Technical Basis for Sampling Plan

ORAUT-RPRT-0078 Rev. 00
Effective Date: 06/17/2016
Supersedes: None

Subject Expert(s): Thomas R. LaBone, Nancy M. Chalmers

Document Owner Approval:	Signature on File Thomas R. LaBone, Document Owner	Approval Date:	06/06/2016
Concurrence:	Signature on File John M. Byrne, Objective 1 Manager	Concurrence Date:	06/03/2016
Concurrence:	Signature on File Edward F. Maher, Objective 3 Manager	Concurrence Date:	06/06/2016
Concurrence:	Dianne Poncio Signature on File for Kate Kimpan, Project Director	Concurrence Date:	06/06/2016
Approval:	Signature on File James W. Neton, Associate Director for Science	Approval Date:	06/17/2016

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
06/17/2016	00	New document initiated to describe the technical basis for sampling of coworker datasets to determine transcription error (typo) rates. Incorporates formal NIOSH review comments. Training required: As determined by the Objective Manager. Initiated by Thomas R. LaBone.

TABLE OF CONTENTS

LIST OF FIGURES

ACRONYMS AND ABBREVIATIONS

AQL	acceptable quality level
LTPD	lot tolerance percent defective
NIOSH	National Institute for Occupational Safety and Health
OC	operational characteristic
ORAU	Oak Ridge Associated Universities
PMF	probability mass function
SRDB Ref ID	Site Research Database Reference Identification (number)

1.0 PURPOSE

Datasets used for coworker modeling are often created by manually transcribing data from scans of the original records into an electronic database. For the Oak Ridge Associated Universities (ORAU) Team Dose Reconstruction Project, the National Institute for Occupational Safety and Health (NIOSH) has specified that (Calhoun 2015):

> *The data acceptance criteria for the coded datasets should be such that the error rate in the analytical results should be less than 1% with the overall error rate (all data fields combined) should be less than 5%.*

For the purpose of this report, the error rates specified above are taken to be "transcription" error rates that quantify the degree to which the electronic dataset agrees with the original hardcopy records. This means that (1) the accuracy and completeness of the original hardcopy records is not a factor in this analysis, and (2) the completeness of transcription from the original hardcopy to the electronic dataset is not addressed. The second statement means that the issue of data in the hardcopies that were inadvertently not transcribed to the electronic dataset (i.e., data that are missing from the electronic dataset) is also not addressed (Section 6.3 explains why). To avoid confusion, because the term "error" encompasses several statistical meanings, this report uses the term "typo rate" to mean "transcription error rate."

This report describes a statistical sampling technique in which a comparison of the data[1] in the electronic dataset to the original data is performed after the transcription is complete to confirm that the specified typo rates have not been exceeded and to generate final typo rates that will be reported to all stakeholders. The sampling plan is used to select a representative sample of the data and to estimate the typo rates. This report gives the technical basis for that sampling plan. Many of the details on how samples will be selected depend on how the original records are organized and will therefore change from dataset to dataset. These details must be filled in when the subject matter experts and statisticians establish the sampling program for a particular dataset; they are therefore not provided here. Finally, it is important to stress that the desired levels of quality in the data are most efficiently achieved through the proper design and application of appropriate data entry procedures, not through repeated application of quality control testing procedures like those discussed in this report.

2.0 INTRODUCTION

The ORAU Team has bioassay records for worker populations that have been transcribed from hard copy into an electronic database. Each record consists of one or more critical fields[2] and one or more noncritical fields. To estimate the typo rates in the critical fields and in all fields (union of critical and noncritical fields), a comparison is made between a sample of records from the electronic dataset and the corresponding entries in the hard copies. This sampling plan addresses:

- How many fields need to be verified, and
- At what point the typo rate is excessive.

The analysis approached these questions in the context of a "lot acceptance sampling program" (Montgomery 2005, Section 14-2). For example, assume there are 5,000 records in the database.

---

[1] Coworker data are the primary focus of this report, but the general approach is applicable to any type of data once the acceptable error rates, unacceptable error rates, and risks (consumer and producer) are specified.

[2] NIOSH has defined a field containing an analytical result to be a "critical field," and a field containing any other information to be a "noncritical field." This does not imply that noncritical fields are unimportant, it is simply a classification for applying different acceptable typo rates.

Each record has one critical field and four noncritical fields, for a total of five fields per record. Therefore, there are N = 5,000 critical fields and an all-field total of N_a = 25,000 fields. This implies that fields in the dataset must be classified as critical, noncritical, or irrelevant. Irrelevant fields are excluded from the dataset for testing purposes before performing the acceptance test.

In this discussion, a typo is defined to be any typo. No allowance is made for potential impact of an error beyond its classification as being in either a critical or noncritical field. For example, if the date of a bioassay is 3/14/1968 and it is entered into the database as 4/14/1968, it is counted as a typo regardless of whether or not it would change any calculations or decisions based on it. Further, a given field either has typos or does not.[3] The number of errors in a field beyond one is irrelevant.

NIOSH has specified that the minimum unacceptable typo rate in the critical field is θ = 0.01 and the minimum unacceptable typo rate in all fields is θ_a = 0.05. These rates refer to the true typo rates in the entire dataset. Thus, if a census (that is, a comparison of every one of the 5,000 records against the hardcopy outputs), resulted in 50 or more typos in critical fields, or 250 or more typos in all fields, the typo rate would be unacceptable and the lot would be rejected.

Rather than a census, the approach given here uses a random sample of the fields, compares them to the hard copies, then estimates the typo rate in the population from the typo rate in the sample. This approach does not give a deterministic answer as a census would, and there is a risk of making two misjudgments. The first is the conclusion, from the sample, that the typo rate is less than the minimum unacceptable typo rate when in reality the rate in the population is greater than the minimum unacceptable rate. This represents the risk of accepting a dataset that has an excessively high typo rate, which is referred to as "consumer's risk" in acceptance sampling. The second type of misjudgment is the conclusion that the typo rate is greater than the minimum unacceptable typo rate when in reality the rate in the population is less than the minimum unacceptable rate. This error represents the risk of rejecting a dataset that has an acceptable typo rate, which is referred to as the "producer's risk." Part of designing the sampling plan is selecting the sample size to balance these two opposing risks. This is the topic of the next section.

3.0 SAMPLING PLAN

Assume that a sample of n = 3,000 critical fields is taken from the population of N = 5,000 records where there is one critical field per record. The sample is selected at random without replacement and checked against the corresponding hard copies. The probability mass function (PMF) for the hypergeometric distribution gives the probability $f(m)$ of observing m typos in this sample:

$$f(m) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} \tag{3-1}$$

where

N = number of critical fields in the population = 5,000
n = number of critical fields in the sample = 3,000
M = the true number of typos in critical fields in the population = θN = 50
m = the observed number of typos in critical fields in the sample

[3] In the language of acceptance sampling, typos are termed "defectives" rather than "defects."

Therefore, according to the PMF in Equation 3-1, in the sample of 3,000 fields there is a probability of

$$f(30) = \frac{\binom{50}{30}\binom{5{,}000-50}{3{,}000-30}}{\binom{5{,}000}{3{,}000}} = 0.1151356 \qquad (3\text{-}2)$$

of observing exactly 30 typos in the critical field. A bar plot of $f(m)$ for m = 0,1,2, . . . ,50 typos is given in Figure 3-1. This plot shows that among repeated random samples of size 3,000 from a population of 5,000 that has 50 typos, there will be between 0 and 50 observed typos in each sample, with each having a different probability of occurring.

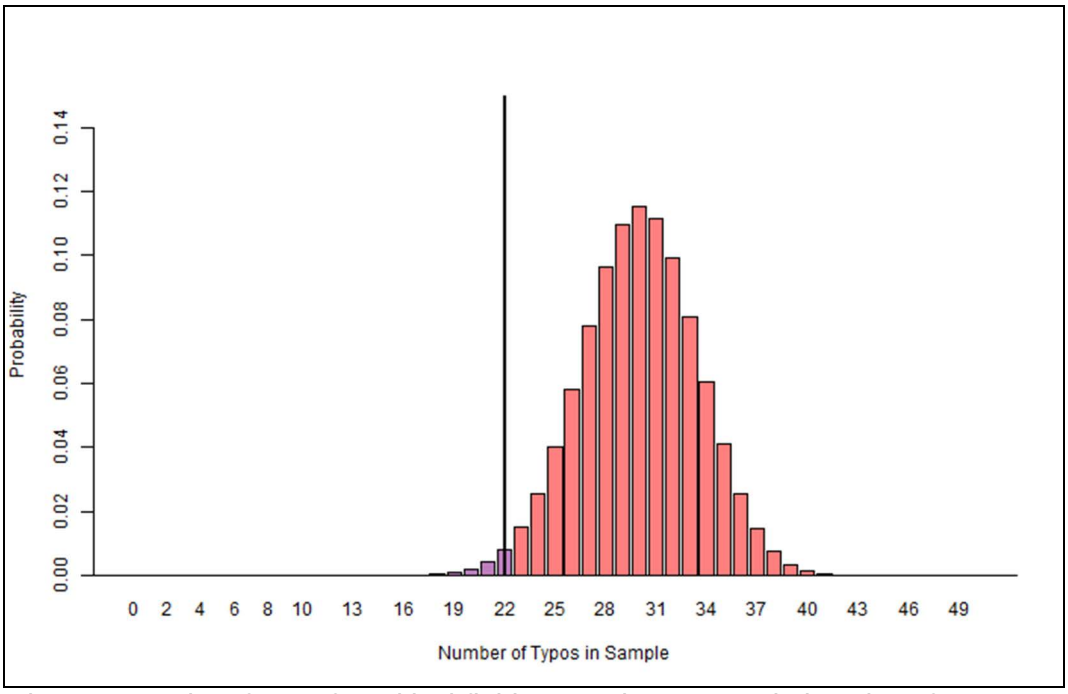

Figure 3-1. Plot of PMF for critical-field typos given a population size of N = 5,000, a sample size of n = 3,000, and an unacceptable typo rate of θ = 0.01. The accept number is 22, which means that 23 or more typos in the sample (the bars in orange) will cause the lot to be correctly rejected.

The standard strategy to follow in this situation is to pick a value for the number of typos in the sample below which that number is unlikely to be observed if the sample was indeed drawn from a population that had 50 typos. In other words, the aim is to define the lowest number of critical-field typos in the sample that is consistent with a critical-field typo rate of θ = 0.01 in the population. To do this, assume the lot is acceptable no more than 2.5% of the time when θ = 0.01.[4] The consumer's risk β is the probability that a bad lot will be accepted, and in this case we chose β = 0.025. There is a probability[5] of 0.00156 of observing 22 or fewer typos in the sample and 0.0307 of observing 23 or fewer typos in the sample if there are 50 typos in the population. Therefore, the maximum number of observed typos that is acceptable for the lot is 22. This is referred to as the "accept number." The

[4] A value of exactly 2.5% would be preferable, but because this is a discrete distribution that is not always achievable, so the criteria must be less than or equal to 2.5%.

[5] Which is the sum of $f(m)$ for m = 0,1,2, . . . ,22.

typo rates in purple in Figure 3-1 are consistent with a typo rate that is less than 0.01 and will result in acceptance of the lot.

In summary, in a random sample of 3,000 from a population of 5,000, there is less than a 2.5% chance of a bad lot – a lot with a typo rate equal to or greater than 1% – being accepted if only those lots with fewer than 23 typos in the sample are accepted. However, this procedure does not provide a unique answer because the same process with a sample size of 2,000 could result in an equally valid (yet different) result. To uniquely specify the sample size, it is necessary to address the other type of error that results from sampling: the error of rejecting a good lot.

3.1 PRODUCER'S RISK

There are typo rates in a population that are considered acceptable and yet can produce a sample that leads to lot rejection. As mentioned previously, this is referred to as the producer's risk and the goal is to limit both it and the competing consumer's risk at the same time. This is accomplished by first defining a typo rate γ that is acceptable.[6] Assume $\gamma = 0.005$ and the probability $\alpha = 0.025$ of rejecting an acceptable lot with that typo rate. Figure 3-2 shows the distribution of typos for this situation.

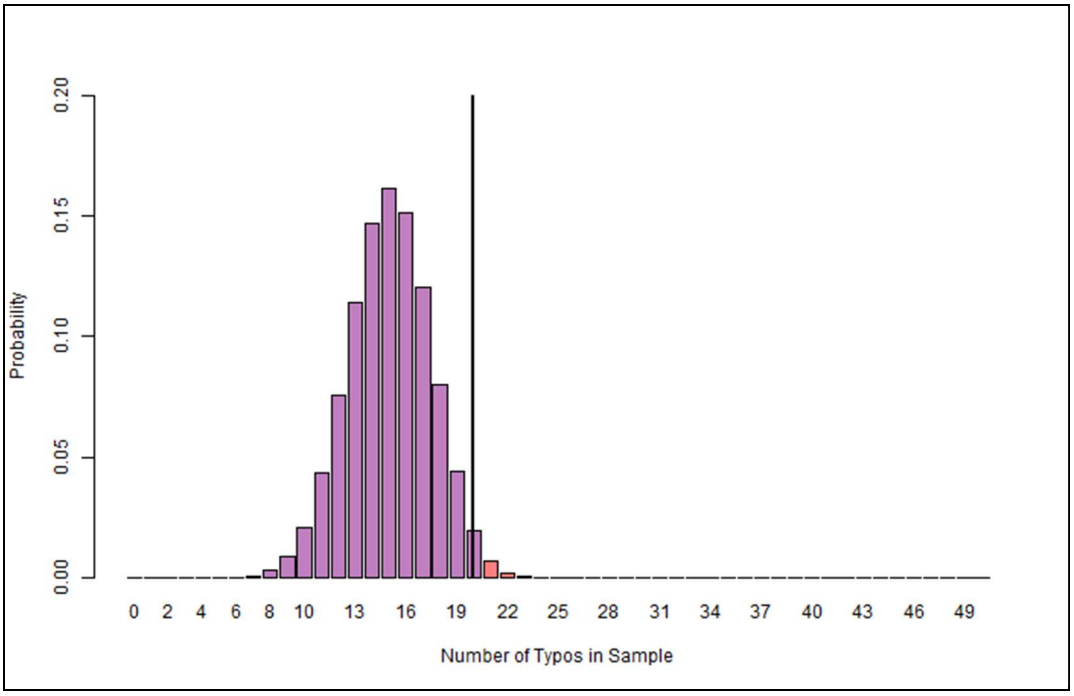

Figure 3-2. Plot of PMF for critical-field typos given a population size of $N = 5{,}000$, a sample size of $n = 3{,}000$, and a typo rate of $\gamma = 0.005$. The accept number is 20, which means that 21 or more typos in the sample (the bars in orange) will cause the lot to be incorrectly rejected.

The goal is to adjust the sample size so that the accept number for the distribution in Figure 3-1 is the same as the accept number for the distribution in Figure 3-2. In other words, we want to adjust the sample size until we arrive at an accept number that leads us to accept a good lot (defined to have a population typo rate of $\gamma = 0.005$) at least 97.5% of the time while at the same time accepting a bad lot

[6] This is referred to as the acceptable quality level (AQL) and its value was defined in conversations with NIOSH.

(defined to have a population typo rate of $\theta = 0.01$) at most 2.5% of the time. As shown in Figure 3-3, this occurs at $n = 2{,}435$, where the acceptance number is 17.

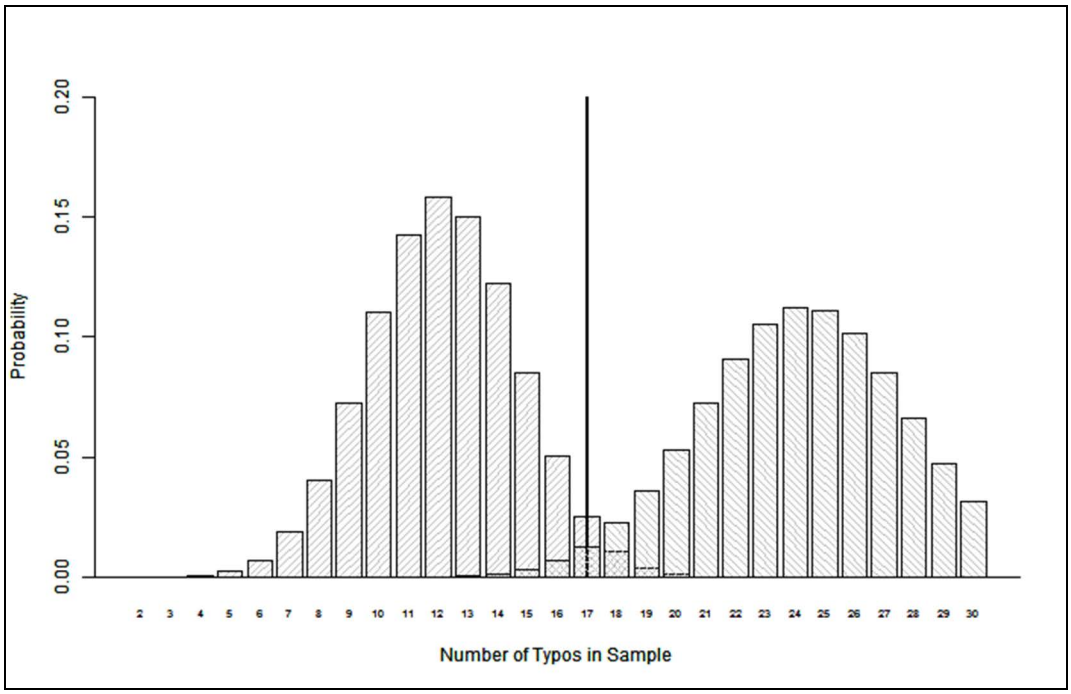

Figure 3-3. Plot of the hypergeometric PMFs generated with sample of size n = 2,435 taken from a population size of N = 5,000. The PMF on the left is for M = 25 typos in the population ($\gamma = 0.005$) and the PMF on the right is for M = 50 typos in the population ($\theta = 0.01$). The accept number of 17 is at approximately the 97.5th percentile of the left PMF and at approximately the 2.5th percentile of the right PMF.

In summary, assume:

- A population size of N = 5,000 critical fields,
- An acceptable error rate (AQL) of $\gamma = 0.005$,
- An unacceptable error rate (LTPD) of $\theta = 0.01$,
- A producer's risk of $\alpha = 0.025$, and
- A consumer's risk of $\beta = 0.025$.

This means a random sample of n = 2,435 critical fields containing 17 or fewer typos is consistent with a population typo rate of 1% or less. Observing 18 or more typos means the population typo rate is greater than 1%. The next section provides details of determining the sample size and acceptance number.

3.2 OPERATIONAL CHARACTERISTIC CURVE

In practice, acceptance sampling plans are often designed with the aid of an operational characteristic (OC) curve like the one shown in Figure 3-4.[7] The red curve is the probability of accepting the lot

[7] Generated with the R package *AcceptanceSampling*.

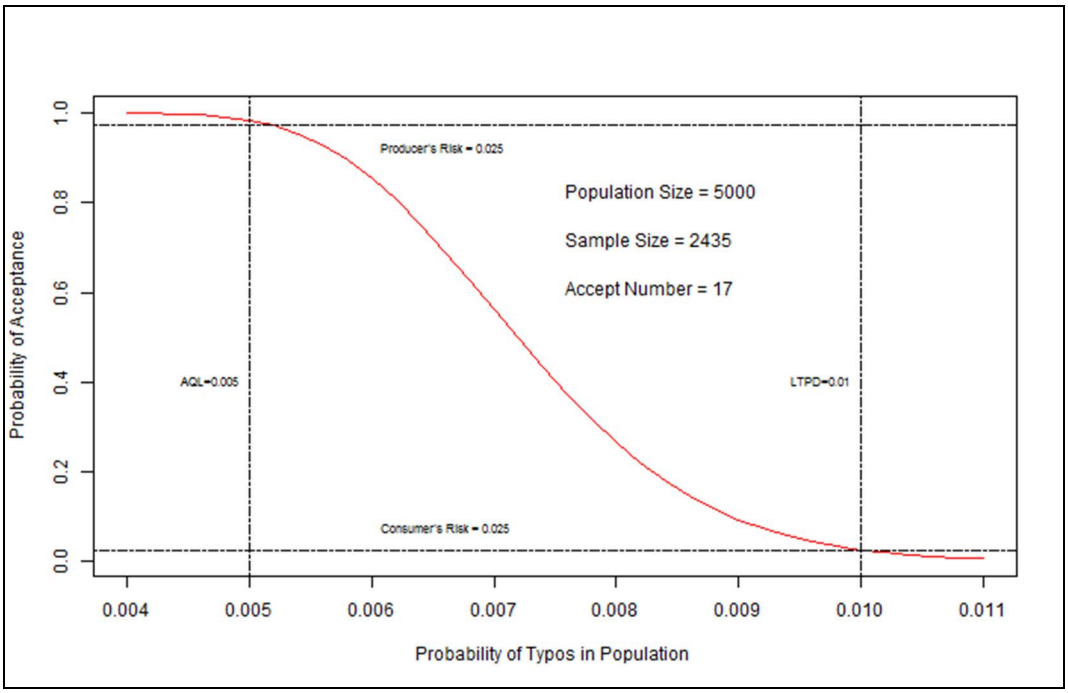

Figure 3-4. OC curve for critical typos based on the hypergeometric distribution (a Type A sampling plan).

(i.e., deciding that the typo rate is acceptable) as a function of the true typo rate in the population. This OC curve summarizes the information in Figure 3-3, and shows that:

- The probability of accepting a bad lot (one in which the true typo rate is ≥0.01,) is ≤0.025.
- There is a gray area between a true typo rate of $\gamma = 0.005$ and a true typo rate of $\theta = 0.01$ where the probability of accepting a lot ranges from 0.975 to 0.025.
- The probability of accepting a good lot (one in which the true typo rate is ≤0.005) is ≥0.975.

In summary, a lot acceptance method is used to specify the size of a simple random sample that should be taken from the population. According to the plan, the lot should be rejected if the number of typos in the sample exceeds the accept number, and accepted otherwise. If this procedure is followed, the risks of accepting a bad lot and rejecting a good lot are held to within prespecified levels.

3.3 CONFIDENCE INTERVALS

After the verification is completed, there is a known number of critical typos *m* actually observed in the sample *n* from the population *N*. Comparing *m* to the accept number to complete the lot acceptance process is basically performing a null hypothesis test in which:

- The null hypothesis H_0 is that the typo rate in the population is $\leq\theta$.
- The alternate hypothesis H_a is that the null hypothesis is false (i.e., the typo rate is $>\theta$).

If *m* is greater than the accept number, the null hypothesis is rejected, the alternate hypothesis is accepted, and as a consequence the lot is rejected. If *m* is less than or equal to the accept number,

the lot is simply accepted.[8] A more useful approach is to first generate a point estimate $\hat{M}$ of the actual number of typos *M* in the population, then generate a confidence interval for *M*. An unbiased point estimate of *M* is given by:

$$\hat{M} = N\left(\frac{m}{n}\right) \tag{3-3}$$

For example, if *m* = 10 observed typos in a sample of *n* = 2,435 critical fields, the point estimate of *M*, the number of critical field typos in the population, would be:

$$\hat{M} = 5{,}000\left(\frac{10}{2{,}435}\right) = 20.53 \tag{3-4}$$

Rounding this value to 21, the estimated critical-field typo rate in the population is:

$$\frac{\hat{M}}{N} = \frac{21}{5{,}000} = 0.0042 \tag{3-5}$$

An exact 95% confidence interval on *M* can be constructed using the *Test Method* (Wright 1991, p. 45; Buonaccorsi 1987).[9] Given *m* observed typos, the lower bound on the confidence interval is the smallest value of *M* that makes this inequality true:

$$\sum_{k=m+1}^{M} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} > \frac{0.05}{2} \tag{3-6}$$

The upper confidence bound is the largest value of *M* that makes this inequality true:

$$\sum_{k=0}^{m} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} > \frac{0.05}{2} \tag{3-7}$$

These inequalities are solved by testing every possible value of *M*, giving a lower bound of the confidence interval on *M* of 13 and an upper bound of 32. The actual coverage in this case is 97.29% instead of 95% because it is not possible to always get the desired coverage with discrete random variables. However, the actual coverage of confidence intervals constructed with the test method will always be conservatively greater than or equal to 95%.[10] Thus, there is 97.29% confidence that the true value of the number of typos in the population is between 13 and 32. This can be restated in terms of the population typo rate: There is 97.29% confidence that the true value of the typo rate in the population is between 0.0026 and 0.0064. Note that a typo rate of $\theta = 0.01$ is not within the confidence interval. This means that a critical typo rate of $\theta = 0.01$ is not a plausible value and since

---

[8] The philosophical issue of whether or not to accept the null hypothesis is disregarded.

[9] "Exact" means using the hypergeometric distribution to generate the confidence interval rather than an approximate distribution such as normal.

[10] This is one reason the exact confidence interval is used rather than a method that uses the normal approximation, which is not guaranteed coverage of at least 95%.

the plausible interval is below $\theta = 0.01$ the lot is accepted. A critical field typo rate of $\theta = 0.01$ in the interval would mean the lot would be rejected as would the plausible interval being above $\theta = 0.01$. In some cases,[11] using the confidence interval can lead to a different conclusion than that reached using the accept number in the null hypothesis test. This is a consequence of working with discrete random variables such as the number of typos in a sample. This is of no practical consequence here because the OC curve is used only to select the sample size, whereas a final decision to accept or reject the lot is made using a point estimate of the typos rate and the associated confidence interval.

4.0 TYPOS IN ALL FIELDS

In the previous sections we calculated the sample size for critical-field typos, which are a subset of typos in all fields. In the example there are five fields for every one critical field. To calculate the sample size for all typos, the OC curve is generated with the design parameters for all typos:

- Population size of $N_a = 5{,}000 \times 5 = 25{,}000$,
- AQL of $\gamma_a = 0.025$,
- LTPD of $\theta_a = 0.05$,
- Producer's risk of $\alpha = 0.025$, and
- Consumer's risk of $\beta = 0.025$.

The resulting OC curve indicates that we need a sample of 846 fields, which are selected at random and may include critical fields, even previously selected critical fields. Note that the sampling plan for all fields is separate and independent of the sampling plan for critical fields.

A sample size of fields is indicated by the OC curve in Figure 4-1. As with the critical-field typo rate, it is useful to express the estimated population all-field typo rate as a confidence interval. For example, if we observe $m = 19$ typos in all fields, there is 96.23% confidence that the true value of the typo rate in the population is between 0.0137 and 0.0346. Because the all-field typo rate $\theta_a = 0.05$ is not within this confidence interval, the lot would be accepted. In summary, the lot is rejected (i.e., the typo rate declared to be too high) if the confidence interval for the typos in critical fields contains 0.01 or if the confidence interval for the typo rate in all fields contains 0.05.

5.0 SAMPLING PLAN FOR LARGE POPULATIONS

When the population is much larger than the sample (e.g., $N > 20n$; Cox 2001) the hypergeometric distribution can be approximated with the binomial distribution. The general approach based on the hypergeometric distribution discussed up to this point is applicable to a sampling plan based on the binomial distribution, but a few details are different. The PMF for the binomial distribution gives the probability $f(m)$ of observing m typos in the sample:

$$f(m) = \binom{n}{m} p^m (1-p)^{n-m} \tag{5-1}$$

where

n = number of critical fields in the sample
m = the observed number of typos in critical fields in the sample
p = probability of a typo occurring, which is $\theta = 0.01$ in this example.

[11] Such as when the confidence interval is generated using a number of typos equal to the accept number.

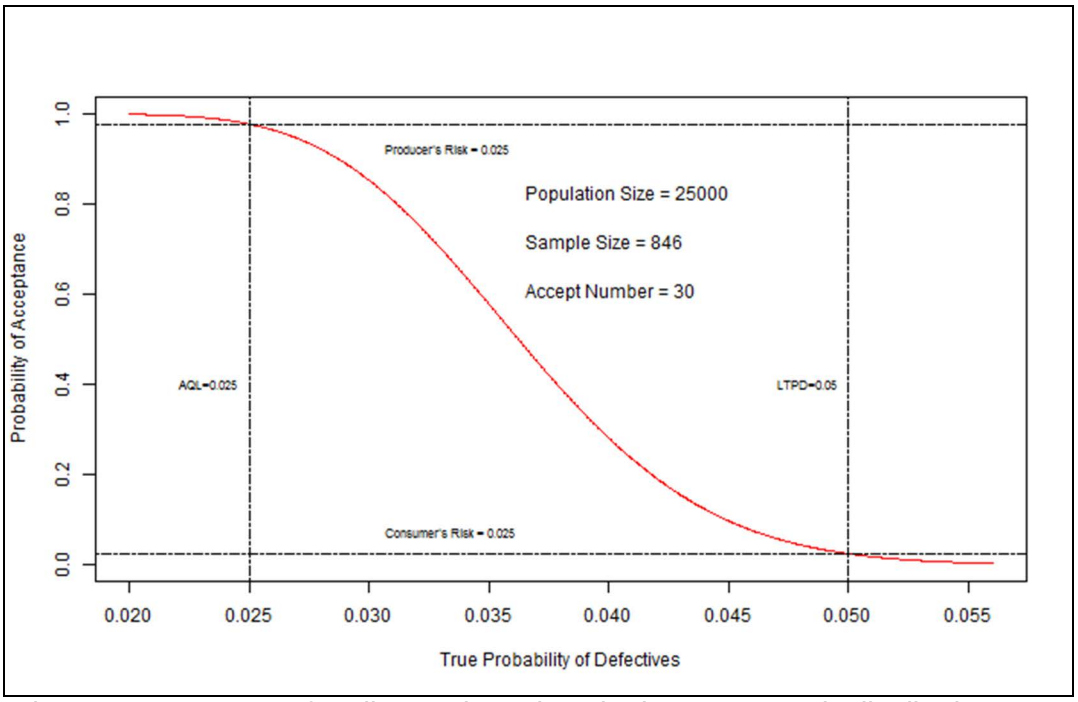

Figure 4-1. OC curve for all typos based on the hypergeometric distribution.

This calculation assumes that the population is infinite in size and taking a sample does not change the probability of success, or equivalently, that the sampling includes replacement. With this information, an OC curve based on the binomial distribution (which is referred to as a Type B curve) can be determined in the same way as the hypergeometric distribution (i.e., the consumer's risk and producer's risk are the same for both curves). Figure 5-1 shows the Type B OC curve calculated with the R software.

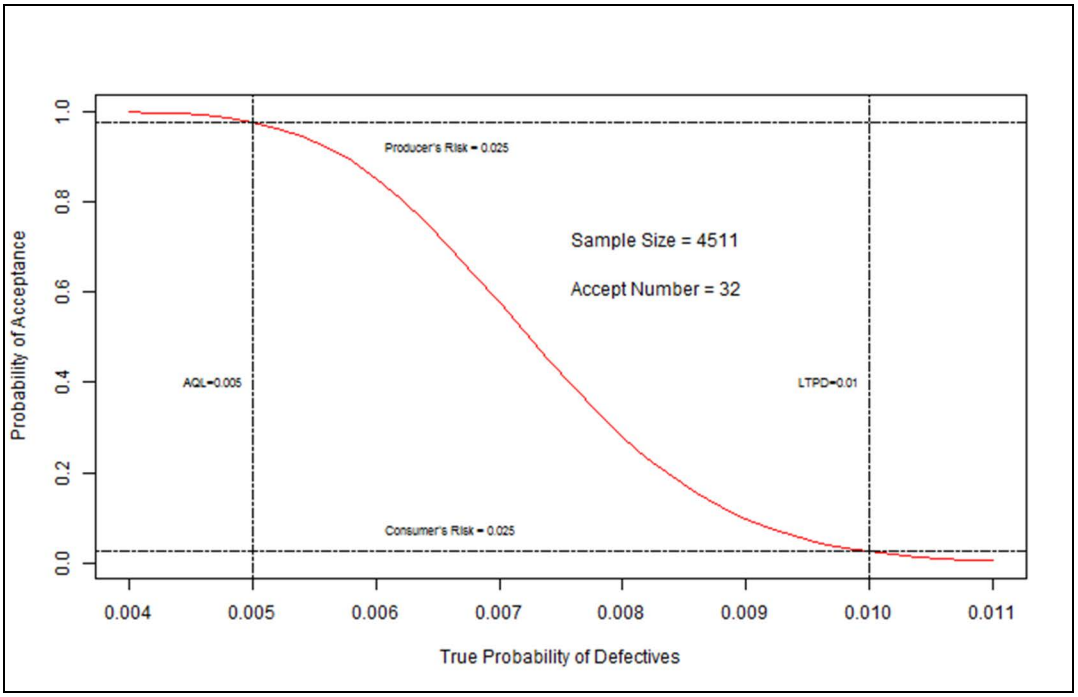

Figure 5-1. OC curve for sampling plan based on the binomial distribution (Type B curve).

Exact confidence intervals are constructed using the Clopper-Pearson method (Agresti 2011) as implemented in the *binGroup* package in R. This method was chosen because it guarantees at least 95% coverage. With large sample sizes such as those usually encountered in coworker datasets, the differences among methods are negligible. For example, if the sample size is n = 4,511 critical fields (as recommended by the OC curve) and there are m = 20 observed typos, there is 95% confidence that the true value of the typo rate in the population is between 0.0027 and 0.0068. Note that a typo rate of 0.01 is not within the confidence interval (i.e., the lot would be accepted).

6.0 PARADOXES AND PHILOSOPHICAL ISSUES

6.1 SAMPLE SIZE

Figure 3-4 details calculations for a sample size of n = 2,435 and a population size of N = 5,000. Here we are going to look at what happens to the sample size if all of the sample plan parameters (AQL, LTPD, consumer's risk, and producer's risk) are held the same as in Figure 3-4 but the population size is increased. Common sense might indicate, incorrectly, that a bigger population always needs a bigger sample. Figure 6-1 presents the sample size as a function of population size. The points are from the Type A (hypergeometric) OC curve, and the line is from the Type B (binomial) OC curve. Note that the sample size from the Type B curve is constant because it assumes that the population size is infinite. This curve demonstrates that as the population size increases the sample size from the Type A curve asymptotically approaches the sample size from the Type B curve, which is 4,511. This leads to the seemingly paradoxical conclusion that, for example, a sample size of around 3,000 from a population of 10,000 has virtually the same probability of accepting a bad lot and rejecting a good lot as a sample size of 3,000 from a population of 100,000,000. The intuitive explanation for this that is often given is that "one sip from a well stirred pot of soup is all that is needed to tell if it has enough salt."

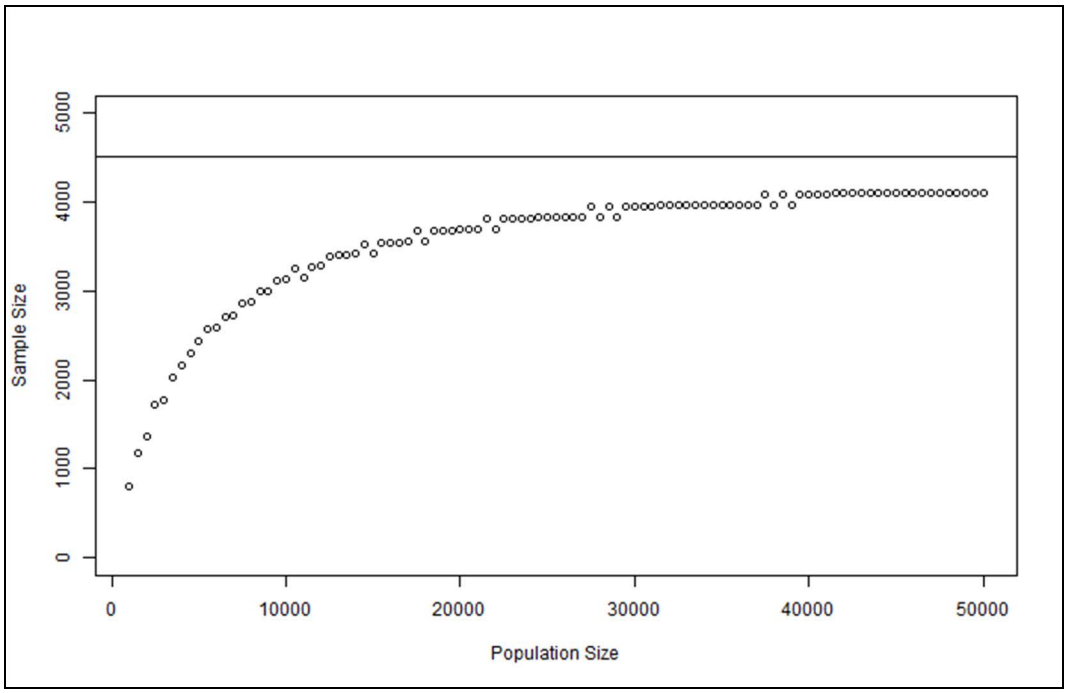

Figure 6-1. Sample size versus population size for given error rates.

6.2 CORRECTION OF TYPOS AND APPROACH TO LOT FAILURE

Typos in a particular sampling should of course be corrected for later use in modeling, but for the purpose of determining the acceptability of the typo rate for the entire electronic dataset, correction of typos observed in the sample would have little effect on the outcome.

There are two possible approaches to take when a lot fails:

1. Use the data “as is” anyway because, on consideration of other issues, it seems to be adequate for the proposed application.

2. Perform a census of the lot, correct all observed errors, and then perform the QA check (as described above) again.

Whether to implement these corrections is a command decision, not a technical decision. Note that a lot is tested and an acceptance decision made without consideration of other lots, which means that no multiple comparison adjustments are made in the calculations. Therefore, if there are two lots and only one fails, no further actions need be taken with the lot that passed.

6.3 SAMPLING FRAME

Two ways of sampling were considered:

- Randomly select fields from the electronic dataset and then pull the corresponding fields from the hard copies, or

- Randomly select fields from the hard copies and then pull the corresponding fields from the electronic dataset.

To develop a probability-based sampling plan as discussed in this report, the probabilities associated with observing any given field must be known. This in turn requires a "sampling frame," which is a complete listing of all fields. The sampling frame for the electronic dataset is immediately available once you have the electronic dataset, but the sampling frame for the hard copies does not exist and would have to be assembled by going through all of the hardcopies and listing the fields. However, the electronic dataset is the sampling frame for the hard copies as long as no fields in the hard copies are missing from the electronic dataset. Therefore, the sampling plan presented in this report uses the electronic dataset as the sampling frame under the assumption that there are no missing data. In other words, the sampling plan is designed to quantify the typo rate for fields that are transcribed from the hardcopy to the electronic dataset, not the rate at which data were not transcribed.

7.0 SUMMARY OF PROCEDURE

This report defines in detail the technical basis for estimating the typo rate in an electronic dataset. The basic approach is to use simple lot acceptance sampling methods to select the number of records from the electronic dataset that should be compared to the corresponding entries in the original hard-copy output to estimate the typo rate with a given level of confidence. This is essentially the same procedure one would use to sample lots of tomatoes to determine if the number of defective tomatoes in the lot was acceptable. When setting up the sampling program, the key decisions are:

- Defining the lot,
- The definition of a typo and how to identify them,
- The definition of the acceptable number (or fraction) of typos, and
- The conditions under which a lot is rejected, and if further actions should be taken when that occurs.

This report offers a precise definition of a typo and, with the guidance of NIOSH, an unacceptable fraction of errors in a lot. There is still some ambiguity in the definition of a lot and what should be done if a lot is rejected. These issues can be resolved on a case-by-case basis depending on the nature of the original records. Based on the technical discussion above, the ORAU Team proposes the following procedure for determining the typo rates in electronic datasets:

1. The first step is to define a lot. This is one of the most difficult steps in the procedure and will likely require the subject matter expert and statistician to collaborate. Two (of many) approaches to the selection of lots are annual lots and a single lot that encompasses the entire time frame of the data in the dataset.

2. The subject matter expert classifies fields as being critical, non-critical, or irrelevant.

3. The statistician determines the number of critical fields and all fields in each lot and constructs an OC curve to determine the sample sizes (for both critical fields and all fields) required for that lot.

4. The statistician pulls fields at random from the electronic dataset until the required number of critical fields is obtained. This procedure is repeated for all fields, and the two lists are consolidated and used for the comparison to the hard-copy records.

5. The numbers of critical field typos and all-field typos are determined.

6. The statistician calculates the confidence intervals for the critical-field and all-field typo rates, to determine if the target typo rates ($\theta = 0.01$ for critical-field typos and $\theta = 0.05$ for all-field typos) are above the upper limit of the respective confidence intervals.

7. Repeat the process for all lots.

8.0 EXAMPLE

An in vitro bioassay dataset from a large U.S. Department of Energy site was obtained that had N = 157,336 critical fields (one critical field per record) from 1952 through 2008. Scans of the original hard-copy data were grouped into 1,658 image files, each of which contained results from one or more years. There were two noncritical fields in each record. Therefore, the number of all fields is N_a = 3 × 157,336 = 472,008. The lot is defined as all the usable records, so the Type B OC curve for critical-field typos in Figure 8-1, which is based on an AQL of γ = 0.005 and an LTPD of θ = 0.01, is applicable here. The required sample size is n = 4,511 critical fields: a random sample of 4,511 critical fields should be drawn from the 157,336 critical fields in the population (which is considered to be infinite). The Type B OC curve for all-field typos is shown in Figure 8-2. This curve results in a sample size of n = 874 fields (critical and noncritical) that should be drawn from the 472,008 fields in the dataset. Note that the two samples are independently drawn, which means that there is a chance that a given critical field can be present in both samples.

Assume there are 4 observed typos in critical fields and 33 in all fields. The 95% confidence interval for the critical-field typo rate is (using the R function *binCI*):

```
binCI(4511,4, conf.level = 0.95,alternative = "two.sided",method = "CP")

95 percent CP confidence interval

 [ 0.0002417, 0.002269 ]

Point estimate 0.0008867
```

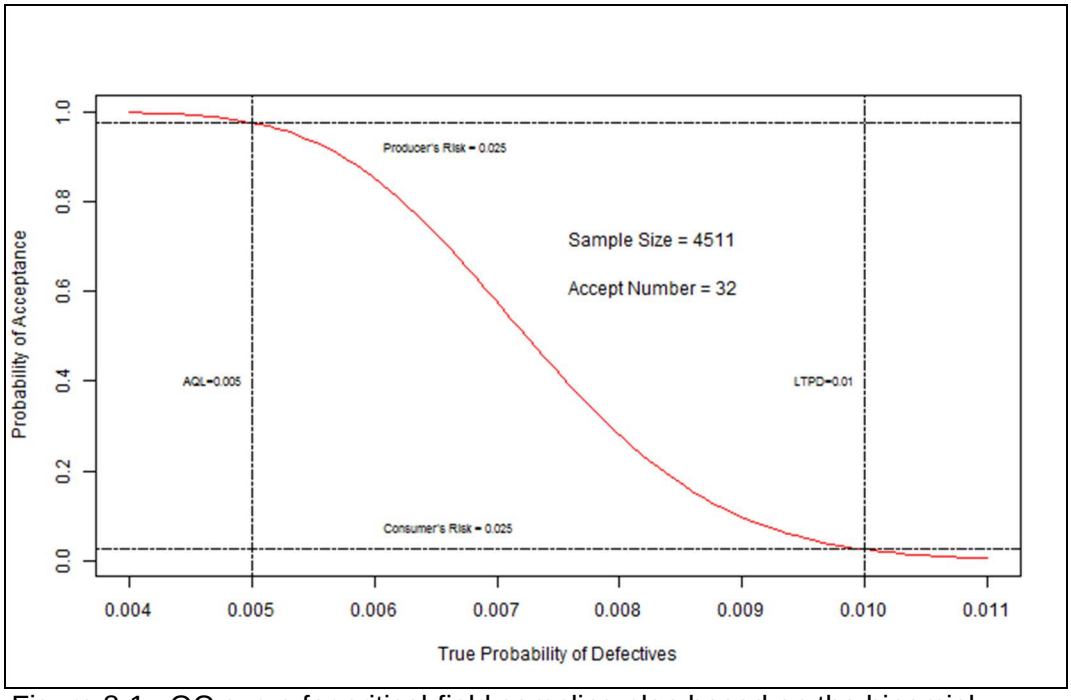

Figure 8-1. OC curve for critical-field sampling plan based on the binomial distribution (Type B curve).

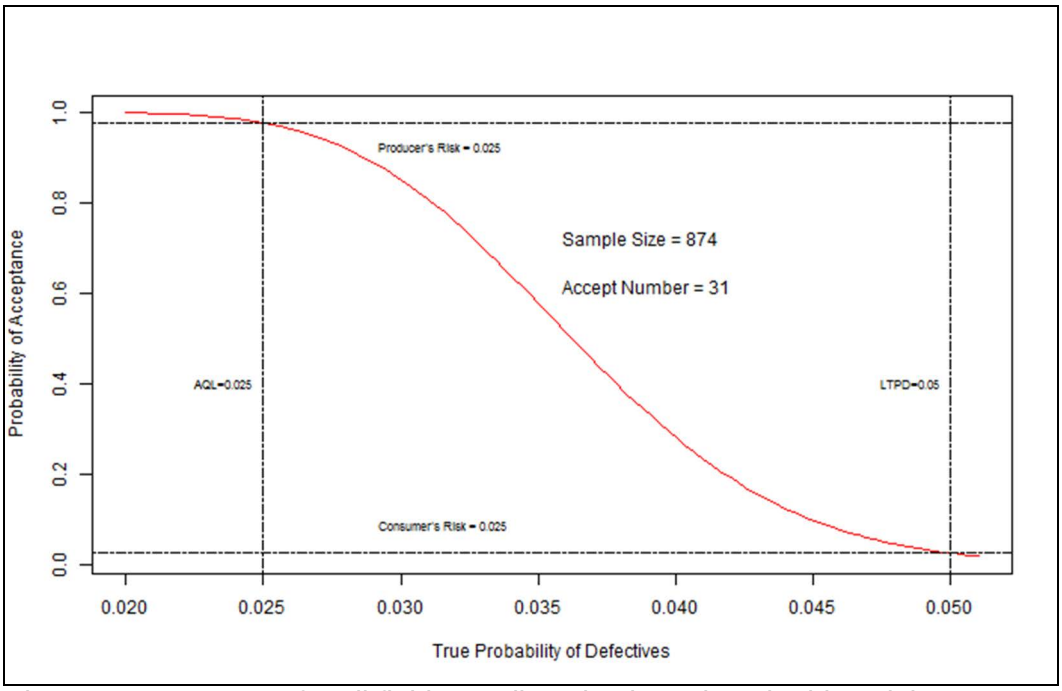

Figure 8-2. OC curve for all-field sampling plan based on the binomial distribution (Type B curve).

This has an upper limit that is less than a typo rate of 1%. However, the 95% confidence interval on the all-fields typo rate is:

```
> binCI(874,33, conf.level = 0.95,alternative = "two.sided",method = "CP")

95 percent CP confidence interval

 [ 0.02613, 0.05262 ]

Point estimate 0.03776
```

This has an upper limit that is greater than a typo rate of 5%. That is, the typo rate of 0.05 is within the confidence interval. The conclusion is that the dataset has typo rates that are unacceptable because of the all-fields typo rate and the dataset is rejected.

REFERENCES

Agresti, A., 2013, *Categorical Data Analysis*, John Wiley & Sons, Hoboken, New Jersey.

Buonaccorsi, J. P., 1987, "A Note on Confidence Intervals for Proportions in Finite Populations," *American Statistician*, volume 41, number 3, pp. 215–218.

Calhoun, G., 2015, "G2K - INL Co-worker model," National Institute for Occupational Safety and Health, Division of Compensation Analysis and Support, Cincinnati, Ohio, August 10. [SRDB Ref ID: 152486]

Cox, D., 2001, "The Binomial Approximation to the Hypergeometric," Rice University, Houston, Texas, February 12. [SRDB Ref ID: 152023]

Montgomery, D. C., 2005, *Introduction to Statistical Quality Control, Fifth Edition*, John Wiley & Sons, Hoboken, New Jersey.

Wright, T., 1991, *Exact Confidence Bounds when Sampling from Small Finite Universes*, Springer-Verlag, New York, New York.