ESTIMATING THE MEAN AND STANDARD DEVIATION FROM PROGRESSIVELY CENSORED NORMAL SAMPLES

BY M. L. TIKU

Department of Applied Statistics,
University of Reading, U.K.
(Received in June 1967)

SUMMARY

A method of obtaining estimators, very nearly as efficient as the M.L. estimators and easier to compute, of the mean and standard deviation from progressively censored normal samples is presented.

1. INTRODUCTION

Let $f(x)=(2\pi)^{-\frac{1}{2}}\exp(-\frac{1}{2}x^2)$ be the p.d.f. of the normal variate x and $F(x)=\int_x^\infty f(x)dx$. Normal distribution is one of the distributions having the interesting property that the points $g(x)=f(x)/F(x)$, over an interval $a\leqslant x\leqslant b$ of finite length, lie very close to the line $\alpha+\beta x$, where

$$\alpha=g(a)-\beta a \quad \text{and} \quad \beta=\{g(b)-g(a)\}/(b-a) \qquad \ldots(1)$$

The equation $g(x)\simeq\alpha+\beta x$ was used by Tiku (1967 *a*, *b*, *c*) to simplify the solutions of the maximum likelihood (ML) equations calculated from censored and truncated samples from normal and log-normal distributions. The resulting estimators are very nearly as efficient as the ML estimators and more efficient than the best linear unbiassed estimators. This equation is used here to simplify the solution of the ML equations calculated from progressively censored normal samples of type I and type II. An example is worked out and the resulting estimates are compared with Cohen's (1963) ML estimates.

2. Progressively Censored Samples

To quote Cohen (1963, p. 328), a progressively censored sample is defined as follows :

> "Let N designate the total sample size and n the number of sample specimens which fail and which therefore result in completely determined life spans. Suppose that censoring occurs progressively in k stages at times T_i, $i=1, 2, \ldots\ldots k$, and that at the ith stage of censoring r_i sample specimens selected randomly from the survivors at time T_i are censored from further observations. It follows that
>
> $$N=n+\sum_{i=1}^{k} r_i \qquad \ldots(2)$$
>
> In type I censoring, the T_i are fixed and the number of survivors at these times are random variables. In type II censoring, the T_i coincide with times of failure and are random variables, whereas the number of survivors at these times are fixed. For both types r_i is fixed."

Let $(y_1, y_2, \ldots\ldots, y_N)$ be a k-stage progressively censored sample of type I from a $N(\mu, \sigma)$.

The samples of this kind are encountered in life and fatigue studies in the context of life testing [see Cohen (1963) and the references given there on page 339]. The log-likelihood function is obtained as

$$L=C-n \log \sigma-\sum_{i=1}^{n} \frac{1}{\sigma^2}(y_i-\mu)^2+\sum_{i=1}^{k} r_i \log F_i(z_i) \qquad \ldots(3)$$

where $z_i=(T_i-\mu)/\sigma$ and $F(z)=(2\pi)^{-\frac{1}{2}} \int_z^{\infty} \exp(-\frac{1}{2}x^2)\, dx$.

The M.L. equations for estimating μ and σ are

$$\frac{\partial L}{\partial \mu}=\frac{n}{\sigma}\left[\frac{1}{\sigma}(\bar{y}-\mu)+\sum_{i=1}^{k}\frac{r_i}{n} g(z_i)\right]=0. \qquad \ldots(4)$$

$$\frac{\partial L}{\partial \sigma}=\frac{n}{\sigma}\left[\frac{1}{\sigma^2}\left\{s^2+(\bar{y}-\mu)^2\right\}-1+\sum_{i=1}^{k}\frac{r_i}{n} z_i\, g(z_i)\right]=0 \qquad \ldots(5)$$

where $\bar{y}=\sum_{i=1}^{n} y_i/n$, $s^2=\sum_{i=1}^{n}(y_i-\bar{y})^2/n$ and $g(z_1)=f(z_i)/F(z_i)$.

Equations (4) and (5) have no explicit solutions and can only be solved by an iterative procedure. From probits, Cohen obtained

the first approximations to be used in such an iterative procedure. The final estimates $\hat{\mu}$ and $\hat{\sigma}$ can thus be obtained in two or three cycles. We present here an alternative method of estimating μ and σ based on (1).

It is to be noted that for moderately large n, $z_i=(T_i-\mu)/\sigma$ is likely to be covered by the interval (a_i, b_i), where

$$a_i=[T_i-(\bar{y}+s/\sqrt{n})]/s \text{ and } b_i=[T_i-(\bar{y}-s/\sqrt{n})]/s \qquad ...(6)$$

Of course, $\bar{y}$ and s^2 are biassed estimators of μ and σ^2 but we are only interested in z_i being covered by (a_i, b_i) not that it should necessarily be the mid-point of this interval. In (4) and (5) we replace $g(z_i)$ by $\alpha_i+\beta_i z_i$, $i=1, 2,..., k$, where α_i and β_i are given by (1) with $a=a_i$ and $b=b_i$, and obtain the following equations with $q_i=r_i/n$:—

$$\frac{\partial L}{\partial \mu} \cong \frac{\partial L'}{\partial \mu}=\frac{n}{\sigma}\left[\frac{1}{\sigma}(\bar{y}-\mu)+\sum_{i=1}^{k} q_i(\alpha_i+\beta_i z_i)\right] \qquad ...(7)$$

$$\frac{\partial L}{\partial \sigma} \cong \frac{\partial L'}{\partial \sigma}=\frac{n}{\sigma}\left[\frac{1}{\sigma^2}\{s^2+(\bar{y}-\mu)^2\}-1+\sum_{i=1}^{k} q_i z_i(\alpha_i+\beta_i z_i)\right] \qquad ...(8)$$

Equating (7) and (8) to zero we obtain the estimators [see also Tiku (1967 *a*, *b*, *c*)].

$$\mu_p=[(\bar{y}+\sum_{i=1}^{k} q_i\beta_i T_i)+\sigma\sum_{i=1}^{k} q_i\alpha_i]/(1+\sum_{i=1}^{k} q_i\beta_i) \qquad ...(9)$$

$$=K+\sigma L, \text{ say}$$

and σ_p is the only positive root of the equation

$$\sigma^2-[q_i\alpha_i(T_i-\mu)]\sigma-[s^2+(\bar{y}-\mu)^2+\sum_{i=1}^{k} q_i\beta_i(T_i-\mu)^2]=0 \qquad ...(10)$$

or

$$\sigma^2-[(1-K)\sum_{i=1}^{k} q_i\alpha_i T_i-L\bar{y}]\sigma-[s^2+\bar{y}^2+\sum_{i=1}^{k} q_i\beta_i T_i^2-(1+\sum_{i=1}^{k} q_i\beta_i)K^2]=0 \qquad ...(11)$$

according as μ is known or unknown. Here

$$\left.\begin{aligned} K&=(\bar{y}+\sum_{i=1}^{k} q_i\beta_i T_i)/(1+\sum_{i=1}^{k} q_i\beta_i)\\ L&=\sum_{i=1}^{k} q_i\alpha_i/(1+\sum_{i=1}^{k} q_i\beta_i)\end{aligned}\right\} \qquad ...(12)$$

Note that (10) and (11) being of the form x^2-bx-c, $c>0$, have only one positive root.

In case k is greater than say 3, it will be time consuming to calculate (α_i, β_i), $i=1, 2, \ldots\ldots, k$. We therefore need to obtain a common value (α, β) replacing (α_i, β_i). This can be done by noting that for moderately large n, T_i's are likely to be covered, all of them, by the interval (a, b), where

$$a=\{T_1-(\bar{y}+s/\sqrt{n})\}/s \text{ and } b=\{T_k-(\bar{y}-s/\sqrt{n})\}/s \qquad \ldots(13)$$

Substituting these values in (1) we obtain the values of α and β and calculate μ_p and σ_p from (9) to (12).

3. Bias in the Estimators

From (9) we obtain

$$E(\mu_p)=\mu+\{E(\bar{y}-\mu)+\sigma\sum_{i=1}^{k} q_i(\alpha_i+\beta_i z_i)\}/(1+\sum_{i=1}^{k} q_i\beta_i)$$

$$=\mu+E\left(\frac{\partial L'}{\partial \mu}\right)/R^2(\mu)$$

where $R^2(\mu)=-E\left(\frac{\partial^2 L'}{\partial \mu^2}\right)$ is given by (16). The exact conditional bias in μ_p is therefore equal to

$$B_1=E(\mu_p/\sigma)-\mu=\left\{E\left(\frac{\partial L'}{\partial \mu}\right)/R^2(\mu)\right\} \qquad \ldots(14)$$

It is difficult to evaluate the bias in σ_p but it follows from (18·31) of Kendall and Stuart (1961, p. 44) that the approximate conditional bias in σ_p is given by

$$B_2=E(\sigma_p/\mu)-\sigma=\left\{E\left(\frac{\partial L'}{\partial \sigma}\right)/R^2(\sigma)\right\} \qquad \ldots(15)$$

where $R^2(\sigma)=-E\left(\frac{\partial^2 L'}{\partial \sigma^2}\right)$ is given by (18).

Unfortunately it is difficult to work out the expressions for (14) and (15) but it is hoped the bias in the estimators is not appreciably large, at least for large n. Some Monte Carlo investigation is obviously needed to study the bias in these estimators.

4. Variances and Covariances

From (7) and (8) we obtain

$$J_{11}=-\frac{\partial^2 L'}{\partial \mu^2}=\frac{n}{\sigma^2}\left[1+\sum_{i=1}^{k} q_i\beta_i\right] \qquad \ldots(16)$$

$$J_{12}=-\frac{\partial^2 L'}{\partial\mu\partial\sigma}=\frac{n}{\sigma^2}\left[\frac{2}{\sigma}(\bar{y}-\mu)+\sum_{i=1}^{k}q_i\alpha_i+2\sum_{i=1}^{k}q_i\beta_i z_i\right] \quad ...(17)$$

$$J_{22}=-\frac{\partial^2 L'}{\partial\sigma^2}=\frac{n}{\sigma^2}\left[\frac{3}{\sigma^2}\{s^2+(\bar{y}-\mu)^2\}-1+2\sum_{i=1}^{k}q_i\alpha_i z_i+3\sum_{i=1}^{k}q_i\beta_i z_i^2\right] \quad ...(18)$$

The approximate variances and covariances of μ_p and σ_p are given by the elements of the matrix

$$[V_{ij}]=[E(J_{ij})]^{-1}\cong[J_{ij}]^{-1}\ \mu=\mu_p,\ \sigma=\sigma_p \quad ...(19)$$

Example. Consider the following example [Cohen (1963, p. 337)]. A total of 316 specimens were placed under observation and their life spans recorded in days. Ten specimens were censored after 36·5 days and ten more censored after 44·5 days. Data for this type I censored sample was summarized as follows :

$N=316$, $n=293$, $q_1=q_2=0{\cdot}0338$, $T_1=36{\cdot}5$, $T_2=44{\cdot}5$,
$\bar{y}=39{\cdot}2703$ and $s^2=20{\cdot}1634$.

Substituting these values in (13), (9) and (11), we obtain

$\alpha=\cdot8030$, $\beta=\cdot6475$ and $\mu_p=39{\cdot}560$ and $\sigma_p=4{\cdot}599$.

The approximate asymptotic variances and covariances of these estimates are given by (16)—(18) as follows :

$V(\mu_p)=0{\cdot}0686$, Cov $(\mu_p, \sigma_p)=0{\cdot}0018$ and $V(\sigma_p)=0{\cdot}0358$.

The estimators μ_p and σ_p were also calculated from (6), (9) and (11) and are substantially the same as above.

The above estimates may be compared with Cohen's M.L estimates

$\hat{\mu}=39{\cdot}583$ and $\hat{\sigma}=4{\cdot}611$ with

$V(\hat{\mu})=0{\cdot}069$, Cov $(\hat{\mu}, \hat{\sigma})=0{\cdot}002$ and $V(\hat{\sigma})=0{\cdot}036$.

Note that the computation of μ_p and σ_p is much easier than $\hat{\mu}$ and $\hat{\sigma}$.

The estimating equations for type II progressively censored samples are identical with the above estimating equations for type I censored samples except that (α_i, β_i), $i=1, 2, \ldots\ldots, k$, are to be obtained as the solutions of the following equations :

$$\alpha_i+\beta_i c_i=g(c_i) \quad \text{and} \quad \alpha_i+\beta_i d_i=g(d_i) \quad ...(20)$$

where c_i and d_i are determined by the following equations with $P(z)=1-F(z)$:

$$P(c_i)=p_i-\sqrt{\left\{\frac{1}{n}\,p_i(1-p_i)\right\}} \text{ and } P(d_i)=p_i+\sqrt{\left\{\frac{1}{n}\,p_i(1-p_i)\right\}} \quad \ldots(22)$$

This is because z_i is a random variable and for $n\to\infty$, $z_i\to t_i$, where $P(t_i)=p_i=i/n$. For moderately large n, z_i is therefore likely to be covered by the interval (c_i, d_i), $i=i, 2,\ldots, k$.

The normal probability integral $F(x)$ is extensively tabulated in Biometrika Tables.

My thanks are due to the referee for helpful comments.

REFERENCES

Cohen, A.C. (1963). Progressively censored samples in life testing. *Technometrics 5, 327-39.*

Kendall; M. G. and Stuart, A. (1961). The Advanced Theory of Statistics, Vol. 2. Charles Griffin and Co. ; London.

Tiku, M.L. (1967*a*). Estimating the mean and standard deviation from a censored normal sample. *Biometrika 54, 155-56.*

Tiku, M.L. (1967*b*). Estimating the parameters of a truncated normal distribution. (paper submitted to Aust. J. Stat.).

Tiku, M.L. (1967*c*). Estimating the parameters of log-normal distribution from censored samples. (paper to appear in J. Amer. Stat. Assn., March 1968).