STA 303H1S / 1002 HS: One-way Analysis of Variance Practice Problems
SOLUTIONS

1. (a) Model: $Y_i = \beta_0 + \beta_1 I_{1,i} + \beta_2 I_{2,i} + \beta_3 I_{3,i} + \beta_4 I_{4,i} + e_i$, $i = 1, \ldots, 30$
where Y_i is the days until healing for the ith observation, e_i is the random error term for the ith observation, and $I_{j,i}$ is 1 if the ith observation received the jth treatment and 0 otherwise.

   (b) In matrix terms: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ where $\mathbf{Y}$ is the vector of length 30 of the days until healing, $\mathbf{e}$ is the vector of length 30 of the random error terms, $\boldsymbol{\beta}$ is the vector of length 5 of the model parameters and $\mathbf{X}$ is the 30×5 matrix with first column all 1's and jth ($j = 2, 3, 4, 5$) column consisting of 1 if the ith row corresponds to an observation from the $(j-1)$th treatment and 0 otherwise.

   (c) $\hat{\beta}_0 = 6.17$ is the estimate of the mean number of days until healing for the 5th treatment.

   (d) $\hat{\beta}_1 = 1.33$, $\hat{\beta}_2 = -1.67$, $\hat{\beta}_3 = -1.83$, $\hat{\beta}_4 = -1.00$.
$\hat{\beta}_j$, $j = 1, 2, 3, 4$ estimates for the jth treatment how much the mean number of days until healing differs from the mean for treatment 5.

   (e) There is evidence of differences among the means of the number of days until healing for the 5 treatments ($p = 0.0136$).

   (f) Confidence intervals for pairwise differences between the means, adjusted by either of Tukey's or Bonferroni's methods to maintain the overall confidence level at 95%, do not include zero for treatments 1 and 3. This is consistent with the pairwise t-tests using Tukey's and Bonferroni's methods to maintain the overall Type I error rate at 0.05: only the test for treatments 1 and 3 have a p-value less than 0.05. From Tukey's procedure there is also weak evidence ($0.05 < p < 0.10$) of differences between the means for treatment pairs 1 and 2, and 1 and 4. The evidence is weaker from the Bonferroni adjusted p-values, as the Bonferroni method is more conservative when sample sizes are equal. Tests for all pairwise comparisons of means, unadjusted for Type I error rate, were also carried out, but the results of these should be viewed with caution since these tests were not pre-planned.

   (g) There is no reason to believe that the observations are correlated. The plot of the residuals versus predicted values shows no outliers and the variance appears to be the same for all treatment groups. The normal quantile plot does not indicate any deviations from normality.

2. Numerically, the pooled t-test has test statistic -5.67 and the analysis of variance F-test has test statistic 32.15 which is $(-5.67)^2$.

   The t test statistic is

$$\frac{\overline{y}_{Spock's} - \overline{y}_{Other}}{s_p \sqrt{\frac{1}{n_{Other}} + \frac{1}{n_{Spock's}}}}$$

   where

$$s_p = \sqrt{\frac{(n_{Other} - 1)s^2_{Other} + (n_{Spock's} - 1)s^2_{Spock's}}{n_{Other} + n_{Spock's} - 2}}.$$

The analysis of variance F-test has test statistic MSReg/MSE where MSE $= s_p^2$ and

$$\begin{aligned}
\text{MSReg} &= \sum_{g=1}^{2} n_g(\overline{y}_g - \overline{y})^2 \\
&= n_{Spock's}(\overline{y}_{Spock's} - \overline{y})^2 + n_{Other}(\overline{y}_{Other} - \overline{y})^2 \\
&= n_{Spock's}\left(\overline{y}^2_{Spock's} - 2\overline{y}_{Spock's}\overline{y} + \overline{y}^2\right) + n_{Other}\left(\overline{y}^2_{Other} - 2\overline{y}_{Other}\overline{y} + \overline{y}^2\right) \\
&= N\overline{y}^2 - 2\overline{y}(\sum_{(Spock's)} y_i + \sum_{(Other)} y_i) + n_{Spock's}\overline{y}^2_{Spock's} + n_{Other}\overline{y}^2_{Other} \\
&= n_{Spock's}\overline{y}^2_{Spock's} + n_{Other}\overline{y}^2_{Other} - N\overline{y}^2
\end{aligned}$$

Substituting

$$\overline{y} = \frac{n_{Spock's}\overline{y}_{Spock's} + n_{Other}\overline{y}_{Other}}{N}$$

gives

$$\begin{aligned}
\text{MSReg} &= \overline{y}^2_{Spock's}\left(n_{Spock's} - \frac{n^2_{Spock's}}{N}\right) + \overline{y}^2_{Other}\left(n_{Other} - \frac{n^2_{Other}}{N}\right) - \frac{2}{N}n_{Spock's}n_{Other}\overline{y}_{Spock's}\overline{y}_{Other} \\
&= \frac{n_{Spock's}n_{Other}}{N}(\overline{y}_{Spock's} - \overline{y}_{Other})^2 \\
&= \frac{(\overline{y}_{Spock's} - \overline{y}_{Other})^2}{\frac{1}{n_{Spock's}} + \frac{1}{n_{Other}}}
\end{aligned}$$

So MSReg/MSE is the square of the t test statistic.

3. Number of groups is $G = 7$
Total number of observations is $N = 5 + 6 + 9 + 2 + 6 + 9 + 9 = 46$
$\overline{y} = (5 * 34.12 + 6 * 33.62 + \cdots + 9 * 14.62)/46 = 26.583$
Pooled estimate of the error variance (MSE) is $(4*11.94^2+\cdots+8*5.04^2)/(4+\cdots+8) = 47.80$
Model Sum of Squares (SSReg) is $5(34.12 - 26.583)^2 + \cdots + 9(14.62 - 26.583)^2 = 1927.856$
This is sufficient to complete the ANOVA table:

Source	df	SS	MS
Model	$7 - 1 = 6$	1927.9	$1927.856/6 = 321.3$
Error	$45 - 6 = 39$	$47.80 * 39 = 1864.1$	47.8
Total	$46 - 1 = 45$	$1927.856 + 1864.1 = 3792.0$	

(This is the same as in SAS output, except for round-off error.)

4. (a) Since observations are uncorrelated,

$$\text{Var}(b_1) = \text{Var}\left(\overline{Y}_A - \overline{Y}_{Spock's}\right) = \frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_{Spock's}}$$

(b) It follows from the fact that all observations are assumed to have the same variance and, for example,

$$\text{E}\left(\sum_{(A)}(Y_i - \overline{Y}_A)^2\right) = (n_A - 1)\text{Var}(Y_i) = (n_A - 1)\sigma^2$$

5. For x_i as defined:
   $\overline{x} = 0$
   $\sum x_i y_i = \overline{y}_{Other} - \overline{y}_{Spock's}$
   and $\sum x_i^2 = \frac{1}{n_{Spock's}} + \frac{1}{n_{Other}}$
   Thus
   $$b_1 = \frac{\overline{y}_{Other} - \overline{y}_{Spock's} - 0}{\frac{1}{n_{Spock's}} + \frac{1}{n_{Other}}}$$
   and
   $$b_0 = \overline{y} - 0$$

6. We want to find the values of θ_g, $g = 1, \ldots, G$ that minimize
   $$S = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{gi} - \theta_g)^2.$$
   Differentiating gives
   $$\frac{\partial S}{\partial \theta_g} = -2 \sum_{i=1}^{n_g} (y_{gi} - \theta_g)$$
   and setting these G derivatives to 0 gives
   $$\hat{\theta}_g = \frac{\sum_{i=1}^{n_g} y_{gi}}{n_g}.$$

7. It is necessary to show that
   $$\sum_{g=1}^{G} \sum_{i=1}^{n_g} (\hat{y}_{gi} - \overline{y})(y_{gi} - \hat{y}_{gi}) = 0.$$
   To do this, consider that $\hat{y}_{gi} = \overline{y}_g$, then
   $$\sum_{g=1}^{G} \sum_{i=1}^{n_g} (\hat{y}_{gi} - \overline{y})(y_{gi} - \hat{y}_{gi}) = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\overline{y}_g - \overline{y})(y_{gi} - \overline{y}_g) = \sum_{g=1}^{G} \left\{ (\overline{y}_g - \overline{y}) \sum_{i=1}^{n_g} (y_{gi} - \overline{y}_g) \right\} = 0$$
   since $\sum_{i=1}^{n_g} y_{gi} = n_g \overline{y}_g$.

8. In the table I've indicated how to calculate each missing number.

Source	df	SS	MS	F
Model	6	1927.1	1927.1/6	6.72
Error	(46-1)-6	MSE * df for Error	MS Model / 6.72	
Total	46-1	1927.1+SSE		

   For the p-value, estimate the F distribution with 6 and 39 degrees of freedom with the F distribution with 6 and 30 degrees of freedom. Since 6.72 is larger than any values in the table, we can say that the p-value is < 0.001.