Model Adequacy Checking in the ANOVA

- **Checking assumptions** is important
- Normality
- Constant variance
- Independence
- Have we fit the right model?
- What to do if some of these assumptions are **violated**

Box–Cox Transformation

- Box–Cox is a **family of power transformations** for strictly **positive** data that helps
    1. stabilize variance
    2. make residuals closer to normal

Definition

For a positive variable $y > 0$ and parameter λ:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log y, & \lambda = 0. \end{cases}$$

It's **monotone** in y, so orderings are preserved. Common "nice" choices: $\lambda \in \{-1, -0.5, 0, 0.5, 1, 2\}$ (reciprocal, inverse-sqrt, log, sqrt, identity, square).

Refit the model on the transformed response and re-check diagnostics.

When to Use Box-Cox

- Data must be **strictly positive**. If you have zeros/negatives, either add a **small offset** c so $y + c > 0$

- Use when residual plots show skewness (violating normal assumption)

- Use when the variability of the response increases (or decreases) systematically with the mean (violating constant variance assumption)

Transformation of Response to Obtain Constant Variance

Let $E(y) = \mu$ be the mean of y, and suppose that the standard deviation of y is proportional to a power of the mean of y such that

$$\sigma_y \propto \mu^\alpha$$

We want to find a transformation on y that yields a constant variance. Suppose that the transformation is a power of the original data, say

$$y^* = y^\lambda \qquad (3.20)$$

Then it can be shown that

$$\sigma_{y^*} \propto \mu^{\lambda+\alpha-1} \qquad (3.21)$$

Clearly, if we set $\lambda = 1 - \alpha$, the variance of the transformed data y^* is constant.

Common Transformations

TABLE 3.9				
Variance-Stabilizing Transformations				
Relationship Between σ_y and μ	**α**	**$\lambda = 1 - \alpha$**	**Transformation**	**Comment**
$\sigma_y \propto$ constant	0	1	No transformation	
$\sigma_y \propto \mu^{1/2}$	1/2	1/2	Square root	Poisson (count) data
$\sigma_y \propto \mu$	1	0	Log	
$\sigma_y \propto \mu^{3/2}$	3/2	−1/2	Reciprocal square root	
$\sigma_y \propto \mu^{2}$	2	−1	Reciprocal	

Randomization Test for Non-normal Data

TABLE 3.3

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

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Between treatments	$SS_{\text{Treatments}} = n \sum_{i=1}^{a} (\bar{y}_{i.} - \bar{y}_{..})^2$	$a - 1$	$MS_{\text{Treatments}}$	$F_0 = \dfrac{MS_{\text{Treatments}}}{MS_E}$
Error (within treatments)	$SS_E = SS_T - SS_{\text{Treatments}}$	$N - a$	MS_E	
Total	$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2$	$N - 1$		

Not F dist anymore, so we cannot obtain p-value

Randomization Test for Non-normal Data

- Consider all possible ways of permuting the observations ($N!$ possible ways)

- For a levels (n replicates each), then $K = \dfrac{N!}{n!...n!}$ possible rearrangements of the observations.

- For each of these rearrangements, calculate the value of the *F*-statistic

$$F = \frac{SS_{\text{Treatments}}/(a-1)}{SS_E/(N-a)} = \frac{MS_{\text{Treatments}}}{MS_E}$$

- The distribution of the *F*-statistic under the <u>null hypothesis</u> is obtained by calculating all possible values of the *F*-statistic under all possible rearrangements of the observed data points.

Randomization Test for Non-normal Data

- Count the number of F values greater than or equal to F_0, denote the count as C

- An estimate of the p-value is

$$C\,/\,K$$

- If $\frac{C}{K} < \alpha$, then reject the null hypothesis.

- Also called (Fisher's) permutation test

- The framework can also be used for t-test

Example: Etch Rate Data

TABLE 3.1

Etch Rate Data (in Å/min) from the Plasma Etching Experiment

Power (W)	Observations					Totals	Averages
	1	2	3	4	5		
160	575	542	530	539	570	2756	551.2
180	565	593	590	579	610	2937	587.4
200	600	651	610	637	629	3127	625.4
220	725	700	715	685	710	3535	707.0

- How to do the randomization test?

Randomization Test for "Big" Data

- It is computationally prohibitive to enumerate the exact randomization distribution when the dataset is a little bigger

- $N!$ or $K = \dfrac{N!}{n!...n!}$ may be too big, and it is computationally expensive to obtain all of the F values.

- In this case, we can randomly consider a fix number of permutations (or rearrangements) of the observations.

Post-ANOVA Comparison of Means

- The analysis of variance tests the hypothesis of equal treatment means

- Assume that residual analysis is satisfactory

- If the null hypothesis in ANOVA is rejected, we don't know **which specific pair of means** are different

- Determining which specific means differ following an ANOVA is called the **multiple comparisons problem (or post hoc tests)**

- There are **lots** of ways to do this

■ TABLE 3.1

Etch Rate Data (in Å/min) from the Plasma Etching Experiment

Power (W)	Observations					Totals	Averages
	1	2	3	4	5		
160	575	542	530	539	570	2756	551.2
180	565	593	590	579	610	2937	587.4
200	600	651	610	637	629	3127	625.4
220	725	700	715	685	710	3535	707.0

TABLE 3.4

ANOVA for the Plasma Etching Experiment

Source of Variation	Sum of Square	Degrees of Freedom	Mean Squares	F_0	P-Value
RF Power	66,870.55	3	22,290.18	$F_0 = 66.80$	< 0.01
Error	5339.20	16	333.70		
Total	72,209.75	19			

Fisher's LSD

$$H_0 : \mu_i = \mu_j$$

$$H_1 : \mu_i \neq \mu_j$$

- **Fisher's Least Significant Difference method**
- Like (pairwise) two-sample t-test, but pay attention to the estimate of variance

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

- Reject when

$$|t_0| > t_{\alpha/2, N-a}$$

Fisher's LSD (A simple way)

μ_i and μ_j would be declared significantly different if

$$|\bar{y}_{i.} - \bar{y}_{j.}| > t_{\alpha/2, N-a}\sqrt{MS_E(1/n_i + 1/n_j)}$$

The procedure is simpler for a balanced design, that is

$$n_1 = n_2 = \cdots = n_a = n$$

$$LSD = t_{\alpha/2, N-a}\sqrt{\frac{2\,MS_E}{n}}$$

If $|\bar{y}_{i.} - \bar{y}_{j.}| > LSD$, we conclude that the population means μ_i and μ_j differ.

To illustrate the procedure, if we use the data from the experiment in , the LSD at $\alpha = 0.05$ is

$$\text{LSD} = t_{.025,16}\sqrt{\frac{2\,MS_E}{n}} = 2.120\sqrt{\frac{2(333.70)}{5}} = 24.49$$

Thus, any pair of treatment averages that differ in absolute value by more than 24.49 would imply that the corresponding pair of population means are significantly different. The differences in averages are

$$\bar{y}_{1.} - \bar{y}_{2.} = 551.2 - 587.4 = -36.2^*$$

$$\bar{y}_{1.} - \bar{y}_{3.} = 551.2 - 625.4 = -74.2^*$$

$$\bar{y}_{1.} - \bar{y}_{4.} = 551.2 - 707.0 = -155.8^*$$

$$\bar{y}_{2.} - \bar{y}_{3.} = 587.4 - 625.4 = -38.0^*$$

$$\bar{y}_{2.} - \bar{y}_{4.} = 587.4 - 707.0 = -119.6^*$$

$$\bar{y}_{3.} - \bar{y}_{4.} = 625.4 - 707.0 = -81.6^*$$

The starred values indicate pairs of means that are significantly different. Clearly, all pairs of means differ significantly.

Tukey's Test

- Issue with Fisher LSD: inflate type-I error α

- Tukey's test statistic for balanced design:

$$q_0 = \frac{|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}|}{\sqrt{MS_E/n}}$$

- Reject when $q_0 > q_\alpha(a, f)$

  a is the number of levels, f is the df of error

- The q-distribution (studentized range distribution) is for the statistic

$$q = \frac{\bar{y}_{\max} - \bar{y}_{\min}}{\sqrt{MS_E/n}}$$

The Q-table

TABLE V

Percentage Points of the Studentized Range Statistic

$$q_{0.01}(p, f)$$

f	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	90	135	164	186	202	216	227	237	246	253	260	266	272	272	282	286	290	294	298
2	14.0	19.0	22.3	24.7	26.6	28.2	29.5	30.7	31.7	32.6	33.4	31.4	34.8	35.4	36.0	36.5	37.0	37.5	37.9
3	8.26	10.6	12.2	13.3	14.2	15.0	15.6	16.2	16.7	17.1	17.5	17.9	18.2	18.5	18.8	19.1	19.3	19.5	19.8
4	6.51	8.12	9.17	9.96	10.6	11.1	11.5	11.9	12.3	12.6	12.8	13.1	13.3	13.5	13.7	13.9	14.1	14.2	14.4
5	5.70	6.97	7.80	8.42	8.91	9.32	9.67	9.97	10.24	10.48	10.70	10.89	11.08	11.24	11.40	11.55	11.68	11.81	11.93
6	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30	9.49	9.65	9.81	9.95	10.08	10.21	10.32	10.43	10.54
7	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55	8.71	8.86	9.00	9.12	9.24	9.35	9.46	9.55	9.65

Tukey's Test Procedure

1. Compute HSD: (honestly significant difference)

$$T_\alpha = q_\alpha(a, f)\sqrt{\frac{MS_E}{n}}$$

2. If $|\bar{y}_{i.} - \bar{y}_{j.}| > T_\alpha$, we conclude that the population means μ_i and μ_j differ.

For unbalanced design (Tukey–Kramer procedure.)

$$T_\alpha = \frac{q_\alpha(a, f)}{\sqrt{2}}\sqrt{MS_E\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

To illustrate Tukey's test, we use the data from the plasma etching experiment in [Example 3.1](#). With $\alpha = 0.05$ and $f = 16$ degrees of freedom for error, [Appendix Table V](#) gives $q_{0.05}(4, 16) = 4.05$. Therefore, from [Equation 3.35](#),

$$T_{0.05} = q_{0.05}(4, 16)\sqrt{\frac{MS_E}{n}} = 4.05\sqrt{\frac{333.70}{5}} = 33.09$$

Thus, any pairs of treatment averages that differ in absolute value by more than 33.09 would imply that the corresponding pair of population means are significantly different.

$$\bar{y}_{1.} - \bar{y}_{2.} = 551.2 - 587.4 = -36.20^*$$

$$\bar{y}_{1.} - \bar{y}_{3.} = 551.2 - 625.4 = -74.20^*$$

$$\bar{y}_{1.} - \bar{y}_{4.} = 551.2 - 707.0 = -155.8^*$$

$$\bar{y}_{2.} - \bar{y}_{3.} = 587.4 - 625.4 = -38.0^*$$

$$\bar{y}_{2.} - \bar{y}_{4.} = 587.4 - 707.0 = -119.6^*$$

$$\bar{y}_{3.} - \bar{y}_{4.} = 625.4 - 707.0 = -81.60^*$$

The starred values indicate the pairs of means that are significantly different. Note that the Tukey procedure indicates that all pairs of means differ. Therefore, each power setting results in a mean etch rate that differs from the mean etch rate at any other power setting.