# A method for avoiding tables for centiles in the case of confidence regions and statistical tests

Ciuiu, Daniel

Technical University of Civil Engineering, Bucharest, Romania, Romanian Institute for Economic Forecasting

6 October 2004

# UNE MODALITÉ D'ÉVITER LES TABLES DES CENTILES DANS LE CAS DES RÉGIONS DE CONFIANCE ET DES TESTS STATISTIQUES

DANIEL CIUIU

RÉSUMÉ. Dans cet article on va determiner des régions de confiance pour les paramètres d'une répartition et des vérsions de quelques tests sans utiliser les tables qui contiennent des centiles.

*Classification AMS 2000 des sujets.* 62F25, 62F03.
*Mots clefs et phrases.* intervalles de confiance, tests statistiques.

## 1. Introduction

Les intervalles de confiance sont déterminés d'habitude en utilisant des tables statistiques qui contiennent des centiles pour quelques répartitions (par exemple les centiles de la répartition normale réduite $N(0,1)$, les centiles de la répartition de Student à $n$ degrés de liberté, $t_n$, les centiles de la répartition de $\chi^2$ à $n$ degrés de liberté ou les centiles de la répartition de Snedecor—Fisher d'ordres $m$ et $n$).

Les tests statistiques, à voir le test de concordance de $\chi^2$ utilisent aussi ces tables avec des centiles. Mais cette chose est dificile à implanter aux ordinateurs, parce qu'on a besoin d'un fichier avec les centiles [1], [2], [3].

Dans tout l'article on note par $\overline{X^j}$ la moyenne de l'échantillon pour $X^j$ et par $\theta$ le vecteur des paramètres qui définissent la répartition de la variable aléatoire $X$. On considére qu'on connaît pour $j = \overline{1,k}$ des formules pour la moyenne du $X^j$ et pour la variance du $X^j$ dépendant de $\theta$. On note ces valeurs par $m_j(\theta)$ et respectivement par $D_j(\theta)$.

Pour déterminer les régions de confiance on utilise l'inégalité de Chébychev et on va résoudre un système d'inéquations.

Pour les tests, on va calculer, s'il est possible, la function de répartition évaluée pour la statistique concernée qui va être comparer avec certains centiles. On utilise le lemme de Hincin qui dit que si la variable aléatoire $X$ a la function de répartition $F(\cdot)$, $F(X)$ suit une loi uniforme sur $[0,1]$.

## 2. Régions de confiance

À présent on a une variable aléatoire $X$ avec la function de répartition $F(x;\theta_1,...,\theta_k)$ dépendant de $k$ paramètres. Nous voulons déterminer une région

de confiance pour $\theta = (\theta_j)_{j=\overline{1,k}}$ avec l'erreur $\varepsilon_1$. On note par $\varepsilon = \frac{\varepsilon_1}{k}$. Si on écrit l'inégalité de Chébychev pour $X^j$ on obtient

$$P\left(\left|\overline{X^j} - m_j(\theta)\right| > \sqrt{\frac{D_j(\theta)}{n\varepsilon}}\right) \leq \varepsilon. \tag{1}$$

En utilisant (1), la probabilité d'exister $j$ de sorte que $1 \leq j \leq k$ et $\left|\overline{X^j} - m_j(\theta)\right| > \sqrt{\frac{D_j(\theta)}{n\varepsilon}}$ est plus petit ou égale à $\varepsilon_1$.

Il résulte qu'on doit résoudre le système d'inéquations

$$\left|\overline{X^j} - m_j(\theta)\right| \leq \sqrt{\frac{D_j(\theta)}{n\varepsilon}},\ j = \overline{1,k}, \tag{2}$$

qui est équivalent à

$$m_j^2(\theta) - 2 \cdot \overline{X^j} \cdot m_j(\theta) - \frac{D_j(\theta)}{n \cdot \varepsilon} + \overline{X^j}^2 \leq 0,\ j = \overline{1,k}. \tag{2'}$$

On considère la variable aléatoire $X$ normale $N\left(m, \sigma^2\right)$ et on a $\theta = \left(m, \sigma^2\right)^T$, $k = 2$. Par des calculs on obtient $m_1(\theta) = m,\ D_1(\theta) = \sigma^2,\ m_2(\theta) = m^2 + \sigma^2$ et $D_2(\theta) = 4 \cdot m^2 \cdot \sigma^2 + 2 \cdot \sigma^4$.

On note par $\alpha = m^2$ et par $\beta = \sigma^2$. L'inéquation du système $(2')$ pour $j = 1$ devient

$$\alpha - \frac{\beta}{n\varepsilon} + \overline{X}^2 \leq 2m\overline{X}. \tag{3}$$

Si $\overline{X} = 0$, la région (3) est

$$\beta \geq n \cdot \varepsilon \cdot \alpha, \tag{3'}$$

qui est dans le système d'axes $\alpha O\beta$ un angle déterminé par $\alpha = 0,\ \beta \geq 0$ et $\beta = n \cdot \varepsilon \cdot \alpha,\ \alpha \geq 0$. Dans le système $mO\beta$ $(3')$ est l'intérieur de la parabole $\beta = n \cdot \varepsilon \cdot m^2$.

Si $\overline{X} \neq 0,$ (3) est

$$\alpha^2 + \frac{1}{n^2\varepsilon^2}\beta^2 - \frac{2}{n\varepsilon}\alpha\beta - 2\overline{X}^2\alpha - \frac{2\overline{X}^2}{n\varepsilon}\beta + \overline{X}^4 \leq 0. \tag{3"}$$

La frontière du (3") est

$$\alpha^2 + \frac{1}{n^2\varepsilon^2}\beta^2 - \frac{2}{n\varepsilon}\alpha\beta - 2\overline{X}^2\alpha - \frac{2\overline{X}^2}{n\varepsilon}\beta + \overline{X}^4 = 0. \tag{4}$$

On calcule les invariants pour (4) et on obtient $I = 1 + \frac{1}{n^2\varepsilon^2},\ \delta = 0$ et $\Delta = -\frac{4\overline{X}^4}{n^2\varepsilon^2}$. Donc (4) est une parabole.

L'axe de la parabole est

$$\beta = n\varepsilon \cdot \alpha - \frac{n\varepsilon\left(n^2\varepsilon^2-1\right)}{n^2\varepsilon^2+1}\overline{X}^2. \tag{5}$$

Le sommet de la parabole est $V\left(\frac{n^4\varepsilon^4}{\left(n^2\varepsilon^2+1\right)^2}\overline{X}^2, \frac{n\varepsilon}{\left(n^2\varepsilon^2+1\right)^2}\overline{X}^2\right)$, et les intersections avec les axes sont $A\left(\overline{X}^2, 0\right)$ et $B\left(0, n\varepsilon\cdot\overline{X}^2\right)$. L'autre axe est

$$\beta = -\frac{1}{n\varepsilon}\alpha + \frac{n\varepsilon}{n^2\varepsilon^2+1}\overline{X}^2. \tag{5'}$$

Parce que $(0,0)$ n'est pas dans (3"), la région de confiance déterminée par la première inéquation est l'intérieur de la parabole.

L'autre inéquation est

$$\alpha^2 + \left(1-\frac{2}{n\varepsilon}\right)\beta^2 + 2\left(1-\frac{2}{n\varepsilon}\right)\alpha\beta - 2\overline{X^2}^2\alpha - 2\overline{X^2}^2\beta + \overline{X^2}^2 \leq 0. \tag{6}$$

La frontière du (6) est

$$\alpha^2 + \left(1-\frac{2}{n\varepsilon}\right)\beta^2 + 2\left(1-\frac{2}{n\varepsilon}\right)\alpha\beta - 2\overline{X^2}^2\alpha - 2\overline{X^2}^2\beta + \overline{X^2}^2 = 0 \tag{7}$$

On calcule les invariants pour (7) et on obtient $I = \frac{2(n\varepsilon-1)}{n\varepsilon}$, $\delta = \frac{2(n\varepsilon-2)}{n^2\varepsilon^2}$ et $\Delta = -\frac{4\overline{X^2}^2}{n^2\varepsilon^2}$. Si $n\varepsilon > 2$ (un fait statistique résonable si on tient compte que $n$ est le volume de l'échantillon) (7) est une ellipse.

Le centre de l'ellipse est $C\left(0, \frac{n\varepsilon}{n\varepsilon-2}\overline{X^2}\right)$ et l'angle de rotation est $\varphi$ de sorte que $\tan(\varphi) = -\frac{1+\sqrt{(n\varepsilon-2)^2+1}}{n\varepsilon-2}$.

Les axes de l'ellipse sont $a$ et $b$ et on a

$$a^2 = \frac{n\varepsilon\left(n\varepsilon-1+\sqrt{n^2\varepsilon^2-4n\varepsilon+5}\right)}{(n\varepsilon-2)^2}\overline{X^2}^2 \text{ et} \tag{8}$$

$$b^2 = \frac{n\varepsilon\left(n\varepsilon-1-\sqrt{n^2\varepsilon^2-4n\varepsilon+5}\right)}{(n\varepsilon-2)^2}\overline{X^2}^2. \tag{8'}$$

Les points d'intersection avec les axes sont $A\left(0, \frac{1-\sqrt{\frac{2}{n\varepsilon}}}{1-\frac{2}{n\varepsilon}}\overline{X^2}\right)$, $B\left(0, \frac{1+\sqrt{\frac{2}{n\varepsilon}}}{1-\frac{2}{n\varepsilon}}\overline{X^2}\right)$ et $C\left(\overline{X^2}, 0\right)$. Donc $O\alpha$ est tangente à l'ellipse.

Quelle que soit $\overline{X}$ ($=0$ ou $\neq 0$) et pour tout $\beta = \sigma^2$ on a une intervalle $[a_\beta, b_\beta]$ avec $a_\beta \geq 0$ de telle sorte que $\alpha = m^2 \in [a_\beta, b_\beta]$. Si $\overline{X} = 0$, on a $m \in \left[\sqrt{a_\beta}, \sqrt{b_\beta}\right] \cup \left[-\sqrt{b_\beta}, -\sqrt{a_\beta}\right]$. Si $\overline{X} \neq 0$, on a $m \in \left[\sqrt{a_\beta}, \sqrt{b_\beta}\right]$ ou $m \in \left[-\sqrt{b_\beta}, -\sqrt{a_\beta}\right]$, dependant si $\overline{X} > 0$, respectivement si $\overline{X} < 0$.

**Exemple 2.1.** *Soit $X$ qui suit une répartition normale $N\left(m, \sigma^2\right)$. On considère un échantillon de volume $n = 1000$ et l'erreur maximale $\varepsilon_1 = 0.2$.*

*Il resulte* $\varepsilon = 0.1$ *et* $n \cdot \varepsilon = 100$. *On a dans cet exemple* $\overline{X^2} = 0.231$.

*La ligne oblique de l'angle est* $\beta = 100\alpha$.

*L'ellipse est* $\alpha^2 + \left(1 - \frac{1}{50}\right)\beta^2 + 2\left(1 - \frac{1}{50}\right)\alpha\beta - 2 \cdot 0.231 \cdot \alpha - 2 \cdot 0.231 \cdot \beta + 0.231^2 = 0$.

*La région de confiance pour* $(\alpha, \beta)$ *se trouve dans la figure suivante.*

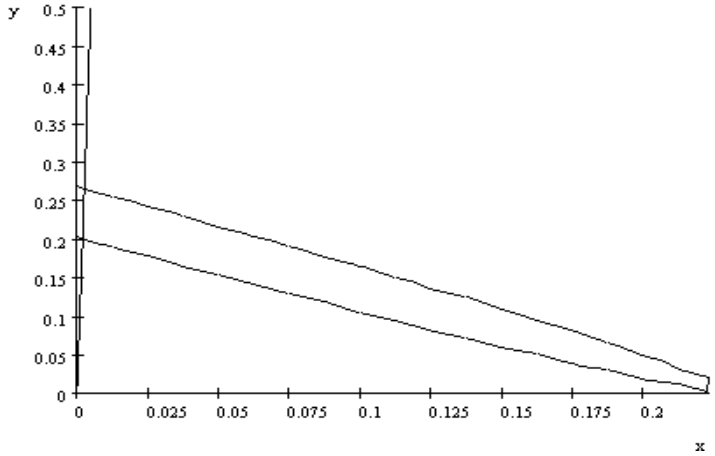


***Fig. 1 :****La région de confiance pour* $(\alpha, \beta)$.

*La région de confiance pour* $\left(m, \sigma^2\right)$ *est déterminée par la parabole* $\beta = 100m^2$ *et* $\left|0.231 - m^2 - \beta\right| = \frac{\sqrt{4 \cdot m^2 \cdot \beta + 2 \cdot \beta^2}}{10}$.

*Elle se trouve dans la figure suivante.*

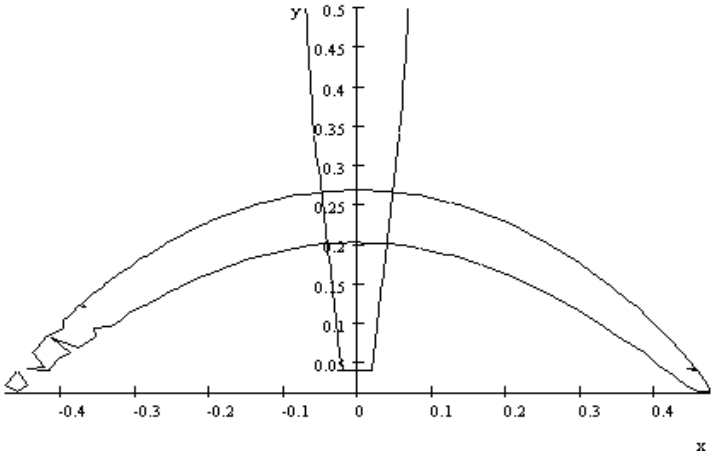


***Fig. 2 :*** *La région de confiance pour* $\left(m, \sigma^2\right)$

**Exemple 2.2.** *Soit* $X$ *qui suit une répartition normale* $N\left(m, \sigma^2\right)$. *On considère un échantillon de volume* $n = 1000$ *et l'erreur maximale* $\varepsilon_1 = 0.2$.

*Il resulte* $\varepsilon = 0.1$ *et* $n \cdot \varepsilon = 100$. *On a dans cet exemple* $\overline{X} = 5.293$ *et* $\overline{X^2} = 28.1$.

*La parabole est*

$\alpha^2 + \frac{1}{100}\beta^2 - \frac{1}{50}\alpha\beta - 2 \cdot 5.293^2 \cdot \alpha - \frac{2 \cdot 5.293^2}{100} \cdot \beta + 5.293^4 = 0$.

*L'ellipse est*

$\alpha^2 + \left(1 - \frac{1}{50}\right)\beta^2 + 2\left(1 - \frac{1}{50}\right)\alpha\beta - 2 \cdot 28.1 \cdot \alpha - 2 \cdot 28.1 \cdot \beta + 28.1^2 = 0$.

*La région de confiance pour $(\alpha, \beta)$ se trouve dans la figure suivante.*

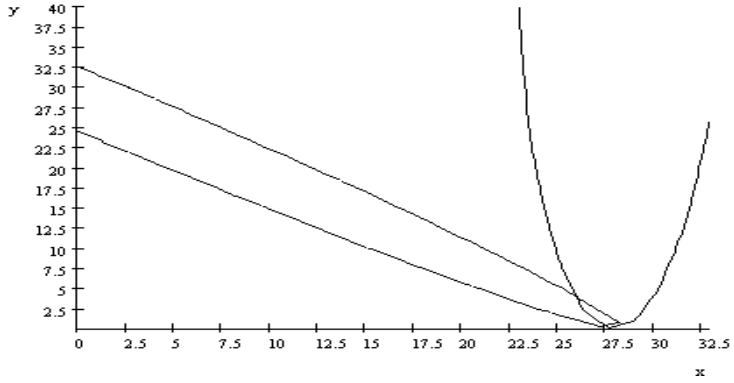


***Fig. 3 :****La région de confiance pour $(\alpha, \beta)$.*

*La région de confiance pour $\left(m, \sigma^2\right)$ est déterminée par $|m - 5.293| = \frac{\sqrt{m^2+\beta}}{10}$ et $\left|\beta + m^2 - 28.1\right| = \frac{\sqrt{4 \cdot m^2 \cdot \beta + 2 \cdot \beta^2}}{10}$. Elle se trouve dans la figure suivante.*

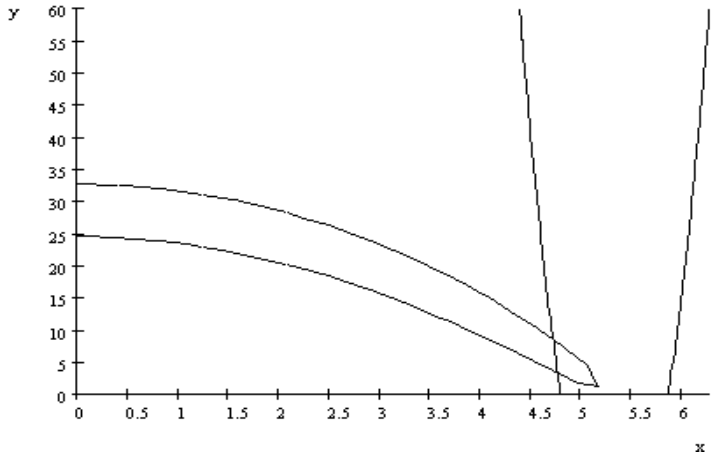


***Fig. 4 :*** *La région de confiance pour $\left(m, \sigma^2\right)$*

## 3. **Versions sans centiles pour quelques tests statistiques**

Le test $T$ bilatéral vérifie l'hypothèse nulle $H_0 : m = m_0$ contre l'hypothèse altérnative $H_1 : m \neq m_0$ avec l'erreur de premier ordre $\varepsilon$. On a un échantillon de volume $n$ de la variable $X$. On calcule $T = \frac{\overline{X} - m_0}{S}\sqrt{n-1}$, où $\overline{X}$ est la moyenne de l'échantillon et $S^2$ est la variance de l'échantillon [1], [3]. On accepte $H_0$ si $|T| < T_{n-1}\left(1 - \frac{\varepsilon}{2}\right)$.

Le test $T$ unilatéral gauche vérifie l'hypothèse nulle $H_0 : m = m_0$ contre l'hypothèse altérnative $H_1 : m < m_0$ avec l'erreur de premier ordre $\varepsilon$. On accepte $H_0$ si $T > T_{n-1}(\varepsilon)$.

Le test $T$ unilatéral droit vérifie l'hypothèse nulle $H_0 : m = m_0$ contre l'hypothèse altérnative $H_1 : m > m_0$ avec l'erreur de premier ordre $\varepsilon$. On accepte $H_0$ si $T < T_{n-1}(1 - \varepsilon)$.

Mais la dénsité de répartition de Student à $n$ degrés de libérté est

$$f(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)\sqrt{n}}\cdot\frac{1}{\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}}. \tag{9}$$

On calcule la function de répartition et on obtient

$$F(x)=\frac{2^n\cdot\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)}\cdot\int_0^{\frac{x+\sqrt{x^2+n}}{\sqrt{n}}}\frac{v^{n-1}}{(v^2+1)^n}dv. \tag{10}$$

Dans (10) on remarque la possibilité de calculer $F(x)$ pour chaque $x$.

Dans le test bilatéral on accepte $H_0$ si $F(T)\in\left(\frac{\varepsilon}{2},1-\frac{\varepsilon}{2}\right)$. Dans le test unilatéral gauche on accepte $H_0$ si $F(T)>\varepsilon$. Dans le test unilatéral droit on accepte $H_0$ si $F(T)<1-\varepsilon$.

Le test de $\chi^2$ bilatéral vérifie l'hypothèse nulle $H_0:\sigma^2=\sigma_0^2$ contre l'hypothèse altérnative $H_1:\sigma^2\neq\sigma_0^2$. On calcule $X^2=\frac{(n-1)\cdot S'^2}{\sigma_0^2}$. On accepte $H_0$ si $X^2\in\left(\chi^2_{n-1}\left(\frac{\varepsilon}{2}\right),\chi^2_{n-1}\left(1-\frac{\varepsilon}{2}\right)\right)$. Le test de $\chi^2$ unilatéral gauche vérifie l'hypothèse nulle $H_0:\sigma^2=\sigma_0^2$ contre l'hypothèse altérnative $H_1:\sigma^2<\sigma_0^2$. On accepte $H_0$ si $X^2>\chi^2_{n-1}(\varepsilon)$ au risque d'erreur de premier ordre $\varepsilon$. Le test de $\chi^2$ unilatéral droit vérifie l'hypothèse nulle $H_0:\sigma^2=\sigma_0^2$ contre l'hypothèse altérnative $H_1:\sigma^2>\sigma_0^2$. On calcule $X^2=\frac{(n-1)\cdot S'^2}{\sigma_0^2}$. On accepte $H_0$ si $X^2<\chi^2_{n-1}(1-\varepsilon)$.

Mais la répartition de $\chi^2_{2n}$ coïncide avec la répartition Erlang $E_{n,\frac{1}{2}}$. La function répartition Erlang $E_{n,\lambda}$ d'ordre $n$ et paramètre $\lambda$ est

$$F_{n,\lambda}(x)=1-e^{-\lambda x}\cdot\sum_{j=0}^{n-1}\frac{\lambda^j\cdot x^j}{j!}. \tag{11}$$

Donc si $n$ n'est pas pair ($n-1$ est pair) on peut calculer la function de répartition du $X^2$. Dans le test de $\chi^2$ bilatéral on accepte $H_0$ si $F_{\frac{n-1}{2},\frac{1}{2}}\left(X^2\right)\in\left(\frac{\varepsilon}{2},1-\frac{\varepsilon}{2}\right)$. Dans le test de $\chi^2$ unilatéral gauche on accepte $H_0$ si $F_{\frac{n-1}{2},\frac{1}{2}}\left(X^2\right)>\varepsilon$. Dans le test de $\chi^2$ unilatéral droit on accepte $H_0$ si $F_{\frac{n-1}{2},\frac{1}{2}}\left(X^2\right)<1-\varepsilon$.

Pour le test de Tukey pour l'égalité des moyennes on a $k$ échantillons des volumes $n$ $(X_{ij})_{1\leq i\leq k,1\leq j\leq n}$ concernant $k$ variables aléatoires normales $N\left(\mu_i,\sigma^2\right)$. On vérifie l'hypothèse nulle $H_0:\mu_i=\mu_j$ pour chaques $i,j$ contre l'hypothèse altérnative $H_1$ : il existent $i,j$ de telle sort que $\mu_i\neq\mu_j$ au risque d'erreur de premier ordre $\varepsilon$. On note par $\overline{X}_i$ la moyenne de l'échantillon $i$, par $\overline{X}_{\min}$ le minimum des $\overline{X}_i$ et par $\overline{X}_{\max}$ le maximum des $\overline{X}_i$. On calcule $S'^2$ un estimateur pour $\sigma^2$ qui suit la répartition de $\chi^2_r$ et $q=\frac{\overline{X}_{\max}-\overline{X}_{\min}}{S'\sqrt{\frac{2}{n}}}$. On accepte $H_0$ si $q\leq t_r\left(\frac{\varepsilon}{2}\right)$.

Mais comment on peut voir dans les tests de Student pour la moyenne si on ne connaît pas la variance, on peut calculer la function de répartition pour $q$, $F(q)$. On accepte $H_0$ si $F(q)\leq\frac{\varepsilon}{2}$.

Pour le test de Hartley pour l'égalité des variances on a $k$ échantillons des volumes $n+1$ $(X_{ij})_{1\leq i\leq k, 1\leq j\leq n+1}$ sur $k$ variables aléatoires normales $N\left(\mu_i, \sigma_i^2\right)$. On vérifie l'hypothèse nulle $H_0 : \sigma_i^2 = \sigma_j^2$ pour chaques $i, j$ contre l'hypothèse altérnative $H_1$ : il existent $i, j$ de sort que $\sigma_i^2 \neq \sigma_j^2$ au risque d'erreur de premier ordre $\varepsilon$.

On note par $\overline{X}_i = \frac{1}{n+1} \cdot \sum\limits_{j=1}^{n+1} X_{ij}$ et par $S_i^2 = \frac{1}{n} \cdot \sum\limits_{j=1}^{n+1} \left(X_{ij} - \overline{X}_i\right)^2$. On note aussi par $S_{\min}^2$ le minimum des $S_i$ et par $S_{\max}^2$ le maximum. On calcule $F_{\max} = \frac{S_{\max}^2}{S_{\min}^2}$. On accepte $H_0$ si $F_{\max} \leq F_{n,n}\left(1 - \frac{\varepsilon}{2}\right)$, où $F_{n,n}\left(1 - \frac{\varepsilon}{2}\right)$ est la centile d'ordre $1 - \frac{\varepsilon}{2}$ pour la repartition de Snedecor—Fisher d'ordres $(n, n)$.

Mais la densité de cette répartition est

$$g_{n,n}(x) = \frac{\Gamma(n)}{\Gamma^2\left(\frac{n}{2}\right)} \cdot \frac{x^{\frac{n}{2}-1}}{(1+x)^n}. \tag{12}$$

On peut calculer la function de répartition et on obtient

$$G_{n,n}(x) = \frac{2 \cdot \Gamma(n)}{\Gamma^2\left(\frac{n}{2}\right)} \cdot \frac{x^{\frac{n}{2}-1}}{(1+x)^n}. \tag{12'}$$

Le test de concordance de $\chi^2$ vérifie l'hypothese nulle $H_0$ : la variable aléatoire $X$ a la function de répartition $F(x; \theta_1, ..., \theta_k)$, où $\theta_1, ..., \theta_k$ sont les paramètres de la répartition, contre l'hypothèse altérnative $H_1$ : la variable aléatoire $X$ n'a pas la function de répartition $F(x; \theta_1, ..., \theta_k)$ avec l'erreur de premier ordre $\varepsilon$. On a un échantillon de volume $n$ de $X$ et $r$ intervalles, $r > k$. On note par $n_i$ le numéro des valeurs de l'échantillon dans l'intervalle $I_i$ et par $n_i' = n \cdot F\left(I_i; \widehat{\theta}_1, ..., \widehat{\theta}_k\right)$, où $\widehat{\theta}_1, ..., \widehat{\theta}_k$ sont des éstimations pour $\theta_1, ..., \theta_k$. On calcule $X^2 = \sum\limits_{j=1}^{r} \frac{\left(n_i - n_i'\right)^2}{n_i'}$ et on accepte $H_0$ si $X^2 < \chi_{r-k-1}^2(1-\varepsilon)$, où $\chi_{r-k-1}^2(1-\varepsilon)$ est la centile de $\chi^2$ à $r-k-1$ degrés de libérté.

Si $r-k-1$ est pair, on peut calculer la function de répartition de $\chi_{r-k-1}^2$ en utilisant (11). Donc on accepte $H_0$ si $F_{\frac{r-k-1}{2}, \frac{1}{2}}\left(X^2\right) < 1 - \varepsilon$.

## 4. **Conclusions**

Pour les régions de confiance l'erreur $\varepsilon$ est maximale (elle peut être plus petite que $\varepsilon$). Dans le cas de $k$ paramètres on doit résoudre un système de $k$ inéquations.

Pour les tests on calcule la function de répartition (s'il est possible) pour la statistique et on compare cette valeur avec l'ordre de centile.

Ces considérations peuvent aider le programmateur à fair des programmes pour résoudre des problems de statistique. On doit choisir si on utilise l'inégalité de Chébychev ou les tables des centiles (si le système est trop difficile à résoudre, on utilise les centiles).

## Références

[1] G. Ciucu, V. Craiu, *Inférence statistique*. Editure Didactique et Pedagogique, Bucarest, 1974. (en roumain)

[2] V. Craiu, V. Preda, *Des tests pour vérifiquer la normalité*. Editure de l'Université de Bucarest, 1981. (en roumain)

[3] A. Popescu, V. Petrehuş, *Probabilitées et statistique*. Editure de l'Université Technique des Constructions, Bucarest, 1997. (en roumain)

(Daniel Ciuiu) Departement des Mathematiques,
Université Technique des Constructions, Bucarest,
Bd. Lacul Tei no. 124, Bucarest, Roumanie
*E-mail address*: dciuiu@yahoo.com