SURV686-HW3

Sagnik Chakravarty

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

Signature:

Date: 02/08/2025

Table of contents

Question 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 youranswer. In this case, the population of interest is those persons who visited the 6 clinics over a specified time period

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

Question 1.a

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

Code

```r
total_population <- 126+100+35+61
cancer <- 126+35
cat("Proportion of Population with lung cancer:\t",
    cancer/total_population)
```

```
Proportion of Population with lung cancer:   0.5
```

Calculation

$$\begin{aligned}\text{Proportion of Population with cancer} &= \frac{\text{Total With Cancer}}{\text{Total Population}} \\ &= \frac{126+35}{126+35+61+100} = \frac{161}{322} = 0.5\end{aligned}$$

But since this is a retrospective study sample proportion is not an unbiased estimate of the population proportion hence **False**

Question 1.b

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

Code

```
smoker <- 126+100
cat("Proportion of Population who smokes:\t",
    smoker/total_population)
```

```
Proportion of Population who smokes:     0.7018634
```

Calculation

$$\begin{aligned}\text{Proportion of Population who smokes} =& \frac{\text{Smoker}}{\text{Total Population}}\\ =& \frac{126+100}{126+35+61+100} = \frac{226}{322} = 0.702\end{aligned}$$

But since this is a retrospective study sample proportion is not an unbiased estimate of the population proportion hence **False**

Question 1.c

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

Code

```
cat('Probaility of a smoker with lung cancer:\t', 126/226)
```

```
Probaility of a smoker with lung cancer:     0.5575221
```

Calculation

$$\begin{aligned} P(\text{lung cancer}|\text{smoker}) =& \frac{n(\text{lung cancer } \cap \text{ smoker})}{n(\text{smoker})} \\ =& \frac{126}{126+100} = \frac{126}{226} = 0.558 \end{aligned}$$

False since not an unbiased estimator

Question 1.d

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

Code

```
cat('Probaility of a non smoker with lung cancer:\t', 35/96)
```

```
Probaility of a non smoker with lung cancer:     0.3645833
```

Calculation

$$\begin{aligned} P(\text{lung cancer}|\text{ non smoker}) =& \frac{n(\text{lung cancer } \cap \text{ non smoker})}{n(\text{non smoker})} \\ =& \frac{35}{35+61} = \frac{35}{96} = 0.365 \end{aligned}$$

False since not an unbiased estimator

Question 1.e

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

Code

```
cat('Relative Risk:\t', 0.558/0.365)
```

```
Relative Risk:   1.528767
```

Calculation

$$
\begin{aligned}
P(Exposed) &= P(\text{lung cancer}|\text{smoker}) = 0.5588\\
P(UnExposed) &= P(\text{lung cancer}|\text{non smoker}) = 0.365\\
\text{Relative Risk} &= \frac{P(Exposed)}{P(UnExposed)} = \frac{0.558}{0.365} = 1.529
\end{aligned}
$$

We can't measure the relative risk since we can't calculate the incident as these are sample data and we cannot measure the population using this hence **False**

Question 1.f

he odds ratio of having lung cancer for smokers relative to non-smokers is (126*61)/(35*100)=2.196.

Code

```
cat("Odds Ratio:\t", (126*61)/(100*35))
```

```
Odds Ratio:  2.196
```

Calculation

$$
\begin{aligned}
\text{Odds Ratio} =&\frac{A \times D}{B \times C}\\
=&\frac{126 \times 61}{100 \times 35}\\
=&\frac{7686}{3500} = 2.196
\end{aligned}
$$

Hence **True**

Question 1.g

Now you must find the 95% CI for the odds ratio from these data.

Code

```
or <- (126*61)/(100*35)
se_or <- sqrt(1/126 + 1/35 + 1/61 + 1/100)
z_score <- qnorm(0.975)
up_bound <- log(or) + z_score*se_or
low_bound <- log(or) - z_score*se_or
cat('The 95% CI for log of Odds ratio are:\t(', low_bound, ',', up_bound, ')\n')
```

```
The 95% CI for log of Odds ratio are:   ( 0.2950757 , 1.278199 )
```

```
cat('The 95% CI for Odds ratio are:\t(', exp(low_bound), ',', exp(up_bound), ')')
```

```
The 95% CI for Odds ratio are:  ( 1.343228 , 3.590169 )
```

Calculation

$$SE = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{126}+\frac{1}{100}+\frac{1}{35}+\frac{1}{61}} = 0.251$$

$$CI = ln(OR) \pm Z_{1-\alpha/2}SE = ln(2.196) \pm 1.96 \times 0.251 = (0.295, 1.278)$$

$$\text{Hence the actual CI} = (e^{0.295}, e^{1.278}) = (1.343, 3.590)$$

Question 2

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

	Bladder Cancer	
Smoker	**Yes**	**No**
Yes	250	99,750
No	125	199,875

Question 2.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?

Solution

To estimate attributable risk (AR) from these data, we must assume that:

The sample is representative of the underlying population at risk.

This means that the proportion of smokers and non-smokers in the control group reflects their true distribution in the source population. In other words, the exposure distribution (smoking) among the controls should approximate the exposure distribution in the general population from which the cases arose.

Why Is This Assumption Important?

- Since case-control studies do not follow a cohort over time, the total population at risk (i.e., the denominator needed for calculating absolute risk) is unknown.
- If the control group is not representative of the population at risk, the calculated attributable risk will be biased.

Question 2.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?

Code

```
a <- 250
b <- 99750
c <- 125
d <- 199875
r_hat <- (a*d)/(b*c)
a_exposed <- (r_hat-1)/r_hat
a_pop <- (a*d-b*c)/(d*(a+c))
v_ln <- a/(c*(a+c)) + b/(d*(b+d))
lcl <- 1-exp(log(1-a_pop) + 1.96*sqrt(v_ln))
ucl <- 1-exp(log(1-a_pop) - 1.96*sqrt(v_ln))
cat('The 95% CI for log of attribution risk for population:\t(', lcl, ',', ucl, ')\n')
```

```
The 95% CI for log of attribution risk for population:  ( 0.4234033 , 0.5669635 )
```

Calculation

$$\hat{R} = \frac{ad}{bc} = \frac{250 \times 199875}{125 \times 99750} = 4.01$$

$$\hat{A}_{Exposed} = \frac{\hat{R}-1}{\hat{R}} = 0.750$$

$$\hat{A}_{pop} = \frac{ad-bc}{d(a+c)} = \frac{250 \times 199875 - 99750 \times 125}{199875(250+125)} = 0.500$$

$$V(ln(1-\hat{A}_{pop})) = \frac{a}{c(a+c)} + \frac{b}{d(b+d)} = \frac{250}{125(250+125)} + \frac{99750}{199875(99750+199875)} = 0.005$$

$$CI \in 1 - exp(ln(1-\hat{A}_{pop}) \mp 1.96 \times \sqrt{V(ln(1-\hat{A}_{pop})})) = (0.423, 0.567)$$

Question 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.

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

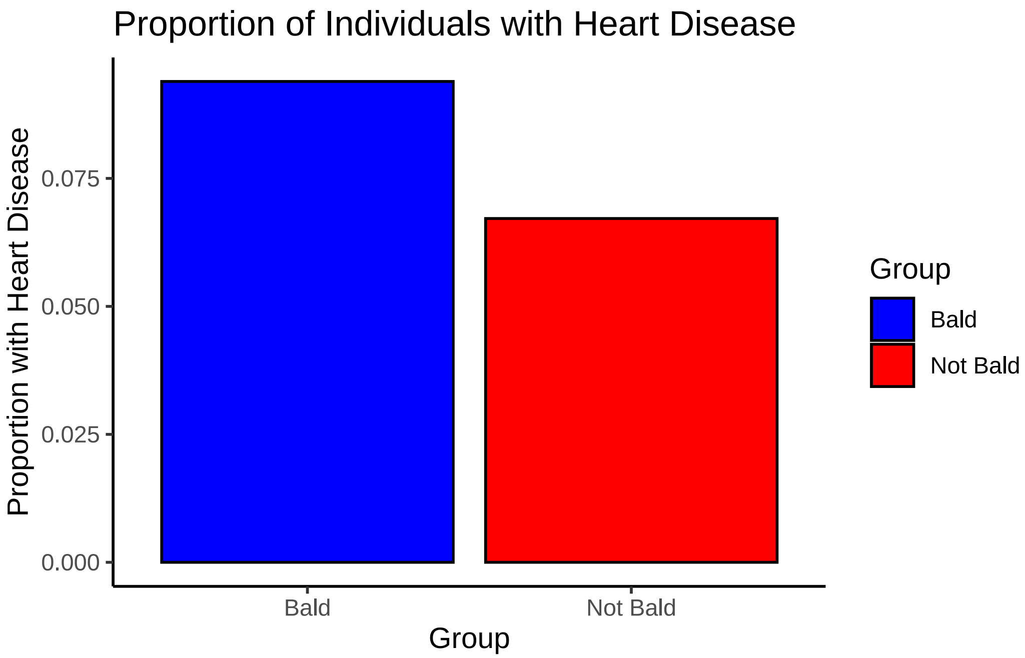
Question 3.a

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

Code

```
data <- data.frame(
  Group = c("Bald", "Not Bald"),
  Proportion = c(127/(127+1225), 548/(548+7611))
)

ggplot(data, aes(x = Group, y = Proportion, fill = Group)) +
  geom_bar(stat = "identity", color = "black") +
  labs(title = "Proportion of Individuals with Heart Disease",
       x = "Group",
       y = "Proportion with Heart Disease") +
  theme_classic() +
  scale_fill_manual(values = c("blue", "red")) # Custom colors
```



Question 3.b

Please estimate the attributable risk for the population of having heart disease due to baldness. What is a 95% confidence interval around this estimate?

Code

```
a <- 127; b <- 1224; c <- 548; d <- 7611
total <- a + b + c + d

# Relative Risk (unchanged)
rr <- (a/(a+b)) / (c/(c+d))

# Attributable Risk Population
a_pop <- (a*d - b*c) / ((a + c)*(c + d))
```

```r
# CORRECTED Variance Calculation
v_ln <- (b + a_pop*(a + d)) / (total * c)  # Added *c in denominator

# Confidence Interval
log_term <- log(1 - a_pop)
lcl <- 1 - exp(log_term + 1.96*sqrt(v_ln))
ucl <- 1 - exp(log_term - 1.96*sqrt(v_ln))

cat('The 95% CI for log of attribution risk for population:\t(', lcl, ',', ucl, ')\n')
```

```
The 95% CI for log of attribution risk for population:  ( 0.02024151 , 0.08605154 )
```

Calculation

$$\begin{aligned}
&\hat{R} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{\frac{127}{1351}}{\frac{548}{8159}} \approx 1.399 \\
&\hat{A}_{\text{Exposed}} = \frac{\hat{R}-1}{\hat{R}} \approx 0.285 \\
&\hat{A}_{\text{pop}} = \frac{ad-bc}{(a+c)(c+d)} = \frac{127 \times 7611 - 1224 \times 548}{(675)(8159)} \approx 0.0537 \\
&V\left(\ln(1-\hat{A}_{\text{pop}})\right) = \frac{b+\hat{A}_{\text{pop}}(a+d)}{t \cdot c} = \frac{1224 + 0.0537 \times 7738}{9510 \times 548} \approx 0.000315 \\
&95\%\ \text{CI} = 1 - \exp\left(\ln(0.9463) \pm 1.96 \times \sqrt{0.000315}\right) \approx (0.020,\ 0.086)
\end{aligned}$$