CHAPTER 16

3. HYPOTHESIS TESTING

I. **Hypothesis**

A. **Hypothesis Defined**–assumption or guess that a researcher or manager makes about some characteristic of the population being investigated

1. In hypothesis testing, a researcher determines whether a hypothesis concerning some characteristic of the population is likely to be true, given the evidence.

2. A statistical hypothesis test allows us to calculate the probability of observing a particular result IF the stated hypothesis is true.

B. Two Basic **Explanations of Observed Differences**

1. **Hypothesis** is **true** and the observed difference is likely due to sampling error

2. **Hypothesis** is **false** and the true value is some other value

II. **Steps in Hypothesis Testing**

A. **Step One: Stating the Hypothesis**

1. Hypotheses are stated using two basic forms:

a. **Ho**–the null hypothesis (sometimes called the *hypothesis of the status quo*) that is tested against its complement

b. **Ha**–the alternative hypothesis (sometimes called the *research hypothesis of interest*) that is the complement of Ho

2. The null and the alternative hypothesis must be stated in a way that both cannot be true at the same time.

B. **Step Two: Choosing the Appropriate Test Statistic**

1. **Exhibit 16.1 Statistical Tests and Their Uses**–provides a guide to selecting the appropriate test for various situations

C. **Step Three: Developing a Decision Rule**

1. As discussed in an earlier chapter, we know it's very unlikely for our sample estimate to exactly equal our population parameter. The issue is whether or not the difference between the sample value and its expected value (based on the hypothesis) could have occurred by chance if the hypothesized value is true.
2. A decision rule, or standard, is needed to determine whether we should accept or reject the null hypothesis.
3. A Significance level (a) that will determine whether to reject or fail to reject the null hypothesis

a. the level of significance—0.01, 0.05, or 0.10—is the probability that is too low to justify acceptance of the null hypothesis.
b. In other words, we reject the null hypothesis if the probability of occurrence of the observed sample value is low if we assume the null value to be true.

D. **Step Four: Calculating the Value of the Test Statistic**

1. Use the appropriate formula to calculate the value of the statistic for the test chosen
2. Compare the value just calculated to the critical value of the statistic (from the appropriate table), based on the decision rule chosen
3. Based on the comparison, determine to either reject or fail to reject the null hypothesis Ho

E. **Step Five: Stating The Conclusion**–the conclusion summarizes the results of the test. It should be stated from the perspective of the original research question

III. **Types of Errors in Hypothesis Testing**

A. **Two General Types of Errors in Hypothesis Testing**

1. **Type I Error** involves situations in which the researcher rejects the null hypothesis when it is, in fact, true. The probability of committing a **type I error** is referred to as the *alpha level*. Conversely, 1 – *alpha* is the probability of making a correct decision by not rejecting the null hypothesis when, in fact, it is true

2. **Type II Error** involves situations in which the researcher fails to reject the null hypothesis when the null hypothesis actually is false. A type II error is referred to as a ***beta error***. The value 1 – *beta* reflects the probability of making a correct decision in rejecting the null hypothesis when, in fact, it is false.

3. The value of *beta* is never set in advance–*alpha* becomes larger when *beta* is made smaller–to minimize type II error, then you choose a larger value for *alpha* in order to make *beta* smaller.

See Exhibit 16.2 Type I and Type II Errors

4. The decision to reject or fail to reject is never 100% certain. There is a probability of being correct and a probability of being incorrect.

5. Hence, Type I Error is set by the researcher after consulting with the client, considering the limitations of the study (e.g., sample size), and considering the implications of a Type I versus Type II error.

6. Note: Type I and Type II errors are NOT complements of each other.

7. Thus in a hypothesis test, there are three critical elements: *alpha*, *beta*, and the sample size, *n*. Unfortunately, the researcher can only control 2 of the 3 elements. The third element is always determined by the other two.

8. When setting either *alpha* or *beta*, it is critical to understand the implications.

a. Let's assume we are testing for a disease using the following hypothesis:

H_o: Test indicated that you do not have the disease

H_a: Test indicates you do have the disease

b. In this case, *a* = the probability of rejecting H_o when it is true, or the test indicates you have the disease when you do not; and *b*= the probability of accepting Ho when it is false, or the test indicates you do not have the disease when you do.

c. If the treatment for the disease is harmless with no risk to someone without the disease but cures you 100% if you do have the disease, which of the two errors is more costly? Clearly Type I error.

9. Typically, we do not set Type II error, b, in advance. For a fixed sample size, when a is made smaller, b becomes larger; and when a is made larger, b becomes smaller.

10. The most commonly accepted range for a is between .01 to .10.

11. In the above example where Type II error is more serious than Type I, one would set *alpha* = .1. If the situation is that Type I error is more serious, than one would set *alpha* = .01. If there is no real difference between Type I and Type II error, than traditionally researchers set *alpha* = .05.

IV. **Decision to Reject or Fail to Reject**

A. **Accepting Ho or Failing to Reject (FTR) Ho**–whether there is enough evidence in the data to conclude that H_a is correct. There is either sufficient evidence to support Ha (reject H_o) or there is not (fail to reject H_o).

1. The real question is whether there is enough evidence in the data to conclude that Ha is correct

2. Failing to reject Ho is saying that the data do to provide sufficient support of the claim made in Ha–not that the statement made in Ho is acceptable

V. **One-Tailed versus Two-Tailed Test**

A. Tests are either one-tailed or two tailed. This decision depends on the nature of the situation and what you are trying to demonstrate.

B. When the researcher is concerned whether there is a difference in a specific direction only (either higher only or lower only), a one-tail test is sufficient. For example, if we're concerned only whether fat content in a product is higher than a preferred level but we're not concerned if it is lower.

C. If the researcher is concerned whether there is pure difference in either direction, a two-tailed test is appropriate. For example, measuring the temperature that trips a fuse or causes a break.

4. COMMONLY USED STATISTICAL HYPOTHESIS TESTS

I. **Independent versus Related Samples**

A. **Independent Samples**–involve situations in which measurement of the variable of interest in one sample has no effect on (i.e., independent) the measurement of the variable in the other sample (e.g., brand awareness of a product for men vs. women).

B. **Related Samples**–are those in which the measurement of the variable of interest in one sample may influence the measurement of the variable of interest in another sample (e.g., brand awareness on a sample group before a new campaign and brand awareness of the same sample group after the new campaign).

II. **Degrees of Freedom**

A. Many statistical tests refer to degrees of freedom.

B. **Degrees of Freedom (D.F.)**–the number of observations in a statistical problem that are not restricted and or are free to vary.

C. **The Number of Degrees of Freedom**–equal to the number of observations minus the number of assumptions or constraints necessary to calculate a statistic.

5. GOODNESS OF FIT

I. **Chi-Square Test**

A. **Chi-square**

1. **Chi-square defined**–chi-square test enables the research analyst to determine whether an observed pattern of frequencies corresponds to or fits an “expected” pattern.

a. **“goodness of fit”**–test the observed distribution to the expected distribution

6. HYPOTHESES ABOUT ONE MEAN

One of the most common goals in marketing research is to make an inference for a single population mean.

I. **Z Test**

A. **Z Test Defined**–When the sample size is large enough ($n \geq 30$), the most common test statistic used for these types of hypotheses is the Z Test.

1. Example of how to apply this test page 538

B. **Steps Involved in the Z Test**

1. Specify the null and alternative hypotheses
2. Specify the level of sampling error allowed
3. Determine the sample standard deviation (S)
4. Calculate the estimated standard error of the mean
5. Calculate the test statistic
6. State the result

II. ***t*-Test**

A. ***t*-Test Defined**– When the sample size is relatively small ($n<30$), the Student *t*-test is more appropriate than the Z Test for hypothesis testing.

1. The *t*-test with $n – 1$ degrees of freedom is appropriate for making statistical inferences in this case.
2. The *t* distribution is theoretically correct for large samples ($n \geq 30$), however when the sample size is large, the t-distribution and the z-distribution become indistinguishable.
3. Although the z-distribution can be used, most statistical computer packages provide the t-test in their results.

B. **Steps Involved in the *t-Test***

1. Specify the null and alternative hypotheses

2. Specify the level of sampling error allowed

3. Determine the sample standard deviation (S)

4. Calculate the estimated standard error of the mean

5. Calculate the *t*-test statistic

6. State the result

See SPSS Jump Start for *t*-Test

7. HYPOTHESES ABOUT TWO MEANS

I. Testing Differences between Groups

A. Marketers are frequently interested in testing differences between groups.

II. Differences Between Groups

A. **Steps for Testing the Differences between Groups**

1. Specify the null and alternative hypotheses

2. Specify the level of sampling error

3. Calculate the estimated standard error of the differences between the two means

4. Calculate the test statistic Z

5. State the result.

8. HYPOTHESES ABOUT PROPORTIONS

I. **Proportion in One Sample**

A. **Hypothesis Test of Proportions Defined**–test to determine whether the difference between proportions is greater than would be expected because of sampling error

1. Example of how to apply this test pages 546-547

B. **The Procedure for the Hypothesis Test of Proportions**

1. Specify the null and alternative hypotheses

2. Specify the level of sampling error allowed

3. Calculate the estimated Standard error, using the value of *P* specified in the null hypothesis

4. Calculate the test statistic

5. State the results

II. **Two Proportions in Independent Samples**

A. **Difference between the Proportions in Two Different Groups**–such as the proportions of people in two different groups who engage in certain activity or have a certain characteristic

1. Specifications Required and the Procedure for Testing this Hypothesis:

a. Specify the null and alternative hypotheses

b. Set the level of sampling error (management decision)

c. Calculate the estimated standard error of the differences between the two proportions

d. Calculate the test statistic

e. State the result

9. ANALYSIS OF VARIANCE (ANOVA)

I. **Analysis of Variance (ANOVA)**

A. **ANOVA Defined**–test for the differences among the means of two or more independent samples

1. It is more commonly used for tests regarding the differences among the means of several (C) independent groups where ($C \geq 3$).

2. It is a statistical technique that permits the researcher to determine whether the variability among or across the C sample means is greater than expected because of sampling error.

3. The Z and *t* tests described earlier normally are used to test the null hypothesis when only two sample means are involved. However, it is inefficient (and statistically unwise) to make multiple, two-at-a-time comparisons.

4. Steps involved in an ANOVA

a. Specify the null and alternative hypotheses

b. Sum the squared differences between each subsample mean ($\overline{X}$j) and the overall sample mean ($\overline{X}$), weighted by sample size (*n j*). This is call the *sum of squares among groups* or among group variation (SSA)

c. Calculate the variation among group means as measured by the *mean sum of squares among groups* (MSA)

d. Sum the squared differences between each observation (Xjj) and its associated sample mean ($\overline{X}$j), accumulated over all C levels (groups). Also called the *sum of squares within groups* or *within group variation*, it is generally referred to as the *sum of squared error* (SSE)

e. Calculate the variation within the sample groups as measured by the mean sum of squares within groups.

Referred to as *mean square error* (MSE), it represents an estimate of the random error in the data.

f. Calculate the *F* statistic

g. State the results

10. p VALUES AND SIGNIFICANCE TESTING

I. ***p* Values and Significance Testing**

A. ***p* Value Defined**–exact probability of getting a computed test statistic assuming the null hypothesis is true.

1. The smaller the ***p* value**, the smaller the probability that the observed result occurred by chance

2. The ***p* value** is the most demanding level of statistical (not managerial) significance that can be met, based on the calculated value of the statistic.

3. An example is provided on page 553 using Exhibit 16.5.

Exhibit 16.5 Sample *t*-Test Output (example of a *p* value calculation)

QUESTIONS FOR REVIEW AND CRITICAL THINKING

1. Explain the notions of mathematical differences, managerially important differences, and statistical significance. Can results be statistically significant and yet lack managerial importance. Explain your answer.

Mathematical differences occur when two measures are not exactly the same. This does not mean that the difference is necessarily statistically significant nor managerially important. Statistical significance is a difference that is large enough to be unlikely to have occurred because of chance. Managerial importance is the concept that the difference is large enough to have meaning in the decision the manager makes.

Yes, results can be statistically significant and yet lack managerial importance. A .05% increase in sales resulting from a 20% increase in advertising may be statistically significant, but the increase may have cost the company an additional $1 million in advertising expense. For this amount, the dollar amount of the increase in profits would have to be greater than $1 million for the manager to think it had any practical meaning.

2. Describe the steps in the procedure for testing hypotheses. Discuss the difference between a null hypothesis and an alternative hypothesis.

In hypothesis testing, the first step is the state the hypothesis. The convention is to state the generally accepted condition, or status quo, as the null hypothesis. Then the alternative hypothesis is stated. The second step involves choosing the appropriate test statistic. Consideration must be given to the situation and the data to determine the correct test statistic. The third step involves developing a decision rule. This is based on the understanding that the sample will not be exactly like the population, and determining what level of difference or error, is acceptable. The acceptable standard is then stated as the rule of when to reject or fail to reject the null hypothesis. In the fourth step, the value of the test statistics is calculated and compared to the critical value stated in the decision rule. Finally, the conclusion from the test is stated.

3. Distinguish between a Type I error and Type II error. What is the relationship between the two?

Type I error, or alpha level, is the probability that the researcher will reject the null hypothesis when it is actually true. Type II error, or beta error, is the probability that the researcher will fail to reject the null hypothesis when it should have been rejected. Type I error plus Type II error does not equal 1 unless the Type I error is 0. That is not a realistic case. The alpha level is set by the researcher, within the constraints of the project. Beta error is not determined by the researcher.

4. What is meant by the terms *independent samples* and *related samples?* Why is it important for a researcher to determine whether an ample is independent?

An independent sample is a sample in which measuring a variable in one population has no effect on measuring that variable in another population. An example would be determining the percentage of voters who are women in Arkansas and the percentage of voters who are women in Texas. A related sample is a sample in which the measurement of a variable in one population may influence the measurement of the variable in the other. For example, if you survey a group about their movie-going experiences and then three months later repeat the survey, the samples are not independent of each other.

5. Your university library is concerned about student desires for library hours on Sunday morning (9:00 a.m. – 12:00 p.m.). It has undertaken a random sample of 1,600 undergraduate students (one-half men, one half women) in each of four status levels (i.e. 400 freshmen, 400 sophomores, 400 juniors, 400 seniors.) If the percentage of students preferring Sunday morning hours are those shown below, what conclusions can the library reach?

Observed Results

	Sen	Jun	Soph	Fresh	**TOTAL**
Women	70	53	39	26	**188**
Men	30	48	31	27	**136**
TOTAL	**100**	**101**	**70**	**53**	**324**

STEP 1: Hypotheses

Ho: There is no relationship between gender and class year.

Ha: There is a significant relationship between gender and class year.

STEP 2: Determine Expected Results

Expected Results if the variables are independent*

	Sen	Jun	Soph	Fresh	**TOTAL**
Women	58.02	58.60	40.62	30.75	**188**
Men	41.98	42.40	29.38	22.25	**136**
TOTAL	**100**	**101**	**70**	**53**	**324**

**Cell values calculated by multiplying each row margin total by each column margin total and dividing by the grand total, e.g., cell 1(woman x sen)=58.02=188*100/324=58.02.*

STEP 3: Calculate Chi-Square Value

Chi-square Calculations*

	Sen	Jun	Soph	Fresh	TOTAL
Women	2.47	0.54	0.06	0.73	
Men	3.42	0.74	0.09	1.02	
TOTAL				c^2=	**9.07**

** Cell values calculated by taking the squared difference between the expected and observed and dividing expected, e.g. cell 1(woman x sen)=(70-58.02)^2/58.02=2.47*

STEP 4: Find Chi Square Table Value based on Significance Level and Degrees of Freedom

Degrees of freedom = (r-1) * (k-1) = (2-1) * (4-1) = 3

The tabulated chi square, 3 df, .05 alpha = 7.814

STEP 5: Compare Results and State Conclusion

The calculated chi square = 9.06

Conclusion: Calculated chi-square > Table chi-square

We can state that, of those who prefer Sunday morning hours, gender and class year are statistically dependent. That is, there is a significant difference in how gender and class status will affect the use of the library on Sunday mornings. We can expect more Senior women than any other group.

6. A consultant has a random sample of 400 usable responses in a database. Included are the following questions:

Household Income Category:
(1) $0 to $24,999 (2) $25,000 to $49,999 (3) $50,000 to $99,999
(4) $100,000 and over

Average Weekly Soft Drink Consumption: ________________________

The consultant wants to determine if there are any statistically significant differences in average weekly soft drink consumption by income category. What statistical test should be applied to give the consultant the information he needs?

Ans: Since the problem will involve comparing more than two means (ratio scale data) for statistically significant differences, 1-Way ANOVA should be used.

7. A market researcher has completed a study of pain relievers. The following table depicts the brand purchased most often broken down by men versus women. Perform a Chi-square test on the data and determine what can be said regarding the crosstabulation.

Pain Reliever Men Women

Pain Reliever	Men	Women
Anacin	40	55
Bayer	60	28
Bufferin	70	97
Cope	14	21
Empirin	82	107
Excedrin	72	84
Excedrin PM	15	11
Vanquish	20	26

STEP 1: Hypotheses

Ho: There is no relationship between gender and the type of brand purchased.

Ha: There is a significant relationship between gender and the type of brand purchased.

STEP 2: Determine Expected Results

The total number of males is 373.

The total number of females is 429

Expected Results If the Variables Are Independent

Pain Reliever	Men	Women	Total
Anacin	44.18	50.82	95
Bayer	40.93	47.07	88
Bufferin	77.67	89.33	167
Cope	16.28	18.72	35
Empirin	87.90	101.10	189
Excedrin	72.55	83.45	156
Excedrin PM	12.09	13.91	26
Vanquish	21.39	24.61	46
Total:	373	429	802

STEP 3: Calculate Chi-Square Value

Pain Reliever	Men	Women
Anacin	0.40	0.34
Bayer	8.89	7.73
Bufferin	0.76	0.66
Cope	0.32	0.28
Empirin	0.40	0.34
Excedrin	0.00	0.00
Excedrin PM	0.70	0.61
Vanquish	0.09	0.08

Chi-square = 21.59

STEP 4: Find Chi Square Table Value based on Significance Level and Degrees of Freedom

Degrees of freedom = (r-1) * (k-1) = (8-1) * (2-1) = 7

The tabulated chi square, 7 df, .05 alpha = 14.067

STEP 5: Compare Results and State Conclusion

The calculated Chi square = 21.59

Conclusion: Calculated chi-square > Table chi-square

The tabular chi square at the .05 level of significance and 7 degrees of freedom is 14.07. The calculated value is higher than the tabular value, so we reject the null hypothesis. We can say with a 95% confidence that there is a significant relationship between sex and the brand purchased.

8. A consultant has collected 700 usable responses via a probability sample concerning political opinions. After analyzing each respondent's political ideology, he/she divides the respondents into two groups, moderate liberals and moderate conservatives. Further, the consultant asks respondents the following:

How likely are you to vote for increasing property taxes in the next election?
(1) very unlikely (2) somewhat unlikely (3) undecided
(4) somewhat likely (5) very likely

The consultant wants to determine if the moderate conservatives differ significantly from moderate liberals concerning how they'll vote on the property tax issue in the next election. How should the data be analyzed?

Ans: The problem involves comparing just two means, and the variable concerning property tax is metric (interval scale), the appropriate method of analysis would be a t-test.

9. American Airlines is trying to determine which baggage handling system it will put in its new hub terminal at San Juan, Puerto Rico. One system is made by Jano Systems and a second baggage handling system is manufactured by Dynamic Enterprises. American has installed a small Jano system and a small Dynamic Enterprises system in two of its low-volume terminals. Both terminals handle approximately the same quantity of baggage each month. American has decided to select the system that will provide the minimum number of instances in which passengers disembarking must wait 20 minutes or longer for baggage. Analyze the data that follow and determine whether there is a significant difference at the .95 level of confidence between the two systems. If there is a difference, which one should American select?

Minutes of Waiting	Jano Systems (Frequency)	Dynamic Enterprise (Frequency)

10-11	4	10
12-13	10	8
14-15	14	14
16-17	4	20
18-19	2	12
20-21	4	6
22-23	2	12
24-25	14	4
26-27	6	13
28-29	10	8
30-31	12	6
32-33	2	8
34-35	2	8
36 or more	2	2

STEP 1: Hypotheses and determine sample proportions

Ho: The difference in the proportions is less than or equal to zero; the proportion of those who waited 20 minutes or more for the Jano system is the same as those who waited 20 or minutes for the Dynamic System.

Ha: The difference in the proportions is greater than zero; the proportion of those who waited 20 minutes or more for the Jano system is the same as those who waited 20 or minutes for the Dynamic System.

	Jano Systems (Frequency)	Dynamic Enterprise (Frequency)
Total Number of people who waited for luggage	88	131
Number of People who waited 20 minutes or more	54	67
Proportion of People who waited 20 minutes or more	61.4%	51.2%

STEP 2: Decide on Significance Level and Look up Appropriate Table Value

Management wants a 95% confidence level, hence an alpha of 0.05. This is a two tailed test, hence, the value from the Z table for alpha = 0.05 is +/-1.96.

STEP 3: Calculate the Pooled Standard Error for the two proportions

$$p = \sqrt{\frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}} \quad = \quad \sqrt{\frac{(.614)88 + (.512)131}{88 + 131}} \quad = \quad .55$$

$$SEp = \sqrt{p(1-p) * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad = \quad \sqrt{.55(.45) * \left(\frac{1}{88} + \frac{1}{131}\right)} \quad = \quad .069$$

Hence, the pooled standard error of the proportions equals .069

STEP 4: Calculate the test statistic

The test statistic equals (.614 - .512)/.069, or 1.49

STEP 5: Compare Results and State Conclusion

Since Zcalc =1.49 < Ztable=1.96, we conclude there is no statistical difference between the proportions.

As a side observation, if instead of comparing those who waited 20 minutes or more you compared those who waited less than 20 minutes, For Jano the proportion is 39% compared to Dynamic's 53%. The resulting zcalc would be -2.15, which is greater than -1.96. In other words, the proportion of people who waited between 10-19 minutes is significantly different for the two systems with Jano system having a lower proportion.

10. Menu space is always limited in fast food restaurants. However, McDonald's has decided that it needs to add one more salad dressing to its menu for its garden salad and check salad. It has decided to test market four flavors: Caesar, Ranch-style, Green Goddess, and Russian. Fifty restaurants were selected in the North-Central region to sell *each* new dressing. Thus, a total of 200 stores were used in the research project. The study was conducted for two weeks and the units of each dressing sold are shown below. As a researcher, you want to know if the differences among the average daily sales of the dressings are larger than can be reasonably expected due to chance. If so, which dressing would you recommend be added to the inventory throughout the United States?

Day	Caesar	Ranch-Style	Green Goddess	Russian	
1	155	143	149	135	
2	157	146	152	136	
3	151	141	146	131	
4	146	136	141	126	
5	181	180	173	115	
6	160	152	170	150	
7	168	157	174	147	
8	157	167	141	130	
9	139	159	129	119	
10	144	154	167	134	
11	158	169	145	144	
13	184	195	178	177	
14	161	177	201	151	Grand Avg.
AVG	158.54	159.69	158.92	138.08	153.81

STEP 1: Hypotheses

H0: M1=M2=M3= M4; mean request for the four salad dressings are equal.

Ha: The variability in group means is greater than would be expected because of sampling error.

STEP 2: Determine Mean Sum of Square Differences BETWEEN Columns

$$SSA = \sum_{j=1}^{C} n_j \left(\overline{X}_j - \overline{X}_t\right)^2 = 13\left[(158.5-153.8)^2 + (159.7-153.8)^2 + (158.9-153.8)^2 + (138.1-153.8)^2\right]$$

SSA= 4298.23

Degrees of Freedom for SSA = C-1 = 4-1 = 3

MSA = SSA / Df = 4298.23 / 3 = 1432.7

STEP 3: Determine Mean Sum of Square Differences WITHIN Columns

$$SSE = \sum_{j=1}^{C} \sum_{i=1}^{n_j} \left(X_{ij} - \overline{X}_j \right)^2$$

SSE=2075.2 +3574.8+4792.9+3126.9 = 13569.8

Degrees of Freedom for SSE= (n1+n2+n3+n4) - C = 52-4 = 48

MSE = SSE / Df = 13569.8 / 48 = 282.7

STEP 4: Calculate the F statistic

F calc = MSA / MSE = 1432.7 / 282.7 = 5.07

F Table Value at alpha = .05 with 3, 35 degrees of freedom = 2.80

STEP 5: Compare Results and State Conclusion

Fcalc > Ftable. Hence, there is at least one mean that is statistically different from the rest. Looking at the results, it would appear that Russian Dressing has the least appeal whereas the other three have almost the same average. Computing confidence intervals for each column mean or other types of direct comparison tests (e.g., Tukey's Test, etc.) would provide more reliable comparisons.