Chi-Square Tests

January 21, 2024.

1. The chi-square test is a statistical test used to determine if there is a significant difference between the **observed values** and the **expected values** in a categorical data set.
2. It measures the **deviation** between the expected and observed frequencies in one or more categories and assesses whether this deviation is statistically significant.
3. The result of a chi-square test is a test statistic, and its p-value is compared against a threshold (e.g. $\alpha = 0.05$) to determine the statistical significance of the observed differences.
4. The test is commonly used in **hypothesis testing**, contingency table analysis, and **goodness-of-fit** testing.

Types of Tests

1. There are several types of chi-square tests. The most popular tests are as follows.

Test	Use
A. Goodness-of-fit test	determine if a sample of data **fits a specified distribution**
B. Independence test	determine if there is a **relationship between two categorical variables**

2. Additional chi-square tests include:

- Homogeneity test: used to determine if different populations have the same distribution of a categorical variable.
- Contingency table test: used to analyze the relationship between two or more categorical variables in a multi-dimensional table.
- McNemar's test: used to determine if the difference between paired nominal data is significant.
- Likelihood-ratio test: used to compare nested models, where the more complex model is tested against a simpler model.
- Mantel-Haenszel test: used to determine if there is a relationship between two categorical variables while controlling for the effect of a third variable.

Chi-Square Goodness of Fit test

A Business Application of the Chi-Square Goodness of Fit test

1. Suppose a **retail** company wants to know **if the gender distribution of their customer base is representative of the general population**. They collect data on the gender of a sample of their customers and compare it to the expected distribution (e.g. 50% male and 50% female).
2. The Chi-Square Goodness of Fit test is then used to evaluate the **null hypothesis** that the observed distribution of gender among the company's customers is the same as the expected distribution.
3. The test could help us potentially conclude that there is a significant difference between the observed and expected distributions, and the company's customer base may not be representative of the general population.
4. For example, **the results might show that a significantly higher proportion of females than expected are customers of the company**. This information could be used by the retail company to tailor their marketing strategies and product offerings to better attract male customers.

Chi-Square Goodness of Fit test – Technicalities

1. The chi-square goodness-of-fit test is a statistical test that is used to determine if a sample of categorical data fits a specified distribution.
2. It compares the **observed frequencies** of different categories in the sample data to the **expected frequencies** based on the specified distribution.
3. The test statistic is calculated as the sum of the squared differences between the observed and expected frequencies, divided by the expected frequencies.
4. **If the test statistic is larger than the critical value from a chi-square distribution table**, it can be concluded that there is a significant difference between the observed and expected distributions
5. The resulting p-value is then compared against a threshold (e.g. alpha=0.05) to determine the statistical significance of the observed differences.
6. It can be used to test hypotheses about the distribution of a categorical variable, or testing if a population is homogeneous with respect to a particular attribute.
7. The test assumes that the sample size is large enough and that the expected frequencies are greater than or equal to 5 for all categories.

Running the Chi-Square Goodness of Fit Test in R

1. The `mtcars` dataset consists of 32 cars, each having 3, 4 or 5 gears.

```
# Load the data
data(mtcars)

# Create a table of the number of cars in each gear category
t0 <- table(mtcars$gear)
t0
```

```
 3  4  5 
15 12  5 
```

2. Graphical display of Contingency table:

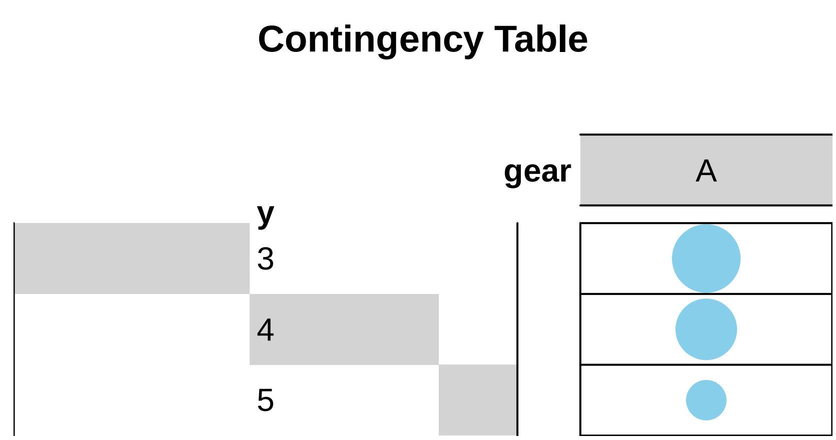
```
library("gplots")
```

```
Attaching package: 'gplots'

The following object is masked from 'package:stats':

    lowess
```

```
# 1. convert the data as a table
dt <- as.table(as.matrix(t0))
# 2. Graph
balloonplot(t(dt), main ="Contingency Table",
            xlab ="gear",
            label = FALSE, show.margins = FALSE)
```



3. Proportions

```r
t1 = prop.table(t0)
t1
```

```
      3       4       5 
0.46875 0.37500 0.15625 
```

4. Running the test

```r
# Specify the distribution we want to test against
expected_probs <- c(0.5, 0.3, 0.2)

# Perform the chi-square goodness-of-fit test
chisq.test(t1, p=expected_probs)
```

```
Warning in chisq.test(t1, p = expected_probs): Chi-squared approximation may be
incorrect

	Chi-squared test for given probabilities

data:  t1
X-squared = 0.030273, df = 2, p-value = 0.985
```

5. This code creates a table of the number of cars in each gear category, specifies the distribution we want to test against (in this case, a distribution where 50% of the cars have 3 gears, 30% have 4 gears, and 20% have 5 gears), calculates the expected frequencies based on the expected probabilities, and then performs the chi-square goodness-of-fit test.
6. The resulting p-value will indicate the significance of the observed differences between the observed and expected frequencies.

Chi-Square Test of Independence

A Business Application of the Chi-Square Test of Independence

1. Suppose a grocery store wants to evaluate the **association** between the **type of product** a customer buys and their **age group**.
2. Suppose the grocery store wants to know if there is a significant association between the **type of product a customer buys** (**e.g. fruits, vegetables, or dairy products**) and their **age group** (e.g. **18-30, 31-45, 46-60, 61 and older**). They collect data on the age group and product type for a sample of customers and create a contingency table.

3. The Chi-square test is then used to evaluate the **null hypothesis** that the **product type and age group are independent**.
4. Depending on the test result, we could potentially conclude that there is a significant association between the two variables.
5. For example, the results might show that **a significantly higher proportion of customers in the 46-60 age group buy fruits compared to the other age groups**.
6. This information could be used by the grocery store to adjust their marketing strategies and product offerings to better cater to their target customers.

Chi-Square Test of Independence – Technicalities

1. The Chi-square test of independence is a statistical test used to determine if there is a significant association between two categorical variables.
2. The test evaluates the **null hypothesis** that the **two variables are independent** and **calculates a test statistic (chi-square)** based on the difference between the observed and expected frequencies of the two variables.
3. If the calculated test statistic is larger than the **critical value from a chi-square distribution table**, then the null hypothesis is rejected and it can be concluded that there is a significant association between the two variables.

Running the Chi-Square Test of Independence Test in R

1. Convert the categorical variables into factor variables in the dataset mtcars

```r
data(mtcars)
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$am  <- as.factor(mtcars$am)
mtcars$gear  <- as.factor(mtcars$gear)
```

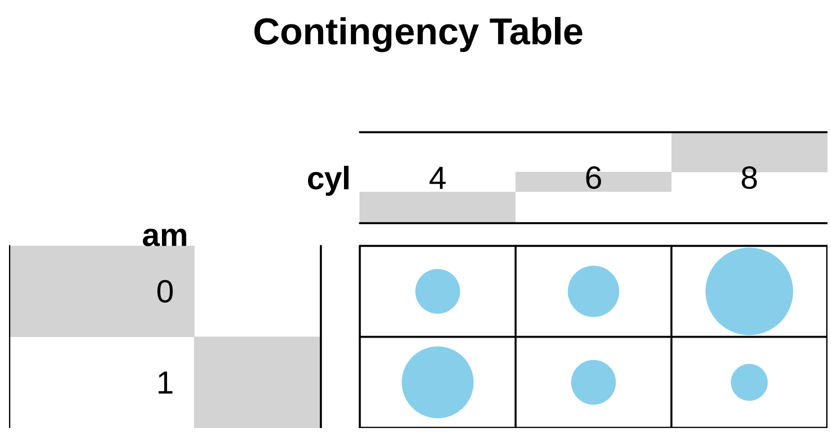
2. Creating a Contingency Table:

```r
ctab <- table(mtcars$am, mtcars$cyl)
ctab
```

```
   4  6  8
0  3  4 12
1  8  3  2
```

3. Graphical display of Contingency table:

```r
library("gplots")
# 1. convert the data as a table
dt <- as.table(as.matrix(ctab))
# 2. Graph
balloonplot(t(dt), main ="Contingency Table",
            xlab ="cyl", ylab="am",
            label = FALSE, show.margins = FALSE)
```



4. Compute Chi-Square test: The Chi-square statistic can be easily computed using the function `chisq.test()` as follow:

```r
chisq <- chisq.test(ctab)
```

```
Warning in chisq.test(ctab): Chi-squared approximation may be incorrect
```

```r
chisq
```

```
    Pearson's Chi-squared test

data:  ctab
X-squared = 8.7407, df = 2, p-value = 0.01265
```

5. In our example, the row and the column variables are statistically significantly associated (p-value = 0.01265).

6. The observed and the expected counts can be extracted from the result of the test as follows:

```r
# Observed counts
chisq$observed
```

```
   4  6  8
0  3  4 12
1  8  3  2
```

```r
# Expected counts
round(chisq$expected, 2)
```

```
     4    6    8
0 6.53 4.16 8.31
1 4.47 2.84 5.69
```

7. Pearson Residuals: **Positive residuals** are positive values in cells specify an attraction (positive association) between the corresponding row and column variables. **Negative residuals** implies a repulsion (negative association) between the corresponding row and column variables.

```r
round(chisq$residuals, 3)
```

```
       4      6      8
0 -1.382 -0.077  1.279
1  1.670  0.093 -1.546
```