SURV686-HW2

Sagnik Chakravarty

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination.

Signature:

Date: 02/02/2025

Table of contents

Question 1

A researcher who wants to study the predictors of lung cancer draws a sample of persons with lung cancer from a database of lung cancer patients and a sample of persons without lung cancer from the general population. Is this a prospective or retrospective study?

Solution

This is a **retrospective study** because the researcher is selecting individuals based on their current lung cancer status (cases and controls) and then looking back at their past exposures or risk factors to determine predictors of lung cancer.

In contrast, a **prospective study** would involve selecting individuals based on exposure status (e.g., smokers and non-smokers) and then following them over time to see who develops lung cancer.

Question 2

A researcher who wants to study the impact of Vitamin D on children's health draws a sample of children, randomly splits them into two groups and gives one group Vitamin D and the other group a placebo. Is this a prospective or retrospective study?

Solution

This is a **prospective study** because the researcher is assigning an exposure (Vitamin D or placebo) to participants and then following them over time to observe the impact on their health.

Question 3

The following data are based on a study (Petrovčič, et al, 2016) that varied the content of an email message asking persons to participate in a survey. One group received a message that included a "plea for help." The other group received a message that did NOT include a plea for help. Here are the results:

data:

Message includes a plea for help	Respond To Survey	
	Yes	**No**
Yes	117	1131
No	94	1158

Question 3a

Estimate the relative risk of responding (plea for help vs not), and report a 95% confidence interval for log-relative risk

Solution

```r
risk_with_plea <- 117/(117 + 1131)
risk_without_plea <- 94/(94 + 1158)
risk_factor <- risk_with_plea/risk_without_plea
log_risk <- log(risk_factor)
se <- sqrt((1-117/(117+1131))/117 + (1 - 94/(94+1158))/94)
u_bound <- log_risk + se*1.96
l_bound <- log_risk - se*1.96

cat("Risk Factor:\t", risk_factor,
    "\nLog Risk:\t", log_risk,
    "\nStandard Error:\t", se,
    "\nConf Interval:\t[", l_bound, ',', u_bound, ']')
```

```
Risk Factor:     1.24867
Log Risk:    0.2220792
Standard Error:  0.1326096
Conf Interval:  [ -0.03783563 , 0.4819939 ]
```

Calculation

1. $$\text{Risk Factor (RR):}\ \frac{\text{Risk with Plea}}{\text{Risk without Plea}} = \frac{\frac{117}{117+1131}}{\frac{94}{94+1158}} = \frac{0.09375}{0.07507987} = 1.25$$

2. Log Risk Factor: $log(RR) = ln(1.25) = 0.22$

3. Standard Error(SE): $SE = \sqrt{\frac{1-\frac{117}{117+1131}}{117} + \frac{1-\frac{94}{94+1158}}{94}} = 0.132$
4. Confidence Interval: $log(RR) \pm 1.96 \times se = (-0.038, 0.48)$

Question 3b

Estimate the odds ratio (plea for help vs not), and report a 95% confidence interval for the log- odds ratio.

Solution

```
odds_ratio <- (117*1158)/(94*1131)
log_odds_ratio <- log(odds_ratio)
se <- sqrt(1/117 + 1/1158 + 1/94 + 1/1131)
u_bound <- log_odds_ratio + se*1.96
l_bound <- log_odds_ratio - se*1.96

cat("Odds Ratio:\t", odds_ratio,
    "\nLog Odds Ratio:\t", log_odds_ratio,
    "\nStandard Error:\t", se,
    "\nConf Interval:\t[", l_bound, ',', u_bound, ']')
```

```
Odds Ratio:  1.274395
Log Odds Ratio:  0.2424713
Standard Error:  0.1446825
Conf Interval:  [ -0.04110645 , 0.5260491 ]
```

Calculation

1. Odds Ratio: $\hat{\theta} = \frac{n_{11}n_{22}}{n_{21}n_{12}} = \frac{117\times1158}{94\times1131} = 1.27$
2. Logs Odds Ratio: $log(\hat{\theta}) = ln(1.27) = 0.24$
3. Standard Error: $\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} + \frac{1}{n_{12}}} = \sqrt{\frac{1}{117} + \frac{1}{1131} + \frac{1}{94} + \frac{1}{1158}} = 0.145$
4. Confidence Interval: $log(\hat{\theta}) \pm 1.96 \times se = 0.24 \pm 1.96 \times 0.145 = (-0.041, 0.526)$

Question 3c

Summarize and interpret your findings from parts a) and b). Does the “plea for help” improve response rates?

Solution

1. Relative Risk (RR) Analysis

- **Estimated RR: 1.25**
- **Log RR: 0.22**
- **Standard Error (SE): 0.132**
- **95% Confidence Interval for Log RR: (-0.038, 0.48)**
- **Interpretation:** The relative risk of responding to the survey when receiving a plea for help is **1.25**, meaning the response probability is **25% higher** in the plea group than in the non-plea group. However, the confidence interval **includes 1** (since exp(-0.038) 0.96 and exp(0.48) 1.61), meaning the result is **not statistically significant** at the 5% level.

2. Odds Ratio (OR) Analysis

- **Estimated OR: 1.27**
- **Log OR: 0.24**
- **Standard Error (SE): 0.145**
- **95% Confidence Interval for Log OR: (-0.041, 0.526)**
- **Interpretation:** The **odds** of responding are **27% higher** when a plea for help is included in the message. However, since the **confidence interval includes 1** (0.96 to 1.69), the odds increase is **not statistically significant** at the 5% level.

Final Conclusion

- Both **relative risk (RR = 1.25)** and **odds ratio (OR = 1.27)** suggest a positive effect of including a plea for help in the email, but the confidence intervals **include 1**, indicating the findings are **not statistically significant**.
- **Does the "plea for help" improve response rates?**
  Potentially yes, but we lack strong statistical evidence to confirm a significant effect at the 5% level. More data may be needed to determine whether the observed increase is meaningful.

Table 3: Expectation table

	yes	no
Voucher	143.5	586.5
Donation	143.5	586.5

Question 4

Question 4a

The following table is loosely based upon a study of the impact of different types of incentives on survey response rates (Deutskens, et al., 2004). Cases were randomized to either receiver a voucher that the respondent could spend at specific online vendors, or a donation would be made on their behalf. The first question is whether vouchers produce lower or higher response rates relative to donations. Calculate the odds ratio of a voucher producing response relative to donation. Calculate the deviance (G^2).

	Respond To Survey	
	Yes	**No**
Voucher	166	564
Donation	121	609

Solution

```
col_total_yes <- 166 + 121
col_total_no <- 564 + 609
row_total_voucher <- 166 + 564
row_total_donation <- 121 + 609
grand_total <- 166 + 564 + 121 + 609

e11 <- row_total_voucher * col_total_yes / grand_total
e12 <- row_total_voucher * col_total_no / grand_total
e21 <- row_total_donation * col_total_yes / grand_total
e22 <- row_total_donation * col_total_no / grand_total

kable(data.frame(yes = c(e11, e21), no = c(e12, e22), row.names = c('Voucher', 'Donation')
      caption = 'Expectation table',
      format = 'latex')
```

```
g_square <- 2 * (166 * log(166 / e11) +
                 564 * log(564 / e12) +
                 121 * log(121 / e21) +
                 609 * log(609 / e22))
odds_ratio <- (166*609)/(121*564)
log_odds_ratio <- log(odds_ratio)
se <- sqrt(1/166 + 1/609 + 1/121 + 1/564)
u_bound <- log_odds_ratio + se*1.96
l_bound <- log_odds_ratio - se*1.96

cat("Odds Ratio:\t", odds_ratio,
    "\nLog Odds Ratio:\t", log_odds_ratio,
    "\nStandard Error:\t", se,
    "\nConf Interval:\t[", l_bound, ',', u_bound, ']',
    "\nG Square:\t", g_square)
```

```
Odds Ratio:  1.481361
Log Odds Ratio:  0.3929613
Standard Error:  0.133055
Conf Interval:  [ 0.1321734 , 0.6537491 ]
G Square:    8.811714
```

Calculation

1. Calculating odds Ratio:
   - Odds Ratio: $\hat{\theta} = \frac{n_{11}n_{22}}{n_{21}n_{12}} = \frac{166\times609}{564\times121} = 1.48$
   - Logs Odds Ratio: $log(\hat{\theta}) = ln(1.48) = 0.39$
   - Standard Error: $\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} + \frac{1}{n_{12}}} = \sqrt{\frac{1}{166} + \frac{1}{564} + \frac{1}{121} + \frac{1}{609}} = 0.13$
   - Confidence Interval: $log(\hat{\theta}) \pm 1.96 \times se = 0.39 \pm 1.96 \times 0.13 = (0.13, 0.65)$
2. Calculating G^2 : $G^2 = 2\sum O_{ij}log(\frac{O_{i,j}}{E_{i,j}})$
   - Calculating the expectation: $\frac{\text{Row Total}\times\text{Col Total}}{\text{Grand Total}}$

	Respond To Survey		**Total**
	Yes	**No**	
Voucher	166	564	730
Donation	121	609	730
Total	287	1173	1460

- - $e_{11} = \frac{730 \times 287}{1460} = 143.5$
  - $e_{12} = \frac{730 \times 1173}{1462} = 586.5$
  - $e_{21} = \frac{287 \times 730}{1462} = 143.5$
  - $e_{22} = \frac{730 \times 1173}{1462} = 586.5$
- $$G^2 = 2 \times (166ln(\frac{166}{143.5}) + 564ln(\frac{564}{586.5}) + 121ln(\frac{121}{143.5}) + 609ln(\frac{609}{586.5})) = 8.81$$

Question 4b

Next, we want to compare vouchers to a lottery. Calculate the odds ratio for a voucher to produce response relative to a lottery. Calculate the deviance (G^2)

	Respond To Survey	
	Yes	**No**
Voucher	166	564
Lottery	132	598

Solution

```
col_total_yes <- 166 + 132
col_total_no <- 564 + 598
row_total_voucher <- 166 + 564
row_total_donation <- 132 + 598
grand_total <- 166 + 564 + 132 + 598

e11 <- row_total_voucher * col_total_yes / grand_total
e12 <- row_total_voucher * col_total_no / grand_total
e21 <- row_total_donation * col_total_yes / grand_total
e22 <- row_total_donation * col_total_no / grand_total

kable(data.frame(yes = c(e11, e21), no = c(e12, e22), row.names = c('Voucher', 'Donation')
      caption = 'Expectation table',
      format = 'latex')
```

```
g_square <- 2 * (166 * log(166 / e11) +
                 564 * log(564 / e12) +
                 132 * log(132 / e21) +
                 598 * log(598 / e22))
```

Table 6: Expectation table

	yes	no
Voucher	149	581
Donation	149	581

```r
odds_ratio <- (166*598)/(132*564)
log_odds_ratio <- log(odds_ratio)
se <- sqrt(1/166 + 1/598 + 1/132 + 1/564)
u_bound1 <- log_odds_ratio + se*1.96
l_bound1 <- log_odds_ratio - se*1.96

cat("Odds Ratio:\t", odds_ratio,
    "\nLog Odds Ratio:\t", log_odds_ratio,
    "\nStandard Error:\t", se,
    "\nConf Interval:\t[", l_bound1, ',', u_bound1, ']',
    "\nG Square:\t", g_square)
```

```
Odds Ratio:  1.333387
Log Odds Ratio:  0.2877224
Standard Error:  0.1305571
Conf Interval:  [ 0.03183054 , 0.5436142 ]
G Square:    4.882633
```

Calculation

1. Calculating odds Ratio:
    - Odds Ratio: $\hat{\theta} = \frac{n_{11}n_{22}}{n_{21}n_{12}} = \frac{166\times598}{564\times132} = 1.33$
    - Logs Odds Ratio: $log(\hat{\theta}) = ln(1.33) = 0.29$
    - Standard Error: $\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} + \frac{1}{n_{12}}} = \sqrt{\frac{1}{166} + \frac{1}{564} + \frac{1}{132} + \frac{1}{598}} = 0.13$
    - Confidence Interval: $log(\hat{\theta}) \pm 1.96 \times se = 0.29 \pm 1.96 \times 0.13 = (0.03, 0.54)$
2. Calculating G^2 : $G^2 = 2\sum O_{ij} log(\frac{O_{i,j}}{E_{i,j}})$
    - Calculating the expectation: $\frac{\text{Row Total}\times\text{Col Total}}{\text{Grand Total}}$

	Respond To Survey		**Total**
	Yes	**No**	

	Respond To Survey		**Total**
Voucher	166	564	730
Lottery	132	598	730
Total	298	1162	1460

-
  - $e_{11} = \frac{730\times298}{1460} = 149$
  - $e_{12} = \frac{730\times1162}{1460} = 581$
  - $e_{21} = \frac{298\times730}{1460} = 149$
  - $e_{22} = \frac{730\times1162}{1460} = 581$
-
$$G^2 = 2 \times (166ln(\frac{166}{149}) + 564ln(\frac{564}{581}) + 132ln(\frac{132}{149}) + 598ln(\frac{598}{581})) = 4.88$$

Question 4c

Describe the results from the analysis of 4a and 4b. Does there appear to be differences in response rates across each of the type of incentive comparisons in 4a and 4b?

Solution

1. 4a: Voucher vs. Donation (Odds Ratio and Deviance)

- **Odds Ratio (OR):**
  The odds ratio of **1.48** indicates that those who received a voucher were **1.48 times more likely** to respond to the survey compared to those who received a donation. This suggests that vouchers had a positive impact on response rates compared to donations.
- **Log Odds Ratio:**
  The log odds ratio of **0.39** corresponds to a moderately positive association between vouchers and survey responses, confirming the increase in the likelihood of responding when receiving a voucher.
- **Standard Error (SE):**
  The standard error of **0.13** reflects the precision of the odds ratio estimate. The smaller the SE, the more reliable the estimate.
- **Confidence Interval for Log Odds Ratio:**
  The 95% confidence interval for the log odds ratio, **(0.13, 0.65)**, does **not** include zero, which provides evidence that the voucher has a significant effect on response rates compared to donations.

- **Deviance (G^2):**
  The deviance statistic of **8.81** suggests that the model fits the data well, as it measures the difference between the fitted and null models. A larger value of G^2 indicates better model fit. Here, the result points to a significant difference between the two groups in terms of survey response.

2. 4b: Voucher vs. Lottery (Odds Ratio and Deviance)

- **Odds Ratio (OR):**
  The odds ratio of **1.33** indicates that those who received a voucher were **1.33 times more likely** to respond to the survey compared to those who received a lottery incentive. The effect is smaller than in 4a (voucher vs. donation), but still suggests that vouchers increase the response rate relative to the lottery incentive.
- **Log Odds Ratio:**
  The log odds ratio of **0.29** similarly reflects a positive relationship between vouchers and survey responses. This suggests that a voucher still has a positive, though somewhat weaker, impact compared to the lottery incentive.
- **Standard Error (SE):**
  The standard error of **0.13** remains the same as in 4a, indicating similarly reliable estimates for both comparisons.
- **Confidence Interval for Log Odds Ratio:**
  The 95% confidence interval for the log odds ratio, **(0.03, 0.54)**, does **not** include zero, meaning the difference between vouchers and lottery incentives is statistically significant. However, the effect is weaker than the voucher vs. donation comparison, which is expected given the smaller odds ratio.
- **Deviance (G^2):**
  The deviance statistic of **4.88** is lower than in 4a, indicating a smaller difference in response rates between the two groups (voucher vs. lottery). The model still fits well, but the effect is less pronounced compared to the voucher vs. donation comparison.

Final Conclusion

- **In 4a**, the voucher significantly increases survey response rates compared to donations, with a **relatively strong odds ratio (1.48)** and a significant deviance ($G^2 = 8.81$).
- **In 4b**, the voucher still increases survey responses relative to the lottery, but the effect is **smaller (OR = 1.33)**, with a corresponding decrease in the deviance ($G^2 = 4.88$), indicating a weaker but still significant effect.

Question 4d

Returning to the data from 4a. The deviance can tell us about association, but not about the direction of that association. Calculate a 95% confidence interval for the odds ratio calculated in 4a. Based on the odds ratio, which form of the incentive has the higher response rate? Is this difference significant?

Solution

```r
cat('The exponentian confidence interval is:\t(', exp(l_bound), ',', exp(u_bound),')')
```

```
The exponentian confidence interval is: ( 1.141306 , 1.922736 )
```

Calculation

new confidence interval: (exp(lower bound), exp(upper bound)) = (1.14, 1.92) **Interpretation**

- **Odds Ratio Interpretation:**
  The odds ratio of **1.48** indicates that individuals who received a voucher are **1.48 times more likely** to respond to the survey than individuals who received a donation. Since the confidence interval does not include **1** (the null value), this suggests that the difference is statistically significant.
- **Significance of the Difference:**
  Given that the confidence interval for the odds ratio is **(1.14, 1.92)** and does **not** include 1, we can confidently conclude that the difference in response rates between the voucher and donation groups is **statistically significant**.

Conclusion

- The group receiving **vouchers** has a higher response rate than the group receiving **donations**.
- The difference is **statistically significant**, as evidenced by the confidence interval for the odds ratio (1.14, 1.92) not including 1.