STAT 5200 Handout #7a
Contrasts & Post hoc Means Comparisons (Ch. 4-5)

Recall CRD means and effects models:

$$\begin{aligned} Y_{ij} &= \mu_i + \epsilon_{ij} \\ &= \mu + \alpha_i + \epsilon_{ij} \end{aligned} \qquad i = 1, \ldots, g \quad ; \quad j = 1, \ldots, n \quad ; \quad \epsilon_{ij}\text{'s iid } N(0, \sigma^2)$$

- If we reject $H_0 : \mu_1 = \ldots = \mu_g$ (based on F_{Trt}), it's natural to look more carefully at why

$H_a: \mu_i \neq \mu_{i'}$ for some $i \neq i'$, or means μ_i not all equal

- How are μ_i's not equal?
  - We usually make pairwise comparisons:

μ_1 vs μ_2 : $H_0: \mu_1 - \mu_2 = 0$

μ_1 vs μ_3 : $H_0: \mu_1 - \mu_3 = 0$

μ_2 vs μ_3 : $H_0: \mu_2 - \mu_3 = 0$ ... (more)

  - This involves two issues:
    - Testing "contrasts" (Ch. 4)
    - Dealing with multiple hypothesis (Ch. 5)

special case of a "contrast"

Contrasts: a way to test any customized research question

- A linear combination of parameters (like μ_i's or α_i's)

$$\psi = \sum_{i=1}^{g} w_i \mu_i \qquad \text{or} \qquad \psi = \sum_{i=1}^{g} w_i \alpha_i$$

w_i's: coeffs. (actual #'s)

where

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

sum to 0

Ex: coeffs. w_i:

$i=1$	$i=2$	$i=3$
0	1	-1

This definition leads to some nice statistical properties

- In a CRD (one-way ANOVA), all contrasts are "estimable":

(i.e., they have unique least squares estimates and SE's)

$$\hat{\psi} = \sum_{i=1}^{g} w_i \hat{\mu}_i \quad = \sum_{i=1}^{g} w_i \hat{\alpha}_i = \sum_{i=1}^{g} w_i \bar{Y}_{i\cdot}$$

(both parameterizations)

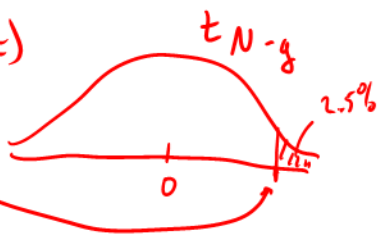
- For means model parameterization (not so clean for effects model):

$$\left(\text{SE of } \hat{\psi}\right) = \sqrt{Var[\hat{\psi}]} = \sqrt{\hat{\sigma}^2 \sum_{i=1}^{g} w_i^2/n_i}$$; recall $\hat{\sigma}^2 = MSE$

- We can use these to get 95% C.I. for ψ:

$$\hat{\psi} \pm (t_{.025,N-g}) \times \left(\text{SE of } \hat{\psi}\right)$$

est ± (Critical Value) × (SE of est)



- Equivalently, test H_0: $\psi = 0$ using

$$\frac{\hat{\psi}}{\left(\text{SE of } \hat{\psi}\right)} \sim t_{N-g} \quad \text{or, equivalently} \quad \left[\frac{\hat{\psi}}{\left(\text{SE of } \hat{\psi}\right)}\right]^2 \sim F_{1,N-g}$$

test statistic — sampling distribution — test statistic

- We look at a different ψ in each pairwise comparison of treatments – but ψ does not need to be a pairwise comparison

HO #7 example: Want to test whether liquid vs. powder matters

H_0: $\mu_{liquid} = \mu_{powder}$ or, equiv: H_0: $\left(\frac{\mu_A + \mu_B + \mu_C}{3}\right) = \mu_D$

Multiple Hypothesis Testing (as in testing all possible pairwise treatment comparisons):

H_0: $\left(\frac{1}{3}\mu_A + \frac{1}{3}\mu_B + \frac{1}{3}\mu_C - \mu_D\right) = 0$

-or, equiv:- $= \psi$

H_0: $\mu_A + \mu_B + \mu_C - 3\mu_D = 0$

$w = \left(\frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3} \quad -1\right)$

Recall in significance testing:

- Type I error:

reject H_0 when H_0 true

Type I error rate:

$\alpha = P\{\text{reject } H_0 \mid H_0 \text{ true}\}$

↑ "given"

- Type II error:

fail to reject H_0 when H_0 false

Type II error rate:

$\beta = P\{\text{fail to reject } H_0 \mid H_0 \text{ false}\}$

- Power: $P\{\text{reject } H_0 \mid H_0 \text{ false}\} = 1 - \beta$

must specify (how) H_0 is false to put a numeric value on power

H_0: diff = 0
H_a: diff ≠ 0 → so what is diff when?

→ what are we willing to accept? (in terms of follow-up effort wasted)

How to set tolerance for a Type I error when testing all K possible pairwise treatment comparisons?

- When we test H_0 at level α, this is an error rate
  reject H_0 when p-value $< \alpha$
- If we test K H_0's (all true), each at level $\alpha = 0.05$, how likely are we to reject any just by chance? ↑ # of tests

$$P\{\text{reject at least one } H_0 \mid \text{all true}\} = 1 - P\{\text{reject none} \mid \text{all true}\}$$

$$= 1 - P\{\text{don't reject \#1 and don't reject \#2 and } \ldots \mid \text{all true}\}$$

$$= 1 - (P\{\text{don't reject } H_0 \mid H_0 \text{ true})^K$$

$$= 1 - (1-\alpha)^K \approx$$ error rate .26 for $K=6$; for $K=100$, this becomes .994

So testing all H_0's at $\alpha = 0.05$ can make at least one Type I error quite likely!
much more likely than 5%

- With multiple H_0's, we need to consider meaningful error rates
  We want to control the error rate so $P\{\text{Type I error}\} \leq \alpha$ controlled at level α

- Type I error rates of major interest: (for a "family" of K nulls $H_{01}, H_{02}, \ldots, H_{0K}$)

1. PCER = $P($ reject $H_{0i} \mid H_{0i}$ true) for single i
   (per-comparison error rate)
   - only one test at a time (ignore multiple tests)
2. FWER = $P($ reject at least one $H_{0i} \mid H_{01}, \ldots, H_{0K}$ true)
   (family-wise error rate, or experimentwise error rate)
   (gets difficult to control when K large)
3. FDR = (# wrongly-rejected H_{0i}'s) / (total # rejected H_{0i}'s)
   (false discovery rate)
4. SFWER = $P($ # wrongly-rejected H_{0i}'s $\geq 1)$
   (strong family-wise error rate)
   ↳ because it allows a broad mixture of true & false nulls (i.e., no "given" in probability)
5. Simultaneous Confidence Intervals
   (set confidence level for family of all K intervals)
   $100 \times (1-\alpha)\%$

Recommendations to control these error rates (i.e., set tolerance for Type I errors):

Error Rate	All Pairwise Comparisons	Other Specialized Comparisons
PCER	~~LSD~~	
FWER	pLSD	
FDR	SNK	
SFWER*	REGWQ	
Simult. CI*	Bonferroni, Tukey	Scheffé, Dunnett

* preferred

NOTES:

- When asking many questions of data from an experiment (including when multiple response variables, or genes, are of interest), the honest thing to do is adjust for multiple testing. (See example on last page of Handout #7.)
- Most criticisms of p-values can be deflected by understanding what a p-value is (and isn't), and by appropriate handling of multiple testing.

Appendix: Multiple Comparison Procedures (Ch. 5)

For each of the error rates mentioned above, there are recommended "multiple comparison procedures". The choice of a multiple comparison procedure (MCP) depends on the error rate to be controlled and the type of comparisons of interest. The table above summarizes the textbook's recommendations. Each of these methods is described in Chapter 5 of the text, and briefly summarized in this appendix to this Handout #7.

Bonferroni

- test each H_{0i} at level α/K
- best suited for small K (2 to 10 or so) since α/K gets too small for larger K; it's too conservative (too hard to reject H_0) for larger K
- also controls SFWER
- a variation to control the FDR (but not SFWER): Benjamini-Hochberg
  - sort p-values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(K)}$
  - starting with largest p-value, work down, and reject H_{0i} if $p_{(j)} \leq j \cdot \alpha/K$ for some $j \geq i$
  - well-suited for large K, but assumes all K tests are statistically independent

Scheffé

- best suited when all possible contrasts are of interest (so data snooping is allowed)
- general idea:
  - For any contrast $\psi = \sum_{i=1}^{g} w_i \mu_i$, compute $\hat{\psi} = \sum_{i=1}^{g} w_i \hat{\mu}_i = \sum_{i=1}^{g} w_i \bar{Y}_{i\cdot}$
  - Sampling distribution of $\hat{\psi}$ is based on $F_{g-1,N-g}$ distribution (note $N-g$ is d.f. for MSE)
  - Reject $H_0 : \psi = 0$ only when $|\hat{\psi}|$ exceeds the Scheffé significant difference:

$$SSD \;=\; \sqrt{(g-1)F^*_{g-1,N-g}} \cdot \sqrt{MSE \cdot \sum_{i=1}^{g} \frac{w_i^2}{n_i}}$$

    where F^* is upper α critical value of appropriate F distribution
  - Downside: low power

Tukey (HSD)

- best suited when all possible pairwise mean comparisons (special case of contrasts) are of interest
- general idea:
  - Let $\bar{Y}_{max}$ and $\bar{Y}_{min}$ be the largest and smallest (respectively) $\bar{Y}_{i\cdot}$ among treatments $i = 1, \ldots, g$, and define statistic

$$q = \frac{\bar{Y}_{max} - \bar{Y}_{min}}{\sqrt{MSE/n}}$$

  - Sampling distribution of q is the "studentized range distribution" (see tables on pages 633-634 of text) – depends on d.f. (same as for MSE; ν in tables) and g (number of factor levels; K in tables)
  - Reject null $H_{0ij} : \mu_i = \mu_j$ only when the difference between $\bar{Y}_{i\cdot}$ and $\bar{Y}_{j\cdot}$ exceeds the honest significant difference (HSD):

$$HSD = \frac{q^*}{\sqrt{2}} \cdot \sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

    where q^* is the α critical value of appropriate studentized range distribution

REGWQ (REGWR)

- example of a "step-down" method, starting with most significant differences
  - Sort factor levels so that $\bar{Y}_{(1)\cdot} < \bar{Y}_{(2)\cdot} < \ldots < \bar{Y}_{(g)\cdot}$
  - For ordered means $\bar{Y}_{(i)\cdot}$ and $\bar{Y}_{(j)\cdot}$ $(i < j)$, "stretch" of means is $j - i + 1$.
  - For all comparisons involving stretch of k means (like $H_0 : \mu_{(i)} = \ldots = \mu_{(i+k-1)}$), use same critical value.
- general idea for REGWQ (author initials: Ryan-Einot-Gabriel-Welsch):
  - Test stretch (1) to (g), and stop if not significant.
  - Otherwise, declare end means (1) and (g) different, and "zoom in" to next smallest stretches
    - Test stretches (1) to $(g-1)$ and (2) to (g); for each, stop if not significant
    - For each, otherwise declare end means different and "zoom in" to next smallest stretches
    - Iterate until stop
  - Declare means $\mu_{(i)}$ and $\mu_{(j)}$ significantly different if reject null for stretch (i) to (j) <u>and</u> if reject null for all stretches containing (i) to (j). [Aside: makes a "closed" procedure]
  - Critical value (q^* in HSD above) involves studentized range distribution (and stretch size k and # factor levels g)

SNK

- Student-Newman-Keuls
- another "step-down" method similar to REGWQ, except critical value (q^* in HSD above) involving studentized range distribution (and stretch size k) does not involve # factor levels g

LSD

- Fisher's least significant difference
- makes no adjustment for multiple comparisons; just do all pairwise mean t-tests (with MSE as pooled variance estimate)
- Reject null $H_{0ij} : \mu_i = \mu_j$ only when the difference between $\bar{Y}_{i\cdot}$ and $\bar{Y}_{j\cdot}$ exceeds the least significant difference (LSD):

$$LSD \;\; = \;\; t^* \cdot \sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

  where t^* is the upper $\alpha/2$ critical value of appropriate student t distribution (with d.f. from MSE)

pLSD

- Fisher's protected least significant difference
- Do LSD tests only when F_{Trt} test of $H_0 : \mu_1 = \ldots = \mu_g$ is significant
- Does <u>not</u> give simultaneous confidence interval
- Controls FWER (under "complete" null $H_0 : \mu_1 = \ldots = \mu_g$, not under a partial null in which some means are equal but others differ); i.e., provides no strong control of FWER, just weak control

Dunnett

- best suited for comparing a "control treatment" (say factor level g) with other treatments (factor levels $1, \ldots, g-1$)
- critical value based on Dunnett's t distribution (see tables on pages 637-638 of text) – depends on d.f. (same as for MSE; ν in tables) and $g-1$ (number of factor levels - 1; K in tables)
- Reject null $H_{0i} : \mu_i = \mu_g$ only when the difference between $\bar{Y}_{i\cdot}$ and $\bar{Y}_{g\cdot}$ exceeds the Dunnett significant difference (DSD):

$$DSD \quad = \quad d^* \cdot \sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_g}\right)}$$

where d^* is the α critical value of appropriate Dunnett's t distribution

Simultaneous confidence intervals

- Recall general C.I. for a parameter θ: $\quad \hat{\theta} \pm \left[(\alpha \text{ Critical Value}) \times \left(\text{SE of } \hat{\theta}\right)\right]$

  Interpret: We are $(1-\alpha) \times 100\%$ confident the interval contains its true value (θ)

- For pairwise comparisons H_0: $\mu_i - \mu_j = 0$, build C.I. for $\mu_i - \mu_j$:

$$(\hat{\mu}_i - \hat{\mu}_j) \quad \pm \quad (\alpha \text{ Critical Value}) \times (SE[\hat{\mu}_i - \hat{\mu}_j])$$

  ($\hat{\mu}_i - \hat{\mu}_j = \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}$; Critical Value depends on multiple comparison procedure;
  $SE[\hat{\mu}_i - \hat{\mu}_j] = \sqrt{MSE \times \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$)

- Reject H_0 if $0 \notin CI$

  Equivalently, if $\quad |\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}| > (\alpha \text{ Critical Value}) \times \sqrt{MSE \times \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$

- Do C.I. for each μ_i vs. μ_j comparison of interest (usually all pairwise); interpret as:

  "For a family of K <u>simultaneous</u> intervals, we are $(1-\alpha) \times 100\%$ confident <u>all K intervals</u> contain their true values $(\mu_i - \mu_j)$"