EGG-EA-5726

February 1982

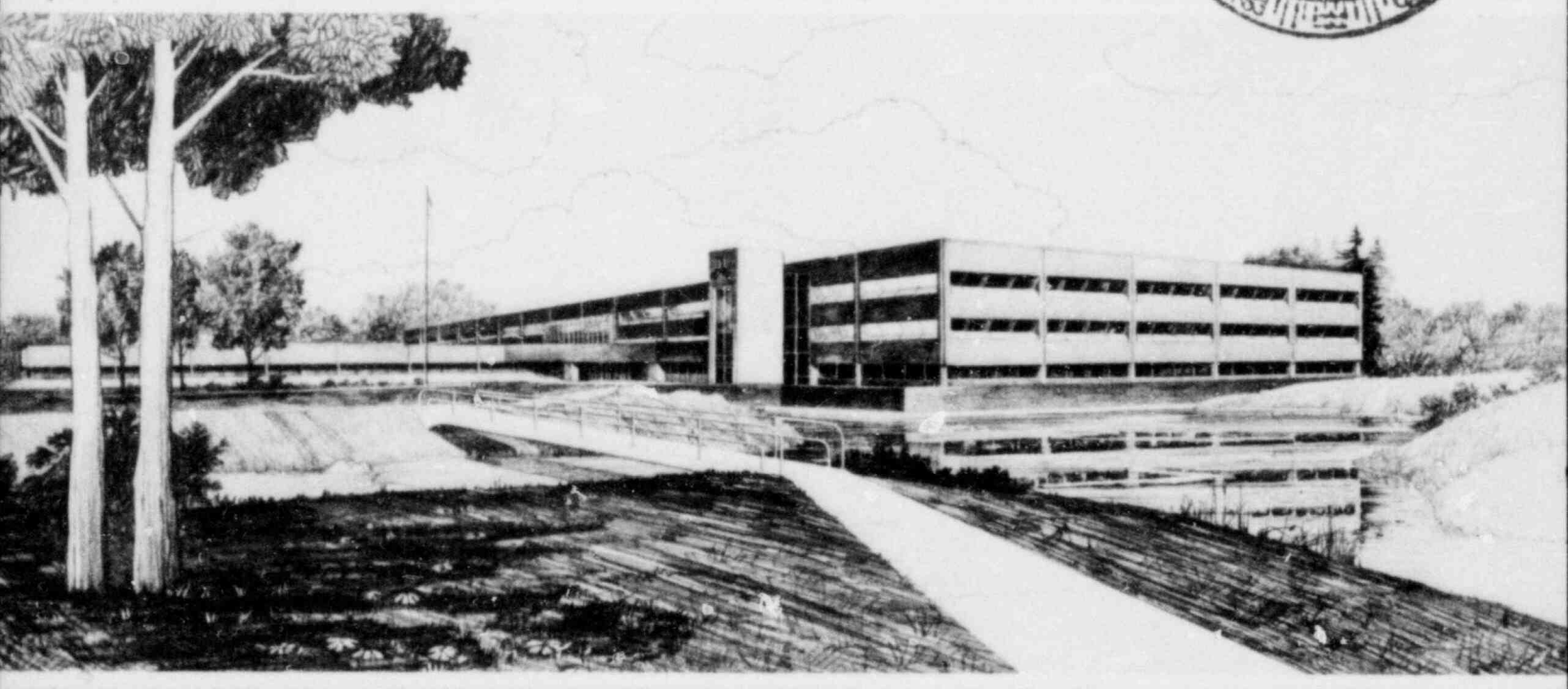
POR

USER'S GUIDE TO HOMOG: A COMPUTER PROGRAM FOR INVESTIGATING HOMOGENEITY OF POISSON DATA SOURCES

NRC Research and/or Technical Assistance Report

Corwin L. Atwood

U.S. Department of Energy
Idaho Operations Office • Idaho National Engineering Laboratory



This is an informal report intended for use as a preliminary or working document

Prepared for the
U.S. Nuclear Regulatory Commission
Under DOE Contract No. DE-AC07-76ID01570
FIN No. A6283

EG&G Idaho

EG&G Idaho, Inc.

FORM EG&G 398
(Rev. 11-81)

INTERIM REPORT

Accession No. _____

Report No. EGG-EA-5726

Contract Program or Project Title:

Common Cause Data Analysis

Subject of this Document:

User's Guide to HOMOG: A Computer Program for Investigating Homogeneity of Poisson Data Sources

Type of Document:

Informal Report

Author(s):

Corwin L. Atwood

Date of Document:

February 1982

Responsible NRC/DOE Individual and NRC/DOE Office or Division:

Leslie E. Lancaster, Division of Risk Analysis

This document was prepared primarily for preliminary or internal use. It has not received full review and approval. Since there may be substantive changes, this document should not be considered final.

EG&G Idaho, Inc.
Idaho Falls, Idaho **83415**

Prepared for the
U.S. Nuclear Regulatory Commission
Washington, D.C.
Under DOE Contract No. **DE-AC07-76ID01570**
NRC FIN No. A6283

INTERIM REPORT

ABSTRACT

Suppose there are various data sources, each corresponding to a count (e.g., number of observed failures) having a Poisson($\lambda_i t_i$) distribution. Here t_i is known and λ_i is unknown. HOMOG is a computer program for investigating the "homogeneity hypothesis" that all the λ_i's are equal. The program prints and plots both a point estimate and a confidence interval for each λ_i. It identifies any outlying data sources, and performs two statistical tests of the homogeneity hypothesis.

SUMMARY

Suppose there are various data sources, each corresponding to a count. For example, the data sources may be plants, and the count for a plant may be the number of recorded failures of a type of pump in a certain time period. It is assumed that for the ith data source, the count is a Poisson($\lambda_i t_i$) random variable, with t_i known and λ_i unknown. HOMOG is a computer program for investigating the "homogeneity hypothesis" that all the λ_i's are equal.

What the program does is explained by showing its output in an example problem. First HOMOG prints and plots both a point estimate and a confidence interval for each λ_i, and for the overall average of the λ_i's. Then it identifies any outlying λ_i's, i.e., those that appear to be so far from the overall average that it is difficult to attribute the difference merely to randomness in the data. Finally it performs two kinds of statistical tests of the homogeneity hypothesis. The first is based on the most extreme outlier, while the second is the Pearson chi-squared test (without the usual requirement of a large sample size.)

There are two types of input to the program. One consists of data, such as failure counts and exposure times. The other consists of parameter values, which control the form of the plots, the accuracy of any approximations, etc. Both types are explained in detail in this report.

Finally, this user's guide contains two example jobs showing how to run HOMOG on the INEL CDC computer. Some details of the programming are mentioned.

Reference 1, giving the mathematical basis for HOMOG, is reprinted as an appendix.

CONTENTS

USER'S GUIDE TO HOMOG: A COMPUTER PROGRAM FOR INVESTIGATING HOMOGENEITY OF POISSON DATA SOURCES

INTRODUCTION

Overview

Suppose that there are different sources of failure counts. For example, there might be five plants reporting failures of a certain kind of pump. HOMOG is a computer program for studying whether the plants all have the same failure rate, i.e., whether there is plant-to-plant homogeneity. The program: (a) calculates and plots a point estimate and a confidence interval for the failure rate of each plant, and for the overall average failure rate; (b) identifies any outlying plants, i.e. plants whose estimated failure rates are so different from the overall rate that it is difficult to attribute the difference merely to randomness of the data; (c) performs statistical tests of the hypothesis that all the plants have the same failure rate.

Of course, the data sources do not have to be plants. They can be systems, vendors, manufacturers, individual components, time periods, or any other sources of data. And "failure" can be defined in any way desired, e.g. as failure to start, or violation of a technical specification. The mathematics of HOMOG uses only the numbers, not their interpretation. What we, for convenience, call failures could really even be successes.

Statistical Assumption

There is one statistical assumption on which HOMOG rests: the failure counts for the data sources are independent Poisson random variables. Independent means that what happens at one data source does not influence

what happens at another source. Poisson means that, in a time interval of length t, the probability of exactly n failures is $e^{-\lambda t} (\lambda t)^n/(n!)$, for some parameter λ, called the failure rate.[a]

---

a. The following conditions give rise to a Poisson random variable. In any small time interval of length dt, the probability of two or more failures is negligible, and the probability of a single failure is approximately $\lambda \cdot dt$. The failure rate, λ, is required to be constant. It does not depend on which time period is considered, or on the number of failures that may have occurred in any other time period. It can also be shown that the failure count has a Poisson(λt) distribution if and only if the time from one failure to the next has an exponential distribution with mean $1/\lambda$.

PROGRAM OUTPUT

Example

This section explains what HOMOG does, using the following example for illustration. There are five plants, with the exposure times and failure counts shown here.

Plant	Exposure	Failures
Plant A	3000 hr	6
Plant B	1000 hr	2
Plant C	7000 hr	1
Plant D	2000 hr	0
Plant E	2000 hr	3

The failures could be those of individual pumps. The exposure time for Plant A then could arise because that plant has three pumps, which each operated for 1000 hours. The same analysis would result if the failures are of a system, which has run for the exposure time shown.

The output from a HOMOG analysis of this data is shown in Figures 1a, 1b, and 2.

In the discussion below, the data sources will be called cells when a general situation is being considered. They will be called plants when the particular example is being discussed.

Analysis of Individual Cells

Let λ_i be the unknown failure rate corresponding to the ith cell. If n_i failures are observed in time t_i for this cell, then λ_i can be estimated by n_i/t_i. This is the maximum likelihood estimate, or MLE. For each cell, HOMOG calculates this estimate of λ_i, and also a

```
PROBLEM    1
EXAMPLE PROBLEM
PLANTS A,B,C,D,E  --  15000 HOURS, 12 FAILURES
```

CELL		EXPOSURE	RELATIVE EXPOSURE	OBSERVED COUNT	SIGNIFICANCE LEVELS LEFT	RIGHT	2-SIDED
1	PLANT A	3000.00	.20000	6	.9961	.0194*	.0194*
2	PLANT B	1000.00	.06667	2	.9586	.1885	.1885
3	PLANT C	7000.00	.46667	1	.0061**	.9995	.0077**
4	PLANT D	2000.00	.13333	0	.1796	1.0000	.2441
5	PLANT E	2000.00	.13333	3	.9354	.2084	.3880
	TOTAL	15000.00	1.00000	12			

CELL		RATES LOWER LIMIT	MLE	UPPER LIMIT
1	PLANT A	.67062E-03	.20000E-02	.39485E-02
2	PLANT B	.35534E-03	.20000E-02	.62980E-02
3	PLANT C	.73276E-05	.14286E-03	.67797E-03
4	PLANT D	0.	0.	.14979E-02
5	PLANT E	.40863E-03	.15000E-02	.38779E-02
	TOTAL	.46150E-03	.80000E-03	.12965E-02

```
*'S CORRESPOND TO SMALL SIGNIFICANCE LEVELS, AS FOLLOWS:
*        (SIGNIFICANCE LEVEL)*(NUMBER OF CELLS) .LE. .1
**       (SIGNIFICANCE LEVEL)*(NUMBER OF CELLS) .LE. .05
***      (SIGNIFICANCE LEVEL)*(NUMBER OF CELLS) .LE. .025
****     (SIGNIFICANCE LEVEL)*(NUMBER OF CELLS) .LE. .010
*****    (SIGNIFICANCE LEVEL)*(NUMBER OF CELLS) .LE. .005
******   (SIGNIFICANCE LEVEL)*(NUMBER OF CELLS) .LE. .0025
```

Figure 1a. Output from first example problem.

```
TESTING OVERALL HOMOGENEITY BASED ON OUTLIERS --
UPPER BOUNDS ON THE ATTAINED SIGNIFICANCE LEVELS, AT WHICH THE HOMOGENEITY HYPOTHESIS WOULD BARELY BE REJECTED,
ARE AS FOLLOWS :

    THE (ONE-SIDED) TEST BASED ON THE MOST SIGNIFICANT LARGE OUTLIER
    REJECTS HOMOGENEITY AT SIGNIFICANCE LEVEL .LE.    .09703 *

    THE (ONE-SIDED) TEST BASED ON THE MOST SIGNIFICANT SMALL OUTLIER
    REJECTS HOMOGENEITY AT SIGNIFICANCE LEVEL .LE.    .03045 **

    THE (TWO-SIDED) TEST BASED ON THE MOST SIGNIFICANT OUTLIER
    REJECTS HOMOGENEITY AT SIGNIFICANCE LEVEL .LE.    .03830 **

TESTING OVERALL HOMOGENEITY BASED ON PEARSON CHI-SQUARED STATISTIC --

    OBSERVED CHI-SQUARED TEST STATISTIC =    .138036E+02
    PROBABILITY (CHI-SQ STATISTIC .GE. OBSERVED)  =  .01382 ***

    THIS CALCULATION IS APPROXIMATE.  THE EXACT PROBABILITY SATISFIES
     .01244  <  PROBABILITY ( CHI-SQ STATISTIC .GE. OBSERVED )  <  .01388 ***

IN ABOVE TESTS OF OVERALL HOMOGENEITY,
*'S CORRESPOND TO SMALL SIGNIFICANCE LEVELS, AS FOLLOWS:
*        SIGNIFICANCE LEVEL .LE. .1
**       SIGNIFICANCE LEVEL .LE. .05
***      SIGNIFICANCE LEVEL .LE. .025
****     SIGNIFICANCE LEVEL .LE. .010
*****    SIGNIFICANCE LEVEL .LE. .005
******   SIGNIFICANCE LEVEL .LE. .0025
```

Figure 1b. Output (continued) from first example problem.

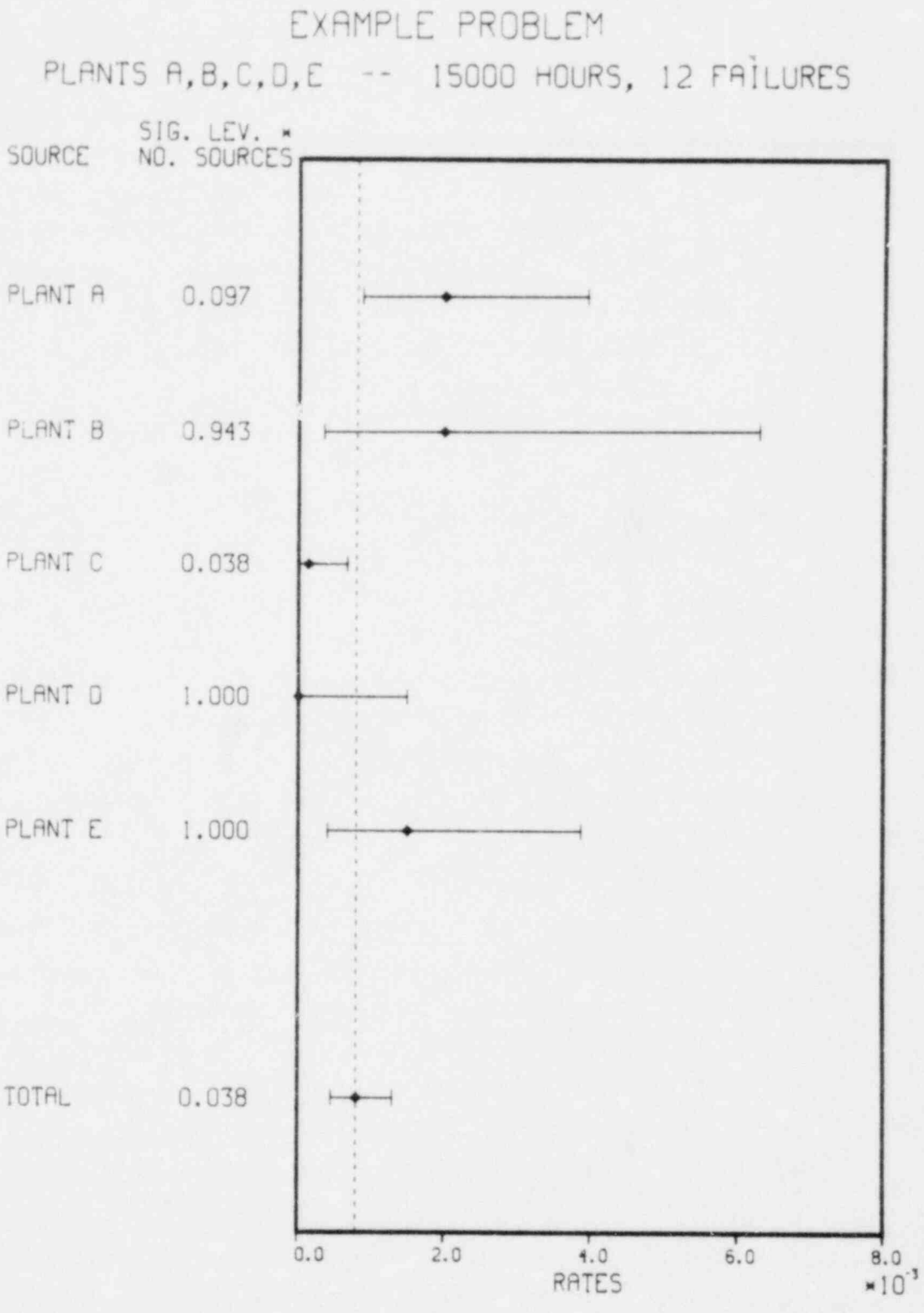


Figure 2. Plot from first example problem.

confidence interval for λ_i. It prints the point estimates and confidence intervals, and plots them so that they can be visually compared to each other. The homogeneity hypothesis is the hypothesis that all the λ_i's are equal to some common value λ. If the homogeneity hypothesis is true, then λ can be estimated by $\sum n_i / \sum t_i$, and a confidence interval for λ can be found. If the homogeneity hypothesis is false, then $\sum n_i / \sum t_i$ estimates the average of the λ_i's (a weighted average if the t_i's are unequal). This average may or may not be of interest to the user. The corresponding confidence interval is for the average of the λ_i's, and does not necessarily indicate the value of any particular λ_i. HOMOG prints and plots this point estimate and confidence interval, based on all the data combined, and lets the user do the interpretation.

Figure 1a shows the estimates and 90% confidence intervals for the failure rates of Plant A through Plant E, on the far right of the figure. Figure 2 is a plot of these estimates and confidence intervals. Also shown in both figures are an estimate and an interval labeled "Total", based on the total exposure time and total number of failures for all the plants. The interval is short, because it is based on all the data. It corresponds to the weighted average of the λ_i's, but not to any individual λ_i unless the homogeneity hypothesis is true.

In the plot of Figure 2, the dashed vertical line goes through the estimated average rate, and is printed to help in visual comparisons of the estimated plant rates with the estimated overall rate.

Note that Plants A and B have the same estimated failure rate, 2/1000. But because the estimate for Plant A is based on more hours (3000 instead of 1000), the confidence interval for Plant A is the shorter of the two.

Note also that the intervals are not symmetrical about the MLE's. This is because the Poisson distribution is skewed.

Outlying Cells

If the homogeneity hypothesis is true, then we would expect the number of observed failures for each cell to be approximately proportional to the exposure time. For example, Plant A has 3000 hours exposure time, and all the plants together have 15000 hours. The relative exposure of Plant A is 3000/15000 = 0.2, so we would expect the number of failures at Plant A to be not too far from 20% of the total, or 0.2 x 12 = 2.4.

An outlying cell (or outlier) is a cell whose observed failure count is far from the expected number. An outlier causes us to question or reject the homogeneity hypothesis. How strongly do we question or reject it? This is discussed informally here. A mathematical treatment appears in Atwood,[1] which is reprinted as an appendix to this user's guide.

The distance between the observed count and the expected count, and the corresponding strength with which we reject the homogeneity hypothesis, are measured by a significance level. A significance level is the probability of getting data as extreme as what we observed. The more unlikely the data, the smaller the signifcance level, and the more strongly we reject the hypothesis. In the present context, extreme data means cell counts that are far from the expected counts. We may be interested only in cell counts that are too high, or only in those that are too low, or in both. This leads to three possible significance levels. The definitions are given here in terms of our example. They are given formally in Reference 1.

In our example, Plant A has relative exposure 0.2, and 6 of the 12 failures. The right significance level for Plant A is the probability that out of some hypothetical 12 failures, 6 or more would occur at Plant A. This probability is calculated assuming that the homogeneity hypothesis is true, i.e., that on the average, 20% of all failures would occur at Plant A. It is printed in Figure 1a as 0.0194. Similarly, the left significance level for Plant A is the probability of 6 or fewer failures at

Plant A, out of 12 in all. The two-sided significance level is defined precisely in Reference 1. Roughly, it is the probability that, out of some hypothetical 12 failures, the number at Plant A would be as unlikely, on either the high side or the low side, as the 6 that actually occurred.

The two-sided level is always at least as large as the smaller of the right level and the left level. Which significance level to use depends on whether we are concerned about departures from homogeneity on the high side, the low side, or both.

A small significance level means that the observed data are unlikely under the homogeneity hypothesis, so there is evidence for rejecting the homogeneity hypothesis. The strength of the evidence is measured by the smallness of the significance level.

There might be some cell of special interest. For example, Plant B may be paying for the study, or there may be reasons, such as design or past history, for suspecting that Plant B is unusual. In that case, a significance level corresponding to Plant B would be of interest.

More commonly, however, all the cells are of equal interest, and an overall test of homogeneity is desired. In that case, we cannot simply look over the list of significance levels for the cells and use the smallest one to test homogeneity. This is because, if there are many cells, then even if all the λ_i's are equal, there will be enough random scatter in the data so that some of the many cells will have small significance levels. It is not easy to calculate the exact significance level for the entire data set. However, a simple upper bound is the smallest significance level for a cell multiplied by the number of cells. For a proof, see Reference 1. Unless there are very few cells, this upper bound is usually quite close to the exact overall significance level.

In Figure 1a, if a significance level multiplied by the number of plants is small, then that significance level is marked by one or more stars. The number of stars is explained in Figure 1a. For each plant,

Figure 2 also shows the two-sided significance level multiplied by the number of plants. If this number happens to be greater than 1, it is printed as 1.000. Notice that small significance levels correspond to confidence intervals that are not close to the estimated average λ.

Figure 1b gives the upper bounds on the significance levels for the data set as a whole. For example, the smallest two-sided significance level for any cell is 0.0077. The corresponding significance level for the data set as a whole is 5 x 0.0077 = 0.03830. (The discrepancy in the arithmetic is due to round-off error.) This is small enough to get two stars. The number of stars assigned is explained at the bottom of Figure 1b. Interpret this number by thinking, "There is only a chance of 0.038 of getting an outlier as extreme as what we have. Therefore our data give fairly strong evidence for rejecting the homogeneity hypothesis."

Figure 2 shows this upper bound on the overall two-sided significance level next to the interval labeled "Total".

The Pearson Chi-Squared Test

A second test of homogeneity may be performed, based on the Pearson chi-squared statistic, defined as

$$\sum(O_i-E_i)^2/E_i.$$

Here O_i is the observed number of failures for the ith cell, and E_i is the expected number. If the observed counts differ greatly from the expected counts, the chi-squared statistic will be large. The significance level for this test is the probability that, in some hypothetical data set with the same total number of failures as actually occurred, the chi-squared statistic would be greater than or equal to the value that was calculated from the actual data. Just as for the other tests, a small significance level means that the data give evidence for rejecting the homogeneity hypothesis.

HOMOG calculates this significance level, either exactly or approximately. If an approximation is given, the upper and lower bounds for the significance level are also given. The algorithm for calculating the significance level was developed especially for HOMOG, and is described in Reference 1. This algorithm does not require a large sample size.

Figure 1b shows the significance level of the chi-squared test to be approximately 0.01382, small enough to be marked by three stars. the exact significance level is between 0.01244 and 0.01388. These bounds are based on a generalization of the Chebyshev inequality, and so are valid but usually very conservative. In this example they are close enough to each other so that greater precision is pointless. HOMOG never puts stars by the lower limit, but it does mark the upper limit with stars if the limit is small enough, as it is in this example. The number of stars assigned is explained in Figure 1b.

In this example, the two-sided outlier test and the chi-squared test agree only to some extent. The outlier test says that, under the homogeneity hypothesis, the probability of observing data as extreme as ours is rather small, about four out of a hundred. The chi-squared test says that the probability is three times smaller, only about one in a hundred. The disagreement is because the tests use different definitions of extreme data. The outlier test rejects homogeneity if a single cell count differs greatly from the expected count. The chi-squared test rejects homogeneity if a weighted sum of the squared differences is large.

PROGRAM INPUT

There are two types of input. One consists of the data, such as failure counts and exposure times. The other consists of values of parameters, which control the form of the plots, the accuracy of any approximations, etc. The data must always be entered; the parameters, on the other hand, all have default values, so need not be changed by the user. The two types of input are described here.

The Data

The data are entered on card images. All the numbers are in free format, i.e. numbers in any columns, separated by blanks and/or commas. Do not use sequence numbers on the cards; they will be read as data!

Card 1. Title, up to 80 characters. The full title will appear on the print-out. However only about 40 characters will fit on the plot, with the exact number depending on the character widths. The rest are truncated.

Card 2. Subtitle, up to 80 characters. Only about 50 characters will fit on the plot.

Card 3. Three numbers, referred to as NCELLS, NEWNAMS, and DIVISOR. NCELLS is the number of cells in this problem, a positive integer. NEWNAMS is an integer, either zero or non-zero. If NEWNAMS is non-zero, then HOMOG expects to read names for the cells. If NEWNAMS is zero, then HOMOG uses default names, defined as follows. If HOMOG just finished another problem as part of this job, and if NCELLS was the same in that problem as in the present problem, then the default names are whatever names were used in the preceding problem. Otherwise, the default names are blank.

The third number on this card, DIVISOR, is a units normalizer. It is useful if the exposure times are to be entered in one set of units but then transformed to another set of units. Every exposure is divided by DIVISOR before any other calculations are done. For example, if the exposures are

entered as hours, and DIVISOR=1000, all the exposures will be divided by 1000, and the rates will have units "events per thousand hours." Even if DIVISOR equals 1, it must be entered. It may be entered as an integer or as a floating point number.

Card 4. Exposure times. There must be NCELLS numbers. If there is room, they may all go on one card. Otherwise, the card may be continued for as many cards as necessary. The numbers may be entered as integers or as floating point numbers.

Card 5. Failure counts. There must be NCELLS integers, on one or more cards.

Card 6...If NEWNAMS was entered as non-zero, the names for the cells must be given here. Each name may have up to ten characters, with only one name to a card, entered in the first ten columns. Therefore, if the names are being entered, there must be NCELLS cards here.

This concludes the data input for a problem. If desired, the sequence may be repeated, as many times as there are problems to do. HOMOG stops when there is no more input to read.

Sometimes unusual spacing is desirable on a plot, with blank lines separating groups of cells, and perhaps headings for the groups. A blank line can be created by entering a negative number for both the exposure time and the corresponding failure count. The "name" corresponding to this cell can be blank, or it can be a heading of up to ten characters. Figure 3 was produced in this way. NCELLS was entered as 8. The exposure time and failure count were entered as -1 for cells 1, 4 and 8. The names for those three cells were entered as OLD PLANTS, NEW PLANTS, and blank. The plant names were entered with two leading blanks, so that they would be indented on the plot. Although NCELLS was entered as 8, the program recognizes that there are only 5 true cells for its calculations of significance levels.

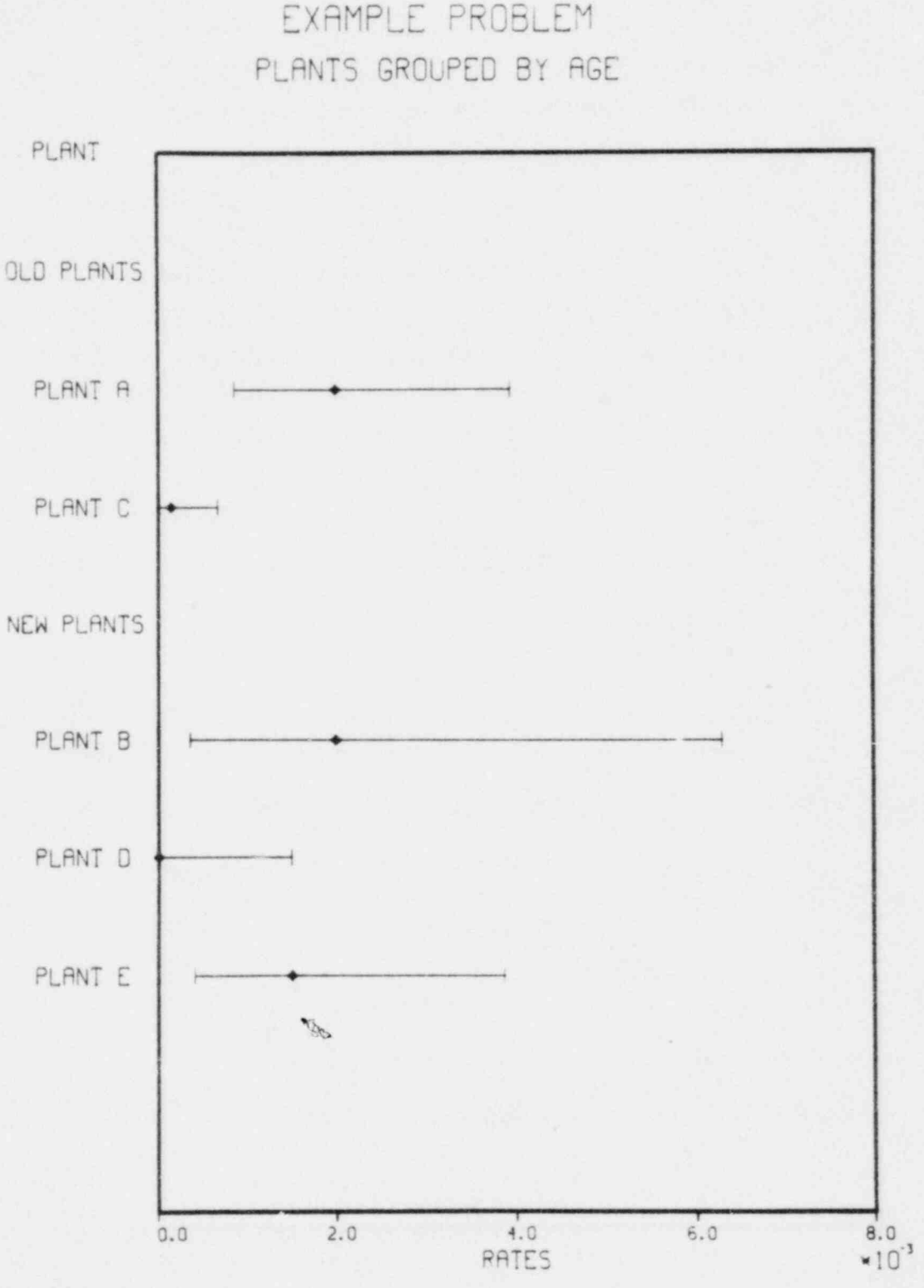


Figure 3. Plot from second example problem.

Parameters

The parameter values are entered as part of the control language statement that calls for execution of HOMOG. For example,

```
HOMOG, NC = 20, GR = 0
```

would define the values of the parameters NC and GR for this HOMOG job. Any parameters that are not defined by the user take their default values. There are four groups of parameters: (a) basic parameters for the computations, (b) parameters affecting the plots, (c) parameters affecting the calculation of the chi-squared test, and (d) parameters for job control. The parameters are defined below in this order.

Basic Parameters for the Computations

Parameter	Default	Meaning
NC	200	NC is an upper bound on the number of cells. It is used as a dimension for arrays in the FORTRAN program.
CONF	90	CONF is the coefficent of the confidence intervals for the failure rates. If CONF = 90, 90% intervals are found, with 5% probability in each tail.

Parameters Affecting the Plots

Parameter	Default	Meaning
GR	1	If GR=0, no graphics plots are produced. If GR=1 or 2, plots are produced, showing the cell names and confidence intervals for the failure rates. If GR=1, for each cell the plot prints the cell's two-sided significance level multiplied by the number of cells. If GR=2, the plot is produced, but tests of homogeneity are regarded as unimportant. Therefore, no significance levels appear on the plot, and the chi-squared test is not performed.
T	1	The interval labeled "Total" is shown on the plot if T≠0. It is omitted if T=0.

Parameter	Default	Meaning
VL	1	A dashed vertical line through the overall estimated failure rate is shown if VL≠0. It is omitted if VL=0.
STR	1	If STR=0, the plot is scaled so that all the confidence intervals fit completely in the border. A cell with an extremely short exposure time will have a very long interval, even if it has no observed failures. Such a noninformative cell can dominate the scaling, making the other intervals hard to compare because they appear so short. If STR≠0, cells with no observed failures are ignored in the scaling. This stretches the confidence intervals, and usually makes the important parts of the plot easier to see.
CL	$SOURCE$	CL is the cell label, the heading for the cell names on the plot. It may have up to ten characters, including blanks, and must be delimited by dollar signs.
CLS	$SOURCES$	CLS is the plural of whatever is used for CL. It appears in the plot in the heading for the significance level multiplied by NCELLS, e.g. in Figure 2. It may have up to seven characters, including blanks, but should have no leading blanks. It must be delimited by dollar signs. The spacing is best if it has exactly seven characters.
NL	25	NL is the maximum number of lines printed per page. If there are more than NL lines, the plot is printed on more than one page, with approximately the same number of lines per page. Note, if T≠0, the total number of lines to be printed is NCELLS+2. If T=0, the total number of lines is NCELLS.

Parameters Affecting Calculation of the Chi-Squared Test

Parameter	Default	Meaning
BIGA	20	The name BIGA is derived from the commonly used notation of α for a significance level. HOMOG stops working on the distribution of the chi-squared statistic if it becomes clear that the exact significance level is greater than BIGA per cent.

Parameter	Default	Meaning
LIM	5000	LIM is the upper limit on the number of possible ways to be considered that the total failure count can be distributed among the cells. If LIM ways have been considered, then HOMOG stops trying to find the significance level of the chi-squared test, and reports upper and lower bounds based on the work done so far.
NP	5	The distribution of the chi-squared statistic is found by decomposing it into pieces. A gamma distribution is used to approximate those pieces for which the expected count for each cell in question is at least NP. See Step 7 of the algorithm in Reference 1 for a fuller explanation.
DEL	25	If HOMOG does not find the exact significance level, then it prints an approximation, and upper and lower bounds on the exact value. DEL is used to set a target for how far apart the upper and lower bounds should be. Usually the final upper and lower bounds are between (1-DEL/100) and (1+DEL/100) times the calculated approximation. A small value of DEL will result in tight bounds, at the cost of possible lengthy calculation.

Parameters for Job Control

Parameter	Default	Meaning
SEE	1	SEE controls how much of the FORTRAN program is printed. If SEE=0, no program listing is printed. If SEE=1, a listing is printed for the main program, but not for any of the subprograms. This may be useful, because the main program contains extensive comments describing the input, only slightly more concisely than what is in this user's guide. If SEE=2, a listing of the entire program is printed (one main program and 19 subprograms, with about 1900 lines). Setting SEE=2 also causes the load map to be printed.
I	INPUT	I is the local name of the file containing the data input.

Parameter	Default	Meaning
ID	000	When plots are generated, a DISSPLA postprocessor sends them to film. It may also be desirable to catalog the PLFILE, so that it can be inspected on a graphics terminal before the film arrives. If ID is set to a valid user ID, the PLFILE will be cataloged as HOMOGPL, with that ID, and a retention period of 2 days.
PL	5000	PL is the line limit for the printer.
DB	0	To use CYBER Interactive Debug, run HOMOG on a terminal with DB≠0.
MAP	0	The three-page load map is printed if MAP≠0 or SEE=2. It is not printed if MAP=0 and SEE<2.

CONSTRUCTING A HOMOG JOB

Accessing HOMOG at INEL

On the INEL CDC computer, the following job will run HOMOG.

```
Job card
Account card
ATTACH,HOMOG,ID=CLA.
HOMOG<,parameter definitions if desired>.
*EOR
Data input
```

If cards are used instead of a terminal, replace *EOR by 7/8/9 punched in column 1.

Examples

Figure 4 shows the job that produced Figures 1a, 1b, 2. All the defaults were used, except SEE was set to 0 so that no program listing would be printed.

Figure 5 shows the job that produced Figure 3. Setting T=0 caused the confidence interval labeled "Total" not to be printed. Setting VL=0 caused the dashed vertical line not to be printed. Setting CL=$ PLANT$ caused the cell names to be headed by the word "Plant" instead of "Source." The two leading blanks caused the heading to be indented. Setting GR=2 caused the significance levels not to be printed on the plot (so CLS, defined as $PLANTS$, was irrelevant). Several exposure times and failure counts were set negative, resulting in blank lines where confidence intervals would normally be. The cell names were entered as they appear on the plot, including indentations for the plant names and a blank for the last name. The final *EOR is unnecessary, but is shown here to emphasize that there is a blank line after PLANT E.

```
CLAHO,T37,P1,STANY.
ACCOUNT,3830,XXXXXXXXX,TM4.
ATTACH,HOMOG,ID=CLA.
HOMOG,SEE=0.
*EOR
EXAMPLE PROBLEM
PLANTS A,B,C,D,E   --   15000 HOURS, 12 FAILURES
5 1 1
3000   1000   7000   2000   2000
6      2      1      0      3
PLANT A
PLANT B
PLANT C
PLANT D
PLANT E
```

Figure 4. Job to run first example problem.

```
CLAHO,T37,P1,STANY.
ACCOUNT,3830,XXXXXXXXX,TM4.
ATTACH,HOMOG,ID=CLA.
HOMOG,SEE=0,CL=$  PLANT$,CLS=$PLANTS$,T=0,VL=0,GR=2.
*EOR
EXAMPLE PROBLEM
PLANTS GROUPED BY AGE
8 1 1
-1   3000   7000   -1   1000   2000   2000   -1
-1   6      1      -1   2      0      3      -1
OLD PLANTS
  PLANT A
  PLANT C
NEW PLANTS
  PLANT B
  PLANT D
  PLANT E

*EOR
```

Figure 5. Job to run second example problem.

DETAILS OF PROGRAMMING

HOMOG consists of a procedure written in CDC CYBER control language under the NOS/BE system, and a program written in CDC FORTRAN 4 Extended. It uses the library IMSL[2] for some of the computations, and DISSPLA[3] for the plots.

On December 17, 1981, the two jobs shown in Figures 4 and 5 were executed on the INEL CDC 176. Each job took about 2.5 CP seconds, and about 43 system seconds, for a cost of $0.66 each.

REFERENCES

1. C. L. Atwood, "Tests of a Simple Multinomial Hypothesis when the Sample Is Not Large," American Statistical Association 1981 Proceedings of the Statistical Computing Section, 1981. (Reprinted as an appendix of this user's guide.)

2. IMSL Library: Reference Manual, Houston, Texas: IMSL, Inc., 1980.

3. DISSPLA User's Manual, San Diego, California: Integrated Software Systems Corporation, 1978.

APPENDIX

TESTS OF A SIMPLE MULTINOMIAL HYPOTHESIS WHEN THE SAMPLE IS NOT LARGE

As in American Statistical Association 1981 Proceedings of the Statistical Computing Section

TESTS OF A SIMPLE MULTINOMIAL HYPOTHESIS WHEN THE SAMPLE IS NOT LARGE

Corwin L. Atwood, EG&G Idaho, Inc.

INTRODUCTION

This note develops tests, valid when n is not large, of H_0, the hypothesis that $(N_1, \ldots, N_k)$ has a multinomial$(n, p_1, \ldots, p_k)$ distribution.

An example from reliability studies is when N_i is the number of failures of a certain kind observed at plant i in time t_i. Assume that N_i has a Poisson$(\lambda_i t_i)$ distribution. Then, conditional on $\Sigma N_i = n$, the N_i's have a multinomial distribution. To test whether λ_i is the same for all i, we can test H_0, the hypothesis that $(N_1, \ldots, N_k)$ is multinomial$(n, p_1, \ldots, p_k)$ with $p_i = t_i/\Sigma t_i$. If the equipment is reliable, then n will not be large, so tests of H_0 cannot use simple asymptotic approximations.

Two kinds of tests are considered: tests based on outlying cells, and the Pearson chi-squared test. The main result is a method for closely approximating the significance level of the Pearson chi-squared test when n is not large.

TESTS BASED ON OUTLYING CELLS

Consider only the ith cell. Under H_0, N_i has a binomial(n, p_i) distribution. One-sided tests based on N_i are easy. Define the attained left significance level and right significance level of cell i as

$$P[N_i \leq n_i \mid H_0] \text{ and } P[N_i \geq n_i \mid H_0].$$

A suitably defined two-sided significance level requires more care because of the discreteness of the distribution. The two-sided level should be the size of a two-sided test of H_0 such that the probabilities of the two tail regions are approximately equal. Accordingly, consider the case when the right significance level, $P[N_i \geq n_i \mid H_0]$, is less than 1/2. Let h be the largest integer satisfying

$$P[N_i \leq h \mid H_0] \leq P[N_i \geq n_i \mid H_0].$$

Define the attained two-sided significance level of source i as

$$P[N_i \leq h \mid H_0] + P[N_i \geq n_i \mid H_0].$$

The definition is similar if the left significance level is less than 1/2. If neither the left nor the right significance level is less than 1/2, then define the two-sided level to be 1.

If no cell is a priori of special interest, then an overall test of H_0 can be performed based on the attained significance levels of all the cells. Either the left, right, or two-sided levels can be used. Let α_i denote the attained significance level for cell i. For some number c, reject H_0 if any α_i is less than or equal to c. The significance level of this overall test is

$$\alpha = P[\alpha_i \leq c \text{ for at least one } i].$$

Observe that

$$P[\alpha_i \leq c] \leq c . \qquad (1)$$

If the data were continuous, the probability would equal c exactly.

An upper bound on α is given by the Bonferroni inequality and (1):

$$\alpha \leq \sum P[\alpha_i \leq c] \leq k\, c.$$

So, for any desired nominal value α_0, a conservative test uses $c = \alpha_0/k$. The overall significance level attained by the data is bounded above by $\min [1, \min_i(k\alpha_i)]$.

A lower bound on α is given by

$$\alpha \geq \max_i P[\alpha_i \leq c].$$

This lower bound is sharp, and may be as small as zero. If c is an attainable significance level, then the lower bound equals c. For large sample sizes and one-sided tests, Fuchs and Kenett (1980) obtain a much larger lower bound. They make essential use of the fact that inequality (1) becomes equality as $n \to \infty$.

THE PEARSON CHI-SQUARED TEST

The Pearson chi-squared statistic, denoted here by X^2, is defined as

$$X^2 = \sum_1^k (N_i - np_i)^2/np_i .$$

When it is necessary to indicate the parameters explicitly, X^2 will be written as $X^2(k, n, p_1, \ldots, p_k)$. As $n \to \infty$, it is well known that the distribution of X^2 is asymptotically $\chi^2(k - 1)$. When n is not large, the approximation is inadequate. The distribution may have large jumps, or it may be nearly continuous in places but have the wrong shape, or both. To handle discreteness, direct calculation of the possibilities is necessary. To handle the wrong shape, approximations other than $\chi^2(k - 1)$ can be tried.

Simple Approximations

Approximations can be based on the moments of X^2. The first four moments were published by Haldane (1937). The mean and variance of X^2 are

$$k - 1$$

and

$2(k - 1) + n^{-1} [-k^2 - 2k + 2 + \sum p_i^{-1}]$.

The skewness and kurtosis are more complicated. One approximation of the distribution is a gamma distribution with the mean and variance matching those of X^2. Another approximation is that Pearson distribution with the first four moments matching those of X^2.

These approximations are shown in an example in Figure 1. The exact distribution is based on five observations, and on ten cells with probabilities 1/1023, 2/1023, 4/1023, ..., 512/1023. The upper portion of the cumulative distribution function (c.d.f.) is shown. In this range, the x^2 distribution woefully overestimates the c.d.f., and so underestimates the significance level. The other two approximations do somewhat better, but none of them satisfactorily matches the bumpiness of the exact distribution. Using four moments is not noticeably better than using only two.

By the way, the variability of the p_i's in this example is not unrealistic. The ratio of the smallest p_i to the largest is 0.002. In reliability studies recently performed at EG&G Idaho, this ratio was often less than 0.01, and sometimes less than 0.001.

Decomposition

To account for the possible bumpiness in the distribution of X^2, decompose the distribution into various cases, conditional on the values of the first several N_i. In particular, for $0 < h < k$, suppose that $N_1, \ldots, N_h$ are fixed, and that

$$\sum_{h+1}^{k} N_i = m, \quad \sum_{h+1}^{k} p_i = c .$$

If $m = 0$ or $h + 1 \geq k$, then the value of X^2 is determined. If, on the other hand, $m > 0$ and $h + 1 < k$, then a little algebra shows that

$$X^2 = \sum_{1}^{h} (N_i - np_i)^2/np_i + (m - nc)^2/nc + (m/nc) \sum_{h+1}^{k} (N_i - mp_i/c)^2/(mp_i/c). \quad (2)$$

The first two terms are constant, conditional on $N_1, \ldots, N_h$, while the summation in the third term is

$$X^2(k - h, m, p_{h+1}/c, \ldots, p_k/c).$$

For short, denote this summation by Y^2. Conditional on $N_1, \ldots, N_h$, the conditional distribution of X^2 either can be approximated as in the preceding section, say by approximating the distribution of Y^2 as a gamma distribution with the first two matching moments, or it can be decomposed further, say by conditioning on N_{h+1}.

To combine these approximations, let E_i symbolically represent an event of the form $N_1 = n_1, \ldots, N_h = n_h$. Let the values of $n_1, \ldots, n_h$ vary, and perhaps let h also vary, to produce mutually exclusive events E_i with $\sum P(E_i) = 1$. Then

$$P(X^2 \leq a) = \sum P(X^2 \leq a \mid E_i) P(E_i).$$

So the approximations of the conditional distributions together yield an approximation of the unconditional distribution.

In several examples, a satisfactory approximation has been obtained by conditioning on those cells with the smallest values of p_i. (Conditioning instead on the cells with the largest p_i seems to give a much less satisfactory approximation.) For the example of Figure 1, a close approximation is given in Figure 2. In this example, there are 2002 possible arrangements of 5 counts in 10 cells, producing 986 distinct possible values of X^2. The approximation of Figure 2 is based on 46 exactly calculated cases (for example, the case with $N_1 = \ldots = N_8 = 0$, $N_9 = 2$, $N_{10} = 3$) and 41 approximate cases (for example, the case with $N_1 = \ldots = N_6 = 0$, $N_7 = 2$, N_8 through N_{10} random). The approximation can be made better or worse by computing more or fewer cases exactly.

Lawal (1980) uses the asymptotic approximation of the decomposition (2), when both $n \to \infty$ and np_i approaches some small value (independent of i) for $i \leq h$. For $\alpha = 0.05$ and 0.01, he presents tables of critical values for selected values of k, h, and the common value $np_1 = \ldots = np_h$. For those whose needs do not justify a computer program using the method of this note, Lawal's paper may be of interest. It does not apply to the example of Figure 1, since too many of the cell expectations are "small" (all but one are less than 2.0). We could try to apply it anyway, treating the first seven expectations as small, acting as if their geometric mean were their common value, and extrapolating from Lawal's tables. Then, the tables would say that the 95% point is approximately 25.3, not too far from the exact value of 27.76, and that the 99% point is approximately 47, far from the exact value of 103.69. I think that manageable tables cannot completely cover the great possible variety of multinomial situations. For some problems, an on-the-spot calculation will be necessary.

Implementing the Decomposition

Suppose now that data have been observed and that we are to find or approximate the attained significance level (i.e., $P(X^2 \geq X_0^2)$, where X_0^2 is the observed value of X^2). The algorithm below considers various grounds for deciding how far to decompose the distribution. The algorithm forms the basis for a computer program now in use at EG&G Idaho. Numbers in square brackets in the description below are the default values now used in the program. Following the algorithm are comments on some of the steps.

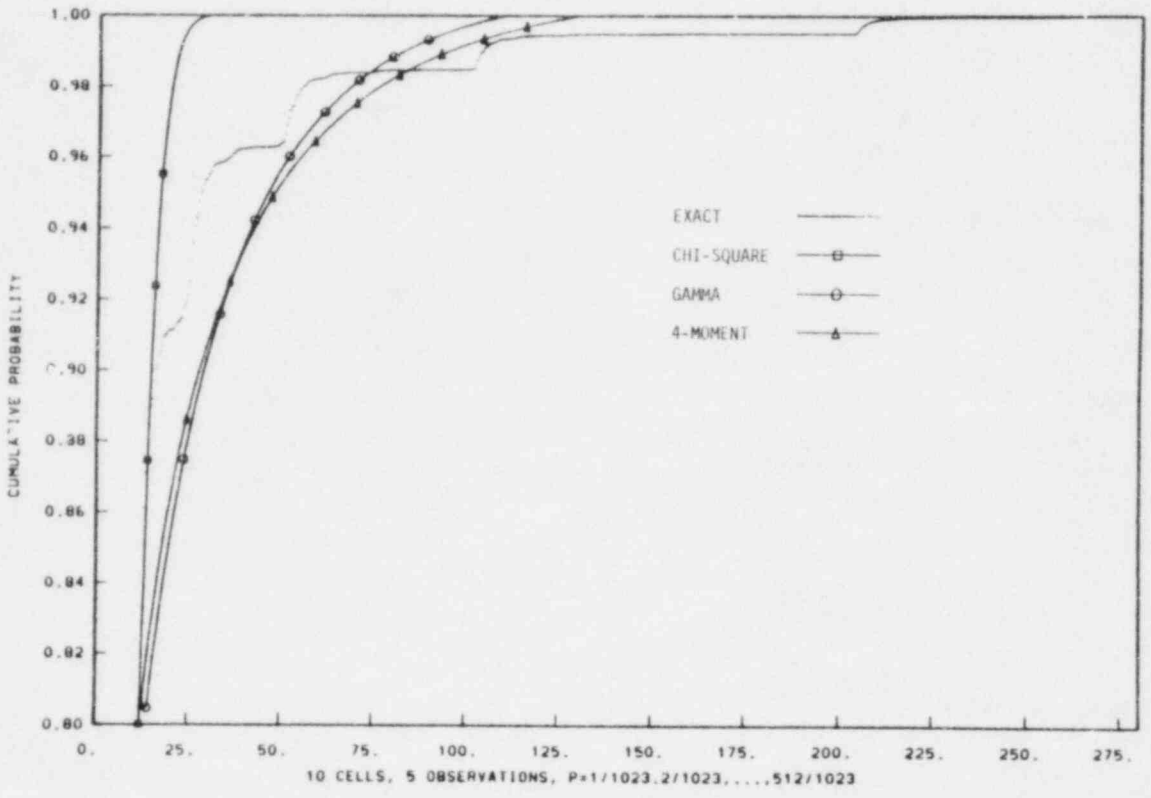


Figure 1. Upper tail of c.d.f. for Pearson chi-square statistic: exact distribution and chi-square, gamma, and four-moment approximations.

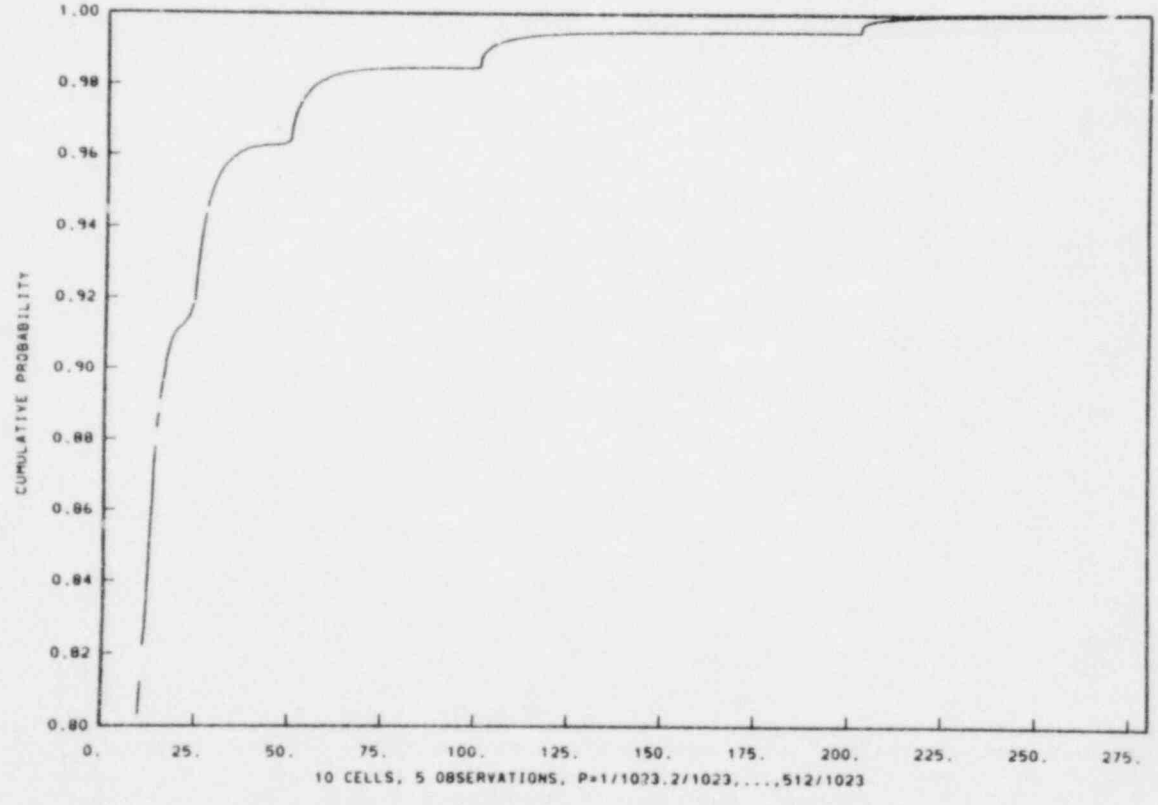


Figure 2. Upper tail of c.d.f. for Pearson chi-square statistic: decomposition approximation.

Algorithm

1. Initialize. Order the cells so that $p_1 \le \ldots \le p_k$. Set $h = 0$, $P_o = 1$, ANSLO = ANS = ANSUP = 0. (Below, if $h = 0$, any sum with index running from 1 to h will be considered zero.)

2. Is the value of X^2 determined? If $h < k - 1$ and $\Sigma_1^h N_i < n$, go to Step 4.

3. The value of X^2 is determined. If $X^2 \ge X_o^2$, set ANSLO = ANSLO + P_o, ANS = ANS + P_o, ANSUP = ANSUP + P_o. Go to Step 11.

4. The value of X^2 is not determined. Let $m = n - \Sigma_1^h N_i$, $c = 1 - \Sigma_1^h p_i$, and let $X^2 = a + (m/nc)\, Y^2$, where Y^2 is defined below (2).

   Let $b = (nc/m)(X_o^2 - a)$, and observe that $X^2 \ge X_o^2$ if and only if $Y^2 \ge b$. From now on, work with Y^2 and b instead of X^2 and X_o^2.

5. Is the probability trivial? Let m_i be the integer closest to mp_i/c, for $i = h + 1, \ldots, k$, and let YMIN = $\Sigma_{h+1}^k (m_i - mp_i/c)^2/(mp_i/c)$. Let YMAX = $m(c - p_{h+1})/p_{h+1}$. Then YMIN $\le Y^2 \le$ YMAX. If YMAX $< b$, go to Step 11. If YMIN $\ge b$, set ANSLO = ANSLO + P_o, ANS = ANS + P_o, ANSUP = ANSUP + P_o, and go to Step 11.

6. Is the probability easy to calculate? If $k > h + 2$, go to Step 7. If $k = h + 2$, then $Y^2 = (N_k - mp_k/c)^2/(p_k\, p_{k-1}/c^2)$, where $N_k \sim$ binomial$(m, p_k/c)$. So calculate $P(Y^2 \ge b)$ exactly, using the binomial distribution. Increase ANSLO, ANS, and ANSUP by $P_o * P(Y^2 \ge b)$, and go to Step 11.

7. At this point, $P(Y^2 \ge b)$ is not easy to find exactly. Steps 7 through 9 consider reasons for deciding whether to approximate the distribution of Y^2 or to decompose it further.

   If m is large [$mp_{h+1}/c \ge 5$], then the gamma approximation is adequate; set FLAG = .TRUE. and go to Step 9. Otherwise, set FLAG = .FALSE.

8. If P_o is large [>1/4], go to Step 12, to be safe.

9. Get upper and lower bounds $P_{LO} \le P(Y^2 \ge b) \le P_{UP}$, based on generalized Chebyshev inequalities. Let P_G be the gamma approximation of $P(Y^2 \ge b)$. If FLAG = .TRUE., go to Step 10. Otherwise, if ANSLO + $P_o * P_{LO}$ is small [$< 0.75*(\text{ANS} + P_o * P_G)$] or ANSUP + $P_o * P_{UP}$ is large [$> 1.25*(\text{ANS} + P_o * P_G)$], go to Step 12.

10. Use the gamma approximation. Set ANSLO = ANSLO + $P_o * P_{LO}$, ANS = ANS + $P_o * P_G$, ANSUP = ANSUP + $P_o * P_{UP}$.

11. Start a new case at the current level of decomposition. Set $N_h = N_h + 1$. If $\Sigma_1^h N_i \le n$, let P_o be the joint probability of $N_1, \ldots, N_h$, and go to Step 2. Otherwise, go to Step 13.

12. Start the next level of decomposition. Set $h = h + 1$, $N_h = 0$. Let P_o be the joint probability of $N_1, \ldots, N_h$. Go to Step 2.

13. Back up one level of decomposition. Set $h = h - 1$. If $h > 0$, go to Step 11. If $h = 0$, then ANSLO $\le P(X^2 \ge X_o^2) \le$ ANSUP, and ANS approximates $P(X^2 \ge X_o^2)$. Print ANSLO, ANS, and ANSUP, and stop.

Comments

Step 7. To save time and avoid microscopic decomposition, we could also set FLAG = .TRUE. if P_o is small [say, $P_o \le 0.01$].

Step 9. Royden (1953) gives generalizations of the Chebyshev inequality for positive random variables with an arbitrary number of known moments. Simple inequalities result from using the mean and variance of Y^2 and the facts that $Y^2 - YMIN \geq 0$ and $YMAX - Y^2 \geq 0$. Use of the first four moments seems to improve the program's execution slightly. Better inequalities on the distribution of Y^2 would enhance the algorithm.

In Step 9, if the condition of FLAG were never used to cause branching to Step 10, then the inequalities on ANSLO and ANSUP would guarantee that, at the end of computation, ANSLO and ANSUP would be close to the calculated value ANS. Using the condition of FLAG speeds up the computation, at the possible cost of an ultimate large difference between ANSLO and ANSUP.

I think that it is advantageous to use FLAG when either m is large or P_0 is small. Even if the resulting spread from ANSLO to ANSUP is large, the values usually tell the user all he needs to know, for a very low computation cost. If, after looking at the output, the user does want a better approximation, he can tighten the parameters in the program and rerun the problem. (The program should, of course, be written so that the relevant parameters are accessible to the user.)

Step 10. If the gamma approximation is good (e.g., if m is very large), then Chebyshev-type inequalities are very conservative. It would then be more realistic to increase ANSLO, ANS, and ANSUP all by the same quantity, $P_0 * P_G$. The cost of this realism is loss of the mathematical certainty that ANSLO and ANSUP bracket the true significance level.

Steps 11-13. If some of the cells have equal probabilities, then the enumeration of cases can be made more efficient, as follows. Group the cells into blocks, with a block consisting of all those cells having a particular probability. Rearranging the counts within a block does not change the value of X^2. Therefore, all such rearrangements can be treated at once by using, say, the arrangement with N_i nonincreasing within each block, and multiplying P_0 by the appropriate factor. The details complicate the algorithm, and are left to the reader.

REFERENCES

1. C. Fuchs and R. Kenett, "A Test for Detecting Outlying Cells in the Multinomial Distribution and Two-way Contingency Tables," J. Am. Stat. Assoc., 75, 1980, pp. 395-398.

2. J. B. S. Haldane, "The Exact Value of the Moments of the Distribution of X^2, Used as a Test of Goodness of Fit, When Expectations are Small," Biometrika, 29, 1937, pp. 133-143.

3. H. B. Lawal, "Tables of Percentage Points of Pearson's Goodness-of-Fit Statistic for Use with Small Expectations," Applied Stat., 29, 1980, pp. 292-298.

4. H. L. Royden, "Bounds on a Distribution Function When its First n Moments are Given," Ann. Math. Stat., 24, 1953, pp. 361-376.