Chapitre 4 : Statistique non–paramétrique : Rudiments

- Introduction
- Estimation de la densité
- Tests non-paramétriques

Problème :

Comment conduire une inférence statistique quand on ne connait pas la loi des observations $X_1, \ldots, X_n$?

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$$

avec F **inconnu**

Problème **non-paramétrique** par opposition au contexte **paramétrique** où $F(\cdot) = G_\theta(\cdot)$ et seul θ est inconnu.

Inférence statistique non–paramétrique

- Estimation d'une quantité dépendant de F

$$\theta(F) = \int h(x)\, dF(x)$$

- Décision à propos d'une hypothèse sur F

$$F \in \mathcal{F}_0\,? \qquad F == F_0\,? \qquad \theta(F) \in \Theta_0\,?$$

- Estimation de fonctions dépendant de F

$$F \qquad f(x) = \frac{dF}{dx}(x) \qquad \mathsf{E}_F[h(X_1)|X_2 = x]$$

Estimation de la densité

Pour estimer

$$f(x) = \frac{dF}{dx}(x)$$

[densité de X]

on peut songer à prendre

$$\hat{f}_n(x) = \frac{d\hat{F}_n}{dx}(x)$$

mais

$\hat{F}_n$ **n'est pas dérivable !**

Estimation par histogramme

Une première solution est de reproduire la représentation en escalier de $\hat{F}_n$ pour f

$$\hat{f}_n(x) = \sum_{i=1}^{k} \omega_i \mathbb{I}_{[a_i, a_{i+1}[}(x) \qquad a_1 < \ldots < a_{k+1}$$

en choisissant les ω_i tels que

$$\sum_{i=1}^{k} \omega_i (a_{i+1} - a_i) = 1 \quad \text{et} \quad \omega_i (a_{i+1} - a_i) = \widehat{P_F}(X \in [a_i, a_{i+1}[)$$

Estimation par histogramme (cont'd)

Par exemple,

$$\begin{aligned} \omega_i(a_{i+1} - a_i) &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{[a_i, a_{i+1}[}(X_i) \\ &= \hat{F}_n(a_{i+1}) - \hat{F}_n(a_i) \end{aligned}$$

[bootstrap]

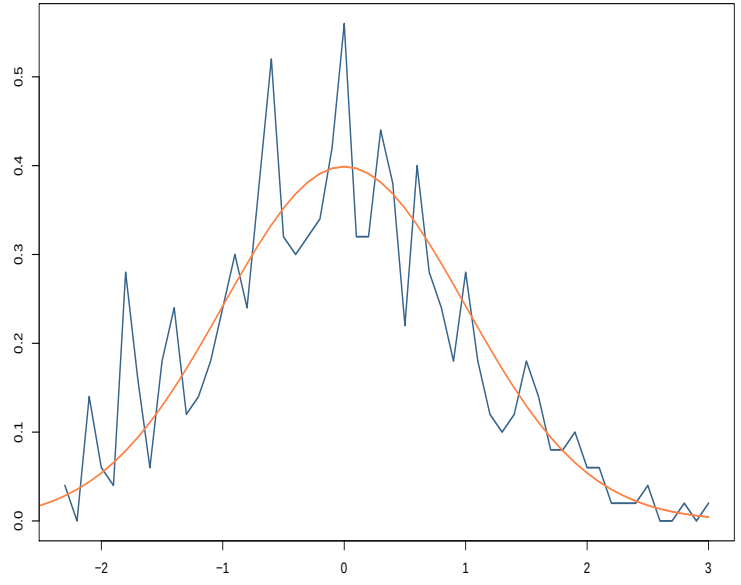
est un estimateur convergent de $P_F(X \in [a_i, a_{i+1}[)$

[Attention aux effets de bord !]

hist(x)$density

En **R**, hist(x)$density donne les valeurs des ω_i et hist(x)$breaks les valeurs des a_i

Il est préférable d'utiliser les valeurs produites par hist(x)$density pour contruire une fonction linéaire par morceaux par plot(hist(x)$density) plutôt qu'une fonction par escalier.



Estimateur par histogramme pour $k = 45$ et 450 observations normales

Interprétation probabiliste

Partant de fonctions en escalier, on aboutit à une représentation de la loi approchée comme somme pondérée d'uniformes

$$\sum_{i=1}^{k} \pi_i \mathcal{U}([a_i, a_{i+1}])$$

Equivalent à une approximation linéaire par morceaux de la fonction de répartition

$$\tilde{F}_n(x) = \sum_{i=1}^{n} \pi_i \frac{x - a_i}{a_{i+1} - a_i} \mathbb{I}_{[a_i, a_{i+1}[}(x)$$

Défauts

- Dépend du choix de la partition $(a_i)_i$, souvent construite en fonction des données (comme dans **R**)
- Problème des extrémités a_1 et a_{k+1} : ils ne peuvent pas être infinis (pourquoi?) mais doivent suffisamment approcher le support de f
- k et $(a_i)_i$ doivent dépendre de n pour permettre la convergence de $\hat{f}_n$ vers f
- **mais...** $a_{i+1} - a_i$ ne doit pas décroître trop vite vers 0 pour que l'estimation π_i soit convergente : il faut suffisamment d'observations par intervalle $[a_i, a_{i+1}]$

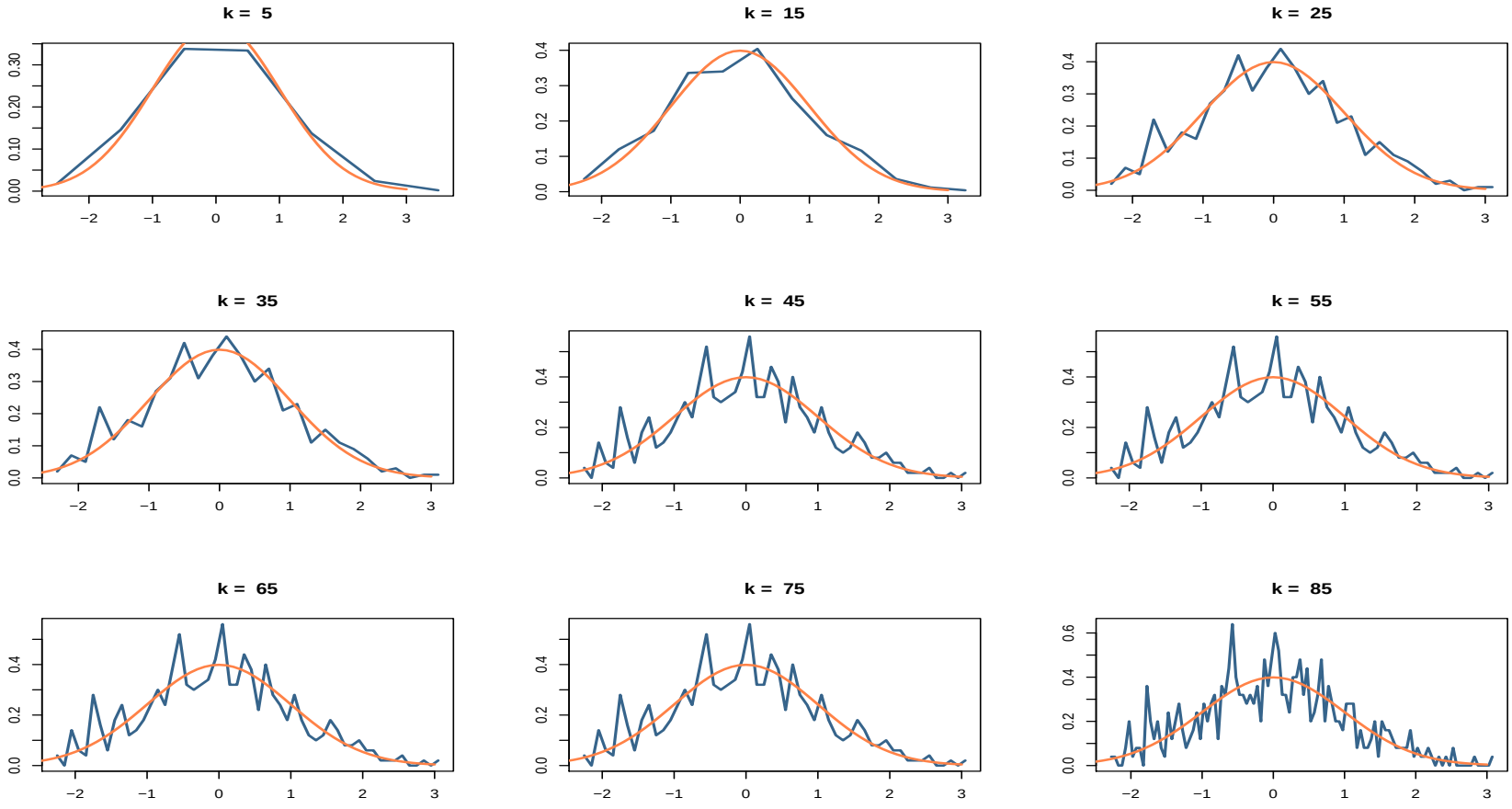
Fenêtres de Scott

Choix "optimal" de la largeur des classes :

$$h_n = 3.5\,\hat{\sigma}\,n^{-1/3} \quad \text{et} \quad h_n = 2.15\,\hat{\sigma}\,n^{-1/5}$$

donnent les bonnes largeurs $a_{i+1} - a_i$ (nclass = range(x) / h) pour $\hat{f}_n$ en escalier et linéaire par morceaux, respectivement. (Et assurent la convergence de $\hat{f}_n$ vers f quand n tend vers ∞.)

[nclass=9 et nclass=12 dans l'exemple suivant]



Variation des estimateurs par histogramme en fonction de k pour un échantillon normal de 450 observations

Estimateur du noyau

Partant de la définition

$$f(x) = \frac{dF}{dx}(x)\,,$$

on peut utiliser l'approximation

$$\begin{aligned}
\hat{f}(x) &= \frac{\hat{F}_n(x+\delta) - \hat{F}_n(x-\delta)}{2\delta} \\
&= \frac{1}{2\delta n}\sum_{i=1}^{n}\left\{\mathbb{I}_{X_i<x+\delta} - \mathbb{I}_{X_i<x-\delta}\right\} \\
&= \frac{1}{2\delta n}\sum_{i=1}^{n}\mathbb{I}_{[-\delta,\delta]}(x - X_i)
\end{aligned}$$

pour δ assez petit.

[Bon point : $\hat{f}$ est une densité]

Interprétation analytique et probabiliste

On a

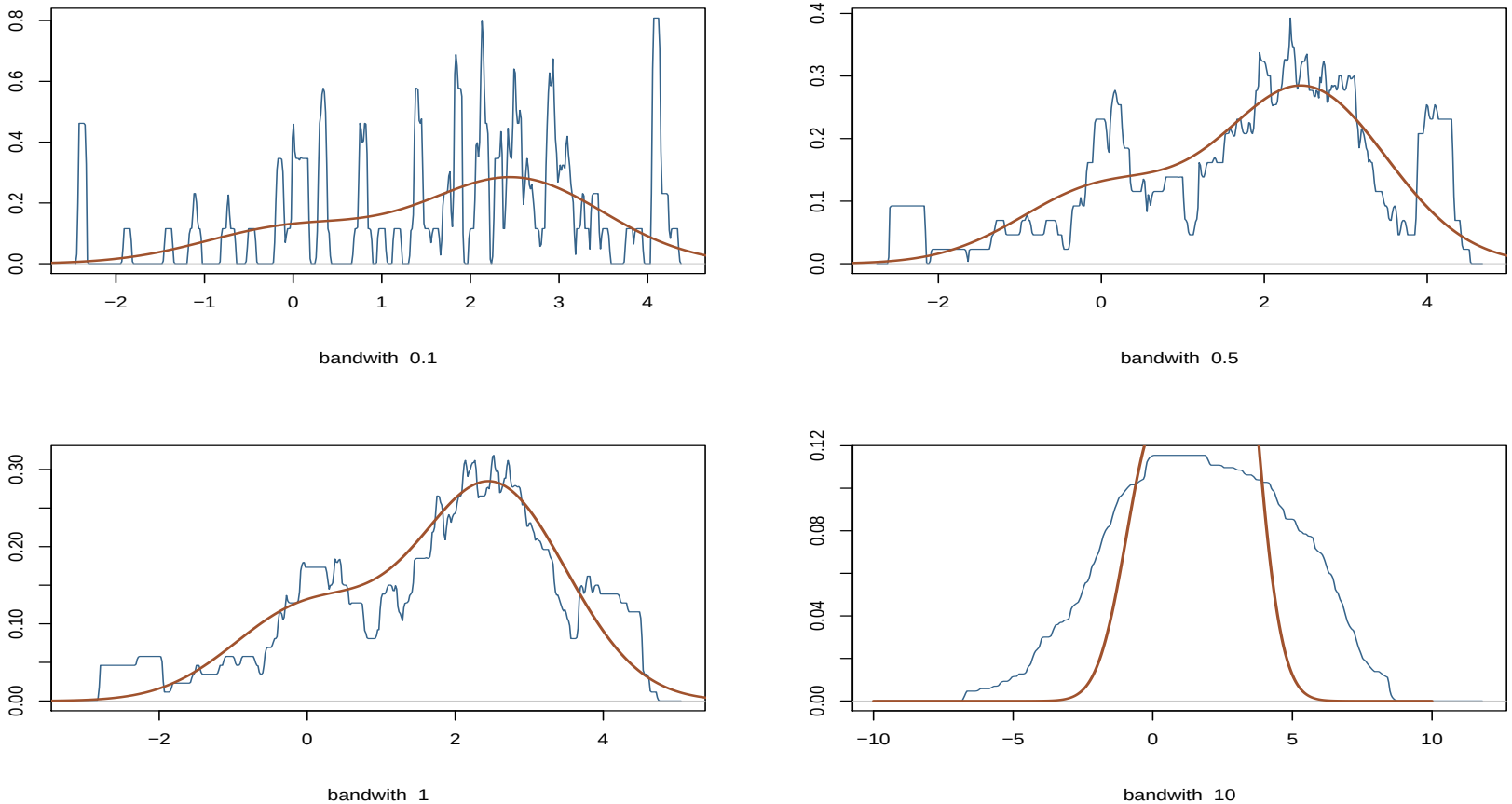
$$\hat{f}_n(x) = \frac{\text{Nb. observations proches de } x}{2\delta n}$$

Cas particulier de l'estimateur par histogramme où les a_i sont de la forme $X_j \pm \delta$

Représentation de $\hat{f}_n$ comme somme pondérée d'uniformes

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{U}([X_i - \delta, X_i + \delta])$$

[Cf. lien avec bootstrap]



Variation des estimateurs du noyau uniforme en fonction de δ pour un échantillon non-normal de 200 observations

Extension

Au lieu de considérer une approximation uniforme autour de chaque X_i, on peut utiliser une distribution plus lisse :

$$\hat{f}(x) = \frac{1}{\delta n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{\delta}\right)$$

où K est une densité de probabilité (**noyau**) et δ un facteur d'échelle "assez" petit.

En **R**, density(x)

Choix de noyaux

Toutes les densités sont en théorie acceptables. On utilise en pratique (et dans **R**)

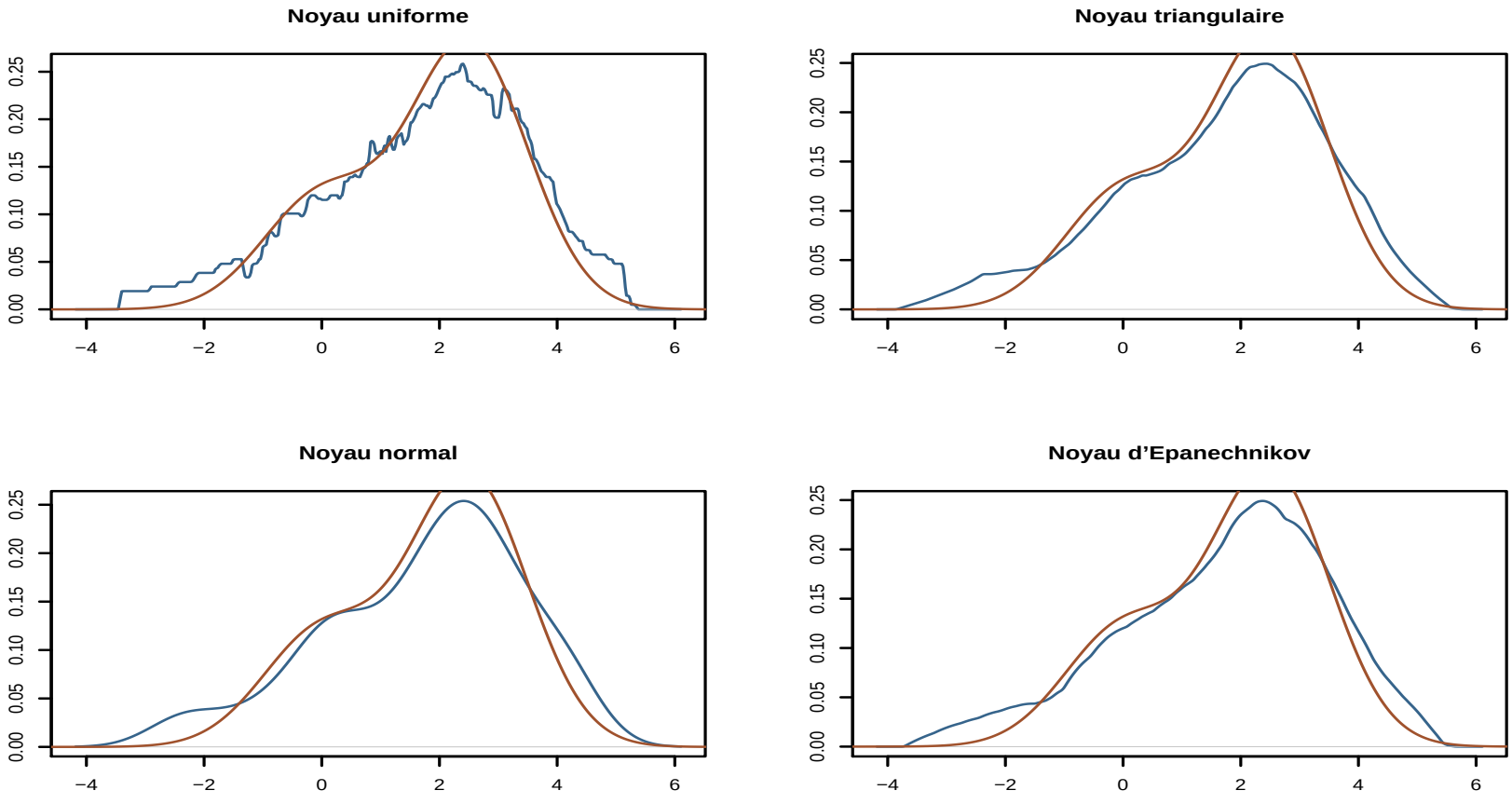
- le noyau normal [kernel="gaussian" ou "g"]
- le noyau d'Epanechnikov [kernel="epanechnikov" ou "e"]

$$K(y) = C\left\{1 - y^2\right\}^2 \mathbb{I}_{[-1,1]}(y)$$

- le noyau triangulaire [kernel="triangular" ou "t"]

$$K(y) = (1+y)\mathbb{I}_{[-1,0]}(y) + (1-y)\mathbb{I}_{[0,1]}(y)$$

Conclusion : Peu d'influence sur l'estimation de f (à l'exception du noyau uniforme [kernel="rectangular" ou "r"]).

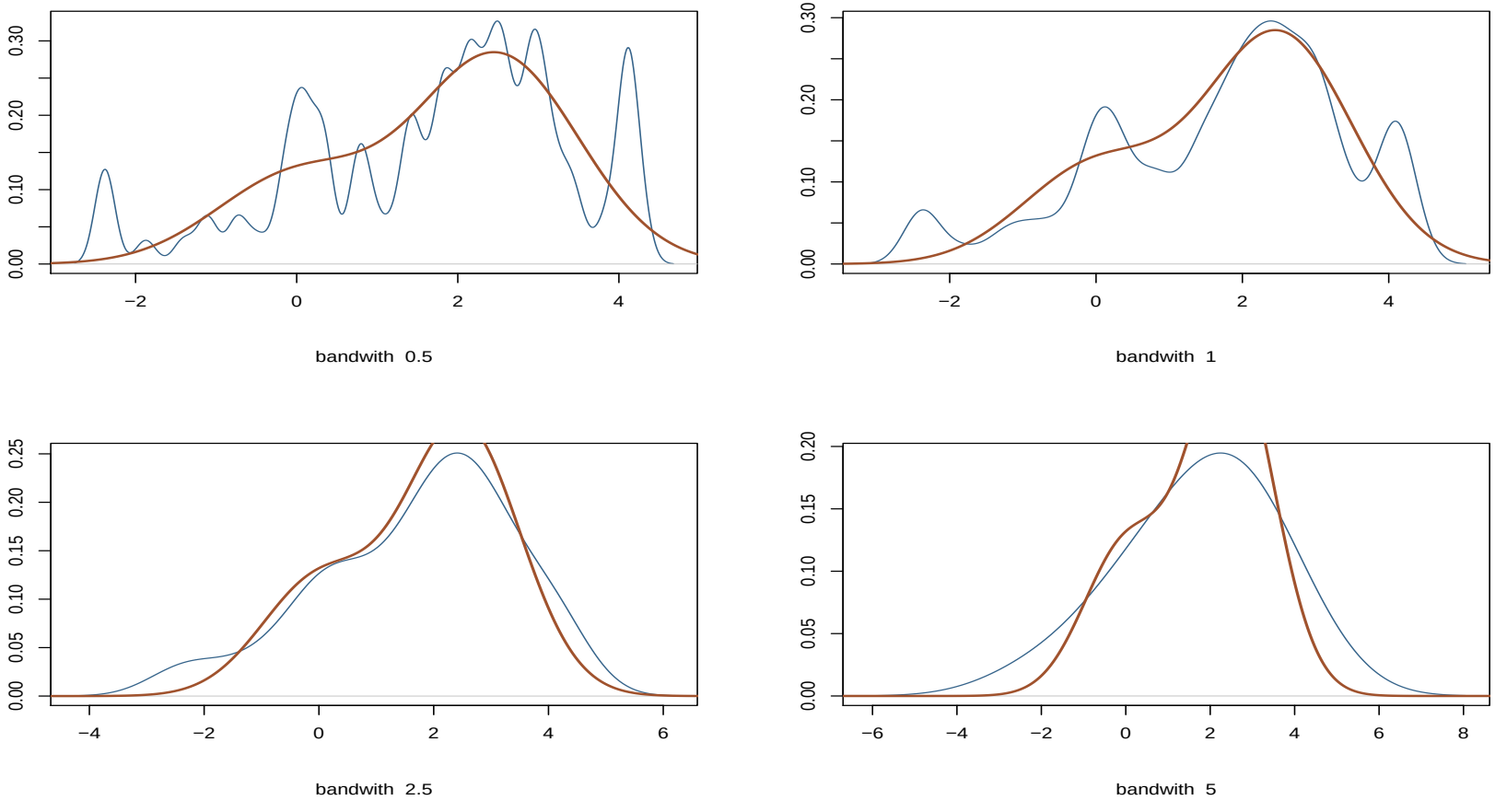


Variation des estimateurs du noyau en fonction du noyau pour un échantillon non-normal de 200 **observations**

Convergence vers f

Choix de la **fenêtre** δ crucial, par contre !

- Si δ grand, beaucoup de X_i contribuent à l'estimation de $f(x)$
  [Over-smoothing]
- Si δ petit, peu de X_i contribuent à l'estimation de $f(x)$
  [Under-smoothing]



Variation de $\hat{f}_n$ en fonction de δ pour un échantillon non-normal de 200 observations

Fenêtre optimale

En étudiant l'erreur moyenne intégrée

$$d(f, \hat{f}_n) = \mathbb{E}\left[\int \{f(x) - \hat{f}_n(x)\}^2 \, \mathrm{d}x\right] ,$$

on peut trouver un choix optimal pour la fenêtre δ, notée h_n pour souligner sa dépendence à n.

Fenêtre optimale (bis)

De la décomposition

$$\int \left\{ f(x) - \mathbb{E}\left[\hat{f}(x)\right] \right\}^2 \, \mathrm{d}x + \int \mathrm{var}\{\hat{f}(x)\} \, \mathrm{d}x \,,$$

[Biais^2+variance]

et des approximations

$$f(x) - \mathbb{E}\left[\tilde{f}(x)\right] \simeq \frac{f''(x)}{2} \, h_n^2$$

$$\mathbb{E}\left[\frac{\exp\{-(X_i - x)^2/2h_n^2\}}{\sqrt{2\pi} h_n} \right] \simeq f(x) \,,$$

[Exercice]

Fenêtre optimale (ter)

on en déduit que le biais est de l'ordre de

$$\int \left\{ \frac{f''(x)}{2} \right\}^2 \mathrm{d}x \, h_n^4$$

et que le terme de variance est approximativement

$$\frac{1}{nh_n\sqrt{2\pi}} \int f(x) \, \mathrm{d}x = \frac{1}{nh_n\sqrt{2\pi}}$$

[Exercice]

Fenêtre optimale (fin)

Par conséquent, l'erreur tend vers 0 quand n tend vers ∞ si

1. h_n tend vers 0 *et*
2. nh_n tend vers l'infini.

La fenêtre optimale est donnée par

$$\hat{h}_n^\star = \left(\sqrt{2\pi} \int \left\{ f''(x) \right\}^2 \, \mathrm{d}x \, n \right)^{-1/5}$$

Fenêtre empirique

Comme la fenêtre optimale dépend de f inconnu, on utilise une approximation de la forme

$$\hat{h}_n = \frac{0.9 \ \min(\hat{\sigma}, \hat{q}_{75} - \hat{q}_{25})}{(1.34n)^{1/5}},$$

où $\hat{\sigma}$ est l'écart-type estimé et $\hat{q}_{25}$ et $\hat{q}_{75}$ sont les quantiles à 25% et à 75% estimés.

Note : Les constantes 0.9 et 1.34 correspondent au noyau normal.

Warning! Cette formule n'est pas celle utilisée par défaut dans **R**

La problématique des tests statistiques

Face à une question sur F, comme

Est ce que F est égale à F_0, connue ?

la réponse statistique se fonde sur les données

$$X_1, \ldots, X_n \sim F$$

pour décider si **oui ou non** la question **[l'hypothèse]** est compatible avec ces données.

La problématique des tests statistiques (bis)

Une **procédure de test** (ou test statistique) $\varphi(x_1, \ldots, x_n)$ est à valeurs dans $\{0, 1\}$ (pour oui/non)

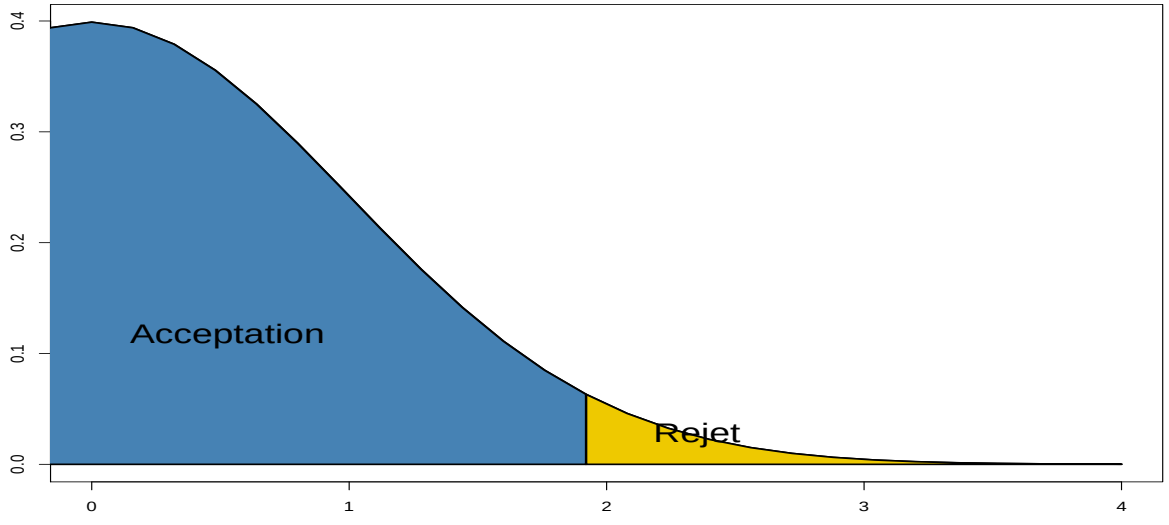
En prenant une décision sur la question sur F, on peut faire deux erreurs :

1. refuser l'hypothèse à tort (Type I)
2. accepter l'hypothèse à tort (Type II)

Il faudrait donc balancer ces deux types d'erreur.

La problématique des tests statistiques (ter)

En pratique, on se concentre sur le type II et on décide de rejeter l'hypothèse seulement si les données semblent **significativement** incompatibles avec cette hypothèse.



Accepter une hypothèse après un test signifie seulement que les données n'ont pas rejeté cette hypothèse !!!

Comparaison de distributions

Example (Deux distributions égales ?)

Soient deux échantillons $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$, de distributions respectives F et G, inconnues.
Comment répondre à la question

$$F == G \ ?$$

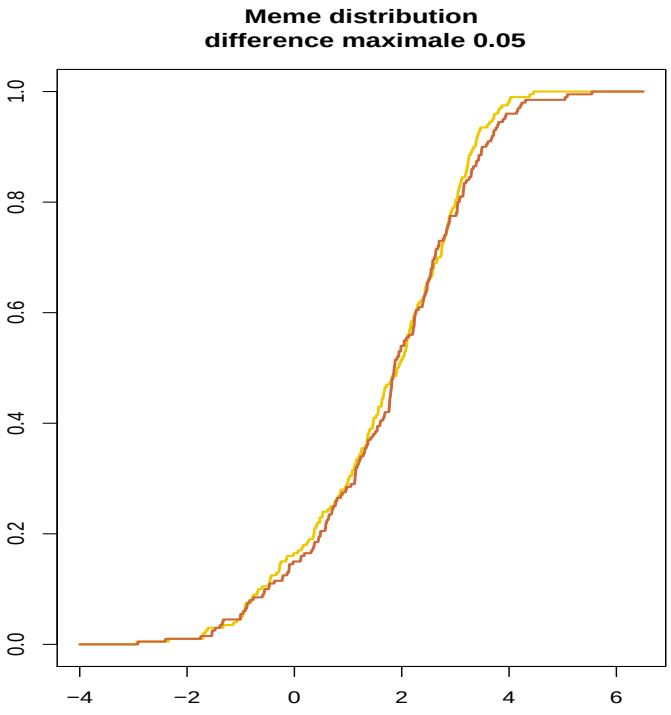
Example (Comparaison de distributions (suite))

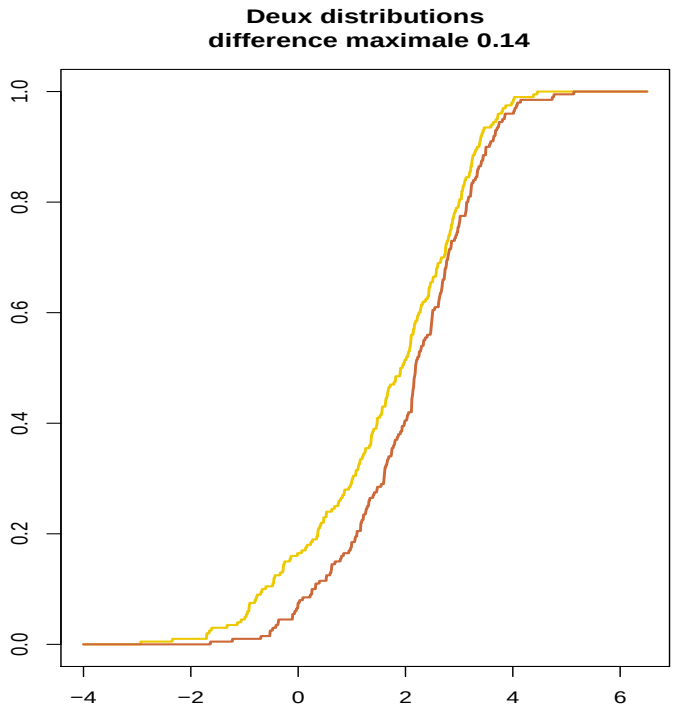
Idée :

Comparer les estimateurs de F et G,

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i \leq x} \qquad \text{et} \qquad \hat{G}_m(x) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}_{Y_i \leq x}$$

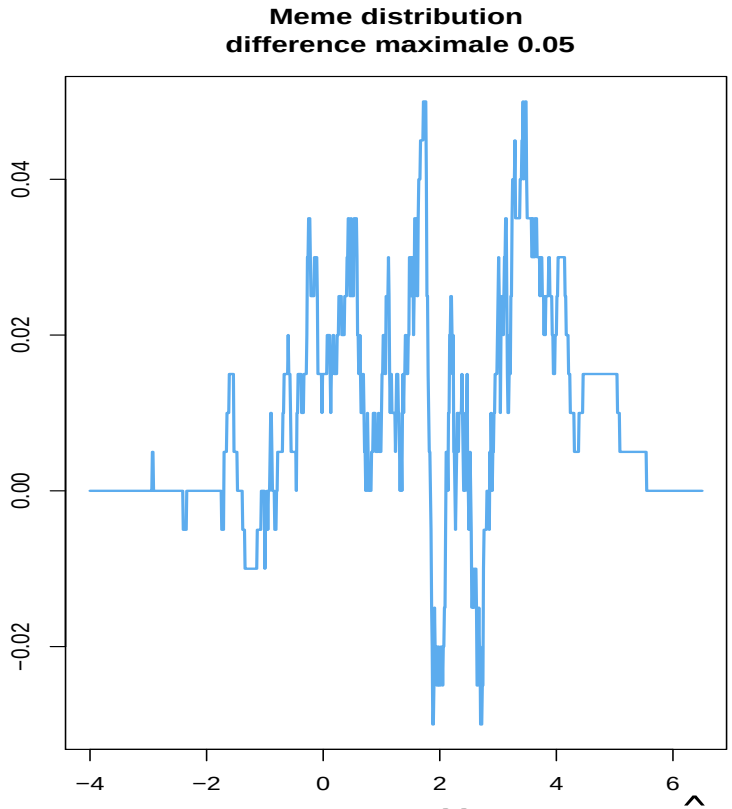
Statistique de Kolmogorov–Smirnov

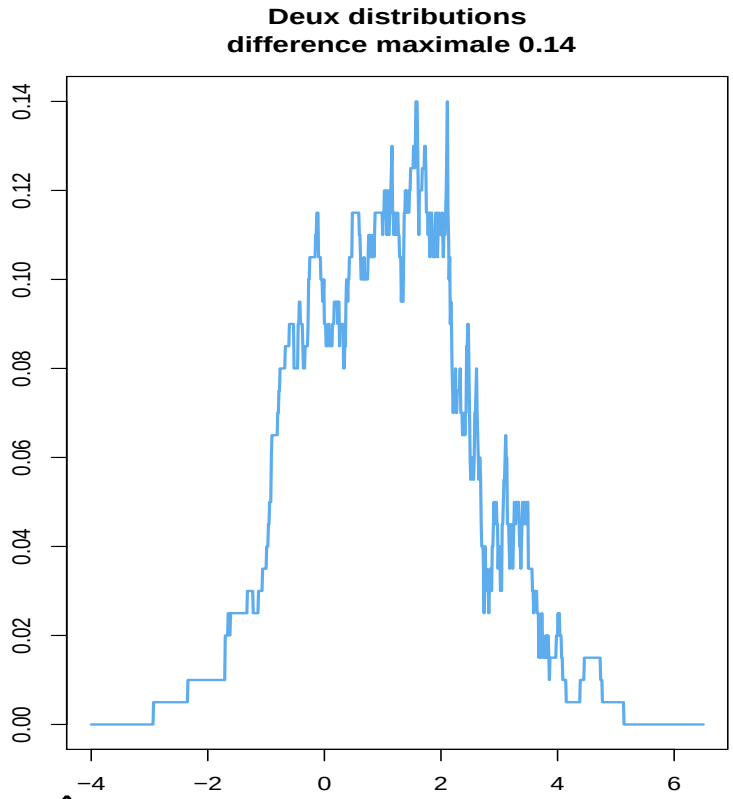




Evaluation via la différence

$$K(m,n) = \max_x \left|\hat{F}_n(x) - \hat{G}_m(x)\right| = \max_{X_i,Y_j} \left|\hat{F}_n(x) - \hat{G}_m(x)\right|$$





Evolution de la différence $\hat{F}_n(x) - \hat{G}_m(x)$ pour deux situations

Statistique de Kolmogorov–Smirnov (suite)

Utilisation :

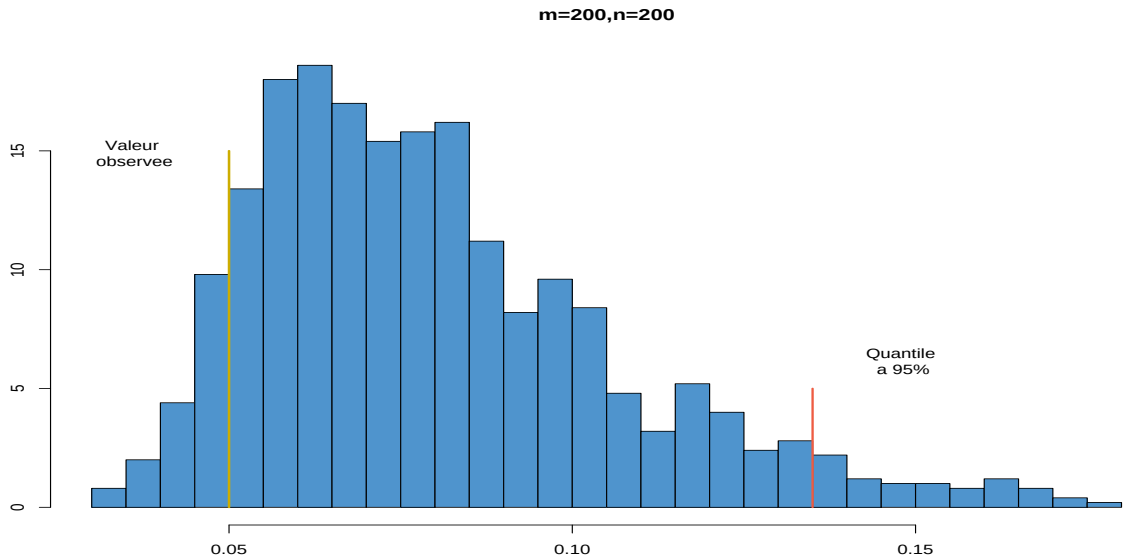
Si $K(m,n)$ "grand", les distributions F et G sont significativement différentes.
Si $K(m,n)$ "petit", on ne peut pas les distinguer au vu des échantillons $X_1,\ldots,X_n$ et $Y_1,\ldots,Y_m$, donc on "accepte" que $F=G$.

[Test de Kolmogorov–Smirnov]

En **R**, ks.test(x,y)

Calibration du test

A m et n donnés, **si** $F = G$, $K(m, n)$ a la même distribution pour tout F.
On peut se ramener à la comparaison de deux échantillons uniformes et utiliser la simulation pour approcher la distribution de $K(m, n)$ et ses quantiles.



Calibration du test (suite)

Si $K(m,n)$ observé dépasse le quantile de $K(m,n)$ sous H_0 à 90 ou 95 %, la valeur est très improbable

$$\text{si } F = G$$

et on rejette l'hypothèse d'égalité des deux distributions.

Calibration du test (suite)

Exemple de sortie R :

Two-sample Kolmogorov-Smirnov test
data: z[, 1] and z[, 2]
D = 0.05, p-value = 0.964
alternative hypothesis: two.sided

p-value = 0.964 signifie que la probabilité que $K(m, n)$ dépasse la valeur observée D = 0.05 est de 0.964, donc la valeur observée est petite pour la distribution de $K(m, n)$: **on accepte l'hypothèse d'égalité.**

Test d'indépendence

Example (Indépendence)

On cherche à tester l'indépendence entre deux v.a. X et Y en observant les couples $(X_1, Y_1), \ldots, (X_n, Y_n)$

Question

$$X \perp Y \ ?$$

Test de rang

Idée :

Si on range les X_i par ordre croissant

$$X_{(1)} \leq \dots X_{(n)}$$

les rangs R_i (ordres après rangement) des Y_i correspondants,

$$Y_{[1]}, \dots, Y_{[n]},$$

doivent être complètement aléatoires.

En **R**, rank(y[order(x)])

Test de rang (suite)

Rang : On appelle

$$\mathfrak{R} = (R_1, \ldots, R_n)$$

la **statistique de rang** de l'échantillon $(Y_{[1]}, \ldots Y_{[n]})$
La **statistique de Spearman** est

$$S_n = \sum_{i=1}^{n} i\, R_i$$

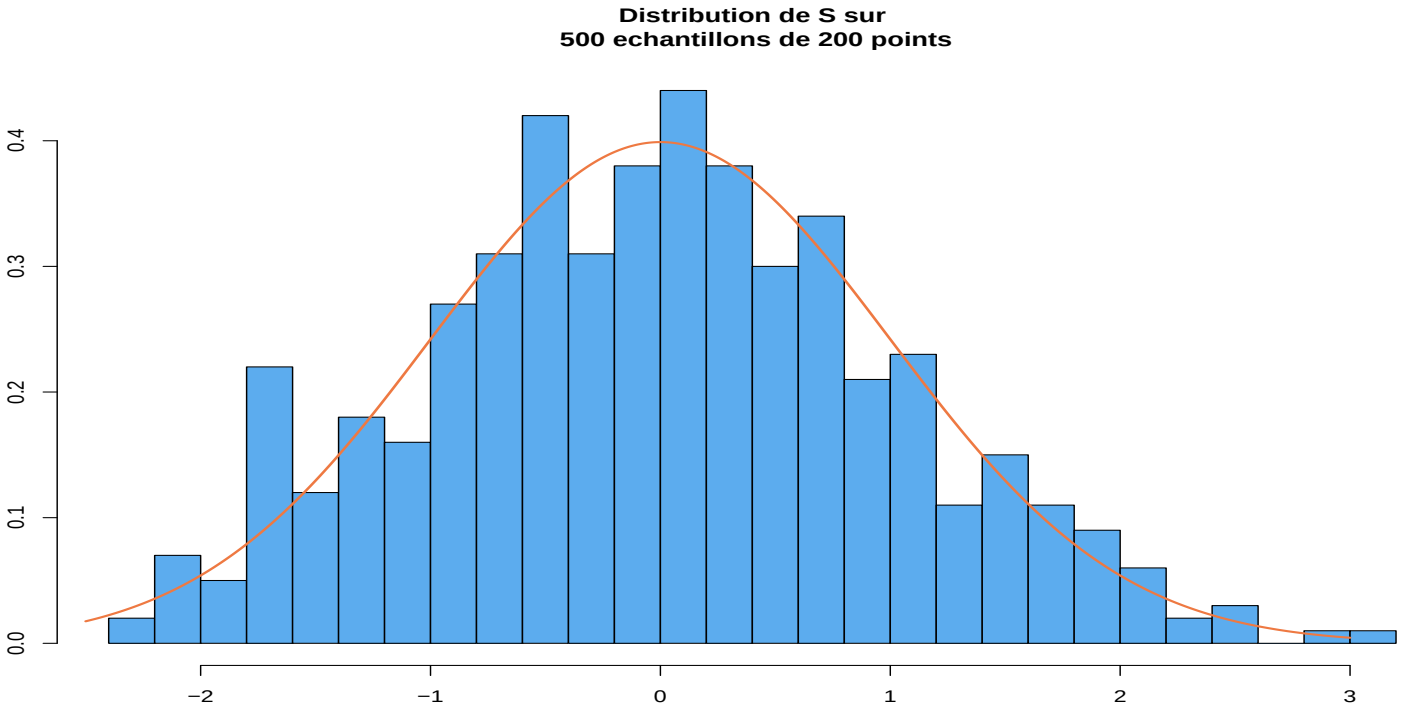
[Corrélation entre i et R_i]

On montre que, **si** $X \perp Y$,

$$\mathsf{E}[S_n] = \frac{n(n+1)^2}{4} \qquad \mathsf{var}(S_n) = \frac{n^2(n+1)^2(n-1)}{144}$$

Statistique de Spearman

Distribution de S_n disponible par simulation [uniforme] ou approximation normale



Version recentrée de la statistique de Spearman et approximation normale

Statistique de Spearman (suite)

On peut donc déterminer les quantiles à 5% et 95% de S_n par simulation et décider si la valeur observée de S_n est à l'intérieur de ces quantiles (= on accepte l'indépendence) ou à l'extérieur (= on rejette l'indépendence)

Tests multinomiaux

Example (Test du chi deux)

Une approche par histogramme permet d'apporter une réponse robuste aux problèmes de test, comme par exemple à la question

L'échantillon $X_1, \ldots, X_n$ est il normal $\mathcal{N}(0,1)$?

Idée:

On remplace le problème par sa forme discrétisée à des intervalles $[a_i, a_{i+1}]$

Est ce que

$$P(X_i \in [a_i, a_{i+1}]) = \int_{a_i}^{a_{i+1}} \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx \stackrel{\text{def}}{=} p_i \ ?$$

Principe

Modélisation multinomiale

On se ramène toujours à un problème d'adéquation à une loi multinomiale

$$\mathcal{M}_k\left(p_1^0, \ldots, p_k^0\right)$$

ou à une famille de lois multinomiales

$$\mathcal{M}_k\left(p_1(\theta), \ldots, p_k(\theta)\right) \qquad \theta \in \Theta$$

Exemples

- Dans le cas de l'adéquation à la loi normale standard, $\mathcal{N}(0,1)$, k est determiné par le nombre d'intervalles $[a_i, a_{i+1}]$ et les p_i^0 par

$$p_i^0 = \int_{a_i}^{a_{i+1}} \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx$$

- Dans le cas de l'adéquation à une loi normale, $\mathcal{N}(\theta,1)$, les $p_i(\theta)$ sont donnés par

$$p_i(\theta) = \int_{a_i}^{a_{i+1}} \frac{\exp(-(x-\theta)^2/2)}{\sqrt{2\pi}} dx$$

Exemples (suite)

- Dans le cas d'un test d'indépendence entre deux variables, X et Y,

$$X \perp Y \ ?$$

k est le nombre de cubes $[a_i, a_{i+1}] \times [b_i, b_{i+1}]$, θ est défini comme

$$\theta_{1i} = P(X \in [a_i, a_{i+1}]) \quad \theta_{2i} = P(Y \in [b_i, b_{i+1}])$$

et

$$\begin{aligned} p_{i,j}(\theta) &\stackrel{\text{def}}{=} P(X \in [a_i, a_{i+1}], Y \in [b_i, b_{i+1}]) \\ &= \theta_{1i} \times \theta_{2j} \end{aligned}$$

Test du chi-deux

L'estimateur naturel des p_i est

$$\hat{p}_i = \hat{P}(X \in [a_i, a_{i+1}]) = \hat{F}_n(a_{i+1}) - \hat{F}_n(a_i)$$

[Cf. bootstrap]

La **statistique du chi-deux** est

$$\begin{aligned} S_n &= n \sum_{i=1}^{k} \frac{(\hat{p}_i - p_i^0)^2}{p_i^0} \\ &= \sum_{i=1}^{k} \frac{(\hat{n}_i - np_i^0)^2}{np_i^0} \end{aligned}$$

si on teste l'adéquation à une loi multinomiale

$$\mathcal{M}_k \left(p_1^0, \ldots, p_k^0\right)$$

Test du chi-deux (suite)

et

$$\begin{aligned} S_n &= n \sum_{i=1}^{k} \frac{(\hat{p}_i - p_i(\hat{\theta}))^2}{p_i(\hat{\theta})} \\ &= \sum_{i=1}^{k} \frac{(\hat{n}_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})} \end{aligned}$$

si on teste l'adéquation à une famille de lois multinomiales

$$\mathcal{M}_k\left(p_1(\theta), \ldots, p_k(\theta)\right) \qquad \theta \in \Theta$$

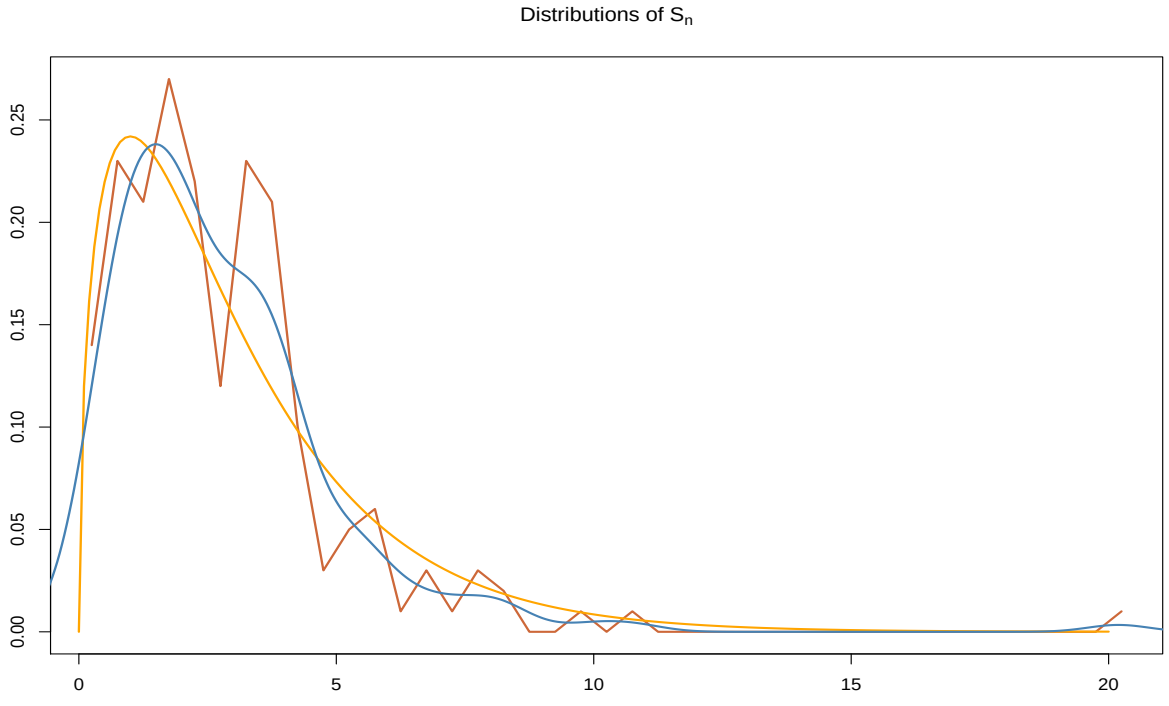
Loi approchée

Pour l'adéquation à une loi multinomiale, la loi de S_n est approximativement (pour n grand)

$$S_n \sim \chi^2_{k-1}$$

et pour l'adéquation à une famille de lois multinomiales, avec $\dim(\theta) = p$,

$$S_n \sim \chi^2_{k-p-1}$$



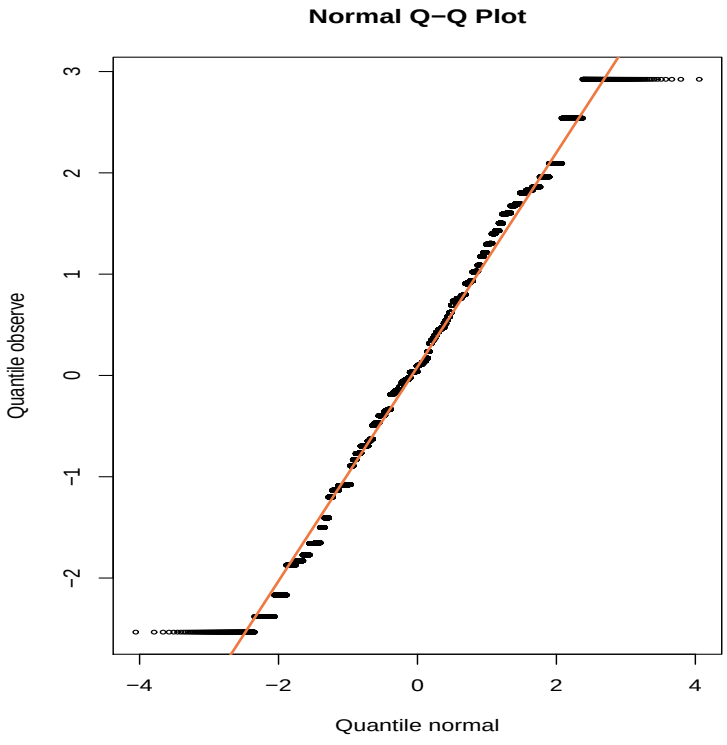
Distribution de S_n pour 200 échantillons normaux de 100 points et un test d'adéquation à $\mathcal{N}(0,1)$ avec $k = 4$

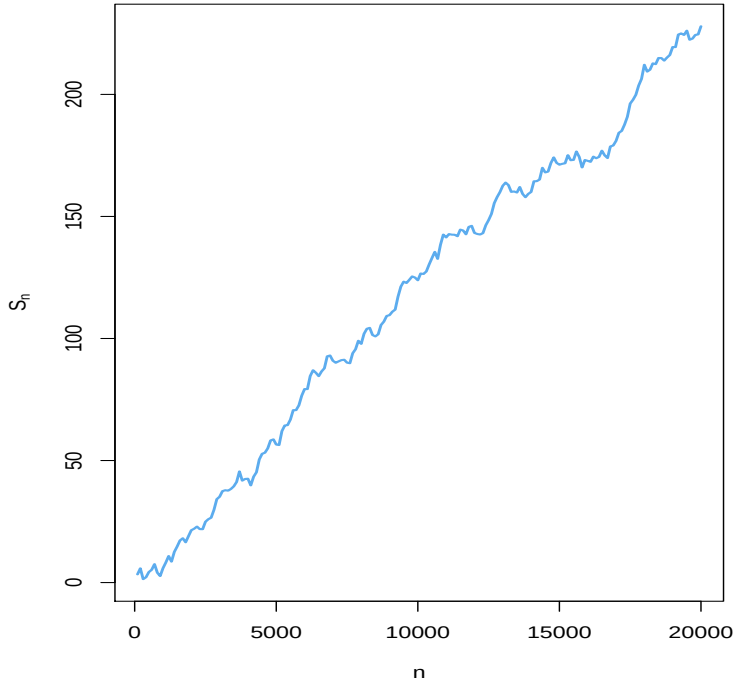
Utilisation et limitations

On rejette l'hypothèse testée si S_n est trop grande pour une loi χ^2_{k-1} ou χ^2_{k-p-1}

[En **R**, pchisq(S)]

La convergence (en n) vers une loi χ^2_{k-1} (ou χ^2_{k-p-1}) n'est établie que pour k et (a_i) fixes. En pratique, on choisit k et (a_i) en fonction des observations, ce qui diminue la validité de l'approximation.





QQ-plot d'un échantillon non-normal et évolution de S_n en fonction de n pour cet échantillon