Lecture 9

Completely Randomized Designs

Reading: Oehlert 2010 Chapter 3; DAE 2017 Chapter 3

DSA 8020 Statistical Methods II

Whitney Huang
Clemson University

Agenda

1. Completely Randomized Designs
2. ANOVA & Multiple Comparisons
3. Checking Model Assumptions

Motivating Example: Maze Training in Mice

Research Question

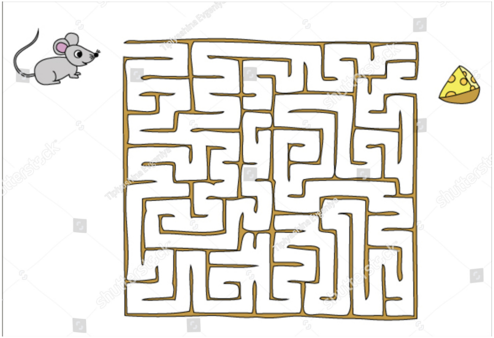
Does maze training improve navigation speed in mice?

Treatments

- 1, 2, 3, or 4 training runs

Response Variable

- Time required to complete the maze



Source: https://www.shutterstock.com/image-vector/find-your-way-cheese-mouse-maze-232569073

Completely Randomized Designs (CRD)

A Completely Randomized Design (CRD) assigns treatments (say g treatments) to experimental units completely at random, where $\sum_{i=1}^{g} n_i = N$.

Why use a CRD?

- Simple experimental structure
- Easy to analyze
- Flexible and robust
- Often sufficient when experimental units are homogeneous

Key Idea:
Randomization helps protect against bias

Goals of Statistical Inference

In a CRD, we are usually interested in:

1. Determining whether treatment means differ overall

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$$

2. Identifying which treatment means differ
3. Estimating treatment means and treatment differences

Main tools:

- ANOVA
- Multiple comparisons
- Confidence intervals

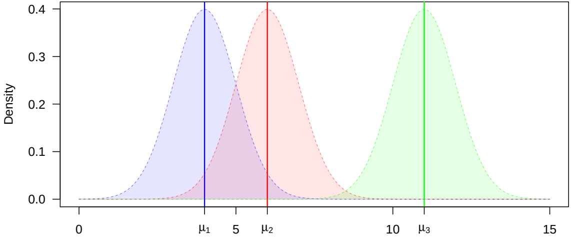
Means Model

The simplest ANOVA model assumes each treatment has its own mean:

$$\boxed{y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \cdots, g, \quad j = 1, \cdots, n_i, \quad \epsilon_{ij} \sim \mathrm{N}(0, \sigma^2),}$$

where:

- μ_i = mean response for treatment i
- ϵ_{ij} = random error



Interpretation:

Each treatment group has its own population mean

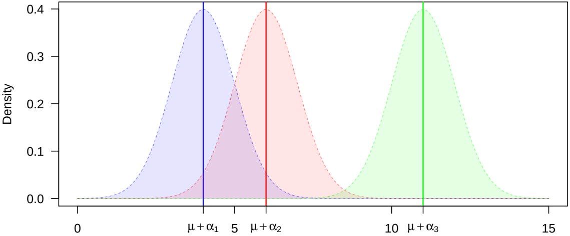
Effects Model

We can rewrite the treatment means as $\mu_i = \mu + \alpha_i$, which gives

$$\boxed{y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \cdots, g, \quad j = 1, \cdots, n_i, \quad \epsilon_{ij} \sim \mathrm{N}(0, \sigma^2),}$$

where:

- μ = baseline mean
- α_i = treatment effect



Overparameterized: a constraint is needed to define the baseline and make the parameters identifiable

Effects Model: Why Do We Need Constraints?

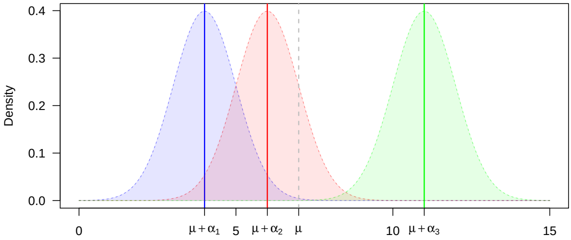
The effects model is overparameterized

Different combinations of μ and α_i can produce the same treatment means.

To make the parameters uniquely estimable, we impose a constraint such as:

$$\sum_{i=1}^{g} \alpha_i = 0$$

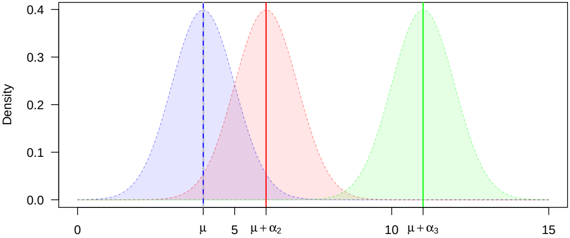
$\Rightarrow \mu$: overall mean



Effects Model Constraints Cont'd

Suppose we let $\alpha_1 = 0$

$\Rightarrow \mu$: mean of the first group (blue group in this example)



Dot Notation

- y_{ij}: observation j from treatment i
- $\bar{y}_{i\cdot}$: sample mean for treatment i
- $\bar{y}_{\cdot\cdot}$: overall sample mean

Example:

$$\bar{y}_{2\cdot}$$

means:

average response for treatment group 2

ANOVA Decomposition

Decomposition of y_{ij}: $y_{ij} = \bar{y}_{\cdot\cdot} + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (y_{ij} - \bar{y}_{i\cdot})$

$$\Rightarrow \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2}_{\text{SS}_T} = \underbrace{\sum_{i=1}^{g} n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}_{\text{SS}_{TRT}} + \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2}_{\text{SS}_E}$$

Total variability can be separated into:

Total Variation = Treatment Variation + Error Variation

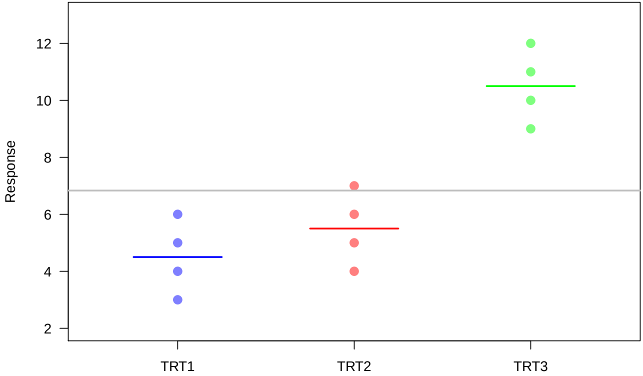
$$SS_T = SS_{TRT} + SS_E$$

Interpretation:

- Treatment variation measures differences between groups
- Error variation measures variability within groups

Graphical Illustration of ANOVA Decomposition

$$\underbrace{\text{Total Variation}}_{SS_T} = \underbrace{\text{Treatment Variation}}_{SS_{TRT}} + \underbrace{\text{Error Variation}}_{SS_E}$$



- Treatment variation measures differences between group means
- Error variation measures variability within groups

ANOVA Table

Source	df	SS	MS	F
Treatment	$g-1$	SS_{TRT}	$MS_{TRT} = \frac{SS_{TRT}}{g-1}$	$\frac{MS_{TRT}}{MS_E}$
Error	$N-g$	SS_E	$MS_E = \frac{SS_E}{N-g}$	
Total	$N-1$	SS_T		

Key Idea:
Large values of

$$\frac{MS_{TRT}}{MS_E}$$

suggest treatment differences

Computational details

$$\text{SS}_T = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{\cdot\cdot}\right)^2 = \sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{\cdot\cdot}^2}{N}$$

$$\text{SS}_{TRT} = \sum_{i=1}^{g} n_i\left(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot}\right)^2 = \sum_{i=1}^{g}\frac{y_{i\cdot}^2}{n_i} - \frac{y_{\cdot\cdot}^2}{N}$$

$$\text{SS}_E = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i\cdot}\right)^2 = \sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{g}\frac{y_{i\cdot}^2}{n_i} = \text{SS}_T - \text{SS}_{TRT}$$

How large is large? ⇒ Use the F-test

The ANOVA F-Test

Testing for treatment effects

$H_0 : \alpha_i = 0$ for all i
$H_a : \alpha_i \neq 0$ for some $i \Rightarrow$ at least one treatment effect differs

Test statistics: $F = \frac{\text{MS}_{TRT}}{\text{MS}_E}$.

Under H_0, the test statistic follows an $F_{g-1,\,N-g}$ distribution

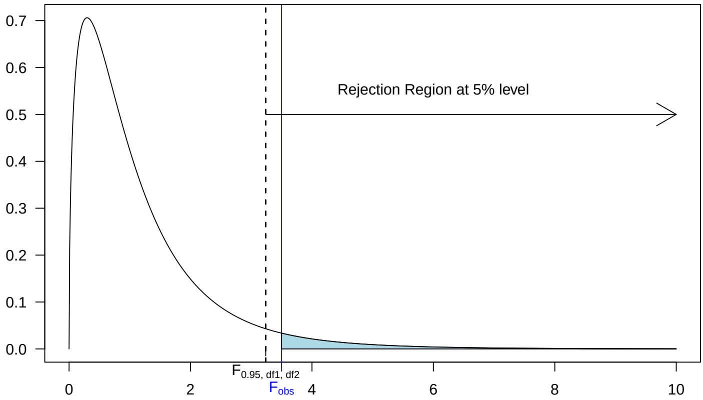
Reject H_0 if

$$F_{obs} > F_{g-1,N-g;\alpha}$$

for an α-level test, $F_{g-1,N-g;\alpha}$ is the upper $100(1-\alpha)\%$ percentile of the $F_{g-1,N-g}$ distribution

p-value $= \mathrm{P}(F > F_{obs})$; reject H_0 if p-value $< \alpha \Rightarrow$ Large F provide evidence against H_0

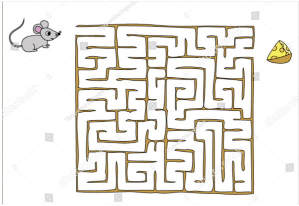
F Distribution and the F-Test



Interpretation:
The shaded region represents values leading to rejection of H_0

Mice Example Revisited

An experiment was conducted to determine if experience has an effect on the time it takes for mice to run a maze. Four treatment groups, consisting of mice having been trained on the maze one, two, three and four times were run through the maze and their times recorded.



Source: https://www.shutterstock.com/image-vector/find-your-way-cheese-mouse-maze-232569073

Training runs	1	2	3	4
n_i	5	5	5	5
$\bar{y}_{i\cdot}$	9.14	7.24	6.76	5.18
s_i^2	0.308	0.418	0.313	0.262

Example Cont'd

Training runs	1	2	3	4
n_i	5	5	5	5
$\bar{y}_{i\cdot}$	9.14	7.24	6.76	5.18
s_i^2	0.308	0.418	0.313	0.262

- Write down the model.
- Fill out the ANOVA table and test whether the time to run the maze is affected by training. Use a significant level of .05.

Model Assumptions

> All models are wrong, but some are useful–G.E.P Box

Model:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \cdots, g, \quad j = 1, \cdots, n_i.$$

We make the following assumptions:

- Errors normally distributed
- Errors have constant variance
- Errors are independent

$$\Rightarrow \epsilon_{ij} \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$$

What If Assumptions are Violated?

If the assumptions are not true, our statistical inferences might not be valid, for example,

- A confidence interval might not cover with the stated coverage rate
- A test with nominal type I error could actually have a larger or smaller type I error rate

We need good strategy for checking model assumptions, i.e., $\epsilon_{ij} \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$.

Checking Model Assumptions

We need to check if these assumptions reasonably met

Model:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

Data:

$$\begin{array}{lccccc} y_{ij} & = & (\bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..})) & + & (y_{ij} - \bar{y}_{i.}) \\ y_{ij} & = & \hat{y}_{ij} & + & \hat{\epsilon}_{ij}\,(r_{ij}) \\ \text{observed} & = & \text{predicted} & + & \text{residual} \end{array}$$

Residuals are our “estimates” of unobservable errors $\epsilon_{ij}' s$

We will conduct model diagnostics using **residual** and **predicted** values.

Residuals

We will use residuals to assess the model assumptions.

- Raw residual:

$$r_{ij} = y_{ij} - \hat{y}_{ij}, \text{ where } \hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i = \bar{y}_{i\cdot}$$

- Standardized residual (internally Studentized residual) adjusts r_{ij} for its estimated standard deviation

$$s_{ij} = \frac{r_{ij}}{\sqrt{\text{MS}_E(1 - \frac{1}{n_i})}}$$

- Studentized residual (externally Studentized residual)

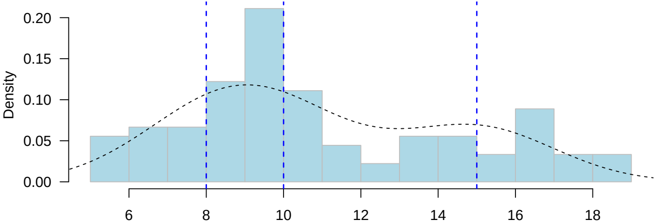
$$t_{ij} = s_{ij}\sqrt{\frac{N-g-1}{N-g-s_{ij}^2}}$$

  $t_{ij} \sim t_{df=N-g-1}$ if the model is correct ⇒ can be used to identify outliers

Assessing Normality

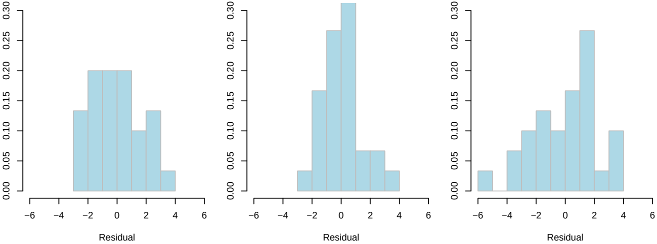
We DO NOT assume all $y'_{ij}s$ come from the same normal distribution, instead we assume $\epsilon'_{ij}s$ come from the same normal distribution ⇒ Not informative to plot a histogram for all the data—-treatment effects lead to non-normality

Example: Suppose $g = 3$, $(\mu_1, \mu_2, \mu_3) = (8, 10, 15)$ and $\epsilon'_{ij}s \sim \mathrm{N}(0, 2^2)$



Assessing Normality Cont'd

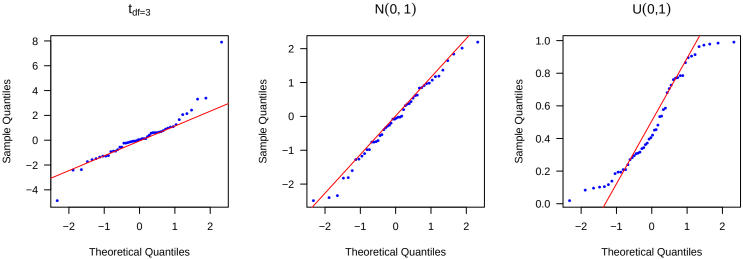
- If sample sizes are large, histograms of residuals can be constructed from each treatment separately



- Also, if sample sizes are large, QQ-plots or normal quantile plots can be generated for each treatment

Normal Quantile Plots

Plots $r_{(k)}$ versus $\Phi^{-1}(\frac{k}{n+1})$, $k = 1, \cdots, n$, where $r_{(k)}$ is the k^{th} ordered residual and $\Phi^{-1}(\frac{k}{n+1})$ is its corresponding (standard) normal score.



Interpretation

- Points near a straight line support normality
- Strong curvature suggests non-normality
- Extreme deviations may indicate outliers

Remarks on Assessing Normality

- Assessing normality
  - Formal tests (e.g., Shapiro–Wilk, Anderson–Darling) are often not very informative
  - Small samples: low power to detect non-normality
  - Large samples: trivial deviations may appear significant
  - Use QQ-plots and histograms for graphical assessment
- Handling non-normality
  - Use nonparametric methods such as the Kruskal–Wallis test
  - Apply transformations such as the Box–Cox transformation
- The ANOVA F-test is reasonably robust to moderate non-normality

Assessing Equal Variance

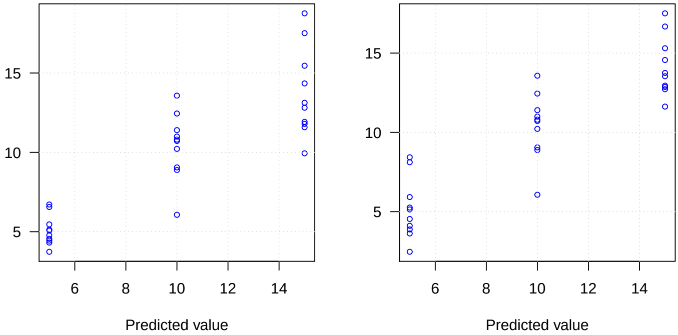
- We can test for equal variance, but some tests rely heavily on normality assumption:
  - Hartley's test
  - Bartlett's test
  - Cochran's C test
- F-test is reasonably robust to unequal variance if $n_i's$ are equal balanced design, or nearly so
- "*If you have to test for equality of variances, your best bet is Levene's test.*" – Gary Oehlert

Levene's Test

1. Compute $r_{ij} = y_{ij} - \bar{y}_{i.}$
2. Treat the $|r_{ij}|$ as data and use the ANOVA F-test to test H_0 that the groups have the same average value of $|r_{ij}|$
3. If $\frac{\text{MS}_{TRT}}{\text{MS}_E} > F_{g-1,N-g-1;\alpha} \Rightarrow$ reject H_0
4. Modified Levene's (Brown-Forsythe) test: use $d_{ij} = |y_{ij} - \tilde{y}_i|$, the absolute deviations from the group medians instead of $|r_{ij}|$

Fairly robust to non-normality and unequal sample size

Diagnostic Plot for Non-Constant Variance



Residuals vs. fitted values plot

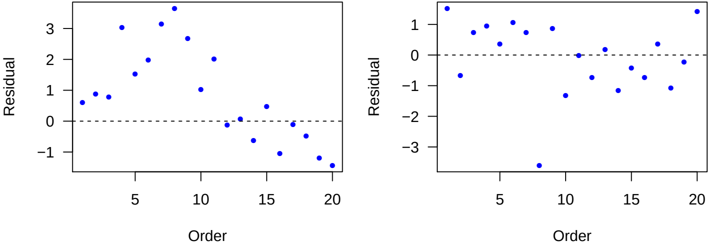
- Assess the constant variance assumption
- Identify possible outliers or unusual patterns

Remarks on the Constant Variance Assumption

- **Assessing constant variance**
  - Use residual plots for qualitative assessment
  - Do not rely solely on formal tests
- **Handling unequal variance**
  - Apply variance-stabilizing transformations
  - Incorporate unequal variances into the model
- The ANOVA F-test is reasonably robust to unequal variances when the design is (nearly) balanced

Assessing Dependence

Randomization helps support independence. Residual plots against run order or spatial location can help detect possible dependence



Durbin–Watson Statistic

A simple diagnostic for serial dependence:

$$\text{DW} = \frac{\sum_{k=1}^{n-1} (r_k - r_{k+1})^2}{\sum_{k=1}^{n} r_k^2}$$

DW $\approx 2:$ independence
DW $< 2:$ positive dependence
DW $> 2:$ negative dependence

Example: Balloon Experiment (DAE 2017, Exercise 3.12)

An experiment was conducted to investigate whether balloon color affects the time required to inflate a balloon to a diameter of 7 inches.

- Four balloon colors from the same manufacturer were tested
- An assistant inflated the balloons
- Inflation time was recorded using a stopwatch
- Color codes:

$$1 = \text{Pink}, \quad 2 = \text{Yellow}, \quad 3 = \text{Orange}, \quad 4 = \text{Blue}$$

Purpose of this example: Assess whether the experimental errors are independent.

⇒ See the accompanying `R` session for the full analysis

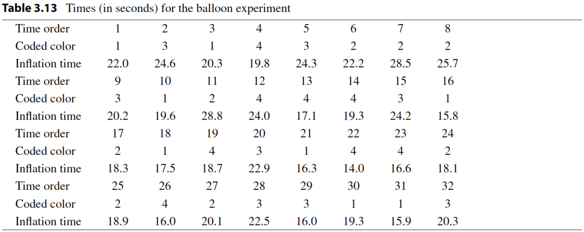


Table 3.13 Times (in seconds) for the balloon experiment

Time order	1	2	3	4	5	6	7	8
Coded color	1	3	1	4	3	2	2	2
Inflation time	22.0	24.6	20.3	19.8	24.3	22.2	28.5	25.7
Time order	9	10	11	12	13	14	15	16
Coded color	3	1	2	4	4	4	3	1
Inflation time	20.2	19.6	28.8	24.0	17.1	19.3	24.2	15.8
Time order	17	18	19	20	21	22	23	24
Coded color	2	1	4	3	1	4	4	2
Inflation time	18.3	17.5	18.7	22.9	16.3	14.0	16.6	18.1
Time order	25	26	27	28	29	30	31	32
Coded color	2	4	2	3	3	1	1	3
Inflation time	18.9	16.0	20.1	22.5	16.0	19.3	15.9	20.3

Summary

Topics Covered

- Completely Randomized Designs (CRD)
- ANOVA decomposition and the F-test
- Multiple comparisons
- Checking model assumptions

R Functions to Know

- **Data preparation:** `as.factor`
- **ANOVA:** `aov`
- **Multiple comparisons:** `LSD.test` (`agricolae`), `TukeyHSD`
- **Diagnostics:** `dwtest`, `levene.test`, `qqnorm`, `qqline`, `hist`