HW1 Catagorical Data

Faizan Khalid Mohsin; 997157570

October 21, 2016

## Question 1:

We look at the aggregate result:

Test statistic:

where the variance is:

Thus the test statistic is:

Which we know has a standard normal distribution.

We find the test statistic to be:

with

n\_T = 997  
y\_T = 299  
p\_T = y\_T/n\_T  
  
n\_c = 995  
y\_c = 352  
p\_c = y\_c/n\_c  
  
z = (p\_T - p\_c) / (p\_T\*(1-p\_T)/n\_T + p\_c\*(1-p\_c)/n\_c)^.5  
z

## [1] -2.56707

p\_T

## [1] 0.2998997

p\_c

## [1] 0.3537688

pvalue=pnorm(z)  
pvalue

## [1] 0.005128099

Now we look at each hospital:

For Princeton Hospital the test statistic is found using:

Doing similar calculations as for the aggregate using R, we find the following statistic and p-values.

with

We do this for each of the other two Hospitals we summarize all of our results in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Hospital |  |  |  |  |
|  | 0.2656 | 0.25 | -0.9999115 | 0.1586767 |
|  | 0.3153153 | 0.3015075 | 0.3563215 | 0.6392001 |
|  | 0.42 | 0.4007286 | 0.4244532 | 0.6643823 |
|  | 0.2998997 | 0.3537688 | -2.56707 | 0.005128 |

Looking at the aggregate results we would conclude that there is statistical evidence (p-value = 0.005128) that a higher proportion of people using the new drug experienced an adverse effect than those who used the usual drug.

However, when we break it down to the individual hospital level, we see that for each hospital there is no statistically significant difference in the proportion of people who experience an adverse effect for those who took the new drug or those who took the usual drug (as all the p-values are greater than 0.05).

Hence, the investigator would arrive at a misleading conclusion if they were to base their analysis on aggregate counts as clearly simson's paradox occurs in this case.

## Question 2

The proportion of people diagnosed with colorectal cancer was among participants with daily intake of processed meat of 80 g or more and among participants with lower intake of processed meat.

The difference in proportion is:

This means that the difference in the proportion of people diagnosed with colorectal cancer who ate 80g or more of processed meat daily and those who who consumed less is 0.000926.

The relative risk is:

This means that people who had a daily intake of processed meat of 80 g or more were 1.34 times more likely to be diagnosed with colorectal cancer compared to those with lower intake of processed meat.

Clearly relative risk is more informative because in a study looking at the probability or risk of being diagnosed with cancer between two different group, one is interested in knowing what is the risk of cancer of a person in one group compared to if he was int the other.

## Question 3

#### Question: State three "real-world" variables X, Y, and Z for which you expect a marginal association between X and Y but conditional independence controlling for Z.

Let be the number of clothing items people are wearing, be the number of car accidents, and be a factor variable for the four seasons. Clearly, in winter there are more accidents because of the ice and snow on the roads but because of winter people also where more cloths. Hence, we expect marginal associations between and , however, if we condition for the seasons, then clearly within winter or any other season, the number of clothing items and the number of car accidents ( and respectively) will be independant.

## Question 4

#### Question: Calculate (show calculations) and interpret a 95% confidence interval, based on the score test, for the proportion of bicyclists wearing a helmet.

The 95% CI based on the score test is:

Where: - number of cyclists sampled, - number of cylists wearing helmets, - is the proportion of cyclists wearing a helmet, as

Plugging these values into the 95% CI formula:

Gives us the CI:

There is a 95% chance that our confidence interval contains the true proportion of Toronto cyclists who wear helmets. Or, the proportion of Toronto cyclists who wear a helmet is anywhere between 0.3474 to 0.4375.

Below is the R code used to calculate the CI.

z = 1.96  
n = 447  
y = 175  
p = y/n  
pc = 1-p  
coef = 1/(n+z^2)  
a = p\*n + z^2/2  
b = z\*(p\*pc\*n + .25\*z^2)^.5  
upperbd = coef\*(a+b)  
lowerbd = coef\*(a-b)

## Question 5

#### Question: Identify a variable for each of the following type of variable form the 2013-2014 dataset for the Canadian Community Health Survey (CCHS):

Note that all the variable were taken from the *Derived and Grouped Variables (PDF) - detailed specifications* document.

1. *Binary Variable:*

* Label: SDCCGT
* Name: Culture/Race flage
  + 1 - white
  + 2 - Non-white

1. *Nominal Variable:*

* Label: DHHGMS
* Name: Marital status
  + 1 - Married
  + 2 - Common Law
  + 3 - Widowed/Divorced/Seperated
  + 4 - Single

1. *Ordinal Variable:*

* Label: ALCDTTM
* Name: Type of Drinker
  + 1 - Regular drinker
  + 2 - Occasional drinker
  + 3 - Did not drink in the last 12 months

1. *categorized Continuous Variable:*

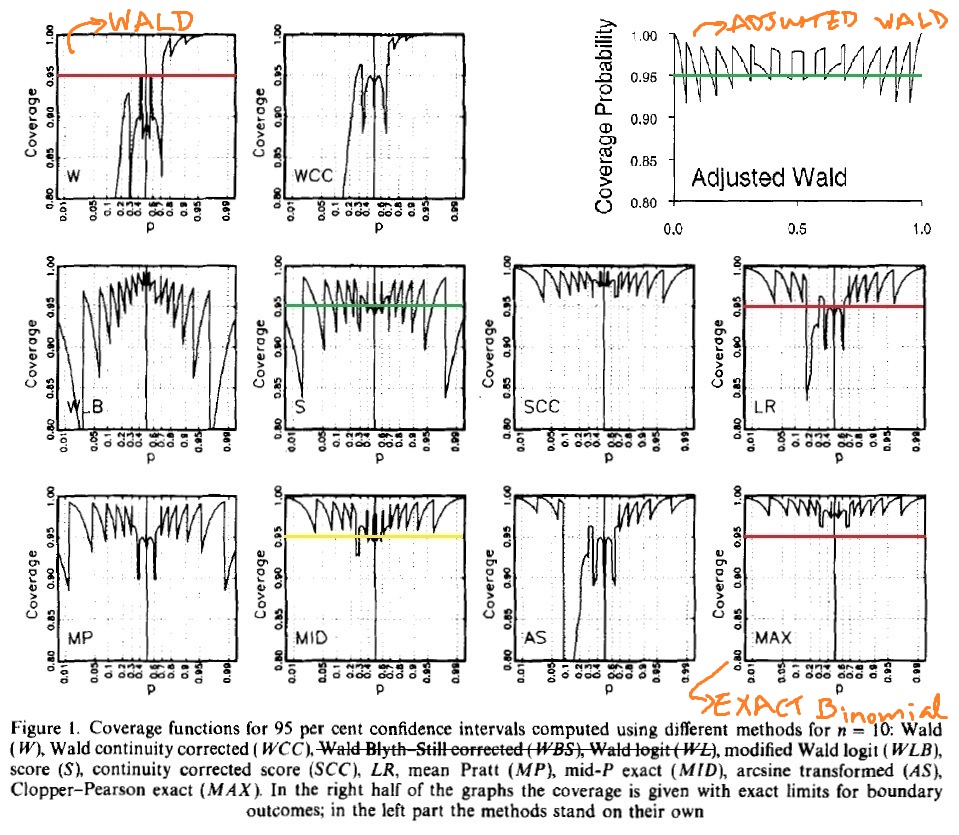
* Label: DHHGAGE
* Name: Age
  + 1 - 12 <= DHH\_AGE <= 14 (Age between 12 and 14)
  + 2 - 15 <= DHH\_AGE <= 17 (etc)
  + 3 - 18 <= DHH\_AGE <= 19
  + 4 - 20 <= DHH\_AGE <= 24
  + 5 - 25 <= DHH\_AGE <= 29
  + 6 - 30 <= DHH\_AGE <= 34
  + 7 - 35 <= DHH\_AGE <= 39
  + 8 - 40 <= DHH\_AGE <= 44
  + 9 - 45 <= DHH\_AGE <= 49
  + 10 - 50 <= DHH\_AGE <= 54
  + 11 - 55 <= DHH\_AGE <= 59
  + 12 - 60 <= DHH\_AGE <= 64
  + 13 - 65 <= DHH\_AGE <= 69
  + 14 - 70 <= DHH\_AGE <= 74
  + 15 - 75 <= DHH\_AGE <= 79
  + 16 - DHH\_AGE >= 80 (Age 80 and older)

## Question 6

#### Question: Using language that an intelligent nonstatistician would understand, describe anticipated relative performance of these intervals for this specific proposal.

As is very small and .

Looking at Figure 1 below:



We see that the Coverage of the 95% CI of the wald test is very bad. For the proportion of 0.9 it is almost 1. As approaches 0 the coverage probability approaches 80%.

For the Exact Binomial test the coverage probability for a 95% CI is alwas greater than 95%.

For the Adjusted Wald we see that for all the values of the coverage is mostly 95%. Hence, for a small sample of 10 people and of 0.9, I would recommend that they go with Adjusted Wald test.

## Question 7

#### Question 7a: Using and , test the hypothesis of independence between party identification and race. Report the P-values and interpret.

Below is the table of data:

## Democrats Independants Republicans  
## Blacks 192 75 8  
## Whites 459 586 471

We know that the statistic is:

$\chi^2 = \mathop{\sum\_{i=1}^{I}\sum\_{j=1}^{J}}\frac{(n\_{ij} - \hat{\mu\_{ij}})^2}{\hat{\mu\_{ij}}}$

with

Where and

And the statistic:

$G^2 = 2\mathop{\sum\_{i=1}^{I}\sum\_{j=1}^{J}}n\_{ij}log(n\_{ij}/\hat{\mu\_{ij}})$

also with

Using the command chisq.test() in R we obtain:

##   
## Pearson's Chi-squared test  
##   
## data: table  
## X-squared = 177.31, df = 2, p-value < 2.2e-16

Using the command G.test in the "RVAideMemoire" package we obtain:

## \*\*\* Package RVAideMemoire v 0.9-60 \*\*\*

##   
## G-test  
##   
## data: table  
## G = 197.39, df = 2, p-value < 2.2e-16

Summarizing the results in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | 177.31 | 2 | p<<0.0001 |
|  | 197.39 | 2 | p<<0.0001 |

We conclude that both test give us evidence (p<<0.0001) to reject the null hypothesis of independence between people's race and political identification. Hence, we would conclude that there is an associate between one's race and the political party he identifies with. In particular, black people tend to identify themselves mostly as democrats.

#### Question 7b: Use standardized residuals to describe the evidence of association.

The standardized residuals for the row and column is:

where and is defined as before.

Using the command chisq.test(table)$stdres in R we present the standardized residuals table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Democrats | Independants | Republicans |
| Blacks | 12.54209 | -3.598555 | -9.70631 |
| Whites | -12.54209 | 3.598555 | 9.70631 |

The standardized residuals give strong evidence of association. Since the standard reisual for black democrats is large (+12.5) we can infer that more black people identify themselves as democrates than if the null hypothesis were true, giving evidence between the association of being black and identifying as a democrat. Similarly since the residual for white Republicans is also a large (9.7) more white people identify themselves as republicans than one would expect under the null hypothesis. Hence, there is evidence of a positive association between a white people identifying as Republicans.

Lastly, since white Democrats and black Republicans have large negative residual values (-12.5, -9.7 respectively) implies that there are fewer people in these catagories than what one would expect if race and political party identification were independent. Hence, there is evidence of a negative association between the two. In lay terms, there is evidence that blacks tend to not be Repulicans.

#### Question 7c: Partition chi-squared into components regarding the choice between Democrat and Independent and between these two combined and Republican. Interpret.

Using the R code below we do the two partionings and then perform the test of independence on them. Also, note that the values of the two statistics add up to our original value.

First looking only at the Democrats and Independants.

# Partition chi-squared into components:  
table[-3,-3]

## Democrats Independants  
## Blacks 192 75  
## Whites 459 586

Performing the test:

G.test(table[-3,-3])

##   
## G-test  
##   
## data: [(table,-3,-3)  
## G = 68.448, df = 1, p-value < 2.2e-16

There is little evidence of a difference between the eclectic and medical schools of thought on the ascribed origin of schizophrenia. Since the p-value is very small again (pvalue << 0.0001) we have strong evidence of a difference between the race of White and black people identifying themselves as Independants or Democrats. Blacks are much more likely to be Democrats rather than Indepedant and whites seem more to identify themselves as Independants.

Now combining Democrats and Independants into a single column verses the Republicans, using the R code below, we get:

n11 = 192  
n12 = 75  
n13 = 8  
n21 = 456  
n22 = 586  
n23 = 471  
  
Dem.and.Ind = table[,1] + table[,2]  
Repubilcans = table[,3]  
table2 = cbind(Dem.and.Ind, Repubilcans)  
table2

## Dem.and.Ind Repubilcans  
## Blacks 267 8  
## Whites 1045 471

Performing the test:

G.test(table2)

##   
## G-test  
##   
## data: table2  
## G = 128.95, df = 1, p-value < 2.2e-16

Again, since the p-value is very small (pvalue << 0.0001) we have evidence against the null hypothesis of independence.

Now, it appears that there are rarely any black Republicans and that most of the Republicans are white. Hence, as p-value is so small we have evidence that that very few, blacks, identify themselves as Republicans, and majority of Republicans are white.

Secondly, we have strong evidence that most white people are likely to identify themselves as either Democrats or Independants.

## Question 8

#### Question 8a: Use logistic regression to model the effect of temperature on the probability of thermal distress. Plot a figure of the fitted model, and interpret.

## [1] 23

## [1] 23

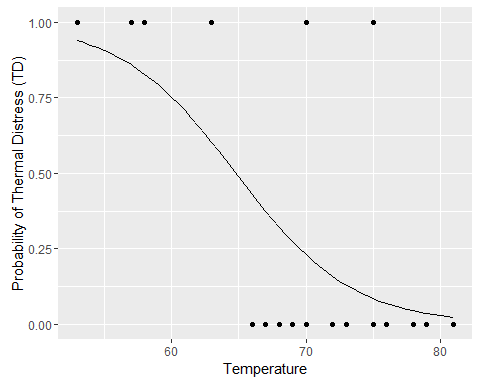
## [1] 23

We perform a logistic regression, with Thermal Distress as the Y variable.

# Do logistic regression  
model = glm(TD ~ Temp, family = "binomial")   
summary(model)

##   
## Call:  
## glm(formula = TD ~ Temp, family = "binomial")  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0611 -0.7613 -0.3783 0.4524 2.2175   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 15.0429 7.3786 2.039 0.0415 \*  
## Temp -0.2322 0.1082 -2.145 0.0320 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 28.267 on 22 degrees of freedom  
## Residual deviance: 20.315 on 21 degrees of freedom  
## AIC: 24.315  
##   
## Number of Fisher Scoring iterations: 5

Since the temperature in the model is statistically significant (p-value = 0.032) with a coefficient of -0.2322, clearly implies that the lower the temperature the higher the odds of at least one O-ring suffering thermal distress.

Now we plot the logistic regression curve:  From the plot it is clear that the lower the temperature the higher the probability of at least one primary O-ring to suffer thermal distress.

#### Question 8b:

Our model is:

This gives us the equation for the probability to be:

Hence, plugging in the values of the beta's we got from the model and for the temperature we get:

beta0 = 15.0429  
beta1 = -0.2322  
temp = 31  
p = exp(beta0 + beta1\*temp)/(1+exp(beta0+beta1\*temp))

A very large p-value , meaning that at that temperature, there as a very large probability of at least one O-ring suffering thermal stress and an accident happening due to that.

#### Question 8c:

##### Question: Construct a confidence interval for the effect of temperature on the odds of thermal distress, and test the statistical significance of the effect.

To construct a 95% confidence interval for the effect of temperature on the odds of thermal distress we use the R built in command confint() and then exponentiate it.

The confint() command gives us the loglikelihood confidence intervals and the reason we use the loglikelihood confidence intervals, is because first they are transformation invariant.

The loglikelihood confidence intervals are transformation invariant hence since we want the confidence interval for the odds and not log odds we need to exponentiate the upper and lower limits of the CI to get the CI of the odds ratio.

confint(model)

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 3.3305848 34.34215133  
## Temp -0.5154718 -0.06082076

exp(confint(model))

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 27.9546841 8.214986e+14  
## Temp 0.5972188 9.409919e-01

The 95% profile likelihood confidence interval for the effect on the log odds of suffereing thermal distress is , corresponding to CI for the odds ratio. The standard error used in the calculation of the CI are based on the expected information from Fisher scoring.

To test the effect of temperature on the probability of suffering thermal distress we use the loglikehood ratio test statistic is:

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: TD  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 22 28.267   
## Temp 1 7.952 21 20.315 0.004804 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We find that for the Loglikelihood ratio test with 1 degrees of freedom had test statistic of 7.952 and a p-value of 0.004804. Hence, the effect of temperature on the probability of suffering thermal distress is statistically significant.

## Question 9

#### Question: Construct the log-likelihood function for the model logit[7r(x)] = a + fix with independent binomial outcomes of yo successes in «o trials at x = 0 and y successes in n i trials at x = 1. Derive the likelihood equations, and show that 0 is the sample log odds ratio.

Since this is two independent binomial outcomes the joint probability function is