**For all problems, show the code, required to answer each of the questions, in blue, and corresponding R output in black. Use copy & paste function in R environment to copy and paste the code and the output.**

**ASSIGNMENT in R (part 1 – Analysis of categorical data)**

**Problem 1** **(10 points)**

Nearly 75% of all death in the United States are attributed to just 10 causes of non - accidental deaths, with the top **four** of these accounting for **56.6 %** of all non -accidental deaths as follows: heart disease (23.4%), cancer (22.5%), respiratory disease (5.6%), and stroke (5.1%).

The rest **43.4%** of all non -accidental deaths are attributed to other diseases.

The study of the causes of n =308 non – accidental deaths at a local hospital gave the following counts.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cause | Heart  disease | Cancer | Respiratory  disease | Stroke | Other  diseases |  |
| Deaths | 78 | 81 | 28 | 16 | 105 | 308 |
|  |  |  |  |  |  |  |

Do the data provide sufficient information to indicate that the number of deaths at this hospital differ significantly from the proportion in the population in large?

Test appropriate hypothesis at 5% level of significance using R. Show the code, output, and make the conclusions based on obtained p -value.

* Null Hypothesis 🡪 The number of deaths at this hospital does not differ significantly from the proportion in the population in large.
* Alternative Hypothesis 🡪 The number of deaths at this hospital differ significantly from the proportion in the population in large.
* **Code: > chisq.test(as.table(c(78,81,28,16,105)),p = c(0.234,0.225,0.056,0.051,0.434))**
* **Output: Chi-squared test for given probabilities**
  + **data: as.table(c(78, 81, 28, 16, 105))**
  + **X-squared = 15.321, df = 4, p-value = 0.00408**
* Since p-value: 0.004<0.05, there is sufficient information to indicate that the number of deaths at this hospital differ significantly from the proportion in the population in large. Thus, null hypothesis is rejected, and alternative hypothesis has been accepted.

**Problem 2 (10 points)**

Accident data were analyzed to determine the number of fatal accidents for the cars of three sizes. The data for 346 accidents are as follows.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Small | Medium | Large |
| Fatal | 67 | 26 | 16 |
| Not Fatal | 128 | 63 | 46 |

Do the data indicate that the frequency of fatal accidents is dependent on the size of the cars? Test appropriate hypothesis using R. Assume 0.05 level of significance.

Show the code, output, and make the conclusions based on obtained p -value.

* Null Hypothesis 🡪 The frequency of fatal accidents is not dependent on the size of the cars.
* Alternative Hypothesis 🡪 The frequency of fatal accidents is dependent on the size of the cars.
* **Code: x1=c(67,26,16)**
  + **> x2=c(128,63,46)**
  + **> df=(data.frame(x1,x2))**
  + **> chisq.test(df)**
* **Output: Pearson's Chi-squared test**
  + **data: df**
  + **X-squared = 1.8857, df = 2, p-value = 0.3895**
* Since the p-value is 0.3895>0.05, the test is not statistically significant and null hypothesis can be accepted with alternative hypothesis being rejected. Thus, there’s insufficient evidence that the frequency of fatal accidents is dependent on the size of the cars here.

**Problem 3 (10 points)**

An engineer of a company wanted to obtain the information about the power failures in a small town. He recorded the number of power failures per day for the period of 102 consecutive days.

He has found that there were 24 days with no failures, 30 days with 1 failure, 31 day with 2 failures,

11 days with 3 failures, and 6 days with 4 or more failures.

The data obtained by the engineer are given in the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Number of power failures  per day | 0 | 1 | 2 | 3 | 4 or more |
| Number of days with given number  of power failures per day | 24 | 30 | 31 | 11 | 6 |
| Expected frequencies |  |  |  |  |  |

Test at the 0.05 level of significance whether the number of power failures in the town per day is a random variable having Poisson distribution with the value of parameter  estimated from the sample data. Use R to answer questions a. – d.

1. Show that estimated value of is 1.46.

**Code: > (0\*24+1\*30+2\*31+3\*11+4\*6)/102**

**Output: [1] 1.460784**

1. Find the vector p of the values of Poisson probability mass function at the values of

X = 0, 1, 2, 3, 4 or more.

**Code: > p = c(dpois(0,1.460),dpois(1,1.460),dpois(2,1.460), dpois(3, 1.460),dpois(4, 1.460),(1-(dpois(0,1.460) + dpois(1,1.460) + dpois(2,1.460) + dpois(3, 1.460) + dpois(4,1.430)))>p**

**Output: [1] 0.23, 0.34, 0.25, 0.12, 0.04**

1. Check that all expected frequencies are no less than 5, and there is no need to combine the cells.
2. Test the hypothesis using R that the number of power failures in the town per day is a random variable having Poisson distribution with the value of parameter = 1.46.

Use chi - square test. Assume 0.05 level of significance.

Show the code, output, and make the conclusions based on obtained p -value.

* Null Hypothesis 🡪 The number of power failures in town per day is not a random variable having Poisson distribution with a value of parameter 𝞴=1.46.
* Alternative Hypothesis 🡪 The number of power failures in town per day is a random variable having Poisson distribution with a value of parameter 𝞴=1.46.
* **Code: > chisq.test(as.table(c(24,30,31,11,6)), p =c(0.25,0.34,0.25,0.12,0.04))**
* **Output: [1] Chi-squared test for given probabilities**
  + **data: as.table(c(24, 30, 31, 11, 6))**
  + **X-squared = 2.490, df = 4, p-value = 0.646**
* Since the p-value is 0.646>0.05, the test is not statistically significant and null hypothesis can be accepted with alternative hypothesis being rejected. Thus, we can conclude that Poisson distribution is appropriate.

**Assignment in R – Part 2 (ANOVA)**

**Problem 4.** **One - way ANOVA, completely randomized design (10 points)**

Four chemical plants, producing the same products and owned by the same company, discharge effluents into streams in the vicinity of their locations. To monitor the extent of pollution created by the effluents and to determine whether this differs from plant to plant to plant, the company collected random samples of liquid waste, five specimens from each plant. The data are given in the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Polluting Effluents  (lb/gal of waste)  Plant | 1 | 2 | 3 | 4 | 5 |
| A | 1.65 | 1.72 | 1.50 | 1.37 | 1.60 |
| B | 1.70 | 1.85 | 1.46 | 2.05 | 1.80 |
| C | 1.40 | 1.75 | 1.38 | 1.65 | 1.55 |
| D | 2.10 | 1.95 | 1.65 | 1.88 | 2.00 |

Use R to answer the following questions.

Do the data provide sufficient sample evidence to indicate a difference in the mean weight of effluents discharged from the four plants? Test appropriate hypothesis using .

* **H0 = means are equal against the alternative hypothesis 🡪 µ1 = µ2 = µ3 = µ4**
* **H1= means are not equal 🡪 µ1 ≠ µ2 ≠ µ3 ≠ µ4**

1. If there are the differences in the mean weights of effluents per gallon in the effluents

discharged from the four plants, indicate which means differ.

Use Tukey’s procedure for multiple comparisons, and draw the diagram for population means.

For this problem you have to present:

* the code required to obtain ANOVA table
* ANOVA table (R output)
* conclusion (based on obtained p -value) about the truth or falsity of null hypothesis of equality of population means for populations of polluting effluents in five different locations.
* the code required to obtain the results of Tukey’s comparisons
* the result of Tukey’s comparisons (R output)
* the diagram for population means of weights of effluents for the plants A, B, C, and D.

To draw the diagram, you have to find all sample means (mean weights of effluents per gallon for each of four plants) and place the sample means in ascending order. Place the population means muA, muB, muC, and muD in the same order. Use p - values obtained as the result of Tukey’s HSD test to decide which population means may be considered as not significantly different. Underline these population means on the diagram.

* **Code:** 
  + **value<c(1.4,1.65,1.7,2.1,1.72,1.75,1.85,1.95,1.38,1.46,1.5,1.65,1.37,1.65,1.88,2.05,1.55,1.6,1.8,2)**
  + **> group1 <- c(rep("x1", 4), rep("x2", 4), rep("x3", 4), rep("x4", 4), rep("x5", 4))**
  + **> my.dataframe<-data.frame(value, group1)**
  + **>** **res.aov <- aov(value ~ group1, data = my.dataframe)**
  + **>** **summary(res.aov)**
* **Output:**

**[1] Df Sum Sq Mean Sq F value Pr(>F)**

**group1 4 0.2311 0.05778 1.22 0.344**

**Residuals 15 0.7106 0.04737**

Since p-value is larger than the significance level of 0.05, H0 is accepted and the averages of all the groups are assumed to be equal. Hence, the difference between these averages is not large enough to be statistically significant.

* Effect of size:
  + **Code: >coeff(res.aov)**
  + **Output** 
    - **[1] (Intercept) group1x2 group1x3 group1x4**

**1.7125 0.1050 -0.2150 0.0250**

**group1x5**

**0.0250**

* + Since the effect observed effect size f is large (1.71). That indicates that the magnitude of the difference between the averages is large.
  + The η2 equals 0.25. It means that the group explains 24.5% of the variance from the average (similar to R2 in the linear regression).

**Problem 5. One - way ANOVA, randomized block design. (10 points)**

Do average automobile insurance costs differ for different insurance companies?

The following are estimates (in dollars) of the cost of 6 -month polices for basic liability coverage for a person with no violations or accidents, who drives between 12,600 and 15,000 miles per year, and who lives in one of indicated areas.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Location  Company | Riverside  1 | San Bernardino  2 | Hollywood  3 | Long Beach  4 |
| 21 st Century A | 736 | 836 | 1492 | 996 |
| Allstate B | 745 | 725 | 1384 | 884 |
| AAA C | 668 | 618 | 1214 | 802 |
| Fireman’s Fund D | 1065 | 869 | 1502 | 1571 |
| State Farm E | 1202 | 1172 | 1682 | 1272 |

Use R to answer the following questions.

1. Do the data provide sufficient sample evidence to indicate that average insurance premiums differ from a company to company? Test appropriate hypothesis using .
2. If there are the differences in the mean insurance premiums for different insurance

companies, indicate which means differ.

Use Tukey’s procedure for multiple comparisons, and draw the diagram.

1. Is there sufficient sample evidence to indicate that insurance premiums differ from location to location? Test appropriate hypothesis using .

For this problem you have to present

* the code required to obtain ANOVA table
* ANOVA table (R output)
* conclusion about the truth or falsity of null hypothesis about the equality of population means for insurance premiums in five different companies, and conclusion about the truth or falsity of null hypothesis about the equality of population means for insurance premiums in four different locations.
* the code required to obtain the results of Tukey’s comparisons
* the result of Tukey’s comparisons (R output) for companies.
* the diagram for population means of insurance premiums for 5 companies (A, B, C, D and E).

To draw the diagram, you have to find all sample means (mean insurance premiums for 5 companies) and place the sample means in ascending order. Place the population means muA, muB, muC, muD, and muE in the same order. Use p -values obtained as the result of Tukey’s HSD test to decide which population means may be considered as not significantly different.

Underline these population means on the diagram.

**Problem 6. Two - way ANOVA, completely randomized design (10 points)**

The data that follow are the ratings for six smart phones from each of four suppliers, three of which cost $650 or more and three of which cost less than $650.

The ratings have a maximum value of 100 and a minimum of 0.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Supplier  Cost | A | B | C | D |
| $650 or more **LargeCost** | 76  74  69 | 74  69  68 | 72  71  71 | 75  73  73 |
| Less than$ 650  **SmallCost** | 69  67  64 | 69  64  60 | 71  71  70 | 72  71  70 |

1. What type of experiment was used to obtain the ratings for smartphones?

What are factors? What are the levels of each factor?

1. Produce an analysis of variance table appropriate for this design using R.
2. Is there significant interaction between the two factors?
3. Is there a main effect due to suppliers?
4. Is there a main effect due to cost? For c- e, test appropriate hypothesis using .

For this problem you have to present

* the answers for question a.
* the code required to obtain ANOVA table and analysis of variance table (R output)
* conclusion about the truth of null hypothesis of no interaction between the factors
* conclusion about the truth or falsity of null hypothesis about the equality of population means for two different types of costs and null hypothesis of equality of population means for rating of four different suppliers.
* the code required to obtain the results of Tukey’s comparisons
* the result of Tukey’s comparisons (R output) for the ratings of four different suppliers.
* the diagram for population means of ratings for 4 supliers (A, B, C, and D).

To draw the diagram, you have to find all sample means (mean ratings for 4 suppliers) and place the sample means in ascending order. Place the population means muA, muB, muC, and muD in the same order. Use p -values obtained as the result of Tukey’s HSD test to decide which population means may be considered as not significantly different. Underline these population means on the diagram.

**ASSIGNMENT in R (part 3 – SIMPLE and MULTIPLE REGRESSION)**

**Problem 7. Simple linear regression (10 points)**

|  |  |  |
| --- | --- | --- |
| Student | Mathematics Achievement Test Score | Final calculus grade |
| 1 | 39 | 65 |
| 2 | 43 | 78 |
| 3 | 21 | 52 |
| 4 | 64 | 82 |
| 5 | 57 | 92 |
| 6 | 47 | 89 |
| 7 | 28 | 73 |
| 8 | 75 | 98 |
| 9 | 34 | 56 |
| 10 | 52 | 75 |

Use R to answer the following questions.

a. Introduce the vectors **testscore** and **examgrade.**

Introduce the object with the name **gradeversusscore,** and assign the outputof **lm(examgrade~testscore)** function to this object**.**

Using the output of **summary(gradeversusscore)** function **write the equation of estimated regression line** and **decide whether the data provide**

**sufficient evidence to indicate that model contributes important**

**information for the prediction of final exam grades?**

1. Construct the scatter plot for given data set. Include the line of best fit.
2. Find a 95% confidence interval (use interval notation to report the answer) for the slope for the calculus grade data.

Show the code (in blue) and output (in black) for parts a, b, and c.

Insert the constructed plot using Plots> CopyClipboard in R studio.

**Problem 8. Multiple linear regression (15 points)**

The price for a newly listed condominium asked by real estate agents may depend on many factors: the living area, number of floors, number of bedrooms, number of bathrooms. It’s assumed that the factors listed above are the main factors, and multiple regression model to be fitted is, where

 = living area in hundreds of square meters,

= number of floors,

= number of bedrooms,

= number of bathrooms

To fit multiple regression model the data set of 15 randomly selected condominiums currently on the market is used. The resulting data set in presented in the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Observation | List price, y  (in $1000) | Living area,  (in hundreds of sq. ft) | Floors, | Bedrooms, | Bathrooms, |
| 1 | 169.0 | 6 | 1 | 2 | 1 |
| 2 | 218.5 | 10 | 1 | 2 | 2 |
| 3 | 216.5 | 10 | 1 | 3 | 2 |
| 4 | 225.0 | 11 | 1 | 3 | 2 |
| 5 | 229.9 | 13 | 1 | 3 | 1.7 |
| 6 | 235.0 | 13 | 2 | 3 | 2.5 |
| 7 | 239.9 | 13 | 1 | 3 | 2 |
| 8 | 247.9 | 17 | 2 | 3 | 2.5 |
| 9 | 260.0 | 19 | 2 | 3 | 2 |
| 10 | 269.9 | 18 | 1 | 3 | 2 |
| 11 | 234.9 | 13 | 1 | 4 | 2 |
| 12 | 255.0 | 18 | 1 | 4 | 2 |
| 13 | 269.9 | 17 | 2 | 4 | 3 |
| 14 | 294.5 | 20 | 2 | 4 | 3 |
| 15 | 309.9 | 21 | 2 | 4 | 3 |

Use R output to answer the following questions

1. Do these data provide sufficient evidence to conclude at the 5% significance level that the model is useful in predicting the list price?
2. Do these data provide sufficient evidence to conclude at the 5% significance level that the list price and the living area are linearly related?
3. Do these data provide sufficient evidence to conclude at the 5% significance level that the list price and the number of the bedrooms are linearly related?

Show the code (in blue) and output (in black) for parts a, b, and c.

Use p-values in output to answer the questions a, b, and c.

**Problem 9. Polynomial regression (15 points)**

In a study of variables that affect productivity in the retail grocery trade, the researcher uses value added per work – hour to measure productivity of retail grocery stores outlets. The value added per work hour is defined as the money generated by the business available to pay for labor, furniture, and equipment. Data consistent with relationship between value added per work – hour y and the size of a grocery outlet are given in the table below for 10 stores. Data points are represented on the plot below the table.

|  |  |  |
| --- | --- | --- |
| Store | Value added per  work – hour, y (in $) | Size of store x (in 1000 **square feet**) |
| 1 | 4.08 | 21.0 |
| 2 | 3.40 | 12.0 |
| 3 | 3.51 | 25.2 |
| 4 | 3.09 | 10.4 |
| 5 | 2.92 | 30.9 |
| 6 | 1.94 | 6.8 |
| 7 | 4.11 | 19.6 |
| 8 | 3.16 | 14.5 |
| 9 | 3.75 | 25.0 |
| 10 | 3.60 | 19.1 |

Use R output to answer the following questions.

1. Do the data provide sufficient evidence to indicate that the model contributes important information for the prediction of *y*?
2. Does the quadratic term contribute important information in the prediction

of y in the presence of linear term?

1. Use R to construct quadratic regression curve.

Show the code (in blue) and output (in black) for parts a, b, and c.

Use p-values in output to answer the questions a and b.

Insert the constructed plot using Plots> CopyClipboard in R studio