Université Paris X - Nanterre
Master ISIFAR, M1
Mathilde Mougeot - Hélène Boistard

Analyse des données
Octobre 2007

TD 2 - Analyse descriptive univariante

But du TD : Utiliser des outils de base dans R pour décrire des données univariantes quantitatives.

1 Description par des paramètres numériques

Soit $(x_i)_{i=1}^n$ un échantillon de n données. On calcule divers paramètres relatifs à cet échantillon.

1.1 Moments

La moyenne $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ s'obtient sur R grâce à la fonction mean.

La variance est $s_n^2 = \frac{1}{n}\sum(x_i - \bar{x}_n)^2$. Sur R, la fonction var calcule un estimateur sans biais de la variance : $\frac{1}{n-1}\sum(x_i - \bar{x}_n)^2$.

On appelle moment d'ordre p la quantité $m_{n,p} = \frac{1}{n}\sum_{i=1}^n x_i^p$.

– Créer une fonction des données et de p qui calcule le moment d'ordre p.

On appelle moment centré d'ordre p la quantité $\mu_{n,p} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^p$. On retrouve la variance pour $p = 2$. Deux paramètres de forme s'obtiennent à partir des moments centrés normalisés d'ordre 3 et 4. $SK_n = \frac{\mu_{n,3}}{(s_n^2)^{3/2}}$ s'appelle le coefficient de dissymétrie ou *skewness*. Une grande valeur de $|SK_n|$ dénote une assymétrie des données, le signe de SK_n montre le type d'assymétrie. $K_n = \frac{\mu_{n,4}}{(s_n^2)^2}$ s'appelle le coefficient d'applatissement ou *kurtosis*. Pour un échantillon gaussien, le coefficient est proche de 3. Si $K_n > 3$, la distribution a des queues plus importantes que la loi normale. Si $K_n < 3$, la distribution a des queues moins importantes que la loi normale.

– Créer une fonction des données et de p qui calcule le moment centré d'ordre p, une fonction skewness et une fonction kurtosis.

– Créer un échantillon N de taille 100 de loi normale de moyenne 2, variance 4, tracer un histogramme et calculer les coefficients de *skewness* et *kurtosis*. Faire la même chose pour un échantillon Exp de loi exponentielle de paramètre 3. Comparer.

1.2 Quantiles

On note $x_{(1)} \leq \cdots \leq x_{(n)}$ l'échantillon ordonné. La proportion de données de l'échantillon qui sont inférieures ou égales à $x_{(1)}$ est donc $1/n$, la proportion de données de l'échantillon inférieures ou égales à $x_{(i)}$ est i/n. La fonction de répartition empirique F_n est définie comme suit : pour x quelconque,

$$F_n(x) = (\text{Prop. de données} \leq x) = i/n \text{ pour } x_{(i)} \leq x < x_{(i+1)}.$$

La fonction quantile empirique, notée F_n^{-1}, se définit pour $t \in (0,1)$ par :

$$F_n^{-1}(t) = x_{(i)} \text{ lorque } \frac{i-1}{n} < t \leq \frac{i}{n}.$$

On distingue des quantiles particuliers, appelés quartiles : $Q_1 = F_n^{-1}(1/4)$, $Q_2 = F_n^{-1}(1/2)$ (qui est aussi la **médiane**), $Q_3 = F_n^{-1}(3/4)$.

– Tracer dans une même fenêtre deux graphiques correspondant à la fonction quantile empirique pour l'échantillon N et pour l'échantillon Exp. Utiliser les fonctions plot avec le type "s", sort, length.

2 Description par des graphiques

2.1 Boîte à moustaches (box-plot)

La boîte à moustaches se calcule à partir des quartiles de l'échantillon. La boîte centrale a pour limites le premier quartile Q_1 et le troisième quartile Q_3. On marque la médiane (à l'intérieur de la boîte) par un trait. Les moustaches montrent la plus petite (resp. la plus grande) observation qui se situe à une distance inférieure à $1,5(Q_3 - Q_1)$ de la boîte. Les données se situant à l'extérieur des moustaches sont considérées comme extrêmes ou *outliers*. En R, on obtient une boîte à moustaches grâce à la commande boxplot.

– Tracer un boîte à moustaches des données N et Exp de la question 1. Commenter.

2.2 Histogrammes

- Avec la fonction hist, former un histogramme de N et de Exp.
- En utilisant le paramètre breaks de hist, proposer un histogramme avec de nouvelles classes.

3 Application à des données disponibles dans la librairie de R

On étudiera au choix les tableaux de données suivants : iris, rock, LifeCycleSavings. Décrire la population étudiée, les données et leur type (qualitative, quantitative, et si quantitative, discrète ou continue). Utiliser les outils suivants pour étudier chaque variable quantitative (en choisir 3) :

- Calculer la moyenne, la médiane, le coefficient de *skewness* et de *kurtosis*. Qualifier les échantillons (symétrie, dispersion).
- Tracer une boîte à moustaches. Commenter.
- Tracer un histogramme des données, commenter. Que dire de la normalité des données, à première vue ?
- Comme dans le TD 1, tracer un graphique quantile-quantile pour étudier la gaussiannité des données. Commenter. Effectuer un test de Shapiro-Wilks et conclure.
- Choisir une variable pour laquelle il peut être intéressant de changer les classes de l'histogramme, et tracer le nouvel histogramme. Commenter.