Computational Statistics & Data Analysis 1 (1983) 167–174
North-Holland

Estimation of the variance of percentile estimates

Morton B. BROWN and Robert A. WOLFE

Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA

Received June 1983

Abstract: The asymptotic formula for the variance of a percentile estimate is inversely proportional to the square of the probability density function evaluated at that percentile. In this note we show, for small and moderate sample sizes, that the estimate of the variance can have a moderate to large coefficient of variation even when the form of the density is known. When the density must be estimated empirically, the coefficient of variation increases substantially. We conclude that the estimate of the variance should *not* be used in either confidence interval estimation or hypothesis testing except for very large sample sizes.

Keywords: Estimation, Percentile estimates.

Introduction

Estimation of the percentiles of a distribution function provides a useful summary for univariate data, but the interpretation of the estimated values is incomplete without an evaluation of the variability of the estimates. When an independent, identically distributed series of observations, $x_1, x_2, \ldots, x_N$, is available from a cumulative distribution, $F(x)$, the estimates of the percentiles are functions of the order statistics. Distribution-free confidence intervals can be obtained for the percentile estimates when the data are complete (David [5]). Emerson [6] and Brookmeyer and Crowley [4] propose two different nonparametric methods to estimate confidence intervals for the median of a distribution when some data are censored.

An alternate method of evaluating the precision of the percentile estimate is to compute its asymptotic variance (Gross and Clark [7])

$$\sigma_a^2 = \frac{1}{f(x_p)^2} \frac{P(1-P)}{N} \tag{1}$$

where x_p is the Pth percentile, $f(x)$ is the probability density function (p.d.f.) and N is the size of the sample. Usually, $f(x)$ must be estimated from the sample. Due to the computational simplicity of this variance estimate, several computer

programs report it both for complete and for censored data such as in a life table.

Our objective is to show that this estimate of the variance is unreliable even with moderately large sample sizes. When the p.d.f. is empirically estimated from the sample and used in place of the theoretical functional form of the p.d.f., the estimated variance is even less reliable. This is demonstrated for a range of Weibull distributions by evaluating the variability of (1) when the percentile is estimated or when the density and the percentile are jointly estimated. We conclude that this estimate of the variance should not be used in hypothesis testing or in confidence interval estimation.

Methods

Since censored data are often modelled by the Weibull family of distributions with c.d.f.

$$F(x) = 1 - \exp(-x^{\gamma}) \tag{2}$$

our study considers this family with

$\gamma = 0.5$, 1.0 (the unit exponential) and 2.0.

Given a sample of size N, the distribution of the jth order statistics $x_{(j)}$ is

$$g(u_j) = \frac{N!}{(j-1)!(N-j)!} u_j^{j-1}(1-u_j)^{N-j}, \quad 0 < u_j < 1, \tag{3}$$

where

$$u_j = F(x_{(j)})$$

and the joint distribution of the ith and jth order statistics ($i < j$) is

$$g(u_i, u_j) = \frac{N!}{(i-1)!(j-i-1)!(N-j)!} u_i^{i-1}(u_j - u_i)^{j-i-1}(1-u_j)^{N-j}, \quad 0 < u_i < u_j < 1, \tag{4}$$

where

$$u_i = F(x_{(i)}); \qquad u_j = F(x_{(j)})$$

(David [5]). By numerical integration we evaluate the variance of the jth order statistic

$$\sigma^2 = \mathrm{Var}_x(x_{(j)}) = \mathrm{Var}_u\left(F^{-1}(u_j)\right)$$
$$= E_u\left\{\left[F^{-1}(u_j)\right]^2\right\} - \left\{E_u\left[F^{-1}(u_j)\right]\right\}^2 \tag{5}$$

where $j = NP$.

Similarly we evaluate the expectation and variance of the asymptotic variance

formula assuming the functional form $f(x)$ is known. That is,

$$E_x\left[\frac{1}{f(x_{(j)})^2}\frac{P(1-P)}{N}\right] = E_u\left[\frac{1}{f\left(F^{-1}(u_j)\right)^2}\frac{P(1-P)}{N}\right] \tag{6}$$

is the expectation and

$$\mathrm{Var}_x\left[\frac{1}{f(x_{(j)})^2}\frac{P(1-P)}{N}\right] = \mathrm{Var}_u\left[\frac{1}{f\left(F^{-1}(u_j)\right)^2}\frac{P(1-P)}{N}\right] \tag{7}$$

where $u_j \doteq F(x_{(j)})$.

Lastly, when the functional form of the p.d.f. $f(x)$ is not assumed to be known, $f(x)$ is estimated as

$$\begin{aligned}\hat{f}(x_{(j)}) &= \frac{\hat{F}(x_{(j+h)}) - \hat{F}(x_{(j-h)})}{x_{(j+h)} - x_{(j-h)}} \\ &= \frac{2h/N}{x_{(j+h)} - x_{(j-h)}}\end{aligned} \tag{8}$$

for appropriately chosen values of h. That is, the density is estimated by numerical differentiation of the c.d.f. in the region of the percentile of interest. The evaluation of the expectation and variance requires numerical integration in two dimensions.

$$\begin{aligned}&E_{\underline{x}}\left[\frac{1}{\left\{\hat{f}(x_{(j)})\right\}^2}\frac{P(1-P)}{N}\right] \\ &= E_{\underline{u}}\left[\frac{\left\{F^{-1}(u_{j+h}) - F^{-1}(u_{j-h})\right\}^2}{(2h/N)^2}\frac{P(1-P)}{N}\right], \\ &0 < u_{j-h} < u_{j+h} < 1,\end{aligned} \tag{9}$$

and similarly for the variance.

The expressions (5), (6), (7) and (9) are evaluated by numerical integration using a 32-point Gaussian quadrature formula (Abramovitz and Stegun [1]). The region of integration (e.g. (0, 1) for the single integral) is divided into one, two, four, six and eight equal subintervals, as necessary, until the results of two successive integrations agree to a relative accuracy of 10^{-5}.

We set sample sizes N equal to 20, 40, 80 and 160 and percentiles P equal to 0.1, 0.25, 0.5, 0.75, 0.9 at which to evaluate the variance (5) and the expectation and variance of the approximation ((6) and (7)).

Due to the greater cost associated with the double integration (9), only a few selected computations are performed.

Results and discussion

The results of the numerical evaluation of the true variance (5) and of the expectation (6) and variance (7) of the asymptotic formula are presented in Tables 1, 2, and 3 for the Weibull distribution with $\gamma = 0.5$, 1.0 and 2.0 respectively. The first column of results presents the true variance σ^2 of $x_{(j)}$ where $j = NP$. The second column reports the ratio of the expectation of the asymptotic formula (6) to σ^2. The following column gives the coefficient of variation (in %) of the estimate due to the asymptotic formula

$$[= 100 \times \text{s.e. (formula)}/E(\text{formula})].$$

The last column presents the ratio of the asymptotic formula σ_a^2 (1), assuming that the density and x_p are known exactly, to the true variance σ^2.

The Weibull with $\gamma = 0.5$ is denser near zero than the other two distributions and sparser than the other two in the upper tail of the distribution. Hence the true variance of the percentiles is relatively small in the lower tail and relatively large in the upper tail. The Weibull with $\gamma = 2.0$ has a much reduced difference in the variances between the two tails.

As expected, the bias of the asymptotic formula decreases as the sample size

Table 1
The true variance and the expected value of the approximation of the variance for the Weibull distribution with $\gamma = 0.5$ when the density is known

P	N	σ^2	$E(\sigma_a^2)/\sigma^2$	c.v. of σ_a^2	σ_a^2/σ^2
0.1	20	0.0005835	0.752	213.9%	0.423
	40	0.0002013	0.833	129.1%	0.613
	80	0.0000803	0.900	84.5%	0.768
	160	0.0000354	0.944	57.5%	0.872
0.25	20	0.007702	1.050	158.5%	0.716
	40	0.003302	1.014	94.6%	0.835
	80	0.001515	1.003	62.0%	0.911
	160	0.000723	1.002	42.3%	0.953
0.5	20	0.10664	1.322	170.1%	0.901
	40	0.05082	1.141	93.7%	0.945
	80	0.02473	1.066	59.8%	0.971
	160	0.01219	1.032	40.4%	0.985
0.75	20	1.1531	1.848	*	1.028
	40	0.5706	1.351	138.3%	1.010
	80	0.2870	1.157	78.8%	1.004
	160	0.1439	1.074	51.0%	1.002
0.9	20	7.7247	*	*	1.235
	40	4.3120	2.119	*	1.107
	80	2.2715	1.418	151.9%	1.050
	160	1.1645	1.184	82.5%	1.024

* The numerical integration did not converge.

Table 2
The true variance and the expected value of the approximation to the variance for the exponential distribution when the density is known

P	N	σ^2	$E(\sigma_a^2)/\sigma^2$	c.v. of σ_a^2	σ_a^2/σ^2
0.1	20	0.005270	1.060	16.3%	1.054
	40	0.002705	1.030	11.0%	1.027
	80	0.001371	1.015	7.6%	1.013
	160	0.000690	1.007	5.3%	1.007
0.25	20	0.015722	1.079	29.0%	1.060
	40	0.008094	1.038	19.3%	1.030
	80	0.004106	1.019	13.3%	1.015
	160	0.002068	1.009	9.2%	1.007
0.5	20	0.046396	1.138	54.2%	1.078
	40	0.024081	1.065	34.5%	1.038
	80	0.012268	1.032	23.3%	1.019
	160	0.006192	1.016	16.2%	1.009
0.75	20	0.13255	1.344	129.8%	1.132
	40	0.07048	1.153	67.0%	1.064
	80	0.03635	1.072	42.4%	1.032
	160	0.01846	1.035	28.6%	1.016
0.9	20	0.34616	2.470	*	1.300
	40	0.19663	1.488	209.9%	1.144
	80	0.10509	1.208	88.0%	1.071
	160	0.05436	1.097	53.5%	1.035

* The numerical integration did not converge.

increases (first column). The percentile with the minimum bias depends upon the distribution. (It may be argued that the estimate of the pth percentile, x_j, should not be based on $j = PN$, but rather on $j = P(N+1)$ or some other function. However, $j = PN$ is commonly used to estimate percentiles in survival analysis.) Serious biases occur when N is small (20 or 40) or P is extreme (0.10 or 0.90).

A more important indicator of the quality of the approximation is the coefficient of variation (c.v.) of the asymptotic formula. When a sample is drawn from the Gaussian distribution, the expected value of the sample variancies σ^2 and the variance of the sample variance is $2\sigma^4/f$ where f is the degrees of freedom of the sample variance. Therefore, the c.v. of the sample variance is of the order of magnitude $100(2/f)^{1/2}$. Therefore, if $f = 50$, the c.v. is approximately 20%; if $f = 18$, c.v. $\simeq$ 33% and if $f = 8$, c.v. $\simeq$ 50%.

The c.v.'s that are reported in Tables 1, 2 and 3 are large compared to the c.v.'s where the estimates are based on Gaussian data. The magnitude of the c.v. is very dependent on the distributional form and on the percentile. For small samples or for the upper percentiles the c.v.'s are large.

Table 4 reports the equivalent results for $N = 160$ and $p = 0.5$ when the density is estimated by (8). The c.v.'s are large compared to those reported in Tables 1, 2,

Table 3
The true variance and the expected value of the approximation to the variance for the Weibull distribution with $\gamma = 2.0$ when the density is known

P	N	σ^2	$E(\sigma_a^2)/\sigma^2$	c.v. of σ_a^2	σ_a^2/σ^2
0.1	20	0.011954	2.044	*	1.103
	40	0.006289	1.346	59.7%	1.048
	80	0.003221	1.148	33.4%	1.023
	160	0.001629	1.069	21.4%	1.011
0.25	20	0.013679	1.232	32.3%	1.059
	40	0.007040	1.104	17.5%	1.029
	80	0.003570	1.049	10.9%	1.014
	160	0.001798	1.024	7.2%	1.007
0.5	20	0.017056	1.111	16.1%	1.057
	40	0.008768	1.053	10.0%	1.028
	80	0.004446	1.026	6.6%	1.014
	160	0.002238	1.013	4.5%	1.007
0.75	20	0.024729	1.179	70.9%	1.094
	40	0.012922	1.084	40.6%	1.047
	80	0.006609	1.041	26.5%	1.023
	160	0.003342	1.020	18.2%	1.012
0.9	20	0.039920	*	*	1.224
	40	0.021969	1.280	*	1.112
	80	0.011568	1.127	64.5%	1.056
	160	0.005942	1.061	40.7%	1.028

* The numerical intergration did not converge.

and 3. That is, if the density is estimated, the quality of the estimate is poor.

It should be noted that some of the commonly-used computer programs do not estimate the density as in (8). The estimate of the density used is often that obtained by an arbitrary division of the data into categories (such as in a life table); the estimate used is the sample density of the category in which the desired sample percentile occurs. This density estimate is conditionally biassed depending upon whether the estimate of the percentile in the original scale is at the upper or

Table 4
The true variance and the expected value of the approximation to the variance when the density is estimated by numerical differentiation when N = 160 and P = 0.5.

Distribution	h	σ^2	$E(\sigma_a^2)/\sigma^2$	c.v. of σ_a^2
$\gamma = 0.5$	5	0.01219	1.121	77.0%
	10	0.01219	1.097	59.2%
$\gamma = 1.0$	5	0.006191	1.099	64.7%
(Exponential)	10	0.006192	1.058	45.4%
$\gamma = 2.0$	5	0.002238	1.096	62.3%
	10	0.002238	1.052	42.5%

lower end of the category. (One leads to overestimation of the density and the other to underestimation depending upon the shape of the density.)

When the data are censored and the percentiles are estimated by the Kaplan–Meier [8] survival curve, the c.v. can be expected to be larger than that reported here.

Berry [3] tested by simulation the equality of percentiles of survival distributions using the formula

$$\frac{\text{estimate}_1 - \text{estimate}_2}{(\text{asymptotic variance}_1 + \text{asymptotic variance}_2)^{1/2}}.$$

She found the empirical test sizes were far from the nominal test sizes. Our results explain her findings.

Emerson [6] and Brookmeyer and Crowley [4] propose two methods of obtaining nonparametric confidence intervals for the median when there are censored data. In their simulation studies each of the methods are compared with the parametric methods of Bartholomew [2] and the use of a variance-stabilizing transformation. Their simulations show that as the shape of the underlying distribution differs more from that of the exponential, the average empirical coverages of the confidence intervals produced by the latter methods differ more greatly from the nominal coverage desired. However, the coverages of the nonparametric methods are not severely affected by the choice of the underlying distribution. As expected, the average lengths of the nonparametric intervals exceed those of the parametric intervals when both have the same average coverage. Both consider only intervals for the median although the extension to other percentiles is direct.

Our recommendation is that the asymptotic variance formula *not* be used except for very large sample sizes. Alternate methods, such as those of Emerson and of Brookmeyer and Crowley should be used to obtain confidence limits rather than standard errors.

Acknowledgement

The computer program used for the numerical integration was programmed by Helen Jean Lieverman.

References

[1] M. Abramovitz and I. Stegun, Editors, *Handbook of Mathematical Functions.* National Bureau of Standards Applied Mathematics Series 55 (U.S. Government Printing Office, Washington, DC, 1964).

[2] D.J. Bartholomew, A problem in life testing, *Journal of the American Statistical Association* **52** (1957) 350–355.

[3] O. Berry, Comparison between two life span distributions based on small samples with censored data. MSc Thesis, Tel-Aviv University (1979) (Hebrew with English summary).

[4] R. Brookmeyer and J. Crowley, A confidence interval for median survival time, *Biometrics* **38** (1982) 29–41.
[5] H.A. David, *Order Statistics* (John Wiley, New York, 1970).
[6] J.D. Emerson, Nonparametric confidence intervals for the median in the presence of right censoring, *Biometrics* **38** (1982) 17–27.
[7] A.M. Gross and V. Clark, *Survival Distributions: Reliability Applications in the Biomedical Sciences* (John Wiley, New York, 1975).
[8] E.L. Kaplan and P. Meier, Nonparametric estimates from incomplete observations, *Journal of the American Statistical Association* **53** (1958) 457–481.