Live Classroom

Module 5 –Analysis of Variance

Topics

- One-way ANOVA
  - Pairwise comparisons
  - Dummy variable creation
- One-way ANCOVA
  - LS means
- Two-way ANOVA
  - Interaction Plots

Introduction

- Earlier in the class, we looked at a way to compare means between two groups.
- But what if we have more than 2 groups?
- ANOVA (Analysis of Variance) can do this.
  - General term which involves breaking down variability in a particular continuous outcome into pieces.
  - It involves comparing the variability after accounting for a characteristic versus the remaining variability.
  - We'll first talk about one factor (one-way ANOVA), and then we'll discuss understanding the role of two factors at the same time.

One-Way ANOVA

- Seek to make comparisons across several groups.

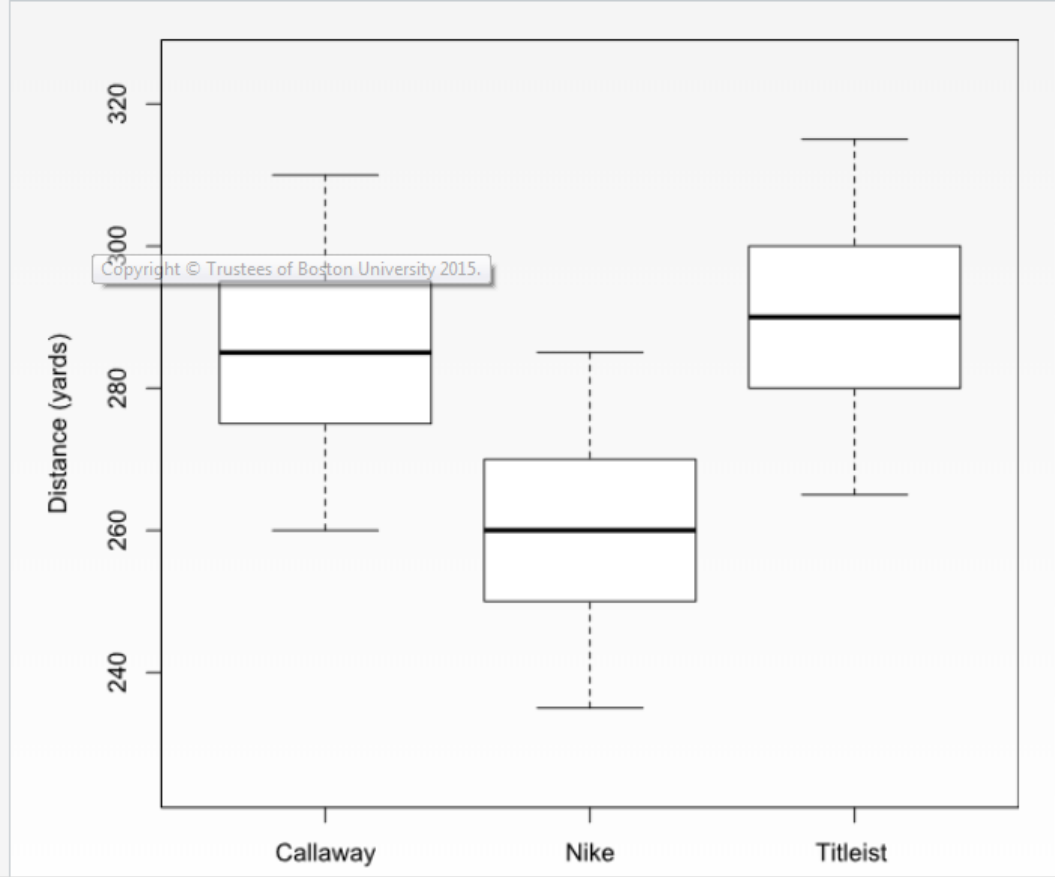
$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_1 : \mu_i \neq \mu_j \text{ for some } i \text{ and } j$$

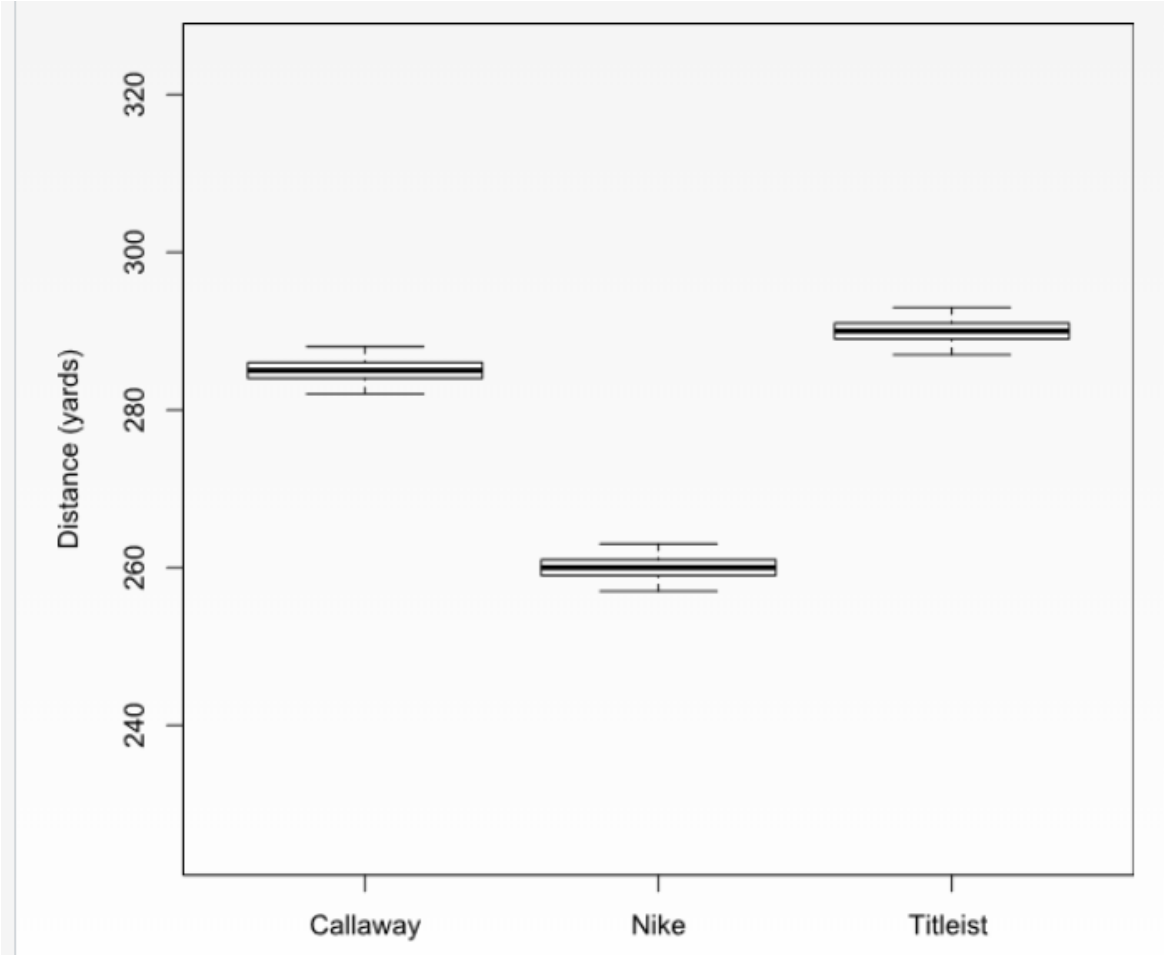
One-Way ANOVA

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 (to improve precision and consistency between hits) hits 5 balls of each of three brands. The distance in yards achieved after each strike is measured.

One-Way ANOVA



One-Way ANOVA



One-Way ANOVA

In general, 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.

On the other hand, if the variability between groups is large in comparison to the variability within each individual group, it is easier to see and conclude that there is a difference.

One-Way ANOVA

$$F = \frac{s_b^2}{s_w^2}$$

$$= \frac{\text{between group variance}}{\text{within group variance}}$$

$$= \frac{\text{mean square between}}{\text{mean square within}}$$

df = k-1 and n-k

The between group variance measures how far the group means are from the overall mean (across all groups).

The within group variance measures the variation among individuals in the same group.

One-Way ANOVA

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

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

One-Way ANOVA

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

where 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 ANOVA

In ANOVA, the F-test derived from the ANOVA table is sometimes referred to as the global test.

For this test, we use the ANOVA table and are interested in testing that at least two of the groups have underlying means that are different from each other.

If this test confirms that there are at least two groups with different means, then subsequent tests can be used to assess which groups have differing underlying means.

	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	$\text{SSW df} = n - k$	$\text{MSW} = \text{SSW}/\text{SSW df} = s_w^2$		
Total	$\text{Total SS} = \text{SSB} + \text{SSW}$				

Evaluating Group Differences

The global F-test above is used to test whether there are differences between the underlying group means. If the global F-test indicates that there are group differences (if the null hypothesis is rejected), then it is often of interest to take the additional steps needed to determine which of the population group means are different.

To determine where the differences lie, we perform testing on each pairwise comparison of interest. In order to test if $\mu_i = \mu_j$, for example, we use a t statistic:

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

Df = n-k

Equal to MSW

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

Issues with Multiple Comparisons

- When we perform tests for each pair of groups, we are performing k(k-1)/2 tests.
- This increases our chance for making a mistake and rejecting the null when we shouldn't!
- Now you can have up to a $\frac{\alpha(k(k-1))}{2}$ chance of making a mistake!
- In order to control the error rate across all tests (family wise error rate or experiment wise error rate), the significance level should be adjusted for each test.
- Bonferroni adjustment is popular and very conservative. Each test has significance level of $\frac{\alpha}{c}$, where c is the number of tests.

Assumptions of ANOVA

- Independent, random samples from each group.
- Outcome of interest in each group is normally distributed.
- Outcome of interest in each group has a similar unknown standard deviation.

** ANOVA is fairly robust with departures from normality. Large sample sizes are preferred if distributions do not look normal.

**ANOVA is also fairly robust from equal standard deviations (especially if sample sizes are similar across groups). Largest sample sd should be no larger than 2 times as large as the smallest sample sd.

Relationship between One way ANOVA and Regression

- We can use linear regression to conduct the same test.
- The explanatory variable indicates group membership.
- Need to create k-1 dummy variables!

$$\text{group}_2 = \begin{cases} 1, & \text{if observation is in group 2} \\ 0, & \text{otherwise} \end{cases}$$

$$\text{group}_3 = \begin{cases} 1, & \text{if observation is in group 3} \\ 0, & \text{otherwise} \end{cases}$$

$$\vdots$$

$$\text{group}_k = \begin{cases} 1, & \text{if observation is in group } k \\ 0, & \text{otherwise} \end{cases}$$

Relationship between One way ANOVA and Regression

- Need to create k-1 dummy variables!

Group	Dummy Variables			
	group_2	group_3	…	group_k
1	0	0		0
2	1	0		0
3	0	1		0
…				
k	0	0		1

$$y = \beta_0 + \sum_{i=2}^{k} \beta_{i-1}\text{group}_i + e$$

One-way ANOVA

- Use the aov() function
  - m<-aov(data$response~data$group)
  - summary(m)

  Note: group must be a factor for this to work
- If overall model is significant, then use pairwise.t.test to compare means
  - pairwise.t.test(data$response, data$group, p.adj = "*method*")

    Note: *method* = "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none"
  - TukeyHSD(m)

One-way ANOVA

- Alternatively, can use lm() function with dummy variables
  - m2<- lm(data$response~data$dummy1+data$dummy2…)
  - summary(m2)
  - This is equivalent to the overall model using the aov statement, but requires the creation of dummy variables

Dummy variable creation

- Indicator variables for each level of a grouping factor
- Regression requires k-1 dummy variables (one less than the number of levels of the variable)
- Level that is not included is the reference group
- Can use the ifelse() function
  - ifelse([*condition*], [*value if true*], [*value if false*])
  - data$dummy0<-ifelse(data$group == 0, 1, 0)

Example – SBP by smoking status

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

Example – SBP by smoking status

- Check if grouping variable (smoking status in this case) is a factor
  - is.factor(data$Group)
  - If result is FALSE, you need to create factor variable
    - data$grp<-factor(data$Group, levels = c(0,1,2,3))
      Note: levels controls the ordering
- Numerical and Graphical Summaries (Modules 1 and 2)
  - Calculate mean, SD of SBP by smoking status
  - Box Plots
- Perform one-way ANOVA and if necessary (based on the results), calculate the associated pairwise comparisons
  - m<-aov(SBP~grp, data = data)
  - summary(m)
  - pairwise.t.test(data$SBP, data$grpnew, p.adj = "bonferroni")
  - TukeyHSD(m)

Example – SBP by smoking status

- Creation of dummy variables
  - data$g0<-ifelse(data$Group == 0, 1, 0)
  - data$g1<-ifelse(data$Group == 1, 1, 0)
  - data$g2<-ifelse(data$Group == 2, 1, 0)
  - data$g3<-ifelse(data$Group == 3, 1, 0)
- One-way ANOVA using lm() function
  - m2<-lm(SBP~g0+g1+g2, data = data)
  - summary(m2)
  - m3<-lm(SBP~g1+g2+g3, data = data)
  - summary(m3)
  - m4<-lm(SBP~g0+g2+g3, data = data)
  - summary(m4)

One-Way ANCOVA

The one-way ANOVA procedure can be extended to adjust or control for one or more continuous or categorical variables. This extension is referred to as a one-way analysis of covariance (ANCOVA) model and can be represented in the regression framework.

In this setting, the model is given by

$$y = \beta_0 + \sum_{i=2}^{k} \beta_{i-1} \text{group}_i + \sum_{i=1}^{j} \beta_{k+i-1} x_i + e$$

Mainly interested in results of overall model, as well as the results of the grouping factor.

First, perform global test! If it's significant, move on to individual test for group.

Report lsmeans (or adjusted means) for each group.

One-way AN**C**OVA

- Do not use aov() function as it will not produce expected output
- Need to use Anova() function from car package to get "Type III" sums of squares
  - install.packages("car")
  - library(car)
  - options(contrasts=c("contr.treatment","contr.poly"))
  - Anova(lm(data$response~data$group+data$covariate), type=3)

One-way ANCOVA

- Generate Least Squares means (Covariate adjusted means) and comparisons
  - install.packages("lsmeans")
  - library(lsmeans)
  - options(contrasts=c("contr.treatment","contr.poly"))
  - lsmeans(lm(data$response~data$group+data$covariate), pairwise ~ group, adjust=[method])

    Note: *method* = "tukey" , "scheffe“, "sidak" , "bonferroni“, "dunnettx" "mvt”, “none”

Example – SBP by smoking status

- Re-run ANOVA adjusting for Age
  - library(car)
  - Anova(lm(SBP~grp+Age, data=data), type=3)
- Least Square Means
  - library(lsmeans)
  - options(contrasts=c("contr.treatment","contr.poly"))
  - lsmeans(lm(SBP ~grp+Age, data=data), pairwise ~ grp, adjust = "none")

Two-way ANOVA

- Use Anova() function
  - Anova([model], type=3)
- **First** test interaction model
  - Use model = lm(data$response~data$group1+data$group2+data$group1*data$group2)
  - Visualize relationship using interaction.plot() function
    - interaction.plot(data$group1, data$group2, data$Response)
- **Then**, if p-value for the interaction is not significant, then can run regular two-way ANOVA
  - Use model = lm(data$response~data$group1+data$group2)

Example – Scores by Gender and Birthday

- Test Interactions
  - Anova(lm(Reading~bday+gender+bday*gender, data=data),type = 3)
  - Anova(lm(Math~bday+gender+bday*gender, data=data),type = 3)
- Generate Interaction plots
  - interaction.plot(data$bday, data$gender, data$Math, xlab = "Birthday Month", ylab="Mean Math Score", trace.label = "Gender")
  - interaction.plot(data$bday, data$gender, data$Reading, xlab = "Birthday Month", ylab="Mean Reading Score", trace.label = "Gender")