X^2 & G Tests for Association in RxC Contingency Tables

By Marc Benison & Melissa Lavides

TERMINOLOGY

Contingency tests use data from categorical (nominal) variables, placing observations in classes

Contingency tables are constructed for comparison of two categorical variables, uses include:

- To show which observations may be simultaneously classified according to the classes.
- To perform tests of association (or tests of independence)

RxC Contingency Table organizes the counts for each categorical variable into rows (R) by columns (C)

- Make tables for both the observed counts and the expected counts
- The expected counts/frequencies provide the "model" against which the observed counts/frequencies can be statistically compared (i.e., the X^2 and G tests for Association of categorical variables)

- A general RxC Contingency Table is organized like this:

	OBSERVED COUNTS		column X	column Y	column Z	
			VARIABLE 2			ROW TOTALS
			X	Y	Z	
row A	VARIABLE 1	A	a	b	c	= row A total a+b+c
row B		B	d	e	f	= row B total d+e+f
row C		C	g	h	i	= row C total g+h+i
	COLUMN TOTALS		= column X total a+d+g	= column Y total b+e+h	= column Z total c+f+i	= GRAND TOTAL N = a+b+c+d+f+g+h+i

- An example RxC Contingency Table of Observed Counts for Sex of Dog (R) by Coat Color (C)

OBSERVED COUNTS		COAT COLOR				ROW TOTALS
		Black	White	Gray	Spotted	
SEX	Male	158	200	46	46	450
	Female	170	282	51	47	550
COLUMN TOTALS		328	482	97	93	1000

THE TESTS

In order to perform statistical tests like the Chi-Square (X^2) & G Tests for Association on the RxC Contingency Tables, the following criteria must apply:

- **Assume** that the observed values of the variables are from a random sample
- **Assume** that the observed values for rows and columns are independent

- **PROBABILITIES:**
  - **P_i= P(R=i)** the probability of counts in the Rows, *i*
  - **P_j = P(C=j)** the probability of counts in the Columns, *j*
  - **P_{ij} = P(R=i, C=j)** the probability of counts in the Rows and the Columns occurring together

- **HYPOTHESES:** the X^2 and G Tests can determine if two variables are independent of each other.

  H_0 – the null hypothesis:
  $P_{ij} = (P_i)(P_j)$
  The probability of the categorical variables in the Rows (*i*) and the Columns (*j*) occurring together equals the product of their separate probabilities.
  Thus, variables in the Rows are independent from variables in the Columns

  H1 – the alternative hypothesis:
  P_{ij} > or < $(P_i)(P_j)$
  The probability of the categorical variables in the Rows (*i*) and the Columns (*j*) occurring together equals is either greater than or less than the product of their separate probabilities.
  Thus, use the Two-Sided Test to show their association.

- **NORMAL APPROXIMATION:** determine if the sample data follows the Normal Distribution
  IF no more than 1/5 of the cells have expected values in each block **is equal to or less than 5**
  AND no cell has expected value **is equal to or less than 1**
  THEN Approximation may be used

- **THE RxC TABLE:** to construct the RxC Contingency Table, include
  - The *observed* counts for each categorical variable
  - The *observed* ROW totals, COLUMN totals and GRAND total

- **EXPECTED VALUES:** calculate the expected frequency for each count occurring in the sample
  Expected frequency (f_e)= $\frac{\textbf{Row Total x Column Total}}{\textbf{Grand Total}}$ i.e. Male with Spotted coat $f_e = \frac{450 \, x \, 93}{1000} = 41.85$

- Construct a table of the EXPECTED frequencies:

EXPECTED FREQUENCY		COAT COLOR				ROW TOTALS
		Black	White	Gray	Spotted	
SEX	Male	147.6	216.9	43.65	$\frac{450 \, x \, 93}{1000}$ = 41.85	450
	Female	$\frac{550 \, x \, 328}{1000}$ = 180.4	265.1	53.35	51.15	550
COLUMN TOTALS		328	482	97	93	1000

- **<u>TEST STATISTIC</u>:** calculate the $X^2$and the G Test statistics

  Let:

  $\boldsymbol{i}$ = 1..R; the Rows $\boldsymbol{O_{i,j}}$ = the observed variable located in cell *i* (Row) by *j* (Column)

  $\boldsymbol{j}$ = 1..C; the Columns $\boldsymbol{E_{i,j}}$ = the expected frequency located in cell *i* (Row) by *j* (Column)

  The **Chi-Square Test Statistic** is the sum of the squares of each cell's (*i,j* or Row, Column) *observed* count minus the *expected* frequency, divided by that cell's *expected* frequency.

  $$X^2 = \sum_i \sum_j \left(\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}\right)$$

  The **G Test Statistic or the "log likelihood statistic"** is 2 times the sum of each cell's *observed* counts times the natural log of that cell's *observed* count divided by its *expected* frequency.

  $$G = 2*\left(\sum_i \sum_j O_{i,j} \cdot \ln\left(\frac{O_{i,j}}{E_{i,j}}\right)\right)$$

  i.e. For the Sex of the Dog (R) by Coat Color (C):

  $$X^2 = \frac{(158-147.6)^2}{147.6} + \frac{(200-216.9)^2}{216.9} + \frac{(46-43.65)^2}{43.65} + \frac{(46-41.85)^2}{41.85} + \frac{(170-180.4)^2}{180.4} + \frac{(282-265.1)^2}{265.1} + \frac{(170-180.4)^2}{180.4} + \frac{(51-53.35)^2}{53.35} + \frac{(47-51.15)^2}{51.15} = 4.7048$$

  $$\mathbf{G} = 2*\left[\left(158*\ln\frac{158}{147.6}\right) + \left(200*\ln\frac{200}{216.9}\right) + \left(46*\ln\frac{46}{43.65}\right) + \left(46*\ln\frac{46}{41.85}\right) + \left(170*\ln\frac{170}{180.4}\right) + \left(282*\ln\frac{282}{265.1}\right) + \left(170*\ln\frac{170}{180.4}\right) + \left(51*\ \ln\frac{51}{53.35}\right) + \left(47*\ln\frac{47}{51.15}\right)\right] = 4.709279$$

- **<u>SAMPLING DISTRIBUTION OF TEST STATISTICS</u>:**

  IF Assumptions hold and H_0 is true,

  THEN $\boldsymbol{X \sim X^2_{((R-1)(C-1))}}$ that is, the Chi-Square Test Statistic for the sample is approximately equal to the test statistic in the normal distribution

  AND $\boldsymbol{G \sim X^2_{((R-1)(C-1))}}$ that is, the G Test Statistic is approximately equal to the $X^2_{((R-1)(C-1))}$ for the sample in the normal distribution (G and $\boldsymbol{X^2}$ should not report significantly different probabilities)

- **<u>CRITICAL VALUE OF THE TEST</u>:**

  $\boldsymbol{\alpha = 0.05}$

  This sets the probability of **Type 1 Error** ("false positive" of rejecting H_0 when it is in fact true; rejecting the truth that the frequencies of the Rows are independent of the frequencies in the Columns).

  df = (R-1)∗(C-1) Let: R= # of Rows and C= # of Columns

  The **degrees of freedom** value is the number of independent values that go into the estimate of a statistical parameter. The df reveals the number of values in the calculated test statistic that are "free" to vary.

  i.e. For the Sex of the Dog (R) by Coat Color (C):

  R = 2 and C = 4 so **df** = (2-1)*(4-1) = 3

  CV = qchisq(1- $\boldsymbol{\alpha}$, df)

  This commands R interpreter to determine the boundary by which a test statistic is "significant" enough to prompt the rejection of H_0 (accept H_1) and which statistic is "insignificant" so that we do not reject H_0. Use the *qchisq* function.

  i.e. For the Sex of the Dog (R) by Coat Color (C):

  R interpreter responds: **CV** = qchisq(1-0.05,3) = 7.814728

- **PROBABILITY VALUE:** Command R interpreter to determine the probabilities that the test statistics do or do not show significance, given the df. Use the *pchisq* function and input the test statistic and the df

  $\boldsymbol{P_\chi}$ **= (1-pchisq(X^2, df))** or $\boldsymbol{P_G}$ **= (1-pchisq(G, df))**

  i.e. For the Sex of the Dog (R) by Coat Color (C):

  $\mathbf{P}_X$ = (1-pchisq(4.7048,9)) = 0.1947

  $\mathbf{P}_\mathbf{G}$ = (1-pchisq(4.709279,9)) = 0.1943656

- **DECISION RULE:**

  ACCEPT H_0 (do not reject) IF:

$X^2 \leq$ **CV** or **G** $\leq$ **CV**	the difference between the observed and the expected frequencies are statistically insignificant
$P_\chi \geq \alpha$ or $P_G \geq \alpha$	the probability of statistical insignificance is equal to/greater than the chance of error

  REJECT H_0 (accept H_1) IF:

$X^2 >$ **CV** or **G** $>$ **CV**	the difference between the observed and the expected frequencies are statistically significant
$P_\chi < \alpha$ or $P_G < \alpha$	the probability of statistical insignificance is less than the chance of error

  i.e. For the Sex of the Dog (R) by Coat Color (C):

  $\mathbf{P}_X$ of 0.1947 or $\mathbf{P}_\mathbf{G}$ of 0.1943656 $\geq$ $\boldsymbol{\alpha}$ of 0.05

  $\mathbf{X^2}$ of 4.7048 or **G** of 4.709279 $\leq$ **CV** of 7.814728

  Therefore, ACCEPT H_0 (do not reject)

  Conclude that the Sex of the Dog is independent from the Coat Color
  The data is consistent with the statistical model, and the differences between the observed and expected frequencies are due to random chance.

PROTOTYPE IN R:

#Using the Sex of the Dog (R) by Coat Color (C) example data table
#Enter the RxC Contingency as a matrix:

```
X=matrix(c(158,200,46,46,170,282,51,47),nrow=2,byrow=T)
X
chisq.test(X)
```

```
> X=matrix(c(158,200,46,46,170,282,51,47),nrow=2,byrow=T)
> X
     [,1] [,2] [,3] [,4]
[1,]  158  200   46   46
[2,]  170  282   51   47
> chisq.test(X)

        Pearson's Chi-squared test

data:  X
X-squared = 4.7048, df = 3, p-value = 0.1947
```

#Calculate G Test Statistic, assigning O as observed and E as expected frequencies

```
O=matrix(chisq.test(X)$observed)
O
E=matrix(chisq.test(X)$expected)
E
G=2*sum(O*log(O/E))
G
```

#Determine the Critical Value using the qchisq() function
#Set the Type 1 Error to alpha = 0.05
#Input the R = # Rows and C= # Columns to calculate the df

```
alpha=0.05
R=2
C=4
df=(R-1)*(C-1)
CV=qchisq(1-alpha,df)
CV
```

#Calculate the Probability for the G and X Test Statistic:

```
PG=(1-pchisq(G,df))
PG
Xsq=sum(((O-E)^2)/E)
PX=(1-pchisq(Xsq,df))
PX
```

#CREATE A RxC CONTINGENCY TABLE IN R FORMAT
#Upload and save the dataset MBML.RxC into your working directory

```
RxC=read.table("c:/R/MBML.RxC.txt")
RxC
attach(RxC)
```

#Use the table function and input the (Row, Column) categorical variables
#Check the data with the chisq.test() function

```
T=table(Gender,Eye.Color)
T
chisq.test(T)
```

```
> attach(RxC)
> T=table(Gender,Eye.Color)
> T
        Eye.Color
Gender   blue brown green
  female   16    23     8
  male     13    35     5
> chisq.test(T)

        Pearson's Chi-squared test

data:  T
X-squared = 3.1367, df = 2, p-value = 0.2084
```