Small-Sample Inference in SAS® Software: Using Exact Methods

Cyrus R. Mehta
Cytel Software Corporation, Cambridge, MA

1 Introduction

Exact tests (or randomization tests as they are sometimes called), are a way to obtain exact p-values or confidence intervals without making any large-sample or parametric assumptions about the data. Thus they provide valid inferences for popular hypothesis tests like the Pearson chi-squared test or the likelihood ratio chi-squared test, even when most statistical packages print out a warning that "the results may not be valid because some of the cell counts in the observed contingency table are small". The method of inference is conceptually simple and may be viewed as a straight-forward extension of Fisher's exact test for the 2×2 table. The hard part is doing the computations. However, the recent development of fast numerical algorithms combined with the rapid advances in computer technology have made it feasible to run exact tests routinely. The StatXact package (Cytel Software Corporation, 1991) provides exact tests and exact confidence intervals for a broad class of nonparametric problems, including one, two, and k-sample problems, ordered or unordered $r \times c$ contingency tables, and stratified $2 \times c$ contingency tables.

In the pharmaceutical industry, StatXact is typically used as a stand-alone package in interactive mode to check the validity of p-values produced by SAS on suspiciously small data sets. There are no tools to interface StatXact with SAS for large production runs in which hundreds of tables are analysed in batch mode and the output submitted as part of a medical report. This is a serious deficiency, for it inhibits the routine use of StatXact. Data can change at the last minute as corrections or omissions are rectified. When that happens the batch-mode SAS production run is usually repeated. But the interactive StatXact spot-checks on small contingency tables may not be repeated for lack of time and resources.

This workshop presents the results of a collaboration between the SAS Institute and Cytel Software Corporation to develop a special version of StatXact for SAS users which can facilitate the integration of StatXact output into a SAS production run. We begin with a brief summary of the statistical principles underlying inference. Next we show how to activate StatXact from SAS, read in a SAS data set, and run StatXact in either interactive mode or batch mode. Finally we show how to submit a batch run from SAS that will automatically invoke StatXact, run a series of exact statistical procedures, and return to SAS with the StatXact output available as a SAS transport file or a formatted ASCII file, suitable for integration into SAS's report generation programs.

2 How Exact Tests Work: The Permutation Principle

At the heart of an exact test is the permutation principle. Under this principle we are required to permute the data in all possible ways and compare what was actually observed with what could have been observed under the null hypothesis. Specifically we calculate the exact p-value, or exact probability that a data set more extreme than the one actually obtained could be generated by chance alone. If this probability is small, we conclude that we have indeed observed a statistically significant outcome. Let us make this idea more precise. The following five-step procedure is the typical way in which the permutation principle is applied to nonparametric problems.

Step 1: Let $\mathbf{X}^*$ represent the observed data set. (For example, $\mathbf{X}^*$ is an $r \times c$ contingency table.) Let H be the null hypothesis of interest. We would like to determine whether the observed $\mathbf{X}^*$ is consistent with this hypothesis, by computing its p-value.

Step 2: Let Γ represent all possible permutations of the observed data set, subject to some constraints. (For instance, Γ could be the set of all contingency tables, $\mathbf{X}$, having the same marginal sums as the observed contingency table, $\mathbf{X}^*$. We refer to Γ as the **reference set.**

Step 3: Order each $\mathbf{X} \in \Gamma$ in accordance with a real valued function d: $X \rightarrow \mathcal{R}$. The statistic $d(\mathbf{X})$ is a measure of the **discrepency** of the table, $\mathbf{X}$, under the null hypothesis. Usually large values of $d(\mathbf{X})$ imply that $\mathbf{X}$ is inconsistent with the null hypothesis. (A good example of a discrepency measure is the Pearson chi-squared statistic for an $r \times c$ contingency table, $\mathbf{X}$.)

Step 4: Let h: $\Gamma \rightarrow [0,1]$ be a probability function such that $h(\mathbf{X})$ is the probability under H of observing $\mathbf{X} \in \Gamma$. (For a very large class of nonparametric problems, $h(\mathbf{X})$ is a generalized hypergeometric probability.)

Step 5 Inference is based on the distribution of the test statistic $d(\mathbf{X})$. The p-value is defined as

$$p = \sum_{d(\mathbf{X}) \geq d(\mathbf{X}^*)} h(\mathbf{X}) .$$

If the p-value is small (say less than 0.05), H is rejected.

The following are interesting aspects of the above permutational principle.

- The inference is nonparametric in the sense that the functional form of $h(\mathbf{X})$ is known without making any distributional assumptions about the population from which $\mathbf{X}$ arose.
- The distribution, or tail area, of $d(\mathbf{X})$ is obtainable by large-sample theory, Monte Carlo algorithms, and exact permutational algorithms. The large sample methods are the easiest but least accurate while exact methods are the most difficult but most accurate. Monte Carlo is often a good compromise between these two alternatives.
- The test statistic may be such that either very small or very large values lead to rejection of H. When this is the case $d(\mathbf{X})$ can furnish both one and two sided p-values.

2.1 A Data Set

We will illustrate the permutation principle through an analysis of the following 5×3 contingency table taken from page 5-45 of the StatXact-Turbo user manual. The tumor regression rates of five chemotherapy regimens, Cytoxan (CTX) alone, Cyclohexyl-chloroethyl nitrosurea (CCNU) alone, Methotrexate (MTX) alone, CTX+MTX, and CTX+CCNU+MTX were compared in a small clinical trial. Tumor regression was measured on a three-point scale: no response, partial response, or complete response. The results are tabulated below.

CHEMO	No Resp.	Partial Resp.	Complete Resp.
CTX	2	0	0
CCNU	1	1	0
MTX	3	0	0
CTX+CCNU	2	2	0
CTX+CCNU+MTX	1	1	4

Small pilot studies like this one are frequently conducted as a preliminary to planning a large-scale randomized clinical trial. We want to test the null hypothesis that there is no association between the chemotherapy regimen and the tumor regression.

2.2 Exact Inference

Let us go through the five steps of the permutation principle outlined above.

Step 1: The observed data, $\mathbf{X}^*$, is the above 5×3 contingency table. The cell counts are very small, and it is unclear whether the five chemotherapy regimens are significantly different.

Step 2: The reference set, Γ, is the set of all 5×3 contingency tables of the form:

$$\begin{array}{ccc|c} x_{11} & x_{12} & x_{13} & 2 = m_1 \\ x_{21} & x_{22} & x_{23} & 2 = m_2 \\ x_{31} & x_{32} & x_{33} & 3 = m_3 \\ x_{41} & x_{42} & x_{43} & 4 = m_4 \\ x_{51} & x_{52} & x_{53} & 6 = m_5 \\ \hline 9 = n_1 & 4 = n_2 & 4 = n_3 & 17 = N \end{array}$$

Step 3: We will now derive the probability $h(x)$ of observing the table $X \in \Gamma$. Let p_{ij} be the probability that a patient treated with the chemotherapy regimen in row i has the tumor response in column j. Then the null hypothesis may be stated as

$$H\colon p_{ij} = p_j \text{ for all } i = 1, 2, \ldots 5.$$

In other words the probability of a response falling in category j does not depend in any way on the treatment regimen, i. Since each row of X has an independent multinomial distribution, the conditional probability of observing any $X \in \Gamma$ is

$$h(X) = \frac{\prod_{i=1}^{5}\left[\left(\frac{m_i!}{\prod_{j=1}^{3} x_{ij}!}\right)\left(\prod_{j=1}^{3} p_{ij}^{x_{ij}}\right)\right]}{\sum_{X\in\Gamma}\prod_{i=1}^{5}\left[\left(\frac{m_i!}{\prod_{j=1}^{3} x_{ij}!}\right)\left(\prod_{j=1}^{3} p_{ij}^{x_{ij}}\right)\right]},$$

which simplifes under H to the generalized hypergeometric probabilty

$$h(X) = \prod_{i=1}^{5}\frac{m_i!}{\prod_{j=1}^{3} x_{ij}!}\Big/\frac{N!}{\prod_{j=1}^{3} n_j!}.$$

Step 4: There are five popular discrepancy measures for ordering the tables in Γ. All these measures capture the departure of a table from the null hypothesis. However the first three measures are 'omnibus' measures and capture departures from H is any general direction. On the other hand, the last two measures are more specialized. They have superior power for capturing departures from H in some very specific directions. We shall see that they are more appropriate for the present data set.

Pearson:

$$d_1(X) = \sum_i\sum_j (x_{ij} - \frac{m_i n_j}{N})^2/(\frac{m_i n_j}{N}).$$

This is the familiar χ^2 statistic. Large values imply that the observed x_{ij} differs markedly from the expected cell value $(m_i n_j/N)$ under H, and are indicative of departures from H.

Fisher:

$$d_2(X) = [h(X)]^{-1}.$$

This discrepancy measure orders the tables according to the inverse of their respective hypergeometric probabilities. Use of this measure leads to the natural generalization of Fisher's exact test for 2×2 tables.

Likelihood Ratio:

$$d_3(X) = 2\sum_i\sum_j x_{ij}\log\left(\frac{x_{ij}N}{m_i n_j}\right).$$

This is also known as the G^2 statistic in popular textbooks.

Notice that in the observed 5×3 contingency table the columns have a natural ordering. That is, a response falling in column j is preferred to a response falling in column $j-1$. If some chemotherapy regimens are more responsive than others, evidence against the null hypothesis must manifest itself by shifts to the right hand columns on some rows of the contingency table relative to others. These departures from the null are captured more effectively by the discrepancy measure

Kruskal-Wallis:

$$d_4(X) = K\sum_{i=1}^{5}\left\{\sum_{j=1}^{3}[r_j x_{ij}] - \frac{m_i(N+1)}{2}\right\}^2/m_i,$$

where,

$$r_j = n_1 + n_2 + \ldots + n_{j-1} + 1/2(n_j + 1),$$

is the mid-rank of the counts falling in column j and

$$K = \frac{12}{N(N+1)[1 - \sum_{j=1}^{3}(n_j^3 - n_j)/(N^3 - N)]}.$$

To see that $d_4(X)$ is a more natural discrepancy measure when the columns are ordered notice that $\sum_{j=1}^{3} r_j x_{ij}$ is the sum of mid-ranks of the observations in group i and $m_i(N+1)/2$ is the expected value of this rank-sum under H. Thus observations with a better response get more weight than those with a poorer response. This discrepancy measure has more power to detect departures from the null hypothesis which occur in the form of response shifts from one regimen to the next. Finally, if the $r \times c$ table has only two rows ($r = 2$), then $d_4(X)$ specializes to the Wilcoxon rank-sum test for tied data.

One could argue that the rows of the observed 5×3 contingency table are also ordered. That is, the chemotherapy regimen in row i is more aggressive than the chemotherapy regimen in row $(i-1)$. In this case evidence against the null hypothesis would manifest by progressively increasing rightward shifts in response as we moved down the rows of the contingency table. A discrepancy measure that is especially powerful for detecting such departures from the null in doubly ordered contingency tables is the

Linear-by-Linear Association:

$$d_5(X) = \frac{[LL(\mathrm{X}) - E(LL(\mathrm{X}))]^2}{var(LL(\mathrm{X}))}$$

where

$$LL(\mathrm{X}) = \sum_{i=1}^{r}\sum_{j=1}^{c} u_i v_j y_{ij} \ ,$$

$u_i, i = 1, 2, \ldots r$, are arbitrary row scores, and $v_j, j = 1, 2, \ldots c$, are arbitrary column scores. (An important special case is where $u_i = i$ and $v_j = j$.) Under the null hypothesis of no row by column interaction the LL statistic has a mean

$$E(LL(\mathrm{X})) = N^{-1}(\sum_{i=1}^{r} u_i m_i)(\sum_{j=1}^{c} v_j n_j) \ ,$$

and a variance

$$var(LL(\mathrm{X})) = (N-1)^{-1}\left[\sum_i u_i^2 m_i - \frac{(\sum_i u_i m_i)^2}{N}\right]\left[\sum_j v_j^2 n_j - \frac{(\sum_j v_j n_j)^2}{N}\right] .$$

The linear-by-linear association test is known in SAS as the Cochran-Mantel-Haenszel test. It is well documented in Agresti, Mehta, and Patel (*JASA*, 85, 453-458, 1990).

Step 5: We finally compute the p-value

$$p = \sum_{d(X) \geq d(X^*)} h(X) \ ,$$

which is equivalent to

$$p = \Pr(d(X) \geq d(X^*)) \ .$$

2.3 Large Sample Theory

All the results given here are fully documented in Chapter 5 of the StatXact user manual.

- The discrepancy measures, $d_1(X), d_2(X)$, and $d_3(X)$ have a limiting chi-squared distribution on $(r-1) \times (c-1)$ degrees of freedom.
- The discrepancy measure $d_4(X)$ has a limiting chi-squared distribution on $(r-1)$ degrees of freedom.
- The discrepancy measure $d_5(X)$ has a limiting chi-squared distribution on 1 degree of freedom.

2.4 Results

The exact and asymptotic p-values for the observed 5×3 contingency table are tabulated below for each discrepancy measure.

Test	Asymptotic	Exact
Fisher	0.2206	0.0987
Pearson	0.0815	0.0691
Likelihood Ratio	0.0443	0.0907
Kruskal-Wallis	0.0695	0.0390
Linear-by-Linear	0.0131	0.0118

Here are some interesting observations about these results.

- There are noticeable differences between the exact and asymptotic p-values.
- There is a consistent pattern in the exact results whereby the p-values get progressively smaller as we start to take advantage of the natural ordering of the rows and columns. Thus the exact p-values are largest if we treat the contingency table as completely unordered, and apply the Fisher, Pearson or Likelihood Ratio tests. If we treat the table as ordered along the columns and apply the Kruskal-Wallis test, the exact p-value is reduced. Finally if we treat the table as doubly ordered and apply the Linear-by-Linear test, the exact p-value is smallest.

Accessing Exact Methods in StatXact® with the SAS® System

StatXact Software, from Cytel Software Corporation, is a statistical package for exact nonparametric inference. It provides exact and monte carlo methods for sparse, heavily tied or skewed data which make results based on large sample theory questionable. There are relatively few exact methods in SAS Software. However, Fisher's exact test for RxC tables and exact p-values for Wilcoxon's signed rank statistic with small N are available.

Because of the need for exact methods, especially in sparse- and small-sample situations, a link between the SAS System and StatXact is needed. There are three key steps in making a such a link between the two:

1) Making SAS data available to StatXact
2) Accessing StatXact from the SAS System
3) Making StatXact results available to the SAS System for report generation or other needs.

The SAS commands presented below are discussed in more detail in *SAS Companion for the Microsoft Windows Environment, Version 6, First Edition*. Details of StatXact commands can be found in the *StatXact User Manual*, especially Appendix F on importing data from other packages.

Making SAS data available to StatXact

Data stored permanently or temporarily in SAS data sets can easily be made available to StatXact by simply creating a permanent SAS transport file which can be read by StatXact. This is done with a LIBNAME statement and a DATA step. For example, if you have a temporary SAS data set named TOX, you can create a permanent SAS transport file on your PC's hard drive with the following statements:

```
libname out xport 'c:\sas2sxt.xpt';
data out.sas2sxt;
  set tox;
  run;
```

It is important to use the XPORT engine name on the LIBNAME statement in Release 6.08. If you are using Release 6.04, use the SASV5XPT engine instead. You may choose any name for the PC file (c:\sas2sxt.xpt, above), libname (out, above) and SAS transport file (out.sas2sxt, above).

StatXact can read SAS transport files. Simply use the SAS option on StatXact's IMport command and specify the SAS variables that represent the row and column variables with the RO= and CO= options. If you have cell-count data, specify the SAS variable containing the cell counts in the WE= option. For example,

```
IM SAS C:\SAS2SXT.XPT RO=DRUG CO=TOXICITY WE=WEIGHT
```

Accessing StatXact from the SAS System

You can invoke StatXact while using the SAS System either interactively or in a batch program. In Release 6.08 of the SAS System under Microsoft® Windows, StatXact can be interactively invoked from

```
options nodate nonumber ls=77 ps=40;
title 'The SAS System';

data tox;
  input
  drug    toxicity    weight; cards;
   1          1          2
   2          1          1
   2          2          1
   3          1          3
   4          1          2
   4          2          2
   5          1          1
   5          2          1
   5          3          4
;

options noxwait;

libname out xport 'c:\sas2sxt.xpt';
data out.sas2sxt;
  set tox;
  run;

data _null_;
  file 'c:\sas2sxt.cmd';
  put 'RE C:\SXT.LOG';

    **** Modify RO, CO and WE as needed ****;
    put 'IM SAS C:\SAS2SXT.XPT RO=DRUG CO=TOXICITY WE=WEIGHT';
    **** Give StatXact commands ****;
    put 'FI EX';
    put 'CH EX';
    put 'LI EX';
    put 'KW EX';
    put 'LL EX';

  put 'QU';
  run;

x 'sxt c:\sas2sxt.cmd';
x 'c:\dos\command.com /c convert c:\sxt.log>c:\sxt.out';

proc format;
  value $testfmt
        'FI_AA'-'FI_ZZ'='Fisher''s'
        'CH_AA'-'CH_ZZ'='Pearson'
        'LI_AA'-'LI_ZZ'='Likelihood Ratio'
        'KW_AA'-'KW_ZZ'='Kruskal-Wallis'
        'LL_AA'-'LL_ZZ'='Linear by Linear' ;
  run;

data sxtout;
  infile 'c:\sxt.out';
  input filename $ test $ item $ value;
  if item='AsmPval' then item ='AsmPval2';
  if item='XctPval' then item ='XctPval2';
  if test='LL_EX' and item='Stat' then delete;
  if test='LL_EX' and item='Std_stat' then do;
     item='Stat'; value=value**2; output;
     item='DF'; value=1;
  end;
  output;
  run;
```

SAS Release 6.04 under DOS

1) You will probably need *expanded* (not *extended*) memory. Once installed, tell the SAS System to make use of it by added this line to your CONFIG.SAS file:

```
-EMS ALL
```

2) For proper execution of the X command in batch program, you should have the following line in your CONFIG.SYS file:

```
DEVICE=C:\DOS\ANSI.SYS
```

3) In the batch program shown above, omit the VAR and LABEL statements, or list only those variable names that actually exist in the data set.