Module 2-Chi Square and ANOVA

ALY6015-Intermediate Analytics

NORTHEASTERN UNIVERSITY

DEVIKA PATIL
MAJOR-PROJECT MANAGEMENT
DATE OF SUBMISSION: 03-05-2024
Prof. Zhi He

Contents

Introduction 3

Analysis 3

- Task 1 3
- Task 2 4
- Task 3 5
- Task 4 7
- Task 5 9
- Task 6 10
- Task 7 11
- Task 8 12
- Task 9.1 (Q1,2 & 3) 14
- Task 9.2 (Q4 & 5) 19
- References: 22

Introduction

The following analysis introduces two statistical tests: the Chi-Square test and ANOVA. We'll learn how to use them in R software and interpret their results for specific tasks.

The Chi-Square test helps determine if two categorical variables (e.g., colors, genders) are related. It comes in three forms: goodness-of-fit, independence, and homogeneity tests.

ANOVA, on the other hand, checks if the average values (means) differ significantly across two or more groups. It helps us understand factors affecting a single variable's variation. In R, ANOVA is used to see if group means are equal. It's like a t-test, but to compare multiple groups simultaneously. Three main types of ANOVA: one-way test, pairwise comparisons test, and two-way test.

Analysis

Task 1

Blood types in hospital

Hypothesis

1.Null Hypothesis (H0): The proportions of individuals in groups A, B, and O, as well as the proportion of individuals belonging to both groups A and B, are equal to 0.20, 0.28, 0.36, and 0.16, respectively.
2. Alternative Hypothesis (H1): At least one of the following statements is true:

- The proportion of individuals in group A is different from 0.20.
- The proportion of individuals in group B is different from 0.28.
- The proportion of individuals in group O is different from 0.36.
- The proportion of individuals belonging to both groups A and B is different from 0.16.

Critical value

```
> #TASK1
> BPDT = c(0.20, 0.28,0.36, 0.16)
> RS= c(12, 8,24, 6)
> a_val = 0.10
> degreesOfFreedom = 3
> critValue <- qchisq(1 - a_val, df = degreesOfFreedom)
> print(paste("Critical_Value = ",critValue))
[1] "Critical_Value =  6.25138863117032"
```

Compute the test value

```
> #computing values
> a_val = 0.10
> BT_result = chisq.test(x = RS, p = BPDT, correct = FALSE)
> BT_result

        Chi-squared test for given probabilities

data:  RS
X-squared = 5.4714, df = 3, p-value = 0.1404
```

The decision

```
> if (BT_result$statistic > critValue) {
+   print("Reject H0")
+ } else {
+   print("Fail to reject H0")
+ }
[1] "Fail to reject H0"
```

Observation: Upon conducting hypothesis testing, we get the statistics value as 5.47, it's apparent that the resulting test value is less that the critical value (6.25) due to which we fail to reject (H_0) Null Hypothesis.

Task 2

Airline Performance

Hypothesis

1.Null hypothesis (H_0): All parameters are equal to: P(on.time) = 0.708, P(system.delay) = 0.082, P(arriving.late) = 0.09, P(other) = 0.12

2.Alternative hypothesis (H_1): At least one parameter is different from the values specified in (H_0) null hypothesis.

Critical value :

```
> #TASK2
> actions_perform = c("On time","National Aviation System Delay","Aircraft Arriv
> government_stats = c(0.708, 0.082, 0.09, 0.12)
> random_sample_2 = c(125,10, 25, 40)
> #critical values
> a_val = 0.05
> degreesOfFreedom = 3
> critValue <- qchisq(1 - a_val, df = degreesOfFreedom)
> print(paste("Critical_Value = ",critValue))
[1] "Critical_Value =  7.81472790325118"
```

Compute the test values

```
> #computing values
> a_val = 0.05
> t2_result = chisq.test(x = random_sample_2, p = government_stats)
> t2_result

	Chi-squared test for given probabilities

data:  random_sample_2
X-squared = 17.832, df = 3, p-value = 0.0004763
```

The decision

```
> if (t2_result$statistic > critValue) {
+   print("Reject H0")
+ } else {
+   print("Fail to reject H0")
+ }
[1] "Reject H0"
```

Observation: The null hypothesis (H0) proposes that the proportions of on-time arrivals, National Aviati on System delays, late aircraft arrivals, and other delays are 0.70, 0.08, 0.9, and 0.12 respectively. Any d eviation from this distribution is considered an "alternative hypothesis." The Chi-square test conducted on these probabilities yields a statistic of 17.83 with a p-value of 0.00047 and 3 degrees of freedom. The refore, with a significance level 0.05, we reject (H_0) Null hypothesis.

Task 3

Ethnicity and Movie Admissions

Hypothesis

1. Null hypothesis: Movie admission rates are the same for all ethnicities.
2. Alternative hypothesis: Movie admission rates differ across different ethnicities.

Critical value

```
> #TASK3
> ethnicities = c("Caucasian", "Hispanic", "African American", "Other")
> population2014 = c(370, 292, 152, 140)
> population2013 = c(724, 335, 174, 107)
> #critical values
> a_val = 0.05
> degreesOfFreedom = 3
> critValue <- qchisq(1 - a_val, df = degreesOfFreedom)
> print(paste("Critical_Value = ",critValue))
[1] "Critical_Value =  7.81472790325118"
```

Compute the test value

```
> #computing values
> a_val = 0.05
> m_3 = matrix(c(population2013, population2014), nrow = 2,
> rownames(m_3) = c("2013", "2014")
> colnames(m_3) = c("Caucasian", "Hispanic", "African Amer
> m_3
     Caucasian Hispanic African American Other
2013       724      335              174   107
2014       370      292              152   140
```

```
> t3_result = chisq.test(m_3)
> t3_result

	Pearson's Chi-squared test

data:  m_3
X-squared = 60.144, df = 3, p-value = 5.478e-13
```

The decision

```
> if (t3_result$statistic > critValue) {
+   print("Reject H0")
+ } else {
+   print("Fail to reject H0")
+ }
[1] "Reject H0"
```

Observation: The null hypothesis states that ethnicity affects movie admission statistics, implying that a dmission to movies is contingent on ethnicity (H_0). Conversely, the alternative hypothesis proposes that admission to movies is unrelated to ethnicity (H_1). The test performed results in p-value of 5.478e-13 an d statistical (X-squared) value of 60.14 which is greater than 7.81 (Critical Value) . Hence, at alpha 0.05 w e reject (H_0) Null Hypothesis

Task 4

Women in Military

Hypothesis

1.Null Hypothesis (H_0): There lies a relationship between rank and branch of the Armed Forces for women in the military.
2. Alternative Hypothesis (H_1): There lies no relationship between rank and branch of the Armed Forces for women in the military.

Critical value:

```
> #Task4
> armyList = c(10791, 62491)
> navyList = c(7816, 42750)
> marineCorpsList = c(932, 9525)
> airForceList = c(11819, 54344)
> #critical values-4
> a_val = 0.05
> degreesOfFreedom = 3
> critValue <- qchisq(1 - a_val, df = degreesOfFreedom)
> print(paste("Critical_Value = ",critValue))
[1] "Critical_Value =  7.81472790325118"
```

Compute the test values

```
> #computing values
> a_val = 0.05
> m4 = matrix(c(armyList, navyList, marineCorpsList, airForceList), ncol = 2, nrow = 4, byrow = TRUE)
> colnames(m4) = c("Officers", "Enlisted")
> rownames(m4) = c("Army", "Navy", "Marine Corps", "Air Force")
> m4
             Officers Enlisted
Army            10791    62491
Navy             7816    42750
Marine Corps      932     9525
Air Force       11819    54344
```

```
> t4_result = chisq.test(m4)
> t4_result

        Pearson's Chi-squared test

data:  m4
X-squared = 654.27, df = 3, p-value < 2.2e-16
```

Decision

```
> if (t4_result$statistic > critValue) {
+   print("Reject H0")
+ } else {
+   print("Fail to reject H0")
+ }
[1] "Reject H0"
```

Observation: Examining the relationship between branch and rank for women in the military, the analysis initially assumed a connection (H0), while the alternative hypothesis (H1) proposed no such link. The Chi-square test, with low p-value (< 2.2e-16), indicates statistically significant relationship at alpha 0.05. Consequently, we reject (H_0)null hypothesis, which suggests that branch and rank aren't randomly distributed for women in the military, meaning they don't experience equal representation across ranks within different branches.

Task 5

Sodium Contents of Foods

Hypothesis

1. Null hypothesis(H_0): All three groups have the same average value (mean)

2. Alternative hypothesis (H_1): At least 1 of 3 groups have a different average value (mean) compared to the others.

```
> a_val = 0.05
> con_df = data.frame("sodium_col" = c(270, 130, 230, 180, 80, 70, 200), "food_col" = rep("condiments",
7), stringsAsFactors = FALSE)
> cer_df = data.frame("sodium_col" = c(260, 220, 290, 290, 200, 320, 140), "food_col" = rep("cereals",
7), stringsAsFactors = FALSE)
> des_df = data.frame("sodium_col" = c(100, 180, 250, 250, 300, 360, 300, 160), "food_col" = rep("dessert
s", 8), stringsAsFactors = FALSE)
> sod_df = rbind(con_df, cer_df, des_df)
> sod_df$food_col = as.factor(sod_df$food_col)
> t5_aov = aov(sodium_col ~ food_col , data = sod_df)
> t5_aov_summary = summary(t5_aov)
> t5_aov_summary
            Df Sum Sq Mean Sq F value Pr(>F)
food_col     2  27544   13772   2.399  0.118
Residuals   19 109093    5742
```

```
> colName = "Pr(>F)"
> pVal_5 = t5_aov_summary[[1]] [[1, colName]]
> pVal_5
[1] 0.1178108
```

#Comparison between p and alpha value

```
> ifelse ( pVal_5 > a_val , "Fail to reject H0", "Reject H0")
[1] "Fail to reject H0"
```

Tukey test

```
> TukeyHSD(t5_aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = sodium_col ~ food_col, data = sod_df)

$food_col
                          diff        lwr       upr     p adj
condiments-cereals  -80.000000 -182.89588  22.89588 0.1456674
desserts-cereals     -8.214286 -107.84279  91.41422 0.9761344
desserts-condiments  71.785714  -27.84279 171.41422 0.1866850
```

Observations: Analyzing sodium content across three food items, the initial assumption (H0) was that all items have the same average sodium level (mean). The H_1 (Alternate hypothesis) proposed that at least 1food item differs from the others. As the test results, with a p-value of 0.117 (greater than the chosen alpha 0.05), we retain the null hypothesis, implying we cannot conclude significant variations in sodium content among the three food items.

Task 6

Sales_Leading_Companies

Hypothesis

1. Null hypothesis (H_0): the average values (means) of all three groups are equal.

2. Alternate hypothesis(H_1): At least 1 group has a different average value compared to the other two groups.

```
> snacksSales_6 = c(578, 320, 264, 249, 237)
> beverageSales_6 = c(261, 185, 302, 689)
> candiesSales_6 = c(311, 106, 109, 125, 173)
> chr_sn = "snacks"
> chr_be = "beverages"
> chr_ca = "candies"
> five_Numeric = 5
> four_Numeric = 4
> a_val = 0.01
> dataFrame_6_Snacks = data.frame("col_sales" = snacksSales_6, "col_category" = rep(chr_sn , five_Numeric), stringsAsFactors = FALSE)
> dataFrame_6_beverages = data.frame("col_sales" = beverageSales_6, "col_category" = rep(chr_be , four_Numeric), stringsAsFactors = FALSE)
> dataFrame_6_Candies = data.frame("col_sales" = candiesSales_6, "col_category" = rep(chr_ca , five_Numeric), stringsAsFactors = FALSE)
> saleData_6 = rbind(dataFrame_6_Snacks, dataFrame_6_Candies, dataFrame_6_beverages)
> saleData_6$col_category = as.factor(saleData_6$col_category)
> t6 = aov(col_sales ~ col_category, data = saleData_6)
> t6Summary = summary(t6)
> t6Summary
             Df Sum Sq Mean Sq F value Pr(>F)
col_category  2 103770   51885   2.172   0.16
Residuals    11 262795   23890
```

#p value

```
> colName = "Pr(>F)"
> pValue_6 = t6Summary[[1]][[1, colName]]
> pValue_6
[1] 0.1603487
```

```
ifelse(pValue_6 > a_val, "Fail to reject H0", "Reject H0")
l] "Fail to reject H0"
```

Tukey Test

```
> TukeyHSD(t6)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = col_sales ~ col_category, data = saleData_6)

$col_category
                     diff        lwr       upr     p adj
candies-beverages -194.45 -474.48983  85.58983 0.1916553
snacks-beverages   -29.65 -309.68983 250.38983 0.9561014
snacks-candies     164.80  -99.22409 428.82409 0.2535458

>
```

Observation: After conducting an test (ANOVA) on the probability statistic, we obtained a p-value of 0.16. Consequently, based on a significance level of 0.01, it is inferred that there lies no statistically relationship between the variables. Therefore, not having sufficient evidence to reject the (H_0) null hypothesis, indicates that sales of the company's products utilize the same mean.

Task 7

Expenditure per pupil

Hypothesis

1.Null Hypothesis (H0): The average values (means) are different in at least one pair of sections compared to the others.

2.Alternative Hypothesis (H1): The average values (means) are equal across all sections of the country.

```
> #Task 7
> #Per-Pupil Expenditures
> a_val = 0.05
> expenditureList_e_3 = c(4946, 5953, 6202, 7243, 6113)
> expenditureList_m_3 = c(6149, 7451, 6000, 6479)
> expenditureList_w_3 = c(5282, 8605, 6528, 6911)
> chr_ET = "Eastern Third"
> chr_MT = "Middle Third"
> chr_WT = "Western Third"
> five_Numeric = 5
> four_Numeric = 4
> east_3_dataframe = data.frame("col_expenditure" = expenditureList_e_3, "col_region" = rep(chr_ET, five_Numeric), stringsAsFactors = FALSE)
> mid_3_dataframe = data.frame("col_expenditure" = expenditureList_m_3, "col_region" = rep(chr_MT, four_Numeric), stringsAsFactors = FALSE)
> west_3_dataframe = data.frame("col_expenditure" = expenditureList_w_3, "col_region" = rep(chr_WT, four_Numeric), stringsAsFactors = FALSE)
> expDataframe_7 = rbind(east_3_dataframe, mid_3_dataframe, west_3_dataframe)
> expDataframe_7$col_region = as.factor(expDataframe_7$col_region)
> t7 = aov(col_expenditure ~ col_region, data = expDataframe_7)
> t7Summary = summary(t7)
> t7Summary
            Df  Sum Sq Mean Sq F value Pr(>F)
col_region   2 1244588  622294   0.649  0.543
Residuals   10 9591145  959114
```

```
> colName = "Pr(>F)"
> pVal_7 = t7Summary[[1]] [[1, colName]]
> pVal_7
[1] 0.5433264
```

#Compare the p value with alpha and make the decision

```
> ifelse( pVal_7 > a_val, "Fail to reject H0", "Reject H0")
[1] "Fail to reject H0"
```

Tukeytest

```
> TukeyHSD(t7)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = col_expenditure ~ col_region, data = expDataframe_7)

$col_region
                             diff       lwr      upr     p adj
Middle Third-Eastern Third  428.35 -1372.582 2229.282 0.7954670
Western Third-Eastern Third 740.10 -1060.832 2541.032 0.5203918
Western Third-Middle Third  311.75 -1586.599 2210.099 0.8954324
```

Observation: Upon conducting an test (ANOVA) on the probability statistics, we obtained a p-value of 0.54. As per the significance level of 0.05, it is inferred that there is no significant relationship between the variables. Consequently, (H_0) null hypothesis cannot be rejected, indicating uniform mean expenditure across all three segments.

Task 8

Increasing Plant Growth

To assess these factors, a statistical analysis, such as a two-way ANOVA test, will be conducted. This analysis will allow us to evaluate the effects of both factors and their interaction on plant growth.

Hypothesis

```
> #Task 8
> lightList <- c(1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2)
> foodList <- c(1, 1, 1, 2, 2, 2, 1, 1, 1, 2, 2, 2)
> growthList <- c(9.2, 9.4, 8.9, 7.1, 7.2, 8.5, 8.5, 9.2, 8.9, 5.5, 5.8, 7.6)
> foodListFactor <- factor(foodList)
> lightListFactor <- factor(lightList)
> dataFrame_task8 <- data.frame(foodListFactor, lightListFactor, growthList)
> t8 <- lm(growthList ~ foodListFactor + lightListFactor + foodListFactor:lightListFactor, data = dataFrame_task8)
> ano_task8Result <- anova(t8)
> ano_task8Result
Analysis of Variance Table

Response: growthList
                               Df  Sum Sq Mean Sq F value   Pr(>F)   
foodListFactor                  1 12.8133 12.8133 24.5623 0.001112 **
lightListFactor                 1  1.9200  1.9200  3.6805 0.091331 . 
foodListFactor:lightListFactor  1  0.7500  0.7500  1.4377 0.264819   
Residuals                       8  4.1733  0.5217                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> summary(t8)

Call:
lm(formula = growthList ~ foodListFactor + lightListFactor + 
    foodListFactor:lightListFactor, data = dataFrame_task8)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.8000 -0.4250 -0.1167  0.2583  1.3000 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       9.1667     0.4170  21.982 1.94e-08 ***
foodListFactor2                  -1.5667     0.5897  -2.657    0.029 *  
lightListFactor2                 -0.3000     0.5897  -0.509    0.625    
foodListFactor2:lightListFactor2 -1.0000     0.8340  -1.199    0.265    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7223 on 8 degrees of freedom
Multiple R-squared:  0.7877,	Adjusted R-squared:  0.7081 
F-statistic: 9.894 on 3 and 8 DF,  p-value: 0.004555
```

Observation: In an effort to improve plant growth, 12 random samples undergo two different treatments involving varied plant feeding and 2 types of grow lights. The hypotheses under examination are as follows:

H_0(Null Hypothesis) - no disparity in plant growth concerning light. H_1(Alternate Hypothesis) - There exists a variation in mean plant growth based on light. With a p-value of (Pr(>F)) 0.0011, we reject the (H_0) null hypothesis as it falls below the alpha (0.05), showing a discernible growth difference attributed to light.

H_0(Null Hypothesis) - no distinction in mean growth based on plant food. H_1(Alternate Hypothesis) - difference in mean growth based on plant food. The obtained p-value of (Pr(>F)) 0.09 exceeds the alpha, hence the (H_0)null hypothesis cannot be dismissed. Therefore, we cannot affirm that plant diet significantly influences mean growth.

H_0(Null Hypothesis) - no interaction effect between plant food and light. H_1(Alternate Hypothesis) - There lies an interaction between plant food and light. With a p-value of (Pr(>F)) 0.26, surpassing the alpha (0.05), we are unable to reject the (H_0) null hypothesis. This implies that there is no interaction between plant food and light

Task 9.1 (Q1,2 & 3)

```
#import into R
data_frame_baseball <- read.csv("baseball.csv")
head(data_frame_baseball)
```

O/P-

```
  Team League Year  RS  RA  W   OBP   SLG    BA Playoffs RankSeason RankPlayoffs   G  OOBP  OSLG
1  ARI     NL 2012 734 688 81 0.328 0.418 0.259        0         NA           NA 162 0.317 0.415
2  ATL     NL 2012 700 600 94 0.320 0.389 0.247        1          4            5 162 0.306 0.378
3  BAL     AL 2012 712 705 93 0.311 0.417 0.247        1          5            4 162 0.315 0.403
4  BOS     AL 2012 734 806 69 0.315 0.415 0.260        0         NA           NA 162 0.331 0.428
5  CHC     NL 2012 613 759 61 0.302 0.378 0.240        0         NA           NA 162 0.335 0.424
6  CHW     AL 2012 748 676 85 0.318 0.422 0.255        0         NA           NA 162 0.319 0.405
```

```
#EDA
summary(data_frame_baseball)
```

O/P-

```
     Team              League               Year            RS               RA               W               OBP              SLG               BA             Playoffs     
 Length:1232        Length:1232        Min.   :1962   Min.   : 463.0   Min.   : 472.0   Min.   : 40.0   Min.   :0.2770   Min.   :0.3010   Min.   :0.2140   Min.   :0.0000  
 Class :character   Class :character   1st Qu.:1977   1st Qu.: 652.0   1st Qu.: 649.8   1st Qu.: 73.0   1st Qu.:0.3170   1st Qu.:0.3750   1st Qu.:0.2510   1st Qu.:0.0000  
 Mode  :character   Mode  :character   Median :1989   Median : 711.0   Median : 709.0   Median : 81.0   Median :0.3260   Median :0.3960   Median :0.2600   Median :0.0000  
                                       Mean   :1989   Mean   : 715.1   Mean   : 715.1   Mean   : 80.9   Mean   :0.3263   Mean   :0.3973   Mean   :0.2593   Mean   :0.1981  
                                       3rd Qu.:2002   3rd Qu.: 775.0   3rd Qu.: 774.2   3rd Qu.: 89.0   3rd Qu.:0.3370   3rd Qu.:0.4210   3rd Qu.:0.2680   3rd Qu.:0.0000  
                                       Max.   :2012   Max.   :1009.0   Max.   :1103.0   Max.   :116.0   Max.   :0.3730   Max.   :0.4910   Max.   :0.2940   Max.   :1.0000  
                                                                                                                                                                              
  RankSeason     RankPlayoffs         G              OOBP             OSLG       
 Min.   :1.000   Min.   :1.000   Min.   :158.0   Min.   :0.2940   Min.   :0.3460  
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:162.0   1st Qu.:0.3210   1st Qu.:0.4010  
 Median :3.000   Median :3.000   Median :162.0   Median :0.3310   Median :0.4190  
 Mean   :3.123   Mean   :2.717   Mean   :161.9   Mean   :0.3323   Mean   :0.4197  
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:162.0   3rd Qu.:0.3430   3rd Qu.:0.4380  
 Max.   :8.000   Max.   :5.000   Max.   :165.0   Max.   :0.3840   Max.   :0.4990  
 NA's   :988     NA's   :988                     NA's   :812      NA's   :812     
> 
```

```
str(data_frame_baseball)
```

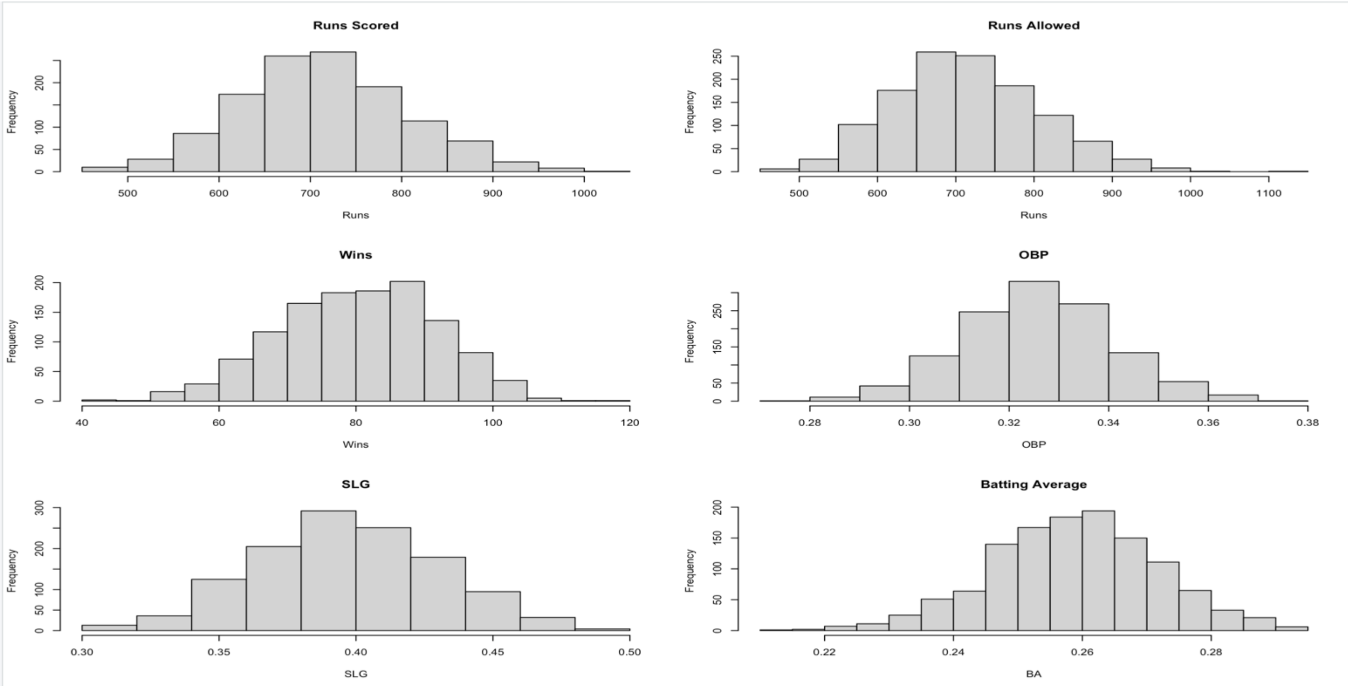
O/P-

```
'data.frame':	1232 obs. of  15 variables:
 $ Team        : chr  "ARI" "ATL" "BAL" "BOS" ...
 $ League      : chr  "NL" "NL" "AL" "AL" ...
 $ Year        : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
 $ RS          : int  734 700 712 734 613 748 669 667 758 726 ...
 $ RA          : int  688 600 705 806 759 676 588 845 890 670 ...
 $ W           : int  81 94 93 69 61 85 97 68 64 88 ...
 $ OBP         : num  0.328 0.32 0.311 0.315 0.302 0.318 0.315 0.324 0.33 0.335 ...
 $ SLG         : num  0.418 0.389 0.417 0.415 0.378 0.422 0.411 0.381 0.436 0.422 ...
 $ BA          : num  0.259 0.247 0.247 0.26 0.24 0.255 0.251 0.251 0.274 0.268 ...
 $ Playoffs    : int  0 1 1 0 0 0 1 0 0 1 ...
 $ RankSeason  : int  NA 4 5 NA NA NA 2 NA NA 6 ...
 $ RankPlayoffs: int  NA 5 4 NA NA NA 4 NA NA 2 ...
 $ G           : int  162 162 162 162 162 162 162 162 162 162 ...
 $ OOBP        : num  0.317 0.306 0.315 0.331 0.335 0.319 0.305 0.336 0.357 0.314 ...
 $ OSLG        : num  0.415 0.378 0.403 0.428 0.424 0.405 0.39 0.43 0.47 0.402 ...
>
```

#Hist Plots

```
> #Hist Plots
> par(mfrow=c(3, 2))
> chr_RS = "Runs Scored"
> chr_RA = "Runs Allowed"
> chr_w = "Wins"
> hist(data_frame_baseball$RS, main=chr_RS, xlab="Runs")
> hist(data_frame_baseball$RA, main=chr_RA, xlab="Runs")
> hist(data_frame_baseball$W, main=chr_w, xlab="Wins")
> hist(data_frame_baseball$OBP, main="OBP", xlab="OBP")
> hist(data_frame_baseball$SLG, main="SLG", xlab="SLG")
> hist(data_frame_baseball$BA, main="Batting Average", xlab="BA")
```

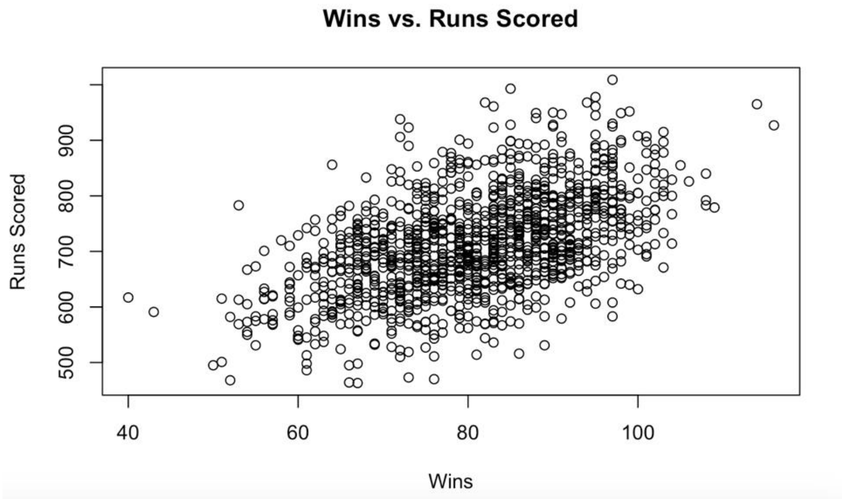
O/P-

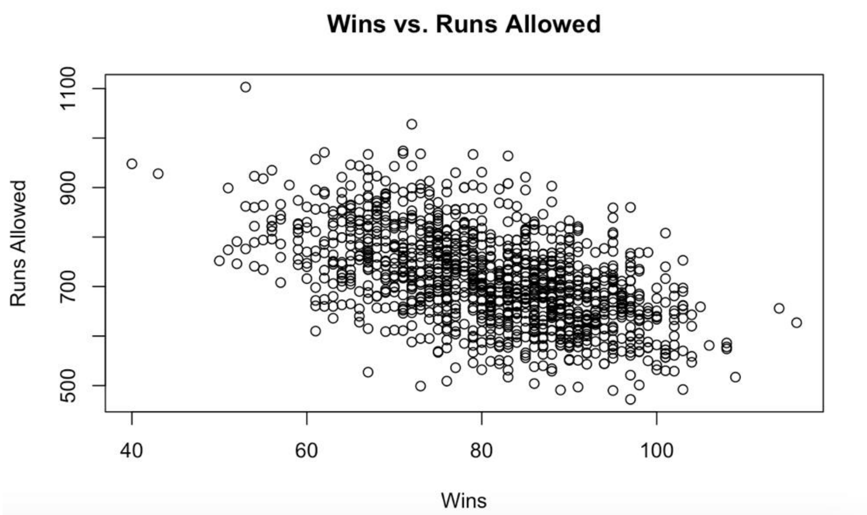


Scatter Plot

```
> # Scatter Plot
> par(mfrow=c(1, 1))
> chr_RS = "Runs Scored"
> chr_RA = "Runs Allowed"
> chr_w = "Wins"
> chr_wVsRS = "Wins vs. Runs Scored"
> chr_wVsRA = "Wins vs. Runs Allowed"
> plot(data_frame_baseball$W, data_frame_baseball$RS, main=chr_wVsRS, xlab=chr_w, ylab=chr_RS)
> plot(data_frame_baseball$W, data_frame_baseball$RA, main=chr_wVsRA, xlab=chr_w, ylab=chr_RA)
```

O/P-





Observation: The baseball dataset provides details on various performance metrics of different baseball teams over time, including runs scored (RS), wins (W), runs allowed (RA), playoff outcomes , on-base percentage (OBP), batting average (BA), and slugging percentage (SLG). Average no. of wins is approximately 80, indicating a consistent win-loss record among most teams. Histograms of runs scored and runs allowed display roughly normal distributions, aligning with expectations in baseball statistics. Scatter plots depict notable relationships, including a +ve correlation between RS and W (runs scored and wins), as well as a -ve correlation between RA and W (runs allowed and wins). This suggests that teams scoring more runs and conceding fewer runs tend to achieve higher win counts.

- Null hypothesis (H_0): distribution of (W) wins across decades is uniform (indicating no variation in wins by decade).
- Alternate hypothesis (H_1): distribution of (W) wins across decades is non-uniform (indicating variation in wins by decade)

```
library(dplyr)
```

```
data_frame_baseball$Decade <- as.factor(10 * (data_frame_baseball$Year %/% 10))
head(data_frame_baseball)
```

O/P-

```
  Team League Year  RS  RA  W   OBP   SLG    BA Playoffs RankSeason RankPlayoffs   G  OOBP  OSLG Decade
1  ARI     NL 2012 734 688 81 0.328 0.418 0.259        0         NA           NA 162 0.317 0.415   2010
2  ATL     NL 2012 700 600 94 0.320 0.389 0.247        1          4            5 162 0.306 0.378   2010
3  BAL     AL 2012 712 705 93 0.311 0.417 0.247        1          5            4 162 0.315 0.403   2010
4  BOS     AL 2012 734 806 69 0.315 0.415 0.260        0         NA           NA 162 0.331 0.428   2010
5  CHC     NL 2012 613 759 61 0.302 0.378 0.240        0         NA           NA 162 0.335 0.424   2010
6  CHW     AL 2012 748 676 85 0.318 0.422 0.255        0         NA           NA 162 0.319 0.405   2010
```

```
winsByDecade <- data_frame_baseball %>%
 group_by(Decade) %>%
 summarize(TotalWins = sum(W))
winsByDecade
```

O/P-

```
  Decade TotalWins
  <fct>      <int>
1 1960       13267
2 1970       17934
3 1980       18926
4 1990       17972
5 2000       24286
6 2010        7289
```

```
totalWins <- sum(winsByDecade$TotalWins)
numDecades <- nrow(winsByDecade)
expectedFreq <- rep(totalWins / numDecades, numDecades)

p = rep(1/length(unique(data_frame_baseball$Decade)), length(unique(data_frame_baseball$Decade)))
chiSqureTest <- chisq.test(winsByDecade$TotalWins, p = p)
chiSqureTest
```

O/P-

```
	Chi-squared test for given probabilities

data:  winsByDecade$TotalWins
X-squared = 9989.5, df = 5, p-value < 2.2e-16
```

```
#---------------------------
#Critical Value from TextBook
criticalValue <- 11.070
#----------------------------
```

```
> #---------------------------
> #Critical Value from TextBook
> criticalValue <- 11.070
> #----------------------------
>
> if (chiSqureTest$statistic > criticalValue) {
+   print("Reject the (H0) Null hypothesis")
+ } else {
+   print("Do not reject the (H0) Null hypothesis")
+ }
[1] "Reject the (H0) Null hypothesis"
>
```

Observation: After conducting the test, we find sufficient evidence to reject (H_0) null hypothesis, showing that distribution of (W) wins across decades is not uniform. This shows that the no. of wins varry across different decades, supporting (H_1) alternative hypothesis which proposes differential distribution of wins across decades.

Task 9.2 (Q4 & 5)

```
#import data into R
data_frame_crop_data <-read.csv("crop_data.csv")
head(data_frame_crop_data)
```

O/P-

```
  density block fertilizer    yield
1       1     1          1 177.2287
2       2     2          1 177.5500
3       1     3          1 176.4085
4       2     4          1 177.7036
5       1     1          1 177.1255
6       2     2          1 176.7783
```

```
summary(data_frame_crop_data)
str(data_frame_crop_data)
```

O/P-

```
> #EDA
> summary(data_frame_crop_data)
    density         block        fertilizer      yield      
 Min.   :1.0   Min.   :1.00   Min.   :1    Min.   :175.4  
 1st Qu.:1.0   1st Qu.:1.75   1st Qu.:1    1st Qu.:176.5  
 Median :1.5   Median :2.50   Median :2    Median :177.1  
 Mean   :1.5   Mean   :2.50   Mean   :2    Mean   :177.0  
 3rd Qu.:2.0   3rd Qu.:3.25   3rd Qu.:3    3rd Qu.:177.4  
 Max.   :2.0   Max.   :4.00   Max.   :3    Max.   :179.1  
> str(data_frame_crop_data)
'data.frame':	96 obs. of  4 variables:
 $ density   : int  1 2 1 2 1 2 1 2 1 2 ...
 $ block     : int  1 2 3 4 1 2 3 4 1 2 ...
 $ fertilizer: int  1 1 1 1 1 1 1 1 1 1 ...
 $ yield     : num  177 178 176 178 177 ...
```

```
> 
> df_factor_crop_d <- within(dataFrame_cd, {
+   density <- factor(density)
+   fertilizer <- factor(fertilizer)
+   block <- factor(block)
+ })
> 
> anovaResult <- aov(yield ~ fertilizer * density, data = df_factor_crop_d)
> summary(anovaResult)
                   Df Sum Sq Mean Sq F value   Pr(>F)    
fertilizer          2  6.068   3.034   9.001 0.000273 ***
density             1  5.122   5.122  15.195 0.000186 ***
fertilizer:density  2  0.428   0.214   0.635 0.532500    
Residuals          90 30.337   0.337                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For the fertilizer:

- H_0: no variation in mean yield across all fertilizer levels.
- H_1: difference in mean yield for at least one fertilizer level.

For the density:

- H_0: no disparity in mean yield across all density levels.
- H_1: difference in mean yield for at least one density level.

Regarding the interaction:

- H_0: impact of fertilizer on yield remains consistent across all density levels, and vice versa.
- H_1: effect of fertilizer on yield varies across at least one density level, or vice versa.

At alpha 0.05:

- For fertilizer: The obtained p-value (0.000273) falls below 0.05, Therefore, rejecting (H_0) null hypothesis. This shows a notable difference in mean yield among fertilizer levels, suggesting a significant impact of fertilizer on yield.
- Concerning the density: The resulting p-value (0.000186) which is below alpha (0.05), leading to the H_0rejection. This signifies influence of density on yield.
- As for the interaction: The obtained p-value (0.5325) surpasses 0.05, indicating no interaction effect (between fertilizer and density) on yield. This implies that the impact of density on yield does not rely on the fertilizer level, and vice versa.

References:

1. Bevans, R. (2023, June 22). *ANOVA in R: A complete step-by-step guide with examples*. Scribbr. https://www.scribbr.com/statistics/anova-in-r/
2. GfG. (2023, December 19). *Chi-square test in R*. GeeksforGeeks. https://www.geeksforgeeks.org/chi-square-test-in-r/
3. Holtz, Y. (n.d.). *Tukey test and boxplot in R*. – the R Graph Gallery. https://r-graph-gallery.com/84-tukey-test.html
4. Nancy E Schoenberg, Yelena N Tarasenko, Claire Snell-Rood, Are evidence-based, community-engaged energy balance interventions enough for extremely vulnerable populations?, *Translational Behavioral Medicine*, Volume 8, Issue 5, October 2018, Pages 733–738, https://doi.org/10.1093/tbm/ibx013