

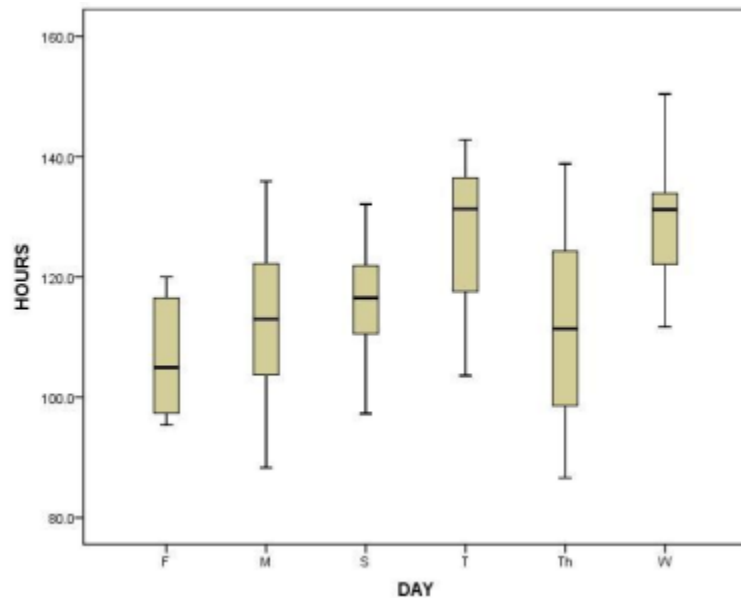
HW5

Name: Naga Satya Silpa Annadevara.

Student ID: A20517818.

let's use the "clerical_Q2.txt" data.

A store manager noticed that the busiest days for clerical staff are Wednesdays and Tuesdays. See enclosed box plot. The manager tries to compare the group means in hours by different days



a). [10] Observe the box plot. Can you confirm that the hours in Tuesday is the highest? Why?

ANSWER:

- From the boxplot, we can observe that there are 6 different groups with days variable on the x-axis and hours variable on the y-axis. For each group, we can observe a 1-box plot.
- From the visualization, we can clearly observe that the variation within the groups (boxplots) are not even/uniform. So, we cannot compare the groups based on their q2 values. Why? Because the within-group variation (IQR) is large. So, we cannot use q2 to compare the groups.
- So, we need to have additional information such as mean values for every group to compare them & give the conclusion.
- As the boxplot cannot provide mean values, we need to use ANOVA technique to say hours on Tuesday is the highest or not.
- So, with the variation difference among the groups & without knowing the mean values for every group, we cannot confirm that hours on Tuesday is the highest.

b). [20] Write down your hypothesis in the ANOVA to compare the group means in hours by different days

ANSWER:

Null hypothesis: All the groups have the same mean. In other words, the means of hours for different days are equal.

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$$

- μ_1 = average of hours on Monday group
- μ_2 = average of hours on Tuesday group
- μ_3 = average of hours on Wednesday group
- μ_4 = average of hours on Thursday group
- μ_5 = average of hours on Friday group
- μ_6 = average of hours on Saturday group

Alternative hypothesis: At least 2 groups (1-pair of the mean) have different means. In other words, at least 2 group's means among the means of hours for different days are not equal.

$$H_a = \text{Not all the } \mu_t \text{ are equal} / \mu_i \neq \mu_j$$

b). [30] Using R to build the ANOVA regression model and help the manager to make the decision whether the group means in hours or different days are the same or not.

ANSWER:

Importing clerical_Q2.txt data into the r-studio:

RStudio interface showing the 'Environment' pane with a dataset named 'data' containing 52 observations and 9 variables. The 'Data' pane displays the first 8 rows of the dataset:

day	hours	mail	cert	acc	change	check	misc	tickets
1 M	128.5	7781	100	886	235	644	56	737
2 T	113.6	7004	110	962	388	589	57	1029
3 W	146.6	7267	61	1342	398	1081	59	830
4 Th	124.3	2129	102	1153	457	891	57	1468
5 F	100.4	4878	45	803	577	537	49	335
6 S	119.2	3999	144	1127	345	563	64	918
7 M	109.5	11777	123	627	326	402	60	335
8 T	128.5	5764	78	748	161	495	57	962

The console shows the following R code and output:

```

> data = read.table(file = "C:/Users/satya/OneDrive/Desktop/Case2_clerical.txt", header = 1, sep = "\t")
> View(data)
> str(data)
'data.frame': 52 obs. of 9 variables:
 $ day : chr "M" "T" "W" "Th" ...
 $ hours : num 128.5 113.6 146.6 124.3 100.4 119.2 109.5 128.5 ...
 $ mail : int 7781 7004 7267 2129 4878 3999 11777 5764 7392 8100 ...
 $ cert : int 100 110 61 102 45 144 123 78 172 126 ...
 $ acc : int 886 962 1342 1153 803 1127 627 748 876 685 ...
 $ change : int 235 388 398 457 577 345 326 161 219 287 ...
 $ check : int 644 589 1081 891 537 563 402 495 823 555 ...
 $ misc : int 56 57 59 57 49 64 60 57 62 86 ...
 $ tickets : int 737 1029 830 1468 335 918 335 962 665 577 ...

```

Building anova regression model using aov () function:

RStudio interface showing the 'Environment' pane with two datasets: 'anova' (List of 13) and 'data' (52 obs. of 9 variables). The console shows the following R code and output:

```

> anova = aov(hours ~ day, data = data)
> summary(anova)
Df Sum Sq Mean Sq F value Pr(>F)
day      5    3876    775.2    4.227 0.00303 **
Residuals 46    8437    183.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
>
> anova = lm(hours ~ day, data = data)
> summary(anova)

Call:
lm(formula = hours ~ day, data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-25.000  -9.606   1.400   9.402  28.856

Coefficients:
(Intercept) 106.625    4.788   22.269  < 2e-16 ***
dayM         6.675    6.581    1.014  0.315722
dayS         9.288    6.771    1.372  0.176846
dayT        20.531    6.581    3.120  0.003121 **
dayTh        3.319    6.581    0.504  0.616367
dayW        23.342    6.581    3.547  0.000909 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.54 on 46 degrees of freedom
Multiple R-squared:  0.3148,    Adjusted R-squared:  0.2403
F-statistic: 4.227 on 5 and 46 DF,  p-value: 0.003033

```

Decision/ conclusion:

- Using 95% confidence level, as the p-value (0.00303) in F-test is smaller than alpha ($p < 0.05\%$) we do have enough evidence to reject null hypothesis.
- In other words, we have enough evidence to reject that all group means in hours or different days are not same. At least 2 group means are different.
- To know which 2 groups means are different, we need to look at the t-test. We can clearly observe that Tuesday and Wednesday group means are not same.

c). [20] Try to interpret the coefficients you got in the ANOVA regression model from part b).

ANSWER:

We cannot see co-efficients in the model which we built using `aov()` function.

So we need to use other modeling approaches such as linear regression (e.g., `lm()`)

So built model using `lm()` function with same variables to see the co-efficients.

The screenshot shows the RStudio interface. The top pane displays a data frame with columns: day, hours, mail, cert, acc, change, check, misc, tickets. The bottom-left pane shows the console with the following code and output:

```
R 4.3.0 ~-/-
> anova = aov(hours ~ day, data = data)
> summary(anova)
          Df Sum Sq Mean Sq F value    Pr(>F)
day           5   3876    775.2    4.227 0.00303 **
Residuals    46   8437    183.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>
>
> anova = lm(hours ~ day, data = data)
> summary(anova)

Call:
lm(formula = hours ~ day, data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-25.000  -9.606   1.400   9.402  28.856

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  106.625     4.788   22.269  < 2e-16 ***
dayM          6.675     6.581    1.014  0.315722
dayS          9.288     6.771    1.372  0.176046
dayT         20.531     6.581    3.120  0.003121 **
dayTh         3.319     6.581    0.504  0.616367
dayW         23.342     6.581    3.547  0.000909 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.54 on 46 degrees of freedom
Multiple R-squared:  0.3148,    Adjusted R-squared:  0.2403
F-statistic: 4.227 on 5 and 46 DF,  p-value: 0.003033

> |
```

Annotations in the image:

- A red box highlights the coefficients table in the `summary(anova)` output for the `lm` model.
- A red arrow points from the text "These are the co-efficients that we need to interpret." to the highlighted coefficients table.
- Red text annotations explain the process: "built anova model using aov() function, but we cannot see co-efficients to interpret in this model." and "so we need to use different approach like linear regression model by using lm() function. so we built another model with same variables, where we can see co-efficients".

Interpretation of co-efficients: as there is no `dayFriday`, we are going to use Friday as baseline.

$$\text{hours} = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_4 * X_4 + \beta_5 * X_5 + e$$

- Intercept (+106.625): $\sim \beta_0$

The intercept represents the estimated mean value for the reference group (Friday) when all other dummy variables are zero. In this case, it suggests that the estimated mean value for Friday is 106.625.

- dayM (+6.675): β_1 $X_1 = \text{dayM}$

The coefficient for "dayM" (+6.675) represents the difference in mean between Monday and Friday. The positive coefficient indicates that, on average, the mean for Monday is 6.675 units higher than the mean for Friday.

- dayS (+9.288): β_2 $X_2 = \text{dayS}$

The coefficient for "dayS" (+9.288) represents the difference in mean between Saturday and Friday. The positive coefficient indicates that, on average, the mean for Saturday is 9.288 units higher than the mean for Friday.

- dayT (+20.531): β_3 $X_3 = \text{dayT}$

The coefficient for "dayT" (+20.531) represents the difference in mean between Tuesday and Friday. The positive coefficient suggests that, on average, the mean for Tuesday is 20.531 units higher than the mean for Friday.

- dayTh (+3.319): β_4 $X_4 = \text{dayTh}$

The coefficient for "dayTh" (+3.319) represents the difference in mean between Thursday and Friday. The positive coefficient indicates that, on average, the mean for Thursday is 3.319 units higher than the mean for Friday.

- dayW (+23.342): β_5 $X_5 = \text{dayW}$

The coefficient for "dayW" (+23.342) represents the difference in mean between Wednesday and Friday. The positive coefficient suggests that, on average, the mean for Wednesday is 23.342 units higher than the mean for Friday.

d). [20] practice for data preprocessing: create N-1 dummy variables for the variable 'DAY'. Convert the variable "mail" to nominal variable by creating 4 groups. Again, paste the codes and snapshots.

ANSWER:

Created N-1 dummy variable for the variable "DAY":

Console:

```
R 4.3.0 ~ ./
> dummy_data <- cbind(data, model.matrix(~ day - 1, data = data))
> str(dummy_data)
'data.frame': 52 obs. of 15 variables:
 $ day : chr "M" "T" "W" "Th" ...
 $ hours : num 128 114 147 124 100 ...
 $ mail : int 7781 7004 7267 2129 4878 3999 11777 5764 7392 8100 ...
 $ cert : int 100 110 61 102 45 144 123 78 172 126 ...
 $ acc : int 886 962 1342 1153 803 1127 627 748 876 685 ...
 $ change : int 235 388 398 457 577 345 326 161 219 287 ...
 $ check : int 644 589 1081 891 537 563 402 495 823 555 ...
 $ misc : int 56 57 59 57 49 64 60 57 62 86 ...
 $ tickets : int 737 1029 830 1468 335 918 335 962 665 577 ...
 $ dayF : num 0 0 0 0 1 0 0 0 0 0 ...
 $ dayM : num 1 0 0 0 0 0 1 0 0 0 ...
 $ dayS : num 0 0 0 0 0 1 0 0 0 0 ...
 $ dayT : num 0 1 0 0 0 0 0 1 0 0 ...
 $ dayTh : num 0 0 0 1 0 0 0 0 0 1 ...
 $ dayW : num 0 0 1 0 0 0 0 0 1 0 ...
> dummy_data$day <- NULL
> str(dummy_data)
'data.frame': 52 obs. of 14 variables:
 $ hours : num 128 114 147 124 100 ...
 $ mail : int 7781 7004 7267 2129 4878 3999 11777 5764 7392 8100 ...
 $ cert : int 100 110 61 102 45 144 123 78 172 126 ...
 $ acc : int 886 962 1342 1153 803 1127 627 748 876 685 ...
 $ change : int 235 388 398 457 577 345 326 161 219 287 ...
 $ check : int 644 589 1081 891 537 563 402 495 823 555 ...
 $ misc : int 56 57 59 57 49 64 60 57 62 86 ...
 $ tickets : int 737 1029 830 1468 335 918 335 962 665 577 ...
 $ dayF : num 0 0 0 0 1 0 0 0 0 0 ...
 $ dayM : num 1 0 0 0 0 0 1 0 0 0 ...
 $ dayS : num 0 0 0 0 0 1 0 0 0 0 ...
 $ dayT : num 0 1 0 0 0 0 0 1 0 0 ...
 $ dayTh : num 0 0 0 1 0 0 0 0 0 1 ...
 $ dayW : num 0 0 1 0 0 0 0 0 1 0 ...
> View(dummy_data)
```

Environment:

- anova: List of 13
- data: 52 obs. of 9 variables
- dummy_data: 52 obs. of 14 variables

Annotations:

- Red box around the dummy variables in the console output: *created N-1 dummy variables for 'day'*
- Red text in the console: *removed original 'day' variable from the data_set*

Note: For creating dummy variable we need to use `model.matrix()` function , As dummies package was removed from the latest version of R.

Convert the variable “mail” to nominal variable by creating 4 groups.:

We need to use `cut()` function to cut the mail into 4 groups:

Console:

```
R 4.3.0 ~ ./
> data$mail = cut(data$mail, breaks = 4, labels = c("Group 1", "Group 2", "Group 3", "Group 4"))
> data$mail
[1] Group 3 Group 3 Group 3 Group 1 Group 2 Group 1 Group 4 Group 2 Group 3 Group 3 Group 2 Group 2 Group 1
[14] Group 3 Group 1 Group 1 Group 1 Group 2 Group 3 Group 3 Group 3 Group 2 Group 3 Group 1 Group 2 Group 2
[27] Group 1 Group 2 Group 2 Group 3 Group 3 Group 3 Group 2 Group 2 Group 2 Group 3 Group 2 Group 1 Group 1
[40] Group 2 Group 2 Group 2 Group 2 Group 2 Group 2 Group 1 Group 2 Group 1 Group 2 Group 2 Group 1 Group 1 Group 3
Levels: Group 1 Group 2 Group 3 Group 4
>
```

Environment:

- data: 52 obs. of 9 variables

Note: as the group values came into scientific notation, I used labels to improve the readability.

