ADVANCED STATISTICAL METHODS FOR FINDING FRAUD

Using more sophisticated methods to unveil suspicious transactions and detect potential fraud, this session will use real examples to show how fraud examiners can go beyond the usual tests and introduce more sophisticated, proactive ways to monitor control systems and identify fraud.

SUNDER GEE
Ottawa, Ontario
Canada

Sunder Gee is a Certified Management Accountant (CMA), with a Bachelor of Commerce degree from Concordia University, Montreal, Quebec. He has been working with a government agency since 1977, where he provides consultation on electronic commerce issues and computer-assisted audit techniques, both within the agency as well as to other authorities worldwide. Sunder also has extensive auditing experience, ranging from sole proprietorships to the largest companies in Canada in the resource and high-technology sectors.

Sunder has developed extensive training material on a wide variety of topics for his employer, private businesses, and local colleges. For example, he has developed presentations and workshops on electronic sales analysis, obtaining electronic data, auditing online vendors, forensic accounting, intelligence gathering, anti-money laundering and data analytics. He has given numerous workshops, presentations, and lectures in these areas to auditors, computer audit specialists, and college students.
Data Analytics

Data analytics is the statistical process used to analyse data that can identify anomalies, trends, patterns, and concerns. It is highly effective when applied to situations that involve large volumes of electronic data.

Traditional analytical methods include:
- Extract
- Sort
- Statistics
- Gaps
- Duplicates
- Aging
- Samples
- Summarize
- Stratify
- Join (Match)
- Compare

In order to effectively apply and interpret the results using traditional and advanced statistical methods, the auditor or investigator must have a good understanding of the business and industry involved as well as be familiar with the software used for the analysis.

Simple analytics can be done using Microsoft Excel or Access. Specialized commercial data mining and analysis software listed in the Fraud Examiners Manual are as follows:
- IDEA
  www.caseware.com/products/idea
- ACL
  www.acl.com/
- ActiveData for Excel
  www.informationactive.com
## Notes

- **AutoAudit**

- **SNAP! Reporter**

- **DataWatch Corporation’s Monarch for Windows**
  - [www.monarchprofessional.com/](http://www.monarchprofessional.com/)

- **Arbutus Query**
  - [www.arbutussoftware.com](http://www.arbutussoftware.com)

- **Oversight Systems**
  - [www.oversightsystems.com](http://www.oversightsystems.com)

### Advanced Statistical Methods and Data Analytics for Detecting Fraud

Many of the advanced methods outlined are discussed in detail in two books by Mark J. Nigrini, Ph.D.

- **Benford’s Law: Applications for Forensic Accounting, Auditing, and Fraud Detection**
  - By Mark Nigrini
  - (Foreword by) Dr. Joseph T. Wells
  - ISBN: 978-1-1181-5285-0
  - Published April 2012

- **Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations**
  - By Mark Nigrini
  - Published May 2011

Dr. Nigrini shows, in detail, how some of the tests can be performed in Access and Excel.

**Benford’s Law**

Benford’s Law states that the digits and digit sequences in a dataset follow a predictable pattern. Benford’s Law performs analyses of digits in numerical data that help identify anomalies such as systematic manipulation of data, potential fraud, and other irregularities. Benford’s
Law identifies unusual or excessive duplication of digits.

In a paper published by Frank Benford in 1938, “The Law of Anomalous Numbers,” the abstract states, “It has been observed that the first pages of a table of common logarithms show more wear than do the last pages, indicating that more used numbers begin with the digit 1 than with the digit 9.”

As seen in the table below, numbers leading with the first digit of 1 should occur 30 percent of the time while numbers starting with the first digit of 9 only occur with a 4.6 percent frequency.

**Expected (Natural) Frequencies**

<table>
<thead>
<tr>
<th>First Digit</th>
<th>Second Digit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>Frequency</td>
</tr>
<tr>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>0.30103</td>
</tr>
<tr>
<td>2</td>
<td>0.17609</td>
</tr>
<tr>
<td>3</td>
<td>0.12494</td>
</tr>
<tr>
<td>4</td>
<td>0.09691</td>
</tr>
<tr>
<td>5</td>
<td>0.07918</td>
</tr>
<tr>
<td>6</td>
<td>0.06695</td>
</tr>
<tr>
<td>7</td>
<td>0.05799</td>
</tr>
<tr>
<td>8</td>
<td>0.05115</td>
</tr>
<tr>
<td>9</td>
<td>0.04576</td>
</tr>
</tbody>
</table>

Data must form a geometric sequence for the digit patterns to conform.
NOTES

First Digit Test

PRIMARY BENFORD’S LAW TEST
- First digit
- First two digits
- First three digits
- Second digit

ADVANCED TESTS
- Summation
  * This test analyses the first two digits in the data by grouping the records of the first two digits and then computes the summation of each group. Using the computed summation
values, the process determines whether a uniform distribution is followed.
* The summation test identifies excessively large numbers compared to the rest of the data.
* This test is based on sums rather than on counts as in the other Benford’s Law tests.
* In theory, the sums of numbers with first two digits should be equal in weight or distribution. In actual data, there are normal abnormal duplications of large numbers whether they are a few very large numbers or a high volume of moderately large numbers. Additional analysis will be needed.
  - Second order
    * This test is based on the first two digits in the data. A numeric field is sorted from the smallest to largest and the value differences between each pair of consecutive records are checked to determine whether the first two digits of the obtained differences meets the expectations of the first two digits distribution.
    * The differences are expected to approximate the digit frequencies of Benford’s Law.

ASSOCIATED TEST
  - Last two digits
    * This test groups the last two digits and computes the grouped frequency to check whether it follows a uniform distribution.
    * The last two digits test is useful in identifying invented or rounded numbers in the data.
<table>
<thead>
<tr>
<th>NOTES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ADVANCED STATISTICAL METHODS FOR FINDING FRAUD</strong></td>
</tr>
<tr>
<td>* It is expected that the right-side two digits be distributed evenly. With 100 possible last two digits numbers (00, 01, 02, …, 98, 99), each should occur approximately 1 percent of the time.</td>
</tr>
<tr>
<td>* Practical applications include inventory counts, weight of fishery catch where a quota applies, vehicle odometer reading associated with warranty work, royalty or percentage of sales/usage, etc.</td>
</tr>
<tr>
<td>− Number duplication</td>
</tr>
<tr>
<td>* The number duplication test identifies specific numbers causing spikes or anomalies in primary and summation tests.</td>
</tr>
<tr>
<td>* Spikes in the primary tests are caused by some specific numbers occurring abnormally too often.</td>
</tr>
<tr>
<td>* Spikes in the summation test are caused by abnormally large numbers in value.</td>
</tr>
<tr>
<td>− Distortion factor model</td>
</tr>
<tr>
<td>* Shows whether the data has an excess of lower digits or higher digits.</td>
</tr>
<tr>
<td>* Assumes that the true number is changed to a false number in the same range as the true number.</td>
</tr>
</tbody>
</table>

*Trend Analysis*

Trend analysis is a statistical technique used to calculate the trend of data over time and to forecast values into the future on the assumption that the trending pattern will continue. One type of trend analysis technique is based on linear regression using the method of least squares. This technique will fit the best-fitting straight line to the data and will forecast ahead using this line.
Trend Analysis of Sales for the Alpha Store

*Time Series Analysis*

If the data has a seasonal component (higher values associated with certain months, days of the week, or times in the day), the time series statistical method should be used instead of trend analysis.

Predictive analysis from both trend analysis and time series can be used in a continuous monitoring environment. The analysis could be used to provide a forecast and the forecast data could then be compared to the actual data immediately after the event. Any difference between the forecast data and the actual data is an indication that the data has deviated from its past trend and some change must have taken place.
Correlation

Correlation is a statistical technique to compare the patterns in numerical values in two data sets. A high level of correlation where the score is close to 1 means that high values in one data set are matched with high values in the second data set. Conversely, low values in one data set are matched with low values in the second data set. Correlation measures how closely the two data sets move in tandem.

The correlation score shows how good this match is. The scores can range from -1 to +1. Where the score is close to -1, a negative correlation is indicated.

A perfect correlation will result in a score of +1 and a perfect reverse correlation will result in a score of -1.
Correlation of sales for five stores

**Z-Score**

The Z-Score statistical calculation standardizes the data and its distribution regardless of the amounts and variance within the data set by using the formula below.

\[
Z\text{-Score} = \frac{\text{Value} - \text{Mean}}{\text{Standard Deviation}}
\]

While indexing or sorting the data will provide the same order as the Z-Score, the Z-Score tells the user how far the data exceeds the norm.

Statistical theory states that in 99.7 percent of the time, the Z-Score will be between -3 and +3.
In Chapter 11 of Dr. Mark Nigrini’s *Forensic Analytics*, he introduces the RSF test, which expands on this concept and states, “In this chapter we compare large amounts to a benchmark to see how large they are relative to some norm, hence the name the *relative size factor test*. The relative size factor test is a powerful test for detecting errors. The test identifies subsets where the largest amount is out of line with the other amounts for that subset. This difference could be because the largest record either (a) actually belongs to another subset, or (b) belongs to the subset in question, but the numeric amount is incorrectly recorded. The relative size factor (RSF) test is an important error-detecting test.”
Outliers may not be the largest amounts in the entire dataset but are large in relation to particular members of the subset. Large differences might be attributed to errors such as the record belongs to another subset or that the amount was posted incorrectly (e.g., shifted decimal point). Large differences may also be an indication of fraudulent activity, such as occupational accounts payable fraud, falsified invoices for HST input tax credits, offset money laundering revenue, or product sales to related companies (offshore transfer pricing).

### Relative Size Factor Test Output

<table>
<thead>
<tr>
<th>SUPNO</th>
<th>SUPNAME</th>
<th>LARGEST_AMT</th>
<th>SECOND_LARGEST_AMT</th>
<th>AVERAGE_X_LARGEST</th>
<th>COUNT</th>
<th>RELATIVE_SIZE_FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>92231</td>
<td>ROBERT BJUM GMBH</td>
<td>32,293.40</td>
<td>545.09</td>
<td>545.09</td>
<td>2</td>
<td>59.24</td>
</tr>
<tr>
<td>92700</td>
<td>SAGE PUBLICATIONS INC.</td>
<td>25,943.21</td>
<td>795.59</td>
<td>795.59</td>
<td>2</td>
<td>32.61</td>
</tr>
<tr>
<td>92241</td>
<td>ROSS RIVER SERVICE CENTRE LTD.</td>
<td>70,509.14</td>
<td>4,073.20</td>
<td>4,073.20</td>
<td>2</td>
<td>17.31</td>
</tr>
<tr>
<td>21175</td>
<td>CORPORATE EXPRESS</td>
<td>9,474.74</td>
<td>1,198.00</td>
<td>1,198.00</td>
<td>2</td>
<td>7.91</td>
</tr>
<tr>
<td>40713</td>
<td>MACKENZIE PETROLEUM LTD.</td>
<td>59,169.76</td>
<td>7,854.97</td>
<td>3,748.19</td>
<td>4</td>
<td>7.57</td>
</tr>
<tr>
<td>21233</td>
<td>BISHOP PUBLISHING INC.</td>
<td>18,692.41</td>
<td>2,713.16</td>
<td>2,713.16</td>
<td>2</td>
<td>6.89</td>
</tr>
<tr>
<td>21490</td>
<td>BULOVA INC.</td>
<td>48,399.72</td>
<td>7,106.92</td>
<td>7,106.92</td>
<td>2</td>
<td>6.81</td>
</tr>
<tr>
<td>92611</td>
<td>MORRIS GOLD INC</td>
<td>21,653.24</td>
<td>3,685.75</td>
<td>3,685.75</td>
<td>2</td>
<td>5.71</td>
</tr>
<tr>
<td>40502</td>
<td>LOCKSHEED SERVICES LTD</td>
<td>5,432.93</td>
<td>1,263.89</td>
<td>1,263.89</td>
<td>2</td>
<td>4.30</td>
</tr>
<tr>
<td>92221</td>
<td>RIO HORA INC</td>
<td>60,539.44</td>
<td>14,370.03</td>
<td>7,784.03</td>
<td>3</td>
<td>4.21</td>
</tr>
<tr>
<td>92508</td>
<td>FOSSIL INC.</td>
<td>37,755.37</td>
<td>9,620.16</td>
<td>9,620.16</td>
<td>2</td>
<td>3.92</td>
</tr>
<tr>
<td>92411</td>
<td>SABA STORES LTD.</td>
<td>86,441.66</td>
<td>29,999.31</td>
<td>16,492.66</td>
<td>3</td>
<td>2.80</td>
</tr>
<tr>
<td>50630</td>
<td>MULTILINGUAL TEK INC.</td>
<td>55,728.78</td>
<td>30,457.59</td>
<td>30,457.59</td>
<td>2</td>
<td>2.81</td>
</tr>
<tr>
<td>50435</td>
<td>JOHN PETERSON</td>
<td>96,166.49</td>
<td>34,259.55</td>
<td>18,238.37</td>
<td>7</td>
<td>2.81</td>
</tr>
<tr>
<td>92211</td>
<td>RICARDO BAL</td>
<td>11,001.66</td>
<td>4,091.17</td>
<td>2,068.26</td>
<td>3</td>
<td>2.69</td>
</tr>
<tr>
<td>92532</td>
<td>WERNER KG</td>
<td>65,992.51</td>
<td>35,097.03</td>
<td>15,641.26</td>
<td>3</td>
<td>2.15</td>
</tr>
<tr>
<td>21600</td>
<td>DIALOG CORPORATION</td>
<td>251.58</td>
<td>144.24</td>
<td>144.24</td>
<td>2</td>
<td>1.74</td>
</tr>
<tr>
<td>21335</td>
<td>D A TOWNLEY &amp; ASSOCIATES LTD</td>
<td>29,033.80</td>
<td>29,033.80</td>
<td>29,033.80</td>
<td>2</td>
<td>1.68</td>
</tr>
<tr>
<td>20928</td>
<td>FABRICA TOSONI</td>
<td>21,632.22</td>
<td>13,041.42</td>
<td>13,041.42</td>
<td>2</td>
<td>1.66</td>
</tr>
<tr>
<td>99999</td>
<td>PENNY CILLIN</td>
<td>97,376.40</td>
<td>86,117.39</td>
<td>19,371.30</td>
<td>81</td>
<td>1.13</td>
</tr>
<tr>
<td>92311</td>
<td>ROYAL TRUST CORPORATION</td>
<td>2,072.54</td>
<td>1,916.80</td>
<td>1,916.80</td>
<td>2</td>
<td>1.08</td>
</tr>
</tbody>
</table>
The Same-Same-Same (SSS) Test
This test identifies records that contain fields of information that are exact duplicates of other records.

Application of this test will assist in detecting duplicate expenses claimed, occurrences of the same payment to vendors made in error, multiple warranty claims, or duplicated service fees paid by private or government health plans.

Results of Applying SSS to the Same Supplier Number, Authorizing Official, and Payment Date

The Same-Same-Different (SSD) Test
This test is used to identify records with near duplicates for fields selected by the users.

In Chapter 12 of Dr. Mark Nigrini’s Forensic Analytics, he states, “The same-same-different test is a powerful test for errors and fraud. This test should be considered for every forensic analytics project.” His experience has shown that, “This test always detects errors in accounts payable data . . . the longer the time period, the higher the chances of SSD detecting errors.”
Results of Applying SSD to the Same Supplier Number and Invoice Date with Different Amounts

**GEL-1 and GEL-2**

The purpose of the GEL tests (GEL-1 and GEL-2) is to detect the relationship or link within the data file as potential indicators of fraud.

The GEL tests establishes the link between two selected fields with the first field being the key field and the second field being the element factor over the entire data set.

For example, GEL tests can be used to detect bribery or improper relationships. Since most entities do not have...
access to the records of the payer company, their own data can be analysed to detect improper relationship patterns.

In the GEL-1 example above, the user decided to further analyse all GEL factors of .6000 and greater. Sales rep 105 falls into this criterion with a GEL factor of .7500 and is used in the GEL-2 test.
The resulting GEL-2 file shows that sales rep 105 dealt with client number 30608 120 times (value of $824,092.78) out of 131 while the client purchased from all other sales reps only 11 times. In addition, this sales rep sold to client number 30501 40 times while the same client purchased from all other sales reps only 14 times.

**Conclusion**

- Data analytics, when used correctly, is a powerful tool for identifying suspect accounts or amounts for further analysis.
- Data analytics is a tool to complement additional tests and tools.
- Users have to gain expertise in interpreting results.

**Interesting Non-Vendor Links**

- [http://Ideascripting.com/](http://Ideascripting.com/)
- [www.caats.ca/](http://www.caats.ca/)
- [http://ezrstats.com/index.htm](http://ezrstats.com/index.htm)
- [http://AuditNet.org](http://AuditNet.org)
- [http://AuditSoftwareVideos.com](http://AuditSoftwareVideos.com)

Sunder Gee
October 2012
gee@ideascripting.com