* Chi-squared test for goodness of fit and as an independence test
* ANOVA
* Correlations

1. A survey is conducted by a gaming company that makes three video games. It wants to know if the preference of game depends on the gender of the player. Total number of participants is 1000. Here is the survey result. (Hint: A case of independence test)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Game A | Game B | Game C | Total |
| Male | 200 | 150 | 50 | 400 |
| Female | 250 | 300 | 50 | 600 |
| Total | 450 | 450 | 100 | 1000 |

tb<-matrix(c(200,150,50,250,300,50),c(2,3),byrow=T)  
chisq.test(tb)

1. A national survey agency conducts a nationwide survey on consumer satisfaction and finds out the response distribution as follows:

Excellent: 8%

Good: 47%

Fair: 34%

Poor: 11%

A store manager wants to find if these results of customer survey apply to customers of super market in her city. So, she interviews 207 randomly selected customers and asked them to rate their responses. The results of this local survey are:

|  |  |
| --- | --- |
| Response | Frequency |
| Excellent | 21 |
| Good | 109 |
| Fair | 62 |
| Poor | 15 |

Determine if the local responses from this survey are the same as expected frequencies of the national survey, at 95% significance.

**#Chi-squared for good-ness of fit**

originalFrequencies<- c(21,109,62,15)

expectedProportions<- c(0.08,0.47,0.34,0.11) #given

expectedFrequencies<- c(16.56,97.29,70.38,22.77) #calculated based on proportions and observed frequencies

chisq.test(originalFrequencies,p=expectedFrequencies/sum(expectedFrequencies))

# Chi-squared test for given probabilities

# data: original

# X-squared = 6.2491, df = 3, p-value = 0.1001

OR

chisq.test(originalFrequencies, p = expectedProportions)

Therefore, we do not have enough evidence to reject null hypothesis.

1. A car crash research team wants to examine the safety of compact cars, intermediate and full size cars. Given below are the hypothetical values of the mean pressure applied to the drivers head during the crash test for each of the car types. Check whether means are equal for each type of these cars.

|  |  |  |  |
| --- | --- | --- | --- |
| Compact | 643 | 655 | 702 |
| Intermediate | 469 | 427 | 525 |
| Full size | 484 | 456 | 402 |

Null Hypothesis: Means are equal for the three cars

Alternate Hypothesis: At least one mean is statistically different

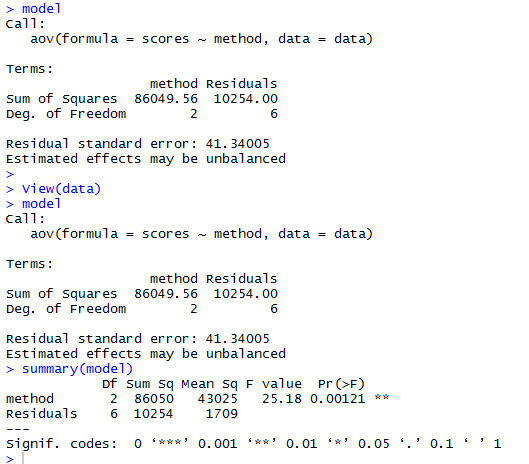
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | σ |
| Compact | 643 | 655 | 702 | 666.67 | 31.18 |
| Intermediate | 469 | 427 | 525 | 473.67 | 49.17 |
| Full Size | 484 | 456 | 402 | 447.33 | 41.68 |

**In R**

data <- data.frame(scores = c(643,655,702,469,427,525,484,456,402),method = factor(rep(c("M1","M2","M3"),c(3,3,3))))

model= aov(scores~method,data=data)

model



1. Given below is the number of cups of coffee ordered in a restaurant in a week. Do the numbers show that people prefer any one coffee over the other?

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **Blue Label** | **Green Label** | **Red Label** |
| 3 | 6 | 9 |
| 5 | 7 | 10 |
| 6 | 9 | 15 |
| 2 | 7 | 12 |
| 1 | 11 | 11 |
| 2 | 6 | 10 |

#Similar to question 4

1. Find the covariance of the eruption duration and waiting time in the data set “faithful” (built-in dataset in R). Observe if there is any linear relationship between the two variables.

View(faithful)

duration = faithful$eruptions # the eruption durations

waiting = faithful$waiting # the waiting period

Covariance

cov(duration, waiting) # apply the cov function

#The covariance of the eruption duration and waiting time is 13.978.

#It indicates a positive linear relationship between the two variables.

#This relation could be observed from the scatter plot of waiting vs duration

plot(faithful$eruptions,faithful$waiting)

Correlation

 cor(duration, waiting)          # apply the cov function   
  
#The correlation coefficient of the eruption duration and waiting time is 0.90081. #Since it is close to 1, we can conclude that the variables are positively linearly #related.

1. Analyzing the linear relation among the attributes in the “Cereals” dataset
   1. compute covariance and correlations on the data
   2. write it to a file
   3. plot the correlations/covariance and obtain the pairs of attributes that are highly correlated

#Similar to question 5