Using Extreme Value Theory to test for Outliers

Nathaniel GBENRO
JMS - 2018, Paris, France

28 mai 2018

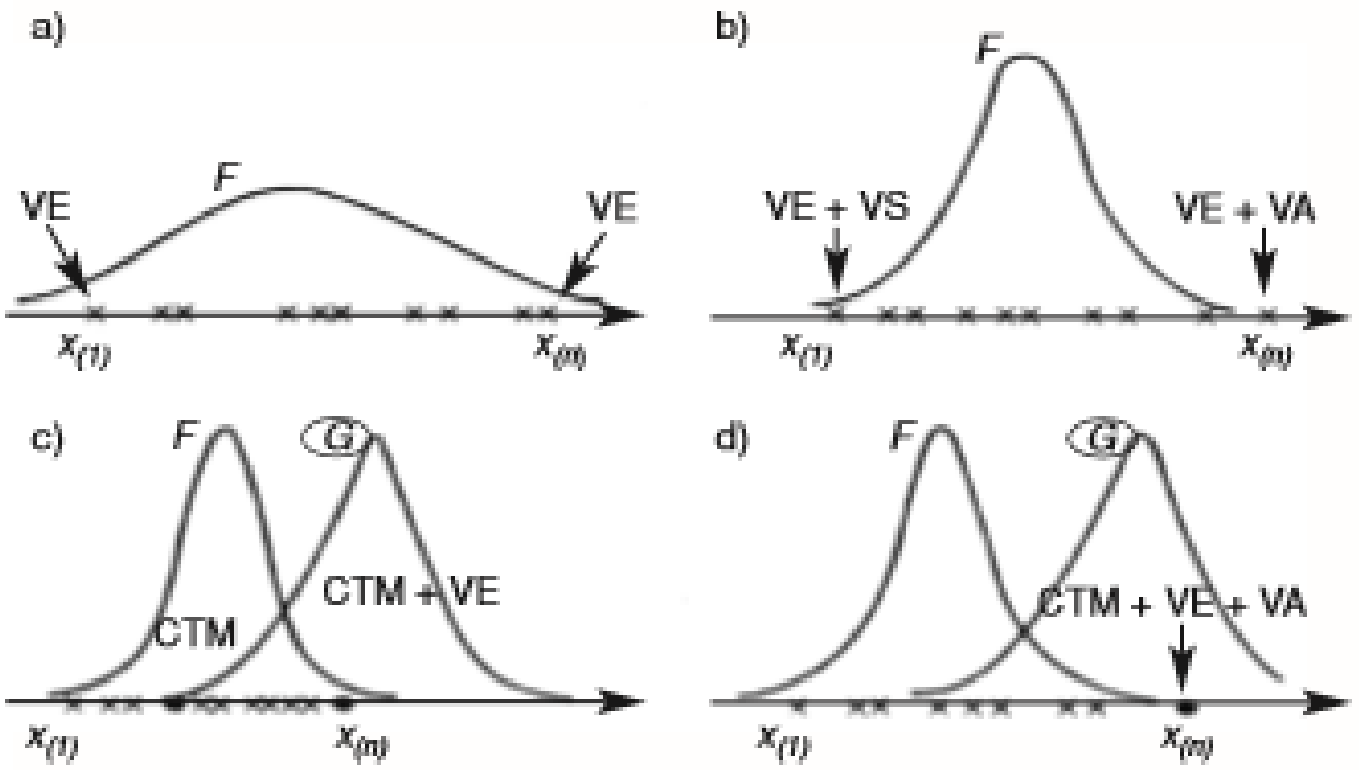
Plan de la présentation

Introduction

- Définition
  Les valeurs extremes sont des observations qui s'écartent du comportement d'ensemble d'un échantillon de données.
- Importance
  - Biais d'analyse (calcul indicateurs, analyse économétrique) Houfi and El Montasser [2009]
  - Modification de politiques (évaluation d'impacts, analyse économique, médicale, etc)

Nuance de définitions

FIGURE – Planchon viviane (2005)
VE : valeurs extremes - VS : valeurs suspectes - VA : valeurs abérrante - CTM : contaminations



Revue de littérature

- statistiques liées excès / volatilité [Dixon, 1950] ;
- statistiques liées amplitude / volatilité [Dixon, 1950] ;
- statistiques liées gap / volatilité, [Thompson, 1935] [Pearson and Sekar, 1936] [Grubbs, 1950] ;
- statistiques liées extrême / position report [Dixon, 1950] ;
- statistiques liées Somme de carré [Dixon, 1950] ;
- statistiques liées aux moments ;
- Shapiro-Wilks W statistic (Shapiro et al, 1968 ; Royston. 1982).

Pearson and Sekar [1936] and Grubbs [1950]

Considérions un échantillon de taille n $x_1, x_2, \ldots, x_n$ généré d'une loi normale de moyenne μ_i et d'écart type σ^2. Test de valeurs abbérantes equivaut à un test d'échantillon de même loi :

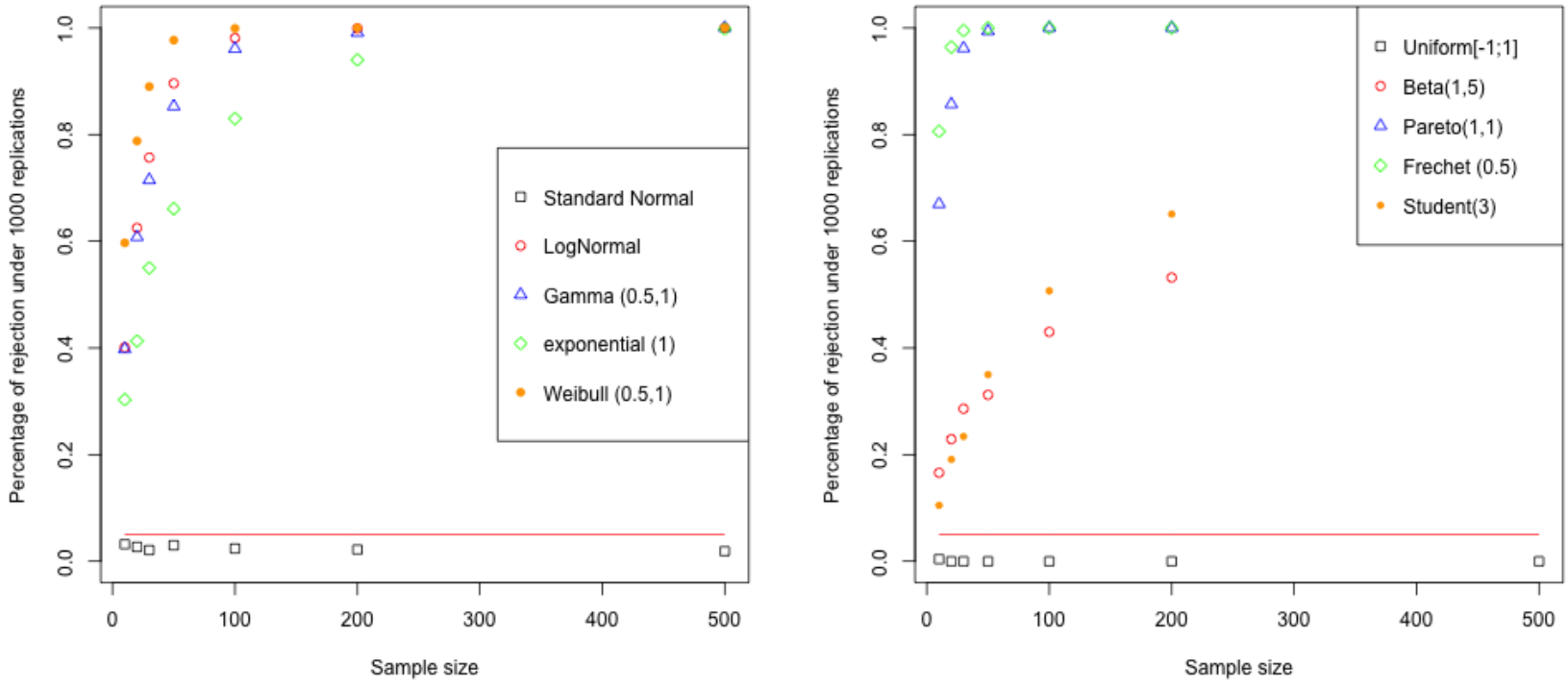
$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_{m-1} = \mu_{m+1} = \ldots = \mu_n = \mu, \mu_m = \mu + d$$

$$H_{a1} = d \neq 0, \qquad H_{a2} = d < 0, \qquad H_{a3} = d > 0.$$

$$\tau = \frac{\delta}{s} \tag{1}$$

$$\delta = x_{(n)} - \bar{x} \quad \textit{and} \quad s = \sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{2}$$

Test de Pearson and Sekar [1936] et Grubbs [1950]



(a) Group 1 (b) Group 2

FIGURE – Application du test de Grubbs sous plusieurs distributions

Généralité

Considérions un échantillon de taille n $x_1, x_2, \ldots, x_n$ généré d'une loi de moyenne μ_i et d'écart type σ^2.

Statistique Naïve de test : $U_+ = \max x_i$

Loi de la statistique :

$$P(U_+ \leq x) = F^n(x)$$

Loi dégénérée :

$$P(U_+ \leq x) \underset{n \to +\infty}{\longrightarrow} \in \{0, 1\} \tag{3}$$

Généralité

EVT indique $\exists a_n$ et $b_n s.t$:

$$P(\frac{x_{(n)} - a_n}{b_n} \leq y) = F^{(n)}(yb_n + a_n) \underset{n \to +\infty}{\longrightarrow} G_\epsilon(y) \tag{4}$$

où

$$G_\epsilon(x) = \begin{cases} exp[-(1 + \epsilon \frac{x-\mu}{\sigma})_+^{-\frac{1}{\epsilon}}] & \text{si } \epsilon \neq 0 \\ exp[-exp(-\frac{x-\mu}{\sigma})] & \text{si } \epsilon = 0 \end{cases} \tag{5}$$

La loi de X $\in$ au domaine d'attraction de G_ϵ.

Généralité

Nous avions l'équivalence suivante :

- X suit un Frechet d'indice ϵ
- $\epsilon^{-1} ln(X)$ suit un Gumbel
- $-X^{-1}$ suit un Weibull

Estimateur Extrem Value Index(EVI)

Il existe plusieurs estimateurs dans la littérature [Embrechts et al., 2013]. Un estimateur général est celui de Dekkers-Einmahl-de Haan. Il est défini par :

$$\epsilon_n^{DEH} = 1 + H_n^{(1)} + \frac{1}{2}\left(\frac{(H_n^{(1)})^2}{H_n^{(2)}} - 1\right)^{-1} \tag{6}$$

où

$$H_n^{(1)} = \frac{1}{k}\sum_{i=0}^{k-1}\left(ln(x_{(n-i)}) - ln(x_{(n-k)})\right)$$

et

$$H_n^{(2)} = \frac{1}{k}\sum_{i=0}^{k-1}\left(ln(x_{(n-i)}) - ln(x_{(n-k)})\right)^2$$

Méthodes d'estimation du seuil

- Mean Excess Function

$$e_{[i]} = \frac{1}{n-i} \sum_{j=0}^{n-i-1} x_{(n-j)} - x_{(i)} \tag{7}$$

- Estimateur de [Pickands III, 1975]

$$\underset{1 \leq j \leq \lfloor \frac{n}{4} \rfloor}{Argmin} \; sup_{0 \leq x < \infty} |\hat{S}_j(x) - \hat{G}_j(x)| \tag{8}$$

- Estimateur de [Neves and Alves, 2004]

$$\underset{2 \leq k \leq n}{Argmin} \frac{1}{k-1} \sum_{i=1}^{k} i^{\delta} (\epsilon_i - \epsilon_k) \tag{9}$$

Hypothesis Test and properties

Notons $x_{(n)}$ (resp. $x_{(1)}$) le max (resp. le min) d'un échantillon de taille n, avec $X_i \sim F_i$. Le test d'hypothèse vise à vérifier si $x_{(n)}$ ou $x_{(1)}$ est un outlier, Nous avons :

$$\begin{cases} H_0 : F_i = F \ \forall \ i \in 1 : n \\ H_a : \exists \text{ m s.t } F_m \neq F \end{cases} . \tag{10}$$

1. $x_{(n)}$ est un outlier, version unilatéral à droite :

$$\begin{cases} H_0 : F_i = F \ \forall \ i \in 1 : n \\ H_a : \exists \text{ m and x* s.t } \ \forall y \geq x^* \ F_m(y) < F(y) \end{cases} \tag{11}$$

2. $x_{(1)}$ est un outlier, version unilatéral à gauche :

$$\begin{cases} H_0 : F_i = F \ \forall \ i \in 1 : n \\ H_a : \exists \text{ m and x* s.t } \ \forall y \leq x^* \ F(y) < F_m(y) \end{cases} \tag{12}$$

Statistique et propriétés

$$\gamma_n^0 = \frac{x_{(n)} - a_n}{b_n} \ pour \ \epsilon = 0 \tag{13}$$

$$\gamma_n^\epsilon = ln \left(\frac{x_{(n)} - a_n}{b_n} \right)^{\frac{1}{\epsilon}} \ pour \ \epsilon > 0 \tag{14}$$

$$\bar{\gamma}_n^\epsilon = ln \left(-\frac{x_{(n)} - a_n}{b_n} \right)^{-\frac{1}{\epsilon}} \ pour \ \epsilon < 0 \tag{15}$$

Test unilateral

Les coéfficients de normalisation sont donnés par [Embrechts et al., 2013] :

Max-Domain	a_n	b_n
Gumbel	$F^{-1}(1-\frac{1}{n})$	$\gamma(a_n)$
Frechet	0	$F^{-1}(1-\frac{1}{n})$
Weibull	x_F	$x_F - F^{-1}(1-\frac{1}{n})$

TABLE – Norming constants

Estimateur des quantiles

1. F $\in$ Domaine d'attraction de Gumbel :

$$\hat{F}^{-1}(1-\frac{1}{n}) = u_n + \gamma(u_n) ln(n_{u_n}) \tag{16}$$

2. F $\in$ Domaine d'attraction de Frechet :

$$\hat{F}^{-1}(1-\frac{1}{n}) = n_{u_n}^{\hat{\epsilon}} u_n \tag{17}$$

3. F $\in$ Domaine d'attraction de Weibull :

$$\hat{F}^{-1}(1-\frac{1}{n}) = x_F - n_{u_n}^{-\hat{\epsilon}}(x_F - u_n) \tag{18}$$

Estimateur du EndPoint

Estimateur de Alves and Neves [2014] :

$$\hat{x_F} = x_{(n)} + \sum_{i=0}^{k-1} a_{ik}(x_{(n-k)} - x_{(n-k-i)}) \tag{19}$$

où

$$a_{ik} = \frac{1}{log2} log(\frac{k+i+1}{k+i})$$

Algorithm Start

1. Use the excess function plot
2. Use DEH estimator.

Algorithm B'($\epsilon = 0$)

1. $j, k \leftarrow n$; $b_1, b_2, iter, tol_0 \leftarrow 0$;t ; m.
   $iter \leftarrow iter + 1$;
2. **While** ($iter \leq m$ and b_1=0) do
3. Compute $x_{(j)} = Max(x_{(1)}, x_{(2)}, \ldots, x_{(j)})$
4. $k_j = j - \min(k_P, k_{RT})$ where k_P is given by (8) and k_{RT} bu (9)
   $tol_0 \leftarrow tol_0 + 1$
5. **While** ($tol_0 \leq t$ and b_2=0) do
   $u_n = x_{(k-k_j)}$
6. Compute γ_j^0 given by (**??**) with
   $n_{u_n} = n - k + k_j$
7. If $\gamma_j^0 > \Lambda_\alpha$ then
   Report "$x_{(j)}$ as outlier"
   $tol_0 \leftarrow 1$
   $b_2 \leftarrow 1$
8. Otherwise (i.e $\gamma_j^0 \leq \Lambda_\alpha$)
   $tol_0 \leftarrow tol_0 + 1$
   $k \leftarrow k - 1$
   End Do ;
   If $tol_0 = t - 1$ then $b_1 = 1$
   $iter \leftarrow iter + 1$;
   $j \leftarrow j - 1$; $k \leftarrow j$;
   $tol_0 \leftarrow 1$;
   $b_2 \leftarrow 0$;
   End Do ;

Algorithm A'($\epsilon \neq 0$)

1. $j, k \leftarrow n$; $b_1, b_2, iter, tol_0 \leftarrow 0$;t ; m.
   $iter \leftarrow iter + 1$;
2. **While** ($iter \leq m$ and b_1=0) do
3. Compute $x_{(j)} = Max(x_{(1)}, x_{(2)}, \ldots, x_{(j)})$
4. $k_j = j - \min(k_P, k_{RT})$ where k_P is given by (8) and k_{RT} bu (9)
   $tol_0 \leftarrow tol_0 + 1$
5. **While** ($tol_0 \leq t$ and b_2=0) do
   $u_n = x_{(k-k_j)}$
6. Compute $\hat{\epsilon_j}$ given by (6) or (**??**) with
   $x_{(k-k_j+1)}, \ldots, x_{(k)}$
7. Compute $\gamma_j^{\hat{\epsilon_j}}$ or $\bar{\gamma}_j^{\hat{\epsilon_j}}$ given by (**??**) and (**??**) with $n_{u_n} = n - k + k_j$ according to the sign of $\hat{\epsilon_j}$
8. If $\gamma_j^{\hat{\epsilon_j}}(resp.\bar{\gamma}_j^{\hat{\epsilon_j}}) > \Lambda_\alpha$ then
   Report "$x_{(j)}$ as outlier"
   $tol_0 \leftarrow 1$
   $b_2 \leftarrow 1$
9. Otherwise (i.e $\gamma_j^{\hat{\epsilon_j}}(resp.\bar{\gamma}_j^{\hat{\epsilon_j}}) \leq \Lambda_\alpha$)
   $tol_0 \leftarrow tol_0 + 1$
   $k \leftarrow k - 1$
   End Do ;
   If $tol_0 = t - 1$ then $b_1 = 1$
   $iter \leftarrow iter + 1$;
   $j \leftarrow j - 1$; $k \leftarrow j$;
   $tol_0 \leftarrow 1$;
   $b_2 \leftarrow 0$;
   End Do ;

Comparaison avec Grubbs

Données simulées

$$Y_i \sim 1_{1 \leq i \leq T-p} N(a, \sigma) + 1_{T-p < i \leq T} N(a+d, \sigma) \qquad (20)$$

- p : nombre de contaminations (0, 1, 2 et 3) ;
- T : taille échantillon (50, 100, 500 et 1000) ;
- a, d, σ : resp. moyennes des PGD (0 et 5), et 1 ;

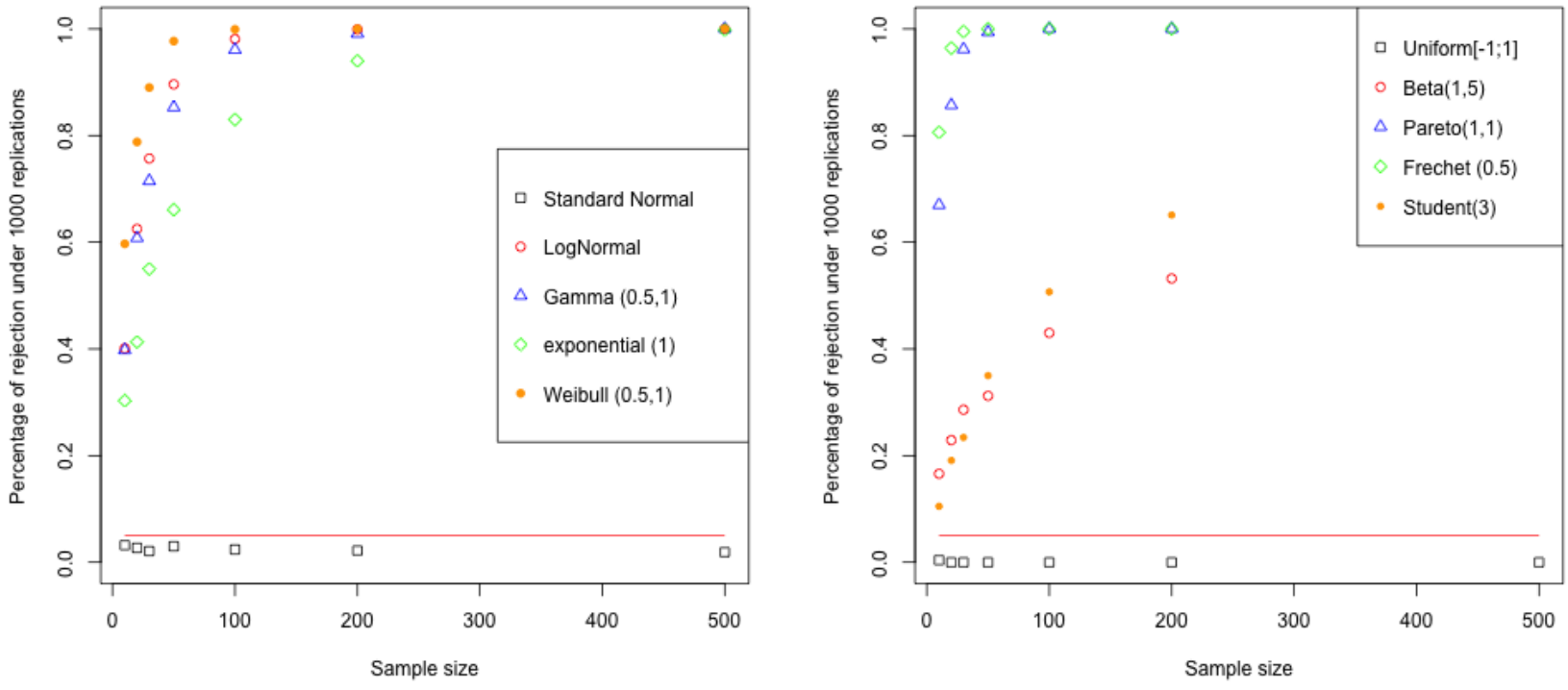
Pour ces paramtrès fixes, 1 000 bootstraps puis taux de rejet de l'hypothèse nulle.

Application 1

N	α (en %)	1-Specificié (p=0)		Sensibilité (p=1)		Sensibilité (p=2)		Sensibilité (p=3)	
		Grubs	EVT	Grubs	EVT	Grubs	EVT	Grubs	EVT
50	1	0.006	0.003	0.81	0.893	0.796	0.991	0.608	0.999
50	5	0.024	0.041	0.919	0.972	0.955	0.997	0.919	1.000
50	10	0.036	0.077	0.939	0.99	0.987	0.999	0.972	1.000
100	1	0.014	0.01	0.807	0.866	0.915	0.975	0.903	0.997
100	5	0.025	0.038	0.923	0.951	0.974	1.000	0.99	0.998
100	10	0.046	0.076	0.935	0.977	0.99	0.999	0.999	1.000
500	1	0.007	0.007	0.768	0.801	0.928	0.951	0.974	0.99
500	5	0.018	0.041	0.859	0.895	0.973	0.987	0.996	1.000
500	10	0.039	0.088	0.903	0.935	0.985	0.994	0.994	1.000
1000	1	0.008	0.011	0.687	0.713	0.913	0.941	0.965	0.98
1000	5	0.019	0.035	0.819	0.865	0.967	0.981	0.992	0.995
1000	10	0.066	0.108	0.875	0.912	0.979	0.993	0.993	0.999

TABLE – Sensibilité et Specificité

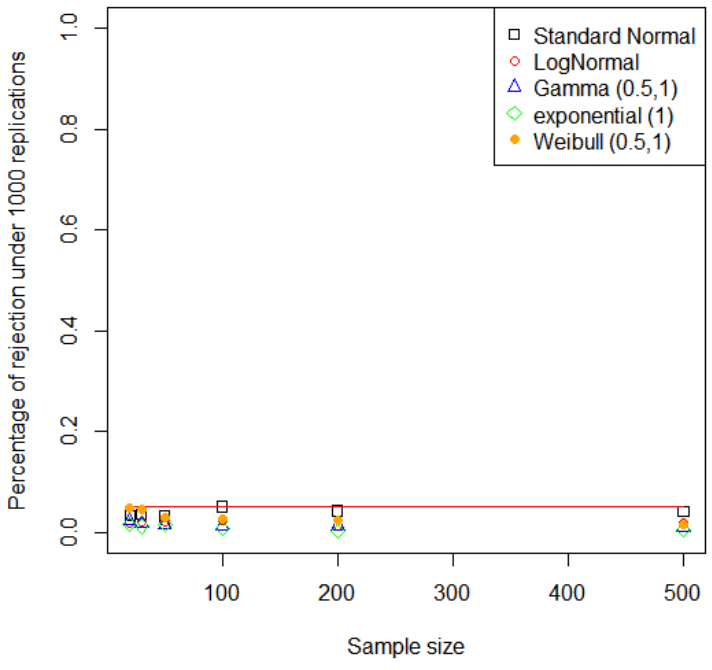
Rappel des résultats du test de Grubbs ...



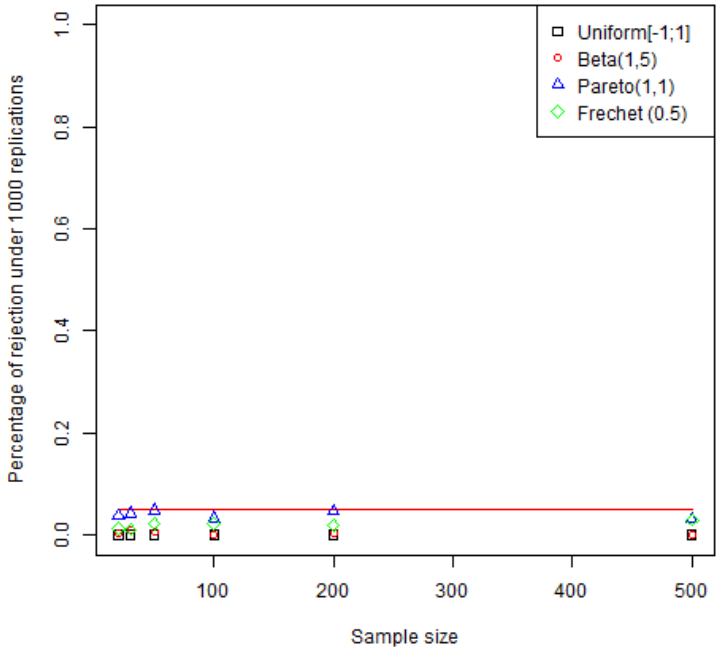
(a) Group 1 (b) Group 2

FIGURE – Test de Grubbs's sous des distributions non normales

... maintenant avec la nouvelle procédure de tes



(a) Gumbel MD.



(b) Frechet and Weibull MD

FIGURE – Application du test basé sur EVT.

Conclusion

- Obj : Identification des valeurs aberrantes
- Echec du test de Grubbs sous la non normalité
- Meilleurs performances sous la normalité et la non normalité
- Resultat robust à la taille de l'echantillon

Bibliographie I

Isabel Fraga Alves and Cláudia Neves. A general estimator for the right endpoint. *arXiv preprint arXiv :1412.3972*, 2014.

Jan Beirlant, Yuri Goegebeur, Jozef Teugels, and Johan Segers. Front matter. *Statistics of Extremes : Theory and Applications*, pages i–xiii, 2004.

Noureddine Benlagha, Michel Grun-Réhomme, et al. Application de la théorie des valeurs extrêmes en assurance automobile. Technical report, ERMES, University Paris 2, 2007.

JACQUES Bernier. Sur l'application des diverses lois limites des valeurs extrêmes au problème des débits de crue. *Houille Blanche*, 11(5), 1956.

Bibliographie II

Peter Burridge and AM Robert Taylor. Additive outlier detection via extreme-value theory. *Journal of Time Series Analysis*, 27 (5) :685–701, 2006.

Frederico Caeiro and M Ivette Gomes. On the bootstrap methodology for the estimation of the tail sample fraction. In *Proceedings of COMPSTAT*, pages 545–552, 2014.

Frederico Caeiro and M Ivette Gomes. Threshold selection in extreme value analysis. *Extreme Value Modeling and Risk Analysis : Methods and Applications*, page 69, 2016.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection : A survey. *ACM computing surveys (CSUR)*, 41(3) : 15, 2009.

Bibliographie III

Denis Cousineau and Sylvain Chartier. Outliers detection and treatment : a review. *International Journal of Psychological Research*, 3(1) :58–67, 2015.

Dipak Dey, Dooti Roy, and Jun Yan. *Extreme Value Modeling and Risk Analysis : Methods and Applications*. CRC Press, 2016.

Wilfred J Dixon. Analysis of extreme values. *The Annals of Mathematical Statistics*, 21(4) :488–506, 1950.

Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling extremal events : for insurance and finance*, volume 33. Springer Science & Business Media, 2013.

B Everitt. *The Cambridge dictionary of statistics/BS Everitt*. Cambridge University Press, Cambridge, UK New York :, 2002.

Bibliographie IV

Ronald Aylmer Fisher and Leonard Henry Caleb Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. Cambridge Univ Press, 1928.

Jesús Gonzalo and José Olmo. Which extreme values are really extreme ? *Journal of Financial Econometrics*, 2(3) :349–369, 2004.

Frank E Grubbs. Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, pages 27–58, 1950.

Mohamed Ali Houfi and Ghassen El Montasser. Effets des points aberrants sur les tests de normalité et de linéarité. applications à la bourse de tokyo. *Revue Congolaise Economie*, 5(21) :2–38, 2009.

Bibliographie V

Malcolm R Leadbetter. Extremes and local dependence in stationary sequences. *Probability Theory and Related Fields*, 65 (2) :291–306, 1983.

Cláudia Neves and MI Fraga Alves. Reiss and thomas, automatic selection of the number of extremes. *Computational statistics & data analysis*, 47(4) :689–704, 2004.

M Nikulin and A Zerbet. Détection des observations aberrantes par des méthodes statistiques. *Revue de statistique appliquée*, 50(3) :25–51, 2002.

Jose Olmo. Extreme value theory filtering techniques for outlier detection. 2009.

ERWIN S Pearson and C Chandra Sekar. The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, 28(3/4) :308–320, 1936.

Bibliographie VI

James Pickands III. Statistical inference using extreme order statistics. *the Annals of Statistics*, pages 119–131, 1975.

Viviane Planchon. Traitement des valeurs aberrantes : concepts actuels et tendances générales. *Biotechnologie, agronomie, société et environnement*, 9(1) :19–34, 2005.

Saad Rais. Outlier detection for the consumer price index. *Statistical Society of Canada Proceedings, Industrial Organization, Finance, and Prices Section, BSMD*, 2008.

N RANGER. R. cleroux j.-m. helbling. *Revue de statistique appliquée*, 38(1) :5–21, 1990.

Carl Scarrott and Anna MacDonald. A review of extreme value threshold es-timation and uncertainty quantification. *REVSTAT–Statistical Journal*, 10(1) :33–60, 2012.

Bibliographie VII

William R Thompson. On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation. *The Annals of Mathematical Statistics*, 6(4) :214–219, 1935.