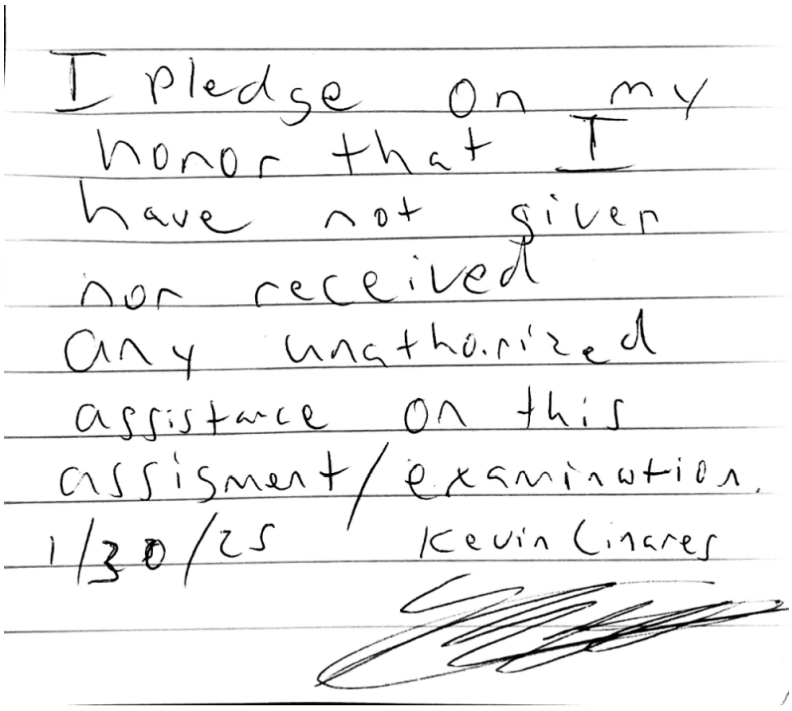
SURV616, Homework 2

Kevin Linares

2025-01-30



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?

    1. Given that the researcher has access to a database of individuals whom have lung cancer, it appears that they are sampling on the outcome of interest (lunch cancer) which suggest that this is a retrospective study. It is so because the event has already occurred and the researcher is investigating potential risk factors for this disease by using existing data.

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?

    1. This is an example of a prospective study where random sampling and random selection is the basis of the study in order to study the effects of an intervention on an outcome of interest that is to be measured at a later point in time. This study likely follows these two child cohorts (control vs treated) to investigate outcome results overtime.

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:

```
matrix_ct <- matrix(
  c(a = 117, b = 1131, c = 94, d = 1158),
  ncol = 2,
  byrow = TRUE
)

matrix_ct_margins <- addmargins(matrix_ct)

dimnames(matrix_ct_margins) <- list(
  plea_help = c("Yes", "No", "colsum"),
  responded = c("Yes", "No", "rowsum")
)

matrix_ct_margins
```

```
         responded
plea_help Yes   No rowsum
   Yes    117 1131   1248
   No      94 1158   1252
   colsum 211 2289   2500
```

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

$$RR = \ln(\frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}) = \frac{\pi_{1|1}}{\pi_{1|2}} = \frac{Pr(R|Y)}{Pr(R|N)}$$

$$\hat{V}\{\ln(\frac{n_{11}/n_{1+}}{n_{21}/n_{2+}})\} = \frac{1-\frac{n_{11}}{n1+}}{n_{11}} + \frac{1-\frac{n_{21}}{n2+}}{n_{21}} = \frac{Pr(\bar{R}|Y)}{n_{11}} + \frac{Pr(\bar{R}|N)}{n_{21}}$$

$$CI_{RR} = 1.96 \times \sqrt{V\{\ln(\hat{\theta})\}}$$

```
# calculate RR w/ formula above
RR <- (matrix_ct_margins[1,1] / matrix_ct_margins[1,3])  /
  (matrix_ct_margins[2,1] / matrix_ct_margins[2,3])

# calculate variance w/ above formula
var_RR <- ((matrix_ct_margins[1,2]/matrix_ct_margins[1,3]) /
               matrix_ct_margins[1, 1]) +
  ((matrix_ct_margins[2,2]/matrix_ct_margins[2,3]) /
     matrix_ct_margins[2, 1])


# sqrt variance & calculate 95% CI
lower <- exp(log(RR) - 1.96 * sqrt(var_RR))
upper <- exp(log(RR) + 1.96 * sqrt(var_RR))
```

- We estimate the relative risk for responding to the survey for the two conditions to be 1.25 with a 95% confidence interval between [0.96,1.62].

- We can check our work by using the epiR package as the example in the lecture.

```
epi_object <- epiR::epi.2by2(matrix_ct)
epi_object$massoc.summary[1,]
```

```
             var     est     lower    upper
1 Inc risk ratio 1.24867 0.9628758 1.619292
```

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

$$Odds_{s_1} = \frac{\pi_{1|1}}{\pi_{2|1}} = \frac{\pi_{11}}{\pi_{12}} = \frac{Pr(R|Y)}{Pr(\bar{R}|Y)} = \frac{Pr(R|Y)}{1 - Pr(R|Y)}$$

$$\hat{\theta} = \frac{\frac{\pi_{1|1}}{\pi_{2|1}}}{\frac{\pi_{1|2}}{\pi_{2|2}}} = \frac{n_{11}n_{22}}{n_{21}n_{1}2}$$

$$\hat{V}\{\ln(\hat{\theta})\} = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

```
# calculate odds ratio based on formula above = odds/odds
OR <- (matrix_ct[1,1] * matrix_ct[2,2]) / (matrix_ct[2,1] * matrix_ct[1,2])

# calculate variance for odds ratio
var_OR <- (1/matrix_ct[1, 1]) + (1/matrix_ct[1, 2]) + (1/matrix_ct[2, 1]) + (1/matrix_ct[2, 2

# calculate 95% CI
lower <- exp(log(OR) - 1.96 * sqrt(var_OR))
upper <- exp(log(OR) + 1.96 * sqrt(var_OR))
```

- We estimate the odds ratio between the odds of responding to the survey when given and not given the treatment to be 1.27 with a 95% confidence interval between [0.96,1.69].

- c) Summarize and interpret your findings from parts a) and b). Does the "plea for help" improve response rates?
- Our odds ratio suggest that respondents who receiving a plea for help were 1.27 times more likely to respond to the survey than those that did not receive the plea. The relative risk of 1.25 has a similar interpretation; however, both 95% confidence intervals included 1 (i.e., there is no difference in the risk [RR], or the odds of the event are the same[OR]). Responding to a survey is not more likely to occur if a plea of help message is given or not given.

4.

a) 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).

- The likelihood ratio statistic G^2 is a statistical method used to compare the goodness-of-fit of two competing statistical models. It approximates the χ^2 test statistic, and is expressed as:

$$G^2 = 2\sum_{i=1}^{k} O_i ln(\frac{O_i}{E_i})$$

```
# replicated table
matrix_ct <- matrix(
  c(a = 166, b = 564, c = 121, d = 609),
  ncol = 2,
  byrow = TRUE
)

dimnames(matrix_ct) <- list(
  treatment = c("Voucher", "Donation"),
  responded = c("Yes", "No")
)

matrix_ct
```

```
          responded
treatment  Yes  No
  Voucher  166 564
  Donation 121 609
```

```
# calculate OR
OR <- (matrix_ct[1,1] * matrix_ct[2,2]) / (matrix_ct[2,1] * matrix_ct[1,2])

# compute varaince and confidence intervals

# calculate variance for odds ratio
var_OR <- (1/matrix_ct[1, 1]) + (1/matrix_ct[1, 2]) +
```

```
  (1/matrix_ct[2, 1]) + (1/matrix_ct[2, 2])

# calculate 95% CI
lower <- exp(log(OR) - 1.96 * sqrt(var_OR))
upper <- exp(log(OR) + 1.96 * sqrt(var_OR))


# we can calculate the expected frequencies using
## (row total * column total) / N
### we can use the chi.square test to extract expected counts
matrix_ct_exp <- chisq.test(matrix_ct)$expected

cat("Printing expected frequencies . . .")
```

```
Printing expected frequencies . . .
```

```
matrix_ct_exp
```

```
          responded
treatment    Yes    No
  Voucher  143.5 586.5
  Donation 143.5 586.5
```

```
# calculate LRT using formula above
LRT <- sum((2*matrix_ct)*log(matrix_ct/ (matrix_ct_exp)))

# confirm G estimate w/ r function from DescTools package
#DescTools::GTest(matrix_ct)[1]

# we can calculate significant using the chi-squred distribution
df <- (1-nrow(matrix_ct)) * (1-ncol(matrix_ct))
LRT_sig <- dchisq(LRT, df)
```

- We calculate the odds ratio using the same formula in 3.b. and determine that respondents that received a voucher were 1.48 times more likely to respond to the survey than those whom were provided a donation on their behalf.

b) 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 (G2).

```r
# replicated table
matrix_ct_2 <- matrix(
  c(a = 166, b = 564, c = 132, d = 598),
  ncol = 2,
  byrow = TRUE
)


dimnames(matrix_ct_2) <- list(
  treatment = c("Voucher", "Lottery"),
  responded = c("Yes", "No")
)

matrix_ct_2
```

```
         responded
treatment Yes  No
  Voucher 166 564
  Lottery 132 598
```

```r
# calculate OR
OR_2 <- (matrix_ct_2[1,1] * matrix_ct_2[2,2]) /
  (matrix_ct_2[2,1] * matrix_ct_2[1,2])

# we can calculate the expected frequencies using
## (row total * column total) / N
### we can use the chi.square test to extract expected counts
matrix_ct_exp_2 <- chisq.test(matrix_ct_2)$expected

message("Printing expected frequencies . . .")
matrix_ct_exp_2
```

```
         responded
treatment Yes  No
  Voucher 149 581
  Lottery 149 581
```

```r
# calculate LRT using formula above
LRT_2 <- sum((2*matrix_ct_2)*log(matrix_ct_2 / (matrix_ct_exp_2)))

# we can calculate significant using the chi-squred distribution
```

```
df_2 <- (1-nrow(matrix_ct_2)) * (1-ncol(matrix_ct_2))
LRT_sig_2 <- dchisq(LRT_2, df_2)
```

- We calculate the odds ratio using the same formula in 3.b. and determine that respondents that received a voucher were 1.33 times more likely to respond to the survey than those whom were submitted for a lottery on their behalf.

c) 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?

- We calculated the $G^2 = 8.81$ goodness of fit statistics and we reject the Null hypothesis p = 0.002; therefore the distribution of responses are different from the heuristic theoretical expected frequencies. We also calculated the $G^2 = 4.88$ goodness of fit statistics for the 4.b. data and again we reject the Null hypothesis p = 0.016; therefore the distribution of responses are different from the heuristic theoretical expected frequencies.

d) 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?

- We calculate the odds ratio using the same formula in 3.b. and determine that respondents that received a voucher were 1.48 times more likely to respond to the survey than those whom were provided a donation on their behalf. We further computed 95% confidence interval in 4.a. to be between 1.14 and 1.92 which does not include 1, and we determine that there is a statistically significant different in survey responses between the two conditions.