The Logic Behind the ANOVA for the Example 3.1 on page 61-81 of Montgomery's 7th Edition INSY7300 Maghsoodloo

In this Example (data on page 62 of Montgomery) there are 4 treatments (or 4 independent populations), whose sample means are to be compared simultaneously in order to test H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$, where μ represents the grand population average of all a = 4 populations (or treatments). For example, the data layout at the bottom of page 62 shows that y_{24} = 579 while y_{43} = 715 $\dot{A}/\text{min}$ (Angstrom/min), etc; further, there are n = 5 random observations from each level (or from each population). Hence, N = 4×5 = 20, and there are a total of 19 degrees of freedom in the entire experiment. In the following development, I will first break down the Total Corrected Sum of Squares (SS_T) into two orthogonal (i.e., additive) components :(1) due to differences in the 4 sample treatment means, (2) due to differences within each treatment. The Total SS (sum of squares) will be obtained from the deviations of individual observations from the grand sample means $\bar{y}_{..} = \frac{y_{..}}{20} = \frac{12355}{20}$ = 617.75, while within SS will be obtained from the variability of the n = 5 observations in the i^th^ treatment from the corresponding means $\bar{y}_{i.}$, i = 1, 2, 3, 4 (RFP = Radio-Frequency Power expressed in Watts). The response variable, y, is Etch Rate.

$$\text{SS(Total)} = SS_T = \sum_{i=1}^{4\text{RFP}} \sum_{j=1}^{n=5} (y_{ij} - 617.75)^2 \equiv \sum_{i=1}^{4} \sum_{j=1}^{n=5} [(y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..})]^2 \equiv \ldots \equiv$$

$$\sum_{i=1}^{4} \sum_{j=1}^{n=5} [(y_{ij} - \bar{y}_{i.})^2] + \sum_{i=1}^{4} \sum_{j=1}^{n=5} [(\bar{y}_{i.} - \bar{y}_{..})^2] \equiv \sum_{i=1}^{4} \sum_{j=1}^{n=5} [(y_{ij} - \bar{y}_{i.})^2] + \sum_{i=1}^{4} n[(\bar{y}_{i.} - \bar{y}_{..})^2]$$

$$\equiv \sum_{i=1}^{4} \sum_{j=1}^{n=5} (e_{ij})^2 + n\sum_{i=1}^{4} (\hat{\tau}_i)^2,$$

where $e_{ij} = y_{ij} - \bar{y}_{i.}$ is the (ij)^th^ residual and $\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}$ is the effect of the i^th^ treatment. Thus, we have the following orthogonal decomposition of SS_T:

SS(Total) = USS − CF = 7704511 −7632301.25 = SS(Within the 4 levels) + SS(Between Treatment Means)

Or:

72209.75 (with 19 df) = 5339.2000 (with 4×4 df) + 66870.55 (with 3 df)

There are two main assumptions in Fixed-Effects ANOVA:

(1) y_{ij}'s ~ NID(μ_i, σ_i^2), and (2) $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_a^2 = \sigma_\epsilon^2 = \sigma^2$.

The assumption 2 above implies that we may pool our sample variances

$$S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n=5}[(y_{ij}-\bar{y}_{i.})^2] = \frac{CSS_i(\text{Within})}{4} = 400.70,\ 280.30,\ 421.30,\ 232.50\ (i = 1, 2,$$

3, 4) from the 4 independent populations to obtain one overall estimate of error variance $\sigma_\epsilon^2 = \sigma^2 = \sigma_y^2$. The above assumption (2) can be tested using Bartlett's test on pp. 79-80 of Montgomery's 7th edition. Assuming that assumption (2) is tenable, we may estimate σ_y^2 from the within treatments as $\hat{\sigma}_y^2 = \dfrac{\sum_{i=1}^{4} CSS_i(W)}{4\times 4} =$

$$\frac{1}{a}\sum_{i=1}^{a} S_i^2 = \frac{1334.80}{4} = 333.70$$

with 16 df. Note that only when the design is balanced (i.e., n_i's = n for all i), then the pooled estimate of σ_ϵ^2 can be computed by averaging the variances of the "*a*" treatments as was done for this case;

otherwise, $\hat{\sigma}_y^2 = S_p^2 = \dfrac{\sum_{i=1}^{a} CSS_i(W)}{\sum_{i=1}^{a}(n_i-1)}$. **The reader is heavily cautioned that averaging the variances is quite wrong when the design is unbalanced and will lead to the wrong value of $\hat{\sigma}_y^2$ = MS(Error) = MS_E, but $MS_{Error} = S_p^2 = \dfrac{\sum_{i=1}^{a} CSS_i(W)}{\sum_{i=1}^{a}(n_i-1)}$ works for all cases. In the balance case this last equation reduces to $\sum_{i=1}^{a} S_i^2 / a$.**

Another estimate of error variance may be obtained by first estimating the variance between the four population means from $\hat{\sigma}^2_{\bar{y}} = \frac{1}{3}\sum_{i=1}^{4}[(\bar{y}_{i.} - \bar{y}_{..})^2] = \frac{13374.1100}{3}$ **= 4458.036667 with 3 df. However, forming the F statistic as the ratio of** $\frac{\hat{\sigma}^2_{\bar{y}}}{\hat{\sigma}^2_{y}}$ **would be exactly like comparing Apples and Oranges because the numerator estimates the variance between means while the denominator estimates the variance amongst the individual measurements! Therefore, we convert the variance of means in the numerator by multiplying it by the size of each sample n = 5 in order to convert** $\hat{V}(\bar{y})$ **to** $\hat{V}(y)$**. This implies that another estimate of** σ^2_y **may be obtained from** $n \times \hat{\sigma}^2_{\bar{y}}$ **= 5×4458.036667 = 22290.18333. Hence, the correct value of the F statistic is**

$$F_0 = \frac{22290.18333}{333.70} = \frac{\text{MS(Between Treatments)}}{\text{MS(Within Treatments)}} = 66.7971,$$

or another way to obtain the correct expression for F_0 is to define it as F_0 = $\hat{V}(\bar{y}_{\text{Between Levls}}) / \hat{V}(\bar{y}_{\text{Within Levls}}) = \frac{4458.036667}{333.70/5}$ **= 66.7971, as before. The corresponding Pr level of testing $H_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$ is given by** ***P-value*** **= $P(F_{3,16} \geq 66.7971)$ =$0.0^8288286$. Since this *P-value* is much smaller than 0.05, we can strongly reject $H_0 : \tau_i = 0$ and conclude that at least two of the 4 population means differ significantly. This implies that the 4 treatments (or RF powers) have a statistically significant impact on the dependent variable tool etch rate** $(\dot{A}/\text{min})$**. However, the significance of the F statistic does not show where the true differences are. Thus, Tukey's Post-ANOVA for fixed-effects must be used to ascertain where the exact differences are only if H_0 is rejected.**

Tukey's Post-ANOVA For Fixed-Effects after H_0 is Rejected

Step 1: Arrange the *a* = 4 level means in ascending order

$\bar{y}_{1.}$ = 551.2 $\bar{y}_{2.}$ = 587.4 $\bar{y}_{3.}$ = 625.4 $\bar{y}_{4.}$ = 707.00

Step 2: Obtain the critical value of Tukey's (Studentized Range) from Table VII (pp. 629-630 where p = *a*) and use it to compute the 95% half confidence interval width as shown below

$w_{ij} = q_{0.05}(4, 16)\times\sqrt{\frac{MS_{Error}(1/n_i + 1/n_j)}{2}} = 4.05\times\sqrt{\frac{333.70(1/5+1/5)}{2}} =$

$4.05\times\sqrt{333.70/5} = 4.05\times se(\bar{y}) = 33.0863$

Step 3: Underline all the means in step 1 which do not differ by as much as w_{ij} = 33.0863 Angstrom/min. Because all pairs of means differ more than 33.0863, then no underlining is needed and all pairs of means are statistically different.

For the Example 3.1 of Montgomery, the raw SS are as follows:

USS = $\sum_{i=1}^{4}\sum_{j=1}^{5} y_{ij}^2$ = 7704511, CF = $(\sum_{i=1}^{4}\sum_{j=1}^{5} y_{ij})^2 / N = 12355^2/20 = 7632301.25$

SS(Treatments) = $(2756^2+2937^2+3127^2+3535^2)/5 - CF = 66870.5500$.

The 4 treatment effects are $\hat{\tau}_1 = \bar{y}_{1.} - \bar{y}_{..} = -66.55$, $\hat{\tau}_2 = \bar{y}_{2.} - 617.75 = -30.35$,

$\hat{\tau}_3 = \bar{y}_{3.} - 617.75 = 7.65$, $\hat{\tau}_4 = \bar{y}_{4.} - 617.75 = 89.25$, and $\sum_{i=1}^{4} n_i\hat{\tau}_i \equiv 0$.

The 1st 3 residuals are $e_{11} = y_{11} - \bar{y}_{1.} = 575 - 551.2 = 23.80$, $e_{12} = y_{12} - \bar{y}_{1.} = 542 - 551.2 = -9.2$, and $e_{13} = 21.20$. The largest residual is $e_{32} = 25.60$, and because $se(e_{ij}) = \sqrt{(n-1)MS_{Error}/n} = 16.3389$, then the largest Studentized residual is $d_{32} = 25.6/16.3389 = 1.5668$.