Trawl survey based abundance estimation using data sets with unusually large catches

R. F. Kappenman


A method for estimating mean catch per unit effort (c.p.u.e.) with groundfish trawl survey data is proposed and examined in detail. The new method is a modification of the arithmetic average of the survey c.p.u.e.s. It is meant to be applied to those cases where one or more of the survey c.p.u.e.s are excessively large and exert too much influence on the arithmetic average, when it is used to estimate mean catch per unit effort. These large c.p.u.e.s are replaced by estimates of their expected values.

Key words: groundfish, survey, c.p.u.e.

Received 16 February 1998; accepted 2 September 1998.

R. F. Kappenman: Alaska Fisheries Science Center, 7600 Sand Point Way NE, Seattle, Washington 98112, USA. Fax: +1 206 526 6723; email: russ.kappenman@noaa.gov

Introduction

There has been some controversy in the literature over how trawl survey catch per unit effort (c.p.u.e.) data should be used to estimate abundance of groundfish and crabs. Generally, the area surveyed is divided into strata. The c.p.u.e. data for any stratum are used to get an estimate of mean c.p.u.e. for the stratum. Abundance for the stratum is then estimated by the product of the area of the stratum and the mean c.p.u.e. estimate. The controversy revolves around how mean c.p.u.e. should be estimated.

Historically, stratum mean c.p.u.e. has been estimated by the arithmetic average of the stratum survey c.p.u.e.s. But in a series of papers, Pennington (1983, 1986, 1991, 1996) has advocated the use of a different estimate of mean c.p.u.e. This estimate is based on the use of the delta distribution as a model for the distribution of c.p.u.e.s. The delta distribution is related to the lognormal distribution in that the lognormal is taken to be the distributional model for the non-zero c.p.u.e.s. With this approach, mean c.p.u.e. is estimated by the minimum variance unbiased estimator of the mean of the delta distribution.

The main reason for considering alternatives to the survey c.p.u.e.s average to estimate mean c.p.u.e. is that this average can be very sensitive to a relatively few inordinately large observed c.p.u.e.s. These large observed c.p.u.e.s, if they occur, can seriously inflate the average observed c.p.u.e. and make it a very poor estimate of mean c.p.u.e. Some examples of this are given by Pennington (1996) who argues for using the delta distribution model approach to alleviate the problem.

However, other authors, such as Myers and Pepin (1990, 1991) and Smith (1990), have contended that the lognormal model may not always be an appropriate model for non-zero c.p.u.e. distributions, because it is not always robust. These authors have pointed out that if very small catches are quite likely to occur, lognormal-based mean estimators can be very poor. They have used the gamma, Weibull, and Birnbaum–Saunders distribution as models to demonstrate this with simulation studies.

The choice of which abundance estimation procedure to use is a very important one, because the two approaches can give markedly different estimates. Two sets of data have been selected to illustrate this. Both data sets result from groundfish trawl surveys conducted in selected subareas of the eastern Bering Sea by the Alaska Fisheries Science Center.

The first set of data, given in Table 1, are c.p.u.e.s for Pacific cod, Gadus macrocephalus, obtained in one of the eastern Bering Sea strata by a 1994 survey. The average of the observations in Table 1 is 52.14, but the value of the minimum variance unbiased estimator for the lognormal distribution mean is 30.84. Thus, one distribution mean estimate is almost 70% larger than the other.

The second set of data, given in Table 2, are the c.p.u.e.s for walleye pollock, Theragra chalcogramma,
Table 1. Pacific cod catch per unit effort (in kg ha$^{-1}$), for one stratum, obtained during a 1994 eastern Bering Sea groundfish trawl survey.

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.35</td>
<td>3.10</td>
<td>4.17</td>
<td>4.23</td>
<td>5.03</td>
<td>5.65</td>
<td>5.66</td>
<td>7.34</td>
<td>8.51</td>
<td>8.71</td>
<td>8.96</td>
</tr>
<tr>
<td></td>
<td>15.22</td>
<td>16.72</td>
<td>17.00</td>
<td>17.08</td>
<td>17.46</td>
<td>17.49</td>
<td>19.45</td>
<td>20.69</td>
<td>21.07</td>
<td>21.67</td>
<td>22.61</td>
</tr>
<tr>
<td></td>
<td>23.18</td>
<td>23.18</td>
<td>23.59</td>
<td>24.59</td>
<td>27.19</td>
<td>28.26</td>
<td>29.78</td>
<td>29.95</td>
<td>31.69</td>
<td>33.61</td>
<td>34.41</td>
</tr>
<tr>
<td></td>
<td>36.70</td>
<td>39.86</td>
<td>40.30</td>
<td>40.48</td>
<td>45.74</td>
<td>47.68</td>
<td>48.22</td>
<td>103.17</td>
<td>115.72</td>
<td>1594.58</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Walleye pollock catch per unit effort (in km ha$^{-1}$), for one stratum, obtained during a 1995 eastern Bering Sea groundfish trawl survey.

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.017</td>
<td>0.020</td>
<td>0.041</td>
<td>0.113</td>
<td>0.195</td>
<td>0.550</td>
<td>0.586</td>
<td>0.823</td>
<td>1.280</td>
<td>2.063</td>
<td>2.412</td>
</tr>
<tr>
<td></td>
<td>31.132</td>
<td>32.021</td>
<td>35.192</td>
<td>35.971</td>
<td>36.070</td>
<td>40.638</td>
<td>41.627</td>
<td>44.069</td>
<td>45.380</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>54.739</td>
<td>55.360</td>
<td>56.800</td>
<td>57.836</td>
<td>60.309</td>
<td>61.285</td>
<td>61.474</td>
<td>63.562</td>
<td>64.995</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>69.712</td>
<td>79.314</td>
<td>79.740</td>
<td>90.574</td>
<td>94.854</td>
<td>103.855</td>
<td>105.352</td>
<td>111.010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>126.350</td>
<td>130.574</td>
<td>155.999</td>
<td>230.658</td>
<td>263.290</td>
<td>268.600</td>
<td>273.620</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>340.460</td>
<td>413.349</td>
<td>464.520</td>
<td>489.493</td>
<td>492.956</td>
<td>954.173</td>
<td>1486.489</td>
<td>3802.413</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

obtained in one stratum by a 1995 survey. The average of the observations in Table 2 is 191.33. The value of the minimum variance unbiased estimator of the lognormal distribution mean is 717.74. The latter c.p.u.e. mean estimate is 3.8 times the former.

When the topic of interest is estimation of mean c.p.u.e., one can confine attention to estimation of the mean of the distribution of the non-zero c.p.u.e.s. Suppose there are zero catches in the survey, as is most often the case, and one has an estimate of the mean of the distribution of the non-zero c.p.u.e.s. Then the estimate of the overall mean c.p.u.e. is simply the product of the non-zero c.p.u.e. distribution mean estimate and the ratio of the number of non-zero catches to the total number of tows. Thus throughout the remainder of this paper, attention will be restricted to estimation of the mean of the distribution of the non-zero c.p.u.e.s, and, from now on, c.p.u.e.s will be assumed to be positive.

The controversy, alluded to earlier, presents the following question. When estimating abundance, should one use the lognormal model for the unknown c.p.u.e. distribution and estimate the distribution mean using this model or should one use the average of the observed c.p.u.e.s to estimate the distribution mean?

There are situations in which one alternative or the other may be a very poor one, and there are situations in which both alternatives can be very poor ones. A few unusually large c.p.u.e.s can make the average of the observed c.p.u.e.s a very poor distribution mean estimate, but a number of very small c.p.u.e.s have no major adverse effect. On the other hand, a few very small c.p.u.e.s can make the lognormal distribution based mean estimate a very poor one. Further, if both a few quite small and a few unusually large c.p.u.e.s are observed, both estimates can be very poor. A major problem with both mean estimators is the lack of suitable answers to the questions, how large is too large and how small is too small?

Thus it appears to be quite evident that good use could be made of another alternative for mean c.p.u.e. estimation, if the alternative overcomes the drawbacks of those estimators currently being used. The purpose of this paper is to promote one for use in certain situations. The mean estimator advocated here is meant to be applied when a set of observed c.p.u.e.s is examined and it is found to contain a few suspiciously large c.p.u.e.s.

This estimator is a modification of the arithmetic average, and it is not sensitive to c.p.u.e.s that are too small or too large. Further, the set of observed c.p.u.e.s itself, dictates whether or not the estimator should be used, that is, whether or not the arithmetic average of the c.p.u.e.s should be modified.

It is to be emphasized that the mean estimator proposed here is not meant to be applied routinely to all sets of trawl survey c.p.u.e. data. It has been developed for use only for those cases where one or more c.p.u.e.s in a data set are suspected of being excessively large and of exerting too much influence on the arithmetic average of the c.p.u.e.s, when this average is used to estimate mean c.p.u.e.

A new mean c.p.u.e. estimator

Let $x_1, \ldots, x_n$ represent a set of $n$ non-zero c.p.u.e.s that are obtained from a trawl survey, and suppose that the $x_i$s are ordered. That is, $x_1 < x_2 < \ldots < x_n$. Statisticians would refer to the $x_i$s as the order statistics for a random sample of size $n$ from some unknown distribution.

One possible estimator of the distributions mean is the sample mean:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$  \hspace{1cm} (1)

Suppose that, after examining the set of c.p.u.e.s represented by the $x_i$s, it is found that the $j$ largest $x_i$s, $x_{i-j+1}, \ldots, x_n$, are suspected of being excessively large and of exerting too much influence on $\bar{x}$, when it is used to estimate mean c.p.u.e. Here $j$ is some small positive integer. By excessively large, it is meant large enough to cause $\bar{x}$ to be a poor estimate of mean c.p.u.e. An example of this is given by the data in Table 1. An examination of that data leads one to suspect that the value of the largest order statistic, 1594.58, may be excessively large. Here $j=1$.

Other possible cases in point are given by the data in Tables 3 and 4. The data given in Table 3 were gathered during a trawl survey for groundfish in the Gulf of
The modification of \( \bar{x} \), proposed here for use in estimating mean c.p.u.e., is that of replacing \( x_{n-j+1}, \ldots, x_n \), in the sum in Equation (1), by estimates of the expected values of these \( j \) order statistics. The estimates are based on all of the data, \( x_1, \ldots, x_n \), obtained, the known mathematical expressions for expectations of order statistics, and on a probability density function estimator known as the kernel estimator. This kernel density function estimator is discussed in the Appendix and in detail by Silverman (1986). The expected value estimates are non-parametric. That is, no distribution is assumed for the sampled c.p.u.e. distribution. This is in contrast with Pennington (1983, 1986, 1991, 1996), who assumes that the lognormal distribution is appropriate.

The modified c.p.u.e. average mean estimator proposed here enjoys the property of not being adversely affected by a few very small c.p.u.e.s, like the c.p.u.e. average does. In addition, inordinately large c.p.u.e.s are replaced, in the sample average, by much more moderate values. Thus the effect of these large c.p.u.e.s on the sample average is often reduced considerably.

### Some examples

To illustrate the latter point, consider the data given in Tables 1, 3, and 4. Table 1 contains 55 observed c.p.u.e.s. Attention is focused on the order statistic \( x_{55} \), because the observed value of this order statistic is suspiciously large. The estimate of the expected value of the fifty-fifth order statistic, for the random sample of size 55 given in Table 1, is 185.01. The new estimate of mean c.p.u.e. would be the average of the first 54 c.p.u.e.s in Table 1 and 185.01. This average is 26.52, whereas the average of all 55 c.p.u.e.s in Table 1 is 52.14.

For the Atka mackerel data, given in Table 3, there are 56 non-zero c.p.u.e.s and interest is focused on \( x_{55} \) and \( x_{56} \), because the observed values of these order statistics appear to be intractably large. The non-parametric estimate of the expected values of these two order statistics are 1712.55 and 2981.50. The corresponding estimate of mean c.p.u.e. is the average of the first 54 c.p.u.e.s and these two expected value estimates. This average is 209.54. The average of all 56 c.p.u.e.s in Table 3 is 1071.55. This is an example of an extreme case where the sample average could be seriously inflated by a few inordinately large c.p.u.e.s.

There are 72 lingcod c.p.u.e.s in Table 4, and the last two are suspiciously large. The estimates of the expected values of \( x_{71} \) and \( x_{72} \), are 54.42 and 100.20, respectively. The average of the first 70 observations in Table 4 and these two expected value estimates is 5.79. The average of all 72 c.p.u.e.s in Table 3 is 15.92.

Another important property of the modified c.p.u.e. average mean estimator proposed here is that it calls for
no adjustment of the sample average, when the sample itself indicates that none should be made. The sample indicates this if the values of the estimates of the expected values of \( x_{n-j+1}, \ldots, x_n \) exceed the observed values of these \( j \) order statistics. If the estimated expected values of these \( j \) order statistics exceed the corresponding observed values, the sample average would be used, unmodified, to estimate mean c.p.u.e.

For example, consider the pollock survey c.p.u.e. data given in Table 2. There are 61 c.p.u.e.s. A non-parametric estimate of the expected value of \( x_{61} \) is 3960.080, while the observed value of this order statistic is 3802.413. Since the estimated expected value is greater than the observed value of the largest order statistic, the sample average would be used unmodified to estimate mean c.p.u.e. This average is 191.33.

### Statistical details

The non-parametric estimators for the expected values of the last \( j \) order statistics, based upon the observed values of the order statistics \( x_1, \ldots, x_n \), are obtained by an iterative procedure. The iterative procedure has been programmed in Visual Basic and APL, and a copy of either program may be obtained from the email address: Tom.Wilderbuer@noaa.gov. The Visual Basic program is in an executable file with appropriate documentation.

For those with a sufficient statistical background, a brief explanation of and a motivation for the procedure will now be given.

The first step in the procedure uses all of the \( x_i \)s to obtain initial estimates for the expected values of the last \( j \) order statistics \( x_{n-j+1}, \ldots, x_n \). If the expected values estimates are larger than the observed values of the last \( j \) order statistics, the procedure is terminated. Otherwise, for the second step, \( x_{n-j+1}, \ldots, x_n \) are replaced, in the sample, by estimates of their expected values, the first step is repeated, and revised estimates for the expected values of the last \( j \) order statistics are obtained. This process is repeated and, at each step, \( x_{n-j+1}, \ldots, x_n \) are replaced by the new \( j \) expected value estimates obtained by the previous step. The iterative procedure is stopped when the changes in the \( j \) expected value estimates are sufficiently small. At each step in the procedure, the first \( n-j \) order statistics play the major role in obtaining estimates of the expected values of the last \( j \) order statistics. The latter order statistics play a much lesser role.

Each step in the iterative procedure makes use of the following calculations. An estimator for the expected value of the \( i \)th order statistic (\( i = n - j + 1, \ldots, n \)) is:

\[
\hat{E}(x_i) = \exp \left\{ \frac{n!}{(n-i)!} \int_0^\infty \frac{u^{n-i-1}}{\{(1 - e^{-y})^{-1}\}^{n-i}} G(u)^{n-i} du \right\} 
- G(u)^{n-i-1}g(u) \, du 
\]

where:

\[
g(u) = \frac{1}{n} \sum_{i=1}^n e^{-u-y_i} \left( 1 - e^{-u/y_i} \right)^{-2} 
(3)
\]

\[
G(u) = \frac{1}{n} \sum_{i=1}^n \left( 1 + e^{-u/y_i} \right)^{-1} 
(4)
\]

and \( y_i = \ln x_i \), for \( i = 1, \ldots, n \). The number \( h \) in Equations (3) and (4) is the value of \( h \) which maximizes:

\[
\sum_{i=1}^{n-j} \ln g(y_i) + j \ln \left[ 1 - G_{n-j}(y_{n-j}) \right] 
(5)
\]

where:

\[
g(y_i) = \frac{1}{h(n-1)} \sum_{k=1, k \neq i}^n e^{-y_i - y_k} \left( 1 - e^{-y_i/y_k} \right)^{-2} 
(6)
\]

and:

\[
G_{n-j}(y_{n-j}) = \frac{1}{n-1} \sum_{k=1, k \neq n-j}^n \left( 1 + e^{y_n/y_k} \right)^{-1} 
(7)
\]

Numerical integration must be used to evaluate the integral in Equation (2), and Hermite integration works very well.

A motivation for Equation (2) being an estimator of the expected value of the \( i \)th order statistic is the following one. Equation (3) gives a non-parametric, kernel-type estimator of the probability density function (p.d.f.) of the random variable \( y = \ln x \), and Equation (4) gives the corresponding distribution function (d.f.) estimator. If Equations (3) and (4) were the actual p.d.f. and d.f. of \( y \), the quantity inside the braces on the right-hand side of Equation (2) would be the expected value of the \( i \)th order statistic for a random sample of observations of \( y \). However, this quantity is, in fact, merely an estimator of the expected value, because \( g \) and \( G \) are estimators of the p.d.f. and d.f. A natural estimator of the expected value of the \( i \)th order statistic for a random sample of observations of \( x = \exp y \) is given by Equation (2).

The quantity \( h \) in Equations (3) and (4) is commonly referred to as a smoothing parameter. Finding the value of \( h \) which maximizes Equation (5) is an ad hoc smoothing parameter value selection procedure. The quantity Equation (5) is an estimate of the logarithm of the joint p.d.f. of the first \( n-j \) order statistics for a random sample of \( n \) observations of the random variable \( y \). The quantities Equations (6) and (7) are cross-validation estimates of the p.d.f. of \( y \) evaluated at \( y = y_i \) and the d.f. of \( y \) evaluated at \( y = y_{n-j} \), respectively.
The iterative procedure for estimating a distribution mean can be summarized by the following algorithm:

1. Set \( m_1 = \frac{1}{n} \sum_{i=1}^{n} x_i \)
2. Compute \( \hat{E}(x_{n-j+1}), \ldots, \hat{E}(x_n) \), where \( \hat{E}(x_i) \), for \( i = n-j+1, \ldots, n \), is given by Equation (2).
3. Compute:
   \[
   m_2 = \frac{1}{n} \left( \sum_{i=1}^{n-j} x_i + \sum_{k=1}^{j} \hat{E}(x_{n-k+1}) \right)
   \]
   If \( m_1 - m_2 \) is sufficiently small go to step 5. Otherwise, go to step 4.
4. Replace \( x_{n-j+1}, \ldots, x_n \) with \( \hat{E}(x_{n-j+1}), \ldots, \hat{E}(x_n) \).
   Set \( m_1 = m_2 \), and go to step 2.
5. If \( m_1 - m_2 \) is negative, the distribution mean estimate is \( m_1 \). Otherwise, it is \( m_2 \). Stop.

An assessment of estimator performance

It is emphasized that the mean estimator, described in the previous section, is not meant to be applied routinely to all sets of trawl survey strata c.p.u.e.s. It has been developed for use only for those cases where one or more c.p.u.e.s are suspected of being excessively large and of exerting too much influence on the c.p.u.e.s arithmetic average when it is used to estimate mean c.p.u.e. Examples of cases where the estimator should be applied are given by the data in Tables 1, 3, and 4.

In order to demonstrate the potential of the new distribution mean estimator for producing markedly improved mean estimates, some simulation studies were performed. Random samples, like those to which the estimator is meant to be applied, were generated from various skewed to the right distributions with heavy right tails. All indications are that the c.p.u.e. random sample obtained, and the new mean estimation procedure was applied to each sample obtained. The averages of the sample average and the root mean squared errors of the two mean estimators over the 1000 “appropriate” samples were determined.

It was found that, for all sampled distribution-sample size combinations, the averages of the new mean estimates were much closer to the distribution mean than were the averages of the sample averages, and the root mean squared errors of the new mean estimates were all much smaller than the root mean squared errors of the sample averages. In fact, the efficiency of the sample average, relative to the new mean estimator, ranged only from about 4% to about 63%.

Simulation study details

The sampling distributions used in the study were the Pareto, f, lognormal, Burr, and log-Laplace distributions. In each case, the scale parameter was taken to be unity. The p.d.f.s and means for these distributions are given in the appendix.

Appropriate random samples were generated from these sampling distributions by making use of a statistical technique for finding unusual spaces or gaps in random samples. The technique is known as gapping. A summary of it is given in the appendix and a complete discussion may be found in Tukey (1971).

Random samples were generated, from the sampling distributions used, until a sample, which contained at least one standardized weighted gap that was at least 10, was found. When such a sample was obtained, the mean estimator \( \hat{\mu} \), given by the algorithm of the previous section, and the sample mean, \( \bar{x} \), were calculated. The number \( j \) (i.e. the number of suspiciously large observations) was determined by counting the number of observations not less than the smallest observation corresponding to a standardized weighted gap of at least 10. This process was repeated until 1000 “appropriate” samples were obtained for each sampling distribution-sample size combination considered.

The results of the simulation studies are summarized by Table 5. For each sampled distribution-sample size combination, this table gives the distribution mean, the averages of \( \bar{x} \) and \( \hat{\mu} \) over the 1000 samples, the root mean squared errors of \( \bar{x} \) and \( \hat{\mu} \) over the 1000 samples, and the ratios of the root mean squared errors of \( \hat{\mu} \) to the corresponding root mean squared errors of \( \bar{x} \).

Discussion

The mean c.p.u.e. estimator advocated here is meant to be applied when a set of observed c.p.u.e.s is examined, and it is found to contain a few suspiciously large c.p.u.e.s. It has been my personal experience that an experienced fish biologist, who is responsible for conducting a trawl survey and/or an analysis of the results, will generally know when one or more unusually large c.p.u.e.s are obtained. If this is not the case, one reasonably good, but informal, rule for applying the
estimator would be to use it when at least one standardized weighted gap of at least 10 is obtained when gapping is applied to the sample. The simulation study discussed previously demonstrates the potential of this rule for producing good results.

When it is appropriately applied, the estimator discussed here overcomes the drawbacks associated with using the arithmetic average of the observed c.p.u.e.s and using the delta distribution-based mean c.p.u.e. estimator. That is, it is not sensitive to a few inordinately large c.p.u.e.s, as the arithmetic average is, and it is not sensitive to a few very small c.p.u.e.s, as is the delta distribution-based estimator. Further, the latter estimator assumes a parametric model for the c.p.u.e. distribution. The estimator discussed here is completely non-parametric. One more important point is the following

Table 5. Averages, root mean squared errors, and the ratios of the root mean squared errors for $\bar{x}$, the sample mean, and $\hat{\mu}$, the distribution mean estimator proposed in this paper.

<table>
<thead>
<tr>
<th>Sampled distribution</th>
<th>Mean A ($\bar{x}$)</th>
<th>Mean A ($\hat{\mu}$)</th>
<th>RMSE A ($\bar{x}$)</th>
<th>RMSE A ($\hat{\mu}$)</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P (2.5)</td>
<td>0.67</td>
<td>1.60</td>
<td>0.96</td>
<td>1.30</td>
<td>0.58</td>
</tr>
<tr>
<td>P (2.0)</td>
<td>1.00</td>
<td>2.72</td>
<td>1.48</td>
<td>3.25</td>
<td>0.52</td>
</tr>
<tr>
<td>P (1.5)</td>
<td>2.00</td>
<td>5.98</td>
<td>2.68</td>
<td>11.58</td>
<td>0.16</td>
</tr>
<tr>
<td>f (5.5)</td>
<td>1.67</td>
<td>4.04</td>
<td>1.86</td>
<td>3.40</td>
<td>0.68</td>
</tr>
<tr>
<td>f (4.4)</td>
<td>2.00</td>
<td>5.37</td>
<td>2.34</td>
<td>6.32</td>
<td>0.10</td>
</tr>
<tr>
<td>f (3.3)</td>
<td>3.00</td>
<td>10.28</td>
<td>3.38</td>
<td>47.64</td>
<td>0.24</td>
</tr>
<tr>
<td>LN (1)</td>
<td>1.65</td>
<td>3.07</td>
<td>1.65</td>
<td>1.70</td>
<td>0.50</td>
</tr>
<tr>
<td>LN (1.5)</td>
<td>3.08</td>
<td>6.61</td>
<td>3.88</td>
<td>5.33</td>
<td>0.45</td>
</tr>
<tr>
<td>LN (2)</td>
<td>7.39</td>
<td>13.62</td>
<td>8.92</td>
<td>14.03</td>
<td>0.52</td>
</tr>
<tr>
<td>B (2.5)</td>
<td>1.32</td>
<td>3.00</td>
<td>1.39</td>
<td>2.18</td>
<td>0.13</td>
</tr>
<tr>
<td>B (2.0)</td>
<td>1.57</td>
<td>4.13</td>
<td>1.73</td>
<td>3.96</td>
<td>0.59</td>
</tr>
<tr>
<td>B (1.5)</td>
<td>2.42</td>
<td>7.72</td>
<td>2.72</td>
<td>11.58</td>
<td>1.49</td>
</tr>
<tr>
<td>LL (0.35)</td>
<td>1.14</td>
<td>2.15</td>
<td>1.17</td>
<td>1.33</td>
<td>0.16</td>
</tr>
<tr>
<td>LL (0.50)</td>
<td>1.33</td>
<td>3.21</td>
<td>1.39</td>
<td>2.57</td>
<td>0.31</td>
</tr>
<tr>
<td>LL (0.65)</td>
<td>1.73</td>
<td>5.23</td>
<td>1.77</td>
<td>6.79</td>
<td>0.62</td>
</tr>
<tr>
<td>n=50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P (2.5)</td>
<td>0.67</td>
<td>1.21</td>
<td>0.86</td>
<td>0.75</td>
<td>0.34</td>
</tr>
<tr>
<td>P (2.0)</td>
<td>1.00</td>
<td>1.95</td>
<td>1.24</td>
<td>1.70</td>
<td>0.49</td>
</tr>
<tr>
<td>P (1.5)</td>
<td>2.00</td>
<td>4.08</td>
<td>2.19</td>
<td>6.00</td>
<td>0.15</td>
</tr>
<tr>
<td>f (5.5)</td>
<td>1.67</td>
<td>3.01</td>
<td>1.80</td>
<td>1.92</td>
<td>0.41</td>
</tr>
<tr>
<td>f (4.4)</td>
<td>2.00</td>
<td>3.84</td>
<td>2.23</td>
<td>3.22</td>
<td>0.67</td>
</tr>
<tr>
<td>f (3.3)</td>
<td>3.00</td>
<td>6.44</td>
<td>3.20</td>
<td>10.28</td>
<td>1.22</td>
</tr>
<tr>
<td>LN (1)</td>
<td>1.65</td>
<td>2.54</td>
<td>1.70</td>
<td>1.06</td>
<td>0.35</td>
</tr>
<tr>
<td>LN (1.5)</td>
<td>3.08</td>
<td>5.06</td>
<td>3.74</td>
<td>3.00</td>
<td>1.54</td>
</tr>
<tr>
<td>LN (2)</td>
<td>7.39</td>
<td>10.77</td>
<td>8.27</td>
<td>9.30</td>
<td>4.58</td>
</tr>
<tr>
<td>B (2.5)</td>
<td>1.32</td>
<td>2.23</td>
<td>1.38</td>
<td>1.16</td>
<td>0.20</td>
</tr>
<tr>
<td>B (2.0)</td>
<td>1.57</td>
<td>2.95</td>
<td>1.68</td>
<td>2.08</td>
<td>0.37</td>
</tr>
<tr>
<td>B (1.5)</td>
<td>2.42</td>
<td>5.27</td>
<td>2.48</td>
<td>6.84</td>
<td>0.86</td>
</tr>
<tr>
<td>LL (0.35)</td>
<td>1.14</td>
<td>1.69</td>
<td>1.18</td>
<td>0.67</td>
<td>0.12</td>
</tr>
<tr>
<td>LL (0.50)</td>
<td>1.33</td>
<td>2.40</td>
<td>1.37</td>
<td>1.83</td>
<td>0.22</td>
</tr>
<tr>
<td>LL (0.65)</td>
<td>1.73</td>
<td>3.74</td>
<td>1.69</td>
<td>7.69</td>
<td>0.41</td>
</tr>
<tr>
<td>n=100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P (2.5)</td>
<td>0.67</td>
<td>0.98</td>
<td>0.79</td>
<td>0.42</td>
<td>0.21</td>
</tr>
<tr>
<td>P (2.0)</td>
<td>1.00</td>
<td>1.54</td>
<td>1.19</td>
<td>0.94</td>
<td>0.36</td>
</tr>
<tr>
<td>P (1.5)</td>
<td>2.00</td>
<td>3.15</td>
<td>2.13</td>
<td>3.30</td>
<td>0.71</td>
</tr>
<tr>
<td>f (5.5)</td>
<td>1.67</td>
<td>2.39</td>
<td>1.77</td>
<td>0.96</td>
<td>0.29</td>
</tr>
<tr>
<td>f (4.4)</td>
<td>2.00</td>
<td>3.14</td>
<td>2.14</td>
<td>2.63</td>
<td>0.46</td>
</tr>
<tr>
<td>f (3.3)</td>
<td>3.00</td>
<td>4.55</td>
<td>3.05</td>
<td>4.21</td>
<td>0.91</td>
</tr>
<tr>
<td>LN (1)</td>
<td>1.65</td>
<td>2.17</td>
<td>1.70</td>
<td>0.62</td>
<td>0.24</td>
</tr>
<tr>
<td>LN (1.5)</td>
<td>3.08</td>
<td>4.18</td>
<td>3.50</td>
<td>1.67</td>
<td>1.05</td>
</tr>
<tr>
<td>LN (2)</td>
<td>7.39</td>
<td>9.08</td>
<td>7.78</td>
<td>5.53</td>
<td>3.31</td>
</tr>
<tr>
<td>B (2.5)</td>
<td>1.32</td>
<td>1.84</td>
<td>1.36</td>
<td>0.67</td>
<td>0.15</td>
</tr>
<tr>
<td>B (2.0)</td>
<td>1.57</td>
<td>2.38</td>
<td>1.64</td>
<td>1.24</td>
<td>0.26</td>
</tr>
<tr>
<td>B (1.5)</td>
<td>2.42</td>
<td>3.95</td>
<td>2.38</td>
<td>3.85</td>
<td>0.56</td>
</tr>
<tr>
<td>LL (0.35)</td>
<td>1.14</td>
<td>1.45</td>
<td>1.17</td>
<td>0.39</td>
<td>0.08</td>
</tr>
<tr>
<td>LL (0.50)</td>
<td>1.33</td>
<td>1.93</td>
<td>1.35</td>
<td>1.01</td>
<td>0.16</td>
</tr>
<tr>
<td>LL (0.65)</td>
<td>1.73</td>
<td>2.79</td>
<td>1.66</td>
<td>3.79</td>
<td>0.30</td>
</tr>
</tbody>
</table>
one. The delta distribution based mean c.p.u.e. estimator will be severely negatively biased if it is routinely applied to samples which come from skewed to the right distributions with heavier right tails than lognormal right tails.

Finally, the lognormal distribution-based mean estimator was not included in the simulation studies discussed here. The reason for this is that this estimator performs extremely poorly, relative to the sample arithmetic average, if the sampling distribution is one of certain members of the gamma, Weibull, Birnbaum–Saunders, and inverse Gaussian distributions families. The authors’ own simulation studies, as well as those of Myers and Pepin (1990), have revealed this. On the other hand, the sample average is an excellent estimator of the mean of these distributions. In fact, it is the minimum variance unbiased estimator of gamma and inverse Gaussian distribution means, and simulation studies that have been conducted indicate that it is close to optimal for estimating Weibull and Birnbaum–Saunders distribution means. Further, intractably large observations very seldom, if ever, occur when sampling is from any of these four distributional families. Thus, the sample average would almost always be used to estimate the distribution mean, if the mean estimation method advocated here is used.

Acknowledgements

The author would like to thank the editor Dr Michael Chadwick and two anonymous referees for their careful reading of the paper and useful suggestions. The author is also very grateful to Dr Jerry Reeves for his development of the Visual Basic program and allowing distribution of the program to anyone interested in using it.


References


Appendix

I. Kernel p.d.f. estimation

The kernel estimate:

\[ \hat{g}(y) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{y - y_i}{h} \right) \]  

(8)

where h is a positive scalar and K is a symmetric p.d.f., is a popular non-parametric estimate of a p.d.f., \( g(y) \), of a random variable \( y \) based on a random sample, \( y_1, \ldots, y_n \), of observations of \( y \). The corresponding kernel estimate of a distribution function is found by integrating Equation (8).

Epanechnikov (1969) showed that although an optimal K can be identified theoretically, the performance of many other appealing kernels are virtually as good as the best, when estimation of a p.d.f. is of concern. Jones (1990) made analogous observations for kernel distribution function estimation. The kernel used here is the logistic one:

\[ K(y) = \frac{e^{-y}}{(1 + e^{-y})^2} \]

primarily because of convenience. Its associated distribution function is in closed form, and it is convenient for estimating a distribution function. The associated distribution function is \( (1 + e^{-y})^{-1} \). Thus Equations (3) and (4) are kernel-type estimators of the probability density function and the distribution function based on a random sample, \( y_1, \ldots, y_n \) and the logistic kernel.

II. Sampling distributions

The sampling distributions used in the simulation studies were the following ones:

1. The Pareto distribution with p.d.f. of the form:

\[ f(x) = \frac{c}{(x + 1)^{c+1}} \]

mean \( 1/(c - 1) \), and denoted by \( P(c) \).
2. The F distribution with p.d.f. of the form:

\[
f(x) = \frac{\Gamma\left(\frac{c_1}{2}\right) \Gamma\left(\frac{c_2}{2}\right) x^{c_1-1}}{\Gamma\left(\frac{c_1 + c_2}{2}\right)} \left(1 + \frac{c_1}{c_2} x^{c_2/c_2 - 2}\right)^{-\frac{c_1 + c_2}{2}}.
\]

mean \(c_2/(c_2 - 2)\), and denoted by \(F(c_1, c_2)\).

3. The lognormal distribution with p.d.f. of the form:

\[
f(x) = \frac{1}{\sqrt{2\pi \sigma^2 x}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}
\]

mean \(e^{\mu^2/2}\), and denoted by \(\text{LN}(\sigma)\).

4. The Burr distribution with p.d.f. of the form:

\[
f(x) = \frac{cx^{c-1}}{(1+x^c)^2}
\]

mean \(\Gamma(1-1/c) \Gamma(1+1/c)\), and denoted by \(B(c)\).

5. The log-Laplace distribution with p.d.f. of the form:

\[
f(x) = \frac{1}{2bx} e^{-\frac{|\ln x|}{b}}
\]

mean \(1 - b^2\), and denoted by \(\text{LL}(b)\).

III. Gapping

Appropriate samples were obtained for the simulation studies by making use of a statistical technique, due to Tukey (1971), for finding unusual spaces or gaps in random samples. The technique, known as gapping, may be described as follows. Suppose that \(x_1, \ldots, x_n\) represent the order statistics for a random sample of size \(n\). Set:

\[
z_{n-i+1} = (n-i)(x_{n-i+1} - x_{n-i})^{1/2}, \text{ for } i = 1, \ldots, n - 1
\]

Let \(z_{25}\) represent the 25% trimmed mean of \(z_2, \ldots, z_n\). That is, \(z_{25}\) is calculated by trimming-off the 25% smallest weighted gaps and the 25% largest, and then taking the mean of the remaining middle 50%. Finally, set \(w_i = z_i/z_{25}\), for \(i = 2, \ldots, n\). This standardizes the distribution of the weighted gaps around a middle value of 1.

Suspiciously large observations in a sample should correspond to large standardized weighted gaps. For example, \(w_{55} = 15.59\), for the data of Table 1, \(w_{56} = 37.75\) and \(w_{56} = 14.26\), for the data of Table 3, and \(w_{71} = 15.92\) and \(w_{72} = 20.11\), for the data of Table 4.