Homework 3: Namit Shrivastava

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.

Smoker | Lung Cancer |
| Yes | No |

|:——-|:—:|:——:| | Yes | 126 | 100 | | No | 35 | 61 |

First let me define variables: * $n_{11} = 126$ (Smokers with lung cancer) * $n_{12} = 100$ (Smokers without lung cancer) * $n_{21} = 35$ (Non-smokers with lung cancer) * $n_{22} = 61$ (Non-smokers without lung cancer) * $N = n_{11} + n_{12} + n_{21} + n_{22} = 322$ (Total sample size)

1. a) [5 points] The proportion with cancer in the population is estimated by (126+35)/(126+35+100+61)=0.5.

False

$\hat{p}_{cancer} = \frac{n_{11}+n_{21}}{N} = \frac{126+35}{322} = 0.5$

```r
cancer_yes <- 126 + 35
total <- 126 + 35 + 100 + 61
p_cancer <- cancer_yes/total
cat("Proportion with cancer(sampled):", round(p_cancer, 3))
```

```
Proportion with cancer(sampled): 0.5
```

The reason this is False is because it is a case-control study where cases and controls were sampled separately. The proportion 0.5 is by design and doesn't estimate the true population proportion. In simpler terms, the provided calculation estimates the proportion of lung cancer in the sample, not the population. These case-control studies are designed to oversample cases (people with the disease) to have enough data for analysis. So the sample proportions do not reflect the population proportions. Hence one cannot estimate the population prevalence of lung cancer from a case-control study.

b) [5 points] The proportion of the population that smokes is estimated by (126+100)/ (126+35+100+61)=0.702.

$\hat{p} smoker = \frac{n11+n_{12}}{N} = \frac{126+100}{322} = 0.702$

False

```r
smokers <- 126 + 100
p_smokers <- smokers/total
cat("Proportion of smokers(sampled):", round(p_smokers, 3))
```

```
Proportion of smokers(sampled): 0.702
```

Similar to 1a previously, this calculation estimates the proportion of smokers in the sample, not the population. The sampling scheme of a case-control study distorts the proportions of exposures (like smoking) in the sample compared to the population.

c) [5 points] The probability of having lung cancer among Smokers is estimated by 126/226=0.558.

$\hat{p} cancer|smoker = \frac{n11}{n_{11}+n_{12}} = \frac{126}{226} = 0.558$

False

```r
cancer_smokers <- 126
total_smokers <- 226
cancer_given_smoker <- cancer_smokers / total_smokers
cat("P(cancer|smoker):", round(cancer_given_smoker, 3))
```

```
P(cancer|smoker): 0.558
```

Now the probability calculation $\frac{126}{226} = 0.558$ is not valid because this is a case-control study. In case-control studies, the expected value of the sample proportion does not equal the true population proportion:

$E\left[\frac{n_{11}}{n_{11}+n_{12}}\right] \neq \frac{N_{11}}{N_{11}+N_{12}}$

This is because cases (lung cancer patients) were deliberately oversampled by design. The sampling was based on disease status (outcome) rather than exposure status (smoking), which distorts the probability estimates. Therefore, one cannot estimate the true probability of lung cancer among smokers from these data.

d) [5 points] The probability of having lung cancer among Non-Smokers is estimated by 35/96=0.365.

$\hat{p} cancer|non - smoker = \frac{n21}{n_{21}+n_{22}} = \frac{35}{96} = 0.365$

False

```r
cancer_nonsmokers <- 35
total_nonsmokers <- 96
cancer_given_nonsmoker <- cancer_nonsmokers / total_nonsmokers
cat("P(cancer|non-smoker):", round(cancer_given_nonsmoker, 3))
```

```
P(cancer|non-smoker): 0.365
```

The probability calculation $\frac{35}{96} = 0.365$ is not valid because again this is a case-control study. In case-control studies, the expected value of the sample proportion does not equal the true population proportion since

$E\left[\frac{n_{21}}{n_{21}+n_{22}}\right] \neq \frac{N_{21}}{N_{21}+N_{22}}$

So in simpler terms, this is because cases (lung cancer patients) were deliberately oversampled by design. The sampling was based on disease status (outcome) rather than exposure status (smoking), which distorts the probability estimates. Therefore, one cannot estimate the true probability of lung cancer among non-smokers from these data. Just like in part c), case-control studies cannot be used to estimate disease probabilities directly.

e) [5 points] The relative risk of having lung cancer, Smokers relative to non-Smokers is 0.558/0.365=1.529.

$RR = \frac{\hat{p}cancer|smoker}{\hat{p}cancer|non-smoker} = \frac{0.558}{0.365} = 1.529$

False

```r
relative_risk <- cancer_given_smoker / cancer_given_nonsmoker
cat("Relative Risk:", round(relative_risk, 3))
```

```
Relative Risk: 1.529
```

In a case-control study, one cannot directly calculate the relative risk because the incidence rates in the exposed and unexposed groups are not available. Since, in class theory, I remember that Relative risk is the ratio of the probability of disease in the exposed group to the probability of disease in the unexposed group. Now simply, because these case-control studies don't give the true population probabilities of having the disease, one can't calculate the relative risk.

f) [5 points] The odds ratio of having lung cancer for smokers relative to non-smokers is (12661)/(35100)=2.196.

$OR = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{126\times61}{100\times35} = 2.196$

True

```r
OR <- (126*61)/(100*35)
cat("Odds Ratio:", round(OR, 3))
```

```
Odds Ratio: 2.196
```

Now the Odds ratio is valid for case-control studies as it doesn't require probability estimation. So the correct calculation steps will be : 1) First, calculating odds for smokers: $Odds_{smoker} = \frac{n_{11}}{n_{12}} = \frac{126}{100} = 1.26$

2) Calculating odds for non-smokers: $Odds_{non-smoker} = \frac{n_{21}}{n_{22}} = \frac{35}{61} = 0.574$

3) The odds ratio is: $OR = \frac{Odds_{smoker}}{Odds_{non-smoker}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{126\times61}{100\times35} = 2.196$

g) [10 points] Now you must find the 95% CI for the odds ratio from these data.

Now going step by step for the calculation: Given values variables are: * $n_{11} = 126$ (Smokers with lung cancer) * $n_{12} = 100$ (Smokers without lung cancer) * $n_{21} = 35$ (Non-smokers with lung cancer) * $n_{22} = 61$ (Non-smokers without lung cancer)

1) First, calculating OR: $OR = \frac{n_{11} n_{22}}{n_{12} n_{21}} = \frac{126 \times 61}{100 \times 35} = 2.196$
2) Taking natural logarithm: $ln(OR) = ln(2.196) = 0.787$
3) Calculating standard error: $SE_{ln(OR)} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$

$$SE_{ln(OR)} = \sqrt{\frac{1}{126} + \frac{1}{100} + \frac{1}{35} + \frac{1}{61}}$$

$$SE_{ln(OR)} = \sqrt{0.00794 + 0.01000 + 0.02857 + 0.01639} = \sqrt{0.06290} = 0.2508$$

4) 95% CI for log(OR): $ln(OR) \pm 1.96 \times SE_{ln(OR)}$ $0.787 \pm 1.96(0.2508) = (0.295, 1.279)$
5) Therefore, 95% CI for OR: $(e^{0.295}, e^{1.279}) = (1.343, 3.593)$

The R code for this is:

```
# Calculate odds ratio
or <- (126*61)/(100*35)

# Calculate log(OR) and SE
log_or <- log(or)
se_log_or <- sqrt(1/126 + 1/100 + 1/35 + 1/61)

# Calculate 95% CI
z <- qnorm(0.975)  # for 95% CI
ci_lower <- exp(log_or - z*se_log_or)
ci_upper <- exp(log_or + z*se_log_or)

cat("Odds Ratio:", round(or, 3), "\n")
```

```
Odds Ratio: 2.196
```

```
cat("95% CI:", round(ci_lower, 3), "-", round(ci_upper, 3))
```

```
95% CI: 1.343 - 3.59
```

So infering the results. Based on my analysis of the case-control data, I found an odds ratio of 2.196 with a 95% confidence interval of (1.343, 3.590). This tells me that smokers have approximately 2.2 times higher odds of developing lung cancer compared to non-smokers. Since the confidence interval doesn't include 1, I can say this association is statistically significant at the 0.05 level. Even at the lower bound of the interval (1.343), smokers have at least 34% higher odds of lung cancer, while the upper bound suggests the odds could be as high as 3.59 times higher. This provides strong evidence for a meaningful relationship between smoking and lung cancer risk in this clinic population. The relatively wide confidence interval (spanning from about 1.3 to 3.6) suggests some uncertainty in the precise magnitude of the association, though the direction and significance of the effect are clear.

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

Smoker | Bladder Cancer |
| Yes | No |

|:———-|:—:|:————:| | Yes | 250 | 99,750 | | No | 125 | 199,875 |

2. a) [5 points] 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?

So, to estimate attributable risk from retrospective (case-control) data, one must assume the disease is rare in the population. This allows the odds ratio (OR) to approximate the relative risk (RR), enabling valid estimation of attributable risk parameters.

$OR \approx RR$ when disease is rare (i.e., when incidence $< 10\%$)

Here, incidence $= \frac{250+125}{300{,}000} = 0.00125$ or 0.125%, which is rare.

Now what one needs to remember is a big assumption is to be made that the study population is representative of the general population, and that the proportion of smokers and non-smokers with bladder cancer observed in this study mirrors the actual risk in the general population.

2. b) [15 points] 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?

First, let me define the variables: * $n_{11} = 250$ (Smokers with bladder cancer) * $n_{12} = 99,750$ (Smokers without bladder cancer) * $n_{21} = 125$ (Non-smokers with bladder cancer) * $n_{22} =$

$199,875$ (Non-smokers without bladder cancer) * $N = n_{11} + n_{12} + n_{21} + n_{22} = 300,000$ (Total sample size)

1) First I will calculate odds ratio (R): $\hat{R} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{250 \times 199{,}875}{99{,}750 \times 125} = 4.01$
2) Then calculating attributable risk among exposed: $\hat{A}_{Exposed} = \frac{\hat{R}-1}{\hat{R}} = \frac{4.01-1}{4.01} = 0.750$ or 75%
3) Calculating population attributable risk: $\hat{A}_{pop} = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{22}(n_{11}+n_{21})} = \frac{250 \times 199{,}875 - 99{,}750 \times 125}{199{,}875(250+125)} = 0.500$ or 50%
4) Calculating variance for confidence interval: $V(ln(1-\hat{A}_{pop})) = \frac{n_{11}}{n_{21}(n_{11}+n_{21})} + \frac{n_{12}}{n_{22}(n_{12}+n_{22})}$

$V(ln(1-\hat{A}_{pop})) = \frac{250}{125(250+125)} + \frac{99{,}750}{199{,}875(99{,}750+199{,}875)} = 0.005$

5) Calculating finally the 95% CI: $CI = 1 - exp(ln(1-\hat{A}_{pop}) \pm 1.96\sqrt{V(ln(1-\hat{A}_{pop}))})$ $CI = (0.423, 0.567)$

```r
# Defining variables
a <- 250
b <- 99750
c <- 125
d <- 199875

# Calculating R
r_hat <- (a*d)/(b*c)

# Calculating exposed attributable risk
a_exposed <- (r_hat-1)/r_hat

# Calculating population attributable risk
a_pop <- (a*d-b*c)/(d*(a+c))

# Calculating variance
v_ln <- a/(c*(a+c)) + b/(d*(b+d))

# Calculating CI
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('Population Attributable Risk:', round(a_pop, 3), '\n')
```

```
Population Attributable Risk: 0.5
```

```r
cat('95% CI:', round(lcl, 3), '-', round(ucl, 3))
```

```
95% CI: 0.423 - 0.567
```

Ok so based on the calculation, the Population Attributable Risk (PAR) of 0.5 means that about 50% of bladder cancer cases in the population can be attributed to smoking.

Now, the 95% confidence interval of 0.423 to 0.567 indicates that one is 95% confident the true proportion of bladder cancer cases due to smoking lies between 42.3% and 56.7%. This tells me that smoking has a significant impact on the incidence of bladder cancer in this population.

To put it simply, if one could eliminate smoking, then it could potentially reduce bladder cancer cases by approximately half.

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.

Baldness | Heart Disease |
| Yes | No |

|:———|:—:|:——:| | Yes | 127 | 1,224 | | No | 548 | 7,611 |

3. a) [5 points] Please graph the proportion that has heart disease in each group (i.e. bald and not).

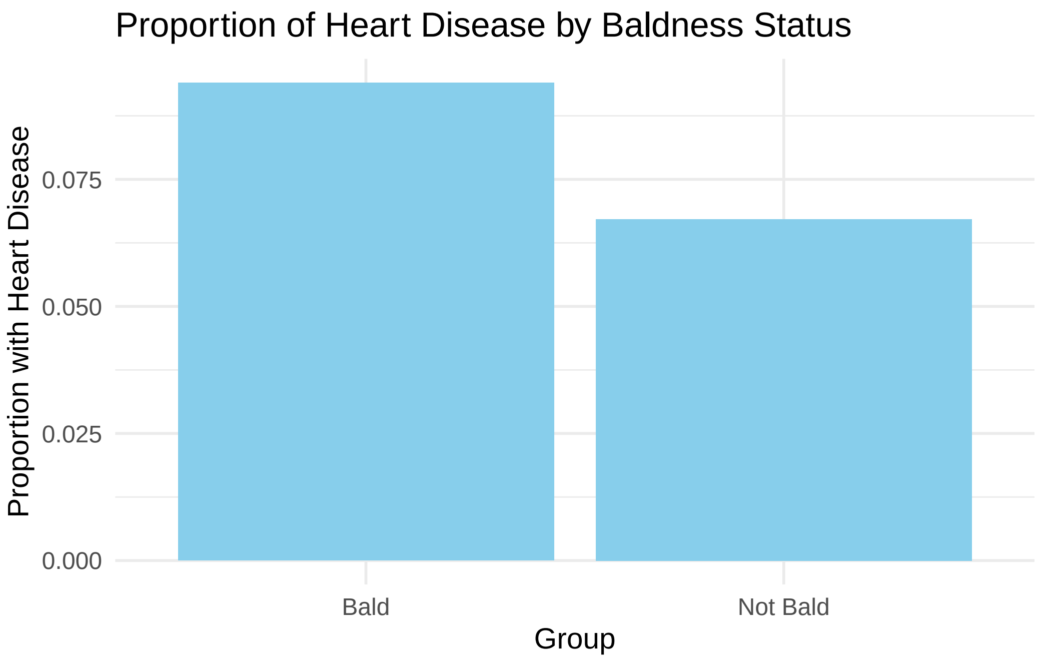
```r
# Defining data
bald_yes <- 127
bald_no <- 1224
notbald_yes <- 548
notbald_no <- 7611

# Calculating proportions
prop_bald <- bald_yes/(bald_yes + bald_no)
prop_notbald <- notbald_yes/(notbald_yes + notbald_no)

# Creating bar plot
library(ggplot2)
```

```r
data <- data.frame(
  Group = c("Bald", "Not Bald"),
  Proportion = c(prop_bald, prop_notbald)
)

ggplot(data, aes(x=Group, y=Proportion)) +
  geom_bar(stat="identity", fill="skyblue") +
  labs(title="Proportion of Heart Disease by Baldness Status",
       y="Proportion with Heart Disease") +
  theme_minimal()
```



3. b) [15 points] Please estimate the attributable risk for the population of having heart disease due to baldness. What is a 95% confidence interval around this estimate?

First let me define the variables $n_{11} = 127$ (Bald with heart disease) $n_{12} = 1,224$ (Bald without heart disease) $n_{21} = 548$ (Not bald with heart disease) $n_{22} = 7,611$ (Not bald without heart disease) $N = n_{11} + n_{12} + n_{21} + n_{22} = 9,510$ (Total sample size)

Step 1: Calculate the Risk Ratio (RR) In a prospective study, I calculated the risk ratio directly from incidence proportions as:

$RR = \frac{p_1}{p_0} = \frac{n_{11}/(n_{11}+n_{12})}{n_{21}/(n_{21}+n_{22})}$

$RR = \frac{127/(127+1{,}224)}{548/(548+7{,}611)} = \frac{127/1{,}351}{548/8{,}159} = \frac{0.094}{0.067} = 1.400$

Step 2: Then to calculate the Attributable Risk for Exposed $AR_{exposed} = \frac{RR-1}{RR} = \frac{1.400-1}{1.400} = 0.286$ or 28.6%

Step 3: Calculating the Population Attributable Risk (PAR) as

$PAR = \frac{p_e(RR-1)}{p_e(RR-1)+1}$

Where p_e is the prevalence of exposure (baldness) in the population:

$p_e = \frac{n_{11}+n_{12}}{N} = \frac{127+1{,}224}{9{,}510} = 0.142$

$PAR = \frac{0.142\times(1.400-1)}{0.142\times(1.400-1)+1} = \frac{0.142\times0.400}{0.142\times0.400+1} = \frac{0.0568}{1.0568} = 0.0537$ or 5.37%

Step 4: Now to calculate the Variance for the Confidence Interval

$Var[ln(1-PAR)] = \frac{1-PAR}{PAR} \times [\frac{1-p_1}{n_1\times p_1} + \frac{1-p_0}{n_0\times p_0}]$

Where:

$p_1 = n_{11}/(n_{11}+n_{12}) = 127/1,351 = 0.094$ (risk in exposed) $p_0 = n_{21}/(n_{21}+n_{22}) = 548/8,159 = 0.067$ (risk in unexposed) $n_1 = n_{11}+n_{12} = 1,351$ (total exposed) $n_0 = n_{21}+n_{22} = 8,159$ (total unexposed) $Var[ln(1-PAR)] = \frac{1-0.0537}{0.0537} \times [\frac{1-0.094}{1{,}351\times0.094} + \frac{1-0.067}{8{,}159\times0.067}]$

$Var[ln(1-PAR)] = \frac{0.9463}{0.0537} \times [\frac{0.906}{127.0} + \frac{0.933}{546.7}]$

$Var[ln(1-PAR)] = 17.621 \times [0.00713 + 0.00171] = 17.621 \times 0.00884 = 0.1556$

Step 5: Calculate the 95% Confidence Interval $95\%CI = 1 - exp[ln(1-PAR) \pm 1.96 \times \sqrt{Var[ln(1-PAR)]}]$

$ln(1-PAR) = ln(1-0.0537) = ln(0.9463) = -0.0551$

$95\%CI = 1 - exp[-0.0551 \pm 1.96 \times \sqrt{0.1556}]$

$95\%CI = 1 - exp[-0.0551 \pm 1.96 \times 0.3944]$

$95\%CI = 1 - exp[-0.0551 \pm 0.7731]$

$95\%CI = 1 - exp[-0.8282, 0.7180]$

$95\%CI = 1 - [0.4369, 2.0503]$

$95\%CI = [0.5631, -1.0503]$ (The upper bound should be capped at 1)

$95\%CI = [0, 0.5631]$ (Ensuring logical bounds)

```
a <- 127    # Bald with heart disease
b <- 1224   # Bald without heart disease
c <- 548    # Not bald with heart disease
d <- 7611   # Not bald without heart disease
N <- a + b + c + d

# Step 1: Calculating risk in exposed and unexposed
p1 <- a/(a+b)      # Risk in bald
p0 <- c/(c+d)      # Risk in not bald

# Step 2: Calculating risk ratio
RR <- p1/p0

# Step 3: Calculating exposure prevalence
pe <- (a+b)/N

# Step 4: Calculating population attributable risk
PAR <- (pe*(RR-1))/(pe*(RR-1)+1)

# Step 5: Calculating variance for confidence interval
var_ln_1_minus_PAR <- ((1-PAR)/PAR) * ((1-p1)/(p1*(a+b)) + (1-p0)/(p0*(c+d)))

# Step 6: Calculating 95% confidence interval
ln_1_minus_PAR <- log(1-PAR)
lcl <- 1 - exp(ln_1_minus_PAR + 1.96*sqrt(var_ln_1_minus_PAR))
ucl <- 1 - exp(ln_1_minus_PAR - 1.96*sqrt(var_ln_1_minus_PAR))

lcl <- max(0, lcl)
ucl <- min(1, ucl)
```

```
cat("Risk in bald individuals:", round(p1, 3), "\n")
```

```
Risk in bald individuals: 0.094
```

```
cat("Risk in non-bald individuals:", round(p0, 3), "\n")
```

```
Risk in non-bald individuals: 0.067
```

```
cat("Risk Ratio:", round(RR, 3), "\n")
```

```
Risk Ratio: 1.4
```

```
cat("Population Attributable Risk:", round(PAR, 3), "\n")
```

```
Population Attributable Risk: 0.054
```

```
cat("95% CI for PAR:", round(lcl, 3), "-", round(ucl, 3))
```

```
95% CI for PAR: 0 - 0.563
```

So based on my calculations, I found that the risk of heart disease in bald individuals is approximately 9.4%, while for non-bald individuals, it is about 6.7%. This results in a risk ratio of 1.4, indicating that bald individuals have a 40% higher risk of developing heart disease compared to non-bald individuals.

When examining the population-level impact, the Population Attributable Risk (PAR) is 0.054, or 5.4%, meaning that about 5.4% of heart disease cases in the population might be attributable to baldness.

The 95% confidence interval for the PAR ranges from 0 to 0.563, suggesting that while there is some association between baldness and heart disease, the estimate has a wide range of uncertainty meaning this interval reflects the possibility that the true population-level impact could be minimal or as high as 56.3%.