УДК 519. 25

CENSORED DATA ANALYSIS WITH USING INFORMATION ABOUT SYMMETRY OF DISTRIBUTION

ZHANNA ZENKOVA
senior lecturer, PhD
NIKOLAY KOLYCHEV

Tomsk (Russia)

The data used at the analysis of economic indicators, engineering, biological and medical researches, can have casual character. Therefore, there is a necessity for estimation of distribution functions of random variables on which basis it is possible to construct various statistical procedures, to find values of many numerical characteristics, for example, an average or a dispersion.

Practically there is a priory information on function of distribution of an investigated random variable take place, for example, its continuity, symmetry, the moments and so forth. The source of this information are experimental conditions, theoretical conclusions, physical sense of a random variable? etc. Hence, there are questions of the account of the available additional information at construction of estimations of distribution functions, and also research of properties received thus the statistics.

The considered problem becomes even more important in case sample is incomplete, truncated or censored [1]. The data such meets in practical work often enough, especially in reliability theory, at carrying out of medical, biological, demographic, economic researches and so forth. Censoring and reduction lead to essential losses of the information, therefore necessity for attraction of additional data on distribution becomes especially actual.

Besides, carrying out of many experiments is expensive or demands a lot of time for reception of results, therefore there is a problem of attraction of aprioristic data on distribution for reduction of quantity of tests and duration of experiences.

Let $\tau \in [0, T]$ is random value with distribution function $F(t)$ and $(X, I) = \{(X_1, I_1), (X_2, I_2), \dots, (X_N, I_N)\}$ is progressive left censored sample with known moment of censoring T_1 .

$$I_i = \begin{cases} 0, \text{ if } X_i \text{ is full value;} \\ 1, \text{ if } X_i \text{ is censored value.} \end{cases}$$

Let's consider the following scheme of censoring: quantity of incomplete values in an interval $[0,T_1]$ is random and numerically equal to a share *g*, 0 <*g* <1, from number of serviceable objects in the end of an interval. Then the estimation of distribution function defined by the formula

$$F_N^c(t)=\begin{cases}0,\ t<0,\\ \frac{1}{N}\sum_{i=1}^{N} I_{[0,t)}(X_i)\bar{I}_i,\ 0\le t\le T_1,\\ \frac{r}{N}+\frac{1}{(1-g)N}\sum_{i=1}^{N} I_{[T_1,t)}(X_i)\bar{I}_i,\ N_1 0\\ \frac{r}{N},\ N_1=0 \quad (1)\\ 1,\ t>T,\end{cases}\ \{T_1 t\le T,\}$$

where for $i=\overline{1,N}$ $\bar{I}_i=1-I_i$, r is a number of complete full value in $[0,T_1]$, $I_A(x)=\{0{:}x\notin A,1{:}x\in A\}$, $N_1=(N-r)(1-g)$.

The estimation (1) is asymptotically unbiased, and

$$\sigma_C^2(t)=\begin{cases}0,\ t\notin[0,T],\\ F(t)(1-F(t)),\ t\in[0,T_1],\\ F(t)(1-F(t))+\frac{g(F(t)-p)(1-F(t))}{(1-p)(1-g)},\ t\in(T_1,T],\end{cases}$$

where $p=F(T_1)$, $p\in(0,1)$, $NDF_N^c(t)$. Here D is variance of random value .

2. S^α-symmetry of distribution function

For $\tau\in R$ let determines S^α*-symmetry* concerning the symmetry center α, if distribution function *F*(*t*) satisfies to a condition:

$$F(t)=1-F(S(t)+0),\ t\in R, \quad (2)$$

where $S(t)$ is continuous, decreasing and $(S)^{-1}(t)=S(t)$, $S(\alpha)=\alpha$. Here $(S)^{-1}(t)$ is inverse to *S*(*t*), $F(\alpha)=0.5$. Note, that if $S(t)=2\alpha-t$ than it is ordinary symmetry concerning a median

$$F(t)=1-F(2\alpha-t+0). \quad (3)$$

Theorem 1. *If F*(*t*) *is continuously increasing, than F*(*t*) *possesses property* (2), *thus for* $\alpha=F^{-1}(0.5)$,

$$S(t)=F^{-1}(1-F(t)), \quad (4)$$

where $F^{-1}(t)$ *is inverse to F*(*t*).

Thus uniform, normal, exponential, lognormal distribution are S^α*-symmetrical.*

3. Using a priory information about S^α-symmetry

Let is known that progressive left censored sample (X,I) is from S^α-symmetrical distribution function *F*(*t*). Than estimation of unknown *F*(*t*),

$$F_N^{cS}(t)=\frac{F_N^c(t)+1-F_N^c(S(t)+0)}{2}, \quad (5)$$

possesses property (2), where $F_N^c(t)$ is defined under the formula (1). Let use the substitution method for reception of estimation of an average, obtain unbiased estimation

$$\theta_N^{cS}=\int_{-\infty}^{+\infty} t dF_N^{cS}(t)= =\frac{1}{2(r+(N-r)(1-g))}\sum_{i=1}^{N}\left(X_i+S(X_i)+0\right)\bar{I}_i, \quad (6)$$

where for $i=\overline{1,N}$ $\bar{I}_i=1-I_i$.

If *g*=0, then $\theta_N^{cS}=\theta_N^S=\frac{1}{2N}\sum_{i=1}^{N}\left(X_i+S(X_i)\right)$ is estimation of average for full sample without censoring. The relation of marks variance is

$$\frac{D\theta_N^{cS}}{D\theta_N^S}=\frac{1-g}{(1-0.5g)}\left(1+\frac{Z}{S^2}\right)<1,$$

where $S^2=D\tau,\theta=\int_0^T t dF(t)$,

$$Z=\int_0^T (x-\theta)(S(x)-\theta)dF(x),$$

and *Z* is decreases on *x*, *g* is a share *g*, 0 <*g* <1, of serviceable objects.

Theorem 2. *Let* $\tau\in[0,T]$ *is random value with distribution function F*(*t*), *h is differentiated in a point* $a=\int g(x)dF(x)$, $0<(h'(a)(<\infty$, $\int g^2(x)dF_0(x)<\infty$. *Then estimating of parameter* θ : $\theta=h\left(\frac{1}{N}\sum_{i=1}^{N} g(x_i)\right)$ *is asymptotical normal estimation with coefficient* $\sigma^2=\left[h'(a)\right]^2\int\left(g(x)-a\right)^2 dF(x)$, i. e. $(\theta-\theta)\sqrt{N}\in N(0,\sigma^2)$.

That allows constructing confidential intervals for unknown average as follows

$$\theta-\frac{\sigma}{\sqrt{N}}t_\gamma<\theta<\theta+\frac{\sigma}{\sqrt{N}}t_\gamma,$$

where $t_\gamma=\Phi^{-1}(\gamma)$ is quantile of confidence level γ.

4. Example

It is considered cost of the stocks which are in a warehouse of one of departments of some large trade enterprise of Tomsk, Russia.

The problem consisted in estimating averages of stocks and constructing confidential intervals, having only the cost of stocks in previous and some incomplete information on stocks in the current period. Thus further it was possible to specify missing data that has allowed to draw conclusions on adequacy of used models (table 1).

It has been defined that data for November is lognormal with $\theta_{Nov}=5032.232$ and $\sigma_{Nov}=8502.45$. This information was applied for censored data on December. In the result it was obtained that mean cost $\theta_N^{cS}=4953.386$ and with confidence level 95%:

$$4742.56<\theta_{Dec}<5164.186.$$

If it was used the estimation without a priory information then $\theta_N^c = 5062.849$ and $4677.89 < \theta_{Dec} < 5447.849$.

After specification of the data it has been received that true value $\theta_{Dec} = 4768.96$, that allows to draw conclusions about improvement of estimation quality by means of attraction of the additional information. ■

Table 1

Data for analysis

Наименование	Ноябрь, тыс. руб.	Декабрь (неполные данные), тыс. руб.	Декабрь, тыс. руб.
Подвес прямой (10шт)	2680	3640	3640
Подвес с зажимом	1605	1699	1699
Подвес евро	17967	20825	20825
Профиль маячк 10мм 3м	4266	4168	4168
Профиль маячк 6мм 3м	7648	5413	5413
Профиль направл ПН 100*40мм 3м	506	678	678
Профиль направл ПН 28*27 3м	26053	––	12709
Профиль направл ПН 50*40 3м	1332	1657	1657
Профиль направл ПН 75*40 3м	4005	––	3027
Профиль потолочн ПП 60*27 3м	36878	––	33594
Профиль стоечный ПС 100*50мм 3м	800	712	712
Профиль стоечный ПС 50*50 3м	1465	––	1886
Профиль стоечный ПС 75*50 3м	3857	3219	3219
Соед-ль 1-уровн краб	13725	––	24466
Соед-ль профилей 2-уровн 60*27	1819	1369	1369
Тяга к подвесу 1000мм	1208	2938	2938
Тяга к подвесу 250мм	735	554	554
Тяга к подвесу 300мм	745	345	345
Тяга к подвесу 500мм	909	1612	1612
Угол 20мм*20мм*3м сталь оцинков	7715	> 1500	7035
Угол 25мм*25мм*3м алюм	2720	> 1500	4180
Угол 25мм*25мм*3м сталь оцинков	4576	41	41
Удлин-ль профилей 60*27	2939	> 1500	4651
ГВЛ влагост 2500*1200*10мм KNAUF	7678	> 1500	4478
ГВЛ влагост 2500*1200*12. 5мм KNAUF	181	122	122
ГКЛ 1500*600*12. 5мм KNAUF (1500*600)	337	385	385
ГКЛ 2000*1200*9. 5мм KNAUF (2000*1200)	336	389	389
ГКЛ 2500*1200*12. 5мм KNAUF	839	885	885
ГКЛ 2500*1200*9. 5мм KNAUF	4927	> 1500	5128
ГКЛ 2500*1200*9. 5мм Пермь	1176	1325	1325
ГКЛ влагост 2500*1200*12. 5мм KNAUF	255	336	336
ГКЛ влагост 2500*1200*9. 5мм KNAUF	678	877	877
Стекломагниевый лист 2500*1220*8мм	1273	1171	1171
Элемент пола 1200*600*20мм KNAUF	147	93	93
Профиль невидимый (багет) 2. 5м	146, 3	375	375
Панель потолочн Devon Eur/Artic 600*600	908	505	505
Панель потолочн Skyfon 600*600	2283	> 1500	1508
Панель потолочн Taurus 600*600 OWA	2022	546	546
Панель потолочн Енисей 600*600	62	38	38
Панель потолочн Эверест 600*600 (24)	31471	33890	33890
Направляющая основн 3. 70м бел	4459	> 1500	5136
Направляющая основн 3. 70м зол	127	96	96
Направляющая основн 3. 70м хром	249	126	126
Направляющая промежуточн 0. 6м бел	26129	25199	25199
Направляющая промежуточн 0. 6м хром	429	88	88
Направляющая промежуточн 1. 2м бел	21620	21976	21976
Направляющая промежуточн 1. 2м хром	401	151	151
Уголок пристен 3м бел	7910	> 1500	4231
Профиль AN 3м бел	1340	759	759
Рейка AN 135/A 4м бел	761	737	737
Рейка AN 85/A 3м бел	1286	670	670
Рейка AN 135/A 3м бел	605	447	447
Рейка AN 85/AC 4м бел	520	670	670