Rapport scientifique No 360

par

Nathalie Huitorel
Luc Perreault
Bernard Bobée

Tests de détection de données singulières pour quelques lois du logiciel AJUSTE-2

Septembre 1992

INRS-Eau
Université du Québec
C.P. 7500
Sainte-Foy, Québec
G1V 4C7

SOMMAIRE

LISTE DES TABLEAUX ET FIGURES

I INTRODUCTION

La planification et le dimensionnement des ouvrages hydrauliques, la gestion et l'opération rationnelle des réservoirs ainsi que la prévision des inondations reposent sur une bonne connaissance des débits de crue et en particulier des débits X_T associés à une période de retour T. L'estimation de ces débits est donc très importante et doit être la plus précise possible afin d'éviter :

- une sur-estimation des crues qui entraine un surdimensionnement des ouvrages hydrauliques et conduit à des coûts de construction supplémentaires (un gain de précision de quelques pourcentages peut conduire à une économie de plusieurs millions de dollars).

- une sous-estimation des crues qui peut causer des inondations et se traduire par des dégâts matériels importants et des pertes de vies humaines.

L'estimation des débits de crue correspondant à une probabilité au dépassement p donnée est donc essentielle. Il est alors important d'un point de vue économique de déterminer le plus précisément possible le débit X_T de période de retour T tel que :

$$p = prob(X > X_T) = \frac{1}{T}$$

où p est également le risque hydrologique associé au dépassement de X_T. L'ajustement de lois statistiques aux séries de débits maximums annuels est un outil privilégié pour déterminer le débit X_T et les intervalles de confiance qui y sont associés.

Le logiciel HFA, créé à l'INRS permet d'estimer les débits maximums X_T de période de retour T, en consirant différentes lois statistiques. Ce logiciel est actuellement utilisé de manière systématique par Hydro-Québec pour :

- le dimensionnement des ouvrages hydrauliques
- la vérification du dimensionnement des ouvrages actuels
- la gestion de ses réservoirs :
  * à court terme (prévision des débits de pointe journalière)
  * à moyen et long terme (évaluation de la distribution des volumes)
- l'évaluation des zones inondables et des risques d'inondations

Pour Hydro-Québec, il est donc important de connaître le plus précisément possible les débits de période de retour T.

C'est pourquoi le logiciel Ajuste, version française de HFA, fait actuellement l'objet d'un projet de partenariat entre Hydro-Québec et l'INRS-Eau afin de développer une nouvelle version, Ajuste-2, plus complète. Ajuste-2 permettra d'estimer automatiquement les débits de manière fiable, rigoureuse et de répondre aux besoins des hydrologues.

L'ajustement d'une distribution statistique à une série d'observations permettant d'estimer X_T repose toutefois sur différentes hypothèses de base (indépendance, homogénéité)que l'on doit vérifier. De plus, un bon ajustement suppose que l'échantillon ne contiennent pas de données singulières.

En effet, il arrive fréquemment qu'un échantillon de débits maximums annuels contienne une valeur extrême qui influence grandement les estimations. On doit donc se demander si cette valeur est réelle ou aberrante (résultant, par exemple, d'une erreur de mesure). La prise en compte de valeurs aberrantes, lors de l'ajustement d'une loi, peut conduire à des estimations inadéquates. Il est donc important, avant d'effectuer un ajustement de loi à un échantillon d'examiner si celui-ci contient des valeurs singulières. On doit ensuite vérifier, à partir de considérations hydrologiques, si cette donnée est réelle ou aberrante afin de l'éliminer éventuellement.

Le présent rapport fait une revue des tests permettant détecter la présence de données singulières afin de retenir ceux qui sont les plus adaptés aux données hydrologiques caractérisées par de faibles tailles d'échantillon et par la qualité réduite des mesures, et aux lois du logiciel. Ces test seront ensuite introduits dans Ajuste-2.

Dans le chapitre II, nous décrivons tout d'abord les objectifs fixés. Le chapitre suivant traite des méthodes utilisées pour atteindre ces objectifs, nous y décrivons le type de tests statistiques considérés. Au chapitre IV, nous présentons les résultats, c'est-à-dire les tests retenus pour la plupart des lois de Ajuste-2, et certains autres tests qui n'ont pas été retenus mais qu'il nous a paru intéressant de présenter.

II OBJECTIFS

Le but de ce travail est de trouver, si possible, des tests de détection de données singulières pour chaque loi de Ajuste-2.

Les différentes lois introduites dans Ajuste-2 sont :

- Gamma
- Log-Gamma
- Pearson type 3
- Log-Pearson type 3
- Gamma généralisée
- Normale
- Log-Normale
- GEV
- Gumbel
- Weibull
- Fréchet
- Exponentielle

Les trois situations considérées ici sont :

- Une donnée singulière supérieure
- Une donnée singulière inférieure
- Un ensemble de données singulières inférieures ou supérieures

Plus précisément, les objectifs à atteindre peuvent se présenter de la façon suivante :

1) Trouver, si possible, un test pour chaque loi pour :
   - Une données singulière supérieure
   - Une donnée singulière inférieure
2) Trouver, si possible, un test pour chaque loi pour plus d'une donnée singulière
3) Trouver, si possible, un test global, c'est-à-dire valable pour un ensemble de lois différentes

III TESTS DE DISCORDANCE

Pour identifier les tests adaptés au problème posé, nous avons été amenés à consulter des livres ou articles concernant les données singulières (les références de ces publications sont données dans la bibliographie au chapitre VII). Nous avons ainsi pu étudier de nombreux tests de détection de données singulières, pour, ensuite, ne retenir que ceux qui semblaient les plus adaptés au contexte de Ajuste-2. Ceux-ci sont présentés dans le chapitre suivant.

Ces tests sont des tests *de discordance*, dont le principe général est le suivant :

Soit un échantillon chronologique x_1, x_2,..., x_n provenant d'une loi F, et l'échantillon $x_{(1)} < x_{(2)} < ... < x_{(n)}$ correspondant rangé par ordre croissant ayant pour valeurs extrêmes $x_{(1)}$ et $x_{(n)}$.

Un test de discordance, est défini par une statistique qui vérifie, par exemple, si $x_{(n)}$ n'est pas seulement une valeur extrême, mais aussi une observation incompatible avec le modèle probabiliste F choisi pour un seuil de signification α donné. Si c'est le cas nous pouvons dire que $x_{(n)}$est une *donnée singulière supérieure* au seuil de signification du test α par rapport au modèle considéré.

Cette procédure ne se limite pas aux données singulières supérieures, nous pouvons également considérer la discordance d'une donnée singulière inférieure $x_{(1)}$, d'un couple de données singulières $(x_{(1)},x_{(n)})$, etc...

Un test de discordance permet donc d'examiner si une ou plusieurs données douteuses, inférieures ou supérieures, doivent être considérées comme faisant partie de la population ou non. Les hypothèses à tester sont les suivantes :

Hypothèse nulle (H_0) :

Les données x_j ($j = 1,...,n$) proviennent d'une même distribution F.

Hypothèse alternative (H_1) :

Un petit nombre k de données parmi les n données observées proviennent d'une version différente de F, dans laquelle un paramètre a été modifié.

IV RESULTATS

Ce chapître n'a pas pour but de montrer les propriétés mathématiques et statistiques des lois ou des différents tests proposés. Il se contente de décrire brièvement les tests adaptés à chaque loi.

Pour les aspects plus théoriques, on se référera aux livres et articles répertoriés dans la bibliographie.

4-1 Explications préalables :

Soit un échantillon $x_1, x_2, ..., x_n$ provenant d'une loi F et l'échantillon ordonné correspondant $x_{(1)} < x_{(2)} < ... < x_{(n)}$. Pour tester l'hypothèse H_0, où toutes les observations proviennent d'une même distribution F, contre l'hypothèse H_1, où un petit nombre k de données proviennent d'une version différente de F, à un seuil de signification α, nous considérons la statistique de discordance T avec comme règle de décision:

Rejeter H_0 si $T \geq c(\alpha)$

où $c(\alpha)$ est une constante telle que :

$\text{Prob}\{T \geq C(\alpha) / H_0 \text{ vraie}\} = \alpha$

et où α est le risque de première espèce ou erreur de type 1

La *probabilité au dépassement* relative à une valeur observée t d'une statistique T est notée P(t) et est définie comme suit :

$$P(t) = Prob(T \geq t)$$

Cette probabilité est utilisée comme règle de décision tout au long du présent rapport. En effet, nous remarquons qu'une grande valeur de P(t) favorise l'hypothèse H_0 alors qu'une petite valeur supporte l'hypothèse H_1. On peut ainsi comparer P(t) au seuil de signification α fixé a priori :

Si $P(t) \geq \alpha$ alors nous adoptons l'hypothèse H_0

Si $P(t) < \alpha$ alors nous adoptons l'hypothèse H_1

Dans ce qui suit, nous présentons brièvement les test de discordance pour différentes lois du logiciel Ajuste-2.

4-2 **Loi Normale** de moyenne μ et d'écart-type σ : $N(\mu, \sigma^2)$

4.2.1 Fonction de densité de probabilité :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

où $-\infty \leq x \leq +\infty$, $\mu \in R$ et $\sigma > 0$.

4.2.2 Test pour une donnée singulière inférieure ou supérieure

- Hypothèses :

$H_0 : x_{(i)} \in N(\mu, \sigma^2)$, $\forall i \in \{1, \ldots, n\}$

$H_1 : x_{(1)}$ ou $x_{(n)} \in N(\mu + a, \sigma^2)$ avec $a \neq 0$

- Statistique : $t = \max\left(\frac{x_{(n)} - \overline{x}}{S}, \frac{\overline{x} - x_{(1)}}{S} \right)$

où $\overline{x}$ est une estimation non biaisée de μ et S est une estimation de σ

- Probabilité au dépassement :

$$P(t) \leq 2n\,prob\left\{ T_{n-2} > \left(\frac{n(n-2)t^2}{(n-1)^2 - nt^2} \right)^{\frac{1}{2}} \right\}$$

Nous avons égalité pour $t > \sqrt{\frac{(n-1)(n-2)}{2n}}$

T_{n-2} est une variable aléatoire distribuée selon une loi de Student à (n-2) degrés de liberté.

- Référence : Barnett et Lewis (1984), Ferguson (1961)

- Propriétés du test :

Ce test maximise la probabilité d'identifier une observation singulière. Il a l'avantage d'utiliser toutes les observations contenues dans l'échantillon, par l'intermédiaire des calculs de la moyenne et de l'écart-type, et de tenir compte de l'effectif de l'échantillon.

4.2.3 Test pour k données singulières supérieures

- Hypothèses :

$H_0 : x_{(i)} \in N(\mu, \sigma^2),\ \forall i \in \{1, \ldots, n\}$

$H_1 : x_{(n-k+1)}, \ldots, x_n \in N(\mu + \alpha, \sigma^2)$ avec $\alpha \neq 0$

- Statistique :

$$t = \frac{\sum_{i=n-k+1}^{n} x_i - k\bar{x}}{S}$$

- Probabilité au dépassement :

$$P(t) \leq C_n^k prob\{T_{n-2} > \left(\frac{n(n-2)t^2}{k(n-k)(n-1)-nt^2}\right)^{\frac{1}{2}}\}$$

nous avons égalité pour $t \geq \left\{\frac{k^2(n-1)(n-k-1)}{nk+n}\right\}^{\frac{1}{2}}$

- Référence : Barnett et Lewis (1984)

- Propriétés du test :

Ce test maximise la probabilité d'identifier ces k observations comme des valeurs singulières.

4-3 Loi log-Normale :

4.3.1 Fonction de densité de probabilité :

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\}$$

où $x > 0$

, $\mu \in R$ et $\sigma > 0$.

4.3.2 Principe :

Si la variable X suit une loi log-Normale, alors ln X suit une loi Normale. Ainsi, pour appliquer un test de discordance pour une ou plusieurs données singulières d'un échantillon provenant d'une loi log-Normale, il suffit de prendre le logarithme des observations et d'appliquer un test approprié à la loi Normale sur l'échantillon transformé.

- Référence : Barnett et Lewis (1984)

4.4 loi Gamma

Deux types de test sont proposés ici : le premier suppose la connaissance du paramètre de forme λ de la loi Gamma et le second, indépendant de ce paramètre, transforme les données pour pouvoir utiliser par la suite un test adapté à la loi Normale.

- Référence des quatre premiers tests : Barnett et Lewis (1984)

- Propriétés de ces quatre tests :

Ce sont des tests du rapport des vraisemblances maximales qui nécessitent la connaissance a priori du paramètre de forme λ.

4.4.1 Fonction de densité de probabilité :

$$f(x) = \frac{\alpha^{\lambda}}{\Gamma(\lambda)} x^{\lambda-1} \exp(-x\alpha)$$

avec $0 < x < \infty$, $\lambda > 0$

et où $$\Gamma(\lambda) = \int_0^{+\infty} e^{-x} x^{\lambda-1} dx$$

4.4.2 Test pour une donnée singulière supérieure :

- Statistique : $t = \frac{x_{(n)}}{\Sigma x_{(i)}}$

avec i=1,...,n

- Probabilité au dépassement :

$$P(t) \leq n Prob\left\{ F_{2\lambda,2(n-1)\lambda} > \frac{(n-1)t}{1-t} \right\}$$

Nous avons égalité pour $t \geq 1/2$

4.4.3 Test pour une donnée singulière inférieure :

- Statistique : $t = \frac{x_{(1)}}{\Sigma x_{(i)}}$

avec i=1,...,n

- Probabilité au dépassement :

$$P(t) \leq n Prob\left\{ F_{2\lambda,2(n-1)\lambda} < \frac{(n-1)t}{1-t} \right\}$$

4.4.4 Test pour k données singulières supérieures :

- Statistique : $t = \frac{\Sigma x_{(j)}}{\Sigma x_{(i)}}$

avec $j = n - k + 1, \ldots, n$ et $i = 1, \ldots, n$

- Probabilité au dépassement :

$$P(t) \leq C_n^k Prob\left\{ F_{2k\lambda, 2(n-k)\lambda} > \frac{(n-k)t}{k(1-t)} \right\}$$

4.4.5 Test pour k données singulières inférieures :

- Statistique : $t = \frac{\Sigma x_{(j)}}{\Sigma x_{(i)}}$

avec $j = 1, \ldots, k$ et $i = 1, \ldots, n$

- Probabilité au dépassement :

$$P(t) < C_n^k Prob\left\{ F_{2k\lambda, 2(n-k)\lambda} < \frac{(n-k)t}{k(1-t)} \right\}$$

4.4.6 Méthode pour une ou plusieurs données singulières inférieures ou supérieures (paramètres inconnus) :

- Propriété : Ce test est indépendant du paramètre de forme λ.

- Principe :

Pour tester la discordance d'une ou plusieurs données singulières d'un échantillon provenant d'une loi Gamma lorsque les deux paramètres sont inconnus, on peut transformer les valeurs x_1, x_2,..., x_n de l'échantillon Gamma en $y_1 = x_1^{1/3}, y_2 = x_2^{1/3}, \ldots, y_n = x_n^{1/3}$ et appliquer aux valeurs ainsi transformées un test de discordance pour un échantillon provenant d'une loi Normale (section 5.2).

En effet, si la variable X suit une loi Gamma, $X^{1/3}$ est distribuée approximativement selon une loi Normale (Kimber 1979) :

- de moyenne : $(\lambda/\alpha)^{\frac{1}{3}}(1 - 1/9\lambda)$
- de variance : $((1/\alpha^2)/\lambda)^{\frac{1}{3}}/9$

Kimber propose, entre autre, d'employer le test utilisant la statistique : (section 4.4.2)

$$t = \max\left(\frac{x_{(n)} - \bar{x}}{s}, \frac{\bar{x} - x_{(1)}}{s}\right)$$

Ce test se comporte bien sauf pour des petites tailles d'échantillon ($n \leq 10$) où il n'est pas satisfaisant lorsqu'il y a seulement des données singulières inférieures.

- Références : Barnett et Lewis (1984), Kimber (1979)

4-5 Loi log-Gamma

- Principe :

Si ln X suit une loi Gamma, alors X suit une loi log-Gamma. Pour tester la discordance d'une ou plusieurs données singulières d'un échantillon provenant d'une loi log-Gamma, nous pouvons donc transformer les valeurs $x_1, \ldots, x_n$ en $y_1 = (\ln x_1)^{1/3}, \ldots, y_n = (\ln x_n)^{1/3}$, et appliquer, sur la variable transformée Y, un test de détection de données singulières pour un échantillon provenant d'une loi Normale (section 5.2).

4-6 Loi Exponentielle

4.6.1 Fonction de densité de probabilité :

$$f(x) = \alpha e^{-\alpha x}$$

La loi Exponentielle est un cas particulier de la loi Gamma, définie en 5.4, où $\lambda = 1$.

4.6.2 Test pour une ou plusieurs données singulières supérieures

- Principe :

On identifie un nombre k de données dont on veut examiner si elles sont conjointement singulières. On applique ensuite le test pour j=k, c'est-à-dire pour examiner la discordance du sous-échantillon $x_n > x_{n-1} > \ldots > x_{n-k+1}$. Si le test est significatif on conclut que les k données extrêmes supérieures sont singulières. Et s'il n'est pas significatif, on applique alors le test pour j=k-1, c'est-à-dire pour examiner la discordance du sous-échantillon $x_n > x_{n-1} > \ldots > x_{n-k+2}$. On continue ainsi la procédure jusqu'à ce que les k données extrêmes soient testées ou qu'un résultat significatif soit obtenu.

- Hypothèses :

H_0 : $x_{(i)} \in E(\alpha)$, $\forall i \in \{1, \ldots, n\}$

H_1 : $x_{(n-k+1)}, \ldots, x_{(n)} \in E(\alpha a)$ avec $a \neq 1$

- Statistique : $$t_j = \frac{x_{n-j+1}}{\sum_{i=j}^{n-j+1} x_i}$$

avec j=1,..., k

- Probabilité au dépassement :

$$P(t_j) = C_n^j \left\{ \frac{1-t_j}{1+jt_j-t_j} \right\}^{n-j}$$

pour $1/2 < t_j < 1$

$$C_n^j \frac{(1-t_j)^{n-j} - j(n-j)(1-2t_j)^{n-j}/(j+1)}{(1+jt_j-t_j)^{n-j}} \leq P(t_j) < C_n^j \left\{ \frac{1-t_j}{1+jt_j-t_j} \right\}^{n-j}$$

pour $(n-j+1)^{-1} < t_j < 1/2$

- Références : Kimber (1982), Barnett et Lewis (1984), Chikkagoudar et Kunchur (1987)

- Propriétés du test :

Kimber (1982) démontre la souplesse de cette procédure par rapport aux autres procédures existantes. Ce test a également l'avantage de ne pas être perturbé par la présence de données singulières supplémentaires si on prend la précaution de choisir k suffisamment grand.

4.6.3 Test pour une ou plusieurs données singulières inférieures

- Principe :

On peut appliquer le test précédent sur des données extrêmes inférieures pour détecter des valeurs singulières inférieures.

- Hypothèses :

$H_0 : x_{(i)} \in E(\alpha)$, $\forall i \in \{1, \ldots, n\}$

$H_1 : x_{(1)}, \ldots, x_{(k)} \in E(\alpha\alpha)$ avec $\alpha \neq 1$

- Statistique : $$t_j' = \frac{x_{j+1}}{\sum_{i=1}^{j+1} x_i}$$

avec j = 1,..., k

Nous avons donc $t_j' = t_{n-j}$ (voir 4.6.2)

- Probabilité au dépassement :

$$P(t_j') = C_n^{n-j} \left\{ \frac{1 - t_j'}{1 + (n-j)t_j' - t_j'} \right\}^j$$

pour $1/2 < t_j' < 1$

$$C_n^{n-j} \left\{ \frac{(1-t_j')^j - (n-j)j(1-2t_j')^j/(n-j+1)}{(1+(n-j)t_j' - t_j')^j} \right\} \leq P(t_j') < C_n^{n-j} \left\{ \frac{1-t_j'}{1+(n-j)t_j' - t_j'} \right\}^j$$

pour $(j+1)^{-1} < t_j' < \frac{1}{2}$

- Référence : Kimber (1982)

- Propriétés du test :

Il a les mêmes propriétés que le test précédent. Il a également l'avantage de n'être que très peu influencé par le comportement des valeurs extrêmes supérieures.

4-7 **Loi Weibull**

4.7.1 Fonction de densité de probabilité :

$$f(x) = \frac{c}{\alpha}\left(\frac{x}{\alpha}\right)^{c-1} \exp\left\{-\left(\frac{x}{\alpha}\right)^{c}\right\}$$

avec $x > 0, \alpha > 0$ et $c > 0$

4.7.2 Tests pour une ou plusieurs données singulières inférieures ou supérieures

- Principe :

Si X suit une loi Weibull de paramètres α et c, la variable $Y = X^c$ suit une loi exponentielle de paramètre $\alpha_0 = \alpha^c$. Si nous connaissons la valeur du paramètre c nous pouvons donc tester la discordance d'une ou plusieurs données singulières en transformant les valeurs x_i en $y_i = (x_i)^c$ et en appliquant ensuite, sur la variable transformée, un test adapté à la loi exponentielle.

- Référence : Barnett et Lewis (1984)

4-8 **Loi Gumbel**

4.8.1 Fonction de densité de probabilité :

$$f(x) = \frac{1}{b}\exp\left\{-\frac{(x-a)}{b} - \exp\left(-\frac{(x-a)}{b}\right)\right\}$$

où $-\infty < x < +\infty, a \in R$ et $b > 0$

4.8.2 Tests pour une ou plusieurs données singulières inférieures ou supérieures :

- Principe :

Si X suit une loi Gumbel de paramètres a et b, alors la variable $Y = \exp(-X/b)$ suit une loi Exponentielle de paramètre $\alpha = \exp(a/b)$.

Ainsi, pour appliquer un test de discordance sur un échantillon provenant d'une loi Gumbel, il suffit de transformer les valeurs x_i en $y_i = \exp(-x_i/b)$, dans la mesure où b est connu, et d'appliquer ensuite aux y_i un test de discordance adapté aux échantillons provenant d'une loi Exponentielle.

- Référence : Barnett et Lewis (1984)

4-9 Tests globaux

Le troisième objectif était de trouver un test global, c'est-à-dire valable pour toutes les lois et pour les différentes hypothèses alternatives H_1 considérées dans le présent rapport.

Deux approches sont envisageables : - L'approche non paramétrique
- L'approche de Carletti

4.9.1 Tests non paramétriques :

les tests non paramétriques sont applicables pour un vaste ensemble de lois . En effet, un test non paramétrique a comme principale propriété d'être défini par une statistique dont la loi est indépendante de celle d'où proviennent les observations.

Certains auteurs ont proposé des tests non paramétriques de détection de données singulières. Citons en particulier Walsh (1953, 1959, 1965), Mosteller (1948) , Walsh et Kelleher (1974) et Tiku (1975). Toutefois, ces tests sont difficilement applicables au type de données qui nous intéresse. En effet, certains de ces tests supposent la symétrie de la loi des observations, d'autres nécessitent de grandes tailles d'échantillon. Or, nous savons que ces caractéristiques sont rarement vérifiées en hydrologie, en particulier pour les études sur les débits maximums annuels.

De plus, le principe même des tests non paramétriques ne nous parait pas adapté au problème des données singulières. En effet, c'est en considérant la façon dont sont générées les données que nous avons des bases pour examiner la discordance d'une donnée. Or, une approche non paramétrique ne fait aucune supposition quant à la façon dont sont générées les données. Il semble donc être contradictoire de chercher à étudier les données singulières par une méthode non paramétrique.

Madansky (1982) et Barnett et Lewis (1984) donnent aussi leur préférence aux méthodes paramétriques pour le problème de détection de données singulières.

C'est pour toutes ces raisons que nous avons décidé de ne pas retenir de tests non paramétriques.

4.9.2 Approche de Carletti :

Carletti (1976 et 1985) propose une méthode générale satisfaisante pour un grand nombre de données de nature différente et dont la distribution est peu ou mal connue. Cette approche, qui répond bien à notre troisième objectif de détermination d'une méthode globale, consiste à appliquer un test de normalité sur les données observées. Si ce test est significatif,c'est-à-dire si l'hypothèse d'acceptation de la normalit èst rejeté, on applique une transformation sur les données pour les normaliser et un test de discordance adapté à la loi normale aux nouvelles observations ainsi obtenues. Si le test n'est pas significatif,c'est-à-dire si les données sont normales, on applique le test de discordance aux données originales.

1) Test de normalité :

Le test de normalité utilisé par Carletti est celui proposé par Pearson et Hartley (1966). Ce test d'adéquation est basé sur les coefficients d'asymétrie C_S et d'aplatissement C_K de l'échantillon. Ces coefficients sont définis comme suit :

$$C_S = \frac{m_3}{S^3} \quad \text{et} \quad C_K = \frac{m_4}{S^4}$$

où $$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

et $$S = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

On rejette alors l'hypothèse de normalité des observations lorsque C_S et C_K sont supérieurs, pour un seuil de signification α donné, aux valeurs correspondantes de la loi Normale $C_S = 0$ et $C_K = 3$.

Evidemment, on peut envisager d'autres tests de normalité qui, pour de faibles tailles d'échantillon, peuvent être plus efficaces que le test de Pearson et Hartley (1966).

Mentionnons en particulier les tests de Shapiro-Wilk et de Kolmogorov-Smirnov qui sont utilisés fréquemment (D'Agostino et Stephens, 1986).

2) Transformation des variables :

Si le test précédent est significatif, c'est-à-dire si l'échantillon ne suit pas une loi normale au seuil α fixé, on applique la transformation puissance de Box et Cox (1964) sur cet échantillon. X étant la variable observée et Z la variable transformée, nous avons :

$$Z = \frac{X^{\lambda}-1}{\lambda} \quad \text{si} \quad \lambda \neq 0$$

$$Z = \ln X \quad \text{si} \quad \lambda = 0$$

La valeur λ peut être déterminée par approximations successives en utilisant la méthode de Draper et Cox (1969) qui a pour objectif d'obtenir l'égalité suivante :

$$C_S(X) = \frac{1}{3} C_V(Z)\{C_K(X)-3\}$$

où $C_V(Z)$ est le coefficient de variation de la variable transformée Z
$C_s(X)$ est le coeffient d'asymétrie de la variable X
$C_K(X)$ est le coefficient d'aplatissement de la variable X

Pour obtenir une valeur plus précise de λ, on peut appliquer la méthode du maximum de vraisemblance telle que proposée par Box et Cox (1964).

3) Test de détection de données singulières :

On applique alors un test de discordance adapté à la loi Normale (section 4-2) sur les X si l'échantillon provient d'une loi Normale. Dans le cas contraire, nous appliquons le test sur les Z et si une donnée transformée est déclarée singulière, cela implique que la donnée originale est elle-même singulière (Carletti, 1976). Le test proposé par Carletti (1976) est le test de Grubbs pour une donnée singulière inférieure ou supérieure (section 4-2-2) dont la statistique est :

$$t = \max\left(\frac{x_{(n)} - \bar{x}}{S}, \frac{\bar{x} - x_{(1)}}{S} \right)$$

Si le test est significatif, on recommence le processus en considérant les observations restantes, c'est-à-dire celles non détectées comme données singulières, comme un tout, et ainsi de suite jusqu'au moment où plus aucune valeur singulière n'est signalée. Un tel procédé se justifie car les valeurs singulières détectées peuvent en masquer d'autres par modification des paramètres de la variable et donc du contrôle lui-même.

Nous remarquons que, par cette méthode, nous utilisons trois fois les données dans certains cas : une première fois pour appliquer les test de normalité, une seconde fois pour transformer les données et une troisième fois pour appliquer le test de discordance. Ces trois utilisations des données peuvent affecter le seuil de signification. De plus, la transformation pour obtenir une distribution normale est approximative.

V ETUDE CONCERNANT LES TESTS ADAPTES A LA LOI GAMMA

5.1 Introduction

Nous avons vu au chapitre IV que, pour certaines lois (Gamma, Weibull, Gumbel), les tests de discordance adaptés à celles-ci supposent la connaissance a priori d'un des paramètres de distribution. Or, en pratique, cette information n'est généralement pas disponible.

En particulier pour la loi Gamma (section 4-4), le paramètre de forme λ doit être connu. Pour pallier ce problème, nous avons proposé une approche qui consiste dans un premier temps à normaliser les données pour, ensuite, appliquer aux données transformées un test adapté à la loi Normale. Ce test, comme nous l'avons vu à la section 4-2, est indépendant des paramètres. Toutefois, cette transformation ne permet d'obtenir qu'une distribution approximativement Normale, ce qui peut, dans certains cas, fausser les résultats du test de discordance.

Une solution naturelle, qui est souvent utilisée dans des situations similaires, serait d'estimer le paramètre λ, en supprimant la ou les données douteuses, et, ensuite, appliquer le tests adapté à la loi Gamma en considérant ce paramètre de forme connu. Il parait évident que le seuil du test ainsi appliqué risque d'être influencé par le fait que nous utilisons les observations deux fois de façon non-indépendante.

Nous avons jugé alors pertinent d'effectuer une simulation de Monte-Carlo afin d'étudier le comportement d'un test de discordance appliqué selon cette dernière approche. Plus particulièrement, nous avons examiné l'influence de l'estimation du paramètre sur le seuil de signification.

Nous avons également effectué cette étude pour l'approche basée sur la normalisation des données. Nous avons pu ainsi comparer les deux méthodes. Nous considérons ici le cas d'une seule donnée singulière supérieure.

Dans le point suivant, nous présentons le plan de simulation que nous avons adopté pour étudier ces deux approches. Nous présentons, ensuite, les résultats que nous avons pu tirer de cette étude et nous comparons les deux méthodes proposées pour tester la discordance d'une donnée douteuse supérieure.

5.2 Plan de simulation

Nous avons considéré quatre populations statistiques différentes pour notre étude. Nous avons généré p = 1000 échantillons de taille n = 20,40,60 et 100, provenant d'une loi Gamma de paramètre de forme λ = 16,4,1.78 et 1. Les valeurs des paramètres correspondent respectivement aux coefficents d'asymétrie que nous avons fixés à Cs = 0.5,1,1.5 et 2. Ces échantillon ont été simulés sous l'hypothèse nulle, c'ext-à-dire sans aucune donnée singulière, puisque nous nous intéressons à l'erreur de type I.

Pour la méthode par transformation, nous avons donc transformé les données et appliqué ,à chaque échantillon, le test adapté à la loi Normale dans le cas d'une seule donnée singulière supérieure (section 4.2.2).

Pour l'autre méthode, nous avons estimé le paramètre de forme λ à l'aide de la méthode du maximum de vraisemblance, en supprimant la donnée suspecte $x_{(n)}$. Nous avons ensuite appliqué le test sur chaque échantillon en y incluant la donnée douteuse et en considérant le paramètre estimé comme étant le paramètre réel.

5.3 Résultats

Pour les deux méthodes, nous avons alors pu compter le nombre d'échantillons dont la probabilité au dépassement est inférieure aux seuils de signification fixés a priori à α_0 = 1%,5%,10%. Nous avons divisé ces nombres par le nombre total d'échantillon, cela nous a donc donné les seuils de signification réels α_1 que nous avons comparé aux seuils de signification fixés a priori α_0.

Nous avons rangé les résultats obtenus dans des tableaux et nous avons représenté les différents seuils de signification réels α_1 sur des graphiques en fonction de la taille d'échantillon pour les quatre valeurs de coefficient d'asymétrie considérés. Les tableaux et figures 1,2 et 3 donnent respectivement les résultats obtenus pour les seuils a priori α_0 = 0.01, 0.05 et 0.1 pour la méthode par estimation et les tableaux et figures 4,5 et 6 donnent les résultats de la méthode par transformation.

Pour la méthode par estimation (tableux et figure 1, 2 et 3), nous pouvons constater que:

- Le seuil de signification réel α_1 est partout supérieur au seuil fixé a priori α_0, ce qui signifie que le test est libéral, c'est-à-dire qu'il rejette plus souvent qu'il le devrait l'hypothèse nulle.
- α_1 tend vers α_0 à mesure que la taille d'échantillon augmente. Ce test est dons asymptotiquement non-biaisé.
- Cette convergence est plus rapide lorsque le coefficient d'asymétrie est élevé.
- L'évolution des seuils réels avec la taille d'échantillon est régulière.

Pour la méthode par transformation (tableaux et figures 4, 5 et 6), nous pouvons voir que :

- Le seuil de signification réel α_1 est inférieur au seuil fixé a priori α_0, quelque soient la taille d'échantillon et le coefficient d'asymétrie. Ce qui veut dire que le test est conservateur, c'est-à-dire qu'il ne rejette pas assez l'hypothèse nulle.
- α_1 converge vers α_0 quand le coefficient d'asymétrie tend vers zéro (par exemple pour Cs=0.5), ce qui est logique car la loi Normale a un coefficient d'asymétrie nul. Toutefois, lorsque Cs est supérieur à 0.5, α_1 ne semble pas converger vers la valeur théorique à mesure que n augmente. Pour de grands coefficients d'asymétrie, ce test est biaisé même asymptotiquement.
- Le seuil réel varie de façon très irrégulière avec la taille d'échantillon.

La méthode par transformation nous donne des résultats peu satisfaisants, en particulier pour des grandes tailles d'échantillon. Ceci est probablement dû au fait que la transformation utilisée approxime mal la loi Normale, principalement pour de grands Cs. Nous préfèreront donc adopter la méthode par estimation qui fournit des résultats plus réguliers et un test asymptotiquement non-biaisé.

Pour améliorer cette approche, il serait peut-être possible de corriger les écarts entre les seuils réels, α_1, et les seuils fixés a priori ,α_0. En effet, nous avons vu que les seuils réels sont supérieurs aux seuils fixés a priori, en particulier les faibles tailles d'échantillon. Pour ce faire, il suffirait d'effectuer le test en choisissant un α_0 plus faible. Touefois, il serait nécessaire de faire une étude plus poussée sur cet aspect du problème pour savoir de combien il faudrait diminuer α_0 suivant la taille de l'échantillon, et peut-être aussi suivant le coefficient d'asymétrie.

Notons finalement que ces résultats obtenus nous font douter de l'efficacité de l'approche de Carletti. En effet, cette approche utlise trois fois les données de façon non-indépendante et utilise une transformation de données, ce qui ne s'avère pas être une bonne méthode.

Tableau 1 : Seuils de signification réels, $\alpha_0 = 0.01$, méthode par estimation, en fonction de la taille

	taille=20	taille=40	taille=60	taille=100
Cs=0.5	0.053	0.037	0.023	0.016
Cs=1	0.049	0.026	0.02	0.017
Cs=1.5	0.038	0.037	0.017	0.012
Cs=2	0.025	0.017	0.014	0.017

Cs = Coefficient d'asymétrie

Figure 1 : Seuils de signification réels, $\alpha_0 = 0.01$, méthode par estimation, en fonction de la taille

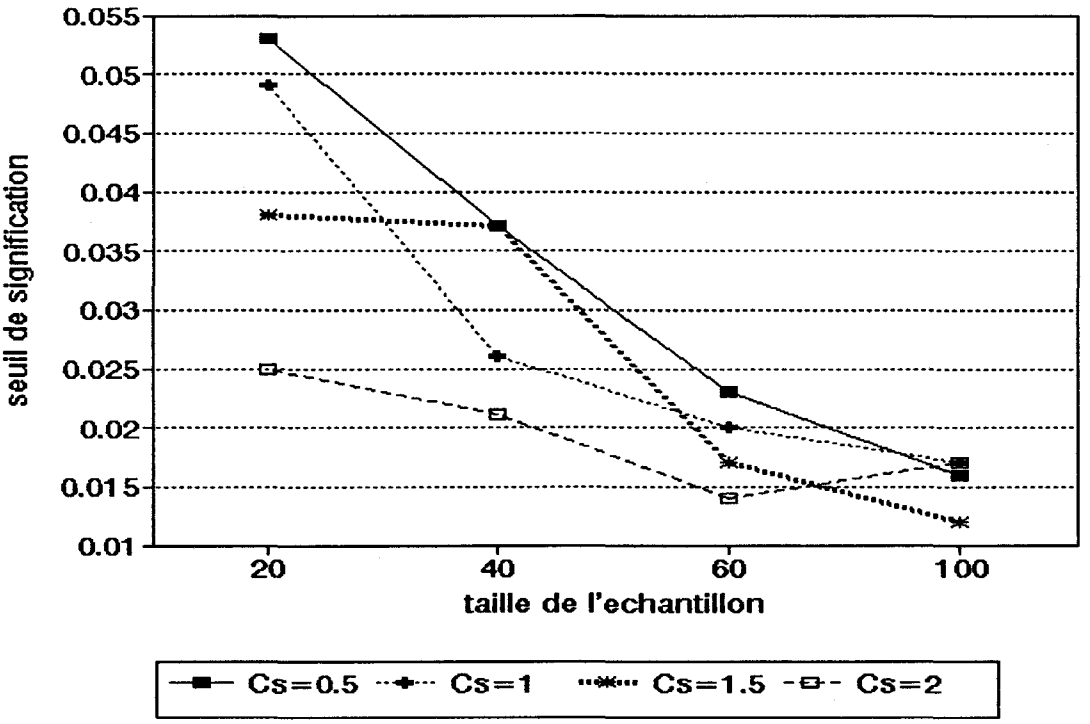


Tableau 2 : Seuils de signification réels, $\alpha_0 = 0.05$, méthode par estimation, en fonction de la taille

	taille=20	taille=40	taille=60	taille=100
Cs=0.5	0.191	0.116	0.082	0.066
Cs=1	0.155	0.096	0.076	0.068
Cs=1.5	0.144	0.104	0.076	0.059
Cs=2	0.103	0.08	0.071	0.048

Cs = coefficient d'asymétrie

Figure 2 : Seuils de signification réels, $\alpha_0 = 0.05$, méthode par estimation, en fonction de la taille

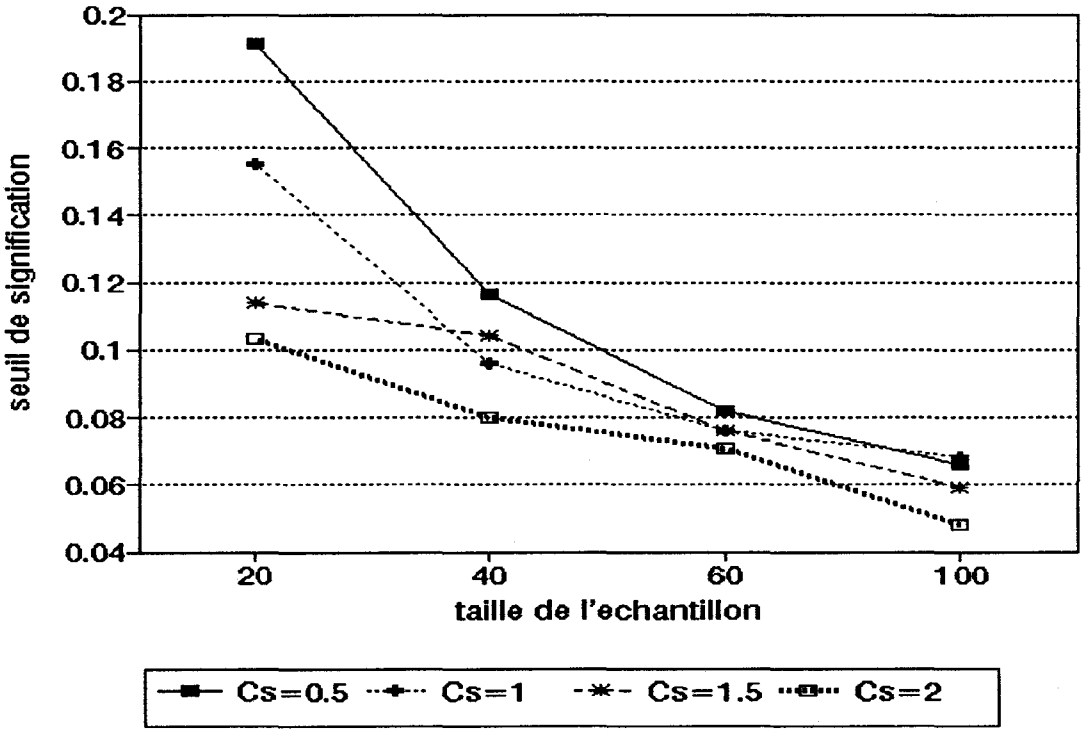


Tableau 3 : Seuils de signification réels, $\alpha_0 = 0.1$, méthode par estimation, en fonction de la taille

	taille=20	taille=40	taille=60	taille=100
Cs=0.5	0.289	0.198	0.164	0.129
Cs=1	0.238	0.166	0.14	0.132
Cs=1.5	0.189	0.158	0.136	0.117
Cs=2	0.137	0.135	0.137	0.106

Cs = coefficient d'asymétrie

Figure 3 : Seuils de signification réels, $\alpha_0 = 0.1$, méthode par estimation, en fonction de la taille

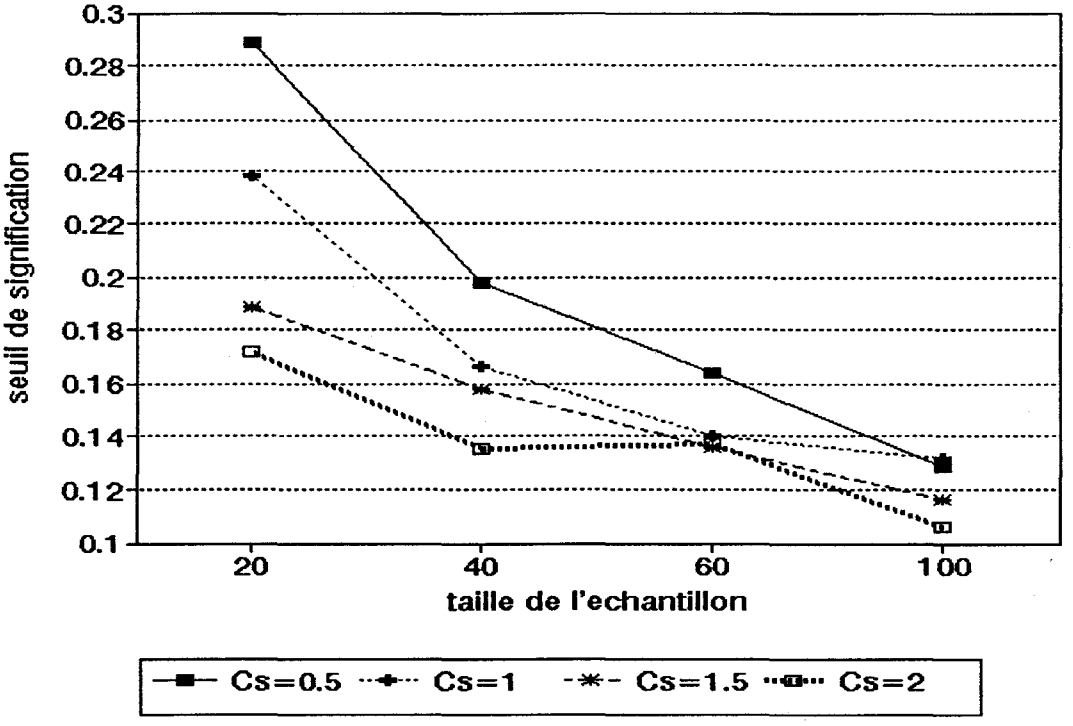


Tableau 4 : Seuils de signification réels, $\alpha_0 = 0.01$, méthode par transformation, en fonction de la taille

	taille=20	taille=40	taille=60	taille=100
Cs=0.5	0.004	0.006	0.005	0.006
Cs=1	0.004	0.002	0.002	0.002
Cs=1.5	0.008	0.001	0.001	0.002
Cs=2	0	0	0.001	0.002

Cs = coefficient d'asymétrie

Figure 4 : Seuils de signification réels, $\alpha_0 = 0.01$, méthode par transformation, en fonction de la taille

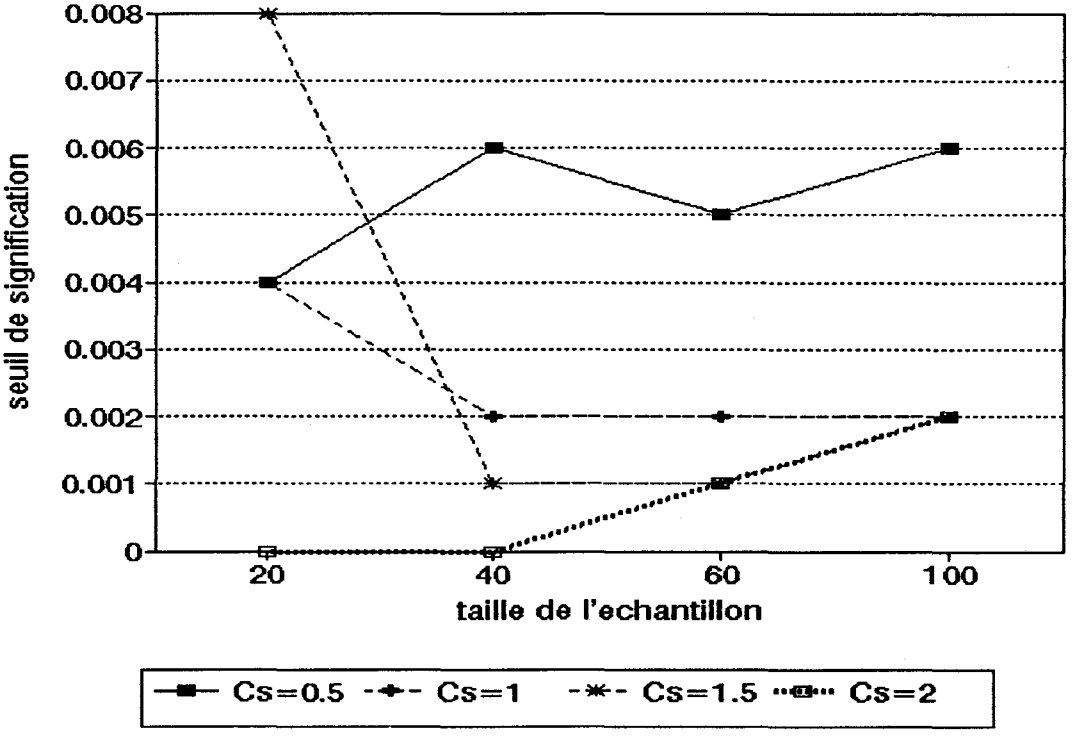


Tableau 5 : Seuils de signification réels, $\alpha_0 = 0.05$, méthode par transformation, en fonction de la taille

	taille=20	taille=40	taille=60	taille=100
Cs=0.5	0.027	0.024	0.026	0.048
Cs=1	0.021	0.015	0.017	0.012
Cs=1.5	0.02	0.017	0.017	0.02
Cs=2	0.012	0.012	0.008	0.01

Cs = coefficient d'asymétrie

Figure 5 : Seuils de signification réels, $\alpha_0 = 0.05$, méthode par transformation, en fonction de la taille

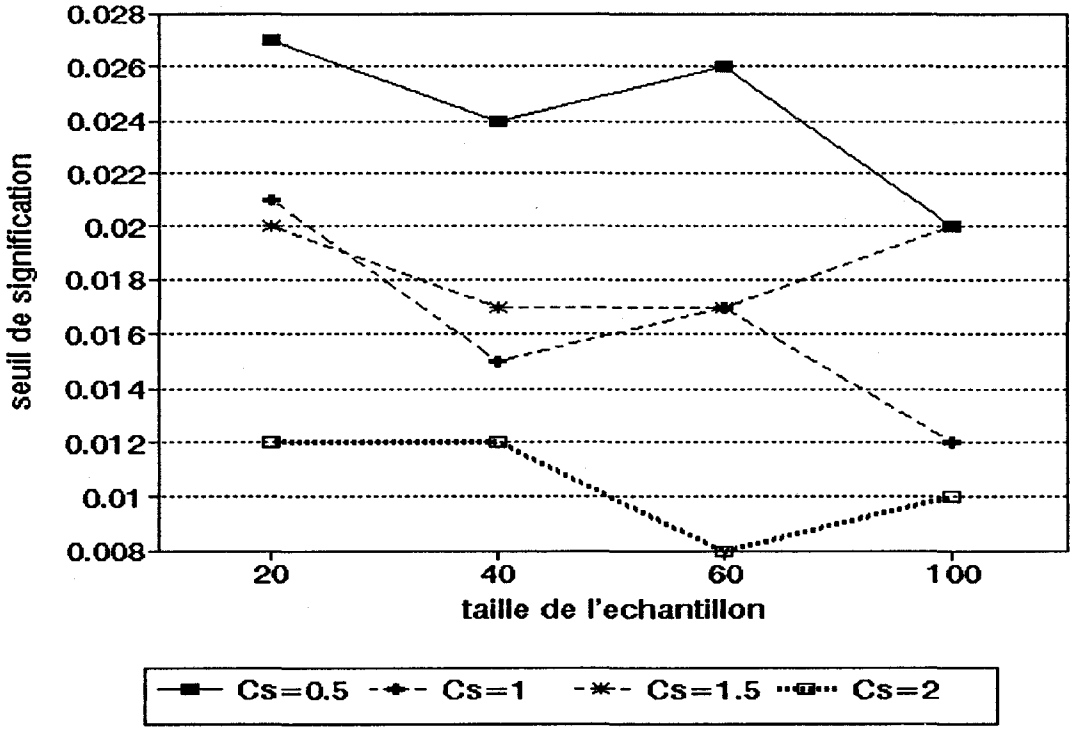
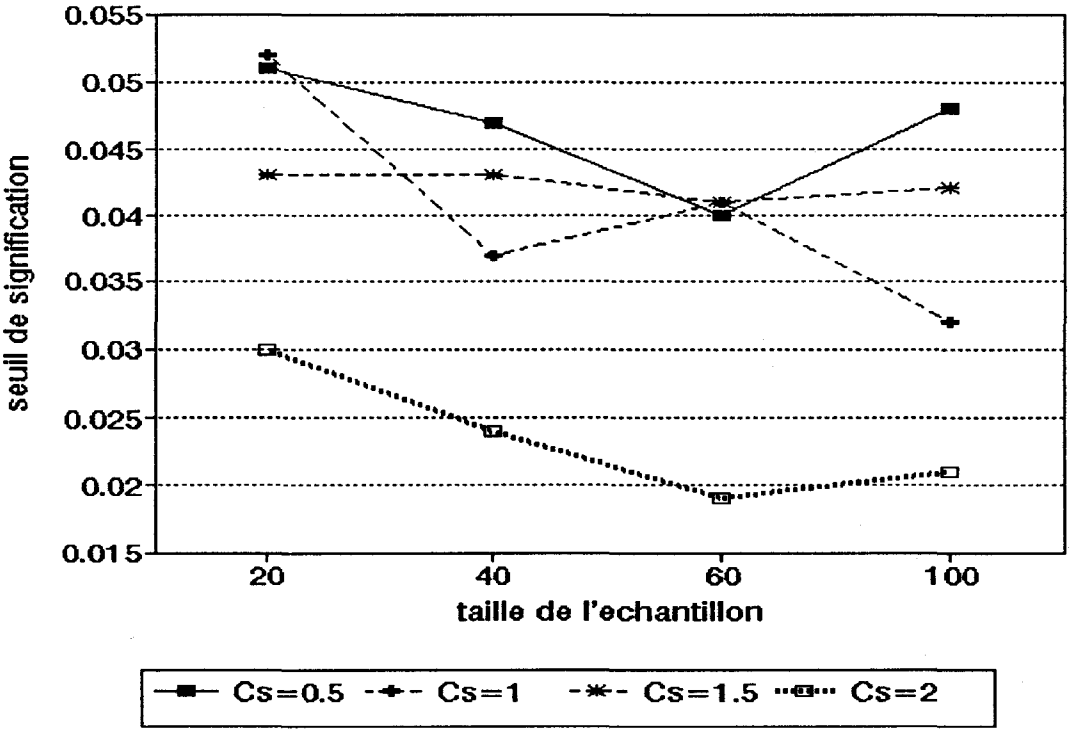


Tableau 6 : Seuils de signification réels, $\alpha_0 = 0.1$, méthode par transformation, en fonction de la taille

	taille=20	taille=40	taille=60	taille=100
Cs=0.5	0.051	0.047	0.04	0.048
Cs=1	0.052	0.037	0.041	0.032
Cs=1.5	0.043	0.043	0.041	0.042
Cs=2	0.019	0.024	0.019	0.021

Cs = coefficient d'asymétrie

Figure 6 : Seuils de signification réels, $\alpha_0 = 0.1$, méthode par transformation, en fonction de la taille



VI CONCLUSION

Les résultats présentés sont les tests qui ont été retenus pour la détection de données singulières pour les différentes lois du logiciel AJUSTE-2. Il s'agit en fait d'une synthèse des principaux résultats que l'on trouve dans la littérature au sujet des données douteuses.

Ce travail a été effectué dans le but d'ajouter ces tests au logiciel AJUSTE-2 dans le cadre d'un projet de partenariat avec Hydro-Québec. Les tests retenus vont donc faire l'objet d'une programmation en langage informatique (le langage C) pour être ensuite implantés dans le logiciel AJUSTE-2, la version actuelle ne contenant que le tests de Grubbs et Beck pour une seule donnée singulière (ce test est présenté en annexe A1).

Les algorithmes des différents tests sont présentés en annexe A2.

Les test de discordance adaptés aux lois Gamma, Gumbel et Weibull supposent la connaissance a priori d'un paramètre, or, en pratique, cette information est rarement disponible. Ces lois étant importantes en hydrologie, il nous a paru intéressant d'étudier le comportement de ces tests lorsque le paramètre en question est estimé. Pour la loi Gamma, nous avons pu constater que cette méthode par estimation était meilleure que la méthode par transformation proposée par Kimber(1979).Nous avons donc choisi d'introduire la méthode par estimation dans la nouvelle version de Ajuste.

Ces résultats nous font penser que la méthode proposée par Barnett et Lewis(1984) pour les loi Gumbel et Weibull devrait nous fournir des résultats similaires. Il serait cependant intéressant de faire une étude plus complète sur ces tests. Par contre, ces résultats nous font douter de l'efficacité de l'approche de Carletti(1976 et 1985).

VII BIBLIOGRAPHIE

7.1 REFERENCES

ABRAMOWITZ M. Et STEGUN I.A., Handbook of mathematical functions, Dover publications, New-York, 1970.

BARNETT V. Et LEWIS T., Outliers in statistical data (second edition), WILEY, New-York, 1984.

BOX G.E.P. Et COX D.R., An analysis of transformations (with discution), J. Roy. Statist. Soc., Ser b, 26, 1964.

CARLETTI G., Détection automatique de valeurs anormales, Revue de statistique appliquée, vol XXIV n°3, 1976.

CARLETTI G., Méthodes de détection de valeurs anormales à une dimension, Biométrie-Praximétrie, n°25, 1985.

CHIKKAGOUDAR M.S. Et KUNCHUR, Comparison of many outlier procedures for exponential samples, Commun. Statist. Theory Meth., Vol 25 n°16, 1987.

D'AGOSTINO R.B. Et STEPHENS M.A., Goodness of fit techniques, Marcel Dekker, New York, 1986.

DRAPPER N.R. Et COX D.R., On dstributions and their transformation to normality , J. Roy. Statit. Soc., Série B, 31, 1969.

FERGUSON T. S., Rules for rejection of outliers, Revue de l'Institut international de statatistique, vol 29 n°3, 1961.

KIMBER A.C., Tests for a single outlier in a gamma sample with unknown shape and scale parameters, Applied statistics, vol 28 n°3, 1979.

GRUBBS F.E. Et BECK G., Extansion of sample sizes and percentage points for significance tests of outlying observations, Technometrics, vol 14 n°4, novembre 1972.

KIMBER A.C., Tests for many outliers in an exponential sample, Applied statistics, vol 31 n°3, 1982.

MADANSKY A., Prescription for working statisticians, Springer-Verlas, New-York, 1982.

MOSTELLER F., A k-sample slippage test for an extreme population, Ann. Math. Statist., 19, 1948.

PEARSON E.S. Et HARTLEY H.O., Biometrika tables for statisticians (vol I), University Press, Cambridge, 1966.

TIKU M.L., A new Statistic for testing suspected outliers, Communications in statistics, vol 4 n°8, 1975.

WALSH J.E., Some nonparametric tests of wether the largest observations of a set are too large or too small, Ann. Math. Statist., 21,1950. Correction, Ann. Math. Statist., 24, 1953.

WALSH J.E., Large sample non-parametric rejections of outlying observations, Ann. Inst. Statist. Math. Tokyo, 10, 1959.

WALSH J.E., Handbook of non-parametric statistics, II, Van Nostrand, Princeton, 1965.

WALSH J.E. Et KELLEHER G.J., Non parametric estimation of mean and variance when a few "sample" values possibly outliers, Ann. Inst. Statist. Math. Tokyo, 25, 1974.

7.2 AUTRES OUVRAGES CONSULTES

BECKMAN R.J. Et COOK R.D., Outlier...s, Technometrics, vol 25 n°2, mai 1983.

BOBEE B. Et ASHKAR F., The gamma family and derived distributions applies in hydrology, Water resources publications, 1991.

Commissariat à l'Energie Atomique, Statistique appliquée à l'exploitation des mesures (Tome I), Masson, Paris, 1978.

DIXON W.J., Analysis of extreme values, The annals of mathematical statistics, vol 21 n°4, 1950.

FIELLER N.R.J. And LEWIS T., A recursive algorithm for null distributions for outliers : I Gamma samples, Technometrics, vol 21 n°3, 1979.

GRUBBS F. E., Sample criteria for testing outlying observations, The annals of mathematical statistics, vol 21 n°1, 1950.

JAIN R.B., PINDEL L.A. AND DAVIDSON J.L., A unifiedapproach for estimation and detection of outliers, commun. statist. theor. meth., Vol 11 n°25, 1982.

KABE D.G., Testing outliers from a exponential population, Metrika, 15, 1970.

KALE B.K., Detections of outliers, Sankhya : The indian journal of statistics, vol 38, series B, pt 4, 1976.

KIMBER A.C., Testing upper and lower outliers pairs in Gamma samples, Commun. Statist. -Simula, vol 17 n°3, 1988.

KIMBER A.C. Et STEVENS H.J., The null distribution of a test for two upper outliers in an exponential sample, Applied statistics, vol 30 n°2, 1981.

MOORE R. H. Et TIETJEN G. L., Some Grubbs-type statistics for the detection of several outliers, Technometrics, vol 14, 1972.

PERREAULT L. Et BOBEE B., Loi Weibull à deux paramètres : propriétés mathématiques et statistiques, estimation des paramètres et des quantiles X_T de période de retour T, Rapport scientifique $n_0$351, INRS-Eau, 1992.

PRESCOTT P., Critical values for a sequential test for many outliers, Applied statistics, vol 28 n°1, 1979.

ANNEXES

A.1 TEST DE GRUBBS ET BECK

Le test de Grubbs et Beck (1972), actuellement utilisé dans Ajuste, est décrit dans ce qui suit :

- Condition d'application :

Ce test est applicable uniquement pour des échantillons tirés d'une population Normale. Pour pouvoir l'appliquer aux données hydrologiques, nous prenons les logarithmes des données observées et nous supposons que les nouvelles données sont normalement distribuées.

- Principe :

Soient X_H et X_L les limites respectivement supérieure et inférieure de ce test. Toute valeur simple supérieure à X_H est considérée comme une donnée singulière supérieure, et celle inférieure à X_L est considérée comme une donnée singulière inférieure. Le logiciel génère un graphique identifiant les valeurs singulières avec les numéros d'observation respectifs.

- Détermination des limites :

$$X_H = \exp(\overline{X} + K_n S) \text{ et } X_L = \exp(\overline{X} - K_n S)$$

où $\overline{X}$ est la moyenne arithmétique des logarithmes des observations
S est l'écart-type des logarithmes des observations
n est la taille de l'échantillon

Avec $\alpha = 10\%$, on peut approximer K_n à l'aide du polynôme suivant :

$$K_n = 3{,}62201 + 6{,}28446 n^{1/4} - 2{,}49835 n^{1/2} + 0{,}491436 n^{3/4} - 0{,}037911 n$$

A.2 ALGORITHMES DES TESTS

Remarque :

Notre règle de décision pour tous les tests est la suivante :
Si $P(t) \geq \alpha$, alors nous adoptons l'hypothèse H_0,
Si $P(t) < \alpha$, alors nous rejetons l'hypothèse H_0,
P(t) étant la probabilité au dépassement relative à une valeur observée t d'une statistique de discordance T et définie comme suit :

$$P(t) = Prob(T \geq t)$$

Or, les tests retenus ne donnent pas souvent une valeur exacte pour P(t), mais davantage une limite supérieure.
Exemple : $P(t) \leq k$ au lieu de $P(t) = k$

Note règle de décision est alors la suivante :
Si $k \geq \alpha$, alors nous adoptons l'hypothèse H_0
Si $k < \alpha$, alors nous adoptons l'hypothèses H_1

De même, il arrive que P(t) soit compris dans un intervalle.
Exemple : $a \leq P(t) < b$

La règle de décision adoptée est alors celle-ci :
Si $a \leq \alpha < b$, alors nous adoptons l'hypothèse H_0
Si $a > \alpha$, alors nous adoptons l'hypothèse H_0
Si $b \leq \alpha$, alors nous adoptons l'hypothèse H_1

Les tests ainsi effectués seront donc plus conservateurs, c'est-à-dire que l'hypothèse H_0 sera moins souvent rejetée qu'elle le devrait.

A.2.1 Loi normale

A.2.1.1 Test pour une donnée singulière inférieure ou supérieure(section 4.2.2)

debut

<< Ranger l'échantillon par ordre croissant >>

<< *Calcul de la statistique de discordance :* >>

$$moy := \frac{\Sigma x_i}{n}$$

$$s := \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n-1}}$$

si $\frac{x_{(n)} - \bar{x}}{s} > \frac{\bar{x} - x_{(1)}}{s}$ faire

$$ext := x_{(n)}$$

$$t := \frac{x_{(n)} - \bar{x}}{s}$$

sinon_faire $ext := x_{(1)}$

$$t := \frac{\bar{x} - x_{(1)}}{s}$$

fin_si

<< *Calcul de la probabilité au dépassement (l'algorithme de la fonction prob-student est donnée en A.2.1.3):* >>

$$\nu := n - 2$$

$$T := \sqrt{\frac{n(n-2)t^2}{(n-1)^2 - nt^2}}$$

prob:=2n[1-prob_student(ν, T)]

< < *Afficher les résultats:* > >

ecrire("La statistique observée du test de discordance est :",t)
ecrire("La probabilité au dépassement est inférieure ou égale à :",prob)

si $prob \geq \alpha$ faire
ecrire("La valeur extrême",ext,"n'est pas une valeur singulière au seuil de signification", α)
sinon-faire
ecrire("La valeur extrême",ext,"est une valeur singulière au seuil de signification", α)
fin-si

fin

A.2.1.2 Test pour k données singulières supérieures(section 4.2.3)

debut
< < Ranger l'échantillon par ordre croissant > >

< < *Détermination du nombre de données douteuses à tester:* > >

ecrire("Donner le nombre de données à tester :")
lire(k)

< < *Calcul de la statistique de discordance:* > >

$$moy := \frac{\Sigma x_i}{n}$$

$$s := \sqrt{\frac{\Sigma (x_i - \overline{x})^2}{n-1}}$$

$$t := \left(\sum_{i=n-k+1}^{n} x_i - k\overline{x} \right) \frac{1}{s}$$

<< *Calcul de la probabilité au dépassement:* >>

$\nu := n-2$

$$comb := \frac{n!}{k!(n-k)!}$$

$$T := \sqrt{\frac{n(n-2)t^2}{k(n-k)(n-1)-nt^2}}$$

prob: = comb*[1-prob-student(ν, T)]

<< *Afficher les résultats:* >>

ecrire ("La statistique observée du test est:",t)
ecrire("La probabilité au dépassement est inférieure ou égale à :",prob)

si $prob \geq \alpha$ faire
    ecrire("Les",k,"données extrêmes supérieures ne sont pas des données singulières au seuil de signification", α)
sinon-faire
    ecrire("Les",k,"données extrêmes supérieures sont des données singulières au seuil de signification",α)
fin-si

fin

A.2.1.3 Loi de Student

<< A étant la probabilité au non-dépassement d'un nombre T donné pour la loi de Student avec un degré de liberté égal à ν, nous avons (Abramowitz et Stegun,1970) :

- pour $\nu > 1$ et impair :

$$A = \frac{2}{\pi}\left\{\theta + \sin\theta\left(\cos\theta + \frac{2}{3}\cos^3\theta + \ldots + \frac{2*4*\ldots*(\nu-3)}{1*3*\ldots*(\nu-2)}\cos^{\nu-2}\theta\right)\right\}$$

- pour $\nu = 1$:

$$A = \frac{2}{\pi}\theta$$

- pour $\nu > 1$ et pair :

$$A = \sin\theta\left\{1 + \frac{1}{2}\cos^2\theta + \frac{1*2}{2*4}\cos^4\theta + \ldots + \frac{1*3*5*\ldots*(\nu-3)}{2*4*6*\ldots*(\nu-2)}\cos^{\nu-2}\theta\right\}$$

avec $\theta = arctan\frac{T}{\sqrt{\nu}}$ >>

<u>debut</u>

$$\theta = arctan\frac{T}{\sqrt{\nu}}$$

<u>si</u> ν impair <u>et</u> $\nu \neq 0$ <u>faire</u>

term := Cosθ

somm := 0

<u>pour</u> expo=3 <u>a</u> $\nu - 2$ <u>faire</u>

$$term := term.\cos^2\theta.\frac{expo-1}{expo}$$

somm := somm + term

expo := expo + 2

<u>fin-pour</u>

$$prob := \frac{2}{\pi}(\theta + (\sin\theta(\cos\theta + somm)))$$

<u>fin-si</u>

si $\nu = 1$ faire

$prob := \frac{2\theta}{\pi}$

fin-si

si ν pair faire

term := 1

somm := 0

pour expo=2 a $\nu - 2$ faire

$term := term\left(\frac{expo-1}{expo}\cos^2\theta\right)$

somm := somm + term

fin-pour

$prob := (1 + somm)\sin\theta$

fin-si

fin

A.2.2 **Loi log-Normale** (section 4.3)

```
debut

< < Ranger l'échantillon par ordre croissant > >

< < Transformer les données: > >
        pour i=1 a n faire
            x(i) := ln x(i)
        fin-pour

< < Demander à l'utilisateur ce qu'il veut tester: > >
        ecrire("Vous voulez tester la discordance d'une donnée douteuse inférieure
        ou supérieure(1) ou d'un ensemble de données douteuses supérieures(2) ? :")
        lire(choix)

        si choix=1 faire
            < < Appliquer le test A.2.1.1 > >
        fin-si

        si choix=2 faire
            < < Appliquer le test A.2.1.2 > >
        fin-si
fin
```

A.2.3 **Loi gamma** (de paramètre de forme λ connu ou estimé)

A.2.3.1 Test pour une donnée singulière supérieure(section 4.4.2)

debut
< < Ranger l'échantillon par ordre croissant > >

< < *Calcul de la statistique de discordance:* > >

$$t := \frac{x_{(n)}}{\Sigma x_{(i)}}$$

< < *Calcul de la probabilité au dépassement* > >

$$\nu_1 := 2\lambda$$

$$\nu_2 := 2(n-1)\lambda$$

$$F := \frac{(n-1)t}{1-t}$$

prob := n * prob_fisher($\nu 1, \nu 2, F$)

< < *Afficher les résultats:* > >

ecrire("La statistique observée du test de discordance est :",t)
ecrire("La probabilité au dépassement est inférieure ou égale à :",prob)

si prob≥ α faire
ecrire("La valeur extrême supérieure",$x_{(n)}$,"n'est pas une donnée singulière au seuil de signification, α)
sinon-faire
ecrire("La valeur extrême supérieure",$x_{(n)}$,"est une valeur singulière au seuil de signification", α)
fin-si

fin

A.2.3.2 Test pour une donnée singulière inférieure(4.4.3)

< < Ranger par ordre croissant > >

< < *Calcul de la statistique de discordance:* > >

$$t := \frac{x_{(1)}}{\Sigma x_{(i)}}$$

< < *Calcul de la probabilité au dépassement:* > >

$$\nu 1 := 2\lambda$$

$$\nu 2 := 2(n-1)\lambda$$

$$F := \frac{(n-1)t}{1-t}$$

prob := n (1 - prob_fisher($\nu 1, \nu 2, F$))

< < *Afficher les résultats:* > >

ecrire("La statistique observée du test de discordance est :",t)
ecrire("La probabilité au dépassement est inférieure ou égale à :",prob)

si prob≥ α faire
ecrire("La valeur extrême inférieure",$x_{(1)}$,"n'est pas une donnée singulière au seuil de signification ,α
sinon-faire
ecrire("La valeur extrême inférieure",$x_{(1)}$,"est une valeur singulière au seuil de signification", α)
fin-si

fin

A.2.3.3 Test pour plusieurs données singulières supérieures(section 4.4.4)

< < Ranger par ordre croissant > >

< < *Détermination du nombre de données douteuses à tester:* > >

ecrire("Donner le nombre de données à tester :")
lire(k)

< < *Calcul de la statistique de discordance* > >

$$t := \left(\sum_{j=n-k+1}^{n} x_{(j)} \right) \div \left(\sum_{i=1}^{n} x_{(i)} \right)$$

< < *Calcul de la probabilité au dépassement:* > >

$$\nu 1 := 2kr$$

$$\nu 2 := 2(n-k)\lambda$$

$$F := \frac{(n-k)t}{k(1-t)}$$

$$comb := \frac{n!}{k!(n-k)!}$$

prob := comb * prob_fisher($\nu 1, \nu 2, F$)

< < *Afficher les résultats:* > >

ecrire("La statistique observée du test de discordance est :",t)
ecrire("La probabilité au dépassement est inférieure ou égale à :",prob)

si $prob \geq \alpha$ faire
    ecrire("Les",k,"données extrêmes supérieures ne sont pas des données singulières au seuil de signification", α)

sinon-faire
ecrire("Les",k,"données extrêmes supérieures sont des données singulières au seuil de signification", α)
fin-si
fin

A.2.3.4 Test pour plusieurs données singulières inférieures (section 4.4.5)

<< Ranger par ordre croissant >>

<< *Déterminer le nombre de données douteuses à tester:* >>
ecrire("Donner le nombre de données à tester :")
lire(k)

<< *Calcul de la statistique de discordance:* >>

$$t := \left(\sum_{j=1}^{k} x_{(j)} \right) \div \left(\sum_{i=1}^{n} x_{(i)} \right)$$

<< *Calcul de la probabilité au dépassement :* >>

$$\nu 1 := 2kr$$

$$\nu 2 := 2(n-k)\lambda$$

$$F := \frac{(n-k)t}{k(1-t)}$$

$$comb := \frac{n!}{k!(n-k)!}$$

prob := comb * prob_fisher($\nu 1, \nu 2, F$)

<< *Afficher les résultats:* >>
ecrire("La statistique observée du test de discordance est :",t)
ecrire("La probabilité au dépassement est inférieure ou égale à :",prob)

si $prob \geq \alpha$ faire

ecrire("Les",k,"données extrêmes inférieures ne sont pas des données singulières au seuil de signification", α)
sinon-faire
ecrire("Les",k,"données extrêmes inférieures sont des données singulières au seuil de signification", α)
fin-si

fin

A.2.3.5 Loi de Fisher

<< P étant la probabilité au dépassement d'un nombre T donné pour la loi de Fisher ayant ν_1 et ν_2 degrés de liberté, nous avons (Abramowitz et Stegun, 1970) :

- Si ν_1 est pair,

$$P = x^{\frac{\nu_1+\nu_2-2}{2}}\left\{1+\frac{\nu_1+\nu_2-2}{2}\left(\frac{1-x}{x}\right)+\frac{(\nu_1+\nu_2-2)(\nu_1+\nu_2-4)}{2*4}\left(\frac{1-x}{x}\right)^2+\ldots+\frac{(\nu_1+\nu_2-2)\ldots(\nu_2+2)}{2*4*\ldots*(\nu_1-2)}\left(\frac{1-x}{x}\right)^{\frac{\nu_1-2}{2}}\right\}$$

- Si ν_1 est impair et ν_2 pair ,

$$P = 1-(1-x)^{\frac{\nu_1+\nu_2-2}{2}}\left\{1+\frac{\nu_1+\nu_2-2}{2}\left(\frac{x}{1-x}\right)+\ldots+\frac{(\nu_1+\nu_2-2)\ldots(\nu_1+2)}{2*4*\ldots*(\nu_2-2)}\left(\frac{x}{1-x}\right)^{\frac{\nu_2-2}{2}}\right\}$$

- Si ν_1 et ν_2 sont impairs,

$$P = 1 - A + B$$

pour $\nu_2 > 1$, $$A=\frac{2}{\pi}\left\{\theta+\sin\theta\left(\cos\theta+\frac{2}{3}\cos^3\theta+\ldots+\frac{2*4*\ldots*(\nu_2-3)}{1*3*\ldots*(\nu_2-2)}\cos^{\nu_2-2}\theta\right)\right\}$$

pour $\nu_1 = 1$, $A = \frac{2}{\pi}\theta$

pour $\nu_2 > 1$, $B = \frac{2}{\sqrt{\pi}}\frac{\left(\frac{\nu_2-1}{2}\right)!}{\left(\frac{\nu_2-2}{2}\right)!}\sin\theta\cos^{\nu_2}\theta\left\{1+\frac{\nu_2+1}{3}\sin^2\theta+\ldots+\frac{(\nu_2+1)(\nu_2+3)\ldots(\nu_1+\nu_2-4)\sin^{\nu_1-3}\theta}{3*5*\ldots*(\nu_1-2)}\right\}$

pour $\nu_1 = 1$, $B = 0$

où $x = \frac{\nu_2}{\nu_2+\nu_1 F}$

et $\theta = arctan\sqrt{\frac{\nu_1}{\nu_2}F}$

>>

<u>debut</u>

<u>si</u> ν_1 pair <u>faire</u>

$x := \frac{\nu_2}{\nu_2+\nu_1 F}$

term := 1

somm := 0

<u>pour</u> exp=1 <u>a</u> $\frac{\nu_1-2}{2}$ <u>faire</u>

$term := term * \frac{\nu_1+\nu_2-2\exp}{2\exp}\left(\frac{1-x}{\exp}\right)$

somm := somm + term

exp := exp + 1

<u>fin-pour</u>

$prob := (1+somm)x^{\frac{\nu_1+\nu_2-2}{2}}$

<u>fin-si</u>

si ν_1 impair et ν_2 pair faire

$x := \frac{\nu_2}{\nu_2 + \nu_1 F}$

term := 1

somm := 0

pour exp=1 a $\frac{\nu_2 - 2}{2}$ faire

$term := term * \frac{\nu_1 + \nu_2 - 2exp}{2exp}\left(\frac{1-x}{exp}\right)$

somm := somm + term

exp := exp + 1

fin-pour

$prob := 1 - \left\{(1 + somm)(1-x)^{\frac{\nu_1 + \nu_2 - 2}{2}}\right\}$

fin-si

si ν_1 impair et ν_2 impair faire

$\theta := arctan\sqrt{\frac{\nu_1}{\nu_2}F}$

si $\nu_2 = 1$ et $\nu_1 \neq 1$ faire

tmp := 1

$\nu_2 := \nu_1$

$\nu_1 := 1$

$F := \frac{1}{F}$

fin-si

si $\nu_2 > 1$ faire

$term := term * \cos\theta$

pour exp=3 a $\nu_2 - 2$ faire

$term := term * \cos^2\theta * \frac{exp-1}{exp}$

somm := somm + term

exp := exp + 1

fin-pour

$A := \frac{2}{\pi}\{\theta + \sin\theta(\cos\theta + somm)\}$

sinon-faire $A := \frac{2\theta}{\pi}$

fin-si

si $\nu_1 \neq 1$ faire

term := 1

somm := 1

pour exp=2 a $\nu_1 - 3$ faire

$term := term \frac{\nu_2 + exp - 1}{exp + 1}\sin^2\theta$

somm := somm+term

exp := exp+2

fin-pour

$B := \frac{2}{\sqrt{\pi}} \frac{\left(\frac{\nu_2-1}{2}\right)!}{\left(\frac{\nu_2-2}{2}\right)!} \sin\theta.(\cos\theta)^{\nu_2}.somm$

sinon-faire B := 0

fin-si

prob := 1-A+B

si tmp=1 faire

prob := 1-prob

fin-si

fin-si

fin

A.2.3.6 Méthode par transformation de variable(section 4.4.6)

```
debut
< < Ranger les données par ordre croissant > >

< < Transformer les données: > >
        pour i=1 a n faire
            x(i) := x(i)^(1/3)
        fin-pour

< < Demander à l'utilisateur ce qu'il veut tester: > >
        ecrire("Vous voulez tester la discordance d'une donnée douteuse inférieure ou
        supérieure(1) ou d'un ensemble de données douteuses supérieures(2) ? :")
        lire(choix)

        si choix=1 faire
            < < Appliquer le test A.2.1.1 > >
        fin-si

        si choix=2 faire
            < < Appliquer le test A.2.1.2 > >fin-si
fin
```

A.2.4 **Loi log-Gamma** (section 4.5)

```
debut

< < Ranger les données par ordre croissant > >

< < Demander à l'utilisateur ce qu'il veut tester: > >
        ecrire("Vous voulez tester la discordance d'une donnée douteuse supérieure(1)
        ou inférieure(2) ou d'un ensemble de données douteuses supérieures(3) ou
        inférieures(4) ? :")
        lire(choix)

        si choix=1 faire
            < < Appliquer le test A.2.3.1 > >
        fin-si

        si choix=2 faire
            < < Appliquer le test A.2.3.2 > >
        fin-si

        si choix=3 faire
            < < Appliquer le test A.2.3.3 > >
        fin-si

        si choix=4 faire
            < < Appliquer le test A.2.3.4 > >
        fin-si
fin
```

A.2.5 Loi Exponentielle

A.2.5.1 Test pour une ou plusieurs données singulières supérieures(section 4.6.2)

debut

< < Ranger l'échantillon par ordre croissant > >

< < *Déterminer le nombre de données singulières à tester:* > >

ecrire("Donner le nombre de données à tester :")
lire(k)

j := k
prob := α
b_sup := α + 1

tant-que j>0 et $prob \geq \alpha$ et $b\_sup > \alpha$ faire

< < *Calcul de la statistique de discordance :* > >

$$t := \frac{x_{(n-j+1)}}{\sum_{i=j}^{n-j+1} x_{(i)}}$$

$$comb := \frac{n!}{j!(n-j)!}$$

< < *Calcul de la probabilité au dépassement:* > >

si 1/2<t<1 faire

$$prob := comb * \left(\frac{1-t}{1+jt-t} \right)^{n-j}$$

sinon-faire

$$b\_inf := comb * \frac{(1-t)^{n-j} - \{j(n-j)(1-2t)^{n-j}/(j+1)\}}{(1+jt-t)^{n-j)}}$$

$$b\_sup := comb * \left(\frac{1-t}{1+jt-t}\right)^{n-j}$$

fin-si

j := j-1

fin-tant-que

< < *Afficher les résultats :* > >

ecrire("La statistique observée du test de discordance est :",t)

si 1/2<t<1 faire

ecrire("La probabilité au dépassement est :",prob)

sinon-faire

ecrire("La probabilité au dépassement est comprise entre ".b_inf," et ",b_sup)

fin-si

si j=0 faire

ecrire("Aucune donnée singulière n'a été détectée au seuil de signification",α)

sinon-faire

si j=1 faire

ecrire("La valeur extrême supérieure est une valeur singulière au seuil de signification",α)

sinon-faire

ecrire("Les",k,"valeurs extrêmes supérieures sont des données singulières au seuil de signification",α)

fin-si

fin-si

fin

A.2.5.2 Test pour une ou plusieurs données singulières inférieures(section 4.6.3)

debut
<< Ranger l'échantillon par ordre croissant >>

<< *Déterminer le nombre de données singulières à tester:* >>
ecrire("Donner le nombre de données à tester :")
lire(k)

j := k
prob := α
b_sup := α + 1

tant-que j>0 et $prob \geq \alpha$ et $b\_sup > \alpha$ faire

<< *Calcul de la statistique de discordance:* >>

$$t := \frac{x_{(j+1)}}{\sum_{i=1}^{j+1} x_{(i)}}$$

$$comb := \frac{n!}{j!(n-j)!}$$

<< *Calcul de la probabilité au dépassement* >>
si 1/2<t<1 faire

$$prob := comb * \left(\frac{1-t}{1+(n-j)t-t} \right)^{j}$$

sinon-faire

$$b\_inf := comb * \frac{(1-t)^{j} - \{j(n-j)(1-2T)^{j}/(n-j+1)\}}{(1+(n-j)t-t)^{j}}$$

$$b\_sup := comb * \left(\frac{1-t}{1+(n-j)t-t} \right)^{j}$$

fin-si

```
            j := j-1

            fin-tant-que

<< Afficher les résultats: >>
            ecrire("La statistique observée du test de discordance est :",t)
            si 1/2<t<1 faire
                ecrire("La probabilité au dépassement est :",prob)
            sinon-faire
                ecrire("La probabilité au dépassement est comprise entre ".b_inf," et ",b_sup)
            fin-si

            si j=0 faire
                ecrire("Aucune donnée singulière n'a été détectée au seuil de
                 signification", α)
            sinon-faire
                si j=1 faire
                    ecrire("La valeur extrême inférieure est une valeur singulière
                    au seuil de signification ",α)
                sinon-faire
                    ecrire("Les",k,"valeurs extrêmes inférieures sont des valeurs
                     singulières au seuil de signification",α)
                fin-si
            fin-si
fin
```

A.2.6 **Loi Weibull** (de paramètre de forme c connu ou estimé) (section 4.7)

debut

< < Ranger les données par ordre croissant > >

< < *Transformer les données:* > >
pour i=1 a n faire
$x_{(i)} := (x_{(i)})^c$
fin-pour

< < *Demander à l'utilisateur ce qu'il veut tester:* > >
ecrire("Vous voulez tester la discordance d'une ou plusieurs données douteuses supérieures (1) ou une ou plusieurs données douteuses inférieures(2) ? :")
lire(choix)

si choix=1 faire
< < Appliquer le test A.2.4.1 > >
fin-si

si choix=2 faire
< < Appliquer le test A.2.4.2 > >
fin-si

fin

A.2.7 **Loi Gumbel** (de paramètre d'échelle b connu ou estimé) (section 4.8)

debut

< < Ranger les données par ordre croissant > >

< < *Transformer les données:* > >

pour i=1 a n faire

$$x_{(i)} := \exp\left(\frac{-x_{(i)}}{b}\right)$$

fin-pour

< < *Demander à l'utilisateur ce qu'il veut tester:* > >

ecrire("Vous voulez tester la discordance d'une ou plusieurs données douteuses supérieures (1) ou une ou plusieurs données douteuses inférieures(2) ? :")
lire(choix)

si choix=1 faire
< < Appliquer le test A.2.4.1 > >
fin-si

si choix=2 faire
< < Appliquer le test A.2.4.2 > >
fin-si

fin