Assignment 2

Hanxiao Du, Jeffery Wei Xuan Su

10/12/2020

Q1. (a)

```
set.seed(1004439529)
pr = c(0.2,0.3,0.3,0.2)
n=10
N=100
rv = rmultinom(n, N, pr)
rv[rv==0]<-0.5
tables = list()
for (i in 1:n)
{
  counts = matrix(rv[,i], ncol=2,byrow=TRUE)
  colnames(counts) = c("A", "B")
  rownames(counts) = c("1", "2")

  counts.table = as.table(counts)
  tables[[i]] = counts.table
}

odd_ratio = numeric(n)
for (i in 1:n)
{
  odd_ratio[i] = (tables[[i]][1]*tables[[i]][4])/
    (tables[[i]][2]*tables[[i]][3])
}
print(odd_ratio)
```

```
##  [1] 0.8500750 0.4861111 0.8513932 0.2317291 0.2317291 0.4761905 0.7058824
##  [8] 0.4769585 0.5146871 0.5681818
```

i. $\theta = \frac{Odds1}{Odds2} = \frac{\frac{0.4}{0.6}}{\frac{0.6}{0.4}} = \frac{4}{9}$

ii.

```
theta = 4/9
count = 0
CI_list = list()
for (i in 1:n) {
  table = tables[[i]]
  SE = sqrt(1/table[1] + 1/table[2]+1/table[3] + 1/table[4])
  CI_log = c(log(odd_ratio[i]) -1.96*(SE), log(odd_ratio[i]) + 1.96*(SE))
  CI = exp(CI_log)
  CI_list[[i]] = CI
}

for (i in 1:n) {
  if (CI_list[[i]][1] <= theta && theta<= CI_list[[i]][2]) {
    count = (count + 1)
  }
}
count
```

```
## [1] 10
```

All 10 intervals contain θ.

iii.

```
# Tables
print(tables)
```

```
## [[1]]
##    A  B
## 1 27 29
## 2 23 21
##
## [[2]]
##    A  B
## 1 20 27
## 2 32 21
##
## [[3]]
##    A  B
## 1 22 19
## 2 34 25
##
## [[4]]
##    A  B
## 1 20 34
## 2 33 13
##
## [[5]]
##    A  B
## 1 20 34
## 2 33 13
##
## [[6]]
##    A  B
## 1 25 24
## 2 35 16
##
## [[7]]
##    A  B
## 1 21 34
## 2 21 24
##
## [[8]]
##    A  B
## 1 18 31
## 2 28 23
##
## [[9]]
##    A  B
## 1 13 29
## 2 27 31
##
## [[10]]
##    A  B
## 1 18 33
## 2 24 25
```

```
# Estimated odd ratios
print(odd_ratio)
```

```
##  [1] 0.8500750 0.4861111 0.8513932 0.2317291 0.2317291 0.4761905 0.7058824
##  [8] 0.4769585 0.5146871 0.5681818
```

```
# C.I.
print(CI_list)
```

```
## [[1]]
## [1] 0.3856532 1.8737751
##
## [[2]]
## [1] 0.218787 1.080064
##
## [[3]]
## [1] 0.3817304 1.8989067
##
## [[4]]
## [1] 0.09936673 0.54040577
##
## [[5]]
## [1] 0.09936673 0.54040577
##
## [[6]]
## [1] 0.2108627 1.0753793
##
## [[7]]
## [1] 0.3173873 1.5699111
##
## [[8]]
## [1] 0.2141011 1.0625328
##
## [[9]]
## [1] 0.2237175 1.1840952
##
## [[10]]
## [1] 0.2547331 1.2673286
```

(b)

i.

```
set.seed(1004439529)
pr = c(0.2,0.3,0.3,0.2)
n=1000000
N=100
rv = rmultinom(n, N, pr)
rv[rv==0]<-0.5
tables = list()
for (i in 1:n)
{
  counts = matrix(rv[,i], ncol=2,byrow=TRUE)
  colnames(counts) = c("A", "B")
  rownames(counts) = c("1", "2")

  counts.table = as.table(counts)
  tables[[i]] = counts.table
}

odd_ratio = numeric(n)
for (i in 1:n)
{
  odd_ratio[i] = (tables[[i]][1]*tables[[i]][4])/
    (tables[[i]][2]*tables[[i]][3])
}

CI_list = list()
for (i in 1:n) {
  table = tables[[i]]
  SE = sqrt(1/table[1] + 1/table[2]+1/table[3] + 1/table[4])
  CI_log = c(log(odd_ratio[i]) -1.96*(SE), log(odd_ratio[i]) + 1.96*(SE))
  CI = exp(CI_log)
  CI_list[[i]] = CI
}

theta = 4/9
count = 0
for (i in 1:n) {
  if (CI_list[[i]][1] <= theta && CI_list[[i]][2] >= theta) {
    count = count + 1
  }
}
count/n
```

```
## [1] 0.949464
```

The proportion of the intervals containing θ is 0.949464.

ii. We have a true coverage probability (94.9464%) of the C.I. which is slightly lower than the targeted coverage probability (95%).

(c)

```r
set.seed(1004439529)
pr = c(0.2,0.3,0.3,0.2)
n=1000000
N=15
rv = rmultinom(n, N, pr)
rv[rv==0]<-0.5
tables = list()
for (i in 1:n)
{
  counts = matrix(rv[,i], ncol=2,byrow=TRUE)
  colnames(counts) = c("A", "B")
  rownames(counts) = c("1", "2")

  counts.table = as.table(counts)
  tables[[i]] = counts.table
}

odd_ratio = numeric(n)
for (i in 1:n)
{
  odd_ratio[i] = (tables[[i]][1]*tables[[i]][4])/
    (tables[[i]][2]*tables[[i]][3])
}

CI_list = list()
for (i in 1:n) {
  table = tables[[i]]
  SE = sqrt(1/table[1] + 1/table[2]+1/table[3] + 1/table[4])
  CI_log = c(log(odd_ratio[i]) -1.96*(SE), log(odd_ratio[i]) + 1.96*(SE))
  CI = exp(CI_log)
  CI_list[[i]] = CI
}

theta = 4/9
count = 0
for (i in 1:n) {
  if (CI_list[[i]][1] <= theta && CI_list[[i]][2] >= theta) {
    count = count + 1
  }
}
count/n
```

```
## [1] 0.972009
```

The proportion of the intervals containing θ is 97.2009% which is higher than true coverage probalility of the case with N=100 (94.9464%). Furthermore, it is higher than the targeted coverage probability (95%) as well.

Q2. (a) $H_0 : \theta = 1$ and $H_a : \theta > 1$

```
data = matrix(c(21,2,15,3), nrow=2, byrow=TRUE,
              dimnames = list(Control = c("Cancer COntrolled", "Cancer Not Controlled"), Treatment=c("Su

p_value = sum(dhyper(21:23, sum(data[1,]), sum(data[2,]), sum(data[,1])))
p_value
```

```
## [1] 0.3808337
```

Conclusion since p-value (0.3808337) is greater than α (0.05), so we failed to reject H_0, which implies that there is no strong evidence that surgery is better than radiation therapy.

(b)

```
prob =dhyper(0:23, sum(data[1,]), sum(data[2,]), sum(data[,1]))
pvalue = sum((prob)*(prob<=prob[data[1]+1]))
pvalue
```

```
## [1] 0.6384258
```

Conclusion: since p-value (0.6384258) is greater than α (0.05), so we failed to reject H_0, which implies that there is no strong evidence that the results are different by using surgery and radiation therapy.

Q3. (a)

```
votes = matrix(c(154,132,180,126,104,131),
                ncol=2, byrow = TRUE,
                dimnames = list(Vote = c("Support", "Oppose", "Do not know"), Collage_Grad =
                                  c("Yes", "No")))
chisq_test <- chisq.test(votes)
chisq_test$p.value
```

```
## [1] 0.003245752
```

Conclusion: since p-value(0.003245752) $<\alpha$ (0.05), thus we reject H_0.So they are not independent to each other, the vote proportion is dependent on if the participant is a college graduate or not.

(b)

```
x = votes[1,]
n = c(sum(votes[,1]), sum(votes[,2]))
result <- prop.test(x, n)
result$p.value
```

```
## [1] 0.7665154
```

Conclusion: since p-value(0.7665154) $>\alpha$ (0.05), thus we do not have enough evidence to reject H_0, meaning the proportion of college graduates supporting of offshore drilling equals to the proportion of non-college graduates supporting offshore drilling with strong evidence.

(c) The Conclusions in part (a) and (b) do not agree. Since we had the third row which is "Do not know", for the two sample test, it assumes that anything other than "Support" is in the other category. However, we have two more categories instead of one.

(d)

```
chisq_test$stdres
```

```
##              Collage_Grad
## Vote                 Yes         No
##   Support      0.3701737 -0.3701737
##   Oppose       2.5879807 -2.5879807
##   Do not know -3.1608057  3.1608057
```

Conclusion: Since the standardized residuals of "Oppose" and "Do not know" are greater than 2 in absolute values. Therefore, the results of the vote for "Oppose" and "Do not know" are more associated with if the participant is a college graduate or not.

Q.4 (a)

```
table_generate = function(x)
{
  return(matrix(x, nrow = 2, byrow = TRUE, dimnames = list(c("Drug", "Control"),
                                                      c("Success", "Failiure"))))
}

odds_ratio = function(matrix)
{
  return((matrix[1,1]*matrix[2,2])/
           (matrix[1,2]*matrix[2,1]))
}

center_data = list(table_generate(c(11, 25, 10, 27)),
table_generate(c(16, 4, 22, 10)),
table_generate(c(14, 5, 7, 12)),
table_generate(c(2, 14, 1, 16)),
table_generate(c(6, 11, 0, 12)),
table_generate(c(1, 10, 0, 10)),
table_generate(c(1, 4, 1, 8)),
table_generate(c(4, 2, 6, 1))
)

adjusted_center_data = list(table_generate(c(11, 25, 10, 27)),
table_generate(c(16, 4, 22, 10)),
table_generate(c(14, 5, 7, 12)),
table_generate(c(2, 14, 1, 16)),
table_generate(c(6, 11, 0.5, 12)),
table_generate(c(1, 10, 0.5, 10)),
table_generate(c(1, 4, 1, 8)),
table_generate(c(4, 2, 6, 1))
)

odds_ratio_1 = odds_ratio(adjusted_center_data[[1]])
odds_ratio_2 = odds_ratio(adjusted_center_data[[2]])
```

```
odds_ratio_5 = odds_ratio(adjusted_center_data[[5]])
print(odds_ratio_1)
```

```
## [1] 1.188
```

```
print(odds_ratio_2)
```

```
## [1] 1.818182
```

```
print(odds_ratio_5)
```

```
## [1] 13.09091
```

(b)

```
marginal_table = Reduce("+", center_data)
marginal_odds_ratio = odds_ratio(marginal_table)
print(marginal_odds_ratio)
```

```
## [1] 1.497872
```

Without considering different centers, odds of Success for Drug curing an infection is higher then the odds of placebo for curing an infection.

(c) Since the success rate is related to the center of the clinical trial. Thus, if the independence test is conducted without considering centers, then there is a potential chance that we can get Simpson's paradox.

(d)

```
n11ks = c()
means = c()
vars = c()
for (i in 1:length(center_data))
{
  matrix = center_data[[i]]
  n11k = matrix[1,1]
  n11ks[i] = n11k
  tk = sum(matrix[,])
  R1k = sum(matrix[1,])
  C1k = sum(matrix[,1])
  R2k = sum(matrix[2,])
  C2k = sum(matrix[,2])
  print(sprintf("T%d: %d", i, tk))
  print(sprintf("R1%d: %d", i, R1k))
  print(sprintf("R2%d: %d", i, R2k))
  print(sprintf("C1%d: %d", i, C1k))
  print(sprintf("C2%d: %d", i, C2k))
  mean_numerator = R1k*C1k
  print(sprintf("mean numerator %d: %d", i, mean_numerator))
```

```r
  mean_denom = tk
  print(sprintf("mean denominator %d: %d", i, mean_denom))
  mean = mean_numerator/mean_denom
  means[i] = mean
  var_numerator = R1k*R2k*C1k*C2k
  print(sprintf("var numerator %d: %d", i, var_numerator))
  var_denom = (tk^2)*(tk-1)
  print(sprintf("var denominator %d: %d", i, var_denom))
  var = var_numerator/var_denom
  vars[i] = var
}
```

```
## [1] "T1: 73"
## [1] "R11: 36"
## [1] "R21: 37"
## [1] "C11: 21"
## [1] "C21: 52"
## [1] "mean numerator 1: 756"
## [1] "mean denominator 1: 73"
## [1] "var numerator 1: 1454544"
## [1] "var denominator 1: 383688"
## [1] "T2: 52"
## [1] "R12: 20"
## [1] "R22: 32"
## [1] "C12: 38"
## [1] "C22: 14"
## [1] "mean numerator 2: 760"
## [1] "mean denominator 2: 52"
## [1] "var numerator 2: 340480"
## [1] "var denominator 2: 137904"
## [1] "T3: 38"
## [1] "R13: 19"
## [1] "R23: 19"
## [1] "C13: 21"
## [1] "C23: 17"
## [1] "mean numerator 3: 399"
## [1] "mean denominator 3: 38"
## [1] "var numerator 3: 128877"
## [1] "var denominator 3: 53428"
## [1] "T4: 33"
## [1] "R14: 16"
## [1] "R24: 17"
## [1] "C14: 3"
## [1] "C24: 30"
## [1] "mean numerator 4: 48"
## [1] "mean denominator 4: 33"
## [1] "var numerator 4: 24480"
## [1] "var denominator 4: 34848"
## [1] "T5: 29"
## [1] "R15: 17"
## [1] "R25: 12"
## [1] "C15: 6"
## [1] "C25: 23"
## [1] "mean numerator 5: 102"
## [1] "mean denominator 5: 29"
## [1] "var numerator 5: 28152"
## [1] "var denominator 5: 23548"
## [1] "T6: 21"
## [1] "R16: 11"
## [1] "R26: 10"
## [1] "C16: 1"
## [1] "C26: 20"
## [1] "mean numerator 6: 11"
## [1] "mean denominator 6: 21"
## [1] "var numerator 6: 2200"
## [1] "var denominator 6: 8820"
## [1] "T7: 14"
## [1] "R17: 5"
## [1] "R27: 9"
## [1] "C17: 2"
## [1] "C27: 12"
## [1] "mean numerator 7: 10"
## [1] "mean denominator 7: 14"
## [1] "var numerator 7: 1080"
## [1] "var denominator 7: 2548"
## [1] "T8: 13"
## [1] "R18: 6"
## [1] "R28: 7"
## [1] "C18: 10"
## [1] "C28: 3"
## [1] "mean numerator 8: 60"
## [1] "mean denominator 8: 13"
## [1] "var numerator 8: 1260"
## [1] "var denominator 8: 2028"
```

```
CMH = sum(n11ks-means)/sqrt(sum(vars))
p.value = 2*(1-pnorm(CMH)) # 2-sided
p.value
```

```
## [1] 0.01151463
```

(e)

```
numerator_fnc = function(matrix, i) {
  sum = sum(matrix)
  cat("n11 for center",i, ":", matrix[1], "\n")
  cat("n22 for center",i, ":", matrix[4], "\n")
  cat("sum of center",i, ":", sum, "\n")
  return((matrix[1] * matrix[4])/sum)
}
denominator_fnc = function(matrix, i) {
  sum = sum(matrix)
  cat("n12 for center",i, ":", matrix[3], "\n")
  cat("n21 for center",i, ":", matrix[2], "\n")
  cat("sum of center",i, ":", sum, "\n")
  return((matrix[2] * matrix[3])/sum)
}

numerator = 0
denominator = 0
for (i in 1:8){
  numerator_value = numerator_fnc(center_data[[i]], i)
  numerator = numerator + numerator_value
  cat("numerator for center", i, ":", numerator_value, "\n")

  denominator_value = denominator_fnc(center_data[[i]], i)
  denominator = denominator + denominator_value
  cat("denominator for center", i, ":", denominator_value, "\n\n")
}
```

```
## n11 for center 1 : 11
## n22 for center 1 : 27
## sum of center 1 : 73
## numerator for center 1 : 4.068493
## n12 for center 1 : 25
## n21 for center 1 : 10
## sum of center 1 : 73
## denominator for center 1 : 3.424658
##
## n11 for center 2 : 16
## n22 for center 2 : 10
## sum of center 2 : 52
## numerator for center 2 : 3.076923
## n12 for center 2 : 4
## n21 for center 2 : 22
## sum of center 2 : 52
## denominator for center 2 : 1.692308
##
## n11 for center 3 : 14
## n22 for center 3 : 12
## sum of center 3 : 38
## numerator for center 3 : 4.421053
## n12 for center 3 : 5
## n21 for center 3 : 7
## sum of center 3 : 38
## denominator for center 3 : 0.9210526
##
## n11 for center 4 : 2
## n22 for center 4 : 16
## sum of center 4 : 33
## numerator for center 4 : 0.969697
## n12 for center 4 : 14
## n21 for center 4 : 1
## sum of center 4 : 33
## denominator for center 4 : 0.4242424
##
## n11 for center 5 : 6
## n22 for center 5 : 12
## sum of center 5 : 29
## numerator for center 5 : 2.482759
## n12 for center 5 : 11
## n21 for center 5 : 0
## sum of center 5 : 29
## denominator for center 5 : 0
##
## n11 for center 6 : 1
## n22 for center 6 : 10
## sum of center 6 : 21
## numerator for center 6 : 0.4761905
## n12 for center 6 : 10
## n21 for center 6 : 0
## sum of center 6 : 21
## denominator for center 6 : 0
##
## n11 for center 7 : 1
## n22 for center 7 : 8
## sum of center 7 : 14
## numerator for center 7 : 0.5714286
## n12 for center 7 : 4
## n21 for center 7 : 1
## sum of center 7 : 14
## denominator for center 7 : 0.2857143
##
## n11 for center 8 : 4
## n22 for center 8 : 1
## sum of center 8 : 13
## numerator for center 8 : 0.3076923
## n12 for center 8 : 2
## n21 for center 8 : 6
## sum of center 8 : 13
## denominator for center 8 : 0.9230769
```

```
cat("\n common odds ratio:", numerator/denominator)
```

```
##
##  common odds ratio: 2.134549
```

(f)

```
center_data_array = array(as.numeric(unlist(center_data)), dim=c(2, 2, 8))
mantelhaen.test(center_data_array, alternative="two.sided", correct = FALSE)
```

```
##
##  Mantel-Haenszel chi-squared test without continuity correction
##
## data:  center_data_array
## Mantel-Haenszel X-squared = 6.3841, df = 1, p-value = 0.01151
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.177590 3.869174
## sample estimates:
## common odds ratio
##          2.134549
```

Q.5 (a)

```
mydata = data.frame(drinks = c(0,0.5,1.5,4,7),
absent = c(17066, 14464, 788, 126, 37),
present = c(48, 38, 5, 1, 1) )
mydata$total = with(mydata, absent + present)
mydata$proportion = with(mydata, present/total)

mydata.linear = glm(cbind(mydata$present, mydata$absent)~drinks, family=binomial(link="identity"), data
summary(mydata.linear)
```

```
##
## Call:
## glm(formula = cbind(mydata$present, mydata$absent) ~ drinks,
##     family = binomial(link = "identity"), data = mydata)
##
## Deviance Residuals:
##       1        2        3        4        5
##  0.6564  -1.0492   0.8631   0.1302   0.8282
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.0025476  0.0003523   7.232 4.77e-13 ***
## drinks      0.0010872  0.0008324   1.306    0.192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6.2020  on 4  degrees of freedom
## Residual deviance: 2.9795  on 3  degrees of freedom
## AIC: 25.606
##
## Number of Fisher Scoring iterations: 10
```

(b)

i. $\hat{\pi}(x) = 0.002548 + 0.001087x$

ii. Coefficient: When mother's alchol consumption is increased by 1, probability of a baby having sex organ malformation is expected to increase by 0.001087. Intercept: when the mother's alchol consumption is 0 the expected probability of a baby having sex organ malformation is 0.002548.

iii. $\hat{\pi}(0) = 0.002548 + 0.001087 \times 0 = 0.002548$ $\hat{\pi}(7) = 0.002548 + 0.001087 \times 7 = 0.010157$

iv. $\hat{RR} = \frac{\hat{\pi}(0)}{\hat{\pi}(7)} = \frac{0.002548}{0.010157} = 0.2508615$ Interpretation: Babies of mother whose alcohol consumption is 7 during pregnancy having sex organ malformation is 0.2508615 as likely for those babies of mother whose alcohol consumption is 0 during pregnancy.

(c) The SE of $\hat{\beta}$ is 0.0008324 from the summary. $\hat{\beta} = 0.0010872$ Thus the Wald C.I. is the following:

```
b_hat = 0.0010872
SE = 0.0008324
ci = c(b_hat - qnorm(0.1/2, lower.tail = F)*SE, b_hat + qnorm(0.1/2, lower.tail = F)*SE)
print(ci)
```

```
## [1] -0.0002819762  0.0024563762
```

(d)

```
mydata = data.frame(drinks = c(0,0.5,1.5,4,7),
absent = c(17066, 14464, 788, 126, 37),
present = c(48, 38, 5, 1, 1) )
mydata$total = with(mydata, absent + present)
mydata$proportion = with(mydata, present/total)

mydata.linear = glm(cbind(mydata$present, mydata$absent)~drinks, family=binomial(link="logit"), data=myd
summary(mydata.linear)
```

```
##
## Call:
## glm(formula = cbind(mydata$present, mydata$absent) ~ drinks,
##     family = binomial(link = "logit"), data = mydata)
##
## Deviance Residuals:
##       1        2        3        4        5
##  0.5921  -0.8801   0.8865  -0.1449   0.1291
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.9605     0.1154 -51.637   <2e-16 ***
## drinks        0.3166     0.1254   2.523   0.0116 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6.2020  on 4  degrees of freedom
## Residual deviance: 1.9487  on 3  degrees of freedom
## AIC: 24.576
##
## Number of Fisher Scoring iterations: 4
```

i. $\hat{\pi}(x) = \frac{exp(-5.9605+0.3166x)}{1+exp(-5.9605+0.3166x)}$

ii. Coefficient: When mother's alcohol consumption is increased by 1, the odds of a baby having sex organ malformation is expected to increase by $e^{0.3166} = 1.372453$. Intercept: when the mother's alcohol consumption is 0 the expected odds of a baby having sex organ malformation is $e^{-5.9605}$.

iii. $\hat{\pi}(0) = \frac{exp(-5.9605+0.3166\times0)}{1+exp(-5.9605+0.3166\times0)} = 0.00257199$ $\hat{\pi}(7) = \frac{exp(-5.9605+0.3166\times7)}{1+exp(-5.9605+0.3166\times7)} = 0.02310568$

iv. $\hat{RR} = \frac{\hat{\pi}(0)}{\hat{\pi}(7)} = \frac{0.00257199}{0.02310568} = 0.1113142$ Interpretation: Babies of mother whose alcohol consumption is 7 during pregnancy having sex organ malformation is 0.1113142 as likely for those babies of mother whose alcohol consumption is 0 during pregnancy.

v.

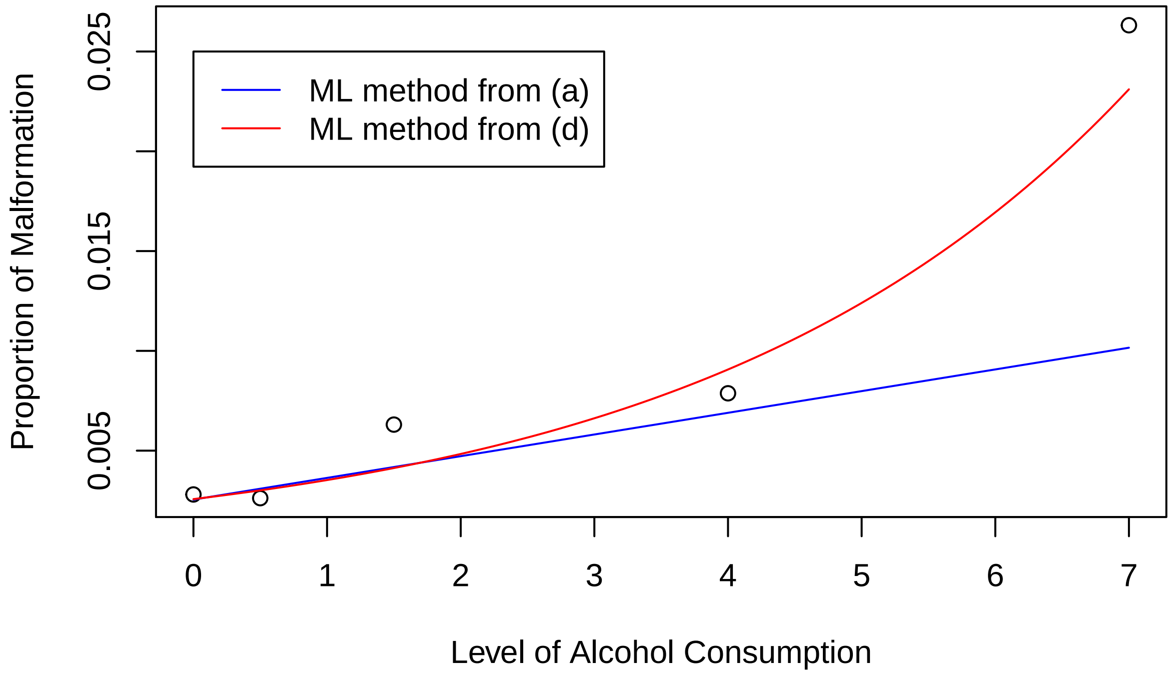
```
(0.02310568/(1-0.02310568))/(0.00257199/(1-0.00257199))
```

```
## [1] 9.17241
```

The odds ratio of malformation of alcohol level 7 vs 0 is 9.17241.

(e)

```
plot(mydata$drinks, mydata$proportion,
     ylab = "Proportion of Malformation",
     xlab = "Level of Alcohol Consumption")
curve(0.002548 + 0.001087*x, add=T, col = "blue")
curve(exp(-5.9605 + 0.3166*x)/(1+exp(-5.9605 + 0.3166*x)), add=T, col = "red")
legend(0, 0.025, legend=c("ML method from (a)", "ML method from (d)"), col=c("blue", "red"), lty=1:1)
```



The slopes of the two straight lines are different bacause the link functions are different, the identity link fits with the linear relationship between $\pi(x)$ and x, however, the logit link function fits with the linear relationship between the log of the odd and x.