Mark Stanley: Linear Programming, Hypothesis Testing: ANOVA & Tukey

2024-04-06

Here is a table about different foods at a buffet.

```
df <- data.frame(Per_Ounce = c("Fat(g)", "Protein(g)", "Carbohydrates(g)", "Sodium(mg)",
↪  "Sugar(g)","Calories","Cost(cents)"), Garden_salad = c(1,0,3,10,12,15,25),
↪  Grilled_veggies = c(1,1,4,12,5,20,40), Pasta = c(3,2,12,15,6,40,30), Meatballs =
↪  c(6,10,6,25,4,60,50), Curried_chicken = c(4,12,3,20,1,50,60))

df
```

```
##          Per_Ounce Garden_salad Grilled_veggies Pasta Meatballs Curried_chicken
## 1           Fat(g)            1               1     3         6               4
## 2       Protein(g)            0               1     2        10              12
## 3 Carbohydrates(g)            3               4    12         6               3
## 4       Sodium(mg)           10              12    15        25              20
## 5         Sugar(g)           12               5     6         4               1
## 6         Calories           15              20    40        60              50
## 7      Cost(cents)           25              40    30        50              60
```

Want to minimize cost with the following constraints:

- fat intake to 40 grams or less.
- minimum of 80 grams protein.
- minimum of 60 grams carbohydrates.
- limit his sodium intake to at most 200 milligrams.
- limit calorie intake to at most 700 calories.

The objective function is:

z = 25 * garden_salad + 40 * grilled_veggies + 30* pasta + 50 * meatballs + 60* curried_chicken

Here is the cost function as a vector in R:

```
cost <- c(25,40,30,50,60)
```

Where z is the cost function of lunch and the variables represent the quantity of food items purchased

The constraints are that

fat_intake <= 40g
protein_intake >= 80g
carb_intake >= 60g
sodium_intake <= 200mg
calorie_intake <= 700cal

Using linear programming we find the optimal combination of foods:

```r
library(lpSolve)

# note that we don't need to include sugar here.

val_matrix <-
↪  matrix(c(1,1,3,6,4,0,1,2,10,12,3,4,12,6,3,10,12,15,25,20,15,20,40,60,50),nrow = 5,
↪  byrow = T)

val_matrix
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    1    3    6    4
## [2,]    0    1    2   10   12
## [3,]    3    4   12    6    3
## [4,]   10   12   15   25   20
## [5,]   15   20   40   60   50
```

```r
constraints <- c(40,80,60,200,700)

constraint_dir <- c("<=",">=",">=","<=","<=")

lp_sol <- lp("min", cost, val_matrix, constraint_dir, constraints)

print(lp_sol$solution)
```

```
## [1] 0.000000 0.000000 2.666667 2.666667 4.000000
```

```r
print(lp_sol$objval)
```

```
## [1] 453.3333
```

We don't need to round up, as the food is weighed, so the optimal quantity of foods is: 2.66 pasta, 2.66 meatballs, and 4 curried chicken salad.

The total cost will be $4.53, and all the constraints will be met.

2. In an experiment to investigate the effect of colour paper (blue, green, orange) on the response rate for questionnaires distributed by the "windshield method" in supermarket partaking lots, 15 lots were chosen, and each colour was assigned at random to five of the lots. The response rate (in %) are given below. Let mu1, mu2, mu3 be the population mean response rates for blue, green and orange questionnaires respectively.
   Blue: 27 25 30 26 34
   Green: 34 29 25 31 29
   Orange: 28 22 24 26 25

The response variable is the response rate of different questionnaires. The factor is the color of the questionnaire paper. The treatments are the colors blue, green and orange and the experimental units are the flyers.

We perform a one-way ANOVA using p value of 10%. The null hypothesis is that the color does not change the response rate, and the alternative hypothesis is that the color of flyers DOES change the response rate:

```r
blue <- c(27, 25, 30, 26, 34)
green <- c(34, 29, 25, 31, 29)
orange <- c(28, 22, 24, 26, 25)

mean(blue)
```

```
## [1] 28.4
```

```r
mean(green)
```

```
## [1] 29.6
```

```r
mean(orange)
```

```
## [1] 25
```

```r
response_rates <- c(blue, green, orange)


colors <- factor(rep(c("Blue", "Green", "Orange"), each = 5))

anova_result <- aov(response_rates ~ colors)

summary(anova_result)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## colors       2  56.93   28.47   2.935 0.0917 .
## Residuals   12 116.40    9.70
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the p value is 0.0917 which is LESS than 0.1, there is sufficient evidence at the 10% level of significance to indicate that there is a difference in the mean response rates under the different colors.

Here is a 90% Tukey simultaneous confidence interval:

```r
tukey_result <- TukeyHSD(anova_result, conf.level = 0.90)

print(tukey_result)
```

```
##   Tukey multiple comparisons of means
##     90% family-wise confidence level
##
## Fit: aov(formula = response_rates ~ colors)
##
## $colors
##              diff       lwr        upr     p adj
## Green-Blue    1.2 -3.262581  5.6625813 0.8178869
## Orange-Blue  -3.4 -7.862581  1.0625813 0.2358108
## Orange-Green -4.6 -9.062581 -0.1374187 0.0888902
```

All this information shows is that Orange is a worse color than green for paper color of flyer response rates. There is no clear winner, as the interval containing green and blue and the interval containing orange and blue contains zero.

Here is the safety of different resorts, with 100 being the safest and 0 the least safe scores possible:

```
df <- data.frame(Resort = c("Xcaret","Moon Palace","Garza Blankca","Riu"), Traveler_1 =
↪  c(40,30,40,50), Traveler_2 = c(60,50,70,80), Traveler_3 = c(60,50,60,60), Traveler_4
↪  = c(20,40,50,60))

df
```

```
##          Resort Traveler_1 Traveler_2 Traveler_3 Traveler_4
## 1        Xcaret         40         60         60         20
## 2   Moon Palace         30         50         50         40
## 3 Garza Blankca         40         70         60         50
## 4           Riu         50         80         60         60
```

Here the response variable is the safety score of the resort. The treatment factors are the different resorts. A block factor is the use of 4 different travelers.

The ANOVA table:

```
# Create a data frame with the provided data
dframe <- data.frame(Traveler =
↪  rep(c("Traveler_1","Traveler_2","Traveler_3","Traveler_4"),each = 4),
  Resort = rep(c("Xcaret", "Moon Palace", "Garza Blanca", "Riu"),times = 4),
  Outcomes = c(40, 30, 40, 50,60, 50, 70, 80, 60, 50, 60, 60,20, 40, 50, 60)
)

dframe
```

```
##      Traveler       Resort Outcomes
## 1  Traveler_1       Xcaret       40
## 2  Traveler_1  Moon Palace       30
## 3  Traveler_1 Garza Blanca       40
## 4  Traveler_1          Riu       50
## 5  Traveler_2       Xcaret       60
## 6  Traveler_2  Moon Palace       50
## 7  Traveler_2 Garza Blanca       70
## 8  Traveler_2          Riu       80
## 9  Traveler_3       Xcaret       60
## 10 Traveler_3  Moon Palace       50
## 11 Traveler_3 Garza Blanca       60
## 12 Traveler_3          Riu       60
## 13 Traveler_4       Xcaret       20
## 14 Traveler_4  Moon Palace       40
## 15 Traveler_4 Garza Blanca       50
## 16 Traveler_4          Riu       60
```

```
model <- aov(Outcomes ~ Resort + Traveler, data = dframe)

summary(model)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Resort        3   1025   341.7    4.92 0.0272 *
## Traveler      3   1725   575.0    8.28 0.0059 **
## Residuals     9    625    69.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Check for difference in mean safety scores of the resorts using critical value approach (alpha = 0.05):

The null hypothesis is that all the resorts have the same mean safety score, and the alternative hypothesis is that at least one resort has a different mean safety score than the others.

```r
critical_F <- qf(0.05, 3,7, lower.tail = F)

critical_F
```

```
## [1] 4.346831
```

As the F value of 4.92 is greater than the critical value of 4.34, there is sufficient evidence to reject the null hypothesis.

Check for difference in mean safety scores the different travellers give using critical value approach (do some travelers just give out higher scores in general?):

The null hypothesis is that all the travelers give the same mean safety score, and the alternative hypothesis is that at least one traveler gives a different mean safety score than the others.

```r
critical_F <- qf(0.05,3,7,lower.tail = F)

critical_F
```

```
## [1] 4.346831
```

As the F value of 8.28 is greater than the critical value of 4.34, there is sufficient evidence to reject the null hypothesis.

Here is a 95% simultaneous Tukey confidence interval to view differences in the mean safety scores of the different resorts:

```r
tukey_result <- TukeyHSD(model,conf.level = 0.95)

tukey_result
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Outcomes ~ Resort + Traveler, data = dframe)
##
## $Resort
##                             diff        lwr        upr     p adj
## Moon Palace-Garza Blanca -12.5 -30.895376  5.8953755 0.2174850
## Riu-Garza Blanca           7.5 -10.895376 25.8953755 0.6005945
## Xcaret-Garza Blanca      -10.0 -28.395376  8.3953755 0.3786531
```

```
## Riu-Moon Palace           20.0   1.604624 38.3953755 0.0332115
## Xcaret-Moon Palace         2.5 -15.895376 20.8953755 0.9728629
## Xcaret-Riu               -17.5 -35.895376  0.8953755 0.0628466
##
## $Traveler
##                         diff         lwr       upr     p adj
## Traveler_2-Traveler_1  25.0   6.6046245 43.395376 0.0095723
## Traveler_3-Traveler_1  17.5  -0.8953755 35.895376 0.0628466
## Traveler_4-Traveler_1   2.5 -15.8953755 20.895376 0.9728629
## Traveler_3-Traveler_2  -7.5 -25.8953755 10.895376 0.6005945
## Traveler_4-Traveler_2 -22.5 -40.8953755 -4.104624 0.0176911
## Traveler_4-Traveler_3 -15.0 -33.3953755  3.395376 0.1183609
```

We observe that Riu gets higher safety scores than the Moon Palace, and that traveler 2 gives higher safety scores that traveler 1 and traveler 4.

The effects of two factors on the response to television advertisements. The first factor is the time of day at which the ad is run, while the second is the position of the ad within the hour.

```r
df <- data.frame(Time_of_day =
↪  c("10AM","10AM","10AM","4PM","4PM","4PM","9PM","9PM","9PM"), On_the_hour =
↪  c(42,37,41,62,60,58,100,96,103), On_the_half_hour = c(36,41,38,57,60,55,97,96,101),
↪  Early_in_program = c(62,68,64,88,85,81,127,120,126), Late_in_program =
↪  c(51,47,48,67,60,66,105,101,107))

df
```

```
##   Time_of_day On_the_hour On_the_half_hour Early_in_program Late_in_program
## 1        10AM          42               36               62              51
## 2        10AM          37               41               68              47
## 3        10AM          41               38               64              48
## 4         4PM          62               57               88              67
## 5         4PM          60               60               85              60
## 6         4PM          58               55               81              66
## 7         9PM         100               97              127             105
## 8         9PM          96               96              120             101
## 9         9PM         103              101              126             107
```

The treatments are the time of day and the position of the advertisement.

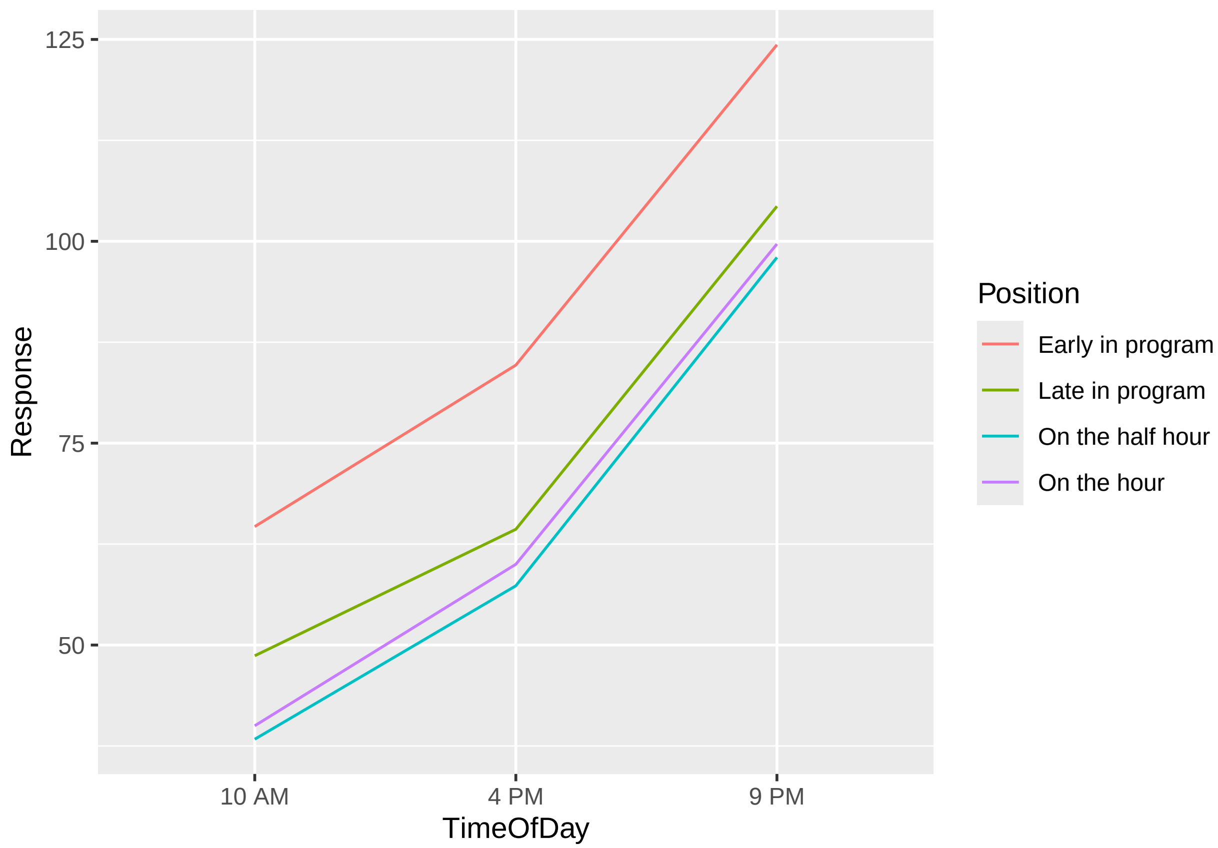
Graphical analysis to check for interaction between time of day and position of advertisement:

```r
library(ggplot2)

# Create a data frame with the provided data
df <- data.frame(
  TimeOfDay = rep(c("10 AM", "4 PM", "9 PM"), each = 12),
  Position = rep(c("On the hour", "On the half hour", "Early in program", "Late in
  ↪  program"), times = 3),
  Response = c(

    ↪  42,36,62,51,37,41,68,47,41,38,64,48,62,57,88,67,60,60,85,60,58,55,81,66,100,97,127,105,96,96,12
  )
```

```r
)

df
```

```
##    TimeOfDay          Position Response
## 1      10 AM       On the hour       42
## 2      10 AM On the half hour       36
## 3      10 AM Early in program       62
## 4      10 AM  Late in program       51
## 5      10 AM       On the hour       37
## 6      10 AM On the half hour       41
## 7      10 AM Early in program       68
## 8      10 AM  Late in program       47
## 9      10 AM       On the hour       41
## 10     10 AM On the half hour       38
## 11     10 AM Early in program       64
## 12     10 AM  Late in program       48
## 13      4 PM       On the hour       62
## 14      4 PM On the half hour       57
## 15      4 PM Early in program       88
## 16      4 PM  Late in program       67
## 17      4 PM       On the hour       60
## 18      4 PM On the half hour       60
## 19      4 PM Early in program       85
## 20      4 PM  Late in program       60
## 21      4 PM       On the hour       58
## 22      4 PM On the half hour       55
## 23      4 PM Early in program       81
## 24      4 PM  Late in program       66
## 25      9 PM       On the hour      100
## 26      9 PM On the half hour       97
## 27      9 PM Early in program      127
## 28      9 PM  Late in program      105
## 29      9 PM       On the hour       96
## 30      9 PM On the half hour       96
## 31      9 PM Early in program      120
## 32      9 PM  Late in program      101
## 33      9 PM       On the hour      103
## 34      9 PM On the half hour      101
## 35      9 PM Early in program      126
## 36      9 PM  Late in program      107
```

```r
ggplot(data=df, aes(x=TimeOfDay, y = Response, col = Position, group = Position)) +
↪  stat_summary(fun = mean, geom = "line")
```



Interaction with alpha = 0.05:

Here the null hypothesis is that both time and position of advertisement have no effect on each other. The alternative hypothesis is that these DO have an effect on each other:

```
model2 <- aov(Response ~ Position*TimeOfDay, data = df)

summary(model2)
```

```
##                    Df Sum Sq Mean Sq  F value   Pr(>F)
## Position            3   3989    1330  149.137 1.19e-15 ***
## TimeOfDay           2  21561   10780 1209.022  < 2e-16 ***
## Position:TimeOfDay  6     25       4    0.474    0.821
## Residuals          24    214       9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the comparison p value of 0.82 is NOT less than 0.05, we accept the null hypothesis, i.e. both time and position of advertisement have no effect on each other.

Significance of time of day effects with alpha = 0.05:

The null hypothesis is that the time of day for an advertisement has no impact on the responses. The alternative hypothesis is that the time of day does have significance on responses. As seen in the ANOVA table, the p value is very small, (2*10^-16) which is less than 0.05, so we reject the null hypothesis. This makes sense, as there is clearly a correlation between the variables as seen on the graph (the slope of the lines are all positive).

Significance of position of advertisement effects with alpha = 0.05:

The null hypothesis is that the position of advertisement has no impact on the number of responses. The alternative hypothesis is that the position of advertisement has significance on the number of responses. As the p value of 1.19*10^-15 is very small and less than 0.05, we reject the null hypothesis. This makes sense, as there is clearly a correlation between the variables (each line has a different height for all times that never crosses).

Pairwise comparison of the four ad positions with Tukey simultaneous 95% confidence interval:

```
TukeyHSD(model2, "Position", conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Response ~ Position * TimeOfDay, data = df)
##
## $Position
##                                          diff        lwr        upr     p adj
## Late in program-Early in program  -18.777778 -22.660936 -14.894619 0.0000000
## On the half hour-Early in program -26.666667 -30.549825 -22.783508 0.0000000
## On the hour-Early in program      -24.666667 -28.549825 -20.783508 0.0000000
## On the half hour-Late in program   -7.888889 -11.772047  -4.005730 0.0000509
## On the hour-Late in program        -5.888889  -9.772047  -2.005730 0.0017611
## On the hour-On the half hour        2.000000  -1.883159   5.883159 0.4991417
```

Pairwise comparison of the morning, afternoon, and evening times with Tukey simultaneous 95% confidence interval:

```
TukeyHSD(model2, "TimeOfDay", conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Response ~ Position * TimeOfDay, data = df)
##
## $TimeOfDay
##                diff      lwr      upr p adj
## 4 PM-10 AM 18.66667 15.62232 21.71101     0
## 9 PM-10 AM 58.66667 55.62232 61.71101     0
## 9 PM-4 PM  40.00000 36.95565 43.04435     0
```

Comparing time of day with advertisement position for maximizing consumer response:

We could compare using the Tukey tables, but the graph shows a much clearer relationship between time of day and advertisement position with consumer response. From this, we can see that earlier in the program has higher response rates. Additionally, we can see that the response rates are higher at 9pm than any other time, and as these variables do not influence each other, the best combination of these variables is just the best selection from both categories. So earlier in the program and 9pm are the best times to get higher response rates for ads.