- ✓ 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)
```

2. 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.

3. 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.

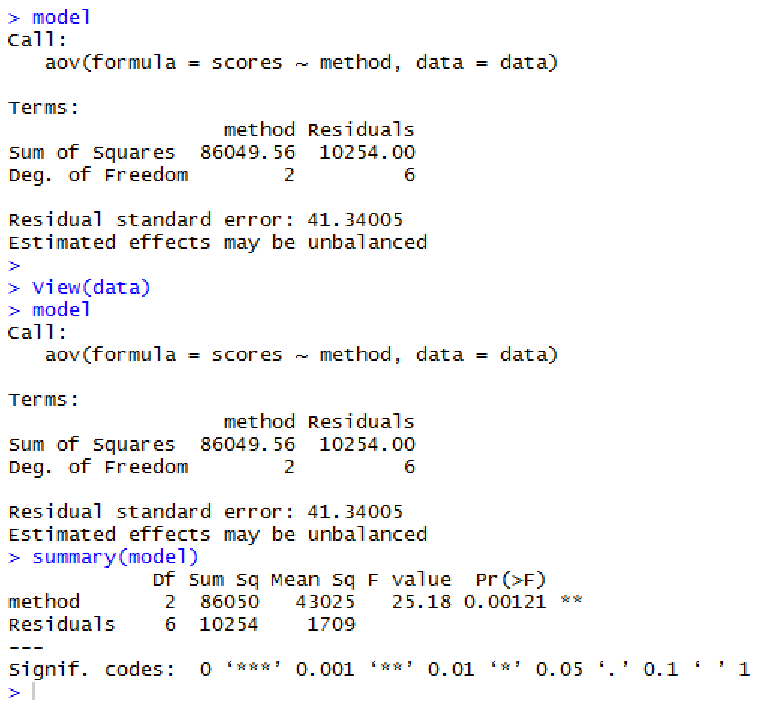
Compact	643	655	702
Intermediate	469	427	525
Full size	484	456	402

Null Hypothesis: Means are equal for the three cars
Alternate Hypothesis: At least one mean is statistically different

				$\bar{X}$	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
```



```
> model
Call:
   aov(formula = scores ~ method, data = data)

Terms:
                  method Residuals
Sum of Squares  86049.56  10254.00
Deg. of Freedom        2         6

Residual standard error: 41.34005
Estimated effects may be unbalanced
>
> View(data)
> model
Call:
   aov(formula = scores ~ method, data = data)

Terms:
                  method Residuals
Sum of Squares  86049.56  10254.00
Deg. of Freedom        2         6

Residual standard error: 41.34005
Estimated effects may be unbalanced
> summary(model)
            Df Sum Sq Mean Sq F value  Pr(>F)
method       2  86050   43025   25.18 0.00121 **
Residuals    6  10254    1709
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

4. 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

5. 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.

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

   #Similar to question 5

7. Suppose that a random sample of n = 5 was selected from the vineyard properties for sale in Sonoma County, California, in each of three years. The following data are consistent with summary information on price per acre for disease-resistant grape vineyards in Sonoma County. Carry out an ANOVA to determine whether there is evidence to support the claim that the mean price per acre for vineyard land in Sonoma County was not the same for each of the three years considered. Test at the 0.05 level and at the 0.01 level.

   1996: 30000 34000 36000 38000 40000
   1997: 30000 35000 37000 38000 40000
   1998: 40000 41000 43000 44000 50000

   Ans: Similar to Q4

8. A business owner had been working to improve employee relations in his company. He predicted that he met his goal of increasing employee satisfaction from 65% to 80%. Employees from four departments were asked if they were satisfied with the working conditions of the company. The results are shown in the following table:

	Finance	**Sales**	**Human Resources**	**Technology**
Satisfied	12	38	5	8
Dissatisfied	7	19	3	1
Total	19	57	8	9

Our first step is to calculate the predicted values so that we can compare them to the actual values from the survey. The predicted number of satisfied employees is 80% of the total number of employees in each department. This leaves the remaining 20% as the number of dissatisfied employees. For example, the predicted number of satisfied employees in the finance department is 0.80(19) = 15.2. The predicted number of dissatisfied employees in the finance department is 0.20(19) = 3.8. The following table shows the observed and expected values for each department. The observed values are in bold and the expected values are in parentheses.

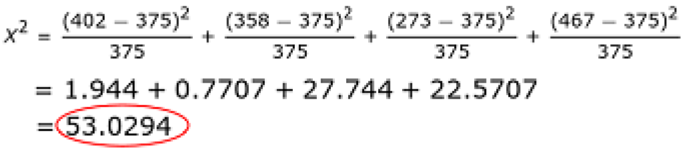
	Finance	**Sales**	**Human Resources**	**Technology**
Satisfied	**12**(15.2)	**38**(45.6)	**5**(6.4)	**8**(7.2)
Dissatisfied	**7**(3.8)	**19**(11.4)	**3**(1.6)	**1**(1.8)
Total	19	57	8	9

$$\begin{aligned}
x^2 &= \frac{(12-15.2)^2}{15.2} + \frac{(38-45.6)^2}{45.6} + \frac{(5-6.4)^2}{6.4} + \frac{(8-7.2)^2}{7.2} \\
&\quad + \frac{(7-3.8)^2}{3.8} + \frac{(19-11.4)^2}{11.4} + \frac{(3-1.6)^2}{1.6} + \frac{(1-1.8)^2}{1.8} \\
&= 0.6767 + 1.2667 + 0.3063 + 0.0889 \\
&\quad + 2.6947 + 5.0667 + 1.2250 + 0.3556 \\
&= \boxed{11.6806}
\end{aligned}$$

9. Many casinos use card-dealing machines to deal cards at random. Occasionally the machine is tested to ensure an equal likelihood of dealing for each suit. To conduct the test, 1,500 cards are dealt from the machine while the number of cards in each suit is counted. Theoretically, 375 cards should be dealt from each suit. As you can see from the results in the table, this is not the case:

We can use chi square to determine if the discrepancies are significant. If the discrepancies are significant, then the game would not be fair. Measures would need to be taken to ensure that the game is fair.

	Spades	Diamond	Clubs	Hearts
Observed	402	358	273	467
Expected	375	375	375	375



$$x^2 = \frac{(402-375)^2}{375} + \frac{(358-375)^2}{375} + \frac{(273-375)^2}{375} + \frac{(467-375)^2}{375}$$

$$= 1.944 + 0.7707 + 27.744 + 22.5707$$

$$= \boxed{53.0294}$$

10. In the dataset *mtcars*, is there significant difference in mileage across no. of cylinders

    Ans: aov(data=mtcars, mpg~cyl)

11. In the dataset ‘airquality’, manually compute covariance and correlation of ‘Solar.R’ & ‘Temp’

    Covariance:

    Sum((Solar.R-mean(Solar.R))*(Temp-mean(Temp)))/(n-1)

12. In the dataset ‘pressure’, manually compute covariance and correlation of ‘temperature’ & ‘pressure’.

    Covariance:

    Sum((temperature-mean(temperature))*(pressure-mean(pressure)))/(n-1)