Tests non-paramétriques de comparaison de distribution

Laurent Ferrara

M1 Modélisation Appliquée
Univ. Paris Ouest
février 2012

Test du χ^2

Objectifs

- Tester si une réalisation observée $(x_1, \ldots, x_n)$ provient d'un n-échantillon $(X_1, \ldots, X_n)$ ayant une loi spécifique

Hypothèses

- H0: $(x_1, \ldots, x_n)$ provient d'une certaine loi de distribution ($\mathcal{N}$, t, Log$\mathcal{N}$, ...)
- H1: $(x_1, \ldots, x_n)$ ne provient pas de cette loi de distribution

Statistique de test

- On range par ordre croissant $(x_{(1)}, \ldots, x_{(n)})$
- On regroupe les données rangées en K classes équidistantes B_j, $j = 1, \ldots, K$, de largeur k, telles que $B_j =]b_{j-1}, b_j]$
- La statistique de test est donnée par:

$$T = n \sum_{j=1}^{K} \frac{(O_j - E_j)^2}{E_j} \tag{1}$$

où $O_j = 1/n \sum_{i=1}^{n} 1_{x_i \in B_j}$ est la fréquence observée pour B_j
où $E_j = F_X(b_j) - F_X(b_{j-1})$ est la fréquence attendue théorique de la loi où $F_x(u) = \int_{-\infty}^{u} f(z)dz$ est la cdf de X

Loi de la statistique sous H0

- Sous l'hypothèse nulle H0 que X suit une certaine loi, on montre que:

$$T \sim \chi^2_{(K-c)} \tag{2}$$

où c est le nombre de paramètres de la loi de distribution (position, dispersion, forme, symétrie, kurtosis ...)

Mise en oeuvre

- Le choix du nombre de classes K, càd le choix de la largeur k Choix standard empirique :

$$k = 0,3 \times \hat{s_n}$$

où $\hat{s_n}$ est l'écart-type empirique

- Les classes ne doivent pas être vides sinon

$$T \sim \chi^2_{(K^*-c)}$$

où $K^* \leq K$ est le nombre de classes non-vides

Généralisation

Soit $(x_1, \ldots, x_n)$ et $(y_1, \ldots, y_n)$ deux réalisations issues de 2 va X et Y

- H0: $\mathscr{L}(X) = \mathscr{L}(Y)$
- H1: $\mathscr{L}(X) \neq \mathscr{L}(Y)$

Test de Kolmogorov-Smirnov

Objectifs

- Tester si une réalisation observée $(x_1, \ldots, x_n)$ provient d'un n-échantillon $(X_1, \ldots, X_n)$ ayant une loi spécifique par comparaison des cdf empirique F_n et théorique F_X

Hypothèses

- H0: $(x_1, \ldots, x_n)$ provient d'une certaine loi de distribution $\mathscr{L}$ ($\mathscr{N}$, t, $Log\mathscr{N}$, ...)
- H1: $(x_1, \ldots, x_n)$ ne provient pas de cette loi de distribution $\mathscr{L}$

Statistique de test

$$T_n = Sup_x \left| F_n(x) - F_X(x) \right|$$

Théorème de Kolmogorov

La suite $\sqrt{n}T_n$ converge en loi de la manière suivante:

$$P(\sqrt{n}T_n < y) \overset{n\to\infty}{\longrightarrow} 1 + 2\sum_{k=1}^{\infty}(-1)^k \exp(-2k^2y^2)$$

Mise en oeuvre

On choisit

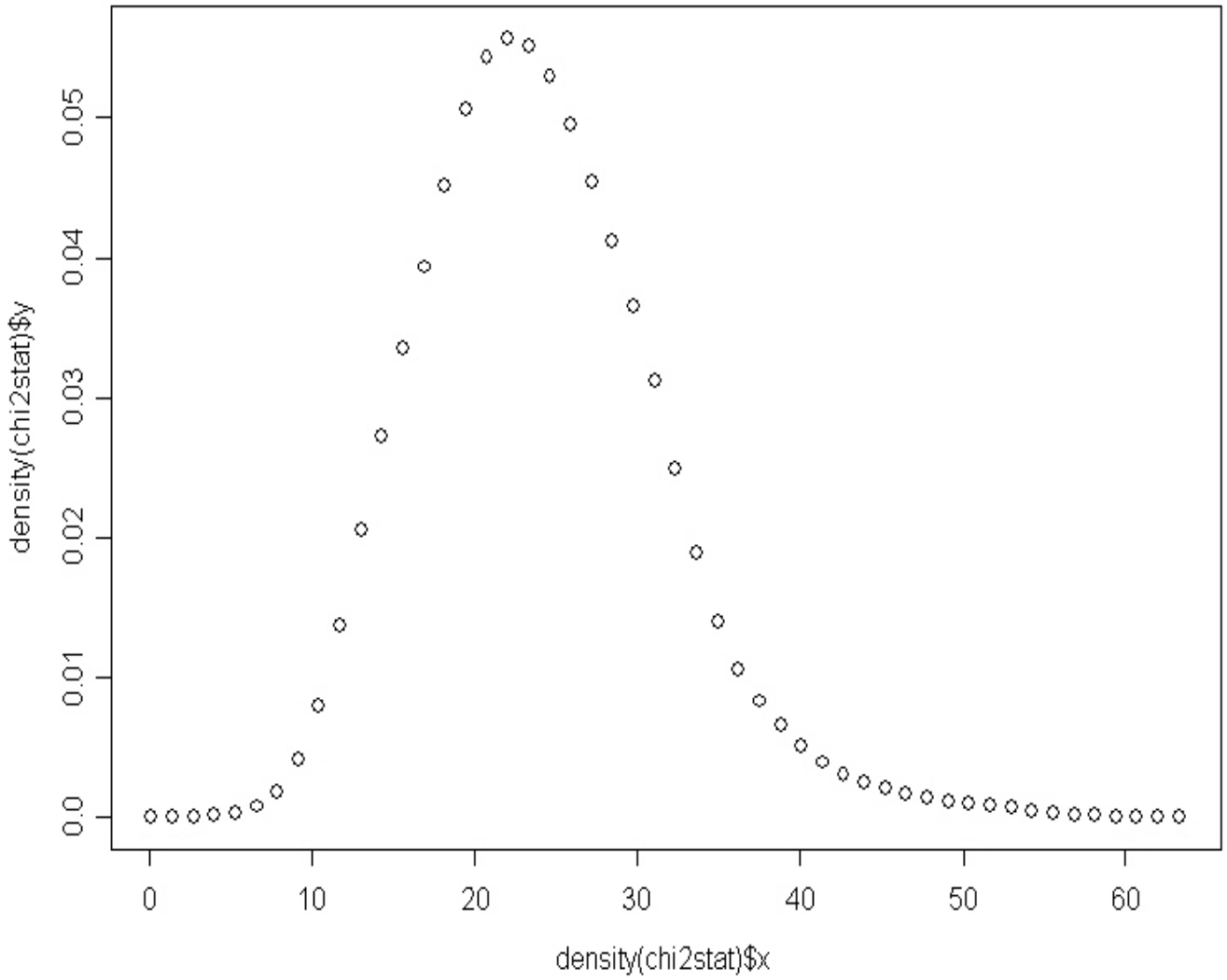
$$T = \max_{1\leq i\leq n} \left| F_n(x_{(i)}) - F_X(x_{(i)}) \right|$$

Sous H0, la loi de la statistique de test est tabulée.
On simule B réplications de la loi sous l'hypothèse H0.
On calcule les quantiles de la loi de T sous H0.

Exemple

On tabule avec $B = 1000$ la statistique pour une loi $\mathcal{N}(0, 1)$ sur un échantillon de $n = 500$ observations.

```
ksstat=numeric(1000)
for(i in 1:1000)
ksstat[i]=ks.gof(rnorm(500),distribution="normal")$statistic
plot(density(ksstat))
quantile(ksstat, probs = c(0.9, 0.95, 0.99))
90% 95% 99%
0.0364 0.0394 0.0444
```


density(chi2stat)$y
0.0
0.01
0.02
0.03
0.04
0.05
0
10
20
30
40
50
60
density(chi2stat)$x

Test de Normalité de Jarque-Bera

Objectifs

- Tester si une réalisation observée $(x_1, \ldots, x_n)$ provient d'un n-échantillon **Gaussien** $(X_1, \ldots, X_n)$ à partir des moments d'ordre 3 et 4. Moment théorique d'ordre k est donné par :

$$m_k = \frac{E(X^k)}{E(X^2)^{k/2}}$$

Hypothèses

- H0: $(x_1, \ldots, x_n)$ provient d'une loi normale $\mathcal{N}$
- H1: $(x_1, \ldots, x_n)$ ne provient pas de cette loi normale $\mathcal{N}$

Sous H0, on a les résultats asymptotiques suivants:

$$\sqrt{n}\hat{m}_3 \xrightarrow{\mathscr{L}} \mathscr{N}(0,6) \tag{3}$$

$$\sqrt{n}\hat{m}_4 \xrightarrow{\mathscr{L}} \mathscr{N}(3,24) \tag{4}$$

$$\hat{m}_3 \text{ ind } \hat{m}_4 \tag{5}$$

Statistique de test :

$$T = n\left(\frac{\hat{m}_3^2}{6} + \frac{(\hat{m}_4 - 3)^2}{24}\right)$$

Sous H0,

$$T \sim \chi^2(2)$$

Les quantiles usuels sont :

$$P(T \geq 4,60) = 0,10$$

$$P(T \geq 5,99) = 0,05$$

$$P(T \geq 9,21) = 0,01$$

Exemples

Si $\hat{T} > 9,21$, on rejette H0 au risque $\alpha = 0,01$
Si $\hat{T} \in [6;9]$, on rejette H0 au risque $\alpha = 0,05$, mais on accepte H0 au risque $\alpha = 0,01$

Série des rendements journaliers du CAC40

```
ksstat=numeric(1000) > ks.gof(cac40.rdt, distribution = "normal")
One sample Kolmogorov-Smirnov Test of Composite Normality
data:  cac40.rdt
ks = 0.0482, p-value = 0
alternative hypothesis:  True cdf is not the normal distn.  with
estimated parameters
sample estimates:
mean of x standard deviation of x
0.03246579 1.327288
```

Série des rendements journaliers du CAC40

```
ksstat=numeric(1000) > ks.gof(cac40.rdt, distribution = "t", df =
6)
One-sample Kolmogorov-Smirnov Test
Hypothesized distribution = t
data:  cac40.rdt
ks = 0.026, p-value = 0.0057
alternative hypothesis:  True cdf is not the t distn.  with the
specified parameters
```

P-value according to df of Student