Problem 1: HairEyeColor Dataset

Dataset Description

The **HairEyeColor** dataset contains the distribution of hair and eye color along with gender in a sample of 592 students. The data categorizes students by hair color, eye color, and sex.

Solution Steps:

Step 1: Load the Dataset and Convert to Data Frame

We will begin by loading the `HairEyeColor` dataset and converting it to a data frame to create a contingency table.

```
# Load the HairEyeColor dataset
data("HairEyeColor")

# Convert it to a data frame
hair_eye_df <- as.data.frame(HairEyeColor)

# Display the data frame
print(head(hair_eye_df))
```

```
##    Hair   Eye  Sex Freq
## 1 Black Brown Male   32
## 2 Brown Brown Male   53
## 3   Red Brown Male   10
## 4 Blond Brown Male    3
## 5 Black  Blue Male   11
## 6 Brown  Blue Male   50
```

Step 2: Create Contingency Table for Hair and Eye Color

Using the `xtabs` function, we'll create a contingency table that cross-tabulates hair and eye colors across genders.

```
# Create a contingency table for Hair and Eye color
hair_eye_table <- xtabs(Freq ~ Hair + Eye, data = hair_eye_df)
```

Exercise 1.a: Verify Marginal Sums

Verify that the sum of the marginals of the rows is equal to the sum of the marginals of the columns, denoted by n.

Confirm that $p_{1+} + p_{2+} = p_{+1} + p_{+2} = 1$.

To add marginals to the table, we use the `addmargins()` function.

```r
# Add marginals to the contingency table
marginal_table <- addmargins(hair_eye_table)
```

Problem 2: Employment Status vs. Sex

Suppose there is a town with 8,000 people, including 800 females. In this town, 1,600 people are employed, and of those employed, only 120 are female.

Solution Steps:

Step 1: Create a Contingency Table of Employment Status vs. Sex.

```r
# Define data for the contingency table
employment_data <- data.frame(
  Sex = rep(c("Female", "Male"), each = 2),
  Status = rep(c("Employed", "Unemployed"), times = 2),
  Count = c(120, 680, 1480, 5720)
)

# Create a contingency table
employment_table <- xtabs(Count ~ Sex + Status, data = employment_data)
```

Step 2: Calculate Expected Frequencies

Calculate the expected frequency e_i for each cell, using the formula:

$$e_i = \frac{n_i}{n}$$

where n is the total count.

```r
# Calculate total count
total_count <- sum(employment_table)

# Calculate expected frequencies
expected_frequencies <- employment_table / total_count
```

Step 3: Calculate Chi-Square Statistic

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.

```r
# Calculate Chi-Square statistic
chi_square_stat <- sum((employment_table - expected_frequencies * total_count)^2 / (expected_frequencies
```