ABSTRACT

Air sampling and analysis methods that are used in determining the amount of airborne contaminants can be given accuracy ratings that describe their expected performance when used by others. For laboratory methods, the National Institute for Occupational Safety and Health methods manual assumes a normal distribution, but finding the point estimate and confidence interval for the accuracy of field data can be difficult because these data are often not normally distributed. Bootstrap sampling can be used to obtain the analogous accuracy measure for non-normal data. SAS® code is presented here to find the point estimate and the confidence intervals using the bootstrap method. SAS version 9.1 was used to create the code, which includes data manipulation in the data set step with arrays and do loops, and PROC MIXED.

INTRODUCTION

The ability of a method to determine the true concentration of sampled particles is called the method’s accuracy. NIOSH standards for the evaluation and analysis of air sampling for airborne contaminants require that results are within 25% of the true concentration 95% of the time [1]. In other words, we would like a 95% confidence interval for results that fall within 25% of the true value.

In laboratory settings, the NIOSH methods manual assumes that airborne particles follow a normal distribution and defines the sampling and method accuracy as a 95% coverage range in terms of two parameters consisting of any uncorrected bias and precision found in the evaluation. To account for evaluation uncertainty, the upper 95% confidence level on this range is used as a performance measure. That approach provides a reliable criterion to recommend the use of the sampling and analytical method. Finding the point estimate and confidence interval for the accuracy of air sampling and analysis methods in field studies, however, can be difficult because field data are often not normally distributed, due to a larger variance and additional noise when compared with laboratory data.

One recent solution for this situation is to take multiple bootstraps of the data. Bootstrapping is a statistical method for estimating distribution parameters by taking multiple samples, with replacement, from the original data set. Here we use this statistical method to take a number of different samples to obtain the analogous accuracy measure for non-normal data. SAS code has been developed to take a data set and run it through this bootstrap simulation, determining the point estimate and 95% confidence interval of the accuracy.

The bootstrap program was written using SAS version 9.1. It relies on data manipulation in the data set, including arrays and do-while loops, as well as using PROC MIXED.

DATA MANIPULATION AND BOOTSTRAP METHOD

This method consists of taking the original data set, typically consisting of around 100 data points, and re-sampling it with replacement. The mean and standard deviation of this new data set, from the first bootstrap, are calculated. Another new data set is then found by bootstrapping the original data set yet again, but creating the new data set with a sample size of 5000. Each one of these 5000 data points of the second bootstrapped sample are compared to the mean and standard deviation of the initial re-sampled data set, as well as a constant, k.

\[ u = \sqrt{\text{bias}^2 + \text{sd}^2} \]

if abs(value) < k*u, for 1.5 < k < 2.6, then keep (value)
The constant k takes the place of the z-value used when finding confidence intervals of normally distributed data. This is equivalent to a non-parametric version of finding confidence intervals, which allows for the broader variance and noise to be taken into account. This method is acceptable when the sample size is fairly large, at n > 30. The number of data points that are less than this comparison are stored in an array, and then a percentage is computed by dividing the number of points that are less than the comparison by 5000 (the number of initial data points), and is typically in the range of .9 to .99, depending on the range of k values used in the program.

This bootstrap is done ninety-nine times for each k-value that is used. In this case, k values ranged from 1.6 to 2.5, to give a range around the normal distribution’s z-value of 1.96, the value used for calculating 95% confidence intervals.

Arrays are set up to hold the original data set (array XRFerror), so it can be re-sampled, for the first re-sampled data set (array XRFerr_), and then to hold the ninety-nine uses of each k value (array q) and the percentage values from the comparison step (array p). The largest do loop is set up to go through all of the k values, while the next is for the ninety-nine bootstraps to find the percentages. The smaller do-loops go through the two actual re-samplings. Also used is the variable nu, the size of the data set, found using PROC MEANS, and the array ranmark, used for the random uniform number generations.

```plaintext
DATA data_1; SET original_data;
nu=length;
array XRFerror{70} err_1 - err_70;
array XRFerr_ (70*0);
array p(990);
array q(990) (99*1.6 99*1.7 99*1.8 99*1.9 99*2.0 99*2.1 99*2.2 99*2.3 99*2.4 99*2.5);
array ranmark_ (70);

a=1;
k=1.6;
do while (k<2.6);
   z=99;
do while (z>0);
      marker = nu;
m = 1;
do while (marker > 0);
      u = ranuni(0);
      u1 = nu*u + 0.5;
      u2 = round(u1,1);
drop u u1;
      ranmark_{m} = u2;
      m = m+1;
      marker=marker-1;
end;
drop m marker u2;
i=1;
n=nu;
do while (n > 0);
   t = ranmark_{i};
   t1 = XRFerror{t};
   XRFerr_{i} = t1;
   n=n-1;
i=i+1;
end;
drop i n;
```

Coders’ Corner
NESUG 2006
bias = mean(of XRFerr_1 - XRFerr_70);
trvar = var(of XRFerr_1 - XRFerr_70);
trsd = sqrt(trvar);
u=sqrt(bias**2 + trsd**2);

i=1;
ntot=5000;
n=5000;
.birth=0;
do while (ntot>0);
   s2 = ranuni(0);
   s1 = nu*s2 + 0.5;
   s = round(s1,1);
drop s2 s1;
   if (abs(XRFerror{s})) < k*u then nless=nless+1;
   ntot=ntot-1;
drop s;
end;
app = nless/n;
p{a}=app;
a=a+1;
end;
k=k+0.1;
end;
run;

USING PROC MIXED TO FIND REGRESSION PARAMETERS

The arrays containing the data with the k values and the percentages are transposed into standard data sets using PROC TRANSPOSE. This new data set is then used in PROC MIXED to find the regression equation relating the k values to the percentages. The parameters from PROC MIXED are given in an ODS output data set and arranged to give the slope and intercept of the regression equation, which will be used to find the point estimate for accuracy.

PROC MIXED data=percentage_and_k_values_from_data_1;
model percent_values=k_values / solution;
ods output solutionf=parm_est;
run;

DATA parm_est_1;SET parm_est;keep estimate;run;
PROC TRANSPOSE data= parm_est_1 out= parm_est_2;run;
DATA parm_est_3;SET parm_est_2;drop _NAME_;rename COL1=intercept COL2=slope;
run;

The same method is used to find the 95% confidence interval. The k value and percentage data sets are ordered and the actual 95th percentile of each k value is found and used in PROC MIXED.

FINDING THE FINAL ANSWERS

The regression parameters from PROC MIXED are used to find the k value that gives the point estimate of the accuracy. This is used in conjunction with the bias and standard deviation that were previously found, and a reference standard deviation to calculate the accuracy point estimate and confidence interval. The reference standard deviation is a parameter that is determined by the type of sampler being used, so no formal equation is used in the program, rather it is a constant inputted at the beginning of the SAS code. (This has been set up with
a macro variable, so it can be more easily changed at the beginning of the program.)  The equations for the accuracy point estimate and 95% confidence interval are given in [2].

```
DATA final;merge parm_3 data_1;
trsdRef = &trsd;    *macro variable for the reference error;
k=(0.95-intercept)/slope;  *parm for point estimate;
k_95=(0.95-intercept_95)/slope_95; *parm for 95% CI;
accuracy = sqrt(k**2 * (bias**2+trsd**2) + 1.960**2 * (.05**2 – trsdRef**2));
CI_95 = sqrt(k_95**2 * (bias**2+trsd**2) + 1.960**2 * (.05**2 – trsdRef**2));
run;
```

**CONCLUSION**

Not every analysis method will achieve the 25% accuracy rating required for air sampling in the workplace. However, this method of bootstrap simulation will take non-normal field data and find the point estimate and 95% confidence interval, and show how close the method comes. The data manipulation steps and procedures used are only of moderate difficulty and should be easily followed by anyone with a basic understanding of SAS.

**REFERENCES**


**ACKNOWLEDGEMENTS**

The authors would like to acknowledge Desta Fekedulegn, Chunlin Dong, and Anna Mnatsakanova for their helpful comments when reviewing this paper.

The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

**CONTACT INFORMATION**

James E. Slaven, Jr  
National Institute for Occupational Health and Safety  
Centers for Disease Control and Prevention  
1095 Willowdale Rd  
MS 4050  
Morgantown, WV, 26505  
Work Phone: 304-285-6299  
Fax: 304-285-6112  
E-mail: jslaven@cdc.gov