Lecture 34.

Recall (ANOVA for single Factor Experiments).

- we have a single variable (the factor) that takes on some small set of values (treatments) that the experimenter chooses.

- Given a many treatments, we assume a linear statistical model

$$\underbrace{y_{ij}}_{\text{random sample}} = \underbrace{\mu}_{\text{common mean}} + \underbrace{\tau_i}_{\text{treatment effects}} + \underbrace{\varepsilon_{ij}}_{\text{Error term}} \qquad \begin{cases} i \in \{1, \dots, a\} \\ j \in \{1, \dots, n\}. \end{cases}$$

Error term $\varepsilon_{ij} \sim N(0, \sigma^2)$.

- Under the assumption that $\sum_{i=1}^{a} \tau_a = 0$, we aim to test the hypotheses

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_a = 0, \quad H_1 : \tau_i \neq 0 \text{ for at least} .$$

- Given a set of data $\{y_{ij} : 1 \le i \le a, \ 1 \le j \le n\}$ we partition the total variation according to the ANOVA Equation

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_{\cdot\cdot}\right)^2}_{SS_T} = \underbrace{n\sum_{i=1}^{a}\left(y_{i\cdot}-\bar{y}_{\cdot\cdot}\right)^2}_{SS_{treatments}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_{i\cdot}\right)^2}_{SS_E}.$$

- From here, we form a test statistic:

$$F_0 = \frac{SS_{treat}/a-1}{SS_E/a(n-1)} = \frac{MS_{treat}}{MS_E}.$$

- The ANOVA f-test results in a rejection of H_0 iff $F_0 = f_0$ is too big, i.e.

$$f_0 > f_{\alpha, n-1, a(n-1)}$$

for a given significance level α.

- The ANOVA calculations are usually summarized in an ANOVA table:

Source of variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments	$SS_{Treatments}$	$a-1$	$MS_{Treat} = SS_{Treat}/a-1$	MS_{Treat}/MS_E
Error	SS_E	$a(n-1)$	$MS_E = \frac{SS_E}{a(n-1)}$	
Total	SS_T	$an-1$		

Confidence Intervals on the treatment means:

- Recall that the treatment means are defined to be $\mu_i := \mu + \tau_i$, $1 \le i \le a$.

- A point estimator for μ_i is $\hat{\mu}_i = \overline{Y}_{i\cdot} = \frac{\sum_{j=1}^{n} Y_{ij}}{n}$.
(i.e. just the usual sample mean)

- If we assume $\varepsilon_{ij} = N(\mu, \sigma^2)$, then $\overline{Y}_i = N(\mu_i, \frac{\sigma^2}{n})$

- If σ^2 is known, we can construct a $100(1-\alpha)\%$ CI based on the normal distribution.

- If σ^2 is unknown, then we can use MS_E as an unbiased estimator of σ^2:

- In this case

$$T = \frac{\bar{Y}_{i\cdot} - \mu_i}{\sqrt{MS_E/n}}$$

follows a t-distribution with $a(n-1)$ degrees of freedom, and so a $100(1-\alpha)\%$ CI for μ_i is

$$\bar{y}_{i\cdot} - t_{\alpha/2, a(n-1)}\sqrt{\frac{MS_E}{n}} \le \mu_i \le \bar{y}_{i\cdot} + t_{\alpha/2, a(n-1)}\sqrt{\frac{MS_E}{n}}.$$

- similarly given any two means μ_i, μ_j, $\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}$ is a point estimator for $\mu_i - \mu_j$, and $V(\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \frac{2\sigma^2}{n}$ (by independence)

Therefore,

$$T = \frac{\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} - (\mu_i - \mu_j)}{\sqrt{2MS_E/n}}$$

has a t-distribution with $a(n-1)$ degrees of freedom, and so a $100(1-\alpha)\%$ CI for $\mu_i - \mu_j$ has critical points

$$(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) \pm t_{\alpha/2, a(n-1)}\sqrt{\frac{2MS_E}{n}}$$

A Note on Unequal Sample Sizes.

- Recall that we assumed that for each $i \in \{1, \dots, a\}$, $n_i = n$. That is, for each treatment, the sample sizes are all the same.
- If we allow the n_i's to be different, we can adjust the ANOVA equation slightly.
- Let $N = \sum_{i=1}^{a} n_i$.
- Then

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}^2 - \frac{y_{\cdot\cdot}^2}{N}$$

$$SS_{Treat} = \sum_{i=1}^{a} \frac{y_{i\cdot}^2}{n_i} - \frac{y_{\cdot\cdot}^2}{N}$$

and

$$SS_E = SS_T - SS_{Treat}.$$

- this is the <u>unbalanced</u> situation.
- <u>balanced</u> is <u>better</u>
  - ↳ power is maximized.
  - ↳ less sensitive to slightly varying σ_i^2.

Last Topic Fisher's LSD test.

- Suppose we have performed an ANOVA test

$H_0: \tau_i = \tau_j = 0 \ \forall i,j$, $H_1: \tau_i \neq 0$ for some i

and we have decided to reject H_0.

- ANOVA doesn't tell us which of the $\tau_i \neq 0$.
- we want a method for comparing $\mu_i - \mu_j$
- there are many techniques, but we will discuss Fisher's "least significant difference" method.

- For each $i \neq j$, we create a statistic

$$t_0 = \frac{y_{i\cdot} - y_{j\cdot}}{\sqrt{\frac{2MS_E}{n}}}$$

(since the order doesn't matter, there are $\binom{a}{2}$ many such stats).

- Assuming $n_i = n_j = n$ (same sample size for each treatment), we say that the difference between two means μ_i and μ_j is significant if

$$|y_{i\cdot} - y_{j\cdot}| > \underbrace{t_{\alpha/2, a(n-1)}\sqrt{\frac{2MS_E}{n}}}_{LSD.}$$