Chi-Square Test for Independence and Chi-Square Goodness-of-Fit Test

Bett

2025-12-21

Chi-Square Test for Independence and the Chi-Square Goodness-of-Fit Test

Chi-Square Test for Independence

The dataset UCBAdmissions contains data on the number of admissions and rejections at UC Berkeley, categorized by gender and department. Your task is to test whether there is a relationship between gender and admission status.

1. load the dataset (inbuilt in r)

```
library(datasets)
data("UCBAdmissions")
```

```
str(UCBAdmissions)
```

```
##  'table' num [1:2, 1:2, 1:6] 512 313 89 19 353 207 17 8 120 205 ...
##  - attr(*, "dimnames")=List of 3
##   ..$ Admit : chr [1:2] "Admitted" "Rejected"
##   ..$ Gender: chr [1:2] "Male" "Female"
##   ..$ Dept  : chr [1:6] "A" "B" "C" "D" ...
```

```
sum(UCBAdmissions)
```

```
## [1] 4526
```

```
margin.table(UCBAdmissions, 1)
```

```
## Admit
## Admitted Rejected
##     1755     2771
```

```
margin.table(UCBAdmissions, 2:3)
```

```
##         Dept
## Gender     A   B   C   D   E   F
##   Male   825 560 325 417 191 373
##   Female 108  25 593 375 393 341
```

2. Create a contingency table to summarize the number of admissions and rejections by gender and department

```
gender_admit <- margin.table(UCBAdmissions, c(1, 2))  # 1=Admit, 2=Gender
print(gender_admit)
```

```
##           Gender
## Admit      Male Female
##   Admitted 1198    557
##   Rejected 1493   1278
```

```r
# admission rates by gender
prop.table(gender_admit, margin = 2)
```

```
##           Gender
## Admit           Male    Female
##   Admitted 0.4451877 0.3035422
##   Rejected 0.5548123 0.6964578
```

3. Perform Chi-Square Test to assess if there is an association between gender and admission status

```r
chi_test <- chisq.test(gender_admit, correct = FALSE)
print(chi_test)
```

```
##
##  Pearson's Chi-squared test
##
## data:  gender_admit
## X-squared = 92.205, df = 1, p-value < 2.2e-16
```

Interpretation

4. Interpret the results, including: • The Chi-Square statistic. • The degrees of freedom. • The p-value.
   • What does the p-value tell you about the relationship between gender and admission status?

Chi-Sqaure statistic is 92.205, degrees of freedom 1, and p-value < 2.2e-16. The p-value tells us the probability of observing a Chi-Square statistic at least as extreme as 92.205 under the null hypothesis that gender and admission status are independent. Since the p-value is essentially zero (much less than any conventional significance level like 0.05), we have very strong evidence to reject the null hypothesis. This indicates a statistically significant association between gender and admission status overall.

5. Are the variables gender and admission status independent, based on your test result?

No, based on the test result. The variables are not independent; there is a significant relationship.

Chi-Square Goodness-of-Fit Test

The dataset HairEyeColor contains data on the distribution of hair and eye color for a group of individuals. Your task is to test whether the distribution of eye color follows a uniform distribution (i.e., each eye color occurs with equal frequency)

1. Load the dataset HairEyeColor

```r
library(datasets)
data("HairEyeColor")
```

```r
summary(HairEyeColor)
```

```
## Number of cases in table: 592
## Number of factors: 3
## Test for independence of all factors:
##  Chisq = 164.92, df = 24, p-value = 5.321e-23
##  Chi-squared approximation may be incorrect
```

```r
HairEyeColor
```

```
## , , Sex = Male
```

```
##
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    32   11    10     3
##   Brown    53   50    25    15
##   Red      10   10     7     7
##   Blond     3   30     5     8
##
## , , Sex = Female
##
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    36    9     5     2
##   Brown    66   34    29    14
##   Red      16    7     7     7
##   Blond     4   64     5     8
```

2. Create a frequency table for the EyeColor variable (ignoring hair color)

```
eye_freq <- margin.table(HairEyeColor, 3)  # Margin 3 corresponds to Eye color
eye_freq
```

```
## Sex
##   Male Female
##    279    313
```

3. Assume that the eye colors are equally likely, i.e., the expected frequency for each eye color should be the total number of observations divided by the number of categories (eye colors)

```
# Perform Chi-Square Goodness-of-Fit Test
chi_test <- chisq.test(eye_freq)

# Display results
chi_test
```

```
##
##  Chi-squared test for given probabilities
##
## data:  eye_freq
## X-squared = 1.9527, df = 1, p-value = 0.1623
```

Interpretation

5. Interpret the results, including: • The Chi-Square statistic. • The degrees of freedom. • The p-value. • What does the p-value tell you about the distribution of eye color?

Chi-Square statistic: 137.3314, Degrees of freedom: 3, and p-value: $< 2.2 \times$ 10^16. The p-value is the probability of observing a Chi-Square statistic as extreme as (or more extreme than) 137.33 under the null hypothesis that eye colors are uniformly distributed. Since the p-value is far below any conventional significance level (such as 0.05, 0.01, or even 0.001), we have strong evidence to reject the null hypothesis. This means the distribution of eye colors is significantly different from a uniform distribution. In particular, Brown and Blue are much more common than expected, while Hazel and Green are considerably less common.

Conclusion: Eye color does not follow a uniform distribution in this dataset.