Année : 2008-2009 1er semestre
Niveau : MASTER IS 1ère année
Cours : Logiciel R
Enseignant : A. Illig

TP4 : Premières simulations en R et échantillons gaussiens

Table des matières

1 Echantillons gaussiens

1.1 Cadre

Soit $x_1, \dots x_n$ des observations de variables aléatoires $X_1, \dots, X_n$ indépendantes et identiquement distribuées (i.i.d.) de loi supposée normale $\mathcal{N}(m, \sigma^2)$. Sur la base de ces observations, on cherche à :

- obtenir une estimation de la moyenne m et de la variance σ^2,
- donner un encadrement de m et de σ^2 pour un niveau de confiance donné,
- effectuer des tests de conformité de la moyenne et de la variance.

La réponse à ces questions repose sur la normalité de l'échantillon. Cependant, dans la pratique, on dispose souvent de n mesures assimilées à des réalisations de variables aléatoires i.i.d dont la loi est inconnue. En raison du théorème de la limite centrale, on suppose fréquement que cette loi est gaussienne lorsque n est assez grand. De manière générale, il convient d'effectuer des tests, préalablement à toute étude statistique, afin de vérifier que la distibution choisie est adaptée aux données. Pour vérifier si l'hypothèse de normalité est raisonnable, on peut tracer l'histogramme des données ou bien estimer la densité de l'échantillon à l'aide d'un estimateur à noyau uniforme ou gaussien comme nous l'avons vu au TP précédent. L'hypothèse de normalité de l'échantillon peut également être vérifiée en traçant la droite de Henry (méthode du qqplot) ou en utilisant le test non paramétrique de Kolmogorov.

1.2 Etude statistique

1.2.1 Estimation ponctuelle

1. La moyenne empirique $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$ est un estimateur sans biais de la moyenne théorique m. D'après la loi des grands nombres, $\overline{X_n}$ est un estimateur fortement consistant. De plus, $\frac{\sqrt{n}(\overline{X_n}-m)}{\sigma}$ suit une loi normale $\mathcal{N}(0,1)$. *La valeur observée* $(\overline{X_n})_{obs} = \overline{x_n} = \frac{1}{n}\sum_{i=1}^n x_i$ *est appelée estimation de* m.
2. La variance empirique $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X_n})^2$ est un estimateur sans biais de la variance σ^2. Il est fortement consistant d'après la loi des grands nombres. De plus, $(n-1)\frac{S_n^2}{\sigma^2}$ suit une loi de Khi Deux $\chi^2(n-1)$ à $n-1$ degrés de liberté. Enfin, puisque $\overline{X_n}$ et S_n^2 sont indépendants, on en déduit que $\frac{\sqrt{n}(\overline{X_n}-m)}{S_n}$ suit une loi de Student $\mathcal{T}(n-1)$ à $n-1$ degrés de liberté. *La valeur observée* $(S_n^2)_{obs} = s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x_n})^2$ *est une estimation de* σ^2.

1.2.2 Estimation par intervalle de confiance

1. L'intervalle de confiance bilatéral de la moyenne m de coefficient de sécurité $1-\alpha$ s'écrit :

$$[\overline{X_n} - t_{1-\alpha/2}\ \frac{S_n}{\sqrt{n}} \quad ; \quad \overline{X_n} + t_{1-\alpha/2}\ \frac{S_n}{\sqrt{n}}]$$

où $t_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ d'une loi de Student $\mathcal{T}(n-1)$:

$$\mathbb{P}\{T \leq t_{1-\alpha/2}\} = 1-\alpha/2$$

pour $T \sim \mathcal{T}(n-1)$.

2. L'intervalle de confiance pour la variance σ^2 de coefficient de sécurité $1-\alpha$ s'écrit :

$$[\frac{(n-1)S_n^2}{q_{1-\alpha/2}} \quad ; \quad \frac{(n-1)S_n^2}{q_{\alpha/2}}]$$

où $q_{\alpha/2}$ (resp. $q_{1-\alpha/2}$) est le quantile d'ordre $\alpha/2$ (resp. $1-\alpha/2$) d'une loi de $\chi^2(n-1)$:

$$\mathbb{P}\{Q \leq q_{\alpha/2}\} = \alpha/2 \quad (\text{resp. } \mathbb{P}\{Q \leq q_{1-\alpha/2}\} = 1-\alpha/2)$$

pour $Q \sim \chi^2(n-1)$.

1.2.3 Test de conformité

1. Pour tester $H_0 : m = 3$ contre $H_1 : m > 3$ (resp. $m < 3$, $m \neq 3$), on décide de rejeter H_0 au niveau $\alpha = \mathbb{P}\{\text{Rejeter } H_0 \mid H_0 \text{ vraie}\}$ (aussi appelé erreur de première espèce) lorsque $T_n = \frac{\sqrt{n}(\overline{X_n}-3)}{S_n} > C$ (resp. $T_n < C$, $|T_n| > C$) avec $C = t_{1-\alpha}$ (resp. $C = -t_{1-\alpha}$, $C = t_{1-\alpha/2}$) pour des quantiles d'une loi $\mathcal{T}(n-1)$.
2. Pour tester $H_0 : \sigma^2 = 2$ contre $H_1 : \sigma^2 > 2$ (resp. $\sigma^2 < 2$), on décide de rejeter H_0 au niveau α lorsque $Q_n = (n-1)\frac{S_n^2}{2} > C$ (resp. $Q_n < C$) avec $C = q_{1-\alpha}$ (resp. $C = q_\alpha$) pour des quantiles d'une loi $\chi^2(n-1)$.

1.3 Hypothèse de normalité

1.3.1 Droite de Henry

D'après la loi des grands nombres, la fonction de répartition empirique d'un échantillon $X_1, \ldots, X_n$ converge vers la fonction de répartition théorique F de l'échantillon :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}} \xrightarrow[n\to\infty]{p.s.} F(x) \quad \forall x \in \mathbb{R}. \tag{1}$$

De même,

$$F_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i < x\}} \xrightarrow[n\to\infty]{p.s.} F(x) \quad \forall x \in \mathbb{R}. \tag{2}$$

Remarquons que si les observations $(x_i)_{i=1\dots n}$ proviennent d'une loi normale $\mathcal{N}(m,\sigma^2)$, alors les $u_i = \frac{x_i - m}{\sigma}$ sont des réalisations d'une loi normale centrée réduite. Pour $i = 1 \dots n$, notons $F_i^* = (F_n^*(x_i))_{obs}$ la proportion de données inférieures strictement à x_i et introduisons (à l'aide de la table d'une loi $\mathcal{N}(0,1)$) le nombre u_i^* tel que $\mathbb{P}(N < u_i^*) = F_i^*$ avec $N \sim \mathcal{N}(0,1)$. D'apres (2), lorsque n est assez grand, F_i^* est une bonne approximation de $F(x_i) = \mathbb{P}(N < u_i)$. Ainsi, sous l'hypothèse de normalité, les valeurs u_i^* et u_i sont proches et par conséquent, les points (x_i, u_i^*) doivent se situer à peu près sur la droite d'équation $y = \frac{x-m}{\sigma}$.

En R, la fonction `qqnorm` représente les points (x_i, u_i^*) et la droite de Henry est obtenue grâce à la commande `qqline` (en précisant en option pour les deux commandes `datax=TRUE`).

1.3.2 Test de Kolmogorov

Le test de Kolmogorov est un test non paramétrique d'ajustement à une loi de fonction de répartition F_0 spécifiée. Il repose sur le théorème de Glivenko et Cantelli précisant que la convergence (1) est uniforme :

$$D_n = \sup_x |F_n(x) - F_0(x)| \xrightarrow[n\to\infty]{p.s.} 0.$$

Par ailleurs, le théorème de Kolmogorov précise la distribution asymptotique de D_n (indépendante de la loi de l'échantillon initial) :

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n} D_n \leq y) = K(y) = \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2 y^2).$$

Pour tester,

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

- au niveau asymptotique α (lorsque $n \geq 80$) : la région critique est définie par $\sqrt{n} D_n > C$ avec C le quantile d'ordre $1-\alpha$ de la loi de fonction de répartition K.
- au niveau α (lorsque $n < 80$) : la région critique est définie par $D_n > d_n$ où d_n est le quantile d'ordre $1-\alpha$ de D_n (tabulée jusqu'à $n = 100$).

En R, on teste l'adéquation d'un vecteur `X` de données à la loi normale centrée réduite, en tapant la commande `ks.test(X,''pnorm'',0,1)` : s'affiche alors la valeur observée $(D_n)_{obs}$ de D_n et la p-value $\mathbb{P}(D_n > (D_n)_{obs})$.

2 Exercices

2.1 Exercice 1

1. Générer un échantillon `X` d'une loi $\mathcal{N}(5,2)$ de taille $n = 100$.
2. Donner une estimation de la moyenne puis une estimation de la variance.
3. Calculer l'intervalle de confiance à 95% pour la moyenne. La moyenne $m = 5$ appartient-elle à l'intervalle obtenu ?
4. Même question pour la variance.
5. Tester $H_0 : m = 4.5$ contre H_1 $m > 4.5$ au niveau 5%.
6. Tester $H_0 : \sigma^2 = 1$ contre $H_1 : \sigma^2 > 1$ au niveau 1%.
7. Générer et enregistrer dans une matrice M, $N = 100$ échantillons de taille $n = 100$ d'une loi $\mathcal{N}(5,2)$. Créer un vecteur `bool` de taille N ne comportant que des zéros. Pour chaque échantillon, calculer l'intervalle de confiance à 95% pour la moyenne. Si la moyenne $m = 5$ appartient au $i$ème intervalle, effectuer `bool[i]=1`. Calculer enfin la moyenne empirique de `bool`. Commenter.

2.2 Exercice 2

1. Générer un échantillon `Y` d'une loi normale $\mathcal{N}(3,1)$ de taille $n = 100$. Ouvrir un dispositif graphique partitionné en 3 fenêtres.
2. Dans la première fenêtre, tracer l'histogramme de `Y`.
3. Dans la deuxième fenêtre, tracer avec des couleurs différentes, la densité d'une loi $\mathcal{N}(3,1)$, les estimations par histogramme et par noyau gaussien.
4. Dans la dernière fenêtre, tracer la droite de Henry et les points (x_i, u_i^*) comme décrit précédement.
5. Effectuer le test de Kolmogorov d'adéquation à une loi normale $\mathcal{N}(3,1)$ puis à une loi normale $\mathcal{N}(4,1)$. Conclusions.