97/100

MATH 536 FINAL EXAM

Gustavo Esparza

11/15/2019

1

A

Consider a multinomial distribution with four cells, sample size n, and probability vector $((1-p), p(1-p), p^2(1-p), p^3)$ for some unknown probability $p, 0 < p < 1$. Let n_j be the number of observations that fall in cell j. Test the hypothesis that $H_0 : p = 0.6$. What is the test statistic and the asymptotic distribution under this hypothesis?

H_0 : p = 0.6 VS H_A : p $\neq$ 0.6

The test statistic is a χ^2 value computed by taking the observed and expected squared differences, then dividing by the expected value:

$$\chi^2 = \sum_{i=1}^{4} \frac{(\hat{p}_j - n_j p_j)^2}{n_j p_j}$$

In this instance, $p_1 = .4$, $p_2 = .24$, $p_3 = .6^2 \times .4 = .144$, $p_4 = .6^3 = .216$. ✓

The asymptotic distribution is a χ^2 dsitribution with 4 - 1 = 3 degrees of freedom.

Using the χ^2 test statistic with 3 degrees, 7.815, we reject the null hypothesis if our observed test statistic exceeds this value.

B

Consider a three way contingency table with probabilities p_{ijk} for cell (i, j, k) for $i = 1, \ldots, I, j = 1, \ldots, J$, and $k = 1, \ldots, K$ such that $\sum_i \sum_j \sum_k p_{ijk} = 1$. Find the degrees of freedom for the χ^2 test for the following hypotheses:

For each instance, the DF under dependence = IJK-1, representing the total amount of free parameters in the saturated model

i

$H_0 : p_{ijk} = p_i q_j r_k$ (Three way independence)

For complete indepdendence, the degrees of freedom is equivalant to the quantity $(I-1)+(J-1)+(K-1)$, as each cell is independent of one another. Thus, the degrees of freedom for a three-way independence test is equal to $(IJK) - (I-1) - (J-1) - (K-1)$. ✓

By expanding and simplyifying, this degrees of freedom becomes $DF = (I-1)(J-1)(K-1)$

ii

$H_0 : p_{ijk} = p_i q_{jk}$ (Independence of the first variable from the other two)

For joint indepdendence of the first variable from the other two, the degrees of freedom is equivalant to the quantity $(I-1)+(JK-1)$,as I is indepdendent of J and K.

Thus the degrees of freedom for a Independence of the first variable from the other two test is equal to $(IJK)-(I-1)-(JK-1)$. ✓

iii

$H_0 : p_{ijk} = p_{i|k} q_{j|k} r_k$ (Conditional independence of the first two variables given the third)

For conditional indepdendence of the first two variables given the third, the degrees of freedom is equivalant to the quantity $(K-1)+K(I-1)+K(J-1)$, as we need to estimate K alone, I given K, and J given K.

Thus, the degrees of freedom for a Conditional independence of the first two variables given the third test is equal to $(IJK)-(K-1)-K(I-1)-K(J-1)$ ✓

2

The table below refers to an experiment on the use of sulfones and streptomycin drugs in the treatment of leprosy. The degree of infiltration at the start of the experiment measures a type of skin damage. The response is the change in the overall clinical condition of the patient after 48 weeks of treatment. The question of interest is whether there is an association between degree of infiltration and clinical change.

```
##       Condition
## Degree Marked Improvement Moderate Improvement Slight Improvement
##   High                  7                    8                 14
##   Low                  12                    9                  5
##       Condition
## Degree Stationary Worse
##   High          6     7
##   Low          10     2
```

A

Perform a chi-square test of independence. State the null and alternative hypotheses. What do you conclude?

Hypothesis: H_0 : Degree of Infiltration and Clinical Condition are independent VS. H_A : Degree of Infiltration and Clinical Condition are dependent.

Test Statistic:

$$\sum_{i=1}^{2}\sum_{j=1}^{5}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

Where expected values are computed by multiplying row and column sums for each cell, then dividing by the total sample size size.

In this instance, our χ^2 value is determined to be 9.2386453 ✓ and follows a χ^2 distribution with (2-1)(5-1) ✓ = 4 degrees of freedom.

P-value: Given our test statistic and distribution, we have a p-value of 0.0554036 ✓.

Conclusion: We proceed to reject the null hypothesis and conclude that there is a relationship between Degree of Infiltration and Clinical Condition.

B

Perform a likelihood ratio test (LRT) of independence. Compare your conclusion with part (a).

Hypothesis: H_0 : Degree of Infiltration and Clinical Condition are independent VS. H_A : Degree of Infiltration and Clinical Condition are dependent.

Test Statistic:

$$LRT = 2\sum_{i=1}^{2}\sum_{j=1}^{5}O_{ij}\times log(\frac{O_{ij}}{E_{ij}})$$

In this instance, our LRT value is determined to be 9.5817072 ✓ and follows a χ^2 distribution with (2-1)(5-1) = 4 degrees of freedom.

P-value: Given our test statistic and distribution, we have a p-value of 0.0480952. ✓

Conclusion: We proceed to reject the null hypothesis and conclude that there is a relationship between Degree of Infiltration and Clinical Condition. Furthermore, we can conclude that both the chi-square and Likelihood Ratio test obtain the same conclusion.

C

Perform a test of independence using Fisher's exact test. Compare your conclusion with parts (a) and (b).

Hypothesis: H_0 : Degree of Infiltration and Clinical Condition are independent VS. H_A : Degree of Infiltration and Clinical Condition are dependent.

Test Statistic: In this instance, fisher's test uses a hypergeometric assumption (fixed rows and columns) to compute the p-value.

P-value: Given our test statistic and distribution, we have a p-value of 0.0578249, ✓ utilziing the fisher.test function in R.

Conclusion: We proceed to reject the null hypothesis and conclude that there is a relationship between Degree of Infiltration and Clinical Condition. Furthermore, we can conclude that both the chi-square and Likelihood Ratio test obtain the same conclusion as fisher's exact test.

D

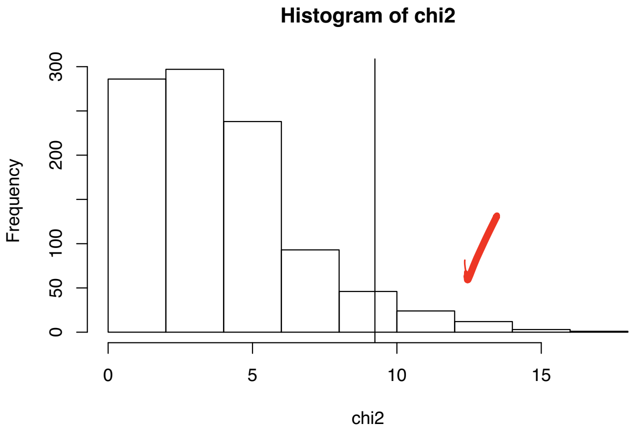
Perform a permutation test for independence two ways:

i

Using the chi-square statistic (explain your procedure in words).

For our permutation test, we will sample from our observed values (without replacement) and compute a new χ^2 test statistic. We will perform this step 1000 times and create a histogram for the test statistics. Finally, we will compare the distribution of χ^2 values and compute a permutation test p-value by observing the proportion of permuted χ^2 values that exceed our original test statistic of 9.2386453 statistic.

Here is a histogram of our permuted χ^2 values, along with a line indicating the original test statistic:



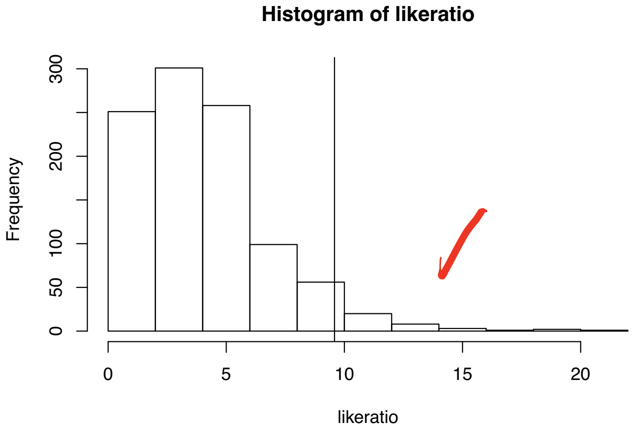
In this instance, our simulated p-value is 0.049. We proceed to reject the null hypothesis and conclude that there is an association between our two variables.

ii

Using the LRT statistic (explain your procedure in words).

For our permutation test, we will sample from our observed values (without replacement) and compute a new LRT test statistic. We will perform this step 1000 times and create a histogram for the test statistics. Finally, we will compare the distribution of LRT values and compute a permutation test p-value by observing the proportion of permuted LRT values that exceed our original test statistic of 9.5817072 statistic.

Here is a histogram of our permuted LRT values, along with a line indicating the original test statistic:



In this instance, our simulated p-value is 0.043. We proceed to reject the null hypothesis and conclude that there is an association between our two variables.

iii

Compare your conclusions from parts (i) and (ii).

It is apparent that both permutation tests provide very similar histograms, p-values and conclusions that indicate a relationship between our two variables. The tails for both of our distributions are also quite similar, with the LRT tail being slightly longer.

E

Provide a 1-paragraph summary comparing all of these tests of independence.

✓

The Chi-square and LRT provided very similar p-values, but have very different ways of computing their respective test statistics.The exact fisher test has a different underlying distribution assumption and computes the respective test statistic and p-value in a different manner as well. Although the same conclusion was reached, only fisher is suitable for a small sample size. When sample size is not an issue, chi-square and likelihood ratio comes down to whether we wish to use the residuals (chi-square) or maximum likelihood estimates (likelihood ratio test.)

3

Consider the following data on prison sentencing. These data report on whether or not an offender received a prison sentence as a function of: (1) Type-whether the crime involved a business or a home, (2) Prior - whether or not the offender had a prior arrest record, and (3) Gender.

```
##   prior     type gender response count
## 1  Some Business   Male        1    28
## 2  None Business   Male        1     7
## 3  Some     Home   Male        1    19
## 4  None     Home   Male        1    27
## 5  Some Business Female        1    14
## 6  None Business Female        1    10
```

A

Fit the following logistic regression model:

$$log(\pi_i/(1-\pi_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

where π_i is the probability that the ith person will receive a prison sentence, $x_{i1} = 1$ if the ith person commits a business crime and $x_{i1} = 0$ if the ith person commits a home crime, and $x_{i2} = 1$ if the ith person has some prior arrest history and $x_i 2 = 0$ if the ith person has no prior arrest history.

Here is a summary of our desired logistic regression model:

```
##
## Call:
## glm(formula = response ~ prior + type, family = binomial, data = prison,
##     weights = count)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.3436  -5.6747   0.3466   7.3608  10.5738
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.9356     0.1374 -14.083  < 2e-16 ***
## priorSome      0.3345     0.1901   1.760  0.07839 .      ✓
## typeBusiness   0.5835     0.1964   2.971  0.00297 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 784.97  on 15  degrees of freedom
## Residual deviance: 769.48  on 13  degrees of freedom
## AIC: 775.48
##
## Number of Fisher Scoring iterations: 5
```

Thus, our fitted model is defined as

$$log(prison/noprison) = -1.9356 + .3345 Prior_{Some} + .5825 Crime_{Business}$$

B

Test the hypothesis for the effect of type of crime on prison sentence. State the null and alternative hypothesis in terms of an odds ratio parameter, give the formula for the test statistic, state the distribution of the test statistic under the null hypothesis, and finally carry out the test using, describing briefly what you did and your conclusions.

Hypothesis: H_0: $\frac{odds_{prison|Business}}{odds_{Prison|home}} = 1$ VS $H_A : \frac{odds_{Prison|Business}}{odds_{Prison|home}} \neq 1$ ✓

IE, this is a test for the independence of type of crime on prison sentence.

Test statistic:

ΔG^2 = Deviance for intercept only model not including type (null deviance) - deviance for model including type (residual deviance) = 784.9743098 - 772.5645194 = 12.4097904 . ✓

Distribution under null: χ^2 with df = 2 parameters - 1 parameter = 1

p-value:0.0004

Thus, we proceed to reject the null hypothesis and conclude that Type of crime is significant to detrmining Prison Sentencing.

C

Estimate the odds ratio for the effect of type of crime on prison sentence controlling for the other variables in the model. Interpret this estimate. Also, compute a 95 % confidence interval for the odds ratio for the effect of type of crime controlling for the other variables in the model. Show your work.

Using the fitted values from the logistic regression model, 0.5835372 and taking the exponent, the odds of prison sentencing for Business related crimes is 1.7923672 times the respective odds for home related crimes.

```
##
## Change in Odds From a change from Home to Business =  1.792367
```

✓

Considering a normal assumption for our regression coefficients, we have the following confidence interval for the logistic regression coeffeient for type,using the provided standard error:

$$.5835 \pm 1.96 \times 0.1964$$

Taking the exponents of our interval endpoints provides the following confidence interval:

```
##
## Confidence Interval = [ 1.219629  ,  2.634064 ]
```

✓

We can observe that our confidence interval contains only values greater than one, so we have great confidence that the odds of prison time are greater for business related crimes than the odds for home related crimes.

D

In the model from part (a), add an interaction term between crime type and prior arrest. Carry out an analysis of deviance for the effect of the interaction term between crime type and prior arrest on the outcome, controlling for the other variables in the model. Make sure to state the null and alternative hypothesis in terms of a model coefficient, give the formula for the test statistic, its distribution, degrees of freedom under the null hypothesis, and report the approximate p-value. Which model should we keep based on your test? Would you conclude based on this test that people who commit business crimes are more likely to have some prior arrest than people who commit home crimes?

Here is a brief summary of our interaction model:

```
##
## Call:  glm(formula = response ~ prior + type + prior * type, family = binomial,
##     data = prison, weights = count)
##
## Coefficients:
##           (Intercept)              priorSome           typeBusiness
##               -1.8943                 0.2261                 0.4101
## priorSome:typeBusiness
##                0.3045
##
## Degrees of Freedom: 15 Total (i.e. Null);  12 Residual
## Null Deviance:       785
## Residual Deviance: 768.9     AIC: 776.9
```

Test for effect of interaction term:

H_0: Model without interaction term VS H_A: Model including interaction term

Test statistic: ΔG^2 with 1 degrees of freedom: 0.5778979

Distribution under null: χ^2 with df = 2 parameters - 1 parameter = 1

p-value: 0.4471375 ✓

Conclusion: We proceed to fail to reject the null hypothesis and conclude that the model does not benefit from an interaction term between type of crime and prior arrests. We can not conclude that that people who commit business crimes are more likely to have some prior arrest than people who commit home crimes.

4

Please answer True or False to the following questions related to overdispersion. If False, please explain why.

A

Overdispersion implies greater variability in the data than expected by the GLM random component.

-3 ✗ False. Overdispersion implies greater variability in the model than expected by the original data.

B

Confidence intervals and hypothesis tests are valid for overdispered data.

False. Overdispersion may cause Confidence Intervals to be too narrow and p-values may be smaller than they truly should be.

✓

C

Overdispersion typically arises from from an error in data collection.

False. The main cause of overdispersion is caused by a poor model without a cause, which can be resolved by adding extra variables from the original data to better explain variability.

✓

5

The table below refers to a clinical trial for the treatment of small cell lung cancer. Patients were randomly assigned to two treatment therapy groups: (1) sequential therapy, which administered the same combination of chemotherapeutic agents in each treatment cycle, and (2) alternating therapy, which had three different combinations alternating from cycle to cycle.

Here is a brief representation of the data:

```
##   gender     therapy            response count
## 1   Male  Sequential Progressive Disease    28
## 2 Female  Sequential Progressive Disease     4
## 3   Male Alternating Progressive Disease    41
## 4 Female Alternating Progressive Disease    12
## 5   Male  Sequential           No Change    45
## 6 Female  Sequential           No Change    12
```

A

Considering a male in sequential therapy as the baseline. Fit a proportional odds model. State the model and the distribution assumptions.

Here is a sumamry of the proportional odds model:

```
## Call:
## polr(formula = response ~ gender + therapy, data = chemo, weights = count)
##
## Coefficients:
##       genderFemale therapyAlternating
##         -0.5413912         -0.5806836
##
## Intercepts:
##        Progressive Disease|No Change          No Change|Partial Remission
##                           -1.3180403                            0.2492343
## Partial Remission|Complete Remission
##                            1.3000568
##
## Residual Deviance: 789.0566
## AIC: 799.0566
```

Then, the model is defined as

✓

$$log(\pi_i/\pi_{i+1}+\cdots+\pi_n) = \alpha_i + .5413(Gender_F) + .5806(Therapy_A)$$

Where α_i changes for each response category, indicated by the cummulative intercepts for Progressive Disease: -1.318, No Change: 0.249, Partial Remission: 1.30

The assumption for the proportional odds model states that the estimated coefficients, except for the intercepts, will stay the same for every cumulative probability (level of the response).

B

Given the baseline, find the estimated probabilities for each of the response categories.

The cummulative probability for each category given the baseline (males in sequential therapy) is defined as

$$p(y \leq i) = \frac{e^{\beta_0 i}}{1 + e^{\beta_{0i}}}, i = 1, 2, 3$$

Where the cummulative probability for "4- Complete remission" is simply 1. Then, using the difference between these cummulative probabilties allows us to compute the exact probabilties for each category.

Thus the probability of our baseline patient being in category "Complete remission" is defined as $1 - p(Y \leq PartialRemission) = 0.2141555$.

The probability of our baseline patient being in category "Partial Remission" is defined as $p(Y \leq PartialRemission) - p(Y \leq NoChange) = 0.2238565$.

The probability of our baseline patient being in category "No Change" is defined as $p(Y \leq NoChange) - p(Y \leq ProgressiveDisease) = 0.3508435$.

Finally, the probability of our baseline patient being in category "Progressive Disease" is defined as $p(Y \leq ProgressiveDisease) = 0.2111445$

We can also compare these computed values with the results from the built in **predict** function:

```
## Progressive Disease          No Change   Partial Remission
##           0.2111445          0.3508435           0.2238565
##  Complete Remission
##           0.2141555
```

✓

C

Find the cumulative odds ratio of gender, given therapy. Interpret this odds ratio.

The odds ratio for gender in this scenario is defined as

$$\frac{odds(Y \leq i|\textbf{Female})}{odds(Y \leq i|\textbf{Male})}$$

The model coeffecients give 0.5413912 as the coeffecient relating to gender. Taking the exponent of this value gives us 1.7183958 as the cumulative odds ratio of gender. Thus, for any constant therapy status, the odds of a female being in or below response group i are 1.7183958 times the equivalent odds for males in the same therapy group.

✓

6

Use the data file puffin.csv for this problem. The data is from the article, "Breeding Success of the Common Puffin on Difference Habitats at Great Island, Newfoundland," Four predictor variables where considered in trying to describe the nesting frequency. The complete data contains 38 observations and 5 variables.

```
##   nesting grass soil angle distance
## 1      16    45 39.2    38        3
## 2      15    65 47.0    36       12
## 3      10    40 24.3    14       18
## 4       7    20 30.0    16       21
## 5      11    40 47.6     6       27
## 6       7    80 47.6     9       36
```

A

What is an appropriate distribution for the y_i (nesting)? Is this distribution part of the exponential family? If so, show this.

A poisson distribution is appropriate because there is no upper bound for the nesting values, and the nesting response is measured in terms of counts.

This distribution is part of the exponential family, explained as follows:

Here is the Likelihood function for the Poisson distribution:

$$L(\theta) = \frac{\theta^{\sum x_i e^{-n\theta}}}{\prod x_i} = \theta^{\sum x_i e^{-n\theta}} \times \frac{1}{\prod x_i}$$

✓

By the factorization theorem, we can observe that the first expression indication by the summation is a sufficient statistic. As this sufficient statistic contains an exponent expression, we have validated that the poisson distribution belongs to the exponential family.

B

What would be an appropriate choice for the link function?

An appropriate choice for the link function would be the log-link since this link would take the exponent of our response values, allowing the response to be expressed in terms of counts that are strictly non-negative. ✓

C

Fit a GLM to the data including all explanatory variables and write the equation representing $\mu \hat{} i$. Interpret the coefficient for the distance variable.

Here is a summary of our Poisson GLM:

```
##
## Call:  glm(formula = nesting ~ grass + soil + angle + distance, family = poisson(link = log),
##     data = puffin)
##
## Coefficients:
```

```
## (Intercept)        grass         soil        angle     distance
##    3.069973     0.005441     0.033441    -0.030077    -0.089399
##
## Degrees of Freedom: 37 Total (i.e. Null);  33 Residual
## Null Deviance:       310.4
## Residual Deviance: 68.76     AIC: 183.4
```

Thus, our model is specified as follows:

$$\mu_i = exp(3.07 + .005(Grass) + .033(Soil) - -.03(Angle) - .089(Distance))$$

Focusing on the coefficient for distance, we can take the exponent to obtain the value of $exp(-.089) =$ 0.9148456. Specifically, we can find the multiplicative term as follows:

$$\frac{\mu(x+c)}{\mu(x)} = \frac{exp(3.07 + .005(Grass) + .033(Soil) - -.03(Angle) - .089(Distance + C))}{exp(3.07 + .005(Grass) + .033(Soil) - -.03(Angle) - .089(Distance))}$$

This simplifies to the following original expression:

$$\frac{\mu(x+c)}{\mu(x)} = exp(C \times -.089)$$

Thus, the mean of our response count will be multiplied by the following exponent for every unit increase in C. As a result, we can observe that the mean count is expected to decrease as distance increases.

Putting this effect in terms of a percent change, we have a -8.5154426 change in the mean nesting count for each unit increase in distance. ✓

D

Conduct a test of lack of fit for part (c) based on the residual deviance. State the null and alternative hypothesis, the test statistic, the distribution of the test statistic, p-value, and your conclusions based on this test.

H_0: Current model VS H_A: Saturated Model
Test statistic: Residual Deviance with 33 degrees of freedom: 68.76, following an underlying χ^2 distribution. ✓
P-value: ≈ 0

Conclusion: We choose to reject our null hypothesis and conclude that the current model alone is not fitting the data properly. An offset may potentially improve the model.

E

Produce a 95 % confidence interval for the effect of the distance variable. Give an interpretation of this confidence interval.

Using the fitted values from the poisson regression model, -0.0893993 and taking the exponent, the multiplicative for distance is 0.9144804.

```
##
## Multiplicative effect =  0.9144804
```

Considering a normal assumption for our regression coefficients, we have the following confidence interval for the poisson regression coeffeient for type,using the provided standard error:

$$-0.0893 \pm 1.96 \times 0.0106$$

Taking the exponents of our interval endpoints provides the following confidence interval:

```
##
## Confidence Interval = [ 0.8955373  ,  0.9338241 ]
```

We can observe that our confidence interval contains only negative values, so we have great confidence that the multiplicative effect of distance lowers the mean nesting count, when all other predictors are constant.

Putting this effect in terms of a percent change, we have a -10.446266 to -6.6175863 percent change in the mean nesting count for each unit increase in distance, with 95 percent confidence.

F

> Examine the effect of grass and soil on the number of nests, given angle and distance are in the model, with an analysis of deviance. State both the null and alternative hypotheses, the test statistic, the p-value, and give your conclusions.

Hypothesis: H_0 : Model only including angle and distance VS H_a : Model including all predictors

Test statistic: Δg^2 = Deviance for model not including grass and soil - deviance for model including all predictors = 16.2847412. Underlying distribution follows χ^2.

Distribution under null: χ^2 with df = (35 df) - (33 df) = 2

p-value: ≈ 0

Thus, we proceed to reject the null hypothesis and conclude that grass and soil are significant to the model, given that angle and distance are included.

Appendix

```
library(MASS)
library(VGAM)
library(vcd)
```

2

```
Degree = c(rep("High",42),rep("Low",38))
Condition = c ( rep("Worse",7),rep("Stationary",6),rep("Slight Improvement",14),
                rep("Moderate Improvement",8),rep("Marked Improvement",7),
                rep("Worse",2),rep("Stationary",10),rep("Slight Improvement",5),
                rep("Moderate Improvement",9),rep("Marked Improvement",12))

df = as.data.frame(cbind(Degree,Condition))
data_freq = table(df)
```

```
data_freq
```

A

```
Observed = c( 7,8,14,6,7,12,9,5,10,2)
T= sum(Observed)
rows = as.numeric(rowSums(data_freq))
cols = as.numeric(colSums(data_freq))

expect =rows %*%t(cols)/T

Expected = c(expect[1,],expect[2,])

X2 = sum((Observed-Expected)^2 /(Expected))
pval = pchisq(X2,4,lower.tail = F)
```

B

```
Observed = c( 7,8,14,6,7,12,9,5,10,2)
T= sum(Observed)
rows = as.numeric(rowSums(data_freq))
cols = as.numeric(colSums(data_freq))

expect =rows %*%t(cols)/T

Expected = c(expect[1,],expect[2,])

Ratio = Observed/Expected
Like.Ratio = Observed*log(Ratio)
lrt = 2 * sum(Like.Ratio)

pval2 = pchisq(lrt,4,lower.tail = F)
```

C

```r
fisher = fisher.test(data_freq)
pval3 = fisher$p.value
```

D

```r
chi2 = rep(0,1000)

for(i in 1:1000) {
  permuted_data =
    as.data.frame(cbind(Degree, sample(Condition,replace = F)))
  perm_freq = table(permuted_data)
  chi2[i] = as.numeric(chisq.test(perm_freq,simulate.p.value = T)$statistic)
}
```

```r
hist(chi2)
abline(v=X2)
sim_pval = sum(chi2>X2)/1000
```

```r
likeratio = rep(0,1000)

for(i in 1:1000) {
  permuted_data = as.data.frame(cbind(Degree, sample(Condition,replace = F)))
  perm_freq = table(permuted_data)
  likeratio[i] =  as.numeric(assocstats(perm_freq)$chisq_tests[1])
}
```

```r
hist(likeratio)
abline(v=lrt)
sim_pval_lrt = sum(likeratio>lrt)/1000
```

3

```r
prison = data.frame(expand.grid(prior=c("Some","None"),
                                   type=c("Business","Home"),
                                    gender=c("Male","Female"),
                                   response=c(1,0)),
        count=c(28,7,19,27,14,10,14,27,59,35,75,200,50,40,100,159))
```

```r
prison$prior = relevel(prison$prior,ref="None")
prison$type = relevel(prison$type,ref="Home")
head(prison)
```

A

Here is a summary of our desired logistic regression model:

```r
logistic_model = glm(response ~ prior + type, weights = count,
                     data = prison, family = binomial)
summary(logistic_model)
```

B

```
type_model = glm(response ~ type, weights = count,
                data = prison, family = binomial)
```

```
pval = pchisq(type_model$null.deviance - type_model$deviance,
      df= type_model$df.null - type_model$df.residual, lower.tail=FALSE)
```

C

```
beta_type= logistic_model$coefficients[3]
se_type= summary(logistic_model)$coefficients[3,2]
OR_type = exp(beta_type)
#Confidence interval
z_star = 1.96
L = exp(beta_type - z_star*se_type)
U = exp(beta_type + z_star*se_type)
```

```
cat("\nChange in Odds From a change from Home to Business = ",OR_type,"")
```

```
cat("\nConfidence Interval = [",L," , ",U,"]")
```

D

Here is a brief summary of our interaction model:

```
interaction_model = glm(response ~ prior + type +prior*type, weights = count,
                        data = prison, family = binomial)
interaction_model
```

```
pval = pchisq(logistic_model$deviance - interaction_model$deviance,
        df=logistic_model$df.residual - interaction_model$df.residual, lower.tail=FALSE)
```

5

```
chemo = data.frame(expand.grid(     gender=c("Male","Female"),
                                    therapy=c("Sequential","Alternating"),
                                    response=c("Progressive Disease","No Change",
                                      "Partial Remission","Complete Remission")),
        count=c(28,4,41,12,45,12,44,7,29,5,20,3,26,2,20,1))
```

Here is a brief representation of the data:

```
head(chemo)
```

A

```
chemo$gender = relevel(chemo$gender,ref = "Male")
chemo$therapy = relevel(chemo$therapy,ref = "Sequential")
```

```
chemo_PO =  polr(response ~gender+therapy, data=chemo,weight=count)
chemo_PO
```

B

```
pi.less.PD = exp(chemo_PO$zeta[1])/(1 + exp(chemo_PO$zeta[1]) )
pi.less.NC = exp(chemo_PO$zeta[2])/(1 + exp(chemo_PO$zeta[2]) )
pi.less.PR = exp(chemo_PO$zeta[3])/(1 + exp(chemo_PO$zeta[3]) )
```

```
probabilities = predict(chemo_PO,
              newdata = data.frame(gender="Male",therapy="Sequential"), type = "probs")
probabilities
```

C

```
odds_gender = -chemo_PO$coefficients[1]

exp_odds_gender = exp(odds_gender)
```

6

```
puffin=read.csv("puffin.csv")
head(puffin)
```

C

```
poisson_model = glm(nesting ~grass + soil + angle + distance,
                    data = puffin, family=poisson(link=log))
poisson_model
```

D

```
pval = pchisq(poisson_model$deviance, poisson_model$df.residual,
              lower.tail = FALSE)
```

E

```
beta = coef(poisson_model)[5]
se = summary(poisson_model)$coefficients[5,2]
L = beta-1.96*se
U = beta+1.96*se
```

```
cat("\nMultiplicative effect = ",exp(beta),"")
```

```r
cat("\nConfidence Interval = [",exp(L)," , ",exp(U),"]")
```

F

```r
reduced_model = glm(nesting ~angle + distance,
                    data = puffin, family=poisson(link=log))

full_model = glm(nesting ~grass + soil + angle + distance,
                 data = puffin, family=poisson(link=log))
```

```r
pval = pchisq(reduced_model$deviance - full_model$deviance,
            df=reduced_model$df.residual - full_model$df.residual, lower.tail=FALSE)
```