MET CS 555 - Data Analysis and Visualization

Module-5: One-Way Analysis of Variance
Lecture - 9

Kia Teymourian

Boston University

Slides last compiled: 02/18/2019, Time: 00:39:35

Table of contents

1. One-Way Analysis of Variance

2. Pairwise Comparisons

3. Type I and type II errors

One-Way Analysis of Variance

One-Way Analysis of Variance

- Analysis of variance is a general term which involves breaking down the **overall variability in a particular continuous outcome into pieces.**
- It involves comparing the variability after accounting for a characteristic versus the remaining variability not explained by the characteristic and just inherent to the outcome.
- In **one-way analysis of variance (ANOVA)**, we study groups that are defined based on the value of **one factor**.
- **The goal of a one-way ANOVA is to compare means across groups.**
- In ANOVA framework, we make comparisons **across several groups** while considering all of the data together.

One-Way Analysis of Variance

To compare data across multiple groups, we will test the null hypothesis that the underlying population means are all equal versus the alternative that at least two of the underlying population means differ.

If we have **k groups** and we denote as the true population **mean for group i**, then the hypotheses for the **one-way ANOVA** can be written as follows:

- $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ (All underlying population means are equal)
- $H_1 : \mu_i \neq \mu_j$ **for some i and j**
  (At least two of the k underlying population means are different or not all of the underlying population means are the same/equal)

Example - One-Way Analysis of Variance

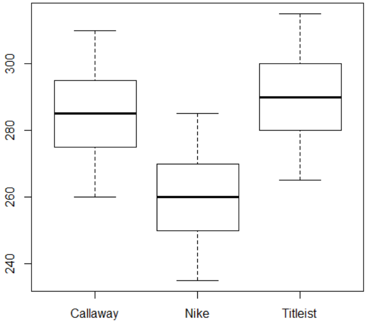
Commercials aired on TV entice potential buyers to consider purchasing specific brand of golf balls by claiming increased driving distances. In order to see which brand of golf balls is best (as measured by distance travelled), an experiment is set up where a mechanical driver hits 5 balls of each of 3 brands. The distance in yards achieved after each strike is measured.

Example - One-Way Analysis of Variance

Commercials aired on TV entice potential buyers to consider purchasing specific brand of golf balls by claiming increased driving distances. In order to see which brand of golf balls is best (as measured by distance travelled), an experiment is set up where a mechanical driver hits 5 balls of each of 3 brands. The distance in yards achieved after each strike is measured.

Table A. Distance by golf ball brand

Observation	Brand		
	Callaway	Nike	Titleist
1	275	235	265
2	310	285	300
3	285	270	280
4	260	250	315
5	295	260	290
Mean	285	260	290
Standard Deviation	19.0	19.0	19.0

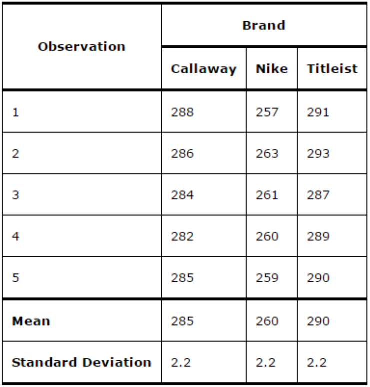


Example - One-Way Analysis of Variance

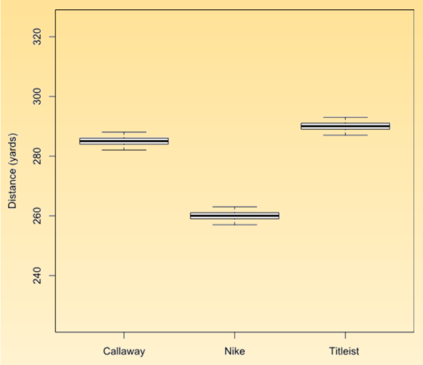
In both cases, the sample means for each of the brands is the same. The difference between the two versions of the examples are the amount of variability in the outcome.

When the variability is small relative to the differences between means, we become increasingly likely to declare that groups differ from each other.

Table B. Distance by golf ball brand



Observation	Brand		
	Callaway	Nike	Titleist
1	288	257	291
2	286	263	293
3	284	261	287
4	282	260	289
5	285	259	290
Mean	285	260	290
Standard Deviation	2.2	2.2	2.2



One-Way Analysis of Variance

If the **variability between groups**

- ▷ **is small** relative to the variability in the measurements **within groups**, we are less inclined to conclude that there **is a difference between them**.
- ▷ **is large** in comparison to the variability **within each group**, it is easier to see and conclude that **there is a difference.**

In ANOVA, we use the F-statistic for this test where the F-statistic is calculated as

$$\begin{aligned} F &= \frac{s_b^2}{s_w^2} \\ &= \frac{\text{between group variance}}{\text{within group variance}} \\ &= \frac{\text{mean square between}}{\text{mean square within}} \end{aligned}$$

One-Way Analysis of Variance

The numerator of F is an estimate of the between group variation. It shows how far the group means are from the overall mean (across all groups).

$$\begin{aligned} s_b^2 &= \text{mean square between} \\ &= \frac{\text{SSB}}{k-1} \\ &= \frac{\text{sum of squares between}}{\text{number of groups} - 1} \\ &= \frac{\sum_{j=1}^{k} n_{.j}(\bar{x}_{.j} - \bar{x}_{..})^2}{k-1} \end{aligned}$$

where

- $n_{.j}$is the number of observations in group j ,
- $\bar{x}_{.j}$ is the sample mean for group j,
- $\bar{x}_{..}$ is the overall mean (using all observations across all groups), and
- k is the number of groups.

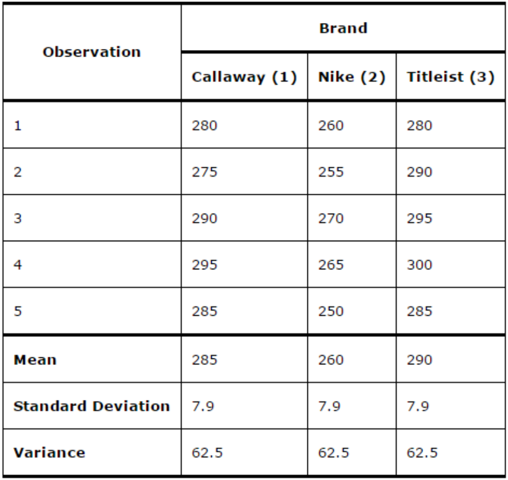
One-Way Analysis of Variance

The denominator of F is an estimate of the within group variation. How far each individual data point is from the corresponding groups mean and take a weighted average.

$$\begin{aligned}
s_w^2 &= \text{mean square within} \\
&= \frac{\text{SSW}}{n-k} \\
&= \frac{\text{sum of squares within}}{\text{number of observations} - \text{number of groups}} \\
&= \frac{\sum\sum(x_{ij} - \bar{x}_{.j})^2}{n-k} \\
\text{where} \qquad &= \frac{\sum_{j=1}^{k}(n_{.j}-1)s_j{}^2}{n-k}
\end{aligned}$$

- ▷ n is the number of observations across all groups,
- ▷ x_{ij} is the ith observation in the group j,
- ▷ $\bar{x}_{.j}$ is the sample mean for group j,
- ▷ $n_{.j}$ is the number of observations in group j, and
- ▷ k is the number of groups.

One-Way Analysis of Variance an example



Observation	Brand		
	Callaway (1)	Nike (2)	Titleist (3)
1	280	260	280
2	275	255	290
3	290	270	295
4	295	265	300
5	285	250	285
Mean	285	260	290
Standard Deviation	7.9	7.9	7.9
Variance	62.5	62.5	62.5

One-Way ANOVA An Example - Mean Square Between

$$
\begin{aligned}
\text{mean square between} &= s_b^2 \\
&= \frac{\text{sum of squares between}}{\text{number of groups} - 1} \\
&= \frac{\text{SSB}}{k-1} \\
&= \frac{\sum_{j=1}^{k} n_{.j}(\bar{x}_{.j} - \bar{x}_{..})^2}{k-1} \\
&= \frac{5 \cdot (285 - 278.33)^2 + 5 \cdot (260 - 278.33)^2 + 5 \cdot (290 - 278.33)^2}{3-1} \\
&= \frac{222.4445 + 1679.9445 + 680.9445}{2} \\
&= \frac{2583.3335}{2} \\
&= 1291.6668
\end{aligned}
$$

One-Way ANOVA An Example - mean square within

mean square within

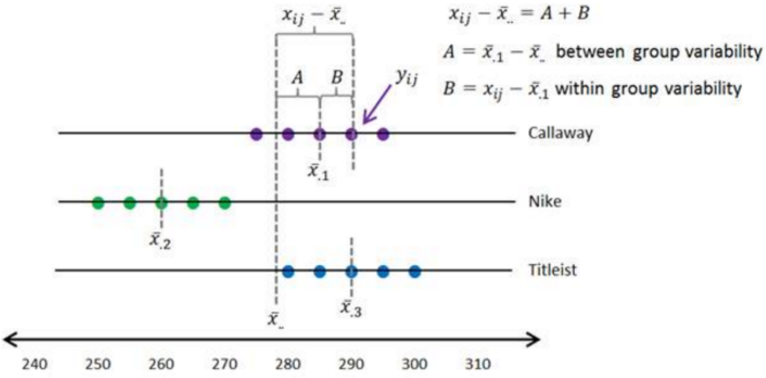
$$
\begin{aligned}
&= \frac{\text{sum of squares within}}{\text{number of observations} - \text{number of groups}} \\
&= \frac{\text{SSW}}{n-k} \\
&= \frac{\sum\sum(x_{ij} - \bar{x}_{.j})^2}{n-k} \\
&= \frac{(280-285)^2 + (275-285)^2 + \cdots + (285-285)^2 + \cdots + (280-290)^2 + (290-290)^2 + \cdots + (285-290)^2}{15-3} \\
&= \frac{750}{12} \\
&= 62.5
\end{aligned}
$$

One-Way ANOVA

The deviation between an individual point and the overall mean is comprised of two parts:

1. deviation between the individual point and the respective group mean,
2. deviation between the group mean and the overall mean

$$x_{ij} - \bar{x}_{..} = x_{ij} - \bar{x}_{.j} + \bar{x}_{.j} - \bar{x}_{..} = (x_{ij} - \bar{x}_{.j}) + (\bar{x}_{.j} - \bar{x}_{..})$$



$x_{ij} - \bar{x}_{..}$
A
B
y_{ij}
$x_{ij} - \bar{x}_{..} = A + B$
$A = \bar{x}_{.1} - \bar{x}_{..}$ between group variability
$B = x_{ij} - \bar{x}_{.1}$ within group variability
Callaway
$\bar{x}_{.1}$
Nike
$\bar{x}_{.2}$
Titleist
$\bar{x}_{.3}$
$\bar{x}_{..}$
240
250
260
270
280
290
300
310

One-Way Analysis of Variance

- ▷ The deviation between the individual point and the respective group mean is representative of the **within group variability**. The deviation between the group mean and the overall mean is representative of **the between group variability**.
- ▷ If the between group variability **is large** and the within group variability **is small**, then the **null hypothesis is often rejected** (in favor of the alternative hypothesis that the underlying means across group are not all equal).

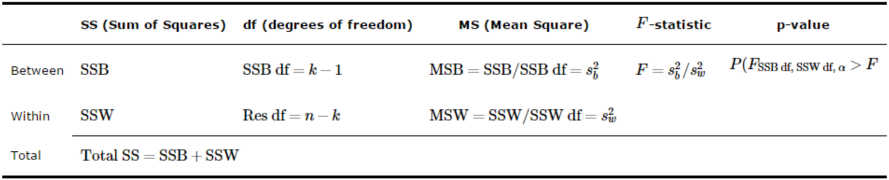
One-Way Analysis of Variance

- **Large values of the F-statistic** indicate that the variation between groups is larger than the variation within each individual group.
- To know how large is large enough to **reject the null hypothesis**, we use the **F-distribution with k-1 and n-k degrees of freedom**.

Inference - F-Test - Global Test

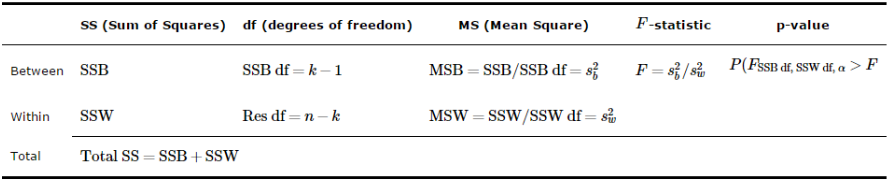
The F-test derived from the ANOVA table is sometimes referred to as the **global test**.

The global F-test only has the ability to conclude that **there are group differences (if the null hypothesis is rejected)**, but **does not allow one to know which groups are different** without taking further steps to investigate where the differences lie.



	SS (Sum of Squares)	**df (degrees of freedom)**	**MS (Mean Square)**	F**-statistic**	**p-value**
Between	SSB	$\mathrm{SSB\ df} = k - 1$	$\mathrm{MSB} = \mathrm{SSB}/\mathrm{SSB\ df} = s_b^2$	$F = s_b^2/s_w^2$	$P(F_{\mathrm{SSB\ df,\ SSW\ df},\,\alpha} > F$
Within	SSW	$\mathrm{Res\ df} = n - k$	$\mathrm{MSW} = \mathrm{SSW}/\mathrm{SSW\ df} = s_w^2$		
Total	$\mathrm{Total\ SS} = \mathrm{SSB} + \mathrm{SSW}$				

Inference - Anova table



	SS (Sum of Squares)	df (degrees of freedom)	MS (Mean Square)	F-statistic	p-value
Between	SSB	$\text{SSB df} = k - 1$	$\text{MSB} = \text{SSB}/\text{SSB df} = s_b^2$	$F = s_b^2/s_w^2$	$P(F_{\text{SSB df, SSW df}, \alpha} > F$
Within	SSW	$\text{Res df} = n - k$	$\text{MSW} = \text{SSW}/\text{SSW df} = s_w^2$		
Total	$\text{Total SS} = \text{SSB} + \text{SSW}$				

$$\text{SSB} = \sum_{j=1}^{k} n_{.j}(\bar{x}_{.j} - \bar{x}_{..})^2$$

$$\text{SSW} = \sum\sum(x_{ij} - \bar{x}_{.j})^2 = \sum_{j=1}^{k}(n_{.j} - 1)s_j{}^2$$

$$\text{Total SS} = \sum\sum(x_{ij} - \bar{x}_{..})^2$$

$$F = \text{MSB}/\text{MSW} = s_b^2/s_w^2$$

F-test for ANOVA: An Example

A random sample (n=19) of current light smokers, current heavy smokers, former smokers, and those who have never smoked was taken to determine if mean systolic blood pressure (SBP) differs across smoking status categories.

1. Set up the hypotheses and select the alpha level

$H_0 : \mu_{heavy} = \mu_{light} = \mu_{former} = \mu_{never}$
(All underlying population means are equal)

$H_1 : \mu_i \neq \mu_j$ for some i and j.
(Not all of the underlying population means are equal)

$\alpha = 0.05$

2. Select the appropriate test statistic

$F = \frac{MSB}{MSW}$ with $k - 1 = 3$ and $n - k = 19 - 4 = 15$ degrees of freedom

F-test for ANOVA: An Example (continued)

3. State the decision rule F-distribution with 3, 15 degrees of freedom and associated with $\alpha = 0.05$.

```
> qf(.95, df1=3, df2=15)
> F3,15,0.05=3.287
```

Decision Rule: Reject H_0 if $F \geq 3.287$

Otherwise, do not reject H_0

4. Compute the test statistic

$F = \frac{MSB}{MSW} = \frac{928.7}{43.2} = 21.49$

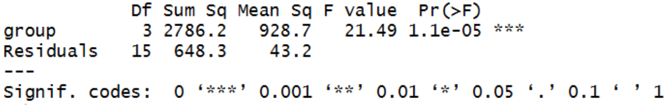
5. Conclusion

Reject H_0 since $21.49 \geq 3.287$.

We have significant evidence at the $\alpha = 0.05$ that there is a difference in SBP among current light smokers, current heavy smokers, former smokers, and those who have never smoked

F-test for ANOVA: An Example (continued) - Using R

```
> data <- read.csv("smoking_SBP.csv")

# Check if the variable group is factor variable in R.
# If not make it a factor variable

> is.factor(data$group)

# aov(data$response~data$group)
> m <- aov(data$SBP~data$group, data=data)

# pass the anova model object to the summary function.
> summary(m)
```



```
            Df Sum Sq Mean Sq F value  Pr(>F)    
group        3 2786.2   928.7   21.49 1.1e-05 ***
Residuals   15  648.3    43.2                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Pairwise Comparisons

Evaluating group differences - t-test

If the global **F-test is significant**, then it is of interest to further determine which of the population group means are different.

We perform testing on each **pairwise comparison of interest**.

We use a t statistic:

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{s_p^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

where

- $\boldsymbol{\bar{x}_i}$ **are** $\boldsymbol{\bar{x}_j}$ the sample mean in groups i and j (respectively),
- $\boldsymbol{s_p^2}$ is the estimate of the variance from the ANOVA model (and is equal to the mean square within, or s_w^2),
- n_i and n_j are the number of observations in groups i and j (respectively), which follows a t-distribution with $n - k$ degrees of freedom under H_0.

Evaluating group differences - t-test

The decision rule for a two-sided level α test is:

- **Reject $\boldsymbol{H_0 : \mu_i = \mu_j}$ if $\boldsymbol{t \geq t_{n-k,\frac{\alpha}{2}}}$ or if $\boldsymbol{t \leq -t_{n-k,\frac{\alpha}{2}}}$**
- Otherwise, do not reject $H_0 : \mu_i = \mu_j$

where $\boldsymbol{t_{n-k,\frac{\alpha}{2}}}$ is the value from the t-distribution table with $n - k$ degrees of freedom and associated with a right hand tail probability of $\alpha/2$.

Evaluating group differences - t-test - Confidence Interval

We can also calculate the two-sided $100\% \times (1 - \alpha)$ confidence interval for the difference between means $(\mu_i - \mu_j)$ using the following formula:

$$(\bar{x}_i - \bar{x}_j) \pm t_{n-k,\frac{\alpha}{2}} \cdot \sqrt{s_p^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

We can say with $100\% \times (1 - \alpha)$ confidence that difference between the underlying means of groups i and j is between the above values.

Example: t-test

The golf ball example.

1. Set up the hypotheses and select the alpha level

$H_0 : \mu_{\text{Titleist}} = \mu_{\text{Callaway}}$

$H_1 : \mu_{\text{Titleist}} \neq \mu_{\text{Callaway}}$

$\alpha = 0.05$

2. Select the appropriate test statistic

$$t = \frac{\bar{x}_{\text{Titleist}} - \bar{x}_{\text{Callaway}}}{\sqrt{s_p^2\left(\frac{1}{n_{\text{Titleist}}} + \frac{1}{n_{\text{Callaway}}}\right)}}$$

3. State the decision rule

Determine the appropriate value from the t-distribution table with n-k=15-3=12 degrees of freedom and associated with a right hand tail probability of $\alpha/2 = 0.05/2 = 0.025$

$t_{n-k,\frac{\alpha}{2}} = t_{12,\ 0.025} = 2.179$

```
> qt(.975, df=12)
> t12,0.025=2.179
```

Example: t-test (continued)

4. Compute the test statistic

$$t = \frac{\bar{x}_{\text{Titleist}} - \bar{x}_{\text{Callaway}}}{\sqrt{s_p^2\left(\frac{1}{n_{\text{Titleist}}} + \frac{1}{n_{\text{Callaway}}}\right)}}$$

$$= \frac{290 - 285}{\sqrt{62.5\left(\frac{1}{5} + \frac{1}{5}\right)}} = \frac{5}{\sqrt{25}} = \frac{5}{5} = 1$$

5. Conclusion

Do not reject H_0 since $1 < 2.179$.

We do not have significant evidence at the $\alpha = 0.05$ level that $\mu_{\text{Titleist}} \neq \mu_{\text{Callaway}}$

That is, we do not have evidence that the mean distances are different between Titleist golf balls and Callaway golf balls (p =0.34 as calculated using a software program).

Example: t-test (continued)

$$
\begin{aligned}
(\bar{x}_{\text{Titleist}} - \bar{x}_{\text{Callaway}}) \pm t_{n-k,\frac{\alpha}{2}} \cdot \sqrt{s_p^2\left(\frac{1}{n_{\text{Titleist}}} + \frac{1}{n_{\text{Callaway}}}\right)} \\
&= (290 - 285) \pm 2.18 \cdot \sqrt{62.5\left(\frac{1}{5} + \frac{1}{5}\right)} \\
&= 5 \pm 2.18 \cdot 5 \\
&= (-5.9, 15.9)
\end{aligned}
$$

We are 95% confident that the true difference between the two is between 5.9 yards (favoring the Callaway balls) and 15.9 yards (favoring the Titleist balls).

Example: t-test R command

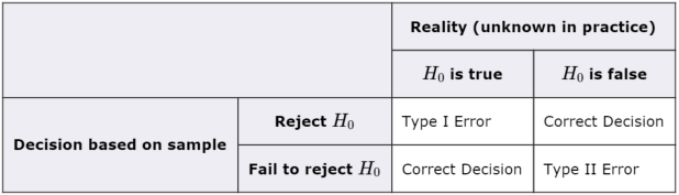
```
> golf <- read.csv("golfball_C.csv")
> golf
> attach(golf)
> aggregate(dist, by=list(brand), summary)
> aggregate(dist, by=list(brand), var)

> pairwise.t.test(dist, brand, p.adj=’none’)

Pairwise comparisons using t tests with pooled SD

data:  dist and brand

         Callaway Nike
Nike     0.00031  -
Titleist 0.33705  6.2e-05

P value adjustment method: none
```

Type I and type II errors

Type I and type II errors

Type I Error: rejecting the null hypothesis when it is true (false positive)

- The probability of making a **Type I Error** is **controlled by the significance level of the test**. It is generally the error rate that we worry the most about.
- The potential implications of this type of error explains **why we specifying small values of alpha** (the significance level).
- The potential implications of this type of error necessitates the **need to report p-values in your summary of results**.



		Reality (unknown in practice)	
		H_0 is true	H_0 is false
Decision based on sample	Reject H_0	Type I Error	Correct Decision
	Fail to reject H_0	Correct Decision	Type II Error

Type I and type II errors

- ▷ **Type II Error: failing to reject the null hypothesis when it is not true (false negative)**
- ▷ The probably of making a **Type II Error** is **controlled by the sample size**. The probability **decreases with increasing sample size**.
- ▷ The **Type II Error** is known as the **power of a test** (the probability of rejecting the null hypothesis when the null hypothesis is not true).

Type I and type II errors

- When we **fail to reject the null hypothesis**, it is often **difficult to know whether or not we didn't have enough power** or if the null hypothesis was indeed true.
- **This is why we dont say we accept the null hypothesis** and we instead state our conclusion as failing to reject the null hypothesis or we do not have sufficient evidence to reject the null hypothesis.
- Since the probability of **type II of error decreases as the sample size increases**, type II error rates of 10% or less are often selected for such settings.
  In most other cases, a type II error rate of 20% is generally considered reasonable for planning purposes.

Issues with multiple comparisons

If there are k groups, the number of possible pairwise comparisons is:

$$c = k \cdot (k - 1)/2$$

- If each of the possible $k \cdot (k-1)/2$ pairwise comparisons are performed at a significance level of α, then the expected number of false positives increases (up to as many as $\alpha \cdot c$).
- To maintain strong **control of the error rate** across all pairwise comparisons (often referred to as the experiment wise or family wise error rate) at a selected level, the significance level of each individual test needs to be adjusted.
- Procedures for controlling the family wise **type I error** rate at a pre-specified level are called multiple comparison procedures.
- These procedures work by essentially making it **“harder” to find differences between groups**. More evidence against the null hypothesis is needed to reject the null hypothesis when these procedures are implemented.

Issues with multiple comparisons - Bonferroni adjustment

Bonferroni adjustment: one of the simplest and most commonly used multiple comparisons procedures

To control the family wise error rate at the level, individual tests are performed at the $\boldsymbol{\alpha^* = \alpha/c}$ level of significance

where

- c is the number of individual comparisons to be performed
- α is the significance level of each individual test.

The **family wise error rate** is calculated using the formula

$$1 - (1 - \alpha) \cdot c$$

Tukey procedure

The Studentized Range test or **Tukeys Test of Honest Significance Test** is another common multiple comparisons procedure which is commonly used to control the family wise **type I error rate for pairwise comparisons**.

It tends to **have more statistical power (is less conservative) than the Bonferroni method**, especially when there are a **large number of pairwise comparisons**.

The procedure is generally implemented using statistical software like R.

Issues with multiple comparisons: the golf example

In the golf ball example, the global F-test showed that there was a difference in mean distance between brands.

Then we can perform three pairwise comparisons *(Titleist versus Callaway, Titleist versus Nike, and Callaway versus Nike).*

However, we did not account for the fact that we were doing all three comparisons and did each at the $\boldsymbol{\alpha = 0.05}$ level when really we had wanted **to control the family wise type I error rate at $\boldsymbol{\alpha = 0.05}$ overall**.

The Bonferroni methodology suggests that individual tests should be performed at the $\alpha^* = \alpha/c$ level of significance,

$$\boldsymbol{\alpha^* = \alpha/c = 0.05/3 \simeq 0.0167}$$

The critical value that we should have used in each comparison should have been

$$\boldsymbol{t_{n-k}, \alpha^*/2 = t_{12,0.00833} = 2.78}$$

instead of $\boldsymbol{t_{12,0.025} = 2.18}$.

Example SBP by smoking status

```
# Check if grouping variable (smoking status) is a factor
> is.factor(data$group)

# Numerical and graphical summaries (Module 1 and 2)
# Calculate mean, SD of SBP by groups
# Box plots and histograms
> aggregate(data$SBP, by=list(data$group), summary)
> aggregate(data$SBP, by=list(data$group), sd)
> boxplot(data$SBP~data$group, data=data, main="SBP by smoking status",
    xlab="group", ylab="SBP", ylim=c(100, 160))

# Perform one-way ANOVA and if necessary, calculate the associated
    pairwise comparisons

> m<- aov(data$SBP~data$group, data=data)
> summary(m)

# bonferroni adjustment
> pairwise.t.test(data$SBP, data$group, p.adj="bonferroni")

# Tukeys Test of Honest Significance Test
> TukeyHSD(m)
```

The golf example R command

Pairwise comparisons using t tests

```
> pairwise.t.test(dist, brand, p.adj='none')
Pairwise comparisons using t tests with pooled SD
data:  dist and brand
         Callaway Nike
Nike     0.00031  -
Titleist 0.33705  6.2e-05
P value adjustment method: none
```

P value adjustment method: bonferroni

```
> pairwise.t.test(dist, brand, p.adj='bonferroni')

Pairwise comparisons using t tests with pooled SD
data:  dist and brand
         Callaway Nike
Nike     0.00093  -
Titleist 1.00000  0.00019
P value adjustment method: bonferroni
```

The golf example R command

Create Anova Model and getting Summary results

```
# golfbal example - distances and different brands

> m <- aov(dist ~ brand, data=golf)
> summary(m)



# Tukey s Test of Honest Significance Test
> TukeyHSD(m)
```