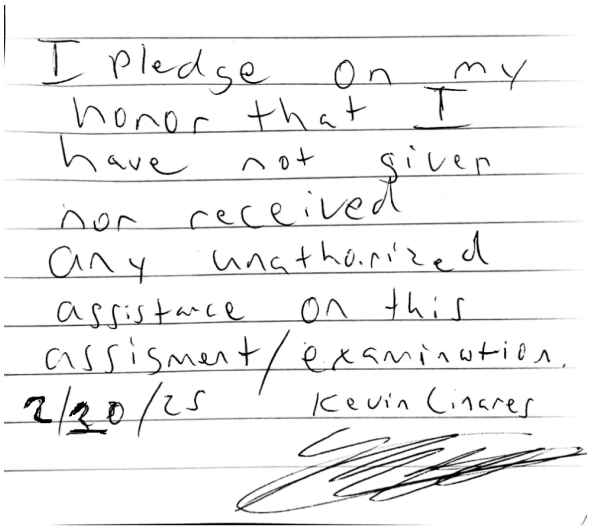
SURV616, Homework 3

Kevin Linares

2025-02-07



1. The following data come from a case-control study. The cases were sampled from a registry of all lung cancer patients at a set of 6 clinics. The controls were sampled from the patients at the 6 clinics who did not have lung cancer. Each group was asked if they had ever been regular smokers. The researchers made the following claims (1a-1f) based upon these data. State whether the claim is TRUE or FALSE and explain your answer. In this case, the population of interest is those persons who visited the 6 clinics over a specified time period.

```
a <- 126
b <- 100
c <- 35
d <- 61

# replicated table
matrix_ct <- matrix(
```

```
  c(a, b, c, d),
  ncol = 2, byrow = TRUE)

dimnames(matrix_ct) <- list(
  Smoker = c("Yes", "No"),
  "Lung Cancer" = c("Yes", "No"))

matrix_ct
```

```
      Lung Cancer
Smoker Yes  No
   Yes 126 100
   No   35  61
```

a) The proportion with cancer in the population is estimated by (126+35)/(126+35+100+61)=0.5.

- False, the proportion of the disease is the overall estimate of the disease in the sample from these clinics, and not the population. This estimate is not an un-biased estimate of the population since this is from a retrospective design. However, the proportion of disease calculation for the sample is correct. $\hat{p}_{disease} = \frac{a+c}{t}$.

b) The proportion of the population that smokes is estimated by (126+100)/ (126+35+100+61)=0.702.

- False, the proportion is the overall estimate of the exposure in the sample from these clinics, and not the population. This estimate is not an un-biased estimate of the population since this is from a retrospective design. However, the proportion of exposure calculation for the sample is correct. $\hat{p}_{exposure} = \frac{a+b}{t}$.

c) The probability of having lung cancer among Smokers is estimated by 126/226=0.558.

- True, the probability of having the disease given the exposure, also known as the incidence proportion of the exposed, is given as $i_{exposed} = \frac{a}{a+b} = .56$

d) The probability of having lung cancer among Non-Smokers is estimated by 35/96=0.365.

- True, the probability of not having the disease given the exposure, also known as the incidence proportion of the unexposed, is given as $i_{unexposed} = \frac{c}{c+d} = .37$

e) The relative risk of having lung cancer, Smokers relative to non-Smokers is 0.558/0.365=1.529.

- False, the relative risk in a retrospective study cannot be calculated correctly, but instead using the odds ratio while making assumptions about the data.

f) The odds ratio of having lung cancer for smokers relative to non-smokers is (126*61)/(*35100)=2.196.

- True, the odds ratio is given as $\hat{o} = \frac{ad}{bc}$, which corresponds with the given calculation of 2.20.

g) Now you must find the 95% CI for the odds ratio from these data.The odds ratio of having lung cancer for smokers relative to non-smokers is (126*61)/(35*100)=2.196.

- The variance for the odds ratio on the natural log is given as $\hat{v}\{\ln(\hat{o})\} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$ while the 95% confidence interval is given as $\ln \hat{o} \pm 1.96 \times \sqrt{\hat{v}\{\ln(\hat{o})\}}$.

```
# compute OR
OR <- (a*d)/(b*c)
print(str_c("Printing Odds Ratio . . . ", OR))
```

```
[1] "Printing Odds Ratio . . . 2.196"
```

```
# 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))
print(str_c("Printing 95% CI . . . [", lower, ", ", upper, "]"))
```

```
[1] "Printing 95% CI . . . [1.34321586157116, 3.59020179702107]"
```

- Smokers are 2.2 [1.3, 3.6] times more likely than non-smokers to develop lung cancer.

2. The following data come from a retrospective study of the association between smoking and bladder cancer.

```r
# replicated table
a <- 250
b <- 99750
c <- 125
d <- 199875

matrix_ct <- matrix(
  c(a, b, c, d),
  ncol = 2,
  byrow = TRUE
)

dimnames(matrix_ct) <- list(
  Smoker = c("Yes", "No"),
  "Bladder Cancer" = c("Yes", "No")
)

matrix_ct
```

```
      Bladder Cancer
Smoker Yes     No
   Yes 250  99750
   No  125 199875
```

a) Given that we cannot estimate the relative risk from these data, what assumption do we need to make in order to estimate the attributable risk from these data?

- In a retrospective design study we cannot easily estimate an unbiased relative risk, particularly in small samples studies. One assumption we can make is that the disease is rare and we can estimate the relative risk using the odds ratio. However, even with this rare disease assumption the odds ratio is not exactly equal to the relative risk, but rather an approximation.

b) Please estimate the attributable risk for the population of having bladder cancer due to smoking. What is a 95% confidence interval around the estimated attributable risk for the population?

- The attributable risk in the population cannot be estimated without the assumption of rare disease. Here we assume that the incident rate for the disease is small and then we

can use the odds ratio as an estimate of the relative risk. The attributable risk in the population is therefore expressed as:

$$\hat{A_{pop}} = \frac{ad - bc}{d(a + c)}$$

```
A_pop <- (a*d - b*c) / (d *(a+c)) ; A_pop
```

```
[1] 0.5003127
```

- We can proceed to calculate the variance of $\hat{A_{pop}}$ as:

$$V(\ln(1 - \hat{A}_{pop}) = \left[\frac{a}{c(a + c)} + \frac{b}{d(b + d)}\right]$$

```
V <-  ( (a / (c*(a+c))) + (b/(d*(b+d))) ) ; V
```

```
[1] 0.005334999
```

- Which we can than use to calculate our 95% confidence intervals as:

$$95\%\text{ C.I.} = 1 - \exp\left(\ln(1 - \hat{A}_{pop}) \pm 1.96\sqrt{V\left(\ln(1 - \hat{A}_{pop})\right)}\right)$$

```
LCL <-  1-exp( log(1-A_pop) + 1.96 * sqrt(V))  ; LCL
```

```
[1] 0.4234033
```

```
UCL <-  1-exp( log(1-A_pop) - 1.96 * sqrt(V))  ; UCL
```

```
[1] 0.5669635
```

- The attributable risk in the population in this retrospective study is an estimated reduction in incidence of 50% if the whole population were unexposed, comparing with actual exposure. Our 95% CI is between 42% and 57%.

3. The following data come from a fictional prospective study of the association between baldness and heart disease. The sample was randomly selected from the population and then followed to see if they developed baldness and/or heart disease.

```
# replicated table
a <- 127
b <- 1224
c <- 548
d <- 7611

t <- sum(a, b, c, d)

matrix_ct <- matrix(
  c(a, b, c, d),
  ncol = 2,
  byrow = TRUE
)

dimnames(matrix_ct) <- list(
  Baldness = c("Yes", "No"),
  "Heart Disease" = c("Yes", "No")
)

matrix_ct
```

```
        Heart Disease
Baldness Yes   No
     Yes 127 1224
     No  548 7611
```

a) Please graph the proportion that has heart disease in each group (i.e. bald and not).

```
# calculate proportion using incidence for exposed & unexposed
Baldness <- a / (a+b)
no_Baldness <- c / (c+d)

# combine proportions and plot
as.data.frame(rbind(Baldness, no_Baldness)) |>
  rownames_to_column() |>
  ggplot(aes(x=rowname, y=V1)) +
  geom_col(fill="dodgerblue", alpha=.7) +
  ggthemes::theme_hc() +
```