Abstract

When testing the goodness-of-fit of actual data against a hypothesized distribution function, the chi-square test statistic requires each expected cell frequency to be greater than or equal to five observations [2]. Often this requirement creates a problem when there are few observations and/or many values of the BY variables. This paper illustrates a SAS algorithm designed to satisfy this requirement.

Introduction

Often in industry, a need arises for testing actual data against a hypothesized distribution function. The examples of the paper are related to queueing models which assume the arrival rate follows a Poisson distribution and the service time follows an exponential distribution. In order to test these underlying assumptions, the number of units arriving in the "system" must be counted and the length of time each unit stays in the system must be measured. Then the chi-square statistic is calculated to test the goodness-of-fit of the actual data to the assumed distribution function. In order for the chi-square statistic to be meaningful, each expected cell must have at least five expected observations.

Base SAS software was used to do the following:

1. Calculate the average rate of arrival (A) or the average length of service time (L).
2. Calculate the expected distributions.
3. Check if the expected distribution has fewer than five observations per cell and collapse the cells if necessary.
4. Calculate the chi-square statistic, degrees of freedom and the probability value.
5. Produce a bar chart and plots of the actual and expected counts.

The focus of this paper is on the algorithm developed to check the number of expected observations per cell. This algorithm is only necessary when a limited amount of data is available. (In fact, four separate SAS software programs were written: 1) No BY variables for the rate of arrival; 2) BY variables included for the rate of arrival; 3) No BY variables for the service time; 4) BY variables for the service time. Only programs 1) and 4) will be shown in Tables 3 and 4, respectively. Note, the programs while highlighting the "collapsing" algorithm, illustrate all five of the steps mentioned above.)

Method

This algorithm uses the DATA step's REPEAT and OUTPUT statements to collapse the cells. To be consistent with SAS notation, "cell" will be referred to as an observation and the "five expected observations" will be referred to as the expected count of the cell variable and (output frequency). Table 1 is a flow diagram of the algorithm. Table 2 illustrates the algorithm with an example. An abbreviated description of the algorithm follows:

First, the algorithm checks an observation to determine if the expected frequency for the observation is greater than or equal to five. If it is, then first the expected frequency is added to any previous observations not yet output to the new data set and the sum is output to the new data set.

If the observation is not greater than or equal to five then its frequency and expected frequency are stored in temporary variables, and REPEAT them. Its frequency and expected frequency are added to the next observation and a check is performed to see if this new expected frequency is greater than five. Once the new expected frequency is greater than five then this observation is output to the data set. The algorithm tests for end-of-file (or for the last value of the BY variable) and outputs the observation to the new data set regardless of the size of its expected frequency. This is done so that all of the actual data contributes to the chi-square statistic.

Summary

The algorithm described above regroups cells such that each cell has at least five expected observations per cell to satisfy the assumptions upon which the chi-square statistics are based. The algorithm could be easily modified to decrease or increase the expected frequency, or to make the minimum expected cell count variable based on the input data.

References

[1] SAS is the registered trademark of SAS Institute Inc., Cary, NC USA.

For Further Information, Please Contact:

Elizabeth R. Zobre
AT&T Communications
Rm. 6235H2
295 N. Maple Ave.
Basking Ridge, N. J. 07920
201-221-3696

820
TABLE 1
Flow Diagram of Algorithm to Collapse Cells
With Fewer Than Five Expected Observations
(With No BY Statements)

INPUT N
FN EXPFREQ;

EXPFRQ < 5
Y
TEMP=FN*TEMP;
TEXP=EXPFREQ+TEXP;

TEXP > 5
OR EOF
N
Y
FN=TEMP;
EXPFREQ=TEXP;
TEMP=0;
TEXP=0;

OUTPUT;

RETAIN TEMP TEXP;

DEFINITIONS:
N= RATE OF ARRIVAL OR SERVICE TIME
FN= ACTUAL FREQUENCY
EXPFRQ= EXPECTED FREQUENCY
TEXP= TEMPORARY EXPECTED FREQUENCY
TABLE 2

An Example to Illustrate the Algorithm to Collapse Cells with Fewer than Five Expected Observations

<table>
<thead>
<tr>
<th>RATE OF ARRIVAL</th>
<th>First</th>
<th>Exp. First</th>
<th>TEMP</th>
<th>TEMP</th>
<th>N</th>
<th>FN</th>
<th>EXPEDO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
<td>0.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>0.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1.4</td>
<td>1</td>
<td>2.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3.3</td>
<td>4</td>
<td>5.3</td>
<td>4</td>
<td>4</td>
<td>5.3</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>6.0</td>
<td>3</td>
<td>6.0</td>
<td>5</td>
<td>3</td>
<td>6.0</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>9.1</td>
<td>5</td>
<td>9.1</td>
<td>6</td>
<td>5</td>
<td>9.1</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>11.9</td>
<td>12</td>
<td>11.9</td>
<td>7</td>
<td>12</td>
<td>11.9</td>
</tr>
<tr>
<td>8</td>
<td>17</td>
<td>13.7</td>
<td>17</td>
<td>13.7</td>
<td>8</td>
<td>17</td>
<td>13.7</td>
</tr>
<tr>
<td>9</td>
<td>22</td>
<td>13.9</td>
<td>22</td>
<td>13.9</td>
<td>9</td>
<td>22</td>
<td>13.9</td>
</tr>
<tr>
<td>10</td>
<td>15</td>
<td>12.8</td>
<td>15</td>
<td>12.8</td>
<td>10</td>
<td>15</td>
<td>12.8</td>
</tr>
<tr>
<td>11</td>
<td>12</td>
<td>10.6</td>
<td>12</td>
<td>10.6</td>
<td>11</td>
<td>12</td>
<td>10.6</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>8.1</td>
<td>6</td>
<td>8.1</td>
<td>12</td>
<td>6</td>
<td>8.1</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>5.7</td>
<td>4</td>
<td>5.7</td>
<td>13</td>
<td>4</td>
<td>5.7</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>3.8</td>
<td>2</td>
<td>3.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>2.3</td>
<td>5</td>
<td>6.1</td>
<td>15</td>
<td>5</td>
<td>6.1</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>2.7</td>
<td>1</td>
<td>2.7</td>
<td>16</td>
<td>1</td>
<td>2.7</td>
</tr>
</tbody>
</table>

822
TABLE 3

******************************************************************************
*  * THIS SAS SOFTWARE PROGRAM TESTS ACTUAL *  
*  DATA AGAINST A POISSON DISTRIBUTION *  
*  USING THE CHI-SQUARE GOODNESS-OF-FIT *  
*  STATISTIC *  
*  *  
*  NOTE: THERE ARE NO BY-VARIABLES *  
******************************************************************************

CMS FT IN DISKR DATAR AR DATA A:

DATA FRQ;
  INPUT ROA COUNT;
PROC PRINT; TITLE1;
TITLE2 ACTUAL DATA BEFORE COLLAPSSING CELLS;
******************************************************************************
*  *  *  *  
*  STEP 1: CALCULATING THE AVERAGE *  
*  ARRIVAL RATE *  
*  *  *  
******************************************************************************

DATA FRQ; SET FRQ;
  NFN=ROA*COUNT;
  N=ROA;
  FN=COUNT;
DUM=1;
DROP ROA COUNT;
PROC MEANS NOPRINT SUM N;
VAR FN NFN;
OUTPUT OUT=NEW SUM=S1 S2 ;
DATA NEW; SET NEW;
LAMBDA=S2/S1;
DUM=1;
DATA FRQ; MERGE FH;! NFN; BY DUM;
******************************************************************************
*  *  *  *  
*  STEP 2: CALCULATING THE EXPECTED *  
*  DISTRIBUTION FUNCTION *  
*  *  *  
******************************************************************************

DATA FRQ; SET FRQ;
  IF EOF THEN PN=1-FOISSON(LAMBDA,N-1);
ELSE
  PN=(LAMBDA**N)*(EXP(-LAMBDA)))/GAMMA(N+1);
  EXPFRQ=FN*PN;
  DUMP DUM;
PROC PRINT; TITLE1;
TITLE2 CALCULATIONS OF EXPECTED OBS. PER CELL;
******************************************************************************
*  *  *  *  
*  STEP 3: CHECK TO SEE IF THE EXPECTED *  
*  DISTRIBUTION HAS FEWER THAN *  
*  FIVE OBSERVATIONS PER CELL AND *  
*  COLLAPSE THE CELLS IF NECESSARY *  
*  *  *  
******************************************************************************

DATA FRQ; SET FRQ;
  IF EXPFRQ < 5.0000 THEN DO;
    TEMP=FN/TEMP;
    EXPFRQ=EXPFRQ + TEMP;
    TEMO=0;
    OUTPUT;
    RETURN;
  ELSE DO;
    FN=FN+TEMP;
    EXPFRQ=EXPFRQ + TEMP;
    TEMO=0;
    OUTPUT;
    RETURN TEMP EXPFRQ; 
END;
PROC PRINT; VAR N FN EXPFRQ; TITLE1;
TITLE2 AFTER GROUPING;
******************************************************************************

******************************************************************************
*  *  *  *  
*  STEP 4: CALCULATE CHI-SQUARE STATISTIC, *  
*  DEGREE OF FREEDOM *  
*  AND PROBABILITY VALUE *  
*  *  *  
******************************************************************************

DATA FRQ; SET FRQ;
  CHI=((FN-EXPFRQ)**2)/EXPFRQ;
PROC SORT; BY LAMBDA ;
PROC MEANS NOPRINT SUM N BY LAMBDA ;
VAR CHI;
OUTPUT OUT=CHI SUM=CHISQ N=V;
DATA CHI; SET CHI;
  DF=V-2;
  PROB1=PROCCHI(CHISQ,DF);  
  PROC PRINT CHI=CHISQ DF=V;
  V1=CHI CHISQ DF=V;
TITLE1; TITLE2 GOODNESS OF FIT TEST;
******************************************************************************
*  *  *  *  
*  STEP 5: PRODUCE BAR CHART AND PLOT *  
*  *  *  
******************************************************************************

PROC CHART DATA=FRQ;
VAR N / FREQ=FN DISCRETE;
LABEL N=RATE OF ARRIVAL;
TITLE1 TITLE2 ACTUAL DATA;
PROC PLOT DATA=FRQ;
  PLOT FN="," EXPFRQ="," /OVERLAY;  
  LABEL N=RATE OF ARRIVAL EXPFRQ=EXP FREQUENCY;
TITLE1; TITLE2 PLOT OF ACTUALS AND EXPECTED FREQUENCIES;

TABLE 3 (CONTINUED)

******************************************************************************
*  DATA FRQ; SET FRQ END=EOF;
*  IF _N=1 THEN TEMP=0;
*  IF _N=1 THEN TEMO=0;
*  IF EXPFRQ < 5.0000 THEN DO;
    TEMP=FN/TEMP;
    EXPFRQ=EXPFRQ + TEMP;
    TEMO=0;
    OUTPUT;
    RETURN;
*  ELSE DO;
    FN=FN+TEMP;
    EXPFRQ=EXPFRQ + TEMP;
    TEMO=0;
    OUTPUT;
    RETURN TEMP EXPFRQ; 
END;

******************************************************************************
*  STEP 4: CALCULATE CHI-SQUARE STATISTIC, *  
*  DEGREE OF FREEDOM *  
*  AND PROBABILITY VALUE *  
*  *  *  
******************************************************************************

DATA FRQ; SET FRQ;
  CHI=((FN-EXPFRQ)**2)/EXPFRQ;
PROC SORT; BY LAMBDA ;
PROC MEANS NOPRINT SUM N BY LAMBDA ;
VAR CHI;
OUTPUT OUT=CHI SUM=CHISQ N=V;
DATA CHI; SET CHI;
  DF=V-2;
  PROB1=PROCCHI(CHISQ,DF);  
  PROC PRINT CHI=CHISQ DF=V;
  V1=CHI CHISQ DF=V;
TITLE1; TITLE2 GOODNESS OF FIT TEST;
******************************************************************************
*  *  *  *  
*  STEP 5: PRODUCE BAR CHART AND PLOT *  
*  *  *  
******************************************************************************

PROC CHART DATA=FRQ;
VAR N / FREQ=FN DISCRETE;
LABEL N=RATE OF ARRIVAL;
TITLE1 TITLE2 ACTUAL DATA;
PROC PLOT DATA=FRQ;
  PLOT FN="," EXPFRQ="," /OVERLAY;  
  LABEL N=RATE OF ARRIVAL EXPFRQ=EXP FREQUENCY;
TITLE1; TITLE2 PLOT OF ACTUALS AND EXPECTED FREQUENCIES;
### TABLE 4

**DATA FRQ;** SET FRQ; BY DAYS;
IF FIRST DAYS THEN TEXP=0;
IF FIRST DAYS THEN TEM=0;
IF EXPFRQ < 5.0000 THEN DO;
TEM=EXP+EXPFRQ;
IF TEXP >= 5.0000 OR LAST DAYS THEN DO;
FN=TEM;
EXPFRQ=EXP;
TEM=0;
EXP=0;
OUTPUT;
END;
REMAIN TEXP;
RETURN;
END;
ELSE DO;
FN=FN+TEM;
EXPFRQ=EXPFRQ + TEM;
TEM=0;
EXP=0;
OUTPUT;
END;
PROC PRINT DATA=CHI; VAR DAYS N FRQ EXPFRQ; TITLE;
TITLE2 DISTRIBUTIONS AFTER COLLAPSING CELLS WITH LESS THAN 5 EXPECTED OBSERVATIONS;

**DATA FRQ;** SET FRQ; BY DAYS;
CHI=(FN-EXPFRQ)**2/EXPFRQ;
PROC SORT DATA=FRQ; BY DAYS;
PROC MEANS NOPRINT SUM N; BY DAYS;
VAR CHI;
DATA CHI; m:-r
DF=V-2;
IF DF LT 1 THEN PROB=.,
ELSE PROB=1-PROBCIII (CHISQ,DF) ;
PROC PRINT DATA=CHI; VAR DAYS SVTIME CHISQ PROB DF; TITLE;
TITLE2 GOODNESS OF FIT TEST;

**DATA FRQ;** SET FRQ; BY DAYS;
VAR N / FRQ=FN DISCRETE;
LABEL N=SERVICE TIME EXPFRQ=REACTION TIME;
PROC PLOT DATA=FRQ; BY DAYS;
PLOT FN=‘N’ EXPFRQ=‘E’ /OVERLAY;
LABEL N=SERVICE TIME EXPFRQ=REACTION TIME;
TITLE;
TITLE2 PLOT OF ACTUALS AND EXPECTED FREQUENCIES;