One Way ANOVA

Lecture 19

Alexandra Chronopoulou

ILLINOIS

COLLEGE OF LIBERAL ARTS & SCIENCES

Department of Statistics
101 Illini Hall, MC-374
725 S. Wright St.
Champaign, IL 61820-5710

Question:

What happens when the F-test leads to the conclusion that the factor level means μ_i differ?

- Analysis of the factor level means of interest using *estimation* techniques.
- Statistical *tests* concerning the factor level means of interest.

Inference for Factor Level Means

- A single factor level mean.

- A difference between two factor level means.

- A contrast among factor level means.

- A linear combination of factor level means.

Single Factor Level Mean

- Estimation of μ_i: $\hat{\mu}_i = \bar{y}_{i\cdot}$
- Distribution of $\hat{\mu}_i$: $E(\hat{\mu}_i) = \mu_i, \quad Var(\hat{\mu}_i) = \frac{\sigma^2}{n_i}$.

  The estimated variance of $\bar{y}_{i\cdot}$ is $s^2_{\bar{y}_{i\cdot}} = \frac{1}{n_i} \cdot \frac{RSS}{n-r}$.
- Under the ANOVA model assumptions

$$\frac{\bar{y}_{i\cdot} - \mu_i}{s_{\bar{y}_{i\cdot}}} \text{ is distributed as } T_{n-r}.$$

- Confidence Interval for μ_i:

$$\mu_i \in \bar{y}_{i\cdot} \pm T_{n-r}(\alpha/2)\, s_{\bar{y}_{i\cdot}}$$

- In order to obtain confidence intervals for the factor level means in R, we fit the *means model* and we use the `confint` command:

```
# CI for single Factor Level Mean
g1=lm(coag~diet-1)
confint(g1)
```

```
##           2.5 %   97.5 %
## dietA 58.53185 63.46815
## dietB 63.98477 68.01523
## dietC 65.98477 70.01523
## dietD 59.25476 62.74524
```

Difference between Two Factor Level Means

The difference between two factor level means (pairwise comparison) is defined as

$$D = \mu_i - \mu_{i'}$$

- Estimation of D: $\hat{D} = \bar{y}_{i\cdot} - \bar{y}_{i'\cdot}$
- Distribution of $\hat{D}$: $E(\hat{D}) = \mu_i - \mu_{i'}$, $Var(\hat{D}) = \sigma^2 \left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)$.

  The estimated variance of $\hat{D}$ is

$$s^2_{\hat{D}} = \frac{RSS}{n-r} \cdot \left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right).$$

- Under the ANOVA model assumptions

$$\frac{\hat{D} - D}{s_{\hat{D}}} \text{ is distributed as } T_{n-r}$$

- Confidence Interval for D: $D \in \hat{D} + T_{n-r}(\alpha/2)\, s_{\hat{D}}$

- Hypothesis Test for D:

$$\left\{ \begin{array}{l} H_0 : \mu_i = \mu_{i'} \\ H_\alpha : \mu_i \neq \mu_{i'} \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} H_0 : \mu_i - \mu_{i'} = D = 0 \\ H_\alpha : \mu_i - \mu_{i'} \neq 0 \end{array} \right.$$

The test statistic is $t = \frac{\hat{D}}{s_{\hat{D}}} \sim T_{n-r}$.

Contrast of Factor Level Means

A contrast is a comparison involving two or more level means:

$$L = \sum_{i=1}^{r} c_i \mu_i, \quad \text{where} \sum_{i=1}^{r} c_i = 0.$$

- Estimation of L: $\hat{L} = \sum_{i=1}^{r} c_i \bar{y}_{i\cdot}$
- Distribution of $\hat{L}$: $E(\hat{L}) = \sum_{i=1}^{r} c_i \mu_i$, $Var(\hat{L}) = \sigma^2 \sum_{i=1}^{r} \frac{c_i^2}{n_i}$

  The estimated variance of $\hat{L}$ is

$$s_{\hat{L}}^2 = \frac{RSS}{n-r} \cdot \sum_{i=1}^{r} \frac{c_i^2}{n_i}$$

- Under the ANOVA model assumptions

$$\frac{\hat{L} - L}{s_{\hat{L}}} \text{ is distributed as } T_{n-r}$$

- Confidence Interval for L: $L \in \hat{L} + T_{n-r}(\alpha/2)\ \ s_{\hat{L}}$
- Hypothesis Testing for L:

$$\begin{cases} H_0 : L = 0 \\ H_\alpha : L \neq 0 \end{cases}$$

The test statistic is $t = \frac{\hat{L}}{s_{\hat{L}}} \sim T_{n-r}$.

Linear Combination of Factor Level Means

$$L = \sum_{i=1}^{r} c_i \mu_i, \quad \text{no restrictions on } c_i' s$$

- Point estimator and estimated variance same as before.
- Single Degree of Freedom Tests

$$\begin{cases} H_0 : L = c \\ H_\alpha : L \neq c \end{cases}$$

The test statistic here is

$$F = t^2 = \left(\frac{\hat{L} - c}{s_{\hat{L}}} \right)^2 \sim F_{1,n-r}$$

Limitations of Inference Procedures

- The confidence coefficient $1 - \alpha$ for the estimation procedures described is a statement confidence coefficient and applies only to a particular estimate, not to a series of estimates.

- Similarly the specified Type I error rate α applies only to a particular test and not to a series of tests.

Bonferroni Correction

When? The family of interest is a **particular set of pairwise comparisons, contrasts, or linear combinations** that is specified by the user.

- Suppose m is the number of statements in the family.
- In order to control the family wise error rate to be α, we need to reduce the error rate for each individual comparison to be α/m.
- That is we need to increase the significance level from $(1-\alpha)$ to $(1-\alpha/m)$.
- Not applicable when m is large, since the CIs would be too wide due to the increase of the significant level.

- In R, we can obtain the *p*-values for the Bonferroni corrections for pairwise differences using the `pairwise.t.test` command.

```
# Bonferroni correction: this test outputs the  p-values for the corresponding differences.
pairwise.t.test(coag, diet, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  coag and diet
##
##   A       B       C
## B 0.02282 -       -
## C 0.00108 0.95266 -
## D 1.00000 0.00518 0.00014
##
## P value adjustment method: bonferroni
```

Tukey's Paired Comparison Procedures

When? the family of interest is a set of **all pairwise comparisons** of factor level means, i.e. it consists of estimates of all pairs $D = \mu_i - \mu_{i'}$.

A confidence interval is given by

$$D \in \hat{D} + \frac{q(\alpha/2;\ r,\ n-r)}{\sqrt{2}}\ s(\hat{D}),$$

where $q(\alpha/2;\ r,\ n-r)$ refers to the $\alpha/2$ upper quantile of the **studentized range** for r means and $n-r$ degrees of freedom.

ILLINOIS

The coverage probability is exact when the sample sizes in each group are identical and is approximate otherwise.

Remark: The **studentized range** refers to the distribution of

$$max_{i \neq j} \sqrt{n}(\bar{y}_i - \bar{y}_j)/\hat{\sigma}$$

where $\bar{y}_i$ and $\bar{y}_j$ are sample means from independent samples of size n from normal distributions with common means and variance σ^2.

– To obtain Tukey family CIs for all pairwise comparisons in R, we use the `TukeyHSD` command.

```r
# Tukey Simultaneous 95% CI for all mean differences
TukeyHSD(aov(coag~diet), data=coagulation)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = coag ~ diet)
##
## $diet
##     diff         lwr       upr     p adj
## B-A    5   0.5932529  9.406747 0.0228300
## C-A    7   2.5932529 11.406747 0.0013858
## D-A    0  -4.2789880  4.278988 1.0000000
## C-B    2  -1.9415144  5.941514 0.4988550
## D-B   -5  -8.7981383 -1.201862 0.0075558
## D-C   -7 -10.7981383 -3.201862 0.0002854
```

Scheffé's Method for Contrasts

When? The family of interest is the set of **contrasts** among the factor level means:

$$L = \sum c_i \mu_i, \text{ where } \sum c_i = 0$$

An confidence interval is given by

$$L \in \hat{L} + (r-1)\, F_{r-1,n-r}(\alpha)\, s_{\hat{L}}$$

– To obtain Scheffé family CIs for all pairwise comparisons in R, we use the `ScheffeTest` in the *DescTools* library.

```r
g2aov=aov(coag~diet)

# If you want all the pairwise comparisons with Scheffe's method:
ScheffeTest(g2aov)
```

```
##
##   Posthoc multiple comparisons of means: Scheffe Test
##     95% family-wise confidence level
##
## $diet
##     diff     lwr.ci    upr.ci    pval
## B-A    5   0.342883  9.657117 0.03233 *
## C-A    7   2.342883 11.657117 0.00210 **
## D-A    0  -4.418129  4.418129 1.00000
## C-B    2  -2.165452  6.165452 0.55494
## D-B   -5  -8.896424 -1.103576 0.00876 **
## D-C   -7 -10.896424 -3.103576 0.00031 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

– If we want to obtain intervals for specific contrasts, such as

$$L_1 = \mu_A - \frac{1}{2}\mu_B - \frac{1}{2}\mu_C \text{ and } L_2 = \mu_B - \frac{1}{2}\mu_C - \frac{1}{2}\mu_D$$

then we can specify this in the `contrasts` argument as follows:

```
ScheffeTest(g2aov, contrasts=matrix(c(1,-0.5,-0.5,0,
                                       0,1,-0.5,-0.5), ncol=2))
```

```
##
##   Posthoc multiple comparisons of means: Scheffe Test
##     95% family-wise confidence level
##
## $diet
##        diff     lwr.ci    upr.ci   pval
## A-B,C -6.0 -10.165452 -1.834548 0.0032 **
## B-C,D  1.5  -2.031434  5.031434 0.6482
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```