Chi-Squared

Table of contents


```
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.2.3
```

```
library(descr)
```

```
Warning: package 'descr' was built under R version 4.2.3
```

0.1 Example using the Iris Data

Alternative datasets:

- mtcars:
  - cyl (Cylinders: 4, 6, 8)
  - am (Transmission: 0 = automatic, 1 = manual)
  - gear (Number of gears)
- warpbreaks
  - wool (A or B)
  - tension (L, M, H)

0.1.1 Load Data

```
data(iris)
str(iris)
```

```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

0.1.2 Data Manipulation

```
# View(iris)
```

0.1.2.1 Species

```
freq(as.ordered(iris$Species), plot = FALSE)
```

```
as.ordered(iris$Species)
           Frequency Percent Cum Percent
setosa            50   33.33       33.33
versicolor        50   33.33       66.67
virginica         50   33.33      100.00
Total            150  100.00
```

0.1.2.2 Sepal Length

```
freq(as.ordered(iris$Sepal.Length), plot = FALSE)
```

```
as.ordered(iris$Sepal.Length)
     Frequency  Percent Cum Percent
4.3          1   0.6667      0.6667
4.4          3   2.0000      2.6667
4.5          1   0.6667      3.3333
4.6          4   2.6667      6.0000
```

```
4.7           2   1.3333      7.3333
4.8           5   3.3333     10.6667
4.9           6   4.0000     14.6667
5            10   6.6667     21.3333
5.1           9   6.0000     27.3333
5.2           4   2.6667     30.0000
5.3           1   0.6667     30.6667
5.4           6   4.0000     34.6667
5.5           7   4.6667     39.3333
5.6           6   4.0000     43.3333
5.7           8   5.3333     48.6667
5.8           7   4.6667     53.3333
5.9           3   2.0000     55.3333
6             6   4.0000     59.3333
6.1           6   4.0000     63.3333
6.2           4   2.6667     66.0000
6.3           9   6.0000     72.0000
6.4           7   4.6667     76.6667
6.5           5   3.3333     80.0000
6.6           2   1.3333     81.3333
6.7           8   5.3333     86.6667
6.8           3   2.0000     88.6667
6.9           4   2.6667     91.3333
7             1   0.6667     92.0000
7.1           1   0.6667     92.6667
7.2           3   2.0000     94.6667
7.3           1   0.6667     95.3333
7.4           1   0.6667     96.0000
7.6           1   0.6667     96.6667
7.7           4   2.6667     99.3333
7.9           1   0.6667    100.0000
Total       150 100.0000
```

```r
# Create a categorical version of Sepal.Length
iris$SepalCat <- cut(iris$Sepal.Length,
                     breaks = 3,
                     labels = c("Short", "Medium", "Long"))
```

```r
freq(as.ordered(iris$SepalCat), plot = FALSE)
```

```
as.ordered(iris$SepalCat)
       Frequency Percent Cum Percent
```

```
Short          59   39.33       39.33
Medium         71   47.33       86.67
Long           20   13.33      100.00
Total         150  100.00
```

0.1.3 Create Table

```r
# Table with Speal Category and Species
table_iris <- table(iris$SepalCat, iris$Species)
print(table_iris)
```

```
        setosa versicolor virginica
  Short      47         11         1
  Medium      3         36        32
  Long        0          3        17
```

0.1.4 Assumptions

```r
# Cells should be greater than 5
chisq.test(table_iris)$expected
```

```
           setosa versicolor virginica
  Short  19.666667  19.666667 19.666667
  Medium 23.666667  23.666667 23.666667
  Long    6.666667   6.666667  6.666667
```

0.1.5 Chi-squared Test

```r
# Cells have less than 5, Fisher's test would be appropriate
# Will proceed with Pearson's to capture test statistic
chisq_test_result <- chisq.test(table_iris)
chisq_test_result
```

```
    Pearson's Chi-squared test

data:  table_iris
X-squared = 111.63, df = 4, p-value < 2.2e-16
```

Interpretation: A chi-square test of independence was conducted to determine if there was an association between species (setosa, versicolor, virginica) and sepal length (short, medium, or long). The results were significant, χ^2 (4)= 111.63, $p < 0.001$.