Problem Setup

Suppose there are 8000 people in a town. Out of the 8000 people:

- 800 are females
- 7200 are males

In that town, 1600 people are employed, of which:

- 120 are females
- 1480 are males

Using this information, we will calculate and analyze various metrics to understand the association between gender and employability.

Step 1: Set Up the Data

First, we define the data for the contingency table and create it to perform our calculations.

```r
# Define data in a contingency table
employment_data <- matrix(c(120, 680, 1480, 5720), nrow = 2, byrow = TRUE,
                          dimnames = list(Gender = c("Female", "Male"),
                                          Employment = c("Employed", "Unemployed")))
employment_data
```

```
##         Employment
## Gender   Employed Unemployed
##   Female      120        680
##   Male       1480       5720
```

Part (a): Confidence Interval for Odds Ratio, θ

The odds ratio (OR) measures the odds of employment for females relative to males. To calculate the confidence interval for the odds ratio, we use the following steps:

Steps to Calculate Confidence Interval for Odds Ratio:

1. **Compute the natural logarithm of the odds ratio (ln(OR))**: This transforms the odds ratio, allowing us to calculate the confidence interval on a symmetrical scale.
2. **Calculate the standard error (SE) of ln(OR)**: Use the formula based on observed counts in the contingency table.
3. **Determine the confidence interval**: For a 95% confidence level, use the formula:

   $$\text{CI} = \exp(\ln(\text{OR}) \pm Z \times \text{SE})$$

   where Z is the Z-score corresponding to the desired confidence level (e.g., 1.96 for 95%).

This process yields the lower and upper bounds of the confidence interval for the odds ratio, indicating the range within which the true odds ratio is likely to fall.

```r
# Calculate odds ratio and its confidence interval
library(epitools)
odds_ratio_result <- oddsratio(employment_data, method = "wald")
odds_ratio <- odds_ratio_result$measure[2, 1]
conf_int_odds_ratio <- odds_ratio_result$measure[2, c(2, 3)]
list(odds_ratio = odds_ratio, conf_int_odds_ratio = conf_int_odds_ratio)
```

```
## $odds_ratio
## [1] 0.682035
##
## $conf_int_odds_ratio
##     lower     upper
## 0.5571157 0.8349643
```

Part (b): Confidence Interval for Difference of Proportion, $\pi_1 - \pi_2$

To assess the difference in employment proportions between females and males, we calculate the confidence interval for the difference of proportions.

Steps to Calculate Confidence Interval for Difference of Proportion:

1. **Calculate the observed proportions** π_1 and π_2: These represent the proportion of employed females and males, respectively.
2. **Determine the standard error (SE) for the difference of proportions**: Use the formula

$$\text{SE} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

   where n_1 and n_2 are the sample sizes for females and males.
3. **Calculate the confidence interval**: For a 95% confidence level, use

$$\text{CI} = (\pi_1 - \pi_2) \pm Z \times \text{SE}$$

   where Z is the Z-score (e.g., 1.96 for 95% confidence).

This interval provides a range for the difference in employment proportions, indicating the potential variation in employment likelihood between females and males.

```r
# Proportion of employed females and males
p1 <- employment_data["Female", "Employed"] / sum(employment_data["Female", ])
p2 <- employment_data["Male", "Employed"] / sum(employment_data["Male", ])
# Difference in proportions
diff_proportion <- p1 - p2
# Confidence interval for difference in proportions
prop_diff_conf <- prop.test(x = c(employment_data["Female", "Employed"], employment_data["Male", "Emplo
                            n = c(sum(employment_data["Female", ]), sum(employment_data["Male", ])),
                            correct = FALSE)$conf.int
list(diff_proportion = diff_proportion, conf_int_diff_proportion = prop_diff_conf)
```

```
## $diff_proportion
```

```
## [1] -0.05555556
##
## $conf_int_diff_proportion
## [1] -0.08200098 -0.02911014
## attr(,"conf.level")
## [1] 0.95
```

Part (c): Confidence Interval for Relative Risk, r

Relative risk compares the probability of employment between females and males. To compute the confidence interval for the relative risk, we use the following steps:

Steps to Calculate Confidence Interval for Relative Risk:

1. **Calculate the observed probabilities** π_1 (for females) and π_2 (for males).
2. **Determine the natural logarithm of the relative risk ($\ln(\mathrm{RR})$)**: This transformation allows calculation on a symmetrical scale.
3. **Calculate the standard error (SE) for $\ln(\mathrm{RR})$**: The standard error is given by:

$$\mathrm{SE} = \sqrt{\frac{1-\pi_1}{n_1\pi_1} + \frac{1-\pi_2}{n_2\pi_2}}$$

   where n_1 and n_2 are the sample sizes for females and males.
4. **Calculate the confidence interval**: For a 95% confidence level, use the formula:

$$\mathrm{CI} = \exp(\ln(\mathrm{RR}) \pm Z \times \mathrm{SE})$$

   where Z is the Z-score corresponding to the desired confidence level (e.g., 1.96 for 95%).

This confidence interval provides a range within which the true relative risk of employment between females and males is likely to fall.

```r
# Calculate relative risk and its confidence interval
relative_risk_result <- riskratio(employment_data)
relative_risk <- relative_risk_result$measure[2, 1]
conf_int_relative_risk <- relative_risk_result$measure[2, c(2, 3)]
list(relative_risk = relative_risk, conf_int_relative_risk = conf_int_relative_risk)
```

```
## $relative_risk
## [1] 0.9346405
##
## $conf_int_relative_risk
##     lower     upper
## 0.9057565 0.9644457
```

Part (d): Comment on Association Based on Confidence Intervals

Using the confidence intervals from Parts (a), (b), and (c), we interpret the association between gender and employment status as follows:

1. **Odds Ratio (OR)**:

- If the confidence interval for the odds ratio includes 1, there is no significant association between gender and employment.
- If the interval is entirely above 1, this indicates a positive association (higher odds for females relative to males).
- If the interval is entirely below 1, it suggests a negative association (lower odds for females relative to males).

2. **Difference of Proportions** ($\pi_1 - \pi_2$):
   - If the confidence interval for the difference in proportions includes 0, there is no significant difference in employment proportions between females and males.
   - If the interval is entirely above 0, this implies females have a higher employment proportion than males.
   - If the interval is entirely below 0, it indicates a lower employment proportion for females relative to males.
3. **Relative Risk (RR)**:
   - If the confidence interval for the relative risk includes 1, it suggests no significant difference in employment probability between genders.
   - An interval entirely above 1 indicates females are more likely to be employed compared to males.
   - An interval entirely below 1 suggests a lower employment likelihood for females compared to males.

Interpretation:

By examining these confidence intervals, we gain insights into the strength and direction of the association between gender and employment. If any of the intervals consistently indicate a positive or negative association, this suggests a potential gender-based disparity in employment status.

Part (e): Chi-Square Statistic, X^2

To assess the discrepancy between observed and expected frequencies, we calculate the Chi-Square statistic:

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i}$$

where f_i is the observed frequency, and e_i is the expected frequency for each cell.

This Chi-Square statistic helps determine if there is a significant difference between the observed and expected counts, providing insight into the association between gender and employment status.

```r
# Calculate Chi-square statistic
chi_square_test <- chisq.test(employment_data, correct = FALSE)
chi_square_stat <- chi_square_test$statistic
p_value_chi_square <- chi_square_test$p.value
list(chi_square_stat = chi_square_stat, p_value = p_value_chi_square)
```

```
## $chi_square_stat
## X-squared
##  13.88889
##
## $p_value
## [1] 0.0001939416
```

Part (f): Wilk's Statistic, G^2

Wilk's statistic, also known as the likelihood ratio test, compares the observed distribution to an expected distribution under the null hypothesis of independence.

Formula for Wilk's Statistic:

$$G^2 = 2 \sum f_i \ln \left(\frac{f_i}{e_i} \right)$$

where f_i is the observed frequency and e_i is the expected frequency for each cell.

Wilk's statistic provides an alternative to the Chi-Square test, often used in cases where data are sparse or when the Chi-Square test assumptions may not hold, offering insight into the association between variables based on likelihood ratios.

```
# Calculate Wilk's G^2 statistic manually

# Observed frequencies
observed <- as.vector(employment_data)

# Expected frequencies under independence assumption
expected <- chisq.test(employment_data, correct = FALSE)$expected

# Calculate G^2 statistic
wilk_stat <- 2 * sum(observed * log(observed / expected))

# Display Wilk's statistic
wilk_stat
```

```
## [1] 14.78517
```

Part (g): Comparison with Chi-Square Cut-Off Value

To determine if we should reject the null hypothesis $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$, which assumes independence between the variables (e.g., gender and employment status), we compare the calculated Chi-square and Wilk's statistics to the Chi-square critical value.

Steps for Comparison:

1. **Determine the Chi-square critical value**: Use the Chi-square distribution table at a chosen significance level (e.g., 0.05) and degrees of freedom based on the contingency table.
2. **Compare the Statistics**:
   - If both the Chi-square statistic (χ^2) and Wilk's statistic (G^2) exceed the critical value, we reject H_0, indicating a statistically significant association.
   - If both are below the critical value, we fail to reject H_0, implying no significant association.

This comparison helps confirm whether the observed association in the data is likely due to chance or indicates a meaningful relationship.

```r
# Chi-square critical value for 1 degree of freedom at 5% significance level
chi_square_critical <- qchisq(0.95, df = 1)
decision_chi_square <- ifelse(chi_square_stat > chi_square_critical, "Reject H0", "Fail to Reject H0")
decision_wilk <- ifelse(wilk_stat > chi_square_critical, "Reject H0", "Fail to Reject H0")
list(chi_square_critical = chi_square_critical,
     decision_chi_square = decision_chi_square,
     decision_wilk = decision_wilk)
```

```
## $chi_square_critical
## [1] 3.841459
##
## $decision_chi_square
##   X-squared
## "Reject H0"
##
## $decision_wilk
## [1] "Reject H0"
```