A 'missing-plot' technique for goodness-of-fit tests with censored data

Arusharka Sen
University of Hyderabad, Department of Mathematics and Statistics
Hyderabad - 500 046
Hyderabad, India
asensm@uohyd.ernet.in

1. Introduction

Suppose $X_i, 1 \leq i \leq n$, are i.i.d random variables with distribution function (d.f.) F. Consider the goodness-of-fit (g.o.f.) test problem $H_0 : F = F_0$ vs. $H_1 : F \neq F_0$, where F_0 is a specified d.f.. A simple and time-honoured test is provided by the Kolmogorov-Smirnov (K-S) statistic $||F_n - F_0|| = \sup_{x \in I\!R} |F_n(x) - F_0(x)|$, where $F_n(x) := 1/n \sum_{i=1}^n \mathbf{1}\{X_i \leq x\}$ is the *empirical distribution function* (e.d.f.). (Here and elsewhere, $\mathbf{1}\{A\}$ denotes the indicator of the event A.)

In survival studies, the (typically non-negative) observations are often subject to some kind of censoring. A few well-known censoring models considered in the literature are: Type-I, Type-II, random censoring, double censoring, interval censoring (Case-I and Case-II) etc. Each of these can be looked upon as special cases of the following, somewhat informally described, general model: instead of X_i, we observe (possibly random) sets $A_i \subseteq I\!R$ such that $X_i \in A_i$, $1 \leq i \leq n$. A_i is typically either a singleton (which means X_i *is* observed), or an interval (possibly depending on other, 'censoring', variables). For example, in random censoring, $A_i = \{X_i\}$ or $A_i = (Y_i, \infty)$; in interval censoring (Case -I), $A_i = [0, Y_i]$ or $A_i = (Y_i, \infty)$. In these two examples, $Y_i, 1 \leq i \leq n$, are the censoring variables.

In this paper, we propose a 'reconstructed' K-S procedure to test H_0 under any of the censoring schemes of the above type. Define the function $\Phi_n(x, v_1, \ldots, v_n) \equiv \Phi_n(x, \mathbf{v}) = 1/n \sum_{i=1}^n \mathbf{1}\{v_i \leq x\}$. Solve the optimization problem:

$$\begin{array}{ll} \text{minimize} & \sup_x |\Phi_n(x, \mathbf{v}) - F_0(x)| \\ \text{subject to} & v_i \in A_i,\ 1 \leq i \leq n. \end{array} \tag{1.1}$$

Let $\mathbf{v}^* = (v_1^*, \ldots, v_n^*)$ be an optimal solution to (1.1). Call $\Phi_n(x, \mathbf{v}^*) \equiv \Phi_n^*(x)$, the *reconstructed* e.d.f., and $D_n^* := ||\Phi_n^* - F_0||$, the *reconstructed* K-S statistic. We then reject H_0 for large values of D_n^*.

Note that this method is analogous to the well-known 'missing-plot' technique devised by F.Yates in the context of missing data in an agricultural experiment where a linear model is used (see, for example, Montgomery (1991)).

Here, we illustrate the method for two important censoring schemes, viz., *random censoring* (see Fleming and Harrington (1991)) and *interval censoring (Case-I)* (see Groeneboom and Wellner (1992)). In Section 2, we give the construction of D_n^* under these censoring schemes. In Section 3, the test-procedure is described, and some simple asymptotic results on the size and power of the test are given.

There are several goodness-of-fit tests in the literature for random censoring, which do or do not reduce to a corresponding 'uncensored' test in the absence of censoring. See, for example, Andersen et al (1993) or Fleming and Harrington (1991). To our knowledge, there is no goodness-of-fit test for interval censoring (Case-I or II) so far. Note also that our method reduces to the usual K-S test when there is no censoring.

2. Construction of D_n^*

Random Censoring: under random censoring, one observes (δ_i, Z_i), where $\delta_i = \mathbf{1}\{X_i \leq Y_i\}$, $Z_i = X_i \wedge Y_i, 1 \leq i \leq n$, $Y_1, \ldots, Y_n$ being the i.i.d. censoring variables with common d.f. G. Let us write Z_i for $F_0(Z_i), 1 \leq i \leq n$, and assume WOLG that $0 \leq Z_1 \leq \cdots \leq Z_n \leq 1$. Then, if there are no ties among $Z_i, 1 \leq i \leq n$, $\Phi_n^*(x), 0 \leq x \leq 1$, is given by

$$\Phi_n^*(x) = \begin{cases} \Phi_n(x, \mathbf{s}), & \text{if } ||\Phi_n(\cdot, \mathbf{s}) - \cdot|| < ||\Phi_n(\cdot, \mathbf{t}) - \cdot||, \\ \Phi_n(x, \mathbf{t}), & \text{otherwise,} \end{cases}$$

where $||\Phi_n(\cdot, \mathbf{s}) - \cdot|| = \sup_{0 \leq x \leq 1} |\Phi_n(x, \mathbf{s}) - x|$, and $\mathbf{s} = (s_1, \ldots, s_n)$, $\mathbf{t} = (t_1, \ldots, t_n)$ are obtained as follows:

definition of $\mathbf{s}$: for $j = 1$, $s_1 = Z_1$, if $\delta_1 = 1$ or if $\delta_1 = 0$ and $Z_1 \geq 1/2n$; $s_1 = \min\{1/2n, Z_{i_1}\}$, if $\delta_1 = 0$ and $Z_1 < 1/2n$, where Z_{i_1} is the smallest *uncensored* $Z_i > Z_1$; for $2 \leq j \leq n$, $s_j = Z_j$, if $\delta_j = 1$ or if $\delta_j = 0$ and $Z_j \geq (2j-1)/2n$; $s_j = \min\{(2j-1)/2n, Z_{i_j}\}$, if $\delta_j = 0$ and $Z_j < (2j-1)/2n$, where Z_{i_j} is the smallest *uncensored* $Z_i > Z_j$;

definition of $\mathbf{t}$: for $j = n$, $t_n = Z_n$, if $H_{1n}(Z_n) = 1$; $t_n = \max\{(2n-1)/2n, Z_n\}$, if $H_{1n}(Z_n) < 1$; for $1 \leq j \leq n-1$, $t_{n-j} = Z_{n-j}$, if $H_{1n}(Z_{n-j}) = (n-j)/n$; $t_{n-j} = \max\{(2(n-j)-1)/2n, Z_{i_{n-j}}\}$, if $H_{1n}(Z_n) < (n-j)/n$, where $Z_{i_{n-j}}$ is the largest *uncensored* $Z_i \leq Z_{n-j}$.

Note that there may be ties among $s_1, \ldots, s_n$ or $t_1, \ldots, t_n$.

Interval Censoring (Case-I): in this case, one observes (δ_i, Y_i), where $\delta_i = \mathbf{1}\{X_i \leq Y_i\}$, $1 \leq i \leq n$, and $Y_1, \ldots, Y_n$ are i.i.d. censoring variables with d.f. G. Here again, write Y_i for $F_0(Y_i)$, $1 \leq i \leq n$, and assume that $0 \leq Y_1 \leq \cdots \leq Y_n \leq 1$. Then $\Phi_n^*(x) = \Phi_n(x, \mathbf{s}), 0 \leq x \leq 1$, where $\mathbf{s} = (s_1, \ldots, s_n)$ is given by: for $1 \leq i \leq n$, $s_i = \min\{(2i-1)/2n, Y_i\}$, if $\delta_i = 1$, $s_i = \max\{(2i-1)/2n, Y_i\}$, if $\delta_i = 0$.

3. Test Procedure

Note that under any censoring scheme of the type described, we have

$$\begin{aligned} D_n^* &= \min_{v_i \in A_i, 1 \le i \le n} \sup_x |1/n \textstyle\sum_{i=1}^n \mathbf{1}\{v_i \le x\} - F_0(x)| \\ &\le \sup_x |1/n \textstyle\sum_{i=1}^n \mathbf{1}\{X_i \le x\} - F_0(x)| = D_n, \text{ say,} \end{aligned} \tag{3.1}$$

where $X_i, 1 \le i \le n$, is the full sample which may not be observable under censoring. Motivated by (3.1), we propose the following test-procedure under any of the above censoring schemes:

$$\text{reject } H_0 \text{ if } n^{1/2} D_n^* > d_\alpha,$$

where $\lim_{n\to\infty} P\{n^{1/2} D_n > d_\alpha | H_0\} = \alpha$. Let α_n be the size of this test. Then it follows that $\text{limsup}_{n\to\infty} \alpha_n \le \alpha$. As for power, we have the following results:

THEOREM 3.1 Consider a *random censoring* scheme where $Y_i, 1 \le i \le n$, are i.i.d. with d.f. G and independent of $X_i, 1 \le i \le n$. Then

$$\lim_{n\to\infty} P\{n^{1/2} D_n^* > d_\alpha | (F, G)\} = 1$$

for (F, G) continuous and $(1 - F)(1 - G) \ge (1 - F_0)$ with strict inequality for some $x > 0$.

THEOREM 3.2 Consider an *interval censoring (Case-I)* scheme where $Y_i, 1 \le i \le n$, are i.i.d. with d.f. G and independent of $X_i, 1 \le i \le n$. Then

$$\lim_{n\to\infty} P\{n^{1/2} D_n^* > d_\alpha | (F, G)\} = 1$$

for (F, G) continuous and $\int (1 - F(y)) \mathbf{1}\{y > x\} G(dy) > 1 - F_0(x)$ for some $x > 0$.

REFERENCES

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes. Springer Verlag.*

Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis. John Wiley & Sons.*

Groeneboom, P. and Wellner, J.W. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation. Birkhäuser Verlag.*

Montgomery, D.C. (1991). *Design and Analysis of Experiments (3rd.Ed.). John Wiley & Sons.*

RESUME

We propose an optimization approach for 'reconstructing' the Kolmogorov- Smirnov statistic under any form of censoring. The reconstructed statistic may then be used for goodness-of-fit tests. The method is illustrated here for *random censoring* and *interval censoring (Case-I).*