DEPARTMENT OF NATIONAL DEFENCE

CANADA

OPERATIONAL RESEARCH DIVISION

DIRECTORATE OF OPERATION RESEARCH (CORPORATE)

CENTRAL OPERATIONAL RESEARCH TEAM
RESEARCH NOTE RN 2004/11

SEPARATION POINT ANALYSIS METHOD (SPAM)

By

D.G. HUNTER

E.J. EMOND

July 2004

OTTAWA, CANADA

OPERATIONAL RESEARCH DIVISION

CATEGORIES OF PUBLICATION

ORD Reports are the most authoritative and most carefully considered publications of the DGOR scientific community. They normally embody the results of major research activities or are significant works of lasting value or provide a comprehensive view on major defence research initiatives. ORD Reports are approved personally by DGOR, and are subject to peer review.

ORD Project Reports record the analysis and results of studies conducted for specific sponsors. This Category is the main vehicle to report completed research to the sponsors and may also describe a significant milestone in ongoing work. They are approved by DGOR and are subject to peer review. They are released initially to sponsors and may, with sponsor approval, be released to other agencies having an interest in the material.

Directorate Research Notes are issued by directorates. They are intended to outline, develop or document proposals, ideas, analysis or models which do not warrant more formal publication. They may record development work done in support of sponsored projects which could be applied elsewhere in the future. As such they help serve as the corporate scientific memory of the directorates.

ORD Journal Reprints provide readily available copies of articles published with DGOR approval, by OR researchers in learned journals, open technical publications, proceedings, etc.

ORD Contractor Reports document research done under contract of DGOR agencies by industrial concerns, universities, consultants, other government departments or agencies, etc. The scientific content is the responsibility of the originator but has been reviewed by the scientific authority for the contract and approved for release by DGOR.

Authors

D. Gregory Hunter, Edward J. Emond

Approved by

D.F. Reding
DOR(Corp)

Abstract

This research note documents an extension of the iterative post-hoc analysis method described by Emond and Massel (2003) to positive integer-value datasets. The original assumption was that the data were drawn from one or more normal distributions. An additional data model has now been added in which the data are assumed to be derived from one or more binomial distributions. Using an appropriate test statistic and likelihood function, this method finds the most likely separation points between datasets that have been shown to have binomial parameters that are statistically significantly different. The method is of particular interest within operational research due to three important properties: it always returns an answer, this answer is never intransitive, and its error accumulation is multiplicative rather than additive. Both the binomial data model and the normal data model of Emond and Massel have been implemented in a MATLAB software application known as SPAM (Separation Point Analysis Method).

Résumé

La présente note de recherche poursuit la description de la méthode d'analyse itérative post-hoc donnée par Emond et Massel (2003) et se penche sur les ensembles de données positives à valeur entière. On présumait initialement que les données avaient été tirées d'une ou de plusieurs courbes de distribution normale. On a donc ajouté un modèle de données dont on pense qu'elles sont dérivées d'une ou de plusieurs courbes de distribution binomiale. Combinant une variable à tester appropriée ainsi qu'une fonction de probabilité, cette méthode peut déterminer plus précisément la limite entre les ensembles de données dont on a démontré qu'ils présentaient des paramètres binomiaux statistiquement très différents. Trois importantes propriétés de cette méthode la rendent particulièrement utile à la recherche opérationnelle : elle accumule les erreurs par multiplication et non par addition, elle fournit toujours une réponse et cette dernière n'est jamais un verbe intransitif. Les modèles de données normales et binomiales d'Emond et Massel ont été intégrés à une application logicielle MATLAB connue sous le nom de « Méthode d'analyse du point de séparation » ou « MAPS ».

Table of contents

List of figures

List of tables

1.   INTRODUCTION

An important quantitative problem in operational research is the comparison of differences between group means in the presence of sampling error or other random effects. Such an analysis is composed of two separate parts. First it must be determined that the observed differences between the group averages are unlikely to be due to chance. This is done by applying formal statistical hypothesis testing. The null hypothesis is that the groups have a common mean and that the observed differences are due to chance variation. The alternate hypothesis is the logical negation of this, namely that there are at least two subgroups with different means. If the null hypothesis is accepted, the analyst concludes that there is insufficient evidence to reject the possibility that the observed differences between the group means are due to chance. Other than calculating the overall average, the analysis is complete.

In the case that the null hypothesis is rejected, the analyst is now faced with determining which subgroups have different means. Just such a problem occurs frequently in studies comparing the effectiveness or value of two or more different things. This type of analysis is called post hoc and has received much attention in the statistical literature[1]. When there are more than two groups involved in the analysis, the more general statistical terminology is the problem of multiple comparisons.

Most of the commonly used pairwise-comparison procedures, including the Tukey HSD (honest significant difference) test, are perplexing for the practitioner because they produce intransitive results in many cases. As a simple example of intransitivity consider a case with three groups labelled A, B and C with population means μ_A , μ_B , and μ_C respectively. We assume for convenience in the discussion that the sample means are ordered and that the null hypothesis of common means has been rejected. Because the test that indicates that the null hypothesis should be rejected is independent of the post hoc analysis, a common result of the post hoc analysis of this case when using pairwise comparisons is that the equalities $\mu_A = \mu_B$ and $\mu_B = \mu_C$ may be indicated while simultaneously the equality $\mu_A = \mu_C$ is rejected. Although this phenomenon is understandable for a theoretical / statistical point of view, it leaves the analyst in the uncomfortable position of reporting an intransitive result.

Other statistical methods that avoid pairwise comparisons may also be difficult for the operational research analyst because of another problem. Given the conservative nature of for example Scheffé's Method [2], it may happen that despite rejecting the null hypothesis of equal means, the post hoc procedure fails to indicate any differences. This again leaves the analyst in an uncomfortable position of stating that there are differences but being unable to determine what they are.

---

[1] For example, Dayton [1] states, "Kirk (1995) described 22 techniques in the latest edition of his widely used textbook. Often, techniques of this type are employed in order to study differences among group means after rejecting the global null hypothesis in, say, a one-way analysis of variance. In particular, pairwise-comparison (PC) procedures are frequently used in this context and Kirk cites 12 such procedures …"

The methodology used in this report has been developed with the goal of providing the operational research analyst with a simple method for post hoc analysis that is statistically sound and above all practical. This means that it must be easy to understand and apply and must give results that avoid the problems of intransitivity and insensitivity noted above.

2. METHODOLOGY

The classical Student's t-test is appropriate for comparing the means of two datasets under mild normality assumptions. Similarly for discrete data there exist statistical tests for comparing the means of two datasets. A problem frequently encountered in ORD studies is the task of separating multiple datasets according to statistically significant differences between the means. Because of the accumulation of Type I errors it is not appropriate to use pair-wise Student's t-tests when making multiple comparisons. The Bonferroni method [3] corrects for this but can result in a test that is so insensitive to differences between datasets that its utility is limited. An analysis of variance (ANOVA) will determine, to the desired confidence level, whether or not the means of the set or some subset of all the datasets are equal, but does not indicate which datasets should be grouped together.

A recent paper by Paul Massel and Edward Emond [4] approaches this problem by iteratively using the analysis of variance to determine if some set of datasets have statistically different means, and then evaluating a likelihood function for each possible way of breaking the set of datasets in two. This method is described in detail in Reference [4].

2.1 The Binomial Data Model

The binomial data model is appropriate for datasets containing positive integer values. Typical sources for data of this type are wargames or trials, where the number of kills or detections achieved must be compared for two or more different cases or situations. Two pieces of data are required for each data point: the number of observed "events" (kills, detections, or whatever measure of performance is being compared) and the maximum possible number of events ("trials"). For example, one pair of values might be an observation of 6 kills out of a maximum possible of 10. Considered as a ratio, these two values form an observed proportion of 0.6, i.e., an observed probability of success of 0.6. The binomial version of the maximum likelihood method of Reference [4] compares the number of observations and trials of a number of datasets, determines whether there is a significant difference between the proportions of the datasets at the specified confidence level and locates the most likely groupings of datasets if differences in proportions are found.

Assume that the data to be analyzed consists of K datasets, where K is a positive integer greater than 1. We will index the K datasets with the letter k so that k is an integer index ranging from 1 to K.

Within dataset k, we have N_k pairs of integer values. We will index the N_k pairs in each dataset with the letter i so that i is an integer index ranging from 1 to N_k. The pair of values will be denoted as x_{ki} and m_{ki} representing the integer data observation and its maximum possible value respectively.

For dataset k, the binomial parameter p_k is estimated using the following formula.

$$p_k = \frac{\sum_{i=1}^{N_k} x_{ki}}{\sum_{i=1}^{N_k} m_{ki}} \tag{2.1}$$

For the entire group of K datasets, the overall binomial parameter p_0 is estimated using the following formula.

$$p_0 = \frac{\sum_{k=1}^{K} \sum_{i=1}^{N_k} x_{ki}}{\sum_{k=1}^{K} \sum_{i=1}^{N_k} m_{ki}} \tag{2.2}$$

We first test the null hypothesis that the K values p_1, p_2, ... , p_K are all equal. The alternative hypothesis is that at least one of the p_k values is different.

$$H_0 : \quad p_1 = p_2 = \ldots = \quad p_K$$

$$H_A : \quad \textit{At least one } p_k \textit{ is different}$$

The test of hypothesis is based on the observed and expected totals for each dataset. The test statistic T is calculated as follows [5]:

$$T = \sum_{k=1}^{K} \frac{\left(p_k \sum_{i=1}^{N_k} m_{ki} - p_0 \sum_{i=1}^{N_k} m_{ki} \right)^2}{p_0 \sum_{i=1}^{N_k} m_{ki}} + \sum_{k=1}^{K} \frac{\left(q_k \sum_{i=1}^{N_k} m_{ki} - q_0 \sum_{i=1}^{N_k} m_{ki} \right)^2}{q_0 \sum_{i=1}^{N_k} m_{ki}} \tag{2.3}$$

where $q_k = 1 - p_k$ and $q_0 = 1 - p_0$.

Under the null hypothesis the test statistic T follows approximately the Chi-square distribution with parameter K-1, with the accuracy of the approximation increasing with sample size. We reject the null hypothesis for values of T that exceed the critical value. For example, if the probability of Type I error (rejecting the null hypothesis when it is true) is taken to be 0.05 (a common value), then the critical value is the 95th percentile of the Chi-square distribution with parameter K-1.

The key phrase of the first sentence of the preceding paragraph is "follows approximately." It is more accurate and practical (especially when sample sizes are small) to calculate the critical test statistic T_c for our desired confidence level c using a Monte Carlo simulation rather than approximate it using the Chi-Square. To calculate

T_c, we create a table of test statistics T containing N^{iter} elements, where N^{iter} is the number of iterations of the Monte Carlo simulation[2]. Each element of the table is populated by randomly generating K datasets, all of which are drawn from a single binomial distribution with binomial parameter p_0. The test statistic T is then calculated and stored in the table. Once the table is filled, it is sorted in ascending order of T and the element corresponding to c is selected as T_c. For example, for c = 95% and N^{iter} = 1000, the element containing T_c is $0.95 \times 1000 = 950$. That is, given a population wherein all datasets are drawn from a single binomial distribution with binomial parameter p_0, we expect the test statistic to exceed T_c 50 times out of 1000.

To test the null hypothesis, we compare the test statistic T calculated from our observed data to T_c. If T is less than or equal to T_c, then we accept the null hypothesis. If the null hypothesis is accepted we conclude that there is insufficient evidence to declare that any of the proportions are different. In this case we are justified in using the value p_0 as the estimator for the common proportion.

Logically, the minimum constraint that rejection of the null hypothesis places on the post-hoc procedure is that there must be at least two sub-groups within the group of K datasets. Let the K datasets be ordered by increasing value of the proportions p_1, p_2, …, p_K. We will consider the K-1 possible ways to break the K datasets into exactly two subgroups based on the ordered datasets. The first breakpoint is between datasets 1 and 2, forming two subgroups consisting of dataset 1 by itself, and datasets 2 through K. The second breakpoint is between datasets 2 and 3, forming two subgroups consisting of datasets 1 and 2 together and datasets 3 through K. The (K-1)[th] breakpoint is between datasets K-1 and K, forming two subgroups consisting of datasets 1 through K-1 and dataset K by itself.

Let the K-1 breakpoints for the ordered datasets be indexed by the letter b so that b is an integer index ranging from 1 to K-1.

We will use the method of maximum likelihood to select from among the K-1 possible breakpoints. For each possible breakpoint indexed by b, we will calculate the likelihood function for the two resulting subgroups of datasets and choose the breakpoint that gives the maximum likelihood.

Consider the two subgroups of datasets resulting from breakpoint b. For each of these subgroups of datasets we calculate the pooled proportions $p_1(b)$ and $p_2(b)$ as follows.

$$p_1(b) \;=\; \frac{\sum\limits_{k=1}^{b}\sum\limits_{i=1}^{N_k} x_{ki}}{\sum\limits_{k=1}^{b}\sum\limits_{i=1}^{N_k} m_{ki}} \qquad p_2(b) \;=\; \frac{\sum\limits_{k=b+1}^{K}\sum\limits_{i=1}^{N_k} x_{ki}}{\sum\limits_{k=b+1}^{K}\sum\limits_{i=1}^{N_k} m_{ki}} \qquad (2.4)$$

The likelihood function for breakpoint b is as follows.

---

[2] The considerations for the selection of N^{iter} are discussed in Chapter 3.

$$L(b) \quad = \quad \prod_{k=1}^{b} \Pr\left[\sum_{i=1}^{N_k} x_{ki} \; ; \; p_1(b) \, , \sum_{i=1}^{N_k} m_{ki}\right] \prod_{k=b+1}^{K} \Pr\left[\sum_{i=1}^{N_k} x_{ki} \; ; \; p_2(b) \, , \sum_{i=1}^{N_k} m_{ki}\right] \qquad (2.5)$$

After taking logarithms, applying the Binomial distribution for the probabilities and simplifying, we have the following function to be calculated for each breakpoint. The breakpoint that maximizes this function is then selected.

$$
\begin{aligned}
L'(b) = \sum_{k=1}^{b} & \left(\left(\sum_{i=1}^{N_k} x_{ki} \right) \ln\, p_1(b) + \left(\sum_{i=1}^{N_k} m_{ki} - \sum_{i=1}^{N_k} x_{ki} \right) \ln\left(1 - p_1(b)\right) \right) \quad + \\
\sum_{k=b+1}^{K} & \left(\left(\sum_{i=1}^{N_k} x_{ki} \right) \ln\, p_2(b) + \left(\sum_{i=1}^{N_k} m_{ki} - \sum_{i=1}^{N_k} x_{ki} \right) \ln\left(1 - p_2(b)\right) \right)
\end{aligned}
\qquad (2.6)
$$

Having divided the K groups into two subgroups, the procedure continues by testing the null hypothesis of equal proportions for each of the subgroups (where appropriate) and using the same methodology to subdivide as outlined above. This procedure continues until the null hypothesis is accepted in all cases.

2.2 Example of Binomial Analysis

We demonstrate the method discussed above by analysing a real dataset. The software implementation of the method (SPAM) has been used to perform this analysis, but only the results are discussed here.

Let us consider a hypothetical wargame comparing the effectiveness of various weapon systems. Part of the output of this wargame was the number of kills with a particular weapon and the number of opportunities with that weapon. The data are tabulated in Table 1 and presented in Figure 1. Based on a careful analysis of the experiment and the form of the data, it has been concluded that the binomial data model is most appropriate for the data. The necessary parameters, the number of iterations per Monte Carlo test of the null hypotheses, and the desired confidence level, have been set to 1000 and 95%, respectively.

Table 1. Example Data for Binomial Data Model

WEAPON SYSTEM	NUMBER OF KILLS	NUMBER OF OPPORTUNITIES	PROPORTION KILLED (OBSERVED P_K)
A	7	75	0.093
B	67	150	0.447
C	30	40	0.750
D	12	15	0.800
E	22	25	0.880

The proportion killed in the rightmost column of Table 1 is calculated by dividing the number of kills by the number of opportunities.
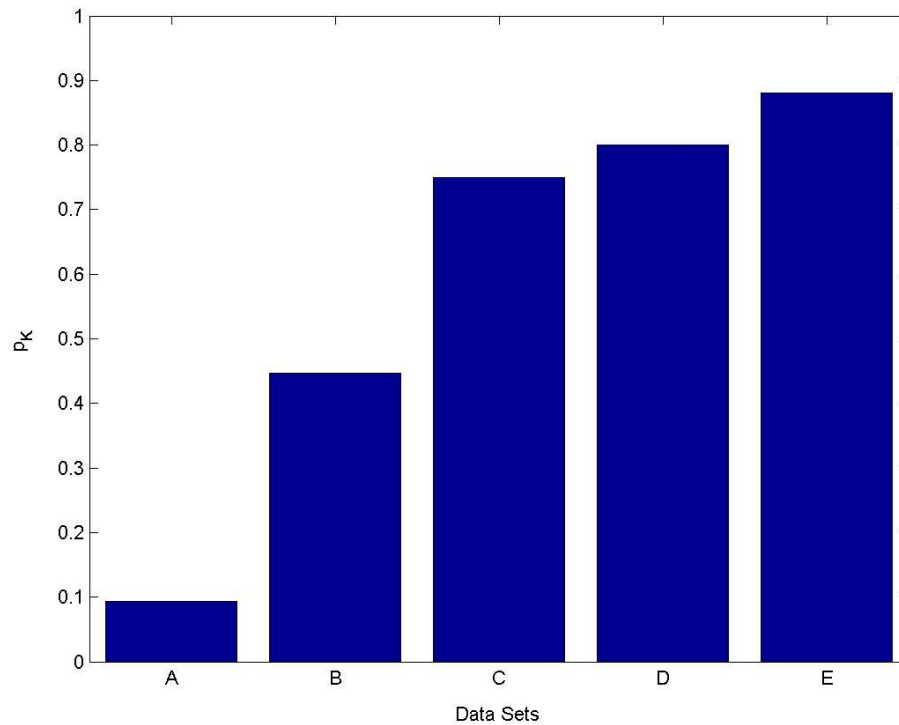
Figure 1. *Binomial Example Plot showing Observed Proportion of Kills by Data Set*

Analyzing the data in Table 1 using the binomial data model in SPAM produced the following output:

```
Maximum Likelihood Separation Points
Data Model: Binomial
```

```
Confidence Level: 95%

Group    Set
-----------
1    A
2    B
3    C
3    D
3    E

  Test Results
----------------
Test # 1:
Sets included in test: A B C D E
p-value for test: 0
Test # 2:
Sets included in test: B C D E
p-value for test: 0
Test # 3:
Sets included in test: C D E
p-value for test: 0.651

Separation Point Information
----------------------------
Separation Point: A - B
Likelihood Values:  -180.4542 -182.5447 -194.414 -199.1197
Separation Point: B - C
Likelihood Values:  -Inf -143.1493 -148.5642 -150.8558
```

As indicated in the header information, a confidence level of 95% (the default value) was selected for this analysis. After the header, the first important piece of information encountered is the grouping of the data sets into larger sets. The table lists the data sets, sorted in ascending order by mean p, along with their group memberships. Group numbers are assigned sequentially in ascending order.

The next group of information contains the results of all of the Monte Carlo Chi-square tests performed during the execution of the algorithm. These are presented in the order in which they were done.

The final group gives more information about each separation point: the data sets considered for each separation point and the likelihood of each separation point. The most likely separation point is that with the largest likelihood value, hence the term "maximum likelihood." Together with the Chi-square test information in the previous section of the analysis report, it is possible to reconstruct the steps taken by the algorithm, which we will do here.

The first Monte Carlo test included all five data sets. The returned p-value was 0, which means that none of the 1000 Monte Carlo iterations resulted in a test statistic as large as the observed test statistic of all of the data sets. Therefore, the null hypothesis that all of the data sets are drawn from the same distribution is rejected, and we proceed to search for a separation point.

The likelihood values are presented for the first separation point are -180.4542, -182.5447, -194.414, and -199.1197. Likelihood values are presented from left to right in the same order that the data sets are sorted. A separation point occurs between two sets, so for K data sets there are K-1 possible separation points. The first likelihood value refers to the separation point between the first and second data sets, the second likelihood value to the separation point between the second and third data sets, and so on. Looking at the four values, we see that that maximum value occurs in the first position, and so the first separation point is located between data sets A and B.

The algorithm now examines each new subgroup to see if there are any more separation points within either subgroup. The "left" subgroup (the subgroup to the left of the most recent separation point, as seen when the data sets are listed from left to right) contains only one data set, A, and obviously contains no more separation points.

The right subgroup, consisting of data sets B, C, D and E, is now subjected to the same Monte Carlo test of the null hypothesis. The p-value returned by the Monte Carlo test is 0, so the null hypothesis is rejected and the algorithm searches the right subgroup for the most likely point of separation.

The likelihood values calculated for the right subgroup are -Inf, -143.1493, -148.5642, and -150.8558. "Inf" is an abbreviation for "infinite," hence "-Inf" indicates a negative infinite value. The length of the likelihood vector is set by the total number of datasets under consideration. Locations representing breakpoints that are not included in the current subgroup are given a likelihood of negative infinity to remove them from the selection process. The maximum likelihood value in this list is the second, meaning that the breakpoint occurs between sets B and C.

The final Monte Carlo test is done on the remaining group of datasets C, D, and E. The resultant test statistic is 0.651, significantly larger than the maximum of 0.05 that we have required to reject the null hypothesis. There are no more subgroups to examine and so the analysis is complete. Figure 2 shows the datasets with lines dividing the groups with equal proportions of kills.
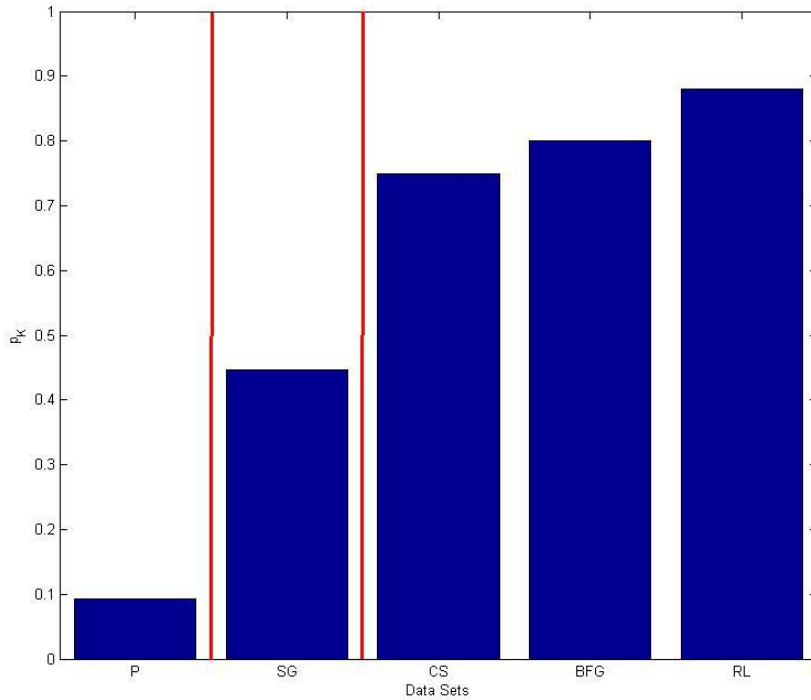
Figure 2. *Binomial Example Plot Showing the Division of the Data Sets into Groups of Equal Proportions of Kills*

2.3   Example of Normal Analysis

Let us consider a field trial experiment comparing a number of sensor systems with a variety of different measures of performance. One of these measures is the range of initial target detection, the data for which is presented in Table 2 and the means graphed in Figure 3. Because the ranges are inherently continuous-valued data, we will use the normal data model for our post-hoc analysis.

**Table 2.** Example Data for Normal Data Model

INITIAL DETECTION RANGE BY SYSTEM									
A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
1539	875	1451	1448	511	1466	1466	1466	470	1540
1453	845	1453	1466	531	1447	1570	1570	471	1538
1438	825	1492	1493	557	1564	1564	1564	461	1538
1467	740	1537	1451	552	1448	1453	1453	453	1539
1488	841	1488	1443	562	1494	1494	1494	474	1444
1453	836	1453	1466	563	1454	1454	1454	471	1470
1540	803	1540	1540	557	1452	1571	1571	471	1565
1438	715	1438	1460	569	1565	1565	1565	465	1490
1492	865	1492	1563	521	1487	1487	1487	471	1540
1538	719	1538	1464	526	1487	1466	1466	471	1543
1465	888	1465	1476	564	1458	1458	1458	474	1541
1480	872	1497	1467	552	1466	1454	1454	465	1540
1542	720	1542	1539	562	1478	1489	1489	474	1469
1453	807	1453	1455	526	1464	1455	1455	471	1486
1463	842	1463	1456	570	1444	1570	1570	469	1540
1472	771	1472	1477	545	1569	1569	1569	474	1490
1486	785	1486	1488	548	1448	1566	1541	469	1474
1486	847	1486	1474	525	1492	1462	1462	473	1493
1490	799	1490	1490	482	1466	1561	1561	438	1480
1538	787	1538	1429	541	1472	1571	1571	474	1542
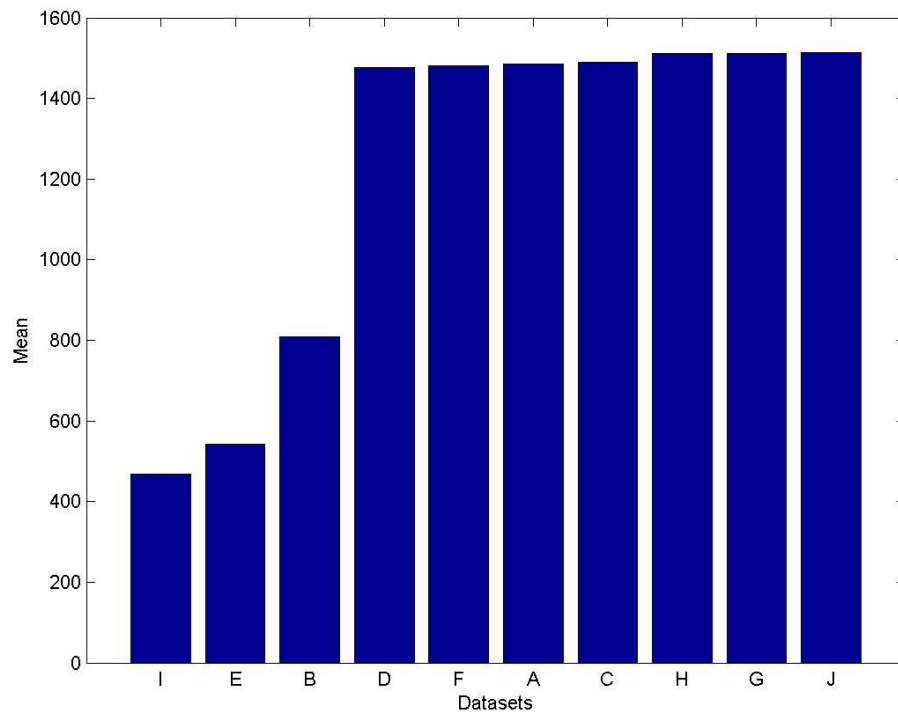
Figure 3. *Mean Initial Detection Ranges by System, Sorted by Ascending Mean*

The SPAM output resulting from a normal analysis of the data in Table 2 with a confidence interval of 95% is presented below. It structure is essentially identical to that of the report for the binomial analysis, with differences noted below.

```
Maximum Likelihood Separation Points
Data Model: Normal
Confidence Level: 95%

Group    Set
-----------
1    I
2    E
3    B
4    D
4    F
4    A
4    C
5    H
5    G
5    J


   Test Results
   ----------------
```

```
Test # 1:
Sets included in test: I E B D F A C H G J
p-value for test: 0
Test # 2:
Sets included in test: I E B
p-value for test: 0
Test # 3:
Sets included in test: I E
p-value for test: 2.2204e-016
Test # 4:
Sets included in test: D F A C H G J
p-value for test: 0.0091745
Test # 5:
Sets included in test: D F A C
p-value for test: 0.75032
Test # 6:
Sets included in test: H G J
p-value for test: 0.98996


Separation Point Information
---------------------------
Separation Point: I - E
Likelihood Values:  -105.8834
Separation Point: E - B
Likelihood Values:  -241.2772 -229.0993
Separation Point: B - D
Likelihood Values:  -1097.7921 -1020.729 -827.0436 -932.0685 -987.0781
-1037.4387 -1084.4179 -1125.0836 -1162.3108
Separation Point: C - H
Likelihood Values:  0 0 0 -523.1857 -521.2091 -518.8123 -515.6676 -
521.5617 -523.5484


Warning: Dataset D has a non-normal distribution with 95% confidence
Warning: Dataset A has a non-normal distribution with 95% confidence
Warning: Dataset C has a non-normal distribution with 95% confidence
Warning: Dataset H has a non-normal distribution with 95% confidence
Warning: Dataset G has a non-normal distribution with 95% confidence
Warning: Dataset J has a non-normal distribution with 95% confidence
```

The table of groups indicates that there are five distinct means within our set of sets. We now examine the list of test results.

The first test includes all sets. The analysis of variance (ANOVA)[3] returns a p-value of 0, so the null hypothesis is rejected, and we search for a separation point. The first separation point is located between sets B and D, where the likelihood function is -827.0436.[4] (The likelihood values for this iteration of the algorithm are located third

---

[3] The normal data models uses the analysis of variance incorporated into the Matlab Statistics Toolbox, and as such the algorithm executes substantially faster than the binomial data model, which uses a Monte Carlo algorithm implemented in code.

[4] Note that there is a discrepancy in the presentation of the separation point details between the normal and binomial data models. The presentation found in the output of the normal data model is a legacy of the original non-GUI version of SPAM. There is no impairment of functionality due to this quirk, and so it will not be changed unless a new version of SPAM is released in the future. The user may determine the correct ordering of the likelihood values by noting the stated breakpoint location and

in the list of separation point details, above.) The datasets are split into two groups and the analysis continues in the following the procedure set out in the pseudocode description (located in Annex B) and illustrated in the binomial example above. The groups with equal means listed at the top of the text report are illustrated in Figure 4.

At the bottom of the normal analysis report is a list of warnings generated by the program. Because the model assumes that the data exhibits a Gaussian distribution, there are a number of statistical tests and tools available within the Matlab Statistics Toolbox for the characterization of the datasets, among them the Lilliefors and Jarque-Bera tests for goodness-of-fit to a normal distribution. One of these tests is applied to each dataset, depending on the number of data points in the set. The Lilliefors test is used for datasets containing less than 1000 data points, the Jarque-Bera test for larger ones. In this example, datasets A, C, D, G, H and J are distributed non-normally with a 95% confidence level. The confidence level used for the goodness-of-fit tests is that specified by the user in the GUI window (See Figure 6).
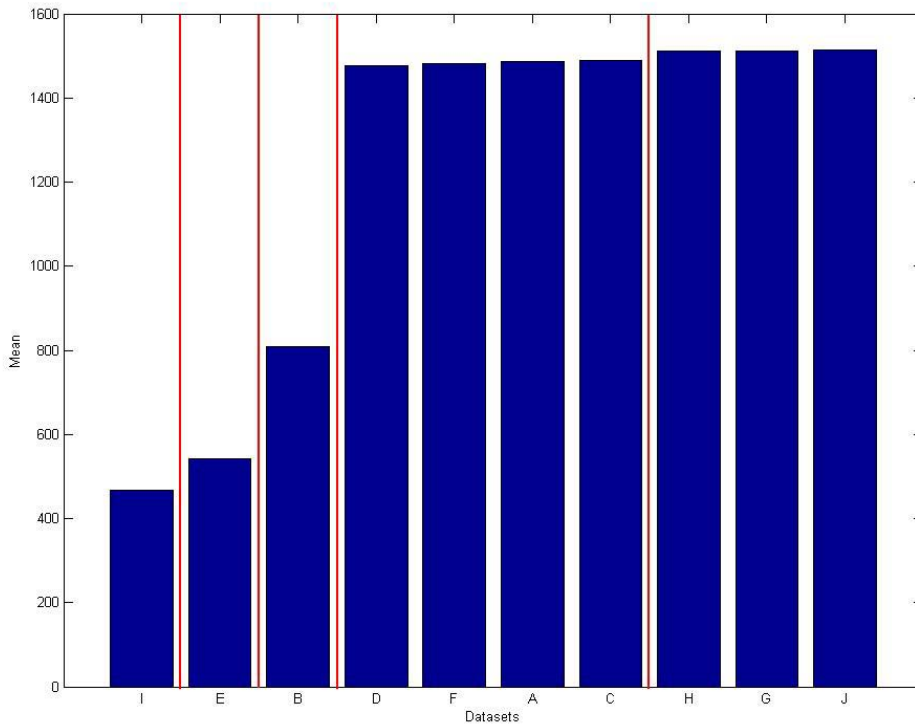
Figure 4. Mean Initial Detection Ranges with Groups of Equal Means Illustrated

correlating the number of values in the likelihood vector to the number of datasets. Likelihood values of zero should not be included in the consideration.

In general, the normal analysis algorithm is robust with respect to most non-normal distributions. Care should be exercised with multi-modally distributed data and data with very a low coefficient of variation[5], as the algorithm may give spurious results in these instances. The latter case is discussed above. The normal data model algorithm sees a multimodal dataset (or any other dataset) only through its mean and variance. The variance in this case is larger than would be the case for the individual component distribution. The effect of this is to decrease the likelihood of separating the multimodal dataset from the other datasets. That is the analysis becomes more conservative. In any event, it behoves the user to understand the characteristics of the data and to make intelligent decisions about the appropriate type of analysis.

---

[5] The coefficient of variation is defined as the ratio of the standard deviation of a distribution to the mean.

3.   SPAM

SPAM is the name of a MATLAB-based graphical user interface (GUI) utility that implements the maximum-likelihood methods for separating datasets into groups of statistically indistinguishable means, as described in this paper and in [4].   The name SPAM is an acronym standing for <u>S</u>eparation <u>P</u>oint <u>A</u>nalysis <u>M</u>ethod. The program is capable of importing data files in a number of different formats.

The primary output is an ASCII text file containing the most likely grouping of datasets, along with information about the statistical tests performed and the likelihood values generated during the analysis.   It is also possible to save a copy of the associated plot of the sorted data sets with illustrated separation points to a file. Several common graphics file formats are supported. A basic working knowledge of MATLAB is assumed in these instructions.

3.1   Installation and Operation

The necessary files for SPAM may be found on the compact disc accompanying this research note. To install SPAM, simply copy all of these files to C:\Matlab*xxxx*\work\, where *xxxx* is the version number of the Matlab installation.

SPAM was written in Matlab version 6.5 and is known to be compatible with Matlab versions 6.5 and 6.5.1.   Compatibility with prior versions is likely good but is not guaranteed.

At any time after copying the necessary files to your working folder, you may start SPAM by typing 'guispam' at the prompt in the MATLAB command window, as illustrated in Figure 5. A new window will open that looks like that shown in Figure 6.
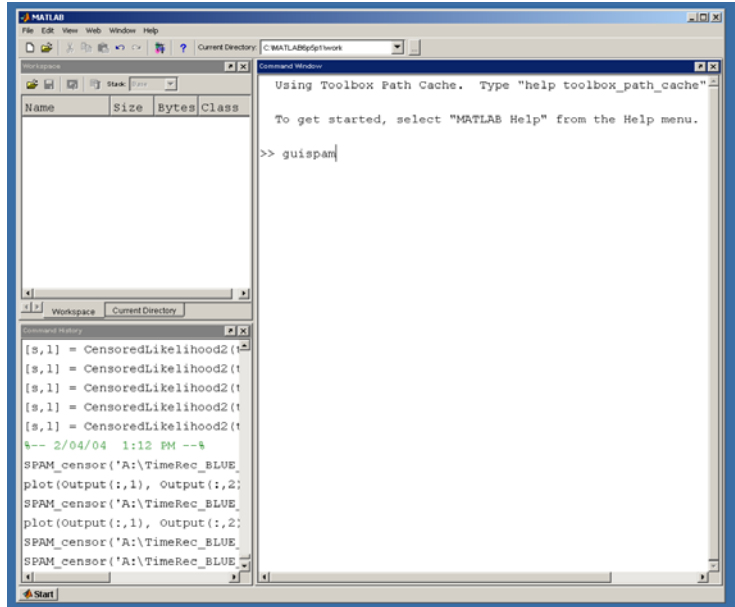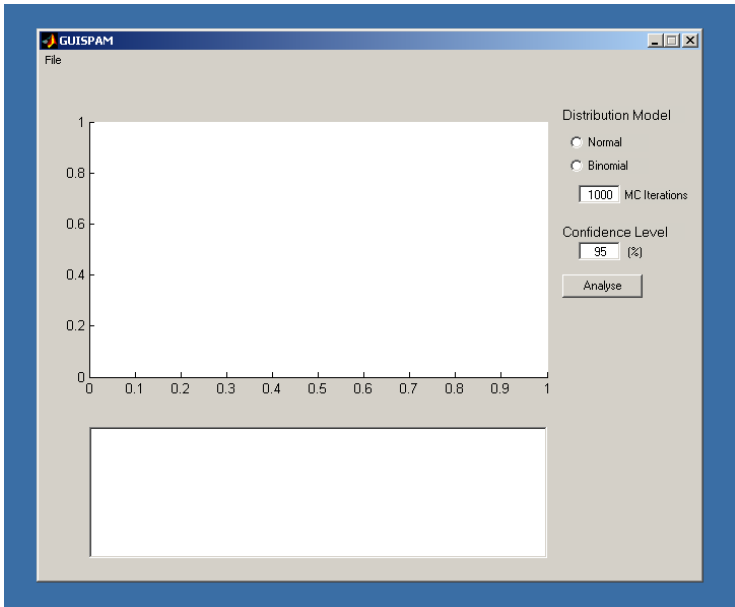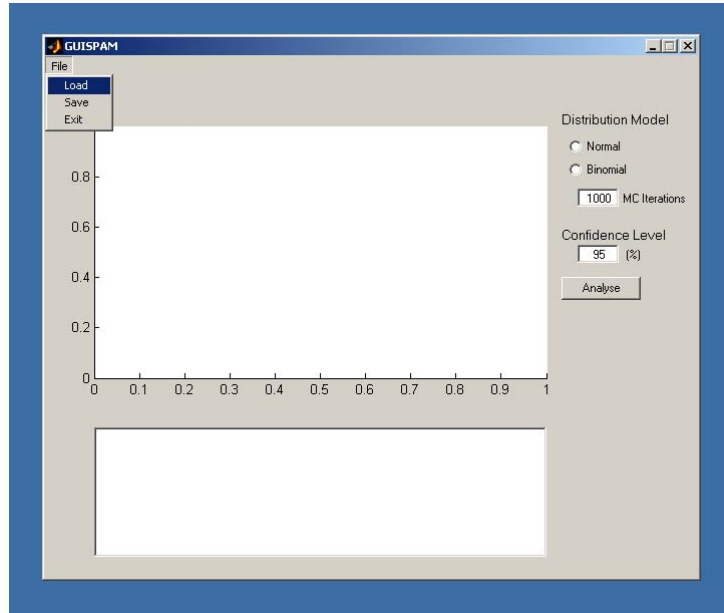
Figure 5. *Starting SPAM*

Figure 6. *SPAM*

3.2 File Menu

Clicking on the File tab of the menu bar produces a dropdown list of selections: Load, Save and Exit (See Figure 7).

Figure 7. *File Menu Options*

3.3 Load

"Load" opens a data file and loads it into SPAM for analysis. At the time of this writing, ASCII text files and Excel files are supported. In both cases, the data must be organised in rows or columns. Column and/or row headers are permitted and encouraged. SPAM can recognize some forms of row or column header and will use these as data set identifiers.

When "Load" is selected from the File menu, a new window will open, as shown in Figure 8. From here, the user can browse until the desired input file is found. Click on the file to select it, and click on "Open" to open it. After opening the data file successfully, SPAM will plot the data sets in the plot in the order in which they were found in the input data file. (See Figure 9.)
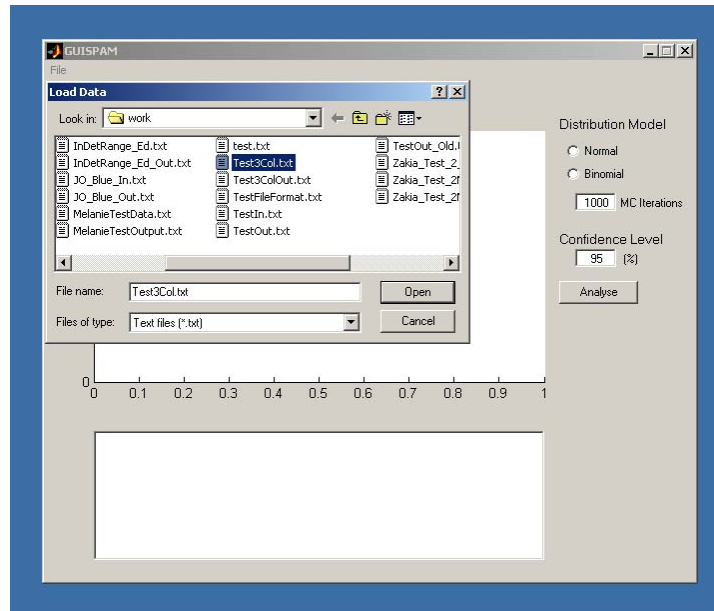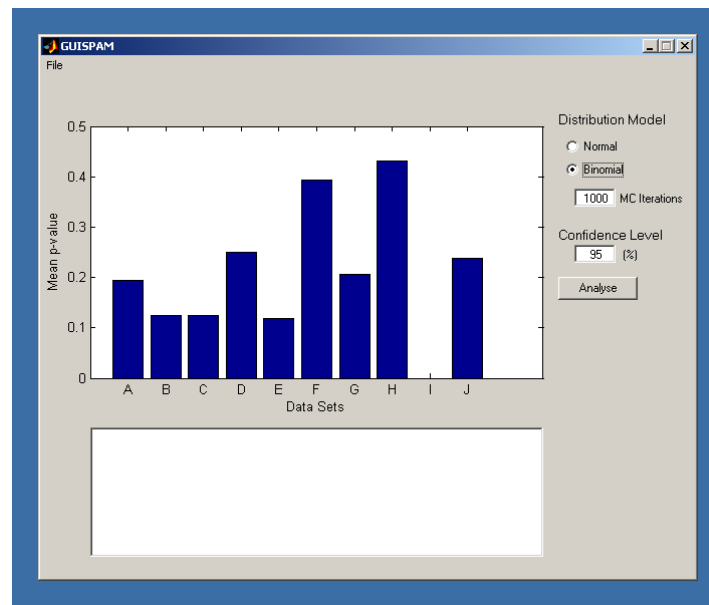
Figure 8. *Load Data Dialog Box*

Figure 9. *SPAM after loading data and selecting the confidence level, distribution model and (in this case) the number of Monte Carlo iterations*

3.4   Distribution Models

Parameters. Prior to starting the analysis, the user must select a distribution model and a confidence level before starting the analysis. SPAM will use the confidence level value entered by the user for everything that requires a confidence level. This includes

the test to determine if there is more than or mean or p-value within a group of sets, but also includes the tests done to check the normality of the data set (if using the normal distribution). The default value for the confidence level is 95%.

The maximum-likelihood method of separating groups of datasets requires that the user make some assumption about the nature of his data. The likelihood function is a function of the distribution function, and the form of the statistical tests used varies with the distribution as well. SPAM has two distribution models, the normal or Gaussian distribution and the binomial distribution. It is very important that the user understand his data and selects the most appropriate model. This section attempts to provide some guidance on distribution selection. Please consult Chapter 2 of this paper and [4] for more details.
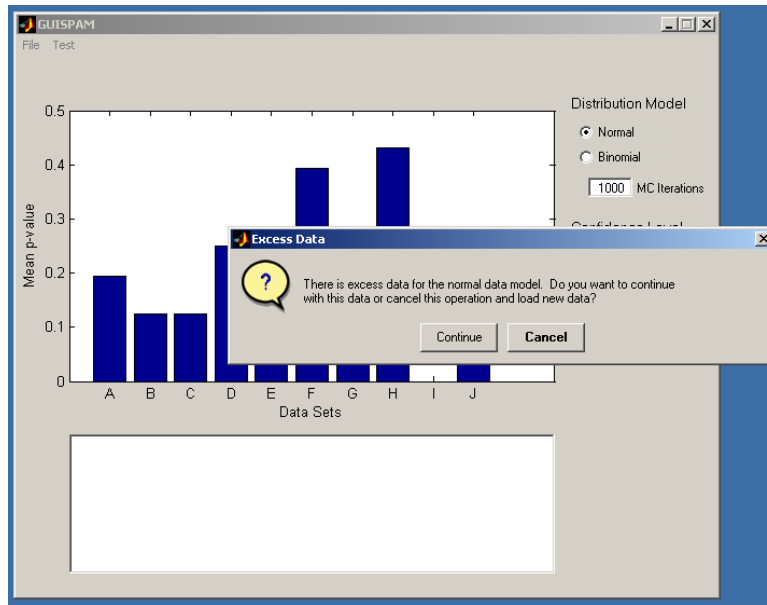
3.4.1 Normal Distribution.

The normal distribution makes the assumption that the data points within each set are distributed normally about the mean value of the set. Notwithstanding that, the method generally gives good results with skewed and even multimodal distributions, although the latter tends to result in very conservative analyses. However, it is **not** reliable in the face of very low variance data sets. As the variance of a dataset approaches zero, the likelihood that that set will be found by SPAM to be a separate dataset approaches 100%, regardless of how small is the difference between its mean and that of any other dataset. If the variance of the dataset is equal to zero, the algorithm for the normal distribution will put the dataset into a separate group even if it has the same mean as another dataset.

In short, the normal distribution is appropriate for nearly all real-valued datasets. In some instances it may also give good results for discrete-valued datasets, but the binomial distribution may be more appropriate. The user, based on the nature of the dataset, must make such a determination.

3.4.1.1 Data Requirements

The normal distribution algorithm requires two pieces of information per data point: an identifier that indicates its set membership and the actual observed value, in that order. Additional information will not be used in the analysis. If extra information is found by SPAM, the user will be prompted to confirm that he or she did indeed want to use the normal distribution. (See Figure 10.)

Figure 10. *Message indicating that there is more data than necessary for normal data model*

3.4.2 Binomial Distribution

The binomial distribution assumes that the data sets contain discrete values with a single p-value. Nonetheless, the algorithm works reasonably well for multimodal distributions. However, it is suitable for use only with discrete data sets. See Chapter 2 for a discussion of the capabilities and limitations of this technique.

Before commencing the analysis use the binomial model, the user must supply SPAM with an extra parameter not required for the normal model, the number of iterations N^{iter} to be used for each Monte Carlo (MC) simulation.

The default and minimum number of iterations is 1000. The maximum value is 50000 iterations. Note that because of the maximum precision implicit in the number of iterations, it is pointless to select a confidence level such that the product of the confidence level and the number of iterations is less than 1. Likewise, the number of iterations and confidence level should be chosen such that their product is an integer value.

3.4.2.1 Data Requirements

The binomial distribution model requires three pieces of information for each data point in the following order: an identifier that indicates its set membership, the observed number of occurrences, the maximum possible number of occurrences.

SPAM will warn the user if sufficient information is not present (See Figure 11.).
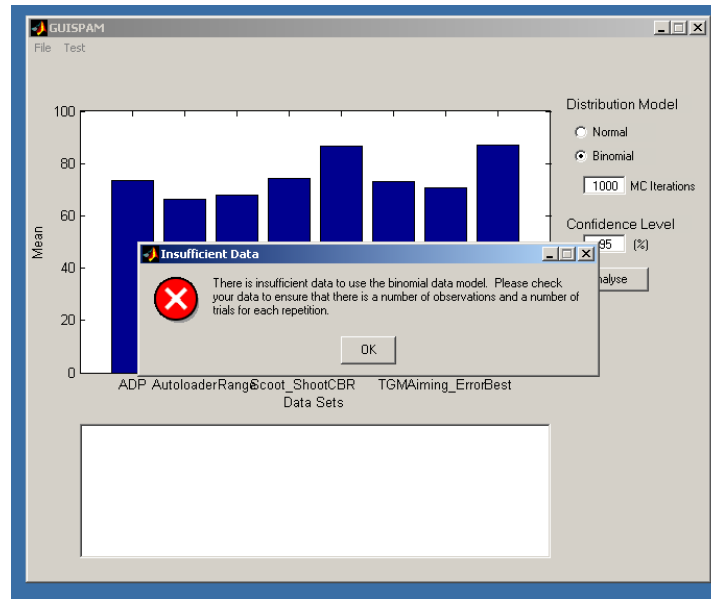
Figure 11. *Error message indicating a lack of data*

3.5 Analysis

Once the necessary parameters have been determined, the user may begin the analysis by pressing the "Analyse" button.   A wait bar will appear that indicates the approximate degree of completion of the analysis, as shown in Figure 12.
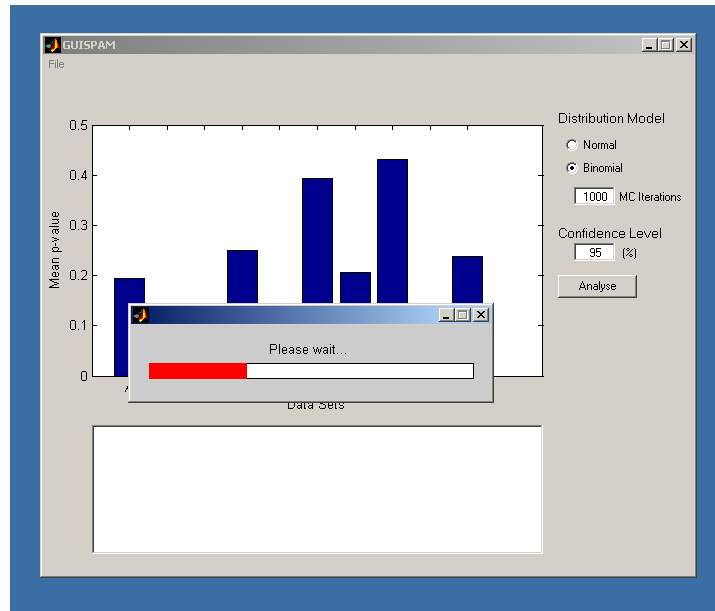
*Figure 12.* During analysis, SPAM gives a rough indication of progress

3.6 Output

SPAM produces both graphical and textual output. As part of its analysis, SPAM sorts the data sets by ascending mean/p-value. Upon completion, the graph is re-plotted in this sorted order (See Figure 13.). Lines drawn between groups of sets illustrate the separation points.

Textual output is located in the text box below the plot. The header indicates data model and confidence level used for the analysis. Below these is a list of the sorted data sets with groupings of sets given. This is not visible in the figure. To view the complete text output, click inside the text box and use the cursor to scroll up or down. The numbers assigned to the groups are arbitrary.

After the set grouping list is information about the resultant p-value of each statistical test SPAM performed and the likelihood values associated with each separations. These allow the user to make his/her own judgements about the results and assess the sensitivity of the results to confidence level and data model.

Finally, depending on the data model, there may also be messages or warnings appended to the end of the text output.
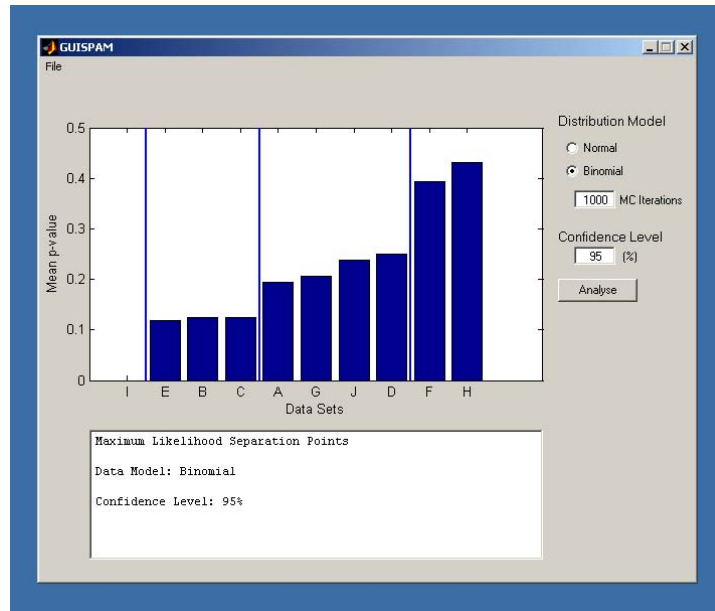
Figure 13. After analysis is complete. Note that the data sets are sorted by ascending p-value, the line separating the significant groupings and the presence of output data in the text box below

3.7 Save

Results can be saved by selecting the Save option from the File menu at the upper left of the screen (See Figure 14). The will launch a save dialog box, such as that shown in Figure 15, where the user can specify the output file name and location. This output file will contain only the textual output found in the text box in ASCII format.

After saving the text output, the user will also be given the option to save the graph in a number of common graphics formats (See Figure 16).
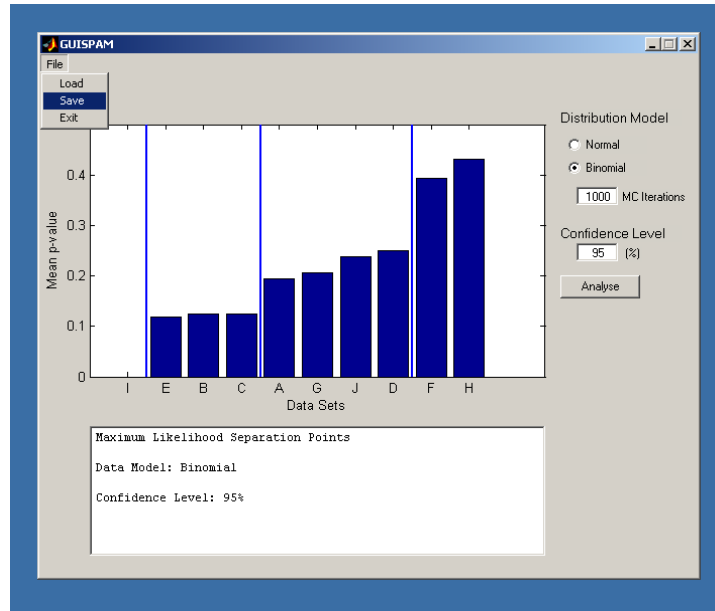
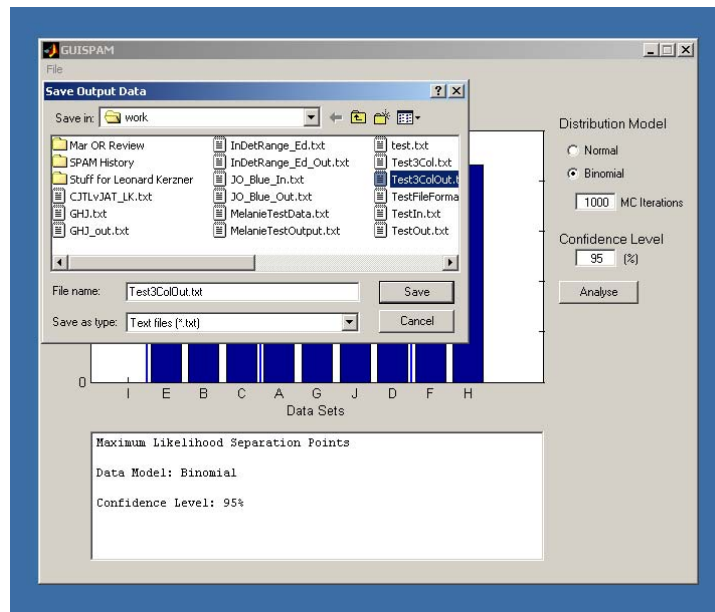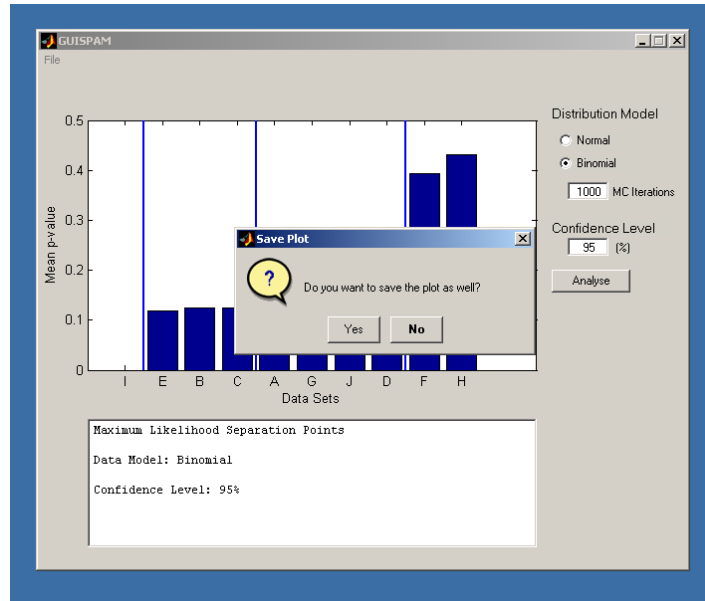**Figure 14.** *Saving results in SPAM*

Figure 15. *Save Data Dialog Box*

Figure 16. Option to Save Diagram Dialog Box

3.8 Exit

The user can exit SPAM by opening the File menu and selecting the Exit option. Upon doing so, the user will be asked if he/she is sure. Clicking "Yes" ends the program. Clicking "No" returns the user to SPAM without any loss of data. Alternatively, clicking on the "x" in the upper right corner o the SPAM window will also end the program, but without the safeguards described above.

3.9 Test Menu

The Test Menu is intended to contain functions to test the operation of the algorithms contained in SPAM. While originally designed to assist with debugging, they may also be employed by the user to gain a better understanding of the analytical sensitivity that can be achieved with SPAM. At the present time, there is only one test function included, Test Binomial. More test functions may be added in future versions if additional data models are incorporated.

3.10 Test Binomial

The Test Binomial function allows to user to specify a set of datasets with varying p-values and sizes, and then tests SPAM's ability to correctly separate the datasets into the appropriate groups of datasets. The user must enter a vector of p-values, a vector of dataset sizes, and a number of iterations for the test, as illustrated in Figure 17. The number of iterations here is not the number of iterations for the Monte Carlo generation of the test statistic described above, but the number of times that SPAM

will create a new set of datasets and test them. The number of iterations and the confidence level specified on the main SPAM window are used for the test. For each iteration of the test SPAM creates a new set of random binomial datasets, based on the provided p-value and size vectors. Note that Test Binomial may take some time to complete, depending on the speed of your computer.
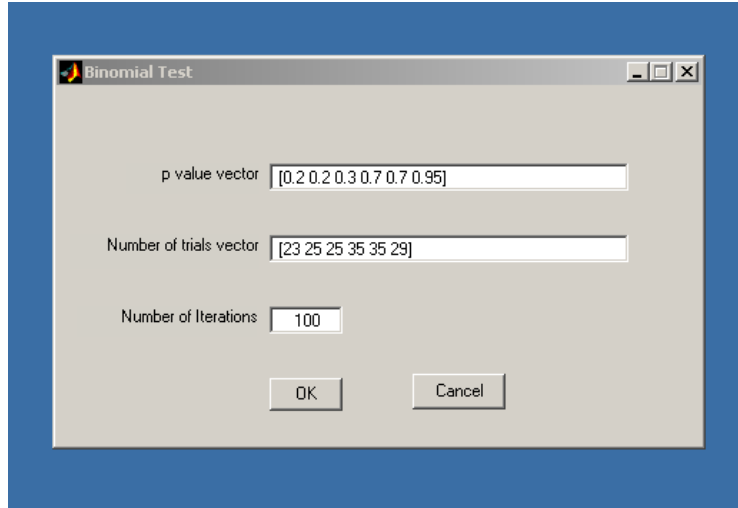
Figure 17. Test Binomial Dialog Box

The output of the Test Binomial function, which is displayed in the Matlab command window, is very spare (see Figure 18). More details can be found by examining the internal operation of the function, and the interested user is encouraged to examine the code.
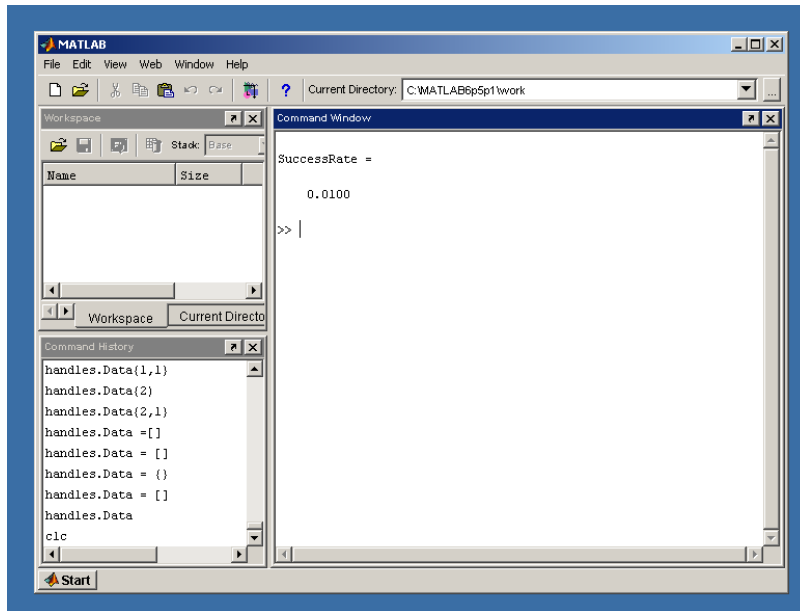
Figure 18. *Test Binomial Output*

4. CONCLUSION

This purpose of this research note was to present two useful additions to the ORD analyst's toolkit. The first is an extension of the post hoc analysis technique presented in [4] to discrete data sets using a binomial data model. The second addition is the software implementation of the procedure of both this research note and that of [4] in the form of SPAM, a Matlab-based utility with a graphical user interface. SPAM is a useful and simple tool for post-hoc analyses, allowing a scientist or analyst to determine quickly and easily the most likely point or points of separation between data sets that have been shown to have means or proportions that are from more than one distribution. Proficient Matlab users can customize it easily to add additional data models, following the model of existing code and the pseudocode outlined in Annex B.

5. References

1. Dayton, C.M. (1998), "Information Criteria for the Paired-Comparisons Problem," *American Statistician*, 52, 144-151.

2. Scheffé, H. (1959). *The Analysis of Variance.* New York: John Wiley & Sons, Inc.

3. Winer, B.J., Brown, D.R., Michels, K.M. (1991) *Statistical Principles in Experimental Design, 3rd Edition*. Boston: McGraw-Hill, Inc.

4. Emond, E.J, and Massel, P.L. (2003) A Post-ANOVA Methodology for Finding Subgroups With Equal Means Using Maximum Likelihood. ORD Research Note RN 2003/05. Ottawa, Ontario: ORD

5. Fleiss, J.L. (1973) *Statistical Methods for Rates and Proportions*. New York: John Wiley & Sons, Inc.

Annexes

Annex A: Simplification of the Likelihood Function for the Normal Data Model

DOR (Corp) Research Note RN 2004/11

July 2004

Massel and Emond [4] give the likelihood function as

$$
\begin{aligned}
L'(b) = &\sum_{k=1}^{b} \sum_{n=1}^{N_k} \left(-Log(s_1) - \frac{(x_{k,n} - \bar{x}_1)^2}{2s_1^2} \right) \\
&+ \sum_{k=b+1}^{K} \sum_{n=1}^{N_k} \left(-Log(s_2) - \frac{(x_{k,n} - \bar{x}_2)^2}{2s_2^2} \right)
\end{aligned}
\tag{A1}
$$

where s_1 and s_2 are the standard deviations of the groups of datasets located to the left and right of breakpoint b, respectively.

This form may be preferable for the explanation of the procedure, but it is not optimal from the standpoint of computer execution of the algorithm. We now show this equation simplifies considerably.

First, we notice that the standard deviations s_1 and s_2 are independent of the summations, and so we factor them out of the summations.

$$
\begin{aligned}
L'(b) = &-Log(s_1) \sum_{k=1}^{b} \sum_{n=1}^{N_k} 1 - \frac{1}{2s_1^2} \sum_{k=1}^{b} \sum_{n=1}^{N_k} (x_{k,n} - \bar{x}_1)^2 \\
&-Log(s_2) \sum_{k=b+1}^{K} \sum_{n=1}^{N_k} 1 - \frac{1}{2s_2^2} \sum_{k=b+1}^{L} \sum_{n=1}^{N_k} (x_{k,n} - \bar{x}_2)^2
\end{aligned}
\tag{A2}
$$

The sum $\left(\sum_{n=1}^{N_k} 1 \right)$ in the first and third terms of the above equation is of course simply equal to N_k.

From Ref. [4], s_1 and s_2 are defined as

$$s_1 = \left(\frac{\displaystyle\sum_{k=1}^{b} \sum_{n=1}^{N_k} \left(x_{k,n} - \overline{x_1} \right)^2}{\left(\displaystyle\sum_{k=1}^{b} N_k \right) - 1} \right)^{\frac{1}{2}} \tag{A3}$$

and

$$s_2 = \left(\frac{\displaystyle\sum_{k=b+1}^{K} \sum_{n=1}^{N_k} \left(x_{k,n} - \overline{x_2} \right)^2}{\left(\displaystyle\sum_{k=b+1}^{K} N_k \right) - 1} \right)^{\frac{1}{2}} \tag{A4},$$

which are of course the standard definitions of the variance. We rearrange these equations:

$$\sum_{k=1}^{b} \sum_{n=1}^{N_k} \left(x_{k,n} - \overline{x_1} \right)^2 = s_1^2 \left(\left(\sum_{k=1}^{b} N_k \right) - 1 \right), \tag{A5}$$

$$\sum_{k=b+1}^{K} \sum_{n=1}^{N_k} \left(x_{k,n} - \overline{x_2} \right)^2 = s_2^2 \left(\left(\sum_{k=b+1}^{K} N_k \right) - 1 \right) \tag{A6}$$

and substitute these back into Equation (A2), which gives

$$L'(b) = 1 - Log(s_1) \sum_{k=1}^{b} N_k - \frac{1}{2} \left(\sum_{k=1}^{b} N_k + \sum_{k=b+1}^{K} N_k \right)$$
$$- Log(s_2) \sum_{k=b+1}^{K} N_k \tag{A7}.$$

Consider the factor inside the brackets of second term of Eq. (A7). This is nothing more than the number of data points in the set under consideration.

$$\sum_{k=1}^{b} N_k + \sum_{k=b+1}^{K} N_k = \sum_{k=1}^{K} N_k = N \tag{A8}$$

It is obviously constant and completely independent of b. Since we interested in the value of b with the maximum likelihood, and not the actual value of the likelihood

function itself, we can subtract (N+1) from L′(b) to form a new likelihood function ℓ(b):

$$\ell(b) = L'(b) - (N + 1) \tag{A9}$$

$$\ell(b) = -Log\,(s_1)\sum_{k=1}^{b} N_k - Log\,(s_2)\sum_{k=b+1}^{K} N_k \tag{A10}$$

Examining Equation (A10), we can see that the likelihood function is essentially the negative of the sum of the natural logarithms of the standard deviations of each subset weighted by the number of data points in each subset.

ANNEX B: Analytical Algorithm Pseudocode

DOR (Corp) Research Note RN 2004/11

July 2004

The basic algorithm used by SPAM is detailed below in pseudocode. The full implementation may be found on the accompanying CD-ROM.

Pseudocode:

1. Select a data model and set necessary parameters

2. Sort datasets by ascending mean/proportion.

3. Calculate appropriate test statistic for the full set of datasets. Compare the test statistic to the critical test statistic for the desired confidence level.

    a. If the datasets are not statistically differentiable within the specified confidence level (i.e., the test statistic is less than or equal to the critical test statistic, and the null hypothesis therefore accepted), end the analysis.

    b. Otherwise, continue.

4. Evaluate likelihood function for all possible break points for current set of sets.

5. Break datasets into two subsets at maximum likelihood point.

6. For each of the two new subsets, repeat step 3 to determine if the means/proportions in the subset are all the same or not. Begin with the subset to the "left" of the break point. ("Left" is defined as the side of the break point with the lower mean(s).)

    a. If the null hypothesis is rejected per step 3a, perform steps 4-6 for this subset.

    b. Repeat step 6 for the right-hand subset.

7. Once there are no more separation points found, output the results.

List of symbols/abbreviations/acronyms/initialisms

DND   Department of National Defence

ANOVA   Analysis of Variance

CF   Canadian Forces

GUI   Graphical User Interface

HSD   Honest Significant Difference

ORD   Operational Research Division

SPAM   Separation Point Analysis Method

This page intentionally left blank.

DOCUMENT CONTROL DATA		
(Security classification of title, body of abstract and indexing annotation must be entered when the overall document is classified)		
1. ORIGINATOR (the name and address of the organization preparing the document. Organizations for whom the document was prepared e.g. Establishment Sponsoring a contractor's report, or tasking agency, are entered in Section 8). Operational Research Division Department of National Defence Ottawa, Ontario K1A 0K2	2. SECURITY CLASSIFICATION (overall security classification of the document, including special warning terms if applicable) UNCLASSIFIED	
3. TITLE (the complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C or U) in parentheses after the title) SEPARATION POINT ANALYSIS METHOD (SPAM) (U)		
4. AUTHORS (last name, first name, middle initial) HUNTER, D. GREGORY, EMOND, EDWARD J.		
5. DATE OF PUBLICATION (month Year of Publication of document) July 2004	6a. NO OF PAGES (total containing information. Include Annexes, Appendices, etc.) 47	6b. NO OF REFS (total cited in document) 5
7. DESCRIPTIVE NOTES (the category of document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.) Research Note		
8. SPONSORING ACTIVITY (the name of the department project office or laboratory sponsoring the research and development. Include the address). None		
9a. PROJECT OR GRANT NO. (if appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.) None	9b. CONTRACT NO. (if appropriate, the applicable number under which the document was written.) None	
10a. ORIGINATOR's document number (the official document number by which the document is identified by the originating activity. This number must be unique to this document.)	10b. OTHER DOCUMENT NOS. (Any other numbers which may be assigned this document either by the originator or by the sponsor.)	
11. DOCUMENT AVAILABILITY (any limitations on further dissemination of the document, other than those imposed by security classification.) (X) Unlimited distribution ( ) Distribution limited to defence departments and defence contractors: further distribution only as approved ( ) Distribution limited to defence departments and Canadian defence contractors; further distribution only as approved ( ) Distribution limited to government departments and agencies; further distribution only as approved ( ) Distribution limited to defence departments; further distribution only as approved ( ) Other (please specify):		
12. DOCUMENT ANNOUNCEMENT (any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in 11) is possible, a wider announcement audience may be selected.)		

13. ABSTRACT (a brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), or (U). It is not necessary to include here abstracts in both official languages unless the test is bilingual).

This research note documents an extension of the iterative post-hoc analysis method described by Emond and Massel (2003) to positive integer-value datasets. The original assumption was that the data were drawn from one or more normal distributions. An additional data model has now been added in which the data are assumed to be derived from one or more binomial distributions. Using an appropriate test statistic and likelihood function, this method finds the most likely separation points between datasets that have been shown to have binomial parameters that are statistically significantly different. The method is of particular interest to operational researchers due to three important properties: it always returns an answer, this answer is never intransitive, and its error accumulation is multiplicative rather than additive. Both the binomial data model and the normal data model of Emond and Massel (2003) have been implemented in a MATLAB software application known as SPAM (Separation Point Analysis Method).

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus-identified . If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

Multiple comparisons, statistical testing, post-hoc analysis

Canada