Practical Interpretation

With a *quantitative* factor, like power in the etch-rate example, typically use *regression* modeling:

\[ y = \beta_0 + \beta_1 x + \epsilon \]

or perhaps

\[ y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon. \]

With a *qualitative* factor, like “method” in the peak discharge rate example, we typically focus on comparisons among means.
Comparing Means

A *contrast* is a linear combination of treatment means:

$$\sum_{i=1}^{a} c_i \mu_i, \quad \text{with} \quad \sum_{i=1}^{a} c_i = 0.$$  

Examples are:

- $\mu_1 - \mu_2$, where $c_1 = 1, c_2 = -1, c_3 = \ldots = c_a = 0$;
- $\mu_1 - \frac{1}{2}(\mu_2 + \mu_3)$, where $c_1 = 1, c_2 = c_3 = -\frac{1}{2}, c_4 = \ldots = c_a = 0$. 

Testing a Contrast

Many hypotheses about treatment means can be written in terms of a contrast:

\[ H_0 : \sum_{i=1}^{a} c_i \mu_i = 0 \]

for appropriate contrast constants \(c_1, c_2, \ldots, c_a\).

We estimate \(\sum_{i=1}^{a} c_i \mu_i\) by

\[ \sum_{i=1}^{a} c_i \hat{\mu}_i = \sum_{i=1}^{a} c_i \bar{y}_i = C. \]
C is an unbiased estimator, with

\[ V(C) = \frac{\sigma^2}{n} \sum_{i=1}^{a} c_i^2, \]

which we estimate by

\[ \hat{V}(C) = \frac{\text{MS}_E}{n} \sum_{i=1}^{a} c_i^2. \]

The test statistic is therefore

\[ t_0 = \frac{C}{\sqrt{\hat{V}(C)}} = \frac{\sum_{i=1}^{a} c_i \bar{y}_i}{\sqrt{\frac{\text{MS}_E}{n} \sum_{i=1}^{a} c_i^2}}. \]
Compare \( t_0 \) with the \( t \)-distribution with \( df = N - a \).

Equivalently, compare \( F_0 = t_0^2 \) with the \( F \)-distribution with \( df = 1, N - a \).

Confidence interval for \( C \):

\[
\sum_{i=1}^{a} c_i \bar{y}_i \pm t_{\alpha/2, N-a} \sqrt{\frac{MS_E}{n} \sum_{i=1}^{a} c_i^2}.
\]
Multiple Contrasts

Sometimes we test several contrasts in the same experiment, or equivalently set up CIs for those contrasts; error rate becomes an issue.

Control *experiment-wise* error rate for *all possible contrasts* using Scheffé’s method: replace \( t_{\alpha/2, N-a} \) with \( \sqrt{(a-1)F_{\alpha, a-1, N-a}} \).

Confidence interval becomes

\[
\frac{\sum_{i=1}^{a} c_i \bar{y}_i \pm \sqrt{(a-1)F_{\alpha, a-1, N-a}}}{\sqrt{\frac{\text{MS}_E}{n} \sum_{i=1}^{a} c_i^2}}.
\]
Pairwise Comparisons

Often the only contrasts we consider are pairwise comparisons $\mu_i - \mu_j$.

To control experiment-wise error rate for all pairwise comparisons, Tukey’s method gives shorter intervals than Scheffé’s: CI is

$$\bar{y}_i - \bar{y}_j \pm q_{\alpha}(a, f) \sqrt{\frac{MSE}{n}},$$

where $q_{\alpha}(a, f)$ is a percent point of the studentized range statistic, and $f = N - a$ is the degrees of freedom in MSE.
Fisher’s Least Significant Difference

The basic $t$-statistic for comparing $\mu_i$ with $\mu_j$ is

$$t_0 = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MS_E \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}}.$$

So we declare $\mu_i$ and $\mu_j$ to be significantly different if

$$|\bar{y}_i - \bar{y}_j| > t_{\alpha/2, N-a} \sqrt{MS_E \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}.$$

That is,

\[ t_{\alpha/2, N-a} \sqrt{MS_E \left( \frac{1}{n_i} + \frac{1}{n_j} \right)} \]

is the least significant difference, or LSD.

Notes

In a balanced design, \( n_i = n_j = n \), so the LSD is the same for every pair \( \mu_i \) and \( \mu_j \).

The LSD method has comparison-wise error rate \( \alpha \); it does not control experimentwise error rate.
R command

TukeyHSD(aov(EtchRate ~ factor(Power), etchRateLong))

Output

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = EtchRate ~ factor(Power), data = etchRateLong)

$'factor(Power)'$

<table>
<thead>
<tr>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>180-160</td>
<td>36.2</td>
<td>69.25438</td>
<td>0.0294279</td>
</tr>
<tr>
<td>200-160</td>
<td>74.2</td>
<td>107.25438</td>
<td>0.0000455</td>
</tr>
<tr>
<td>220-160</td>
<td>155.8</td>
<td>188.85438</td>
<td>0.0000000</td>
</tr>
<tr>
<td>200-180</td>
<td>38.0</td>
<td>71.05438</td>
<td>0.0215995</td>
</tr>
<tr>
<td>220-180</td>
<td>119.6</td>
<td>152.65438</td>
<td>0.0000001</td>
</tr>
<tr>
<td>220-200</td>
<td>81.6</td>
<td>114.65438</td>
<td>0.0000146</td>
</tr>
</tbody>
</table>
Graph

plot(TukeyHSD(aov(EtchRate ~ factor(Power), etchRateLong)))
Sample Size

Using Operating Characteristic curves: $\beta$ is the probability of type II error:

$$\beta = 1 - P\{\text{Reject } H_0 | H_0 \text{ is false}\}$$
$$= 1 - P\{F_0 > F_{\alpha, a-1, N-a} | H_0 \text{ is false}\}.$$ 

Charts plot $\beta$ against $\Phi$, where

$$\Phi^2 = \frac{n \sum_{i=1}^{a} \tau_i^2}{a\sigma^2}.$$

So if we know $\sigma^2$ and a set of treatment effects $\tau_1, \tau_2, \ldots, \tau_a$ for which we want the type II error to be $\beta$, we can find the smallest acceptable $n$. 

Note: often the resulting $n$ is too large for the experiment to be feasible. The experimenter must accept higher $\beta$ or larger $\tau$’s.

Alternative to using OC charts: using length of confidence interval; for comparing two means, confidence interval is

$$
\bar{y}_i - \bar{y}_j \pm t_{\alpha/2, N-a} \sqrt{MSE} \times \frac{2}{n},
$$

If we know $\sigma^2$, can choose $n$ to give desired width.

Still often gives infeasible (too large) $n$. 
Recall the “effects” model:

\[ y_{i,j} = \mu + \tau_i + \epsilon_{i,j}, \quad i = 1, 2, \ldots, a, \quad j = 1, 2, \ldots, n_i. \]

Note that we now allow for unbalanced data (unequal \( n_i \)).
Suppose we put all the $y$’s in a *response vector*:

$$
Y = \begin{pmatrix}
    y_{1,1} \\
    y_{1,2} \\
    \vdots \\
    y_{1,n_1} \\
    y_{2,1} \\
    y_{2,2} \\
    \vdots \\
    y_{2,n_2} \\
    \vdots \\
    y_{a,1} \\
    y_{a,2} \\
    \vdots \\
    y_{a,n_a}
\end{pmatrix}
$$
To keep track of the group that an observation belongs to, we also create a design matrix $X$.

The row for an observation in group $i$ consists of all zeroes, except for a single 1 at position $i$.

For convenience, we add a first column with all 1’s.
\( X = \begin{pmatrix}
1 & 1 & 0 & \ldots & 0 \\
1 & 1 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & 0 & \ldots & 0 \\
1 & 0 & 1 & \ldots & 0 \\
1 & 0 & 1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 0 & \ldots & 1 \\
1 & 0 & 0 & \ldots & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 0 & \ldots & 1
\end{pmatrix} \)
We also put all the parameters $\mu, \tau_1, \tau_2, \ldots, \tau_a$ into a parameter vector:

$$\beta = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_a \end{pmatrix}$$

With an error vector $\epsilon$ constructed like $Y$, we have

$$Y = X\beta + \epsilon.$$
This is a *multiple regression* equation.

The least squares estimates $\hat{\mu}, \hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_a$ satisfy the *least squares normal equations*

$$X'X\hat{\beta} = X'Y.$$ 

If $X'X$ were non-singular, we could solve these equations:

$$\hat{\beta} = (X'X)^{-1}X'Y.$$ 

But it isn’t, so we can’t...
The first column of $\mathbf{X}$ is the sum of the other columns, which makes one column redundant.

More formally, if

$$
\mathbf{b} = \begin{pmatrix}
+1 \\
-1 \\
-1 \\
\vdots \\
-1
\end{pmatrix},
$$

then

$$
(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'(\mathbf{X}\mathbf{b}) = 0.
$$

So $\mathbf{X}'\mathbf{X}$ is singular.
If we leave out one column of $X$, the remaining columns have no redundancy, and the reduced $X'X$ matrix is non-singular.

Leaving out a column means leaving the corresponding parameter out of the model.

Leave out the first column $\Leftrightarrow$ leave out $\mu$ $\Leftrightarrow$ the means model.

Leave out the second column $\Leftrightarrow$ leave out $\tau_1$ $\Leftrightarrow$ the effects model with the first level as the baseline.

Leave out the last column $\Leftrightarrow$ leave out $\tau_a$ $\Leftrightarrow$ the effects model with the last level as the baseline.
The R function \texttt{lm()} is a general-purpose multiple regression function, and uses this approach to estimation.

Because the formula \texttt{EtchRate \sim factor(Power)} has a \textit{factor} on the right hand side, \texttt{lm()} internally creates a design matrix $X$ with one column for each level of the factor except the first (thus imposing the constraint $\tau_1 = 0$), and solves the normal equations.

This looks complicated, but it does not require balanced data, and generalizes easily to models with more than one factor.