The Third International Conference "Problems of Cybernetics and Informatics"
September 6-8, 2010, Baku, Azerbaijan. The plenary paper
www.pci2010.science.az/4/06.pdf

ON THE HYPOTHESIS TESTING CONCERNING THE TYPE OF THE SURVIVAL FUNCTION

Evgeniy Chepurin[1], Innesa Chepurina[2]

Moscow State University, Moscow, Russia
[1]*echepurin@mail.ru*, [2]*uchsov@cs.msu.ru*

The aim of this paper is to discuss the problems of the testing hypothesis concerning a survival distribution function form.

Let the random variable T be a lifetime of a person which exists at the moment $t = 0$. In other words, the random variable T is the future lifetime of this entity measured from $t = 0$. The probability that this person's lifetime is greater than t, i.e.

$$S(t) = \mathbf{P}\{T > t\},$$

is called *survival distribution function*. It is supposed that

a) $T \geq 0$;

b) $S(0) = 1$;

c) $S(t)$ is a non-increasing function;

d) $S(\infty) = 0$.

Suppose also that

e) $S(t)$ is a differentiable function.

In this paper it is assumed that any other information about survival distribution function except the characteristics a) – e) is unknown.

Let us denote by

$$p_x = \mathbf{P}\{T > x+1 \,|\, T > x\} = \frac{S(x+1)}{S(x)} \tag{1}$$

and

$$q_x = 1 - p_x = \frac{S(x) - S(x+1)}{S(x)}. \tag{2}$$

Consider the sequence of the embedded events

$$\{T > k\} \supset \{T > k+1\} \supset \ldots \{T > x-1\} \supset \{T > x\}, \tag{3}$$

where k and x are integers. From the properties of conditional distributions, the embedding (3) and the assumption that $x > k$, follows that

$$S(x) = S(x-1)\mathbf{P}\{T > x \,|\, T > x-1\} = p_{x-1}S(x-1), \tag{4}$$

$$S(x) = S(k)\prod_{j=k}^{x-1} p_j, \tag{5}$$

and

$$S(x) = S(k)\prod_{j=k}^{x-1}(1 - q_j). \tag{6}$$

Let $S_0(t)$ be a veritable survival distribution function which generates our random life data concerning the life durations of some people group, for example, the participants of some pension fund. In reality an actual functional form of $S_0(t)$ is unknown.

At the same time often it is necessary to test hypothesizes $\Gamma_1 : S_0(t) \equiv S_1(t)$ for $t_1 \leq t \leq t_2$, under alternatives $\Gamma_2 : S_0(t) \neq S_1(t)$ for $t_1 \leq t \leq t_2$ on the basis of special censored

sample of size N. Here $S_1(t)$ will be well known survival distribution function, t_1 and t_2 will be known constants. It is supposed that $S_1(t)$ is either parametric function with known parameters or is defined by the life table, for integers $k : t_1 \le k \le t_2$. To obtain survival data for the members of pension plan we use the following sample scheme: participant i, $i = \overline{1, N}$, is observed only in the time interval $[H_1, H_2]$, where H_1 and H_2 are calendar dates for the beginning and the termination of the observation period in which life status information of statistical data members can be obtained.

Let $Y = (A_1, A_2, \ldots, A_N)$ be a sample data, where

$A_i = (W_i, \tilde{J}_i, J_i, B_i, D_i, y_i, z_i)$,

W_i is calendar dates of withdrawal from observation (by reason other than by death),

$\tilde{J}_i$ is calendar dates of joining to pension plan,

J_i is observable calendar dates of joining to group under observation,

B_i is calendar dates of birth,

D_i is calendar dates of death (if $D_i < W_i$),

y_i is calendar dates of age under joining to observation group $(y_i = J_i - B_i)$,

z_i is scheduled age at termination for $i^{\underline{th}}$ sample member.

In our case

$$J_i = \begin{cases} \tilde{J}_i & if \quad \tilde{J}_i \ge H_1, \\ H_1 & if \quad \tilde{J}_i < H_1, \end{cases}$$

and

$$z_i = \begin{cases} H_2 - B_i & if \quad W_i > H_2, \\ W_i - B_i & if \quad W_i \le H_2. \end{cases}$$

Then let be

$y_{(1)} = \min\, y_i$,

$z_n = \max\, z_i$,

$\underline{x} = \min\limits_{t_1 \le x \le t_2} \{integer\ x \ge y_{(1)}\}$,

$\overline{x} = \max\limits_{t_1 \le x \le t_2} \{integer\ x \le z_{(n)}\}$,

$A(x) = \{A_i : for\ x\ integer\ (y_i \le x, \quad z_i \ge x+1) \cup (x < y_i < x+1) \cup$

$\cup (x < z_i < x+1) \cup (x \le y_i < z_i < x+1)\}$,

n_x is the number of $A_i \in A(x)$.

For $A_i \in A(x)$ let us define

$$\upsilon_i = \upsilon_i(x) = \begin{cases} z_i - x & when \quad x < z_i < x+1, \\ 1 & when \quad z_i \ge x+1, \end{cases}$$

and

$$r_i = r_i(x) = \begin{cases} y_i - x & when \quad x < y_i < x+1, \\ 0 & when \quad y_i \le x, \end{cases}$$

ν_x is the number $A_i \in A(x)$ died in interval $(x, x+1]$.

The sample design is such that $N, n_x, \tilde{J}_i, J_i, B_i, H_1, H_2, y_i, z_i$ are non random (deterministic) variables known to H_x.

At the same time events connected with D_i are random events.

It is known [1] that

$$\hat{q}_x = \frac{\nu_x}{\sum\limits_{i:A_i \in A(x)} \left(\upsilon_i(x) - r_i(x)\right)} \tag{7}$$

and

$$\hat{S}(x) = S_0(t_1) \prod_{j=t_1}^{x-1} \hat{p}_j, \tag{8}$$

where $\hat{p}_j = 1 - \hat{q}_j$, is unbiased moment estimator of q_x and $S_0(x)$.

If x is exposure

$$\mathcal{E}_x = \sum_{i:A_i \in A(x)} \left(\upsilon_i(x) - r_i(x)\right) \to \infty, \quad N \to \infty, \tag{9}$$

Then it can be shown in the usual way that $\hat{q}_x$ and $\hat{S}_x$ are consistent and asymptotically normal estimators. If under Γ_1 the value $S_0(t_1)$ is known then it is possible to construct a test statistic by using the estimator (2). In the opposite case the estimator (1) can be used.

From the conditions (3) and

$$\frac{\sum\limits_{i:A_i \in A(x)} \left(\upsilon_i(x) - r_i(x)\right)^2}{\sum\limits_{i:A_i \in A(x)} \left(\upsilon_i(x) - r_i(x)\right)} \to 1, \quad N \to \infty, \tag{10}$$

it follows that

$$V_x = \frac{\hat{p}_x \hat{q}_x}{\sum\limits_{i:A_i \in A(x)} \left(\upsilon_i(x) - r_i(x)\right)} \tag{11}$$

are consistent estimators for $\boldsymbol{D}\hat{q}_x$. If under Γ_1 the conditions (3), (4) are fulfilled and $\hat{q}_x \equiv q_{1x}$ then we have

$$S_x = \frac{\hat{q}_x - q_{1x}}{\sqrt{V_x}} \overset{d}{=} N(0,1) + o_d(1). \tag{12}$$

Using graphic methods [2], we find a set $\mu = \{x\}$ such that the difference between $\hat{q}_x$ and q_{1x} will be obviously considerable. Let m be the number of integer points $\tilde{x} \in \mu$. Denote by $\Im(\mu)$ a test statistic for testing the hypothesis Γ_1 against Γ_2.

Lets assume that

$$\Im(\mu) = \sum_{\tilde{x} \in \mu} \delta_{\tilde{x}}^2 \tag{13}$$

But it is impossible to find the correct value of its observed significance level. The problem is that $\delta_{\tilde{x}}$ are complicated dependent random variables. However, observed significance level can be estimated.

We have

$$\lim_{N \to \infty} p_{obs}\left(\Im(\mu)\right) \le 1 - KHI\left(\frac{\left(\Im(\mu)\right)}{m}; m\right). \tag{14}$$

Here $KHI(u;m)$ is the distribution function of χ_m^2 random variable. The criterion with test statistic (13) is consistent.

References

1. London D. Survival models and their estimation. AXTEX Publications. Winsted and Avon, Connecticut, 1986.
2. Чепурин Е.В. Об аналитико-компьютерных методах разведочного анализа данных. Колмогоров и современная математика. Тезисы докладов. Москва, МГУ, 2003.