J. Analysis of Variance

1. Steel Data: Does oxygen content vary by day analyzed?

2. Setup:
   a. Groups of observations from \( k \) different samples
   i. Categorical variable indicating group is called a factor
   ii. Values it takes on are called levels.
   b. Observations in one group all have same mean
   c. Summarize:
      \[
      \begin{align*}
      \text{Population or Treatment} & \quad 1 \quad 2 \quad \cdots \quad k \\
      \text{Group mean} & \quad \mu_1 \quad \mu_2 \quad \cdots \quad \mu_k \\
      \text{Group variances} & \quad \sigma_1^2 \quad \sigma_2^2 \quad \cdots \quad \sigma_k^2 \\
      \text{Sample size} & \quad n_1 \quad n_2 \quad \cdots \quad n_k \\
      \text{Sample mean} & \quad \bar{x}_1 \quad \bar{x}_2 \quad \cdots \quad \bar{x}_k \\
      \text{Sample variance} & \quad s_1^2 \quad s_2^2 \quad \cdots \quad s_k^2
      \end{align*}
      \]
   d. Let overall mean be \( \bar{x} \).

3. Test null hypothesis \( H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \) vs. alternative \( H_A : \) not all group means are equal.
   a. Test statistic is how far group means are from \( \bar{x} \): Square them and add them up:
      \[
      SS_{Tr} = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \cdots + n_k(\bar{x}_k - \bar{x})^2
      \]
   b. Presupposes \( \sigma \)'s are all equal; assume this.
   c. Want to divide by common \( \sigma^2 \), but we don’t know what it is.
   d. Take average of \( s_k^2 \)'s:
      \[
      SS_{Resid} = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2;
      \text{estimate} \quad s^2 = SS_{Resid}/(n_1 + \cdots + n_k - k)
      \]

4. Example: Oxygen content
   a. Recall that regression of oxygen content on position was non-significant.
   b. Data file also contains day on which assays were performed
   c. Look for dependence of assay on day (Fig. 43). 
   d. Analysis of Variance Table:
      \[
      \begin{array}{rrrrrr}
      \text{Day} & \text{Df} & \text{Sum Sq} & \text{Mean Sq} & \text{F value} & \text{p} \\
      \text{Residuals} & 35 & 41.372 & 1.182 & & \\
      \end{array}
      \]
   e. Do not reject null hypothesis of no day-to-day assay variation.
   f. Parallel Multiple Regression Analysis gives same results.
   g. If you take only first two days, you can get the same results as from earlier techniques for two samples.
   i. Anova table:

\[
\text{Day} \quad \text{Sum Sq} \quad \text{Mean Sq} \quad \text{F value} \quad \text{Pr(}F\text{)}
\begin{array}{cccc}
1 & 0.9134 & 0.9134 & 0.6599 & 0.4312 \\
\end{array}
\]

\[
\text{Residuals} & 13 & 17.9959 & 1.3843 \\
\]

ii. If you do two-sample \( t \) test,
   - 7 observations from day 1, 8 observations from day 2.
   - get \( t = 0.8019, df = 11.782, p\text{-value} = 0.4385 \).
   - Note almost exact equality:
     - \( p\text{-value with ANOVA} \) \( p\text{-value} \)
     - \( F\text{ statistic from the ANOVA with square of } t \text{ statistic.} \)
     - Residual degrees of freedom from the ANOVA
with degrees of freedom from the $t$ statistic.

iii. If you do two-sample $t$ test assuming equality of variance,
- get $t = 0.8123$, df = 13, p-value = 0.4312
- Note exact equality:
  - $p$-value with ANOVA $p$-value
  - $F$ statistic from the ANOVA with square of $t$ statistic.
  - Residual degrees of freedom from the ANOVA with degrees of freedom from the $t$ statistic.
- Exact equality always holds between ANOVA and two sample pooled $t$-test when there are only two groups.