Chi-squared tests in Relability and Survival analysis

Mikhail Nikulin

IBM, University Victor Segalen, Bordeaux, France

jointly with

V.Bagdonavicius and R.Tahir

Abstract

We give chi-squared goodness-of fit tests for **parametric models** including **various regression models** such as **accelerated failure time**, **proportional hazards**, **generalized proportional hazards**, **frailty models**, **linear transformation models**, **models with cross-effects of survival functions**. Choice of random grouping intervals as data functions is considered.

Résumé

On propose des test d'ajustement du type chi carré pour des modèls paramétriques y compris les modèls de la vie accélerée, de Cox, de risques proportionnels généralisés, de transformations et d'autres. Le choix de limites des intervales de groupement comme fonctions des données est considéré.

1 Introduction

The famous **chi-squared test of Pearson** is well know, but different modifications of this test are not so well known. The same time we may say that the **theory of chi-squared type tests** is developping very actively till now. It is enough to say some names to see how much efforts were done to find good statistics to construct good chi-squared type tests.

a) K.Pearson, Fisher, Cramer, Feller, Cochran, Yates, Lancaster,Yarnold, Larntz, Lawal, Tumanian, E.Pearson, Neyman, Placket, Lindly, Rao, Bolshev, Kambhampati, Chibisov, Kruopis, Von Mises, Mann, Wald, Cohen, Roy, Chernoff, Lehmann, Dahiya, Gurland, Moore, Watson, Vessereau, Patnaik, Park, Holst, Dzaparidze, Nikulin, Dudley, (**1900-1975**) ;

b) Bhapkar, Stephens, Greenwood, Drost, Kallenberg, Rao, Robson, McCulloch, McCullagh, Spruill, D'Agostino, Hsuan, Cressie, Lockhart, Moore, Nikulin, Read, Mirvaliev, Aguirre, Oller, Heckman, Voinov, Nair, Pya, Chichagov, Charmes, LeCam, Singh, Lemeshko, Chimitova, Postovalov, Balakrishnan, Van der Vaart, Rayner, Best, Hollander, Pena, (**1970-2000**) ;

c) Habib, Thomas, Kim, Solev, Singh, Zhang, Karagrigoriou, Mattheou, Anderson, Boero, Eubank, Koehler, Gan, Doss, Ducharm, Akritas, Hjort, Bagdonavicius, Nikulin, etc., (**1995-2010**).

Today one can find easy different interesting applications of this theory in **reliability, survival analysis, demography, insurance, sport**,. There are constructed many different **modifications of standard statistic of Pearson** in dependence on situations. We shall discuss a little the situation with the **chi-squared type tests** today.

1) Lancaster, H.O. (**1969**) . The Chi-Squared Distributions. J.Wiley : NY.
2) Drost, F. (**1988**). Asymptotics for Generalized Chi-Squared Goodness-of-fit Tests. Center for Mathematics and Computer Sciences. CWI Tracts,vol.48, Amsterdam.
3) Greenwood, P.E. and Nikulin, M.S. (**1996**). A Guide to chi-squared testing. J.Wiley : NY.
4) Van der Vaart, A. (**1998**). Asymptotic statistics, Cambridge University Press, UK.
5) Bagdonavicius,V., Kruopis, J., Nikulin,M. ((**2010**). Nonparametric tests for complete data. ISTE/WILEY.
6) Bagdonavicius,V., Kruopis, J., Nikulin,M. ((**2010**). Nonparametric tests for censored data. ISTE/WILEY.

Notations and Models

Suppose that n independent objects are observed. Let us consider the hypothesis $\mathbf{H_0}$ stating that the **survival functions**

$$S_i(t) = \mathbf{P}(T_i > t), \quad t > 0, \quad (i = 1, ..., n)$$

of these objects are **specified functions** $S_i(t, \theta)$ of time t and **finite-dimensional parameter** $\theta \in \Theta \subset \mathbf{R}^s$.

For any $t > 0$ the value $S_i(t, \theta)$ of the i-th survival function is the probability for i-th object **not to fail** to time t.

We suppose that the survival distributions of all n subjects are absolutely continuous and we denote $f_i(t, \theta)$ its densities.

The hypothesis $\mathbf{H_0}$ can be also formulated in terms of **the hazard functions**

$$\lambda_i(t,\theta) = \frac{f_i(t,\theta)}{S_i(t,\theta)} = -\frac{S_i'(t,\theta)}{S_i(t,\theta)} = \lim_{h\downarrow 0}\frac{1}{h}\mathbf{P}_\theta\{t < T_i \le t+h | T_i > t\}$$

or **the cumulative hazard functions**

$$\Lambda_i(t,\theta) = \int_0^t \lambda_i(u,\theta)du, \quad S_i(t,\theta) = e^{-\Lambda_i(t,\theta)}.$$

The value $\lambda_i(t,\theta)$ of the hazard rate characterizes the **failure risk just after the time** t for the i-th object survived to this time.

Let us consider examples of such hypotheses :

1) **Composite hypothesis**

$$H_0 : S_i(t) = S_0(t, \theta),$$

where $S_0(t, \theta)$ is the **survival function** of a specified parametric class of survival distributions, for example, **a class** of **Weibull, Generalized Weibull, Inverse-Gaussian, Loglogistic, Lognormal, Gompertz, Birnbaum-Saunders distributions.** The distribution is the same for any object.

2) Parametric **Accelerated Failure Time** (**AFT**) model :

$$S_i(t) = S_0(\int_0^t e^{-\beta^T z_i(u)} du, \gamma), \quad z_i(\cdot) \in E,$$

where $z_i(t) = (z_{i1}(t), ..., z_{im}(t))^T$ is a vector of **possibly time dependent covariates**, $\beta = (\beta_1, \cdots, \beta_m)^T$ is a vector of **unknown regression parameters**, the **baseline survival function** S_0 does not depend on z_i and belongs to a **specified class** of survival functions :

$$S_0(t, \gamma), \quad \gamma = (\gamma_1, ..., \gamma_q)^T \in G \subset R^q.$$

The set E is called the set of all **possible or admissibles covariates**, (**stress**).

If explanatory variables are **constant over time**, $z(\cdot) = z$, then the parametric **AFT** model on E_1 has the form

$$S_i(t) = S(t|z_i) = S_0\left(e^{-\beta^T z_i}\, t, \gamma\right), z_i \in E_1 \subseteq E.$$

Under this model the logarithm of the failure time T under stress $z \in E_1$ may be written as

$$\ln\{T\} = \beta^T z + \epsilon, \quad \epsilon \sim S(t) = S_0(\ln t), \quad z \in E_1.$$

If ϵ is normally distributed random variable then **AFT** model is **standard multiple linear regression model**.

3) Parametric **proportional hazards** (**Cox**) model :

$$\lambda_i(t) = e^{\beta^T z_i(t)} \lambda_0(t, \gamma), \quad z_i(\cdot) \in E,$$

where $\lambda_0(t, \gamma)$ is a **baseline hazard function** of **specified parametric form**.

4) Parametric **generalized proportional hazards** (**GPH**) models (including parametric **frailty and linear transformations models**), with $\theta = (\beta^T, \gamma^T, \nu^T)^T$:

$$h(\Lambda_i(t), \gamma) = \int_0^t e^{\beta^T z_i(u)} \lambda_0(u, \nu) du,$$

where the function $h(x, \gamma)$ and the **baseline hazard function** $\lambda_0(t, \nu)$ have **specified** forms. In particular, if

$$h(x, \gamma) = \frac{(1+x)^\gamma - 1}{\gamma}, \quad h(x, \gamma) = \frac{1 - e^{-\gamma x}}{\gamma}, \quad h(x, \gamma) = x + \frac{\gamma x^2}{2},$$

we have **generalizations of positive stable, gamma, and inverse gaussian frailty models** with **explanatory variables**.

V.Bagdonavicius, M.Nikulin (2002) *Accelerated Life Models : modeling and estimation*. Chapman&Hall/CRC.

5) Models with **cross effects of survival functions** :
$\lambda_i(t) = g(z_i, \beta, \gamma, \Lambda_0(t, \nu))$, where the **baseline cumulative hazard** $\Lambda_0(t, \nu)$ has a specified form and the function $g(z_i, \beta, \gamma, x)$ has one of the following forms :

$$g(z_i, \beta, \gamma, x) = e^{\beta^T z_i} \left[1 + e^{(\beta+\gamma)^T z_i} x\right]^{e^{-\gamma^T z_i} - 1}$$

$$g(z_i, \beta, \gamma, x) = \frac{e^{\beta^T z_i + x e^{\gamma^T z_i}}}{1 + e^{(\beta+\gamma)^T z_i} \left[e^{x e^{\gamma^T z_i}} - 1\right]}.$$

In complete data case several well known **modifications of the classical chi-squared tests** studied by Lehman and Chernoff (1954), Chibisov (1971), Nikulin (1973), Dzaparidze and Nikulin (1974), Rao and Robson (1974), Moore (1975), Le Cam et al (1983), Drost (1988), Van der Vaart (1998), Voinov (2003),.. which are based on the differences between two estimators of the probabilities to fall into grouping intervals : one estimator is based on **the empirical distribution function**, other – on **the maximum likelihood estimators of unknown parameters of the tested model** using initial non-grouped data. Goodness-of-fit tests for linear regression have been studied by Mukantseva [16], Pierce and Kopecky [18], Loynes [15], Koul [13].

Habib and Thomas [10], Hollander and Peña [12] , Solev (1999) did considered natural modifications of the **Pearson** statistic to the case of censored data. These tests are also based on the differences between two estimators of the probabilities to fall into grouping intervals : one is based on the **Kaplan-Meier estimator** of the cumulative distribution function, other – on **the maximum likelihood estimators** of unknown parameters of the tested model using initial non-grouped censored data.

The idea of comparing observed and expected numbers of failures in time intervals is dew to Akritas [2] and was developed by Hjort [11]. In censored data case Hjort [11] considered goodness-of-fit for parametric Cox models, Gray and Pierce [8], Akritas and Torbeyns [3] – for linear regression models.

We give **chi-squared type goodness-of fit tests for general hypothesis $\mathbf{H_0}$**. Choice of random grouping intervals as data functions is considered.

Right censored data

Suppose that n objects are observed and the value of the failure time T_i of the i-th objects is known if $T_i \leq C_i$; here $C_i \geq 0$ is a random variable or a constant known as **the right censoring time.** Otherwise, the value of the random variable T_i is **unknown**, but it is known that this value is **greater** than the known value of the censoring time C_i.

Example.

Surgical operation are done at consecutive times $t_1, t_2, ..., t_n$. The purpose of the statistical analysis is at time $t > \max t_i$ to conclude the survival time of patient after the operation. If the i-th patient dies at time t then his **life duration** from the operation to death (i.e. the failure T_i) **is known**. If he is **alive** at time t then the failure time is **unknown**, but it is known that the failure time is greater than $C_i = t - t_i$. So **the data are right censored**.

We say that we observe **random right censoring** when $C_1, C_2, ..., C_n, T_1, T_2, ..., T_n$ are **independent** random variables.

We say that we observe **independent right censoring** if for all $i = 1, 2, ..., n$, for any t such that $\mathbf{P}\{X_i > t\}$, and for almost $s \in [0, t]$

$$\lambda_{ci}(t) := \lim_{h \downarrow 0} \mathbf{P}\{T_i \in [s, s+h) | X_i \geq s\} = \lambda(s).$$

So independent right censoring signifies that for almost all $s \in [0, t]$ the probability of failure just after time s **for non-failed and non-censored objects** to time s coincides with the probability of failure just after time s when there is **no censoring**. So **independent censoring** has **no influence** on the survival of objects.

2 Data, MLE and the chi-squared test

We shall give chi-squared tests for the hypothesis **H_0** from **right censored data** (right-censored sample)

$$(X_1, \delta_1, z_1(s)), \cdots, (X_n, \delta_n, z_n(s)), \quad 0 \leq s \leq \tau,$$

where

$$X_i = T_i \wedge C_i, \quad \delta_i = \mathbf{1}_{\{T_i \leq C_i\}},$$

T_i being **failure times**, C_i -**censoring times**, and $z_i = z_i(t) = (z_{i1}(t), ..., z_{im}(t))^T$ – the **possibly time depending covariates** (this third component is absent in the case of the first example), the random variable δ_i is **the indicator** of the event $\{T_i \leq C_i\}$, and τ is the **finite time of the experiment**. Denote by $\bar{G}_i$ **the survival function** of the censoring time C_i. Denote by $g_i(t)$ the density of $\bar{G}_i$.

Right censored samples are often presented by different way, which is well adapted to the problems considered in survival analysis and reliability. In particular we shall use these processes for construction the chi-squared type tests for mentioned above models. Set

$$N_i(t) = \mathbf{1}_{\{X_i \le t, \delta_i = 1\}}, \quad Y_i(t) = \mathbf{1}_{\{X_i \ge t\}},$$

$$N(t) = \sum_{i=1}^{n} N_i(t), \quad Y(t) = \sum_{i=1}^{n} Y_i(t).$$

The process $N(t)$ shows for any $t > 0$ the **number of observed failures** in the interval $[0, t]$.

The process $Y(t)$ shows the number of objects which are **at risk** (**not failed, not censored**) **just prior** the time $t < \tau$.

It is evident that two above data presentations are equivalent, but there is a very important advantage of data presentation in terms of introduced stochastic processes N_i and Y_i since they show the dynamics of failures and censoring history throughout the experiment and

$$\{N_i(s), Y_i(s), 0 \leq s \leq t, \quad i = 1, 2, ..., n\}$$

to time t. **The notion of history** is formalized by introduction of the notion of **filtration** generated by stochastic processes N_i and Y_i.

Suppose that

1) the processes N_i, Y_i, z_i are observed **finite time** τ ;

2) survival distribution of all n objects given z_i are absolutely continuous with the survival function S_i and the hazard rates λ_i ;

3) censoring is **non informative** and **the multiplicative intensities model** is verified : **the compensators** of the counting processes N_i with respect to **the history of the observed processes** are $\int_0^t Y_i \lambda_i ds$, where $\lambda_i(t, \theta) = \lambda(t, \theta, z_i)$.

In this case for non-informative and independent censoring we may give the following expressions for the **likelihood function**

$$L(\theta) = \prod_{i=1}^{n} f_i^{\delta_i}(X_i, \theta) S_i^{1-\delta_i}(X_i, \theta)\, \bar{G}_i^{\delta_i}(X_i)\, g_i^{1-\delta_i}(X_i), \quad \theta \in \Theta.$$

Since the problem is to estimate the parameter θ, we can skip the multipliers which do not depend on this parameter. So under non-informative censoring the likelihood function has the next form :

$$L(\theta) = \prod_{i=1}^{n} f^{\delta_i}(X_i, \theta) S^{1-\delta_i}(X_i, \theta), \quad \theta \in \Theta.$$

Using the relation $f_i(t, \theta) = \lambda_i(t, \theta) S_i(t, \theta)$ the **likelihood function** can be written

$$L(\theta) = \prod_{i=1}^{n} \lambda_i^{\delta_i}(X_i, \theta) S_i(X_i, \theta), \quad \theta \in \Theta.$$

The estimator $\hat{\theta}$, maximizing the likelihood function $L(\theta), \theta \in \Theta$, is called **maximum likelihood estimator**. We denote $\hat{\theta}$ the ML estimator of θ under $\mathbf{H_0}$. We remind that $\theta = (\beta^T, \gamma^T)^T$.

Chi-squared type tests construction

Divide the interval $[0, \tau]$ into k smaller intervals $I_j = (a_{j-1},\ a_j]$ with $a_0 = 0$, $a_k = \tau$, and denote by

$$U_j = N(a_j) - N(a_{j-1})$$

the number of observed failures in the j-th interval, $j = 1, 2...., k$.
What is **"expected" number of observed failures** in the interval I_j under the $\mathbf{H_0}$?
Set $\lambda_i(t, \theta) = \lambda(t, \theta, z_i)$ the hazard function of T_i under z_i. Under $\mathbf{H_0}$ and regularity conditions the equality

$$\mathbf{E}N_i(t) = \mathbf{E}\int_0^t \lambda_i(u, \theta)Y_i(u)du$$

holds, since

$$N_i(t) - \int_0^t \lambda_i(u,\theta) Y_i(u) du$$

is **a martingale of counting process** $N_i(t)$. It suggests that we can **"expect"** to observe

$$e_j = \sum_{i=1}^{n} \int_{I_j} \lambda_i(u,\hat{\theta}) Y_i(u) du$$

failures in the interval I_j ; here $\hat{\theta}$ is the **ML** estimator of θ under $\mathbf{H_0}$. So a test can be based on the statistic

$$Z = (Z_1, ..., Z_k)^T, \quad Z_j = \frac{1}{\sqrt{n}}(U_j - e_j), \quad j = 1, ..., k,$$

of differences between the numbers of **observed** and **"expected" failures** in the intervals $I_1, ..., I_k$.

3 Asymptotic properties of the test statistics

To investigate the properties of the statistic Z we need properties of the stochastic process

$$H_n(t) = \frac{1}{\sqrt{n}}(N(t) - \sum_{i=1}^{n} \int_0^t \lambda_i(u, \hat{\theta}) Y_i(u) du).$$

To obtain these properties we use the properties of the ML estimators which are well known.

Conditions A :

$$\hat{\theta} \xrightarrow{P} \theta_0, \quad \frac{1}{\sqrt{n}} \dot{\ell}(\theta_0) \xrightarrow{d} N_m(0, i(\theta_0)),$$

$$-\frac{1}{n} \ddot{\ell}(\theta_0) \xrightarrow{P} i(\theta_0),$$

and

$$\sqrt{n}(\hat{\theta} - \theta_0) = i^{-1}(\theta_0) \, \frac{1}{\sqrt{n}} \dot{\ell}(\theta_0) + o_P(1),$$

where

$$\dot{\ell}(\theta) = \sum_{i=1}^{n} \int_0^\tau \frac{\partial}{\partial \theta} \ln \lambda_i(u, \theta) \, \{dN_i(u) - Y_i(u) \lambda_i(u, \theta) du\}$$

is the **score function**.

The conditions A for **consistency and asymptotic normality** of the **ML** estimator $\hat{\theta}$ hold, for example, if conditions VI.1.1 given in Andersen et al [4] hold. Set

$$S^{(0)}(t,\theta) = \sum_{i=1}^{n} Y_i(t)\lambda_i(t,\theta), \quad S^{(1)}(t,\theta) = \sum_{i=1}^{n} Y_i(t)\frac{\partial \ln \lambda_i(t,\theta)}{\partial \theta}\lambda_i(t,\theta),$$

$$S^{(2)}(t,\theta) = \sum_{i=1}^{n} Y_i(t)\frac{\partial^2 \ln \lambda_i(t,\theta)}{\partial \theta^2}\lambda_i(t,\theta).$$

For more details see Andersen, Borgan, Gill and Keiding (1993), Van der Vaart (1998).

Conditions B : There exist a neighborhood θ of θ_0 and continuous bounded on $\theta \times [0, \tau]$ functions

$$s^{(0)}(t, \theta), \quad s^{(1)}(t, \theta) = \frac{\partial s^{(0)}(t, \theta)}{\partial \theta}, \quad s^{(2)}(t, \theta) = \frac{\partial^2 s^{(0)}(t, \theta)}{\partial \theta^2},$$

such that for $j = 0, 1, 2$

$$\sup_{t \in [0,\tau], \theta \in \theta} || \frac{1}{n} S^{(j)}(t, \theta) - s^{(j)}(t, \theta) || \xrightarrow{P} 0 \quad \text{as} \quad n \to \infty.$$

The conditions B imply that **uniformly** for $t \in [0, \tau]$

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^t \lambda_i(u,\theta_0)Y_i(u)du \xrightarrow{P} A(t), \quad \frac{1}{n}\sum_{i=1}^{n}\int_0^t \dot{\lambda}_i(u,\theta_0)Y_i(u)du \xrightarrow{P} C(t),$$

where A and C are finite quantities.

Theorem 3.1 *Under the* **conditions A and B** *the following convergence holds :*

$$H_n \xrightarrow{d} H \quad on\ D[0,\tau];$$

here H *is* **zero mean Gaussian martingale** *such that for all* $0 \leq s \leq t$

$$cov(H(s), H(t)) = A(s) - C^T(s) I^{-1}(\theta_0) C(t),$$

$D[0,\tau]$ *is the* **space of cadlag fonctions with Skorokhod metric**.

For $i = 1, ..., s;\ j, j' = 1, ..., k$ set

$$V_j = H(a_j) - H(a_{j-1}), \quad v_{jj'} = cov(V_j, V_{j'}), \quad A_j = A(a_j) - A(a_{j-1}),$$

$$C_{ij} = C_i(a_j) - C_i(a_{j-1}), C_j = (C_{1j}, ..., C_{s,j})^T, \quad V = [v_{jj'}]_{k\times k}, \quad C = [C_{ij}]_{s\times k},$$

and denote by A a $k \times k$ **diagonal matrix** with the diagonal elements $A_1, ..., A_k$.

Theorem 3.2 *Under conditions A and B*

$$Z \xrightarrow{d} Y \sim N_k(\mathbf{0}, V) \quad as\ n \to \infty,$$

where

$$V = A - C^T i^{-1}(\theta_0) C.$$

Set

$$G = i - CA^{-1}C^T.$$

The formula

$$V^- = A^{-1} + A^{-1}C^T G^- CA^{-1}$$

implies that we need **to inverse** only diagonal $k \times k$ matrix A and find the **general inverse** of the $s \times s$ matrix G.

Theorem 3.3 *Under conditions A and B the following estimators of* A_j, C_j, $I(\theta_0)$ *and* V *are consistent :*

$$\hat{A}_j = U_j/n, \quad \hat{C}_j = \frac{1}{n}\sum_{i=1}^{n}\int_{I_j} \frac{\partial}{\partial\theta} \ln \lambda_i(u,\hat{\theta}) dN_i(u),$$

and

$$\hat{i} = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \frac{\partial \ln \lambda_i(u,\hat{\theta})}{\partial\theta} \left(\frac{\partial \ln \lambda_i(u,\hat{\theta})}{\partial\theta}\right)^T dN_i(u), \quad \hat{V} = \hat{A} - \hat{C}^T \hat{i}^{-1} \hat{C}.$$

4 Chi-squared goodness-of-fit test

The theorems 1 and 2 imply that a test for the hypothesis $\mathbf{H_0}$ can be based on the statistic

$$Y^2 = Z^T \hat{V}^- Z,$$

where

$$\hat{V}^- = \hat{A}^{-1} + \hat{A}^{-1}\hat{C}^T\hat{G}^-\hat{C}\hat{A}^{-1}, \quad \hat{G} = \hat{I} - \hat{C}\hat{A}^{-1}\hat{C}^T.$$

This statistic can be written in the form

$$Y^2 = \sum_{j=1}^{k} \frac{(U_j - e_j)^2}{U_j} + Q,$$

where

$$U_j = \sum_{i:X_i \in I_j} \delta_i, \quad e_j = \sum_{i:X_i > a_{j-1}} [\Lambda_0(a_j \wedge X_i); \hat{\gamma}) - \Lambda_0(a_{j-1}; \hat{\gamma})],$$

$$Q = W^T \hat{G}^- W,$$

$$W = \hat{C}\hat{A}^{-1} z = (W_1, ..., W_s)^T, \quad \hat{G} = [\hat{g}_{ll'}]_{s \times s},$$

$$\hat{g}_{ll'} = \hat{i}_{ll'} - \sum_{j=1}^{k} \hat{C}_{lj} \hat{C}_{l'j} \hat{A}_j^{-1},$$

$$\hat{C}_j = \frac{1}{n} \sum_{i: X_i \in I_j} \delta_i \frac{\partial}{\partial \theta} \ln \lambda_i(X_i, \hat{\theta}),$$

$$\hat{i} = \frac{1}{n} \sum_{i=1}^{n} \delta_i \frac{\partial \ln \lambda_i(X_i, \hat{\theta})}{\partial \theta} \left(\frac{\partial \ln \lambda_i(X_i, \hat{\theta})}{\partial \theta} \right)^T,$$

$$W_l = \sum_{j=1}^{k} \hat{C}_{lj} \hat{A}_j^{-1} Z_j, \quad l, l' = 1, ..., s.$$

The **limit distribution of the statistic** X^2 is **chi-square** with $r = rank(V^-) = Tr(V^- V)$ **degrees of freedom**. If the matrix G is **non-degenerate** then $r = k$.

Test for the hypothesis $\mathbf{H_0}$: the hypothesis is rejected with approximate significance level α if $Y^2 > \chi^2_\alpha(r)$.
Note that for all examples 1-5 the rank $r = k - 1$ in the case of exponential, Weibull, Gompertz **baseline distribution**, and also for distribution with hyperbolic baseline hazard function, $r = k$ for lognormal, loglogistic baseline distribution. In the case of composite hypothesis of Example 1 and exponential distribution the second quadratic form in the expression of the test statistic is equal to zero.

5 Choice of random grouping intervals

Let us consider **choice of the limits of grouping intervals** as random data functions. Define

$$E_k = \sum_{i=1}^{n} \int_0^{\tau} \lambda_i(u, \hat{\theta}) Y_i(u) du = \sum_{i=1}^{n} \Lambda_i(X_i, \hat{\theta}), \quad E_j = \frac{j}{k} E_k, \quad j = 1, ..., k.$$

So we seek $\hat{a}_j$ to have equal numbers of expected failures (not necessary integer) in all intervals. So $\hat{a}_j$ verify the equalities

$$g(\hat{a}_j) = E_j, \quad g(a) = \sum_{i=1}^{n} \int_0^{a} \lambda_i(t, \hat{\theta}) Y_i(u) du.$$

Denote by $X_{(1)} \leq ... \leq X_{(n)}$ the ordered sample from $X_1, ..., X_n$.
Note that the function

$$g(a) = \sum_{i=1}^{n} \Lambda_i(X_i \wedge a, \hat{\theta}) = \sum_{i=1}^{n} \left[\sum_{l=i}^{n} \Lambda_{(l)}(a, \hat{\theta}) + \sum_{l=1}^{i-1} \Lambda_{(l)}(X_{(l)}, \hat{\theta}) \right] \mathbf{1}_{[X_{(i-1)}, X_{(i)}]}(a)$$

is continuous and increasing on $[0, \tau]$; here $X_{(0)} = 0$, and we understand $\sum_{l=1}^{0} c_l = 0$. Set

$$b_i = \sum_{l=i+1}^{n} \Lambda_{(l)}(X_{(i)}, \hat{\theta}) + \sum_{l=1}^{i} \Lambda_{(l)}(X_{(l)}, \hat{\theta}).$$

If $E_j \in [b_{i-1}, b_i]$ then $\hat{a}_j$ is the **unique solution** of the equation

$$\sum_{l=i}^{n} \Lambda_{(l)}(\hat{a}_j, \hat{\theta}) + \sum_{l=1}^{i-1} \Lambda_{(l)}(X_{(l)}, \hat{\theta}) = E_j,$$

We have $0 < \hat{a}_1 < \hat{a}_2 ... < \hat{a}_k = \tau$. Under this choice of the intervals $e_j = E_k/k$ for any j.

Theorem 5.1 *Under conditions A and B and random choice of the endpoints of grouping intervals the limit distribution of the statistic* Y^2 *is* **chi-square** *with* r *degrees of freedom.*

Références

[1] Aalen, O. (1978). Nonparametric inference for the family of counting processes, *Ann. Statist.*, **6**, 701–726.

[2] Akritas, M.G. (1988). Pearson-type goodness-of-fit tests : the univariate case, *J.Amer.Statist.Assoc.*, **83**, 222–230.

[3] Akritas, M.G., Torbeyns, A.F., (1997). Pearson-type goodness-of-fit tests for regression, *Can.J.Statist.*, **25**, 359–374.

[4] Andersen P.K., Borgan O., Gill R.D. and Keiding N. (1993). *Statistical Models Based on Counting Processes*, Springer-Verlag : New York.

[5] Bagdonavičius, V. and Nikulin, M. (2002).*Accelerated Life Models.* Chapman and Hall/CRC, Boca Raton.

[6] Bagdonavičius V., Kruopis, J., Nikulin, M. (2011). *Non-parametric tests for censored data.* ISTE&WILEY :London

[7] Billingsley, P. (1968). *Convergence of Probability Measures*, Wiley : New York.

[8] Gray, R.J., and Pierce, D.A. (1985). Goodness of fit tests for censored survival data, *Ann. Statist.*, **13**, 552–563.

[9] Greenwood, P., Nikulin, M.S. (1996). *A Guide to Chi-squared Testing*, Wiley : New York.

[10] Habib, M.G., Thomas, D.R. (1986). Chi-squared goodness-of-fit tests for randomly censored data, *The Annals of Statistics*, **14**, 759–765.

[11] Hjort, N.L. (1990). Goodness of fit tests in models for life history data based on cumulative hazard rates, *The Annals of Statistics*, **18**, 1221–1258.

[12] Hollander, M., Pena, E. (1992). Chi-square goodness-of-fit test for randomly censored data, *JASA*, **417**, 458–463.

[13] Koul, H. (1984). Tests of goodness-of-fit in linear regression, *Colloq. Math. Soc. Janos Bolyai*, **45**, 279–315.

[14] LeCam, L., Mahan,C., Singh, A. (1983). An extension of a Theorem of H.Chernoff and E.L.Lehmann. In : *Recent advances in statistics*, Academic Press, Orlando, 303–332.

[15] Loynes, R.M. (1980). The empirical distribution function of residuals from generalized regression, *Ann. Statist.*, **8**, 285–298.

[16] Mukantseva,L .A. (1977). Testing normality in one-dimensional and multi-dimensional linear regression, *Theory Probab. Appl.*, **22**, 591–602.

[17] Nikulin, M.S. (1973). Chi-square test for continuous distribution with shift and scale parameters, *Theory of Probability and its Application*, **19**, 559–568.

[18] Pierce, D., and Kopecky, K. (1979). Testing goodness of fit for the distribution of errors in regression models, *Biometrika*, textbf66, 1–6.

[19] Rao, K.C., and Robson, D.S. (1974). A chi-square statistic for goodness-of-fit tests within the exponential family, *Communications in Statistics*, textbf3, 1139–1153.