Si F_0 est continue, alors*

$$\lim_{n\to\infty} P_0\Big(D_n(\mathbf{X}) \leq x\Big) = \sum_{k=-\infty}^{\infty} (-1)^k \exp(-2k^2x^2).$$

- Programmer le test de Kolmogorov pour

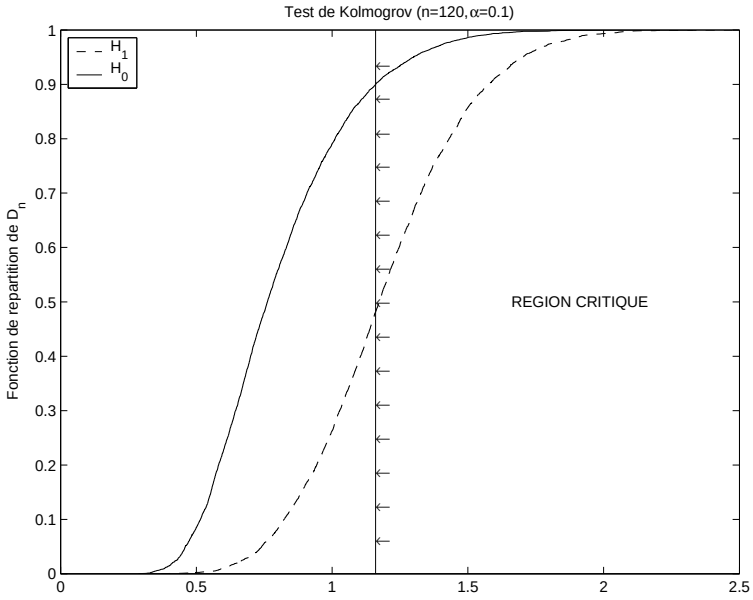
$$F_0(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-u^2/2}\, du.$$

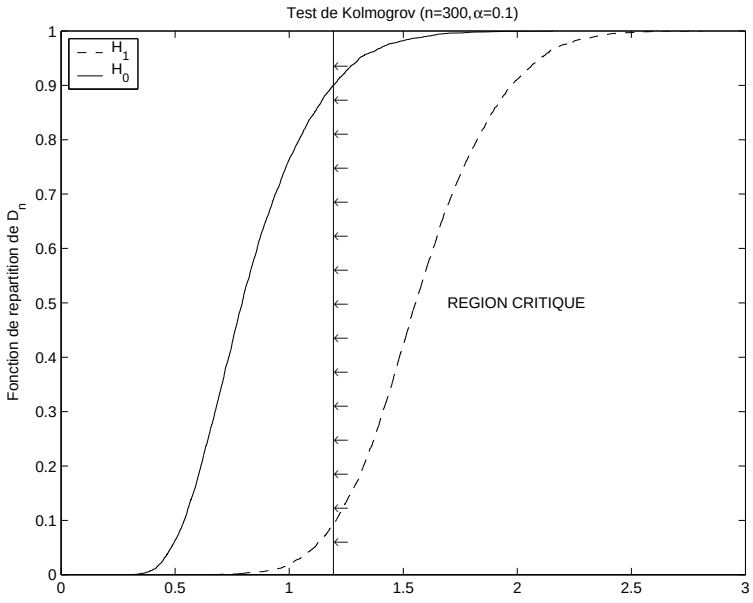
- Trouver h_α en utilisant la méthode de Monte Carlo.

- Pour $n = \{100, 150, 200, 250, 300\}$ tracer deux fonctions de répartition

$$P_0(D_n \leq x) \text{ et } P(D_n \leq x),$$

où P est la loi de Laplace avec la moyenne 0 et la variance 1. Sur les graphiques indiquer la région critique pour $\alpha = 0.1$ Visualiser vos résultats comme suit


Test de Kolmogrov (n=120, α=0.1)
H1
H0
REGION CRITIQUE
Fonction de repartition de Dn
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
1.5
2
2.5


Test de Kolmogrov (n=300, α=0.1)
H1
H0
REGION CRITIQUE
Fonction de repartition de Dn
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
1.5
2
2.5
3

3 Régression linéaire simple

Leçons :
ensembles de confiance,
régression linéaire

On va considérer le modèle de la régression linéaire simple

$$Y_i = a + bX_i + \varepsilon_i, \quad i = 1, \dots, n \tag{2}$$

où a, b sont les paramètres inconnus et ε_i sont des variables aléatoires indépendantes de $\mathbf{E}\varepsilon_i = 0,\ \mathbf{E}\varepsilon_i^2 = \sigma^2$. On suppose que la variance σ^2 n'est pas connue.

Notre but consiste à trouver dans $\mathbb{R}^2$ (en se basant sur les données $(X_i, Y_i)\ i = 1, \dots, n$) un ensemble de confiance $A_\alpha(\mathbf{X}, \mathbf{Y})$ de niveau α. C'est-à-dire

$$\mathbf{P}\Big(a + bx \notin A_\alpha(\mathbf{X}, \mathbf{Y})\Big) \leq \alpha.$$

Pour construire $A_\alpha(\mathbf{X}, \mathbf{Y})$ on utilise l'estimateur du maximum de vraisemblance de la fonction $a + bx$

$$\hat{R}(x, \mathbf{X}, \mathbf{Y}) = \hat{a}(\mathbf{X}, \mathbf{Y}) + \hat{b}(\mathbf{X}, \mathbf{Y})x$$

où

$$\begin{aligned} \hat{b} =& \frac{\langle \mathbf{X}, \mathbf{Y} \rangle - \bar{\mathbf{X}}\bar{\mathbf{Y}}}{\mathbf{var}(\mathbf{X})} \\ \hat{a} =& \bar{\mathbf{Y}} - \hat{b}\bar{\mathbf{X}} \end{aligned}$$

avec

$$\begin{aligned} \bar{\mathbf{X}} =& \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \\ \langle \mathbf{X}, \mathbf{Y} \rangle =& \frac{1}{n}\sum_{i=1}^{n} X_i Y_i, \quad \mathbf{var}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{\mathbf{X}}^2. \end{aligned}$$

Pour estimer la variance inconnue σ^2 on va utiliser l'estimateur sans biais

$$\hat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}X_i)^2.$$

Le résultat suivant constitue la base théorique pour la construction de l'ensemble de confiance $A_\alpha(\mathbf{X}, \mathbf{Y})$.

Théorème. *La variable aléatoire*

$$\frac{(\hat{a} + \hat{b}x) - (a + bx)}{\Sigma(x)}$$

où

$$\Sigma(x) = \sqrt{\frac{\hat{\sigma^2}}{n}\left[1 + \frac{(x - \bar{\mathbf{X}})^2}{\mathbf{var}(\mathbf{X})}\right]}$$

suit la loi de Student à $n-2$ degrés de liberté.

C'est pourquoi on définit

$$A_\alpha(\mathbf{X}, \mathbf{Y}) = \Big\{(x, y) : \ \hat{a} + \hat{b}x - h_\alpha \Sigma(x) \leq y \leq \hat{a} + \hat{b}x + h_\alpha \Sigma(x)\Big\}$$

où h_α est le quantile symétrique d'ordre α de la loi de Student à $n-2$ degrés de liberté, c'est-à-dire

$$\mathbf{P}\Big(|S_{n-2}| \geq h_\alpha\Big) = \alpha$$

1. Programmer un fichier script qui calcule l'ensemble de confiance $A_\alpha(\mathbf{X}, \mathbf{Y})$.
   - Générer les données $(X_i, Y_i),\ i = 1, \ldots, n$ (voir (2)) avec
     $$n = 20, \quad a = 1.5, \quad b = -1.0, \quad \sigma = 0.5$$
     et X_i indépendants uniformément distribués sur $[0, 1]$. N'oubliez pas de les ordonner!
   - Calculer $\hat{a}, \hat{b}, \hat{\sigma^2}$.
   - Trouver la valeur h_α pour $\alpha = 0.1$ (Pour le faire on peut utiliser la méthode de Monte-Carlo)
2. Tracer sur le même graphique les données (X_i, Y_i), la fonction à estimer $a + bx$ et l'ensemble de confiance $A_\alpha(\mathbf{X}, \mathbf{Y})$ comme sur la figure 2.
3. Lancer votre programme plusieurs fois pour vérifier que la région de confiance couvre la régression.

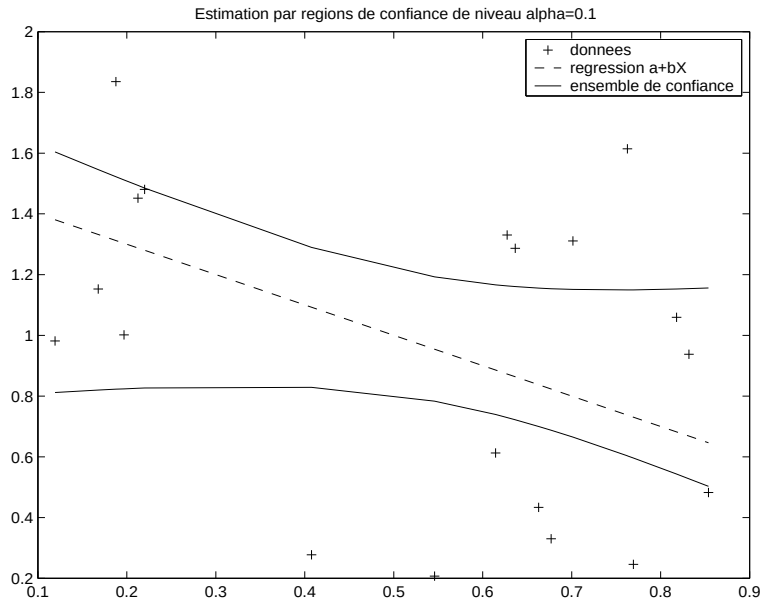


Figure 2:

4 Estimation d'un paramètre de loi exponentielle

Leçons :
ensembles de confiance,
loi exponentielle,
théorème de la limite centrale et loi des grands nombres

Le but de cet exercice consiste à modéliser un problème d'estimation paramétrique par intervalles de confiance. On considère le problème d'estimation du paramètre inconnu λ en se basant sur un n-échantillon $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbf{R}^n$ de loi exponentielle

$$\mathbf{P}_\lambda(X_i < x) = (1 - \mathrm{e}^{-x/\lambda})\mathbf{1}\{x \geq 0\}.$$

Pour une valeur $\alpha \in (0, 1)$ donnée on veut trouver un intervalle de confiance de niveau $1 - \alpha$

$$[t_1(\mathbf{X}), t_2(\mathbf{X})]$$

tel que pour tout $\lambda > 0$

$$\mathbf{P}_\lambda\{t_1(\mathbf{X}) \le \lambda \le t_2(\mathbf{X})\} \ge 1 - \alpha$$

On va construire les fonctions $t_1(\mathbf{X}), t_2(\mathbf{X})$ à l'aide de l'estimateur du maximum de vraisemblance défini par

$$\hat{\lambda}(\mathbf{X}) = \arg\max_\lambda \left\{ \prod_{i=1}^n p_\lambda(X_i) \right\}, \quad p_\lambda(X_i) = \frac{d\mathbf{P}_\lambda(x)}{dx}.$$

Il est facile de voir que

$$\hat{\lambda}(\mathbf{X}) = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

On va supposer que

$$t_1(\mathbf{X}) = t_1[\hat{\lambda}(\mathbf{X})], \quad t_2(\mathbf{X}) = t_2[\hat{\lambda}(\mathbf{X})].$$

Alors on cherche deux fonctions $t_1(\cdot)$ et $t_2(\cdot)$ telles que pour tout $\lambda > 0$

$$\mathbf{P}_\lambda\{\lambda < t_1[\hat{\lambda}(\mathbf{X})]\} \le \frac{\alpha}{2}, \quad \mathbf{P}_\lambda\{\lambda > t_2[\hat{\lambda}(\mathbf{X})]\} \le \frac{\alpha}{2}.$$

Il y a deux approches pour trouver de telles fonctions:

- **Méthode générale.** Supposons que la fonction de répartition de $\hat{\lambda}(\mathbf{X})$

  $$F_\lambda(x) = \mathbf{P}_\lambda\{\hat{\lambda}(\mathbf{X}) < x\}$$

  est connue pour tout $\lambda > 0$. Alors on peut trouver deux fonctions $x_1(\lambda)$ et $x_2(\lambda)$ t. q.

  $$F_\lambda(x_1(\lambda)) = \frac{\alpha}{2}, \quad F_\lambda(x_2(\lambda)) = 1 - \frac{\alpha}{2}.$$

  On prend

  $$t_1(\lambda) = x_2^{-1}(\lambda), \quad t_2(\lambda) = x_1^{-1}(\lambda).$$

- **Approche asymptotique.** On utilise le théorème de la limite centrale pour calculer les fonctions $x_1(\lambda)$ et $x_2(\lambda)$. Dans notre cas

  $$\frac{\sqrt{n}(\hat{\lambda} - \lambda)}{\lambda} \xrightarrow{\mathcal{L}} \xi$$

où ξ suit la loi $\mathcal{N}(0,1)$. C'est pourquoi

$$x_2(\lambda) \approx \lambda\left(1+\frac{t_{\alpha/2}}{\sqrt{n}}\right), \quad x_1(\lambda) \approx \lambda\left(1-\frac{t_{\alpha/2}}{\sqrt{n}}\right)$$

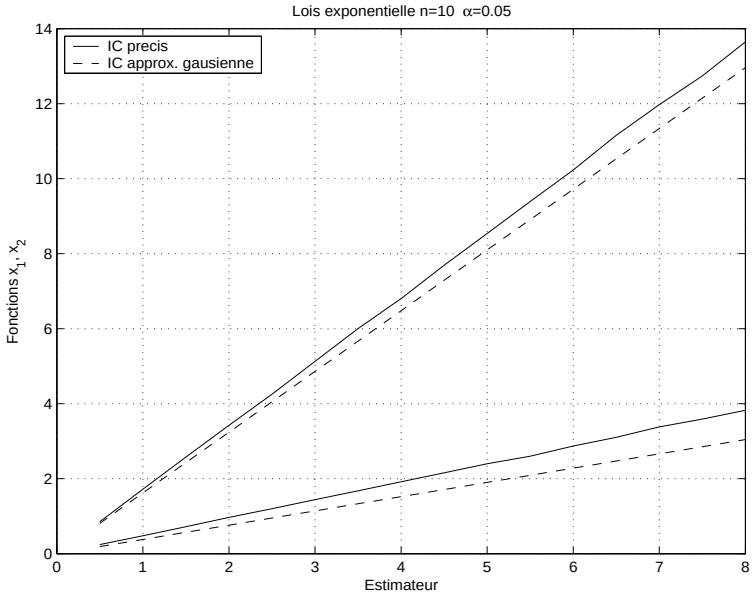
où t_α est défini par

$$\frac{1}{\sqrt{2\pi}}\int_{t_\alpha}^{\infty} \mathrm{e}^{-x^2/2}\,dx = \alpha.$$

1. Programmer un fichier script qui calcule les fonctions $x_1(\lambda)$ et $x_2(\lambda)$ par la méthode de Monte-Carlo. En statistique cette méthode est appelée *rééchantillonage* ou *bootstrap*.

   - Générer un n-échantillon de loi exponentielle de paramètre λ
   $$X_1, X_2, \ldots, X_n$$
   - Calculer l'estimateur du maximum de vraisemblance de λ
   $$\hat{\lambda} = \bar{X}$$
   - Générer une $n \times m$-matrice ($m = 20000$) dont les éléments X^*_{ij} sont des v.a. indépendantes de loi exponentielle de paramètre $\hat{\lambda}$.
   - Pour chaque colonne de cette matrice calculer la moyenne
   $$\bar{X}^*_j = \frac{1}{n}\sum_{i=1}^{n} X^*_{ij}$$
   - Soit $\bar{F}_{\hat{\lambda}}(x)$ la fonction de répartition empirique de $\bar{X}^*_j$. Trouver les valeurs x_1 et x_2 t.q.
   $$\bar{F}_{\hat{\lambda}}(x_1) = \frac{\alpha}{2} \quad \text{et } \bar{F}_{\hat{\lambda}}(x_2) = 1-\frac{\alpha}{2}$$
   Notons que x_1, x_2 sont des fonction de $\hat{\lambda}$.

2. Pour $\alpha = 0.05$ et $n = 5, 10, 20, 40$ faire une série de graphiques de $x_1(\hat{\lambda})$, $x_2(\hat{\lambda})$ et
$$\hat{\lambda}\left(1-\frac{t_{\alpha/2}}{\sqrt{n}}\right), \quad \hat{\lambda}\left(1+\frac{t_{\alpha/2}}{\sqrt{n}}\right)$$
sur l'intervalle $[0,8]$. Illustrer vos résultats comme suit


Lois exponentielle n=10 α=0.05
IC precis
IC approx. gausienne
Fonctions x1, x2
Estimateur
0
2
4
6
8
10
12
14
0
1
2
3
4
5
6
7
8