STAT 3119

Week3: 9/12/2019 @GWU

Outline Week 3

- A. Analysis of factor levels means (Ch 17.1-3)
- B. Simultaneous Inference Procedures
- Next week: **a)** Tuesday: review Chapter 17, and topics for Quiz #2 (analysis of factor level means) ; and start Chapter 18; **b)** Thursday: **Quiz #2**; Review on homeworks#1 and #2 and Quiz #1 by TA;

Review: Inference of factor level means

1. For these inference, because the denominator is based on MSE with $n_T - r$ df, all these t tests about the factor level means in one-factor ANOVA model follows the same t-distribution.
2. For each parameter of interest that we discussed, e.g. θ, we need to know what is the point estimator $\hat{\theta}$ and the estimator of the standard deviation $s(\hat{\theta})$, then we can make inference: (1- α) two-sided CI is

$$\hat{\theta} \pm t(1-\alpha/2, n_T - r)s(\hat{\theta})$$

The test statistic for the $H_0 : \theta = \theta_0$ would be

$$t^* = (\hat{\theta} - \theta_0)/s(\hat{\theta}) \sim t(n_T - r) \text{ distribution}$$

Summary table: find estimator and est SD to plug-in.

Parameter θ	Estimator $\hat{\theta}$	Estimated SD $s(\hat{\theta})$	(1-α) confidence interval $\hat{\theta} \pm t(1-\alpha/2, n_T - r)s(\hat{\theta})$	Test statistic for $H_0 : \theta = \theta_0$ $(\hat{\theta} - \theta_0)/s(\hat{\theta}) \sim t(n_T - r)$
μ_i	$\hat{\mu}_i = \overline{Y}_{i.}$	$s^2\{\overline{Y}_{i.}\} = \frac{MSE}{n_i}$	$\overline{Y}_{i.} \pm t(1-\alpha/2; n_T - r)s\{\overline{Y}_{i.}\}$	H_0: $\mu_i = c$ $t^* = \frac{\overline{Y}_{i.} - c}{s\{\overline{Y}_{i.}\}}$
$D = \mu_i - \mu_{i'}$	$\hat{D} = \overline{Y}_{i.} - \overline{Y}_{i'.}$	$s^2\{\hat{D}\} = MSE\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)$	$\hat{D} \pm t(1-\alpha/2; n_T - r)s\{\hat{D}\}$	H_0: $\mu_i - \mu_{i'} = 0$ $t^* = \frac{\hat{D}}{s\{\hat{D}\}}$
$L = \sum_{i=1}^{r} c_i \mu_i$	$\hat{L} = \sum_{i=1}^{r} c_i \overline{Y}_{i.}$	$s^2\{\hat{L}\} = MSE \sum_{i=1}^{r} \frac{c_i^2}{n_i}$	$\hat{L} \pm t(1-\alpha/2; n_T - r)s\{\hat{L}\}$	H_0: $\sum c_i \mu_i = c$ $t^* = \frac{\hat{L} - c}{s(\hat{L})}$

The linear combination L is the most general form, the single/pair difference/contrast of the factor level means μ_i are its special case.

Need for Simultaneous Inference Procedures (Ch 17.4)

The procedures for estimating and testing factor level means discussed up to this point have two important limitations:

- The confidence coefficient $1-\alpha$ for the estimation 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 more tests.

$\Longrightarrow$ Therefore, we can only make inference for a specific estimate, or its corresponding test at level α. If we make inferences on multiple CIs or multiple test results, the overall (experimentwise) type I error of these study results would be inflated.

$\Longrightarrow$ Assume we perform m **independent** tests for $H_{0j}, j = 1, ..., m$, each at α, then the probability to make **at least one** false rejection is given by

$$\alpha_{overall} = 1 - (1-\alpha)^m$$

Assume, $\alpha = 0.05$, when $m = 5$, $\alpha_{overall} = 0.23$; When $m = 10$, $\alpha_{overall} = 0.40$. The $\alpha_{overall}$ increases quickly with the numbers of tests performed.

- The confidence coefficient $1-\alpha$ and the specified significance level α are appropriate only if the estimate or test was not suggested by the data.

$\Longrightarrow$ Therefore, it is not appropriate to do "data snooping", e.g. only compare the observed extreme pairs of the factor means. This also increases the type I error.

Single vs. simultaneous inference

For example, we take $\alpha = 0.05$, and consider 95% confidence interval here. (Same idea for any α).

- Previously, for a single parameter θ (such as μ_i, D , L), we have 95% two-sided CI as $CI_0 = \hat{\theta} \pm t(.975, n_T - r)s(\hat{\theta})$, i.e.,
$$Pr(\theta \in CI_0) = 95\%$$
- If we have several or many parameters, $\theta_1, ..., \theta_m$, and we calculate their individual 95% CIs with the above formula, then we can NOT make the same claim that

$$Pr(\theta_1 \in CI_1, ..., \theta_m \in CI_m) = 95\%$$

This statement is not true from basic statistical theory: CI_i's would be too narrow to cover all parameters simultanously with 95% confidence.

- What we will discuss next is to find the adjustment for those CI_i's to satisfy

$$Pr(\theta_1 \in CI_1^A, ..., \theta_m \in CI_m^A) \geq 95\%$$

So the collection of adjusted CI_i^A's is a family of confidence intervals with confidence coefficent of 95%.

- For mutiple testing problem, we adjust the test statistics and rejection rules, so we can control the overall type I error of 5% for all the test results.

Common procedures for multiplicity adjustment

Different choices are available to adjust for multiple comparisons and they roughly fall into two categories:

- Change α level in the t-distribution
- Use a different distribution

To get the $(1-\alpha)$ simultaneous confidence interval, we have very similar formulation but using different multiplier:

$$\hat{\theta} \pm \mathbf{C}_{adjusted} * s(\hat{\theta})$$

We will consider 4 commonly-used testing/comparison procedures.

1. Fisher's LSD (Least Significant Difference)
2. Bonferroni
3. Tukey's HSD (honestly significant difference)
4. $Scheffé$

Fisher's LSD procedure

LSD (Least Significant Difference) is the "least conservative" procedure of the four procedures. It has 2 steps:

1. F-test for H_0 (equality of the means) is performed first.
2. Followed by t-tests among all pairs of means at level of α **only if F-test rejects the null hypothesis**. (Not to perform any pairwise comparison if we don't reject F-test.)

Note: The F-test provides the overall protection against rejecting H_0 when it is true. When the F-test rejects, the subsequent pairwise t-tests are each performed at α level and thus likely will reject more than they should.

Despite this drawback Fisher's LSD is still used since it has overall α level protection for H_0, and offers simplicity to understand and interpret.

Bonferroni procedure (17.7)

Bonferroni adjustment is most simple way to make simultaneous inference. If there are g parameters to make inference, we use a new adjusted $\alpha_B = \alpha/g$ level for the CI or testing.

For **CI Estimation**: if there are g linear combinations $L = \{L_1, ..., L_g\}$, the $(1-\alpha)$ confidence intervals are :

$$\hat{L} \pm Bs\{\hat{L}\} \qquad \textbf{(17.46)}$$

where:

$$B = t(1-\alpha/2g; n_T - r) \qquad \textbf{(17.46a)}$$

Note: $1-\alpha/(2g)$ is used for the t-critical value using Bonferroni adjustment instead of $(1-\alpha/2)$ without the adjustment.

Because $t(1-\alpha/(2g), df)$ will be bigger than $t(1-\alpha/2, df)$ (further away in the tail of t-distribution from the center), the simultaneous CIs for all $L_1, ...L_g$ are wider than the corresponding individual unadjusted CIs.

Example:

- When df=20, α=0.05, t(1- $\alpha/2$, 20)= t(.975, 20)= 2.086.
- If g=10, we use $\alpha_B = \alpha$/g=0.05/10=0.005, then t(1- $\alpha_B/2$, 20)= t(.9975, 20)= 3.153.

Bonferroni procedure (2)

For **Testing** of $H_0 : L = c$, we use the same test statistic,

$$t^* = (\hat{L} - c)/s(\hat{L})$$

but we compare $|t^*|$ with the $t(1 - \alpha/(2g), n_T - r)$.

Note:

1. Bonferroni correction is very simple to use if we only want to do comparisons for a small number of pairs of treatment means (pre-specified without "data snooping") or do a small number of tests.
2. Example: in a clinical trials of 2 active drugs (A, B) and one control (C) treatment, we may be primarily interesting to compare drug A vs. C, and drug B vs. C, but not A vs. B. In this case, we can take g = 2, and set the adjusted $\alpha/2 = 0.025$ for each test using the Bonferroni correction.
3. It is not necessary to use the same adjusted $\alpha_B = \alpha/g$ for the g comparisons/test. We can "budget" to spend the overall type I error α differently if some inferences are more important to give more weights. The overall α is protected if $\alpha_1 + ... + \alpha_g = \alpha$. For two tests, instead of $\alpha_1 = \alpha_2 = 0.025$, we can pre-specify to use $\alpha_1 = 0.04$ and $\alpha_1 = 0.01$ before looking at the data.

Bonferroni vs. LSD for pairwise comparisons

In the food examples that we previous discussed. We study the differences in Sales ~ 4 package designs and we reject the null hypothesis that all the packages have the same effects on sales using an ANOVA analysis.

R examples

If the LSD procedure is chosen, since the F-test rejects null, we can then perform the pairwise tests each at $\alpha = 0.05$ level, as we showed in last class.

```
library(emmeans)
Ex16 =  read.table(url("https://raw.githubusercontent.com/npmldabook/Stat3119/master/Week2/CH16_TA01.tx
names(Ex16) =  c("sales", "package", "stores")
Ex16$package =  as.factor(Ex16$package)

fit<- aov(sales~ package, data=Ex16)

Est.mean<- emmeans(fit, ~ package)
pairs(Est.mean, adjust = "none" )
```

```
##  contrast estimate   SE df t.ratio p.value
##  1 - 2         1.2 2.05 15  0.584  0.5677
##  1 - 3        -4.9 2.18 15 -2.249  0.0399
##  1 - 4       -12.6 2.05 15 -6.135  <.0001
##  2 - 3        -6.1 2.18 15 -2.800  0.0135
##  2 - 4       -13.8 2.05 15 -6.719  <.0001
##  3 - 4        -7.7 2.18 15 -3.534  0.0030
```

Now, if we want to use Bonferroni adjustment for the six pairwise tests, each testing at $\alpha = 0.05/6$, we can set the adjust="bonferroni" in the pairs() function.

```
pairs(Est.mean, adjust = "bonferroni" )
```

```
##  contrast estimate   SE df t.ratio p.value
##  1 - 2         1.2 2.05 15  0.584  1.0000
##  1 - 3        -4.9 2.18 15 -2.249  0.2397
##  1 - 4       -12.6 2.05 15 -6.135  0.0001
##  2 - 3        -6.1 2.18 15 -2.800  0.0808
##  2 - 4       -13.8 2.05 15 -6.719  <.0001
##  3 - 4        -7.7 2.18 15 -3.534  0.0180
##
## P value adjustment: bonferroni method for 6 tests
```

Note: As we can see, the **p.value** (max=1) with Bonferroni correction is six times the p-value without the correction.

Tukey multiple comparison procedure (Ch 17.5)

- Tukey method: *Honestly Significant Difference* (**HSD**) Test
- The family test of interest include all pairwise comparisons of factor level means $D = \mu_i - \mu_{i'}$, which is $\binom{r}{2}$=$r(r-1)/2$ pairs.
- Tukey procedure is based on the *studentized range distribution*, denoted by $q(r, n_T - r)$.

We obtain the Tukey simultaneous confidence intervals for all pairwise comparisons with overall type I error of α as follows (same estimate and SD, but adjust for the multiple of SD).

$$\hat{D} \pm \mathbf{T} s(\hat{D})$$

where $T = 1/\sqrt{2} * q(1-\alpha; r, n_T - r)$.

Simultaneous Testing of all pairwise comparisons

- We can simply determine for each interval whether zero is contained in the interval. If zero is contained, then we can't conclude the parameters in the comparisons were significantly different.
- or use test statistic:

  $$q^* = \frac{\sqrt{2}\hat{D}}{s(\hat{D})} \sim q(r, n_T - r)$$

  We reject H_0 if $|q^*| \geq q(1-\alpha; r, n_T - r)$, which we can find from the probability distribution table or from R.

R example: get critical value for $q(1-\alpha; r, v)$ from qtukey.

```
# use qtukey for tail value of Studentized Range Distribution
qtukey(.95, nm=4, df=36)
```

```
## [1] 3.808798
```

```
qtukey(.90, nm=4, df=15)
```

```
## [1] 3.539891
```

Example 1: Rust inhibitor example

In a study of the effectiveness of different rust inhibitors, four brands (A, B, C, D) were tested. Altogether, 40 experimental units were randomly assigned to the four brands, with 10 units assigned to each brand. The higher the coded value, the more effective is the rust inhibitor. This study is a completely randomized design, where the levels of the single factor correspond to the four rust inhibitor brands.

TABLE 17.2 Data and Analysis of Variance Results—Rust Inhibitor Example (data are coded).

(a) Data

	Rust Inhibitor Brand			
	A	B	C	D
j	$i = 1$	$i = 2$	$i = 3$	$i = 4$
1	43.9	89.8	68.4	36.2
2	39.0	87.1	69.3	45.2
3	46.7	92.7	68.5	40.7
...	...	...	...	...
8	38.9	88.1	65.2	38.7
9	43.6	90.8	63.8	40.9
10	40.0	89.1	69.2	39.7
$\bar{Y}_{i\cdot}$	43.14	89.44	67.95	40.47

$\bar{Y}_{\cdot\cdot} = 60.25$

(b) Analysis of Variance

Source of Variation	*SS*	*df*	*MS*
Between brands	15,953.47	3	5,317.82
Error	221.03	36	6.140
Total	16,174.50	39	

Rust example (2): ANOVA

We can use R to run the ANOVA model to fit this model.

```
# read data from week3 folder online
Ex17 =read.table(
        url("https://raw.githubusercontent.com/npmldabook/Stat3119/master/Week3/CH17_TA02.txt"))
```

```r
names(Ex17) =  c("response", "brand", "units")
Ex17$brand =  as.factor(Ex17$brand)

# relabel level from 1:4 to A to D
levels(Ex17$brand)<- LETTERS[1:4]
str(Ex17)
```

```
## 'data.frame':    40 obs. of  3 variables:
##  $ response: num  43.9 39 46.7 43.8 44.2 47.7 43.6 38.9 43.6 40 ...
##  $ brand   : Factor w/ 4 levels "A","B","C","D": 1 1 1 1 1 1 1 1 1 1 ...
##  $ units   : int  1 2 3 4 5 6 7 8 9 10 ...
```

```r
fit<- aov(response~ brand, data=Ex17)
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## brand        3  15953    5318   866.1 <2e-16 ***
## Residuals   36    221       6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rust example (3): simultaneous inference.

TABLE 17.3 Simultaneous Confidence Intervals and Tests for Pairwise Differences Using the Tukey Procedure—Rust Inhibitor Example.

	Test		
Confidence Interval	H_0	H_a	q^*
$43.3 \le \mu_2 - \mu_1 \le 49.3$	$\mu_2 = \mu_1$	$\mu_2 \ne \mu_1$	58.99
$21.8 \le \mu_3 - \mu_1 \le 27.8$	$\mu_3 = \mu_1$	$\mu_3 \ne \mu_1$	31.61
$-.3 \le \mu_1 - \mu_4 \le 5.7$	$\mu_1 = \mu_4$	$\mu_1 \ne \mu_4$	3.40
$18.5 \le \mu_2 - \mu_3 \le 24.5$	$\mu_2 = \mu_3$	$\mu_2 \ne \mu_3$	27.37
$46.0 \le \mu_2 - \mu_4 \le 52.0$	$\mu_2 = \mu_4$	$\mu_2 \ne \mu_4$	62.39
$24.5 \le \mu_3 - \mu_4 \le 30.5$	$\mu_3 = \mu_4$	$\mu_3 \ne \mu_4$	35.01

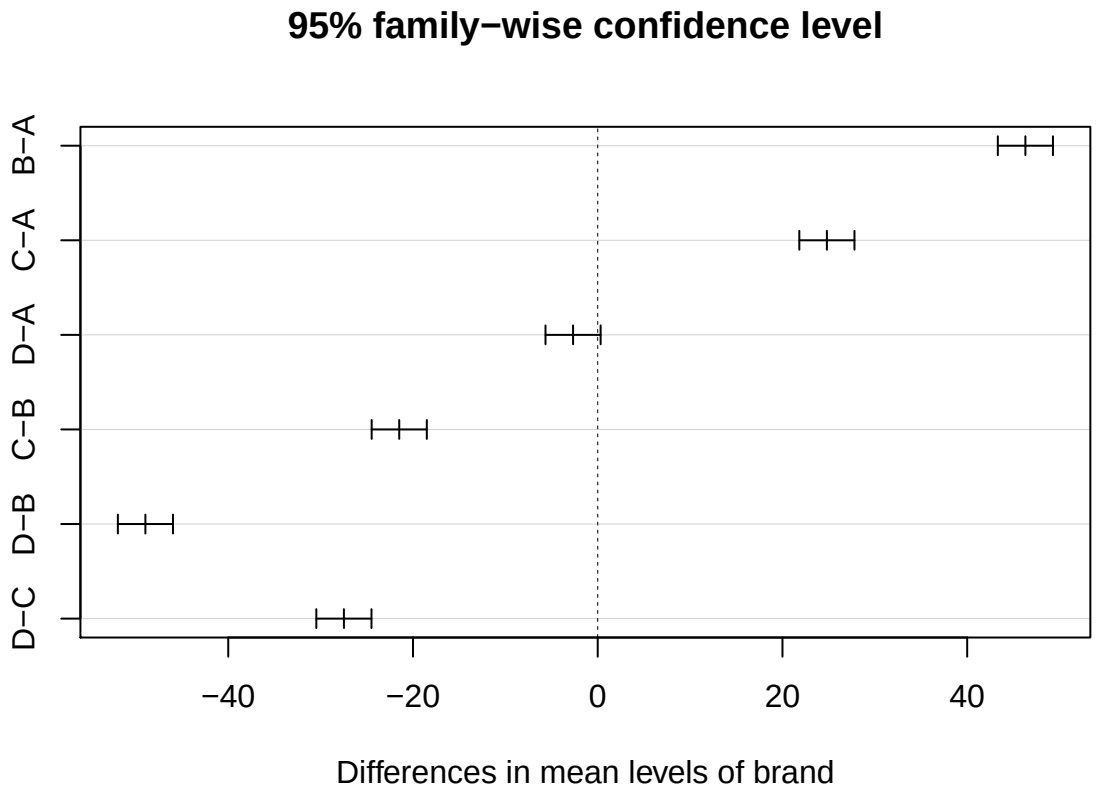
R example: we can use a build-in function TukeyHSD.

```r
## Tukey HSD with built-in function, including plots:
TukeyHSD(fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = response ~ brand, data = Ex17)
##
```

```
## $brand
##       diff        lwr         upr     p adj
## B-A  46.30  43.315536  49.2844635 0.0000000
## C-A  24.81  21.825536  27.7944635 0.0000000
## D-A  -2.67  -5.654464   0.3144635 0.0933303
## C-B -21.49 -24.474464 -18.5055365 0.0000000
## D-B -48.97 -51.954464 -45.9855365 0.0000000
## D-C -27.48 -30.464464 -24.4955365 0.0000000
```

```
plot(TukeyHSD(fit))
```



From the R output, we can find the p-value adjusted for the multiple comparisons, showed pair (A, D) are not significantly different (D-A pair overlaps with the vertical line at 0).

Also this plots show the order of the effects for the four brand from left to the right with biggest positive difference to the biggest negative difference: (1) B » A, C>A; (2) D~A (3) C< B, D« B, D < C. Therefore, B is best, C is second best, and D and A follow substantially behind with little difference between D & A.

Example 2: Kenton Food Company example with unequal n_i

When the Tukey procedure is used with unequal sample sizes, it is sometimes called the **Tukey-Kramer procedure**.

We have studied this example and reject the F-test that all 4 packages are the same. The simultaneous inference shown in example 2 (page 750).

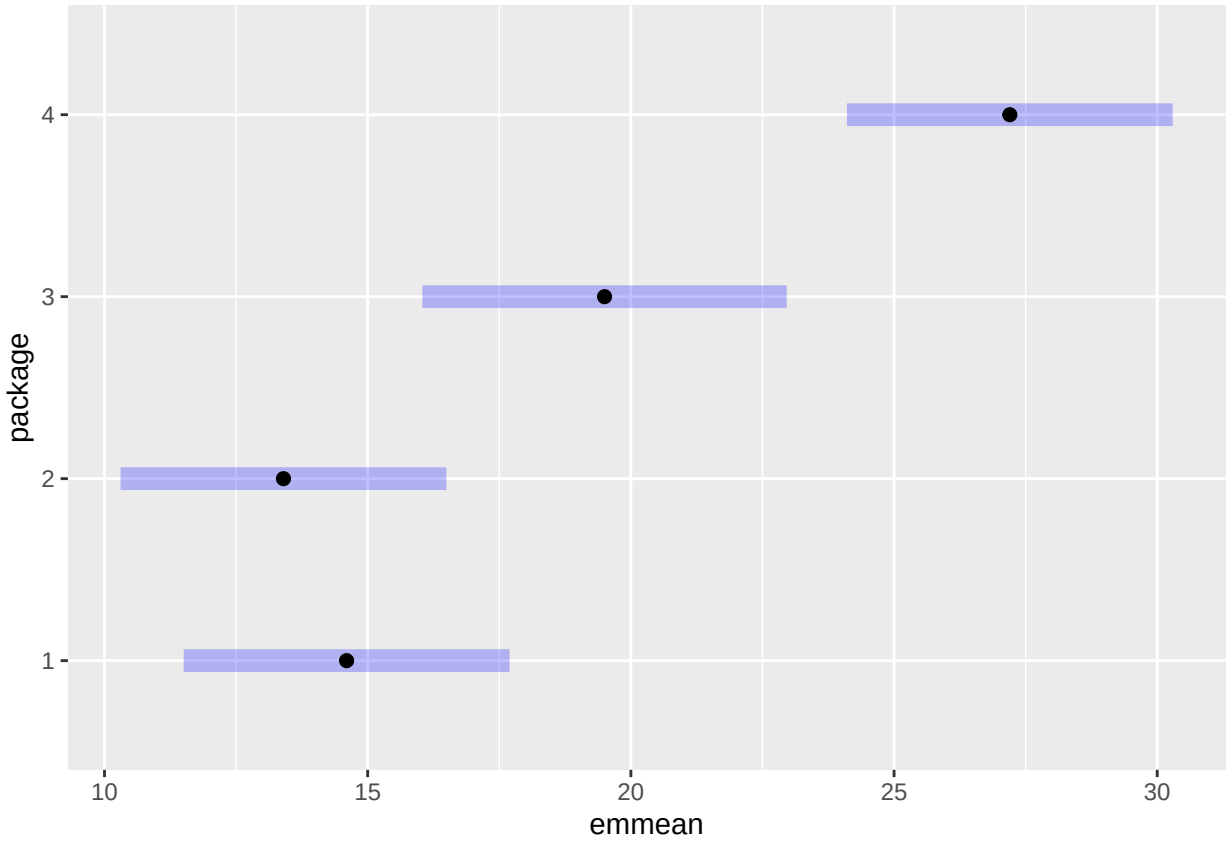
```r
fit2<- aov(sales~ package, data=Ex16)

## Tukey HSD with built-in function, including plots,
#   reset conf.level=0.90 (0.95 by default)
TukeyHSD(fit2, conf.level = 0.90)
```

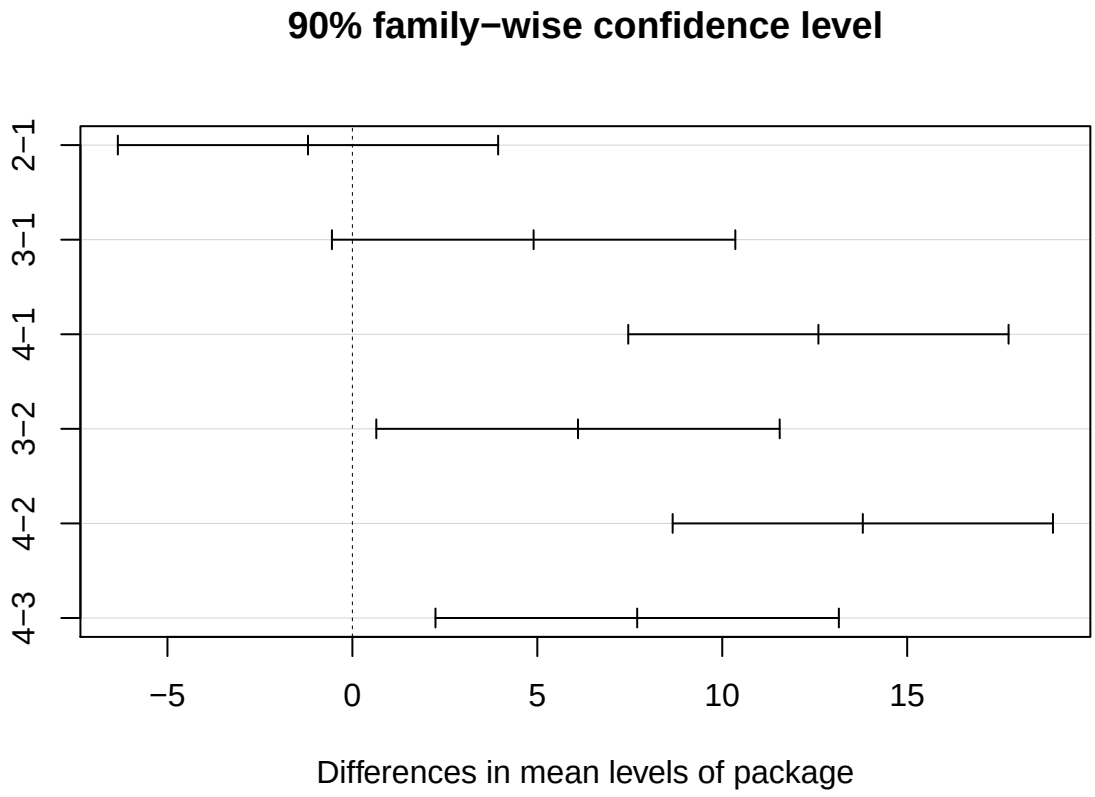
```
##   Tukey multiple comparisons of means
##     90% family-wise confidence level
##
## Fit: aov(formula = sales ~ package, data = Ex16)
##
## $package
##     diff        lwr       upr     p adj
## 2-1 -1.2 -6.3411769  3.941177 0.9352978
## 3-1  4.9 -0.5530416 10.353042 0.1548895
## 4-1 12.6  7.4588231 17.741177 0.0001013
## 3-2  6.1  0.6469584 11.553042 0.0582866
## 4-2 13.8  8.6588231 18.941177 0.0000368
## 4-3  7.7  2.2469584 13.153042 0.0142180
```

```r
# plot the mean with unadjusted CI and pairwise comparison#
# library(emmeans) # load library first

plot(emmeans(fit2, ~ package))
```



```
plot(TukeyHSD(fit2, conf.level = 0.90))
```



From this plot, we can see the comparison of 4-3, 4-2, 4-1 and 3-2 are on the right side of vertical 0 line. This indicate that design 4 was best, and design 3 is more effective than design 2 but may not be more effective than design 1, which in turn may not be more effective than design 2.

Scheffe Procedure (17.6)

- For this procedure, the family of interest is the set of **all possible contrasts** among the factor level means and the family confidence level for the Scheffe procedure is exactly $1 - \alpha$.
- If the Scheffe test (F) rejects the null hypothesis at level α, then there exists at least one contrast which would be rejected using the Scheffe procedure at level α regardless how many contrasts are tested.

We obtain the Scheffe simultaneous confidence intervals for all contrasts L with overall type I error of α as follows (same estimate and SD, but adjust for the multiple of SD).

$$\hat{L} \pm \mathbf{S}s(\hat{L})$$

where $S = \sqrt{(r-1)F(1-\alpha; r-1, n_T - r)}$.

Simultaneous Testing of all contrasts

$$F^* = \frac{\hat{L}^2}{(r-1)s^2\{\hat{L}\}} \qquad (17.45)$$

Conclusion H_0 in (17.44) is reached at the α family significance level if $F^* \leq F(1-\alpha; r-1, n_T - r)$; otherwise, H_a is concluded.

Some quick comparison of different procedures

1. LSD protect the overall α for ANOVA test, but not individual tests; LSD is not preferred if to make simultaneous inferences for the individual tests.
2. If only all pairwise comparisons are to be made, the **Tukey method** will result in a narrower confidence limit, which is preferable. Both Bonferroni and Scheffe's methods are more conservative than Tukey's procedure.
3. Bonferroni method is simple and have a narrower CI than Tukey or Scheffe methods when only a few pairs or a few contrasts are to be performed or controlled for multiple testing.
4. Because Scheffe methods controls for all possible contrasts, it is often more conservative than Tukey/ Bonferroni. The Bonferroni procedure will be better than the Scheffe procedure when the number of contrasts of interest is about the same as or less than the number of factor levels. Indeed, the number of contrasts must exceed the number of factor levels a lot before the Scheffe procedure becomes better.
5. All three procedures are ofthe form "estimator +/- multiplier×SE." The difference among the three procedures is the multiplier. In any given problem, one may compute all these multiples and then select the one that is smallest. This choice is proper (depending on experiment setting $[r, n_i, n_T]$, not depending on the observed data).

R example: paired comparison for the Food example (Ex16)

A. Calculate the multiplier:

```
# Tukey multiple
1/sqrt(2)* qtukey(.95, nm=4, df=15)
```

```
## [1] 2.882149
```

```
#Bonferroni multiple, df=n_T- r=19-4=15
g=6 # choose(4, 2)
qt(1- 0.05/(2*g), 15)
```

```
## [1] 3.036283
```

```
# Scheffe multiple r-1=3, n_T-r=15
sqrt(3*qf(.95, 3, 15))
```

```
## [1] 3.140405
```

It shows Tukey multiple < Bonferroni multiple < Scheffe multiple, and Tukey method is preferable to give narrower CI.

B. Using pairs and look at the adjusted p-value

```r
## Tukey HSD with built-in function, including plots:
fit2<- aov(sales~ package, data=Ex16)
Est.mean<- emmeans(fit2, ~ package)
pairs(Est.mean, adjust = "Tukey" )
```

```
##  contrast estimate   SE df t.ratio p.value
##  1 - 2         1.2 2.05 15  0.584  0.9353
##  1 - 3        -4.9 2.18 15 -2.249  0.1549
##  1 - 4       -12.6 2.05 15 -6.135  0.0001
##  2 - 3        -6.1 2.18 15 -2.800  0.0583
##  2 - 4       -13.8 2.05 15 -6.719  <.0001
##  3 - 4        -7.7 2.18 15 -3.534  0.0142
##
## P value adjustment: tukey method for comparing a family of 4 estimates
```

```r
pairs(Est.mean, adjust = "bonferroni" )
```

```
##  contrast estimate   SE df t.ratio p.value
##  1 - 2         1.2 2.05 15  0.584  1.0000
##  1 - 3        -4.9 2.18 15 -2.249  0.2397
##  1 - 4       -12.6 2.05 15 -6.135  0.0001
##  2 - 3        -6.1 2.18 15 -2.800  0.0808
##  2 - 4       -13.8 2.05 15 -6.719  <.0001
##  3 - 4        -7.7 2.18 15 -3.534  0.0180
##
## P value adjustment: bonferroni method for 6 tests
```

```r
pairs(Est.mean, adjust = "scheffe" )
```

```
##  contrast estimate   SE df t.ratio p.value
##  1 - 2         1.2 2.05 15  0.584  0.9507
##  1 - 3        -4.9 2.18 15 -2.249  0.2125
##  1 - 4       -12.6 2.05 15 -6.135  0.0002
##  2 - 3        -6.1 2.18 15 -2.800  0.0895
##  2 - 4       -13.8 2.05 15 -6.719  0.0001
##  3 - 4        -7.7 2.18 15 -3.534  0.0248
##
## P value adjustment: scheffe method with dimensionality 3
```

We can see the Tukey's P-value are smallest, then the Bonferroni, and the Scheffe adjustment is most conservative.

Summary this week and homework

- Reading: Chapter 16.9; Chapter 17.1-17.7
- Homework assignment #2: **17.8, 17.11 and 17.14** (due 9/19 by 6 pm before class; submit online from blackboard).
- This homework used the same data sets as the last homework.
- Quiz #2 : next Thursday (inferences on factor levels means and linear combinations, multiple comparison procedures)