United States Environmental Protection Agency	Office of Research and Development Washington, D.C. 20460	EPA/600/R-96/085 December 1997

Data Quality Evaluation Statistical Toolbox (DataQUEST) User's Guide

EPA QA/G-9D

QA96 Version

DISCLAIMER

The Data Quality Evaluation Statistical Toolbox (DataQUEST) Software and documentation are provided “as is” without guarantee or warranty of any kind, expressed, or implied. The Quality Assurance Division, U.S. Environmental Protection Agency, or the United States Government will not be liable for any damages, losses, or claims consequent to use of the software or documentation.

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not constitute or imply its endorsement, recommendation, or favoring by the U.S. Environmental Protection Agency or the United States Government.

TABLE OF CONTENTS

Chapter **Page**

LIST OF FIGURES ii

LIST OF TABLES ii

1. BACKGROUND 1

2. INSTALLATION AND STARTUP 3
2.1 OVERVIEW OF QA96 VERSION 3
2.2 SYSTEM REQUIREMENTS 3
2.2 INSTALLATION 3
2.3 STARTING DATAQUEST 3
2.4 EXITING DATAQUEST 4
2.6 DATA FILES 4
2.7 USER'S GUIDE 4
2.8 TROUBLESHOOTING 4

3. PROGRAM OVERVIEW 5
3.1 USING DATAQUEST 5
3.2 DATA FILES 6
3.2.1 Creating a Data File 6
3.2.2 Sample Data Files 8
3.2.3 Changing the Data File 8
3.2.4 Viewing the Data File 8
3.3 SAVING THE OUTPUT TO A FILE 8
3.3.1 Saving Text Files 9
3.3.2 Saving Data Files 10

4. STATISTICAL QUANTITIES AND GRAPHICAL REPRESENTATIONS 11
4.1 STATISTICAL QUANTITIES 11
4.2 GRAPHICAL REPRESENTATIONS 12

5. HYPOTHESIS TESTS 15

6. STATISTICAL TOOLS 17
6.1 Distributional Assumptions 17
6.2 Independence/Trends 17
6.3 Outliers 17
6.4 Dispersion 19
6.5 Transformations 19
6.6 Data Below Detection Limit 20

LIST OF FIGURES

Page

1. Relationship Between DQA Guidance (EPA QA/G-9), DataQUEST (EPA QA/G-9D), and the DQA Process 2
2. Example Statistical Quantities Screen 11
3. Example Graphical Representations Screen 12
4. Primary Hypothesis Tests Screen 15
5. Statistical Tools Screen 17

LIST OF TABLES

Page

1. Keystrokes for DataQUEST 5
2. Creating a Data File 7
3. Sample Data 8
4. Example Creating the File 9
5. Statistical Quantities 12
6. Graphical Representations 13
7. Inputs for the Primary Hypothesis Tests 15
8. One-Sample Hypothesis Tests Available in DataQUEST 16
9. Tests for Distributional Assumptions 18
10. Tests for Outliers Available in DataQUEST 19

1. BACKGROUND

The Data Quality Evaluation Statistical Toolbox (DataQUEST) is a quality assurance management tool for performing baseline Data Quality Assessment (DQA), i.e., to determine whether a set of data is adequate for its intended use. DataQUEST is designed to provide a quick and easy way for managers and analysts or those not familiar with standard statistical packages to review data and verify assumptions that are important in implementing the DQA process.

DQA is a quantitative process that employs statistical methods to determine whether a set of data will support a particular decision with an acceptable level of confidence. DQA consists of 5 steps:

1. Review the Data Quality Objectives (DQOs) and sampling design

2. Conduct a preliminary data review

3. Select the statistical test

4. Verify the assumptions of the statistical test

5. Draw conclusions from the data

EPA's Quality Assurance Division (QAD) has developed a document[1] that provides guidance on performing DQA. DataQUEST supplements this guidance by implementing the techniques discussed in the technical portions of the guidance. Both the DQA guidance and DataQUEST will be updated periodically to allow topics to be expanded and revised. These updates will be labeled QA96, QA98, etc. The current version of the DQA guidance is QA96. Similarly, this is the QA96 version of DataQUEST.

DataQUEST implements the procedures covered in Chapters 2-4 of the DQA guidance (see Figure 1). The technical sections of these chapters cover preliminary data review, performing a hypothesis test, and verifying statistical assumptions, respectively. Preliminary data review includes graphical representations of data and statistical quantities that are used to evaluate data. Preliminary review should be performed whenever data are used regardless of whether or not they are used to support a decision. When data are used to support a decision, then a statistical hypothesis test should be performed, and the assumptions underlying the test should be verified. Statistical assumptions fall into several categories, including distribution, independence, and dispersion. Trends and outliers also need to be identified and considered while selecting a statistical test. Data sets containing data below the detection limits may need to be modified before a statistical test is performed. DataQUEST contains all of these baseline DQA capabilities.

DataQUEST has been developed to provide simple statistical tools to a wide audience. Although standard statistical packages contain simple statistical tools, these packages are often difficult to use, especially when they contain their own programming language. Common spreadsheet packages deal more with discreet data in their graphics capabilities and do not contain many standard statistical graphs

[1] U.S. Environmental Protection Agency, 1996. *Guidance for Data Quality Assessment: Practical Methods for Data Analysis, EPA QA/G-9 QA96 Version*. EPA/600/R-96/084. Office of Research and Development.

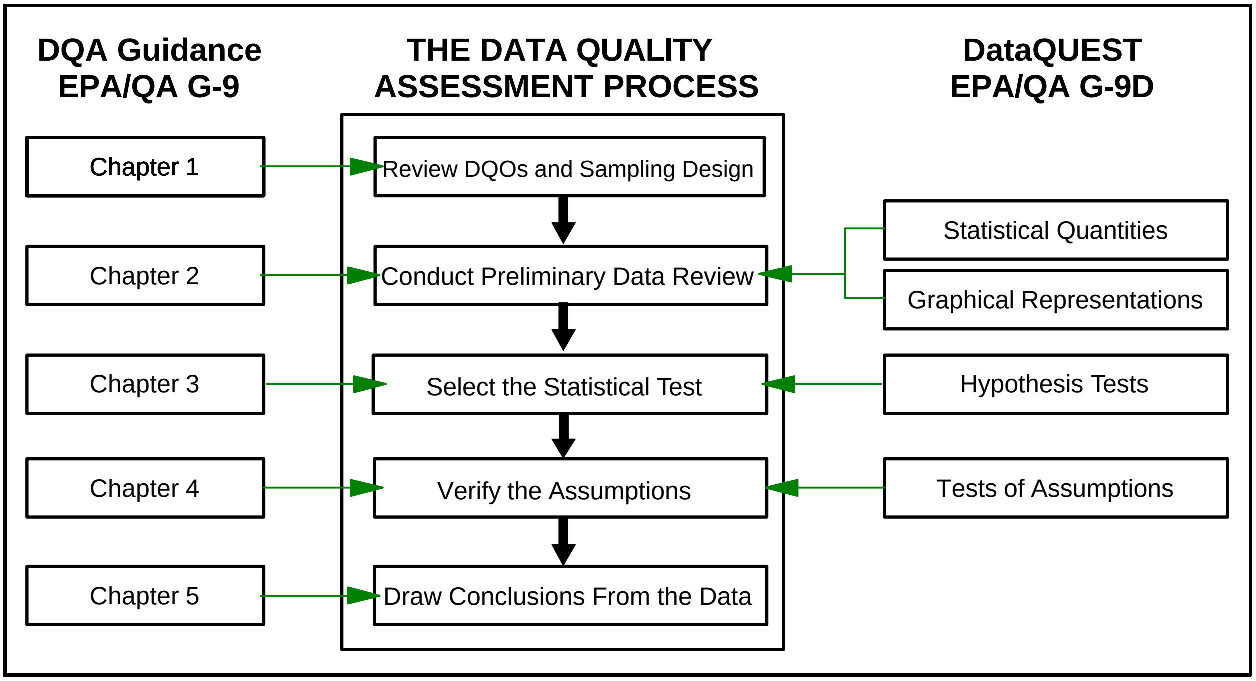


Figure 1. Relationship Between DQA Guidance (EPA QA/G-9), DataQUEST (EPA QA/G-9D), and the DQA Process

(for example, probability plots) or statistical analysis methods. The DataQUEST software is designed for minimal start-up effort on the user's part since it does not require any special language or commands.

There are many large general-purpose commercial statistical software packages designed for statistical analysis (e.g., SAS, Statistica, S Plus, etc.). DataQUEST is not designed to replace any of these full scale statistical packages since DataQUEST was developed to provide simple statistical tools to a wide audience. These packages contain more in-depth data analysis and data manipulation capabilities that DataQUEST does not attempt to duplicate. For complex data analysis, it is recommended that the analyst use a package that contains full statistical capabilities for that particular statistical application.

This software is not designed for statisticians or those who routinely perform in-depth data analysis. DataQUEST is designed to allow any manager or analyst to perform baseline DQA in order to determine when additional help (from a statistician) or additional investigation of a set of data is necessary. For preliminary data quality assessment, DataQUEST should be useful for the manager or analysts in quickly generating graphical displays and summary statistics for a set of data or verify assumptions underlying a statistical hypothesis tests.

This user's guide contains detailed instructions on how to use the DataQUEST software. This user's guide does not give in-depth instructions on performing DQA; for more information on this topic, consult the DQA guidance. In addition, the DataQUEST software was designed to supplement the DQA guidance. Therefore, descriptions of the statistical routines contained in the software are not repeated in the user's guide. It is strongly recommended that the DQA guidance be consulted when using the DataQUEST software.

2. INSTALLATION AND STARTUP

2.1 OVERVIEW OF QA96 VERSION

The QA96 version of DataQUEST contains a subset of the tools contained in the QA96 version of the DQA guidance. The two-sample tests, the graphical representations of two or more variables, and some statistical tools for assumption verification are not implemented. These will be completed in the QA98 version of this software.

The current version of DataQUEST will only allow one variable with no more than 150 values. The data can not have any missing values and are assumed to belong to a simple random sample or a systematic simple random sample with or without compositing. Data generated from a more complex sampling design (e.g., a stratified random sample) require more sophisticated analysis techniques. However, DataQUEST may be used to analyze the data from these more complex designs by treating each complete unit of randomization as a separate data set. For example, DataQUEST may be used to create graphical representations or test hypotheses concerning each individual stratum.

2.2 SYSTEM REQUIREMENTS

DataQUEST should run on any IBM PC-compatible computer with 8 MBs of RAM using DOS or Windows-based operating system. The QA96 version of the software is DOS-based; however, the QA98 version of this software will be a Windows version.

2.2 INSTALLATION

The DataQUEST software can either be run from a floppy disk or the hard drive. It is recommended that the software be installed on a hard drive since this will significantly speed up the start-up time and provide a directory for storing data files. To run the software using the hard drive, first install the software. To do this, insert the DataQUEST floppy disk into either drive ‘a’ or drive ‘b,’ then type the following at the DOS prompt:

```
prompt>  a:<return>  (b:<return>)
prompt>  install a  (install b)
```

The installation program will install the DataQUEST software in the directory “c:\dquest.”

2.3 STARTING DATAQUEST

If DataQUEST is installed on the hard drive of the computer, start the software by typing the following at the DOS prompt:

```
prompt>  c:<return>
prompt>  cd \dquest<return>
prompt>  quest<return>
```

If DataQUEST is not installed on the hard drive of the computer, place the DataQUEST floppy disk into drive ‘a’ or drive ‘b.’ Then, at the DOS prompt type:

```
prompt>  a:<return>  (b:<return>)
prompt>  quest<return>
```

2.4 EXITING DATAQUEST

The user may exit DataQUEST by entering ‘X’ or ‘x’ on the four main screens: Statistical Quantities (Section 4.1), Graphical Representations (Section 4.2), Hypothesis Tests (Chapter 5), and Statistical Tools (Chapter 6).

2.6 DATA FILES

The DataQUEST software is designed to perform baseline analysis of a data file that the user must first create. Directions for creating a data file are contained in Section 3.2.1.

2.7 USER’S GUIDE

A copy of this user’s guide is contained with the software in the file “read.me.” This file can be read or printed using any word processing package. Note that the “read.me” version of the user’s guide does not contain all the figures and tables contained in the official printed version.

2.8 TROUBLESHOOTING

Below are some errors that the user may encounter while running the DataQUEST software and their solutions. If the software exits with an error message not listed below or exits unexpectedly, please contact the EPA Quality Assurance Division (QAD) at (202) 564-6881 or by e-mail at “ord-qad@epamail.epa.gov”.

• “Error: can’t set video mode.” - If this error appears, use another monitor. The software should run on most EGA, CGA, and VGA color and monochrome monitors. However, it will not run on a MDPA (monochrome) monitor and on other adapters/monitors that do not support graphics modes.

• “Error: can’t register fonts.” - The file “helvb.fon” must be in either the DataQUEST directory (c:\dquest) or on the floppy disk. If this file is missing, the software will not run. Contact QAD for another copy of this file.

• “Error reading NAME datafile.” - The software uses several data files. If a file is missing, the DataQUEST software will exit with this message when the tool that requires that data file is selected. Contact QAD for another copy of named data file.

3. PROGRAM OVERVIEW

3.1 USING DATAQUEST

The DataQUEST software consists of four main sections: statistical quantities, graphical representations, primary hypothesis tests[2], and statistical tools. Surrounding these sections are the input and output routines and the menu options. Each main section is contained on a screen, i.e., there is a Statistical Quantities Screen, a Graphical Representations Screen, a Primary Hypothesis Tests Screen, and a Statistical Tools Screen. At the bottom of each screen is a list of options that include selecting a different screen, changing the data file, or exiting the program. Table 1 describes the keystrokes to select different options and where these topics are described in both this User's Guide and the DQA guidance.

Table 1. Keystrokes for DataQUEST

Statistical Routines				
Option	**Menu Bar**	**Key-Stroke[a]**	**DataQUEST Chapter**	**DQA Guidance Chapter**
Statistical Quantities	(Q)uantities	Q	4	2.2
Graphical Representations	(G)raph Data	G	4	2.3
Hypothesis Tests	(H)ypo Tests	H	5	4
Statistical Tools	(T)ools	T	6	5
Input/Output Routines				
Option	**Menu Bar**	**Keystroke[a]**	**DataQUEST Section**	
New Data	(N)ew File	N	3.2.2	
View Data	(V)iew File	V	3.2.3	
Save Results/Save Data	(S)ave	S	3.3	
Exit DataQUEST	E(X)it	X	2.4	

[a] These keystrokes are not case-sensitive so either capital or lowercase letters may be used.

[2] Primary hypotheses refer to the statistical hypotheses that correspond to the user's decision. Other statistical hypotheses can be formulated to formally test the assumptions that underlie the specific calculations used to test the primary hypotheses. For simplicity, primary hypotheses will be referred to under Hypothesis Tests, the other hypotheses will be considered under Statistical Tools.

The DataQUEST software begins by prompting the user for the name of the file that contains the data; directions for creating this file are contained in Section 3.2, Data Files. DataQUEST then proceeds to read the data from the file. If the file does not exist, the software provides an error message and allows the user to enter a different name. If the name entered was correct, the user should either check the name of the data file or check which directory the file is in. If the file is not in the directory “dquest,” the user must move the file to the current directory. The DataQUEST software comes with two sample data files, “DATA1.DAT” and “DATA2.DAT”; these files contain 75 data points and 24 data points, respectively. These files will be used to describe the screens and outputs of the DataQUEST software in the sections below.

As the data are read, DataQUEST automatically computes the statistical quantities. Once the statistical quantities are calculated, DataQUEST displays the Statistical Quantities Screen (see Section 4.1). This screen is automatically the first screen shown in DataQUEST once the name of the data file is entered. After viewing the Statistical Quantities Screen, the user has the option of saving this screen, moving to the different screen, viewing the data, or exiting the software.

From any of the four main screens, the user has the option of selecting one of the statistical routines, selecting one of the other main screens, viewing the data, or exiting the program (see Table 1). Some of the options on the four main screens may lead to submenus which are discussed in the appropriate chapter of this user’s guide.

3.2 DATA FILES

3.2.1 Creating a Data File

The DataQUEST software is designed to perform baseline analysis of a data set. To create a data file containing the data set, enter the data values into a file separating each value by a space or by placing each observation on a separate line. The first value in the data file should be the sample size. The data file can be created using WordPerfect®, Lotus 1-2-3®, or other standard word processing or spreadsheet packages. After the data are entered, the file containing the data should be saved as an ASCII file (DOS text file) and be placed in the same directory as the DataQUEST software. Table 2 contains detailed directions and a example of creating a data file.

If the sample size entered in the first row of the data file (the written sample size) is greater than the actual sample size, the software will warn the user of the discrepancy and adjust the sample size accordingly. If the written sample size is less than the actual sample size, DataQUEST will only recognize the written sample size, not the actual number of data points. If the sample size is greater than 150 (the maximum number of data points), DataQUEST will truncate the data after 150 points.

Since DataQUEST will only recognize the actual sample size when the written sample size is greater than the actual sample size, the DataQUEST software does not recognize missing values. Therefore, if the data contain values below the detection limit, substitute the reported value (if available) or any easily recognizable number to represent these values. For example, substitute the detection limit, ½ the detection limit, or 0 for any non-detects when the reported value is not available. This way, the user can easily apply the methods described in Section 4.7 of the DQA guidance for dealing with data below the detection limit. However, it is recommended that the user substitute the reported value when it is available since tied values (identical values) may be useful for non-parametric statistical tests.

Table 2. Creating a Data File

A data file can be created using WordPerfect®, Lotus 1-2-3®, or other standard word processing or spreadsheet packages. To create a data file, perform the following steps:

STEP 1: Count the number of data values and enter this number as the first value in the data file. For example, if there are 7 data values, 7 will be the first number in the file.

STEP 2: Identify any missing values or data below the detection limit. If necessary, decide whether to substitute a fixed value as a marker for these data or to delete the observations from the data set. For example, all data reported as <4 where 4 is the detection limit should be included in the data file as 4, or 3, or any other number less than 4. However, any reported value for data below the detection limit should be used when available. Methods for dealing with data below the detection limit are described in Section 6.6.

STEP 3: Enter the data values into the data file. These values may be separated by a space, tab, or a return. For example, consider the following 6 data points: 1, 2, 3, 3, 2, 1. A file of this data may look like:

```
6 1 2 3 3 2 1
```

or

```
6   1       2       3       3       2       1
```

or may look like:

```
6
1
2
3
3
2
1
```

STEP 4: Save the data file as an ASCII file (DOS text file). This can be done by using <CTRL> F5, 1 in WordPerfect® 5.1, “Save As” in WordPerfect 6.1, or /PF in Lotus® 1-2-3. Most spreadsheet or word processing packages have an option to save a file as an ASCII file (DOS file).

STEP 5: Copy the file to the same directory as the DataQUEST software. If the software is installed on the hard drive, the data file should be placed in the directory “c:\dquest.” If the software will be run off a floppy disk, the data file should be placed on that disk.

3.2.2 Sample Data Files

The DataQUEST software contains two sample data files, "DATA1.DAT" and "DATA2.DAT." These two data files will be used in this user's guide to demonstrate the screens and outputs for DataQUEST. Table 3 shows the data contained in the two sample files; Table 4 contains an example of creating the data file "DATA2.DAT".

Table 3. Sample Data

DATA1.DAT contains the following 75 observations:

11.00	11.75	10.45	13.18	10.37	10.54	11.55	11.01	10.23	15.63	11.00
11.75	10.45	13.18	10.37	10.54	11.55	11.01	10.23	10.37	10.54	11.55
11.01	10.23	10.23	11.00	11.75	10.45	13.18	10.37	10.54	11.55	11.01
10.23	15.63	11.00	11.75	10.45	13.18	10.37	10.54	11.55	11.01	10.23
10.37	10.54	11.55	11.01	10.23	10.23	11.00	11.75	10.45	13.18	10.37
10.54	11.55	11.01	10.23	15.63	11.00	11.75	10.45	13.18	10.37	10.54
11.55	11.01	10.23	10.37	10.54	11.55	11.01	10.23	10.23		

DATA2.DAT contains the following 24 observations:

1850	1760	725	1710	1575	1475	1780	1790	1780	725	1790
1800	725	1800	1840	1820	1860	1780	1760	1800	1900	1770
1790	1780									

3.2.3 Changing the Data File

The user can select a different datafile by selecting 'N' or 'n' on any of the main four screens. The user is then asked to input the name of the data file in the same manner as when starting the software (see Section 3.1). DataQUEST then proceeds to the Statistical Quantities Screen using the new data.

3.2.4 Viewing the Data File

The user can view the data by selecting 'V' or 'v' on any of the four main screens. While viewing the data, the user is presented the option of saving the data to a file (see Section 3.3.2). From this screen, the user may proceed to any of the four main screens or exit DataQUEST.

3.3 SAVING THE OUTPUT TO A FILE

The DataQUEST software allows the user to save the results of the data analysis to a file. There are two separate types of output that can be saved: a text file (for statistical quantities and hypotheses test results) and a data file (for transformed data).

Table 4. Example Creating the File "DATA2.DAT"

Sulfate concentrations were measured for 24 data points. The detection limit was 1,450 mg/L and 3 of the 24 values were below the detection level. The 24 values are 1850, 1760, < 1450 (ND), 1710, 1575, 1475, 1780, 1790, 1780, < 1450 (ND), 1790, 1800, < 1450 (ND), 1800, 1840, 1820, 1860, 1780, 1760, 1800, 1900, 1770, 1790, 1780 mg/L. This example corresponds to the data used in Box 4.7-2 of the DQA guidance.

STEP 1: The number 24 will be the first value in the data file.

STEP 2: There are 3 values below the detection limit. Any number less than 1450 could be used to replace these values. In this example, the value 725 will be used to identify these data (725 = ½*Detection Limit)

STEP 3: A file of this data may look like:

```
24      1850   1760    725   1710   1575   1475   1780   1790
1780     725   1790   1800    725   1800   1840   1820   1860
1780    1760   1800   1900   1770   1790   1780
```

In this file, the data have been separated using tabs. In the example data file, DATA1.DAT, the data have been separated using returns.

STEP 4: This file has been saved as DATA2.DAT.

STEP 5: The file has been placed on the floppy disk containing DataQUEST.

This data file can now be used in DataQUEST. Since there are data less than the detection limit, the statistical tools for adjusting for non-detects should be used before performing data analysis.

3.3.1 Saving Text Files

DQA will save the statistical quantities and the results from the hypothesis tests and assumption verification in an output file. On screens where this option is available, the user can select 'S' or 's' to save the results to a file. The first time this option is selected, the user will be prompted to enter a filename. DOS name conventions are used (up to 8 characters plus an optional 3-character extension separated by a period). If a file already exists under the name entered by the user, the program will ask the user to either select a new name or to overwrite the existing file. Once a filename has been selected, the user may use this option to save any additional outputs to this file. The output file will be an ASCII file (DOS text file) that can be imported into any standard word processing software package.

3.3.2 Saving Data Files

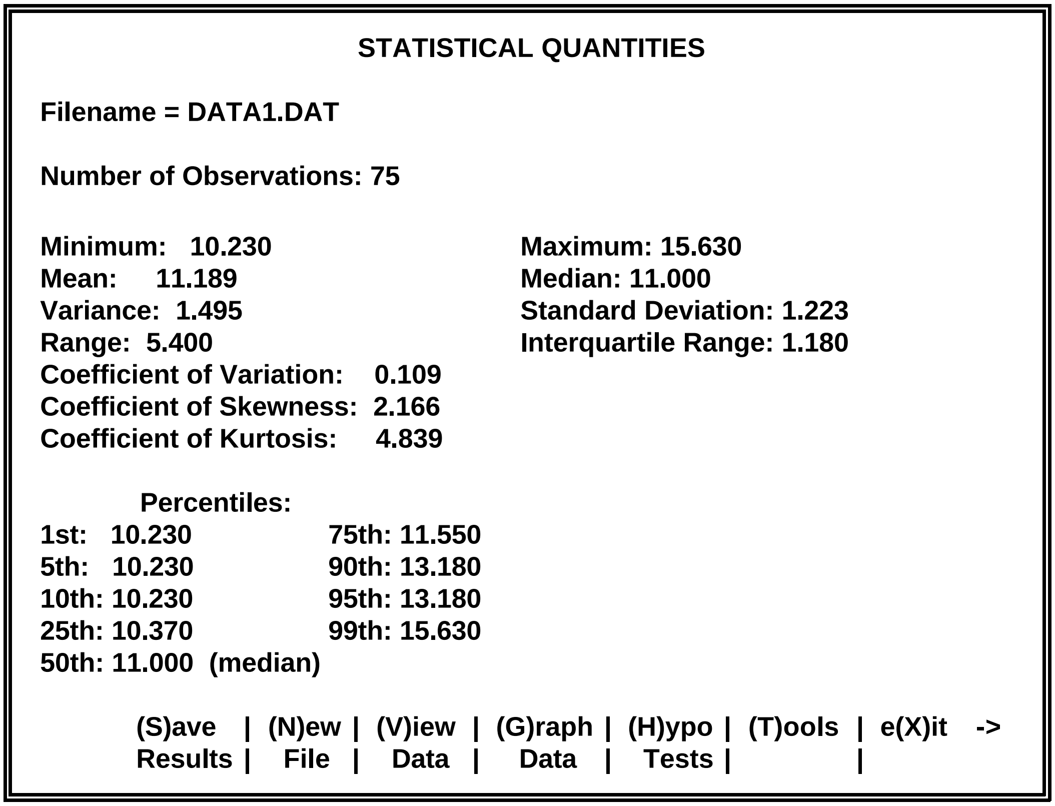
Transforming the data or using a substitution method to account for data below the detection limit creates a new data set. In this case, DataQUEST will ask if the data should be saved in a new data file. If so, the user will be prompted to enter a filename. An alternate way to save the data is available when viewing the data (select ‘V’ or ‘v’ on any of the four main screens.) At the bottom of the screen is an option to save the data along with the option to proceed to any of the four main screens or to exit the software. If the user chooses to save the data, DataQUEST will prompt the user to enter a filename.

DOS name conventions are used (up to 8 characters plus an optional 3-character extension separated by a period) when saving data files. If a file already exists under the name entered by the user, the program will ask the user to either select a new name or else to overwrite the existing file. The output file will be an ASCII file (DOS text file) that can be imported into any standard word processing or spreadsheet packages or can be used in DataQUEST.

4. STATISTICAL QUANTITIES AND GRAPHICAL REPRESENTATIONS

4.1 STATISTICAL QUANTITIES

The Statistical Quantities Screen (Figure 2) displays statistical quantities described in Chapter 2, Section 2 of the DQA guidance document. This screen is automatically the first screen shown in DataQUEST once the name of the data file is entered. In addition, the user can return to this screen by entering ‘Q’ or ‘q’ on any of the other main screens.



```
                         STATISTICAL QUANTITIES

Filename = DATA1.DAT

Number of Observations: 75

Minimum:   10.230                   Maximum: 15.630
Mean:     11.189                    Median: 11.000
Variance:  1.495                    Standard Deviation: 1.223
Range:  5.400                       Interquartile Range: 1.180
Coefficient of Variation:    0.109
Coefficient of Skewness:  2.166
Coefficient of Kurtosis:     4.839

          Percentiles:
1st:   10.230              75th: 11.550
5th:   10.230              90th: 13.180
10th: 10.230               95th: 13.180
25th: 10.370               99th: 15.630
50th: 11.000  (median)

        (S)ave   | (N)ew | (V)iew | (G)raph | (H)ypo | (T)ools | e(X)it  ->
        Results |  File  |  Data  |  Data   | Tests |         |
```

Figure 2. Example Statistical Quantities Screen

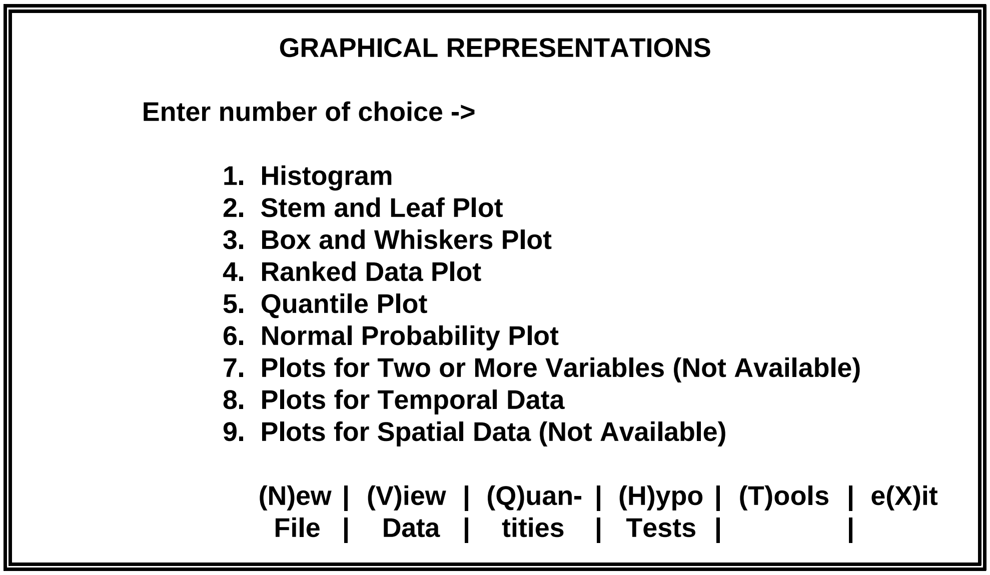
All statistical quantities described in the DQA guidance are implemented in DataQUEST except the correlation coefficient. These quantities are listed in Table 5. In addition to statistical quantities, the Statistical Quantities Screen also shows the filename, the sample size, number of values below the detection limit, the minimum and maximum of the data set, and the coefficients of skewness and kurtosis. The detection limit is initially set at zero. If a detection limit is set in the statistical tools section (see Section 6.6), this screen will show this value in addition to the number of non-detects. The correlation coefficient is not available in this version because only one variable is recognized.

Table 5. Statistical Quantities

Statistical Quantities	DQA Guidance Section
Coefficient of Variation	2.2.3
Interquartile Range	2.2.3
Median	2.2.2
Mean	2.2.2
Percentiles	2.2.1
Range	2.2.3
Standard Deviation	2.2.3
Variance	2.2.3

4.2 GRAPHICAL REPRESENTATIONS

The Graphical Representations Screen (Figure 3) of DataQUEST contains the graphical representations described in Chapter 2, Section 3 of the DQA guidance. On this screen, selecting any of the first six items produces that graphical representation. The selection of items 7-9 results in a sub-menu that lists the graphical representations relevant to that topic. The user can select the Graphical Representations Screen by entering ‘G’ or ‘g’ on any of the other main screens.



GRAPHICAL REPRESENTATIONS

Enter number of choice ->

1. **Histogram**
2. **Stem and Leaf Plot**
3. **Box and Whiskers Plot**
4. **Ranked Data Plot**
5. **Quantile Plot**
6. **Normal Probability Plot**
7. **Plots for Two or More Variables (Not Available)**
8. **Plots for Temporal Data**
9. **Plots for Spatial Data (Not Available)**

```
(N)ew | (V)iew | (Q)uan- | (H)ypo | (T)ools | e(X)it
 File |  Data  |  tities |  Tests |         |
```

Figure 3. Example Graphical Representations Screen

All graphs and plots described in the DQA guidance are implemented except the plots for two or more variables and the plots for spatial data since the current version of DataQUEST only recognizes one variable and these graphical representations require a minimum of two variables. The graphical representations in this version of DataQUEST are listed in Table 6 along with information needed to interpret these graphs.

Table 6. Graphical Representations

Graphical Representation	DQA Section	Key-Stroke	Notes
Histogram	2.3.1	1	The difference between the histogram and the frequency plot is minor for equal sized boxes so DataQUEST will only display a histogram. Minimum sample size is 9. The minimum number of boxes is 3, the maximum number of boxes is 7. The endpoint convention is that a data point is put into the largest box containing the observation. For example, the data point 2 would be placed in the box 2-4, instead of 0-2. The division of data into boxes for use in histograms depends on the subject under investigation. Interpretation of the meaning of a histogram becomes difficult when unfamiliar groupings have been made. For example, if the data concerned the number of children per family, interpretation of the data would be easy if the center point of each histogram box was a whole number (i.e., 2 children per family), but relatively difficult if an unfamiliar value was used (i.e., 2.71 children per family). DataQUEST notes the difficulty with interpreting unfamiliar groupings with the message "Care should be taken to ensure that the box sizes have contextual meaning."
Stem and Leaf Plot	2.3.2	2	The maximum number of stems is 10. DataQUEST automatically computes the number of stems based on the sample range. The range of the data must be less than 10,000 or greater than .0001 for this plot. For larger ranges and small ranges the data may be adjusted by factors of 10 for clarity in the plot. If so, the correction factor is noted at the bottom of the screen. The correct division of data into stems depends on the subject matter under investigation. It is important to note that interpretation of the meaning of diagrams becomes difficult when unfamiliar groupings have been made. This is especially true when default options are allowed to dictate the size and number of stems into which the data will be divided. DataQUEST notes the difficulty with interpreting unfamiliar groupings with the message "Care should be taken to ensure that stems sizes have contextual meaning."

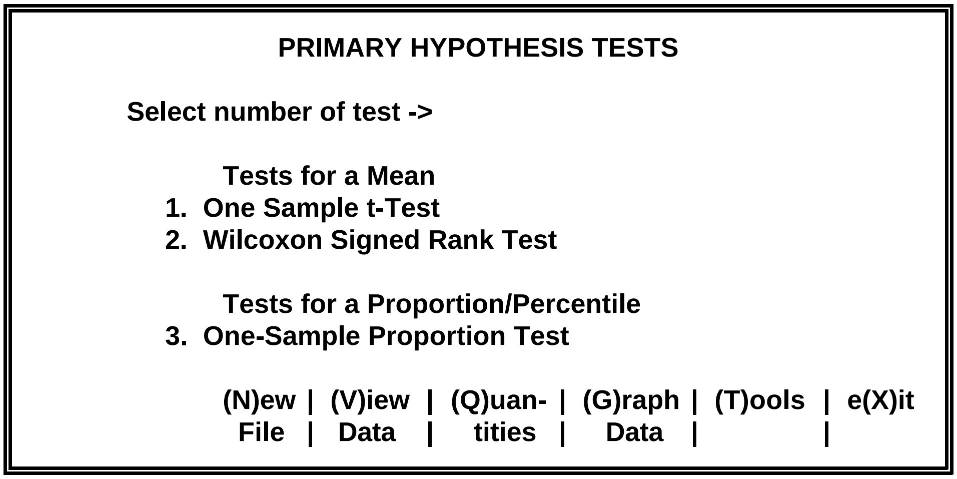
Table 6. Graphical Representations

Graphical Representation	DQA Section	Key-Stroke	Notes
Box and Whiskers Plot	2.3.3	3	
Ranked Data Plot	2.3.4	4	
Quantile Plot	2.3.5	5	
Normal Probability Plot	2.3.6	6	Note that the software does not use Normal Probability Paper to develop this plot. Therefore, data that are normally distributed will have an 'S' shape instead of forming a straight line.
Plots for Two or More Variables	2.3.7	7	Since this version (QA96) of DataQUEST only recognizes one variable, the plots for two or more variables are not available in this edition.
Plots for Temporal Data	2.3.8	8	
• Time Plot	2.3.8.1	8,1	
• Correlogram	2.3.8.2	8,2	
Plots for Spatial Data	2.3.9	9	This version of DataQUEST only uses one variable so the plots for spatial data are not available in this edition. The graphical representations are, however, available in the software Geo-EAS. Geo-EAS is a PC-based software package developed by the U.S. EPA Characterization Research Division for performing two-dimensional geostatistical analyses of spatially distributed data. The software package includes the capabilities for data file management, data transformations, univariate statistics, variogram analysis, cross validation, kriging, contour mapping, post plots, and line/scatter graphs. Geo-EAS is designed to produce 2-dimensional grids and contour maps of interpolated (kriged) estimates from sample data. Geo-EAS can read ASCII files and provides the spatial data analysis capabilities not contained in DataQUEST. Version 1.2.1 (EPA/600/8-91/008, April 1991) is available from NTIS (703) 487-4650, PB93-504967AS.

5. HYPOTHESIS TESTS

5. HYPOTHESIS TESTS

The Primary Hypothesis Tests Screen (Figure 4) contains the primary statistical tests described in Chapter 3 of the DQA guidance. The user should select the test to perform by entering the corresponding number. Once a statistical test has been performed, the user has the option of saving the results of the test in a file (see Section 3.3.1). The user can select the Primary Hypothesis Tests Screen by entering 'H' or 'h' on any of the other main screens.



```
                    PRIMARY HYPOTHESIS TESTS

         Select number of test ->

                Tests for a Mean
           1.  One Sample t-Test
           2.  Wilcoxon Signed Rank Test

                Tests for a Proportion/Percentile
           3.  One-Sample Proportion Test

                (N)ew |  (V)iew  |  (Q)uan-  |  (G)raph |  (T)ools  |  e(X)it
                 File  |   Data    |    tities   |    Data   |              |
```

Figure 4. Primary Hypothesis Tests Screen

In order to perform a statistical test, some basic information is necessary, depending on the specific test selected. This information should have been identified in Step 1 of the DQA process. Inputs required for the statistical tests contained in DataQUEST are described in Table 7. In this table, default value refers to the default significance level that appears the first time an input is required for a statistical test. The second time that input is required, the default value is the previously entered value. For example, if a t-test is first implemented and a significance level of 0.10 is selected, 0.10 will be the default value for the next test selected. In this table, the range is the limits on the valid entries.

Table 7. Inputs for the Primary Hypothesis Tests

Input	Default Value	Range	DQA Guidance Section
Action Level (AL)	Midpoint between Maximum and Minimum of the Data	None	1.1.2
Null Hypothesis	None	1. $\theta \leq$ AL 2. $\theta \geq$ AL 3. $\theta =$ AL	1.2
Type I Error Rate (α)	0.05	$0 < \alpha \leq 0.5$	1.1.3
Contaminant Level to Create Proportion	Midpoint between Maximum and Minimum of the Data	None	3.2.2

Hypotheses concerning population proportions and percentiles require that the data be reformatted into binary data (e.g., in 0's and 1's) in order to estimate the sample proportion. To translate the data, DataQUEST will prompt the user to enter the concentration level to use in developing the sample proportion. For example, if the user needs to determine if more that 90% of the concentration levels exceed 2 ppb, then 2 ppb is the contaminant level used to develop the sample proportion. Any values over 2 ppb are temporarily changed to 1 and any values under 2 ppb are temporarily changed to 0. (Note that 90% is the action level.) If the data are already binary coded (e.g., not present and present), then any value between 0 and 1 can be used so that the data are not reformatted; for example, a value of 0.5 may be used.

For the hypothesis tests implemented in DataQUEST, the data are assumed to belong to a simple random sample or a systematic simple random sample with or without compositing. Data generated from a more complex sampling design (e.g., a stratified random sample) require more sophisticated analysis techniques. However, the hypothesis tests in DataQUEST may be applied to data generated with a more complex design by treating each complete unit of randomization as a separate data set. For example, a hypothesis test may be performed on each individual stratum from a stratified random sample in order to make decisions regarding that individual stratum.

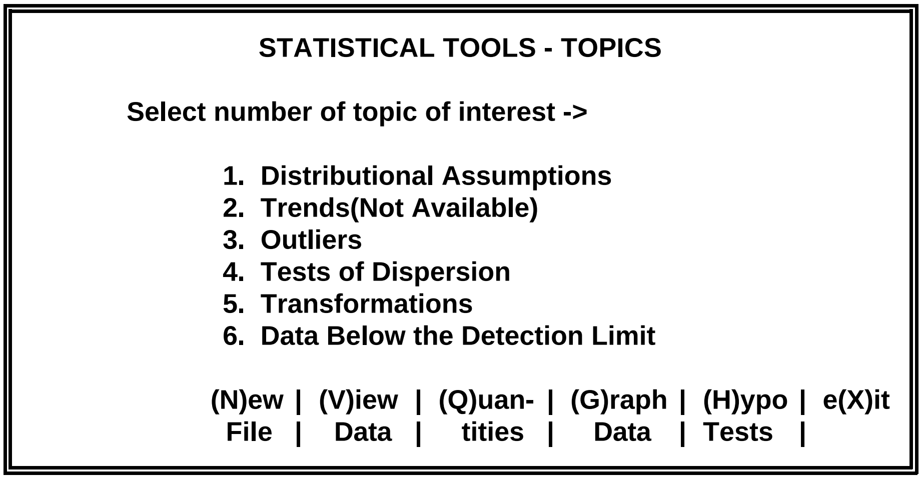
The one-sample tests available in DataQUEST are described in Table 8. Since this version (QA96) of DataQUEST only recognizes one variable, the two-sample primary hypothesis tests described in Section 3.3 of the DQA guidance are not available; these tests will be available in the QA98 version of DataQUEST.

Table 8. One-Sample Hypothesis Tests Available in DataQUEST

Test	Parameter	Required Inputs	DQA Guidance
One-Sample t-test	Mean	Action Level Type I Error Rate Null Hypothesis	Section 3.2.1.1 Directions: Box 3.2-1 Example: Box 3.2-2
Wilcoxon Signed Rank Test	Mean Median	Action Level Type I Error Rate Null Hypothesis	Section 3.2.1.2 Directions: Box 3.2-5, Box 3.2-7 Example: Box 3.2-6
One-Sample Proportion Test	Proportion Percentile	Action Level Type I Error Rate Null Hypothesis Concentration Level to Estimate Sample Proportion	Section 3.2.2.1 Directions: Box 3.2-8 Example: Box 3.2-9

6. STATISTICAL TOOLS

The Statistical Tools Screen of DataQUEST (Figure 5) contains a menu of the statistical tools topics described in Chapter 4 of the DQA guidance for verifying the assumptions underlying a primary hypothesis test. Each topic on this screen leads to a menu that list tools specific to that topic. Once a specific tool has been performed, the user has the option of saving the results of the test in a file (see Section 3.3.1). From any of the sub-topic screens, the user may select 'R' or 'r' to return to the Statistical Tools Screen.



```
                    STATISTICAL TOOLS - TOPICS

           Select number of topic of interest ->

                  1.  Distributional Assumptions
                  2.  Trends(Not Available)
                  3.  Outliers
                  4.  Tests of Dispersion
                  5.  Transformations
                  6.  Data Below the Detection Limit

         (N)ew |  (V)iew  | (Q)uan- | (G)raph | (H)ypo |  e(X)it
          File |   Data   |  tities |   Data  | Tests  |
```

Figure 5. Statistical Tools Screen

6.1 Distributional Assumptions

The distributional assumptions tools parallel Section 4.2 of the DQA guidance. Similar to the DQA guidance, DataQUEST contains tools for testing the assumption that the data are either normally or lognormally distributed. After selecting the distribution assumptions on Statistical Tools Screen, the DataQUEST software will ask whether the user wishes to test for normality or lognormality. When testing lognormality, all data must be greater than zero; if not, the software will provide an error message to the user. Once a distribution is selected, the software then displays a menu of applicable tests based on sample size. Table 9 contains a lists of the tests available for testing for distributional assumptions.

For most of the tests in this section, the user must select a significance level of either 1% or 5%. DataQUEST will prompt the user to select a significance level when this is necessary. The software then displays the results of the statistical test which can then be saved (see Section 3.3.1).

6.2 Independence/Trends

The tests for independence and trends described in Section 4.3 of the DQA guidance are not implemented in this version; these tests will be available in the QA98 version of DataQUEST.

6.3 Outliers

The outliers tools parallel Section 4.4 of the DQA guidance. Selecting this screen will produce a menu of statistical tests that can be used to test for statistical outliers, depending on the sample size.

Table 9. Tests for Distributional Assumptions

Test	Sample Size	Significance Level	DQA Guidance
Shapiro-Wilk Test	≤ 50	1. 0.01 2. 0.05	Section 4.2.2
Filliben's Statistic	≤100	1. 0.01 2. 0.05	Section 4.2.3
Coefficient of Variation Test	Any	Not Applicable	Section 4.2.4 Directions: Box 4.2-1 Example: Box 4.2-1
Coefficients of Skewness and Kurtosis Tests	> 50	1. 0.01 2. 0.05	Section 4.2.5
Studentized Range Test	≤ 1000	1. 0.01 2. 0.05	Section 4.2.6 Directions: Box 4.2-2 Example: Box 4.2-2
Geary's Test	> 50	1. 0.01 2. 0.05	Section 4.2.6 Directions: Box 4.2-3 Example: Box 4.2-4

From this menu screen, the user can either select one of the listed statistical tests or press 'R' to return to the main Statistical Tools Screen. Table 10 contains a list of all the statistical tests for outliers available in DataQUEST, what sample sizes that may be used with, the number and type of outliers they may be used to test for, inputs required to apply these tests, and where more information may be found on these tests in the DQA guidance.

In DataQUEST, there must be at least three data points in order to test for outliers, although a larger sample size is recommended. Additionally, the software will also not allow the user to test whether more than 50% of the data are statistical outliers. In this case, the user should consult with a statistician.

If a data point is found to be an outlier, the analyst may either: 1) correct the data point; 2) discard the data point from analysis; or 3) use the data point in all analyses. This decision should be based on scientific reasoning *in addition to* the results of the statistical test. One should never discard an outlier based solely on a statistical test. Discarding an outlier from a data set should be done with extreme caution, particularly for environmental data sets, which often contain legitimate extreme values. If an outlier is discarded from the data set, all statistical analysis of the data should be applied to both the full and truncated data set so that the effect of discarding observations may be assessed. For these reasons, the DataQUEST software will not allow the user to delete any data values found to be statistical outliers through the use of a statistical test for outliers. To delete a value, the user must exit DataQUEST, create a new data set (using the directions for creating a data set in Section 3.2.1), then rerun the DataQUEST software.

Table 10. Tests for Outliers Available in DataQUEST

Statistical Test	Sample Size (n)	Significance Level	Number and Type of Values to Test	DQA Guidance
Extreme Value Test (Dixon's Test)	n≤25	1. 0.01 2. 0.05	1. Minimum 2. Maximum	Section 4.4.3 Directions: Box 4.4-1 Example: Box 4.4-2
Discordance Test	n≤50	1. 0.01 2. 0.05	1. Minimum 2. Maximum	Section 4.4.4 Directions: Box 4.4-3 Example: Box 4.4-4
Rosner's Test	n≥25	1. 0.01 2. 0.05	1. Most Extreme Value 2. 2 Most Extreme Values 3. 3 Most Extreme Values 4. 4 Most Extreme Values 5. 5 Most Extreme Values	Section 4.4.5 Directions: Box 4.4-5 Example: Box 4.4-6
Walsh's Test	n≥50	0.10	1. Smallest Values (from 1 to ½n) 2. Largest Values (from 1 to ½n)	Section 4.4.6 Directions: Box 4.4-7

6.4 Dispersion

The only tool contained in Section 4.5 of the DQA guidance implemented in this version of DataQUEST is the confidence intervals for a single variance as the remaining tests, the F-test, Bartlett's Test, and Levene's Test, all require more than one variable. These tests will be implemented in the QA98 version of DataQUEST. The only input required for the confidence intervals is the confidence level. This level must be greater than 0 and less than 0.5; the default value is 0.05. Confidence intervals for a single variance are described in Section 4.5.1 of the DQA guidance.

6.5 Transformations

The transformation section of DataQUEST contains common transformations that may be applicable to environmental data including:

- Logarithmic Transformation, ln(x);
- Exponential Transformation, e^x;
- n^{th} Power Transformation, x^n ;
- n^{th} Root Transformation, $\sqrt[n]{x}$; and
- Reciprocal Transformation, (1/x).

These transformations are described in Section 4.6 of the DQA guidance.

Once the data have been transformed, all statistical analysis must be performed on the transformed data. No attempt should be made to transform the data back to the original form because this can lead to biased estimates. For example, estimating quantities such as means, variances, confidence limits, and regression coefficients in the transformed scale typically leads to biased estimates when transformed back into original scale.

If the user selects either the logarithmic, exponential, or reciprocal transformations, DataQUEST will ask the user to verify that the transformation should be performed because the original data will not be saved. The user may either enter 'Y' to perform the transformation or 'N' to select a different transformation or no transformation at all. If a logarithmic transformation is selected and the data contain negative values, DataQUEST will provide the user with an error message. Once the transformation is performed, the user has the option of saving the transformed data to a data file; see Section 3.3.2, Saving Data Files, for more information on this option.

If the user selects a power or root transformation, DataQUEST will ask the user to enter either the power to which the data should be raised or the root to be taken (i.e., '*n*' in the equations above). This number, *n*, can not be larger than nine. In addition, if an even root of the data is to be taken (e.g., the square root or fourth root), the data must all be greater than or equal to zero (positive). At this time, DataQUEST also provides the warning that the data may not be transferred back into its original form. Once the transformation is performed, the user has the option of saving the transformed data to a data file. See Section 3.3.2, Saving Data Files, for more information on this option.

6.6 Data Below Detection Limit

The "Data Below the Detection Limit" tools implement the methods described in Section 4.7 of the DQA guidance. Similar to the DQA guidance, DataQUEST contains substitution methods for replacing non-detects with a fixed value and methods for estimating means and standard deviations that account for the data below the detection limit.

After the user selects a method for dealing with data below the detection limit, DataQUEST prompts the user to enter the detection limit. The default value is the minimum value of the data set so that the default is that there are no data below the detection limit in the data set. There are no lower bounds on the detection limit entered by the user. If the detection limit is smaller than all the data, DataQUEST will tell the user that there are no data below the detection limit and return to the detection limit menu. The detection limit is bounded above by the median of the data set. DataQUEST will not allow more than 50% of the values to be less than the detection limit. If this is the case, the user should consult a statistician before proceeding with the data analysis.

For the substitution methods (Section 4.7.1 of the DQA guidance), it is important to note that once the data have been transformed, DataQUEST will not allow the user to reverse the substitution. This means that all further analysis will be performed on the new data. DataQUEST does not save the original data because substitution for data below the detection limit should not be a matter of trial and error. Therefore, if the user presses 'Y,' the original data are deleted from memory and all further analysis is then performed with all the data below the detection limit substituted with either the detection limit, ½ the detection limit, or a value input by the user.