Computer Assignment 1

Gabriel Lindqvist, Jiahui Li

2024-11-22

Exercise1:1

Task 1

Calculate the percentage in favor and against legal abortion for men and women separately.

```
survey_data<- as.table(rbind(c(309, 191), c(319, 281)))
dimnames(survey_data) <- list(gender = c("women", "men"),
                              opinion = c("favor","against"))

women_in_favor <- survey_data[1,1]
men_in_favor <- survey_data[1,2]
women_total <- sum(survey_data[1, ])
men_total <- sum(survey_data[2, ])

# Calculate percentages for women
women_percent_in_favor <- (women_in_favor / women_total) * 100
women_percent_against <- 100 - women_percent_in_favor

# Calculate percentages for men
men_percent_in_favor <- (men_in_favor / men_total) * 100
men_percent_against <- 100 - men_percent_in_favor

# Print results
cat("Women: In favor =", women_percent_in_favor,
    "%, Against =", women_percent_against, "%\n")
```

```
## Women: In favor = 61.8 %, Against = 38.2 %
```

```
cat("Men: In favor =", men_percent_in_favor,
    "%, Against =", men_percent_against, "%\n")
```

```
## Men: In favor = 31.83333 %, Against = 68.16667 %
```

The proportion of women who are in favor of legal abortion is evidently larger than men's. ## Task 2

```
##
##  Pearson's Chi-squared test
##
## data:  survey_data
## X-squared = 8.2979, df = 1, p-value = 0.003969
```

```
##        opinion
## gender     favor  against
##   women 285.4545 214.5455
##   men   342.5455 257.4545

## Pearson's Chi-squared statistic (X^2): 8.297921

## Likelihood ratio statistic (G^2): 8.32232

## P-value (from Chi-squared test): 0.003969048

## Conclusion: Reject the null hypothesis.
##  There is a significant
##       association between gender and opinion on legal abortion.
```

From the output above we can see that the test statistics are large with small p-values (less than 0.01). This provides strong evidence against the null hypothesis of independence, indicating a significant association between gender and their opinions toward legal abortion.

Task 3

Next we calculate the odds ratio for women and men.

```r
# Calculate odds for women and men
odds_women <- survey_data[1, 1] / survey_data[1, 2]
odds_men <- survey_data[2, 1] / survey_data[2, 2]

# Odds ratio
odds_ratio <- odds_women / odds_men

# Log odds ratio and its standard error
log_odds_ratio <- log(odds_ratio)
se_log_odds_ratio <- sqrt(1 / survey_data[1, 1] + 1 / survey_data[1, 2] +
                          1 / survey_data[2, 1] + 1 / survey_data[2, 2])

# 95% confidence interval for the log odds ratio
ci_log_lower <- log_odds_ratio - 1.96 * se_log_odds_ratio
ci_log_upper <- log_odds_ratio + 1.96 * se_log_odds_ratio

# Transform back to get the confidence interval for the odds ratio
ci_lower <- exp(ci_log_lower)
ci_upper <- exp(ci_log_upper)

# Print results
cat("Odds for women (In favor / Against):", odds_women, "\n")
```

```
## Odds for women (In favor / Against): 1.617801
```

```r
cat("Odds for men (In favor / Against):", odds_men, "\n")
```

```
## Odds for men (In favor / Against): 1.135231
```

```r
cat("Odds Ratio (Women vs Men):", odds_ratio, "\n")
```

```
## Odds Ratio (Women vs Men): 1.425085
```

```r
cat("95% Confidence Interval for Odds Ratio: [", ci_lower, ",", ci_upper, "]\n")
```

```
## 95% Confidence Interval for Odds Ratio: [ 1.119477 , 1.814121 ]
```

```r
# Interpretation
if (1 < ci_lower || 1 > ci_upper) {
  cat("Interpretation: The odds of being in favor of legal abortion",
      "\n", "significantly differ between women and men.\n")
} else {
  cat("Interpretation: The odds of being in favor of legal abortion",
      "\n", "do not significantly differ between women and men.\n")
}
```

```
## Interpretation: The odds of being in favor of legal abortion
##  significantly differ between women and men.
```

- The calculation uses **manual formulas** for:
  - OR: $\frac{(a \cdot d)}{(b \cdot c)}$
  - Standard Error (SE): $\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$
  - CI: Based on the standard normal approximation ($1.96 \times SE$).

From the output above we can see that the estimated odds ratio is OR = 1.425085 with a 95% C.I as (1.119477, 1.814121). This suggests that women are more likely than men to support legal abortion. Furthermore, the C.I does not include 1, which supports that the difference is statistically significant.

Task 4

Next we calculate the risk ratio for women and men.

```r
# Calculate risks (probabilities of being in favor)
risk_women <- survey_data[1, 1] / women_total
risk_men <- survey_data[2, 1] / men_total

# Risk ratio
risk_ratio <- risk_women / risk_men

# Log risk ratio and its standard error
log_risk_ratio <- log(risk_ratio)
se_log_risk_ratio <- sqrt((1 / survey_data[1, 1]) - (1 / women_total) +
                          (1 / survey_data[2, 1]) - (1 / men_total))

# 95% confidence interval for the log risk ratio
ci_log_lower <- log_risk_ratio - 1.96 * se_log_risk_ratio
ci_log_upper <- log_risk_ratio + 1.96 * se_log_risk_ratio
```

```r
# Transform back to get the confidence interval for the risk ratio
ci_lower <- exp(ci_log_lower)
ci_upper <- exp(ci_log_upper)

# Print results
cat("Risk for women (In favor / Total):", risk_women, "\n")
```

```
## Risk for women (In favor / Total): 0.618
```

```r
cat("Risk for men (In favor / Total):", risk_men, "\n")
```

```
## Risk for men (In favor / Total): 0.5316667
```

```r
cat("Risk Ratio (Women vs Men):", risk_ratio, "\n")
```

```
## Risk Ratio (Women vs Men): 1.162382
```

```r
cat("95% Confidence Interval for Risk Ratio: [", ci_lower, ",", ci_upper, "]\n")
```

```
## 95% Confidence Interval for Risk Ratio: [ 1.049742 , 1.287109 ]
```

```r
# Interpretation
if (1 < ci_lower || 1 > ci_upper) {
  cat("Interpretation: The relative risk of being in favor of legal abortion",
      "\n", "significantly differs between women and men.\n")
} else {
  cat("Interpretation: The relative risk of being in favor of legal abortion",
      "\n", "does not significantly differ between women and men.\n")
}
```

```
## Interpretation: The relative risk of being in favor of legal abortion
##  significantly differs between women and men.
```

- Risk Ratio:

$$RR = \frac{\text{Risk for Women}}{\text{Risk for Men}}$$

- Confidence Interval:

$$CI = \exp\left[\log(RR) \pm 1.96 \cdot SE\right]$$

  where:

$$SE = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}$$

From the output above we can see that the estimated risk ratio RR = 1.162382 with a 95% C.I as (1.049742, 1.287109). This suggests that women's relative risk of being in favor of legal abortion is 1.162 times that of men. Furthermore, the C.I does not include 1, which supports that the difference is statistically significant relative to gender. ## Task 5 ### Run the code given by the file to address Q1-4

```r
#Generate a frequency table and calculate row percentage
tab1<- as.table(rbind(c(309, 191), c(319, 281)))
dimnames(tab1) <- list(gender = c("women", "men"),
                       opinion = c("favor","against"))
addmargins(tab1)
```

```
##        opinion
## gender  favor against  Sum
##   women   309     191  500
##   men     319     281  600
##   Sum     628     472 1100
```

```r
addmargins(prop.table(tab1,1),2)
```

```
##        opinion
## gender      favor   against       Sum
##   women 0.6180000 0.3820000 1.0000000
##   men   0.5316667 0.4683333 1.0000000
```

```r
#Calculate X2, G2 and p-values
chisq.test(tab1,correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab1
## X-squared = 8.2979, df = 1, p-value = 0.003969
```

```r
library(MASS)
loglm(~gender+opinion,tab1)
```

```
## Call:
## loglm(formula = ~gender + opinion, data = tab1)
##
## Statistics:
##                       X^2 df    P(> X^2)
## Likelihood Ratio 8.322320  1 0.003916088
## Pearson          8.297921  1 0.003969048
```

The test statistics are equal to the results generated form the basic R code, which also indicates a significant association between gender and the opinions toward legal abortion.

```r
#Calculate odds ratio and confidence interval
library(epitools)
oddsratio(tab1, method = "wald", rev="neither")$measure
```

```
##        odds ratio with 95% C.I.
## gender  estimate    lower    upper
##   women 1.000000       NA       NA
##   men   1.425085 1.119482 1.814113
```

Both the basic R calculation and the `oddsratio()` function provide similar results for the **Odds Ratio (OR)**, but there are small differences in the confidence interval (CI) and methodological details. Differences are due to implementation precision, but they are negligible and do not impact the interpretation.

```
#Calculate risk ratio and confidence interval
riskratio(tab1,rev="both")$measure
```

```
##        risk ratio with 95% C.I.
## gender  estimate    lower    upper
##   men   1.000000       NA       NA
##   women 1.162382 1.049744 1.287107
```

Both the basic R calculation and the `riskratio()` function provide similar results for the **Risk Ratio (RR)**, but there are small differences in the confidence interval (CI) and methodological details. Differences are due to implementation precision, but they are negligible and do not impact the interpretation.

Exercise 1:2

Task 1

First we enter the given data in R and generate a contingency table.

```
tab2 <- as.table(rbind(c(557, 1835 - 557), c(1198, 2691 - 1198)))
dimnames(tab2) <- list(gender = c("women", "men"),
                       admission = c("admitted", "not admitted"))
addmargins(tab2)
```

```
##        admission
## gender  admitted not admitted  Sum
##   women      557         1278 1835
##   men       1198         1493 2691
##   Sum       1755         2771 4526
```

```
addmargins(prop.table(tab2, 1), 2)
```

```
##        admission
## gender   admitted not admitted       Sum
##   women 0.3035422    0.6964578 1.0000000
##   men   0.4451877    0.5548123 1.0000000
```

Next, we want to do the same analysis as in Exercise 1:1. Firstlym we test whether there is an association between gender and admission using χ^2 and G^2 statistics.

```
## Call:
## loglm(formula = ~gender + admission, data = tab2)
##
## Statistics:
##                       X^2 df P(> X^2)
## Likelihood Ratio 93.44941  1        0
## Pearson          92.20528  1        0
```

From the output above we can see that the test statistics are large with small p-values (less than 0.01). This provides strong evidence against the null hypothesis of independence, indicating a significant association between gender and admission.

Next, we calculate the odds ratio of admission for men relative relative to women.

```
##        odds ratio with 95% C.I.
## gender  estimate    lower    upper
##   women  1.00000       NA       NA
##   men    1.84108 1.624377 2.086693
```

From the output above we can see that the estimated odds ratio is OR $= 1.841$ with a 95% C.I as $(1.624, 2.087)$. This suggests that that men have higher odds of being admitted compared to women. Furthermore, the CI does not include 1, which supports that the difference is statistically significant.

Lastly, we compute the risk ratio for men relative to women.

```
##        risk ratio with 95% C.I.
## gender  estimate   lower    upper
##   women 1.000000      NA       NA
##   men   1.466642 1.35235 1.590592
```

From the output above we can see that the estimate risk ratio RR $= 1.467$ with a 95% C.I as $(1.352, 1.591)$. This indicates that men are 46.7% more likely of being admitted compared to women. The C.I does not include 1 here, which further supports that there is a distinct difference in admission relative to gender.

Task 2

The next task is to investigate the effect of replacing all values in the contingency table with one-tenth of the original values, and then comment on what we observe.

```
##        admission
## gender  admitted not admitted Sum
##   women       56          128 184
##   men        120          149 269
##   Sum        176          277 453

##        admission
## gender   admitted not admitted       Sum
##   women 0.3043478    0.6956522 1.0000000
##   men   0.4460967    0.5539033 1.0000000

## Call:
## loglm(formula = ~gender + admission, data = tab2_2)
##
## Statistics:
##                       X^2 df    P(> X^2)
## Likelihood Ratio 9.364274  1 0.002212556
## Pearson          9.240912  1 0.002366671

##        odds ratio with 95% C.I.
## gender  estimate    lower    upper
##   women 1.000000       NA       NA
##   men   1.840844 1.239547 2.733825
```

```
##        risk ratio with 95% C.I.
## gender  estimate    lower    upper
##   women 1.000000       NA       NA
##   men   1.465746 1.134883 1.893069
```

First we observe that the Pearson's and LR statistics values change by 1/10, which yields higher p-values. Secondly we observe that the estimates for the odds and risk ratios are still the same. However, the 95% C.I are much wider compared to using the original data. This is expected, because smaller sample sizes introduces greater uncertainty, ultimately reducing the precision of the estimates.

Next, we repeat by replacing the numbers with one hundredth of the original values.

```
##        admission
## gender  admitted not admitted Sum
##   women        6           13  19
##   men         12           15  27
##   Sum         18           28  46
```

```
##        admission
## gender   admitted not admitted       Sum
##   women 0.3157895    0.6842105 1.0000000
##   men   0.4444444    0.5555556 1.0000000
```

```
## Call:
## loglm(formula = ~gender + admission, data = tab2_2)
##
## Statistics:
##                        X^2 df  P(> X^2)
## Likelihood Ratio 0.7833632  1 0.3761145
## Pearson          0.7749930  1 0.3786768
```

```
##        odds ratio with 95% C.I.
## gender  estimate     lower    upper
##   women 1.000000        NA       NA
##   men   1.733333 0.5068344 5.927862
```

```
##        risk ratio with 95% C.I.
## gender  estimate     lower    upper
##   women 1.000000        NA       NA
##   men   1.407407 0.6420765 3.084984
```

The estimated odds and risk ratios are roughly the same as the original ones, which is expected since we just scale the data entries with 1/100 and then round the values. Further, the statistics have also been scaled with approximately 1/100 compared to the statistics in Task 1, which is why the p-values are now larger. Lastly the 95% C.I are much wider now and does also include 1 in them (which is expected with smaller sample sizes).

Task 3

In this task we want to create a new two-way contingency table that satisfy the following conditions:

- The sample odds ratio $\hat{\theta}$ is within the interval $(0.99, 1.01)$.

- The population odds ratio θ is significantly different from 1.

Finally, we will comment on the relevance of declaring 'statistical significance' in a situation the one like above.

Creating the Contingency Table

In order for $\hat{\theta}$ to satisfy the first condition, we know that each cell count $n_{ij},\ i,j = 1,2$ needs to be approximately the same such that

$$\hat{\theta} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}} \approx 1.$$

To satisfy the second condition, we need larger n_{ij}, as larger sample sizes increase statistical power, making small imbalances more likely to yield significant results. With the reasoning above, we construct the following contingency table

```
##          outcome
## group     Outcome 1 Outcome 2  Total
##   Group 1    200000    199000 399000
##   Group 2    199000    200000 399000
##   Total      399000    399000 798000
```

```
##          odds ratio with 95% C.I.
## group     estimate   lower    upper
##   Group 1 1.000000      NA       NA
##   Group 2 1.010076 1.00125 1.018979
```

From above, we observe that the sample odds ratio satisfies the first condition, with $\hat{\theta} \approx 1.01$. Next, we test whether the second condition holds.

```
loglm(~group + outcome, tab3)
```

```
## Call:
## loglm(formula = ~group + outcome, data = tab3)
##
## Statistics:
##                       X^2 df   P(> X^2)
## Likelihood Ratio 5.012537  1 0.02516441
## Pearson          5.012531  1 0.02516449
```

The test output shows that the null hypothesis of independence is rejected at the 5% significance level. Thus satisfying the second condition.

Comments on Above Results

The results above highlights the fact that statistical tests are sensitive to small imbalances when sample sizes are large, even if the sample odds ratio is weak. In the case above, the sample odds ratio indicates that there is no association, yet the large sample size leads to a significant test. This demonstrates that statistical significance does not always equate to practical relevance, which is why relying solely on statistical significance can be misleading in cases like the one above. Instead, both the sample odds ratio and the tests should be included when reporting the results, whilst weighing the relevance of the test.