Rappels de statistique mathématique (compléments du Chapitre 3)

Yves Aragon*
Université Toulouse 1 Capitole

7 mars 2011

Introduction

On a rassemblé ici des notions élémentaires de statistique mathématique constamment utilisées dans l'analyse des séries temporelles : les propriétés de base de la loi normale et un rappel sur les tests d'hypothèse paramétrique.

3.1 Loi normale et loi de χ^2

Loi normale. Soit $\mathbf{X} = [X_1, \cdots, X_n]'$ un vecteur aléatoire (v.a.). $\mathbf{X}$ a une distribution normale multidimensionnelle (ou multivariée) de paramètres μ, $n \times 1$ et $\Sigma = \Sigma_{\mathbf{X},\mathbf{X}}$ $n \times n$, définie positive, et on écrit $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$, si la densité de probabilité du vecteur $\mathbf{X}$ est :

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-n/2} \det(\Sigma)^{-1/2} \exp[-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)] \tag{3.1}$$

On montre que $\mathsf{E}(\mathbf{X}) = \mu$ et que la matrice des covariances de $\mathbf{X}$ est $\mathsf{cov}(\mathbf{X}) = \Sigma$.
Loi de χ^2. La loi de χ^2 est une loi à un paramètre, pas nécessairement entier, et qui admet une densité de probabilité. On a les propriétés suivantes.

Propriété 3.1
(1) La somme des carrés de k v.a. i.i.d. $\mathcal{N}(0,1)$ suit une loi de χ^2 à k degré de liberté (ddl), on la note $\chi^2(k)$.
(2) La somme de deux variables aléatoires indépendantes, distribuées respectivement suivant des lois $\chi^2(k_1)$ et $\chi^2(k_2)$ est distribuée suivant une loi de $\chi^2(k_1+k_2)$.

*aragon@cict.fr

Propriétés de la loi normale.

Propriété 3.2

Si $\boldsymbol{X} \sim \mathcal{N}(\mu, \Sigma)$, $\mathbf{B}$ *est une matrice* $m \times n$, *de rang* m, *et* a *un vecteur réel* $m \times 1$, *alors le vecteur aléatoire*

$$\boldsymbol{Y} = a + \mathbf{B}\boldsymbol{X}$$

suit une loi normale. Sa moyenne est $a + \mathbf{B}\mu$ *et sa matrice des covariances :* $\mathbf{B}\Sigma\mathbf{B}'$.

Propriété 3.3

Σ *étant définie positive, sa factorisation de Choleski est définie :* $\Sigma = \Sigma^{1/2}(\Sigma^{1/2})'$ *où* $\Sigma^{1/2}$ *est une matrice triangulaire inférieure. Alors la variable :* $\boldsymbol{Z} = [Z_1, \cdots, Z_n] = \Sigma^{-1/2}(\boldsymbol{X} - \mu)$ *est de moyenne* 0, *de matrice des covariances,* $\Sigma^{-1/2}\Sigma(\Sigma^{-1/2})' = \mathbb{I}_n$, $\boldsymbol{Z} \sim \mathcal{N}(0_{n,1}, \mathbb{I}_n)$. *On appelle cette loi, loi normale mulivariée standardisée. La densité de* $\boldsymbol{Z}$ *est*

$$f_{\boldsymbol{Z}}(z) = (2\pi)^{-n/2} \exp[-\frac{1}{2} z' z] = \{(2\,\pi)^{-1/2} \exp[-\frac{1}{2} z_1^2]\} \cdots \{(2\pi)^{-1/2} \exp[-\frac{1}{2} z_n^2]\} \quad (3.2)$$

On reconnaît le produit des densités de n *v.a. i.i.d.* $\mathcal{N}(0,1)$.

Propriété 3.4

De (??), on voit que $\boldsymbol{Z}'\boldsymbol{Z} \sim \chi^2(n)$, *mais*

$$\boldsymbol{Z}'\boldsymbol{Z} = (\Sigma^{-1/2}(\boldsymbol{X} - \mu))'\Sigma^{-1/2}(\boldsymbol{X} - \mu) = (\boldsymbol{X} - \mu)'\Sigma^{-1}(\boldsymbol{X} - \mu) \quad (3.3)$$

qui n'est autre que l'exposant de la densité de la loi normale (??) où les valeurs $\boldsymbol{x}$ *du vecteur ont été remplacées par le v.a.* $\boldsymbol{X}$. *On énonce parfois ce résultat ainsi :* l'exposant de la densité d'une v.a. normale $\mathcal{N}(\mu, \Sigma)$ suit une loi $\chi^2(rang(\Sigma))$.

Notes.

Nor-1. Dans ce livre, v.a. est une abréviation de "variable aléatoire" comme de "vecteur aléatoire".

Nor-2. A' désigne la matrice transposée de la matrice A.

Nor-3. On peut définir une loi normale même si la matrice des covariances n'est pas inversible, mais seulement semi-définie positive. On dit alors que la loi est *dégénérée*.

Loi normale conditionnelle Considérons un vecteur normal $\mathbf{X}$, une partition de ses composantes et les partitions associées des moyenne et matrice de covariance :

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix} \underset{n_1 \times 1}{\mathbf{X}^{(1)}}, \underset{n_2 \times 1}{\mathbf{X}^{(2)}}, \ \mu = \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix} \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

La proposition suivante est souvent utilisée :

Propriété 3.5

1 $\boldsymbol{X}^{(1)}$ *et* $\boldsymbol{X}^{(2)}$ *sont indépendants si et seulement si* $\Sigma_{21} = 0$

2 *La distribution conditionnelle de* $\boldsymbol{X}^{(1)}$ *sachant que* $\boldsymbol{X}^{(2)} = \boldsymbol{x}^{(2)}$ *est*

$$\mathcal{N}(\mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \quad (3.4)$$

3.2 Test d'une hypothèse paramétrique

Situation pratique courante. Soit X une v.a.. On s'intéresse à une caractéristique de la loi de probabilité de X : moyenne, 1er quartile, variance... Notons θ cette caractéristique. C'est un nombre (ou un vecteur) non aléatoire inconnu. On dispose d'autre part d'un échantillon d'observations [1] $x_1, \cdots, x_T$ de X, d'où on tire un estimateur de θ, $\widehat{\theta}_T$. Dans beaucoup de situations, on sait par le théorème central limite, que si le nombre d'observations T est suffisamment grand, on a

$$\begin{aligned} \widehat{\theta}_T &\sim AN(\theta, \mathsf{var}(\widehat{\theta}_T)) \\ \mathsf{var}(\widehat{\theta}_T) &\xrightarrow[T\to\infty]{} 0 \end{aligned} \tag{3.5}$$

AN signifiant "approximativement normal", et enfin on dispose d'une estimation $\widehat{\mathsf{var}}(\widehat{\theta}_T)$ de $\mathsf{var}(\widehat{\theta}_T)$.

Test sur un paramètre unidimensionnel On veut tester une hypothèse sur θ du genre

$$H_0 : \theta = \theta_0, \text{ contre, par exemple, } H_1 : \theta < \theta_0, \tag{3.6}$$

H_0 est l'hypothèse nulle et H_1 l'hypothèse alternative, θ_0 est une valeur particulière de θ. Dans la situation (??), si H_0 est vraie, la statistique de test

$$Z = \frac{\widehat{\theta}_T - \theta_0}{s_{\widehat{\theta}_T}}, \tag{3.7}$$

où $s_{\widehat{\theta}_T} = \widehat{\mathsf{var}}(\widehat{\theta}_T)^{0.5}$, suit approximativement une loi $\mathcal{N}(0,1)$, $Z \sim AN(0,1)$. Si Z prend une valeur exceptionnellement faible pour une variable $\mathcal{N}(0,1)$, par exemple inférieure à -2.5, $\widehat{\theta}$ prenant des valeurs proches de la vraie valeur de θ, on conclut que la valeur θ_0 qu'on a retranchée est plus élevée que la vraie valeur, et donc on doit alors rejeter H_0 au profit de H_1 dans (??). La zone de rejet ou région critique (RC) [2] est donc, pour le couple (??), de la forme

$$Z < z_0.$$

Si on prend comme valeur z_0, la valeur z_{obs} observée pour Z sur l'échantillon, la probabilité de rejeter l'hypothèse nulle alors qu'elle est vraie, est approximativement $Pr(Z < z_{\text{obs}} | Z \sim \mathcal{N}(0,1))$. On appelle cette probabilité *niveau de signification empirique* ou *p-value*. La statistique Z est appelée t-statistique quand on choisit $\theta_0 = 0$ dans H_0.

1. Ces observations sont indépendantes en statistique classique, mais dépendantes en statistique des séries temporelles.

2. La RC est l'ensemble des valeurs de la statistique de test pour lesquelles on rejette l'hypothèse nulle.

Test sur un paramètre multidimensionnel. Parfois le test porte sur un paramètre à plusieurs composantes, c'est le cas dans le test du portemanteau où l'on veut tester qu'un vecteur dont la loi est approximativement normale, est de moyenne nulle. C'est également le cas en régression, quand on teste qu'un vecteur de paramètres prend une certaine valeur. On s'en sert au chapitre 12, *Hétéroscédasticité conditionnelle*.
Appelons $\underset{k\times 1}{\theta}$ le paramètre pour lequel on dispose d'un estimateur approximativement normal et sans biais :

$$\widehat{\theta}_T \sim AN(\theta, \Sigma_{\widehat{\theta}_T}),\ \Sigma_{\widehat{\theta}_T} \xrightarrow[T\to\infty]{} 0$$

L'hypothèse nulle est :

$$H_0 : \theta = \theta_0$$

et l'hypothèse alternative :

$$H_1 : \theta \neq \theta_0.$$

Autrement dit, sous l'hypothèse nulle, $\widehat{\theta}_T$ est de moyenne θ_0. On peut tester cette hypothèse à l'aide de la statistique de test

$$(\widehat{\theta}_T - \theta_0)' \Sigma^{-1}_{\widehat{\theta}_T} (\widehat{\theta}_T - \theta_0) \tag{3.8}$$

qui suit approximativement, sous H_0, vu (??), une loi de $\chi^2(k)$. On rejette l'hypothèse nulle pour de grandes valeurs de la statistique de test. On appelle souvent (??), *distance du* χ^2 entre l'estimation et la valeur théorique du paramètre. On peut la voir comme une distance euclidienne pondérée. La distance du χ^2 est définie également si la loi est dégénérée ; la matrice des covariances n'étant alors pas inversible, on en prend une inverse généralisée qui remplace $\Sigma^{-1}_{\widehat{\theta}_T}$ dans (??), et le nombre de ddl est le rang de la matrice des covariances $\Sigma_{\widehat{\theta}_T}$, comme on l'a noté après (??).
Il existe d'autres tests d'hypothèse sur un paramètre multidimensionnel qui prennent en compte le fait que la matrice des covariances est estimée.

3.3 Mesures et tests de normalité

Rappelons d'abord les notions d'aplatissement et d'asymétrie.

Asymétrie. L'*asymétrie* (skewness) d'une distribution de probabilité est mesurée par le coefficient d'asymétrie :

$$S = \frac{\mu_3}{\mu_2^{3/2}},$$

où $\mu_k = \mathsf{E}((X - \mathsf{E}(X))^k)$ est le moment centré d'ordre k. S est sans dimension, nul pour une distribution symétrique, comme c'est le cas de la loi normale. Un coefficient positif indique une distribution peu dispersée vers la gauche avec une queue

de distribution étalée vers la droite : on dit que la distribution est *positivement asymétrique* (*positively skewed*), c'est le cas de la loi log-normale. Dans une distribution positivement asymétrique, des valeurs supérieures à la moyenne ont plus de chances d'apparaître que des valeurs inférieures à la moyenne.

Aplatissement. L'*aplatissement* (*kurtose* ou *kurtosis*) d'une distribution est mesuréS par le coefficient d'aplatissement :

$$K = \frac{\mu_4}{\mu_2^2}.$$

K est positif et sans dimension. Il vaut 3 pour une distribution normale, et 1.8 pour une distribution uniforme continue. Un coefficient d'aplatissement élevé indique que la distribution est plutôt pointue à sa moyenne, avec nécessairement des queues de distribution épaisses (*fat tails*). En effet, comme l'intégrale sous la densité vaut toujours 1, plus la distribution est pointue près de la moyenne plus les queues de la distribution sont chargées, et donc plus le moment d'ordre 4 est important par rapport au carré du moment d'ordre 2. La distribution normale servant de référence, on a introduit l'excès de kurtosis : $K-3$. Une distribution plus pointue que la normale est dite *platykurtique* et une distribution moins pointue est dite *leptokurtique*.
Etant donné un échantillon de taille n d'une distribution, on fabrique les estimations $\widehat{S}$ et $\widehat{K}$ de S et K en remplaçant dans leur expression, les moments théoriques μ_k par les moments empiriques

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k.$$

Illustration. Simulons un vecteur x de 1000 observations i.i.d. $\mathcal{N}(0,1)$ et calculons le coefficient d'asymétrie et l'excès d'aplatissement de x et de $\log(x)$ (échantillon tiré suivant la loi log-normale).

```
> require(fUtilities)
> set.seed(5923)
> x= rnorm(1000)
> xr = exp(x)
> c(skewness(x),kurtosis(x))
[1] -0.01105251 -0.06899848
> c(skewness(xr),kurtosis(xr))
[1]  6.158772 63.372464
```

On obtient, comme on s'y attendait, des valeurs assez proches de 0 pour l'échantillon d'une variable normale, et, pour l'échantillon d'une v.a. log-normale, une asymétrie très positive et un aplatissement très élevé. L'examen des estimations non-paramétriques des densités des deux échantillons portées (fig. **??**), permet d'associer une forme de distribution aux ordres de grandeur de ces coefficients.

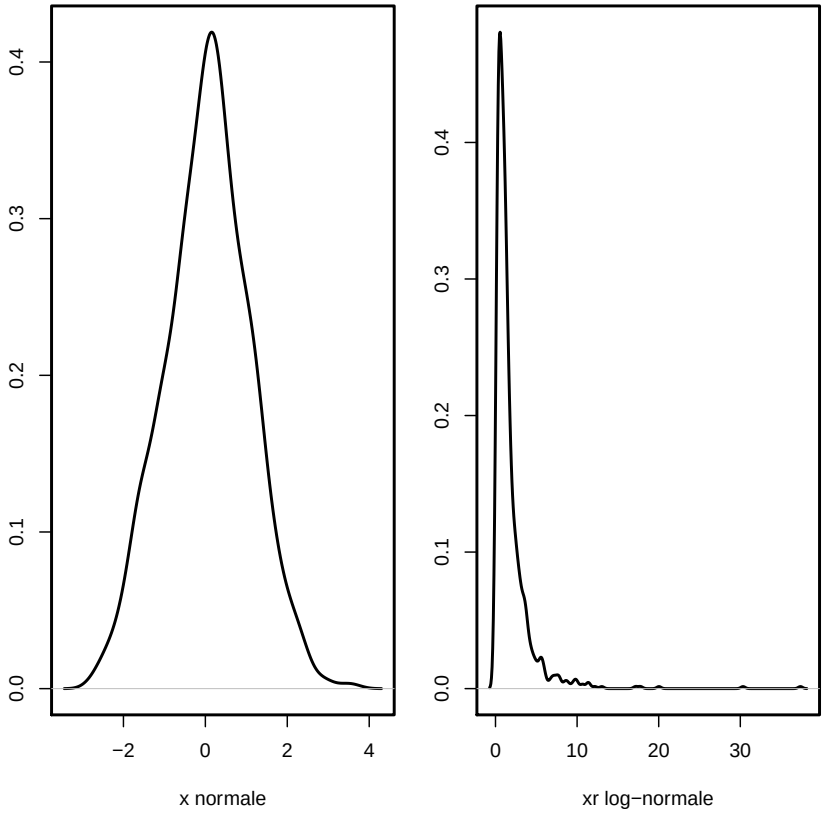

Fig. 1 – Densités d'échantillons tirés dans une loi normale (gauche) et dans une loi log-normale (droite).

Tests de normalité d'une distribution. Il existe de nombreux tests de normalité basés par exemple sur l'écart entre la distribution empirique de l'échantillon et une distribution normale, ou entre des caractéristiques de la distribution empirique de l'échantillon et les caractéristiques théoriques correspondantes de la distribution normale. Leurs puissances [3] dépendent de ce qu'est la vraie distribution et de la taille de l'échantillon. D'Agostino et Stephens (1986) en contient un exposé détaillé. Dans R on trouve des tests de normalité dans **fBasics**, **nortest** et **stats** notamment. Nous allons examiner deux tests, basés sur l'écart du couple $(\widehat{S},\widehat{K})$ d'une distribution empirique, à la valeur théorique de ce couple pour une distribution normale. Mais d'abord, nous considérons une évaluation graphique de la normalité : le QQ-plot de normalité.

QQ-plot de normalité. On se sert d'un Q-Q plot, littéralement diagramme quantile-quantile, pour vérifer visuellement si un échantillon $x_1, x_2, \cdots, x_n$ provient d'une distribution théorique supposée.
Principe. La plus petite observation de l'échantillon est le quantile empirique d'ordre $1/n$, la deuxième plus petite est le quantile d'ordre $2/n$, $\cdots$, la plus grande est le quantile d'ordre 1. (On corrige parfois ces ordres pour ne pas avoir, pour l'ordre 1, un quantile empirique fini correspondant à un quantile théorique infini, comme ce serait le cas pour la loi normale.) La distribution théorique supposée donne les quantiles théoriques en ces mêmes ordres. On construit un diagramme de dispersion de n points, un point par ordre quantile. Un point a pour abscisse le quantile théorique sous la loi supposée, et pour ordonnée le quantile empirique de même ordre. Plus les points d'un Q-Q plot sont alignés, plus la distribution supposée coïncide avec la distribution empirique de l'échantillon. Si la distribution supposée est normale, on parle de Q-Q-plot de normalité. Le coefficient de corrélation des points du Q-Q plot fournit une mesure empirique de la normalité de l'échantillon.
Le choix de la distribution supposée n'est pas restreint à la loi normale. Si nécessaire, on estime d'après l'échantillon les paramètres de la loi supposée. L'usage des Q-Q plots est intuitif. On se forge cette intuition en examinant des Q-Q plots d'échantillons simulés qu'on construit sous différentes distributions hypothétiques. On comprend ainsi ce qu'indique leur déformation par rapport à l'alignement. Les exemples de `qqnorm()` et de `qq.plot()` de **car** sont instructifs. La fonction `qreference()` de **DAAG** tire un certain nombre d'échantillons, 5 par défaut, dans la loi normale, de même moyenne et variance que l'échantillon, et dessine les Q-Q plots de normalité (SiteST).

Tests basés sur l'excès de kurtosis et sur le coefficient d'applatissement
Test dit de "Jarque-Bera". Sous l'hypothèse que l'échantillon, de taille T, est

3. La puissance d'un test est la probabilité de rejeter l'hypothèse nulle alors qu'elle est fausse.

tiré d'une distribution normale, on peut montrer que

$$JB = \frac{T}{6}(\widehat{S}^2 + \frac{(\widehat{K}-3)^2}{4})$$

suit approximativement une loi de $\chi^2(2)$. On rejette l'hypothèse nulle de normalité, pour de grandes valeurs de JB. Ce test est connu sous le nom de test de Jarque-Bera, Jarque et Bera (1980), mais D'Agostino dans D'Agostino et Stephens (1986) indique qu'il a été examiné par Bowman et Shenton en 1975. Ceux-ci, remarquant la lenteur de la convergence de $\widehat{S}$ vers la normalité, en déconseillent l'emploi. C'est un test *omnibus* : il rejette la normalité sans distinguer si ce rejet est dû à l'asymétrie ou à l'aplatissement.

Test de D'Agostino. Le test de normalité de D'Agostino est basé sur des transformations des coefficients d'asymétrie et d'aplatissement. Il évalue trois causes d'écart à la normalité : par l'aplatissement, par l'asymétrie ou la réunion des deux, dans ce cas le test est omnibus.

Nous utilisons ces tests, notamment au chapitre 12 à propos de la modélisation des rendements d'actions. Wikipedia (2010) présente un bon panorama critique des tests de normalité.

3.3.1 Transformation préalable des données

Transformation de Box-Cox. La transformation de Box-Cox est une technique pour obtenir des données plus normales que les données initiales. Elle est définie pour une variable positive par :

$$y_t(\lambda) = \frac{y_t^\lambda - 1}{\lambda},$$

où λ est un paramètre > 0.

On peut choisir λ à l'aide d'un *graphique de normalité de Box-Cox* : on porte en abscisse une grille de valeurs de λ et en ordonnée le coefficient de corrélation du Q-Q-plot de normalité pour la série transformée par les λ correspondants ; on choisit le λ qui correspond au maximum. Notons que si $\lambda \to 0$, la transformation de Box-Cox tend vers la transformation $\log(.)$.

De façon moins sophistiquée, on peut chercher, en cas de non-normalité, parmi des transformations comme : $\log(.)$, $\sqrt{.}$, une transformation qui, à vue, améliore la normalité.

Stabilisation de la variance. Le chronogramme des séries temporelles montre souvent une évolution en entonnoir : la série est croissante et les fluctuations ont de plus en plus d'ampleur au cours du temps. Schématiquement, on a affaire à une série $\{Y_t\}$ dont la moyenne, μ_t, varie avec le temps de façon déterministe et dont la variance dépend du niveau moyen :

$$Y_t = \mu_t + U_t$$

avec $\mathsf{var}(U_t) = h^2(\mu_t)\sigma^2$ pour une certaine fonction h, c'est une forme d'hétéroscédasticité. Pour traiter cette situation on cherche une transformation $g()$ telle que $\mathsf{var}(g(Y_t)) \simeq$ `constante`. C'est la technique dite de *stabilisation de la variance*. Par linéarisation, c'est-à-dire par un développement de Taylor à l'ordre 1 au voisinage de la moyenne de Y_t, et sous certaines conditions, on a :

$$g(Y_t) \simeq g(\mu_t) + (Y_t - \mu_t)g'(\mu_t)$$

et

$$\mathsf{var}(g(Y_t)) \simeq [g'(\mu_t)]^2 \mathsf{var}(Y_t)$$

On cherche donc $g()$ telle que $g'(x) = 1/h(x)$. Par exemple, pour $h(x) = x$, $g'(x) = 1/x$ et donc $g(x) = \log(x)$, pour $h(x) = \sqrt{x}$, $g'(x) = 1/\sqrt{x}$ et donc $g(x) = \sqrt{x}$. Notons qu'ici l'aspect temporel joue un rôle mineur, le seul aspect important est la dépendance de la variance par rapport à la moyenne. On choisit $h(.)$ d'après le chronogramme de la série.
Observons que ces transformations, Box-Cox ou autres, ne sont pas linéaires et donc la moyenne de la série transformée n'est pas la transformée de la moyenne de la série initiale. Ce point sera repris dans le traitement de la série `log(kwh)`, au chapitre 10, *Consommation d'électricité.*

Bibliographie

Jarque C.M. et Bera A.K. (1980). Ecient tests for normality, homoscedasticity and serial independence of regression residuals. Economics Letters, 6, 255 259.

D'Agostino R.B. et Stephens M., (Eds.) (1986). Goodness-of-Fit Techniques. Marcel Dekker.