AMS 572 Data Analysis I
Analysis of Single Factor Experiments

Pei-Fen Kuan

Applied Math and Stats, Stony Brook University

Multiple Comparisons

- Suppose we do n independent tests, each with probability α of making a type I error
- Suppose all n null hypotheses are true
- What is the probability of making at least one type I error?

Multiple Comparisons

- Probability of rejecting at least one null hypothesis when n independent tests are carried out at the α level and each null hypothesis is true

	α		
n	0.01	0.05	0.10
1	0.01	0.05	0.10
2	0.02	0.10	0.19
3	0.03	0.14	0.27
4	0.04	0.19	0.34
5	0.05	0.23	0.41
10	0.10	0.40	0.65
20	0.18	0.64	0.88
100	0.63	0.99	1.00

Multiple Comparisons

- The probability of incorrectly rejecting at least one of the true null hypotheses in an experiment involving one or more tests or comparisons is called the *per experiment error rate (PEER)*
- PEER is also known as the *family-wise error rate (FWE)*

ANOVA and Multiple Comparisons

- Rejection of $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ does not indicate where the inequalities are
- For example,

$$H_a : \mu_1 = \mu_2 = \cdots = \mu_{K-1} \neq \mu_K$$

or

$$H_a : \mu_1 \neq \mu_2 \neq \cdots \neq \mu_{K-1} \neq \mu_K$$

- Usually we want to identify the inequalities

ANOVA

- Need a multiple comparisons method to test

$$H_0 : \mu_i = \mu_j \quad (i \neq j)$$

- Popular methods:
  - Scheffé
  - Tukey
  - Bonferroni (Sidak, Holm, Hochberg)

ANOVA: Scheffé

- For each pair of means, compute

$$t_{ij} = \frac{\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}}{\sqrt{\mathrm{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

- Rejection region

$$C_\alpha = \left\{ t_{ij} : |t_{ij}| > \sqrt{(K-1)F_{K-1,N-K,\alpha,U}} \right\}$$

- Note: if $n_i = n\ \forall i$, then the denominator of t_{ij} is always $\sqrt{2\mathrm{MSE}/n}$
- So we could also write the critical region in terms of the *minimum significant difference*

$$C_\alpha = \{ |\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}| > \sqrt{(K-1)F_{K-1,N-K,\alpha,U} \times 2\mathrm{MSE}/n}$$

Example: A study was conducted to compare the lung function of groups of smokers and non-smokers. Test the hypothesis if the lung function differs by smoking status.

Group	n_i	Mean (L/sec)	sd (L/sec)
Non-smokers	200	3.78	0.79
Passive smokers	200	3.30	0.77
Non-inhalers	50	3.32	0.86
Light smokers	200	3.23	0.78
Mod. smokers	200	2.73	0.81
Heavy smokers	200	2.59	0.82

Scheffé: Passive Smoking Example

Comparison	t_{ij}	Significant
NS-PS	6.02	yes
NS-NI	3.65	yes
NS-LS	6.90	yes
NS-MS	13.17	yes
NS-HS	14.92	yes
PS-NI	−0.16	no
PS-LS	0.88	no
PS-MS	7.15	yes
PS-HS	8.90	yes
NI-LS	0.71	no
NI-MS	4.68	yes
NI-HS	5.79	yes
LS-MS	6.27	yes
LS-HS	8.03	yes
MS-HS	1.76	no

ANOVA: Scheffé

- For each pair of means, we can also compute multiplicity adjusted confidence intervals using Scheffé's method

$$\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm \sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \times \sqrt{(K-1)F_{K-1,N-K,\alpha,U}}$$

- What happens when $K = 2$?

ANOVA: Tukey

- Alternative multiple comparisons approach to Scheffé
- Critical region

$$C_\alpha = \left\{ t_{ij} : |t_{ij}| > (\tilde{q}_{K,N-K,\alpha})/\sqrt{2} \right\}$$

  where $\tilde{q}_{k,m,\alpha}$ is the upper α quantile (1-CDF) of the *studentized range* given in Table A.7 of your textbook, where $P(Q_{k,m} > \tilde{q}_{k,m,\alpha}) = \alpha$.
- Multiplicity adjusted CIs

$$\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm \sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \times (\tilde{q}_{K,N-K,\alpha})/\sqrt{2}$$

ANOVA: Tukey

- What is the studentized range?
- Suppose $Y_1, \ldots, Y_k$ iid $N(\mu, \sigma^2)$
- Let s be an estimator for σ with m degrees of freedom, $s \perp Y_1, \ldots, Y_k$
- Then

$$\frac{Y_{(k)} - Y_{(1)}}{s}$$

has a studentized range distribution with parameters k and m

- Tukey is preferred to Scheffé in balanced designs where all pairwise comparisons are being considered.
- This is because Tukey confidence intervals will be narrower, thus easier to reject;
- Otherwise, use Scheffé or Bonferroni-type method (later in this section)

Bonferroni Method

- Let $A_1, A_2, \ldots, A_n$ be a set of events
- Bonferroni inequality

$$\Pr(A_1 \cup A_2 \cup \cdots \cup A_n) \leq \sum_{i=1}^{n} \Pr(A_i)$$

- Let A_i be the event that we reject H_{0i} when H_{0i} is true for $i = 1, 2, \ldots, n$

$$\Pr(A_i) = \alpha_i$$

Bonferroni Method

- Probability of at least one Type I error

$$\Pr(A_1 \cup A_2 \cup \cdots \cup A_n) \leq \sum_{i=1}^{n} \alpha_i$$

- If $\alpha_i = \alpha^*$ for all i,

$$\sum_{i=1}^{n} \alpha_i = n\alpha^*$$

- If we want $\Pr(A_1 \cup \cdots \cup A_n) \leq \alpha$, choose $\alpha^* = \alpha/n$
- For ANOVA with K groups,

$$\alpha^* = \frac{\alpha}{\binom{K}{2}}$$

Bonferroni Method: Passive Smoking Example

Bonferroni Method

- Definition: The significance level at which each test or comparison is carried out in an experiment is call the *per comparison error rate (PCER)*
- Bonferroni uses
$$\text{PCER} = \frac{\alpha}{\binom{K}{2}}$$
to ensure
$$\text{PEER} \leq \alpha$$
- Bonferroni-type improvements (Sidak, Holm, Hochberg, Westfall and Young) available.

Example: A deer hunter prefers to practice with several different rifles before deciding which one to use for hunting. The hunter has chosen five particular rifles to practice with this season. In one test to see which rifles could shoot the farthest and still have sufficient knock-down power, each rifle was fired six times and the distance the bullet traveled recorded. A summary of the sample data is listed below, where the distances are recorded in yards.

Rifle	Mean	Std. Dev.
1	341.7	40.8
2	412.5	23.6
3	365.8	62.2
4	505.0	28.3
5	430.0	38.1

(a) Are these rifles equally good? Test at $\alpha = 0.05$

(b) If these rifles are not equally good, use Tukey's procedure with α=0.05 to make pairwise comparisons among the five population means.

Example: Fifteen subjects were randomly assigned to three treatment groups X, Y and Z (with 5 subjects per treatment). Each of the three groups has received a different method of speed-reading instruction. A reading test is given, and the number of words per minute is recorded for each subject. The following data are collected:

X	Y	Z
700	480	500
850	460	550
820	500	480
640	570	600
920	580	610

Write a SAS and R program to answer the following questions.

(a) Are these treatments equally effective? Test at $\alpha = 0.05$.

(b) If these treatments are not equally good, use Tukey's procedure with $\alpha = 0.05$ to make pairwise comparisons.

SAS Code

This is one-way ANOVA with 3 samples and 5 observations per sample.

```
data reading;
input group $ words @@;
datalines;
X 700 X 850 X 820 X 640 X 920 Y 480 Y 460 Y 500
Y 570 Y 580 Z 500 Z 550 Z 480 Z 600 Z 610
;
run ;

proc anova data=reading ;
title Analysis of Reading Data ;
class group;
model words = group;
means group / tukey;
run;
/* replace tukey with scheffe for Scheffe's procedure*/
```

SAS Output

```
                              The ANOVA Procedure

Dependent Variable: WORDS

                                        Sum of
       Source                 DF       Squares     Mean Square    F Value    Pr > F

       Model                   2   215613.3333     107806.6667      16.78    0.0003

       Error                  12    77080.0000       6423.3333

       Corrected Total        14   292693.3333


                     R-Square     Coeff Var      Root MSE    WORDS Mean

                     0.736653      12.98256      80.14570      617.3333


       Source                 DF       Anova SS    Mean Square    F Value    Pr > F

       GROUP                   2   215613.3333     107806.6667      16.78    0.0003
^^L
                          Analysis of Reading Data
```

SAS Output

```
                         The ANOVA Procedure

            Tukey's Studentized Range (HSD) Test for WORDS

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher
                       Type II error rate than REGWQ.


            Alpha                                      0.05
            Error Degrees of Freedom                     12
            Error Mean Square                      6423.333
            Critical Value of Studentized Range     3.77289
            Minimum Significant Difference           135.23


         Means with the same letter are not significantly different.


          Tukey Grouping          Mean      N    GROUP

                       A        786.00      5    X

                       B        548.00      5    Z
                       B
                       B        518.00      5    Y
```

1. The p-value of the ANOVA F-test is 0.0003, less than the significance level $\alpha = 0.05$. Thus we conclude that the three reading methods are not equally good.
2. Furthermore, the Tukey's procedure with $\alpha = 0.05$ shows that method X is superior to methods Y and Z, while methods Y and Z are not significantly different.

R Code and Output

```
> nword <- c(700,850,820,640,920,480,460,500,570,580,
500,550,480,600,610)
> trt <- rep(c("X","Y","Z"),each=5)
> fit <- aov(nword~trt)
> summary(fit)
            Df Sum Sq Mean Sq F value   Pr(>F)
trt          2 215613  107807   16.78 0.000334 ***
Residuals   12  77080    6423
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1
```

R Code and Output

```
> TukeyHSD(fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = nword ~ trt)

$trt
    diff       lwr       upr     p adj
Y-X -268 -403.2303 -132.7697 0.0005212
Z-X -238 -373.2303 -102.7697 0.0013898
Z-Y   30 -105.2303  165.2303 0.8270221
```

R Code and Output

```
> library(DescTools)
>
> ScheffeTest(fit)

  Posthoc multiple comparisons of means: Scheffe Test
    95% family-wise confidence level

$trt
    diff    lwr.ci     upr.ci    pval
Y-X -268 -409.2984 -126.70163 0.00073 ***
Z-X -238 -379.2984  -96.70163 0.00192 **
Z-Y   30 -111.2984  171.29837 0.84144

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Simple linear regression vs one way ANOVA

- One way ANOVA can be fitted using linear regression model, where the covariate X is the factor variable representing the K groups.
- In linear regression, a factor variable will be converted into dummy variable.
- For the reading test example, one can also fit the model using `lm` function in `R`:

```
> nword <- c(700,850,820,640,920,480,460,500,570,580,
500,550,480,600,610)
> trt <- rep(c("X","Y","Z"),each=5)
> fit <- aov(nword~trt) ### this is what we did previously
> ### alternative fit using lm()
> fit2 <- lm(nword~trt)
> anova(fit2)
Analysis of Variance Table

Response: nword
          Df Sum Sq Mean Sq F value    Pr(>F)
trt        2 215613  107807  16.784 0.0003336 ***
Residuals 12  77080    6423
```

Read Chapter 14