Chapter 12

One-Way Analysis of Variance

Introduction to the Practice of

STATISTICS EIGHTH EDITION

Moore / McCabe / Craig

Lecture Presentation Slides

Chapter 12
One-Way Analysis of Variance

12.1 Inference for One-Way Analysis of Variance

12.2 Comparing the Means

12.1 Inference for One-Way Analysis of Variance

- The idea of ANOVA
- Comparing several means
- The problem of multiple comparisons
- The ANOVA *F* test

Introduction

The two-sample *t* procedures of Chapter 7 compared the means of two populations or the mean responses to two treatments in an experiment.

In this chapter, we'll compare *any number* of means using **analysis of variance.**

> **Note:** We are comparing ***means,*** even though the procedure is called analysis of *variance.*

The Idea of ANOVA

When comparing different populations or treatments, the data are subject to sampling variability. We can pose the question for inference in terms of the *mean* response.

Analysis of variance (ANOVA) is the technique used to compare several means.

One-way ANOVA is used for situations in which there is only one factor, or only one way to classify the populations of interest.

Two-way ANOVA is used to analyze the effect of two factors.

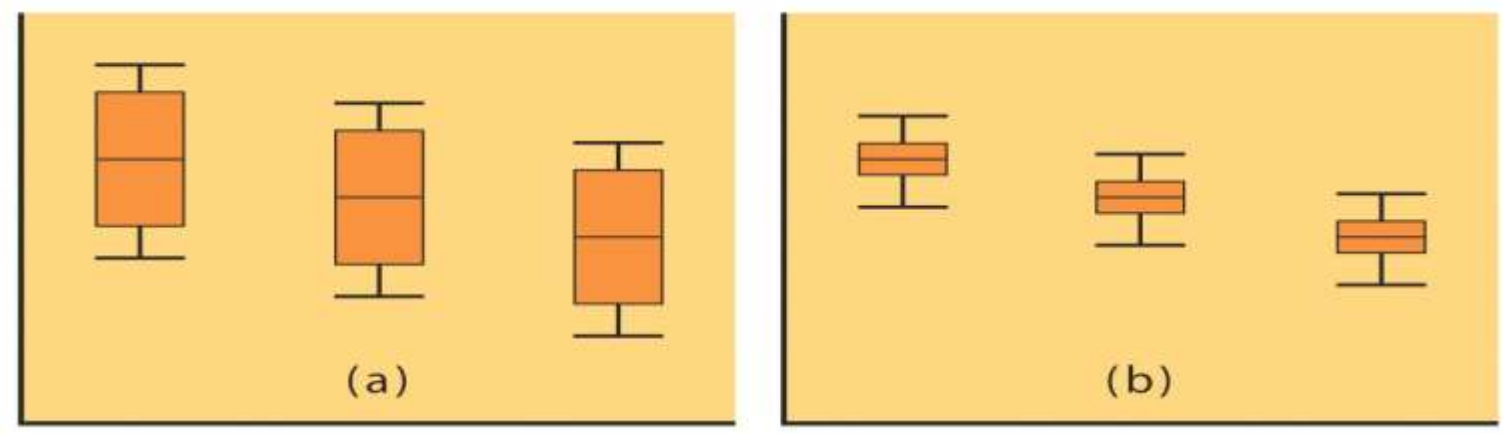
The Idea of ANOVA

The details of ANOVA are a bit daunting. The main idea is that, when we ask if a set of means gives evidence for differences among the population means, what matters is not how far apart the sample means are, but how far apart they are *relative to the variability of individual observations.*

The Analysis of Variance Idea

Analysis of variance compares the variation due to specific sources with the variation among individuals who should be similar. In particular, ANOVA tests whether several populations have the same means by comparing *how far apart the sample means are* with *how much variation there is within a sample*.

The Idea of ANOVA



- ✓ The sample means for the three samples are the same for each set.
- ✓ The variation *among sample means* for (a) is identical to that for (b).
- ✓ The variation *among the individuals within* each of the three samples is much less for (b).

CONCLUSION: The samples in (b) contain a larger amount of variation among the sample means *relative to* the amount of variation within the samples, so ANOVA will find *more significant differences among the means in (b)*

- assuming equal sample sizes here for (a) and (b).
- **Note:** Larger samples will find more significant differences.

Comparing Several Means

TABLE 22.1 Highway gas mileage for 2003 model vehicles

Midsize Cars		Sport-Utility Vehicles		Standard Pickup Trucks	
Model	MPG	Model	MPG	Model	MPG
Acura 3.5TL	29	Cadillac Escalade	18	Chevrolet Silverado	20
Audi A6	27	Chevrolet Avalanche	18	Dodge Dakota	19
Audi A8	25	Chevrolet Blazer	23	Dodge Ram	20
Buick Century	29	Chevrolet Suburban	18	Ford Explorer	20
Buick Regal	27	Chevrolet Tahoe	18	Ford F150	20
Cadillac CTS	26	Chevrolet Tracker	26	Ford Ranger	26
Cadillac Seville	27	Chevrolet Trailblazer	22	GMC Sierra	20
Chevrolet Monte Carlo	32	Chrysler PT Cruiser	25	Lincoln Blackwood	17
Chrysler Sebring	30	Dodge Durango	19	Mazda B2300	26
Honda Accord	33	Ford Escape	28	Mazda B3000	22
Hyundai Sonata	30	Ford Expedition	19	Mazda B4000	21
Hyundai XG350	26	Ford Explorer	19	Nissan Frontier	23
Infiniti I35	26	Honda CR-V	28	Toyota Tacoma	25
Jaguar S-Type	26	Hyundai Santa Fe	27	Toyota Tundra	19
Jaguar Vanden Plas	24	Infiniti QX4	21		
Kia Optima	30	Isuzu Axiom	21		
Lexus ES300	29	Isuzu Rodeo	23		
Lexus GS300	25	Jeep Grand Cherokee	21		
Lincoln LS	26	Jeep Liberty	24		
Mercedes-Benz E320	27	Kia Sorento	20		
Mercedes-Benz E500	23	Lincoln Navigator	17		
Mitsubishi Diamante	25	Mazda Tribute	28		
Mitsubishi Galant	27	Mitsubishi Montero	22		
Nissan Altima	29	Nissan Pathfinder	21		
Nissan Maxima	26	Nissan Xterra	24		
Saab 9-5	29	Saturn Vue	28		
Saturn L200	32	Suzuki Vitara	25		
Saturn L300	29	Toyota 4Runner	20		
Toyota Camry	32	Toyota Highlander	27		
Volkswagen Passat	31	Toyota Rav4	29		
Volvo S80	28	Volvo XC 90	24		

Do SUVs and trucks have lower gas mileage than midsize cars?

Response variable: gas mileage (mpg)

Groups: vehicle classification

- 31 midsize cars
- 31 SUVs
- 14 standard-size pickup trucks

Data from the Environmental Protection Agency's *Model Year 2003 Fuel Economy Guide*, www.fueleconomy.gov.

Comparing Several Means

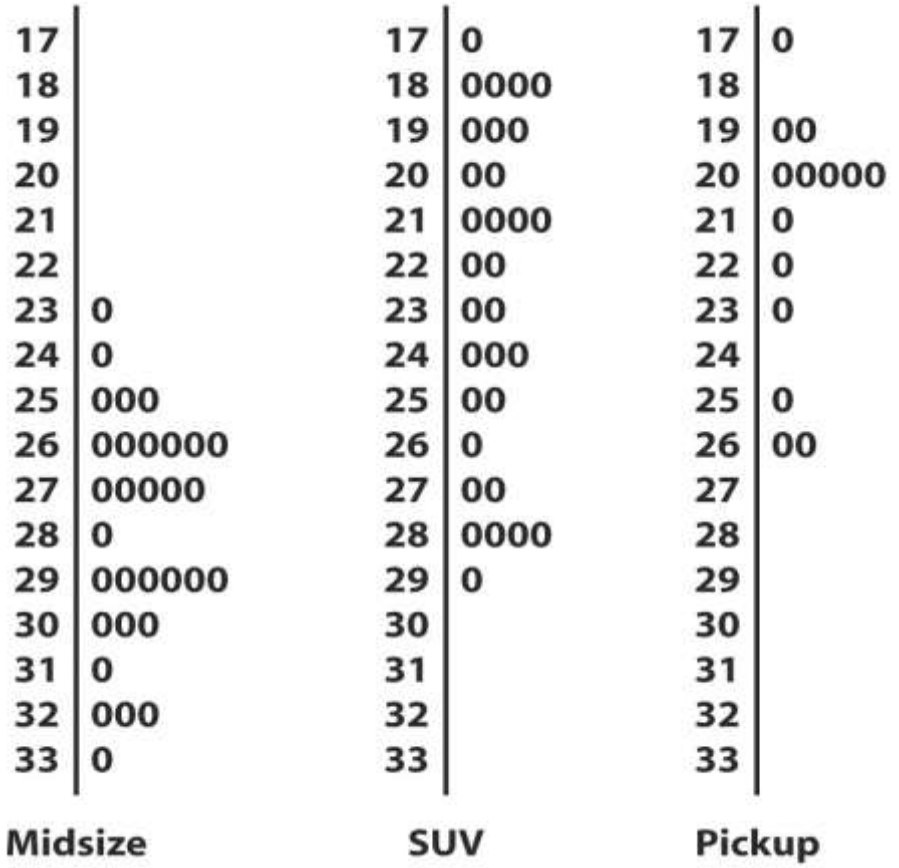
Means:

Midsize: 27.903

SUV: 22.677

Pickup: 21.286

- ✓ Mean gas mileage for SUVs and pickups appears less than for midsize cars.
- ✓ Are these differences statistically significant?



Midsize		SUV		Pickup	
17		17	0	17	0
18		18	0000	18	
19		19	000	19	00
20		20	00	20	00000
21		21	0000	21	0
22		22	00	22	0
23	0	23	00	23	0
24	0	24	000	24	
25	000	25	00	25	0
26	000000	26	0	26	00
27	00000	27	00	27	
28	0	28	0000	28	
29	000000	29	0	29	
30	000	30		30	
31	0	31		31	
32	000	32		32	
33	0	33		33	

Comparing Several Means

Means:

Midsize:	27.903
SUV:	22.677
Pickup:	21.286

Null hypothesis:
The true means (for gas mileage) are the same for all groups (the three vehicle classifications).

We could look at separate *t* tests to compare each pair of means to see if they are different:

27.903 vs. 22.677,	27.903 vs. 21.286, and	22.677 vs. 21.286
$H_0: \mu_1 = \mu_2$	$H_0: \mu_1 = \mu_3$	$H_0: \mu_2 = \mu_3$

However, this gives rise to the problem of **multiple comparisons**!

Problem of Multiple Comparisons

We have the problem of how to do many comparisons at the same time with some overall measure of confidence in all the conclusions.

Statistical methods for dealing with this problem usually have two steps:

- An **overall test** to find any differences among the parameters we want to compare
- A detailed **follow-up analysis** to decide which groups differ and how large the differences are

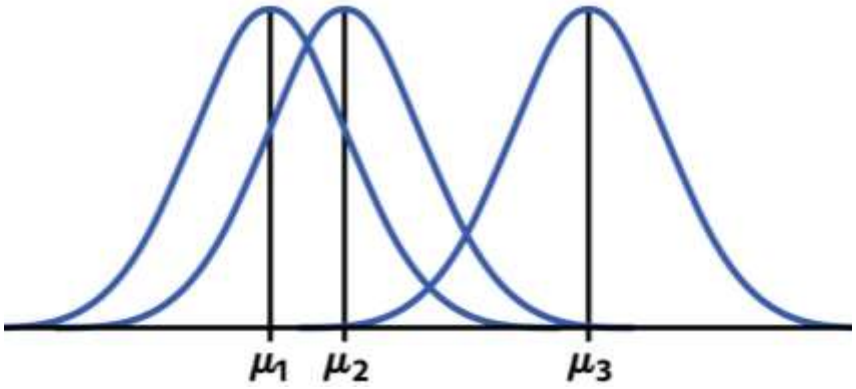
Follow-up analyses can be quite complex. For now, we will look only at the overall test for a difference in several means and examine the data to make follow-up conclusions.

The One-Way ANOVA Model

Random sampling always produces chance variations. Any "factor effect" would thus show up in our data as the factor-driven differences plus chance variations ("error"):

Data = factor effect + error

The one-way ANOVA model analyzes data x_{ij} where chance variations are normally distributed $N(0,\sigma)$:



$$x_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

for $I = 1,\ldots,I$ and $j = 1,\ldots,n_i$. The ε_{ij} are assumed to be from a $N(0,\sigma)$ distribution. The **parameters of the model** are the population means $\mu_1,\mu_2,\ldots,\mu_I$ and the common standard deviation σ.

μ_i can be broken into the overall mean μ and the effect of level i of the factor

The ANOVA *F* Test

We want to test the null hypothesis that there are *no differences* among the means of the populations.

The basic conditions for inference are that we have a random sample from each population and that each population is Normally distributed.

The alternative hypothesis is that there is *some difference*. That is, not all means are equal. This hypothesis is not one-sided or two-sided. It is "many-sided."

This test is called the **analysis of variance *F* test (ANOVA).**

Conditions for ANOVA

Like all inference procedures, ANOVA is valid only in some circumstances. The conditions under which we can use ANOVA are:

Conditions for ANOVA Inference

- We have ***I* independent SRSs**, one from each population. We measure the same response variable for each sample.
- The *i*th population has a **Normal distribution** with unknown mean μ_i.
- All the populations have the **same standard deviation** σ, whose value is unknown.

Checking Standard Deviations in ANOVA

- The results of the ANOVA *F* test are approximately correct when the largest sample standard deviation is no more than twice as large as the smallest sample standard deviation.

The ANOVA *F* Statistic

To determine statistical significance, we need a test statistic:

The ANOVA *F* Statistic

The **analysis of variance *F* statistic** for testing the equality of several means has this form:

$$F = \frac{\textit{variation among the sample means}}{\textit{variation among individuals in the same sample}}$$

- *F* is always zero or positive
  - *F* is zero only when all sample means are the same.
  - *F* gets larger as means move further apart.
- Large values of *F* are evidence against H_0*: equal means*
- The *F* test is upper-one-sided.

F Distributions

The *F* distributions are a family of right-skewed distributions that take only values greater than 0. A specific *F* distribution is determined by the degrees of freedom of the numerator and denominator of the *F* statistic.

When describing an *F* distribution, always give the numerator degrees of freedom first. Our brief notation will be *F*(df1, df2) with df1 degrees of freedom in the numerator and df2 degrees of freedom in the denominator.

Degrees of Freedom for the *F* Test

We want to compare the means of I populations. We have an SRS of size n_i from the i^{th} population, so that the total number of observations in all samples combined is:

$$N = n_1 + n_2 + \cdots + n_I$$

If the null hypothesis that all population means are equal is true, the ANOVA *F* statistic has the *F* distribution with $I - 1$ degrees of freedom in the numerator and $N - I$ degrees of freedom in the denominator.

The ANOVA *F* Test

$$F = \frac{\textit{variation among the sample means}}{\textit{variation among individuals in the same sample}}$$

The measures of variation in the numerator and denominator are ***mean squares:***

- Numerator: **Mean Square for Groups** (MSG)

$$MSG = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \cdots + n_I(\bar{x}_I - \bar{x})^2}{I - 1}$$

- Denominator: **Mean Square for Error** (MSE)

$$MSE = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_I - 1)s_I^2}{N - I}$$

  - MSE is also called the **pooled sample variance**, written as s_p^2 (s_p is the ***pooled standard deviation***).
  - s_p^2 estimates the common variance σ^2.

The ANOVA Table

Source of variation	Sum of squares SS	DF	Mean square MS	*F*	*P*-value	*F* crit
Among or between "groups"	$\sum_i n_i(\bar{x}_i - \bar{x})^2$	I -1	MSG= SSG/DFG	MSG/MSE	Tail area above F	Value of F for α
Within groups or "error"	$\sum_i (n_i - 1)s_i^2$	$N - I$	MSE= SSE/DFE			
Total	SST=SSG+SSE $\sum_i (x_i - \bar{x})^2$	$N - 1$				

R^2 = SSG/SST — **Coefficient of determination**

$\sqrt{\text{MSE}} = s_p$ — **Pooled standard deviation**

The sums of squares represent different sources of variation in the data:

SST=SSG + SSE.

The degrees of freedom mirror the sums of squares:

DFT = DFG + DFE.

Data ("Total") = Factor effect ("Groups") + Error ("Error")

Example

One-way ANOVA: Midsize, SUV, Pickup

Analysis of Variance

Source	DF	SS	MS	F	P
Factor	2	606.45	303.22	31.61	0.000
Error	73	700.34	9.59		
Total	75	1306.79			

***P*-value < .05 significant differences**

There is significant evidence that the three types of vehicle do not all have the same gas mileage.

However, this does not tell us which groups have different means.

12.2 Comparing the Means

- Contrasts
- Multiple comparisons
- Power of the *F* test*

Introduction

You have calculated a *P*-value for your ANOVA test. Now what?

If you found a significant result, you still need to determine which treatments are different from which.

- You can gain insight by looking at plots, such as side by side boxplots.
- There are several tests of statistical significance designed specifically for multiple tests. You can choose **contrasts,** or **multiple comparisons.**
- You can find a confidence interval for each mean μ_i that is shown to be significantly different from the others.

Introduction

- **Contrasts** can be used only when there are clear expectations BEFORE starting an experiment, and these are reflected in the experimental design. Contrasts are **planned comparisons.**
  - Patients are given either drug A, drug B, or a placebo. The placebo is meant to provide a baseline against which the other drugs can be compared.
  - ***It is NOT appropriate to use a contrast test when that test is suggested only AFTER the data are collected.***
- **Multiple comparisons** should be used when the pattern of differences between means is unknown beforehand. Such comparisons are **pair-wise tests** of significance.
  - We compare gas mileage for eight brands of SUVs. We have no prior knowledge to expect any brand to perform differently from the rest. Pair-wise comparisons should be performed here, but only if an ANOVA test on all eight brands was statistically significant.

Contrasts

When an experiment is designed to test a specific hypothesis that some treatments are different from other treatments, we can use a contrast to test for significant differences between these specific treatments.

- Contrasts are more powerful than multiple comparisons because they are more specific. They are better able to pick up a significant difference.
- You can use a *t*-test on the contrast or calculate a *t* confidence interval.
- Since the hypothesis is pre-specified, the corresponding *t*-test is typically done in lieu of the multiple sample ANOVA test.

Contrasts

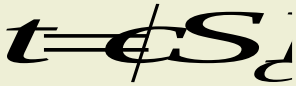
A contrast is a combination of population means of the form:

where the coefficients a_i have sum 0.

The corresponding sample contrast is:

The standard error of c is:

To test the null hypothesis H_0: $\psi = 0$ use the t statistic:



with degrees of freedom **DFE** that is associated with s_p. The alternative hypothesis can be one- or two-sided.

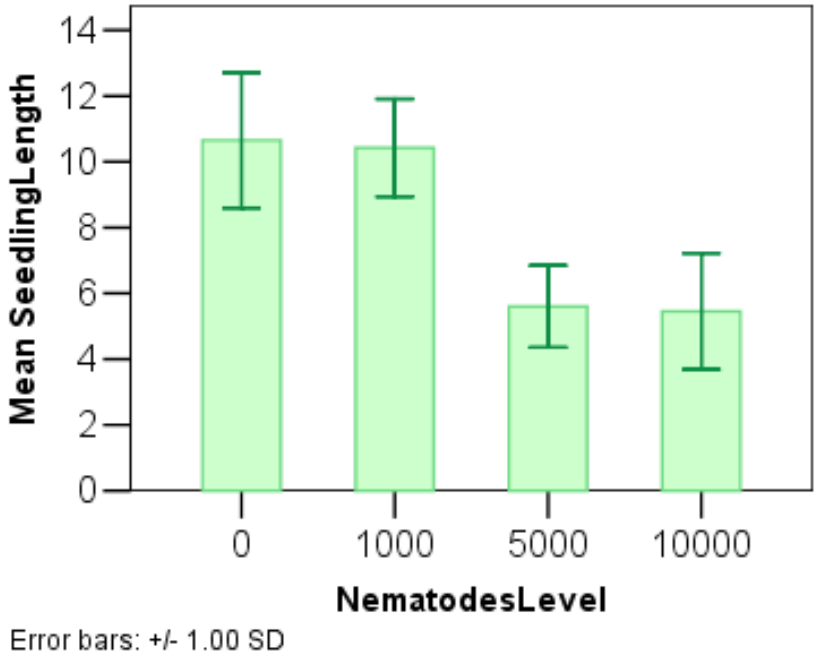
A level C confidence interval for the difference ψ is:

where t^* is the critical value defining the middle C% of the t distribution with **DFE** degrees of freedom.

Example

Do nematodes affect plant growth? A botanist prepares 16 identical planting pots and adds different numbers of nematodes into the pots. Seedling growth (in mm) is recorded two weeks later.

Nematodes	Seedling growth				$\bar{x}_i$
0	10.8	9.1	13.5	9.2	10.65
1,000	11.1	11.1	8.2	11.3	10.43
5,000	5.4	4.6	7.4	5	5.6
10,000	5.8	5.3	3.2	7.5	5.45
overall mean 8.03					



One group contains no nematodes at all. If the botanist planned this group as a baseline/control, then a contrast of all the nematode groups against the control would be valid.

Example

ANOVA: H_0: all μ_i are equal vs. H_a: not all μ_i are equal

ANOVA

SeedlingLength

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	100.647	3	33.549	12.080	.001
Within Groups	33.328	12	2.777		
Total	133.974	15			

➔ not all μ_i are equal

Planned comparison:

H_0: $\mu_1 = 1/3\ (\mu_2 + \mu_3 + \mu_4)$ vs.

H_a: $\mu_1 > 1/3\ (\mu_2 + \mu_3 + \mu_4)$ → *one-tailed*

Contrast coefficients: (+3 −1 −1 −1)

Contrast Coefficients

	NematodesLevel			
Contrast	0	1000	5000	10000
1	-3	1	1	1

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
SeedlingLength	Assume equal variances	1	-10.4750	2.88650	-3.629	12	.003
	Does not assume equa	1	-10.4750	3.34823	-3.129	4.139	.034

Nematodes result in significantly shorter seedlings (α = .01).

Planned vs. Post-Hoc Tests

- *Planned tests* are tests formulated prior to the data collection. These are contrasts that are of interest to the researchers on top of the overall *F*-test.
- *Post-hoc tests* are tests formulated after ANOVA has been performed.
- More specifically, after a significant *F*-test has been obtained, different contrasts that are of interest to the researchers may be formulated. In most cases, contrasts in the form of pairwise comparisons are formulated.
- Some people proceed with testing the different contrasts without considering when the contrasts were formulated – *before* or *after* the data collection – or how many contrasts involved.

Planned vs. Post-Hoc Tests

- This approach inflates the overall Type I error (i.e., false positive rate) because it only controls for what is called the *testwise* Type I error rate.
- This means that the test for each contrast will have an error rate of α.
- However, the error rate for all the different tests taken together (known as *familywise* Type I error rate) is not known, but can be assumed to be higher than α
- To keep the familywise Type I error at most at α, some adjustments need to be adopted to make the individual tests more conservative.

The Bonferroni Procedure

The **Bonferroni procedure** is an example of a procedure *that adjusts for simultaneously performing many tests at the same time*.

The Bonferroni procedure can be done in either of two equivalent ways. In either approach, one chooses an overall significance level α and then does a number of pair-wise t-tests. Let k be the total number of t-tests performed.

- Use the procedure on the previous page with t^{**} being the t-value with degrees of freedom DFE and an area of $\alpha/(2k)$ to its right.
- Compute a P-value for each t-test in the usual way. Conclude that a particular pair of means is significantly different only when ***k* times that *P*-value** is no larger than α.

The Bonferroni Procedure

- ✓ Generally speaking, when one does multiple comparisons **the chance of committing at least one type I error increases with the number of tests done.**
- ✓ To compensate for this tendency, the Bonferroni procedure **lowers** the working significance level of each test to control the probability of making at least one type I error among all tests performed.
- ✓ As a consequence, the more pair-wise comparisons you perform, the more difficult it will be to show statistical significance for each test.
- ✓ The Bonferroni procedure is typically used with planned contrasts.

Scheffé's Method

- Scheffé's method keeps the Type I error at α for all possible (post-hoc) contrasts.
- Consequently, sleuthing or fishing for significant results is allowed.
- That is, we can try as many contrasts as we want after getting a significant *F*-test.
- However, there's no guarantee that the significant results for the contrasts would be particularly interesting.
- Scheffé's method requires that a different distribution by used.
- The details are beyond the coverage of this class, but Scheffé's method adjust the critical t^{**} as a function of the F distribution $F(I - 1, N - I)$ and the number of groups I

Scheffé's Method

- Scheffé's method is the most conservative method that can be used without getting more conservative than necessary.
- A related method for post-hoc tests is Tukey's method.
  - This gives a Type I experimentwise error rate of α for all possible *pairwise* comparisons, and
  - Is less conservative than Scheffé's method.
- As long as we adjust for (1) the number of contrasts or (2) when the contrasts are formulated, the procedure will be more conservative (e.g., more likely to retain the null) than if we do not make any adjustments at all.

Chapter 12
One-Way Analysis of Variance

12.1 Inference for One-Way Analysis of Variance

12.2 Comparing the Means