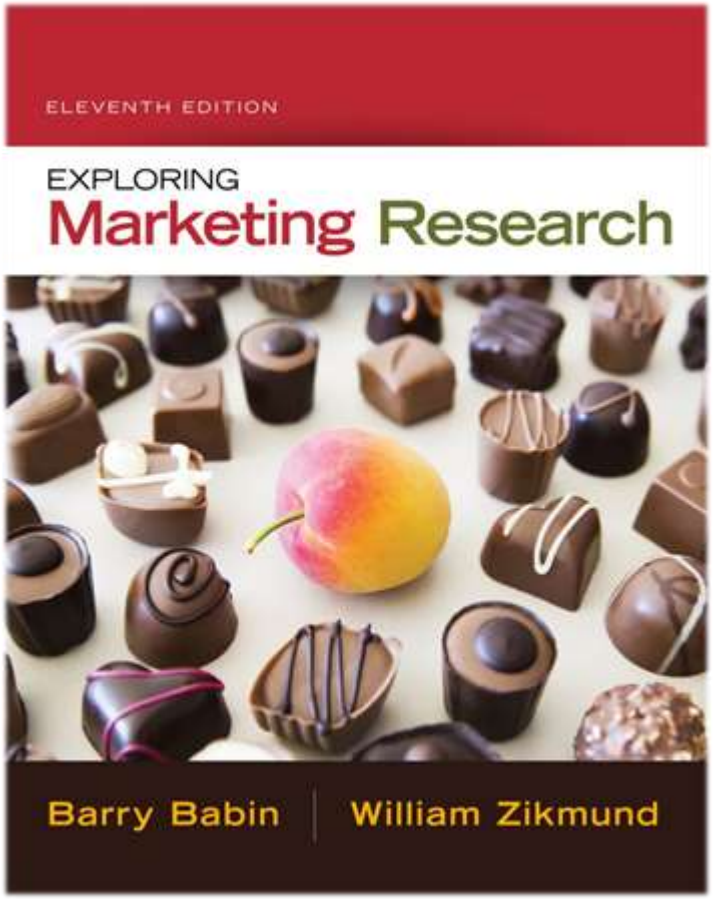


Chapter 15 Testing for Differences Between Groups and for Predictive Relationships

LEARNING OUTCOMES

After studying this chapter, you should

1. Choose an appropriate statistic based on data characteristics
2. Compute a $\chi 2$ statistic for cross-tab results
3. Use a *t*-test to compare a difference between two means
4. Conduct a one-way analysis of variance test (ANOVA)
5. Appreciate the practicality of modern statistical software packages
6. Understand how the General Linear Model (GLM) can predict a key dependent variable

Introduction

- A surprising number of inferences involve two variables
- The automated search for relationships between two variables provides the backbone for automated big data searchers
- Sometimes, the marketing analyst reduces a more complex analysis involving multiple variables to a series of two-variable comparison because presenting the results becomes very simple

What Is the Appropriate Test Statistic?

- Researchers commonly test hypotheses stating that two groups differ
  - Such tests are bivariate tests of differences when they involve only two variables
- Both the type of measurement and the number of groups to be compared influence the type of bivariate statistical analysis

What Is the Appropriate Test Statistic? (cont'd.)

- Two questions help determine the analytical approach
  - How many independent variables (IV) and dependent variables (DV) are involved in the analysis?
  - What is the scale level of the independent and dependent variables involved in the analysis?

EXHIBIT 15.2 Choosing the Right Statistic

Independent Variables:	Dependent Variables: 1 Nominal / Ordinal DV	1 at Least Interval DV	More than 1 DV
1 Nominal/Ordinal IV	Cross-Tabulation with χ^2 Test	*t*-Test or One-Way ANOVA	Multivariate Analysis
2 or More Nominal/ Ordinal IVs	Cross-Tabulation with χ^2 Test	n-Way ANOVA	Multivariate ANOVA
At Least 1 Nominal/ Ordinal IV and at Least 1 Interval or Ratio IV	Multivariate Analyses—(Logistic Regression)	Full-Factorial ANCOVA	Multivariate MANCOVA
1 Interval/Ratio IV	*t*-Test	Simple Regression	Multivariate Regression
1 or More Interval/ Ratio IVs	Multivariate Analyses—(Logistic Regression)	Multiple Regression	Multivariate Analyses such as Path Model

Color Code:

- Beyond the Scope of this Text
- Variations of the GLM illustrated in Chapter
- Independent Variable Condition
- Dependent Variable Condition
- Bivariate test illustrated in Chapter

Source: © Cengage Learning 2013.

Cross-Tabulation Tables: The χ^2 Test for Goodness-of-Fit

- One of the most widely used and simplest techniques for describing sets of relationships is the cross-tabulation
  - A cross-tabulation, or contingency table, is a joint frequency distribution of observations on two or more nominal or ordinal variables
  - The χ^2 distribution provides a means for testing the statistical significance of contingency tables
  - The test involves comparing the observed frequencies (O_i) with the expected frequencies (E_i) in each cell of the table

Cross-Tabulation Tables: The χ^2 Test For Goodness-Of-Fit (cont'd.)

- The goodness- (or closeness-) of-fit of the observed distribution with the expected distribution is captured by this statistic
- The test allows us to conduct tests for significance in the analysis of the *R* x *C* contingency table (*R* = row and *C* = column)
  - Formula: $E_{ij} = \frac{R_i C_j}{n}$
    - where
      R_i = total observed frequency count in the ith row
      C_j = total observed frequency count in the jth column
      n = sample size

Computing χ^2 and Hypothesis Testing

- To compute a chi-square, the same formula as before is used, except that we calculate degrees of freedom as the number of rows minus one, times the number of columns minus one
- Testing the hypothesis involves two key steps
  - Examine the statistical significance of the observed contingency table
  - Examine whether the differences between the observed and expected values are consistent with the hypothesized prediction
    - Proper use of the chi-square test requires that each expected cell frequency (*E*) have a value of at least 5

The *t*-Test for Comparing Two Means

- Independent samples *t*-test
  - A *t*-test is appropriate when a researcher needs to compare means for a variable grouped into two categories based on some less-than interval variable
    - One way to think about this is as testing the way a dichotomous (two levels) independent variable is associated with changes in a continuous dependent variable
    - Most typically, the researcher will apply the independent samples *t*-test which tests the differences between means taken from two independent samples or groups
    - This test assumes the two samples are drawn from normal distributions and that the variances of the two populations are approximately equal (homoscedasticity)

Independent Samples *t*-Test Calculation

- The *t*-test actually tests whether or not the differences between two means is zero
  - The null hypothesis is normally stated as $\mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$
  - However, since this is inferential statistics, we test the idea by comparing two sample means $(\overline{X}_1 - \overline{X}_2)$
  - Thus, the t-value is a ratio with information about the differences between means (provided by the sample) in the numerator and the standard error in the denominator
    - The question is whether the observed differences have occurred by chance alone

Independent Samples *t*-Test Calculation (cont'd.)

- A pooled estimate of the standard error is a better estimate of the standard error than one based on the variance from either sample
  - A higher *t*-value is associated with a lower p-value, and as the *t* gets higher and the p-value gets lower, the researcher has more confidence that the means are truly different
- In a test of two means, the degrees of freedom are calculated as follows:
  - $df = n - k$ (where $n = n_1 + n_2$ and k = number of groups)

Practically Speaking

- In practice, computer software is used
  - Interpretation of the *t*-test is made by focusing on either the p-value or the confidence interval and the group means
  - Basic steps
    - Examine the difference in means to find the "direction" of any difference
    - Compute or locate the computed *t*-test value
    - Find the p-value associated with this *t* and the corresponding degrees of freedom

Practically Speaking (cont'd.)

- Note that the means may be the same due to the variance
  - The *t*-statistic is a function of the standard error, which is a function of the standard deviation
    - Check for outliers
    - Consider increasing the sample size and test again
  - As samples get larger, the *t*-test and *Z*-test will tend to yield the same result
    - A *t*-test can be used with large samples
    - A *Z*-test should not be used with small samples
    - A *Z*-test can be used in instances where the population variance is known ahead of time

EXHIBIT 15.3 Independent Samples *t*-Test Results

Group Statistics

	rel	N	Mean	Std. Deviation	Std. Error Mean
Price	Catholic	57	61.00	43.381	5.746
	Protestant	43	50.27	64.047	9.767

1. Shows mean, standard deviation, and standard error for each group (Catholic and Protestant)

NOTE: Top row shows results assuming equal variances. Bottom row assumes variance is different in each.

Independent Samples Test

		Levene's Test for Equality of Variances		*t*-Test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	*t*	*df*	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Price	Equal variances assumed	.769	.383	.998	98	.321	10.734	10.752	−10.603	32.070
	Equal variances not assumed			.947	69.829	.347	10.734	11.332	−11.868	33.336

2. Computed *t*-test value shown in this column ($t = 0.998$).

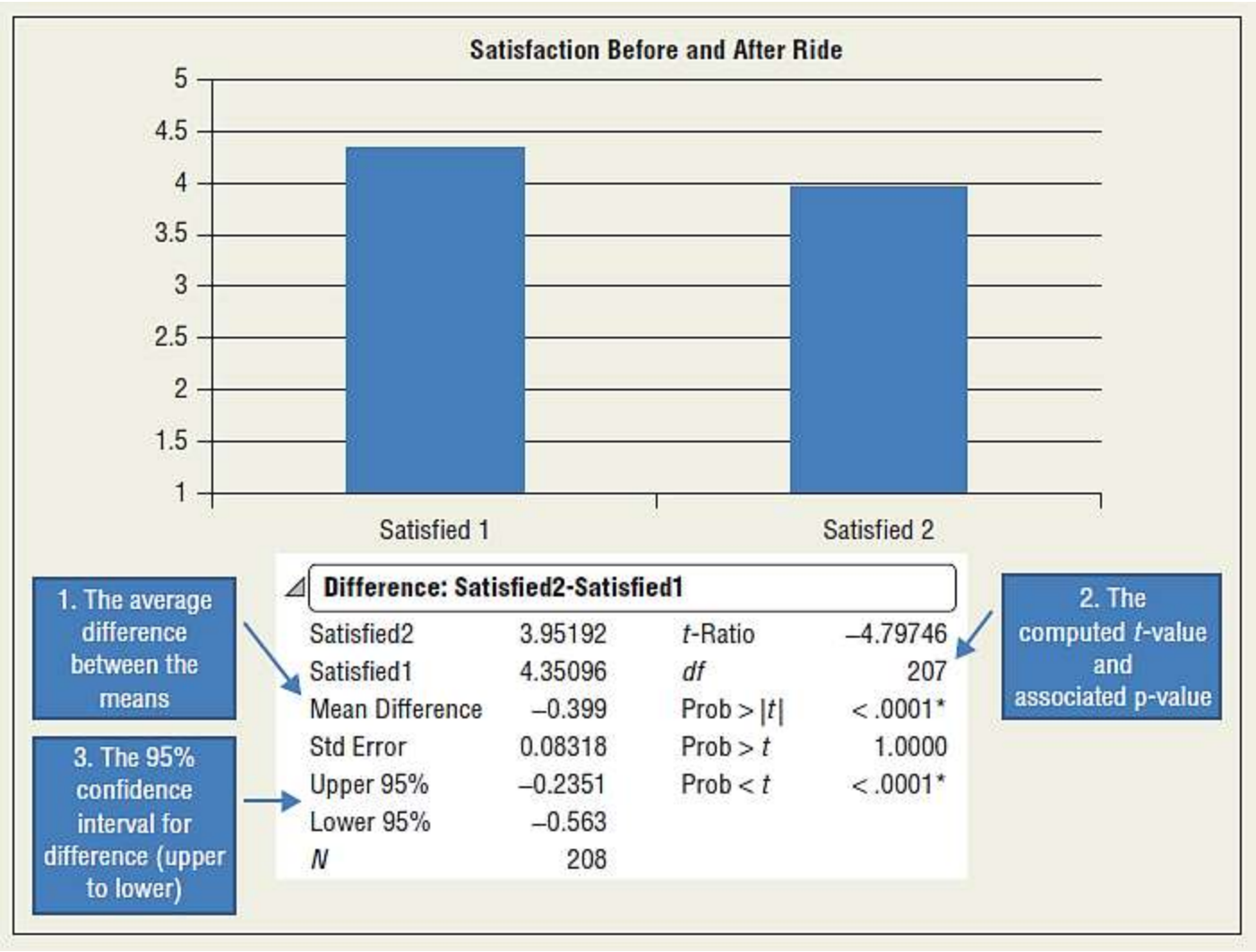
3. P-value for *t*-value and associated degrees of freedom ($t = 0.998$, 98 *df*)

4. Confidence intervals for $\alpha = 0.05$ (100% − 95%). In this case, it includes 0.

Paired Samples *t*-Test

- A paired samples *t*-test is appropriate when means that need to be compared are not from independent samples (i.e., the same respondent is measured twice)
- When a paired samples *t*-test is appropriate, the two numbers being compared are usually scored as separate variables

EXHIBIT 15.4 Illustration of Paired-Samples *t*-Test Results



The *Z*-Test For Comparing Two Proportions

- The *Z*-test for differences of proportions is used to test the hypothesis that the two proportions will be significantly different for two independent samples or groups
  - Requires a sample size greater than 30
- Appropriate for a hypothesis of this form:
  - $H_0: \pi_1 \neq \pi_2$ or $H_0: \pi_1 - \pi_2 = 0$
  - The comparison of the observed sample proportions p_1 and p_2 allows the researcher to ask whether the differences from two *large* random samples occurred due to chance alone

The *Z*-Test Formula

- $Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{S_{p1-p2}}$
  - where
    - p_1 = sample proportion of successes in group 1
    - p_2 = sample proportion of successes in group 2
    - $\pi_1 - \pi_2$ = hypothesized population proportion 1 minus hypothesized population proportion 2
    - S_{p1-p2} = pooled estimate of the standard error of differences in proportions
  - The statistic works on the assumption that the value of $\pi_1 - \pi_2$ is zero
    - Note the similarity between this and the paired-samples *t*-test

Standard Error of Differences in Proportions Formula

- $S_{p1-p2} = \sqrt{\overline{pq}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$
  - where
    - $\overline{p}$ = pooled estimate of proportion of successes in a sample
    - $\overline{q}$ = 1 $-\overline{p}$, or pooled estimate of proportion of failures in a sample
    - n_1 = sample size for group 1
    - n_2 = sample size for group 2
  - To calculate the pooled estimator, $\overline{p}$:
    - $\overline{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

One-Way Analysis of Variance (ANOVA)

- When the means of more than two groups or populations are to be compared, one-way analysis of variance (ANOVA) is the appropriate statistical tool
  - ANOVA involving only one grouping variable is often referred to as *one-way* ANOVA because only one independent variable is involved
  - Another way to define ANOVA: the appropriate statistical technique to examine the effect of a less than interval independent variable on an at least interval dependent variable

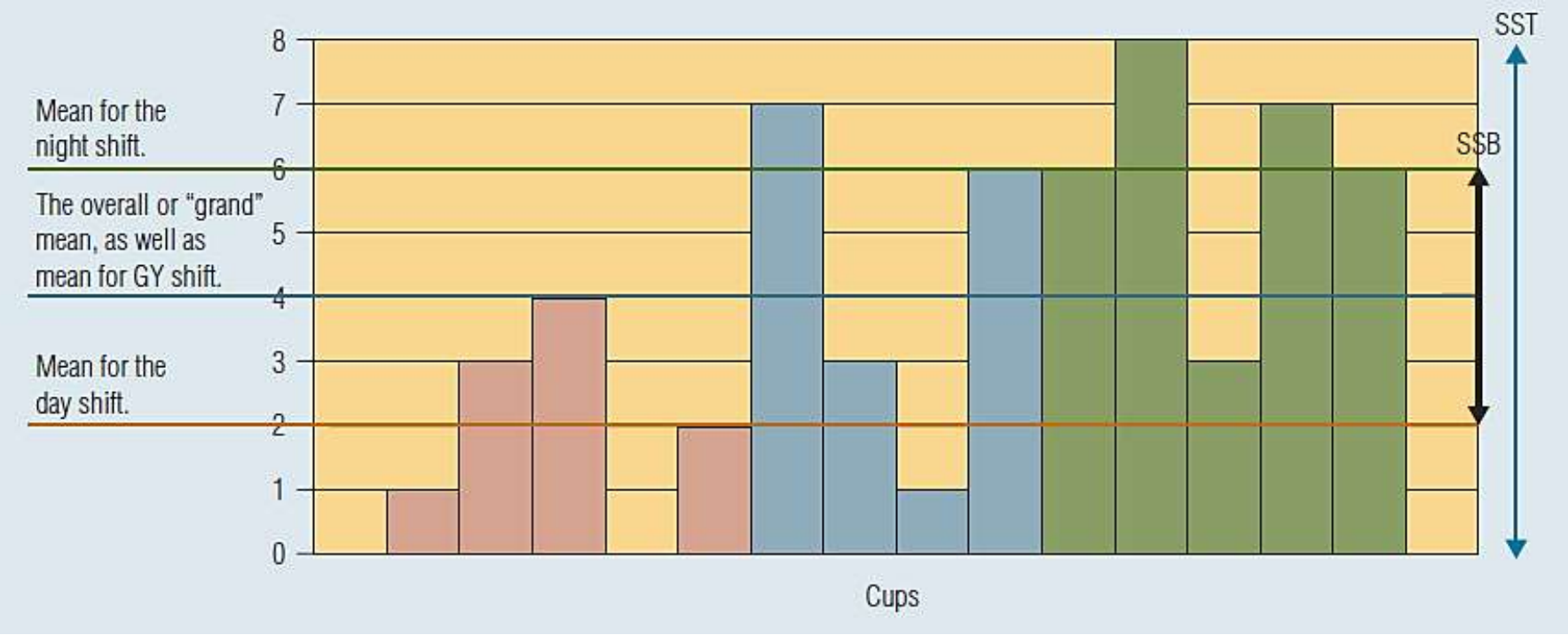
One-Way Analysis of Variance (ANOVA) (cont'd.)

- An independent samples *t*-test can be thought of as a special case of ANOVA in which the independent variable has only two levels
  - When more levels exist, the *t*-test alone cannot handle the problem
- The null hypothesis in such a test is that all the means are equal
- The substantive hypothesis tests in ANOVA is: at least one group mean is not equal to another group mean
  - As the term analysis of variance suggests, the problem requires comparing variances to make inferences about the means

Simple Illustration of ANOVA

- Data are given describing how much coffee respondents report drinking each day based on which shift they work (i.e., day, night, graveyard)
- A table displaying the means for each group and the overall mean is given, and Exhibit 15.5 plots each observation with a bar and lines corresponding to the variances

EXHIBIT 15.5 Illustration of ANOVA Logic



Partitioning Variance in ANOVA

- Total variability
  - An implicit question with the use of ANOVA is "How can the dependent variable best be predicted?"
    - Absent any additional information, the error in predicting an observation is minimized by choosing the central tendency, or mean, for an interval variable
    - The total error (or variability) that would result from using the grand mean, meaning the mean over all observations, can be thought of as:
    - *SST* = Total of (observed value – grand mean)^2
  - Although the term error is used, this really represents how much total variation exists among the measures

Between-Groups Variance

- ANOVA tests whether "grouping" observations explains variance in the dependent variable
  - The between groups variance can be found by taking the total sum of the weighted difference between group means and the overall mean:
    - *SSB* = Total of n_{group}(group mean – grand mean)^2
    - The weighting factor (n_{group}) is the specific group sample size
    - The total SSB represents the variation explained by the experimental or independent variable

Within-Group Error

- While the group means explain the variation between the total mean and the group mean, the distance from the group mean and each individual observation remains unexplained, and this distance is called within-group error or variance
  - The values for each observation can be found by:
    - $SSE = Total\ of\ (Observed\ Mean\ - Group\ Mean)^2$
  - The term total error variance is sometimes used to refer to SSE since it is variability not accounted for by the group means

EXHIBIT 15.6 Interpreting ANOVA

Tests of Between-Subjects Effects (Dependent Variable: Coffee)

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	40.000[a]	2	20.000	4.400	.039
Intercept	221.538	1	221.538	48.738	.000
Shift	40.000	2	20.000	4.400	.039
Error	50.000	11	4.545		
Total	314.000	14			

[a]R Squared = .444 (Adjusted R Squared = .343)

1. This row shows overall *F*-value testing whether all group means are equal. The sums of squares column calculates the SST, SSE, and SSB (shift row).

Shift	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Day	2.000	.953	2.099	4.099
GY	4.000	1.066	1.654	6.346
Night	6.000	.953	3.901	8.099

2. This column shows the group means for each level of the independent variable.

The *F*-Test

- The *F*-Test is the key statistical test for an ANOVA model
  - Determines whether there is more variability in the scores of one sample than in another sample
  - The key question is whether the two sample variances are different from each other or whether they are from the same population
  - The *F*-statistic (of *F*-ratio) can be obtained by taking the larger sample variance and dividing by the smaller sample variance
  - Degrees of freedom must be specified

Using Variance Components to Compute *F*-Ratios

- Three forms of variation
  - SSE – variation of scores due to random error or within-group variation due to individual differences from the group mean
  - SSB – systematic variation of scores between groups due to manipulation of an experimental variable or group classifications of a measured independent variable or between-group variance
  - SST – the total observed variation across all groups and individual observations

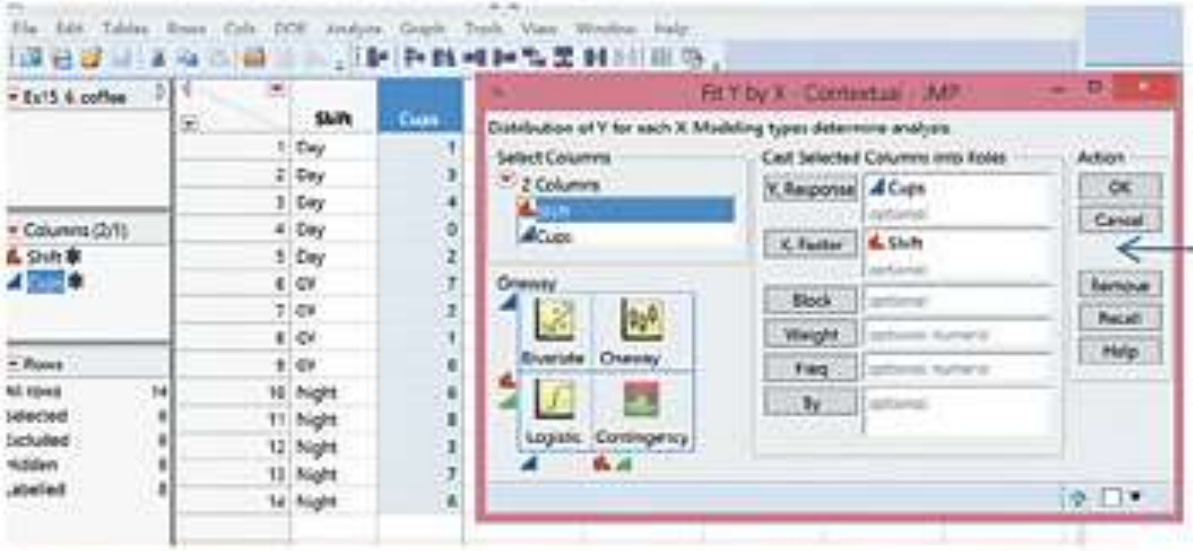
Using Variance Components to Compute *F*-Ratios (cont'd.)

- The *F*-distribution is a function of the ratio of these two sources of variance:
  - $F = f\left(\frac{SSB}{SSE}\right)$
  - A larger ratio of variance between groups to variance within groups implies a greater value of *F*
  - If the *F*-value is large, the results are likely to be statistically significant

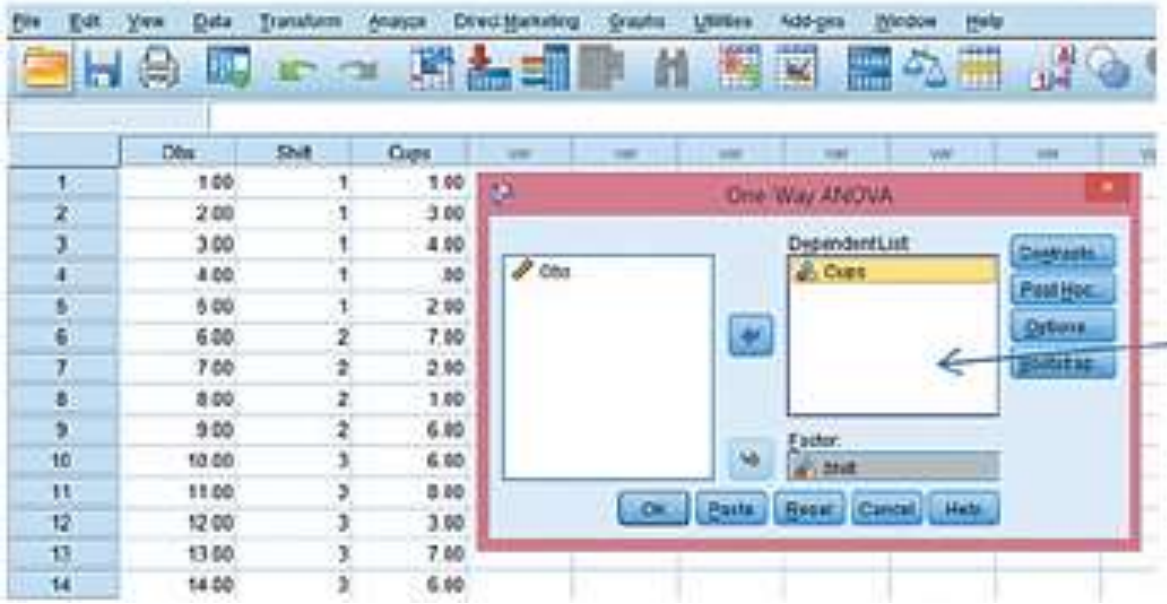
A Different But Equivalent Representation

- *F* also can be thoughts of as a function of the between group variance and total variance
  - $F = f\left(\frac{SSB}{SST - SSB}\right)$

EXHIBIT 15.7 How to Do One-Way ANOVA using SAS JMP, SPSS, and EXCEL

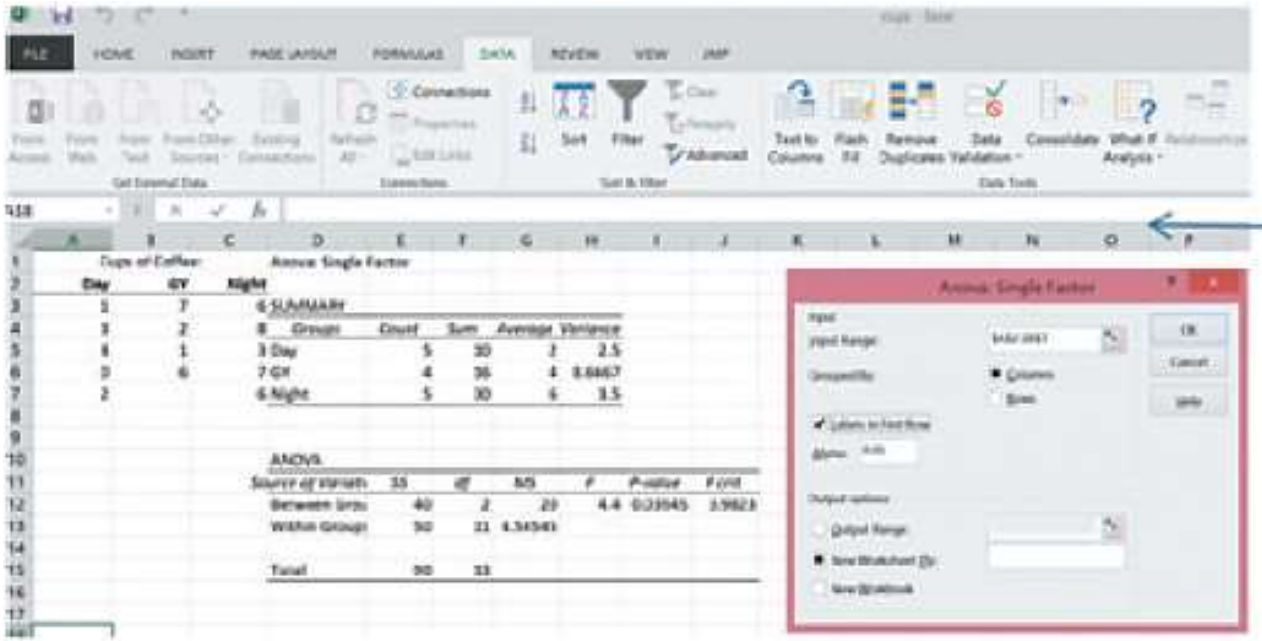


JMP: Click Analyze and then fit Y by X. Enter interval or ratio Y response and nominal or ordinal X Factor. Then click OK.



SPSS: Click Analyze and then choose "Compare Means." Select "One-Way ANOVA" from the options and enter data as shown here. Then click OK.

EXHIBIT 15.7 How to Do One-Way ANOVA using SAS JMP, SPSS, and EXCEL (cont'd.)



EXCEL: After adding the data analysis pack, select ANOVA and enter data as shown here. Then click OK.

Practically Speaking

- The first thing to check is whether or not the overall model *F* is significant
  - Second, the researcher must examine the actual means from each group to properly interpret the result

Statistical Software

- The marketing research analyst has access to statistical software that facilitates statistical analysis by quickly and easily providing results for *t*-tests, cross-tabulations, ANOVA, GLM, and more
  - Some data mining routines even automate some of this analysis
  - Some of the most common statistical software packages are SPSS, now owned by IBM, SAS, and its new user friendly product called JMP

Statistical Software (cont'd.)

- Excel includes basic data analysis functions and an add-in data analysis function that contains procedures like ANOVA
- However, for marketing researchers, packages like SPSS and JMP offer an easy to use interface and a standardized approach to statistics

General Linear Model

- Multivariate dependence techniques are variants of the general linear model (GLM)
  - The GLM is a way of modeling some process based on how different variables cause fluctuations from the average dependent variable
- Fluctuations can come in the form of group means that differ from the overall mean as is in ANOVA or in the form of a significant slope coefficient as in regression

GLM Equation

- $\hat{Y}_i = \overline{Y} + \Delta X + \Delta F + \Delta XF$
  - Here, $\overline{Y}$ represents a constant, which can be thought of as the overall mean of the dependent variable, *ΔX* and *ΔF* represent changes due to main effect independent variables (such as experimental variables) and blocking independent variables (such as covariates or grouping variables), respectively, and *ΔXF* represents the change due to the combination (interaction effect) of those variables
    - Y_i in this case could represent multiple dependent variables, just as *X* and *F* could represent multiple independent variables
    - This form is an ANOVA representation

ANCOVA

- An ANCOVA representation would add a continuous covariate (Xc):
  - $\hat{Y}_i = \overline{Y} + \Delta X + \Delta F + \Delta XF + BX_C$
  - *B* is a regression coefficient

Regression Analysis

- Simple regression investigates a straight-line relationship of the type:
  - $Y = \alpha + \beta X$
    - Where Y is a continuous dependent variable and X is an independent variable that is usually continuous (although a dichotomous nominal or ordinal variables can be included in the form of a dummy variable)
    - Alpha (α) and beta (β) are two parameters that must be estimated so that the equation best represents a given set of data
    - These two parameters determine the height of the regression line and the angle of the line relative to horizontal
    - When these parameters change, the line changes

Interpreting Multiple Regression Analysis

- Multiple regression analysis allows one dependent variable to be explained by more than one independent variable
  - When trying to explain sales, plausible independent variables include prices, economic factors, advertising intensity, and consumers' incomes in the area
  - A simple regression equation can be expanded to represent multiple regression analysis:
  - $Y_i = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_nX_n + e_i$

Parameter Estimate Choices

- The estimates for α and β are the key to regression analysis
  - In most business research, the estimate of β is most important
    - The explanatory power of regression rests with β because this is where the direction and strength of the relationship between the independent and dependent variable is explained
    - A *Y*-intercept term is sometimes referred to as a constant because α represents a fixed point
    - An estimated slope coefficient is sometimes referred to as a regression weight, regression coefficient, parameter estimate, or sometimes even as a path estimate

Parameter Estimate Choices (cont'd.)

- Researchers often explain regression results by referring to a standardized regression coefficient (β)
  - A standardized regression coefficient, like a correlation coefficient, provides a common metric allowing regression results to be compared to one another no matter what the original scale range may have been
    - Due to the mathematics involved in standardization, the standardized *Y*-intercept term is always 0

Using Shorthand Regression Coefficients as Either "Raw" or "Standardized"

- The most common shorthand is as follows:
  - B_0 or b_0—raw (unstandardized) *Y*-intercept term; an estimate of what was referred to as α in the previous slide
  - B_1 or b_1—raw regression coefficient or estimate
  - β_1—standardized regression coefficients
    - The bottom line is that when the actual units of measurement are the focus of analysis, such as might be the case in trying to forecast sales during some period, raw (unstandardized) coefficients are most appropriate
    - Use standardized regression when the size of the relationship for each IV can be compared directly

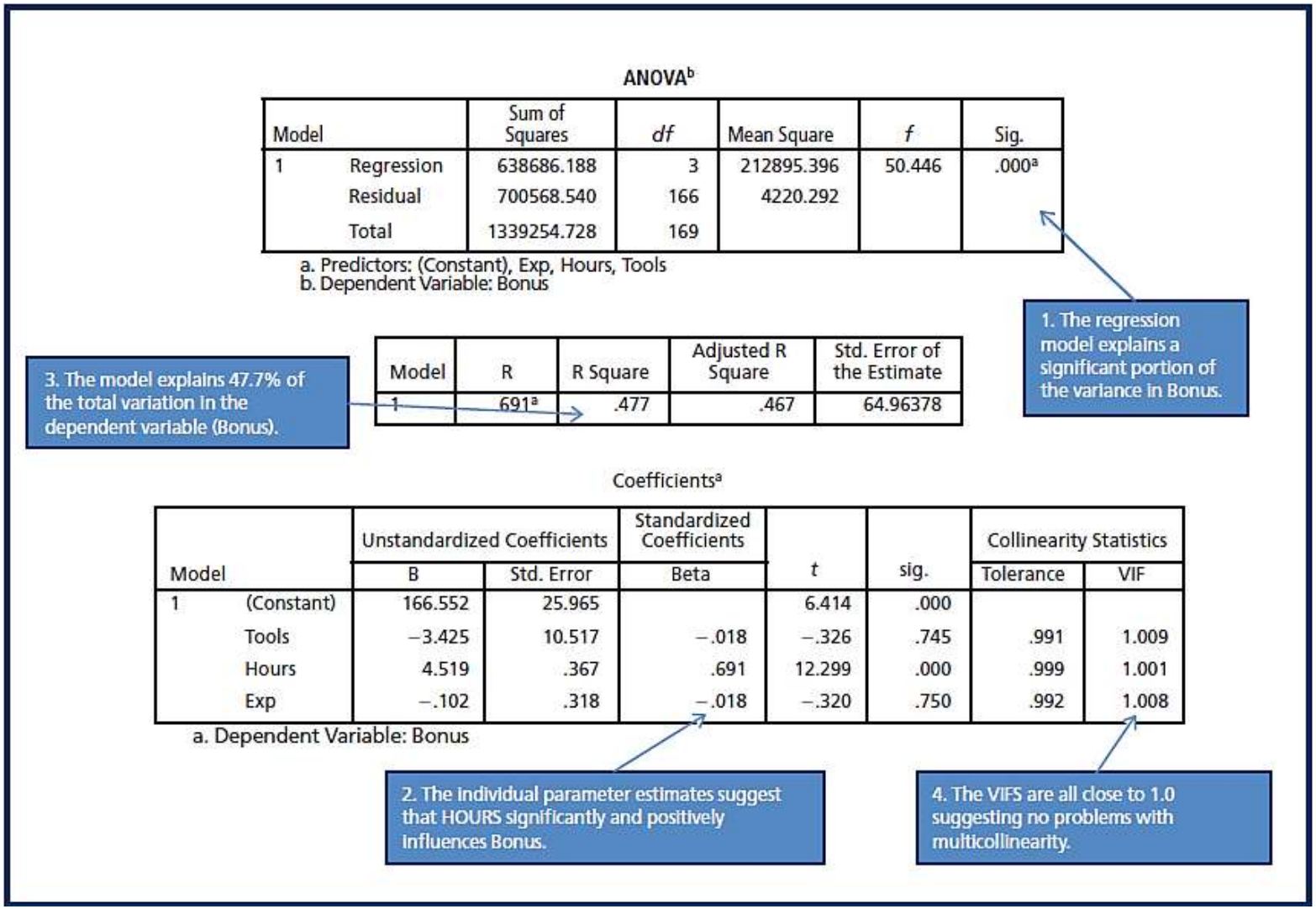
Steps in Interpreting A Multiple Regression Model

- Examine the model *F*-test
  - If the result is not significant, the model should be dismissed
- Examine the individual statistical tests for each parameter estimate
  - Independent variables with significant results can be considered a significant explanatory variable

Steps in Interpreting A Multiple Regression Model (cont'd.)

- Examine the model R^2
  - No cutoff values exist that can distinguish an acceptable amount of explained variation across all regression models
  - However, the absolute value of R^2 is more important when the researcher is more interested in prediction than explanation
- A next step would be to diagnose multicollinearity
  - This is the extent to which the independent variables are redundant

EXHIBIT 15.8 Illustration of Steps for Interpreting a Multiple Regression Model



ANOVA[b]

Model		Sum of Squares	*df*	Mean Square	*f*	Sig.
1	Regression	638686.188	3	212895.396	50.446	.000[a]
	Residual	700568.540	166	4220.292		
	Total	1339254.728	169			

a. Predictors: (Constant), Exp, Hours, Tools
b. Dependent Variable: Bonus

1. The regression model explains a significant portion of the variance in Bonus.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.691[a]	.477	.467	64.96378

3. The model explains 47.7% of the total variation in the dependent variable (Bonus).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	*t*	sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	166.552	25.965		6.414	.000		
	Tools	−3.425	10.517	−.018	−.326	.745	.991	1.009
	Hours	4.519	.367	.691	12.299	.000	.999	1.001
	Exp	−.102	.318	−.018	−.320	.750	.992	1.008

a. Dependent Variable: Bonus

2. The individual parameter estimates suggest that HOURS significantly and positively influences Bonus.

4. The VIFS are all close to 1.0 suggesting no problems with multicollinearity.