GR5291 Advanced Data Analysis Problem Set 2b

Francis Zhang

September 20, 2024

Question

Consider the ToothGrowthdata in R, concerning the Effect of Vitamin C on Tooth Growth in Guinea Pigs.

Assume that if "len" is above 21, it is classified as "HIGH"; and "LOW", otherwise.

a. Ignore 'dose', and determine whether there is a significant difference in the proportions of the two groups classified as "HIGH" using a suitable test and a 95% confidence interval.
b. Repeat the above, taking into account "dose".
c. Comment on the results in terms of the assumptions you made about the tests and their validity.

Solution

a:

1.Classify "len" as HIGH or LOW:

If the tooth length (len) is above 21, classify it as "HIGH"; otherwise, classify it as "LOW".

2.Perform a suitable test for proportions:

Since we are comparing proportions between two groups (OJ and VC), we can use a Chi-square test or Fisher's exact test to determine if there is a significant difference between the proportions of "HIGH" in the two groups.

3.Calculate the 95% confidence interval:

We can use a proportion confidence interval to estimate the confidence interval for the difference in proportions.

```r
# Load the data
data("ToothGrowth")

# Classify 'len' as HIGH (above 21) or LOW (21 or below)
ToothGrowth$len_class <- ifelse(ToothGrowth$len > 21, "HIGH", "LOW")

# Create a contingency table for proportions of HIGH between OJ and VC groups
table_len_supp <- table(ToothGrowth$supp, ToothGrowth$len_class)

# Perform a Chi-square test for proportions
chi_test_result <- chisq.test(table_len_supp)

# View the result
chi_test_result
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_len_supp
## X-squared = 3.2812, df = 1, p-value = 0.07008
```

```
# Calculate 95% confidence interval for the difference in proportions
prop_table <- prop.table(table_len_supp, margin = 1)
prop_diff <- prop_table[1, "HIGH"] - prop_table[2, "HIGH"]
prop_se <- sqrt((prop_table[1, "HIGH"] * (1 - prop_table[1, "HIGH"])
                 / sum(ToothGrowth$supp == "OJ")) +
                  (prop_table[2, "HIGH"] * (1 - prop_table[2, "HIGH"])
                   / sum(ToothGrowth$supp == "VC")))
conf_interval <- prop_diff + c(-1, 1) * 1.96 * prop_se

# View the confidence interval
conf_interval
```

```
## [1] 0.02337857 0.50995476
```

b:

1.Stratify by dose:

We can create a 2x2 contingency table for each dose level (0.5, 1.0, and 2.0 mg) and then test whether the proportions of “HIGH” differ between OJ and VC within each dose group.

2.Perform the test for each dose level:

For each dose, perform a Chi-square test or Fisher’s exact test to see if there is a significant difference in proportions of “HIGH” between the two groups.

```
# Perform the Chi-square test for each dose level
for (dose_level in unique(ToothGrowth$dose)) {
  # Subset the data for the current dose
  subset_data <- subset(ToothGrowth, dose == dose_level)

  # Create a contingency table
  table_len_supp_dose <- table(subset_data$supp, subset_data$len_class)

  # Perform a Chi-square or Fisher's exact test
  if (all(table_len_supp_dose >= 5)) {
    chi_test_result_dose <- chisq.test(table_len_supp_dose)
  } else {
    chi_test_result_dose <- fisher.test(table_len_supp_dose)
  }

  # Print the results for the current dose
  print(paste("Dose:", dose_level))
  print(chi_test_result_dose)

  # Calculate 95% confidence interval for the difference in proportions
  prop_table_dose <- prop.table(table_len_supp_dose, margin = 1)
  prop_diff_dose <- prop_table_dose[1, "HIGH"] - prop_table_dose[2, "HIGH"]
  prop_se_dose <- sqrt((prop_table_dose[1, "HIGH"]
                        * (1 - prop_table_dose[1, "HIGH"])
                        / sum(subset_data$supp == "OJ"))
                       + (prop_table_dose[2, "HIGH"]
                          * (1 - prop_table_dose[2, "HIGH"])
                          / sum(subset_data$supp == "VC")))
  conf_interval_dose <- prop_diff_dose + c(-1, 1) * 1.96 * prop_se_dose
```

```r
  print("95% Confidence Interval for difference in proportions:")
  print(conf_interval_dose)
}
```

```
## [1] "Dose: 0.5"
## 
##  Fisher's Exact Test for Count Data
## 
## data:  table_len_supp_dose
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.02564066        Inf
## sample estimates:
## odds ratio 
##        Inf 
## 
## [1] "95% Confidence Interval for difference in proportions:"
## [1] -0.08594193  0.28594193
## [1] "Dose: 1"
## 
##  Fisher's Exact Test for Count Data
## 
## data:  table_len_supp_dose
## p-value = 0.01977
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##     1.387127 1039.290968
## sample estimates:
## odds ratio 
##   17.27587 
## 
## [1] "95% Confidence Interval for difference in proportions:"
## [1] 0.260518 0.939482
## [1] "Dose: 2"
## 
##  Fisher's Exact Test for Count Data
## 
## data:  table_len_supp_dose
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.02564066        Inf
## sample estimates:
## odds ratio 
##        Inf 
## 
## [1] "95% Confidence Interval for difference in proportions:"
## [1] -0.08594193  0.28594193
```

c:

a. Ignoring Dose

- Assumptions:

- The Chi-square test assumes that the expected frequency in each cell of the contingency table is at least 5. If this assumption is violated, Fisher's exact test is a better choice because it doesn't rely on large sample size assumptions.

- The test assumes that the observations are independent and that the proportions are compared between two distinct groups (OJ and VC).

- Validity:

- In our case, if the expected cell counts are sufficient, the Chi-square test is valid. If not, Fisher's exact test provides a valid alternative.

- The 95% confidence interval for the difference in proportions gives us an estimate of how different the proportions of "HIGH" cases are between OJ and VC.

- Remedial Measures:

- If the expected counts in any cell of the table are less than 5, Fisher's exact test should be used to avoid violating the assumptions of the Chi-square test.

b. Taking Dose into Account

- Assumptions:

- Similar to the analysis without dose, each contingency table for a specific dose level must meet the assumptions of the Chi-square test (expected cell counts ≥ 5) for the test to be valid. Otherwise, Fisher's exact test should be used.

- The test also assumes that the groups (OJ and VC) within each dose level are independent, and the proportions of "HIGH" are compared within each dose group.

- Validity:

- For each dose level, the tests are valid as long as the expected cell counts are adequate or Fisher's exact test is used when the counts are low.

- If the results show a significant difference in proportions for some dose levels but not others, this indicates that the effect of the supplement on tooth growth may depend on the dose administered.

- Remedial Measures:

- For any dose group where expected cell counts are low, using Fisher's exact test ensures the validity of the analysis.

- If significant differences are found only in certain dose groups, further investigation into how different doses affect the relationship between supplement type and tooth growth would be useful.

Summary:

By performing both a Chi-square test (or Fisher's exact test if necessary) and calculating confidence intervals, we can determine whether there is a significant difference in the proportions of guinea pigs with "HIGH" tooth growth between OJ and VC groups. When dose is taken into account, the results may vary across dose levels, highlighting the importance of considering dosage when analyzing the effect of Vitamin C on tooth growth.