Ch 14 ::: ANOVA

Analysis of Variance

Think of this . . .

A study compared the size of ants from different colonies. Researchers measured the masses (in milligrams) of random samples of ants from three different colonies. Do the population mean sizes differ among the three ant colonies?

You have a candy that you could produce in blue, red, or green color. You sell it in three different test markets – blue in one, and so on. After a month, you get the resulting sales in each market. Assuming everything else being equal, does it make a difference which color you use?

A farmer treats four parts of his farm with four different types of pesticides. Are the pesticides any different?

The number of times each car went through the carwash before its wax deteriorated

Observation	Wax Type 1	Wax Type 2	Wax Type 3
1	27	33	29
2	30	28	28
3	29	31	30
4	28	30	32
5	31	30	31
Sample Mean	29.0	30.4	30.0

How does ANOVA work?

Sep 1st, 2016 CNN.COM

*(CNN)*An experimental drug shattered and removed toxic plaques in the brains of patients with early-stage Alzheimer's disease, researchers said Thursday.

Given to patients once a month for a year, infusions of the drug aducanumab cleared the brain of the deposits, which experts believe play a crucial role in disrupting cellular processes and blocking communication among nerve cells.

Although most aging brains contain some plaques, the brains of Alzheimer's patients tend to have much more. The disease, the most common form of dementia, has no cure, although some treatments are available to alleviate symptoms. Treatments to slow the progression or reverse it have not panned out.
Cambridge, Massachusetts-based Biogen developed the drug aducanumab and funded the study, which primarily tested its safety in humans and was not designed to test for cerebral benefits for patients. Still, the condition of some patients who received the drug showed less decline than patients receiving a placebo.

The study, funded by the makers of aducanumab, split 165 participants into groups and treated them with monthly intravenous infusions of either aducanumab or a placebo over 54 weeks. Four groups of patients received the drug in four separate doses.

As measured by PET brain scans, treatment with aducanumab reduced brain plaques based on both duration and dose; all groups showed more reduction in plaques over time, and the highest-dose group showed the greatest reduction of all.

The study has too few patients to prove that the drug actually works, wrote Eric M. Reiman, executive director of the Banner Alzheimer's Institute in Phoenix, in a commentary on the research, published in the journal Nature. He added that many other Alzheimer's drugs have looked promising early on but ended in failure.

However, "confirmation of a cognitive benefit would be a game-changer," said Reiman, a psychiatrist who is unaffiliated with the current study.

Of all the patients, 125 completed the treatment. Among the 40 who discontinued it, most withdrew due to negative side effects, which included fluid building up in the brain as a result of the removal of plaques. In some cases, this can cause brain bleeds.

Larger trials of the drug involving Alzheimer's patients are in progress and planned to run until at least 2020.

In their conclusion, the researchers observed that it may have taken up to 20 years for the plaques to have accumulated to the levels seen in patients, so the removal of it within a 12-month period "appears encouraging."

Example: Best way to advertise

A company is just not sure if the means to advertise (newspaper, radio, TV, mail) makes any difference. So, in 4 test markets, it uses the 4 strategies and collects 17 weeks worth of data. Use ANOVA to find out if the choice of advertising medium makes any difference.

Dataset: Anova Example

Analysis of variance

The null hypothesis looks like this:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

i.e. there are no differences between population means (they all are equal).

Our alternative hypothesis becomes:

H_a: at least two means differ

We use F-distribution to test the hypothesis.

Requirements for Performing Analysis of Variance

- 1. Each of the k populations is normally distributed.
- 2. The variances (σ^2) of the populations are all approximately equal. That is, smallest sample standard deviation must be less than twice the largest standard deviation.
- 3. The samples are independently drawn.

Analysis of Variance for the Equality of k Population Means: p-Value Method continued

- F_{data} follows *F* distribution with $df_1 = k - 1$ and $df_2 = n_t - k$ if required conditions are satisfied
- n_t represents total sample size

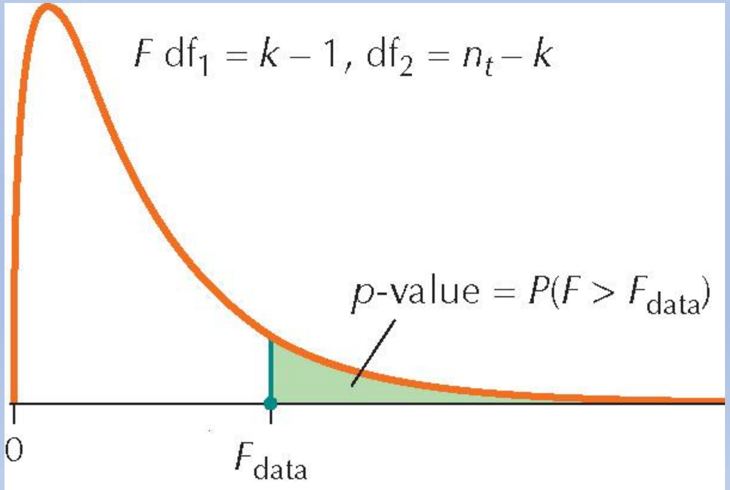


FIGURE 12.8 *p*-Value for the ANOVA *F* test.

Properties of the *F* Curve

1. Total area under the *F* curve equals 1.

2. F random variable is never negative; F curve starts at 0, extends indefinitely to the right, and approaches but never meets the horizontal axis.

3. *F* curve is right-skewed.

4. Different *F* curve for each different pair df_1 and df_2.

Mean Square Treatment (MSTR)

- Measures the variability in the sample means. Also called MST.

Mean Square Error (MSE)

- Measures the variability within the samples

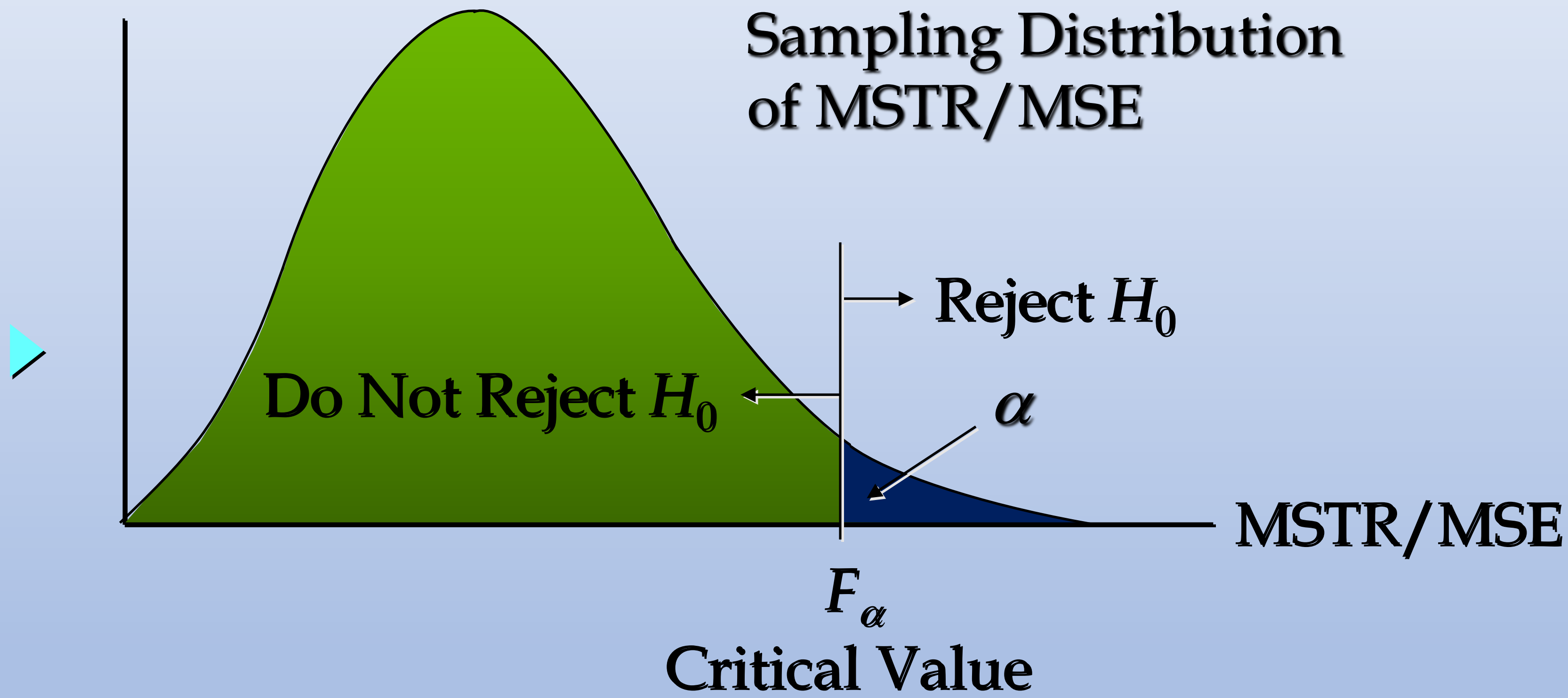
Test Statistic for Analysis of Variance

$$F_{data} = \frac{MSTR}{MSE}$$

- F_{data} measures the variability among the sample means, compared to the variability within the samples.
- F_{data} follows an *F* distribution with $df_1 = k - 1$ and $df_2 = n_t - k$.
- We compare F_{data} F-critical and decide whether to reject the hypothesis or not. We can also use p-value to draw the same conclusion.
- In our class, we will use Excel (data analysis) which will give us the required p-value, and F_{data}.

Comparing the Variance Estimates: The *F* Test

- Sampling Distribution of MSTR/MSE



Example: Investment & age

In the last decade stockbrokers have drastically changed the way they do business. It is now easier and cheaper to invest in the stock market than ever before.

What are the effects of these changes?

To help answer this question a financial analyst randomly sampled 366 American households and asked each to report the age of the head of the household and the proportion of their financial assets that are invested in the stock market.

Excel File 1 (Anova Investment)

Example

The age categories are

Young (Under 35)

Early middle-age (35 to 49)

Late middle-age (50 to 65)

Senior (Over 65)

The analyst was particularly interested in determining whether the ownership of stocks varied by age.

Do these data allow the analyst to determine that there are differences in stock ownership between the four age groups?

Example

Terminology

Percentage of total assets invested in the stock market is the response variable; the actual **percentages** are the responses in this example.

Population classification criterion is called a ***factor***.

The **age category** is the factor we're interested in. This is the ***only*** factor under consideration (hence the term "one way" analysis of variance).

Each population is a ***factor level***.

In this example, there are four factor levels: **Young, Early middle age, Late middle age,** and **Senior**.

Example

The null hypothesis in this case is:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

i.e. there are no differences between population means.

Our alternative hypothesis becomes:

H_1: at least two means differ

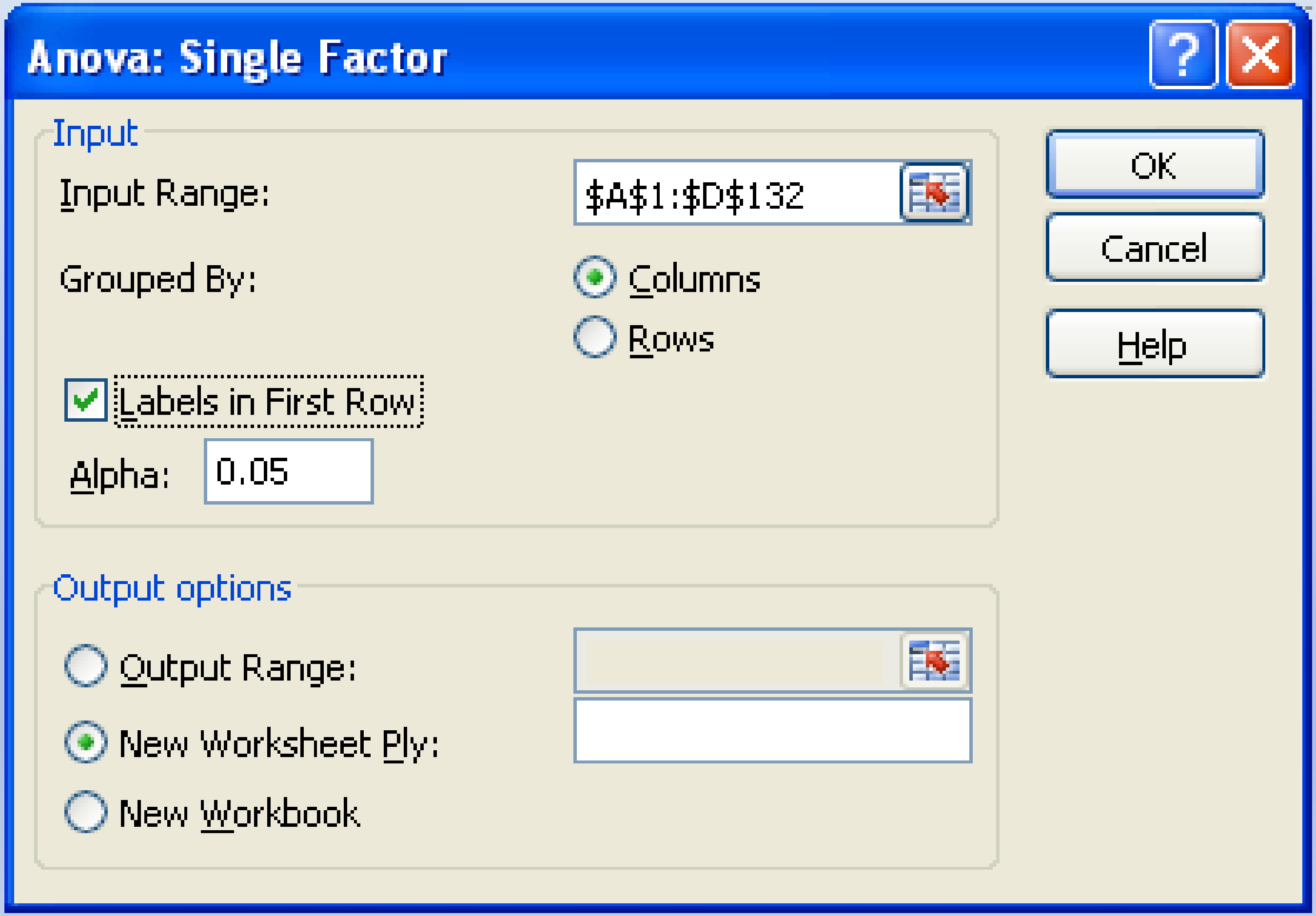
OK. Now we need some test statistics…

COMPUTE

Example

Using Excel:

Click Data, Data Analysis, Anova: Single Factor



COMPUTE

Example

	A	B	C	D	E	F	G
1	Anova: Single Factor						
2							
3	SUMMARY						
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5	Young	84	3729.5	44.40	386.55		
6	Early Middle Age	131	6873.9	52.47	469.44		
7	Late Middle Age	93	4755.9	51.14	471.82		
8	Senior	58	3006.6	51.84	444.79		
9							
10							
11	ANOVA						
12	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
13	Between Groups	3741.4	3	1247.12	2.79	0.0405	2.6296
14	Within Groups	161871.0	362	447.16			
15							
16	Total	165612.3	365				

Table 12.4 ANOVA table

- Convenient way to display the various statistics
- Quantities in the "Mean square" column equal the ratio of the two columns to its left

Source of variation	Sum of squares	Degrees of freedom	Mean square	F-test statistic
Treatment	SSTR	$df_1 = k - 1$	$MSTR = \frac{SSTR}{k-1}$	$F_{data} = \frac{MSTR}{MSE}$
Error	SSE	$df_2 = n_t - k$	$MSE = \frac{SSE}{n_t - k}$	
Total	SST			

Example: Best way to advertise

A company is just not sure if the means to advertise (newspaper, radio, TV, mail) makes any difference. So, in 4 test markets, it uses the 4 strategies and collects 17 weeks worth of data. Use ANOVA to find out if the choice of advertising medium makes any difference.

Dataset: Anova Example

Randomized Block Analysis of Variance

The purpose of designing a randomized block experiment is to ***reduce*** the ***within-treatments variation*** to more easily detect ***differences between the treatment means***.

Randomized Block Design

- Example: Crescent Oil Co.

  Five automobiles have been tested using each of the three gasoline blends and the miles per gallon ratings are shown on the next slide.

  Factor . . . Gasoline blend

  Treatments . . . Blend X, Blend Y, Blend Z

  Blocks . . . Automobiles

  Response variable . . . Miles per gallon

Randomized Block Design

Automobile (Block)	Type of Gasoline (Treatment)			Block Means
	Blend X	Blend Y	Blend Z	
1	31	30	30	30.333
2	30	29	29	29.333
3	29	29	28	28.667
4	33	31	29	31.000
5	26	25	26	25.667
Treatment Means	29.8	28.8	28.4	

Example

Many North Americans suffer from high levels of cholesterol, which can lead to heart attacks. For those with very high levels (over 280), doctors prescribe drugs to reduce cholesterol levels. A pharmaceutical company has recently developed four such drugs. To determine whether any differences exist in their benefits, an experiment was organized.

The company selected 25 groups of four men, each of whom had cholesterol levels in excess of 280. In each group, the men were matched according to age and weight. The drugs were administered over a 2-month period, and the reduction in cholesterol was recorded. Do these results allow the company to conclude that differences exist between the four new drugs?

Excel File 2

IDENTIFY

Example

The hypotheses to test in this case are:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1: At least two means differ

H_0: All the groups are the same

H_1: At least two groups are different

Example

IDENTIFY

Each of the four drugs can be considered a ***treatment***.

Each group) can be ***blocked***, because they are matched by age and weight.

By setting up the experiment this way, we eliminates the ***variability*** in cholesterol reduction ***related*** to different combinations of age and weight. This helps ***detect differences*** in the mean cholesterol reduction ***attributed to the different*** drugs.

The Data

Example

Group	Drug 1	Drug 2	Drug 3	Drug 4
1	6.6	12.6	2.7	8.7
2	7.1	3.5	2.4	9.3
3	7.5	4.4	6.5	10.0
4	9.9	7.5	16.2	12.6
Block 5	13.8	6.4	8.3	10.6
6	13.9	13.5	5.4	15.4

Treatment

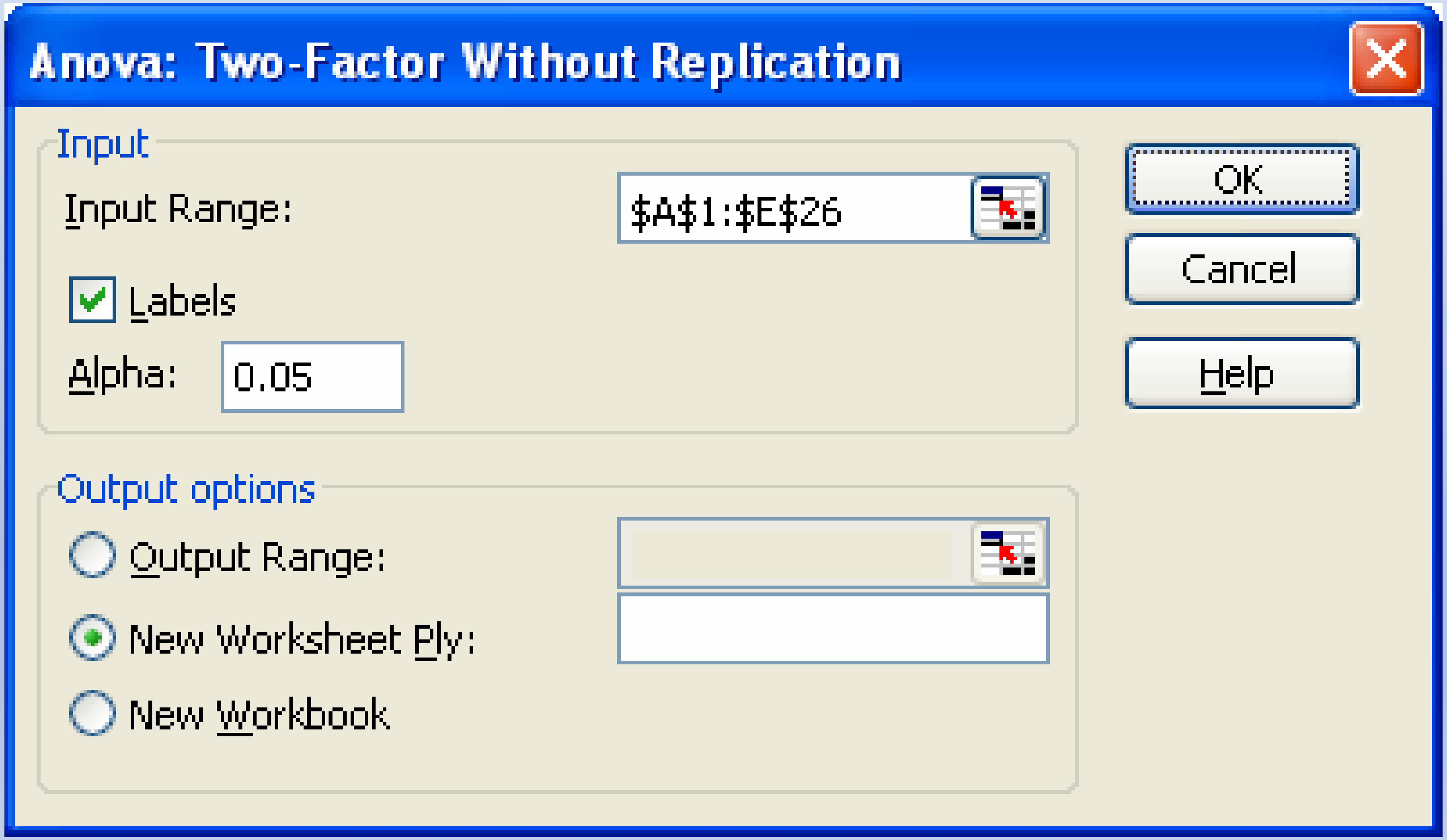
There are b = 25 blocks, and
k = 4 treatments in this example.

Example

COMPUTE

Click Data, Data Analysis, Anova: Two Factor Without Replication

a.k.a. Randomized Block



Example

COMPUTE

	A	B	C	D	E	F	G
1	Anova: Two-Factor Without Replication						
2							
3	*SUMMARY*	*Count*	*Sum*	*Average*	*Variance*		
4	1	4	30.60	7.65	17.07		
5	2	4	22.30	5.58	10.20		
25	22	4	112.10	28.03	5.00		
26	23	4	89.40	22.35	13.69		
27	24	4	93.30	23.33	7.11		
28	25	4	113.10	28.28	4.69		
29							
30	Drug 1	25	438.70	17.55	32.70		
31	Drug 2	25	452.40	18.10	73.24		
32	Drug 3	25	386.20	15.45	65.72		
33	Drug 4	25	483.00	19.32	36.31		
34							
35							
36	ANOVA						
37	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
38	Rows	3848.7	24	160.36	10.11	0.0000	1.67
39	Columns	196.0	3	65.32	4.12	0.0094	2.73
40	Error	1142.6	72	15.87			
41							
42	Total	5187.2	99				