Journal of Modern Applied Statistical Methods

Volume 2 | Issue 2 Article 20

11-1-2003

A Note On MLEs For Normal Distribution Parameters Based On Disjoint Partial Sums Of A Random Sample

W. J. Hurley
Royal Military College of Canada, hurley-w@rmc.ca

Journal of Modern Applied Statistical Methods
November, 2003, Vol. 2, No. 2, 475-477

1538 – 9472/03/$30.00

Brief Reports

A Note On MLEs For Normal Distribution Parameters Based On Disjoint Partial Sums Of A Random Sample

W. J. Hurley
Royal Military College of Canada

Maximum likelihood estimators are computed for the parameters of a normal distribution based on disjoint partial sums of a random sample. It has application in the disaggregation of financial data.

Introduction

Motivation

The Canadian Forces conducts much of its army individual training at the Combat Training Center (CTC) in eastern Canada. Over the 2001-2002 Training Year, 97 serials (a "serial" is an instance of a "course") were run for a total of 2008 students. The overall expenditure on ammunition was $28.8 million. The Commander, CTC, was interested in developing a model of the ammunition dollar cost for each type of course in order to help him assess the risk of over-expending his annual ammunition budget for a given slate of serials. At the point of budgetary deliberations for a given fiscal year, the ammunition cost for any serial is uncertain due primarily to uncertain course enrollments, uncertain student failure rates, and uncertain weather (ranges are closed when it gets dry due to the threat of forest fires).

As a first pass, we conceptualized the ammunition cost of a course as a normal random variable. To estimate its mean and variance, it would be reasonable to use historical data. For some courses this is what we did. However there were some high demand courses where a number of serials were run each year, and unfortunately, ammunition expenditures for these individual serials were aggregated into a single number for the year.

Bill Hurley is Professor of Business Administration. His research interests are military operations research, decision analysis, game theory, logistics modeling and the application of OR to problems in sport. Contact him at hurley-w@rmc.ca

The expenditures for individual serials were not tracked. Hence, for these high demand courses, the problem was to estimate the normal distribution parameters using this aggregated data.

With this background in mind, suppose the ammunition cost for a particular course is a normal random variable with mean μ and variance σ^2. Let

$$X = \{X_1, X_2, ..., X_n\}$$

be an iid sample from this distribution. Unfortunately we cannot observe individual elements of this sample. Rather, we can only observe a sample of disjoint partial sums. Suppose the sample is partitioned into sets $S_1, S_2, ..., S_m$ with cardinalities $k_1, k_2, ..., k_m$ where

$$S_1 \cup S_2 \cup ... \cup S_m = X$$
$$S_i \cap S_j = 0 \quad \textit{for all } i \neq j \quad \textit{and}$$
$$k_1 + k_2 + ... + k_m = n.$$

Let $\kappa(i)$ be the set of indices of the elements of S_i. For instance if $S_2 = \{X_2, X_3, X_7\}$, then $\kappa(i) = \{2,3,7\}$. Then we observe the set of partial sums, $Y = \{Y_1, Y_2, ..., Y_m\}$, where

$$Y_i = \sum_{j \in \kappa(i)} X_j \quad \textit{for } i = 1,2,...,m.$$

We want to compute MLEs for μ and σ^2 using Y rather than X.

There has been a lot of research on grouped and combined datasets. See, for example, the work of Rao (1973). However, to my knowledge, the estimation problem described above has not been mentioned in the literature.

Solution

Note first that Y_i is normally distributed with mean $k_i\mu$ and variance $k_i\sigma^2$. Also, the Y_i are independent since the partial sums are disjoint. Hence the likelihood function is

$$L(\mu,\sigma^2)=\prod_i \frac{1}{\sqrt{2\pi k_i\sigma^2}}\exp\left[-\sum_i \frac{1}{2}\frac{(y_i-k_i\mu)^2}{k_i\sigma^2}\right].$$

Maximizing the ln of this likelihood function gives

$$\hat{\mu}=\bar{y}=\frac{\sum_i y_i}{\sum_i k_i}=\frac{\sum_i y_i}{n}$$

and

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

Note that for the special case $m=n$ (we are working at the level of the iid sample), the last equation returns the usual MLE for variance.

As for the properties of these estimators, the MLE for the mean is unbiased,

$$E\left(\bar{Y}\right)=\mu,$$

but, not surprisingly, the estimator for the variance is biased:

$$E\left(\frac{1}{m}\sum_{i=1}^{m}\frac{(Y_i-k_i\bar{Y})^2}{k_i}\right)=\frac{m-1}{m}\sigma^2.$$

Hence, the estimator

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

is an unbiased estimate of the variance.

Another aspect of this problem is how to revise these estimates as new data becomes available. At the CTC, this new data will not be aggregated. Suppose the new sample is $Z=\{Z_1,Z_2,...,Z_p\}$. What now are the maximum likelihood estimates (MLEs) of μ and σ^2 based on Y and Z? The answer is a straightforward application of the previous development. We simply think of Z_i as an additional element of Y having cardinality $k_i=1$. Hence we have that

$$\hat{\mu}_{Y\cup Z}=\bar{y}^*=\frac{\sum_i y_i+\sum_i z_i}{\sum_i k_i+p}=\frac{\sum_i y_i+\sum_i z_i}{n+p}$$

and

$$\hat{\sigma}^2_{Y\cup Z}=\frac{1}{m-p}\left[\sum_{i=1}^{m}\frac{(y_i-k_i\bar{y}^*)^2}{k_i}+\sum_{j=1}^{p}(z_j-\bar{y}^*)^2\right].$$

Another Example

Returning to the CTC problem, suppose we have the following data set for a given course:

Fiscal Year	***#Serials***	***Total Ammunition Dollars Expended***
2001	3	713,316
2002	2	486,345
2003	3	728,408
2004	3	700,843
2005	2	462,004

The MLEs for the mean and standard deviation are

$$\hat{\mu}=\frac{\sum_i y_i}{\sum_i k_i}=\frac{3{,}090{,}916}{13}=237{,}763$$

and

$$\hat{\sigma}^2 = \frac{1}{m-1}\sum_{i=1}^{m}\frac{(y_i - k_i\bar{y})^2}{k_i}$$
$$= 11,691$$

respectively.

Discussion

This analysis suggests that it would be easy to find maximum likelihood estimators for the parameters of other underlying distributions. The main requirement is to identify the distributions of sums of these random variables.

An interesting extension would be to calculate maximum likelihood estimators in the case where the partial sums overlapped. In this case the Y_i are no longer independent, and hence the likelihood function is more difficult to calculate.

References

Rao, C. R. (1973). *Linear statistical inference and its applications*. NY: Wiley