

Calculation of the Mean and Variance of Lognormal Data Which Contains Left-Censored Observations

Stephen J. Ganocy
The Goodyear Tire & Rubber Company
Akron, Ohio

Abstract

The mean and variance of data which not only is distributed lognormally but also contains left-censored observations, e.g. measurements which fall below the detection limit of a measuring instrument, can be estimated by means of the delta distribution. Furthermore, using the delta distribution allows estimation of these parameters when the data also contains true zeroes. This paper describes how this distribution can be used for this purpose and presents a SAS[®] macro for doing the calculations.

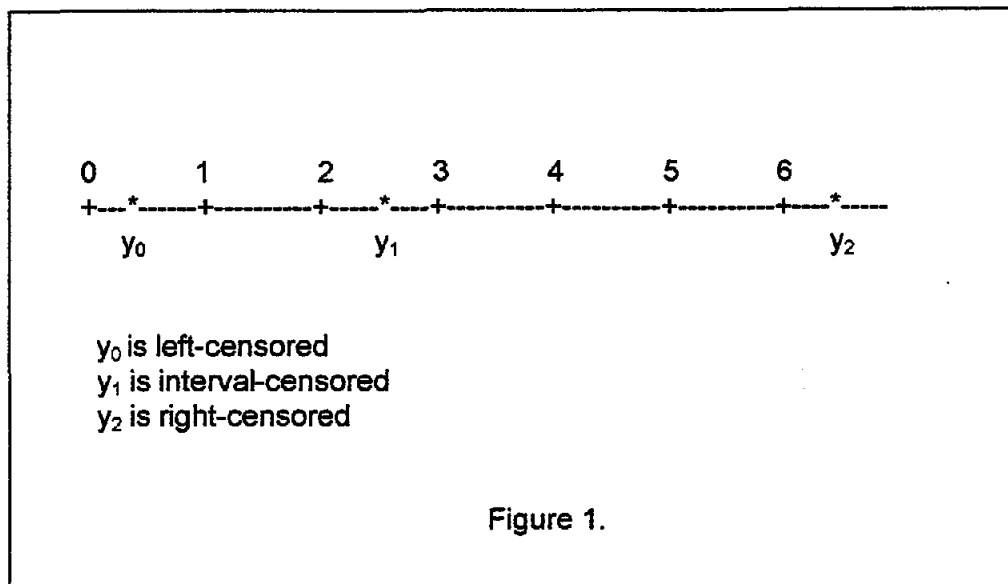
Introduction

There are three basic types of censored data, right-censored, left-censored and interval-censored. In order to illustrate these three types of censoring imagine a device which can record measurements only to the nearest integer value. A ruler which is marked with smallest subdivisions of one inch could be thought of as an example of such a device. Furthermore suppose the total length of the ruler is 6 inches. Note that for this hypothetical device, the 1 inch subdivisions coincide with the integer values.

An object that is $\frac{3}{16}$ " long would be considered a left-censored observation as measured by this hypothetical device; it has some positive length, but the length is only known to fall somewhere between 0 and 1 inch. An object that is longer than 6 inches would be right-censored, i.e. that its length is at least 6 inches. Finally, the length of any object measured by this ruler which is not exact to the nearest inch, and is neither left- or right-censored is called interval-censored. Figure 1 graphically illustrates these concepts.

Time-to-failure data frequently contains right-censored observations. This occurs because not all experimental units necessarily fail within the time allotted to conducting the entire experiment. Furthermore, if the units are only inspected periodically for failure, then some of the observations may also be interval-censored.

Left-censored observations can also occur with time-to-failure data if a failure event happens before the first inspection. Typically, however, this type of censoring is usually associated with observations which are below the detection limit of some measuring device or method. A voltmeter with analog readout is a good example of a device which can generate left-censored observations. Industrial hygiene monitoring techniques are also potential sources of this phenomenon.



From the data analysis perspective the difficulty in handling data with censored observations occurs in the complication of the estimation of parameters in known distributions such as the normal (or Gaussian), Weibull, gamma, etc. The reason for this is that an observation that has a value of say, 10, should not really be given the same weight as one that is somewhere between 9 and 10 . . . or one that is greater than 10. In this case what value should be assigned to an observation that is > 10 ? Should it be 10? 10.5? Infinity?

Fortunately mathematical techniques have been developed and distributions have been explored which enable us to analyze data that has these censoring characteristics.

The Delta Distribution

A distribution which has been studied for analysis of data that contains both true zeroes and left-censored observations is the delta distribution. One other requirement of the data is that the nonzero values must be lognormally distributed. If this condition does not hold then it is not appropriate to use the delta distribution; other techniques may be helpful, however.

Suppose that a data set does satisfy the preceding conditions. Mathematically, if we let W^* denote the nonzero values of a distribution, W , and W^* is lognormal with parameters μ and σ^2 , i.e. $W^* \sim \text{LN}(\mu, \sigma^2)$, and if δ is the proportion of observations with value zero, then $W \sim \Delta(\delta, \mu, \sigma^2)$.

The mean, κ , of this distribution is:

$$\kappa = (1 - \delta) \exp(\mu + \sigma^2/2). \quad (1)$$

Aitchison (1955) provides a minimum variance unbiased estimator (MVUE) of κ which is evidently better than either the moment or maximum likelihood estimators (Owen and DeRouen, 1980).

Given a sample of n observations from $\mathbf{W} \sim \Delta(\delta, \mu, \sigma^2)$ where the sample contains n_0 zero values and $n_1 = n - n_0$ nonzero values and $y_i = \ln(w_i)$, the Aitchison estimator, M_A , of the mean is:

$$M_A = (1 - \hat{\delta}) \exp(\hat{\mu}) \psi_{n_1}(\hat{\sigma}^2 / 2) \quad (2)$$

where,

$$\hat{\delta} = \frac{n_0}{n}$$

$$\hat{\mu} = \frac{1}{n_1} \sum_{i=1}^{n_1} y_i = \bar{y} \quad (3)$$

$$\hat{\sigma}^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (y_i - \bar{y})^2$$

and $\psi_m(x)$ is a Bessel function:

$$\psi_m(x) = 1 + \sum_{k=1}^{\infty} c_k(m) x^k. \quad (4)$$

It can be shown that the coefficients, $c_k(m)$, of x in the infinite series (4) can be expressed recursively as:

$$c_k(m) = \left[\frac{1}{k(m+2k-3)} \right] \frac{(m-1)^2}{m} c_{k-1}(m) \quad k = 1, 2, 3, \dots \quad (5)$$

$$c_0(m) = 1$$

which enables us to obtain an estimate of (4) to any desired precision (within the limits of the computer on which the calculations are being done).

The asymptotic variance of M_A is given by:

$$\text{VAR}(M_A) = \frac{1}{n} \exp(2\hat{\mu} + \hat{\sigma}^2) \left[\hat{\delta}(1 - \hat{\delta}) + \frac{1}{2}(1 - \hat{\delta})(2\hat{\sigma}^2 + \hat{\sigma}^4) \right]. \quad (6)$$

The papers by Aitchison (1955) and Owen & DeRouen (1980) contain further mathematical discussion of all the above subject matter.

Example

This example is the same as that found in Owen & DeRouen (1980) which contains monitoring data for chlorine exposure to workers.

Chlorine Monitoring Data	
Measurement	Chlorine (PPM)
1	6.0
2	0.0
3	6.0
4	9.0
5	6.5
6	0.0
7	0.0
8	0.0
9	1.0
10	0.5
11	2.0
12	2.0
13	0.0
14	0.0
15	1.0

For this data all the zero values are to be counted as left-censored; there are no "true" zeroes in this set. Thus we have $n = 15$, $n_0 = 6$ and $n_1 = 9$.

From the data we can directly calculate (3):

$$\hat{\delta} = 0.4$$

$$\hat{\mu} = 0.9273$$

$$\hat{\sigma}^2 = 1.05683$$

The SAS macro %DELTA yields the following parameter estimates:

$$M_A = 2.37775$$

$$\text{Var}(M_A) = 1.48188$$

which agrees with the values given in Owen & DeRouen (1980).

Invoking asymptotic normality also allows estimation of confidence intervals for M_A , e.g. for this data the 90% confidence interval for M_A is (0.37525, 4.38025).

The following illustrates a short SAS program used to analyze the above data:

```
%include 'c:\sas\mwsug95\delta.sas';
data chlorine;
  input ppm @@;
  if ppm = 0 then ppm = .; /* Set left-censored values to missing! */
  if ppm = . then l_ppm = .;
  else l_ppm = log(ppm);
datalines;
6 0 6 9 6.5 0 0 0 1 0.5 2 2 0 0 1
;
run;
%delta(data=chlorine, yvar=l_ppm);
```

References

Aitchison, J. (1955), "On the Distribution of a Positive Random Variable Having a Discrete Probability Mass at the Origin", *Journal of the American Statistical Association*, 50, 901-908.

Owen, W. J. and DeRouen, T. A. (1980), "Estimation of the Mean for Lognormal Data Containing Zeroes and Left-Censored Values, with Applications to the Measurement of Worker Exposure to Air Contaminants", *Biometrics*, 36, 707-719.

SAS Macro

```
%macro delta(data=_last_  /* Input data set          */
               yvar= );    /* Variable to be analyzed */

proc means data=&data noprint;
  var &yvar;
  output out=_stats(rename=( _freq_ =n)) mean=ybar var=s2 n=m nmiss=r;
run;
proc print data=_stats;
run;
data _delta;
  set _stats;
  delta = r / n;
  x = s2 / 2;
```

```

%psifcn(m,x);

kappa = (1 - delta) * psi * exp(ybar);
ni = 1 / n;
d1 = 1 - delta;
varkappa = ni*(delta*d1+.5*d1*(2*s2+s2*s2))*exp(2*ybar+s2);
lcl95 = kappa - 1.645 * sqrt(varkappa);
output;
keep n r ybar s2 delta m iter psi kappa varkappa lcl95;
run;
proc print data=_delta d;
run;
%mend delta;
%macro psifcn(m,x);
  k = ((&m-1)*(&m-1))/&m;
  iter = 1;
  c = k / (&m-1);
  s = c * &x;
  psi = 1 + s;
STEP:
  iter = iter + 1;
  if iter > 50 then go to fin;
  cnew = (1 / (iter * (&m+2*iter-3))) * k * c;
  s = cnew * (&x**iter);
  psinew = psi + s;
  if abs(psinew - psi) < .000001 then do;
    go to FIN;
  end;
  c = cnew;
  psi = psinew;
  go to STEP;
FIN:
%mend psifcn;

```

Author

Stephen J. Ganocy
 The Goodyear Tire & Rubber Company
 Technical Center
 P. O. Box 3531
 Akron, OH 44309-3531
 E-mail: sjganocy@aol.com

Acknowledgement

SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.