



TESTING A HYPOTHESIS ABOUT TWO INDEPENDENT MEANS

Chapter Preview

Research question	Do women do more housework than men in the United States?
Null hypothesis	There is no difference in the hours of housework done by men and women in the United States.
Test	Two independent-samples <i>t</i> test
Types of variables	One continuous variable: hour of household work (rhhwork) One categorical variable with two categories: 1 "male," 2 "female" (sex)
When to use	Two samples Two populations Population standard deviation is unknown
Assumptions	Independence of observations Normal distribution Homogeneity of variances
Additional tests needed	Equality of variances
Stata code: generic	ttest continuousvar, by(categoricalvar)
Stata code: example	ttest rhhwork, by(sex)

9.1 INTRODUCTION

ARTICLE 9.1

Bloomberg

Women in the U.S. Still Do Way More Housework Than Men

Women in the U.S. Still Do Way More Housework Than Men

The latest data from the Bureau of Labor Statistics reveal how little has changed

by Sheelah Kolhatkar
 sheelahk

June 26, 2015 – 10:53 AM EDT



f

Countless panels, conferences, studies, task forces, and books (see Anne Marie Slaughter's forthcoming blockbuster) pop up every year to address a seemingly intractable problem: why there are so few women CEOs, senators, law firm partners, venture capitalists, and hedge fund moguls, not to mention female executives lower down the chain. Once you attain the highest levels of powerful institutions in America, ample evidence shows, the scenery becomes overwhelmingly male.

For all the resources dedicated to untangling why this is, though, the answer may be relatively simple. Newly released data remind us that a large part of the answer lies at home.

Source: Kolhatkar (2015). Used with permission of Bloomberg L.P. Copyright 2017. All Rights Reserved.

According to Article 9.1, women have been doing more housework than men have since the Bureau of Labor Statistics began its annual Time Use Survey in 2003. A higher percentage of women engage in housework and they work longer hours. As we learned in Chapter 2 on questionnaire design, we should ask how they defined “housework.” Does that include only cleaning and cooking in the home or does it include errands outside of the home such as grocery shopping? Is child care part of housework? As we also learned in Chapter 2, we would want to know more about the data. How large was the sample size? How were participants selected? Finally, we would want to know if there is a statistically significant difference in the mean score between the two groups.

In this chapter, we will learn how to test for a statistically significant difference between two independent-sample means drawn from two populations. This type of test is called the two independent-samples t test. It is used with one continuous variable (hours of housework in the example above) and one categorical variable with two categories (men and women). Although we don’t have access to the time-use survey, we will use the General Social Survey to examine the number of hours of housework that men and women do each week on average. Other examples along with a review of assumptions, procedures, and interpretation of the output are included below.

9.2 WHEN TO USE A TWO INDEPENDENT-SAMPLES t TEST

There are many situations when we may want to compare two means. In this chapter, we only consider cases where there are two independent samples. This means that individuals or objects are assigned to one of two groups. Table 9.1 offers examples from different fields and identifies the continuous variable and the categorical variable with two groups.

We can also consult the decision tree in Figure 8.1 and Appendix 3 when we are unsure about which test to use. Since we are comparing the means, we would follow the path on the left to “comparing means.” Next, we would choose “comparing sample means” since we now have two sample means in this case—the average number of hours worked by men and those worked by women. Underneath the “two independent sample means” is the two independent-samples t test.

TABLE 9.1 ■ EXAMPLES OF TWO INDEPENDENT-SAMPLES t TEST

Field	Research Question	Null Hypothesis	Continuous Variable	Categorical Variable
Criminal Justice	Are men more likely to commit delinquent offenses than women?	There is no difference in the number of delinquent offenses committed by men and by women.	Number of offenses committed	1. Men 2. Women
Economics	Do men earn more than women in the same job with the same set of skills?	There is no difference between salaries of men and women in the same job with the same skill level.	Annual salary	1. Men 2. Women
Political Science	Are Democratic voters younger than Republican voters?	There is no difference in the average age of Democrats and Republicans.	Age	1. Democrats 2. Republicans
Psychology	Does multitasking while studying for an exam have an impact on the final score?	There is no difference in the final scores among students who multitask and those who do not.	Exam score	1. Those who multitask 2. Those who do not multitask
Public Health	Do women who smoke give birth to infants with a lower birth weight?	There is no difference in birth weight of children of pregnant mothers who smoke and those who do not.	Birth weight	1. Pregnant mothers who smoke 2. Pregnant mothers who do not smoke
Sociology	Do Catholics or Protestants spend more time volunteering for community work?	There is no difference in the number of hours per week that Catholics and Protestants spend volunteering.	Hours per week volunteering	1. Catholics 2. Protestants

9.3 CALCULATING THE t STATISTIC

To test for a significant difference between the two means, we must calculate a t statistic. Although Stata will calculate the t statistic in the example that follows in this section, it is important to understand how it is calculated in order to interpret its meaning. It is expressed in Equation 9.1 below.

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{S_{\bar{X}_1 - \bar{X}_2}} \quad (9.1)$$

The numerator is simply the observed difference between the two means and how much greater it is than zero, which is the hypothesized difference. The denominator is the standard error of the mean difference. This is calculated as follows:

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad (9.2)$$

where

S_1^2 is the variance for the first sample

S_2^2 is the variance for the second sample

n_1 is the sample size for the first sample

n_2 is the sample size for the second sample

Combined, the full formula tells us how many standard error units the observed difference is from zero. As described in Chapter 7, this indicates how unusual our results are if the true difference is zero.

9.4 CONDUCTING A t TEST

Using the 2012 General Social Survey, we will examine the number of hours that men and women spend on housework each week.¹ We can begin by generating a summary of these values using the commands (or menus) that we learned in Chapter 6 as illustrated in Output 9.1.

Based on these results, we can see that on average women spend more time on housework than do men, but they have a much larger standard deviation. We now want to

¹For simplicity's sake, we will ignore the sampling weights.



OUTPUT 9.1: AVERAGE AND STANDARD DEVIATION OF HOURS SPENT ON HOUSEHOLD CHORES BY MEN AND WOMEN IN THE UNITED STATES

```
. table sex, c(mean rhhwork sd rhhwork) format(%3.2fc)
```

Sex	mean (rhhwork)	sd (rhhwork)
Male	8.32	9.43
Female	11.86	12.70

use the data to test whether this is a statistically significant difference beginning with our research question below.

Research question

Do men and women spend the same number of hours on housework each week?

Null hypothesis

Men and women spend the same number of hours on housework each week.

Variables

Continuous variable—hours per week spent on housework (rhhwork)

Categorical variable—sex (*male* = 1, *female* = 2)

Assumptions

As described earlier, to use the two independent-samples *t* test, you must have one continuous variable and one categorical variable with two categories. We also make the following assumptions to generate valid results.

1. *Independence of observations:* Each individual can appear in only one of the two groups. In addition, they can only appear once in each group.
2. *Normal distribution:* The dependent variable, hours of housework in this example, should be approximately normally distributed within each category. It only needs to be approximately normally distributed since minor violations of normality do not affect the results. Normality can be tested with the Shapiro–Wilk test.

3. *Homogeneity of variances*: The variances of the two groups must be equal. This is tested with Levene's test. If the variances are not equal, Stata will generate output to show the results with unequal variances assumed with the command **unequal**.

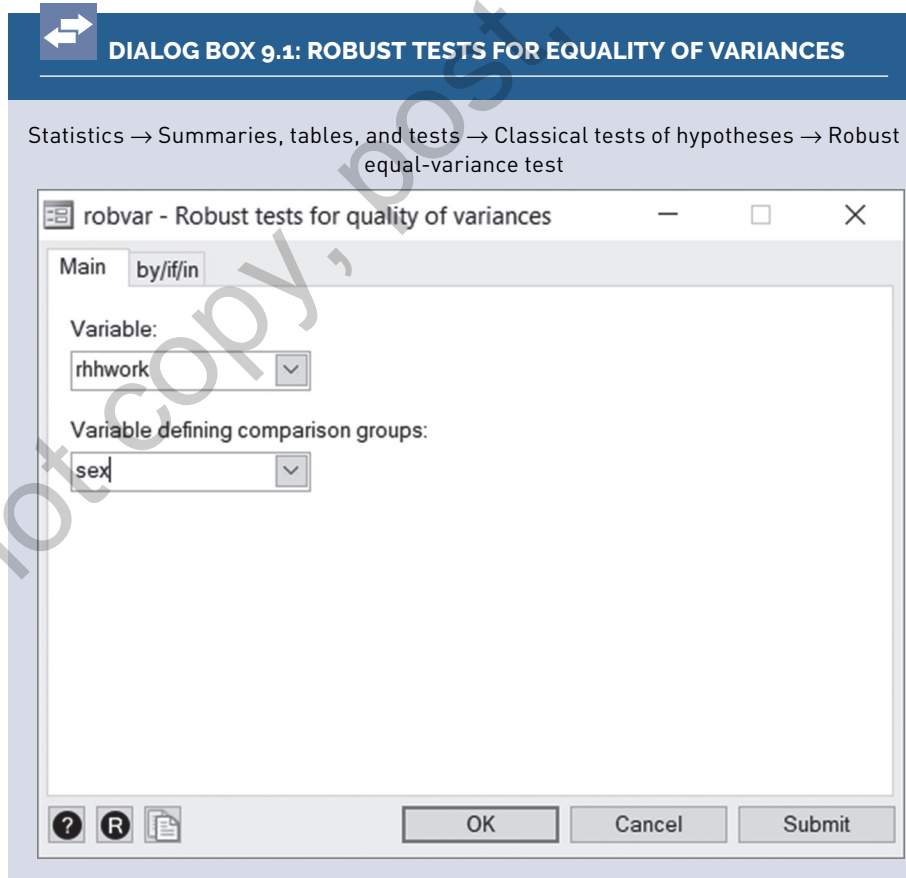
Stata code for doing a t test

Using a do-file, we would run the commands below.

```
robvar rhhwork, by(sex)
ttest rhhwork, by(sex) unequal
esize twosample rhhwork, by(sex) cohensd unequal
```

Menus for doing a t test

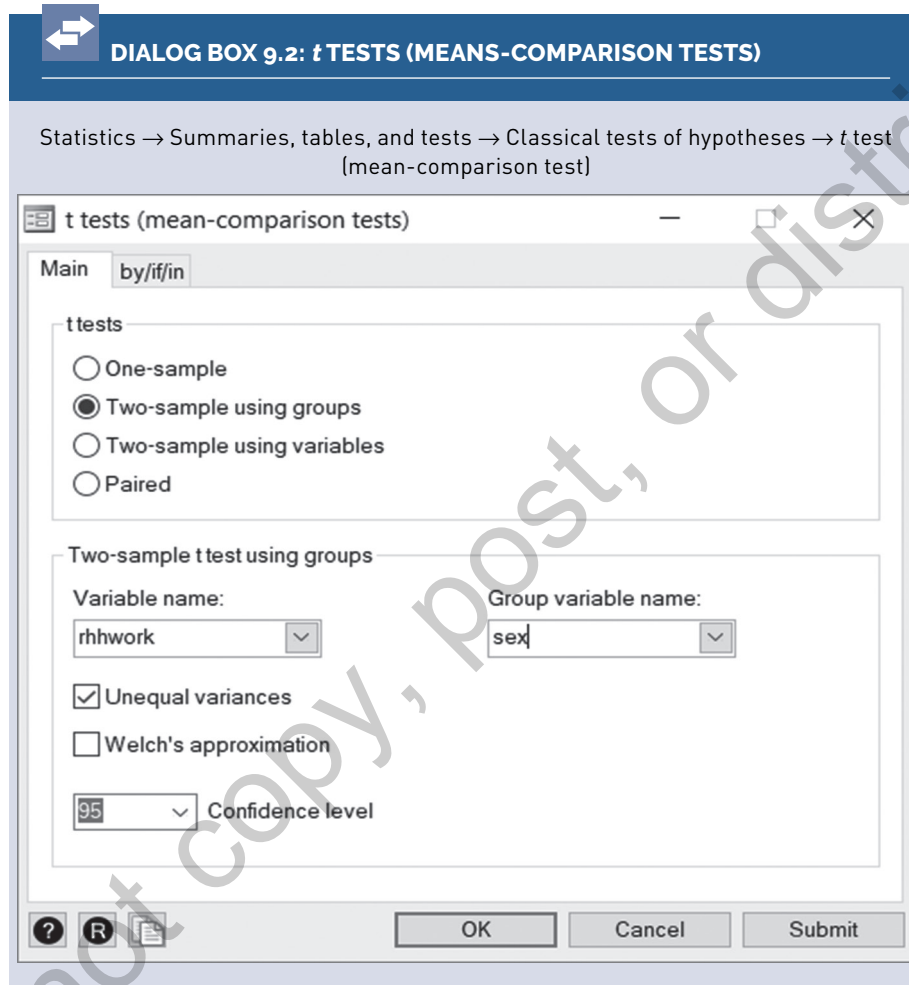
Using menus in Stata, we would click on the sequence listed below that would bring us to Dialog Box 9.1. In the dialog box, we would select the variables “rhhwork” and “sex” in the two drop-down menus as displayed.



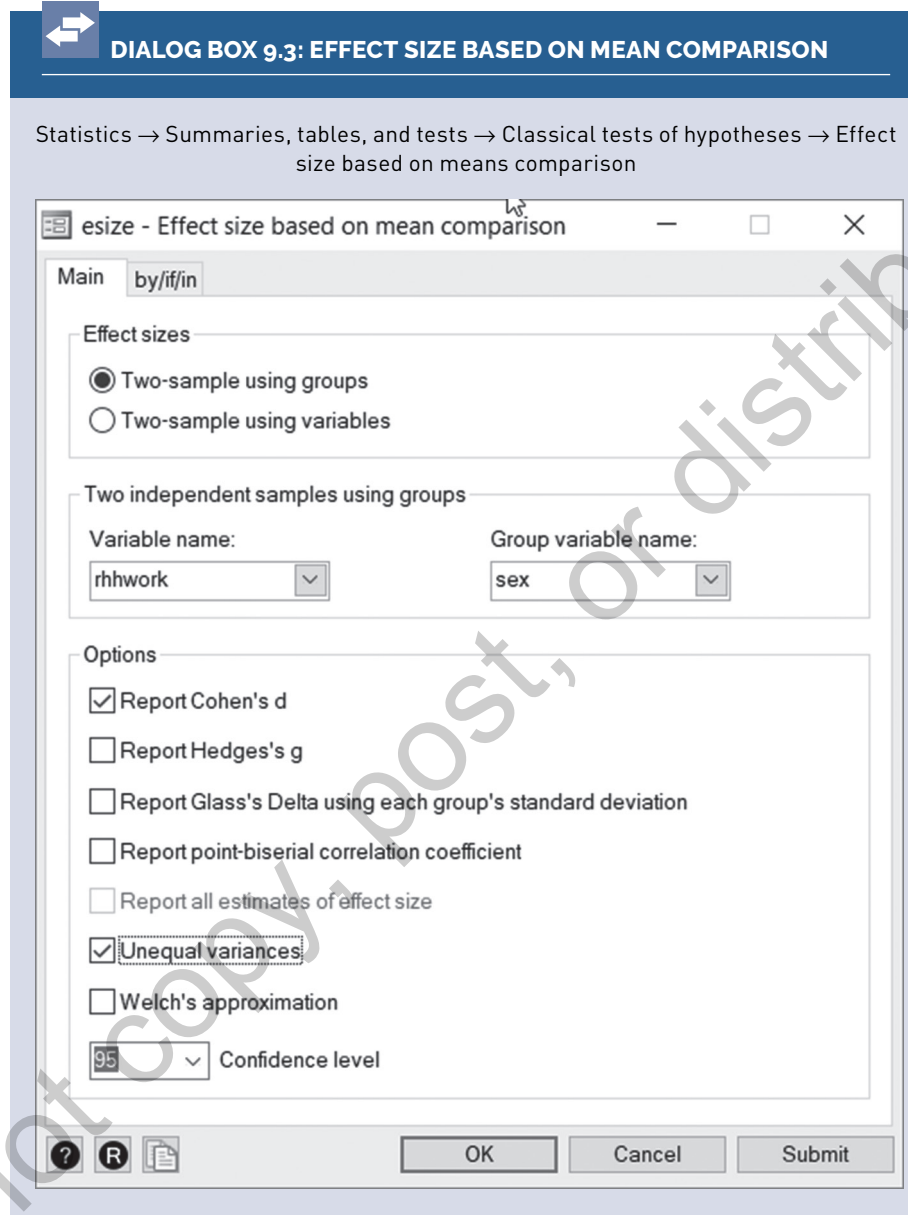
Copyright ©2020 by SAGE Publications, Inc.

This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

We would then click on the following sequence to bring us to Dialog Box 9.2. Depending on the results from the equality of variance test, we would leave the box “unequal variances” either checked or unchecked. This is explained further in Section 9.4 on interpreting the output.



Finally, we could click on the sequence below and fill in the variable names in the drop boxes as shown in Dialog Box 9.3.




9.5 INTERPRETING THE OUTPUT

The first step to determine if there is a significant difference in the number of hours is to check for equality of variances. As we saw in the preview to the chapter, “homogeneity of variances” is one of the assumptions for this test. If the variances are not equal, this will increase the likelihood of rejecting the null hypothesis when it is true. We, therefore, first test the assumption and then make a correction if the variances are unequal.

To test for equality of variances, we run the robust equal-variance test with the null hypothesis that the two variances are the same. Output 9.2 shows the results. In this example, we only need to interpret the p -value (labeled as Pr) at the end of the row labeled W0. Because the value is less than 0.05, we reject the null hypothesis that the variances are equal.

Once we have determined that the variances are equal or unequal, we then run the t test. In this example, we specify unequal variance either in the commands or in the dialog box if we are using menus. The results are illustrated in Output 9.3. In the first column, we see the average number of hours for men (8.3) and women (11.9) and the overall average number of hours (10.2). The difference in the average hours worked by men and women is listed as “diff” at the bottom of the “Mean” column. To test

 **OUTPUT 9.2: STATA OUTPUT FOR VARIANCE RATIO TEST**

```
. robvar rhhwork, by(sex)
```

Summary of how many hours a week does r spend on hh work			
Sex	Mean	Std. Dev.	Freq.
Male	8.3207547	9.4333454	583
Female	11.861472	12.702049	693
Total	10.24373	11.458679	1,276

W0 =	31.424330	df(1, 1274)	Pr > F = 0.00000003
W50 =	21.521522	df(1, 1274)	Pr > F = 0.00000386
W10 =	23.090941	df(1, 1274)	Pr > F = 0.00000173

When the null hypothesis is true (that the variances are equal), the probability of observing an F value at least as large as 31.42 is less than 0.05.

OUTPUT 9.3: STATA OUTPUT FOR TWO-SAMPLE t TEST WITH UNEQUAL VARIANCES

```
. ttest rhhwork, by(sex) unequal
```

Two-sample t test with unequal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
Male	583	8.320755	.3906892	9.433345	7.553422	9.088087
Female	693	11.86147	.4825109	12.70205	10.91411	12.80883
combined	1,276	10.24373	.3207814	11.45868	9.614413	10.87305
diff		-3.540717	.6208501		-4.758735	-2.322699

diff = mean(Male) - mean(Female) t = -5.7030
 Ho: diff = 0 Satterthwaite's degrees of freedom = 1255.28

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
 Pr(T < t) = 0.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 1.0000

The probability of observing a t statistic greater than 5.703 or less than -5.703 is less than 0.05.

whether this difference is statistically significant, we examine the t value and the significance level. According to the output, the probability of observing a t value greater than 5.7 or less than -5.7 is less than 5%. We therefore reject the null hypothesis and say that there is a statistically significant difference in the average hours spent on housework by men and women.

We can also examine the confidence interval from Output 9.3, which we learned how to generate in Chapter 8. Notice that the 95% confidence interval for the mean difference is from -4.76 to -2.32. This suggests that we are 95% confident that the true value of the difference is within that range.

In addition to examining the significance level of the difference in the two means, we may also want to examine the “effect size” or the magnitude of the difference between the two groups. Although there are several measures that can estimate the effect size, Cohen’s d is commonly used. It is calculated as the difference between two means divided by the pooled standard deviation for the two independent samples. The results are illustrated in Output 9.4. According to Cohen (1988), effect sizes are defined as small when $d = 0.2$, medium when $d = 0.5$, and large if $d = 0.8$. Since the absolute value of Cohen’s d in Output 9.4 is 0.3, it is between a small and medium effect.

✓ **OUTPUT 9.4: COHEN'S *d***

```
. esize twosample rhhwork, by(sex) cohensd unequal
```

Effect size based on mean comparison, unequal variances

Obs per group:
 Male = 886
 Female = 1,088

Effect Size	Estimate	[95% Conf. Interval]	
Cohen's <i>d</i>	-.3126059	-.4233679	-.2017208

Satterthwaite's degrees of freedom = **1.3e+03**

9.6 PRESENTING THE RESULTS

Presenting the results for a nontechnical audience

To present these results to a lay audience who may not be familiar with statistical tests, we could write the following:

On average, adults spend 10.24 hours per week on household chores. Our results show, however, that there is a statistically significant difference between the average hours spent on household chores per week by men and by women. Men spend 8.3 hours on average per week compared with women, who spend 11.9 hours, which is slightly more than a small difference in means.

Presenting the results in a scholarly journal

To present these results in a peer-reviewed scholarly journal, we would need to include more information. This could be written as follows:

To test the hypothesis that men and women spend the same amount of time on housework each week, we used a two independent-samples *t* test. The

results indicated that on average, men spend 8.32 hours per week ($SD = 9.43$), compared with women, who spend 11.86 hours ($SD = 12.7$). This was a statistically significant difference at the 0.05 level ($t(1,255) = -5.7, p < 0.001$). Examining the effect size or magnitude of the difference, Cohen's d revealed that the difference between the means is between a small and medium effect ($d = -0.31$).

9.7 SUMMARY OF COMMANDS USED IN THIS CHAPTER

As described in Chapter 4, this last section of each chapter summarizes all of the Stata code used in the chapter. In addition, all Stata code used throughout the book is summarized in Appendix 1.

Table

```
table sex, c(mean rhhwork) format(%3.2fc)
```

Test for equal variances

```
robvar rhhwork, by(sex)
```

Two independent-means test

```
ttest rhhwork, by(sex) unequal
```

Cohen's d effect size test

```
esize twosample rhhwork, by(sex) cohensd unequal
```

EXERCISES

1. You want to determine if men and women watch the same number of hours of television per week. Assume that the robust variance test determined that there was no statistically significant difference in the variances to answer this question.

	Mean Hours of TV Watched per Week	Sample Size	Variance
Men	14	20	100
Women	6	8	88

- a. Based on the information in the table, determine if this is a statistically significant difference. (Hint: The degrees of freedom would be equal to $n_1 + n_2 - 2$).
 - b. Use the information to calculate a 95% confidence interval of the mean difference.
2. Use the National Survey on Drug Use and Health 2015 to determine if there is a difference in age when men and women first try alcohol by following the instructions below.
- a. Determine if there is a significant difference in the average age when men and women (irsex) first try alcohol (alctry). For each command that you use, you will need to eliminate all observations above the age of 71 since there are observations with large numeric codes that represent bad data, individuals who never used alcohol, and individuals who didn't know or refused to answer. To do this, include the code, **if alctry < 72** at the end of each command line.
 - b. Use Cohen's *d* to examine the effect size, again using **if alctry < 72** at the end of the command line.
 - c. Write the results of your findings for a nontechnical audience.
 - d. Write the results of your findings for a journal article.
3. One of the arguments for school uniforms is that they will deter crime and increase student safety. We can explore this by using the School Survey on Crime and Safety data set from the 2015–2016 school year, which offers data on school characteristics, crimes, practices, and policies. The data represent 2,092 public schools in the United States. In particular, we can look at the total number of disciplinary actions required at schools that do and do not require uniforms. One question, however, is whether uniforms lead to fewer incidents (a negative relationship) or more incidents lead schools to require uniforms (a positive relationship). The possibility that two variables may influence each other makes it difficult to identify and measure the causal relationship, a problem called “endogeneity” that is discussed in Chapters 12 and 13.

Using the `pu_ssocs16` data set, generate a table that shows the average, the standard deviation, and the sample size of the total number of disciplinary actions required (`DISTOT16`) among schools that require and those that do

not require uniforms (C0134). Format the table so that there are two digits to the right of the decimal point.

- a. Determine if there is a significant difference in the average number of disciplinary actions required between schools that require and those that do not require uniforms.
- b. What is the null hypothesis?
- c. Can you reject the null hypothesis? Use statistics to support your conclusion.

REFERENCES

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Kolhatkar, S. (2015, June 26). *Women in the U.S. still do way more housework than men*. Retrieved from www.bloomberg.com/news/articles/2015-06-26/women-in-the-u-s-still-do-way-more-housework-than-men