# STAT 2509B4
Assignment 4

Krystian Wojcicki, 101001444

Winter 2020

1. **Christmas week is a critical period for most ski resorts. Because many students and adults are free, they are able to spend several days indulging in their favorite pastime, skiing. A ski resort in Vermont wanted to determine the effect that weather had on their sales of lift tickets. The manager of the resort collected the number of lift tickets sold during the Christmas week (y), the total snowfall ($x_1$) and the average temperature ($x_2$) for the past 20 years ($x_3$). The TSS for the full model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$ is : TSS $= 56601012$ We decided to screen the independent variables to determine the best set for predicting the lift tickets sales. The sums of squares for all possible regression models were found to be as follows:**

| Independent variables in the model | SSR | SSE | d.f. $_{SSE}$ | MSE |
|---|---|---|---|---|
| $X_1$ | 6 215 561 | 50 385 451 | 18 | 2 799 192 |
| $X_2$ | 665 121 | 55 935 892 | 18 | 3 107 550 |
| $X_3$ | 35 788 320 | 20 812 692 | 18 | 1 156 261 |
| $X_1, X_2$ | 6 793 798 | 49 807 214 | 17 | 2 929 836 |
| $X_1, X_3$ | 41 296 990 | 15 304 022 | 17 | 900 237 |
| $X_2, X_3$ | 36 518 115 | 20 082 897 | 17 | 1 181 347 |
| $X_1, X_2, X_3$ | 41 940 217 | 14 660 795 | 16 | 916 300 |

(a) **Determine the subset of variables that is selected as best by the Forward Selection Procedure using $F_0^* = 4.2$ (to-add-variable). Show your steps.**

(1) Fit all one term models: $y = \beta_0 + \beta_1 x_j + \epsilon$ for $j = 1, 2, 3$
$\text{SSR}(X_1) = 6215561$
$\text{SSR}(X_2) = 665121$
$\text{SSR}(X_3) = 35788320 \Rightarrow \max$

Therefore $F_3 = \frac{\text{MSR}(X_3)}{\text{MSE}(X_3)} = \frac{\text{SSR}(X_3)/1}{\text{SSE}(X_3)/18} = \frac{35788320/1}{20812692/18} = 30.952$
Since $F_3 = 30.952 > F_0 = 4.2$ we keep $X_3$

(2) Fit all two term models $y = \beta_0 + \beta_1 x_3 + \beta_2 x_j + \epsilon$ for $j = 1, 2$

Calculate $\text{SSR}(X_j|X_3)$

$\text{SSR}(X_1|X_3) = \text{SSR}(X_1, X_3) - \text{SSR}(X_3) = 41296990 - 35788320 = 5508670 \Rightarrow \max$
$\text{SSR}(X_2|X_3) = \text{SSR}(X_2, X_3) - \text{SSR}(X_3) = 36518115 - 35788320 = 729795$
Therefore $F_1 = \frac{\text{MSR}(X_1|X_3)}{\text{MSE}(X_1,X_3)} = \frac{[\text{SSR}(X_1,X_3) - \text{SSR}(X_3)]/[df_{SSR(X_1,X_3)} - df_{SSR(X_3)}]}{\text{SSE}(X_1,X_3)/df_{SSE(X_1,X_3)}} = \frac{5508670/(2-1)}{15304022/17} = 6.1191358716$
Since $F_1 = 6.1191358716 > 4.2$ we keep $X_1, X_3$

(3) Fit the full model $y = \beta_0 + \beta_1 x_3 + \beta_2 x_1 + \beta_3 x_3 + \epsilon$
Calculate $\text{SSR}(X_2|X_1, X_3) = \text{SSR}(X_1, X_2, X_3) - \text{SSR}(X_1, X_3) = 41940217 - 41296990 = 643227$
Therefore $F_2 = \frac{\text{MSR}(X_2|X_1,X_3)}{\text{MSE}(X_1,X_2,X_3)} = \frac{[\text{SSR}(X_1,X_2,X_3) - \text{SSR}(X_1,X_3)]/[df_{SSR(X_1,X_2,X_3)} - df_{SSR(X_1,X_3)}]}{\text{SSE}(X_1,X_2,X_3)/df_{SSE(X_1,X_2,X_3)}} = \frac{643227/(3-2)}{14660795/16} = 0.701983214416$
Since $F_2 \leq F_0^*$ we do not add $X_2$ and only keep $\{X_1, X_3\}$ and terminate the process.

Therefore the best set is $\{X_1, X_3\}$ and the model is $y = \beta_0 + \beta_1 x_3 + \beta_2 x_1 + \epsilon$.

(b) **Determine the subset of variables that is selected as best by the Backward Elimination Procedure using $F_0^{**} = 4.1$ (to-delete-variable). Show your steps. NOTE: $(t_0^{**})^2 = F_0^{**}$**

Fit the full model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$ and check whether model is significant or not at $\alpha = 5\%$.

$F = \frac{\text{MSR}_f}{\text{MSE}_f} = \frac{\text{SSR}_f/3}{\text{SSE}_f/16} = \frac{41940217/3}{14660795/16} = 15.2570960397$

Since $F = 15.2570960397 > F_{(3,16);0.05} = 3.24$, we can conclude that at 5% level of sigificance the full model is significant and can be used

(1) Calculate $F_j = (t_j)^2 = \frac{\text{MSR}(X_j|\text{all}X\text{'s except}X_j)}{\text{MSE}(X_1,X_2,X_3)} = \frac{[SSR_f - SSR(\text{all}X\text{'s except}X_j)]/df}{MSE_f}$ for $j = 1, 2, 3$

$F_1 = \frac{MSR(X_1|X_2,X_3)}{MSE(X_1,X_2,X_3)} = \frac{[SSR(X_1,X_2,X_3) - SSR(X_2,X_3)]/[df_{SSR(X_1,X_2,X_3)} - df_{SSR(X_2,X_3)}]}{SSE(X_1,X_2,X_3)/df_{SSE(X_1,X_2,X_3)}} = \frac{[41940217 - 36518115]/[3-2]}{14660795/16} = 5.91738933666$

$F_2 = \frac{MSR(X_2|X_1,X_3)}{MSE(X_1,X_2,X_3)} = \frac{[SSR(X_1,X_2,X_3) - SSR(X_1,X_3)]/[df_{SSR(X_1,X_2,X_3)} - df_{SSR(X_1,X_3)}]}{SSE(X_1,X_2,X_3)/df_{SSE(X_1,X_2,X_3)}} = \frac{[41940217 - 41296990]/[3-2]}{14660795/16} = 0.701983214416 \Leftarrow \min$

$F_3 = \frac{MSR(X_3|X_1,X_2)}{MSE(X_1,X_2,X_3)} = \frac{[SSR(X_1,X_2,X_3) - SSR(X_1,X_2)]/[df_{SSR(X_1,X_2,X_3)} - df_{SSR(X_1,X_2)}]}{SSE(X_1,X_2,X_3)/df_{SSE(X_1,X_2,X_3)}} = \frac{[41940217 - 6793798]/[3-2]}{14660795/16} = 38.3569038378$

Since $F_2 \not> 4.1$ we remove $x_2$ from the model.

(2) Fit model $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon$

$F_1 = \frac{MSR(X_1|X_3)}{MSE(X_1,X_3)} = \frac{SSR(X_1,X_3) - SSR(X_3)}{SSE(X_1,X_3)/df_{SSE(X_1,X_3)}} = \frac{(41296990 - 35788320)}{(15304022/17)} = 6.1191 \Leftarrow \min$

$F_3 = \frac{MSR(X_3|X_1)}{MSE(X_1,X_3)} = \frac{SSR(X_1,X_3) - SSR(X_1)}{SSE(X_1,X_3)/df_{SSE(X_1,X_3)}} = \frac{(41296990 - 6215561)}{(15304022/17)} = 38.9691215159$

Since $F_1 = 6.1191 > 4.1$ we remove nothing and keep the current model as the final model and terminate the procedure.

Therefore the best set is $\{X_1, X_3\}$ and the model is $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon$.

(c) **Determine the subset of variables that is selected as best by the Stepwise Regression Procedure using $F_0^* = 4.2$ (to-add) and $F_0^{**} = 4.1$ (to-delete). Show your steps.**

(1) Fit all one term models: $y = \beta_0 + \beta_1 x_j + \epsilon$ for $j = 1, 2, 3$ As we saw in forward selection in part (a) we know that we keep $X_3$

(2) Fit all two term models $y = \beta_0 + \beta_1 x_3 + \beta_2 x_j + \epsilon$ for $j = 1, 2$ As we saw in forward selection in part (a) we know that we keep $X_1$

Now check if after adding $X_1$, $X_3$ became insignificant.

$F_3 = \frac{SSR(X_3|X_1)}{MSE(X_1,X_3)} = \frac{SSR(X_1,X_3) - SSR(X_1)}{SSE(X_1,X_3)/df_{SSE(X_1,X_3)}} = \frac{41296990 - 6215561}{15304022/17} = 38.9691215159$

Since $F_3 = 38.9691215159 > F_0^{**} = 4.1$ we keep $X_3$ and keep $X_1$.

(3) Fit the full model: $y = \beta_0 + \beta_1 x_3 + \beta_2 x_1 + \beta_3 x_2 + \epsilon$. As we saw in forward selection in part (a) we know that we don't add $X_2$ and there is no need to check if adding $X_2$ makes any variable redundant. Therefore the best set is $\{X_1, X_3\}$.

2. **A quality engineer in a company manufacturing electronic audio equipment was inspecting a new type of battery that was being considered for use. A batch of 20 batteries was randomly assigned to four groups (so that there were five batteries per group). Each group of batteries was then subjected to a particular pressure level – low, normal, high, very high. The batteries were simultaneously tested under these pressure levels and the times to failure (in hours) were recorded and are given below:**

| Pressure | | | |
|---|---|---|---|
| LOW | NORMAL | HIGH | VERY HIGH |
| 8.0 | 7.6 | 6.0 | 5.1 |
| 8.1 | 8.2 | 6.3 | 5.6 |
| 9.2 | 9.8 | 7.1 | 5.9 |
| 9.4 | 10.9 | 7.7 | 6.7 |
| 11.7 | 12.3 | 8.9 | 7.8 |

**Establish whether the average times to battery failure are the same for the four pressure levels; if not, do a follow-up analysis to determine which are the same and which differ. (Use $\alpha = 0.10$). List**

**all necessary assumptions and indicate which might be suspect. Also perform a nonparametric analysis. Verify your results using SAS.**

C.R.D Assume

1) 4 independent random samples of batteries (given)

2) 4 normally distributed battery groups

3) with equal variance, $\sigma^2$ (?, potentially need to check)

To check the assumption of equal variance using Hartleys test we need $s_i^2$'s for $j = 1, 2, 3, 4$ where $n_1 = n_2 = n_3 = n_4 = 5, k = 4, \bar{n} = 5, [\bar{n}] = 5, n = 20$

$s_1^2 = \frac{\sum_{j=1}^{n_1} y_{1j}^2 - \frac{(\sum_{j=1}^{n_1} y_{1j})^2}{n_1}}{n_1 - 1} = \frac{439.5 - 46.4^2/5}{4} = 2.227$

$s_2^2 = \frac{\sum_{j=1}^{n_2} y_{2j}^2 - \frac{(\sum_{j=1}^{n_2} y_{2j})^2}{n_2}}{n_2 - 1} = \frac{491.14 - 48.8^2/5}{4} = 3.713 \Leftarrow \max$

$s_3^2 = \frac{\sum_{j=1}^{n_3} y_{3j}^2 - \frac{(\sum_{j=1}^{n_3} y_{3j})^2}{n_3}}{n_3 - 1} = \frac{264.6 - 36^2/5}{4} = 1.35$

$s_4^2 = \frac{\sum_{j=1}^{n_4} y_{4j}^2 - \frac{(\sum_{j=1}^{n_4} y_{4j})^2}{n_4}}{n_4 - 1} = \frac{197.91 - 31.1^2/5}{4} = 1.117 \Leftarrow \min$

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$
$H_a$ : atleast one of the $\sigma^2$'s $\neq$
$\alpha = 0.01$ or $0.05$

Test statistic $F_{max} = \frac{s_{max}^2}{s_{min}^2} = 3.713/1.117 = 3.324$

Rejection Region we reject $H_0$ if $F_{max} > F_{max(k,[\bar{n}]-1);\alpha} = \begin{cases} F_{max(4,4);0.01} = 49 \\ F_{max(4,4);0.05} = 20.6 \end{cases}$

Since 3.324 is not greater than 20.6 (or 49) we can conclude that at a 1% (or 5%) level of significance there is no evidence to say that the variances are not equal (i.e we have equal variance). Therefore we may proceed with the main test

$G.T. = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} = \sum_{i=1}^{k} T_i = 162.3$
$TSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{(G.T)^2}{n} = 1393.15 - \frac{162.3^2}{20} = 76.0855$
$SST_r = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - \frac{(G.T)^2}{n} = [(46.4)^2/5 + (48.8)^2/5 + (36^2)/5 + (31.1)^2/5] - \frac{162.3^2}{20} = 42.4575$
$SSE = TSS - SST_r = 76.0855 - 42.4575 = 33.628$
$MST_r = \frac{SST_r}{k-1} = 42.4575/3 = 14.1525$
$MSE = \frac{SSE}{n-k} = 33.628/16 = 2.10175$
$F_T = \frac{MST_r}{MSE} = 14.1525/2.10175 = 6.73367431902 = 6.733$

| Source | d.f | SS | MS | F |
|---|---|---|---|---|
| Regression | 3 | 42.4575 | 14.1525 | 6.733 |
| Error | 16 | 33.628 | 2.10175 | |
| Total | 19 | 76.0855 | | |

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
$H_a$ : at least one of the $\mu' s \neq$
$\alpha = 0.10$

Test statistic $F_T = \frac{MST_r}{MSE} = 6.733$

Rejection Region: we reject $H_0$ if $F_T > F_{(k-1,n-k);\alpha} = F_{(3,16);0.10} = 2.460$

Since $F_T = 6.733 > 2.460$ we reject $H_0$ and conclude that at a 10% level of significance there is evidence to say the mean time to battery failure is different among the four different pressure levels.

**Which are different? Use Tukey's h.s.d**

1) $\binom{k}{2} = \binom{4}{2} = 6$ pairs of $|\bar{y_i} - \bar{y_j}|$
$H_0 : \mu_i = \mu_j$
$H_a : \mu_i \neq \mu_j$
for i,j = 1,2,3,4 and i ≠ j

2) h.s.d $= q_{\alpha(k,v)} \times \sqrt{\frac{MSE}{2}(\frac{1}{n_i} + \frac{1}{n_j})} = q_{0.10(4,16)} \times \sqrt{\frac{2.10175}{2}\frac{2}{5}} = (3.52003)(0.64834404447) = 2.28219048686 = 2.282$
$\bar{y_1} = \frac{T_1}{n_1} = 46.4/5 = 9.28$
$\bar{y_2} = \frac{T_2}{n_2} = 48.8/5 = 9.76$
$\bar{y_3} = \frac{T_3}{n_3} = 36/5 = 7.2$
$\bar{y_4} = \frac{T_4}{n_4} = 31.1/5 = 6.22$

3) $|\bar{y_1} - \bar{y_2}| = 0.48 < 2.282 \Rightarrow \mu_1 = \mu_2$
$|\bar{y_1} - \bar{y_3}| = 2.08 < 2.282 \Rightarrow \mu_1 = \mu_3$
$|\bar{y_1} - \bar{y_4}| = 3.06 > 2.282 \Rightarrow \mu_1 \neq \mu_4$
$|\bar{y_2} - \bar{y_3}| = 2.56 > 2.282 \Rightarrow \mu_2 \neq \mu_3$
$|\bar{y_2} - \bar{y_4}| = 3.54 > 2.282 \Rightarrow \mu_2 \neq \mu_4$
$|\bar{y_3} - \bar{y_4}| = 0.98 < 2.282 \Rightarrow \mu_3 = \mu_4$

Therefore there are differences in average time for battery failure for (low and very high), (normal and high), (normal and very high) groups. At the 10% level of significance.

**None-parametric Analysis (Kruskal-Wallis test)** Assume:

1) C.R.D (4 independent random samples from 4 treatment populations) with (given this)
2) approximately the same shape and spread (assume this)

First we need to rank the observations from smallest to the largest

| Low | Normal | High | Very High |
|---|---|---|---|
| 8.0 (11) | 7.6 (8) | 6.0 (4) | 5.1 (1) |
| 8.1 (12) | 8.2 (13) | 6.3 (5) | 5.6 (2) |
| 9.2 (15) | 9.8 (17) | 7.1 (7) | 5.9 (3) |
| 9.4 (16) | 10.9 (18) | 7.7 (9) | 6.7 (6) |
| 11.7 (19) | 12.3 (20) | 8.9 (14) | 7.8 (10) |

$T_{R_1} = 73, T_{R_2} = 76, T_{R_3} = 39, T_{R_4} = 22$
Check $\frac{n(n+1)}{2} = \frac{20\times21}{2} = 210$
$\sum_i T_{R_i} = 73 + 76 + 39 + 22 = 210$

$H_0 : Md_1 = Md_2 = Md_3 = Md_4$
$H_a$: at least one of the $Md$'s ≠.
$\alpha = 0.10$
$H = \frac{12}{n(n+1)}[\sum_{i=1}^{k} \frac{T_{R_i}^2}{n_i}] - 3(n+1) = \frac{12}{20\times21}[\frac{73^2}{5} + \frac{76^2}{5} + \frac{39^2}{5} + \frac{22^2}{5}] - 3(21) = 11.91$
Rejection region, we reject $H_0$ if $H > \chi^2_{(k-1);\alpha} = \chi^2_{3;0.10} = 6.251$
Since $H = 11.91 > 6.251$ we reject $H_0$. and conclude that at a 10% level of significance there is evidence to say that the medians vary among the 4 pressures.

**Which pressures differ? Dunes procedure:**

1) Calculate $\binom{k}{2} = \binom{4}{2} = 6$ pairs of $|\bar{R}_i - \bar{R}_j|$ for $H_0 : Md_i = Md_j$ vs $H_a : Md_i \neq Md_j$
for $i, j = 1, 2, 3, 4$ and $i \neq j$

2) Critical range $= z_{\frac{\alpha}{k(k-1)}} \times \sqrt{\frac{n(n+1)}{12}(\frac{1}{n_i}+\frac{1}{n_j})} = z_{\frac{0.1}{4(3)}} \times \sqrt{\frac{20(21)}{12}\frac{2}{5}} = 2.395(3.74) = 8.9573$

$\bar{R_1} = \frac{T_{R_1}}{n_1} = 73/5 = 14.6$
$\bar{R_2} = \frac{T_{R_2}}{n_2} = 76/5 = 15.2$
$\bar{R_3} = \frac{T_{R_3}}{n_3} = 39/5 = 7.8$
$\bar{R_4} = \frac{T_{R_4}}{n_4} = 22/5 = 4.4$

3) $|\bar{R_1} - \bar{R_2}| = 0.6 < 8.96 \Rightarrow Md_1 = Md_2$
$|\bar{R_1} - \bar{R_3}| = 6.8 < 8.96 \Rightarrow Md_1 = Md_3$
$|\bar{R_1} - \bar{R_4}| = 10.2 > 8.96 \Rightarrow Md_1 \neq Md_4$

$|\bar{R_2} - \bar{R_3}| = 7.8 < 8.96 \Rightarrow Md_2 = Md_3$
$|\bar{R_2} - \bar{R_4}| = 10.8 > 8.96 \Rightarrow Md_2 \neq Md_4$

$|\bar{R_3} - \bar{R_4}| = 3.4 < 8.96 \Rightarrow Md_3 = Md_4$

Therefore there are differences in medians of battery failure for (low and very high), (normal and very high) groups. At the 10% level of significance.

## The ANOVA Procedure

| Class Level Information | | |
|---|---|---|
| Class | Levels | Values |
| group | 4 | high low normal very_hig |

| | |
|---|---|
| Number of Observations Read | 20 |
| Number of Observations Used | 20 |

Krystian Wojcicki, 101001444

**The ANOVA Procedure**

**Dependent Variable: pressure**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 42.45750000 | 14.15250000 | 6.73 | 0.0038 |
| Error | 16 | 33.62800000 SSE | 2.10175000 | <- MSE | |
| Corrected Total | 19 | 76.08550000 | <- TSS | | |

| R-Square | Coeff Var | Root MSE | pressure Mean |
|---|---|---|---|
| 0.558024 | 17.86496 | 1.449741 | 8.115000 |

treatments ->

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| group | 3 | 42.45750000 | 14.15250000 | 6.73 | 0.0038 |

SST_R MST_R F_T

**Krystian Wojcicki, 101001444**

Values match values obtained in ANOVA table done by hand.

## The ANOVA Procedure

### Tukey's Studentized Range (HSD) Test for pressure

**Note:** This test controls the Type I experimentwise error rate.

| | |
|---|---|
| Alpha | 0.1 |
| Error Degrees of Freedom | 16 |
| Error Mean Square | 2.10175 |
| Critical Value of Studentized Range | 3.52003 |
| Minimum Significant Difference | 2.2822 |

alpha

q(0.10)(4, 12)
h.s.d

Comparisons significant at the 0.1 level are indicated by ***.

| group Comparison | Difference Between Means | Simultaneous 90% Confidence Limits | | |
|---|---|---|---|---|
| normal - low | 0.4800 | -1.8022 | 2.7622 | |
| normal - high | 2.5600 | 0.2778 | 4.8422 | *** |
| normal - very_hig | 3.5400 | 1.2578 | 5.8222 | *** |
| low - normal | -0.4800 | -2.7622 | 1.8022 | |
| low - high | 2.0800 | -0.2022 | 4.3622 | |
| low - very_hig | 3.0600 | 0.7778 | 5.3422 | *** |
| high - normal | -2.5600 | -4.8422 | -0.2778 | *** |
| high - low | -2.0800 | -4.3622 | 0.2022 | |
| high - very_hig | 0.9800 | -1.3022 | 3.2622 | |
| very_hig - normal | -3.5400 | -5.8222 | -1.2578 | *** |
| very_hig - low | -3.0600 | -5.3422 | -0.7778 | *** |
| very_hig - high | -0.9800 | -3.2622 | 1.3022 | |

As we saw in the calculations above, we conclude there is a difference in mean times of (normal vs high), (normal vs very high) and (low vs very high) at the 0.1 level of significance.

**Krystian Wojcicki, 101001444**

**The SAS System** 12:07 Thursday, April 2, 2020 4

**The NPAR1WAY Procedure**

| Wilcoxon Scores (Rank Sums) for Variable pressure Classified by Variable group | | | | | | |
|---|---|---|---|---|---|---|
| group | N | Sum of Scores | Expected Under H0 | Std Dev Under H0 | Mean Score | |
| low | 5 | 73.0 | 52.50 | 11.456439 | 14.60 | <- R_1 |
| normal | 5 | 76.0 | 52.50 | 11.456439 | 15.20 | <- R_2 |
| high | 5 | 39.0 | 52.50 | 11.456439 | 7.80 | <- R_3 |
| very_hig | 5 | 22.0 | 52.50 | 11.456439 | 4.40 | <- R_4 |

T_Ri

| Kruskal-Wallis Test | |
|---|---|
| Chi-Square | 11.9143 |
| DF | 3 |
| Pr > Chi-Square | 0.0077 |

<- chi^2 approximation to H

**Krystian Wojcicki, 101001444**

3. **An experiment was conducted by a private research corporation to investigate the toxic effects of three chemicals (I, II and III) used in the tire-manufacturing industry. In this experiment 1-inch squares of skin on rats were treated with the chemicals and then scored from 0 to 10, depending on the degree of irritation. Three adjacent 1-inch squares were marked on the back of each of eight rats, and each of the three chemicals was applied to each rat. The data are as shown in the table.**

| | Rat Number | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Chemical** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| I | 6 | 9 | 6 | 5 | 7 | 5 | 6 | 6 |
| II | 5 | 9 | 9 | 8 | 8 | 7 | 7 | 7 |
| III | 3 | 4 | 3 | 6 | 8 | 5 | 5 | 6 |

**Is it possible to be 95% certain that the toxic effects of three chemicals are not equal? Conduct the appropriate follow-up analysis (use $\alpha = 0.05$) to establish which means are significantly different. List all necessary assumptions and indicate which might be suspect. Also perform a non-parametric analysis. Verify your results using SAS.**

R.B.D Assume

1) random samples of 3 different chemicals randomly assigned to 8 different rats (given)
2) populations corresponding to each chemical-rat combination are normally distributed
3) with equal variance, $\sigma^2$ (?, potentially need to check)
4) no interactions between chemicals and rats

To check the assumption of equal variance using Hartley's test we need $s_i^2$'s for $i = 1, 2, 3$ where $n_1 = n_2 = n_3 = 8, k = 3, b = 8, \bar{n} = n = 8, [\bar{n}] = 8, n = 24$

$s_1^2 = \frac{\sum_{j=1}^{b} y_{1j}^2 - \frac{(\sum_{j=1}^{b} y_{1j})^2}{b}}{b-1} = \frac{324 - 50^2/8}{7} = 1.6428 \Leftarrow min$

$s_2^2 = \frac{\sum_{j=1}^{b} y_{2j}^2 - \frac{(\sum_{j=1}^{b} y_{2j})^2}{b}}{b-1} = \frac{462 - 60^2/8}{7} = 1.71428571429$

$s_3^2 = \frac{\sum_{j=1}^{b} y_{3j}^2 - \frac{(\sum_{j=1}^{b} y_{3j})^2}{b}}{b-1} = \frac{220 - 40^2/8}{7} = 2.85714285714 \Leftarrow max$

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$
$H_a$ : atleast one of the $\sigma^2$'s $\neq$
$\alpha = 0.05$
Test statistic $F_{max} = \frac{s_{max}^2}{s_{min}^2} = 2.85714285714/1.6428 = 1.73919092838$
Rejection Region we reject $H_0$ if $F_{max} > F_{max(k,[\bar{n}]-1);\alpha} = F_{max(3,7);0.05} = 6.94$
Since 1.739 is not greater than 6.94 we can conclude that at a 5% level of significance there is no evidence to say that the variances are not equal (i.e we have equal variance). Therefore we may proceed with the main test

$G.T. = \sum_{i=1}^{k} \sum_{j=1}^{b} y_{ij} = \sum_{i=1}^{k} T_i = 150$
$TSS = \sum_{i=1}^{k} \sum_{j=1}^{b} y_{ij}^2 - \frac{(G.T)^2}{bk} = 1006 - \frac{150^2}{24} = 68.5$
$SST_r = \sum_{i=1}^{k} \frac{T_i^2}{b} - \frac{(G.T)^2}{bk} = [(50)^2/8 + (60)^2/8 + (40)^2/8] - \frac{150^2}{24} = 25$
$SSB = \sum_{j=1}^{b} \frac{B_j^2}{k} - \frac{(G.T)^2}{bk} = [(14)^2/3 + (22)^2/3 + (18)^2/3 + (19)^2/3 + (23)^2/3 + (17)^2/3 + (18)^2/3 + (19)^2/3] - \frac{150^2}{24} = 18.5$
$SSE = TSS - SST_r - SSB = 68.5 - 25 - 18.5 = 25$

$MST_r = \frac{SST_r}{k-1} = 25/2 = 12.5$
$MSB = \frac{SSB}{b-1} = 18.5/7 = 2.643$
$MSE = \frac{SSE}{(b-1)(k-1)} = 25/14 = 1.786$
$F_T = \frac{MST_r}{MSE} = 12.5/1.786 = 7.00$
$F_B = \frac{MSB}{MSE} = 2.643/1.786 = 1.48$

| Source | d.f | SS | MS | F |
|---|---|---|---|---|
| Treatments | 2 | 25 | 12.5 | 7 |
| Blocks | 7 | 18.5 | 2.643 | 1.48 |
| Error | 14 | 25 | 1.786 | |
| Total | 23 | 68.5 | | |

**Test for difference in block means**

$H_0 : \beta_1 = ... = \beta_8$
$H_a$ : Not all $\beta$'s are equal
$\alpha = 0.05$

Test statistic $F_B = \frac{MSB}{MSE} = 1.48$

Rejection Region: we reject $H_0$ if $F_B > F_{(b-1)(bk-k-b+1);\alpha} = F_{(7)(14);0.05} = 2.76$

Since $F_B = 1.48 < 2.76$ we do not have enough evidence to reject $H_0$ at a 5% level of signifiance and conclude that there is no difference in mean between rats. This suggests that RBD is not more effective than CRD in this case.

**Test for treatment means**

$H_0 : \mu_1 = \mu_2 = \mu_3$
$H_a$ : at least one of the $\mu's \neq$
$\alpha = 0.05$

Test statistic $F_T = \frac{MST_r}{MSE} = 7$

Rejection Region: we reject $H_0$ if $F_T > F_{(k-1,(b-1)(k-1));\alpha} = F_{(2,14);0.05} = 3.740$

Since $F_T = 7 > 3.740$ we reject $H_0$ and conclude that at a 5% level of significance there is evidence to say that there are differences between the three chemicals means.

**Which are different? Use Tukey's h.s.d**

1) $\binom{k}{2} = \binom{3}{2} = 3$ pairs of $|\bar{y_i} - \bar{y_j}|$
   $H_0 : \mu_i = \mu_j$
   $H_a : \mu_i \neq \mu_j$
   for i,j = 1,2,3 and i $\neq$ j
2) h.s.d $= q_{\alpha(k,(b-1)(k-1))} \times \sqrt{\frac{MSE}{b}} = q_{0.05(3,14)} \times \sqrt{\frac{1.786}{8}} = (3.70128)(0.4725) = 1.7487$
   $\bar{y_1} = \frac{T_1}{b} = 50/8 = 6.25$
   $\bar{y_2} = \frac{T_2}{b} = 60/8 = 7.5$
   $\bar{y_3} = \frac{T_3}{b} = 40/8 = 5$
3) $|\bar{y_1} - \bar{y_2}| = 1.25 < 1.7487 \Rightarrow \mu_1 = \mu_2$
   $|\bar{y_1} - \bar{y_3}| = 1.25 < 1.7487 \Rightarrow \mu_1 = \mu_3$
   $|\bar{y_2} - \bar{y_3}| = 2.5 > 1.7487 \Rightarrow \mu_2 \neq \mu_3$

Therefore there are differences in mean irritation for chemical groups (2 & 3). At a 5% level of significance.

**Non-parametric Analysis Friedman-Rank Test**
Assume

1) R.B.D (given)
2) in each chemical-rat number combination we have populations with approximately the same shape and spread
3) no interactions between chemical and rat number

First we need to rank the observations from smallest to the largest within each block

| Chemical | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| I | 6 (3) | 9 (2.5) | 6 (2) | 5 (1) | 7 (1) | 5 (1.5) | 6 (2) | 6 (1.5) $T_{R_1} = 14.5$ |
| II | 5 (2) | 9 (2.5) | 9 (3) | 8 (3) | 8 (2.5) | 7 (3) | 7 (3) | 7 (3) $T_{R_2} = 22$ |
| III | 3 (1) | 4 (1) | 3 (1) | 6 (2) | 8 (2.5) | 5 (1.5) | 5 (1) | 6 (1.5) $T_{R_3} = 11.5$ |

Check $\frac{bk(k+1)}{2} = \frac{24(4)}{2} = 48$
$\sum_{i=1}^{3} T_{R_i} = 14.5 + 22 + 11.5 = 48$

$H_0 : Md_1 = Md_2 = Md_3$
$H_a$ : at least one of the $Md$'s $\neq$
$\alpha = 0.05$

Test statistic $F_R = \frac{12}{bk(k+1)}[\sum_{i=1}^{3} T_{R_i}^2] - 3b(k+1) = \frac{12}{96}[14.5^2 + 22^2 + 11.5^2] - 3(8)(4) = 7.3125$

Rejection region, we reject $H_0$ if $F_R > \chi^2_{(k-1);\alpha} = \chi^2_{(2;0.05)} = 5.991$

Since $F_R = 7.3125 > 5.991$ we reject $H_0$ and conclude that at a 5% level of significance that there is enough evidence to indicate that median of the 3 chemicals differ.

**Which differ? Nemenyi's procedure.**

1) Calculate $\binom{k}{2} = \binom{3}{2} = 3$ pairs of $|\bar{R}_i - \bar{R}_j|$ for $H_0 : Md_i = Md_j$ vs $H_a : Md_i \neq Md_j$ for all $i, j = 1, 2, 3$ and $i \neq j$.

2) Critical value $N_e = q_{\alpha(k,\infty)} \times \sqrt{\frac{k(k+1)}{12b}} = q_{0.05(3,\infty)} \times \sqrt{\frac{3(4)}{12\times 8}} = 3.314(0.3536) = 1.1717$
$\bar{R}_1 = \frac{T_{R_1}}{b} = 14.5/8 = 1.8125$
$\bar{R}_2 = \frac{T_{R_2}}{b} = 22/8 = 2.75$
$\bar{R}_3 = \frac{T_{R_3}}{b} = 11.5/8 = 1.4375$

3) $|\bar{R}_1 - \bar{R}_2| = 0.9375 < 1.1717 \Rightarrow Md_1 = Md_2$
$|\bar{R}_1 - \bar{R}_3| = 0.375 < 1.1717 \Rightarrow Md_1 = Md_3$
$|\bar{R}_2 - \bar{R}_3| = 1.3125 > 1.1717 \Rightarrow Md_2 \neq Md_3$

Therefore there are differences in the median of the irritation for (chemical 2 and chemical 3). At the 5% level of significance.

## The ANOVA Procedure

| Class Level Information | | |
|---|---|---|
| Class | Levels | Values |
| rat_number | 8 | 1 2 3 4 5 6 7 8 |
| chemical | 3 | 1 2 3 |

| | |
|---|---|
| Number of Observations Read | 24 |
| Number of Observations Used | 24 |

Krystian Wojcicki, 101001444

The ANOVA Procedure

Dependent Variable: irritation

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 9 | 43.50000000 | 4.83333333 | 2.71 | 0.0463 |
| Error | 14 | 25.00000000 | 1.78571429 | <- MSE | |
| Corrected Total | 23 | 68.50000000 | <- TSS | | |

SSE

| R-Square | Coeff Var | Root MSE | irritation Mean |
|---|---|---|---|
| 0.635036 | 21.38090 | 1.336306 | 6.250000 |

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| rat_number | 7 | 18.50000000 | 2.64285714 | 1.48 | 0.2518 |
| chemical | 2 | 25.00000000 | 12.50000000 | 7.00 | 0.0078 |

blocks
treatments

SSB MSB F_B
SST_R MST_R F_T

**Krystian Wojcicki, 101001444**

Values match values obtained in ANOVA table done by hand.

### The ANOVA Procedure

### Tukey's Studentized Range (HSD) Test for irritation

**Note:** This test controls the Type I experimentwise error rate.

| | |
|---|---|
| Alpha | 0.05 |
| Error Degrees of Freedom | 14 |
| Error Mean Square | 1.785714 |
| Critical Value of Studentized Range | 3.70128 |
| Minimum Significant Difference | 1.7487 |

alpha

q(0.05)(3,14)

h.s.d

Comparisons significant at the 0.05 level are indicated by ***.

| chemical Comparison | Difference Between Means | Simultaneous 95% Confidence Limits | | |
|---|---|---|---|---|
| 2 - 1 | 1.2500 | -0.4987 | 2.9987 | |
| 2 - 3 | 2.5000 | 0.7513 | 4.2487 | *** |
| 1 - 2 | -1.2500 | -2.9987 | 0.4987 | |
| 1 - 3 | 1.2500 | -0.4987 | 2.9987 | |
| 3 - 2 | -2.5000 | -4.2487 | -0.7513 | *** |
| 3 - 1 | -1.2500 | -2.9987 | 0.4987 | |

As we concluded previously the difference in means occurs for chemicals (2 & 3)

**Krystian Wojcicki, 101001444**

The FREQ Procedure

Summary Statistics for chemical by irritation
Controlling for rat_number

| Cochran-Mantel-Haenszel Statistics (Based on Rank Scores) | | | | |
|---|---|---|---|---|
| Statistic | Alternative Hypothesis | DF | Value | Prob |
| 1 | Nonzero Correlation | 1 | 0.6429 | 0.4227 |
| 2 | Row Mean Scores Differ | 2 | 8.3571 | 0.0153 |

Chi^2 approx of F_R

Total Sample Size = 24

```
Footnote 'Krystian Wojcicki, 101001444';
data battery;
input group$ pressure @@;
cards;
    low 8.0 low 8.1 low 9.2 low 9.4 low 11.7
    normal 7.6 normal 8.2 normal 9.8 normal 10.9 normal 12.3
    high 6.0 high 6.3 high 7.1 high 7.7 high 8.9
    very_high 5.1 very_high 5.6 very_high 5.9 very_high 6.7 very_high 7.8
run;

ods pdf file="a4-output.pdf";
ods graphics off;

proc anova;
    class group;
    model pressure=group;
    means group/tukey cldiff alpha=0.10;
run;
proc NPAR1WAY WILCOXON;
    class group;
run;


data chemicals;
input chemical rat_number irritation @@;
cards;
    1 1 6 1 2 9 1 3 6 1 4 5 1 5 7 1 6 5 1 7 6 1 8 6
    2 1 5 2 2 9 2 3 9 2 4 8 2 5 8 2 6 7 2 7 7 2 8 7
    3 1 3 3 2 4 3 3 3 3 4 6 3 5 8 3 6 5 3 7 5 3 8 6
run;
proc anova;
    class rat_number chemical;
    model irritation=rat_number chemical;
    means chemical/tukey cldiff alpha=0.05;
run;
proc freq;
    tables rat_number*chemical*irritation/CMH2 SCORES=RANK NOPRINT;
run;

ods pdf close
```