United States Environmental Protection Agency

Office of Research and Development

National Human Exposure Assessment Survey (NHEXAS)

Maryland Study

Quality Systems and Implementation Plan for Human Exposure Assessment

Emory University
Atlanta, GA 30322

Cooperative Agreement CR 822038

Standard Operating Procedure **NHX/SOP-D05**

Title: Exploratory Data Analysis and Summary Statistics

Source: Harvard University/Johns Hopkins University

U.S. Environmental Protection Agency
Office of Research and Development
Human Exposure & Atmospheric Sciences Division
Human Exposure Research Branch

Notice: *The U.S. Environmental Protection Agency (EPA), through its Office of Research and Development (ORD), partially funded and collaborated in the research described here. This protocol is part of the Quality Systems Implementation Plan (QSIP) that was reviewed by the EPA and approved for use in this demonstration/scoping study. Mention of trade names or commercial products does not constitute endorsement or recommendation by EPA for use.*

1. Title of Standard Operating Procedure
Harvard University/Johns Hopkins University Standard Operating Procedure:
D05 Exploratory Data Analysis and Summary Statistics, Rev 1.1

2. Overview and Purpose
This SOP describes the methods and procedures for two types of QA procedures: spot checks of hand entered data and QA procedures for co-located and split samples. The spot checks will be used to determine whether the error rate goal for the input of hand entered data is being attained and correct any errors which are discovered. The QA procedures for co-located and split samples will check the variability in field equipment and the accuracy of analysis of samples performed by the relevant labs.

This SOP also describes the procedures which will be used to generated summary statistics which will be used as a preliminary representation of the data which can be used to look for unusual behavior in the data. These summary statistics will be of interest in themselves as well.

3. Discussion
After all the quality assurance procedures have been taken for the different methods of data entry, a final check will be performed to determine the success of these quality assurance procedures. These fall into two general categories: checking for errors in data entry and checking on the reliability of field and lab equipment. Spot checks will be used to check data entry and co-located and split samples will be used to check on the reliability of field and lab equipment.

[Not intended to be part of final SOP.]
This is a discussion of choice of frequency of spot checks and number and type of items to be selected. Arguments can be made whether spot checks need to made more frequently at the beginning or at the end of the study. Personally, I think evenly spaced spot checks will be just fine. I suggest either every half Cycle or every Cycle of questionnaire data, etc. Assume that the error rate for data entry is p. The number of items being entered does not really matter, beyond the fact that it is large. My approximation is that we will be entering about 300,000-500,000 items (depending on whether you count stuff we skip over) into the questionnaire and logsheet tables. Taking the smallest number, if we have p=0.01 we will end up with 3000 errors in the tables. Even if p=0.0001 there will be 30 errors in the tables. I do not think you want or need to be using percent of data set as the criterion for determining the number of items to be checked.

Not all entry values are created equal. Assume the double entry method is used and if the second person tries to enter something different from the first person the software immediately flags the error. This will certainly to be used by the Data Entry Company (and maybe by us?). For this type of double entry, mistakes will occur only when both people make the same mistake. This will be extremely rare for fill in the blank questions relative to other types of variables. I would imagine that Yes/No questions, Multiple Response questions, and Multiple choice questions will be the most likely source of errors and spot checks should focus on these. Admittedly this does not narrow the search much, but should help with checking the logsheets.

How many items need to be checked to be confident that the actual error rate is less than the maximum acceptable error rate? The test is simple: reject the null hypothesis (the value of which we don't really care about here) if any of the items checked is in error. The power calculation is straightforward:

$$P(\text{at least one error} \mid p) = 1-(1-p)^n,$$

where n is the number of items in the spot check and p is the maximum acceptable error rate. An example: if n=150 and you find at least one error, then there is an 80% chance that p>0.01. If you want to find how many samples are needed to check the Data Entry Company's claim for their error rate, pick your power probability (0.8, etc.) and try a few values of n for the value of p they are claiming.

Since we are doing several spot checks during the study we will have two power values, the individual spot check value and the overall spot check value. Presume that a random selection of items are to checked and they are independent from one spot check to another. This makes for a nice model for determining sample size. Let us start with the overall spot check rate. The following table gives error rates and sample size needed to detect those error rates for powers 0.8 and 0.9.

p=	0.99	0.995	0.999	0.9995	0.9999
Power=0.8	150	325	1750	3500	>10,000
Power=0.9	250	500	2500	5000	>>10,000

Sample sizes are rounded off. This shows that for around 2000 to 4000 observations total you can detect whether the error rate is worse than 0.001 with high probability. It also shows that to detect an error rate of 0.0001 or larger you will need to sample a really large number of items. The table also shows that samples of 150 will detect error rates of 0.01 with high probability. My suggested sampling schemes are given in the next table. They show the power of a single sample of sizes 175 and 250 and the overall power for a sequence of 8 or 16 spot checks.

Strategy	175 single	250 single		175 x 16 2800 overall	250 x 8 2000 overall	250 x 16 4000 overall
p=0.005	0.57	0.7	p=0.0005	0.75	0.63	0.86
p=0.01	0.80	0.90	p=0.001	0.9	0.86	0.98

In the spot checks you also need to keep in mind that going through the items will be tedious and the chance of overlooking an error will increase with the number of items checked at one sitting. I think 250 is probably close to an upper limit in the tedium scale. If one wants to make sure the error rate is less than 0.0001 I would suggest using a triple entry method and skip the spot checks. [This is the end of the comments on spot check not intended for final use. Pick your favorite numbers and include where appropriate.]

A special case of Exploratory Data Analysis is quality assurance testing of lab results. There are three types of quality assurance measures which will be studied: replicate analysis, duplicate analysis and lab comparisons. The labs will do the replicate analysis so no working data set will be needed. See SOP D03 for description on creating these data files.

Duplicate analysis measures the variability in taking observations in the field. The results needed for duplicate analysis will be stored in a worksheet derived from queries of the CDS. The only variables needed are: Sample ID, cycle, compound, media, original result and duplicate result. Using the duplicate analysis files, see SOP D03 for details on these files, tests will be carried out to check the precision of the field equipment. These files are Excel workbooks and Excel has the statistical capabilities to carry out the necessary tests so the test can be done with the duplicate analysis files. The tests will determine whether differences between co-located samples are significantly different from zero and the amount of variability in field equipment.

A small number (less than 10%) of field samples will be split and sent to at least two other labs besides the main lab for analysis. The reason for this is to judge the accuracy of the results from the main lab. The main lab samples and outside lab samples will be recombined into one file with the observations from different labs as different records (rather than different variables as is the case in duplicate analysis). Variables included are: Sample ID, cycle, compound, media, lab ID, and result. Separate worksheets will be created for each compound and media in the same manner as the duplicate analysis database. The lab ID is the variable which will be used to test for differences in the mean. A simple one way ANOVA with three groups (if three labs are involved) will be carried out to test whether any of the labs are different. Excel has the statistical capabilities to carry out the necessary tests so the test can be done with the testlab files.

4. Personnel Responsibilities

The Project Data Coordinator is responsible for:

- creating the list of items to be used in the spot check.
- modifying data entry protocol if necessary, in writing
- notifying the Data Entry Supervisor of any modifications that affect working databases
- carrying out the analysis on co-located and split samples
- reviewing summary statistics and carrying out any further analysis as warranted

The Data Entry Supervisor is responsible for

- training and supervising Data Entry Assistants
- assigning spot checks to Data Entry Assistants and maintaining Spot Check logbook
- resolving problems and ambiguities that Assistants are unable to handle

Data Entry Assistants are responsible for

- performing the spot checks
- generating the summary statistics
- retrieving the data sets requested by the Project Data Coordinator

5. Required Equipment and Reagents

- PC with Borland Paradox V5.0 for Windows
- PC with Microsoft Excel V5.0 for Windows
- files associated with the CDS, see SOP D02 and D03 for a list.
- Completed Questionnaires and Logsheets with data that has been entered into the CDS.

6. Procedure

Spot Checks of Questionnaire and Logsheet Data Entry

1. At periodic intervals, see discussion above, the Project Data Coordinator will select a number of randomly chosen items from the questionnaire and logsheet tables.
2. For each of these items the Data Entry Assistant will compare the original data sheets with the CDS to check for errors.
3. The number and description of error, if any, will be logged into the Spot Check logbook by the Data Entry Assistant.
4. Any errors will be brought to the attention of the Data Entry Supervisor and corrected in the CDS using the methods described in the relevant SOP D02 or SOP D04.
5. The Data Entry Supervisor will bring the errors to the attention of the Project Data Coordinator who will determine which avenue to take to correct problems in the data entry methods.

Lab Data Processing and Quality Assurance Testing
The Project Data Coordinator will measure the variability in co-located samples when sufficient data has returned from the labs. The first measurements will be performed soon after results from 5 to 10 co-located samples have been returned. The tests will be repeated periodically after that, depending on the rate of return of results from the respective labs. The Project Data Coordinator will measure the variability in differences between co-located samples and test to make sure that the mean difference is not significantly different from zero.

The Project Data Coordinator will test for differences in the mean of split samples when sufficient data has returned from the labs. The first tests will be performed soon after results from 5 to 10 split samples have been returned. The tests will be repeated periodically after that, depending on the rate of return of results from the respective labs. The Project Data Coordinator will use ANOVA to test for differences between the labs.

Collection of Summary Statistics
The collection of summary statistics will not be a formal part of the quality assurance procedure. The results are viewed as part of preliminary analysis and will be carried out approximately once per Cycle.

The Data Entry Assistant will:
1. run the *Paradox* queries which have been constructed and tested by the Project Data Coordinator. The Data Entry Assistant will run the queries at the beginning of each Cycle starting with Cycle 2.
2. store the results in the summary directory and copy the file to the summary back-up disk.
3. give the results to the Data Entry Supervisor.

The Data Entry Supervisor will in turn give the results to the Project Data Coordinator who will review the results and select useful items to disseminate among the other researchers so they can see some of the fruits of their labor.

7. Training of Data Entry Assistants
See SOP D04 for training methods to be used to familiarize Data Entry Assistants with the manipulation of data sets.

The Data Entry Supervisor will show the Data Entry Assistants a sample list of selected items for the spot check and how to enter error information into the Spot Check logbook. The Data Entry Supervisor will also show them how to sign off on the spot check.

The data entry Assistant will then take the sample list and compare the sample questionnaire with the sample data set and check for errors.

The data set will contain errors so the Data Entry Assistant will be able to practice entering errors in the Spot Check logbook.

The Data Entry Assistants will be considered trained when they can perform these tasks without error.

8. Quality Assurance Procedures

These procedures are themselves quality assurance procedures.

The Data Entry Assistant will be properly trained.

The Project Data Coordinator will have previous experience in the statistical methods needed for testing of co-located samples and split samples.

9. References

Borland Paradox Relational Database V5.0 User's Guide and Online Help.
Microsoft Excel V5.0 User's Guide and Online Help.
Harvard University/Johns Hopkins University Standard Operating Procedures:
D01 Data Flow Procedures, Rev 1.1
D02 Questionnaire Data Entry and Preparation, Rev 1.1
D03 Lab Results Data Entry and Preparation, Rev 1.1
D04 Logsheet and Confidential Questionnaire Data Entry and Preparation, Rev 1.1