STAT 705 Chapter 17: Analyzing factor level means

Adapted from Timothy Hanson

Department of Statistics, University of South Carolina

Stat 705: Data Analysis II

Inference for group means

Once the model is fit, we are typically interested in inference regarding group means $\mu_1, \ldots, \mu_r$.

In particular, if we reject the overall F-test of $H_0 : \mu_1 = \cdots = \mu_r$, we often want to know which *pairs* of means are significantly different. That is, we look at CIs for $\mu_i - \mu_j$ and tests of $H_0 : \mu_i = \mu_j$.

If one looks at all possible pairs, the number of comparisons is $\binom{r}{2} = \frac{r(r-1)}{2}$. For $r = 3$, this entails looking at $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, and $\mu_2 - \mu_3$.

Alternatively, one might be interested in differences such as $\mu_1 - \frac{1}{2}(\mu_2 + \mu_3)$. Here level 1 is placebo and levels 2 and 3 are two different doses of the same allergy medicine.

17.3 Comparing factor levels

Model is $Y_{ij} = \mu_i + \epsilon_{ij}$, where $\epsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2)$.

We have mean parameters $\mu_1, \ldots, \mu_r$. Most functions of interest are linear combinations of means:

$$L = L(\mathbf{c}) = \sum_{i=1}^{r} c_i \mu_i,$$

where $\mu_i = E\{Y_{ij}\}$. These include

- each mean, e.g. $L = \mu_2$
- differences, e.g. $L = \mu_3 - \mu_7$
- general contrasts, e.g. $L = \mu_1 - \frac{1}{3}\mu_2 - \frac{1}{3}\mu_3 - \frac{1}{3}\mu_4$
- general linear forms, e.g. $L = \mu_1 + 2\mu_2 - 10\mu_3$

A linear combination is called a *contrast* if $\sum_{i=1}^{r} c_i = 0$.

Estimation of L

Since $\bar{Y}_{i\cdot}$ is unbiased estimate of μ_i, $\hat{L} = \sum_{i=1}^r c_i \bar{Y}_{i\cdot}$ is unbiased estimate of L.

Note that $\bar{Y}_{i\cdot} \overset{ind.}{\sim} N(\mu_i, \sigma^2/n_i)$. Then

$$\hat{L} = \sum_{i=1}^r c_i \bar{Y}_{i\cdot} \sim N\left(\sum_{i=1}^r c_i \mu_i, \sigma^2 \sum_{i=1}^r \frac{c_i^2}{n_i}\right).$$

The estimated standard error of L is

$$\hat{\sigma}(\hat{L}) = \sqrt{\text{MSE} \sum_{i=1}^r \frac{c_i^2}{n_i}}.$$

When the model is true, we have

$$\frac{\hat{L} - L}{\hat{\sigma}(\hat{L})} \sim t(n_T - r).$$

CI and hypothesis test

Recall $\hat{L} = \sum_{i=1}^{r} c_i \bar{Y}_{i\cdot}$ estimates $L = \sum_{i=1}^{r} c_i \mu_i$ and $\hat{\sigma}(\hat{L})$ estimates $\sigma(\hat{L})$.

A 95% CI for L is $\hat{L} \pm se(\hat{L}) t(0.975, n_T - r)$.

To test $H_0 : L = L_0$, obtain p-value $P\left\{|t(n_T - r)| > |\frac{\hat{L} - L_0}{\hat{\sigma}(\hat{L})}|\right\}$.

Both of these can be computed in SAS procedures via `test`, `contrast`, or `estimate`.

Example: CI for μ_8

pp. 737–738.

Take $c_8 = 1$ and $c_i = 0$ for $i \neq 8$.

A $(1-\alpha)100\%$ CI is

$$\bar{Y}_{8\cdot} \pm \sqrt{\frac{MSE}{n_8}}\, t(1 - \frac{\alpha}{2}, n_T - r).$$

Difference $\mu_1 - \mu_2$

pp. 739–740.

Take $c_1 = 1$, $c_2 = -1$, and $c_i = 0$ for $i = 3, \ldots, r$.

Then

$$\frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot} - (\mu_1 - \mu_2)}{\sqrt{MSE(\frac{1}{n_1} + \frac{1}{n_2})}} \sim t(n_T - r).$$

To test $H_0 : L = 0 \Leftrightarrow H_0 : \mu_1 = \mu_2$, note that if H_0 is true then

$$t^* = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{MSE(\frac{1}{n_1} + \frac{1}{n_2})}} \sim t(n_T - r).$$

Reject at level α if $|t^*| > t(1 - \frac{\alpha}{2}; n_T - r)$.

Two-sample t-test w/ refined estimate of σ^2 (when $r > 2$).

Feral hog rooting activity example

For the feral hog study, one possible contrast would be $L = \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) - \mu_4$, comparing upland sites vs. floodplain sites

```
data rooting;
input habitat $ activity @@;
rootroot=sqrt(activity);
datalines;
BLH 139 BLH 228 BLH 275 CTS 45 CTS 127 CTS .
U 0 U 45 U 16  MS 145 MS 124 MS 240
 ;
 run;

proc glm data=rooting; class habitat;
 model rootroot=habitat / solution clparm; * solution not needed;
 lsmeans habitat; * not needed;
 estimate "Upland vs. Floodplain" habitat 1 1 1 -3 / divisor=3;
run;

proc glimmix data=rooting; class habitat;
 model rootroot=habitat;
 lsmestimate habitat 1 1 1 -3/ cl divisor=3;
run;
```

Is rooting activity greater in the floodplain? By how much?

If we obtain several 95% CI's for $L_1, \ldots, L_g$ separately, the probability that each L_i will be in its interval *simultaneously* will actually be (typically much) less than 95%:

$$P(L_1 \in I_1, L_2 \in I_2, \ldots, L_g \in I_g) \leq 0.95.$$

Question: what would this probability be if the intervals are independent?

Question: what would this probability be if the intervals are perfectly correlated in that $L_i \in I_i \Leftrightarrow L_j \in I_j$ for all $i \neq j$?

Simultaneous inference

We need CI's for linear combinations $L_1, \ldots, L_g$ such that probability that $L_1, \ldots, L_g$ are *simultaneously* in their respective CI's is at least $1 - \alpha$.

For example, say $r = 3$, $\boldsymbol{\beta} = (\mu_1, \mu_2, \mu_3)$ and we want to look at three pairwise differences $L_{12} = \mu_1 - \mu_2$, $L_{13} = \mu_1 - \mu_3$, $L_{23} = \mu_2 - \mu_3$. We want intervals I_{12}, I_{13}, I_{23} such that

$$P(L_{12} \in I_{12}, L_{13} \in I_{13}, L_{23} \in I_{23}) \geq 1 - \alpha.$$

We'll look at (1) Tukey, (2) Scheffe, and (3) Bonferroni procedures. All three procedures produce confidence intervals that look like

$$\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm \hat{\sigma}(\hat{L}_{ij}) \times \text{stat},$$

where stat is a critical value that depends on the method.

For Tukey,

$$\text{stat} = \frac{1}{\sqrt{2}} q(1-\alpha; r, n_T - r)$$

where q is the studentized range distribution (p. 746). Table B-9 has these values, but we'll just get them automatically from SAS. There are several examples on pp. 748–752.

- Unequal sample sizes ($n_i \neq n_j$ for some $i \neq j$) gives overall confidence greater than $1-\alpha$ (Tukey-Kramer). Equal sample sizes $n_1 = \cdots = n_r$ gives exact overall confidence of $1-\alpha$.
- Can be used for data "snooping" or data "dredging" – letting data suggest L's of interest.
- Derivation of the studentized range on next slide...

Derivation of Tukey intervals

Assume $n_1 = n_2 = \cdots = n_r = n$, so $n_T = rn$. Let $X_i = \bar{Y}_{i\cdot} - \mu_i$. Let $X_{(i)}$ be the ith order statistic.

$$X_1, \ldots, X_r \overset{iid}{\sim} N(0, \sigma^2/n).$$

Define

$$Q = \frac{X_{(r)} - X_{(1)}}{\sqrt{MSE/n}} \sim q(r, n_T - r).$$

This is the definition of the studentized range distribution. Then

$$\begin{aligned}
1-\alpha &= P\left\{ \frac{X_{(r)} - X_{(1)}}{\sqrt{MSE/n}} \le q(1-\alpha; r, n_T - r) \right\} \\
&= P\left\{ X_{(r)} - X_{(1)} \le \sqrt{MSE/n}\, q(1-\alpha; r, n_T - r) \right\} \\
&\ge P\left\{ |X_i - X_j| \le \sqrt{MSE/n}\, q(1-\alpha; r, n_T - r) \text{ for all } i, j \right\} \\
&= P\left\{ \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} - \hat{\sigma}(\hat{L}_{ij}) \times \text{stat} \le \mu_i - \mu_j \le \bar{Y}_{j\cdot} - \bar{Y}_{i\cdot} + \hat{\sigma}(\hat{L}_{ij}) \times \text{stat for all } i, j \right\}.
\end{aligned}$$

where stat $= \frac{1}{\sqrt{2}} q(1-\alpha; r, n_T - r)$.

Tukey example

```
* Tukey example ;

data rooting;
input habitat $ activity @@;
rootroot=sqrt(activity);
datalines;
BLH 139 BLH 228 BLH 275 CTS 45 CTS 127 CTS .
U 0 U 45 U 16  MS 145 MS 124 MS 240
 ;

proc glm data=rooting; class habitat;
 model rootroot=habitat;
 lsmeans habitat/ pdiff adjust=tukey alpha=0.05 cl lines;
run;
```

The subcommand `lines` adds a lines plot illustrating which levels are not significantly different.

17.6 Scheffe multiple comparisons

Recall $L(\mathbf{c}) = \sum_{i=1}^{r} c_i\mu_i$. Scheffe's method works for any number of arbitrary contrasts $L_1, \ldots, L_g$. The ith interval I_i among the g simultaneous intervals $I_1, \ldots, I_g$ has endpoints

$$\hat{L}(\mathbf{c}_i) \pm \hat{\sigma}\{\hat{L}(\mathbf{c}_i)\}\sqrt{(r-1)F(1-\alpha; r-1, n_T - r)}.$$

These intervals have the property,

$$P(L_1 \in I_1, L_2 \in I_2, \ldots, L_g \in I_g) \geq 1 - \alpha.$$

Example, pp. 754–755.

Comments on Scheffe

- Works for *all possible* contrasts, including differences in means.
- Okay for data snooping!
- If only pairwise differences are to be looked at, Tukey is better.
- If $H_0 : \mu_1 = \cdots = \mu_r$ is rejected, Scheffe's method guarantees at least one significant contrast out of all possible (p. 755).
- Here, stat $= \sqrt{(r-1)F(1-\alpha; r-1, n_T - r)}$.

17.7 Bonferroni procedure (p. 756)

Recall from STAT 712, if you have events $E_1, E_2, \ldots, E_g$, where $P(E_i) = \alpha$ for $i = 1, \ldots, g$, then

$$P(E_1^C \cap E_2^C \cap \cdots \cap E_g^C) \geq 1 - g\alpha.$$

We define our events to be $E_i = \{L(\mathbf{c}_i) \neq I_i\}$ and let I_i have endpoints

$$\hat{L}(\mathbf{c}_i) \pm t(1 - \frac{\alpha}{2g}, n_T - r)\hat{\sigma}\{\hat{L}(\mathbf{c}_i)\}.$$

Then $P(E_i) = \frac{\alpha}{g}$ and

$$P\{L(\mathbf{c}_1) \in I_1, \ldots, L(\mathbf{c}_g) \in I_g\} \geq 1 - g(\frac{\alpha}{g}) = 1 - \alpha.$$

Read this over several times to make sure you understand!

A bit more detail...

Draw a Venn diagram to convince yourself

$$P\left(\cup_i E_i\right) \leq \sum_i P(E_i).$$

This implies

$$1 - P\left(\cup_i E_i\right) \geq 1 - \sum_i P(E_i).$$

De Morgan implies

$$(\cup_i E_i)^c = \cap_i E_i^c.$$

Finally,

$$P(\cap_i E_i^c) = 1 - P\left(\cup_i E_i\right) \geq 1 - \sum_i P(E_i) = 1 - g\alpha.$$

Comments on Bonferroni

- Now the $\mathbf{c}_i$'s don't even have to be contrasts – all linear combinations work.
- Here, stat $= t(1 - \frac{\alpha}{2g}, n_T - r)$.
- If all pairwise differences in means are to be considered, use Tukey, else Bonferroni may or may not be better.
- Bonferroni usually beats Scheffe for comparison of contrasts (provides smaller intervals) unless looking at MANY L_i's. Note that Bonferroni's method has g in $t(1 - \frac{\alpha}{2g}, n_T - r)$, whereas Scheffe's method does not have g in $\sqrt{(r-1)F(1-\alpha; r-1, n_T - r)}$.
- Not good for snooping. Need to have $L_1, \ldots, L_g$ defined before analyzing data.

General comments

- If looking at handful g of pairwise comparisons, can calculate

$$\frac{1}{\sqrt{2}} q(1-\alpha; r, n_T - r), \ \sqrt{(r-1)F(1-\alpha; r-1, n_T - r)}, \ t(1 - \frac{\alpha}{2g}, n_T - r),$$

  and see which is smallest!
- In `estimate` command in `proc glm`, SAS will give you $\hat{L}$ and $\hat{\sigma}(\hat{L})$ for any $L = \sum_{i=1}^{r} c_i \mu_i$. Need to use `lsmestimate` with `cl` in `proc glimmix` to get CI automatically.

For feral hog example, interest is on

- $L_1 = \frac{1}{2}(\mu_1 + \mu_2 + \mu_3) - \mu_4$, comparing Upland vs Floodplain
- $L_2 = \frac{1}{2}(\mu_1 + \mu_4) - \frac{1}{2}(\mu_2 + \mu_3)$, comparing year-long wet habitats vs. dry habitats.
- $L_3 = \mu_2 - \mu_3$, comparing the two wettest habitats.

Feral hog rooting activity SAS code

```
* Scheffe example, p. 734 & pp. 754-755      ;
* glimmix does simultaneous testing and CI's ;
* use either adjust=scheffe or adjust=bon    ;

proc glimmix data=rooting; class habitat;
 model rootroot=habitat;
 lsmestimate habitat 'Floodplain vs upland' -1 -1 -1 3,
                 'Dry vs Wet' 1 -1 -1 1,
                 'Cypress-Tupelo vs. Muck Swamp' 0 1 -1 0 / adjust=scheffe alpha=0.1 cl divisor=3;
run;
```