DOE Review Release 09/21/2017

Internal Dosimetry Coworker Data Completeness Test

ORAUT-RPRT-0086 Rev. 00
Effective Date: 09/18/2017
Supersedes: None

Subject Expert(s): Thomas R. LaBone

Document Owner Approval:	Signature on File Thomas R. LaBone, Document Owner	Approval Date:	09/12/2017
Concurrence:	Signature on File John M. Byrne, Objective 1 Manager	Concurrence Date:	09/12/2017
Concurrence:	Signature on File Scott R. Siebert, Objective 3 Manager	Concurrence Date:	09/12/2017
Concurrence:	Vickie S. Short Signature on File for Kate Kimpan, Project Director	Concurrence Date:	09/12/2017
Approval:	Signature on File James W. Neton, Associate Director for Science	Approval Date:	09/18/2017

FOR DOCUMENTS MARKED AS A TOTAL REWRITE, REVISION, OR PAGE CHANGE, REPLACE THE PRIOR REVISION AND DISCARD / DESTROY ALL COPIES OF THE PRIOR REVISION.

☒ New ☐ Total Rewrite ☐ Revision ☐ Page Change

PUBLICATION RECORD

EFFECTIVE DATE	REVISION NUMBER	DESCRIPTION
09/18/2017	00	New document initiated to evaluate the completeness of internal dosimetry data by providing a method to calculate the proportion of missing data for a given set of coworkers. Incorporates formal internal and NIOSH review comments. Training required: As determined by the Objective Manager. Initiated by Thomas R. LaBone.

TRADEMARK INFORMATION

TABLE OF CONTENTS

LIST OF FIGURES

ACRONYMS AND ABBREVIATIONS

AQL	acceptable quality level
DCAS	Division of Compensation Analysis and Support
LTPD	lot tolerance percent defective
NIOSH	National Institute for Occupational Safety and Health
NOCTS	NIOSH-DCAS Claims Tracking System
OC	operational characteristic
ORAU	Oak Ridge Associated Universities
SRDB Ref ID	Site Research Database Reference Identification (number)
SSE	site-supplied electronic
SSH	site-supplied hard-copy

1.0 INTRODUCTION

The *Interim Guidance Criteria for the Evaluation and Use of Coworker Datasets* (NIOSH 2015) established data quality criteria for bioassay[1] data from monitored coworkers that is used to estimate intake rates for unmonitored workers. These are predominately health physics criteria that relate to the adequacy of the programs that put workers on monitoring programs and performed the bioassay. In other words, the criteria address the general question of whether the right people were in the right program at the right time. That answer is ultimately a judgment call based on the maturity and rigor of the health physics program at the site.

However, the question of whether the data are complete enough to produce usable results for dose reconstruction can be answered through statistical analysis. Bioassay data from the sites (which can be referred to as "source data") are generally in one or more of the following forms:

- Site-supplied hard-copy (SSH). Dosimetry data for a population of workers in the form of electronic copies of the original hard-copy records and reports. These data are not normally in a form that is readable by a computer program and must usually be transcribed into a database before use in coworker models.

- Site-supplied electronic (SSE). Dosimetry data for a population of workers in the form of a spreadsheet, text file, or relational database. These data are in a form that can be organized for input into a database and do not require transcription.

- National Institute for Occupational Safety and Health (NIOSH) Division of Compensation Analysis and Support (DCAS) Claims Tracking System (NOCTS). Dosimetry data for individual claimants in the form of electronic copies of the original hard-copy records and reports. These data are intended for dose reconstructions and also must be transcribed into a database before use in coworker models.

Once transcribed, the best available computer-readable representation of the SSH, SSE, or NOCTS datasets is referred to as the "original dataset."

Note that no effort is made to correct transcription errors or interpret results as the original dataset is assembled. Appropriate corrections and interpretations of the original dataset are performed to produce the "coworker dataset," which can be used to develop a coworker model. Although not specifically addressed in the draft criteria, there are concerns with:

- The completeness of the source datasets (i.e., to what extent these datasets contain the bioassay results from the site), and

- The accuracy of the transcription of the data from the source datasets into the original datasets.

The accuracy question can be readily answered with a statistically based test as described in ORAUT-RPRT-0078, *Technical Basis for Sampling Plan* (ORAUT 2016a; RPRT-0078 in this document). The question of completeness is more difficult to answer because of the need to estimate the quantity of missing data. The general approach for demonstrating completeness is to examine different compilations of the data and see if they contain results that are not in the original dataset. The method in this report compares the data from NOCTS to the data in the original dataset[2]. This

[1] The interim guidance document also mentions external dosimetry data. Many of the methods discussed in this document would also be applicable to other types of data, external dosimetry data for example.

[2] The original dataset may have been derived from NOCTS files. In such a case this turns into a test of whether all of the data were transcribed from NOCTS to the electronic dataset and some steps of this procedure may have to be modified.

can shed some light on completeness because the NOCTS and site-supplied datasets are usually assembled independently and at different times and the monitored individuals in NOCTS are assumed to be a random sample of the population of monitored workers.

If all individuals with data in the NOCTS dataset have the same data in the original dataset (i.e., if NOCTS is a subset of the original dataset), the data is complete[3]. However, if this is not the case, a method must be developed to test the completeness by estimating what proportion of data is missing from the original dataset.

The test that answers this question, the "data completeness test," is the subject of this report. This test is complementary to and would usually be performed before the data accuracy test in RPRT-0078.

2.0 PURPOSE

This report gives a discussion of the statistical methods that can be used to select a sample of people from NOCTS and use their data to calculate an interval estimate of the missing data proportion for the population of all monitored workers at a site. Many details of exactly how the methods are implemented at a given site are dependent on how the datasets are structured at that site, so the discussion provided here is generic by necessity. The omitted details of the test procedure will be provided in the site-specific reports that document the application of this report.

3.0 TEST OVERVIEW

The data completeness test is performed in two parts. First, the NOCTS file for every claimant from the site is reviewed to see if it contains bioassay data, and the original dataset is searched to see if it contains bioassay data for the claimants. As discussed in ORAUT-OTIB-0075, *Use of Claimant Datasets for Coworker Modeling* (ORAUT 2016b), the individuals who make a claim (i.e., have data in NOCTS) are considered to be a random sample of the entire monitored worker population at that site for a given timeframe. There are four possible outcomes for a given individual:

1. The claimant has bioassay data in the original dataset and the NOCTS dataset.
2. The claimant has bioassay data in the original dataset but not in the NOCTS dataset.
3. The claimant has bioassay data in the NOCTS dataset but not in the original dataset.
4. The claimant has bioassay data in neither the NOCTS dataset nor the original dataset.

This comparison addresses only the presence of data, not the completeness or accuracy of transcription. Outcome #1 places the individual in the pool of claimants that are eligible for the next phase of the test; they are the "sampling frame." Outcomes #2 and #4 require no further action at this time but should be recorded and reported. Outcome #3 requires a determination of whether or not it is reasonable to expect the claimant to have bioassay data in the original dataset. If it is reasonable to expect that data should be in the original dataset, we must determine why it is not there and provide arguments to support why this does not indicate that the original dataset is incomplete.

Second, a number of individuals in the sampling frame are selected at random, and all of the records in their NOCTS datasets are checked to see if they are present in the original dataset.[4] From these data the proportion of missing data in the population and its 95% confidence interval are calculated. Therefore, the main tasks in the second phase of the completeness test are to:

- Determine the number of individuals to sample,

[3] The implicit assumption here is that the NOCTS dataset is not missing any data.
[4] Note that these data are not checked for the accuracy of transcription.

- Compare the results in the NOCTS files to the original dataset and determine (1) the total number of results and (2) the number of missing results in the sample, and

- Calculate the estimated proportion of data missing in the population and its 95% confidence interval.

The sample of individuals should be large enough to give reasonably precise answers yet small enough to minimize the amount of effort in collecting and analyzing the sample. As in RPRT-0078, a lot acceptance plan should be used to select the appropriate sample size. However, the statistical methods in RPRT-0078 to select the sample do not apply to these data because, for reasons of efficiency, they are collected in "clusters" (or "people")[5] rather than individual data fields. Arriving at an appropriate sample size can be difficult because the number of results per person is not the same for all people. The calculation of the estimated proportion of missing data in the population, given the proportion of missing data in the sample, is well defined and so is presented first before moving on to the sample size calculation.

4.0 ESTIMATE OF PROPORTION OF DATA MISSING

The sample selection methods discussed later will specify the NOCTS files to be used in the completeness test. Prior to performing the test the subject matter expert shall specify the data in NOCTS that is being considered and define exactly what it means for something to be missing. Then, all the results in the selected NOCTS files are matched up with results in the original dataset and the number of missing results are determined.

The estimate of the proportion of data missing for the population $\hat{\bar{y}}_r$ is calculated (see Section 5.2.3 of Lohr (2010)):

$$\hat{\bar{y}}_r = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i} \tag{4-1}$$

where

n = number of clusters (people) in sample
M_i = number of results (records) in ith cluster
$\bar{y}_i$ = mean result (proportion of records missing) for the ith cluster

The standard error $SE\left(\hat{\bar{y}}_r\right)$ is given by:

$$SE\left(\hat{\bar{y}}_r\right) = \sqrt{\left(1-\frac{n}{N}\right)\left(\frac{1}{n\bar{M}^2}\right)\left(\frac{1}{n-1}\right)\sum_{i=1}^{n} M_i^2\left(\bar{y}_i - \hat{\bar{y}}_r\right)^2} \tag{4-2}$$

where

N = number of clusters in population

[5] In this context, the people are the clusters.

$\bar{M}$ = the mean number of results in the *n* clusters.

The 95% confidence interval (the Wald interval) for the population proportion of data missing is calculated as:

$$\hat{\bar{y}}_r \pm t_{0.975,df} SE\left(\hat{\bar{y}}_r\right) \quad (4\text{-}3)$$

where the degrees of freedom in the Student's t quantile equals the number of clusters minus 1 (i.e., $df = n - 1$). This is a straightforward calculation and illustrative examples using three different statistical software packages are given in Attachment A. In this report the data are ungrouped, each line consisting of a 1 if a record is missing or a 0 if it not. The mean of such data is the probability that a record is missing, or equivalently, the proportion of records that are missing. The calculation of the confidence interval in Equation (4-3) assumes that the proportion has a normal distribution, which might produce a confidence interval that includes values outside of the interval of 0 to 1, which are not physically possible for a proportion. Therefore, a different method that always has upper and lower limits in the interval of 0 to 1, the "beta" method in the *svyciprop* function from the R *survey* package (Korn and Graubard 1998), is used in this report to calculate the confidence interval for the proportion of data missing. Before this calculation can be performed, the test must determine an appropriate sample size, which is discussed in the next section.

5.0 SPECIFYING A SAMPLE SIZE

RPRT-0078 defines how to select a simple random sample of individual fields in the original dataset and then go back to the source datasets to verify that these fields were accurately transcribed from the source datasets to the original dataset. The original dataset is organized in a matrix (e.g., a spreadsheet) that clearly defines the sampling frame and facilitates selection of simple random samples. In the data completeness test, the data in NOCTS are mostly from hard-copy pages organized by person. The most efficient way to sample the NOCTS data is to pull individual files at random and compare all of the data in the file to the original dataset. This is referred to as a "single-stage cluster sampling scheme" (Lohr 2010). In addition, there are a variable number of results in each NOCTS file and that number will probably be indeterminate until after the file is reviewed. The general assumption therefore is that the number of records for each person in the original dataset is a reasonable estimate[6] of the number of records per person in NOCTS.

Given the list of individuals in NOCTS and the number of records for each, a simulation can be used to estimate the sample size that has, on the average, the desired properties. The simulation reproduces the sampling process and repeats it many times. The results of the simulation provide information on how quantities like the number of records in a group of files will vary, which allows probability-based decisions on the necessary size of the sample. In many respects, the simulation is very similar to the calculations in RPRT-0078. First, the simulation must have specific sampling parameters. In the language of lot acceptance testing, these parameters (which are the same parameters specified in RPRT-0078 for all-field errors) are:

- AQL = 0.025; the acceptable error rate or acceptable quality level (AQL), which is the percentage of defects at which the consumer is willing to accept the lot as good.
- LTPD = 0.05; the unacceptable error rate or lot tolerance percent defective (LTPD), which is the upper limit of the percentage of defects in a lot that the consumer is willing to accept.

[6] Good enough to estimate the necessary size of the sample.

- α = 0.025; the producer's risk, which is the probability that a good lot containing defects equal to AQL will be rejected on the basis of sample data.

- β = 0.025; the consumer's risk, which is the probability that a bad lot containing defects equal to LTPD will be accepted on the basis of sample data.

In the context of the data completeness test these parameters imply that if the proportion of missing data in the original dataset is 2.5% (where anything less than 5% is acceptable), the conclusion is that it is not acceptable (a Type 1 error) less than 2.5% of the time. If the proportion of missing data is less than 2.5%, then the Type 1 error will be less than 2.5%. For missing proportions between 2.5% and 5%, the probability of a Type 1 error will be greater than 2.5%. On the other end of the test, if the proportion of missing data in the original dataset is 5% (where anything greater than or equal to 5% is unacceptable), the conclusion is that it is acceptable (a Type 2 error) less than 2.5% of the time. If the proportion of missing data is greater than 5%, then the Type 2 error will be less than 2.5%.

5.1 COMPUTER SIMULATION

To illustrate the simulation, assume the first part of the data completeness test identified 2,875 individuals in NOCTS who have bioassay data in the original dataset. In addition, assume that, for the purpose of arriving at the sample size, the number of records per person in the original dataset is an acceptable estimate of the number of records per person in the NOCTS dataset. This dataset composed of claim numbers and the associated number of records is the sampling frame. The density plot in Figure 5-1 shows the distribution of the number of records per individual in the sampling frame. It is important to note that there are 187 individuals who have a single record (the minimum) and 1 individual who has 1,019 records (the maximum). This affects the test as explained in Section 5.3.

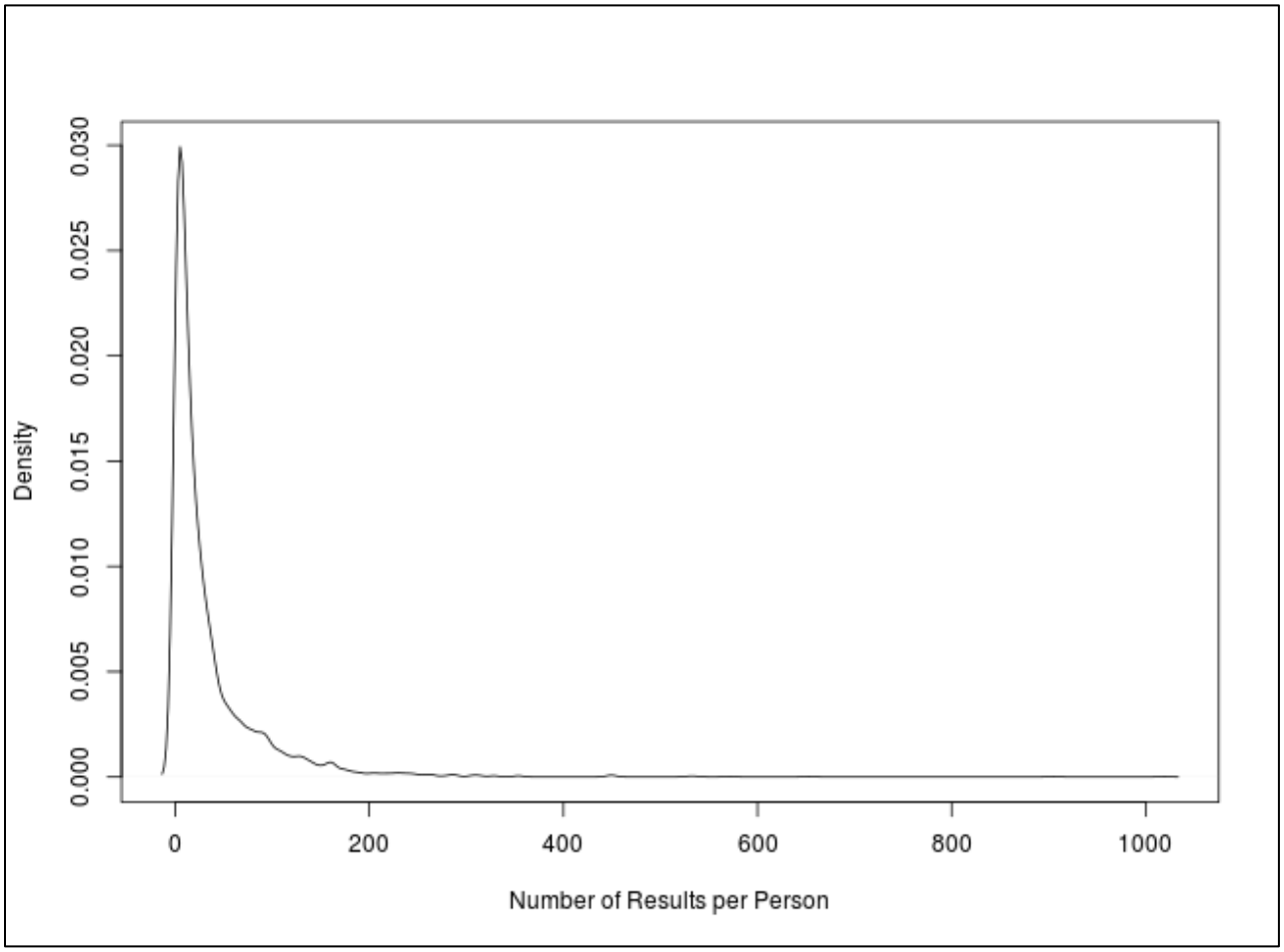

Figure 5-1. Distribution of records per person in the sampling frame.

The following is an outline of the simulation:

1. Create a long-form dataset from the sampling frame in which each row of the dataset has a claim number and a result of 0, which means that the result is not missing from the original dataset. In our example with 2,875 individuals, the long-form dataset has 101,303 lines.

2. Assign a value of 1, which indicates the result is missing, to randomly selected rows (selected without replacement) so that 2.5% of the results in the population are missing from the original dataset. This produces a proportion of missing data in the population equal to the AQL.

3. Select the data from *n* individuals at random without replacement from the 2,875 individuals in the sampling frame.

4. Apply the method discussed in Section 4.0 to the sample to estimate the proportion of data missing in the population. Save this result.

5. Go to step #1 and repeat steps 1 through 4 *m* times.

The simulation produces *m* values of the proportion of data missing it calculates for *n* individuals using the same method (which properly accounts for cluster sampling) as that for the actual sample. The simulation is then repeated with a proportion of missing data in the population of 0.05 (i.e., the LTPD). As an example, Figure 5-2 shows the results of the simulation for $n = 24$ and $m = 3 \times 10^4$. The density curves show the distribution of the missing proportions that are calculated from $m = 3 \times 10^4$

samples when the true missing data proportion is 0.025 (the lower distribution in black) and $m = 3 \times 10^4$ when the true missing data proportion is 0.05 (the upper distribution in red). The vertical lines in Figure 5-2 denote the empirical 97.5 percentile of the lower distribution (in black) and the empirical 2.5 percentile of the upper distribution (in red). These are referred to as the "critical values."

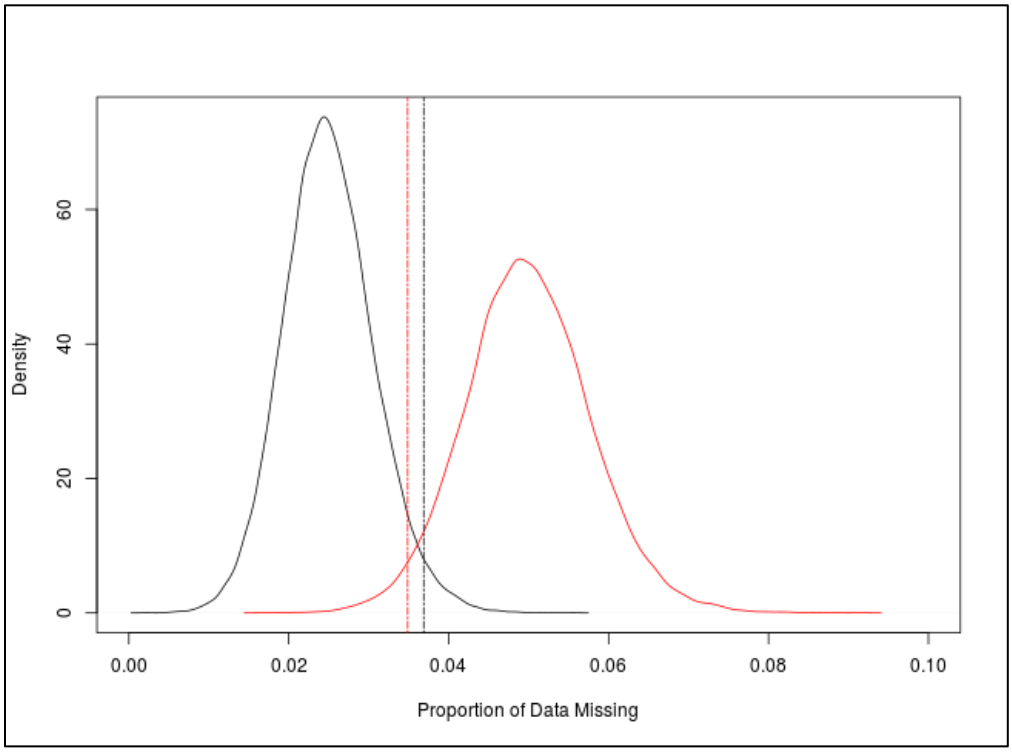

Figure 5-2. Distribution of missing data proportion observed in the sample $n = 24$ for 2.5% (left black curve) and 5% (right red curve) missing data proportions in the population.

As the sample size increases, both distributions will become narrower; the critical value of the lower distribution moves to the left, and the critical value of the upper distribution moves to the right. The desired approach is to increase the sample size to the point where the critical value of the lower (AQL) distribution is equal to the critical value of the upper (LTPD) distribution. The value of the missing proportion at which these two critical values are equal is called the "accept number." At the accept number there is an $\alpha = 0.025$ probability of rejecting a good lot and a $\beta = 0.025$ probability of accepting a bad lot.[7] In RPRT-0078 the sample size that gives the accept number is calculated by trying out all possible sample sizes until one fits. This is not a feasible approach for this test because of the time[8] needed to perform the simulations. Therefore, the approach of this method is to run the simulation for a relatively small number of sample sizes over a range that includes the sample size associated with the accept number. Two separate linear regressions of critical values on sample size are then performed, resulting in two lines. The point where the lines cross is the sample size where the 2.5% and 97.5% critical values are equal. The desired sample size is the value of the *x*-axis where the two regression lines cross, and the accept number is the value of the *y*-axis where the two lines cross. This process is illustrated in Figure 5-3, where, for example, the critical values for $n = 25$ are those shown in Figure 5-2. The calculated sample size is 27.14 (value of *x*-axis where lines

[7] This is analogous to the situation in Figure 3-3 of RPRT-0078, except the sampling distributions come from simulations rather than the exact hypergeometric distribution of RPRT-0078.

[8] The calculations in this report took about 3.5 hours to complete on a capable workstation.

cross) and the associated accept number is 0.0359 (value of *y*-axis where the lines cross). In practice, the sample size is rounded upward to the whole number 28.

5.2 OC CURVE

So far, fairly standard lot acceptance methods have provided a sample size of 28 and an accept number of 0.0359. In practice the analysis would use these values by randomly selecting 28 individuals from the sampling frame and comparing their NOCTS data to the original dataset. If the proportion of missing data in the sample was less than 0.0359 the lot is accepted. Otherwise it is rejected. The way the sampling plan was constructed ensures that whatever the decision is about the proportion of missing data, in the long run the acceptance or rejection would be wrong 2.5% of the time if the population missing proportion was equal to either AQL or the LTPD. The performance of the sampling plan for population missing proportions between 2.5% and 5% is captured in an operational characteristic (OC) curve[9] like the one shown in Figure 5-4.

The OC curve provides the necessary information about the sampling plan. For example, this OC curve indicates that:

- If the proportion of missing data in the population is 2.5%, in the long run the lot is acceptable 97.5% of the time (i.e., the missing data proportion is less than 5%);
- If the proportion of missing data in the population is 5%, in the long run the lot is acceptable 2.5% of the time (i.e., the missing data proportion is less than 5%); and
- If the proportion of missing data in the population is 3.5%, in the long run the lot is acceptable 57.3% of the time, which is represented by the vertical blue line (i.e., the missing data proportion is less than 5%).

[9] Analogous to the OC curve in Figure 3-4 of RPRT-0078.

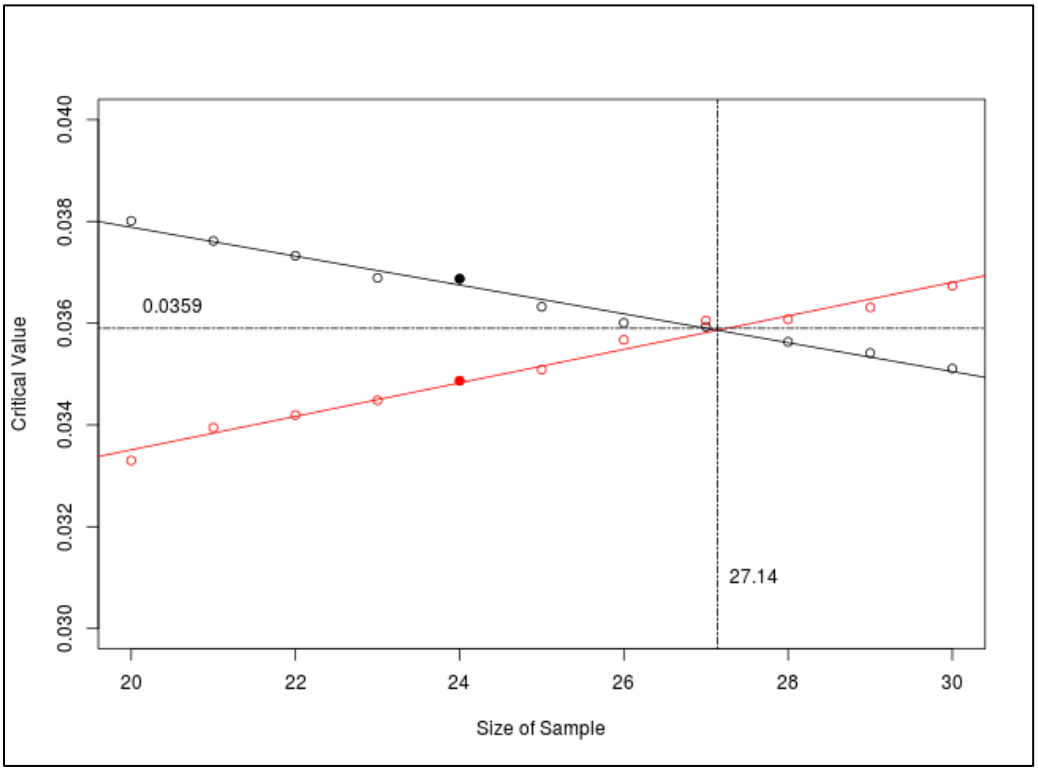

Figure 5-3. Regression of critical values on sample size when there is 2.5% of the data missing (black line) and when there is 5% of the data missing (red line). The sample size and accept number are where the two lines intersect. The black dot corresponds to the black vertical line in Figure 5-2 and the red dot to the vertical red line.

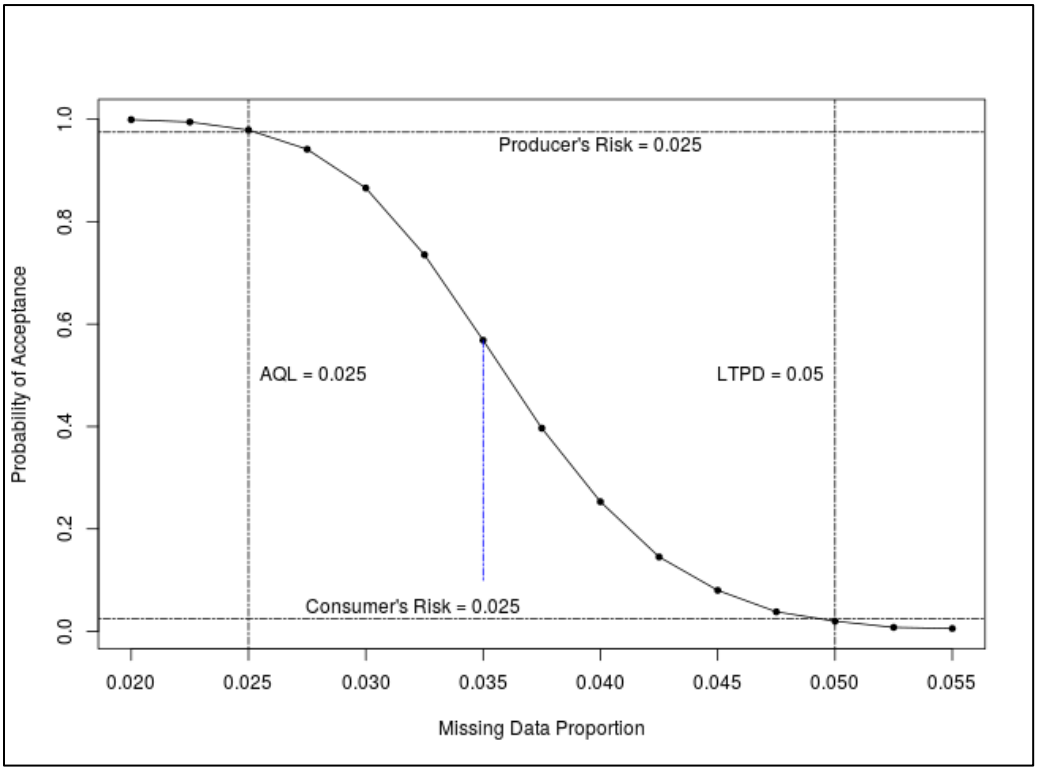

Figure 5-4. OC curve based on accept number for a sample of 28 people.

5.3 OTHER SAMPLE SIZE CONSIDERATIONS

The distribution of records per person from the simulation using 28 people is shown in Figure 5-5. As mentioned before, there are 187 individuals in NOCTS who have a single record. While very unlikely, it is possible for a sample of 28 individuals to yield exactly 28 records, which is clearly not a useful sample for estimating anything about the population. The sampling plan does not offer absolute protection against selecting such pathological samples, so some minimum number of records in the sample is necessary in addition to the number of people specified by the sampling plan.

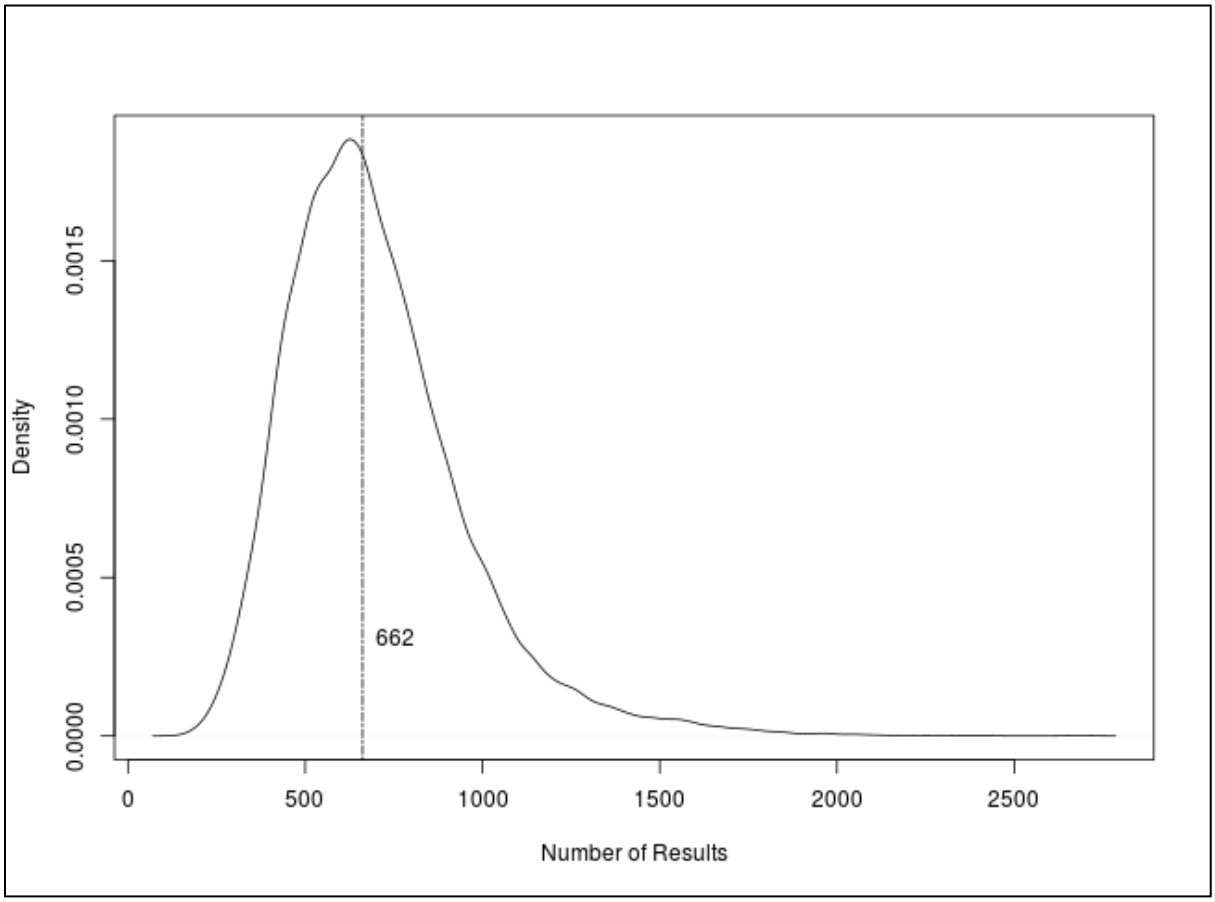

Figure 5-5. Distribution of the number of records in the files of 28 people.

One reasonable candidate for such a minimum is the median of the number of records observed from repeated sampling of 28 people from the sampling frame, which is 662 in this case (the vertical line in Figure 5-5). In practice, the analysis would examine the number of records in the actual sample of 28 people. If the number of records is less than 662, the records for more randomly selected individuals from the sampling frame would be added to the sample. This process is repeated until the desired number of records in the sample is at least 662.

The method discussed in this report is intended to apply to a variety of as-yet unseen datasets, the diversity of which has not been bounded in any way. To protect against possible samples that might be considered to be too small, a final non-statistical design criteria for the sample is that it consist of no fewer than 30 individuals, regardless of the outcome of the sampling plan calculation and the minimum acceptable number of records. In summary, the sample size is determined as follows:

1. Use the lot acceptance sampling plan to select a sample size *n* that meets the desired AQL, LTPD, consumer's risk, and producer's risk.

2. At this point the sample must contain at least 30 individuals. If $n < 30$, increase the number of individuals until $n = 30$.

3. Determine the median number of records n_r for samples of size n. If needed, increase the sample size until the sample contains at least n_r records.

These three steps produce the sample size (the number of individuals n) who should be selected at random from the sampling frame. The data from these n individuals is then compared to the data from the original dataset and the proportion of data missing in the sample is calculated.

6.0 CONFIDENCE INTERVALS

In RPRT-0078 two equivalent ways of testing for the number of typographical errors (typos) were presented:

1. A lot was accepted if the number of typos in the sample was less than or equal to the accept number and rejected if greater than the accept number.

2. A lot was accepted if the upper limit of the 95% confidence interval for the number of typos in the population was less than the LTPD and rejected if it was above the LTPD.

Both methods result in the same verdict for a given sample (accept or reject), but the confidence interval approach was used because it provides additional information beyond that simple decision. For the same reason, this method uses the confidence interval instead of the accept number to make the acceptance decision. For the completeness test there is another reason to use the confidence interval, which relates to the decision to increase the sample size beyond that dictated by the acceptance number as discussed in the previous section. For example, if the sample size is increased from 28 to 40 to achieve the minimum number of records, the OC curve in Figure 5-4 is no longer valid and, in fact, it is not possible to make an OC curve based on the accept number without changing the design specifications of the test. Thus, the Producer's Risk and Consumer's Risk in the OC curve shown in Figure 6-1 are slightly less than the desired 2.5%, i.e., we are more likely to accept a good lot and reject a bad lot with a sample of 40 individuals than we were with a sample of 28 individuals.

7.0 GRANULARITY OF MISSING DATA

The sample size calculation in Section 4 that yielded $n = 28$ individuals assumed that the missing records were randomly distributed throughout all of the 101,303 records for 2,875 individuals in the NOCTS dataset. This is the "best-case" scenario and will result in the smallest sample size for a given set of lot acceptance parameters. The "worst case" scenario occurs when the missing records are compressed into the smallest possible number of individuals so that practically all of their records are missing. In the example, the worse-case sample size calculations will give a sample size of $n = 1,259$ using the same lot acceptance parameters used in the best case. The first phase of the completeness test makes this worst-case scenario implausible, but it helps delimit the range of possibilities. In practice, the degree of granularity in the distribution of missing records will vary from dataset to dataset and therefore must be specified by the SME in consultation with the statistician. Misspecification of the degree of granularity of the missing data may result in the confidence interval on the final result being too wide (giving an answer that is not as definitive as desired) or too narrow (resulting in wasted effort).

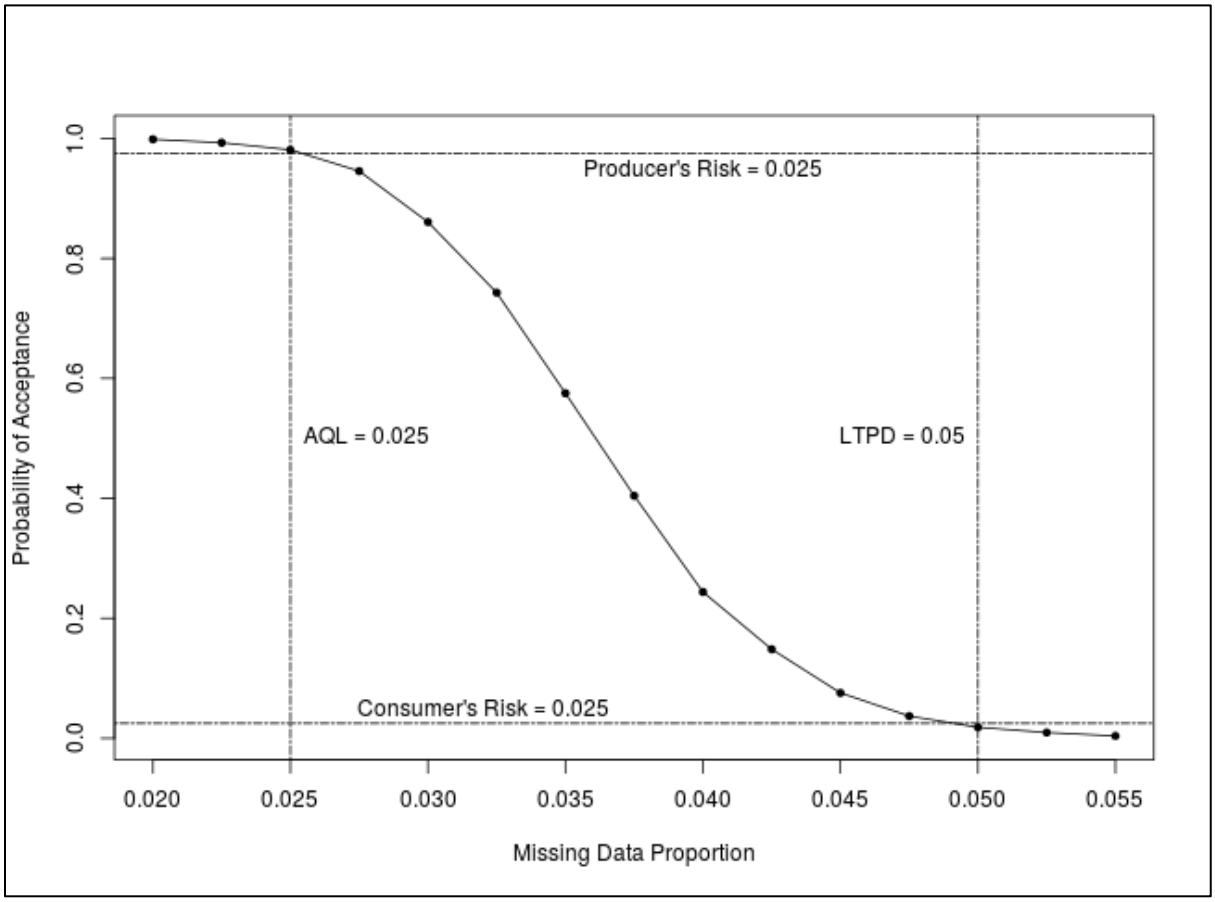

Figure 6-1. OC curve for a sample of 40 people.

The granularity of the missing data in an actual sample can be visualized using a plot of the cumulative missed data versus cumulative data, which is typically used in a Wald sequential sampling acceptable test (Montgomery 2005, p. 669). Such a plot is shown in Figure 7-1 for a sample of 40 individuals having 1,693 records, where 17 of the records are missing completely at random (i.e., low granularity). Rug plots are presented on the top and bottom of the plot, where a tic mark on the rug plot denotes the value of the x axis for the last datum in a person's results. Thus, each tic separates the data of one person from the data of the next person. Consecutive jumps in the vertical direction for one person (which end up looking like large jumps) would indicate that missing results are more granular (concentrated in individuals).

8.0 NOCTS AS A RANDOM SAMPLE

The sampling scheme in Section 4 assumes that the monitored workers in NOCTS are a random sample of all the monitored workers at the site during a specified time period. A random sample of the workers in NOCTS is taken and analyzed (a random sample of a random sample). Thus, the sampling scheme is:

all workers → workers in NOCTS → workers in random sample of NOCTS.

Ideally, the random sample would be taken directly from the worker population, bypassing NOCTS.

all workers → workers in random sample of population.

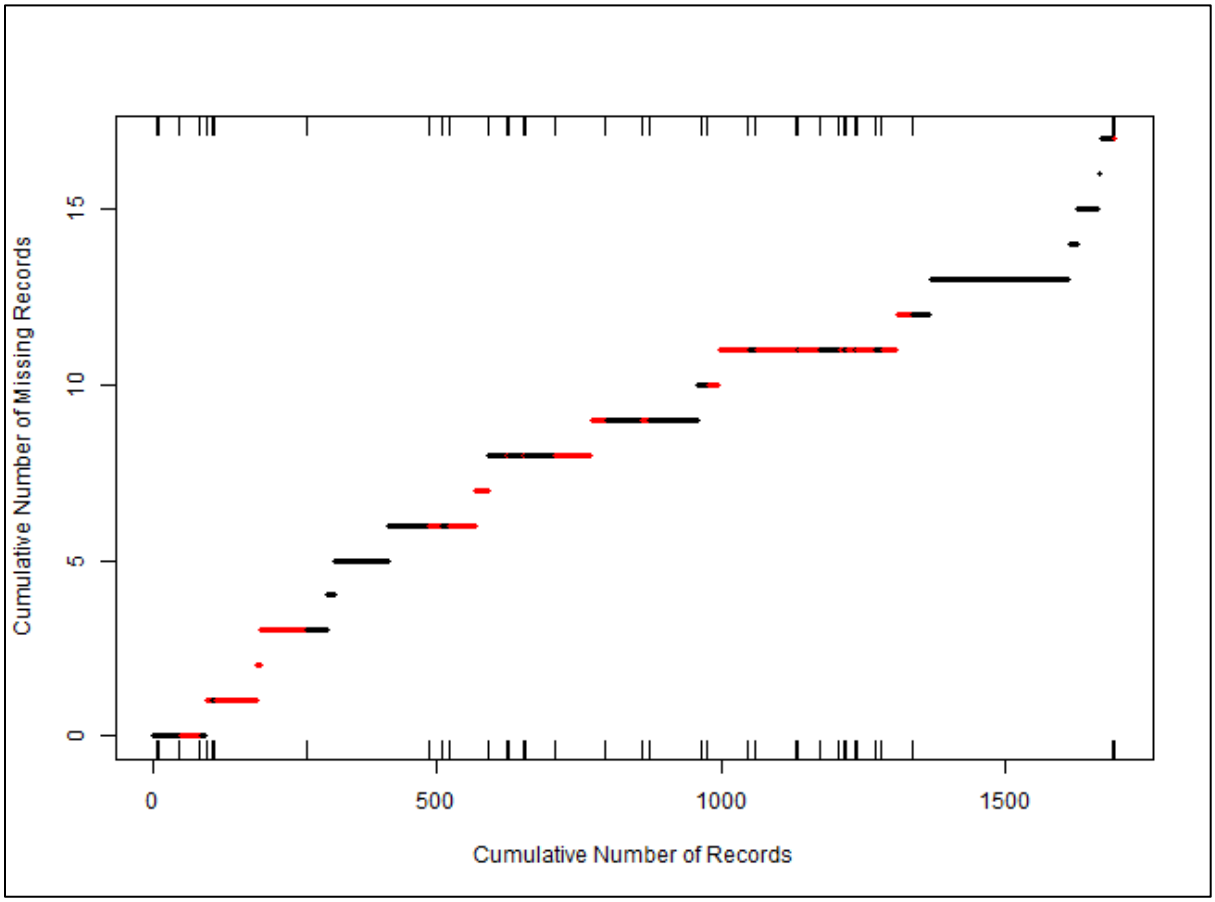

Figure 7-1. Sequential plot of missing records in a sample of 40 individuals having 1,693 records, 17 of which are missing. The color of the dots is alternated from red to black going from one person to the next.

The sampling scheme presented here includes the intermediate NOCTS step because dosimetry files are available for NOCTS but not for all workers in general. The point estimates of the proportion of data missing are considered to be equivalent for both sampling schemes, but the uncertainty in the estimate would be smaller for the second scheme. This is not considered to be an issue because:

- Of how the end result is used,
- Acceptance is based on the upper 95% confidence limit of the point estimate rather than the point estimate, and
- The alternative (i.e., getting all the dosimetry files for all workers) is not feasible.

9.0 SUMMARY

The method proposed in this report can be used to select a sample of individuals from NOCTS and use their data to estimate the proportion of data missing from the original dataset. A lot acceptance approach is used, which is similar to the one in RPRT-0078 for quantifying typos. The primary difference between the two methods is that this method takes into account the data being clustered by individual rather than being independent fields as in RPRT-0078. It is assumed that the sampling frame (i.e., the list of individuals in NOCTS who should have data in the original dataset and the number of records in each person's file), has been assembled. Once the sampling frame is available

the following procedure[10] should be used to estimate the proportion of data missing from the original dataset:

1. Get listing of claim numbers and number of records for each claim in the sampling frame.

2. Run the simulation to select the sample size *n* based on accept number. The calculations are performed here with $m = 3 \times 10^4$, but acceptable results can be achieved with fewer iterations.

3. Make a plot showing how *n* was selected (Figure 5-3).

4. Generate the OC curve for a sample of *n* people (Figure 5-4).

5. Use a simulation to determine the median number of records (n_r) for the *n* individuals (Figure 5-5).

6. Increase the sample size to 30 individuals if *n* calculated in the previous step is less than 30.

7. If the number of records is less than n_r (calculated in Step 5), then increase the sample size until the number of records exceeds the number of records n_r. In the long run, this will likely result in the need to increase the sample size about half of the time based on this requirement.

8. Generate the OC curve for the final sample size (Figure 6-1).

9. Pull the required number of NOCTS files at random without replacement and determine the number of missing records.

10. Calculate the point estimate of the proportion of data missing in the population and its 95% confidence interval using the methods discussed in Section 4.0 and generate the sequential sampling plot (Figure 7-1).

This method determines if the upper limit of the 95% confidence interval on the proportion of missing data exceeds 0.05. The lot would typically be rejected if it does, but DCAS makes the final decision to accept or reject the lot.

[10] The calculations in this report were performed with R, and can be reproduced using the script in ORAUT (2017).

REFERENCES

Korn, E. L., and B. I. Graubard, 1998, "Confidence Intervals for Proportions with Small Expected Number of Positive Counts Estimated from Survey Data, *Survey Methodology*, volume 24, pp. 193–201. [SRDB Ref ID: 166973]

Lohr, S., 2010, *Sampling Design and Analysis, Second Edition*, Brooks/Cole, New York, New York. [SRDB Ref ID: 139758]

Montgomery, D. C., 2005, *Introduction to Statistical Quality Control*, *Fifth Edition*, John Wiley & Sons, Hoboken, New Jersey. [SRDB Ref ID: 167353]

NIOSH (National Institute for Occupational Safety and Health), 2015, *Interim Guidance for the Evaluation and Use of Coworker Datasets*, Division of Compensation Analysis and Support, Cincinnati, Ohio, April 2. [SRDB Ref ID: 167199]

ORAUT (Oak Ridge Associated Universities Team), 2016a, *Technical Basis for Sampling Plan*, ORAUT-RPRT-0078, Rev. 00, Oak Ridge, Tennessee, June 17. [SRDB Ref ID: 156949]

ORAUT (Oak Ridge Associated Universities Team), 2016b, *Use of Claimant Datasets for Coworker Modeling*, ORAUT-OTIB-0075, Rev. 01, Oak Ridge, Tennessee, March 30. [SRDB Ref ID: 157060]

ORAUT (Oak Ridge Associated Universities Team), 2017, ORAUT-RPRT-0086 Rev.00_Support Files.zip, Oak Ridge, Tennessee, July 26. [SRDB Ref ID: 166972]

ATTACHMENT A
BENCHMARK PROBLEM FOR MEAN AND CONFIDENCE INTERVAL CALCULATION

SECTION **TITLE** **PAGE**

LIST OF TABLES

TABLE **TITLE** **PAGE**

ATTACHMENT A
BENCHMARK PROBLEM FOR MEAN AND CONFIDENCE INTERVAL CALCULATION
(continued)

Example 5.6 from (Lohr 2010) gives data for algebra test scores from 12 schools randomly selected from a population of 187 schools.[11] In this dataset N = 187 is the number of clusters in the population and wt = 187/12 = 15.58333333 is the weight applied to each cluster. Estimates of the population mean test score and its 95% confidence interval were calculated from the "algebra" dataset by hand and with the statistical software packages R, SAS, and Stata. All four methods generated identical results for the mean and 95% confidence interval.

A.1 HAND CALCULATION

The estimate of the mean test score and its standard error are calculated below by hand using these equations and the summary statistics in Table A-1:

$$\hat{\bar{y}}_r = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i} \tag{A-1}$$

$$SE\left(\hat{\bar{y}}_r\right) = \sqrt{\left(1-\frac{n}{N}\right)\left(\frac{1}{n\left(\bar{M}\right)^2}\right)\left(\frac{1}{n-1}\right)\left(\sum_{i=1}^{n} M_i^2\left(\bar{y}_i - \hat{\bar{y}}_r\right)^2\right)} \tag{A-2}$$

Table A-1. Summary statistics used in the calculation of the mean and standard error of the mean.

Class Number	Number of results in *i*th cluster	Mean result for the *i*th cluster	Number of results times mean for the *i*th cluster	Product of the square of the number of results multiplied by the squared difference of the mean for the *i*th cluster minus the mean of the population.
23	20	61.5	1,230	456.7298
37	26	64.2	1,670	1,867.74
38	24	58.4	1,402	9,929.22
39	34	58	1,972	24,127.75
41	26	58	1,508	14,109.31
44	28	64.9	1,816	4,106.28
46	19	55.2	1,048	19,825.39
51	32	72.1	2,308	93,517.32
58	17	58.2	989	5,574.94
62	21	66.6	1,398	7,066.12
106	26	62.3	1,621	33.4386
108	26	67.2	1,746	14,212.7867
Totals	**299**	N/A	**18,708**	**194,827.04**

[11] The dataset is available from SRDB 166972.

ATTACHMENT A
BENCHMARK PROBLEM FOR MEAN AND CONFIDENCE INTERVAL CALCULATION
(continued)

$$\hat{\bar{y}}_r = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i} = \frac{1.8708 \times 10^4}{299} = 62.569 \qquad \text{(A-3)}$$

$$SE\left(\hat{\bar{y}}_r\right) = \sqrt{\left(1 - \frac{12}{187}\right)\left(\frac{1}{12\left(24.917\right)^2}\right)\left(\frac{1}{12-1}\right)\left(1.94827 \times 10^5\right)} = 1.492 \qquad \text{(A-4)}$$

The lower and upper limits on the 95% confidence interval are:

L_{low} = 62.569 – t(0.975, df = 11) (1.492) = 59.286
L_{up} = 62.569 + t(0.975, df = 11) (1.492) = 65.852

A.2 R

The estimated population mean score and its 95% confidence interval as given by R Version 3.3.3 are shown below along with the data for the first two classes:

```
> data.in <- read.csv("algebra.csv",as.is=TRUE)
> data.in$N <- 187
> head(data.in,n=46)
   class Mi score       wt   N
1     23 20    57 15.58333 187
2     23 20    90 15.58333 187
3     23 20    56 15.58333 187
4     23 20    57 15.58333 187
5     23 20    46 15.58333 187
6     23 20    55 15.58333 187
7     23 20    62 15.58333 187
8     23 20    66 15.58333 187
9     23 20    78 15.58333 187
10    23 20    76 15.58333 187
11    23 20    57 15.58333 187
12    23 20    84 15.58333 187
13    23 20    27 15.58333 187
14    23 20    70 15.58333 187
15    23 20    49 15.58333 187
16    23 20    82 15.58333 187
17    23 20    52 15.58333 187
18    23 20    82 15.58333 187
19    23 20    59 15.58333 187
20    23 20    25 15.58333 187
21    37 26    57 15.58333 187
22    37 26    34 15.58333 187
23    37 26    68 15.58333 187
24    37 26    99 15.58333 187
25    37 26    24 15.58333 187
26    37 26    83 15.58333 187
27    37 26    40 15.58333 187
28    37 26    98 15.58333 187
29    37 26    90 15.58333 187
30    37 26    50 15.58333 187
31    37 26    60 15.58333 187
```

ATTACHMENT A
BENCHMARK PROBLEM FOR MEAN AND CONFIDENCE INTERVAL CALCULATION
(continued)

```
32    37 26    64 15.58333 187
33    37 26    58 15.58333 187
34    37 26   100 15.58333 187
35    37 26    68 15.58333 187
36    37 26    52 15.58333 187
37    37 26    77 15.58333 187
38    37 26    69 15.58333 187
39    37 26    64 15.58333 187
40    37 26    64 15.58333 187
41    37 26    73 15.58333 187
42    37 26    46 15.58333 187
43    37 26    44 15.58333 187
44    37 26    95 15.58333 187
45    37 26    28 15.58333 187
46    37 26    65 15.58333 187
>
> dclus1 <- svydesign(id=~class, weights=~wt, fpc=~N, data=data.in)
> dclus1
1 - level Cluster Sampling design
With (12) clusters.
svydesign(id = ~class, weights = ~wt, fpc = ~N, data = data.in)
>
> svymean(~score, dclus1)
        mean     SE
score 62.569 1.4916
>
> confint(svymean(~score, dclus1),df=degf(dclus1))
         2.5 %  97.5 %
score 59.28562 65.8515
```

A.3 SAS

The text provides SAS output for the calculations. An abbreviated version of the SAS 9.4 code from the program website, which also calculates the 95% confidence interval, is shown below:

```
proc import datafile =" algebra . csv " out = algebra dbms = csv replace
; getnames = yes ;
run ;
proc  surveymeans  data = algebra  total  =  187  nobs  mean  sum  clm
clsum  df; cluster    class ;
    var  score ;
    weight wt;
    ods output Statistics = myout ;
run ;
```

ATTACHMENT A
BENCHMARK PROBLEM FOR MEAN AND CONFIDENCE INTERVAL CALCULATION
(continued)

The SAS output was:

The SURVEYMEANS Procedure

Data Summary	
Number of Clusters	12
Number of Observations	299
Sum of Weights	4659.41667

			Statistics			
Variable	N	DF	Mean	Std Error of Mean	95% CL for Mean	
score	299	11	62.568562	1.491578	59.2856211	65.8515026

A.4 STATA

Stata 14.2 was used to calculate the mean and its 95% confidence interval. The code is shown below:

```
import  delimited  " algebra . csv ", clear
svyset  [pweight = wt],  psu(class) fpc(n)
svy : mean score
```

The output of Stata was:

```
Survey: Mean estimation

Number of strata =       1        Number of obs    =         299
Number of PSUs   =      12        Population size = 4,659.4166
                                  Design df        =          11
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
score	62.56856	1.491578	59.28562	65.8515