ASYMPTOTICS OF THE GENERALIZED STATISTICS FOR TESTING THE HYPOTHESIS UNDER RANDOM CENSORING

A.A. Abdushukurov & N.S. Nurmuhamedova
National University of Uzbekistan, Tashkent

ABSTRACT
In this article in a model of random censorship from both sides, we consider research two statistics for testing the composite hypotheses, which have in limit a chi-square distribution with appropriate degrees of freedom. First one is the generalized chi-square statistics, for the construction of which we use the power estimate distributions of function(d.f.). The second statistics is twice the logarithm of the likelihood ratio statistics (LRS) of model of random censorship from both sides. Both of these statistics can be used to construct an asymptotic tests of chi-square type for the composite hypotheses.

Keywords: *chi-square statistics, likelihood ratio statistics ,maximum likelihood estimate, random censoring.*

1. INTRODUCTION

Let $\{(X_i, Y_{1i}, Y_{2i}), i \geq 1\}$ sequence of independent and identically distributed (i.i.d) random vectors with mutually independent components and marginal d.f.-s F and G_k for random variables (r.v.) X_i and Y_{ki}, $k = 1,2;$ $i \geq 1$, respectively. Consider the case, when r.v. X_i subject to random censoring from both sides by variables Y_{ki} . On n-th stage of the experiment we observe the sample of size n: $S^{(n)} = \{(Z_i, \delta_{0i}, \delta_{1i}, \delta_{2i}), 1 \leq i \leq n\}$, where $Z_i = Y_{1i} \vee (X_i \wedge Y_{2i})$ $\delta_{0i} = I(X_i \wedge Y_{2i} < Y_{1i})$, $\delta_{1i} = I(Y_{1i} \leq X_i \leq Y_{2i})$, $\delta_{2i} = I(Y_{1i} \leq Y_{2i} < X_i)$. Here for numbers a and b: $a \wedge b = \min(a,b)$, $a \vee b = \max(a,b)$. In a sample $S^{(n)}$ r.v. X_i observed only when $\delta_{1i} = 1$. In this model of random censorship from the both sides of the problem consists in estimating of conditional survival function $1 - F_\tau(x) = P(X_i \geq x / X_i \geq \tau), x \geq \tau,$ from sample $S^{(n)}$ under nuisance pair (G_1, G_2) for specific number τ . In this article, we consider the problem of testing the composite hypothesis $H_0: F \in \mathrm{F}$, where $\mathrm{F} = \{F(\cdot;\theta), \theta \in \Theta\}$ - family of distribution depends on unknown parameter $\theta = (\theta_1, ..., \theta_s) \in \Theta$ and Θ - an open set in R^s. Consider two statistical tests for verify H_0 with a limit of chi-square distribution.

2. GENERALIZED CHI-SQUARE STATISTICS

For build statistics of chi-square test we consider the nonparametric estimates of $1 - F_\tau(x)$ from [1]:

$$1 - F_{\tau n}(x) = \left[\frac{q_n(x)}{q_n(\tau)}\right]^{R_n(x;\tau)}, x \geq \tau ,$$

where

$$R_n(x;\tau) = \Lambda_n(x;\tau)(L_n(x;\tau))^{-1}, \quad \Lambda_n(x;\tau) = -\int_{[\tau;x]} (q_n(u))^{-1} dH_n^{(1)}(u),$$

$$L_n(x;\tau) = -\int_{[\tau;x]} (q_n(u))^{-1} dq_n(u), \quad q_n(x) = G_{1n}(x) - H_n(x) + \frac{1}{n},$$

$$H_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(Z_i < x) = H_n^{(0)}(x) + H_n^{(1)}(x) + H_n^{(2)}(x),$$

$$H_n^{(m)}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{mi} I(Z_i < x), \quad m = 0,1,2,$$

$$G_{1n}(x)=\exp\left\{-\int\limits_{[x;\infty)}\left(H_n(u)+\frac{1}{n}\right)^{-1}dH_n^{(0)}(u)\right\},\ x\ge\tau.$$

In order to construct test statistics we introduce the conditions

(C1) Let d.f.-s F and G_1 continuous and the numbers τ, T such that, $\tau<T$ and

$$\inf_{\tau\le x\le T}P(Y_{1i}\le x\le X_i\wedge Y_{2i})>0;$$

(C2) Support $N_F=\{x:0<F(x;\theta)<1\}$ independent on θ;

(C3) There is a density $f(x;\theta)$ with d.f. $F(x;\theta)$, it has continuous derivatives: $\frac{\partial^2 f(x;\theta)}{\partial\theta_i\partial\theta_j}$, $i,j=\overline{1,s}$ and $\int\limits_{-\infty}^{+\infty}\left|\frac{\partial^2 f(x;\theta)}{\partial\theta_i\partial\theta_j}\right|dx<\infty;\ i,j=\overline{1,s}$;

(C4) Information matrix of Fisher $\mathrm{I}(\theta)=\left\|I_{ij}\right\|_{i,j=\overline{1,s}}$ is positive definite and continuous by θ, where

$$I_{ij}(\theta)=-\int\limits_{-\infty}^{\infty}\frac{\partial^2\log f(x;\theta)}{\partial\theta_i\partial\theta_j}(G_1(x)-G_2(x))dF(x;\theta)-\int\limits_{-\infty}^{\infty}\frac{\partial^2\log F(x;\theta)}{\partial\theta_i\partial\theta_j}F(x;\theta)dG_1(x)-$$
$$-\int\limits_{-\infty}^{\infty}\frac{\partial^2\log(1-F(x;\theta))}{\partial\theta_i\partial\theta_j}(1-F(x;\theta))dG_2(x);$$

(C5) There is a maximum likelihood estimate (MLE) $\hat\theta_n=(\hat\theta_{1n},...,\hat\theta_{sn})$, for parameter $\theta=(\theta_1,...,\theta_s)$, obtained by solving the system of equations

$$\frac{\partial\log p_n(\theta)}{\partial\theta_i}=0,\ i=1,...,s,$$

where $p_n(\theta)=\prod\limits_{i=1}^{n}(F(Z_i;\theta))^{\delta_{0i}}(f(Z_i;\theta))^{\delta_{1i}}(1-F(Z_i;\theta))^{\delta_{2i}}$ - the truncated likelihood function of the model. Moreover, the MLE $\hat\theta_n$ can be represented by $n\to\infty$

$$n^{1/2}(\hat\theta_n-\theta)=\mathrm{I}^{-1}(\theta)A_n(\theta)+o_p(1),$$

where $A_n(\theta)=n^{-1/2}\frac{\partial\log p_n(\theta)}{\partial\theta}$ is normalized contribution function.

Remark 1. It should be noted that the conditions (C2) and (C3) ensure the existence of second-order derivatives of functions $\log f(x;\theta)$, $\log F(x;\theta)$ and $\log(1-F(x;\theta))$. Indeed, for all $i,j=\overline{1,s}$ and $(x;\theta)\in N_F\times\Theta$:

$$\left|\frac{\partial^2 \log f(x;\theta)}{\partial\theta_i\partial\theta_j}\right| \le \frac{1}{f(x;\theta)}\left|\frac{\partial^2 f(x;\theta)}{\partial\theta_i\partial\theta_j}\right| + \frac{1}{f^2(x;\theta)}\left|\frac{\partial f(x;\theta)}{\partial\theta_i}\right|\left|\frac{\partial f(x;\theta)}{\partial\theta_j}\right|,$$

$$\left|\frac{\partial^2 \log F(x;\theta)}{\partial\theta_i\partial\theta_j}\right| \le \frac{1}{F(x;\theta)}\int_{-\infty}^{+\infty}\left|\frac{\partial^2 f(x;\theta)}{\partial\theta_i\partial\theta_j}\right|dx + \frac{1}{F^2(x;\theta)}\int_{-\infty}^{+\infty}\left|\frac{\partial f(x;\theta)}{\partial\theta_i}\right|dx\int_{-\infty}^{+\infty}\left|\frac{\partial f(x;\theta)}{\partial\theta_j}\right|dx\ ,$$

$$\left|\frac{\partial^2 \log(1-F(x;\theta))}{\partial\theta_i\partial\theta_j}\right| \le \frac{1}{1-F(x;\theta)}\int_{-\infty}^{+\infty}\left|\frac{\partial^2 f(x;\theta)}{\partial\theta_i\partial\theta_j}\right|dx +$$

$$+\frac{1}{(1-F(x;\theta))^2}\int_{-\infty}^{+\infty}\left|\frac{\partial f(x;\theta)}{\partial\theta_i}\right|dx\int_{-\infty}^{+\infty}\left|\frac{\partial f(x;\theta)}{\partial\theta_j}\right|dx\ .$$

We present the asymptotic properties of estimates $F_{\tau n}$ from [1]. We define a sequence of processes $\left\{V_n(x) = n^{1/2}(F_{\tau n}(x) - F_\tau(x)), x \ge \tau, n \ge 1\right\}$. For these processes the sequence of approximating processes is $\left\{M_n(x) = (1-F_\tau(x))N_n(x)\right\}$, where

$$N_n(x) = \int_\tau^x \frac{(B_n(u) - \beta_n^*(u))dH^{(1)}(u)}{(G_1(u) - H(u-))^2} + \frac{B_n^{(1)}(x)}{G_1(x) - H(x-)} -$$

$$-\frac{B_n^{(1)}(\tau)}{G_1(\tau) - N(\tau-)} - \int_\tau^x \frac{B_n^{(1)}(u)d(G_1(u) - H(u-))}{(G_1(u) - H(u-))^2},$$

$$\beta_n^*(x) = -G_1(x)\left(\int_\tau^{+\infty}\frac{B_n(u)dH^{(0)}(u)}{H^2(u)} + \frac{B_n^{(0)}(x)}{H(x)} - \int_\tau^{+\infty}\frac{B_n^{(0)}(u)dH(u)}{H^2(u)}\right).$$

Here, for each n: $H(x) = P_{\theta_0}\left(Z_1 < x\right) = EH_n(x)$, $H^{(m)}(x) = P_{\theta_0}\left(Z_1 < x, \delta_{m1} = 1\right) = EH_n^{(m)}(x)$; $B_n(u) \overset{D}{=} B(H(u))$, $B_n^{(m)}(u) \overset{D}{=} B(H^{(m)}(u))$, $m = 0,1,2$ and $\left\{B(y), 0 \le y \le 1\right\}$ is process of a Brownian bridge. Note, that the processes $M_n(x)$ are linear functional of the Brownian bridge, and thus are Gaussian processes with zero mean. We present the following theorem from [1 , Theorem 2.1.2].

Theorem A. [1]. Under condition (C1) we have an approximation

$$P\left(\sup_{\tau\le x\le T}\left|V_n(x) - M_n(x)\right| > Rn^{-1/2}\log n\right) \le Qn^{-\varepsilon},$$

where ε, $R = R(\varepsilon)$ and Q (absolute) positive constants.

Remark 2. In conditions of Theorem A for $\varepsilon > 1$ by lemma of Borel-Cantelli we have the strong approximation

$$\sup_{\tau \le x \le T} \left| V_n(x) - M_n(x) \right| \overset{\text{a.s.}}{=} O\left(n^{-1/2} \log n \right).$$

From here we have the weak convergence

$$V_n(x) \overset{D}{\Rightarrow} M(x) \text{ in } D[\tau;T], \tag{1}$$

where $M_n(\cdot) \overset{D}{=} M(\cdot)$ for each n and Gaussian process $M(x)$ obtained from $M_n(x)$ by replacement of $B_n(u)$ and $B_n^{(m)}(u)$, $m = 0,1,2$ by the appropriate Brownian bridges with arguments $H(u)$ and $H^{(m)}(u), m = 0,1,2$ respectively.

We introduce the random processes $\varphi_n(x;\theta) = n^{1/2}(F_{\tau n}(x) - F_\tau(x;\theta))$, $\hat{\varphi}_n(x) = \varphi_n(x;\hat{\theta}_n)$. Let $\tau = x_o < x_1 < ... < x_{r-1} < x_r < T < \infty$ possible random partition for a given probability p_i, satisfying the equality $F(t_i;\hat{\theta}_n) = p_i$. Consider a random vector $\Phi_n = \left(\hat{\varphi}_n(x_1), ..., \hat{\varphi}_n(x_r) \right)^T$. The next result generalizes (1).

Theorem 1. Let for all $\theta \in \Theta$ the conditions (C1)-(C5) hold. Then the random process $\left\{ \hat{\varphi}_n(x), \tau \le x \le T \right\}$ converges weakly to the Gaussian process $\hat{\varphi}_n(x)$ with zero mean and covariance with $\tau < x \le y < T$:

$$Cov_\theta \left(\hat{\varphi}(x), \hat{\varphi}(y) \right) = Cov_\theta \left(M(x), M(y) \right) - \left(\frac{\partial F(x;\theta)}{\partial \theta} \right)^T I^{-1}(\theta) \frac{\partial F(y;\theta)}{\partial \theta},$$

where $\frac{\partial F(x;\theta)}{\partial \theta} = \left(\frac{\partial F(x;\theta)}{\partial \theta_1}, ..., \frac{\partial F(x;\theta)}{\partial \theta_s} \right)^s$.

Proof of the Theorem 1. Expand the process $\hat{\varphi}_n(x)$ in the neighborhood of $\hat{\theta}_n$ and then we have

$$\hat{\varphi}_n(x) = V_n(x) + V_n^*(x) + R_n(x), \tag{2}$$

where $V_n^*(x) = \frac{\partial F(x;\theta)}{\partial \theta} n^{1/2} \left(\hat{\theta}_n - \theta \right)$ and under $n \to \infty$ from Theorem A

$$\sup_{\tau \le x \le T} \left| R_n(x) \right| = o_p(1).$$

Consider a finite-dimensional distribution of sums $V_n(x)+V_n^*(x)$. For an arbitrary partition $\tau < x_1 < ... < x_{r-1} < x_r < T$, let $V_n = \left(V_n(x_1),...,V_n(x_r)\right)^T$ and $V_n^* = \left(V_n^*(x_1),...,V_n^*(x_r)\right)^T$ such that $V_n^* = Bn^{1/2}\left(\hat{\theta}_n - \theta\right)$, where $B = \left\|B_{ij}\right\|$ is matrix with elements $B_{ij} = \frac{\partial F(t_i;\theta)}{\partial \theta_j}$, $i=\overline{1,r}$, $j=\overline{1,s}$. By (1), condition (C5) and multivariate central limit theorem

$$\left(V_n, n^{1/2}\left(\hat{\theta}_n - \theta\right)\right)^T \underset{n\to\infty}{\overset{D}{\Rightarrow}} \left(V, \xi_0\right)^T \overset{D}{=} N(0;\Sigma), \qquad (3)$$

where

$$\Sigma = \begin{Vmatrix} \mathrm{M} & \Delta \\ \Delta & I^{-1} \end{Vmatrix},$$

$\mathrm{M} = \left\|Cov_\theta\left(M(x_i), M(x_j)\right)\right\|_{i,j=\overline{1,r}}$, $I^{-1}(\theta)$ is inverse matrix of $I(\theta)$. Thus, $V_n + V_n^* \overset{D}{\Rightarrow} \tilde{\Phi} = M + B\cdot\xi_0 \overset{D}{=} N\left(0;\Sigma_0\right)$. Under conditions (C2)-(C5) from results of [3] follows that $\tilde{\Phi}$ and r.v. ξ_0 is independent and therefore $M = \Sigma_0 + B I^{-1} B^T$. Hence, we have

$$M_\theta\left[\tilde{\Phi}^T\tilde{\Phi}\right] = \Sigma_0 = M - B I^{-1} B^T. \qquad (4)$$

In view of (2)-(4), using the technique of proof of Theorem 4 in [2], we see that the weak convergence of the sum $V_n(x)+V_n^*(x)$ follows from the weak convergence to the continuous limits of individual summands. Convergence $V_n(x)$ follows from (1), process $V_n^*(x)$ consists of product of non-random function to an asymptotically normal sequence of random variables and therefore its convergence to the limit is obvious. This completes the proof of Theorem 1.

Let $\tilde{M}$ and $\hat{\Sigma}_0$ estimates of the matrices M= and Σ_0, obtained by replacing θ on MLE $\hat{\theta}_n$. The functions $G_1, H^{(m)}$ and H replaced by their nonparametric estimates G_{1n}, $H_n^{(m)}$ and $H_n, m=0,1,2$. Following the general principles of construction of chi-square statistics (see [4,5]), we consider the statistics

$$\Omega_n\left(\hat{\theta}_n\right) = \tilde{\Phi}_n^T \hat{\Sigma}_0 \tilde{\Phi}_n,$$

where $\tilde{\Phi}_n = \left(\hat{\varphi}_n(x_1),...,\hat{\varphi}_n(x_r)\right)^T$. Then we have

Theorem 2. Let the conditions (C1)-(C5) hold and $rang\left(\hat{\Sigma}_0\right) = r$. Then

$$\mathscr{L}\left(\Omega_n\left(\hat{\theta}_n\right)/H_0\right) \underset{n\to\infty}{\to} K_r,$$

where K_r is chi-square distribution with degrees of freedom r.

3. CHI-SQUARE TEST BASED ON THE LIKELIHOOD RATIO STATISTICS

First**,** let θ is scalar parameter. Consider a simple hypothesis $H_0:\theta=\theta_0$ against the composite alternative $H_1:\theta\in\Theta_1$, where $\Theta=\{\theta_0\}\cup\Theta_1$. Let $\hat{\theta}_n$ is MLE, satisfying the condition (C5) and consider LRS

$$L_n=\frac{p_n\left(\hat{\theta}_n\right)}{p_n\left(\theta_0\right)}.$$

We also consider the condition

(C6) There is a third derivative on θ of density $f(x;\theta)$, exists the independent of θ function h_n such that

$$\left|\frac{d^3\log p_n(\theta)}{d\theta^3}\right|\le h_n;\quad M_\theta h_n<\infty.$$

From the general theory of MLE (see [6]) follows that under conditions (C2) - (C4), (C6), there exists a unique consistent MLE $\hat{\theta}_n$ and at $n\to\infty$

$$\mathscr{L}\left(n^{1/2}\left(\hat{\theta}_n-\theta_0\right)/H_0\right)\to N\left(0,I^{-1}(\theta_0)\right).$$

Then by Taylor's formula

$$\log L_n=\log p_n(\theta_n)-\log p_n(\theta_0)=\frac{\partial\log p_n(\theta_0)}{\partial\theta}\left(\hat{\theta}_n-\theta_0\right)+$$

$$+\frac{1}{2}\frac{\partial^2\log p_n(\theta_0)}{\partial\theta^2}\left(\hat{\theta}_n-\theta_0\right)^2+\frac{1}{3!}\frac{\partial^3\log p_n(\theta_*)}{\partial\theta^3}\left(\hat{\theta}_n-\theta_0\right)^3 \quad (5)$$

and

$$\frac{\partial\log p_n(\hat{\theta}_n)}{\partial\theta}-\frac{\partial\log p_n(\theta_0)}{\partial\theta}=\frac{\partial^2\log p_n(\theta_{**})}{\partial\theta^2}\left(\hat{\theta}_n-\theta_0\right)+\frac{1}{2}\frac{\partial^3\log p_n(\theta_0)}{\partial\theta^3}\left(\hat{\theta}_n-\theta_0\right)^2, \quad (6)$$

where $|\theta_*-\theta_0|\vee|\theta_{**}-\theta_0|\le\left|\hat{\theta}_n-\theta_0\right|$. Since, $\frac{\partial\log p_n(\theta_0)}{\partial\theta}=0$, then substituting the expression for $\frac{\partial\log p_n(\hat{\theta}_n)}{\partial\theta}$ from (6) into (5) we have

$$\log L_n = -\frac{1}{2}\frac{\partial^2 \log p_n(\theta_0)}{\partial\theta^2}\left(\hat{\theta}_n - \theta_0\right)^2 + q_n,$$

where $q_n = o_p(1)$ at $n \to \infty$. Now, using the low of large numbers and central limit theorem, we find that under hypothesis H_0 at $n \to \infty$ statistics

$$-\frac{1}{n}I^{-1}(\theta_0)\frac{\partial^2 \log p_n(\theta_0)}{\partial\theta^2}(nI(\theta_0))\left(\hat{\theta}_n - \theta_0\right)^2$$

have a chi-square distribution K_1 with one degree of freedom, i.e.

$$\mathscr{L}\left(2\log L_n / H_0\right) \to K_1. \tag{7}$$

Thus we have proved

Theorem 3. Under conditions (C2) - (C6) and the hypothesis H_0

$$2\log L_n \overset{D}{\Rightarrow} \chi_1^2. \tag{8}$$

Theorems 2 and 3 can be used to construct an asymptotic chi-square tests for the hypotheses H_0.

Remark 3. It should be noted that in the case s dimensional parameter $\theta = (\theta_1, \ldots, \theta_s)$ in the limit in (7) we will have a chi-square distribution K_s, i.e. instead of (8) we have $2\log L_n \overset{D}{\Rightarrow} \chi_s^2$. This result is proved similarly to convergence (7), using a multivariate central limit theorem for statistics

$$2\log L_n = -(\theta_n - \theta_0)^T\left(\frac{\partial \log p_n(\theta_0)}{\partial\theta}\right)\left(\frac{\partial \log p_n(\theta_0)}{\partial\theta}\right)^T(\theta_n - \theta_0) + q_n,$$

where $q_n = o_p(1), n \to \infty$.

4. REFERENCES

[1]. Abdushukurov A.A. Estimators of unknown distributions from incomplete observations and its properties. LAMBERT Academic Publishing. 2011. 299 p. (In Russian).

[2]. Breslow N., Crowley J. A large sample study of the life table and product limit estimates under random censorship.// Ann. Statist. 1974. v.2. p.437-453.

[3]. Piece D.A. The asymptotic effect of substituting estimators for parameters in certain types of statistics. // Ann. Statist. 1982. v.10. p.475-478.

[4]. Voinov V.G., Nikulin M.S. Unbiased estimates and there applications. –M.: Science. 1989.-440p. (In Russian).

[5]. Voinov V., Naumov A., Pya N. Some recent advances in chi-squared testing. // Proceeding of Internat. Conf. “Advances in Statistical Inferential Methods” : Almaty. Kazakhstan. 2003. p.233-247.

[6]. Zacks Sh. The theory oy statistical inference John Wiley & Sons, Inc. New York-London-Sydney-Toronto. 1971. -776p.