Fraud Pattern Detection using the XL Audit Commander

Version 1.54 (March 2008)

XL Audit Commander
Detection of Fraud Patterns

Detection of certain types of fraud patterns can be achieved using the XL Audit The primary purpose of the tool is to support forensic accounting and fraud investigations, although it may find other uses in research and as a general purpose accounting support tool. This is *free software*, there is no license cost, and it may be used without restriction for any purpose, including commercial and educational.

The tool is installed as an Excel add-in, and as such requires Excel 2000 or later. The tool works only on Windows operating systems.

The techniques described here for the detection of fraud patterns require that the data to be analyzed be in "tab separated value" format and also be sorted. If the data to be analyzed is not in tab separated value format, there is an explanation of how to convert data in the white paper located at http://www.ezrstats.com/doc/Process_To_Convert_Data_To_TSV.pdf.

This paper is divided into two sections –1) Procedure for the detection of fraud patterns and 2) Example output reports produced using the data provided with the software installation.

XL Audit Commander

Detection of Fraud Patterns

XL Audit Commander
Detection of Fraud Patterns

Fraud Pattern detection

Overview

The purpose of the fraud pattern detection routines is to divide the population into components and then quantitatively assess to what extent each component differs from the population as a whole, using one of seven types of attributes or tests:

1. Existence of round numbers
2. Benford's Law
3. Days of the week
4. Occurrence of U.S. Federal holidays
5. Coefficient of variation
6. Data Stratification
7. Trend lines ("best fit")

By automating the process of identification of population components which differ significantly from the remainder of the population, the process is speeded up and the likelihood of identifying error/fraud transactions is increased, making the auditor's work more efficient and effective.

The approach taken is as follows for each type of test performed:

> Using the test specified, determine the attributes of the population as a whole
> Using the sort sequence specified, separate the population into unique groups using the sort parameters, i.e. each group extracted will consist of elements which match using the sort parameters specified.

For each group, perform the same test as that of the population as a whole. Then compare the attributes of the selected group with that of the population as a whole. Differences can be quantified using one of three measures (which are explained in more detail below):

1. The d-statistic for the counts
2. The d-statistic for the amounts
3. The chi squared value for the counts

(Note that for some tests, there may be no "amount" which can be tested, in which case only the two "count" tests may be performed.)

All of the tests are done using data organized as a histogram (described more fully below)

XL Audit Commander

Detection of Fraud Patterns

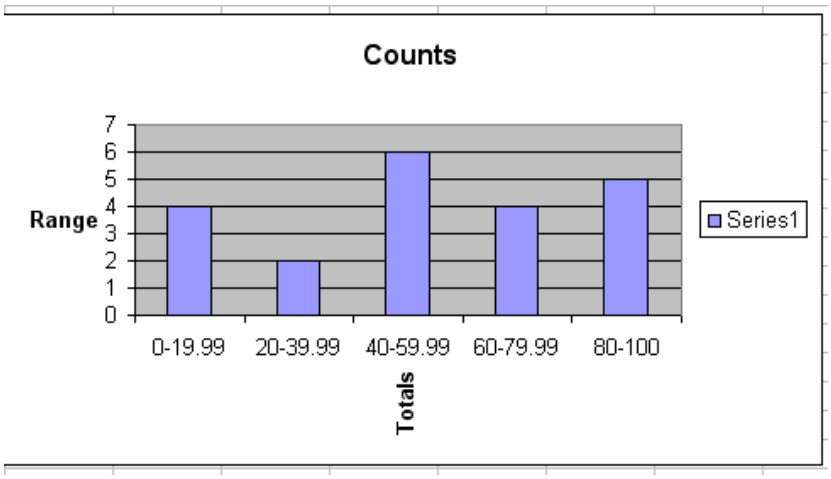
Histograms

A histogram is a means to classify numeric data. Each value or observation in the population is classified by its value and placed into a "bin". A "bin" is either a single value or a range between two values. If the observed value fits within this "bin", then the count is updated by one and a total value for the bin (if applicable) is updated for the observation's amount or value.

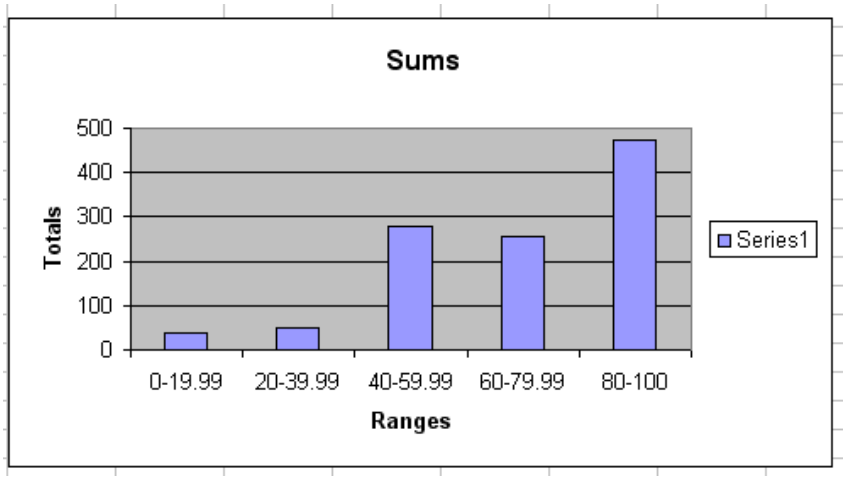
A very simple example is provided below to illustrate the process. The population consists of 21 values, which were randomly generated values between 0 and 100. These values are then classified as falling into one of five bins (0-20, 20-40, 40-60, 60-80 and 80-100). The population classification then consists of the following counts:

B	C	D	E	F	G
Value	**0-19.99**	**20-39.99**	**40-59.99**	**60-79.99**	**80-100**
46.00552			1		
61.2835				1	
86.42871					1
37.49362			1		
45.94952			1		
91.34961					1
65.63498				1	
22.25758		1			
12.84337	1				
55.86754			1		
96.26757					1
64.98203				1	
9.681697	1				
5.074137	1				
54.23275			1		
40.97227			1		
9.836638	1				
26.75297		1			
62.01787				1	
98.60277					1
98.95376					1
1092.488	4	2	6	4	5

Of the 21 values, we can see that four (4) fall in the first bucket, two (2) in the second bucket, six (6) in the third bucket, etc. This information can then be charted as follows:



It is also possible to chart the total amounts in a similar fashion, which results in the following:

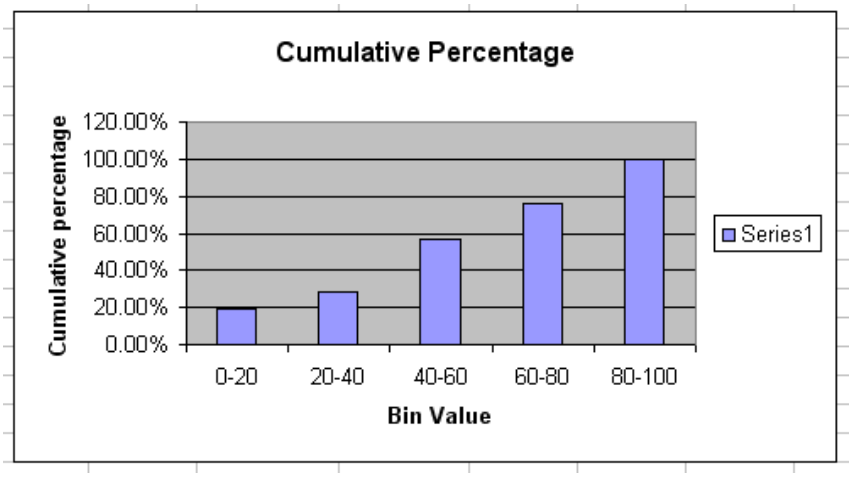


Instead of charting the values in the traditional format, they can also be charted according to their cumulative percentage value. By plotting in this manner it becomes easier to compare two distributions. Let's begin by plotting the counts using their cumulative percentage values.

Range	Count	Percentage	Cumulative Pct
0-20	4	19.05%	19.05%
20-40	2	9.52%	28.57%
40-60	6	28.57%	57.14%
60-80	4	19.05%	76.19%
80-100	5	23.81%	100.00%
Total	21	100.00%	

We can now chart these values as a traditional histogram:

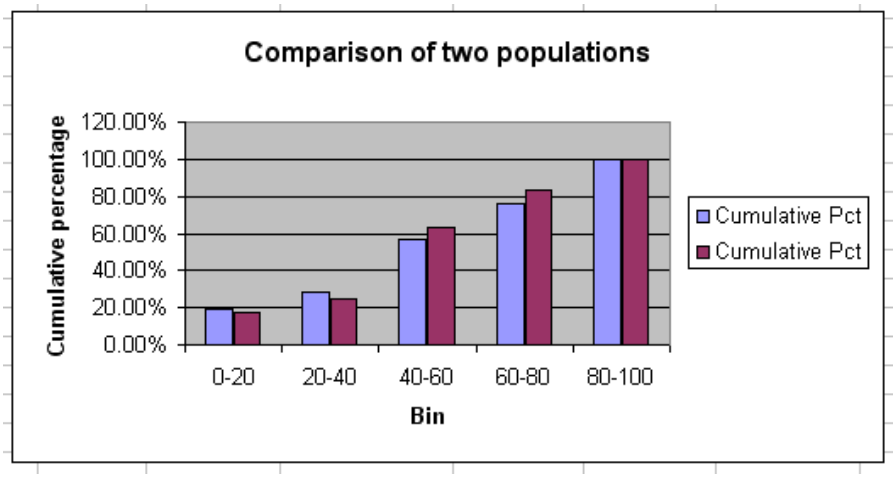
XL Audit Commander
Detection of Fraud Patterns



Let's assume that we have two populations which we wish to compare. The other population has more values, but all are within the same range (0 – 100). Let's plot these values side by side with the values we already have.

Range	Count	Percentage	Cumulative Pct	Count	Percentage	Cumulative Pct
0-20	4	19.05%	19.05%	7	17.07%	17.07%
20-40	2	9.52%	28.57%	3	7.32%	24.39%
40-60	6	28.57%	57.14%	16	39.02%	63.41%
60-80	4	19.05%	76.19%	8	19.51%	82.93%
80-100	5	23.81%	100.00%	7	17.07%	100.00%
Total	21	100.00%		41	100.00%	

Charting the two cumulative percentages yields the following:



XL Audit Commander

Detection of Fraud Patterns

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test (K-S test) is used to determine if two groups of transactions differ significantly. One of the main advantages of the test is that it makes no assumptions about how the data is distributed. The test was developed by two Russian mathematicians, Andrey Nikolaevich Kolmogorov, and V. I. Smirnov.

The test is designed for use in a typical experiment where one set of data has been gathered (e.g. "control group") and another set of data has been gathered for testing (a "treatment group"). The purpose of the test is to determine if there is any significant difference between the two groups. Rarely would the two groups be identical, thus there should be a method wherein the difference between the two groups can be quantified, and then a value assigned to determine statistically the likelihood that the groups do in fact have the same distribution.

How the test works

There is a very good explanation of the K-S statistic at the St. Benedict St. John's University Physics Department web site at http://www.physics.csbsju.edu/stats/KS-test.html. What follows is a synopsis taken from this site.

The basis for the test is the "cumulative fraction function" which is also known as the "empirical distribution function". Expressed in audit terms, this simply the result of performing the following analysis on a group of amounts:

1. Foot the amounts to determine their total value.
2. For each amount, express its value as a fraction of the total value.
3. For each amount, add its amount expressed as a fraction to the cumulative value of all preceding amounts.
4. Plot this distribution as a line chart which will result in a line starting at zero and going up to 1.0 (100%).

Perform this identical procedure for the second group of amounts to be tested, which results in another cumulative line which can be plotted.

Now place both of these lines on the same chart and identify the point where the distance between each of the cumulative values is the greatest. This distance is termed the "K-S d-statistic" and can be used to determine how close the two groups of transactions are.

The "d-statistic"

XL Audit Commander

Detection of Fraud Patterns

One of the measures of the difference between the cumulative percentage values for two populations is the "d-statistic'. This type of comparison, when done of certain types of distributions is referred to as the "Kolmogorov-Smirnov" (KS) d-statistic.

The d-statistic is the greatest vertical difference between two distributions. In the current case, the d-statistic is 6.74% as computed below (difference is for the fourth bin, 76.19% vs. 82.93%, a difference of 6.74%):

Range	Cumulative	Cumulative Pct	d-statistic
0-20	19.05%	17.07%	1.97%
20-40	28.57%	24.39%	4.18%
40-60	57.14%	63.41%	6.27%
60-80	76.19%	82.93%	6.74%
80-100	100.00%	100.00%	0.00%
Total			

An example illustration using Excel

The process will be illustrated by taking hypothetical comparative income statements at two points in time and quantifying the extent to which they differ. Note that all sub-totals and totals have been removed in order that only the income statement amounts themselves are analyzed.

For this purpose, an Excel sheet will be prepared which reflects the following eight columns:

1. Account name (used only for identification of amounts)
2. Balance or amount at period one
3. Balance or amount at period two
4. Period one amount expressed as a decimal fraction of the total for period 1
5. Period two amount expressed as a decimal fraction of the total for period 2
6. The cumulative amount for period one as a decimal fraction
7. The cumulative amount for period two as a decimal fraction
8. The difference between the two cumulative amounts expressed as an absolute value (i.e. sign is disregarded).

The d-statistic will then be the largest difference between the periods.

Comparative income Statements

Below is the application of the K-S test to hypothetical comparative income statements for year end 1999 and 2000. The largest difference (i.e. d-statistic) is for taxes paid, 17/3% shown as follows:

XL Audit Commander

Detection of Fraud Patterns

				Percentages		Cumulative		
For the year ended 31 December	1999	2000	2001	1999	2000	1999	2000	Difference
Net cash inflow from operating activities	10,290	13,246	22,409	66.4%	81.0%	66.4%	81.0%	14.6%
Dividends from joint ventures	949	645	104	6.1%	3.9%	72.5%	85.0%	12.5%
Dividends from associated undertakings	219	394	528	1.4%	2.4%	73.9%	87.4%	13.4%
Interest received	179	444	256	1.2%	2.7%	75.1%	90.1%	15.0%
Interest paid	-1,065	-1,354	-1,282	-6.9%	-8.3%	68.2%	81.8%	13.6%
Dividends received	34	42	132	0.2%	0.3%	68.4%	82.1%	13.6%
Dividends paid to minority shareholders	-151	-24	-54	-1.0%	-0.1%	67.5%	81.9%	14.5%
Net cash outflow from servicing of finance a	-1,003	-892	-948	-6.5%	-5.5%	61.0%	76.5%	15.5%
UK corporation tax	-559	-869	-1,058	-3.6%	-5.3%	57.4%	71.1%	13.8%
Overseas tax	-701	-5,329	-3,602	-4.5%	-32.6%	52.9%	38.6%	14.3%
Tax paid	-126	-619	-4,660	-0.8%	-3.8%	52.0%	34.8%	17.3%
Sale of tangible and intangible fixed assets	6,371	8,837	-12,142	41.1%	54.0%	93.1%	88.8%	4.3%
Payments for fixed assets – investments	-86	-1,200	-39	-0.6%	-7.3%	92.6%	81.5%	11.1%
Proceeds from the sale of fixed assets	1,149	3,029	2,365	7.4%	18.5%	100.0%	100.0%	0.0%
Totals	15,500	16,350		100.0%	100.0%		**d-statistc**	17.3%

The large difference between the two would indicate that there is a significant difference between the income statements for the two periods.

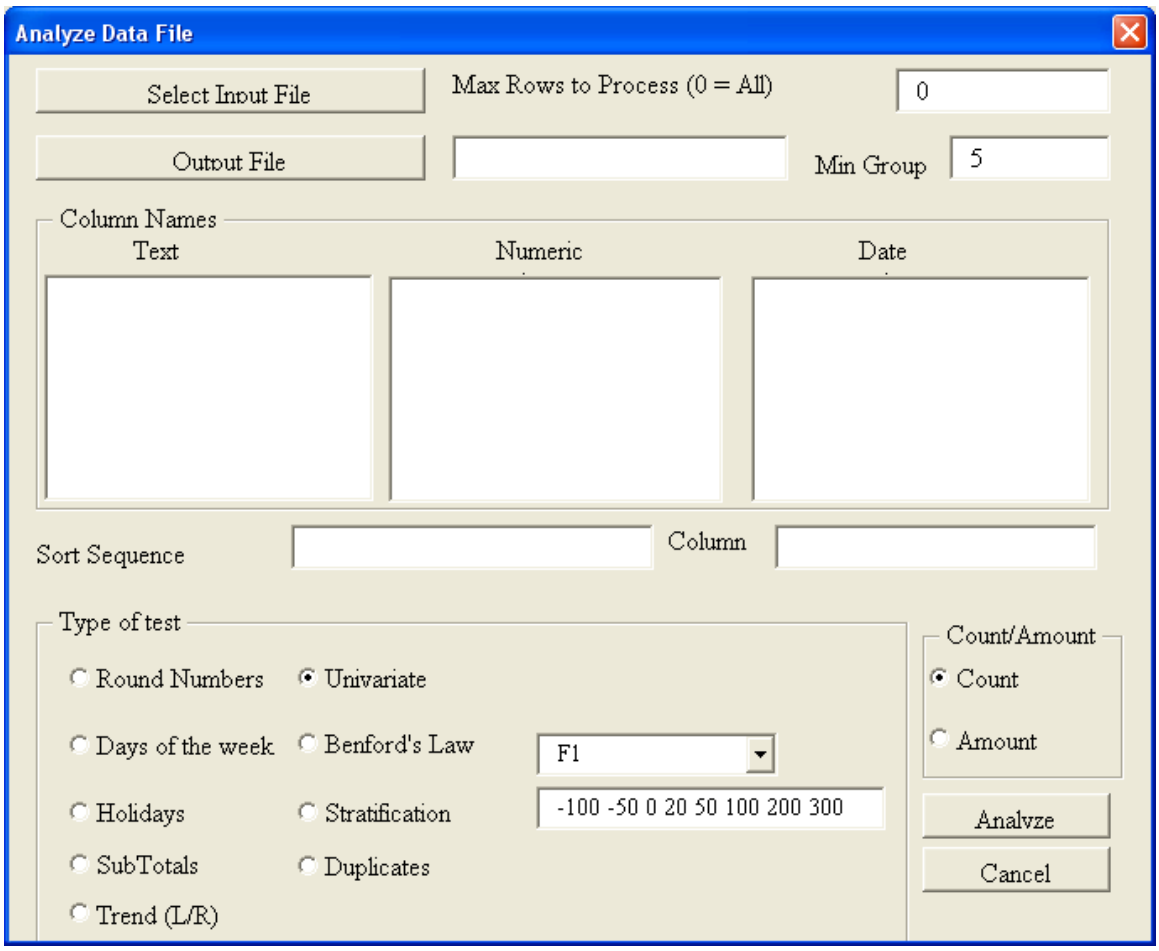
Chi Square

The Chi Square statistic is a useful measure of association between columns of data, each containing counts, such as attributes. There is extensive information on both how to calculate as well as how to use the statistic.

Form

XL Audit Commander

Detection of Fraud Patterns



Column Meanings

Form Element	Description
Select Input File	Click this button in order to select the input file to be analyzed
Max Rows to process	To specify that all rows be analyzed, enter "0", otherwise specify the limit, i.e. the first "N" records, e.g. 1000
Output file	The name of the file to contain the results of the analysis
Min Group	For some tests, it may be best to exclude results for small groups, as the results may not be meaningful. This text box allows the specification of the minimum number of observations before the specified test if performed.
Column names – text	This information is displayed (not entered) and shows the names of all text columns in the input file to be analyzed
Column names – numeric	This information is displayed (not entered) and shows the names of all numeric columns in the input file to be

	analyzed
Column names – date	This information is displayed (not entered) and shows the names of all date columns in the input file to be analyzed
Sort sequence	This is the sort sequence for the input data (as provided during a prior sort)
Column	This is the column to be used for analysis, generally either a numeric column or a date column depending upon the type of test being performed
Round numbers option	Select this option if the desired test to be performed is the analysis of round number usage
Univariate option	Select this option if the desired test to be performed is the determination of coefficient of variation
Days of week option	Select this option if the desired test to be performed is the analysis of counts of days of the week
Benford's Law option	Select this option if the desired test to be performed is conformity with benford's Law
Holidays option	Select this option if the desired test to be performed is the analysis of counts of federal holidays
Stratification option	Select this option if the desired test to be performed is the analysis by data stratification
Subtotals option	Select this option if the subtotals are needed for extract
Duplicates option	Select this option if the desired test to be performed is the analysis of round number usage
Trend (L/R)	Select this option for a trend using "best fit" linear regression
Count option	Select this option if the analysis is to be based upon counts of observations
Amount option	Select this option if the analysis is to be based upon amounts (sums) of observations
Analyze button	When all the parameters have been entered, click "analyze" to run the tests
Cancel button	If no further analysis is needed, click cancel to close the form

XL Audit Commander
Detection of Fraud Patterns

System Limitations

1. The maximum number of columns which may be specified for merging is 200.
2. Each numeric value in the specified merge sequence must be between -920 trillion and positive 920 trillion.
3. For date values specified as part of the merge sequence, only dates between AD 100 and AD 9999 are handled.

Sort Sequence

Sort sequence is the same as the sort parameters used when the input files were sorted.

Processing Output

For each test performed, there are two outputs:

1. A status bar message indicating the two sub groups which differ the most from the population as a whole, as well as their associated d-statistic.
2. An output file which contains detail information for each sub-group, including d-statistic and chi square value, etc.

Note that the term sub-group means a group of observations which has the same sort key values for each of the specified sort parameters. For example, if the input file was sorted by Vendor Number, then each subgroup would consist of all the transactions for any particular vendor number.

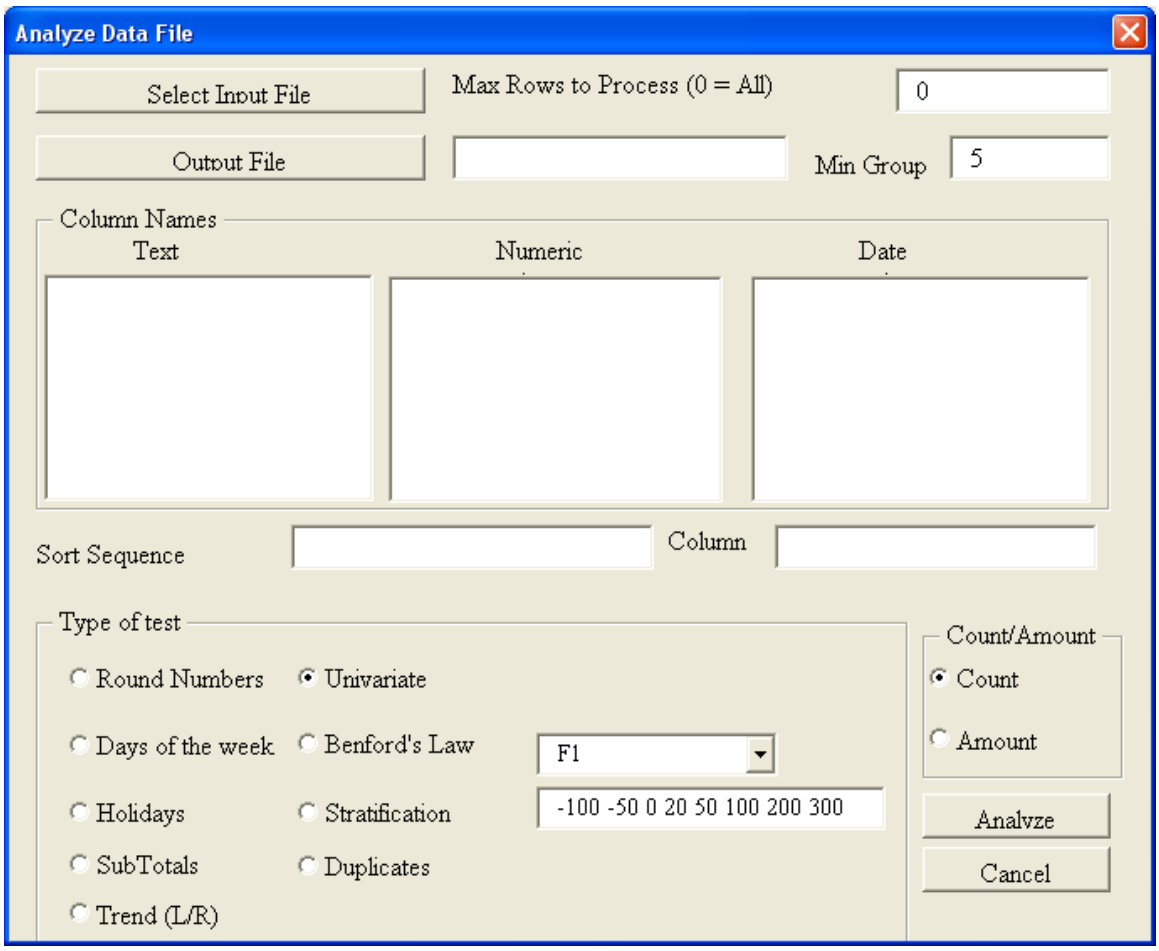
Examples of processing (including output)

The first example illustrates the use of the test for round numbers. The input file has been sorted by vendor number. Tests for round number usage are to be done on the column named “Paid Amount”.

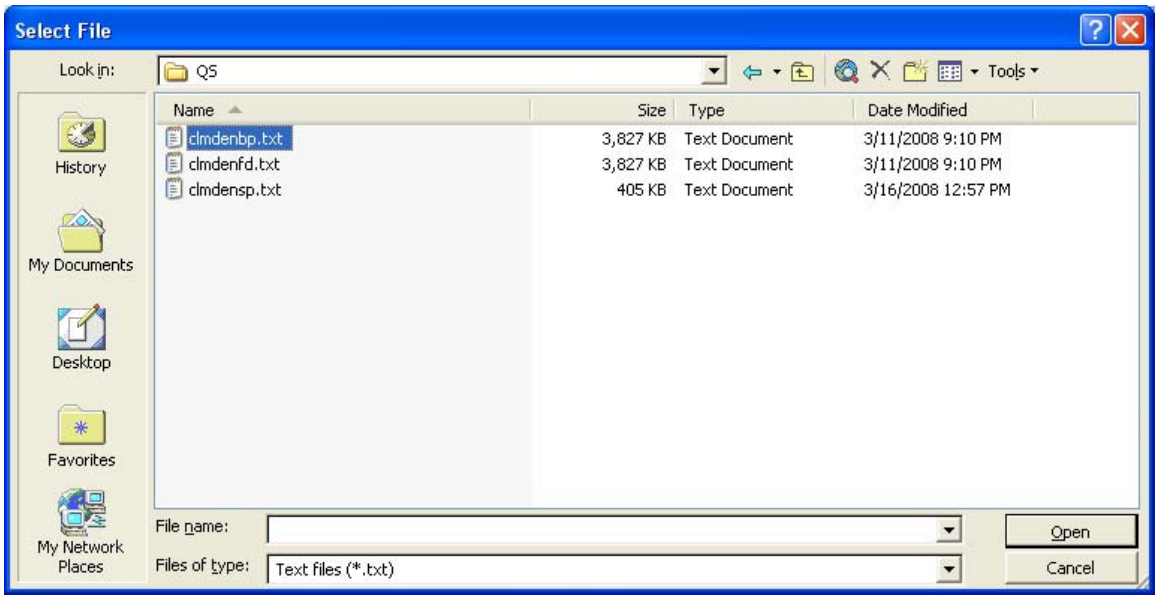
The process starts with a blank form, where the input file may be specified.

XL Audit Commander

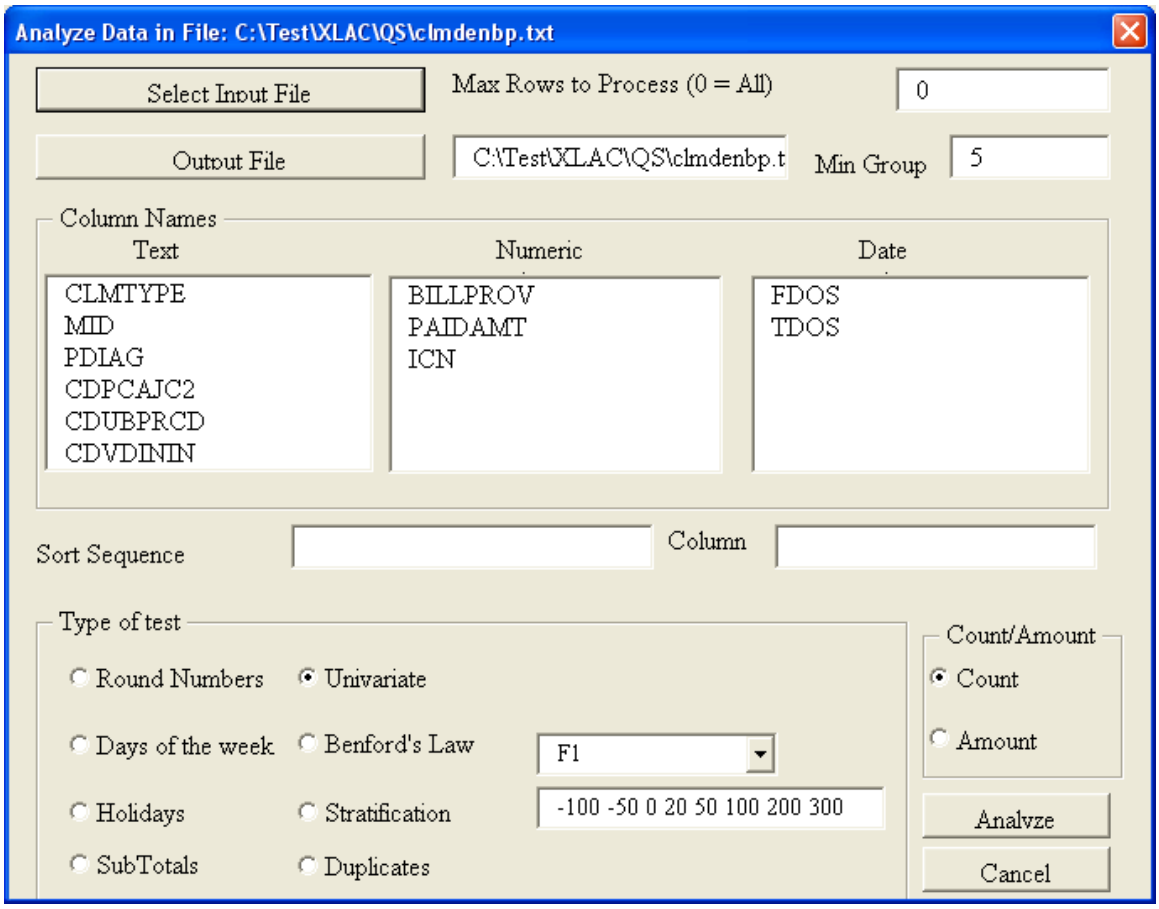
Detection of Fraud Patterns



The data file to be selected is called “clmdenbp.txt” and consists of a number of medical claims details which have been sorted in provider order. This file is selected by clicking the “Select Input File” button and browsing to the proper file directory.



Once the file name is clicked, the caption on the input form changes to reflect it, and a default name for the output file is provided (using the same file name, but with an extension of "tmp"). Also, all the available column names from the input file are classified and displayed (as shown below).



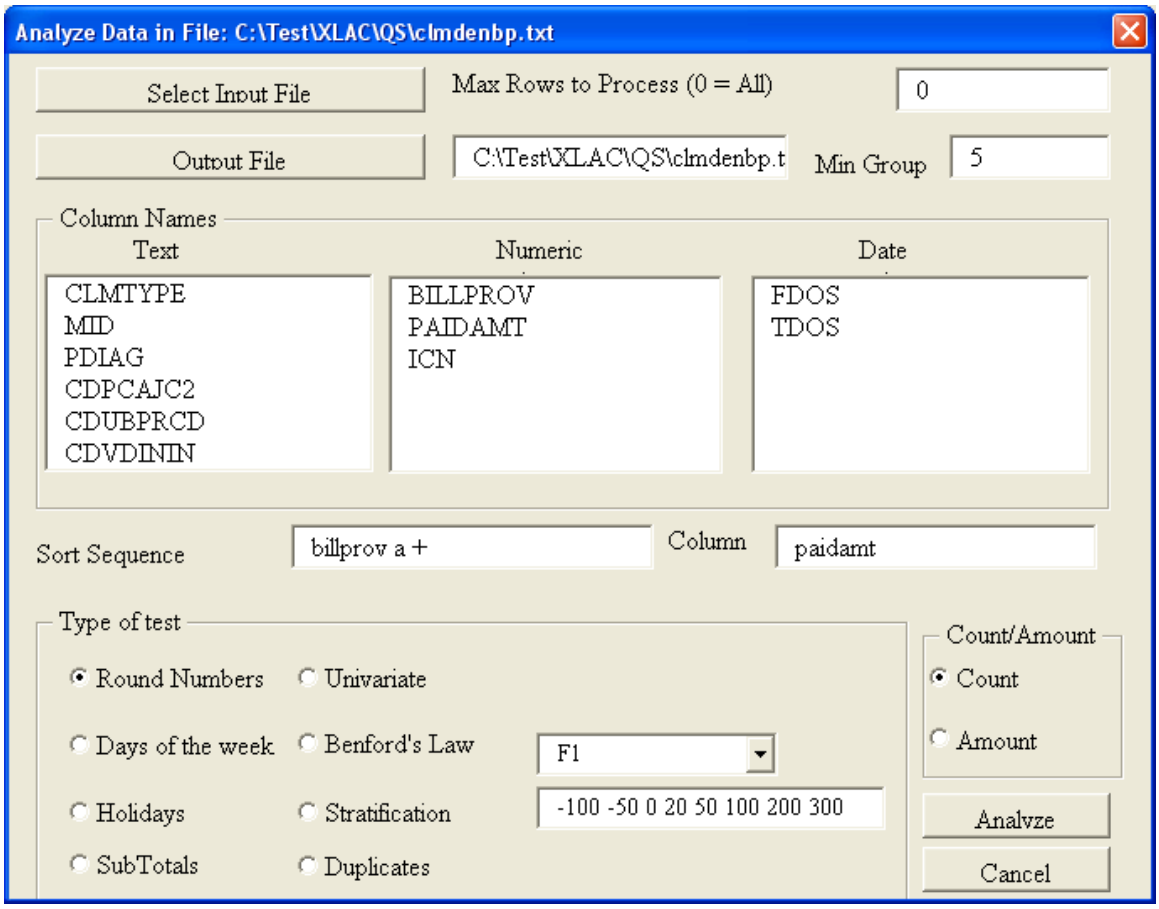
The number of rows to be processed will be left at "0", indicating that all rows are to be processed. The default value of a minimum of 5 observations per group will be left as is.

The next step is to specify the sort sequence. This file was sorted in provider number sequence (BILLPROV). Even though the billing provider number is numeric, we will treat it as text for this purpose. The sort sequence was ascending. Thus, the Sort sequence will be specified as "billprov a +". (Note that none of the sort sequence parameters are case sensitive). The test to be performed will be the round numbers test and the column to be analyzed will be the one named "paidamt".

The updated form now appears as follows:

XL Audit Commander

Detection of Fraud Patterns

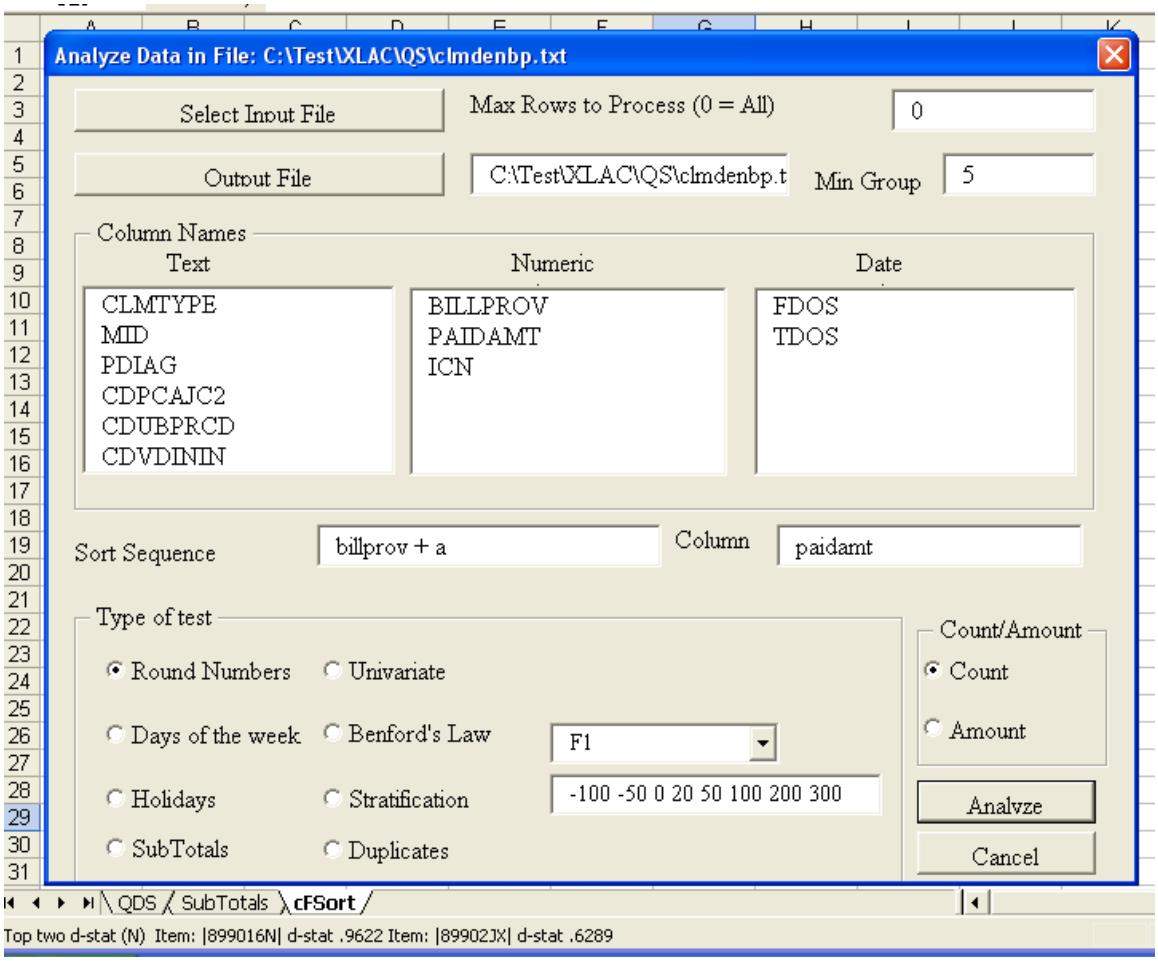


The test can now be run by clicking the "Analyze" button.

The status bar indicates the two groups which had the largest difference (899016N and 89902JX), these are shown below

XL Audit Commander

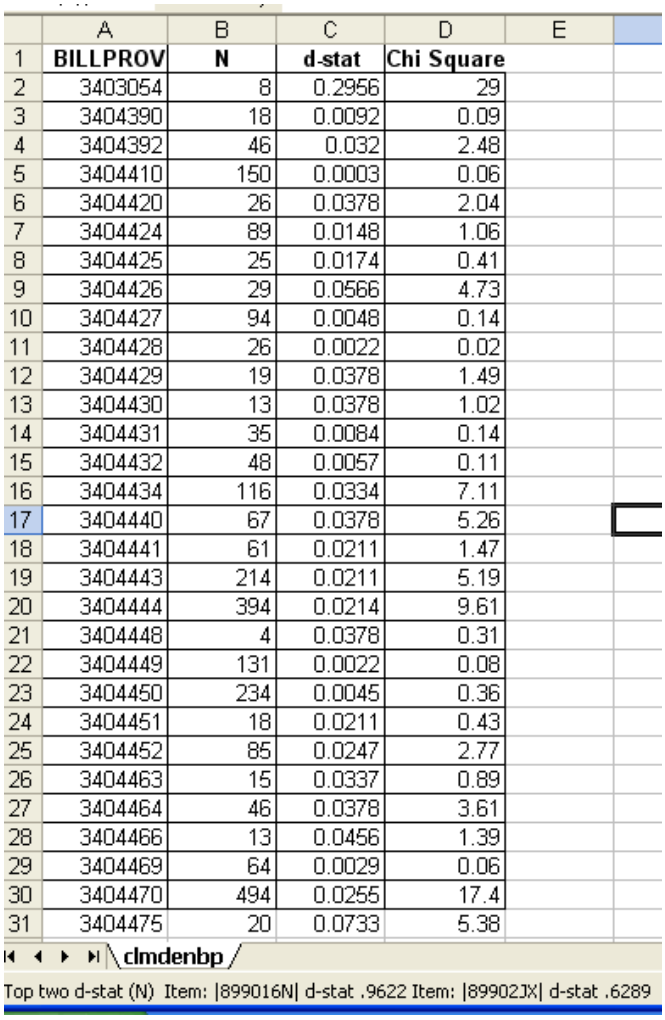
Detection of Fraud Patterns



The first part of the output file is shown below (not sorted):

XL Audit Commander

Detection of Fraud Patterns



	A	B	C	D	E
1	BILLPROV	N	d-stat	Chi Square	
2	3403054	8	0.2956	29	
3	3404390	18	0.0092	0.09	
4	3404392	46	0.032	2.48	
5	3404410	150	0.0003	0.06	
6	3404420	26	0.0378	2.04	
7	3404424	89	0.0148	1.06	
8	3404425	25	0.0174	0.41	
9	3404426	29	0.0566	4.73	
10	3404427	94	0.0048	0.14	
11	3404428	26	0.0022	0.02	
12	3404429	19	0.0378	1.49	
13	3404430	13	0.0378	1.02	
14	3404431	35	0.0084	0.14	
15	3404432	48	0.0057	0.11	
16	3404434	116	0.0334	7.11	
17	3404440	67	0.0378	5.26	
18	3404441	61	0.0211	1.47	
19	3404443	214	0.0211	5.19	
20	3404444	394	0.0214	9.61	
21	3404448	4	0.0378	0.31	
22	3404449	131	0.0022	0.08	
23	3404450	234	0.0045	0.36	
24	3404451	18	0.0211	0.43	
25	3404452	85	0.0247	2.77	
26	3404463	15	0.0337	0.89	
27	3404464	46	0.0378	3.61	
28	3404466	13	0.0456	1.39	
29	3404469	64	0.0029	0.06	
30	3404470	494	0.0255	17.4	
31	3404475	20	0.0733	5.38	

clmdenbp

Top two d-stat (N) Item: |899016N| d-stat .9622 Item: |89902JX| d-stat .6289

The report lists, for each sort key, the number of items in the group (N), the d-statistic for the difference between the group and the population overall, as well as the Chi Square value comparing the group and the overall population. This report can be sorted in various orders to identify the key value groups which may be of interest.

XL Audit Commander
Detection of Fraud Patterns

Example Reports

All of the reports were obtained using the test data provided with the software installation, and named "clmdenbp.txt". This data is a set of hypothetical insurance claims submitted by "billing providers – BILLPROV". The data includes the claim amount paid (PAIDAMT), the date of service "FDOS" and other data.

This data has been sorted in billing provider sequence.

Each of the seven types of tests was run against the test data.

Benford's Law

Tests of population components using Benford's Law enables the auditor to focus on those transaction groups which differ the most from expected, as measured using values expected per Benford's Law. This enables the auditor to more quickly identify suspect population components and focus on those which may be the most likely to contain errors. There is wide applicability of Benford's Law, including tax returns, insurance claims, inventory counts, invoice amounts, etc.

BILLPROV	N	Chi Squared	DF	D-Stat	High Digit
3403054	8	0	0	0	0
3404390	18	5.14	8	0.183	1
3404392	46	18.88	8	0.1957	2
3404410	150	20.49	8	0.086	7
3404420	26	14.99	8	0.2493	2
3404424	89	23.88	8	0.1	7
3404425	25	10.54	8	0.095	3
3404426	29	10.74	8	0.2759	3
3404427	94	14.1	8	0.0882	4
3404428	26	10.92	8	0.2892	2
3404429	19	22.45	8	0.3158	1
3404430	13	14.98	8	0.3571	4
3404431	35	14	8	0.2286	1
3404432	48	13.41	8	0.1973	2
3404434	116	32.81	8	0.1421	2
3404440	67	23.49	8	0.1584	1
3404441	61	31.3	8	0.2575	1

Sort in descending sequence by Chi Squared (using Excel sort on column)

XL Audit Commander

Detection of Fraud Patterns

BILLPROV	N	Chi Squared	DF	D-Stat	High Digit
File Totals	65264	5,585.07	8	0.0663	7
8996806	420	496.43	8	0.6156	5
8990187	1585	468.67	8	0.1452	6
89012P6	1424	404.38	8	0.2108	5
8994865	678	298.58	8	0.1667	6
89014NG	982	271.82	8	0.2031	3
7996465	686	251.9	8	0.2915	4
8.90E+11	2177	246.53	8	0.0809	7
89013MP	1703	216.3	8	0.0892	7
899019P	206	197.53	8	0.5194	5
8998934	142	189.06	8	0.6408	5
89013U4	1002	185.48	8	0.1513	3
89015CV	897	183.91	8	0.1717	3
899014C	124	182.63	8	0.724	5
7901933	1070	171.71	8	0.0887	7
89015C2	1026	171.08	8	0.0898	7
8990168	534	163.79	8	0.1891	1

Sort in descending sequence by d-stat (using Excel sort on column)

BILLPROV	N	Chi Squared	DF	D-Stat	High Digit
89902TT	10	17.66	8	0.8182	6
8993443	90	142.17	8	0.7556	5
899014E	30	48.92	8	0.7333	5
899014C	124	182.63	8	0.724	5
8995790	85	125.8	8	0.7202	5
8994771	64	93.25	8	0.7153	5
8992155	14	24.27	8	0.7143	5
8990267	103	137.21	8	0.6796	5
8998797	11	17.35	8	0.6667	5
8997763	17	25.58	8	0.6471	4
8994644	13	15.12	8	0.6429	4
8998934	142	189.06	8	0.6408	5
7999405	25	28.19	8	0.6317	5
8902986	41	49.26	8	0.6287	5
8992243	94	112.82	8	0.6254	5
8995839	21	25.57	8	0.619	5
8996806	420	496.43	8	0.6156	5

Univariate Statistics

The coefficient of variation (CV) is the ratio of the standard deviation divided by the average and presents a comparable statistic which enables the comparison of two groups of amounts. The larger the value of the CV, the larger the relative variation.

XL Audit Commander

Detection of Fraud Patterns

Tests of population components using univariate statistics based on a numeric column value enables the auditor to focus on those transaction groups which differ the most from expected. Generally two approaches can be taken – focus on those groups with a low coefficient of variation, or focus on those groups with a high coefficient of variation. .

BILLPROV	N	CV
3403054	8	5.61%
3404390	18	60.79%
3404392	46	62.05%
3404410	150	90.94%
3404420	26	79.66%
3404424	89	63.24%
3404425	25	66.16%
3404426	29	69.66%
3404427	94	65.77%
3404428	26	44.38%
3404429	19	48.17%
3404430	13	15.57%
3404431	35	52.48%
3404432	48	51.42%
3404434	116	59.08%
3404440	67	65.02%
3404441	61	54.51%
3404443	214	76.82%

Sort by coefficient of variation descending.

BILLPROV	N	CV
8998924	2	N/A
899016N	6	N/A
899018G	6	N/A
7907032	1	488.84%
89015FT	75	366.26%
7998702	1	338.73%
8993722	1	269.72%
6901778	11	250.23%
8994741	1	246.52%
8999020	13	231.38%
8995048	46	227.96%
89902JX	5	223.61%
7907002	53	211.53%
89902AT	10	209.82%
8998139	1	194.93%
8991770	23	184.48%
89902R3	45	183.53%
8995839	21	180.18%

Note that in certain instances, the average of the group being measured may be zero, in which case the coefficient of variation can not be measured. Also, where averages are less than one, the coefficient of variation will exceed 100%.

Round Numbers

Round numbers are indicative of estimates. The application of round number tests has found extensive use in the testing of journal entries, as well as identification of suspect invoices where procurement procedures are being circumvented through the use of split invoices or invoices submitted under the limit which requires competitive bidding.

BILLPROV	N	d-stat	Chi Square
3403054	8	0.2956	29
3404390	18	0.0092	0.09
3404392	46	0.032	2.48
3404410	150	0.0003	0.06
3404420	26	0.0378	2.04
3404424	89	0.0148	1.06
3404425	25	0.0174	0.41
3404426	29	0.0566	4.73
3404427	94	0.0048	0.14
3404428	26	0.0022	0.02
3404429	19	0.0378	1.49
3404430	13	0.0378	1.02
3404431	35	0.0084	0.14
3404432	48	0.0057	0.11
3404434	116	0.0334	7.11
3404440	67	0.0378	5.26
3404441	61	0.0211	1.47

Sorting the results of the round number test by descending values for chi square results in the following:

XL Audit Commander

Detection of Fraud Patterns

BILLPROV	N	d-stat	Chi Square
890219W	131	0.044	1,744.26
799002P	40	0.0733	996.51
8995084	86	0.471	709.67
8997596	7	0.235	464.05
8997826	55	0.2563	153.9
899016N	6	0.9622	153.48
8993165	203	0.1124	122.39
8992534	41	0.2638	120.95
8907029	77	0.1748	106.99
8993910	161	0.1164	103.9
89015FV	252	0.0415	101.03
8993912	163	0.0981	76.13
89902JX	5	0.6289	65.6
8994751	246	0.0728	64.75
89015CV	897	0.0355	62.11
8997337	13	0.3307	57.42
89012P6	1426	0.0257	50.89
8993292	115	0.0952	50.86

This report indicates that for provider number 890219W, the chi squared value for the round number test (1,744.26) is significantly larger (almost double) than the second largest value (996.51) for provider number 799002P.

XL Audit Commander

Detection of Fraud Patterns

Day of Week Analysis

Many transaction groups can be tested based upon transaction counts by day of week, and especially where there should be a clear pattern. For example, some businesses are closed on weekends, others (e.g. retail) may have the bulk of their activity occurring on weekends. Thus this test, since it measures activity by day of the week, can be used to identify suspect transactions in certain cases.

BILLPROV	N	d-stat	Chi Square
3403054	8	0.3137	410.72
3404390	18	0.1012	416.97
3404392	46	0.0698	408.15
3404410	150	0.0546	425.73
3404420	26	0.1248	426.8
3404424	89	0.0613	419.15
3404425	25	0.1493	421.37
3404426	29	0.158	424.12
3404427	94	0.1009	436.75
3404428	26	0.0675	412.9
3404429	19	0.2523	436.94
3404430	13	0.3988	453.5
3404431	35	0.0988	426.61
3404432	48	0.0492	410.34
3404434	116	0.1358	451.59
3404440	67	0.159	478.34
3404441	61	0.1995	445.79

XL Audit Commander
Detection of Fraud Patterns

Holiday Analysis

Certain businesses are open on holidays, and others are not. This test can be used to quantify transactions as to whether they occur on Federal Holidays (U.S. only). For example, retail operations may be open on Federal Holidays, whereas certain government agencies and financial institutions are not.

BILLPROV	N	d-stat	Chi Square
|3403054	8	0.0192	0.16
|3404390	18	0.0192	0.35
|3404392	46	0.0192	0.9
|3404410	150	0.0192	2.94
|3404420	26	0.0192	0.51
|3404424	89	0.035	19.12
|3404425	25	0.0192	0.49
|3404426	29	0.0192	0.57
|3404427	94	0.0192	1.84
|3404428	26	0.0192	0.51
|3404429	19	0.0192	0.37
|3404430	13	0.0192	0.25
|3404431	35	0.0192	0.69
|3404432	48	0.0192	0.94
|3404434	116	0.0231	13.53
|3404440	67	0.0192	1.31

Sorting these values in descending order by d-stat yields the following results:

BILLPROV	N	d-stat	Chi Square
|8992984	4	0.4818	198.42
|8994031	6	0.4813	5,267.98
|8995579	4	0.4104	126.92
|899012V	5	0.3568	110.33
|8991324	12	0.3502	251.15
|8995584	4	0.3151	65.05
|8991956	7	0.2818	87.11
|8993141	10	0.2773	127.54
|8990016	12	0.2318	95.39
|8990324	25	0.2232	338.42
|8996695	23	0.1984	134.99
|89902EJ	8	0.1915	43.68
|799002P	40	0.1485	72.32
|890207P	15	0.1474	1,255.55
|8995955	66	0.1464	193.31
|8997060	11	0.1454	32.76

These statistics could be valid for providers which operate on holidays, such as hospital emergency rooms, etc.

Stratification

Commonly transactions will tend to be stratified in a similar pattern throughout the population of transaction amounts. This test can be used to identify components of populations whose transaction values are "skewed" significantly from that of the population as a whole.

One potential application is the identification of "up-coding" of medical procedures for insurance purposes, i.e. coding a more expensive procedure, when a less expensive procedure was performed. Generally this may be identified because the less expensive procedures consist of a lower portion of the population transactions than the more expensive "up-coded" transactions. There is wide applicability of analysis using transaction stratification in other areas as well.

BILLPROV	N	d-stat	Chi Square
3403054	8	0.4657	12.93
3404390	18	0.1563	4.27
3404392	46	0.0961	1.9
3404410	150	0.0763	14.85
3404420	26	0.1965	6.39
3404424	89	0.1563	26.03
3404425	25	0.1136	3.96
3404426	29	0.3622	17.72
3404427	94	0.2742	44.74
3404428	26	0.2734	9.92
3404429	19	0.2025	11.82
3404430	13	0.4657	21
3404431	35	0.1563	8.26
3404432	48	0.1147	5.96
3404434	116	0.0488	2.73
3404440	67	0.1373	12.05
3404441	61	0.1706	31.07
3404443	214	0.0535	5.65
3404444	394	0.2906	139

Sorted in descending order buy chi squared value:

BILLPROV	N	d-stat	Chi Square
899012W	15	0.9943	3,327.00
6901778	11	0.3525	2,640.40
89015FT	75	0.4303	1,857.60
8995048	46	0.3857	1,736.32
8997026	8	0.4943	1,721.05
899016N	6	0.9943	1,150.41
8993276	16	0.4318	749.25
89902JX	5	0.7943	615.95
89902K4	35	0.3294	531.99
790242G	153	0.1808	530.09
8990078	24	0.3277	501.32
8996806	420	0.4105	496.68
8995047	11	0.4489	433.6
8991983	36	0.299	422.31
899018Y	179	0.2235	307.28
8990187	1585	0.1741	296.8
899018G	6	0.677	290.52
890206H	312	0.0681	290.1
8992027	6	0.4943	289.63

Reviewing the report, it can be seen that the vendors on the third and fourth row have very high chi-squared value, in addition to reasonable size counts ("N"). These vendors may merit a look to see why their stratified amounts differ so much from the others.

Trend

Trend lines by sub-group can often indicate areas which may merit further attention. For example, if most of the population has an increasing trend, then it may be of value to look at any sub-groups which have a decreasing trend (or the trend which has the lowest value). A test of the opposite condition could also have merit.

The system fits a trend line using least squares "best fit" for each subgroup (Y = B = M*X) and reports several items:

N the count of observations in the trend (must be at least the minimum specified when the analysis was run)
M The slope
B The intercept (where the left most value crosses the Y axis)
SE The standard error of the slope, a measure of the variability (or fit) of the trend line

These values, once extracted, may then be sorted in a variety of ways in order to identify those values which are of potential interest.

XL Audit Commander

Detection of Fraud Patterns

In the example below, the vendor invoice amounts have been sorted in ascending order using the sort parameter "billprov + a paidamt + n". This sorted file is then analyzed using trend line, specifying the sort sequence as "billprov + a" and the numeric column to analyze as "paidamt".

This type of analysis is also useful for identifying key-punch errors where a vendor invoice has been entered as an order of magnitude too large.

BILLPROV	N	M	B	SE
3403054	8	0.569524	23.05968	0.09
3404390	18	5.467915	2.970341	0.25
3404392	46	2.047309	-0.275546	0.02
3404410	150	1.060299	-14.48903	0
3404420	26	3.396564	-2.243111	0.09
3404424	89	1.438339	-0.524854	0.01
3404425	25	5.313762	-7.22125	0.15
3404426	29	2.042882	5.199214	0.05
3404427	94	0.744641	-1.63181	0
3404428	26	2.09226	10.01914	0.05
3404429	19	2.717316	13.326	0.11
3404430	13	0.842637	24.31143	0.06
3404431	35	2.549218	6.765899	0.04
3404432	48	1.954447	8.086539	0.02
3404434	116	0.802796	2.022093	0
3404440	67	1.316347	2.050027	0.01
3404441	61	1.247203	8.874032	0.01
3404443	214	0.619185	-6.20206	0

Sorting this report file (using Excel) in descending order by standard error (SE) gives the following:

XL Audit Commander

Detection of Fraud Patterns

BILLPROV	N	M	B	SE
899019A	5	168.567	-251.5425	53.31
8990146	5	158.434	-209.64	50.1
89014RC	6	91.992	-111.7046	21.99
89902PC	7	80.43536	-125.9813	15.2
89014UU	7	60.70036	-56.835	11.47
899014T	7	57.05536	-43.64125	10.78
8996510	8	63.4519	-78.66762	9.79
89902CP	5	30.003	-5.489167	9.49
8999102	7	47.16964	-50.8325	8.91
8996286	7	44.31857	-16.3925	8.38
7902249	5	24.747	-17.20417	7.83
8993709	7	41.25607	-60.3525	7.8
6901778	11	75.35745	-318.9927	7.19
899019W	6	26.65771	-23.51029	6.37
899007X	8	40.50524	-78.54095	6.25
8994612	6	25.16314	-17.99371	6.02
8991846	5	17.789	-1.295833	5.63
8995054	5	17.56	-10.535	5.55

Sorting the report by slope ("M"), using Excel, provides the following information:

BILLPROV	N	M	B	SE
899019A	5	168.567	-251.5425	53.31
8990146	5	158.434	-209.64	50.1
89014RC	6	91.992	-111.7046	21.99
89902PC	7	80.43536	-125.9813	15.2
6901778	11	75.35745	-318.9927	7.19
8996510	8	63.4519	-78.66762	9.79
89014UU	7	60.70036	-56.835	11.47
899014T	7	57.05536	-43.64125	10.78
89902AT	10	47.75327	-126.8473	5.26
8999102	7	47.16964	-50.8325	8.91
8996286	7	44.31857	-16.3925	8.38
8993709	7	41.25607	-60.3525	7.8
899007X	8	40.50524	-78.54095	6.25
8995753	11	36.19409	-86.85917	3.45
8998797	11	36.00773	-72.17	3.43
899019T	11	30.67536	-60.83617	2.92
89902CP	5	30.003	-5.489167	9.49
8998519	9	29.78	-50.821	3.84

As can be seen, most of the providers identified have relatively low counts of observations (here under 12), so it may be more useful to rerun this test and specify a larger "Minimum Group" in order to exclude those providers who have relatively few observations.

XL Audit Commander

Detection of Fraud Patterns

Summary and Recap

The approach described here for the identification of various sub-groups of transactions which may differ significantly from the overall population is based upon objective, quantifiable methods. Two measures have been presented, both of which are well rooted within statistics: 1) the chi squared value and 2) the K-S d-statistic.

The approach is an efficient and effective method of identifying statistically different sub-groups because 1) the test run fairly fast on a desk top computer and 2) a variety of tests are available, each applicable within different situations depending upon the type of transactions being tested.

The purpose of the tests is not to provide any statistical estimate as to the difference between any two groups, but instead to highlight groups which may merit further investigation.

The software and documentation is available for download from http://ezrstats.com/downloads.htm.