Categorical Data Analysis - Assignment 1

Hyungyeong Hong (2022KU0104)

2022/3/30

Note 1) Show your code and output
Note 2) Conducting a test includes
a) write the hypothesis, b) calculate the test statistic, c) find the p-value, and d) write the conclusion

```
setwd("/Users/hyungyeonghong/Desktop/2022_Spring/CDA_2022/Assignment_1")
rm(list = ls(all = TRUE))
```

[Question 1]

When the 2010 General Social Survey asked subjects in the US whether they would be willing to accept cuts in their standard of living to protect the environment, 486 of 1374 subjects said yes. Construct a 99% Wald confidence interval for this proportion.

- Computing the sample proportion

```
pi.hat = 486/1374
```

- Computing the standard error

```
std.error = sqrt(pi.hat*(1-pi.hat) / 1374)
```

- Lower bound for the 99% Wald confidence interval for the proportion

```
pi.hat - qnorm(1-0.005)*std.error
```

```
## [1] 0.320487
```

- Upper bound for the 99% Wald confidence interval for the proportion

```
pi.hat + qnorm(1-0.005)*std.error
```

```
## [1] 0.3869365
```

[Question 2]

From Question 1, conduct a significance test.

a) Setting the hypothesis
- The hypothesis should be $H_0 : \pi = 0.5 \quad vs \quad H_1 : \pi \neq 0.5$

```
pi.null = 0.5
```

b) Calculating the test statistic

```
std.dev = sqrt((pi.null*(1-pi.null)/1374))
test.stat = (pi.hat - pi.null) / std.dev; test.stat
```

```
## [1] -10.84508
```

c) Finding the p-value

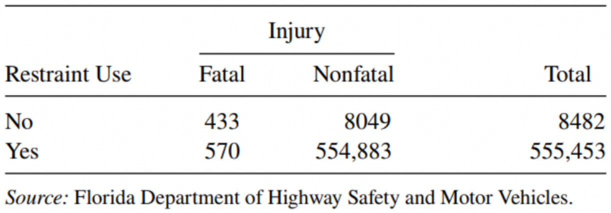
```
2*(pnorm(test.stat))
```

```
## [1] 2.104559e-27
```

d) Conclusion
- Since the p-value is less than 0.01, reject the null hypothesis $H_0 : \pi = 0.5$
- Therefore, we can conclude that π, the proportion of the population who would say 'yes', is different from 0.5.

[Questions 3-6]

Following table shows fatality results for drivers and passengers in auto accidents in Florida in 2015, according to whether the person was wearing a shoulder and lap belt restraint versus not using one. Let π_1 be the proportion of fatal injury who use restraint and π_2 be the proportion of fatal injury who does not use restraint.



	Injury		
Restraint Use	Fatal	Nonfatal	Total
No	433	8049	8482
Yes	570	554,883	555,453

Source: Florida Department of Highway Safety and Motor Vehicles.

```
rm(list = ls(all = TRUE))
```

[Question 3]

Find the (approximate) 95% Wald CI for $\pi_1 - \pi_2$
- Calculating the sample proportions and the difference between the two proportions

```
pi1.hat = 570/555453
pi2.hat = 433/8482
diff.hat = pi1.hat - pi2.hat
```

- Calculating the standard error for the difference in the two proportions

```
std.err = sqrt((pi1.hat*(1-pi1.hat)/555453)+(pi2.hat*(1-pi2.hat)/8482))
```

- Lower bound for the (approximate) 95% Wald confidence interval for the proportion

```
diff.hat - qnorm(1-0.025)*std.err
```

```
## [1] -0.05470783
```

- Upper bound for the (approximate) 95% Wald confidence interval for the proportion

```
diff.hat + qnorm(1-0.025)*std.err
```

```
## [1] -0.04533835
```

- The confidence interval does not include 0 and indicates that the group with restraints have significantly less fatal injuries.

[Question 4]

Conduct a significance test of $H_0 : \pi_1 - \pi_2 = 0$

a) Setting the hypothesis
- The hypothesis is given as $H_0 : \pi_1 - \pi_2 = 0 \quad vs \quad H_1 : \pi_1 - \pi_2 \neq 0$

```r
diff.null = 0
```

b) Calculating the test statistic
- For the significance testing, we need to compute the pooled standard error under the null hypothesis.

```r
pi.common = (570+433)/(555453+8482)
std.err = sqrt(pi.common*(1-pi.common)*(1/555453 + 1/8482))
```

- Compute Wald statistics by using the standard error calculated above.

```r
test.stat = (diff.hat - diff.null)/std.err; test.stat
```

```
## [1] -108.5124
```

c) Finding the p-value

```r
2*(pnorm(test.stat))
```

```
## [1] 0
```

d) Conclusion
- Since the p-value is much smaller than 0.05, reject the null hypothesis $H_0 : \pi_1 - \pi_2 = 0$.
- In other words, we can conclude that significant association exists between the restraint use and the cases of fatal injuries.

[Question 5]

Conduct a Pearson's Chi-squared test

a) Setting the hypothesis
- The hypothesis is given as $H_0 : \pi_1 - \pi_2 = 0 \quad vs \quad H_1 : \pi_1 - \pi_2 \neq 0$

```r
diff.null = 0
```

b) Calculating the test statistic
- The chi-square test statistic is the same as the square of the test statistic that we have computed previously.

```r
Chi.stat = test.stat^2; Chi.stat
```

```
## [1] 11774.94
```

c) Finding the p-value

```r
1-pchisq(Chi.stat, df = 1)
```

```
## [1] 0
```

d) Conclusion
- Since the p-value is much smaller than 0.05, reject the null hypothesis $H_0 : \pi_1 - \pi_2 = 0$.
- Thus, significant association exists between the restraint use and the cases of fatal injuries.

We can use `chisq.test()` function for the calculations above.

```r
Table1= matrix(c(570, 433, 554883, 8049), nrow = 2, ncol = 2)
chisq.test(Table1, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Table1
## X-squared = 11775, df = 1, p-value < 2.2e-16
```

- Since the p-value is much smaller than 0.05, reject the null hypothesis $H_0 : \pi_1 - \pi_2 = 0$.
- In other words, we can conclude that the group with restraints have significantly less fatal injuries.
- The result is the same as the result of Wald significance test with 95% confidence level.

Or, we can use `prop.test()` function.
- Here, `c(570, 433)` represents the number of fatal injuries and `c(555453, 8482)` represents the total cases.
- Since we are going to find the 95% Wald CI, the confidence level should be set to 0.95.
- We are not going to use continuity correction here by setting as `correst = FALSE`

```
prop.test(c(570, 433), c(555453, 8482), conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(570, 433) out of c(555453, 8482)
## X-squared = 11775, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.05470783 -0.04533835
## sample estimates:
##      prop 1      prop 2
## 0.001026189 0.051049281
```

- From the result above, the 95% Wald confidence interval for $\pi_1 - \pi_2$ is (0.04533835, 0.05470783), which is the same as we have computed above in Question 3.
- We can reject the null hypothesis $H_0 : \pi_1 - \pi_2 = 0$ since the p-value is < 2.2e-16, which is significantly less than 0.05.

[Question 6]

Find a relative risk and odds ratio

a) Relative Risk

```
pi1.hat = 570/555453
pi2.hat = 433/8482
RR = pi1.hat / pi2.hat; RR
```

```
## [1] 0.02010194
```

b) Odds Ratio

```
odds.pi1 = pi1.hat / (1-pi1.hat)
odds.pi2 = pi2.hat / (1-pi2.hat)
OR = odds.pi1 / odds.pi2; OR
```

```
## [1] 0.01909534
```

- Odds ratio can also be calculated by dividing multiplication of diagonal elements by multiplication of off-diagonal elements in the 2X2 contingency table.

```
(570*8049)/(433*554883)
```

```
## [1] 0.01909534
```

[Questions 7-9]

The following table shows the results of a study comparing radiation therapy with surgery in treating cancer of the larynx. The response indicates the cancer was controlled for at least two years following treatment. Let θ be the odds ratio. Significance level 0.05 is used for testing.

	Controlled	Not controlled
Surgery	21	2
Radiation therapy	8	9

```
rm(list = ls(all = TRUE))
```

[Question 7]

Perform a Fisher's exact test for testing $H_0 : \theta = 1 \quad vs \quad H_1 : \theta \neq 1$

- `fisher.test()` function can be used for Fisher's exact test.
- Since the alternative hypothesis is $H_1 : \theta \neq 1$, we use `alternative = "two.sided"`, which is the default value of the argument.

```
Table2 = matrix(c(21, 8, 2, 9), nrow = 2, ncol = 2)
fisher.test(Table2, alternative = "two.sided")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  Table2
## p-value = 0.003444
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    1.755377 126.314516
## sample estimates:
## odds ratio
##   10.98843
```

- As shown above, the null hypothesis is rejected since the p-value is much smaller than 0.05.
- Therefore, the odds ratio is not equal to 1, which means that the two variables are associated.

[Question 8]

Perform a likelihood ratio test for testing $H_0 : \theta = 1 \quad vs \quad H_1 : \theta \neq 1$

a) Setting the hypothesis

- The hypothesis is given as $H_0 : \theta = 1 \quad vs \quad H_1 : \theta \neq 1$

b) Calculating the test statistic

- The test statistic can be calculated by using the formula $2 \sum n_{ij} \log(\frac{n_{ij}}{\hat{\mu}_{ij}})$

```
LRT.stat = 2*(21*log(21*40/(29*23)) + 2*log(2*40/(11*23)) +
                8*log(8*40/(29*17)) + 9*log(9*40/(11*17))); LRT.stat
```

```
## [1] 9.955153
```

c) Finding the p-value

```
1-pchisq(LRT.stat, df = 1)
```

```
## [1] 0.001603998
```

d) Conclusion
- Since the p-value is much smaller than 0.05, reject the null hypothesis $H_0 : \theta = 1$.
- Thus, we can conclude that the odds ratio is not equal to 1, which means that the two variables are associated.

- `chisq.test()` function can also be used for likelihood ratio test.
- We first generate the table and get likelihood ratio test statistic by using the outputs generated from `chisq.test()` function.

```
Table2 = matrix(c(21, 8, 2, 9), nrow = 2, ncol = 2)
Chi.test = chisq.test(Table2)
```

```
## Warning in chisq.test(Table2): Chi-squared approximation may be incorrect
```

```
LRT.stat = 2*sum(Chi.test$observed*log(Chi.test$observed/Chi.test$expected)); LRT.stat
```

```
## [1] 9.955153
```

- Since there exists an expected frequency less than 5, the Chi-squared approximation may be poor for LRT statistic.

```
Chi.test$expected
```

```
##        [,1]  [,2]
## [1,] 16.675 6.325
## [2,] 12.325 4.675
```

- Then, the p-value can be computed from the right-side tail of Chi-square distribution.

```
1 - pchisq(LRT.stat, df = Chi.test$parameter)
```

```
## [1] 0.001603998
```

- The result is the same as we have computed above without using `chisq.test()` function.

[Question 9]

Find the 95% Wald CI for θ
- First of all, we compute sample odds ratio.

```
pi1.hat = 21/23
pi2.hat = 8/17

odds.pi1 = pi1.hat/(1-pi1.hat)
odds.pi2 = pi2.hat/(1-pi2.hat)

OR = odds.pi1 / odds.pi2
```

- Then we construct a 95% Wald CI for log OR.

```
log.OR = log(OR)
std.err = sqrt(1/21 + 1/2 + 1/8 + 1/9)
lower.bound = log.OR - qnorm(1-0.025)*std.err
upper.bound = log.OR + qnorm(1-0.025)*std.err
```

- 95% Wald CI for OR can be computed by taking exponential on the lower bound and upper bound of 95% Wald CI for log OR.
- The lower bound of 95% Wald CI for OR is,

```
exp(lower.bound)
```

```
## [1] 2.083462
```

- The upper bound of 95% Wald CI for OR is,

```
exp(upper.bound)
```

```
## [1] 66.97274
```