Chapter 3

Analysis of one-way (fixed) factor level effects

Two major questions for one-way classification

(1) Determine whether or not the factor level means are the same. A recommended strategy is

$$Diagnostics \longrightarrow Transformation \longrightarrow ANOVA\ table$$

We are done if $\mathcal{H}_0$ is accepted.

(2) If the factor level means differ (i.e. $\mathcal{H}_a$ is true), examine

(i) how they differ
(ii) what the implications of the difference are

Inferences for factor level effects are generally concerned with one or more of the following

(i) A single factor level mean μ_i

(ii) A difference between two factor level means $\mu_i - \mu_{i'}$

(iii) A contrast among factor level means $\sum_{i=1}^{r} c_i\mu_i$ where $\sum_{i=1}^{r} c_i = 0$

(iv) A linear combination of factor level means $\sum_{i=1}^{r} c_i\mu_i$

3.1 Single factor level mean μ_i

Point estimator of μ_i:

$$\hat{\mu}_i = \bar{Y}_{i\cdot}$$

Two pivotal quantities for μ_i:

$$(i) \qquad \frac{\bar{Y}_{i\cdot} - \mu_i}{S_{i\cdot}/\sqrt{n_i}} \sim t_{n_i - 1} \qquad (\textit{one sample case})$$

$$(ii) \qquad \frac{\bar{Y}_{i\cdot} - \mu_i}{\sqrt{MSE/n_i}} \sim t_{n_T - r} \qquad (\textit{from the ANOVA table})$$

From these we can construct two CIs for μ_i. The question is then: which do you prefer to?

3.2 Difference between two factor level means $\mu_i - \mu_{i'}$

Point estimator of $\mu_i - \mu_{i'}$: $\quad \bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}$

Two pivitol quantities for $\mu_i - \mu_{i'}$:

$$(i) \quad \frac{\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot} - (\mu_i - \mu_{i'})}{S_{i\cdot i'}\sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}} \sim t_{n_i + n_{i'} - 2} \quad (two - sample\ case)$$

$S_{i \cdot i'}$ *is the pooled variance*

$$(ii) \quad \frac{\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot} - (\mu_i - \mu_{i'})}{\sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_{i'}})}} \sim t_{n_T - r} \quad (from\ the\ ANOVA\ table)$$

For a pairwise comparison, it is often to construct a (1-α)100% confidence interval for $\mu_i - \mu_{i'}$:

$$\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot} \pm qt(1 - \alpha/2, n_T - r) * \sqrt{MSE\Big(\frac{1}{n_i} + \frac{1}{n_{i'}}\Big)}$$

Kenton Food Company Example, page 677. There are 4 factor levels: package designs 1-4, with samples 5, 5, 4 & 5.

```
> y <- read.table("CH16TA01.DAT")
> data <- y[,1]
> package <- factor(rep(LETTERS[1:4],c(5,5,4,5)))
> food.df <- data.frame(package,data)
```

After checking assumptions via various graphs and tests, we obtain the ANOVA table

```
> anova <- aov(data~package,food.df)
> summary(anova)
             Df  Sum Sq  Mean Sq  F value   Pr(>F)
 package      3  588.22   196.07   18.591   2.585e-05
 Residuals   15  158.20    10.55
```

Since p-value=2.585e-05 is very small, the factor level means differ. The next step is to undertake the analysis of factor level effects. As an example, consider $\mu_3 - \mu_4$. For this, we need `model.tables(anova,type="means")`

```
> model.tables(anova,type="means")
 Tables of means
 Grand mean     18.63158

 package    A    B    C    D
         14.6 13.4 19.5 27.2
     rep   5.0  5.0  4.0  5.0

> meanC <- 19.5        #  <--- model.tables(anova,type="means")
> meanD <-  27.2
> nC <- 4
> nD <- 5
> MSE <- 10.55         #  <--- summary(anova)

> tval <- qt(1-.05/2, 15)
   [1] 2.131450

> ci <- c(meanC-meanD- sqrt(MSE * (1/nC+1/nD))* tval,
          meanC-meanD+sqrt(MSE * (1/nC+1/nD))* tval)
  [1]   -12.344164 -3.055836          # 95% CI
```

You may try the two sample method.

```
> t.test(y[,1][y[,2]==3], y[,1][y[,2]==4], var.equal=TRUE, conf.level=1-.05)

        Two Sample t-test

   data:  y[, 1][y[, 2] == 3] and y[, 1][y[, 2] == 4]
   t = -3.3175, df = 7, p-value = 0.01281
   alternative hypothesis: true difference in means is not equal to 0
   95 percent confidence interval:
      -13.188345  -2.211655
   sample estimates:
   mean of x mean of y
     19.5      27.2
```

Two methods give different answers. What information they provide for the difference? Which one would you prefer to?

3.3 Tukey multiple comparison procedure

$$\begin{array}{rcl} \text{The family of interest} & \triangleq & \{\text{all pairwise comparisons of factor level means}\} \\ & = & \{\mu_i - \mu_{i'} : \ \ i \neq i', i, i' \in \{1, \ldots, r\}\} \\ \text{Questions of interest} & & \\ & (i) & \text{simultaneous confidence intervals for all pairs } \mu_i - \mu_{i'} \\ & (ii) & \text{simultaneous tests of form } \mathcal{H}_0 : \mu_i - \mu_{i'} = 0 \end{array}$$

(1-α)100 % Tukey simultaneous confidence intervals for all pairwise comparisons $\mu_i - \mu_{i'}$:

$$\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot} \pm T * \sqrt{MSE\Big(\frac{1}{n_i} + \frac{1}{n_{i'}}\Big)}, \qquad i \neq i', i, i' \in \{1, \ldots, r\}$$

where

$$T \triangleq \frac{1}{\sqrt{2}} q(1-\alpha; r, n_T - r), \qquad \text{(Table B.9)}$$

For the balanced case (i.e., all sample sizes are equal, $n_1 = \ldots = n_r = n$), the probability statement is

$$P\Big\{\Big|\frac{(\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}) - (\mu_i - \mu_{i'})}{\sqrt{\frac{MSE}{n}}}\Big| \leq q(1-\alpha; r, n_T - r), \quad i, i' \in \{1, \ldots, r\}\Big\} = 1 - \alpha,$$

where $q(\cdot; r, n_T - r)$ is the quantile function of the studentized range distribution[1] with $r, n_T - r = r(n-1)$ for parameters. Thus, the family confidence coefficient for the Tukey method is exactly $1 - \alpha$ and the family significance level is exactly α.

For the unbalanced case, the Tukey procedure is conservative in the sense that the family confidence coefficient for the Tukey method is greater than $1 - \alpha$ and the family significance level is less than α.

[1] $\frac{\max_{1 \leq i \leq r} Z_i - \min_{1 \leq i \leq r} Z_i}{\sqrt{\chi_v^2 / v}}$, where $Z_1, \ldots, Z_r$ are independent normal $N(\cdot, 1)$ random variables, and independent of χ_v^2

Below is an example to write your own function

```
> tukey <- function(x,y){
            n1_length(x);
            n2_length(y);
            tukey <- c(mean(x)-mean(y)-Tval*sqrt(MSE*(1/n1+1/n2)),
                       mean(x)-mean(y)+Tval*sqrt(MSE*(1/n1+1/n2)));
            tukey
            }
```

where `MSE` is obtained from the ANOVA table, and `qval` is obtained from Table B.9.

Rust inhibitor example, page 712.

```
> y  <- read.table("CH17TA02.DAT")
> y1 <- y[,1][y[,2]==1]
> y2 <- y[,1][y[,2]==2]
> y3 <- y[,1][y[,2]==3]
> y4 <- y[,1][y[,2]==4]

> data  <- y[,1]
> brand <- factor(rep(LETTERS[1:4],c(10,10,10,10)))
> rust.df <- data.frame(brand,data)
```

After checking assumptions via various graphs and tests, we obtain the ANOVA table via

```
>  summary(aov(data~brand, rust.df))
```

or alternatively,

```
> anova(lm(data~brand, rust.df))
    Analysis of Variance Table

    Response: data
              Df   Sum Sq   Mean Sq  F value    Pr(>F)
     brand      3   15953.5  5317.8   866.12     < 2.2e-16
     Residuals 36    221.0     6.1
```

Now pick up `MSE` from the above table

```
>  MSE <-  anova(lm(data~brand, rust.df))[2,3]
```

Let $\alpha = 5\%$. From Table B.9, `q(1-0.05;4, 36)`=3.79 —

```
> Tval <- 1/sqrt(2)*3.79
```

Apply `tukey` to obtain $\binom{4}{2} = 6$ CIs

```
> tukey(y1, y2)
  -49.26973 -43.33027
> tukey(y1, y3)
  -27.77973 -21.84027
> tukey(y1, y4)
  -0.2997337  5.6397337
> tukey(y2, y3)
  18.52027 24.45973
> tukey(y2, y4)
  46.00027 51.93973
> tukey(y3, y4)
  24.51027 30.44973
```

Interpret these intervals appropriately.

Derivation (for the balanced case)

A necessary and sufficient condition that the inequalities

$$\frac{|(\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}) - (\mu_i - \mu_{i'})|}{\sqrt{MSE}} \le c$$

be satisfied for all $i, i' \in \{1, \ldots, r\}$ is for

$$\frac{\max_{i,i'} |(\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}) - (\mu_i - \mu_{i'})|}{\sqrt{MSE}} \le c,$$

or

$$\frac{\max_{i,i'} |(\bar{Y}_{i\cdot} - \mu_i) - (\bar{Y}_{i'\cdot} - \mu_{i'})|}{\sqrt{MSE/n}} \le c_1$$

to hold. Notice that

$$\max_{i,i'} |(\bar{Y}_{i\cdot} - \mu_i) - (\bar{Y}_{i'\cdot} - \mu_{i'})| = \max_{1 \le i \le r} (\bar{Y}_{i\cdot} - \mu_i) - \min_{1 \le i \le r} (\bar{Y}_{i\cdot} - \mu_i),$$

the range of r independent $\mathcal{N}(0, \sigma^2/n)$ r.v.s. It is independent of MSE. And $\frac{SSE}{\sigma^2} \sim ?$ Thus,

$$\frac{\max_{i,i'} |(\bar{Y}_{i\cdot} - \mu_i) - (\bar{Y}_{i'\cdot} - \mu_{i'})|}{\sqrt{MSE/n}} = \frac{\frac{\max_{i,i'} |(\bar{Y}_{i\cdot} - \mu_i) - (\bar{Y}_{i'\cdot} - \mu_{i'})|}{\sigma/\sqrt{n}}}{\sqrt{\frac{SSE}{\sigma^2}/(r(n-1))}}$$

follows a studentized range distribution with parameters $r, r(n-1)$.

Question 1: What do parameters r and $r(n-1)$ refer to ?

Question 2: At which step the above argument is not applicable to the unbalanced case?

3.4 Scheffé multiple comparison procedure

$$\begin{aligned} \text{The family of interest} \quad &\triangleq \quad \{\text{all possible contrasts among the factor level means}\} \\ &= \quad \{L = \sum_{i=1}^{r} c_i \mu_i : \quad \sum_{i=1}^{r} c_i = 0\} \end{aligned}$$

Questions of interest

(i) simultaneous confidence intervals for all possible contrasts L

(ii) simultaneous tests of form $\mathcal{H}_0 : L = 0$

(1-α)100 % Scheffé simultaneous confidence intervals for the family of contrasts L:

$$\sum_{i=1}^{r} c_i \bar{Y}_{i.} \pm S * \sqrt{MSE \sum_{i=1}^{r} \frac{c_i^2}{n_i}},$$

where

$$S^2 \triangleq (r-1) * qF(1-\alpha; r-1, n_T - r).$$

These simultaneous CIs may be defined as

```
>  scheffe <- function(coef){
             s.value    <- sqrt((r-1)*qf(1-alpha, r-1, nT-r));
             lowbound   <- sum(coef*level.mean)-
                            s.value*sqrt(MSE*sum(coef^2/level.size));
             upperbound <- sum(coef*level.mean)+
                           s.value*sqrt(MSE*sum(coef^2/level.size));
             scheffe    <- c(lowbound,  upperbound);
             scheffe
             }
```

where `MSE` is obtained from the ANOVA table, and `level.mean` and `level.size` are from `model.tables`.

Ex. 17.15 (c), Data set: Ch16pr10.dat

```
> y  <- read.table("CH16PR10.DAT")
> y1 <- y[,1][y[,2]==1]
> y2 <- y[,1][y[,2]==2]
> y3 <- y[,1][y[,2]==3]

> data  <- y[,1]
> size <- factor(rep(LETTERS[1:3],c(length(y1),length(y2), length(y3))))
> improv.df <- data.frame(data=data, size)
```

After checking assumptions via various graphs and tests, we obtain the ANOVA table via

```
> anova(lm(data~size, improv.df))
  Response: data
             Df   Sum Sq   Mean Sq   F value    Pr(>F)
  size       2    20.1252  10.0626   15.720    4.331e-05
  Residuals  24   15.3622   0.6401
```

Now pick up MSE from the above table

```
>  MSE <-  anova(lm(data~size, improv.df))[2,3]
```

The sample means and sizes are also needed.

```
>  model.tables(aov(data~size, improv.df), type="means")
   Grand mean  7.951852

   size       A      B   C
           6.878  8.133 9.2
   rep     9.000 12.000 6.0
 > level.mean <- c(6.878,  8.133, 9.2)     # or  c(mean(y1),mean(y2),mean(y3))
 > level.size <- c(9,12,6)                 # or  c(length(y1),length(y2),length(y3))
```

Let $\alpha = 5\%$. Apply scheffe to obtain CIs for various contrasts

```
 > scheffe(c(0,-1,1))           # for mu3-mu2
   0.02308538 2.11024795
 > scheffe(c(.5,.5,-1))         # for 1/2(mu1+mu2)-mu3
   -2.662847 -0.726042
```

Interpret these intervals appropriately.

Derivation

An unbiased estimator of $L = \sum_{i=1}^{r} c_i\mu_i$ is $\hat{L} = \sum_{i=1}^{r} c_i\bar{Y}_{i\cdot} \sim \mathcal{N}(L,\ \sigma^2 \sum_{i=1}^{r} \frac{c_i^2}{n_i})$. Thus,

$$\frac{\sum_{i=1}^{r} c_i\bar{Y}_{i\cdot} - \sum_{i=1}^{r} c_i\mu_i}{\sigma\sqrt{\sum_{i=1}^{r}\frac{c_i^2}{n_i}}} \sim \mathcal{N}(0,1).$$

As before,

$$\frac{SSE}{\sigma^2} \sim \chi^2_{n_T-r}.$$

Notice that $\hat{L}$ and SSE are independent. A pivotal quantity for a single contrast L is

$$\frac{\sum_{i=1}^{r} c_i\bar{Y}_{i\cdot} - \sum_{i=1}^{r} c_i\mu_i}{\sqrt{MSE\sum_{i=1}^{r}\frac{c_i^2}{n_i}}} \sim t_{n_T-r}.$$

What we really need is simultaneous confidence intervals for the family of contrasts L. A necessary and sufficient condition that the inequalities

$$\frac{\left|\sum_{i=1}^{r} c_i\bar{Y}_{i\cdot} - \sum_{i=1}^{r} c_i\mu_i\right|}{\sqrt{MSE\sum_{i=1}^{r}\frac{c_i^2}{n_i}}} \le c$$

be satisfied for all possible contrasts L is for[2]

$$\frac{\sum_{i=1}^{r} n_i(\bar{Y}_{i\cdot} - \mu_i)^2}{MSE} \le c^2$$

to hold. Check

$$\frac{\sum_{i=1}^{r} n_i(\bar{Y}_{i\cdot} - \mu_i)^2/(r-1)}{MSE} \sim F_{r-1,n_T-r}.$$

The probability statement for all possible contrasts is

$$P\Big(\frac{\left|\sum_{i=1}^{r} c_i\bar{Y}_{i\cdot} - \sum_{i=1}^{r} c_i\mu_i\right|}{\sqrt{MSE\sum_{i=1}^{r}\frac{c_i^2}{n_i}}} \le (r-1) * qF(1-\alpha; r-1, n_T-r), \forall\text{contrasts }\Big) = 1-\alpha$$

[2] *Lemma*: Let $c > 0$. Then

$$\frac{|\sum_{i=1}^{r} a_i y_i|}{\sqrt{\sum_{i=1}^{r} a_i^2}} \le c,\ \forall(a_1,\ldots,a_r) \iff \sum_{i=1}^{r} y_i^2 \le c^2$$

3.5 Bonferroni multiple comparison procedure

$$\begin{aligned} \text{The family of interest} &\triangleq \{\text{specified pairwise comparisons, contrasts, or linear combinations among the factor level means}\} \\ &= \{L = \sum_{i=1}^{r} c_i\mu_i\} \end{aligned}$$

Questions of interest

(i) simultaneous confidence intervals for g statements L

(ii) simultaneous tests of form $\mathcal{H}_0 : L = 0$

(1-α)100 % Bonferroni simultaneous confidence intervals for the g linear combinations L:

$$\sum_{i=1}^{r} c_i\bar{Y}_{i\cdot} \pm B * \sqrt{MSE\sum_{i=1}^{r}\frac{c_i^2}{n_i}},$$

where

$$B \triangleq qt(1-\frac{\alpha}{2g}; n_T - r).$$

Similar to Scheffé CIs, the Bonferroni simultaneous CIs may be defined as

```
>  bonferroni <- function(coef){
                 B.value    <- qt(1-alpha/(2*g), nT-r);
                 lowbound   <- sum(coef*level.mean)-
                               B.value*sqrt(MSE*sum(coef^2/level.size));
                 upperbound <- sum(coef*level.mean)+
                               B.value*sqrt(MSE*sum(coef^2/level.size));
                 bonferroni <- c(lowbound,  upperbound);
                 bonferroni
                 }
```

where `MSE` is obtained from the ANOVA table, and `level.mean` and `level.size` are from `model.tables`.

Derivation

Suppose that there are g statements in the group,

$$L^{(k)} = \sum_{i=1}^{r} c_i^{(k)} \mu_i, \qquad k = 1, \ldots, g,$$

where $c_i^{(k)}, i = 1, \ldots, r$ are coefficients.

A pivotal quantity for the kth statement $L^{(k)}$ is

$$\frac{\sum_{i=1}^{r} c_i^{(k)} \bar{Y}_{i\cdot} - \sum_{i=1}^{r} c_i^{(k)} \mu_i}{\sqrt{MSE \sum_{i=1}^{r} \frac{(c_i^{(k)})^2}{n_i}}} \quad \sim \quad t_{n_T - r}.$$

The corresponding probability statement for the kth statement is

$$P\Bigg(\frac{\left| \sum_{i=1}^{r} c_i^{(k)} \bar{Y}_{i\cdot} - \sum_{i=1}^{r} c_i^{(k)} \mu_i \right|}{\sqrt{MSE \sum_{i=1}^{r} \frac{(c_i^{(k)})^2}{n_i}}} \leq qt(1 - \alpha/(2 * g); n_T - r) \Bigg) = 1 - \alpha/g.$$

The question here is how to put all these into a single statement. To this end, we employ the **Bonferroni inequality**.[3] The probability statement for the g statements L is

$$P\Bigg(\frac{\left| \sum_{i=1}^{r} c_i^{(k)} \bar{Y}_{i\cdot} - \sum_{i=1}^{r} c_i^{(k)} \mu_i \right|}{\sqrt{MSE \sum_{i=1}^{r} \frac{(c_i^{(k)})^2}{n_i}}} \leq qt(1 - \alpha/(2 * g); n_T - r), \ k = 1, \ldots, g \Bigg) \geq 1 - \alpha$$

Question 1. What is the difference between the Scheffeé multiple comparison procedure and the Bonferroni multiple comparison procedure?

Question 2. What will happen for the Bonferroni multiple comparison procedure if g is large enough?

[3]The **Bonferroni inequality** for g events $A_1, \ldots, A_g$,

$$P(A_1 \cap A_2 \cap \cdots \cap A_g) \geq 1 - \sum_{i=1}^{g} P(\bar{A}_i).$$