Estimating Geometric Means Using Data from a Complex Survey Sampling Design

Overview

Geometric means are widely used in a variety of scientific disciplines. They are the natural parameter of interest for a lognormal random variable because a ratio of lognormal random variables has a known lognormal distribution, and the geometric mean of a lognormal ratio is equal to the ratio of the individual geometric means. Some common uses for the geometric mean with survey data include estimating average population growth rates, bacterial contamination rates, and chemical concentration rates.

If you use SAS/STAT® 12.1 or later, you can estimate geometric means by specifying the ALLGEO statistic keyword in the PROC SURVEYMEANS statement. If you use a release prior to SAS/STAT 12.1, none of the SAS/STAT survey procedures directly compute geometric means. However, with a little programming you can still estimate a geometric mean and its variance from sample survey data.

The SAS source code for this example is available as an attachment in a text file. In Adobe Acrobat, right-click the icon in the margin and select Save Embedded File to Disk. You can also double-click to open the file immediately.

Analysis

Following Wolter (1985), suppose \( \bar{Y} \) denotes the population mean of a characteristic \( y \) and \( \bar{y} \) denotes an estimator of \( \bar{Y} \) based on a sample of fixed size \( n \). The natural estimator for the exponential function \( \theta = e^{\bar{Y}} \) is

\[
\hat{\theta} = e^{\bar{y}}
\]

Suppose \( v(\bar{y}) \) denotes an estimator of the variance of \( \bar{y} \) that is appropriate to the particular sampling design. Then, the Taylor series estimator of variance is

\[
v(\hat{\theta}) = e^{2\bar{y}} v(\bar{y})
\]

These results can be applied directly to the problem of estimating the geometric mean of a finite population characteristic \( x \) because the geometric mean is the exponentiation of the mean of the natural logarithm. That
is, the geometric mean can be expressed as a finite population quantity by

\[ \theta = \exp \left( \frac{1}{N} \sum_{i=1}^{N} \ln(x_i) \right) \]

where \( x_i > 0 \) for \( i = 1, \ldots, n \). Substituting \( y_i = \ln(x_i) \), then \( \theta = e^{\bar{y}} \) and the natural estimator of \( \theta \) is \( \hat{\theta} = e^{\bar{y}} \).

You can estimate the variance of \( \hat{\theta} \) by

\[ v(\hat{\theta}) = \hat{\theta}^2 v(\bar{y}) \]

where the estimator \( v(\bar{y}) \) is appropriate to the particular sampling design and estimator and is based on the variable \( y = \ln(x) \). You can use PROC SURVEYMEANS to estimate the arithmetic mean of \( \ln(x) \); then using the result, you can compute the geometric mean of \( x \). Similarly, you can use the variance of the arithmetic mean of \( \ln(x) \) and the computed estimate of the geometric mean of \( x \) to compute the variance and standard error of the geometric mean of \( x \).

---

**Example**

This example uses hypothetical data that represents survey results from an inspection of meat processing plants. The sampling plan is designed to estimate the levels of the *Clostridium perfringens* bacteria in processed ground beef. The survey has a stratified, two-stage sampling design and includes the variables PLANT, SHIFT, CLOSTRIDIUM, and WGT. PLANT identifies the strata, SHIFT identifies the clusters, WGT contains the sampling weights, and CLOSTRIDIUM measures the levels of bacteria per gram of product.

```plaintext
data example;
  input plant shift clostridium wgt;
datalines;
  1 1  60   10.3676
  1 1 129  11.4145
  1 1   5  10.3055
  1 2  159 10.3626
  1 2   38 11.1399
  1 2   72 10.7285
  2 1 6335 11.2469
  2 1  605 10.7137
  2 1   1 11.4907
```

After you put your data in a SAS data set, generate a variable that contains the natural logarithm of the variable CLOSTRIDIUM:

```
data example;
  set example;
  logClostridium = log(clostridium);
run;
```

Use PROC SURVEYMEANS to estimate the arithmetic mean of the newly generated variable LOG-CLOSTRIDIUM. The WEIGHT statement specifies that the variable WGT contains the sampling weights. The STRATA statement specifies that the variable PLANT identifies the strata. The CLUSTER statement specifies that the variable PLANT identifies the clusters. The VAR statement specifies the variable LOG-CLOSTRIDIUM as the variable whose mean you want to estimate. An ODS OUTPUT statement generates a data set that contains the estimation results.

```
proc surveymeans data=example;
  weight wgt;
  strata plant;
  cluster shift;
  var logClostridium;
  ods output statistics = estimates;
run;
```

![Figure 1](image)

**Figure 1** PROC SURVEYMEANS Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Error of Mean</th>
<th>95% CL for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>logClostridium</td>
<td>168</td>
<td>3.277231</td>
<td>0.176325</td>
<td>2.90942422 3.64503856</td>
</tr>
</tbody>
</table>

Table 1 shows the contents of the ODS output data set ESTIMATES.
Table 1  ODS Output Data Set ESTIMATES

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>VarName</td>
<td>Variable name</td>
</tr>
<tr>
<td>N</td>
<td>Number of observations</td>
</tr>
<tr>
<td>Mean</td>
<td>Estimated mean</td>
</tr>
<tr>
<td>StdErr</td>
<td>Standard error</td>
</tr>
<tr>
<td>LowerCLMean</td>
<td>Lower confidence limit</td>
</tr>
<tr>
<td>UpperCLMean</td>
<td>Upper confidence limit</td>
</tr>
</tbody>
</table>

Use the following DATA step to perform the required transformations on the estimates in the ESTIMATES output data set. The first assignment statement replaces the contents of the variable MEAN with its exponentiated value; this transforms the estimated arithmetic mean of the variable LOGCLOSTRIDIUM into the geometric mean of the variable CLOSTRIDIUM. The next assignment statement transforms the standard error of the arithmetic mean of LOGCLOSTRIDIUM into the standard error of the geometric mean of CLOSTRIDIUM. The next two assignment statements replace the lower and upper confidence limits of the arithmetic mean of LOGCLOSTRIDIUM with their exponentiated values; this transforms the confidence limits for the arithmetic mean of LOGCLOSTRIDIUM into confidence limits for the geometric mean of CLOSTRIDIUM. The last assignment statement changes the contents of VARNAME from LOGCLOSTRIDIUM to CLOSTRIDIUM. The LABEL statement relabels the variables Mean, StdErr, LowerCLMean, and UpperCLMean.

```sas
data estimates;
  set estimates;
  Mean = exp(Mean);
  StdErr = sqrt((Mean**2)*(StdErr**2));
  LowerCLMean = exp(LowerCLMean);
  UpperCLMean = exp(UpperCLMean);
  VarName = 'Clostridium';
  label Mean='Geometric Mean'
    StdErr='Standard Error'
    LowerCLMean='Lower 95% Confidence Limit'
    UpperCLMean='Upper 95% Confidence Limit';
run;
```

Finally, print the transformed output data set. The output from the PRINT procedure in Figure 2 displays the variable name, the estimate of the geometric mean, the standard error of the estimated geometric mean, and the lower and upper 95% confidence limits.

```sas
proc print data=estimates label noobs;
run;
```

**Figure 2** Geometric Mean Estimation

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>N</th>
<th>Geometric Mean</th>
<th>Standard Error</th>
<th>Lower 95% Confidence Limit</th>
<th>Upper 95% Confidence Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clostridium</td>
<td>168</td>
<td>26.502297</td>
<td>4.673013</td>
<td>18.3462321</td>
<td>38.2842492</td>
</tr>
</tbody>
</table>
The estimated geometric mean for *Clostridium perfringens* bacteria per gram of product is 26.50, the standard error of the estimate is 4.67, and a 95% confidence interval for the estimate is (18.34, 38.28).

If you are using SAS/STAT 12.1 or later, you estimate the geometric mean of CLOSTRIDIUM by simply specifying the ALLGEO statistic keyword in the PROC SURVEYMEANS statement; this requests all available statistics associated with geometric means.

```sas
proc surveymeans data=example allgeo;
  weight wgt;
  strata plant;
  cluster shift;
  var Clostridium;
run;
```

![Figure 3 PROC SURVEYMEANS Results](image)

The estimated geometric mean for *Clostridium perfringens* bacteria per gram of product is 26.50, the standard error of the estimate is 4.67, a two-sided 95% confidence interval for the estimate is (18.34, 38.28), and a one-sided 95% confidence interval for the estimate is (19.55, 35.92).

**References**