Feuille d'Exercices,éléments de correction

Vincent Guillemot

6 février 2009

Exercice 1 : Variables aléatoires continues

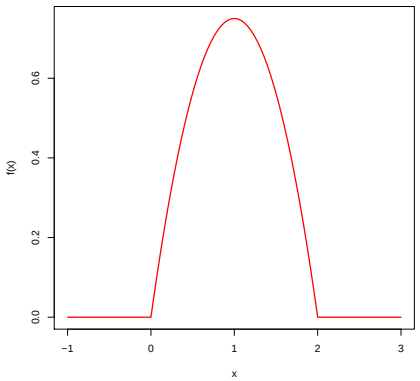
Exemple

Une variable aléatoire X admet pour densité de probabilité : $f(x) = ax(2-x)$ pour $x \in [0;2]$ et 0 ailleurs.

1. Calculer a.
2. Déterminez la moyenne, la variance et l'écart type de cette variable.
3. Tracez cette densité avec le R. Indication : utilisez la fonction `plot` pour tracer et la commande `seq(from=0,to=2,length=100)` .

Correction

1. On se sert de la propriété
$$\int_{-\infty}^{+\infty} f = \int_0^2 ax(2-x)dx = 1$$
pour trouver $a = 3/4$.
2. D'après les formules du cours, $E(X) = 1, V(X) = 1/5$ et l'écart-type de X vaut $1/\sqrt{5}$.
3. La commande `plot(function(x) 3/4*x*(2-x))` (à quelques fioritures près) nous donne le graphe suivant :



Remarquez dans le code source l'utilisation de la fonction `ifelse` pour garantir que $f(x) = 0$ quand $x \notin [0;2]$.

Densités connues

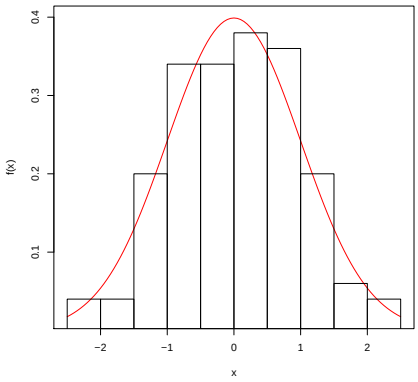
Pour chaque type de variable aléatoire à densité que nous avons vue, tracez sa densité (cherchez dans l'aide en ligne de R) et lui superposez l'histogramme d'un échantillon de taille n. Indications :

- variable uniforme, trouvez l'aide de la commande `runif`,
- variable du Khi-deux, trouvez l'aide de la commande `rchisq`,
- variable de Fisher, trouvez l'aide de la commande `rf`,
- variable de Student, trouvez l'aide de la commande `rt`.

Un exemple pour une variable aléatoire normale vous est proposé.

Correction

Il faut bien faire attention au fait que `rnorm(n=100,mean=0,sd=1)` est une fonction utilisée pour générer un échantillon gaussien (ici $\mathcal{N}(0,1)$) de taille 100, alors que c'est la fonction `dnorm(x,mean=0,sd=1)` qui est utilisée pour tracer la densité d'une variable $\mathcal{N}(0,1)$. Le graphe obtenu ressemble à celui ci-dessous. Notez dans le code source la petite modification qui permet que l'histogramme ne « déborde » pas.



Il n'y a pas de subtilité particulière pour obtenir le même graphe avec des densités uniforme, du Khi-deux, de Fisher et de Student, il faut juste bien faire attention aux paramètres à spécifier :

- a et b ($a < b$) pour la densité uniforme,
- le nombre de degré de liberté ν (« degrees of freedom ») pour la densité du Khi-deux,
- les nombres de degrés de liberté ν_1 et ν_2 pour la densité de Fisher,
- le nombre de degrés de liberté ν pour la densité de Student.

Exercice 2 : Estimation

Propriétés des estimateurs de la moyenne et de la variance

Dans une population de taille $N = 4$, une variable Y peut prendre uniquement les valeurs suivantes de façon équiprobable

$$-4, -2, 2, 4$$

1. Calculez la moyenne μ et la variance σ^2 de Y dans cette population.
2. On effectue des prélèvements d'échantillons de taille $n = 2$ sans remise dans cette population. Énumérez tous les échantillons possibles. Pour chacun d'entre eux, calculez leur moyenne empirique et leur variance empirique.

3. Vérifiez que la moyenne empirique est un estimateur sans biais de la moyenne μ.
4. Calculez la variance de la moyenne empirique.

Correction

Dans cet exercice, la variable décrite est discrète à valeurs dans un espace fini à 4 éléments $\Omega = \{-4, -2, 2, 4\}$. Il faudra donc bien faire attention à l'interprétation des résultats, puisque nous nous sommes contentés, dans le cours, de ne regarder que les propriétés des moyenne et variance de variables aléatoires à densité.

1. D'aprés le cours,

$$\mu = \frac{1}{4}(-4 - 2 + 2 + 4) = 0 \text{ et } \sigma^2 = \frac{1}{4}(16 + 4 + 4 + 16) = 10$$

.

2. Il est facile d'énumérer les échantillons possibles (attention, l'ordre n'a pas d'importance) et de calculer la moyenne et la variance empiriques pour un échantillon (y_1, y_2) :

$$\overline{y} = \frac{1}{2}(y_1 + y_2) \text{ et } S^2 = \frac{1}{2-1}\left((y_1 - \overline{y})^2 + (y_2 - \overline{y})^2\right).$$

Les résultats sont rassemblés dans le tableau ci-dessous.

y_1	y_2	$\overline{y}$	S^2
-4	-2	-3	2
-4	2	-1	18
-4	4	0	32
-2	2	0	8
-2	4	1	18
2	4	3	2

3. On vérifie aisément que la moyenne de $\overline{y}$ vaut bien μ : on a vérifié que $\overline{y}$ est un estimateur non biaisé de μ ! On a vu dans le cours que S^2 est un estimateur non biaisé de σ^2, on est donc surpris de trouver une valeur de moyenne égale à peu près à 10.33 ! L'explication vient du fait que quand Ω est un ensemble fini de cardinal N (ici $N = 2$),

$$E(S^2) = \frac{N}{N-1}\sigma^2.$$

4. La variance de la moyenne empirique nous donne toujours le résultat

$$V(\overline{Y}) = \frac{1}{6}\left((-3)^2 + (-1)^2 + 0^2 + 0^2 + 1^2 + 3^2\right) = \frac{10}{3}.$$

Encore une fois, on est surpris de ne pas retrouver la valeur que l'on a vu en cours pour des variables continues $V(Y)/n = 5$. Cela est dû au fait que, pour une population finie de taille N dans laquelle on choisit des échantillons de taille n,

$$V(\overline{Y}) = \frac{\sigma^2}{n}\frac{N-n}{N-1} = \frac{10}{2}\frac{2}{3} = \frac{10}{3}.$$

L'intérêt de cet exercice est justement de mettre en valeur les différences entre variables continues et discrètes (pour un espace fondamental fini) : attention donc à ces valeurs de N et n dans vos analyses statistiques.

Étude de la moyenne empirique sur des données simulées

Générez, avec le logiciel R, 1000 échantillons de taille 47 et stockez les dans une matrice (matrix) à 1000 lignes et 47 colonnes, échantillons qui seront gaussiens de moyenne nulle et de variance 1.

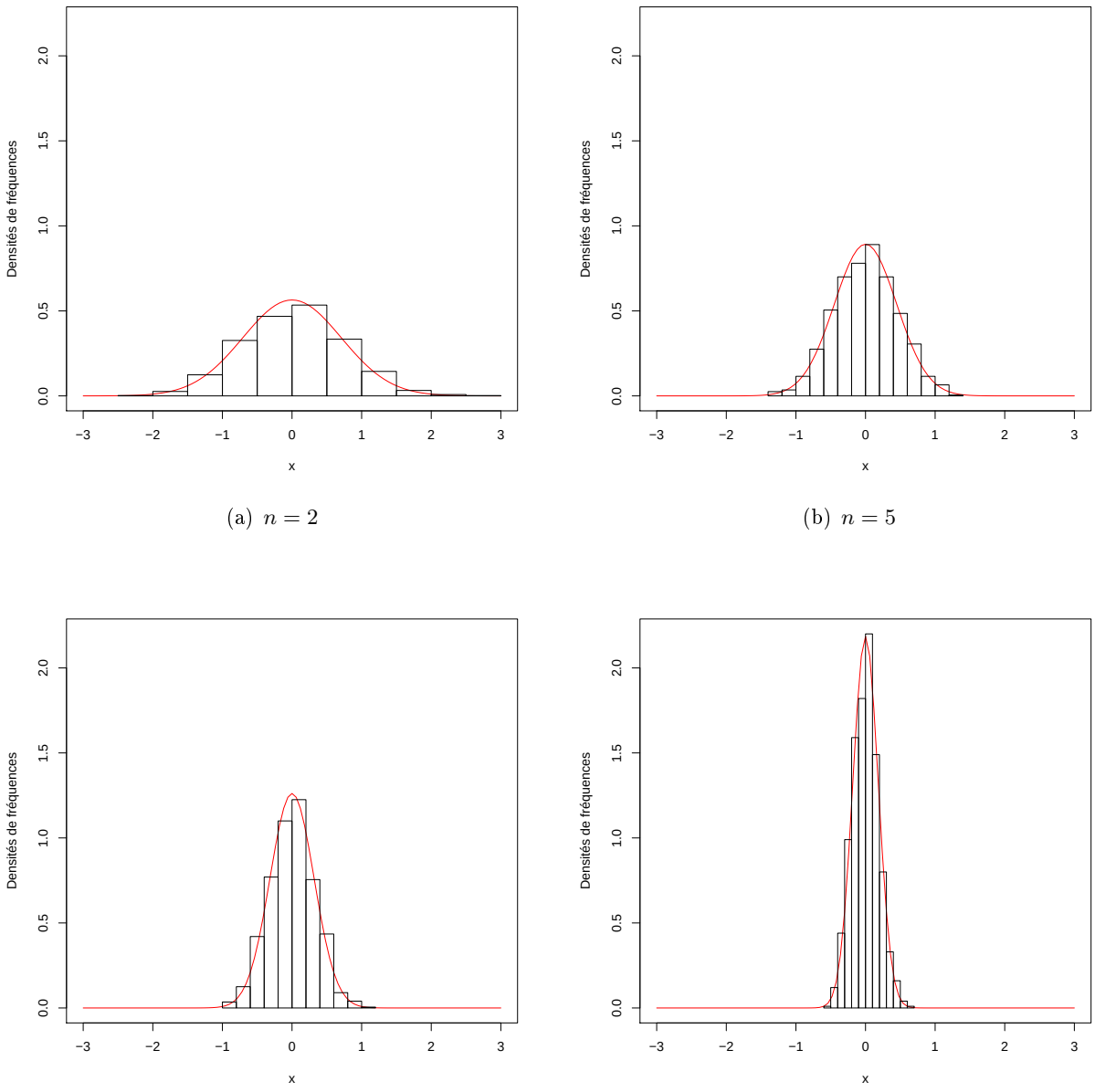
1. On veut calculer la moyenne empirique d'échantillons de taille 2, 5, 10 et 30. Utilisez les 2 premières colonnes pour calculer 1000 moyennes sur 2 individus, les colonnes 3 à 7 pour 5 individus, les colonnes 8 à 17 pour 10 individus et 18 à 47 pour 30 individus.
2. Construisez un histogramme de ces 1000 moyennes pour chaque taille d'échantillon.
3. Reprendre les mêmes questions pour une loi uniforme sur $[-1, 1]$.

Correction

Le but de l'exercice est juste de vérifier sur des simulations que plus on a d'individus pour estimer une moyenne, plus l'estimation est précise. Seuls les histogrammes sont représentés sur le graphe ci-après (Figure 1), voir le code pour plus de détails.

Les paramètres graphiques utilisés peuvent sembler abscons, les fonction essentielles à utiliser pour cet exercice sont

- `matrix`, qui permet de créer un objet matriciel,
- `rowMeans`, qui permet de calculer les moyennes de toutes les lignes d'une matrice,
- `hist`, que nous avons déjà vue précédemment, ne pas oublier de fixer la valeur du paramètre `freq` à `FALSE` : cela permet de pouvoir superposer l'histogramme et la densité associée à la variable aléatoire continue observée.



(a) $n = 2$

(b) $n = 5$

(c) $n = 10$

(d) $n = 30$

FIGURE 1 – 1000 moyennes sont générés à partir d'échantillons de taille n. On remarque que la variance de l'estimateur utilisé diminue quand n augmente. (l'histogramme est représenté en noir et la densité théorique de l'estimateur en rouge)