

Auditor's Guide to Data Analysis and Testing

**Using automated processes
to perform analytical audit procedures**

Auditor’s Guide to Data Analysis and Testing

OVERVIEW OF DATA ANALYSIS.....	2
OVERVIEW OF TYPICAL AUDIT STEPS.....	2
TYPES OF AUDIT TASKS SUITABLE FOR EZ-R STATS FOR WINDOWS	3
DETERMINING IF THE POPULATION CONFORMS WITH THAT EXPECTED USING BENFORD’S LAW	3
IS THE DATA REASONABLE? – A STATISTICAL VIEW	3
WHAT IS THE DEGREE OF VARIABILITY	3
WHAT ARE THE POPULATION CHARACTERISTICS?	3
TO WHAT EXTENT IS THE POPULATION SKEWED FROM A NORMAL DISTRIBUTION	3
DOES THE POPULATION CONTAIN DATA OUTLIERS?	4
ARE THERE LINEAR RELATIONSHIPS BETWEEN THE VARIABLES?	4
HOW CAN THE POPULATION BEST BE CHARACTERIZED?	4
DOES THE TIME SERIES DATA CONTAIN ANY “SPIKES” OR UNEXPECTED FLUCTUATIONS?	4
WHAT ARE THE RANGE OF VALUES FOR EACH DATA ELEMENT AND THEIR FREQUENCY?	4
WHAT ARE THE KEY MEASURES FOR EACH NUMERIC DATA ELEMENT.....	4
WHAT IS THE BEST WAY TO DRAW A RANDOM SAMPLE IN ORDER TO PERFORM TESTS AND DRAW A CONCLUSION ABOUT THE POPULATION AS A WHOLE	4
HOW CAN THE DATA BE CROSS-TABULATED IN ORDER TO DETERMINE IF THERE ARE ANY UNDERLYING RELATIONSHIPS?	5
WHAT IS THE BEST WAY TO APPROACH A “NEEDLE IN A HAYSTACK” TYPE INVESTIGATION?	5
HOW CAN THE DATA BEST BE SUMMARIZED?	5
WHAT THE TOP 10 (BOTTOM 10) VALUES?	5
DOES THE POPULATION CONTAIN ANY DUPLICATE VALUES?	5
IS THE POPULATION COMPLETE, I.E. DOES THE POPULATION HAVE MISSING SEQUENCE NUMBER VALUES? ..	5
UNUSUAL RELATIVE VALUES.....	5

Overview of Data Analysis

Data analysis is a technique used by auditors, researchers, investigators and others to reach a conclusion as to the character of the data being reviewed. Data analysis techniques consist of a number of various procedures which can provide answers to typical questions such as:

1. Do the numbers in the population conform with what would be expected from Benford's law, i.e. does the population contain unusual or unexpected values?
2. Is the data reasonable? For example are the maximum and minimum values both reasonable?
3. How variable is the data? Is this variation within that which would be expected?
4. What are the average, median and mode values of the population?
5. To what extent is the population skewed from that which would be expected from a normal distribution?
6. Does the data population contain any extreme values, i.e. values which are well outside the statistical norm?
7. Are there any linear relationships between the data elements?
8. How is the population characterized, as shown in a histogram of values.
9. Does the data contain any unexpected fluctuations or "spikes" in values?
10. What are the possible ranges of values for any single data element (domain)?
11. What are the statistics on each numeric element in the population (e.g. min, max, average, standard deviation, etc.)
12. How can a random selection be drawn in order to sample the population and characterize it based upon the sample?
13. How can data be cross tabulated between variables and does the data indicate that there is a relationship or correlation?
14. "Needle in a haystack" searches for unusual conditions.
15. Data summarization
16. What are the "top 10" (top 50) instances/ values?
17. Does the population contain duplicates?
18. Does the population have missing sequence numbers or other missing values?

Each of these areas are discussed below. Generally, there are a number of software tools which can facilitate the researcher's investigation. Many of these tools work in a similar fashion. Here we will illustrate these data analysis techniques using the investigative tool "EZ-R Stats for Windows". These same techniques may also apply to other software tools as well.

Overview of typical audit steps

The analytical process used in most audits will include the following eight steps:

Auditor's Guide to Data Analysis and Testing

1. Identification of data to be analyzed (i.e. data sources and data elements)
2. Transforming that data into a format suitable for import into a database
3. Define the tables needed to hold the data
4. Identify key fields which should be indexed to facilitate queries
5. Load the data into the tables
6. Design and test the queries
7. Review the output of the queries
8. Repeat/ refine the steps above

Types of audit tasks suitable for EZ-R Stats for Windows

Determining if the population conforms with that expected using Benford's Law

19. Do the numbers in the population conform with what would be expected from Benford's law, i.e. does the population contain unusual or unexpected values?

Is the data reasonable? – A Statistical View

20. Is the data reasonable? For example are the maximum and minimum values both reasonable?

What is the degree of variability.

21. How variable is the data? Is this variation within that which would be expected?

What are the population characteristics?

22. What are the average, median and mode values of the population?

To what extent is the population skewed from a normal distribution

23. To what extent is the population skewed from that which would be expected from a normal distribution?

Does the population contain data outliers?

24. Does the data population contain any extreme values, i.e. values which are well outside the statistical norm?

Are there linear relationships between the variables?

25. Are there any linear relationships between the data elements?

How can the population best be characterized?

26. How is the population characterized, as shown in a histogram of values.

Does the time series data contain any "spikes" or unexpected fluctuations?

27. Does the data contain any unexpected fluctuations or "spikes" in values?

What are the range of values for each data element and their frequency?

28. What are the possible ranges of values for any single data element (domain)?

What are the key measures for each numeric data element.

29. What are the statistics on each numeric element in the population (e.g. min, max, average, standard deviation, etc.)

What is the best way to draw a random sample in order to perform tests and draw a conclusion about the population as a whole

30. How can a random selection be drawn in order to sample the population and characterize it based upon the sample?

How can the data be cross-tabulated in order to determine if there are any underlying relationships?

31. How can data be cross tabulated between variables and does the data indicate that there is a relationship or correlation?

What is the best way to approach a "needle in a haystack" type investigation?

32. "Needle in a haystack" searches for unusual conditions.

How can the data best be summarized?

33. Data summarization

What the Top 10 (Bottom 10) values?

34. What are the "top 10" (top 50) instances/ values?

Does the population contain any duplicate values?

35. Does the population contain duplicates?

Is the population complete, i.e. does the population have missing sequence number values?

36. Does the population have missing sequence numbers or other missing values?

Unusual relative values

Do any elements of the population seem to have unusual values, as when compared with other values. For example, in a population of invoices from one vendor, does the largest invoice exceed the others by a significant factor?