Analysis of Variance (ANOVA)

• Suppose we want to compare more than two means?

For example, suppose an engineer is investigating the relationship between the RF power setting and the etch rate for this tool. The objective of an experiment like this is to model the relationship between etch rate and RF power, and to specify the power setting that will give a desired target etch rate. The response variable is etch rate.

If there are 2 different RF power settings (say, 160W and 180W), then a z-test or t-test is appropriate:

\[ H_0: \mu_1 = \mu_2 \]
\[ H_1: \mu_1 \neq \mu_2 \]

Comparing > 2 Means

• What if there are 3 different power settings (say, 160, 180, and 200 W)?

\[ H_0: \mu_1 = \mu_2 \quad \text{and} \quad H_0: \mu_1 = \mu_3 \quad \text{and} \quad H_0: \mu_2 = \mu_3 \]
\[ H_1: \mu_1 \neq \mu_2 \quad \text{and} \quad H_1: \mu_1 \neq \mu_3 \quad \text{and} \quad H_1: \mu_2 \neq \mu_3 \]

• How about 4 different settings (say, 160, 180, 200, and 220 W)?

All of the above, **PLUS**

\[ H_0: \mu_1 = \mu_4 \quad \text{and} \quad H_0: \mu_2 = \mu_4 \quad \text{and} \quad H_0: \mu_3 = \mu_4 \]
\[ H_1: \mu_1 \neq \mu_4 \quad \text{and} \quad H_1: \mu_2 \neq \mu_4 \quad \text{and} \quad H_1: \mu_3 \neq \mu_4 \]

• What about 5 settings? 10?
Comparing > 2 means

• Also, suppose $\alpha = 0.05$
  
  - $(1 - \alpha) = P(\text{accept } H_0 \mid H_0 \text{ is true}) = 0.95$
  
  - 4 settings: $(0.95)^4 = 0.814$
  
  - 5 settings: __________
  
  - 10 settings: __________

• Instead, use Analysis of Variance (ANOVA)
  
  - $treatment, factor, independent variable$: that which is varied ($a$ levels)
  
  - $observation, response, dependent variable$: the result of concern ($n$ per treatment)
  
  - $randomization$: performing experimental runs in random order so that other factors don’t influence results.

Randomizing the data collection

• Do we need to $stratify$ the data collection?
  
  - Is it physically possible and feasible to run in random order?
  
  - If not, how do we manage our strata so as to avoid introducing variability that we don’t intend?
    
    - blocking?
  
• Use Excel or Minitab to determine a random order for our data, e.g. … (example)
Our example (See pg. 62)

**Table 3.1**
Ethch Rate Data (in Å/min) from the Plasma Etching Experiment

<table>
<thead>
<tr>
<th>Power (W)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Totals</th>
<th>Averages</th>
</tr>
</thead>
<tbody>
<tr>
<td>160</td>
<td>573</td>
<td>542</td>
<td>530</td>
<td>539</td>
<td>570</td>
<td>2756</td>
<td>551.2</td>
</tr>
<tr>
<td>180</td>
<td>565</td>
<td>595</td>
<td>590</td>
<td>579</td>
<td>610</td>
<td>2937</td>
<td>587.4</td>
</tr>
<tr>
<td>200</td>
<td>609</td>
<td>651</td>
<td>619</td>
<td>637</td>
<td>620</td>
<td>3127</td>
<td>625.4</td>
</tr>
<tr>
<td>220</td>
<td>725</td>
<td>700</td>
<td>715</td>
<td>685</td>
<td>710</td>
<td>3535</td>
<td>707.0</td>
</tr>
</tbody>
</table>

**Figure 3.2**
Box plots and scatter diagram of the etch rate data

There are a couple of basic questions we’d like to answer …

- Does changing the power change the mean etch rate?
- Is there an optimum level for power?

We would like to have an objective way to answer these questions
The Analysis of Variance (Sec. 3.2, pg. 62)

- In general, there will be \( a \) levels of the factor, or a treatments, and \( n \) replicates of the experiment, run in random order...a completely randomized design (CRD)
- \( N = an \) total runs
- We consider the fixed effects case...the random effects case will be discussed later
- Our objective is to test hypotheses about the equality of the \( a \) treatment means

<table>
<thead>
<tr>
<th>Treatment (Level)</th>
<th>Observations</th>
<th>Totals</th>
<th>Averages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( y_{11} )</td>
<td>( y_{12} )</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>( y_{21} )</td>
<td>( y_{22} )</td>
<td>...</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>( a )</td>
<td>( y_{a1} )</td>
<td>( y_{a2} )</td>
<td>...</td>
</tr>
</tbody>
</table>

What does it mean?

- The name “analysis of variance” stems from a \textit{partitioning} of the total variability in the response variable into components that are consistent with a model for the experiment
- The basic single-factor ANOVA model is

\[
y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad \lbrace i = 1, 2, \ldots, a \rbrace \\
\mu = \text{an overall mean, } \tau_i = \text{ith treatment effect,} \\
\epsilon_{ij} = \text{experiment al error, } NID(0, \sigma^2)
\]
Models for the data

There are several ways to write a model for the data:

\[ y_{ij} = \mu + \tau_i + \varepsilon_{ij} \] is called the effects model

Let \( \mu_i = \mu + \tau_i \), then

\[ y_{ij} = \mu_i + \varepsilon_{ij} \] is called the means model

Regression models can also be employed

Variability is measured by sums of squares …

- Total sums of squares is partitioned as …
  \[ SS_{\text{total}} = \sum_{i=1}^{\alpha} \sum_{j=1}^{n_i} y^2_{ij} - \frac{\sum_{i=1}^{\alpha} \sum_{j=1}^{n_i} y^2_{ij}}{N} = \] …

  \[ SS_{\text{treat}} = \sum_{i=1}^{\alpha} \frac{\sum_{j=1}^{n_i} y^2_{ij}}{n_i} - \frac{\sum_{i=1}^{\alpha} \sum_{j=1}^{n_i} y^2_{ij}}{N} = \] …

  \[ SS_E = SS_{\text{total}} - SS_{\text{treat}} = \] …

- A large value of \( SS_{\text{Treatments}} \) reflects large differences in treatment means
- A small value of \( SS_{\text{Treatments}} \) likely indicates no differences in treatment means
Variability is measured by sums of squares …

• Formal statistical hypotheses are:

• While sums of squares cannot be directly compared to test the hypothesis of equal means, \textit{mean squares} can be compared.

• A mean square is a sum of squares divided by its degrees of freedom, so …

\[ df_{treat} = a - 1 = _____ \]

\[ df_E = a(n - 1) = _____ \]

\[ df_{total} = an - 1 = _____ \]

Determining the Difference

• Mean Square, \( MS = SS/df \)

\[ MS_{treat} = \] ____________

\[ MS_E = \] ____________

• If the treatment means are equal, the treatment and error mean squares will be (theoretically) equal.

• If treatment means differ, the treatment mean square will be larger than the error mean square.

• The statistic is: \( F_0 = MS_{treat}/MS_E = \) ____________
The analysis of variance is summarized in a table

**TABLE 3.3**
The Analysis of Variance Table for the Single-Factor, Fixed Effects Model

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square</th>
<th>( F_a )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between treatments</td>
<td>( SS_{Treatment} = n \sum_{i=1}^{e} (\bar{y}_i - \bar{y})^2 )</td>
<td>( a - 1 )</td>
<td>( MS_{Treatment} )</td>
<td>( F_0 = \frac{MS_{Treatment}}{MS_E} )</td>
</tr>
<tr>
<td>Error (within treatments)</td>
<td>( SS_E = SS_T - SS_{Treatment} )</td>
<td>( N - a )</td>
<td>( MS_E )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( SS_T = \sum_{i=1}^{e} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 )</td>
<td>( N - 1 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The **reference distribution** for \( F_0 \) is the \( F_{a-1,a(n-1)} \) distribution
- **Reject** the null hypothesis (equal treatment means) if
  \[ F_0 > F_{a,a-1,a(n-1)} \]

---

ANOVA table

- For our example, this looks like …

**TABLE 3.4**
ANOVA for the Plasma Etching Experiment

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square</th>
<th>( F_0 )</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF Power</td>
<td>66,870.55</td>
<td>3</td>
<td>22,290.18</td>
<td>( F_0 = 66.80 )</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Error</td>
<td>5339.20</td>
<td>16</td>
<td>333.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>72,209.75</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **How do we interpret this?**
The reference distribution:

- **FIGURE 3.3** The reference distribution ($F_{3,16}$) for the test statistic $F_0$ in Example 3.1

---

Try this on Excel and Minitab

- Excel …

- Minitab …
A (very) little humor...

**ANOVA: ANALYSIS OF VALUE**

Is your research worth anything?

The test involves computation of the $F$-ratio:

$$F' = \frac{\text{sum of people who care about your research}}{\text{world population}}$$

This ratio is compared to the $F$-distribution with $I-1, N_I$ degrees of freedom to determine a $p$ (in your pants) value. A low $p$ (in your pants) value means you're on to something good (though statistically improbable).

**Hypothesis Testing**

$$H_0: \mu_1 = \mu_2$$

where,

- $H_0$: the Null Hypothesis
- $\mu_1$: significance of your research
- $\mu_2$: significance of a monkey typing randomly on a typewriter in a forest where no one hears it.

Which means are different?

- **Graphical methods**
  - Box plots
  - Dot diagrams
  - etc.

- **Numerical methods**
  - Tukey’s test (available on Minitab)
  - Duncan’s Multiple Range test