Assignment 2 - Categorical Data Analysis (for Biostatistics)

Name: Soufiane Fadel - Pablo Señas Peón
Date: October 17, 2018

Exercise 1

(a)

$\pi_{+1} - \pi_{1+}$ measures if the success is more likely to happen on first or second throw. That is: if $\pi_{+1} - \pi_{1+} > 0$ then the success is more likely to happen on the second throw. If $\pi_{+1} - \pi_{1+} < 0$ then the success is more likely to happen on the first. When $\pi_{+1} - \pi_{1+} = 0$ ($\pi_{+1} = \pi_{1+}$), then the probability of success is the same in either the first and the second throw. If this happens we have *marginal homogeneity*.

(b & c) **R Code :**

```
library ( package = PropCIs )
c.table <- array ( data = c(251, 48, 34, 5) , dim = c(2 ,2) , dimnames =
           list (First = c(" made ", " missed ") , Second =
           c(" made ", " missed ")))


#the Wald confident interval----------------------------------------
diffpropci.Wald.mp(b = c.table [1 ,2] , c = c.table [2 ,1] , n =sum (c.table
     ), conf.level = 0.95)


#the Agresti-Min confident interval---------------------------
diffpropci.mp(b = c.table [1 ,2] , c = c.table [2 ,1] , n =sum (c.table ),
    conf.level = 0.95)

#McNemar test
mcnemar.test(c.table,correct=FALSE)
```

Output:

```
         Second
First made missed
made 251 34
missed 48 5

data:
#Wald
95 percent confidence interval:
 -0.01090344 0.09374367
sample estimates:
[1] 0.04142012

data:
#Agresti-Min
95 percent confidence interval:
 -0.01115884 0.09351178
sample estimates:
[1] 0.04117647
```

```
McNemar's Chi-squared test

data: c.table
McNemar's chi-squared = 2.3902, df = 1, p-value = 0.1221
```

(d)

As we have seen both of our confidence intervals contain 0. On the other hand, the value of the statistic does not allow us to reject H_0 to accept H_1, as $p-value = 0.1221$. As a consequence, we do not have evidence against marginal homogeneity. This means that we do not have any evidence indicating that the probability of success in the second throw is greater than the probability of success in the first.

(e)

The warming up effect hypothesis states that the probability of attempt in the second throw will be greater than the probability of attempt in the first. The hypothesis test $\{H_0 : \pi_{+1} - \pi_{1+} = 0, H_a : \pi_{+1} - \pi_{1+} \neq 0\}$ does not entirely correspond to a warming up effect. We could reject H_0 if we have strong evidence supporting that $\pi_{+1} - \pi 1+ \leq 0$, which is completely contrary to the idea of the warming up hypothesis. This idea is, precisely, that $\pi_{+1} - \pi_{1+} \geq 0$, and that is why the best hypothesis test that we can use if we want to look for evidence supporting the warm up hypothesis is $\{H_0 : \pi_{+1} - \pi_{1+} \leq 0, H_a : \pi_{+1} - \pi_{1+} \geq 0\}$.
As we want a significance of 95% we will use $Z_{1-0.05}$ instead of $Z_{1-\frac{0.05}{2}}$ as our critical value, which is approximately 1.6449. Our test statistic, on the other hand, is

$$\frac{n_{21} - n_{12}}{\sqrt{n_{21} + n_{12}}} = \frac{48 - 34}{\sqrt{48 + 34}} \approx 1.5460$$

As our test statistic is lesser than the critical value, then we do not reject H_0 and again, we do not have strong evidence supporting the warm up hypothesis. We note that this result is expected, because in this case, the McNemar test is equivalent to what we have just done, and we rejected it.

Exercise 2

As we know, the Delta-method allows us to compute the variance of a non linear random variable, $g(\theta) = g((\theta_1, \cdots, \theta_p))$. We do this considering a linear approximation of $g(\hat{\theta})$, the estimator of $g(\theta)$ (this is just, as we can note, Taylor of first order):

$$g(\hat{\theta}) \approx g(\theta) + \sum_{k=1}^{p} g'(\theta_k)(\hat{\theta_k} - \theta_k)$$

In this case, $p = 2$, $\theta = OR$ and $g(\theta) = g(OR) = \log\left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right)$. Taking variances in the linear approximation expression, considering the development shown and class and taking into account that the two binomial experiments are independent (i.e., $Cov(\hat{\pi_1}, \hat{\pi_2}) = 0$), we have that

$$\hat{V}(\log(\widehat{OR})) = \left[\left(\frac{\partial \log(OR)}{\hat{\pi_1}}\right)(\hat{\pi}_1, \hat{\pi_2})\right]^2 Var(\hat{\pi}_1) + \left[\left(\frac{\partial \log(OR)}{\hat{\pi_1}}\right)(\hat{\pi}_2, \hat{\pi_2})\right]^2 Var(\hat{\pi}_2).$$

As π_1, π_2 are the probabilities of success of binomial experiments, we have that

$$Var(\hat{\pi}_i) = Var\left(\frac{w_i}{n_i}\right) = \frac{1}{n_i^2} Var(w_i) = \frac{n_i \hat{\pi}_i(1 - \hat{\pi}_i)}{n_i^2} = \frac{\hat{\pi}_i(1 - \hat{\pi}_i)}{n_i}.$$

We also have that

$$\log(OR) = \log\left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right) = \log(\pi_1(1-\pi_2)) - \log(\pi_2(1-\pi_1).$$

If we calculate the partial derivatives:

$$\frac{\partial \log(OR)}{\partial \pi_1} = \frac{1-\pi_2}{\pi_1(1-\pi_2)} - \frac{\pi_2(1-\pi_1)}{-\pi_2} = \frac{1}{\pi_1} + \frac{1}{1-\pi_1} = \frac{1}{\pi_1(1-\pi_1)}$$

$$\frac{\partial \log(OR)}{\partial \pi_2} = \frac{-\pi_1}{\pi_1(1-\pi_2)} - \frac{1-\pi_1}{\pi_2(1-\pi_1)} = -\frac{1}{1-\pi_2} - \frac{1}{\pi_2} = -\frac{1}{\pi_2(1-\pi_2)}$$

We finally have that

$$\begin{aligned}\hat{V}(\log(\hat{OR})) &= \frac{\hat{\pi}_1(1-\hat{\pi_1})}{n_1(\hat{\pi}_1(1-\hat{\pi_1}))^2} + \frac{\hat{\pi}_2(1-\hat{\pi_2})}{n_2(\hat{\pi}_2(1-\hat{\pi_2}))^2} = \frac{1}{n_1\hat{\pi}_1(1-\hat{\pi}_1)} + \frac{1}{n_2\hat{\pi}_2(1-\hat{\pi}_2)} \\ &= \frac{1}{n_1\frac{w_1}{n_1}(1-\frac{w_1}{n_1})} + \frac{1}{n_2\frac{w_2}{n_2}(1-\frac{w_2}{n_2})} = \frac{n_1}{w_1(n_1-w_1)} + \frac{n_2}{w_2(n_2-w_2)} \\ &= \frac{1}{w_1} + \frac{1}{n_1-w_1} + \frac{1}{w_2} + \frac{1}{n_2-w_2}\end{aligned}$$

Exercise 3

(a)

$$\begin{aligned}odds_1 = \frac{P(A|B)/P(A'|B)}{P(A|B')P(A'|B')} &= \frac{\frac{P(B|A)P(A)}{P(B)} / \frac{P(B|A')P(A')}{P(B)}}{\frac{P(B'|A)P(A)}{P(B')} / \frac{P(B'|A')P(A')}{P(B')}} \\ &= \frac{P(B|A)/P(B'|A)}{P(B|A')P(B'|A')} = odds_2\end{aligned}$$

(b) In part (a) we have seen that estimating estimating the odds of success of X as a function of Y is the same as estimating the odds of success of Y as a function of X. This is extremely useful when we have a variable Y as a function of X but we want to get some insight about X as a function of Y. The odds ratio allows us to do that, as we have seen. For example, this property can be applied in case control studies, as we will see.

Exercise 4

(a)

Identify the response variable and the explanatory variable.

In this experiment we first have selected people according to their lung cancer state and then, in both groups, we have studied whether they are smokers or not. That said, we conclude that the explanatory variable is their lung cancer condition (if they have it or not) and the response variable is if they are smokers.

(b)

Identify the type of study this was.

This was a case control study, also known as a retrospective study. In this type of studies the researches proceed to take 2 groups of people,one with a specific medical condition and the other without it, and the behaviour of another variable in both groups. This allows the researchers to establish relations between the disease and the variable studied. In this example we are trying to see if there is any relation between having lung cancer and smoking.

(c)

Can you use these data to compare smokers with nonsmokers in terms of the proportion who suffered lung cancer? Explain why or why not.

We cannot do that. According to the textbook (page 41), *the proportions of cases and controls in the sample [...] is chosen by the researcher and does not typically match the proportion of cases in the population.* For this reason if we were to compare smokers with nonsmokers in terms of the proportion, we will be most likely commiting a mistake.

(d)

Summarize the association with a descriptive parameter. Explain how to interpret it. Also give a 95% confidence interval for this parameter.

The best descriptive parameter that we can use in this situation is the odds ratio. If $\widehat{OR}$ is the estimator of OR, the interpretation is that the odds of smoking are $\widehat{OR}$ greater for those who have lung cancer than for those who do not. However, as we have seen in class and formalized in Exercise 3, this is equivalent to stating that the odds of having lung cancer are $\widehat{OR}$ greater for those who smoke than for those who do not smoke. As a consequence, the odds ratio allows us to *invert* the experiment and, if we were first studying the smoking behaviour according to having lung cancer (or not), we are now analyzing if there is any relation between smoking and having lung cancer (we note that this is extremely powerful in the case of rare diseases, where we are able to approximate the risk ratio using the odds ratio). We proceed to compute the odds ratio and its 95% confidence interval:

R code :

```
library(Epi)

c.table <- array(data = c(688, 21, 650, 59) , dim = c(2 ,2) , dimnames =
                   list ("Have smoked" = c(" yes ", " no ") , "Lung cancer"
                       =
                           c("cases", "controls")))
twoby2(c.table, alpha = 0.05)
```

Output :

```
2 by 2 table analysis:
------------------------------------------------------
Outcome : cases
Comparing : yes vs. no

     cases controls P(cases) 95% conf. interval
 yes 688 650 0.5142 0.4874 0.5409
 no 21 59 0.2625 0.1778 0.3694

                                95% conf. interval
            Relative Risk: 1.9589 1.3517 2.8387
        Sample Odds Ratio: 2.9738 1.7867 4.9494
Conditional MLE Odds Ratio: 2.9716 1.7556 5.2107
    Probability difference: 0.2517 0.1427 0.3398

            Exact P-value: 0
       Asymptotic P-value: 0
------------------------------------------------------
```

We get that $\widehat{OR} = 2.9738 \approx 3$. As we have just mentioned, we can interpret this saying that the odds for a smoker person to have lung cancer are 3 times larger than the odds for a non smoker person. The 95% confidence interval for $\widehat{OR}$ is $(1.7867, 4.9494)$. If we consider the hypothesis test $\{H_0 : \text{OR}= 1,\ H_a : \text{OR}\neq 1\}$ we would definitely reject the null hypothesis, and we can conclude that smoking is a determinant factor to lung cancer.

Exercise 5

(a)

Find the maximum likelihood estimates for β_0 and β_1. State the estimated response function.

R code :

```
connection <- textConnection( "deposit ; solde ; returned
                      2.0 ; 500.0 ; 72.0
                      5.0 ; 500.0 ; 103.0
                      10.0 ; 500.0 ; 170.0
                      20.0 ; 500.0 ; 296.0
                      25.0 ; 500.0 ; 406.0
                      30.0 ; 500.0 ; 449.0")
bottles <- read.csv(connection, sep=";")
mod <- glm( returned/solde ~ deposit , weights = solde, family = binomial(
    link = logit), data = bottles)
summary(mod)
```

Output :

```
Call:
glm(formula = returned/solde ~ deposit, family = binomial(link = logit),
    data = bottles, weights = solde)

Deviance Residuals:
     1 2 3 4 5
 0.1754 0.4330 0.5784 -2.9193 1.2710
     6
 1.2209

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.076565 0.084839 -24.48 <2e-16
deposit 0.135851 0.004772 28.47 <2e-16

(Intercept) ***
deposit ***
---
Signif. codes:
0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1108.171 on 5 degrees of freedom
Residual deviance: 12.181 on 4 degrees of freedom
AIC: 53.419

Number of Fisher Scoring iterations: 3
```

The maximum likelihood estimates for β_0 and β_1 are:

$$\begin{cases} \hat{\beta}_0 \approx -2.076 \\ \hat{\beta}_1 \approx 0.135 \end{cases}$$

Let denote π_i the probability that bottles will be returned :

$$\begin{aligned} logit\,(\hat{\pi}_i) &= \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) \\ &= \beta_0 + \beta_1 x_i \\ &= -2.076 + 0.135 x_i \end{aligned}$$

As a result :

$$\hat{\pi}_i = \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1 x_i\}}{1+\exp\{\hat{\beta}_0 + \hat{\beta}_1 x_i\}} = \frac{\exp\{-2.076 + 0.135 x_i\}}{1+\exp\{-2.076 + 0.135 x_i\}}$$

(b)

Obtain a scatter plot of the sample proportions against the level of the deposit, and superimpose the estimated logistic response onto the plot.

```
# plot proportions versus x
with(bottles, plot( x = deposit, y = returned/solde,
                    xlab = "Deposit level (in cents)", ylab = "Proportion
                        returned",
                    panel.first = grid(col = "gray", lty = "dotted")))

#Put estimated logistic response on the plot
curve(expr = predict(object = mod, newdata = data.frame( deposit = x ), type
     = "response"), col = "red", add = TRUE)
```

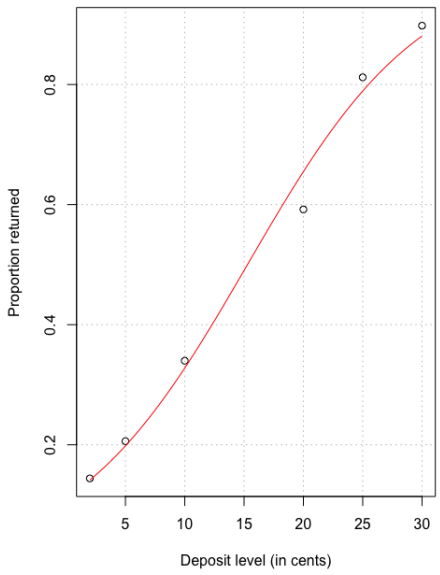

Figure 1: Scatter plot of the sample proportions against the level of the deposit

Does the fitted logistic response function appear to fit well?

From the plot we see that the sample proportions (scatter) is close to the estimated logistic response (red line), so we can say that out logistic model fits our data well.

(c)

Obtain $\exp(\hat{\beta}_1)$ and interpret this number.

```
#the estimated odds ratio for deposit
exp(mod$coefficients[2])
#Output
# deposit
#1.145511
```

Interpretation :
In class we have seen that e^{β_j} is related to the odds ratio of the response variable to x_i when we increment x_i by 1. As a consequence, we estimate that the odds of returned bottles increases by 14.55% for every cent in the deposit.

(d)

What is the estimated probability that a bottle will be returned when the deposit is 15 cents?

```
as.numeric(predict(mod,newdata = data.frame(deposit=15),type = "response"))
#[1] 0.4903005
```

$$\hat{\pi}(15) = \frac{\exp\{-2.076 + 0.135(15)\}}{1 + \exp\{-2.076 + 0.135(15)\}} = 0.4903$$

(e)

Estimate the amount of deposit for which 75% of the bottles are expected to be returned.

$$logit(\hat{\pi}_i) = \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_i \Longrightarrow \hat{x}_i = \frac{\text{logit}(\pi_i) - \hat{\beta}_0}{\hat{\beta}_1} =: g(\hat{\beta}_0, \hat{\beta}_1)$$

Replacing by values of estimated parameters the last equation:

$$\hat{x}_i = \frac{\text{logit}(0.75) - \hat{\beta}_0}{\hat{\beta}_1} = \frac{\log\left(\frac{0.75}{0.25}\right) - (-2.076)}{0.135}$$

gives an estimate of a deposit of 23.52 cents for 75% of bottles to be returned.

(f)

In part (e), we have an estimate $\hat{x} = g(\hat{\beta}_0, \hat{\beta}_1)$ for the level of the deposit that corresponds to $\pi = 75\%$ of the bottles are returned. This estimator is a non-linear function of $\hat{\beta}_0$ and $\hat{\beta}_1$. Use the delta-method to find an asymptotic estimated standard error for this estimate.

Same as in exercise 2, the delta-method allows as to compute the variance of a non linear random variable, here we will use the vectorial version of this method :

$$\widehat{Var}(g(\hat{\beta}_0, \hat{\beta}_1)) \approx \nabla g(\hat{\beta}_0, \hat{\beta}_1)' \Gamma_{\hat{\beta}_0, \hat{\beta}_1} \nabla g(\hat{\beta}_0, \hat{\beta}_1)$$

with

$$\begin{cases} \Gamma_{\hat{\beta}_0, \hat{\beta}_1} \text{it's the variance-covariance matrix of } \hat{\beta}_0, \hat{\beta}_1 \\ g(\hat{\beta}_0, \hat{\beta}_1) := \dfrac{\text{logit}(\pi_i) - \hat{\beta}_0}{\hat{\beta}_1} \end{cases}$$

$$\begin{aligned} \nabla g(\hat{\beta}_0, \hat{\beta}_1) &= \begin{bmatrix} \dfrac{\partial g}{\partial \beta_0} \\ \dfrac{\partial g}{\partial \beta_1} \end{bmatrix}_{(\beta_0, \beta_1) = (\hat{\beta}_0, \hat{\beta}_1)} \\ &= \begin{bmatrix} -1/\hat{\beta}_1 \\ -(logit(\pi_i) - \hat{\beta}_0)/\hat{\beta}_1^2 \end{bmatrix} \end{aligned}$$

```
logit.pi <- log(0.75/(1-0.75))
#Gradient of g
grad <- as.numeric(c(-1/mod$coefficients[2], -(logit.pi - mod$coefficients[1] )/
                     mod$coefficients[2]**2))

#variance-covariance matrix of parameters
vbeta <- vcov(mod)
vg <- t(grad) %*% vbeta %*% grad
sqrt(vg)

          [,1]
[1,] 0.4392811
```

Exercise 6

(a)

Find the maximum likelihood estimates of β_0, β_1, and β_2. State the estimated response function.

R code :

```
cars<-read.csv("CarPurchase.csv")
Mod <- glm( y ~ x1 + x2 , family = binomial(link = logit), data = cars)
summary(Mod)
```

Output :

```
Call:
glm(formula = y ~ x1 + x2, family = binomial(link = logit), data = cars)

Deviance Residuals:
    Min 1Q Median 3Q Max
-1.6189 -0.8949 -0.5880 0.9653 2.0846

Coefficients:
           Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) -4.73931 2.10195 -2.255 0.0242 *
x1 0.06773 0.02806 2.414 0.0158 *
x2 0.59863 0.39007 1.535 0.1249
---
Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.987 on 32 degrees of freedom
Residual deviance: 36.690 on 30 degrees of freedom
AIC: 42.69

Number of Fisher Scoring iterations: 4
```

The maximum likelihood estimates for β_0 , β_1 and β_2 are :

$$\begin{cases} \hat{\beta}_0 \approx -4.73931 \\ \hat{\beta}_1 \approx 0.06773 \\ \hat{\beta}_2 \approx 0.59863 \end{cases}$$

let denote $\pi_i = P(Y_i = 1)$ the probability that a family i will purchase a new car next year

$$\begin{aligned} logit\,(\hat{\pi}_i) &= \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) \\ &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \\ &= -4.73931 + 0.06773 x_{1i} + 0.59863 x_{2i} \end{aligned}$$

as a result :

$$\hat{\pi}_i = \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}\}}{1 + \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}\}} = \frac{\exp\{-4.73931 + 0.06773 x_{1i} + 0.59863 x_{2i}\}}{1 + \exp\{-4.73931 + 0.06773 x_{1i} + 0.59863 x_{2i}\}}$$

(b)

Obtain $\exp(\hat{\beta}_1)$ and $\exp(\hat{\beta}_2)$ and interpret these numbers.

R code :

```
#the estimated odds ratio for x_1 (annual family income in thousands of
    dollars)
> exp(Mod$coefficients[2])
     x1
1.070079

#the estimated odds ratio for x_2 (the current age of the oldest family
    automobile in years)
> exp(Mod$coefficients[3])
     x2
1.819627
```

Interpreting $\exp(\hat{\beta}_1)$ **:** We see that the odds ratio corresponding to income is 1.070. As we have seen before, this implies that if we fix the age of the oldest car, increasing family income by one thousand dollars will increase the odds of purchasing a new car by 7%.

Interpreting $\exp(\hat{\beta}_2)$ **:** On the other hand, we see that the odds ratio corresponding to age is 1.8196. This implies that if we fix the annual family income, increasing the current age of the oldest family automobile by one year will increase the odds of purchasing a new car by 81.96 %.

(c)

What is the estimated probability that a family with annual income of $50 thousand and an oldest car of 3 years will purchase a new car next year?

R code :

```
> as.numeric(predict(Mod,newdata = data.frame(x1=50, x2=3),type = "response"
    ))
[1] 0.6090245
```

This tells us that the predicted probability that a family with annual income of $50 thousand and an oldest car of 3 years will purchase a new car next year is : 0.6090245.