MÉTHODES DE DÉTECTION D'UNE RUPTURE DANS DES ÉCHANTILLONS DE PETITE TAILLE SUIVANT DES LOIS EXPONENTIELLES.

Narayanaswamy Balakrishnan[1], Laurent Bordes[2], Christian Paroissin[3] & Jean-Christophe Turlot[4]

[1] *McMaster University, Department of Mathematics and Statistics, Hamilton, Ontario, Canada L8S 4K1, E-mail : bala@mcmaster.ca*
[2] *Université de Pau et des Pays de l'Adour, Laboratoire de Mathématiques et de leurs Applications- UMR CNRS 5142, Avenue de l'Université, 64013 Pau cedex E-mail : laurent.bordes@univ-pau.fr*
[3] *Université de Pau et des Pays de l'Adour, Laboratoire de Mathématiques et de leurs Applications- UMR CNRS 5142, Avenue de l'Université, 64013 Pau cedex, E-mail : cparoiss@univ-pau.fr*
[4] *Université de Pau et des Pays de l'Adour, Laboratoire de Mathématiques et de leurs Applications UMR CNRS 5142, Avenue de l'Université, 64013 Pau cedex, E-mail : turlot@univ-pau.fr*

Résumé. On s'intéresse au problème de détection d'une rupture dans le taux de défaillance observé sur une série courte d'observations. Plus précisément, il s'agit de décider si les instants séparant les défaillances successives sur une série courte de n observations consécutives ont un même taux de défaillance, ou s'il existe un instant $k \in \{1,2,...,n\}$ tel que ce taux, constant jusqu'à une date k inconnue, prenne à partir de $k+1$ jusqu'à n une autre valeur constante correspondant à une augmentation de la fréquence de rupture. On suppose les observations indépendantes. Les tests statistiques que nous proposons sont fondés sur le rapport des moyennes empiriques sous l'hypothèse classique de distributions exponentielles. Ils sont confrontés au test non paramétrique de Wilcoxon-Mann-Whitney qui ne nécessite aucune hypothèse paramétrique sur la loi du taux de défaillance. La loi des statistiques proposées ne dépend pas de la distribution inconnue sous l'hypothèse nulle d'homogénéité des n dates de défaillance, ce qui permet de calculer les valeurs critiques des tests suggérés par la méthode de Monte Carlo pour de petits échantillons. Des études de puissance sont réalisées dans un cadre un peu plus large, en considérant la famille des lois de Weibull.

Mots-clés. Rupture, loi exponentielle, test de Wilcoxon-Mann-Whitney

Abstract. In this paper, we address the problem of deciding if either n consecutive independent failure times have the same failure rate or if there exists some $k \in \{1,...,n\}$ such that the common failure rate of the first k failure times is different from the common failure rate of the last $n-k$ failure times, based on an exponential lifetime distribution. The statistical test we propose is based on the empirical average ratio under the assumption of exponentiality distributed. It is compared to the one based on the Wilcoxon-Mann-Whitney statistic for which no parametric assumption on the underlying distribution is necessary. The proposed statistics are free of the unknown underlying distribution under the null hypothesis of homogeneity of the n failure times which allows computing critical values of the suggested tests by Monte Carlo methods for small sample size.

Keywords. Change-point, Exponential distribution, Wilcoxon-Mann-Whitney test

1 Introduction

Les entreprises industrielles ont besoin de surveiller la qualité et la fiabilité des produits qu'elles fabriquent. On dispose généralement des dates de défaillance de chacune des pièces. Cependant lorsqu'il s'agit de composants hautement fiables, les taux de panne sont extrêmement faibles, aussi dispose t'on d'échantillons de petite taille ou de taille modérée. Le contrôle en ligne et la détection rapide d'une augmentation du taux de panne constitue un vrai problème. Parmi les méthodes proposées, nombreuses sont celles fondées sur une statistique de type maximum. Cependant, nous sommes face à deux problèmes : le premier est la difficulté de prendre une décision sur la base de résultats asymptotiques dont on dispose pour ce type de statistique ; le second est dû au fait que la distribution de ces statistiques dépend de la loi sous-jacente des observations sous l'hypothèse nulle. Dans ce travail, nous proposons deux types de méthodes : la première approche fait l'hypothèse que les intervalles de temps séparant deux défaillances successives sont de nature exponentielle ; la seconde basée sur la statistique de Wilcoxon-Mann-Whitney est de type non paramétrique.

2 Les méthodes proposées

2.1 Le principe méthodologique général

Il est le suivant :

1. On divise l'échantillon en deux sous-échantillons : $(X_1,..,X_k)$et $(X_{k+1},..,X_n)$ pour $k \in \{m,..,n-m\}$
2. On calcule la valeur de la statistique $S_{n,k}$ de comparaison de deux échantillons pour chacun des points de séparation en faisant varier k.
3. On utilise l'ensemble de ces statistiques pour prendre une décision sur l'hypothèse d'une modification dans la durée de vie.

La détection de la date k n'est pas la préoccupation principale dans ce travail. On note $S_{n,k}$ une statistique qui mesure la distance entre les distributions des deux sous-échantillons et l'on suggère l'utilisation de plusieurs statistiques globales fondées sur les $n-2m+1$ statistiques $S_{n,k}$.

- Une statistique de type maximum : $M_n = \max_{m \le k \le n-m} \dfrac{S_{n,k}}{\sqrt{Var(S_{n,k})}}$
- Une statistique de type χ^2 d'un premier genre : $\chi_n^2 = \sum_{k=m}^{n-m} \dfrac{S_{n,k}^2}{Var(S_{n,k})}$
- Une statistique de type χ^2 d'un second genre : $\tilde{\chi}_n^2 = \sum_{k=m}^{n-m} \left(\dfrac{S_{n,k}}{Var(S_{n,k})} \right)^2$
- Une statistique de type quadratique : $Q_n = \underline{S}_n^T \Sigma^{-1} \underline{S}_n$

où $\underline{S}_n^T = (S_{n,m},..,S_{n,n-m})$ et Σ est la matrice des covariances de $\underline{S}_n$.

- Une statistique linéaire : $U_n = \sum_{k=m}^{n-m} w_{k,n} S_{n,k}$

où les poids $w_{m,n},..,w_{n-m,n}$ sont choisis de sorte à minimiser la variance de U_n sous la contrainte $EU_n = 1$.

2.2 Le cas d'une distribution exponentielle

Nous proposons pour la comparaison des moyennes des deux sous-échantillons le ratio des deux

statistiques suivantes :

$$\frac{T_k}{k} = \frac{\sum X_i}{k} \qquad \text{et} \qquad \frac{T_n - T_k}{n-k} = \frac{\sum X_i}{n-k}$$

Pour supprimer le paramètre inconnu sous l'hypothèse nulle, on considère le ratio de ces statistiques :

$$S_{n,k}^{(1)} = \frac{n-k-1}{k} \frac{T_k}{T_n - T_k}$$

dont l'espérance vaut 1 sous l'hypothèse nulle et dont on explicite la variance et la matrice des covariances, ainsi que les poids de la méthode linéaire.

Ces cinq statistiques définies dans le cadre de distributions exponentielles (EAR ou Exponential Average Ratio) sont notées $M_n^{(1)}, \chi_n^{2(1)}, \tilde{\chi}_n^{2(1)}$, $Q_n^{(1)}$ et $U_n^{(1)}$.

Remarque : on montre que pour tout $k \in \{m,...,n-m\}$ avec $m \geq 2$, on a :

$$\text{var}\, S_{n,k}^{(1)} = \frac{n-1}{k(n-k-2)}$$

et, pour tout couple $(k,k') \in \{m,...,n-m\}^2$ tel que $k' > k$ et $m \geq 3$, on a :

$$\text{cov}\left(S_{n,k}^{(1)}, S_{n,k'}^{(1)}\right) = \frac{n-1}{k'(n-k-2)}$$

La valeur de m doit donc être supérieure ou égale à 3 pour cette famille de statistiques.

2.3 L'approche non paramétrique

Elle est fondée sur la statistique de Wilcoxon Mann et Whitney (WMW) :

$$S_{n,k}^{(2)} = \sum_{i=1}^{k} \sum_{j=k+1}^{n} 1_{\{X_j < X_i\}}$$

Les expressions de l'espérance et de la variance de $S_{n,k}^{(2)}$ sont bien connues, l'expression de la covariance de $S_{n,k}^{(2)}$ peut être explicitée. On peut en déduire l'expression des cinq statistiques précédentes dans le cadre non paramétrique ; elles sont notées $M_n^{(2)}, \chi_n^{2(2)}, \tilde{\chi}_n^{2(2)}$, $Q_n^{(2)}$ et $U_n^{(2)}$.

3 Les résultats

Des simulations par Monte-Carlo ont été réalisées pour identifier les valeurs critiques de ces tests, puis pour comparer leur puissance. On a considéré quatre types de modèles :

- Cas 1 : Les variables aléatoires du premier sous-échantillon suivent une loi exponentielle de moyenne $\mu_1 = \{2,3,5\}$ alors que celles du second échantillon sont distribuées avec une moyenne $\mu_2 = 1$
- Cas 2 : Les variables aléatoires du premier sous-échantillon suivent une loi exponentielle de moyenne $\mu_1 = \{2,3,5\}$ alors que celles du second échantillon sont distribuées selon une Weibull avec pour paramètre de forme 1.2 et pour paramètre d'échelle la valeur donnant $\mu_2 = 1$.
- Cas 3 : Les variables aléatoires du premier sous-échantillon suivent une loi de Weibull avec paramètre de forme 1.2 et pour paramètres d'échelle les valeurs donnant $\mu_1 = \{2,3,5\}$, alors que pour le second échantillon le paramètre de forme est inchangé, mais le paramètre d'échelle fixé pour que $\mu_2 = 1$.
- Cas 4 : Les variables aléatoires du premier sous-échantillon suivent une loi de Weibull avec 3 comme paramètre de forme et pour paramètres d'échelle les valeurs donnant $\mu_1 = \{2,3,5\}$. Les variables aléatoires du second échantillon sont distribuées comme le mélange de deux lois de Weibull avec comme proportions de mélange les valeurs $(0.15, 0.85)$: la première composante du mélange conserve la distribution avant rupture, alors que la seconde conserve la même valeur de forme et un paramètre d'échelle fixé de sorte à obtenir une moyenne $\mu_2 = 1$.

Deux indicateurs de performance ont été calculés : $\bar{\hat{\pi}}$ qui est un indicateur de puissance moyenne et $\bar{\hat{l}}$ qui est un indicateur de perte moyenne. Soit $\pi(k, F_\theta, G_\eta, S_n, \phi)$ la fonction puissance du test dont les arguments sont : k la date de rupture ; F_θ, G_η les distributions inter-pannes avant et après la rupture selon les modèles décrits ci-dessus ; S_n la statistique de comparaison des sous-échantillons et ϕ la statistique choisie (§2.1). La puissance moyenne s'écrit :

$$\bar{\hat{\pi}}(F_\theta, G_\eta, S_n, \phi) = \frac{1}{n-2m+1} \sum_{k=m}^{n-m} \hat{\pi}(k, F_\theta, G_\eta, S_n, \phi)$$

Pour définir le second indicateur, on détermine à k, F_θ, G_η, S_n fixés, laquelle des statistiques ϕ est la meilleure globalement : $\phi_k^* = \arg\max_\phi \hat{\pi}(k, F_\theta, G_\eta, S_n, \phi)$, ce qui nous permet de définir la perte de puissance pour les différentes statistiques globales :

$$\hat{l}(k, F_\theta, G_\eta, S_n, \phi) = \hat{\pi}(k, F_\theta, G_\eta, S_n, \phi_k^*) - \hat{\pi}(k, F_\theta, G_\eta, S_n, \phi_k)$$

Comme la date de rupture est inconnue, on considère la perte moyenne selon l'ensemble des valeurs de k :

$$\bar{\hat{l}}(F_\theta, G_\eta, S_n, \phi) = \frac{1}{n-2m+1} \sum_{k=m}^{n-m} \hat{l}(k, F_\theta, G_\eta, S_n, \phi)$$

La taille de l'échantillon est fixée à $n = 20$, la date de rupture en un point $k \in \{3,...,17\}$, le risque de première espèce est fixé à $\alpha = 0.05$. Les valeurs critiques des tests ainsi que les puissances associées aux différentes alternatives considérées ont été calculées sur la base de 50.000 simulations.

Les résultats des calculs de la puissance des différents tests sont confinés dans le tableau 1. D'autres statistiques non paramétriques ont été confrontées à EAR. Nous discutons ces résultats sur une application issue de données publiées sur des pannes dans le système d'air conditionné de dix avions Boeing.

Bibliographie

[1] Pettitt A (1980). Some results on estimating a change-point using non-parametric type statistics. Appl. Statist. 11: 261-272.

[2] Proschan F (1963). Theoretical explanation of observed decreasing failure rate. Technometrics 5: 375-383.

[3] Sen A, Srivastava M (1975). On tests for detecting change in mean. Ann Statist 3: 98-108.

[4] Page E (1957). On problems in which a change in parameters occurs at an unknown point. Biometrika 44: 248-252.

[5] Csörgö M, Horvath L (1997). Limit theorems in change-point analysis. Wiley, New York.

[6] Arnold BC, Balakrishnan N, Nagajara AN (2008). A first course in order statistics. SIAM, Philadelphia.

[7] Lombard F (1987). Rank tests for change-point problems. Biometrika 74: 615-624.

μ_1/μ_2	Case	Indicator	EAR statistic					WMW statistic				
			M_n	χ^2_n	$\tilde{\chi}^2_n$	Q_n	U_n	M_n	χ^2_n	$\tilde{\chi}^2_n$	Q_n	U_n
2	1	$\hat{\pi}$	0.270	0.293	**0.295**	0.257	0.223	0.223	**0.228**	0.225	0.217	0.224
		$\hat{\ell}$	0.043	0.019	0.018	0.056	0.089	0.019	0.015	0.018	0.026	0.018
	2	$\hat{\pi}$	0.188	0.238	**0.246**	0.149	0.167	0.279	**0.283**	0.279	0.269	0.278
		$\hat{\ell}$	0.073	0.023	0.015	0.112	0.094	0.025	0.021	0.025	0.035	0.026
	3	$\hat{\pi}$	0.225	0.261	**0.268**	0.203	0.186	0.202	**0.205**	0.203	0.196	0.202
		$\hat{\ell}$	0.055	0.019	0.012	0.077	0.094	0.015	0.011	0.014	0.021	0.015
	4	$\hat{\pi}$	0.020	0.080	**0.098**	0.001	0.070	**0.690**	0.683	0.686	0.672	0.672
		$\hat{\ell}$	0.097	0.037	0.018	0.116	0.046	0.074	0.081	0.078	0.093	0.092
3	1	$\hat{\pi}$	0.525	**0.541**	0.538	0.503	0.416	0.394	**0.396**	0.392	0.378	0.388
		$\hat{\ell}$	0.055	0.039	0.043	0.077	0.164	0.038	0.037	0.041	0.055	0.045
	2	$\hat{\pi}$	0.477	0.522	**0.524**	0.413	0.379	**0.490**	0.488	0.485	0.469	0.479
		$\hat{\ell}$	0.086	0.041	0.039	0.150	0.184	0.050	0.052	0.055	0.071	0.061
	3	$\hat{\pi}$	0.495	0.523	**0.525**	0.462	0.383	0.376	**0.377**	0.373	0.360	0.370
		$\hat{\ell}$	0.066	0.038	0.037	0.100	0.178	0.035	0.034	0.038	0.051	0.041
	4	$\hat{\pi}$	0.472	0.528	0.527	0.289	**0.536**	**0.790**	0.780	0.788	0.778	0.771
		$\hat{\ell}$	0.145	0.088	0.090	0.328	0.081	0.075	0.084	0.076	0.086	0.093
5	1	$\hat{\pi}$	**0.809**	0.797	0.780	0.791	0.691	**0.604**	0.598	0.596	0.581	0.587
		$\hat{\ell}$	0.042	0.053	0.071	0.060	0.159	0.065	0.071	0.072	0.088	0.081
	2	$\hat{\pi}$	**0.814**	0.808	0.788	0.770	0.703	**0.702**	0.694	0.696	0.683	0.683
		$\hat{\ell}$	0.051	0.057	0.077	0.096	0.162	0.078	0.086	0.083	0.097	0.097
	3	$\hat{\pi}$	**0.804**	0.795	0.778	0.778	0.682	**0.593**	0.588	0.587	0.570	0.577
		$\hat{\ell}$	0.048	0.057	0.074	0.074	0.170	0.064	0.069	0.070	0.086	0.079
	4	$\hat{\pi}$	**0.854**	0.835	0.796	0.816	0.830	**0.805**	0.795	0.804	0.795	0.787
		$\hat{\ell}$	0.012	0.031	0.071	0.050	0.037	0.073	0.083	0.074	0.083	0.092

Tableau 1 : calcul des puissances pour les différentes alternatives 1-4 et les différentes valeurs de $\mu_1/\mu_2 \in \{2,3,5\}$.