Multiple comparisons and contrasts

Professor: Hammou El Barmi
Columbia University

Multiple Comparisons

- The ANOVA F-test checks whether all the population means are equal.
- Multiple comparisons are often used as a follow up to a significant ANOVA F-test to determine which population means are different.
- We will discuss Fisher's, Bonferroni's and Tukey's mehods for comparing all pairs of means
- In general to get more detailed information, we do inferences for treatment contrast which are differences between treatment means or more generally any linear contrast of treatment means, that is

$$\sum_{i=1}^{k} c_i \mu_i \qquad \text{where} \qquad \sum_{i=1}^{k} c_i = 0.$$

Least significant difference (LSD) method

- Fisher (1935) described a procedure for pairwise comparisons called the least significant difference (LSD) test.
- This test is to be used only if the hypothesis that all means are equal is rejected by the overall F test.
- If the overall test is significant, a procedure analogous to ordinary Student's t test is used to test any pair of means.
- If the overall F ratio is not significant, no further tests are performed. When it is used, the two treatments will be declared different if the absolute difference between two sample means that LSD. What is LSD?

Least significant difference (LSD) method

Fisher's least significant difference method (LSD) is a two step process

(1) Carry out the ANOVA F-test of $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$. If H_0 is not rejected stop and conclude that there insufficient evidence to claim differences among the population means. If H_0 is rejected, go to step 2

(2) Compare each pair of means using a pooled two sample t-test at the γ level using $s_{pooled} = \sqrt{MSE}$ from the ANOVA table and $df = df(SSE)$, that is test $H_0 : \mu_i = \mu_j$ against $H_a : \mu_i \neq \mu_j$ for all pair (i, j) using

$$t = \frac{\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}}{\sqrt{MSE}\sqrt{1/n_i + 1/n_j}}$$

and reject H_0 if $|t| > t_{n-k}(\gamma/2)$. of equivalently if

$$|\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}| > t_{n-k}(\gamma/2)\sqrt{MSE}\sqrt{1/n_i + 1/n_j}$$

- The minimum absolute difference between $\bar{Y}_{i\bullet}$ and $\bar{Y}_{j\bullet}$ needed to reject H_0 is the LSD, the quantity on the right hand side of the equation above
- If $n_1 = n_2 = \cdots = n_k$

$$LSD = t_{n-k}(\gamma/2)\sqrt{MSE}\sqrt{2/n_1}$$

Least significant difference (LSD) method

1. In our example $s_{spooled} = \sqrt{MSE} = \sqrt{67} = 8.18, n - k = 20$ and if $\alpha = 0.05, t_{20}(0.025) = 2.086$. Since $n_1 = n_2 = n_3 = n_4 = 6$,

$$LSD = 2.086 \times 8.18 \times \sqrt{2/6} = 9.85.$$

2. Any two sample means that differ by at least 9.85 in magnitude are significantly different at 5%.
3. One way to get Fisher comparisons in R uses pairwise.t.test() with p.adjsut.method.
4. The resulting summary of multiple comparisons is in terms of p-values for all pairwise two sample t-tests using the pooled standard deviation from the ANOVA using pool.sd=TRUE.

Least significant difference (LSD) method

```
> pairwise.t.test(data[,2],data[,1],pool.sd=TRUE,p.adjust.method="none" )

Pairwise comparisons using t tests with pooled SD

data:  data[, 2] and data[, 1]

     fat1  fat2    fat3
fat2 0.038 -       -
fat3 0.944 0.044   -
fat4 0.015 9.3e-05 0.013

P value adjustment method: none
```

Least significant difference (LSD) method

There are $c = 4(4-1)/2 = 6$ comparisons of two fats

Comparison	Absolute difference in means	Exceeds LSD	p-value
1 versus 2	10.50	Yes	0.038
1 versus 3	0.33	No	0.944
1 versus 4	12.50	Yes	0.015
2 versus 3	10.17	Yes	0.044
2 versus 4	23.00	Yes	9.3×10^{-5}
3 versus 4	12.83	Yes	0.013

There are three groups here $\{4\}, \{1,3\}$ and $\{2\}$

Bonferroni method

- If the F-test indicates that a factor is significant, then any pair of means that differ by at least LSD are considered to be different.
- This is the least conservative of all the procedures, because no adjustment is made for multiple comparisons (so when doing lots of comparisons this makes Type I errors likely)
- The Bonferroni method controls the probability of type I error by reducing the individual comparison rate
- The probability of type I error is guaranteed to be no larger than a pre-specified amount say γ by setting the individual error rate for each of the $k(k-1)/2$ comparisons of interest equal to

$$\gamma^* = \frac{\gamma}{k(k-1)}$$

- To implement the Bonferroni adjustment in R use p.adjust.method="bonf"

Bonferroni method

```
> pairwise.t.test(data[,2],data[,1],pool.sd=TRUE,p.adjust.method="bonf" )

Pairwise comparisons using t tests with pooled SD

data:  data[, 2] and data[, 1]

     fat1    fat2    fat3
fat2 0.22733 -       -
fat3 1.00000 0.26241 -
fat4 0.09286 0.00056 0.07960

P value adjustment method: bonferroni
```

Tukey method

- The LSD and Bonferroni methods comprise the ends of the spectrum of multiple comparisons methods
- Among multiple comparisons procedure, the LSD method is the most likely to find differences whether real of due to variation while Bonferroni is often the most conservative method
- The Bonferroni method is conservative but tend to work well when the number of comparisons is small, say 4 or less
- For $r > 4$, Bonferroni starts to get much more conservative than necessary

Tukey method

- Another multiple comparisons procedure is Tukey's method (a.k.a. Tukey's Honest Significance Test). The function TukeyHSD() creates a set of confidence intervals on the differences between means with the specified family-wise probability of coverage.
- The general form is TukeyHSD(fit, conf.level = 0.95). Here fit is a fitted model object (e.g., an aov.fit) and conf.level is the confidence level.
- Tukey's method is designed for equal sample sizes but can be used for different sample sizes too.
- The method rejects the equality of a pair of means based on the studentized range distribution. To implement this method at α, reject $H_0 : \mu_i = \mu_j$ when

$$|\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}| > \frac{q(1-\gamma, k, n-k)}{\sqrt{2}} \sqrt{MSE} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where $q(1-\gamma, k, n-k)$ is the γth level critical value of the studentized range distribution

Tukey method

```
>fit<-aov(data[,2] ~ data[,1])
> TukeyHSD(fit)
  Tukey multiple comparisons of means
    95\% family-wise confidence level

Fit: aov(formula = data[, 2] ~ data[, 1], data = data)

$`data[, 1]`
                  diff        lwr        upr     p adj
fat2-fat1  10.5000000  -2.719028 23.7190277 0.1510591
fat3-fat1   0.3333333 -12.885694 13.5523611 0.9998693
fat4-fat1 -12.5000000 -25.719028  0.7190277 0.0679493
fat3-fat2 -10.1666667 -23.385694  3.0523611 0.1709831
fat4-fat2 -23.0000000 -36.219028 -9.7809723 0.0004978
fat4-fat3 -12.8333333 -26.052361  0.3856944 0.0590077
```

Inference for a contrast of level means

- Suppose that chainsaws A & B were homeowner models and C & D were industrial grade. Now additional comparisons can be made
  - Homeowner vs. Industrial
    $$H_0 : \mu_A + \mu_B = \mu_C + \mu_D$$
  - Model A vs Model C
    $$H_0 : \mu_A = \mu_C$$
  - Model A vs Model C
    $$H_0 : \mu_B = \mu_C$$
- In all these cases H_0 can be expressed as

$$L = c_1\mu_A + c_2\mu_B + c_3\mu_C + c_4\mu_D = 0 \quad \text{where} \quad c_1 + c_2 + c_3 + c_4 = 0$$

Inference for a contrast of the level means

- A contrast L is defined as a linear combination of the level means where the coefficient add up to zero. That is

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

- Examples:
  1. $L = \mu_2 - \mu_1$
  2. $L = \mu_3 - (\mu_1 + \mu_2)/2$
  3. $L = (\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4)/2$

Inference for a contrast of the level means

- We estimate $L = \sum_{i=1}^{k} c_i \mu_i$ by

$$\hat{L} = \sum_{i=1}^{k} c_i \bar{Y}_{i\bullet}$$

- We have

$$E(\hat{L}) = \sum_{i=1}^{k} c_i E(\bar{Y}_{i\bullet}) = \sum_{i=1}^{k} c_i \mu_i = L \quad (\hat{L} \text{ is an unbiased estimator of L})$$

and

$$Var(\hat{L}) = \sum_{i=1}^{k} c_i^2 Var(\bar{Y}_{i\bullet}) = \sigma^2 \sum_{i=1}^{k} \frac{c_i^2}{n_i}$$

This implies that

$$SE(\hat{L}) = \sqrt{MSE} \sqrt{\sum_{i=1}^{k} \frac{c_i^2}{n_i}}$$

Inference for a contrast of the level means

- A $100(1-\gamma)\%$ confidence interval for L is

$$\hat{L} \pm t_{n-k}(\gamma/2) SE(\hat{L})$$

- To test $H_0 : L = 0$ against $H_a : L \neq 0$, the test statistic is

$$t = \frac{\hat{L} - 0}{SE(\hat{L})}$$

and we reject H_0 is

$$|t| > t_{n-k}(\gamma/2)$$

or of

$$t^2 > F_{\gamma}(1, n-k)$$

Same technique works for linear combinations. Later we will look at multiple contrasts.

Inference for a contrast of the level means

```
> contrasts(brand)<-cbind(c(1,-1,-1,+1), c(1,0,0, -1), c(0,1,-1,0))
> fit<-aov(angle~brand, contrasts=contrasts(brand))
> summary.lm(fit)

Call:
aov(formula = angle ~ brand, contrasts = contrasts(brand))

Residuals:
   Min     1Q Median     3Q    Max
-16.00  -8.25   0.00   7.25  18.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   39.000      2.250  17.333 8.58e-12 ***
brand1        -7.000      2.250  -3.111  0.00672 **
brand2         1.000      3.182   0.314  0.75738
brand3        -3.000      3.182  -0.943  0.35980
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 10.06 on 16 degrees of freedom
Multiple R-squared:    0.4,Adjusted R-squared:  0.2875
F-statistic: 3.556 on 3 and 16 DF,  p-value: 0.03823
```

- Suppose we have k population with medians $\eta_1, \eta_2, \cdots, \eta_k$.
- Test

  $H_0 : \eta_1 = \eta_2 = \ldots = \eta_k$ against H_a : at least two of these medians are not equal

- We apply the Kruskal-Wallis test. And to do so we pool the responses from all groups and rank them; then we apply one way ANOVA to the ranks, not to the original observations.
- If $R_{i\bullet}$ = sum of the ranks corresponding to the data from ith sample, the Kruskal-Wallis test statistic is

$$KW = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_{i\bullet}^2}{n_i} - 3(n+1)$$

  and we reject H_0 if $KW > \chi^2_{k-1}(\alpha)$ or if $p - value < \alpha$.

```
> kruskal.test(data[,2]~data[,1])

Kruskal-Wallis rank sum test

data:  data[, 2] by data[, 1]
Kruskal-Wallis chi-squared = 13.249, df = 3, p-value = 0.004128
```