Stats Exam 2016: Question 1

Sergii Gladchuk

December 23, 2016

Background

Professor L.R. Plupp has his main research area in the north of Sweden, where he studies his favourite animals: lemmings (Lemmus lemmus).Over the years, he has noticed a difference in the colouring of the lemmings. Some individuals are redder than others, which are darker, or black. Moreover, he had the feeling that he saw black lemmings more often in overgrown habitats, and red lemmings in open habitat. To test this hypothesis he set out traps to collect lemmings. He set some traps in open habitat and some traps in overgrown habitats. He collected in total 80 individuals and checked them all for colour. He also noted the sex of all individuals. You find the results in the file "lemmingcolours" (.sav and .csv).

This is clear from description that analysis of frequencies should be used in order to test the hypothesis of professor L.R. Plupp.
First things first - upload and check the data:

```
lemming <- read.csv('~/Documents/courses/stats/Exam/lemmingcolours.csv', sep=';')
str(lemming)
```

```
## 'data.frame':    80 obs. of  3 variables:
##  $ habitat: Factor w/ 2 levels "open","overgrown": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex    : Factor w/ 2 levels "female","male": 1 1 1 1 2 1 1 1 1 1 ...
##  $ colour : Factor w/ 2 levels "black","red": 2 2 2 2 2 2 2 2 2 2 ...
```

All variables are factors and no missing data.

Sub-question a)

If you ignore sex, is there a significant difference in the colour morph distribution between the two habitat types?

To test hypothesis that color is dependent on habitat type, the frequency data should be converted into contingency table and then χ^2-test, G-test and Fishers' exact test performed.

```
counts <- table(lemming$habitat,lemming$colour)
counts
```

```
##
##               black red
##    open          12  23
##    overgrown     31  14
```

```
chisq.test(counts)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  counts
## X-squared = 8.1418, df = 1, p-value = 0.004326
```

```r
library(DescTools)
GTest(counts)
```

```
##
##  Log likelihood ratio (G-test) test of independence without
##  correction
##
## data:  counts
## G = 9.6504, X-squared df = 1, p-value = 0.001893
```

```r
fisher.test(counts)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  counts
## p-value = 0.003154
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.08242649 0.66381708
## sample estimates:
## odds ratio
##  0.2403215
```

The null hypothesis is that *color* of lemmings is independent of *habitat*. All three tests p-value less than 0.05, so null hypothesis can be rejected. In other words professor was right - there is clear dependence between color and habitat.

```r
hab.exp <- matrix(NA,2,2)
for (i in 1:length(counts[1,])) {
  for (j in 1:length(counts[,1])){
    hab.exp[i,j] <- sum(counts[i,])/sum(counts) *
      (sum(counts[,j])/sum(counts)) * sum(counts);
  }
}
hab.exp
```

```
##         [,1]    [,2]
## [1,] 18.8125 16.1875
## [2,] 24.1875 20.8125
```

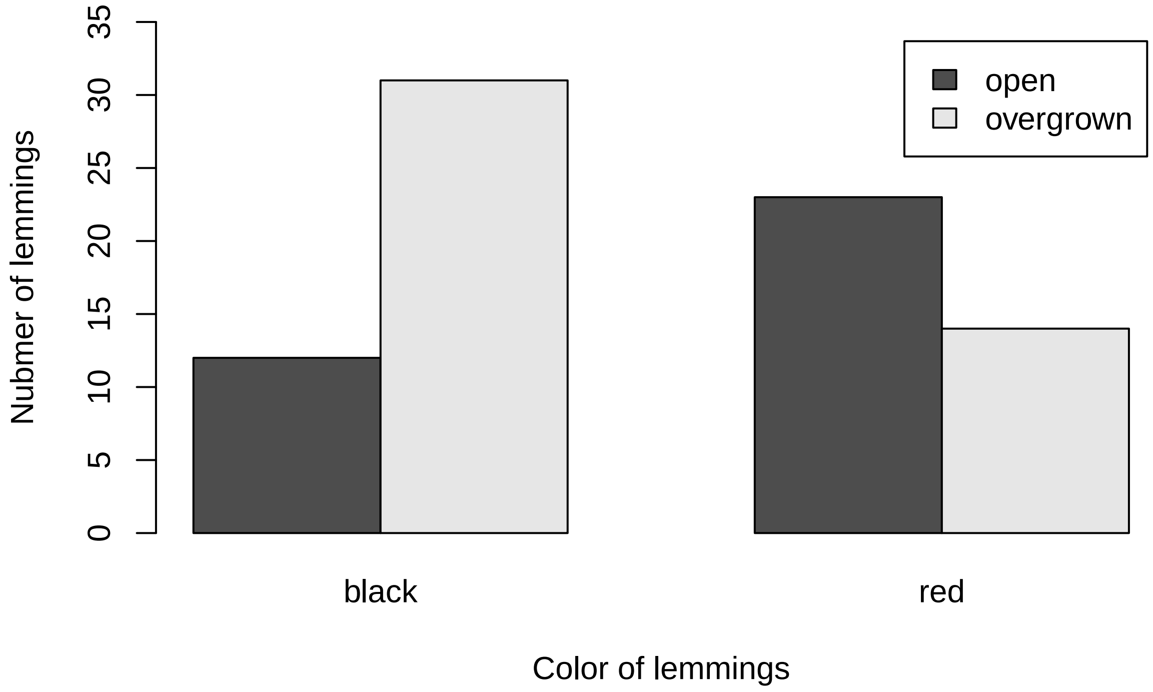
In this particular case (since there are no expected counts with less than 5) only G-test can be used here.

Sub-question b)

Make a suitable illustration of your result in a)

Bar-plot is a nice choice to visualize the result

```
barplot(counts, xlab='Color of lemmings', ylab='Nubmer of lemmings',
        beside=TRUE, legend=rownames(counts), ylim=c(0,35))
```



The plot clearly shows that proportion of open and overgrown habitat for different colors are opposite - so there is dependence.

Sub-question c)

Ignoring the habitat, is there a colour difference between the sexes?

New contingency table is needed for color and sexes

```
counts2 <- table(lemming$sex,lemming$colour)
counts2
```

```
##
##          black red
##   female     6  24
##   male      37  13
```

```
chisq.test(counts2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
```

```
##
## data:  counts2
## X-squared = 19.875, df = 1, p-value = 8.267e-06
```

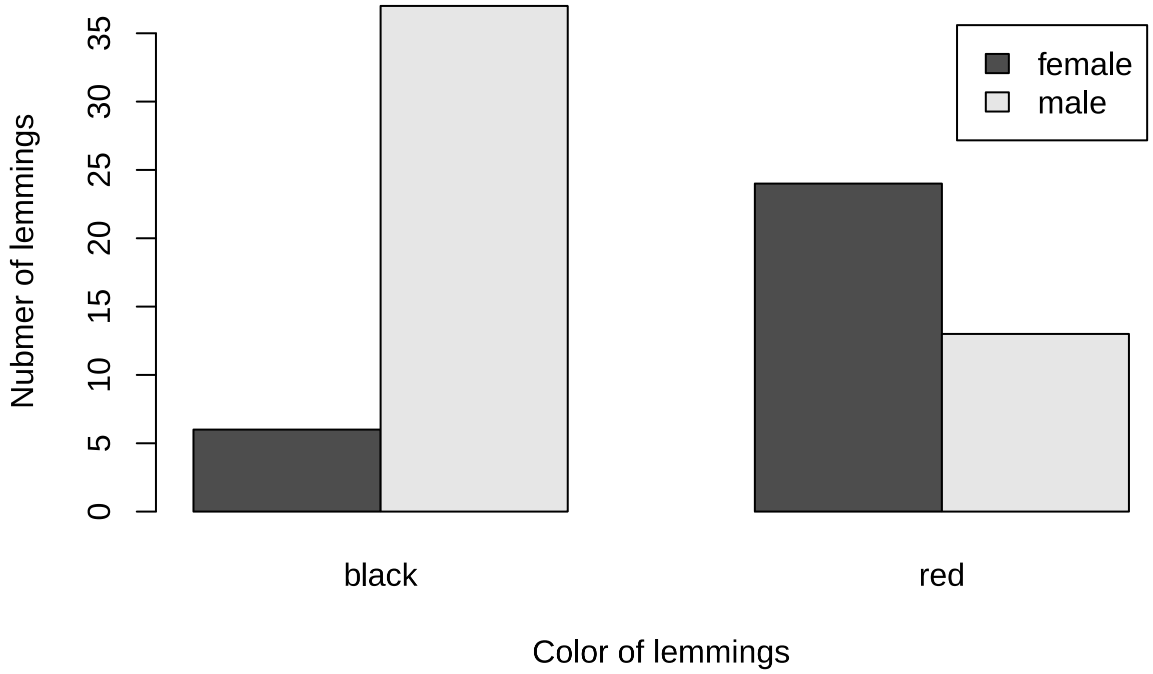
```
library(DescTools)
GTest(counts2)
```

```
##
##  Log likelihood ratio (G-test) test of independence without
##  correction
##
## data:  counts2
## G = 23.123, X-squared df = 1, p-value = 1.519e-06
```

```
fisher.test(counts2)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  counts2
## p-value = 2.905e-06
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.02458519 0.29152789
## sample estimates:
## odds ratio
## 0.09119479
```

```
barplot(counts2, xlab='Color of lemmings', ylab='Nubmer of lemmings',
        beside=TRUE, legend=rownames(counts2))
```



Dependence between *sex* and *color* is even more significant than in previous pair. The p-values are around e-06 for all 3 tests. The bar-plot supports the alternative hypothesis that there is more red females and black males than red males and black females.

```
hab.exp2 <- matrix(NA,2,2)
for (i in 1:length(counts2[1,])) {
  for (j in 1:length(counts2[,1])){
    hab.exp2[i,j] <- sum(counts2[i,])/sum(counts2) *
      (sum(counts2[,j])/sum(counts2)) * sum(counts2);
  }
}
hab.exp2
```

```
##        [,1]   [,2]
## [1,] 16.125 13.875
## [2,] 26.875 23.125
```

Same as previously, since there are no expected counts with less than 5 only G-test can be used here.

Sub-question d)

Using all the data, what are your conclusions? Do red lemmings prefer open habitat or can you find other relationships in the data? (Here some sort of model selection procedure would be appropriate)

Based on pairwise analysis of color dependence on sex and habitat, there might be a good chance of interaction between sex and habitat or some other iterations and dependencies. So analysis of table using log-linear

models would be appropriate.
So procedure will be:

1. Convert frequency data to three-way table
2. Run model selection routine with *glm* function based on full model. Poisson distribution should be used, since the data is counts.
3. Compere reduced best.model to full model

```
lemming$n <- 1
three.way.table <- aggregate(lemming$n,
                             by=list(lemming$habitat,lemming$sex,lemming$colour), sum)
colnames(three.way.table) <- colnames(lemming)
three.way.table
```

```
##     habitat    sex colour  n
## 1      open female  black  5
## 2 overgrown female  black  1
## 3      open   male  black  7
## 4 overgrown   male  black 30
## 5      open female    red 20
## 6 overgrown female    red  4
## 7      open   male    red  3
## 8 overgrown   male    red 10
```

```
library(glmulti)
```

```
## Loading required package: rJava
```

```
full.model <- glm(n~habitat*sex*colour, data=three.way.table, family = poisson)
summary(full.model)
```

```
##
## Call:
## glm(formula = n ~ habitat * sex * colour, family = poisson, data = three.way.table)
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0  0  0
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      1.609e+00  4.472e-01   3.599  0.00032
## habitatovergrown                -1.609e+00  1.095e+00  -1.469  0.14177
## sexmale                          3.365e-01  5.855e-01   0.575  0.56554
## colourred                        1.386e+00  5.000e-01   2.773  0.00556
## habitatovergrown:sexmale         3.065e+00  1.173e+00   2.612  0.00899
## habitatovergrown:colourred      -4.676e-11  1.225e+00   0.000  1.00000
## sexmale:colourred               -2.234e+00  8.522e-01  -2.621  0.00877
## habitatovergrown:sexmale:colourred -2.513e-01  1.452e+00  -0.173  0.86263
##
## (Intercept)                     ***
## habitatovergrown
## sexmale
## colourred                          **
## habitatovergrown:sexmale           **
## habitatovergrown:colourred
## sexmale:colourred                  **
## habitatovergrown:sexmale:colourred
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance:  6.2558e+01  on 7  degrees of freedom
## Residual deviance: -4.2188e-15  on 0  degrees of freedom
## AIC: 45.789
##
## Number of Fisher Scoring iterations: 3
```

```
model.sel <- glmulti(full.model, level=2, crit = 'aic')
```

```
## Initialization...
## TASK: Exhaustive screening of candidate set.
## Fitting...
## Completed.
```

```
summary(model.sel)
```

```
## $name
## [1] "glmulti.analysis"
##
## $method
## [1] "h"
##
## $fitting
## [1] "glm"
##
## $crit
## [1] "aic"
##
## $level
## [1] 2
##
## $marginality
## [1] FALSE
##
## $confsetsize
## [1] 100
##
## $bestic
## [1] 27.61266
##
## $icvalues
##  [1] 27.61266 56.90885 62.33809 62.47613 63.30690 63.88326 64.22654
##  [8] 64.33198 64.42453 64.48420 65.72792 65.82769 66.07536 66.24537
## [15] 66.27944 67.04038 67.67125 68.32562
##
## $bestmodel
## [1] "n ~ 1 + habitat + sex + colour + sex:habitat + colour:habitat + "
## [2] "    colour:sex"
##
## $modelweights
##  [1] 9.999994e-01 4.349248e-07 2.880512e-08 2.688404e-08 1.774574e-08
##  [6] 1.330267e-08 1.120461e-08 1.062922e-08 1.014854e-08 9.850254e-09
## [11] 5.289041e-09 5.031676e-09 4.445613e-09 4.083315e-09 4.014346e-09
## [16] 2.743972e-09 2.001651e-09 1.443090e-09
##
## $includeobjects
## [1] TRUE
```

```r
weightable(model.sel)
```

```
##                                                                          model
## 1  n ~ 1 + habitat + sex + colour + sex:habitat + colour:habitat + colour:sex
## 2                   n ~ 1 + habitat + sex + colour + sex:habitat + colour:sex
## 3                                         n ~ 1 + habitat + sex + sex:habitat
## 4                                                                       n ~ 1
## 5               n ~ 1 + habitat + sex + colour + sex:habitat + colour:habitat
## 6                                                                 n ~ 1 + sex
## 7                                n ~ 1 + habitat + sex + colour + sex:habitat
## 8                                                             n ~ 1 + habitat
## 9                                                              n ~ 1 + colour
## 10                                          n ~ 1 + sex + colour + colour:sex
## 11                                                      n ~ 1 + habitat + sex
## 12                                                       n ~ 1 + sex + colour
## 13               n ~ 1 + habitat + sex + colour + colour:habitat + colour:sex
## 14                                n ~ 1 + habitat + sex + colour + colour:sex
## 15                                                   n ~ 1 + habitat + colour
## 16                                  n ~ 1 + habitat + colour + colour:habitat
## 17                                             n ~ 1 + habitat + sex + colour
## 18                            n ~ 1 + habitat + sex + colour + colour:habitat
##         aic      weights
## 1  27.61266 9.999994e-01
## 2  56.90885 4.349248e-07
## 3  62.33809 2.880512e-08
## 4  62.47613 2.688404e-08
## 5  63.30690 1.774574e-08
## 6  63.88326 1.330267e-08
## 7  64.22654 1.120461e-08
## 8  64.33198 1.062922e-08
## 9  64.42453 1.014854e-08
## 10 64.48420 9.850254e-09
## 11 65.72792 5.289041e-09
## 12 65.82769 5.031676e-09
## 13 66.07536 4.445613e-09
## 14 66.24537 4.083315e-09
## 15 66.27944 4.014346e-09
## 16 67.04038 2.743972e-09
## 17 67.67125 2.001651e-09
## 18 68.32562 1.443090e-09
```

```
best.model <- glm(n ~ habitat + sex + colour + sex:habitat
                  + colour:habitat + colour:sex,
                  data=three.way.table, family = poisson)
summary(best.model)
```

```
##
## Call:
## glm(formula = n ~ habitat + sex + colour + sex:habitat + colour:habitat +
##     colour:sex, family = poisson, data = three.way.table)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7
##  0.05489  -0.11711  -0.04574   0.02225  -0.02716   0.06149   0.07125
##        8
## -0.03834
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  1.5848     0.4292   3.692 0.000222 ***
## habitatovergrown            -1.4699     0.7089  -2.074 0.038118 *
## sexmale                      0.3784     0.5352   0.707 0.479618
## colourred                    1.4170     0.4718   3.003 0.002672 **
## habitatovergrown:sexmale     2.9039     0.6878   4.222 2.42e-05 ***
## habitatovergrown:colourred  -0.1765     0.6558  -0.269 0.787853
## sexmale:colourred           -2.3230     0.6857  -3.388 0.000704 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 62.558371  on 7  degrees of freedom
## Residual deviance:  0.030381  on 1  degrees of freedom
## AIC: 43.82
##
## Number of Fisher Scoring iterations: 3
```

```
anova(best.model, test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: n
##
## Terms added sequentially (first to last)
##
##
##                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                7     62.558
## habitat          1    1.253         6     61.305  0.262928
## sex              1    5.053         5     56.252  0.024577 *
## colour          1    0.450         4     55.801  0.502134    
## habitat:sex     1   32.576         3     23.225 1.146e-08 ***
## habitat:colour  1    9.650         2     13.574  0.001893 ** 
## sex:colour      1   13.544         1      0.030  0.000233 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
second.best.model <- glm(n  ~ habitat + sex + colour + sex:habitat + colour:sex,
                         data=three.way.table, family = poisson)
summary(second.best.model)
```

```
## 
## Call:
## glm(formula = n ~ habitat + sex + colour + sex:habitat + colour:sex, 
##     family = poisson, data = three.way.table)
## 
## Deviance Residuals: 
##        1         2         3         4         5         6         7  
##  0.00000   0.00000  -0.14840   0.07336   0.00000   0.00000   0.24208  
##        8  
## -0.12484  
## 
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                1.6094     0.4163   3.866 0.000111 ***
## habitatovergrown          -1.6094     0.4899  -3.285 0.001019 ** 
## sexmale                    0.3920     0.5295   0.740 0.459049    
## colourred                  1.3863     0.4564   3.037 0.002388 ** 
## habitatovergrown:sexmale   2.9957     0.6042   4.959 7.10e-07 ***
## sexmale:colourred         -2.4323     0.5588  -4.352 1.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 62.55837  on 7  degrees of freedom
## Residual deviance:  0.10159  on 2  degrees of freedom
## AIC: 41.891
## 
## Number of Fisher Scoring iterations: 4
```

```
anova(second.best.model, test='Chisq')
```

```
## Analysis of Deviance Table
## 
## Model: poisson, link: log
## 
## Response: n
## 
## Terms added sequentially (first to last)
## 
## 
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
## NULL                            7     62.558
## habitat      1    1.253         6     61.305   0.26293
## sex          1    5.053         5     56.252   0.02458 *
## colour       1    0.450         4     55.801   0.50213
## habitat:sex  1   32.576         3     23.225 1.146e-08 ***
## sex:colour   1   23.123         2      0.102 1.519e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
third.best.model <- glm(n ~ habitat + sex + sex:habitat,
                        data=three.way.table,family = poisson)
summary(third.best.model)
```

```
##
## Call:
## glm(formula = n ~ habitat + sex + sex:habitat, family = poisson,
##     data = three.way.table)
##
## Deviance Residuals:
##       1        2        3        4        5        6        7        8
## -2.4160  -1.0805   0.8430   2.0804   1.9494   0.8718  -0.9670  -2.4773
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 2.5257     0.2000  12.629  < 2e-16 ***
## habitatovergrown           -1.6094     0.4899  -3.285  0.00102 **
## sexmale                    -0.9163     0.3742  -2.449  0.01433 *
## habitatovergrown:sexmale   2.9957     0.6042   4.959  7.1e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 62.558  on 7  degrees of freedom
## Residual deviance: 23.675  on 4  degrees of freedom
## AIC: 61.465
##
## Number of Fisher Scoring iterations: 5
```

```
anova(third.best.model, test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: n
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                            7     62.558
## habitat      1    1.253         6     61.305   0.26293
## sex          1    5.053         5     56.252   0.02458 *
## habitat:sex  1   32.576         4     23.675 1.146e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(full.model)
```

```
## [1] 45.78948
```

```
AIC(best.model)
```

```
## [1] 43.81986
```

```
AIC(second.best.model)
```

```
## [1] 41.89107
```

```
AIC(third.best.model)
```

```
## [1] 61.46478
```

So final best model is **second.best.model**! It includes all factors and two two-way interactions sex:habitat and sex:colour. It has lowest AIC value 41.89107.
The two interactions mean that both *habitat* and *color* depends on *sex*. This model is simpler than full model it lacks one two-way interaction *colour:habitat*, and three-way interaction between all the factors. The exclusion of *colour:habitat* interaction, which is significant in anova output for best model (0.001893) but not so as other interactions, can be explained that *color* is determined by *sex*, and *habitat* is determined by *sex*. So conclusion is that males are more frequently black than red and prefer to live in overgrown habitat. As a result professor observes more black lemmings in overgrown habitat.