Les différentes versions du test ET,

test d'adéquation à la queue de distribution.

Myriam Garrido (INRIA Rhône-Alpes).

contexte : $(X_1, \ldots, X_n)$ échantillon de données.

but : estimation de $x \mapsto P(X < x)$ pour $x \geq X_{(n)}$.

$\rightarrow$ bien ajuster une loi à la queue de distribution des données.

Hypothèse : $H_0 : F = F_\theta$ où $F_\theta \in$ DA(Gumbel).

But : Tester l'adéquation de la loi $F_{\hat{\theta}}$ du point de vue des extrêmes.

Principe du test ET : comparaison de deux façons d'estimer un quantile extrême d'ordre $1 - p$ (avec $p \leq 1/n$ i.e. $q_{1-p} \geq X_{(n)}$).

Méthode des excès pour l'estimation des queues de distribution.

Soit u, un réel suffisamment élevé appelé $seuil$.

$$P(X-u \leq y | X > u) = F_u(y) = \frac{F(u+y)-F(u)}{1-F(u)}.$$

Théorème de Pickands (1975) : Si $F \in$ DA(Gumbel),

$$\lim_{u\to\infty} \sup_{0<y<\infty} \left| F_u(y) - \left(1 - \exp\left(-\frac{y}{\sigma(u)}\right)\right)\right| = 0.$$

Calcul de $\widehat{q}_{ET}$, l'estimation ET du quantile.

On choisit $u = X_{(n-m_n)}$, m_n petit devant n.

Les excès : $Y_j = (X_{i_j} - u)$ pour $X_{i_j} > u$.

On estime le paramètre σ par $\widehat{\sigma}_n = (1/m_n) \sum_{j=1}^{m_n} Y_j$.

$$p = P(X > q_{1-p}) \simeq \exp\left(-\frac{q_{1-p} - u}{\sigma(u)}\right) \frac{m_n}{n}.$$

D'où l'estimation du quantile extrême d'ordre $1 - p$

$$\widehat{q}_{ET} = X_{(n-m_n)} + \widehat{\sigma}_n \ln\left(\frac{m_n}{np}\right).$$

Erreur entre le vrai quantile q_{1-p} et $\widehat{q}_{ET}$.

On introduit

$$q_{ET} = F^{\leftarrow}\left(1 - \frac{m_n}{n}\right) + \sigma(u_n) \ln\left(\frac{m_n}{np}\right).$$

Alors on peut décomposer

$$q_{1-p} - \widehat{q}_{ET} = \underbrace{q_{1-p} - q_{ET}}_{\text{erreur d'approximation } \delta} + \underbrace{q_{ET} - \widehat{q}_{ET}}_{\text{erreur d'estimation}}$$

δ dépend de la loi F et de ses paramètres, du seuil u_n, de n, m_n, et p.

Ici, on remplace F qui est inconnue par $F_{\widehat{\theta}}$, la loi sous H_0.

<u>Test basé sur la loi asymptotique de $\widehat{q}_{ET}$.</u>

On peut montrer que

$$\sqrt{m_n}\frac{\widehat{q}_{ET} - q_{ET}}{\sigma_n\sqrt{A_n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \quad \text{où} \quad A_n = 1 + \ln^2\left(\frac{m_n}{np}\right).$$

On en déduit un intervalle de confiance pour q_{ET}.
Une translation de δ donne un intervalle de confiance pour q_{1-p} :

$$IC = \left[\widehat{q}_{ET} + \delta \pm \sqrt{\frac{\widehat{\sigma}_n^2}{m_n}\left(1 + \ln^2\left(\frac{m_n}{np}\right)\right)} z_{1-\alpha/2}\right].$$

<u>Test ET</u> (version 1) : On accepte H_0 si $\widehat{q}_{param} = F^{\leftarrow}_{\widehat{\theta}_n}(1-p) \in IC$.

Pertinence du test basé sur la loi asymptotique de $\widehat{q}_{ET}$.

Problèmes avec la normalité asymptotique de $\sqrt{m_n}\dfrac{\widehat{q}_{ET} - q_{ET}}{\sigma_n\sqrt{A_n}}$.

Calcul du taux de recouvrement de q_{1-p} par son intervalle de confiance

$\rightarrow$ valeurs de m_n à utiliser pour les lois normale, Weibull, gamma et exp.

PB : Test non applicable aux fonctions à décroissance lente à l'infini :

– Weib(α, β) avec $\beta < 1$
– $\Gamma(a, \lambda)$ avec $a < 1$
– Lognormale(μ, σ).

<u>Méthode du bootstrap paramétrique.</u>

Sous H_0, on approxime F par $F_{\widehat{\theta}}$.

On génère N échantillons iid de loi $F_{\widehat{\theta}}$, de taille n.
$\rightarrow N$ estimateurs $\widehat{\delta}^* = \tilde{q}^*_{param} - \widehat{q}^*_{ET}$.
$\rightarrow$ intervalle de confiance empirique $IC_{\delta,BP}$ pour δ.

<u>Test ET-BP</u> (version 2) : On accepte H_0 si $\widehat{q}_{param} - \widehat{q}_{ET} \in IC_{\delta,BP}$.

<u>Test ET-BP simplifié</u> (version 3) : On ne bootstrappe plus q_{param}.
$\rightarrow$ intervalle de confiance $IC_{ET,BP}$ pour q_{ET}.
On accepte H_0 si $\widehat{q}_{ET} \in IC_{ET,BP}$.

Pertinence du test basé sur la méthode du bootstrap paramétrique.

Calcul du taux de recouvrement de q_{1-p} par son intervalle de confiance pour la version (3) simplifiée.

$\rightarrow$ valeurs de m_n (plages très étendues) à utiliser pour toutes les lois.

Attention : $\widehat{q}_{ET}$ est un estimateur de q_{1-p} pour m_n petit.

$\rightarrow$ utiliser de petites valeurs de m_n.
$\rightarrow$ utiliser les valeurs conseillées pour la version 1 du test.

Comparaison entre les deux versions du test ET-BP.

Données de loi gamma(3,1/3). $n = 100,\ m_n = 15,\ \alpha = 0.05,\ p = 0.01,\ N = 500.$

loi	IC simplifié	IC non simplifié	$\widehat{\delta}$
Norm.	[-2.51 , 4.16]	[-1.71 , 3.66]	6.29
	[-2.73 , 3.99]	[-2.11 , 3.23]	
	[-2.37 , 3.89]	[-1.86 , 3.26]	
Exp.	[-12.35 , 14]	[-10.14 , 10.68]	-14.1
	[-12.05 , 14.26]	[-10.42 , 11.62]	
	[-12.44 , 14.46]	[-10.14 , 12.1]	
Gamma	[-5.23 , 6.69]	[-3.82 , 4.89]	1.68
	[-4.6 , 7.15]	[-3.87 , 4.82]	
	[-5.23 , 7.65]	[-4.12 , 6.19]	
Logn.	**[-10.36 , 13.15]**	**[-8.77 , 8.59]**	-6.29
	[-9.73 , 13.58]	**[-9.01 , 8.5]**	
	[-11.34 , 14.47]	**[-8.92 , 8.1]**	
Weibull	**[-3.96 , 5.26]**	**[-2.05 , 3.95]**	3.8
	[-4.47 , 5.81]	**[-2.56 , 4.08]**	
	[-3.82 , 5.83]	[-2.52 , 3.72]	

Données de loi gamma(3,1/3). $n = 100,\ m_n = 15,\ \alpha = 0.05,\ p = 10^{-4},\ N = 500.$

loi	IC simplifié	IC non simplifié	$\widehat{\delta}$
Norm	[-3.17 , 13.76] [-2.26 , 16.53] [-2.58 , 14.95]	[-1.99 , 13.2] [-1.96 , 15.48] [-1.6 , 14.21]	23.35
Exp	[-29.94 , 32.46] [-29.49 , 35.11] [-29.09 , 35.96]	[-25.17 , 30.12] [-26.79 , 30.94] [-29.26 , 31.71]	-31.69
Gamma	[-9.27 , 16.65] [-11.24 , 20.04] [-11.55 , 20.24]	[-11.71 , 19.01] [-8.67 , 18.63] [-9.34 , 17.36]	8.42
Logn	**[-45.51 , 14.29]** **[-44.11 , 18.69]** **[-45.07 , 13.65]**	**[-48.05 , 1.44]** **[-48.14 ,7.22]** **[-49.06 , 6.66]**	-32.41
Weibull	**[-5.13 , 20.52]** **[-5.22 , 19.13]** **[-6.49 , 18.79]**	[-2.95 , 16.79] [-3.11 , 16.65] [-4.08 , 16.68]	16.91

Jeu de données réelles, de taille $n = 24$.

Résultats du test d'Anderson-Darling.

loi	Norm	Exp	Weib	Logn	Gamma
0.25	—	—	—	$\times$	$\times$
0.1	$\times$	—	—	$\times$	$\times$
0.05	$\times$	—	$\times$	$\times$	$\times$
0.025	$\times$	—	$\times$	$\times$	$\times$
0.01	$\times$	—	$\times$	$\times$	$\times$

Résultats du test ET-BP pour $m_n = 6$ excès, un niveau de test de 5%, $p = 1/n = 0.0417$ et $N = 500$.

loi	résultat	$IC_{\delta,BP}$	$\widehat{q}_{ET} - \widehat{q}_{param}$
Norm	rejetée	[- 0.3501 , 0.3898]	0.5459
Exp	rejetée	[- 2.6444 , 2.9690]	-4.4434
Weib	rejetée	[- 0.2248 , 0.3552]	0.5208
Logn	**acceptée**	[- 0.5723 , 0.4839]	0.3155
Gamma	rejetée	[- 0.3888 , 0.4536]	0.4633

Conclusion.

Problème : le niveau n'est pas exact $\to$ le contrôler.

Travail en cours :
- valeurs de m_n à retenir pour le test ET-BP complet.
- étudier la puissance (simulations et théorie).

Suite :
- Test GPD basé sur le bootstrap paramétrique.