DESIGN and ANALYSIS of EXPERIMENTS

The goals of the next portion of the course are to develop the methods to (1) determine if proposed process/equipment modifications improve or impact the results of concern and (2) devise experiments to aid in modeling or optimization of the process.

- Comparison of Treatments
- Blocking and Randomization
- Reference Distributions
- ANOVA
- MANOVA
- Factorial Designs
- Two-Level Factorials
- Fractional Factorials
- Regression Analysis
- Robust (Taguchi) Design
A SIMPLE EXPERIMENT: COMPARISON OF TREATMENTS

A new process, process B, is to be compared against the process of record — process A. Process 10 lots w/ Process A and 10 w/ B, and measure the yield for each lot:

<table>
<thead>
<tr>
<th>time order</th>
<th>method</th>
<th>yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>89.7</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>81.4</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>84.5</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>84.8</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>87.3</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>79.7</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>85.1</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>81.7</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
<td>83.7</td>
</tr>
<tr>
<td>10</td>
<td>A</td>
<td>84.5</td>
</tr>
<tr>
<td>11</td>
<td>B</td>
<td>84.7</td>
</tr>
<tr>
<td>12</td>
<td>B</td>
<td>86.1</td>
</tr>
<tr>
<td>13</td>
<td>B</td>
<td>83.2</td>
</tr>
<tr>
<td>14</td>
<td>B</td>
<td>91.9</td>
</tr>
<tr>
<td>15</td>
<td>B</td>
<td>86.3</td>
</tr>
<tr>
<td>16</td>
<td>B</td>
<td>79.3</td>
</tr>
<tr>
<td>17</td>
<td>B</td>
<td>82.6</td>
</tr>
<tr>
<td>18</td>
<td>B</td>
<td>89.1</td>
</tr>
<tr>
<td>19</td>
<td>B</td>
<td>83.7</td>
</tr>
<tr>
<td>20</td>
<td>B</td>
<td>88.5</td>
</tr>
</tbody>
</table>

\[ \bar{y}_a = 84.24, \quad \bar{y}_b = 85.54 \]

\[ \bar{y}_b - \bar{y}_a = 1.30 \]

**QUESTION:** Is Process B really better than Process A?

\[ H_0: \mu_A = \mu_B \quad vs. \quad H_1: \mu_A < \mu_B \]

**APPROACHES:**
1. External historical data — empirical distribution
2. External t-distribution
3. External normal distribution with random sampling assumption
4. Internal estimate of standard deviation — assume random sampling from normal populations
**Approach 1: External Reference Distribution**

- Suppose we have historical data for process A.

- Since our experiment is comparing \( \bar{Y}_A \) to \( \bar{Y}_B \), where each constructed by 10 consecutive runs, we can build a reference distribution of all possible differences between adjacent runs of 10 from historical data.

"Randomization Distribution"

Reference distribution of 191 differences between averages of adjacent sets of 10 observations.

\[
\Rightarrow \text{EMPIRICAL level of significance/confidence:}
\]

\[
\frac{9}{191} \quad \text{runs gave a difference} \quad > 1.30 = \bar{Y}_B - \bar{Y}_A
\]

**: difference is statistically significant at} \; \alpha = 0.047 \; \text{level.}

**Note:** No assumption of independence!
Approach 2: External - $t$ distribution

- Can build up a set of 10 differences in averages between non-overlapping historical data, with small gaps in between.

What we gain:
1. 10 differences Normally distributed - CLT
2. Approximate independence of differences in averages ... even if some autocorrelation in individual runs

- Compute a sample std. dev. of mean differences (assuming population mean of 0):

<table>
<thead>
<tr>
<th>Ten nearly independent differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>observed results</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>from past records</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>from plant trial</td>
</tr>
<tr>
<td>variance of differences</td>
</tr>
<tr>
<td>standard deviation of differences</td>
</tr>
</tbody>
</table>

Now we base our test on difference between means with estimated population variance:

$$t_0 = \frac{(\bar{y}_B - \bar{y}_A) - (\mu_B - \mu_A)}{s}$$

Under $H_0$, $\mu_B = \mu_A$

$$= \frac{1.30}{0.60} = \frac{s}{\sqrt{n}} = 2.17 \Rightarrow \alpha = 0.028 \text{ level of confidence}$$

$t$ with 10 d.o.f.
**Approach 3:** Assume random sampling from normal distrb, with external value for $\sigma$

- Now we're close to one of our more standard statistical tests $\Rightarrow$ inferences on sampling distributions

  $n_A = 10, \quad \text{Var}(\bar{y}_A) = \frac{\sigma^2}{n_A}$
  
  $n_B = 10, \quad \text{Var}(\bar{y}_B) = \frac{\sigma^2}{n_B}$

- How about the sampling distribution for the difference in means?

  $\text{Var}(\bar{y}_B - \bar{y}_A) = \frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B} = \sigma^2 \left( \frac{1}{n_A} + \frac{1}{n_B} \right)$

  $s_{\bar{y}_B - \bar{y}_A} = \sigma \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \quad \text{standard error}$

- Even if original process moderately nonnormal, distribution of $\bar{y}_B - \bar{y}_A \approx$ normal by CLT

  $z = \frac{(\bar{y}_B - \bar{y}_A) - (\mu_B - \mu_A)}{\sigma \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$

  Using $s = \sigma$ from historical (individual) data

  $\sigma \sqrt{\frac{1}{10} + \frac{1}{10}} = \frac{2.88}{\sqrt{5}} = 1.29$

  Under $H_0$, $\mu_B = \mu_A$, so

  $z_0 = \frac{1.30}{1.29} = 1.01 \Rightarrow Pr(z > z_0) = 0.156$

- Technically, we're estimating $\sigma$; really, we ought to use

  $t_0 = 1.01$ with $N = 210$ or 209 degrees of freedom

- Disadvantage: Still requires external historical data
Approach 4: Random Sampling with Internal $t$ Estimate

- Now suppose we do not have historical data - only $n_A = 10$ & $n_B = 10$ runs.

- Task 1: Estimate $\sigma$ from samples
  - Individual variances - $V_A = n_A - 1$ & $V_B = n_B - 1$
  - $s_A^2 = \frac{1}{V_A} \sum (y_A - \bar{y}_A)^2$ & $s_B^2 = \frac{1}{V_B} \sum (y_B - \bar{y}_B)^2$

  - Pooled variance -
    $$s^2 = \frac{\sum (y_A - \bar{y}_A)^2 + \sum (y_B - \bar{y}_B)^2}{n_A + n_B - 2} = \frac{V_A s_A^2 + V_B s_B^2}{V_A + V_B}$$

- Task 2: Form $t$ test using pooled estimate for variance
Comparison of Approaches: Testing Experiment Mean Difference

(A) External reference distribution

\( \overline{y}_B - \overline{y}_A = 1.30 \)

\( \overline{y}_B - \overline{y}_A = 1.30 \)

\( \overline{y}_B - \overline{y}_A = 1.30 \)

\( \overline{y}_B - \overline{y}_A = 1.30 \)

(B) External t reference distribution

(C) Reference distribution based on random sampling model, external value for \( \sigma \).

(D) Reference distribution based on random sampling model, internal estimate of \( \sigma \).

Four reference distributions for industrial experiment, with significance levels shown at right.

Observations:

- The more rough our estimate of distribution, the less confidence we have in mean shift.

- Random sampling - IID - assumption is crucial to use of internal data in inferences!
Effect of Autocorrelation on Variance Estimates

Recall our definitions & properties

\[ X = a_1 y_1 + a_2 y_2 + a_3 y_3 = \sum_{i=1}^{3} a_i y_i \]

\[ E(x) = \mu_x = a_1 \mu_1 + a_2 \mu_2 + a_3 \mu_3 \]

\[ \text{Var}(x) = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + a_3^2 \sigma_3^2 \]

\[ + 2a_1a_2 \sigma_1 \sigma_2 \rho_{12} + 2a_1a_3 \sigma_1 \sigma_3 \rho_{13} + 2a_2a_3 \sigma_2 \sigma_3 \rho_{23} \]

If \( y_1, y_2, y_3 \) are not independent

\[ = \sum_{i=1}^{n} a_i^2 \sigma_i^2 + 2 \sum_{i=1}^{n} a_i a_j \sigma_i \sigma_j \rho_{ij} \]

\[ \text{cov}(y_i, y_j) \]

Suppose \( y_i \) have \( \mu_i = \mu, \sigma_i = \sigma \), and a lag 1 autocorrelation \( \rho_1 = \rho_{i-1,i} \neq 0 \)

\[ \text{Var}(x) = \text{Var}(Z'y) \]

\[ = \text{Var}(n y) \]

Since \( a_i = 1, \sigma_i = \sigma \), \( \rho_1 \neq 0 \), use above to find

\[ \text{Var}(ny) = n \sigma^2 + 2 \sigma^2 (n-1) \rho_1 \]

\[ = \sigma^2 \left[ n + 2(n-1) \rho_1 \right] \]

Since \(-1 \leq \rho \leq 1\), the sampling variance can be either inflated or deflated quite largely.

\[ \text{Var}(\bar{y}) = \frac{\sigma^2}{n} \left[ 1 + \frac{2(n-1) \rho_1}{n} \right] = \frac{\sigma^2}{n} C \]

E.g., for \(-\frac{1}{2} \leq \rho \leq \frac{1}{2}\), \( n = 10 \) \( \Rightarrow 0.1 \leq C \leq 1.9 \)
**RANDOMIZATION & BLOCKING**

- Randomization: to ensure validity in face of unknown disturbances
- Blocking: to eliminate unwanted sources of variability

(1) Randomization
- We depend on an experimental realization being representative of the "randomization distribution" formed by the initial pool of data ⇒ allow use of \( t \)

- Important to ensure that treatments are applied randomly in time and/or space
  
  E.g., \( n_A, n_B = 10 \) process experiments should be randomly intermixed to avoid correlation, trend issues

- Suppose we've subjected half of 24 wafers each to 2 different implants. In the subsequent furnace anneal (batch), how should we order the wafers?

- Especially important for dealing with time or space trends

(2) Blocking
- Experimental precision can often be greatly increased if we can make comparisons within matched pairs of experimental items.

- Example: Compare recipes A & B across five machines

  \[
  \begin{array}{c}
  \text{AB} \\
  \text{BA} \\
  \text{BA} \\
  \text{AB} \\
  \text{BA}
  \end{array}
  \]

  ⇒ Run each recipe on each machine
  ⇒ Randomize order of runs within machines
- Advantage: Can now examine differences (B - A) across the tools.

\[ \bar{d} = \frac{d_1 + d_2 + d_3 + d_4 + d_5}{5} \]

Better resolution in distribution of paired differences

\[ \bar{d} - \delta \sim t_{n-1}, \text{ to test for probability of observing deviation given } H_0: \delta = 0 \]
ANALYSIS OF VARIANCE (ANOVA)
Comparing Several Treatment Means

Suppose we want to compare 4 different process options - A, B, C, and D. How do we tell if these "treatments" has any effect?

Diet

Coagulation Time (min)

80
70
60

τ₁
τ₂
τ₃
τ₄

n₁
n₂
n₃
n₄

Conceptual Model

Population

- True difference in means?

Experiment Samples
- Estimates of population parameters

BETWEEN
vs.
WITHIN
GROUP
Variation.
Data for diet/coagulation example

Run order in time shown in parentheses

Coagulation time (seconds) for blood drawn from 24 animals randomly allocated to four different diets

<table>
<thead>
<tr>
<th>diet (treatment)</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>62(20)</td>
<td>63(12)</td>
<td>68(16)</td>
<td>56(23)</td>
</tr>
<tr>
<td></td>
<td>60(3)</td>
<td>67(9)</td>
<td>66(17)</td>
<td>62(3)</td>
</tr>
<tr>
<td></td>
<td>63(11)</td>
<td>71(13)</td>
<td>71(11)</td>
<td>60(6)</td>
</tr>
<tr>
<td></td>
<td>59(10)</td>
<td>64(14)</td>
<td>67(17)</td>
<td>61(18)</td>
</tr>
<tr>
<td></td>
<td>65(4)</td>
<td>68(13)</td>
<td>63(22)</td>
<td>64(19)</td>
</tr>
<tr>
<td></td>
<td>66(8)</td>
<td>68(21)</td>
<td>63(5)</td>
<td>59(24)</td>
</tr>
</tbody>
</table>

| treatment average | 61 | 66 | 68 | 61 |
| grand average     | 64 |    |    |    |
(1) Within Group Variation

Assume that each group is normally distributed and share a common $\sigma^2$

\[ S_t^2 = \text{sum of squares, within } t\text{'th group} = \sum_{j=1}^{n_t} (y_{tj} - \bar{y}_t)^2 \]

where $n_t = \# \text{ of samples for } t\text{'th treatment}$

\[ S_t^2 = \frac{S_t^2}{n_{t-1}} \quad \text{since } v_t = n_t - 1 \]

Pooling these to get estimate of common, w/in group $\sigma^2$

\[ S_R^2 = \frac{v_1 S_1^2 + v_2 S_2^2 + \ldots + v_k S_k^2}{v_1 + v_2 + \ldots + v_k} = \frac{S_R^2}{N-k} \]

\[ \hat{\text{WITHIN TREATMENT MEAN SQUARE}} \]

(2) Between Group Variation

$\Rightarrow$ We will be testing hypotheses $\mu_1 = \mu_2 = \ldots = \mu_k$

In this case, a 2nd estimate of $\sigma^2$ would be

\[ S_T^2 = \frac{\sum_{t=1}^{k} n_t (\bar{y}_t - \bar{y})^2}{k-1} = \frac{S_T^2}{v_T} \]

\[ \hat{\text{BETWEEN TREATMENT MEAN SQUARE}} \]

(3) Key Question: What if treatments are different?

Then $S_T^2$ estimates $\sigma^2 + \left[ \frac{\sum_{t=1}^{k} n_t \tau_t^2}{(k-1)} \right]$

where $\tau_t = \mu_t - \mu$

$\Rightarrow$ That is, $S_T^2$ is inflated by some factor related to the difference between treatments!
(4) Formal Test for Treatment Significance

Reject H₀ if \( \frac{S_T^2}{S_R^2} \) is significantly > 1

⇒ use F distribution, since \( \frac{S_T^2}{S_R^2} \sim F_{k-1, N-k} \)

(5) Total Variation

\[ S_D = \sum_{t=1}^{K} \sum_{i=1}^{n_t} (Y_{ti} - \bar{y})^2 \]

\[ s_D^2 = \frac{S_D}{N-D} = \frac{S_D}{N-1} \text{ Total sample variance} \]

(6) Table:

<table>
<thead>
<tr>
<th>source of variation</th>
<th>sum of squares</th>
<th>degrees of freedom</th>
<th>mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>between treatments</td>
<td>( S_T = 228 )</td>
<td>( v_T = 3 )</td>
<td>( s_T^2 = 76.0 )</td>
</tr>
<tr>
<td>within treatments</td>
<td>( S_R = 112 )</td>
<td>( v_R = 20 )</td>
<td>( s_R^2 = 5.6 )</td>
</tr>
<tr>
<td>total about the</td>
<td>( S_D = 340 )</td>
<td>( v_D = 23 )</td>
<td>( s_D^2 = 14.8 )</td>
</tr>
<tr>
<td>grand average</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ANOVA AS DECOMPOSITION OF VARIATION

- Algebraic identity:
  \[
  \sum_{t=1}^{k} \sum_{i=1}^{n_t} (y_{ti} - \bar{y})^2 = \sum_{t=1}^{k} n_t (\bar{y}_t - \bar{y})^2 + \sum_{t=1}^{k} \sum_{i=1}^{n_t} (y_{ti} - \bar{y}_t)^2
  \]
  \[
  S_D = S_T + S_R
  \]
  Total Sum of Squares of deviations = between-treatment + within-treatment sum of squares
  Squares of deviations from grand average = sum of treatment squares + sum of treatment squares

- One can further decompose the total sum of squares:
  \[
  S' = S'_A + S'_D
  \]
  \[
  \sum_{t=1}^{k} \sum_{i=1}^{n_t} y_{ti}^2 = N \bar{y}^2 + \sum_{t=1}^{k} \sum_{i=1}^{n_t} (y_{ti} - \bar{y})^2
  \]
  Total Sum of Squares (about zero origin) = sum of squares due to the average + Total sum of squares of deviations from average

- So
  \[
  S = S'_A + S_T + S_R
  \]
  d.o.f.: \[N = 1 + k - 1 + N - k\]

Full analysis of variance table

<table>
<thead>
<tr>
<th>source of variation</th>
<th>sum of squares</th>
<th>degrees of freedom</th>
<th>mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>average</td>
<td>98,304</td>
<td>1</td>
<td>98,304</td>
</tr>
<tr>
<td>between treatments</td>
<td>228</td>
<td>3</td>
<td>76.0</td>
</tr>
<tr>
<td>within treatments</td>
<td>112</td>
<td>20</td>
<td>5.6</td>
</tr>
<tr>
<td>total</td>
<td>98,644</td>
<td>24</td>
<td></td>
</tr>
</tbody>
</table>
ANOVA: CHECKING THE MODEL & RESIDUAL ANALYSIS

- A key assumption in resolving differences between variation components is that the residuals are "random" - IID & N(0, \sigma^2)

- Assumed mathematical models:
  \[ y_{ti} = \mu + \varepsilon_{ti} \]
  \[ \uparrow \quad \uparrow \]
  treatment residuals
  mean

- Checks:
  1. Plot residuals against time order
     - Attempt to catch any time trends. While it is possible to randomize against such trends, we lose resolving power if such a trend is large.

  2. Examine distribution of residuals
     - Check IID N(0, \sigma^2) assumption; look for gross non-normality
     \[ \Rightarrow \] examine residuals for each treatment group

  3. Plot residuals vs. estimates
     - Be especially alert to dependencies on size of estimate (e.g. proportional vs. absolute errors)

  4. Plot residuals vs. other variables of interest
     - Consider other variables of possible relevance (e.g. environmental factors)
Suppose we carefully structure our experiment to randomize our treatments, but also to block against some undesired source of variation (not of interest). How do we analyze our results?

More generally: Our treatments (e.g., process A, B) are one factor. Our blocks (e.g., tools 1-5) are another factor. How analyze experiments with 2 or more factors?

Assumed Model:

\[ y_{ti} = \mu + \beta_i + \tau_t + \varepsilon_{ti} \]

With decomposition

\[ y_{ti} = \bar{y} + (\bar{y}_i - \bar{y}) + (\bar{y}_t - \bar{y}) + (y_{ti} - \bar{y}_i - \bar{y}_t + \bar{y}) \]

\[ \bar{y} = \bar{A} + \bar{B} + \bar{t} + \bar{r} \]

Each with \( N = nk \) elements

\[ S = S_A + S_B + S_T + S_R \]

and d.o.f. \( nk = 1 + (n-1) + (k-1) + (n-1)(k-1) \)

Results from randomized block design, general case

<table>
<thead>
<tr>
<th>block</th>
<th>1</th>
<th>2</th>
<th>( \ldots )</th>
<th>t</th>
<th>( \ldots )</th>
<th>k</th>
<th>treatment average</th>
<th>( \bar{y}_i )</th>
<th>( \bar{y} ) = grand average</th>
</tr>
</thead>
<tbody>
<tr>
<td>( i )</td>
<td>( y_{i1} )</td>
<td>( y_{i2} )</td>
<td>( \ldots )</td>
<td>( y_{it} )</td>
<td>( \ldots )</td>
<td>( y_{ik} )</td>
<td>( \bar{y}_i )</td>
<td>( \bar{y} )</td>
<td></td>
</tr>
<tr>
<td>( i )</td>
<td>( y_{i1} )</td>
<td>( y_{i2} )</td>
<td>( \ldots )</td>
<td>( y_{it} )</td>
<td>( \ldots )</td>
<td>( y_{ik} )</td>
<td>( \bar{y}_i )</td>
<td>( \bar{y} )</td>
<td></td>
</tr>
<tr>
<td>( i )</td>
<td>( y_{i1} )</td>
<td>( y_{i2} )</td>
<td>( \ldots )</td>
<td>( y_{it} )</td>
<td>( \ldots )</td>
<td>( y_{ik} )</td>
<td>( \bar{y}_i )</td>
<td>( \bar{y} )</td>
<td></td>
</tr>
<tr>
<td>( n )</td>
<td>( y_{n1} )</td>
<td>( y_{n2} )</td>
<td>( \ldots )</td>
<td>( y_{nt} )</td>
<td>( \ldots )</td>
<td>( y_{nk} )</td>
<td>( \bar{y}_i )</td>
<td>( \bar{y} )</td>
<td></td>
</tr>
</tbody>
</table>
The corresponding ANOVA table is constructed:

### Algebraic decomposition of sums of squares for the randomized block design, general formulas

<table>
<thead>
<tr>
<th>source of variation</th>
<th>sum of squares</th>
<th>degrees of freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>average (correction factor)</td>
<td>$S_A = nk\bar{y}^2$</td>
<td>1</td>
</tr>
<tr>
<td>between blocks</td>
<td>$S_B = k\sum_i^n(\bar{y}_i - \bar{y})^2$</td>
<td>$n-1$</td>
</tr>
<tr>
<td>between treatments</td>
<td>$S_T = n\sum_i^k(\bar{y}_i - \bar{y})^2$</td>
<td>$k-1$</td>
</tr>
<tr>
<td>residuals</td>
<td>$S_R = \sum_i^n\sum_{i=1}^k(y_{ii} - \bar{y}_i - \bar{y} + \bar{y})^2$</td>
<td>$(n-1)(k-1)$</td>
</tr>
<tr>
<td>total</td>
<td>$S = \sum_i^n\sum_{i=1}^k y_{ii}^2$</td>
<td>$N = nk$</td>
</tr>
</tbody>
</table>

With expected values

$$S_B^2 = \sigma^2 + k\sum_{i=1}^k \beta_i^2 / (k-1)$$

$$S_T^2 = \sigma^2 + n\sum_{t=1}^T \tau_t^2 / (n-1)$$

So again, we can test the significance of observed "inflation" in block or treatment std. dev.

Assumptions to keep in mind:

1. IID, Normal residuals
2. Additivity of effects

Deviation or "corrected total ANOVA": $S = S_A + S_D$

Summary table: residuals and treatment and block deviations for a general randomized block design

<table>
<thead>
<tr>
<th>block</th>
<th>1</th>
<th>2</th>
<th>……</th>
<th>t</th>
<th>……</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>……</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

deviation of treatment averages from grand average

*… (\(y_{ii} - \bar{y}_i - \bar{y} + \bar{y}\)) …*

\((\bar{y}_i - \bar{y})\)

\(\bar{y} = \) grand average
Example: Particle Contamination

Two LPCVD tubes, three gas suppliers. Does supplier matter in average particle counts on wafers?

Experiment: 3 lots on each tube, for each gas, report average # particles added

<table>
<thead>
<tr>
<th>Treatment (Gas)</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block (Tube)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>36</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>44</td>
<td>18</td>
</tr>
<tr>
<td>10</td>
<td>40</td>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>

Decompose according to the equation:

\[
y_{ti} = \mu + \beta_i + \tau_t + \epsilon_{ti}
\]

\[
y_{ti} = \bar{y} + (\bar{y}_i - \bar{y}) + (\bar{y}_t - \bar{y}) + (y_{ti} - \bar{y}_t - \bar{y}_i + \bar{y})
\]

\[
S = S_A + S_B + S_T + S_e
\]

### MANOVA for our Example

<table>
<thead>
<tr>
<th>Analysis of Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
</tr>
<tr>
<td>Model</td>
</tr>
<tr>
<td>Error</td>
</tr>
<tr>
<td>C Total</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effect Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
</tr>
<tr>
<td>Block</td>
</tr>
<tr>
<td>Treatment</td>
</tr>
</tbody>
</table>