Categorical data

Self-test answers

- Using what you have learnt about data entry in **R**, can you work out how you would enter these data directly into **R**?

This question is a bit tricky because there are unequal numbers of rows in the different combinations of categories, which means that we can't use the *gl()* function to generate the categories for us. Instead we need to use the *rep()* function to create the requires number of values. For example, for the **Training** variable we need 38 instances of 'Food as Reward' and 162 of 'Affection as reward'. We can achieve this goal by executing:

```
Training<-c(rep(0, 38), rep(1, 162))
Training<-factor(Training, labels = c("Food as Reward", "Affection as Reward"))
```

The first command creates a variable called **Training**, which consists of 38 zeros followed by 162 ones. The second command then converts this variable to a factor in which the first level is labelled *Food as Reward* and the second *Affection as Reward*. We can create the **Dance** variable in a similar way. Remember that within the first 38 rows (i.e., *Food as Reward*) we want 28 rows labelled as 'Yes' and 10 labelled as 'No'. Similarly, within the final 162 rows (i.e., *Affection as Reward*), we want 48 Yes responses and 114 Nos. Therefore, we have to use the *rep()* function four times as follows:

```
Dance<-c(rep(0, 28), rep(1, 10), rep(0, 48), rep(1, 114))
Dance<-factor(Dance, labels = c("Yes", "No"))
```

The first command creates a variable called **Dance**, which consists of 28 zeros followed by 10 ones, followed by 48 zeros, followed by 114 ones. The second command then converts this variable to a factor in which the first level is labelled *Yes* and the second *No*. We can bind these variables into a dataframe by executing::

```
catsData<-data.frame(Training, Dance)
```

- Run a multiple regression analysis using **CatRegression.dat** with **LnObserved** as the outcome, and **Training, Dance** and **Interaction** as your three predictors.

Open the data and create the model as follows (remember to set your working directory to be the location of the data file):

```
catsRegression<-read.delim("CatRegression.dat", header = TRUE)

catModel<-lm(LnObserved ~ Training + Dance + Interaction, data = catsRegression)
summary(catModel)
```

The regression parameters are shown in the book.

Richard Leigh 8/11/11 13:04
Deleted: k

Richard Leigh 8/11/11 13:05
Formatted: Justified

Richard Leigh 8/11/11 13:06
Formatted: Font:Italic

Richard Leigh 8/11/11 13:05
Formatted: Justified

Richard Leigh 8/11/11 13:07
Deleted: 4

Richard Leigh 8/11/11 13:05
Formatted: Justified

Richard Leigh 8/11/11 13:08
Deleted: s

- Run another multiple regression analysis using **CatRegression.dat**; this time the outcome is the log of expected frequencies (**LnExpected**) and **Training** and **Dance** are the predictors (the interaction is not included).

Richard Leigh 8/11/11 13:09
Deleted: To show that this all actually works, r

Richard Leigh 7/11/11 06:02
Deleted: sav

Richard Leigh 8/11/11 13:09
Formatted: Justified

You should have the data loaded already, but if not execute:

```
catsRegression<-read.delim("CatRegression.dat", header = TRUE)
```

The multiple regression model can be obtained by executing:

```
catModel2<-lm(LnExpected ~ Training + Dance, data = catsRegression)
summary(catModel2)
```

The resulting regression parameters are:

```
Coefficients:
             Estimate Std. Error   t value Pr(>|t|)
(Intercept) 2.670e+00  2.427e-15 1.100e+15   <2e-16 ***
Training    1.450e+00  2.797e-15 5.183e+14   <2e-16 ***
Dance       4.895e-01  2.261e-15 2.165e+14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that b_0 = 2.67, the beta coefficient for the type of training is 1.45 and the beta coefficient for whether they danced is 0.49. All of these values are consistent with those calculated in the book chapter.

Richard Leigh 8/11/11 13:13
Formatted: Justified

- Use the *subset()* function to run a chi-square test on **Dance** and **Training** for dogs and cats separately.

Richard Leigh 8/11/11 13:10
Formatted: Font:Italic, English (UK)

First, to create different dataframes for cats and dogs, execute:

Richard Leigh 8/11/11 13:10
Deleted: ~

```
justCats = subset(catsDogs, Animal=="Cat")
justDogs = subset(catsDogs, Animal=="Dog")
```

To run the chi-square tests, execute:

Richard Leigh 8/11/11 13:10
Formatted: Indent: First line: 0 cm

```
CrossTable(justCats$Training, justCats$Dance, chisq = TRUE, fisher = TRUE, sresid = TRUE, format = "SPSS")
CrossTable(justDogs$Training, justDogs$Dance, chisq = TRUE, fisher = TRUE, sresid = TRUE, format = "SPSS")
```

- Calculate the odds ratio for dogs by hand.

ddsdancing after food=umber that had food and dancedumber that had food but didn't dance

=2014

=1.43

ddsdancing after affection=umber that had affection and dancedumber

that had affection but didn't dance

=297

=4.14

dds Ratio=ddsdancing after foodddsdancing after affection

=1.434.14

=0.35

Oliver Twisted

Please sir, can I have more … tables?

The other option for tabulating more than two variables is to use the *table()* function, which can handle more than two variables, but doesn't have very nice output. We give the *table()* function the three variables from the example (i.e., the *catsDogs* dataframe):

```
table(catsDogs$Animal, catsDogs$Training, catsDogs$Dance)
```

The resulting output is:

```
, ,  = No

     Affection as Reward Food as Reward
Cat                  114              10
Dog                    7              14

, ,  = Yes

     Affection as Reward Food as Reward
Cat                   48              28
Dog                   29              20
```

Another option is the *xtabs()* function. This allows you to enter as many variables as you like. The way it works looks a bit like the predictor side of *lm()*, with no outcome variables. So, we begin with a ~, and then list any variables we want tabulated, separating them with a '+'. We can therefore execute this command to get a table of the three variables in our example:

```
xtabs(~Animal + Training + Dance, data = catsDogs)
```

The resulting output is:

```
, , Dance = No
```

Richard Leigh 8/11/11 13:13
Formatted: Justified

Richard Leigh 8/11/11 13:13
Deleted: and execute

Richard Leigh 8/11/11 13:13
Formatted: Normal Indent, Justified

Richard Leigh 8/11/11 13:13
Deleted: , t

Richard Leigh 8/11/11 13:13
Deleted: a

Richard Leigh 8/11/11 13:14
Deleted: separeting

```
      Training
Animal Affection as Reward Food as Reward
   Cat                 114             10
   Dog                   7             14

, , Dance = Yes

      Training
Animal Affection as Reward Food as Reward
   Cat                  48             28
   Dog                  29             20
```

In both the *table()* and the *xtabs()* commands we put the **Animal** variable first, because it makes sense to split tables by this variable.

Richard Leigh 8/11/11 13:14
Formatted: Font:Italic

Richard Leigh 8/11/11 13:14
Formatted: Font:Italic

Richard Leigh 8/11/11 13:14
Formatted: Normal Indent

Labcoat Leni's real research

Is the black American happy?

Problem

Beckham, A. S. (1929). *Journal of Abnormal and Social Psychology*, *24*, 186–190.

Richard Leigh 8/11/11 13:14
Formatted: Font:Italic

Richard Leigh 8/11/11 13:14
Formatted: Font:Italic

Richard Leigh 8/11/11 13:14
Deleted: -

Richard Leigh 8/11/11 13:15
Formatted: Justified

When I was doing my psychology degree I spent a lot of time reading about the civil rights movement in the USA. Although I was supposed to be reading psychology, I became more interested in Malcolm X and Martin Luther King Jr. This is why I find Beckham's 1929 study of black Americans such an interesting piece of research. Beckham was a black American academic who founded the psychology laboratory at Howard University, Washington, DC, and his wife Ruth was the first black woman ever to be awarded a Ph.D. (also in psychology) at the University of Minnesota. The article needs to be placed within the era in which it was published. To put some context on the study, it was published 36 years before the Jim Crow laws were finally overthrown by the Civil Rights Act of 1964, and in a time when black Americans were segregated, openly discriminated against and were victims of the most abominable violations of civil liberties and human rights. For a richer context I suggest reading James Baldwin's superb novel *The Fire Next Time*. Even the language of the study and the data from it are an uncomfortable reminder of the era in which it was conducted.

Beckham sought to measure the psychological state of black Americans with three questions put to 3443 black Americans from different walks of life. He asked them whether they thought black Americans were happy, whether they personally were happy as a black American, and whether black Americans *should* be happy. They could answer only *yes* or *no* to each question. By today's standards the study is quite simple, and he did no formal statistical analysis of his data (Fisher's article containing the popularized version of the chi-square test was published only 7 years earlier in a statistics journal that would not have been read by psychologists). I love this study, though, because it demonstrates that you do not need elaborate methods to answer important and far-reaching questions; with just three questions, Beckham told the world an enormous amount about very real and important psychological and sociological phenomena.

The frequency data (number of yes and no responses within each employment category) from this study are in the file **Beckham(1929).dat**. Labcoat Leni wants you to carry out three chi-square tests (one for each question that was asked). What conclusions can you draw?

Richard Leigh 7/11/11 06:02
Deleted: sav

Richard Leigh 8/11/11 13:17
Deleted:

Solution

Are black Americans happy?

First of all, read in the data by executing:

```
americanData<-read.delim("Beckham(1929).dat", header = TRUE)
```

We can view the data by executing the name of the dataframe, which I have called *americanData*:

```
americanData
```

```
          Profession  Response  Happy  You_Happy    Should_be_Happy
1    College Students       Yes    390       1822             141
2    College Students        No   1610         48            1810
3  Unskilled Laborers       Yes    378        305             396
4  Unskilled Laborers        No    122        195             104
5           Preachers       Yes     35        230             264
6           Preachers        No    265          0              36
7          Physicians       Yes    159        203             174
8          Physicians        No     51          7              36
9          Housewives       Yes     78         17              90
10         Housewives        No    122        146             120
11    School Teachers       Yes    108         79              75
12    School Teachers        No     38         28              33
13            Lawyers       Yes     11         30               7
14            Lawyers        No     64          0              57
15           Musician       Yes     31         16              36
16           Musician        No     19         34              14
```

Looking at the data above we can see that the data are not in the correct format for the analysis that we need to do.

We could enter a contingency table for the **Happy** data called *happyTable* by executing:

Richard Leigh 8/11/11 13:18
Formatted: Normal Indent

```
College <- c(1610, 390)
Laborers <- c(122, 378)
Preachers <- c(265, 35)
Physicians <- c(51, 159)
Housewives <- c(122, 78)
Teachers <- c(38, 108)
Lawyers <- c(64, 11)
Musician <- c(19, 31)

happyTable <- cbind(College, Laborers, Preachers, Physicians, Housewives, Teachers,
Lawyers, Musician)
```

If we then execute:

```
happyTable
```

We can see that the data are now in the correct format to run the chi-square test (**NB**, I have entered the data so that 1 = No and 2 = Yes, it doesn't matter which way round you do it as long as you make a note of it, as it will be important when we are interpreting the results later on).

Richard Leigh 8/11/11 13:18
Deleted: chi

```
College  Laborers  Preachers Physicians Housewives Teachers Lawyers Musician
[1,]    1610       122        265          51         122        38       64        19
[2,]     390       378         35         159          78       108       11        31
```

Now we can run the chi-square for the question 'Are Black Americans happy?' (**Happy**) by executing:

Richard Leigh 8/11/11 13:19
Formatted: Normal Indent

Richard Leigh 8/11/11 13:19
Deleted: chi

```
CrossTable(happyTable, fisher = TRUE, chisq = TRUE, expected = TRUE, sresid = TRUE,
format = "SPSS")
```

```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|            Total Percent |
|           Std Residual |
|-------------------------|

Total Observations in Table:  3481

             |
             |    College  |   Laborers  |  Preachers  | Physicians  | Housewives  |   Teachers  |    Lawyers  |   Musician  | Row Total
             |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
       [1,]  |        1610 |         122 |         265 |          51 |         122 |          38 |          64 |          19 |        2291 |
             |    1316.288 |     329.072 |     197.443 |     138.210 |     131.629 |      96.089 |      49.361 |      32.907 |             |
             |      65.538 |     130.302 |      23.115 |      55.029 |       0.704 |      35.117 |       4.342 |       5.877 |             |
             |      70.275% |       5.325% |      11.567% |       2.226% |       5.325% |       1.659% |       2.794% |       0.829% |      65.814% |
             |      80.500% |      24.400% |      88.333% |      24.286% |      61.000% |      26.027% |      85.333% |      38.000% |             |
             |      46.251% |       3.505% |       7.613% |       1.465% |       3.505% |       1.092% |       1.839% |       0.546% |             |
             |       8.096 |     -11.415 |       4.808 |      -7.418 |      -0.839 |      -5.926 |       2.084 |      -2.424 |             |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
       [2,]  |         390 |         378 |          35 |         159 |          78 |         108 |          11 |          31 |        1190 |
             |     683.712 |     170.928 |     102.557 |      71.790 |      68.371 |      49.911 |      25.639 |      17.093 |             |
             |     126.174 |     250.859 |      44.501 |     105.943 |       1.356 |      67.607 |       8.359 |      11.315 |             |
             |      32.773% |      31.765% |       2.941% |      13.361% |       6.555% |       9.076% |       0.924% |       2.605% |      34.186% |
             |      19.500% |      75.600% |      11.667% |      75.714% |      39.000% |      73.973% |      14.667% |      62.000% |             |
             |      11.204% |      10.859% |       1.005% |       4.568% |       2.241% |       3.103% |       0.316% |       0.891% |             |
             |      -11.233 |      15.839 |      -6.671 |      10.293 |       1.164 |       8.222 |      -2.891 |       3.364 |             |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
Column Total |        2000 |         500 |         300 |         210 |         200 |         146 |          75 |          50 |        3481 |
             |      57.455% |      14.364% |       8.618% |       6.033% |       5.745% |       4.194% |       2.155% |       1.436% |             |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|

Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  936.1395     d.f. =  7     p =  7.523435e-198


Error in fisher.test(t, alternative = "two.sided") :
  FEXACT error 7.
LDSTP is too small for this problem.
Try increasing the size of the workspace.
```

Hopefully you can read the output above, I had to make it very small so that it would fit on the page.

The chi-square test is highly significant, $\chi^2(7) = 936.14$, $p < .001$. This indicates that the profile of yes and no responses differed across the professions. Looking at the standardized residuals, the only profession for which these are non-significant are housewives, who showed a fairly even split of whether they thought black Americans were happy (40%) or not (60%). Within the other professions all of the standardized residuals are much higher than 1.96, so how can we make sense of the data? What's interesting is to look at the direction of these residuals (i.e. whether they are positive or negative). For the following professions the residual for 'no' (1) was positive but for 'yes' (2) was negative; these are therefore people who responded more than we would expect that black Americans were not happy and less than expected that black Americans were happy: college students, preachers and lawyers. The remaining professions (labourers, physicians, school teachers and musicians) show the opposite pattern: the residual for 'no' (1) was negative but for 'yes' (2) was positive; these, are therefore, people who responded less than we would expect that black Americans were not happy and more than expected that black Americans were happy.

Are they happy as black Americans?

Richard Leigh 8/11/11 13:17
Deleted: *Happy*

To run a chi-square on the variable **You_Happy** we need to enter the data as a contingency table as we did for **Happy** above. I am going to call this *you_happyTable*. We can create this contingency table by executing:

```
College <- c(48, 1822)
Laborers <- c(195, 305)
Preachers <- c(0, 230)
Physicians <- c(7, 203)
Housewives <- c(146, 17)
Teachers <- c(28, 79)
Lawyers <- c(0, 30)
Musician <- c(34, 16)
you_happyTable <- cbind(College, Laborers, Preachers, Physicians, Housewives,
Teachers, Lawyers, Musician)
```

We can then view the contingency table by executing:

```
you_happyTable
```

```
     College Laborers Preachers Physicians Housewives Teachers Lawyers Musician
[1,]      48       195         0          7         146       28        0       34
[2,]    1822       305       230        203          17       79       30       16
```

As before, I have entered the data so that 1 = No and 2 = Yes; it doesn't matter which way round you do it as long as you make a note of it, as it will be important for interpretation.

Next we can run the chi-square test for **You_Happy** by executing:

```
CrossTable(you_happyTable, fisher = TRUE, chisq = TRUE, expected = TRUE, sresid =
TRUE, format = "SPSS")
```

```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|            Std Residual |
|-------------------------|

Total Observations in Table:  3160

             |
             |    College   |   Laborers   |  Preachers   | Physicians   | Housewives   |   Teachers   |    Lawyers   |   Musician   | Row Total
|
-------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
       [1,]  |           48 |          195 |            0 |            7 |          146 |           28 |            0 |           34 |          458 |
             |      271.032 |       72.468 |       33.335 |       30.437 |       23.625 |       15.508 |        4.348 |        7.247 |              |
             |      183.532 |      207.180 |       33.335 |       18.047 |      633.901 |       10.062 |        4.348 |       98.765 |              |
             |       10.480% |       42.576% |        0.000% |        1.528% |       31.878% |        6.114% |        0.000% |        7.424% |       14.494% |
             |        2.567% |       39.000% |        0.000% |        3.333% |       89.571% |       26.168% |        0.000% |       68.000% |              |
             |        1.519% |        6.171% |        0.000% |        0.222% |        4.620% |        0.886% |        0.000% |        1.076% |              |
             |      -13.547 |       14.394 |       -5.774 |       -4.248 |       25.177 |        3.172 |       -2.085 |        9.938 |              |
-------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
       [2,]  |         1822 |          305 |          230 |          203 |           17 |           79 |           30 |           16 |         2702 |
             |     1598.968 |      427.532 |      196.665 |      179.563 |      139.375 |       91.492 |       25.652 |       42.753 |              |
             |       31.110 |       35.118 |        5.650 |        3.059 |      107.449 |        1.706 |        0.737 |       16.741 |              |
             |       67.432% |       11.288% |        8.512% |        7.513% |        0.629% |        2.924% |        1.110% |        0.592% |       85.506% |
             |       97.433% |       61.000% |      100.000% |       96.667% |       10.429% |       73.832% |      100.000% |       32.000% |              |
             |       57.658% |        9.652% |        7.278% |        6.424% |        0.538% |        2.500% |        0.949% |        0.506% |              |
             |        5.578 |       -5.926 |        2.377 |        1.749 |      -10.366 |       -1.306 |        0.858 |       -4.092 |              |
-------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
Column Total |         1870 |          500 |          230 |          210 |          163 |          107 |           30 |           50 |         3160 |
             |       59.177% |       15.823% |        7.278% |        6.646% |        5.158% |        3.386% |        0.949% |        1.582% |              |
-------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|

Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  1390.740     d.f. =  7     p =  3.891606e-296
```

Looking at the output above, we can see that the chi-square test is highly significant, $\chi^2(7) = 1390.74$, $p < .001$. This indicates that the profile of yes (2) and no (1) responses differed across the professions. Looking at the standardized residuals, these are significant in most cells with a few exceptions: physicians, lawyers and school teachers saying 'yes'. Within the other cells all of the standardized residuals are much higher than 1.96. Again, we can look at the direction of these residuals (i.e. whether they are positive or negative). For labourers, housewives, school teachers and

Richard Leigh 8/11/11 13:22
Deleted: .

Richard Leigh 8/11/11 13:20
Deleted: chi

Richard Leigh 8/11/11 13:20
Deleted: ,

Richard Leigh 8/11/11 13:21
Deleted: .

Richard Leigh 8/11/11 13:21
Formatted: Normal Indent

Richard Leigh 8/11/11 13:21
Deleted: chi

Richard Leigh 8/11/11 13:23
Formatted: Normal, Justified

musicians the residual for 'no' (1) was positive but for 'yes' (2) was negative; these are therefore people who responded more than we would expect that they were not happy as black Americans and less than expected that they were happy as black Americans. The remaining professions (college students, physicians, preachers and lawyers) show the opposite pattern: the residual for 'no' (1) was negative but for 'yes' (2) was positive; these are therefore people who responded less than we would expect that they were not happy as black Americans and more than expected that they were happy as black Americans. Essentially, the former group are in low-paid jobs in which conditions would have been very hard (especially in the social context of the time). The latter group are in much more respected (and probably better-paid) professions. Therefore, the responses to this question could say more about the professions of the people asked than their views of being black Americans.

Should black Americans be happy?

To run a chi-square on the variable **Should_be_Happy** we need to enter the data as a contingency table as we did for **Happy** and **You_Happy** above. I am going to call the contingency table *should_happyTable*. We can create this contingency table by executing:

Richard Leigh 8/11/11 13:22
Deleted:

Richard Leigh 8/11/11 13:23
Formatted: Justified

Richard Leigh 8/11/11 13:21
Deleted: chi sq

```
College <- c(1810, 141)
Laborers <- c(104, 396)
Preachers <- c(36, 264)
Physicians <- c(36, 174)
Housewives <- c(120, 90)
Teachers <- c(33, 75)
Lawyers <- c(57, 7)
Musician <- c(14, 36)
should_happyTable <- cbind(College, Laborers, Preachers, Physicians, Housewives,
Teachers, Lawyers, Musician)
```

We can then view the contingency table by executing:

```
should_happyTable
```

```
     College  Laborers Preachers  Physicians  Housewives Teachers Lawyers Musician
[1,]    1810       104        36          36         120       33       57       14
[2,]     141       396       264         174          90       75        7       36
```

As before, I have entered the data so that 1 = No and 2 = Yes.

Next we can run the chi-square test for **Should_be_Happy** by executing:

Richard Leigh 8/11/11 13:23
Deleted:

Richard Leigh 8/11/11 13:23
Formatted: Normal Indent

Richard Leigh 8/11/11 13:21
Deleted: chi sq

```
CrossTable(should_happyTable, fisher = TRUE, chisq = TRUE, expected = TRUE, sresid =
TRUE, format = "SPSS")
```

```
   Cell Contents
|-------------------------|
|                       Count |
|              Expected Values |
|  Chi-square contribution |
|               Row Percent |
|            Column Percent |
|              Total Percent |
|               Std Residual |
|-------------------------|

Total Observations in Table:  3393

             |
             |    College  |   Laborers  |  Preachers  | Physicians  | Housewives  |   Teachers  |    Lawyers  |   Musician  | Row Total
|
-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
       [1,]  |        1810 |         104 |          36 |          36 |         120 |          33 |          57 |          14 |        2210 |
             |     1270.766 |     325.670 |     195.402 |     136.782 |     136.782 |      70.345 |      41.686 |      32.567 |             |
             |      228.817 |     150.882 |     130.035 |      74.257 |       2.059 |      19.826 |       5.626 |      10.585 |             |
             |       81.900% |       4.706% |       1.629% |       1.629% |       5.430% |       1.493% |       2.579% |       0.633% |      65.134% |
             |       92.773% |      20.800% |      12.000% |      17.143% |      57.143% |      30.556% |      89.062% |      28.000% |             |
             |       53.345% |       3.065% |       1.061% |       1.061% |       3.537% |       0.973% |       1.680% |       0.413% |             |
             |       15.127 |      -12.283 |      -11.403 |       -8.617 |       -1.435 |       -4.453 |        2.372 |       -3.254 |             |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
       [2,]  |         141 |         396 |         264 |         174 |          90 |          75 |           7 |          36 |        1183 |
             |      680.234 |     174.330 |     104.598 |      73.218 |      73.218 |      37.655 |      22.314 |      17.433 |             |
             |      427.460 |     281.867 |     242.922 |     138.721 |       3.846 |      37.037 |      10.510 |      19.775 |             |
             |       11.919% |      33.474% |      22.316% |      14.708% |       7.608% |       6.340% |       0.592% |       3.043% |      34.866% |
             |        7.227% |      79.200% |      88.000% |      82.857% |      42.857% |      69.444% |      10.938% |      72.000% |             |
             |        4.156% |      11.671% |       7.781% |       5.128% |       2.653% |       2.210% |       0.206% |       1.061% |             |
```

```
             |   -20.675  |    16.789  |    15.586  |    11.778  |     1.961  |     6.086  |    -3.242  |     4.447  |            |
-------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
Column Total |       1951 |        500 |        300 |        210 |        210 |        108 |         64 |         50 |       3393 |
             |     57.501% |    14.736% |     8.842% |     6.189% |     6.189% |     3.183% |     1.886% |     1.474% |            |
-------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|

Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  1784.226     d.f. =  7     p =  0
```

The chi-square test is highly significant, $\chi^2(7) = 1784.23$, $p < .001$. This indicates that the profile of yes and no responses differed across the professions. Looking at the standardized residuals, these are nearly all significant. Again, we can look at the direction of these residuals (i.e. whether they are positive or negative). For college students and lawyers the residual for 'no' was positive but for 'yes' was negative; these are therefore people who responded more than we would expect that they thought that black Americans should not be happy and less than expected that they thought black Americans should be happy. The remaining professions show the opposite pattern: the residual for 'no' was negative but for 'yes' was positive; these are therefore people who responded less than we would expect that they did not think that black Americans should be happy and more than expected that they thought that black Americans should be happy.

Richard Leigh 8/11/11 13:23
Formatted: Normal, Justified

What is interesting here and in question 1 is that college students and lawyers are in vocations in which they are expected to be critical about the world. Lawyers may well have defended black Americans who had been the subject of injustice and discrimination or racial abuse, and college students would likely be applying their critically trained minds to the immense social injustice that prevailed at the time. Therefore, these groups can see that their racial group should not be happy and should strive for the equitable and just society to which they are entitled. People in the other professions perhaps adopt a different social comparison.

Richard Leigh 8/11/11 13:23
Formatted: Justified

It's also possible for this final question that the groups interpreted the question differently: perhaps the lawyers and students interpreted the question as 'should they be happy given the political and social conditions of the time?' whereas the others interpreted the question as 'do they *deserve* happiness?'

It might seem strange to have picked a piece of research from so long ago to illustrate the chi-square test, but what I wanted to demonstrate is that simple research can sometimes be incredibly illuminating. This study asked three simple questions, yet the data are utterly fascinating. It raises further hypotheses that could be tested, it unearths very different views in different professions, and it illuminates a very important social and psychological issue. There are other studies that sometimes use the most elegant paradigms and the highly complex methodologies, but the questions they address are utterly meaningless for the real world. They miss the big picture. Albert Beckham was a remarkable man, trying to understand important and big real-world issues that mattered to hundreds of thousands of people.

Smart Alex's solutions

Richard Leigh 8/11/11 13:26
Deleted: Solutions

Task 1

- Certain editors at Sage like to think they're a bit of a whiz at football (soccer if you prefer). To see whether they are better than Sussex lecturers and postgraduates we invited various employees of Sage to join in our football

Richard Leigh 8/11/11 13:27
Formatted: Justified

matches (oh, sorry, I mean we invited them down for important meetings about books). Every player was only allowed to play in one match. Over many matches, we counted the number of players who scored goals. The data are in the file **SageEditorsCan'tPlayFootball.dat**. Do a chi-square test to see whether more publishers or academics scored goals. We predict that Sussex people will score more than Sage people.

Richard Leigh 8/11/11 13:27
Deleted: that

First of all, we need to load in the data by executing:

```
sageFootball<-read.delim("SageEditorsCan'tPlayFootball.dat", header = TRUE)
```

If we then view the dataframe, we can see that the data are not in the correct format for carrying out a chi-square test:

Richard Leigh 8/11/11 13:21
Deleted: chi sq

```
                Job            Score   Frequency
1    Sage Publications   Yes          5
2    Sage Publications    No         19
3 University of Sussex   Yes         23
4 University of Sussex    No         30
```

One way to overcome this issue is to enter the data as a contingency table by hand. We can do this by executing:

Richard Leigh 8/11/11 13:29
Formatted: Justified

```
Sage_Publications <- c(5, 19)
University_of_Sussex <- c(23, 30)
sagefootball_Table <- cbind(Sage_Publications, University_of_Sussex)
```

If we then look at the *sagefootball_Table* contingency table, we can see that the data are in the correct format for running a chi-square test (**NB**: I have entered the data so that 1 = Yes and 2 = No. It doesn't matter which way round you do this, but it is important to make a note of it, as it will be important in interpreting the data).

Richard Leigh 8/11/11 13:29
Formatted: Justified

Richard Leigh 8/11/11 13:28
Deleted: *table*

Richard Leigh 8/11/11 13:21
Deleted: chi sq

```
     Sage_Publications University_of_Sussex
[1,]                 5                   23
[2,]                19                   30
```

Now it's time for the best bit – we can run the chi-square by executing:

Richard Leigh 8/11/11 13:29
Deleted: -

Richard Leigh 8/11/11 13:21
Deleted: chi sq

Richard Leigh 8/11/11 13:29
Formatted: Normal Indent

```
CrossTable(sagefootball_Table, fisher = TRUE, chisq = TRUE, expected = TRUE,
sresid = TRUE, format = "SPSS")
```

```
 Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|            Std Residual |
|-------------------------|

Total Observations in Table:  77

             |
             |   Sage_Publications  | University_of_Sussex |            Row Total |
-------------|----------------------|----------------------|----------------------|
       [1,] |                    5 |                   23 |                   28 |
             |                8.727 |               19.273 |                      |
             |                1.592 |                0.721 |                      |
             |               17.857% |               82.143% |              36.364% |
             |               20.833% |               43.396% |                      |
             |                6.494% |               29.870% |                      |
```

```
             |              -1.262  |               0.849  |                      |
-------------|----------------------|----------------------|----------------------|
        [2,] |                19    |                30    |                49    |
             |            15.273    |            33.727    |                      |
             |             0.910    |             0.412    |                      |
             |            38.776%   |            61.224%   |             63.636%  |
             |            79.167%   |            56.604%   |                      |
             |            24.675%   |            38.961%   |                      |
             |             0.954    |            -0.642    |                      |
-------------|----------------------|----------------------|----------------------|
Column Total |                24    |                53    |                77    |
             |            31.169%   |            68.831%   |                      |
-------------|----------------------|----------------------|----------------------|

Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  3.634237     d.f. =  1     p =  0.05660253

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  2.724597     d.f. =  1     p =  0.09881305


Fisher's Exact Test for Count Data
------------------------------------------------------------
Sample estimate odds ratio:  0.3478333

Alternative hypothesis: true odds ratio is not equal to 1
p =  0.07478225
95% confidence interval:  0.08802237 1.157545

Alternative hypothesis: true odds ratio is less than 1
p =  0.04713642
95% confidence interval:  0 0.9855414

Alternative hypothesis: true odds ratio is greater than 1
p =  0.9867113
95% confidence interval:  0.1090477 Inf


       Minimum expected frequency: 8.727273
```

The crosstabulation table produced by **R** contains the number of cases that fall into each combination of categories. We can see that in total 28 people scored goals (36.37% of the total), of whom 5 were from Sage Publications (17.9% of the total who scored) and 23 were from Sussex (82.1% of the total who scored); 49 people didn't score at all (63.6% of the total) and, of those, 19 worked for Sage (38.8% of the total who didn't score) and 30 were from Sussex (61.2% of the total who didn't score).

Before moving on to look at the test statistic itself it is vital that we check that the assumption for chi-square has been met. The assumption is that in 2 × 2 tables (which is what we have here), all expected frequencies should be greater than 5. The smallest expected count is 8.7 (for Sage editors who scored). This value exceeds 5 and so the assumption has been met.

Pearson's chi-square test examines whether there is an association between two categorical variables (in this case the job and whether the person scored or not). The Pearson chi-square statistic tests whether the two variables are independent. If the significance value is small enough (conventionally it must be less than .05) then we reject the hypothesis that the variables are independent and accept the hypothesis that they are in some way related. The value of the chi-square statistic is given in the table (and the degrees of freedom), as is the significance value. The value of the chi-square statistic is 3.63. This value has a two-tailed significance of .057, which is bigger than

Richard Leigh 8/11/11 13:29
Formatted: Normal, Justified

Richard Leigh 8/11/11 13:30
Deleted: s

Richard Leigh 8/11/11 13:30
Deleted: and of these,

Richard Leigh 8/11/11 13:30
Deleted: that

Richard Leigh 8/11/11 13:30
Deleted: that

Richard Leigh 8/11/11 13:30
Deleted: that

Richard Leigh 8/11/11 13:30
Deleted: that

Richard Leigh 8/11/11 13:29
Formatted: Justified

Richard Leigh 8/11/11 13:31
Deleted: s

Richard Leigh 8/11/11 13:31
Deleted: Sig.

Richard Leigh 8/11/11 13:31
Formatted: Font:Not Italic

.05 (hence non-significant). However, we made a specific prediction (that Sussex people would score more than Sage people), hence we can halve this value. Therefore, the chi-square is significant (one-tailed) because $p = .0285$, which is less than .05. The one-tailed significance values of the other statistics are also less than .05 so we have consistent results.

The significant result indicates that there is an association between the type of job someone does and whether they score goals. This significant finding reflects the fact that for Sussex employees there is about a 50% split of those that scored and those that didn't, but for Sage employees there is about a 20–80 split with only 20% scoring and 80% not scoring. This supports our hypothesis that people from Sage, despite their delusions, are crap at football!

Calculating an effect size

The odds of someone scoring given that they were employed by Sage is 5/19 = 0.26, and the odds of someone scoring given that they were employed by Sussex University is 23/30 = 0.77. Therefore, the odds ratio is 0.26/0.77 = 0.34. In other words, the odds of scoring if you work for Sage are about a third as high as they are if you work for Sussex; to put it another way, if you work for Sussex, the odds of scoring are about three times as high as they are if you work for Sage!

Reporting the results of chi-square

We could report:

- There was a significant association between the type of job and whether or not a person scored a goal, $\chi^2(1) = 3.63$, $p < .05$ (one-tailed). This represents the fact that, based on the odds ratio, Sage employees were only one-third as likely to score compared to Sussex employees.

Task 2

- I wrote much of this update while on sabbatical in the Netherlands (I have a real soft spot for Holland). However, living there for three months did enable me to notice certain cultural differences between Holland and England. The Dutch are famous for travelling by bike; they do it much more than the English. However, I noticed that many more Dutch people cycle while steering with only one hand. I pointed this out to one of my friends, Birgit Mayer, and she said that I was being a crazy English fool and that Dutch people did not cycle one-handed. Several weeks of me pointing at one-handed cyclists and her pointing at two-handed cyclists ensued. To put it to the test I counted the number of Dutch and English cyclists who ride with one or two hands on the handlebars (**Handlebars.dat**). Can you work out which one of us is right?

First of all we must load in the data by executing:

```
handlebarsData<-read.delim("Handlebars.dat", header = TRUE)
```

If we then view the dataframe, we can see that the data are not in the correct format for carrying out a chi-square test:

```
    Hands      Nationality Frequency
1 One Handed       Dutch       120
2 One Handed     English        17
```

Richard Leigh 8/11/11 13:32
Deleted: -

Richard Leigh 8/11/11 13:32
Formatted: Justified

Richard Leigh 8/11/11 13:36
Deleted: 0.34 times higher than

Richard Leigh 8/11/11 13:37
Deleted: a better way to express this is that

Richard Leigh 8/11/11 13:37
Deleted: Sage

Richard Leigh 8/11/11 13:38
Deleted: 1/0.34 = 2.95 lower than if you work for Sussex

Richard Leigh 8/11/11 13:38
Formatted: Justified

Richard Leigh 8/11/11 13:39
Deleted: 2.95 times less

Richard Leigh 8/11/11 13:39
Deleted: than

Richard Leigh 8/11/11 13:40
Formatted: Justified

Richard Leigh 8/11/11 13:40
Deleted: to

Richard Leigh 8/11/11 13:41
Deleted: ether Birgit or I am

Richard Leigh 8/11/11 13:41
Formatted: Justified

Richard Leigh 8/11/11 13:21
Deleted: chi sq

```
3 Two Handed       Dutch       578
4 Two Handed     English       154
```

One way to overcome this issue is to enter the data as a contingency table by hand. We can do this by executing:

```
Dutch <- c(120, 578)
English <- c(17, 154)
handlebars_Table <- cbind(Dutch, English)
```

If we then look at the *handlebars* *Table* contingency table, we can see that the data are now in the correct format for running a chi-square test (**NB**: I have entered the data so that 1 = One Handed and 2 = Two Handed. It doesn't matter which way round you do this, but it is important to make a note of it, as it will be important when interpreting the data).

```
      Dutch  English
[1,]   120       17
[2,]   578      154
```

Now it's time to run the chi-square test. We can do this by executing:

```
CrossTable(handlebars_Table, fisher = TRUE, chisq = TRUE, expected = TRUE, sresid =
TRUE, format = "SPSS")
```

Richard Leigh 8/11/11 13:41
Deleted: *table*

Richard Leigh 8/11/11 13:21
Deleted: chi sq

Richard Leigh 8/11/11 13:41
Deleted: handed

Richard Leigh 8/11/11 13:42
Deleted: handed

Richard Leigh 8/11/11 13:42
Formatted: Normal Indent

Richard Leigh 8/11/11 13:21
Deleted: chi sq

```
Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|            Std Residual |
|-------------------------|

Total Observations in Table:  869

             |
             |     Dutch   |   English   | Row Total |
-------------|-------------|-------------|-----------|
       [1,]  |       120   |        17   |       137 |
             |   110.041   |    26.959   |           |
             |     0.901   |     3.679   |           |
             |    87.591%  |    12.409%  |    15.765% |
             |    17.192%  |     9.942%  |           |
             |    13.809%  |     1.956%  |           |
             |     0.949   |    -1.918   |           |
-------------|-------------|-------------|-----------|
       [2,]  |       578   |       154   |       732 |
             |   587.959   |   144.041   |           |
             |     0.169   |     0.689   |           |
             |    78.962%  |    21.038%  |    84.235% |
             |    82.808%  |    90.058%  |           |
             |    66.513%  |    17.722%  |           |
             |    -0.411   |     0.830   |           |
-------------|-------------|-------------|-----------|
Column Total |       698   |       171   |       869 |
             |    80.322%  |    19.678%  |           |
-------------|-------------|-------------|-----------|

Statistics for All Table Factors
```

```
Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  5.437138     d.f. =  1     p =  0.01971294

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  4.904869     d.f. =  1     p =  0.02678109


Fisher's Exact Test for Count Data
------------------------------------------------------------
Sample estimate odds ratio:  1.879536

Alternative hypothesis: true odds ratio is not equal to 1
p =  0.01911155
95% confidence interval:  1.085101 3.437067

Alternative hypothesis: true odds ratio is less than 1
p =  0.9946394
95% confidence interval:  0 3.119562

Alternative hypothesis: true odds ratio is greater than 1
p =  0.01091851
95% confidence interval:  1.173343 Inf


      Minimum expected frequency: 26.95857
```

The crosstabulation table produced by **R** contains the number of cases that fall into each combination of categories. We can see that in total 137 people rode their bike one-handed, of which 120 (87.6%) were Dutch and only 17 (12.4%) were English; 732 people rode their bike two-handed, of which 578 (79%) were Dutch and only 154 (21%) were English.

Before moving on to look at the test statistic itself it is vital that we check that the assumption for chi-square has been met. The assumption is that in 2 × 2 tables (which is what we have here), all expected frequencies should be greater than 5. If you look at the expected counts in the crosstabulation table, it should be clear that the smallest expected count is 27 (for English people who ride their bike one-handed). This value exceeds 5 and so the assumption has been met.

The value of the chi-square statistic is 5.44. This value has a two-tailed significance of .020, which is smaller than .05 (hence significant). This suggests that the pattern of bike riding (i.e. relative numbers of one- and two-handed riders) significantly differs in English and Dutch people.

The significant result indicates that there is an association between whether someone is Dutch or English and whether they ride their bike one- or two-handed. Looking at the frequencies, this finding seems to show that the ratio of one- to two-handed riders differs in Dutch and English people. In Dutch people 17.2% ride their bike one-handed compared to 82.8% who ride two-handed. In England, though, only 9.9% rode their bike one-handed (almost half as many as in Holland), and 90.1% rode their bikes two-handed. If we look at the standardized residuals we can see that the only cell with a residual approaching significance (a value that lies outside of ±1.96) is the cell for English people riding one-handed ($z = -1.9$). The fact that this value is negative tells us that *fewer* people than expected fell into this cell.

Calculating an effect size

The odds of someone riding one-handed if they are Dutch are 120/578 = 0.21, and the odds of someone riding one-handed if they are English is 17/154 = 0.11. Therefore, the odds ratio is 0.21/0.11 = 1.9. In other words, the odds of riding one-handed if you

Richard Leigh 8/11/11 13:42
Deleted: s

Richard Leigh 8/11/11 13:42
Formatted: Justified

Richard Leigh 8/11/11 13:42
Deleted: s

Richard Leigh 8/11/11 13:44
Formatted: Justified

Richard Leigh 8/11/11 13:43
Deleted: is

are Dutch are 1.9 times higher than if you are English (or the odds of riding one-handed if you are English are about half that of a Dutch person).

Richard Leigh 8/11/11 13:45
Deleted: is

Reporting the results of chi-square

We could report:

- There was a significant association between nationality and whether the Dutch or English rode their bike one- or two-handed, $\chi^2(1) = 5.44$, $p < .05$. This represents the fact that, based on the odds ratio, the odds of riding a bike one-handed were 1.9 times higher for Dutch people than English people. This supports Field's argument that there are more one-handed bike riders in the Netherlands than in England and utterly refutes Mayer's theory that Field is a complete idiot. These data are in no way made up.

Richard Leigh 8/11/11 13:45
Deleted: arse

Task 3

- I was interested in whether horoscopes are just a figment of people's minds. Therefore, I got 2201 people, made a note of their star sign (this variable, obviously, has 12 categories: Capricorn, Aquarius, Pisces, Aries, Taurus, Gemini, Cancer, Leo, Virgo, Libra, Scorpio and Sagittarius) and whether they believed in horoscopes (this variable has two categories: believer or unbeliever). I then sent them a horoscope in the post of what would happen over the next month: everybody, regardless of their star sign, received the same horoscope, which read 'August is an exciting month for you. You will make friends with a tramp in the first week of the month and cook him a cheese omelette. Curiosity is your greatest virtue, and in the second week, you'll discover knowledge of a subject that you previously thought was boring, statistics perhaps. You might purchase a book around this time that guides you towards this knowledge. Your new wisdom leads to a change in career around the third week, when you ditch your current job and become an accountant. By the final week you find yourself free from the constraints of having friends, your boy/girlfriend has left you for a Russian ballet dancer with a glass eye, and you now spend your weekends doing loglinear analysis by hand with a pigeon called Hephzibah for company.' At the end of August I interviewed all of these people and I classified the horoscope as having come true, or not, based on how closely their lives had matched the fictitious horoscope. The data are in the file **Horoscope.dat**. Conduct a loglinear analysis to see whether there is a relationship between the person's star sign, whether they believe in horoscopes and whether the horoscope came true.

Richard Leigh 8/11/11 13:48
Formatted: Justified

First of all set your working directory and load in the data:

```
horoscopeData<-read.delim("Horoscope.dat", header = TRUE)
```

We can view the data by executing the name of the dataframe (*horosocpeData*)

```
    Star_Sign    Believe            True                Frequency
1    Capricorn Unbeliever Horoscope Didn't Come True        56
2    Capricorn Unbeliever        Horoscope Came True        46
3    Capricorn   Believer Horoscope Didn't Come True        50
4    Capricorn   Believer        Horoscope Came True        60
5     Aquarius Unbeliever Horoscope Didn't Come True        26
6     Aquarius Unbeliever        Horoscope Came True        20
```

```
7     Aquarius   Believer Horoscope Didn't Come True        22
8     Aquarius   Believer        Horoscope Came True        29
9       Pisces Unbeliever Horoscope Didn't Come True        55
10      Pisces Unbeliever        Horoscope Came True        51
11      Pisces   Believer Horoscope Didn't Come True        64
12      Pisces   Believer        Horoscope Came True        70
13       Aries Unbeliever Horoscope Didn't Come True        42
14       Aries Unbeliever        Horoscope Came True        36
15       Aries   Believer Horoscope Didn't Come True        70
16       Aries   Believer        Horoscope Came True        54
17      Taurus Unbeliever Horoscope Didn't Come True        56
18      Taurus Unbeliever        Horoscope Came True        42
19      Taurus   Believer Horoscope Didn't Come True        41
20      Taurus   Believer        Horoscope Came True        50
21      Gemini Unbeliever Horoscope Didn't Come True        65
22      Gemini Unbeliever        Horoscope Came True        53
23      Gemini   Believer Horoscope Didn't Come True        40
24      Gemini   Believer        Horoscope Came True        48
25      Cancer Unbeliever Horoscope Didn't Come True        84
26      Cancer Unbeliever        Horoscope Came True        76
27      Cancer   Believer Horoscope Didn't Come True        96
28      Cancer   Believer        Horoscope Came True        83
29         Leo Unbeliever Horoscope Didn't Come True        14
30         Leo Unbeliever        Horoscope Came True        23
31         Leo   Believer Horoscope Didn't Come True        12
32         Leo   Believer        Horoscope Came True        20
33       Virgo Unbeliever Horoscope Didn't Come True        69
34       Virgo Unbeliever        Horoscope Came True        55
35       Virgo   Believer Horoscope Didn't Come True        49
36       Virgo   Believer        Horoscope Came True        66
37       Libra Unbeliever Horoscope Didn't Come True        27
38       Libra Unbeliever        Horoscope Came True        26
39       Libra   Believer Horoscope Didn't Come True        22
40       Libra   Believer        Horoscope Came True        36
41     Scorpio Unbeliever Horoscope Didn't Come True        32
42     Scorpio Unbeliever        Horoscope Came True        20
43     Scorpio   Believer Horoscope Didn't Come True        24
44     Scorpio   Believer        Horoscope Came True        32
45 Sagittarius Unbeliever Horoscope Didn't Come True        56
46 Sagittarius Unbeliever        Horoscope Came True        41
47 Sagittarius   Believer Horoscope Didn't Come True        42
48 Sagittarius   Believer        Horoscope Came True        50
```

As you can see from looking at the data above, they are not in the same format as the **CatsandDogs.dat** data from the book chapter. This is really not a problem, though, we don't need to reshape the data or anything! We can still create a contingency table using the same command that we used in the book chapter, except we need to put the variable containing the values (in this case **Frequency**) before the tilde. Therefore, to generate our contingency table using *xtabs()* for the *horoscopeData*, we could execute:

```
horoscopeContingencyTable<-xtabs(Frequency ~ Star_Sign + Believe + True, data =
horoscopeData)
```

This takes the original dataframe (*horoscopeDate*) and creates a contingency table based on the variables **Star_Sign**, **Believe** and **True**. The resulting contingency table is stored as *horoscopeContingencyTable*, which is what we'll use in the loglinear analysis; it looks like this:

```
True = Horoscope Came True

            Believe
Star_Sign    Believer Unbeliever
  Aquarius         29         20
  Aries            54         36
  Cancer           83         76
  Capricorn        60         46
  Gemini           48         53
  Leo              20         23
  Libra            36         26
```

Richard Leigh 8/11/11 13:49
Formatted: Justified

Richard Leigh 8/11/11 13:50
Deleted: execpt

Richard Leigh 8/11/11 13:51
Deleted:

Richard Leigh 8/11/11 13:51
Deleted: beofre

Richard Leigh 8/11/11 13:50
Deleted: tidle

Richard Leigh 8/11/11 13:50
Formatted: Font:Italic

Richard Leigh 8/11/11 13:51
Formatted: Normal, Justified

```
  Pisces            70         51
  Sagittarius       50         41
  Scorpio           32         20
  Taurus            50         42
  Virgo             66         55

, , True = Horoscope Didn't Come True

             Believe
Star_Sign     Believer Unbeliever
  Aquarius          22         26
  Aries             70         42
  Cancer            96         84
  Capricorn         50         56
  Gemini            40         65
  Leo               12         14
  Libra             22         27
  Pisces            64         55
  Sagittarius       42         56
  Scorpio           24         32
  Taurus            41         56
  Virgo             49         69
```

We start by estimating the saturated model, which we know will fit the data perfectly with a chi-square equal to zero. We'll call the model *horoscopeSaturated*. We can create this model in the same way as we did in the book chapter:

```
horoscopeSaturated<-loglm(Frequency ~ Star_Sign*Believe*True, data =
horoscopeContingencyTable)

summary(horoscopeSaturated)
```

The first command creates the model called *horoscopeSaturated* based on all main effects and interactions in the contingency table called *horoscopeContingencyTable*. The second command summarizes this model; the output below shows the main statistics, and, as we expect, it has a likelihood ratio of 0 and a *p*-value of 1, because it is a perfect fit of the data.

```
Formula:
Frequency ~ Star_Sign * Believe * True

Statistics:
                X^2 df P(> X^2)
Likelihood Ratio   0  0        1
Pearson            0  0        1
```

Next we'll fit the model with all of the main effects and two-way interactions. In other words, we'll remove the three-way interaction; because this model tells us the effect of removing the three-way interaction we'll call it *threeWay*. We could create this model by respecifying the model with all terms except the three-way interaction:

```
threeWay <- loglm(Frequency ~ Star_Sign + Believe + True + Star_Sign:Believe +
Star_Sign:True + True:Believe, data = horoscopeContingencyTable)
```

This command uses the same format as before to create a model called *threeWay*. The only difference (apart from that we have changed the name of the model) is that the three-way interaction isn't included. This is a lot of typing, so you could also consider using the *update()* function. Remember that this function allows us to take an existing model and 'update' it. In the past we have updated models by adding in new variables, but we can also remove them using this function. For example, to remove the three-way interaction from the saturated model we would execute:

```
threeWay<-update(horoscopeSaturated, .~. -Star_Sign:Believe:True)
```

Richard Leigh 8/11/11 13:50
Formatted: Justified

Richard Leigh 8/11/11 13:51
Formatted: Normal, Justified

Richard Leigh 8/11/11 13:50
Deleted: summarises

Richard Leigh 8/11/11 13:53
Deleted: ,

Richard Leigh 8/11/11 13:53
Deleted: two

Richard Leigh 8/11/11 13:53
Deleted: three

Richard Leigh 8/11/11 13:53
Deleted: three

Richard Leigh 8/11/11 13:53
Deleted: -

Richard Leigh 8/11/11 13:53
Deleted: three

Richard Leigh 8/11/11 13:54
Formatted: Normal, Justified

Richard Leigh 8/11/11 13:54
Deleted: three way

Richard Leigh 8/11/11 13:54
Deleted: three way

We can summarize this model by executing:

```
summary(threeWay)
```

The pertinent parts of the resulting output are below. You can see that both chi-square and likelihood ratio tests agree that removing this interaction will not significantly affect the fit of the model (because the probability value is greater than .05).

```
Formula:
Frequency ~ Star_Sign + Believe + True + Star_Sign:Believe +
    Star_Sign:True + Believe:True

Statistics:
                      X^2 df  P(> X^2)
Likelihood Ratio 8.841279 11 0.6365403
Pearson          8.850313 11 0.6357069
```

We can compare the saturated model to the model without the three-way interaction by executing:

```
anova(horoscopeSaturated, threeWay)
```

The resulting output shows the difference between these models. We're interested in the part called Delta: delta is Greek letter Δ, which is the equivalent of D, and is often used in statistics to mean 'difference'. In the column labelled *P(> Delta(Dev)* we see the *p*-value of the difference between the models. This value is greater than .05 and therefore is non-significant. What this is actually telling us is that the three-way interaction is not significant: removing it from the model does not have a significant effect on how well the model fits the data.

```
LR tests for hierarchical log-linear models

Model 1:
 Frequency ~ Believe + Star_Sign + True
Model 2:
 Frequency ~ Believe + Star_Sign + True

          Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   8.841279 11
Model 2   0.000000  0   8.841279        11        0.63654
Saturated 0.000000  0   0.000000         0        1.00000
```

Next, let's create models that systematically remove the two-way interactions:

```
BelieveTrue<-update(threeWay, .~. -Believe:True)
Star_SignTrue<-update(threeWay, .~. -Star_Sign:True)
Star_SignBelieve<-update(threeWay, .~. -Star_Sign:Believe)
```

The first command creates a model called *BelieveTrue* that takes the *threeWay* model and removes the **Believe** × **True** interaction (i.e., it does not include either this interaction or the three-way interaction). The second does the same but removes the **Star_Sign** × **True** interaction. The final command again takes the *threeWay* model but this time removes the **Star_Sign** × **Believe** interaction. We can compare all of these models to the model without the three-way interaction using the *anova()* function:

```
anova(threeWay, BelieveTrue)
anova(threeWay, Star_SignTrue)
anova(threeWay, Star_SignBelieve)
```

Richard Leigh 8/11/11 13:55
Deleted: summarise

Richard Leigh 8/11/11 13:56
Formatted: Indent: First line: 0.2 cm

Richard Leigh 8/11/11 13:56
Deleted: If you look at the two columns labelled *Prob* then y

Richard Leigh 8/11/11 13:55
Formatted: Justified

Richard Leigh 8/11/11 13:57
Formatted: Normal Indent, Justified

Richard Leigh 8/11/11 13:57
Deleted: one

Richard Leigh 8/11/11 13:57
Formatted: Normal, Justified

Richard Leigh 8/11/11 13:57
Deleted: go

Richard Leigh 8/11/11 13:57
Deleted: non

Richard Leigh 8/11/11 13:58
Formatted: Justified

Richard Leigh 8/11/11 13:58
Formatted: Normal, Justified

Richard Leigh 8/11/11 13:58
Deleted: that

Richard Leigh 8/11/11 14:02
Formatted: Font:Bold

Richard Leigh 8/11/11 14:02
Formatted: Font:Bold

Richard Leigh 8/11/11 14:02
Formatted: Font:Bold

Richard Leigh 8/11/11 14:02
Formatted: Font:Bold

Richard Leigh 8/11/11 14:02
Formatted: Font:Bold

Richard Leigh 8/11/11 14:02
Formatted: Font:Bold

Richard Leigh 8/11/11 13:54
Deleted: three way

The output below shows the result of the first comparison, which shows the effect of removing the **Believe** × **True** interaction: the likelihood ratio difference (or delta) is 12.54 with 1 degrees of freedom. This difference is significant, at p =.0004, and therefore we cannot remove the **Believe** × **True** interaction from the model without the fit getting worse (in other words, this interaction is significant).

```
LR tests for hierarchical log-linear models

Model 1:
 Frequency ~ Believe + Star_Sign + True
Model 2:
 Frequency ~ Believe + Star_Sign + True

           Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   21.382129 12
Model 2    8.841279 11  12.540851         1         0.00040
Saturated  0.000000  0   8.841279        11         0.63654
```

The next output below shows the effect of removing the **Star_Sign** × **True** effect. Now we get a likelihood ratio difference of 10.74, with 11 df. The p-value is greater than .05, and therefore we can remove the **Star_Sign** × **True** effect. What this is actually telling us is that the **Star_Sign** × **True** interaction is not significant: removing it from the model does not have a significant effect on how well the model fits the data.

```
LR tests for hierarchical log-linear models

Model 1:
 Frequency ~ Believe + Star_Sign + True
Model 2:
 Frequency ~ Believe + Star_Sign + True

           Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   19.581741 22
Model 2    8.841279 11  10.740462        11         0.46526
Saturated  0.000000  0   8.841279        11         0.63654
```

The next output below shows the effect of removing the **Star_Sign** × **Believe** interaction. The difference here is 20.67, with 11 df. This is significant (p < .05), and therefore this effect cannot be removed from the model without making the fit worse.

```
LR tests for hierarchical log-linear models

Model 1:
 Frequency ~ Believe + Star_Sign + True
Model 2:
 Frequency ~ Believe + Star_Sign + True

           Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   29.507059 22
Model 2    8.841279 11  20.665780        11         0.03700
Saturated  0.000000  0   8.841279        11         0.63654
```

To summarize, the **Star_Sign** × **Believe** (p = .037) and **Believe** × **True** (p ≤ .001) interactions are significant, but the **Star_Sign** × **True** interaction (p = .465) is not. Therefore, the non-significant **Star_Sign** × **True** interaction can be deleted, leaving the remaining two-way interactions in the model. Therefore, the final model is the one that retains all main effects and these two interactions. As neither of these interactions can be removed without affecting the model, and these interactions involve all three of the main effects (the variables **Star_Sign**, **True** and **Believe** are all involved in at least one of the remaining interactions), the main effects are not examined (because their effect is confounded with the interactions that have been retained).

Richard Leigh 8/11/11 13:58
Formatted: Justified
Richard Leigh 8/11/11 13:58
Deleted: us
Richard Leigh 8/11/11 14:02
Formatted ... [1]
Richard Leigh 8/11/11 13:59
Formatted: Justified
Richard Leigh 8/11/11 14:02
Formatted ... [2]
Richard Leigh 8/11/11 13:59
Deleted: >
Richard Leigh 8/11/11 13:59
Deleted: *
Richard Leigh 8/11/11 14:02
Formatted: Font:Bold
Richard Leigh 8/11/11 14:02
Formatted ... [3]
Richard Leigh 8/11/11 13:59
Deleted: *
Richard Leigh 8/11/11 14:02
Formatted: Font:Bold
Richard Leigh 8/11/11 13:59
Formatted: Justified
Richard Leigh 8/11/11 14:02
Formatted ... [4]
Richard Leigh 8/11/11 14:00
Formatted: Justified
Richard Leigh 8/11/11 13:59
Deleted: summarise
Richard Leigh 8/11/11 14:03
Deleted: **star**
Richard Leigh 8/11/11 14:03
Formatted: Font:Bold
Richard Leigh 8/11/11 14:03
Formatted ... [5]
Richard Leigh 8/11/11 14:04
Deleted: = …0…star sign ... [6]
Richard Leigh 8/11/11 14:03
Formatted: Font:Bold
Richard Leigh 8/11/11 14:03
Deleted: 0… …star sign ... [7]
Richard Leigh 8/11/11 14:03
Formatted: Font:Bold
Richard Leigh 8/11/11 14:04
Deleted: star sign
Richard Leigh 8/11/11 14:04
Formatted ... [8]

Let's look at the significant interaction between **Believe** and **True** by doing a chi-square on these variables. First we need to generate a contingency table for the variables believe and true by executing:

```
BelieveTrue_ContingencyTable<-xtabs(Frequency ~ Believe + True, data = horoscopeData)
```

We can then do the chi-square for **Believe × True** by executing:

```
CrossTable(BelieveTrue_ContingencyTable, fisher = TRUE, chisq = TRUE, expected = TRUE,
sresid = TRUE, format = "SPSS")
```

```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|            Std Residual |
|-------------------------|

Total Observations in Table:  2201

             | True
     Believe |        Horoscope Came True | Horoscope Didn't Come True|                 Row Total |
-------------|----------------------------|---------------------------|---------------------------|
    Believer |                        598 |                       532 |                      1130 |
             |                    558.069 |                   571.931 |                           |
             |                      2.857 |                     2.788 |                           |
             |                     52.920% |                    47.080% |                    51.340% |
             |                     55.014% |                    47.756% |                           |
             |                     27.169% |                    24.171% |                           |
             |                      1.690 |                    -1.670 |                           |
-------------|----------------------------|---------------------------|---------------------------|
  Unbeliever |                        489 |                       582 |                      1071 |
             |                    528.931 |                   542.069 |                           |
             |                      3.015 |                     2.941 |                           |
             |                     45.658% |                    54.342% |                    48.660% |
             |                     44.986% |                    52.244% |                           |
             |                     22.217% |                    26.443% |                           |
             |                     -1.736 |                     1.715 |                           |
-------------|----------------------------|---------------------------|---------------------------|
Column Total |                       1087 |                      1114 |                      2201 |
             |                     49.387% |                    50.613% |                           |
-------------|----------------------------|---------------------------|---------------------------|

Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  11.60103     d.f. =  1     p =  0.000659153

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  11.31232     d.f. =  1     p =  0.0007699444


Fisher's Exact Test for Count Data
------------------------------------------------------------
Sample estimate odds ratio:  1.337640

Alternative hypothesis: true odds ratio is not equal to 1
p =  0.0007502112
95% confidence interval:  1.127319 1.587629

Alternative hypothesis: true odds ratio is less than 1
p =  0.9997199
95% confidence interval:  0 1.545397

Alternative hypothesis: true odds ratio is greater than 1
p =  0.0003830682
95% confidence interval:  1.158060 Inf


      Minimum expected frequency: 528.9309
```

Looking at the resulting output above, we can see that the chi-square is highly significant. To interpret the **Believe × True** interaction we could consider calculating some odds ratios. First, the odds of the horoscope coming true given that the person was a believer were 598/532 = 1.12. However, the odds of the horoscope coming true given that the person was an unbeliever were 489/582 = 0.84. Therefore, the odds ratio is 1.12/0.84 = 1.33. We can interpret this by saying that the odds that a horoscope would come true were 1.33 as high in believers than non-believers. Given

Richard Leigh 8/11/11 14:00
Deleted: .

Richard Leigh 8/11/11 14:04
Deleted: first

Richard Leigh 8/11/11 14:04
Formatted: Font:Bold

Richard Leigh 8/11/11 14:04
Formatted: Font:Bold

Richard Leigh 8/11/11 14:04
Formatted: Normal Indent, Justified

Richard Leigh 8/11/11 14:00
Deleted: .

Richard Leigh 8/11/11 13:21
Deleted: chi sq

Richard Leigh 8/11/11 14:05
Deleted: believe*True

Richard Leigh 8/11/11 14:05
Formatted: Justified

Richard Leigh 8/11/11 14:05
Deleted: believe x true

Richard Leigh 8/11/11 14:05
Deleted: was

Richard Leigh 8/11/11 14:05
Deleted: was

Richard Leigh 8/11/11 14:05
Deleted: times

Richard Leigh 8/11/11 14:06
Deleted: er

that the horoscopes were made-up twaddle, this might be evidence that believers behave in ways to make their horoscopes come true!

Let's also do a chi-square to look at the significant interaction between **Star_Sign** and **Believe**. First, generate a contingency table by executing:

```
Star_SignBelieve_ContingencyTable<-xtabs(Frequency ~ Star_Sign + Believe, data =
horoscopeData)
```

We can then do the chi-square by executing:

```
CrossTable(Star_SignBelieve_ContingencyTable, fisher = TRUE, chisq = TRUE, expected =
TRUE, sresid = TRUE, format = "SPSS")
```

```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|            Std Residual |
|-------------------------|

Total Observations in Table:  2201

             | Believe
   Star_Sign |   Believer | Unbeliever |  Row Total |
-------------|------------|------------|------------|
    Aquarius |         51 |         46 |         97 |
             |     49.800 |     47.200 |            |
             |      0.029 |      0.031 |            |
             |     52.577% |     47.423% |      4.407% |
             |      4.513% |      4.295% |            |
             |      2.317% |      2.090% |            |
             |      0.170 |     -0.175 |            |
-------------|------------|------------|------------|
       Aries |        124 |         78 |        202 |
             |    103.707 |     98.293 |            |
             |      3.971 |      4.189 |            |
             |     61.386% |     38.614% |      9.178% |
             |     10.973% |      7.283% |            |
             |      5.634% |      3.544% |            |
             |      1.993 |     -2.047 |            |
-------------|------------|------------|------------|
      Cancer |        179 |        160 |        339 |
             |    174.044 |    164.956 |            |
             |      0.141 |      0.149 |            |
             |     52.802% |     47.198% |     15.402% |
             |     15.841% |     14.939% |            |
             |      8.133% |      7.269% |            |
             |      0.376 |     -0.386 |            |
-------------|------------|------------|------------|
   Capricorn |        110 |        102 |        212 |
             |    108.841 |    103.159 |            |
             |      0.012 |      0.013 |            |
             |     51.887% |     48.113% |      9.632% |
             |      9.735% |      9.524% |            |
             |      4.998% |      4.634% |            |
             |      0.111 |     -0.114 |            |
-------------|------------|------------|------------|
      Gemini |         88 |        118 |        206 |
             |    105.761 |    100.239 |            |
             |      2.983 |      3.147 |            |
             |     42.718% |     57.282% |      9.359% |
             |      7.788% |     11.018% |            |
             |      3.998% |      5.361% |            |
             |     -1.727 |      1.774 |            |
-------------|------------|------------|------------|
         Leo |         32 |         37 |         69 |
             |     35.425 |     33.575 |            |
             |      0.331 |      0.349 |            |
             |     46.377% |     53.623% |      3.135% |
```

Richard Leigh 8/11/11 14:05
Deleted: "

Richard Leigh 8/11/11 14:07
Formatted: Normal Indent, Justified

Richard Leigh 8/11/11 14:06
Deleted:

Richard Leigh 8/11/11 14:06
Formatted: Font:Bold

Richard Leigh 8/11/11 14:06
Deleted: believe

Richard Leigh 8/11/11 14:06
Deleted: believe

Richard Leigh 8/11/11 14:06
Deleted: "

Richard Leigh 8/11/11 14:06
Formatted: Font:Bold

Richard Leigh 8/11/11 14:06
Formatted: Font:Bold

Richard Leigh 8/11/11 14:06
Deleted: chi sq

```
             |      2.832% |      3.455% |             |
             |      1.454% |      1.681% |             |
             |     -0.575  |      0.591  |             |
-------------|-------------|-------------|-------------|
       Libra |          58 |          53 |         111 |
             |      56.988 |      54.012 |             |
             |       0.018 |       0.019 |             |
             |      52.252% |      47.748% |       5.043% |
             |       5.133% |       4.949% |             |
             |       2.635% |       2.408% |             |
             |       0.134 |      -0.138 |             |
-------------|-------------|-------------|-------------|
      Pisces |         134 |         106 |         240 |
             |     123.217 |     116.783 |             |
             |       0.944 |       0.996 |             |
             |      55.833% |      44.167% |      10.904% |
             |      11.858% |       9.897% |             |
             |       6.088% |       4.816% |             |
             |       0.971 |      -0.998 |             |
-------------|-------------|-------------|-------------|
 Sagittarius |          92 |          97 |         189 |
             |      97.033 |      91.967 |             |
             |       0.261 |       0.275 |             |
             |      48.677% |      51.323% |       8.587% |
             |       8.142% |       9.057% |             |
             |       4.180% |       4.407% |             |
             |      -0.511 |       0.525 |             |
-------------|-------------|-------------|-------------|
     Scorpio |          56 |          52 |         108 |
             |      55.448 |      52.552 |             |
             |       0.006 |       0.006 |             |
             |      51.852% |      48.148% |       4.907% |
             |       4.956% |       4.855% |             |
             |       2.544% |       2.363% |             |
             |       0.074 |      -0.076 |             |
-------------|-------------|-------------|-------------|
      Taurus |          91 |          98 |         189 |
             |      97.033 |      91.967 |             |
             |       0.375 |       0.396 |             |
             |      48.148% |      51.852% |       8.587% |
             |       8.053% |       9.150% |             |
             |       4.134% |       4.453% |             |
             |      -0.612 |       0.629 |             |
-------------|-------------|-------------|-------------|
       Virgo |         115 |         124 |         239 |
             |     122.703 |     116.297 |             |
             |       0.484 |       0.510 |             |
             |      48.117% |      51.883% |      10.859% |
             |      10.177% |      11.578% |             |
             |       5.225% |       5.634% |             |
             |      -0.695 |       0.714 |             |
-------------|-------------|-------------|-------------|
Column Total |        1130 |        1071 |        2201 |
             |      51.340% |      48.660% |             |
-------------|-------------|-------------|-------------|

Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  19.63405     d.f. =  11     p =  0.05061914
```

Looking at the output above, we can see that this chi-square is borderline significant (two-tailed, but then again we had no prediction so we need to look at the two-tailed significance). It doesn't make a lot of sense to compute odds ratios because there are so many star signs (although we could use one star sign as a base category and compute odds ratios for all other signs compared to this category). However, the obvious general interpretation of this effect is that the ratio of believers to unbelievers in certain star signs is different. For example, in most star signs there is a roughly 50–50 split of believers and unbelievers, but for Aries there is a 40–60 split, and it is probably this difference that is most contributing to the effect. However, it's

Richard Leigh 8/11/11 14:07
Formatted: Justified

Richard Leigh 8/11/11 14:07
Deleted: :

Richard Leigh 8/11/11 14:07
Deleted: :

important to keep this effect in perspective. It may not be that interesting that we happened to sample a different ratio of believers and unbelievers in certain star signs (unless you believe that certain star signs should have more cynical views of horoscopes than others!). We actually set out to find out something about whether the horoscopes would come true, and it's worth remembering that this interaction ignores the crucial variable that measured whether or not the horoscope came true!

Finally, we can evaluate the final model by running the loglinear analysis on the *horoscopeContingencyTable* as we did before but including only the main effects, the **Believe** × **True** and **Star_Sign** × **Believe** interactions. We can do this by executing:

```
horoscopeFinal<-loglm(Frequency ~ Star_Sign + Believe + True + Believe:True +
Star_Sign:Believe, data = horoscopeContingencyTable)

summary(horoscopeFinal)


Formula:
Frequency ~ Star_Sign + Believe + True + Believe:True + Star_Sign:Believe

Statistics:
                     X^2 df  P(> X^2)
Likelihood Ratio 19.58174 22 0.6091847
Pearson          19.53319 22 0.6122144
```

We're looking for a non-significant test statistic, which indicates that the expected values generated by the model are not significantly different from the observed data (put another way, the model is a good fit to the data). In this case the result is very non-significant, indicating that the model is a good fit to the data.

Reporting the results

For this example we could report:

- ✓ The three-way loglinear analysis produced a final model that retained the star sign × believe and believe × true interactions. The likelihood ratio of this model was $\chi^2(22) = 19.58$, $p = 0.61$. The star sign × believe interaction was significant, $\chi^2(11) = 20.67$, $p < .05$. This interaction indicates that the ratio of believers and unbelievers was different across the 12 star signs. In particular, the ratio in Aries (38.6–61.4 ratio of unbelievers to believers) was quite different than the other groups, which consistently had a roughly 50–50 split. The believe × true interaction was also significant, $\chi^2(1) = 12.54$, $p < .001$. The odds ratio indicated that the odds of the horoscope coming true were 1.33 times as high in believers as in non-believers. Given that the horoscopes were made-up twaddle, this might be evidence that believers behave in ways to make their horoscopes come true.

Task 4

- ✓ On my statistics course students have weekly classes in a computer laboratory. These classes are run by postgraduate tutors but I often pop in to help out. I've noticed in these sessions that many students are studying Facebook rather more than they are studying the very interesting statistics assignments that I have set them. I wanted to see the impact that this behaviour had on their exam performance. I collected data from all 260 students on my course. First I checked their **Attendance** and classified them as having attended either more

Richard Leigh 8/11/11 14:07
Deleted: -

Richard Leigh 8/11/11 14:08
Deleted: *

Richard Leigh 8/11/11 14:08
Deleted: *

Richard Leigh 8/11/11 14:08
Formatted: Font:Not Bold

Richard Leigh 8/11/11 14:08
Formatted: Normal, Justified

Richard Leigh 8/11/11 14:08
Deleted: of

Richard Leigh 8/11/11 14:09
Deleted: of

Richard Leigh 8/11/11 14:10
Formatted: Justified

Richard Leigh 8/11/11 14:10
Deleted: :

Richard Leigh 8/11/11 14:10
Deleted: to

Richard Leigh 8/11/11 14:10
Deleted: :

Richard Leigh 8/11/11 14:11
Deleted: more likely

Richard Leigh 8/11/11 14:11
Deleted: than

Richard Leigh 8/11/11 14:11
Formatted: Justified

or less than 50% of their lab classes. Next, I classified them as being either someone who looked at **Facebook** during their lab class, or someone who never did. Lastly, after the Research Methods in Psychology exam, I classified them as having either passed or failed (**Exam**). The data are in **Facebook.dat**. Do a loglinear analysis on the data to see if there is an association between studying Facebook and failing your exam.

Richard Leigh 1/11/11 13:51
Deleted: (RMiP)

First of all set your working directory and load in the data:

```
facebookData<-read.delim("Facebook.dat", header = TRUE)
```

We can view the data by executing the name of the dataframe (*facebookData*)

```
    Attendance                 Facebook  Exam    Frequency
1 More than 50%       Looked at Facebook Pass        39
2 More than 50%       Looked at Facebook Fail        30
3 More than 50% Did Not Look at Facebook Pass        98
4 More than 50% Did Not Look at Facebook Fail         5
5 Less than 50%       Looked at Facebook Pass         5
6 Less than 50%       Looked at Facebook Fail        30
7 Less than 50% Did Not Look at Facebook Pass        26
8 Less than 50% Did Not Look at Facebook Fail        27
```

As you can see from looking at the data above, they are not in the same format as the **CatsandDogs.dat** data from the book chapter. This is really not a problem, though, we don't need to reshape the data or anything! We can still create a contingency table using the same command that we used in the book chapter, except we need to put the variable containing the values (in this case **Frequency**) before the ~. Therefore, to generate our contingency table using *xtabs()* for the *facebookData*, we could execute:

Richard Leigh 8/11/11 14:12
Formatted: Justified

Richard Leigh 8/11/11 14:13
Deleted: execpt

Richard Leigh 8/11/11 14:13
Deleted:

Richard Leigh 8/11/11 14:13
Deleted: beofre

```
facebookContingencyTable<-xtabs(Frequency ~ Attendance + Facebook + Exam, data =
facebookData)
```

This takes the original dataframe (*facebookData*) and creates a contingency table based on the variables **Attendance**, **Facebook** and **Exam**. The resulting contingency table is stored as *facebookContingencyTable*, which is what we'll use in the loglinear analysis. It looks like this:

Richard Leigh 8/11/11 14:13
Formatted: Justified

Richard Leigh 8/11/11 14:13
Deleted: ; i

```
Exam = Fail

              Facebook
Attendance      Did Not Look at Facebook Looked at Facebook
  Less than 50%                       27                 30
  More than 50%                        5                 30

, , Exam = Pass

              Facebook
Attendance      Did Not Look at Facebook Looked at Facebook
  Less than 50%                       26                  5
  More than 50%                       98                 39
```

We start by estimating the saturated model, which we know will fit the data perfectly with a chi-square equal to zero. We'll call the model *facebookSaturated*. We can create this model in the same way as we did in the book chapter:

Richard Leigh 8/11/11 14:13
Formatted: Normal Indent, Justified

```
facebookSaturated<-loglm(Frequency ~ Attendance*Facebook*Exam, data =
facebookContingencyTable)

summary(facebookSaturated)
```

The first command creates the model called *facebookSaturated* based on all main effects and interactions in the contingency table called *facebookContingencyTable*. The second command summarizes this model; the output below shows the main statistics, and, as we expect, it has a likelihood ratio of 0 and a *p*-value of 1, because it is a perfect fit to the data.

```
Formula:
Frequency ~ Attendance * Facebook * Exam
Statistics:
                X^2 df P(> X^2)
Likelihood Ratio   0  0        1
Pearson            0  0        1
```

Next we'll fit the model with all of the main effects and two-way interactions. In other words, we'll remove the three-way interaction; because this model tells us the effect of removing the three-way interaction we'll call it *threeWay*. We could create this model by respecifying the model with all terms except the three-way interaction:

```
threeWay <- loglm(Frequency ~ Attendance + Facebook + Exam + Attendance:Facebook +
Attendance:Exam + Exam:Facebook, data = facebookContingencyTable)
```

This command uses the same format as before to create a model called *threeWay*. The only difference (apart from that we have changed the name of the model) is that the three-way interaction isn't included. This is a lot of typing, so you could also consider using the *update()* function. Remember that this function allows us to take an existing model and 'update' it. In the past we have updated models by adding in new variables, but we can also remove them using this function. For example, to remove the three-way interaction from the saturated model we would execute:

```
threeWay<-update(facebookSaturated, .~. -Attendance:Facebook:Exam)
```

We can summarize this model by executing:

```
summary(threeWay)
```

The pertinent parts of the resulting output are below. You can see that both chi-square and likelihood ratio tests agree that removing this interaction will not significantly affect the fit of the model (because the probability value is greater than .05).

```
Formula:
Frequency ~ Attendance + Facebook + Exam + Attendance:Facebook +
    Attendance:Exam + Facebook:Exam
Statistics:
                      X^2 df  P(> X^2)
Likelihood Ratio 1.572777  1 0.2098041
Pearson          1.628314  1 0.2019364
```

We can compare the saturated model to the one without the three-way interaction by executing:

```
anova(facebookSaturated, threeWay)
```

The resulting output shows the difference between these models. We're interested in the part called 'Delta'. In the column labelled *P(> Delta(Dev)* we see the *p*-value of the difference between the models. This value is greater than .05 (p = .21) and therefore is non-significant. What this is actually telling us is that the three-way interaction is not significant: removing it from the model does not have a significant effect on how well the model fits the data.

Richard Leigh 8/11/11 14:13
Formatted: Justified

Richard Leigh 8/11/11 14:14
Deleted: summarises

Richard Leigh 8/11/11 14:14
Deleted: ,

Richard Leigh 8/11/11 14:14
Deleted: of

Richard Leigh 8/11/11 14:14
Formatted: Normal Indent, Justified

Richard Leigh 8/11/11 14:14
Deleted: two way

Richard Leigh 8/11/11 13:54
Deleted: three way

Richard Leigh 8/11/11 13:54
Deleted: three way

Richard Leigh 8/11/11 14:15
Deleted: -

Richard Leigh 8/11/11 13:54
Deleted: three way

Richard Leigh 8/11/11 14:15
Formatted: Justified

Richard Leigh 8/11/11 13:54
Deleted: three way

Richard Leigh 8/11/11 13:54
Deleted: three way

Richard Leigh 8/11/11 14:15
Deleted: summarise

Richard Leigh 8/11/11 14:15
Deleted: If you look at the two columns labelled *Prob* then y

Richard Leigh 8/11/11 14:15
Formatted: Justified

Richard Leigh 8/11/11 14:16
Formatted: Normal Indent, Justified

Richard Leigh 8/11/11 14:16
Formatted: Normal, Justified

Richard Leigh 8/11/11 14:16
Deleted: : delta is Greek letter Δ, which is the equivalent of D, and is often used in statistics go mean 'difference'

Richard Leigh 8/11/11 14:16
Deleted: non

```
LR tests for hierarchical log-linear models

Model 1:
 Frequency ~ Facebook + Attendance + Exam
Model 2:
 Frequency ~ Facebook + Attendance + Exam

          Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   1.572777  1
Model 2   0.000000  0   1.572777         1         0.2098
Saturated 0.000000  0   0.000000         0         1.0000
```

Next, let's create models that systematically remove the two-way interactions:

```
FacebookExam<-update(threeWay, .~. -Facebook:Exam)
AttendanceExam<-update(threeWay, .~. -Attendance:Exam)
AttendanceFacebook<-update(threeWay, .~. -Attendance:Facebook)
```

The first command creates a model called *FacebookExam* that takes that *threeWay* model and removes the **Facebook** × **Exam** interaction (i.e., it does not include either this interaction or the three-way interaction). The second does the same but removes the **Attendance** × **Exam** interaction. The final command again takes the *threeWay* model but this time removes the **Attendance** × **Facebook** interaction. We can compare all of these models to the model without the three-way interaction using the *anova()* function:

```
anova(threeWay, FacebookExam)
anova(threeWay, AttendanceExam)
anova(threeWay, AttendanceFacebook)
```

The output below shows the result of the first comparison, which shows us the effect of removing the **Facebook** × **Exam** interaction. The **Facebook** × **Exam** interaction was significant, $\chi^2(1) = 49.77$, $p < .001$, indicating that whether you looked at Facebook or not affected exam performance. Therefore, we cannot remove the **Facebook** × **Exam** interaction from the model without the fit getting worse.

```
LR tests for hierarchical log-linear models

Model 1:
 Frequency ~ Facebook + Attendance + Exam
Model 2:
 Frequency ~ Facebook + Attendance + Exam

           Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   51.339103  2
Model 2    1.572777  1  49.766325         1         0.0000
Saturated  0.000000  0   1.572777         1         0.2098
```

The next output below shows the effect of removing the **Attendance** × **Exam** effect. The **Attendance** × **Exam** interaction was significant, $\chi^2(1) = 61.80$, $p < .0001$, indicating that whether you attended more or less than 50% of classes affected exam performance. Therefore we cannot remove the **Attendance** × **Exam** interaction from the model without the fit getting worse.

```
LR tests for hierarchical log-linear models

Model 1:
 Frequency ~ Facebook + Attendance + Exam
Model 2:
 Frequency ~ Facebook + Attendance + Exam

           Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   63.375333  2
```

Richard Leigh 8/11/11 14:16
Formatted: Normal Indent, Justified

Richard Leigh 8/11/11 14:17
Formatted: Justified

Richard Leigh 8/11/11 14:17
Formatted: Font:Bold

Richard Leigh 8/11/11 14:17
Formatted: Font:Bold

Richard Leigh 8/11/11 14:17
Formatted: Font:Bold

Richard Leigh 8/11/11 14:17
Formatted: Font:Bold

Richard Leigh 8/11/11 14:17
Formatted: Font:Bold

Richard Leigh 8/11/11 14:17
Formatted: Font:Bold

Richard Leigh 8/11/11 13:54
Deleted: three way

Richard Leigh 8/11/11 14:17
Formatted: Normal Indent, Justified

Richard Leigh 8/11/11 14:17
Formatted: Font:Bold

Richard Leigh 8/11/11 14:17
Formatted: Font:Bold

Richard Leigh 8/11/11 14:17
Formatted: Font:Bold

Richard Leigh 8/11/11 14:17
Formatted: Font:Bold

Richard Leigh 8/11/11 14:17
Formatted: Font:Bold

Richard Leigh 8/11/11 14:17
Formatted: Font:Bold

Richard Leigh 8/11/11 14:18
Formatted: Normal Indent, Justified

Richard Leigh 8/11/11 14:18
Formatted: Font:Bold

Richard Leigh 8/11/11 14:18
Formatted: Font:Bold

Richard Leigh 8/11/11 14:18
Formatted: Font:Bold

Richard Leigh 8/11/11 14:18
Formatted: Font:Bold

Richard Leigh 8/11/11 14:18
Formatted: Font:Bold

Richard Leigh 8/11/11 14:18
Formatted: Font:Bold

```
Model 2     1.572777  1  61.802555          1          0.0000
Saturated   0.000000  0   1.572777          1          0.2098
```

The final output below shows the effect of removing the **Attendance × Facebook** interaction. The difference here is 11.90, with 1 df. This is significant ($p < .001$), and therefore this effect cannot be removed from the model without making the fit worse.

```
LR tests for hierarchical log-linear models

Model 1:
 Frequency ~ Facebook + Attendance + Exam
Model 2:
 Frequency ~ Facebook + Attendance + Exam

           Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   13.469934  2
Model 2    1.572777  1  11.897156          1        0.00056
Saturated  0.000000  0   1.572777          1        0.20980
```

To summarise, the two-way interactions were all found to be significant, and therefore the final model is the one that retains all main effects and two-way interactions. However, the main effects are not examined because we have significant higher-order two-way effects, which are more interesting than the main effects.

To interpret the two-way interactions, we can do some chi-square tests. Let's first look at the significant interaction between **Facebook** and **Exam**. First, generate a contingency table by executing:

```
FacebookExam_ContingencyTable<-xtabs(Frequency ~ Facebook + Exam, data = facebookData)
```

We can then do the chi-square by executing:

```
CrossTable(FacebookExam_ContingencyTable, fisher = TRUE, chisq = TRUE, expected =
TRUE, sresid = TRUE, format = "SPSS")
```

```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|            Std Residual |
|-------------------------|

Total Observations in Table:  260

                         | Exam
                Facebook |      Fail |      Pass | Row Total |
-------------------------|-----------|-----------|-----------|
Did Not Look at Facebook |        32 |       124 |       156 |
                         |    55.200 |   100.800 |           |
                         |     9.751 |     5.340 |           |
                         |   20.513% |   79.487% |   60.000% |
                         |   34.783% |   73.810% |           |
                         |   12.308% |   47.692% |           |
                         |    -3.123 |     2.311 |           |
```

Richard Leigh 8/11/11 14:18
Deleted: next

Richard Leigh 8/11/11 14:18
Formatted: Normal Indent, Justified

Richard Leigh 8/11/11 14:19
Formatted: Font:Bold

Richard Leigh 8/11/11 14:19
Formatted: Font:Bold

Richard Leigh 8/11/11 14:19
Formatted: Normal Indent, Justified

Richard Leigh 8/11/11 14:19
Deleted: ,

Richard Leigh 8/11/11 14:14
Deleted: two way

Richard Leigh 8/11/11 14:19
Deleted: higher

Richard Leigh 8/11/11 14:14
Deleted: two way

Richard Leigh 8/11/11 14:20
Deleted:

Richard Leigh 8/11/11 14:20
Formatted: Justified

Richard Leigh 8/11/11 14:14
Deleted: two way

Richard Leigh 8/11/11 13:21
Deleted: chi sq

Richard Leigh 8/11/11 14:20
Deleted: let's

Richard Leigh 8/11/11 14:20
Deleted:

Richard Leigh 8/11/11 13:21
Deleted: chi sq

Richard Leigh 8/11/11 14:20
Deleted:

```
-------------------------|-----------|-----------|-----------|
      Looked at Facebook |        60 |        44 |       104 |
                         |    36.800 |    67.200 |           |
                         |    14.626 |     8.010 |           |
                         |    57.692% |    42.308% |    40.000% |
                         |    65.217% |    26.190% |           |
                         |    23.077% |    16.923% |           |
                         |     3.824 |    -2.830 |           |
-------------------------|-----------|-----------|-----------|
            Column Total |        92 |       168 |       260 |
                         |    35.385% |    64.615% |           |
-------------------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  37.72602     d.f. =  1     p =  8.141145e-10

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  36.11742     d.f. =  1     p =  1.857787e-09


Fisher's Exact Test for Count Data
------------------------------------------------------------
Sample estimate odds ratio:  0.1906368

Alternative hypothesis: true odds ratio is not equal to 1
p =  1.326512e-09
95% confidence interval:  0.1050906 0.3397684

Alternative hypothesis: true odds ratio is less than 1
p =  8.798576e-10
95% confidence interval:  0 0.3119569

Alternative hypothesis: true odds ratio is greater than 1
p =  1
95% confidence interval:  0.1151419 Inf


       Minimum expected frequency: 36.8
```

The output above shows that those who looked at Facebook had a much lower chance of passing their exam (58% failed) than those who didn't look at Facebook during their lab classes (around 80% passed).

Next let's do a chi-square to look at the significant interaction between **Attendance** and **Exam**. First, generate a contingency table by executing:

```
AttendanceExam_ContingencyTable <-xtabs(Frequency ~ Attendance + Exam, data =
facebookData)
```

We can then do the chi-square by executing:

```
CrossTable(AttendanceExam_ContingencyTable, fisher = TRUE, chisq = TRUE, expected
= TRUE, sresid = TRUE, format = "SPSS")
```

```
  Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|            Std Residual |
|-------------------------|
```

Richard Leigh 8/11/11 14:20
Formatted: Justified

Richard Leigh 8/11/11 14:20
Deleted:

Richard Leigh 8/11/11 14:21
Formatted: Normal Indent, Justified

Richard Leigh 8/11/11 14:21
Deleted:

Professor Andy Field 14/10/13 12:48
Deleted: Facebook +

Richard Leigh 8/11/11 13:21
Deleted: chi sq

```
Total Observations in Table:  260

              | Exam
   Attendance |      Fail |      Pass | Row Total |
--------------|-----------|-----------|-----------|
Less than 50% |        57 |        31 |        88 |
              |    31.138 |    56.862 |           |
              |    21.479 |    11.762 |           |
              |    64.773% |    35.227% |    33.846% |
              |    61.957% |    18.452% |           |
              |    21.923% |    11.923% |           |
              |     4.635 |    -3.430 |           |
--------------|-----------|-----------|-----------|
More than 50% |        35 |       137 |       172 |
              |    60.862 |   111.138 |           |
              |    10.989 |     6.018 |           |
              |    20.349% |    79.651% |    66.154% |
              |    38.043% |    81.548% |           |
              |    13.462% |    52.692% |           |
              |    -3.315 |     2.453 |           |
--------------|-----------|-----------|-----------|
 Column Total |        92 |       168 |       260 |
              |    35.385% |    64.615% |           |
--------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  50.2482     d.f. =  1     p =  1.354788e-12

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  48.32401     d.f. =  1     p =  3.613004e-12


Fisher's Exact Test for Count Data
------------------------------------------------------------
Sample estimate odds ratio:  7.130104

Alternative hypothesis: true odds ratio is not equal to 1
p =  4.218392e-12
95% confidence interval:  3.902149 13.30689

Alternative hypothesis: true odds ratio is less than 1
p =  1
95% confidence interval:  0 12.09419

Alternative hypothesis: true odds ratio is greater than 1
p =  2.172352e-12
95% confidence interval:  4.263635 Inf


       Minimum expected frequency: 31.13846
```

The output above shows that those who attended more than half of their classes had a much better chance of passing their exam (80% passed) than those attending less than 50% of classes (only 35% passed). All of the standardized residuals are significant, indicating that all cells contribute to this overall association.

The three-way **Facebook** × **Attendance** × **Exam** interaction was not significant, $\chi^2(1) = 1.57$, $p = .21$. This result indicates that the effect of Facebook (described above) was the same (roughly) in those who attended more than 50% of classes and those that attended less than 50% of classes. In other words, although those attending less than 50% of classes did worse than those attending, within that group, those looking at Facebook did relatively worse than those not looking at Facebook.

Richard Leigh 8/11/11 14:21
Formatted: Normal, Justified

Richard Leigh 8/11/11 14:21
Formatted: Font:Bold

Richard Leigh 8/11/11 14:21
Formatted: Justified

Richard Leigh 8/11/11 14:21
Formatted: Font:Bold

Richard Leigh 8/11/11 14:21
Formatted: Font:Bold

Richard Leigh 8/11/11 14:21
Deleted: