Problem 1: We will use for this problem 1 a dataset collected from the 2014 World Cup Soccer in Brazil. Using your favorite search engine (yahoo, google, bing, ...), get the frequency of various number of goals scored by a team during the first round of play. (Similar to our discussion from class).

For this problem, the Colombian team was chosen. Data for the team’s goals during the first round were obtained from the official FIFA website at <http://www.fifa.com/worldcup/archive/brazil2014/matches>.

During the first round, Colombia played a total of 3 matches in Group C: The first match was won 3-0 against Greece. The second match was won 2-1 against Ivory Coast. The third match was won 4-1 against Japan. Thus, the Colombian team scored 3, 2, and 4 goals.

In *R*, k is defined as the number of goals scored in a match and obs as the number of matches in which k goals were scored. Since the Colombian team scored at most 4 goals in a single match, only included k values from 0 to 4 are included. The two variables are bound as dat to better illustrate the setup:

> ###Input values

> k<-c(0,1,2,3,4) ###k: number of goals in a match

> obs<-c(0,0,1,1,1) ###obs: observed goal frequency

> ###Put k and obs in dat

> dat<-cbind(k,obs);dat

k obs

[1,] 0 0

[2,] 1 0

[3,] 2 1

[4,] 3 1

[5,] 4 1

1. Find the mean number (or average number) of goals scored during the first round of the 2014 World Cup.

Using *R* to find mn, the mean number of goals scored by the Colombian team, we first calculate sumgr, the total number of goals scored by the Colombian team during the first round, by adding the products of k and obs at each level. We then divide sumgr by the number of matches played in the first round, n, which is computed as the sum of obs for all levels of k.

> ###Calculate E(x)

> sumgr<-sum(k \* obs);sumgr ###sumgr: Σ(k \* obs)

[1] 9

> n<-sum(obs);n ###n: Σ(obs)

[1] 3

> mn<-sumgr/n;mn ###mn: E(x) = (Σ(k \* obs))/n

[1] 3

Hence, the Colombian team scored an average of 3 goals during the first round of the 2014 Brazil World Cup.

1. Find the Poisson probabilities at each level.

To find the Poisson probabilities at each level of k, it is useful to have the mean number of goals scored by the Colombian team—3. Hence in *R*, the Poisson probabilities were computed as px with parameters k and mn.

For values of k greater than 4, pgreater is defined as the Poisson probability associated with scoring more than 4 goals in a match. It is calculated by subtracting from 1 the sum of all Poisson probabilities from k = 0 to k = 4, and then added to the px variable. The k variable wis adjusted to include the string “>4”, representing values of k above 4. Finally, a dataframe is used to visualize the probabilities.

> ###Calculate Poisson probabilities

> px<-dpois(k,mn); ###px: X~Pois(k, mn)

> ###Probabilities greater than 4

> pgreater<-(1 - sum(px)); ###pgreater: p(X > 4)

> px<-c(px, pgreater)

> k<-c(k, ">4"); ###Modify k for k > 4

> ###View on a dataframe

> data.frame(k,px)

k px

1 0 0.04978707

2 1 0.14936121

3 2 0.22404181

4 3 0.22404181

5 4 0.16803136

6 >4 0.18473676

1. Compute the expected frequencies

To calculate the expected frequencies in *R*, the variable expec is created to hold the values for the expected frequency of each level of k. The value of expec for each level is calculated as the product of k and its associated Poisson probability. Again, a dataframe helps to visualize the expected frequencies for each level.

> ###Compute expected frequencies

> expec<-(n\*px); ###E(f) = np(X = k)

> ###View on a dataframe

> data.frame(k,expec)

k expec

1 0 0.1493612

2 1 0.4480836

3 2 0.6721254

4 3 0.6721254

5 4 0.5040941

6 >4 0.5542103

1. Perform a goodness-of-fit test to check if your Poisson model is appropriate for this data.

To check if the Poisson model is an appropriate fit for the data, we must perform a goodness-of-fit test. The hypotheses for this test are as follows:

* *H*0: For all *i*, (The Poisson model is appropriate.)
* *Ha*: At least one value of differs from its expected value. (The Poisson model is not appropriate.)

For the test, we use a cutoff of α = .05. To begin, the k variable is given the placeholder “5” instead of “>4” to be used as a scalar variable. Then, a matrix called datamatrix is created to hold values of k and obs. From here, it is tempting to simply conduct the χ2 test in *R*:

> ###Readjust k

> k<-c(0,1,2,3,4,5); ###5 used as placeholder for k > 4

> ###Put data in datamatrix

> datamatrix = matrix(c(k, obs), nrow = 6, ncol = 2);

>

> ###Chi-square test

> chisq.test(obs, p = px)

Chi-squared test for given probabilities

data: obs

X-squared = 1.9594, df = 5, p-value = 0.8547

Warning message:

In chisq.test(obs, p = px) : Chi-squared approximation may be incorrect

From the output given by *R*, the observed value of χ2(5) = 1.96. Its *p*-value is .8547, which is not significant at the 5% level.

However, from the output in the previous part of the problem, we see than none of the expected frequencies () are above 5. Thus, as noted by the warning message in *R*, our approximations of the χ2 statistic may be incorrect.

Nonetheless, we can use the Monte Carlo simulation to estimate a simulated *p*-value for χ2:

> chisq.test(obs, p = px,simulate.p.value = TRUE)

Chi-squared test for given probabilities with simulated p-value (based on

2000 replicates)

data: obs

X-squared = 1.9594, df = NA, p-value = 0.945

From the output given by *R*, the simulated *p*-value is .9450, which is not significant at the 5% level. This result indicates that there is not sufficient evidence to reject the null hypothesis that the Poisson model is an appropriate fit for the data.

Given the problems associated with the χ2 goodness-of-fit test at small values of , without gathering more data, Fisher’s exact test must be used instead of the χ2 test. The hypotheses for this test are as follows:

* *H*0: θ = 1 (The Poisson model is appropriate.)
* *Ha*: θ ≠ 1 (The Poisson model is not appropriate.)

Fisher’s exact test was conducted on this matrix:

> ###Fisher's exact test

> fisher.test(datamatrix, alternative = "two.sided")

Fisher's Exact Test for Count Data

data: datamatrix

p-value = 0.7855

alternative hypothesis: two.sided

Given our one-sided *p*-value of .7855, we do not have sufficient evidence to reject the null hypothesis that the Poisson model is an appropriate fit for the data.

Problem 2: In a study of heart disease among males, the 356 subjects were classified according to socioeconomic status and smoking habits. The study recognized three levels of socioeconomic status (high, middle and low) and three smoking categories (current smoker, never smoked, former smoker).

To begin, we must first input our values into a 3 × 3 contingency table. In *R*, this was accomplished by creating an array called table. Smoking status was coded as Smoking, while socioeconomic status was labelled as SES.

> ###Input values

> table<-array(data = c(51,92,68,22,21,9,43,28,22), dim = c(3,3))

> dimnames(table)<-list(Smoking = c("current","former","never"), SES = c("high","medium","low"));table

SES

Smoking high medium low

current 51 22 43

former 92 21 28

never 68 9 22

For the cutoff value of , we must calculate the degrees of freedom associated with the table. For these tests, it will be *v* = (*I* – 1)(*J* – 1) = (2 – 1)(2 – 1) = 4. In *R*:

> ###Degrees of freedom

> v<-(nrow(table)-1)\*(ncol(table)-1);v ###df

[1] 4

Thus, the cutoff will be (4) = 9.49.

1. Using two different methods, at the significance level of 5%, do these data show that smoking habits and socioeconomic status are dependent or independent?

The two tests that will be used to determine if the data indicate that smoking habits and socioeconomic status are either dependent or independent are the χ2 and *G*2 tests for independence.

The hypotheses for this test are as follows:

* *H*0: for all *i* and *j* (Smoking and socioeconomic status are independent.)
* *Ha*: At least one (Smoking and socioeconomic status are dependent.).

However, before using either tests, we must ensure that *n* is sufficiently large for our interpretations to be valid. In essence, we must verify that all of our expected frequencies are at least 5, assuming that *H*0 is true. To calculate expected frequencies (), we first compute the row (*ni*+), column (*n*+*j*), and overall (*n*) totals. In *R*, these are coded as rowtot, coltot, and n respectively:

> ###Totals

> rowtot<-apply(table,1,sum);rowtot ###ni+

current former never

116 141 99

> coltot<-apply(table,2,sum);coltot ###n+j

high medium low

211 52 93

> n<-sum(rowtot);n ###n

[1] 356

Another matrix—muhat—is created to hold the values of expected frequencies. The values are calculated in *R* as for each cell:

> ###Expected frequencies

> muhat<-rowtot%\*%t(coltot)/n ###E(f)

> dimnames(muhat)<-list(Smoking = c("current","former","never"), SES = c("high","medium","low"));muhat

SES

Smoking high medium low

current 68.75281 16.94382 30.30337

former 83.57022 20.59551 36.83427

never 58.67697 14.46067 25.86236

As is evident, no value of is below 5. Thus, we can proceed with the tests for independence.

* Pearson’s χ2 Test for Independence

Using the table matrix to conduct the χ2 test for independence, the test statistic is calculated as:

In *R*, the test was conducted using the chisq.test() function:

> chisq.test(table)

Pearson's Chi-squared test

data: table

X-squared = 18.51, df = 4, p-value = 0.0009808

Our obtained value of χ2(4) is 18.51, which is greater than (4) = 9.49 and thus significant at the α = .05 level. In fact, its *p*-value is *p* < .001. Hence, using the χ2 test, there is sufficient evidence at the 5% level to reject the null hypothesis that socioeconomic status and smoking are independent.

* *G*2 Likelihood Ratio Test for Independence

Using the table matrix to conduct the χ2 test for independence, the test statistic was calculated as:

In *R*:

> ###G-square test

> g<-2\*sum(table\*log(table/muhat));g ###G^2 statistic

[1] 18.6635

> pval<-1-pchisq(g,v);pval ###p-value

[1] 0.000915055

Our obtained value of *G*2(4) is 18.66, which is greater than (4) = 9.49 and thus significant at the α = .05 level. In fact, its *p*-value is *p* < .001. Hence, using the *G*2 test, there is sufficient evidence at the 5% level to reject the null hypothesis that socioeconomic status and smoking are independent.

1. Use residuals to describe the evidence of association.

Standardized Pearson residuals of each cell are calculated according to the following formula which adjusts for the non-standardized residuals’ low variance:

To simply calculations in *R*, the element in the formula is coded as qi,

while is coded as qj.

> ###pi+ and p+j

> qi<-c(1 - (rowtot/n));qi ###qi: qi+ = 1 - pi+

current former never

0.6741573 0.6039326 0.7219101

> qj<-c(1 - (coltot/n));qj ###qj: q+j = 1 - p+j

high medium low

0.4073034 0.8539326 0.7387640

From here, we calculate the standardized residuals according to the formula:

> ###Standardized Pearson residuals

> z<-c((table - muhat)/(sqrt(muhat \* qi \* qj))) ###z: adjusted Pearson residual

> z<-array(data = c((table - muhat)/(sqrt(muhat \* qi \* qj))), dim = c(3,3))

> dimnames(z)<-list(Smoking = c("current","former","never"), SES = c("high","medium","low"));z

SES

Smoking high medium low

current -4.085848 2.3441058 4.401527

former 1.284057 0.1241138 -2.026929

never 1.666591 -1.9663352 -1.039978

An analysis of the standardized Pearson residuals reveals outliers in the current/high and current/low cells, as well as possible outliers in the current/medium, former/low, and never/medium cells.

In comparison to a model of association between socioeconomic status and smoking status, a model of independence would expect:

* the frequency of current/high cell to be significantly higher than it is.
* the frequency ofcurrent/low cell to be significantly lower than it is.
* the frequency ofcurrent/medium cell to be somewhat lower than it is.
* the frequency of former/low cell to be somewhat higher than it is.
* the frequency of never/medium cell to be somewhat higher than it is.

1. Use the partitioning of the chi-square to interpret the association.

Since partitioning only works approximately with the χ2 statistic, partitioning will be done with the *G*2 likelihood ratio statistic. Since we have 4 degrees of freedom, there will be 4 partitions. In other words:

Note that the following procedures will be similar to the calculation of *G*2. As such, variables will follow the same naming conventions used before, but “gn.” will prefix the variable name, where n is the partition index. Hence for the first partition, values will be stored in the g1.table variable.

* : (current & former) × (high & medium)

For the first partition, we only consider current and former smokers with high and medium socioeconomic statuses. In *R*, we represent the corresponding 2 × 2 marginal table as the array g1.table, which holds the observed frequencies.

> ###G1

> g1.table<-array(data = c(51,92,22,21), dim = c(2,2))

> dimnames(g1.table)<-list(Smoking = c("current","former"), SES = c("high","medium"));g1.table

SES

Smoking high medium

current 51 22

former 92 21

From here, we recalculate the row and column totals to find the expected frequencies as done before globally:

> g1.rowtot<-apply(g1.table,1,sum) ###ni+

> g1.coltot<-apply(g1.table,2,sum) ###n+j

> g1.n<-sum(g1.rowtot)###n

> g1.muhat<-g1.rowtot%\*%t(g1.coltot)/g1.n; ###E(f)

> dimnames(g1.muhat)<-list(Smoking = c("current","former"), SES = c("high","medium"));g1.muhat

SES

Smoking high medium

current 56.12366 16.87634

former 86.87634 26.12366

After this, we simply compute the test statistic and its *p*-value. From *R*:

> g1<-2\*sum(g1.table\*log(g1.table/g1.muhat));g1###G1 partition

[1] 3.275386

> g1.pval<-1-pchisq(g1,1);g1.pval###p-value

[1] 0.07032638

The output indicates that (1) = 3.28. Its *p*-value is .07, which prevents (1) from being significant at the 5% level. The result of this testsuggests that there is sufficient evidence to reject the null hypothesis that socioeconomic status and smoking status are independent for the (current & former) × (high & medium) marginal table. In other words, the frequencies for current and former smokers are not significantly different for the high and medium columns.

* : (current & former) × (high/medium & low)

For the second partition, we collapse the high and medium socioeconomic status columns on the previous marginal table and compare them to the low socioeconomic status column for current and former smokers.

In *R*, we follow the exact same process as before: First, we create the array g2.table to hold the observed frequencies for the combined high/medium column and the low column. Secondly, we recalculate the row and column totals for the current and former columns to find the expected frequencies. Lastly, we calculate the value of :

> ###G2

> g2.table<-array(data = c(73,113,43,28), dim = c(2,2))

> dimnames(g2.table)<-list(Smoking = c("current","former"), SES = c("high/medium","low"));g2.table

SES

Smoking high/medium low

current 73 43

former 113 28

> g2.rowtot<-apply(g2.table,1,sum) ###ni+

> g2.coltot<-apply(g2.table,2,sum) ###n+j

> g2.n<-sum(g2.rowtot) ###n

> g2.muhat<-g2.rowtot%\*%t(g2.coltot)/g2.n; ###E(f)

> dimnames(g2.muhat)<-list(Smoking = c("current","former"), SES = c("high/medium","low"));g2.muhat

SES

Smoking high/medium low

current 83.95331 32.04669

former 102.04669 38.95331

> g2<-2\*sum(g2.table\*log(g2.table/g2.muhat));g2###G2 partition

[1] 9.426977

> g2.pval<-1-pchisq(g2,1);g2.pval ###p-value

[1] 0.002138165

The output indicates that (1) = 9.43. Its *p*-value is .002, which allows for (1) to be significant at the 5% level. The result of this testsuggests that there is sufficient evidence to reject the null hypothesis that socioeconomic status and smoking status are independent for the (current & former) × (high/medium & low) marginal table. In other words, the frequencies for current and former smokers are significantly different for the combined high/medium columns and the low column.

* : (current/former & never) × (high & medium)

For the third partition, we collapse the current and former rows to compare them against the never column on frequencies of high and medium socioeconomic status.

In *R*, we follow the exact same process as before: First, we create the array g3.table to hold the observed frequencies for the combined current/former row and the never row. Secondly, we recalculate the row and column totals for the high and low columns to find the expected frequencies. Lastly, we calculate the value of :

> ###G3

> g3.table<-array(data = c(143,68,43,9), dim = c(2,2))

> dimnames(g3.table)<-list(Smoking = c("current/former","never"), SES = c("high","medium"));g3.table

SES

Smoking high medium

current/former 143 43

never 68 9

> g3.rowtot<-apply(g3.table,1,sum) ###ni+

> g3.coltot<-apply(g3.table,2,sum) ###n+j

> g3.n<-sum(g3.rowtot) ###n

> g3.muhat<-g3.rowtot%\*%t(g3.coltot)/g3.n; ###E(f)

> dimnames(g3.muhat)<-list(Smoking = c("current/former","never"), SES = c("high","medium"));g3.muhat

SES

Smoking high medium

current/former 149.22433 36.77567

never 61.77567 15.22433

> g3<-2\*sum(g3.table\*log(g3.table/g3.muhat));g3###G3 partition

[1] 4.855609

> g3.pval<-1-pchisq(g3,1);g3.pval ###p-value

[1] 0.02755639

The output indicates that (1) = 4.86. Its *p*-value is .03, which allows for (1) to be significant at the 5% level. The result of this testsuggests that there is sufficient evidence to reject the null hypothesis that socioeconomic status and smoking status are independent for the (current/former & never) × (high & medium) marginal table. In other words, the frequencies for high and medium socioeconomic status are significantly different for the combined current/former rows and the never row.

* : (current/former & never) × (high/medium & low)

For the fourth and last partition, we collapse the current and former rows to compare them against the never column on frequencies of combined high/medium and low socioeconomic status.

In *R*, we follow the exact same process as before: First, we create the array g3.table to hold the observed frequencies for the combined current/former row and the never row. Secondly, we recalculate the row and column totals for the high/medium and low columns to find the expected frequencies. Lastly, we calculate the value of :

> ###G4

> g4.table<-array(data = c(186,77,71,22), dim = c(2,2))

> dimnames(g4.table)<-list(Smoking = c("current/former","never"), SES = c("high/medium","low"));g4.table

SES

Smoking high/medium low

current/former 186 71

never 77 22

> g4.rowtot<-apply(g4.table,1,sum) ###ni+

> g4.coltot<-apply(g4.table,2,sum) ###n+j

> g4.n<-sum(g4.rowtot) ###n

> g4.muhat<-g4.rowtot%\*%t(g4.coltot)/g4.n; ###E(f)

> dimnames(g4.muhat)<-list(Smoking = c("current/former","never"), SES = c("high/medium","low"));g4.muhat

SES

Smoking high/medium low

current/former 189.86236 67.13764

never 73.13764 25.86236

> g4<-2\*sum(g4.table\*log(g4.table/g4.muhat));g4###G4 partition

[1] 1.105524

> g4.pval<-1-pchisq(g4,1);g4.pval ###p-value

[1] 0.293057

The output indicates that (1) = 1.11. Its *p*-value is .3, which prevents (1) from being significant at the 5% level. The result of this testsuggests that there is not sufficient evidence to reject the null hypothesis that socioeconomic status and smoking status are independent for the (current/former & never) × (high/medium & low) marginal table. In other words, the frequencies for high/medium and low socioeconomic status are not significantly different for the combined current/former rows and the never row.

* Note

The marginal tables that gave the most evidence of an association between socioeconomic status and smoking status were:

* (current & former) × (high/medium & low)
* (current/former & never) × (high & medium)

The marginal tables that gave the most evidence of an independence model between socioeconomic status and smoking status were:

* (current & former) × (high & medium)
* (current/former & never) × (high/medium & low)

Lastly, we can verify in *R* that we partitioned *G*2 correctly by checking that *G*2 is exactly the sum of its individual partitions by summing them together:

> ###Check for G

> print("G-square equals");g

[1] "G-square equals"

[1] 18.6635

> print("G-square equals");sum(g1,g2,g3,g4)

[1] "G-square equals"

[1] 18.6635

1. Summarize association by constructing a 95% confidence interval for the odds ratio between current or never and whether high or low. Interpret.

In *R*, we first create an array called odds.table to hold the values for the (current & never) × (high & low) marginal table:

> ###Input values

> odds.table<-array(data = c(51,68,43,22), dim = c(2,2));

> dimnames(odds.table)<-list(Smoking = c("current","never"), SES = c("high","low"));odds.table

SES

Smoking high low

current 51 43

never 68 22

Next, we calculate the observed sample odds ratio as . In *R*:

> odds.estimate<-c((51\*22)/(68\*43));odds.estimate ###OR estimate

[1] 0.3837209

To form a confidence interval, we must calculate the asymptotic standard error (ASE) of :

Likewise in *R*, we calculate ASE as the variable ase:

> ase<-sqrt(c(1/51 + 1/68 + 1/43 + 1/22));ase ###ASE of lnOR

[1] 0.3209737

From here, our confidence interval is essentially in the format:

Thus, to proceed we compute in *R* as ln.odds.estimate:

> ln.odds.estimate=log(odds.estimate);ln.odds.estimate ###lnOR estimate

[1] -0.9578397

Assuming a cutoff of α = .05, we calculate with α/2 = .025 as z\_alpha.

> z\_alpha=qnorm(0.025,lower.tail = F);z\_alpha###z(.025)

[1] 1.959964

Next, for the confidence interval, we create two variables: ln.theta\_low and ln.thetha\_high. These are used to store the lower and upper bounds respectively of the confidence interval:

> ln.theta\_low=(ln.odds.estimate - (z\_alpha \* ase));ln.theta\_low ###lnOR - (z\_alpha \* ASE)

[1] -1.586937

> ln.theta\_high=(ln.odds.estimate +(z\_alpha \* ase));ln.theta\_high ###lnOR + (z\_alpha \* ASE)

[1] -0.3287429

Thus, with the information we have so far, the interval is reduced to:

ln(*θ*).95 = [-1.587, -0.329]

Exponentiation of ln.theta\_low and ln.thetha\_high yields a confidence interval for the odds ratio alone. Thus in *R*, we calculate a new interval for the odds ratio bounded by theta\_low and theta\_high, which are simply the exponentials of the bounds used in the previous confidence interval:

> ###OR CI

> theta\_low=exp(ln.theta\_low);theta\_low

[1] 0.2045513

> theta\_high=exp(ln.theta\_high);theta\_high

[1] 0.719828

Hence, the 95% confidence interval for the odds ratio is *θ*.95 = [0.205, 0.720].

From this information, we can ascertain with 95% confidence that the odds of a current smoker being at a high socioeconomic status are between 0.205 and 0.720 times the odds of a non-smoker being at a high socioeconomic status.