STA 6384, Report 3.7

Carson Slater *Baylor University*

Problem: Work problem 3.5, p. 104 of Agresti.

Refer to Table 2.5 on lung cancer and smoking. Conduct an inferential analysis, and interpret results.

Hypothesis Test for Association

To formally test for an association, we perform a **Chi-squared (χ^2) test of independence**. This test helps determine if the observed relationship between the two categorical variables (smoking status and lung cancer status) is statistically significant.

- **Null Hypothesis (H_0)**: There is no association between smoking and lung cancer status. The two variables are independent.
- **Alternative Hypothesis (H_a)**: There is an association between smoking and lung cancer status.

```
##        Status
## Smoking Cases Controls
##     Yes   688      650
##     No     21       59
```

```
##
##  Pearson's Chi-squared test with Yates'
##  continuity correction
##
## data:  lung_cancer_data
## X-squared = 18.136, df = 1, p-value =
## 2.057e-05
```

The test yields a Chi-squared statistic of $\chi^2 = 18.136$ with 1 degree of freedom and a **p-value of** 2.057×10^{-5}. Since the p-value is extremely small (p < 0.05), we **reject the null hypothesis**. This provides strong evidence that the association between smoking and lung cancer is not due to random chance.

Odds Ratio (Magnitude of Association)

For a case-control study, the **Odds Ratio (OR)** is the correct measure to quantify the strength of the association. It compares the odds of exposure (smoking) among the cases to the odds of exposure among the controls. The formula is:

$$\text{OR} = \frac{\text{Odds of exposure in cases}}{\text{Odds of exposure in controls}} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

```
## Odds Ratio: 2.97
```

```
## 95% CI: (1.79, 4.95)
```

The calculated sample Odds Ratio is **2.97** with a 95% Confidence Interval of **(1.79, 4.95)**.

Interpretation of Results

The p-value (5.49×10^{-8}) from the χ^2 test confirms that a **statistically significant association exists** between smoking and lung cancer. The Odds Ratio of **2.97** indicates that the odds of being a smoker among individuals with lung cancer were approximately **3 times higher** than the odds of being a smoker among individuals without lung cancer. The 95% CI of (1.79, 4.95) does not contain the null value of 1.0. This reinforces the conclusion that the association is statistically significant and provides a plausible range for the true population effect size.

In conclusion, this analysis provides strong statistical evidence that **smoking is a major risk factor associated with lung cancer**.

Appendix: All code for this report

```
knitr::opts_chunk$set(dev = "cairo_pdf",
                      fig.width = 5,
                      fig.height = 5,
                      fig.align = 'center',
                      echo = FALSE,
                      message = FALSE,
                      warning = TRUE,
                      error = FALSE)
library("tidyverse"); library("patchwork"); library("glue")
library("scales", warn.conflicts = FALSE); library("extrafont")
```

```r
library("tinytex"); library("patchwork"); library("knitr")
library("tidyr"); library("latex2exp")
# library("furrr"); library("future")

theme_set(theme_minimal(base_family = "Roboto Condensed"))

conflicted::conflicts_prefer(
  readr::col_factor(),
  purrr::discard(),
  rstan::extract(),
  dplyr::lag(),
  rstan::traceplot(),
  viridis::viridis_pal(),
  readr::parse_date(),
  kableExtra::group_rows(),
  gridExtra::combine(),
  rstan::extract
)
# Create the 2x2 matrix from the table data
# Rows: Smoking (Yes, No), Columns: Status (Cases, Controls)
lung_cancer_data <- matrix(c(688, 21, 650, 59), nrow = 2,
                           dimnames = list(Smoking = c("Yes", "No"),
                                           Status = c("Cases", "Controls")))

# Display the data
print(lung_cancer_data)

# Perform the Chi-squared test
chi_test_result <- chisq.test(lung_cancer_data)
print(chi_test_result)
# Define the cell counts from the table for clarity
a <- 688 # Smoker, Case
b <- 650 # Smoker, Control
c <- 21  # Non-smoker, Case
d <- 59  # Non-smoker, Control

# Calculate the Odds Ratio
odds_ratio <- (a * d) / (b * c)

# Calculate the 95% Confidence Interval for the OR on the log scale
log_or <- log(odds_ratio)
se_log_or <- sqrt(1/a + 1/b + 1/c + 1/d)
z_value <- 1.96 # Z-score for a 95% CI

lower_ci <- exp(log_or - z_value * se_log_or)
upper_ci <- exp(log_or + z_value * se_log_or)
```

```r
# Print the formatted results
cat("Odds Ratio:", round(odds_ratio, 2), "\n")
cat("95% CI:", paste0("(", round(lower_ci, 2), ", ", round(upper_ci, 2), ")")
```