Statistics for Engineers Lecture 8
One-Way Analysis of Variance

Chong Ma

Department of Statistics
University of South Carolina
chongm@email.sc.edu

April 10, 2017

Outline

One-Way ANOVA

Recall: In the last lecture, we talked about confidence intervals for the difference of two population means $\mu_1 - \mu_2$. More importantly, we saw that the design of the experiment or study completely determined how the analysis should proceed such as **(two) independent-sample design** and a **matched-pairs design**. In fact, the purpose of an **experiment** is to investigate differences between or among two or more treatments. In a statistical framework, we do this by comparing the population means of the responses to each treatment.

- In order to detect treatment mean differences, we must try to **control** the effects of errors so that any variation we observe can be attributed to the effects of the treatments rather than to structural differences among individuals.
- There may be a **systematic source of variation** arising from the ages of employees in the recycling project. Age of employees could be a **confounding effect** that lead to significant difference in two independent-sample design.

One-Way ANOVA

Terminology: Designs involving meaningful grouping of individuals, that is, blocking, can help reduce the effects of experimental error by identifying systematic components of variation among individuals. The matched-pairs design for comparing two treatments is an example of such a design.
Situation: When there are more than two treatments (populations) for an experiment, we pursue the **one-way classification model**. Such kind of experiments set up as follows.

- Obtain one random sample of individuals and then randomly assign individuals to treatments (i.e., different experimental conditions).
- In an **observational** study (where no treatment is physically applied to individuals), individuals are inherently different to begin with. Therefore we simply take random samples from each treatment populations.
- Do not attempt to group individuals according to some other factors (e.g., location, gender, weight, race, etc.).

One-Way ANOVA

Main point: In one-way classification, the only way individuals are "classified" is by the treatment group assignment. When individuals are thought to be "basically alike" (other than the possible effect due to treatment), experimental error consists only of the variation among the individuals themselves. There are no other **systematic** sources of variability.

Example Mortar mixes are usually classified on the basis of compressive strength and their bonding properties and flexibility. In a building project, engineers wanted to compare specially the population mean strengths of four types of mortars:

1. ordinary cement mortar (OCM)
2. polymer impregnated mortar (PIM)
3. resin mortar (RM)
4. polymer cement mortar (PCM)

One-Way ANOVA

Random samples of specimens of each mortar type were taken; each specimen was subjected to a compression test to measure strength (MPa). An initial question that engineers may have is the following

"Are the population mean mortar strengths equal among the four types of m

This initial question can be framed statistically as the following **hypothesis test:**

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \text{ v.s. } \mathrm{H}_1 : \text{at least one not equal}$$

Goal: We now develop a **statistical inference** procedure that allows us to test this type of hypothesis in a one-way classification.

One-Way ANOVA

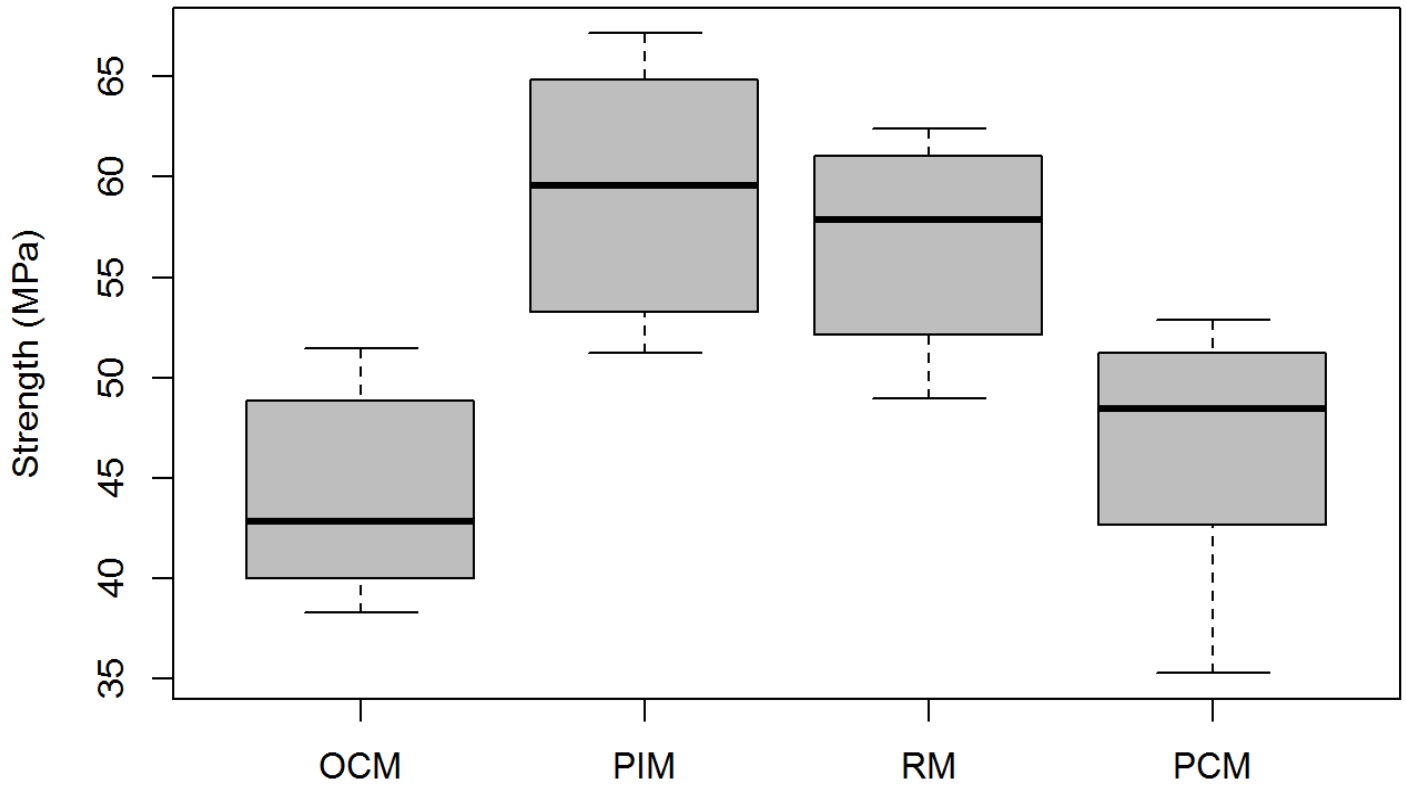

Figure 1: boxplots of strength data (MPa) for four mortar types

Outline

Overall F-Test

Denote by t the number of treatments (populations) to be compared. Define

$$Y_{ij} = \text{response on the jth individual in the ith treatment group}$$

for $i = 1, 2, \ldots, t$ and $j = 1, 2, \ldots, n_i$.

- n_i is the number of observations for the ith treatment. When $n_1 = n_2 = \ldots = n_t$, we say the design is **balanced**; otherwise, the design is **unbalanced**.
- Denote by $N = n_1 + n_2 + \ldots + n_t$ the total number of individuals measured. If the design is balanced, then $N = nt$.
- Define the statistics

$$\bar{Y}_{i+} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}, \quad S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i+})^2$$

$$\bar{Y}_{++} = \frac{1}{N}\sum_{i=1}^{t}\sum_{j=1}^{n_i} Y_{ij}$$

Overall F-Test

- The statistics $\bar{Y}_{i+}$ and S_i^2 denote the **sample mean** and the **sample variance**, respectively, of the ith treatment group. The **overall sample mean** $\bar{Y}_{++}$ is the sample mean of all the data (aggregated across all t treatment groups).
- The **null hypothesis** H_0 says that there is "no treatment difference", that is, all t populations means are the same.
- The **alternative hypothesis** H_1 syas that a difference among the t population means exits "somewhere". It does not specify how the means are different.
- When performing a hypothesis test, we basically decide which hypothesis is more supported by the data.

Overall F-Test

Setting: Suppose that we have t independent random samples:

$$\text{Sample 1: } Y_{11}, Y_{12}, \ldots, Y_{1n_1} \overset{i.i.d}{\sim} \mathcal{N}(\mu_1, \sigma^2)$$
$$\text{Sample 2: } Y_{21}, Y_{22}, \ldots, Y_{2n_2} \overset{i.i.d}{\sim} \mathcal{N}(\mu_2, \sigma^2)$$
$$\ldots\ldots\ldots \qquad \ldots\ldots\ldots\ldots$$
$$\text{Sample t: } Y_{t1}, Y_{t2}, \ldots, Y_{tn_t} \overset{i.i.d}{\sim} \mathcal{N}(\mu_t, \sigma^2)$$

Assumption: Note the statistical assumption we are making

1. the t random samples are **independent**
2. the t population distributions are **normal**
3. the t population distributions have the **same variance** σ^2

If we are trying to learn about how the population means compare, why is the statistical inference procedure designed to do this called "the analysis of variance"?

Overall F-Test

We learn about the population means by estimating the common variance σ^2 in two different ways. The two estimators are formed by

Within Estimator: measuring variability of the observations **within** each treatment group

Across Estimator: measuring variability of the sample means **across** the treatment groups

The two estimators tend to be similar when H_0 is true. The second estimator tends to be larger than the first estimates when H_1 is true.

Within Estimator: calculate the **residue sum of squares**:

$$\begin{aligned} SS_{res} &= (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \ldots + (n_t - 1)S_t^2 \\ &= \sum_{i=1}^{t} \underbrace{\sum_{j=1}^{n_i} (Y_ij - \bar{Y}_{i+})^2}_{(n_i - 1)S_i^2} \end{aligned}$$

Overall F-Test

Within Estimator

- The **residual mean squares** $MS_{res} = \frac{SS_{res}}{N-t}$ is an unbiased estimator of σ_2 regardless of whether H_0 or H_1 is true.
- The sample variance S_i^2 estimates the population parameters σ^2 (which assumed to be common across all t populations) from **within** the ith sample.
- The **within estimator** MS_{res} is a generalization of the pooled sample variance estimator S_P^2 in two-sample inference.

Across Estimator: We assume a common sample size $n_1 = n_2 = \ldots = n_t = n$ to simplify notation (i.e., a balanced design). The unbiased estimator of σ^2 (**when H_0 is true**) is

$$MS_{trt} = \frac{1}{t-1} \underbrace{\sum_{i=1}^{t} n(\bar{Y}_{i+} - \bar{Y}_{++})^2}_{SS_{trt}}$$

Overall F-Test

Recall that the sample mean is also normally distributed when the sample arise from a normal population. Therefore, the sample mean of the ith treatment group

$$\bar{Y}_{i+} \sim \mathcal{N}(\mu_i, \frac{\sigma^2)}{n}$$

When the null hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_t$ is true, we have

$$\bar{Y}_{1+}, \bar{Y}_{2+}, \ldots, \bar{Y}_{t+} \overset{i.i.d}{\sim} \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

Consider $\bar{Y}_{1+}, \bar{Y}_{2+}, \ldots, \bar{Y}_{t+}$ as a random sample from the $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ population distribution, the sample variance of this "random sample" is

$$\frac{1}{t-1}\sum_{i=1}^{t}(\bar{Y}_{i+} - \bar{Y}_{++})^2$$

and is an unbiased estimator of σ^2/n.

Overall F-Test

Therefore,

$$MS_{trt} = \frac{1}{t-1} \underbrace{\sum_{i=1}^{t} n(\bar{Y}_{i+} - \bar{Y}_{++})^2}_{SS_{trt}}$$

is an unbiased estimator of σ^2 when H_0 is true. If we have different sample size n_i, we simply adjust MS_{trt} to

$$MS_{trt} = \frac{1}{t-1} \underbrace{\sum_{i=1}^{t} n_i(\bar{Y}_{i+} - \bar{Y}_{++})^2}_{SS_{trt}}$$

This is still an unbiased estimator for σ^2 when H_0 is true.

Overall F-Test

Summary

1. **When H_0 is true**(the population means are equal), then

$$E(MS_{trt}) = \sigma^2, E(MS_{res}) = \sigma^2 \Rightarrow F = \frac{MS_{trt}}{MS_{res}} \approx 1$$

2. **When H_1 is true**(the population means are different), then

$$E(MS_{trt}) > \sigma^2, E(MS_{res}) = \sigma^2 \Rightarrow F = \frac{MS_{trt}}{MS_{res}} > 1$$

Sampling Distribution: When H_0 is true, the statistic

$$F = \frac{MS_{trt}}{MS_{res}} \sim F_{t-1,N-t}$$

Overall F-test

Recall the mean of an F distribution is around 1, therefore

- Values of F in the center of this distribution are consistent with H_0.
- Large values of F (out in the right tail) are consistent with H_1.
- Unusually small values of F (close to zero) are not necessarily consistent with either hypothesis. This is more likely to occur when there is a violation of our statistical assumptions such as correlated individuals within/across samples, unequal population variances, normality departures, etc.

Mortar data: Use R to calculate the F statistics.

```
> anova(lm(strength ~ mortar.type))
Analysis of Variance Table
Response: strength
             Df  Sum Sq Mean Sq F value    Pr(>F)
mortar.type   3 1520.88  506.96  16.848 9.576e-07
Residuals    32  962.86   30.09
```

Overall F-test

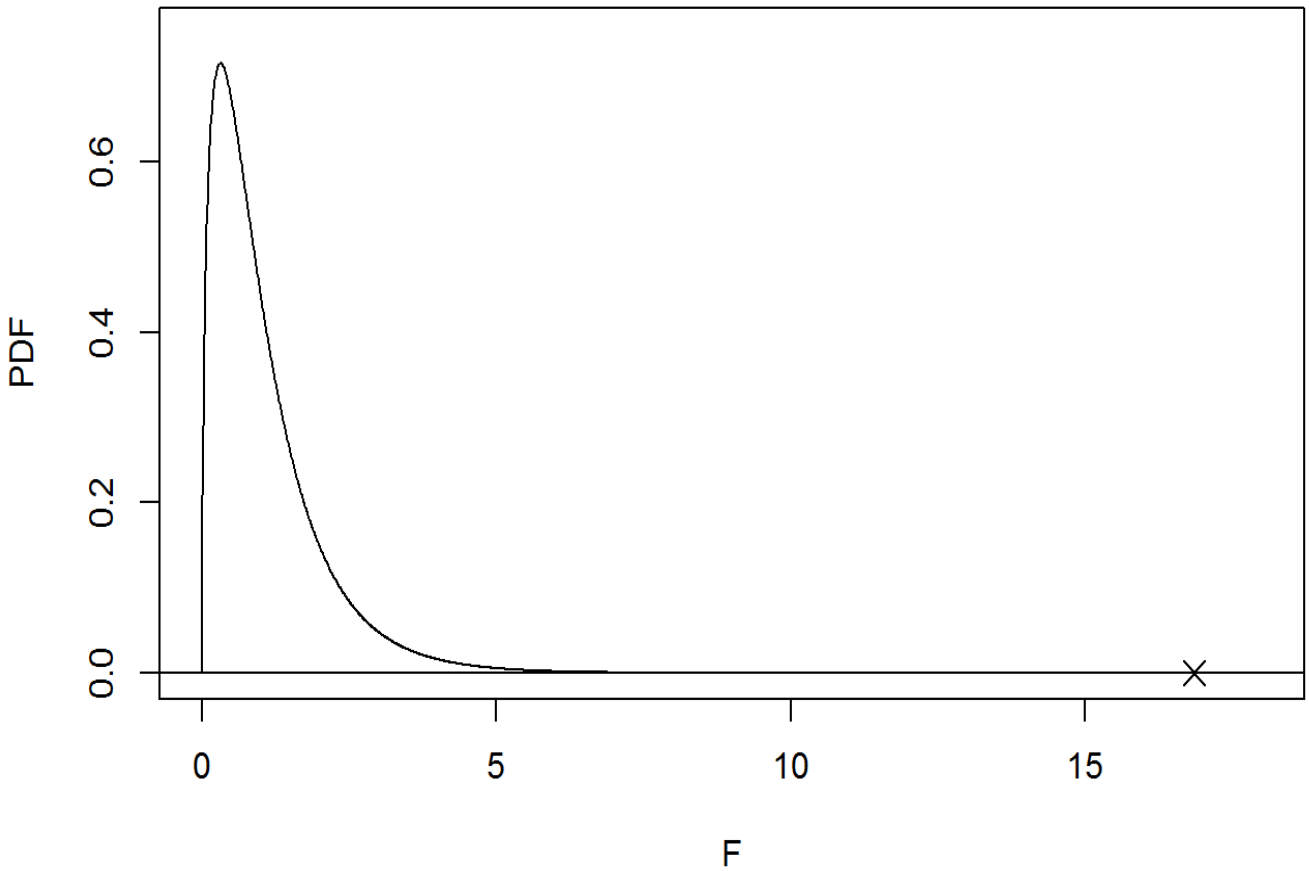

Figure 2: $F_{3,32}$ pdf. This is the sampling distribution of F for the mortar data when H_0 is true. An "$\times$" at $F = 16.848$ has been added.

Overall F-test

The form of the ANOVA one-way classification table is given as follows.

Source	df	SS	MS	F
Treatments	t-1	SS_{trt}	$MS_{trt} = \frac{SS_{trt}}{t-1}$	$F = \frac{MS_{trt}}{MS_{res}}$
Residuals	N-t	SS_{res}	$MS_{res} = \frac{SS_{res}}{N-t}$	
Total	N-1	SS_{total}		

- In general, it is easy to show that

$$\begin{aligned} SS_{total} &= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{++})^2 \\ &= \sum_{i=1}^{t} n_i(\bar{Y}_{i+} - \bar{Y}_{++})^2 + \sum_{i=1}^{t}\sum_{j=1}^{n_i}(\bar{Y}_{ij} - \bar{Y}_{i+})^2 \\ &= SS_{trt} + SS_{res} \end{aligned}$$

Overall F-test

- SS_{total} measures how observations vary about the overall mean, without regard to treatment groups; that is, SS_{total} measures the total variation in all the data
- SS_{total} can be partitioned into two components:
  - SS_{trt} measures how much of the toal variation is due to the treatment groups.
  - SS_{res} measures what is left over, which we attribute to inherent variation among individuals.
- The **probability value (p-value)** for a hypothesis test measures how much evidence we have against H_0. It is important to remember

the smaller the p-value $\Rightarrow$ the more evidence against H_0

The p-value is a probability. For the mortar data, the p-value can be interpreted as **If H_0 is true, the probability we should get a test statistic equal to or larger than $F = 16.848$ is 9.576×10^{-8}.**

Overall F-test

P-value Rules: Probability values are used in more general hypothesis test settings in statistics.

- Common values of α chosen beforehand are $\alpha = 0.1$ and $\alpha = 0.05$ (the most common).
- The smaller the α is chosen to be, the more evidence one requires to reject H_0.
- The value of α chosen by the experimenter determines how small the p-value must get before H_0 is ultimately rejected.
- For the mortar data, there is no ambiguity. For other situations, like p-value=0.06, the decision may not be as clear cut.

Assumptions/Robustness: There are three main assumptions when performing an analysis of variance:

Overall F-test

1. **Independent random samples.** This assumption holding is largely up to the experimenter/investigator, like drawing random samples from the different populations independently (in the case of an observational study) or using randomization to assign individuals to treatments (in an experiment).
2. **Normality.** Each of the t population distributions is normal. The one-way ANOVA analysis is robust to normality departure.
3. **Equal population variances**. This is the most important assumption.
    - A one-way ANOVA analysis is not robust to departures from this assumption, and it is very critical.
    - If you suspect the population variances may be markedly different, then you should not use a one-way ANOVA analysis.

Outline

Multiple Comparisons

In a one-way classification, the overall F test is used to test:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_t \text{ v.s. } H_1 : \mu_i \text{ are not all equal.}$$

If we reject H_0 in favor of H_1, we conclude that at least one population mean is different. The the follow-up analysis becomes determining which population means are different and how they are different. To do this, we will construct **Tukey pairwise confidence intervals** for all population treatment mean differences $\mu_i - \mu_{i'}, 1 \leq i < i' \leq t$. If there are t treatments, then there are

$$\binom{t}{2} = \frac{t(t-1)}{2}$$

pairwise confidence intervals to construct.

Multiple Comparisons

In the mortar strength study, there are $t = 4$ populations and therefore 6 pairwise comparisons.

$$\mu_1 - \mu_2 \quad \mu_1 - \mu_3 \quad \mu_1 - \mu_4 \quad \mu_2 - \mu_3 \quad \mu_2 - \mu_4 \quad \mu_3 - \mu_4$$

where

$$\begin{aligned}
\mu_1 &= \text{population mean strength for mortar type OCM} \\
\mu_2 &= \text{population mean strength for mortar type PIM} \\
\mu_3 &= \text{population mean strength for mortar type RM} \\
\mu_4 &= \text{population mean strength for mortar type PCM}
\end{aligned}$$

Problem: If we construct multiple confidence intervals(here 6 of them), and if we construct each one using $100(1-\alpha)$ percent confidence level, then the overall confidence level in the 6 intervals together will be less than $100(1-\alpha)$ percent.

Multiple Comparisons

A well-known inequality in probability called **Bonferroni's Inequality** which states that if we have events $A_1, A_2, \ldots, A_J$, the probability that each event occurs

$$P(\bigcap_{j=1}^{J} A_j) \geq \sum_{j=1}^{J} P(A_j) - (J-1)$$

To see how this inequality can be used in our current discussion, define the event

$$A_j = \{\text{jth confidence interval includes its population mean difference}\}$$

for $j = 1, 2, \ldots, J$. The event

$$\bigcap_{j=1}^{J} A_j = \{\text{each of the J intervals includes its population mean difference}\}$$

Multiple Comparisons

In this light, consider the following table, which contains a lower bound on how small this probability can be (for different values of t and J). This table assumes that each pairwise interval has been constructed at the nominal $1 - \alpha = 0.95$ level.

# of treatment (t)	# of intervals $J = \binom{t}{2}$	Lower bound
3	3	$3(0.95) - 2 = 0.85$
4	6	$6(0.95) - 5 = 0.70$
5	10	$10(0.95) - 9 = 0.50$
$\vdots$	$\vdots$	$\vdots$
10	45	$45(0.95) - 44 = -1.25$!!

For $t = 4$ treatments(populations), the probability that each of the 6 95 confidence intervals will contain its population mean difference can be as low as 0.7! For larger experiments with more treatments, this probability is even lower!!

Multiple Comparisons

Goal: Construct confidence intervals for all pairwise intervals $\mu_i - \mu_{i'}, 1 \leq i < i' \leq t$, and have our **family-wise confidence level** still be at $100(1-\alpha)$ percent. By "family-wise", we mean that out level of confidence applies to the collection of all $\binom{t}{2}$ intervals (not to the intervals individually).

Solution: Increase the confidence level associated with each individual interval. **Tukey's method** is designed to do this. The intervals are of the form:

$$(\bar{Y}_{i+} - \bar{Y}_{i'+}) \pm q_{t,N-t,\alpha} \sqrt{MS_{res}(\frac{1}{n_i} + \frac{1}{n_{i'}})}$$

where $q_{t,N-t,\alpha}$ is the **Tukey quantile** that guarantees a **family-wise confidence level** of $100(1-\alpha)$ percent.

Multiple Comparisons

Mortar data: We use R to construct the Tukey confidence intervals. The family-wise confidence level is 95 percent.

```
> TukeyHSD(aov(lm(strength ~ mortar.type)),conf.level=0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lm(strength ~ mortar.type))

$mortar.type
             diff       lwr       upr     p adj
PCM-OCM   2.48000 -4.950955  9.910955 0.8026758
PIM-OCM  15.21575  8.166127 22.265373 0.0000097
RM-OCM   12.99875  5.949127 20.048373 0.0001138
PIM-PCM  12.73575  5.686127 19.785373 0.0001522
RM-PCM   10.51875  3.469127 17.568373 0.0016850
RM-PIM   -2.21700 -8.863448  4.429448 0.8029266
```

Multiple Comparisons

In the R output, the columns labeled lwr and upr give, respectively, the lower and upper limits of the pairwise confidence intervals.

- PCM-OCM: We are (at least)95 percent confident that the difference in the population mean strengths for the PCM and OCM mortars is between -4.95 and 9.91 MPa.
- PIM-OCM: We are (at least)95 percent confident that the difference in the population mean strengths for the PIM and OCM mortars is between 8.17 and 22.27 MPa.
- Interpretations for the remaining 4 confidence intervals are written similarly.
- If a pairwise confidence interval (for two population means) includes 0, then these population means are not declared to be different; otherwise, the population means are declared to be different.
- Had we not used an adjusted analysis based on Tukey's method, our overall confidence level would have been much lower.

Multiple Comparisons