REVUE DE STATISTIQUE APPLIQUÉE

G. CARAUX
O. GASCUEL

Approximations et majorations optimales des statistiques d'ordre d'un échantillon

Revue de statistique appliquée, tome 39, n° 1 (1991), p. 21-35
<http://www.numdam.org/item?id=RSA_1991__39_1_21_0>

Rev. Statistique Appliquée, 1991, **XXXIX** (1), 21-35

APPROXIMATIONS ET MAJORATIONS OPTIMALES DES STATISTIQUES D'ORDRE D'UN ÉCHANTILLON

G. CARAUX*†, O. GASCUEL†

RÉSUMÉ

Les valeurs extrêmes d'un échantillon sont de bons indicateurs de la validité d'un modèle [11]. Nous nous en servirons pour détecter des valeurs atypiques dans le cadre d'un modèle posé a priori. Plus précisément, nous chercherons ici à faciliter la mise en oeuvre de tests statistiques portant sur les statistiques d'ordre d'un échantillon. A cette fin nous présenterons des résultats généraux non spécifiques d'une loi de probabilité particulière. Nous étudierons, dans un premier temps, des approximations de la fonction de répartition des statistiques d'ordre quand les variables aléatoires de l'échantillon sont indépendantes. Dans un deuxième temps, nous aborderons le cas général (non indépendance) pour lequel nous exhiberons un système de minoration et de majoration de la fonction de répartition. Les bornes trouvées pourront être utilisées dans la construction d'un test statistique portant sur les statistiques d'ordre d'un échantillon.

Mots-clés : *Valeurs extrêmes, Statistiques d'ordre, Borne exacte, Approximation, Dépendance.*

1. Introduction

L'analyse des valeurs extrêmes d'un échantillon de variables aléatoires est un sujet d'étude ancien en statistique (pour un bon aperçu de ce domaine consulter [4], [6] et [11]). Il fait encore l'objet d'une attention particulière [27] et trouve de nombreuses applications dans des domaines variés : météorologie [25], détection de seuil d'alerte [24], contrôle de qualité [14], résistance des matériaux [27], étude des processus stochastiques [18], analyse statistique non paramétrique [2] et plus particulièrement traitement des données aberrantes (outliers), suspectes ou douteuses [1] [3].

Ce dernier domaine est l'un des plus riches et des plus féconds. Une littérature ancienne [21] [26] [15] et abondante sur ce domaine est là pour en témoigner. Ces travaux portent sur des échantillons de variables aléatoires indépendantes et identiquement distribuées. Une attention toute particulière y est donnée au cas où les variables aléatoires rencontrées sont Gaussiennes [19] ou exponentielles.

* Unité de Biométrie, Ecole Nationale Supérieure Agronomique, Place P. Viala, 34060 Montpellier Cédex (France). E. mail : CARAUX@Montpellier.inra.fr.

† CRIM, 860 rue de Saint Priest, 34100 Montpellier (France).

En laissant de côté l'étude des propriétés asymptotiques des valeurs extrêmes [8] [12] [16], ces travaux peuvent grossièrement se regrouper en deux classes non disjointes [1]. La première couvre les problèmes posés par la recherche d'estimateurs robustes, et l'autre ceux liés à l'identification d'observations aberrantes (une ou plusieurs).

Notre contribution se rattache ici à cette dernière classe et plus particulièrement à la mise en oeuvre de tests statistiques capables de mettre en évidence des situations atypiques sous une hypothèse fixée a priori.

Après avoir exposé, par un exemple illustratif, notre problème nous le situerons dans un contexte formel plus vaste que nous étudierons. Nous serons alors amené à nous intéresser aux valeurs extrêmes d'un échantillon. Dans le cas où les variables aléatoires de l'échantillon sont indépendantes, nous proposerons des approximations simplifiant la mise en oeuvre de tests statistiques. Dans le cas général nous exhiberons une minoration et une majoration de la fonction de répartition des valeurs extrêmes.

Enfin nous conclurons notre propos après avoir discuté nos résultats.

2. Motivations et exposé du problème

Ce travail prend sa source en apprentissage inductif à partir d'exemples. C'est un domaine de recherche partagé par la reconnaissance des formes [20] et l'intelligence artificielle [7] [13]. Les problèmes abordés y sont classiques et portent sur la discrimination, la prédiction de variables quantitatives, la classification automatique, etc ... L'originalité de l'approche dévelopée réside dans la manière dont les exemples sont décrits et dans les méthodes employées. A titre d'exemple on peut citer :

- En induction grammaticale les exemples sont des mots (ou chaîne de caractères). On cherche à caractériser un langage (ou ensemble des mots admissibles) à l'aide d'exemples positifs (appartenant au langage) et négatifs. On suppose que le langage peut être défini ou approché par une grammaire formelle (qui a la capacité de reconnaître ou de rejeter un mot) appartenant à une classe donnée et généralement finie de grammaire. On va donc explorer tout ou partie de cette classe de grammaires pour extraire celle qui, au vu des exemples et au sens d'un certain critère, approche au mieux le langage.
- En analyse d'image on rencontre une situation analogue lorsque, comme dans le cas précédent, on cherche à faire de la discrimination entre deux classes. On réalise une "feature extraction" ou extraction de caractéristiques, qui consiste à tester un nombre variable de caractéristiques qui sont données a priori et peuvent être de tout ordre, puis à retenir celles qui semblent les meilleurs au sens d'un certain critère.

Ainsi, dans les deux situations décrites ci-dessus (et dans bien d'autres), on dispose d'un ensemble $(x_1, \ldots, x_n)$ de différentes valeurs d'un certain critère. Etant dans une problématique de sélection, on s'intéresse aux valeurs extrêmes de celui-ci.

Compte tenu de nos applications, il est légitime de poser que l'ensemble de valeurs $(x_1, \ldots, x_n)$ est la réalisation d'un échantillon aléatoire $(X_1, \ldots, X_n)$. Cependant nous ne pouvons, par construction, poser ici l'indépendance des variables aléatoires de cet échantillon.

Aussi, pour mettre en oeuvre des procédures de sélection appropriées, nous allons établir un certain nombre de résultats sur les valeurs extrêmes. Ces résultats devront être applicables au cas où les variables aléatoires de l'échantillon sont dépendantes.

Notations

Soit $(X_1, \ldots, X_n)$ un échantillon de n variables aléatoires non nécessairement indépendantes.

Posons l'hypothèse a priori H_0 que ces variables aléatoires suivent toutes une même loi de probabilité et soit H_1 l'hypothèse alternative.

Soit $F(x)$ la fonction de répartition de la distribution commune sous H_0.

Dans le cas général on ne présuppose aucune structure de dépendance mutuelle des variables aléatoires de l'échantillon étudié.

On notera [10] l'échantillon réordonné en valeurs croissantes par :

$$(X_{1:n}, \ldots, X_{j:n}, \ldots, X_{n:n})$$

Ainsi :

$$X_{1:n} \leq \cdots \leq X_{j:n} \leq \cdots \leq X_{n:n}$$

Avec cette écriture $X_{r:n}$ est la statistique de la $r^{ième}$ valeur ordonnée de l'échantillon $(X_1, \ldots, X_n)$.

Notons la fonction de répartition de cette variable aléatoire :

$$F_{X_{r:n}}(x) = \mathrm{P}(X_{r:n} \leq x)$$

Avec nos notations nous avons :

$$X_{1:n} = \mathrm{Min}\,(X_1, \ldots, X_n)$$
$$X_{n:n} = \mathrm{Max}\,(X_1, \ldots, X_n)$$

3. Etude dans le cadre d'une hypothèse d'indépendance

3.1. Etude de $X_{r:n}$

Si les variables aléatoires X_j sont indépendantes et identiquement distribuées nous savons que :

$$\begin{aligned} F_{X_{n:n}}(x) &= \mathrm{P}(X_1 \leq x, \ldots, X_n \leq x) \\ &= F(x)^n \end{aligned} \tag{1}$$

de même nous avons :

$$F_{X_{1\ n}}(x) = 1 - (1 - F(x))^n \tag{2}$$

Dans le cas général l'événement $\{X_{r:n} \leq x\}$ signifie qu'au moins r événements de la forme $\{X_j \leq x\}$ sont vérifiés. Ainsi, de par l'expression de la loi binomiale de paramètres n et $F(x)$, nous avons :

$$F_{X_{r\ n}}(x) = \sum_{j=r}^{n} \binom{n}{j} F(x)^j (1 - F(x))^{n-j} \tag{3}$$

Cette dernière expression est peu exploitable dans la recherche de la valeur de $F(x)$ associée à une valeur prédéterminée de $F_{X_{r\ n}}(x)$. Aussi allons nous proposer une approximation de l'expression de $F_{X_{r\ n}}(x)$. Elle pourra utilement se substituer à l'équation (3).

3.2. Approximation de $F_{X_{r\ n}}(x)$

L'expression du développement limité d'ordre 1 des expressions (1) et (2) au voisinage de respectivement $F(x) = 1$ et $F(x) = 0$ permet d'écrire :

$$\begin{array}{lll} \text{Si } F(x) \text{ est au voisinage de } 1 & \text{alors} & 1 - F_{X_{n\ n}}(x) \approx n\,(1 - F(x)) \\ \text{Si } F(x) \text{ est au voisinage de } 0 & \text{alors} & F_{X_{1\ n}}(x) \approx n\,F(x) \end{array} \tag{4}$$

Plus généralement nous allons montrer :

Proposition 1 *Etant donné un échantillon de variables aléatoires* $(X_1, \ldots, X_n)$ *de même fonction de répartition* $F(x)$ *nous avons* $\forall r$ *:*

$$F(x) \text{ au voisinage de } 1 \implies 1 - F_{X_{r\ n}}(x) \approx \binom{n}{r-1} (1 - F(x))^{n-r+1} \tag{5}$$

De même :

$$F(x) \text{ au voisinage de } 0 \implies F_{X_{r\ n}}(x) \approx \binom{n}{r} F(x)^r \tag{6}$$

Pour établir la relation (5) nous allons exhiber le développement limité d'ordre $n - r + 1$ de $F_{X_{r\ n}}(x)$ au voisinage de $F(x) = 1$.

Compte tenu de l'équation (3) nous avons :

$$1 - F_{X_{r\ n}}(x) = \sum_{j=0}^{r-1} \binom{n}{j} F(x)^j (1 - F(x))^{n-j}$$

Chacun des r termes de cette somme est d'ordre infinitésimal $n - j$, par rapport à $1 - F(x)$, et a respectivement $\binom{n}{j}(1 - F(x))^{n-j}$ pour partie principale.

De plus la somme toute entière a l'ordre infinitésimal identique à celui de ses termes d'ordre infinitésimal le plus faible.

Ainsi :

$$F_{X_{r:n}}(x) = 1 - \binom{n}{r-1}(1 - F(x))^{n-r+1} + o\big((F(x) - 1)^{n-r+1}\big)$$

L'écriture de l'approximation (5) est donc justifiée au voisinage de $F(x) = 1$.

L'approximation (6) de la proposition se déduit facilement de la première en utilisant l'égalité suivante, présentée dans [18] :

$$F_{X_{r:n}}(x) = 1 - F_{-X_{n-r+1:n}}(-x)$$

3.3. Approximation d'un seuil de confiance de niveau α

Le résultat précédent a un intérêt pratique direct dans le calcul d'un seuil u_α, solution de $F_{X_{r:n}}(u_\alpha) = 1 - \alpha$.

En effet si la loi de probabilité des X_j est tabulée nous pouvons facilement trouver u_α en exploitant l'approximation :

$$1 - F(u_\alpha) \approx \sqrt[n-r+1]{\frac{1 - F_{X_{r:n}}(u_\alpha)}{\binom{n}{r-1}}} \tag{7}$$

Exemple :

Supposons que $n = 25$ et que les variables aléatoires de l'échantillon sont indépendantes et suivent une loi de probabilité Gaussienne centrée réduite. Nous recherchons une constante u_α telle que la probabilité α que u_α soit dépassée 5 fois soit de $\alpha = 5\%$

$$F_{X_{20:25}}(u_\alpha) = 95\%$$

Par un algorithme d'approximations successives nous trouvons pour solution de l'équation (3) :

$$F(u_\alpha) = 0.890 \quad \text{et} \quad u_\alpha = 1.226$$

Par un calcul, utilisant l'approximation (7), nous trouvons :

$$F(u_\alpha) = 0.919 \quad , \quad u_\alpha = 1.398$$

3.4. Niveau de risque

Dans un test statistique unilatéral basé sur $X_{r:n}$ la quantité α, introduite ci-dessus, représente le niveau de risque du test.

Soit μ_α tel que

$$F(\mu_\alpha) = 1 - \sqrt[n-r+1]{\frac{\alpha}{\binom{n}{r-1}}}$$

En retenant μ_α comme seuil critique de niveau α, plutôt que u_α, on modifie le niveau de risque α fixé a priori. Soit $\hat{\alpha}$ le risque réellement encouru après substitution de u_α par μ_α :

$$\mu_\alpha = u_{\hat{\alpha}}$$

Nous avons :

$$\hat{\alpha} = 1 - \sum_{j=r}^{n} \binom{n}{j} F(\mu_\alpha)^j (1 - F(\mu_\alpha))^{n-j} \tag{8}$$

Dans le cas simple de $X_{n:n}$, où $r = n$, nous avons :

$$\hat{\alpha} = 1 - \left(1 - \frac{\alpha}{n}\right)^n$$

$\hat{\alpha}$ est dans ce cas une fonction décroissante de n telle que $\hat{\alpha} = \alpha$ pour $n = 1$ et :

$$\lim_{n \to \infty} \hat{\alpha} = 1 - e^{-\alpha}$$

Ainsi dans ce cas particulier nous en déduisons :

$$1 - e^{-\alpha} \leq \hat{\alpha} \leq \alpha \tag{9}$$

Dans le cas général, le développement de l'équation (8) nous donne une expression qui se simplifie quand $n \to \infty$ avec r ($n - r = \mathrm{C}^{te}$) :

$$\lim_{n \to \infty} \hat{\alpha} = e^{-a} \sum_{j=0}^{n-r} \frac{a^j}{j!} \tag{10}$$

où :

$$a = \sqrt[n-r+1]{(n-r+1)!\,\alpha}$$

Illustrations

Cas du maximum :

Si $\alpha = 5\,\%$ d'après (9) : $\widehat{\alpha} \in [4.8771\,\%, 5\%]$

Exemples : $n = 10$ $\widehat{\alpha} = 4.8890\,\%$,
$n = 25$ $\widehat{\alpha} = 4.8818\,\%$.

Cas général :

Si $\alpha = 1\,\%$ nous trouvons :

Si :		alors, d'après (10),	
	$r = n$		$\widehat{\alpha} \in [0.995\,\%, 1\,\%]$
	$r = n-1$		$\widehat{\alpha} \in [0.911\,\%, 1\,\%]$
	$r = n-2$		$\widehat{\alpha} \in [0.750\,\%, 1\,\%]$
	$r = n-3$		$\widehat{\alpha} \in [0.580\,\%, 1\,\%]$

La dégradation des bornes de $\widehat{\alpha}$ vient du fait que, pour une valeur fixée de α, $F(\mu_\alpha)$ s'éloigne du voisinage de 1, et donc des conditions d'application des approximations introduites plus haut, quand $n - r$ augmente.

4. Etude sans hypothèse d'indépendance

Sans poser l'hypothèse d'indépendance des variables aléatoires de l'échantillon, nous ne pouvons trouver de résultats génériques permettant d'identifier la loi des valeurs extrêmes.

Cependant nous allons exhiber un système de bornes qui encadreront $F_{X_{r:n}}(x)$. Elles seront utiles, par exemple, dans la mise en oeuvre d'un test statistique.

4.1. Recherche de bornes encadrant $F_{X_{r:n}}(x)$

Le cas du maximum et du minimum est simple à traiter. En effet nous pouvons facilement écrire :

$$\begin{aligned} F(x) \;\geq\; F_{X_{n:n}}(x) &= 1 - \mathrm{P}\Big(\{X_1 > x\} \vee \ldots \vee \{X_n > x\}\Big) \\ &\geq 1 - n\,(1 - F(x)) \end{aligned} \tag{11}$$

Par un raisonnement symétrique on trouve :

$$F(x) \leq F_{X_{1:n}}(x) \leq n\,F(x)$$

Nous allons élargir ces résultats au cas où r est quelconque. Pour cela nous allons établir :

Proposition 2 *Quelle que soit la loi de probabilité des variables aléatoires d'un échantillon équidistribué de loi $F(x)$ et quelle que soit la structure de dépendance mutuelle de ces variables aléatoires, nous avons pour toutes valeurs de x :*

$$F_{X_{r:n}}(x) \geq 1 - \frac{n}{(n-r+1)}(1 - F(x)) \tag{12}$$

et :

$$F_{X_{r:n}}(x) \leq \frac{n}{r} F(x) \tag{13}$$

Soit $\nu_n(x)$ le nombre d'événements de la forme $\{X_j \leq x\}$, $1 \leq j \leq n$, à être réalisés. Nous avons alors :

$$F_{X_{r:n}}(x) = \mathrm{P}(\nu_n(x) \geq r) \tag{14}$$

$$= \sum_{j=r}^{n} \mathrm{P}(\nu_n(x) = j) \tag{15}$$

Le terme $\nu_n(x)$ est une variable aléatoire telle que :

$$\mathrm{E}\Big(\nu_n(x)\Big) = n\, F(x) \tag{16}$$

En effet soit $1_{\{X_j \leq x\}}$ la variable indicatrice associée à l'événement $\{X_j \leq x\}$, $(1 \leq j \leq n)$. Nous avons alors :

$$\nu_n(x) = \sum_{j=1}^{n} 1_{\{X_j \leq x\}}$$

Ainsi :

$$\begin{aligned} \mathrm{E}\Big(\nu_n(x)\Big) &= \sum_{j=1}^{n} \mathrm{E}\Big(1_{\{X_j \leq x\}}\Big) \\ &= \sum_{j=1}^{n} \mathrm{P}\Big(\{X_j \leq x\}\Big) \\ &= n\, F(x) \end{aligned} \tag{17}$$

Or, d'après l'inégalité de Markov [22], $\nu_n(x)$ étant une variable aléatoire positive, pour tout $\lambda > 0$:

$$\mathrm{P}\Big(\nu_n(x) \geq \lambda\, \mathrm{E}\Big(\nu_n(x)\Big)\Big) \leq \frac{1}{\lambda} \tag{18}$$

donc en prenant $\lambda = r\,\mathrm{E}\Big(\nu_n(x)\Big)^{-1}$ et en utilisant (14) et (16) on obtient immédiatement la relation (13) de la proposition.

De façon symétrique on établit l'inéquation (12) en utilisant, comme plus haut, la relation :

$$F_{X_{r,n}}(x) = 1 - F_{-X_{n-r+1:n}}(-x)$$

On notera que les inégalités, que nous venons d'établir, ne sont pertinentes qu'à la condition d'être plus fines que : $0 \leq F_{X_{r:n}}(x) \leq 1$. C'est le cas pour (12) si x est tel que $F(x) \geq (r-1)/n$ et si dans (13) x est tel que $F(x) \leq r/n$

On remarquera également que l'on peut étendre facilement le résultat acquis plus haut au cas plus général où les variables aléatoires de l'échantillon ne sont pas équidistribuées. En effet si $F_j(x)$ est la fonction de répartition de la variable aléatoire X_j on montre, d'après l'équation (17) que :

$$\mathrm{E}\Big(\nu_n(x)\Big) = \sum_{j=1}^{n} F_j(x)$$

et donc par l'inégalité de Markov citée plus haut nous trouvons :

$$F_{X_{r:n}}(x) \geq 1 - \frac{\sum_{j=1}^{n}(1 - F_j(x))}{(n-r+1)}$$

et :

$$F_{X_{r:n}}(x) \leq \frac{\sum_{j=1}^{n} F_j(x)}{r}$$

4.2. Application

Comme nous l'avons fait plus haut nous pouvons exploiter la proposition précédente pour l'approximation d'un seuil u_α tel que $F_{X_{r:n}}(u_\alpha) = 1 - \alpha$. Nous utiliserons pour cela les inégalités :

$$\frac{r}{n} F_{X_{r:n}}(u_\alpha) \leq F(u_\alpha) \leq 1 - \frac{n-r+1}{n}\left(1 - F_{X_{r:n}}(u_\alpha)\right)$$

déduites directement des inéquations (12) et (13).

Exemple :

Si nous reprenons l'exemple 3.3, en abandonnant ici l'hypothèse d'indépendance, nous trouvons :

$$F(u_\alpha) \leq 0.988 \quad \text{et} \quad u_\alpha \leq 2.25$$

En prenant pour u_α la valeur 2.25 on retient un seuil plus élevé que celui prévu dans le cadre de l'indépendance. Si α est un niveau de risque fixé a priori, prendre un seuil plus élevé revient à prendre un risque plus faible.

On perçoit empiriquement, par comparaison avec les résultats trouvés plus haut, les répercussions de l'abandon de l'hypothèse d'indépendance.

4.3. Propriété des bornes de $F_{X_{r:n}}(x)$

Nous allons montrer que pour toute distribution de probabilité $F(x)$, et pour r fixé, il existe un n-uplet $(X_1, \ldots, X_n)$ de variables aléatoires équidistribuées selon $F(x)$, tel que les bornes établies plus haut soient atteintes.

Proposition 3 *Pour tout échantillon de variables aléatoires* $(X_1, \ldots, X_n)$ *de même fonction de répartition* $F(x)$*, la minoration* (12) *(respectivement la majoration* (13)*) peut être effectivement atteinte pour* $F(x) \geq (r-1)/n$ *(respectivement* $F(x) \leq r/n$*).*

Rappelons, comme mentionné plus haut, qu'en dehors de ces intervalles les inéquations (12) et (13) sont sans intérêt puisque moins fines que la condition structurelle : $0 \leq F_{X_{r:n}}(x) \leq 1$.

Nous n'examinerons ici que le cas de la minoration (12). Le raisonnement sur la majoration (13) s'en déduira par symétrie.

Nous allons tout d'abord exhiber une situation où la borne (12) est atteinte pour le cas particulier où les variables aléatoires de l'échantillon sont uniformément distribuées sur l'intervalle $[0, 1]$:

Soit π une variable aléatoire uniformément distribuée sur l'intervalle $[0, 1]$, posons :

$$\begin{aligned} U_j^* &= \frac{j-1}{n} + \frac{1}{n}\pi & \forall j \in \{1, \ldots, r-1\} \\ U_j^* &= \frac{r-1}{n} + \frac{n-r+1}{n}\pi & \forall j \in \{r, \ldots, n\} \end{aligned}$$

Ainsi les variables aléatoires U_j^* sont uniformément distribuées sur les intervalles $[(j-1)/n, j/n]$, quand $j \in \{1, \ldots, r-1\}$, et sur l'intervalle $[(r-1)/n, 1]$, sinon.

Notamment, on montre facilement :

$$\begin{matrix} \forall j \in \{r, \ldots, n\} \\ \forall u \in [(r-1)/n, 1] \end{matrix} \qquad F_{U_j^*}(u) = 1 - \frac{n}{n-r+1}(1-u) \tag{19}$$

Soit $\sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ une permutation aléatoire de l'ensemble $\{1, 2, \ldots, n\}$. Celle-ci peut être définie comme une épreuve consistant au tirage, sans remise, des valeurs de $\{1, 2, \ldots, n\}$. Dans ce cadre σ_j représente le résultat du $j^{ième}$ tirage de cette épreuve. Posons :

$$U_j = U_{\sigma_j}^*$$

On peut facilement montrer que les U_j sont équidistribuées et suivent une loi uniforme sur $[0,1]$. En effet :

$$\forall j \in \{1,\dots,n\} \qquad \mathrm{P}(U_j \le u) = \mathrm{P}\Big(\bigcup_{k=1}^{n}(\{\sigma_j = k\} \cap \{U_k^* \le u\})\Big)$$
$$= \frac{1}{n}\sum_{k=1}^{n}\mathrm{P}(U_k^* \le u)$$

$$\text{Si} \quad u \in \Big[\frac{j-1}{n}, \frac{j}{n}\Big] \quad \text{alors :} \quad \mathrm{P}(U_j \le u) = \frac{1}{n}\Big((j-1) + nu - (j-1)\Big)$$
$$= u$$

$$\text{et si} \quad u \in \Big[\frac{r-1}{n}, 1\Big] \quad \text{alors :} \quad \mathrm{P}(U_j \le u) = \frac{1}{n}\Big((r-1) + (n-r+1)\frac{nu-(r-1)}{n-r+1}\Big)$$
$$= u$$

Par construction nous avons :

$$U_{r:n} = U_r^*$$

Donc $U_{r:n}$ suit une loi uniforme sur l'intervalle $[(r-1)/n, 1]$ et d'après (19) nous pouvons écrire :

$$\forall u \in \Big[\frac{r-1}{n}, 1\Big] \quad F_{U_{r:n}}(u) = 1 - \frac{n}{n-r+1}(1 - F(u))$$

L'échantillon $(U_1, U_2, \dots, U_n)$ a les propriétés recherchées.

Dans le cas général, où la fonction de répartition $F(x)$ est quelconque, nous étendrons l'exemple exhibé ci-dessus par anamorphose en posant :

$$X_j = F^{-1}(U_j)$$

où :

$$F^{-1}(u) = \inf\{\, x \mid F(x) \ge u\}$$

Chacune des variables aléatoires X_j est alors distribuée selon $F(x)$.

De même nous aurons :

$$X_{r:n} = F^{-1}(U_{r:n})$$

et

$$F_{X_{r:n}}(x) = 1 - \frac{n}{n-r+1}(1 - F(x))$$

4.4. Condition nécessaire et suffisante de saturation des bornes de $F_{X_{r:n}}(x)$

Pour compléter nos résultats nous allons montrer que :

Proposition 4 *Une condition nécessaire et suffisante pour que la borne inférieure de* $F_{X_{r:n}}(x)$ *soit atteinte (pour* $F(x) \geq (r-1)/n$*) est que :*

$$P\Big(F\left(X_{r-1:n}\right) \leq \frac{r-1}{n}\Big) = 1 \tag{20}$$

et :

$$F_{X_{r:n}}(x) = F_{X_{r+1:n}}(x) = \cdots = F_{X_{n:n}}(x) \tag{21}$$

Notons tout d'abord que nous pourrions établir, par symétrie, un résultat similaire pour la borne inférieure de $F_{X_{r:n}}(x)$.

$F(x)$ étant fixée, soit :

$$p_j = \mathrm{P}(\nu_n(x) = j) \qquad \forall j \in \{1, \ldots, n\}$$

D'après (16) on peut écrire :

$$\begin{aligned} n\,F(x) &= \sum_{j=0}^{r-1} j\,p_j + \sum_{j=r}^{n} j\,p_j \\ &= (r-1)\sum_{j=0}^{r-1} p_j - \sum_{j=0}^{r-1}(r-1-j)\,p_j + n\sum_{j=r}^{n} p_j - \sum_{j=r}^{n}(n-j)\,p_j \\ &= (r-1)\big(1 - F_{X_{r:n}}(x)\big) + n\,F_{X_{r:n}}(x) - \mathcal{K} \end{aligned}$$

Où :

$$\mathcal{K} = \sum_{j=0}^{r-1}(r-1-j)\,p_j + \sum_{j=r}^{n}(n-j)\,p_j$$

Nous en déduisons :

$$F_{X_{r:n}}(x) = 1 - \frac{n}{(n-r+1)}\big(1 - F(x)\big) + \frac{\mathcal{K}}{(n-r+1)}$$

Ainsi, pour que la minoration (12) soit atteinte, pour tout x tel que $F(x) \geq (r-1)/n$, il faut et il suffit que $\mathcal{K} = 0$ pour ces mêmes valeurs de x.

$\mathcal{K}$ étant une somme de termes positifs et $\sum_{j=1}^{n} p_j = 1$ on établit facilement :

$$\mathcal{K} = 0 \qquad \Longleftrightarrow \qquad p_{r-1} + p_n = 1$$

Cette équation s'écrit, d'après nos notations :

$$\mathrm{P}\Big(\nu_n(x) = r-1\Big) + \mathrm{P}\Big(\nu_n(x) = n\Big) = 1$$

Ainsi, en utilisant (15) on trouve :

$$\mathcal{K} = 0 \iff \begin{cases} F_{X_{r-1:n}}(x) = 1 \\ F_{X_{r:n}}(x) = \cdots = F_{X_{n:n}}(x) \end{cases} \qquad \forall x \text{ tel que : } F(x) \geq \frac{r-1}{n}$$

Or :

$$F_{X_{r-1:n}}(x) = 1 \ \ \forall x \text{ tel que : } F(x) \geq \frac{r-1}{n} \iff \mathrm{P}\Big(F\left(X_{r-1:n}\right) \leq \frac{r-1}{n}\Big) = 1$$

Ainsi avons nous montré l'équivalence énoncée dans la proposition 4.

5. Conclusion

Nous avons opposé plus haut le cas où les variables aléatoires de l'échantillon étaient indépendantes à celui où cette hypothèse ne pouvait être admise. On a pu alors noter que les résultats acquis pour $r = n$, c'est-à-dire pour $X_{n:n} = \mathrm{Max}\,(X_1, \ldots, X_n)$, étaient voisins pour chacune des deux hypothèses. En effet les relations (4) et (11) présentent de fortes similitudes. De plus, dans le cas de l'indépendance, nous avons vu (9) que l'adoption de l'approximation (4) ne modifiait que faiblement les niveaux de risques acceptés a priori.

Ainsi, peut on conclure qu'en règle générale l'hypothèse d'indépendance des variables aléatoires de l'échantillon est peu utile dans la recherche des seuils critiques, de niveau α, de $X_{n:n}$ (respectivement et par symétrie de $X_{1:n}$). Ceci est d'autant plus vrai que le risque α posé a priori est faible.

Dans le cas général (r quelconque), les bornes des inégalités (12) et (13) se différencient nettement des valeurs trouvées dans le cadre de l'hypothèse d'indépendance (cf. : (5) et (6)) et ce d'autant plus que r s'éloigne de n. Ainsi les bornes que nous proposons dans le cas général peuvent être sur-dimensionnées et inadaptées à des situations s'écartant peu de l'hypothèse d'indépendance. Aussi cherchera-t-on, chaque fois que ce sera possible, à expliciter la structure de dépendance des variables aléatoires de l'échantillon et à exhiber des résultats spécifiques et mieux adaptés. On évitera ainsi toute perte inutile de puissance dans la mise en oeuvre de tests statistiques portant sur $X_{r:n}$.

Remerciements : Ce travail a fait l'objet d'une relecture attentive de G. Celeux et a bénéficié des remarques amicales de X. Milhaud. Qu'ils en soient ici remerciés.

Références

[1] BECKMAN, R.J., COOK, R.D. (1983), "Outlier..... s", *Technometrics*, Vol. 25, No. 2, pp. 119-149.

[2] CAPÉRAÀ, P., VAN CUTSEN, B. (1988), *Méthodes et modèles en statistique non paramétrique*, Dunod-Presses de l'Université Laval, Paris.

[3] CEA (1969), *Recherche des valeurs aberrantes*, Dunod, Paris.

[4] DAVID, H.A. (1981), *Order Statistics*, 2^{nd}Ed, Wiley, New York.

[5] DEHEUVELS, P. (1974), "Majoration et minoration presque sûre optimale des éléments de la statistique ordonnée d'un échantillon croissant de variables aléatoires indépendantes", *Rend. Acad. Nazionale dei Lincei Ser*, VIII, Vol. LVI, 707-719.

[6] DEHEUVELS, P. (1981), "Univariate extreme values - Theory and Applications", *43rd Session of the International Statistical Institute (November 30th-December 11th)*, I.P. 7.1, Buenos Aires, Argentina.

[7] DIETTRICH, G.T., MICHALSKI, R.S. (1983), "A comparative Review of Selected Methods for Learning from exemples", *Machine Learning 1*, eds. MICHALSKI, R.S., CARBONELL, J.G., MITCHELL, T.M., Tioga publishing company, Palo Alto, California.

[8] FISHER, R.A., TIPPETT, L.H.C. (1928), "Limiting forms of the frequency distributions of the largest or smallest member of a sample", *Proc. Cambridge Philos. Soc.*, No.24, 180-190.

[9] FOURGEAUD, C., FUCHS, A. (1967), *Statistique*, Dunod, Paris.

[10] GALAMBOS, J., KOTZ, S. (1978), "Characterizations of Probability Distributions", *Lecture Notes in Mathematics*, No. 675, eds. DOLD, A. and ECKMANN, B., Springer Verlag, New-York.

[11] GALAMBOS, J. (1984), "Statistical extremes : Theory and applications, motivation and perspectives", *Statistical Extremes and Applications*, eds. TIAGO DE OLIVEIRA, J., D. Reidel Publishing Company, Dordrecht, Holland.

[12] GALAMBOS, J. (1987), *The Asymptotic Theory of Extreme Order Statistics*, 2^{nd}Ed, Robert E. Krieger Publishing Compagny, Malabar, Florida.

[13] GASCUEL, O., GUENOCHE, A., (1990), "Approches symboliques/numériques en apprentissage", *PRC-GDR Intelligence artificielle, Actes des 3e journées nationales*, eds. BOUCHON-MEUNIER, B.., 91-110, Hermes, Paris.

[14] GRANT, E.L., LEAVENWORTH, R.S. (1972), *Statistical quality control*, Mc Graw Hill.

[15] GRUBBS, F.E. (1950), "Sample criteria for testing outlying observations", *Ann. Math. Statist.*, 21, 27-58.

[16] GUMBEL, E.J., (1958), *Statistics of extremes*, Columbia University Press, New York – London.

[17] KENDALL, M.G., STUART, A. (1969), *The Advanced Theory of Statistics*, Vol.1, Charles Griffing, London.

[18] LEADBETTER, M.R., LINDGREN, G., ROOTZÉN, H. (1983), *Extremes and Related Properties of Random Sequences and Processes.*, Springer-Verlag, New York.

[19] MAHFOUDI, A. (1980), *Une approche unifiée de la détection des observations aberrantes dans le contexte gaussien*, Thèse, USTL, Montpellier, France.

[20] MICLET, L. (1984) *Méthodes structurelles pour la reconnaissance des formes*, Dunod, Paris.

[21] PEIRCE, B. (1852), "Criterion for the Rejection of Doubtful Observations", *Astronomical Journal*, 2, 161-163.

[22] RÉNYI, A. (1966), *Calcul des Probabilités*, Dunod, Paris.

[23] SARHAN, A.E., GREENBERG, B.G. (1962), *Contributions to order statistics*, Wiley, New York.

[24] SMITH, R.L. (1989), "Extreme Value Analysis of Environnemental Time Series : An Application to Trend Detection in Ground-Level Ozone", *Statistical Science*, Vol. 4, No. 4, 367-393.

[25] SNEYERS, R. (1984), "Extremes in Meteorology", *Statistical Extremes and Applications*, eds. TIAGO DE OLIVEIRA, J., D. Reidel Publishing Company, Dordrecht, Holland.

[26] STONE, E.J. (1868), "On rejection of Discordant Observations", *Monthly Notices of the Royal Astronomical Society*, 28, 165-168.

[27] TIAGO DE OLIVEIRA, J. (1984), *Statistical Extremes and Applications*, D. Reidel Publishing Company, Dordrecht, Holland.

[28] WILKS, S.S. (1962), *Mathematical statistics*, Wiley, New York.