TMA4255 Applied Statistics
Solution to Exercise 9

Problem 1

$$A_1, \ldots, A_4 = \text{workers (added as } 1, \ldots, 4 \text{ in C2)}$$
$$M_1, \ldots, M_4 = \text{machines (added as } 1, \ldots, 4 \text{ in C3)}$$

a) We assume that the skills of the workers do not influence the production units. This means we have one-way grouping, and we assume the model

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}, \quad \sum_j \alpha_j = 0$$

Here:

- Y_{ij}: number of produced units by machine j and worker i.
- $E(Y_{ij}) = \mu + \alpha_j$.
- ϵ_{ij} assumed independent and $\sim N(0, \sigma^2)\ \forall i, j$.
- α_j is a factor which is special for machine j.
- μ: "average effect"

Wish to test wether the machines have different capacities:

$$\begin{aligned} H_0 &: \quad \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0 \\ H_1 &: \quad \text{at least one not equal.} \end{aligned}$$

The total variation in the data $SS_{\text{tot}} = \sum_{j=1}^4 \sum_{i=1}^4 (Y_{ij} - \bar{Y}_{\cdot\cdot})^2$, can be written as a sum of two sums of squares:[Theorem. 13.1]

$$SS_{\text{tot}} = SS_A + SS_E = \sum_{j=1}^4 4(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2 + \sum_{j=1}^4 \sum_{i=1}^4 (Y_{ij} - \bar{Y}_{\cdot j})^2$$

It can be shown that [Teorem 13.2]

$$\begin{aligned} E(SS_A) &= (4-1)\sigma^2 + \sum_{i=1}^4 4\alpha_i^2 = 3\sigma^2 + 4\sum \alpha_i^2 \\ E(SS_E) &= (16-4)\sigma^2 \\ F &= \frac{MS_A}{MS_E} = \frac{SS_A/(4-1)}{SS_E/(16-4)} \sim F_{(4-1),(16-4)} = F_{3,12} \end{aligned}$$

We see that if H_0 is correct, we can expect an $F_{0\ \text{obs}}$ of about 1. If H_0 is wrong, we can expect a big value of $F_{0\ \text{obs}}$.
Minitab gives us:

```
One-way Analysis of Variance
Analysis of Variance for Data

Source     DF        SS        MS        F        P
M           3      72,0      24,0     1,58    0,245
Error      12     182,0      15,2
Total      15     254,0
                                   Individual 95% CIs For Mean
                                   Based on Pooled StDev
Level       N      Mean     StDev  -+---------+---------+---------+-----
1           4    72,000     2,944  (----------*----------)
2           4    75,000     3,162          (---------*----------)
3           4    77,000     4,243               (---------*----------)
4           4    72,000     4,899  (----------*----------)
                                   -+---------+---------+---------+-----
Pooled StDev =    3,894          68,0      72,0      76,0      80,0
```

Here we have that:

$$F_{0\text{ obs}} = \frac{SS_A/3}{SS_E/12} = \frac{24.0}{15.2} = 1.58$$

the p-value:

$$p = P(F_{3,12} > F_{0\text{ obs}}) = P(F_{3,12} > 1.58) = 0.245$$

): p is larger than any reasonable significance level α, which means we can not reject H_0, and claim that there is a difference between the machines.

b) Now we assume that skills of the workers have an influence. Model:

$$X_{ij} = \mu + \alpha_j + \beta_i + \epsilon_{ij}, \quad \sum_j \alpha_j = \sum_i \beta_i = 0$$

We have:

- X_{ij}: The number of produced units with machine j and worker i.
- ϵ_{ij} assumed independent and $\sim N(0, \sigma^2)\ \forall i, j$.
- α_j is a factor which is special for machine j.
- β_i is a factor which is special for worker i.
- μ: "average effect"

We have the same hypothesis test as in **a)**: H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ against H_1: at least one is different.

We split the total variation into three sums of squares

$$SS_{\text{tot}} = SS_{\text{mask}} + SS_{\text{arb}} + SS_E$$

$$\Updownarrow$$

$$\sum_{j=1}^{4}\sum_{i=1}^{4}(X_{ij} - \bar{X}_{\cdot\cdot})^2 = 4\sum_{j=1}^{4}(\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})^2 + \sum_{i=1}^{4}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 + \sum_{j=1}^{4}\sum_{i=1}^{4}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot})^2$$

The same type of argument as in **a)** tells us that we can expect a big value of $F_{0\text{ obs}}$ if H_0 is wrong.

Here

$$F = \frac{SS_{\text{mask}}/(4-1)}{SS_E/((4-1)(4-1))} \sim F_{4-1,(4-1)(4-1)} = F_{3,9}$$

Minitab gives:

```
Two-way Analysis of Variance

Analysis of Variance for Data
Source        DF        SS         MS        F        P
A              3    160,00      53,33    21,82    0,000
M              3     72,00      24,00     9,82    0,003
Error          9     22,00       2,44
Total         15    254,00
```

And we see that

$$F_{0\text{ obs}} = \frac{72.0/3}{22.0/9} = 9.82,$$

and this gives p-value: $p = P(F_{3,9} > 9.82) = 0.003$.

): We have a small p and we reject H_0.

c) Expected number of produced units from machine M_2:

$$\mu_{\cdot 2} = E(X_{\cdot 2}) = \mu + \alpha_2$$

Estimator: $\hat{\mu}_{\cdot 2} = \frac{1}{4}\sum_{i=1}^{4} X_{i2}$. This gives the point estimate: $\hat{\mu}_{\cdot 2} = \frac{1}{4}(77 + 71 + 78 + 74) = 75$.

We have:

$$E(\hat{\mu}_{\cdot 2}) = \frac{1}{4}\sum_{i=1}^{4} E(X_{i2}) = \frac{1}{4}\sum_{i=1}^{4}(\mu + \alpha_2 + \beta_i) = \mu + \alpha_2 + \frac{1}{4}\sum_{i=1}^{4}\beta_i = \mu + \alpha_2$$

and

$$\begin{aligned}
\text{Var}(\hat{\mu}_{\cdot 2}) &= E[((\hat{\mu}_{\cdot 2}) - E(\hat{\mu}_{\cdot 2}))^2] = E\left[\left(\frac{1}{4}\sum(Y_{i2} - E(Y_{i2}))\right)^2\right] \\
&= E\left[\left(\frac{1}{4}\sum_{i=1}^{4}\epsilon_{i2}\right)^2\right] = \frac{1}{16}\sum_{i=1}^{4} E(\epsilon_{i2})^2 = \left(\frac{1}{16}\right)^2 \sum_{i=1}^{4}\text{Var}(\epsilon_{i2}) \\
&= \frac{1}{4}\sigma^2
\end{aligned}$$

Therefore we get $\hat{\mu}_{\cdot 2} \sim N(\mu + \alpha_2, \frac{1}{4}\sigma^2) \Rightarrow \frac{\hat{\mu}_{\cdot 2} - (\mu + \alpha_2)}{\sigma/2} \sim N(0, 1)$.

σ^2 is estimated in **b)** as $S^2 = \frac{1}{9}SS_E$.

Now we have:

$$\frac{\hat{\mu}_{\cdot 2} - (\mu + \alpha_2)}{S/2} \sim T_9$$

(same number of degrees of freedom as SS_E).

$(1-\alpha)\cdot 100$ % confidence interval:

$$P\left(-t_{\alpha/2,9} \leqslant \frac{\hat{\mu}_{\cdot 2} - (\mu+\alpha_2)}{S/2} \leqslant t_{\alpha/2,9}\right) = 1-\alpha$$
$$\Updownarrow$$
$$P(\hat{\mu}_{\cdot 2} - t_{\alpha/2,9}S/2 \leqslant \mu+\alpha_2 \leqslant \hat{\mu}_{\cdot 2} + t_{\alpha/2,9}S/2) = 1-\alpha$$

With numbers: $\hat{\mu}_{\cdot 2} = 75$, $\alpha = 0.1$, $t_{0.05,4} = 1.83$, $S^2 = 2.444$.

): 90 % confidence interval $\mu+\alpha_2$: $[73.6, 76.4]$.

Problem 2

a) We do the analysis in MINITAB:

```
Estimated Effects and Coefficients for C9 (coded units)

Term       Effect    Coef
Constant           17,544
A           8,837   4,419
B          -2,512  -1,256
C          -1,087  -0,544
D           0,112   0,056
A*B        -0,762  -0,381
A*C         1,013   0,506
A*D         0,212   0,106
B*C         1,012   0,506
B*D         0,262   0,131
C*D        -0,162  -0,081
A*B*C       0,213   0,106
A*B*D      -0,038  -0,019
A*C*D       1,387   0,694
B*C*D       0,288   0,144
A*B*C*D    -0,263  -0,131


S = *   PRESS = *


Analysis of Variance for C9 (coded units)

Source              DF   Seq SS   Adj SS   Adj MS  F  P
Main Effects         4  342,437  342,437  85,6094  *  *
2-Way Interactions   6   11,089   11,089   1,8481  *  *
3-Way Interactions   4    8,217    8,217   2,0544  *  *
4-Way Interactions   1    0,276    0,276   0,2756  *  *
Residual Error       0        *        *        *
Total               15  362,019
```

$$\hat{A} = 8.84$$
$$\hat{B} = -2.51$$
$$\hat{C} = -1.09$$
$$\hat{D} = 0.11$$
$$\vdots \quad \vdots$$
$$\widehat{ABCD} = -0.262$$

From the normal plot in figure(1) it looks like A and B are the most important factors.

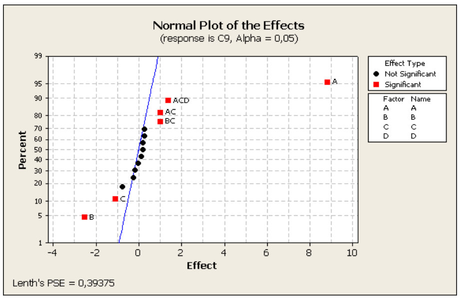

Figure 1: Normal plot a)

b) The corresponding regression model is

$$\begin{aligned} Y = {} & \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4 && (1) \\ & + \beta_{12} z_1 z_2 + \beta_{13} z_1 z_3 + \beta_{14} z_1 z_4 && (2) \\ & + \beta_{23} z_2 z_3 + \beta_{24} z_2 z_4 + \beta_{34} z_3 z_4 && (3) \\ & + \beta_{123} z_1 z_2 z_3 + \beta_{124} z_1 z_2 z_4 + \beta_{134} z_1 z_3 z_4 && (4) \\ & + \beta_{234} z_2 z_3 z_4 + \beta_{1234} z_1 z_2 z_3 z_4 + \epsilon && (5) \end{aligned}$$

And the estimated effects are of the kind

$$\hat{A} = 2b_1 \tag{6}$$

where b_1 is the least squares estimator of β_1. Same goes for the other effects.

c) In the analysis in a) we have 16 equations and 16 coefficients to estimate. Therefore there are no degrees of freedom left to estimate the variance. If we assume that the variance is known it is possible to make inference about the effects. For factor A we have:

$$\left.\begin{aligned} & \hat{A} = \tfrac{1}{8}(-Y_1 + Y_2 - \cdots - Y_{15} + Y_{16}) \\ & \text{Var}(\hat{A}) = \tfrac{1}{64} 16\sigma^2 = \tfrac{\sigma^2}{4} \end{aligned}\right\} \Rightarrow (\hat{A} \sim N(\mu_A, \frac{\sigma^2}{4})$$

95 % confidence interval for μ_A:

$$\hat{A} \pm z_{0.025} \frac{\sigma}{2} = (6.88, 10.80)$$

95 % confidence interval for μ_B:

$$\hat{B} \pm z_{0.025}\frac{\sigma}{2} = (-4.47, -0.5)$$

d) If there are good reasons to assume that the 3- and 4-factor interactions are 0, we have enough degrees of freedom to estimate the variance.
From MINITAB we get:

```
Fractional Factorial Fit

Estimated Effects and Coefficients for Response (coded units)

Term         Effect      Coef  StDev Coef        T      P
Constant              17,544      0,3258    53,84  0,000
A             8,837    4,419      0,3258    13,56  0,000
B            -2,512   -1,256      0,3258    -3,86  0,012
C            -1,087   -0,544      0,3258    -1,67  0,156
D             0,112    0,056      0,3258     0,17  0,870
A*B          -0,762   -0,381      0,3258    -1,17  0,295
A*C           1,012    0,506      0,3258     1,55  0,181
A*D           0,212    0,106      0,3258     0,33  0,758
B*C           1,012    0,506      0,3258     1,55  0,181
B*D           0,262    0,131      0,3258     0,40  0,704
C*D          -0,162   -0,081      0,3258    -0,25  0,813

Analysis of Variance for Response (coded units)

Source                DF      Seq SS      Adj SS      Adj MS      F      P
Main Effects           4     342,437     342,437      85,609  50,40  0,000
2-Way Interactions     6      11,089      11,089       1,848   1,09  0,473
Residual Error         5       8,493       8,493       1,699
Total                 15     362,019
```

We see that the estimator for σ^2 is now:

$$s^2 = MS_E = \frac{s_{ABC} + \cdots + s_{BCD} + s_{ABCD}}{5} = \frac{8.22 + 0.28}{5} = 1.70$$

where 8.22 is 3-way Seq SS, and 0.28 is 4-way Seq SS from the full analysis in section a. The variance of the effects is thus estimated by

$$s^2_{effect} = \frac{4s^2}{n} = 0.425$$

We can also obtain this estimate of σ^2_{effect} directly by using the estimated effects

$$s^2_{effect} = \frac{\widehat{ABC}^2 + \cdots + \widehat{BCD}^2 + \widehat{ABCD}^2}{5} = \frac{0.213^2 + 0.038^2 + 1.387^2 + 0.288^2 + 0.263^2}{5} = 0.425$$

Now we can do a T-test or an equivalent F-test to decide which of the effects are significant. We use the results and do an F-test:

$$F_A = \frac{MS_A}{MS_E} = \frac{s_A}{1.7} = \frac{(n/2\hat{A})^2/n}{1.7} = \frac{312.37}{1.7} = 183.74$$

$$F_B = \frac{MS_B}{MS_E} = \frac{s_B}{1.7} = \frac{25.26}{1.7} = 14.86,$$

and get the p-values:

$$p = P(F_{1,5} > 183.74) = 2P(T_5 > 13.56) \approx 0$$

$$p = P(F_{1,5} > 14.86) = 2P(T_5 > 3.85) = 0.012$$

Use that

$$\boxed{F_{1,\nu} = T_\nu^2}$$

): A has effect and B is significant at all levels > 0.012.