MODULE 1

Checking the Representativeness of a Sample

DATA FILES FOR THIS MODULE

module1_county_weighted.sav
module1_disability_standard.sav

KEY LEARNING OBJECTIVES

The student will learn to

- use the weighted cases approach to conduct a one-sample chi-square test that determines how well the sample generalizes to the population
- use the standard approach to conduct a one-sample chi-square test that determines how well a subgroup of students that participates in a particular aspect of the study generalizes to all students in the sample.

A. Description of Researcher's Study

Dr. Porter is an educator and researcher who was asked to evaluate the first year of an initiative that provides support to students with Individual Education Plans (IEPs) in three rural counties. Support during the first year focused on students whose primary disability was either learning, mental, or emotional. Dr. Porter created a comprehensive evaluation plan involving quantitative and qualitative data analysis. Data were collected from school records, artifacts, questionnaires, and interviews.

The evaluation required that Dr. Porter travel across the three counties. Due to the time-intensive nature of collecting the data as well as the time and monetary expenses involved in visiting schools that are widespread across the rural region, it was not feasible to travel to all schools. Therefore, Dr. Porter used a cluster sampling approach to obtain his sample. Using this approach, the schools in each county were considered to be the "clusters." His first step was to randomly select half of all schools in each county, and then he contacted each school to seek permission from the administrators. He received permission to collect evaluation data from all but two schools. In these two cases, he randomly chose another school in the same county. Ultimately, the sample consisted of 7 of the 14 total schools in the three counties. In addition, within each school, parents or guardians of students in the initiative were asked to provide consent for their child to be interviewed. In all, 87% of parents/guardians in the sample agreed to allow their child to participate in this portion of the evaluation study.

This module describes how Dr. Porter examined the representativeness of his sample in terms of student distribution by county and by disability type using two forms of a one-sample chi-square analysis.

B. A Look at the Data

The entire population of students with IEPs who have a learning, mental, or emotional disability as their primary type was 463 across the three counties. Table 1.1 shows the total number of districts and schools in each county as well as the frequencies of students.

Table 1.1 Frequencies in the Population by County

	Districts (*n*)	**Schools (*n*)**	**Students (*n*)**
County 1	1	2	191
County 2	2	7	203
County 3	4	5	69
Total	7	14	463

Dr. Porter's sample contained 190 students across seven schools. This sample represents 41% of the total population of students with IEPs[1] and learning, mental, or emotional disabilities. Table 1.2 shows the number of districts, schools, and students in his sample by county.

Table 1.2 Frequencies in the Sample by County

	Districts (*n*)	Schools (*n*)	Students (*n*)
County 1	1	1	31
County 2	2	3	115
County 3	3	3	44
Total	6	7	190

In addition to the aggregated data, Dr. Porter also has a data file that includes the primary disability (learning, mental, emotional) for each of the 165 students in the sample whose parents provided permission for interviews. It also includes the county in which they reside. Codes for the two variables in the file are listed in Table 1.3.

Table 1.3 Variables in the Disability File

Type of Disability by County for the Sample	
Variable Name	**Description**
county	codes of 1, 2, and 3 for each county
disability	indicates type of disability for each student (1 = learning; 2 = mental; 3 = emotional)

The variable names Dr. Porter created for his two variables, *county* and *disability,* are succinct and adequate descriptors. In early versions of SPSS, variable names were limited to eight characters, but now they can be up to 64 characters in length. However, the best practice is to keep them as short and simple as possible. There are several rules for naming variables in SPSS. For example, each name must be unique, spaces are not allowed, and the first character cannot be a number or certain symbols such as % or &. An underscore, _, is useful for separating letters, numbers, or words in a variable name, but it is best to avoid ending a variable name with this symbol. A combination of upper- and lowercase letters may be used. For a complete description of the rules, go to the **Help menu** in SPSS and select **Topics**. Then, in the **Index tab,** type “variable names” and click on **rules**.

Researchers often develop their own naming conventions and try to follow them as closely as possible. For demographic or background variables, they may use one word that best describes the variable, and whenever that variable appears in another data file, they are consistent in assigning the same name. This practice is beneficial if a study involves merging multiple data files (see Section 5). Consistency is also useful for variables that represent responses on a questionnaire or scores on a test. Using names such as *q1, q2, q3,* and so on allow for easy identification of these variables.

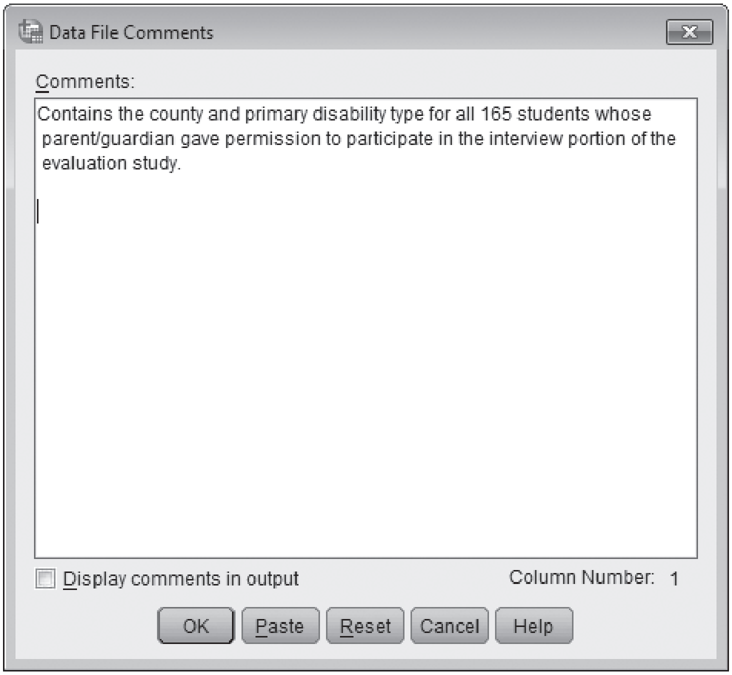
Another best practice is to document all pertinent information about the attributes of each variable. This is sometimes called a data dictionary. Brief descriptions for each variable, or at least for those variables needing additional explanation, should be created. In addition, numerical codes (e.g., 1, 2, and 3) for each categorical variable should be defined by creating value labels (a word or two) that describe each category. In SPSS, assigning variable descriptions and value labels can be performed in the **Variable View** window of the **Data Editor**. Other variable attributes in this window include data type (numeric, string, date, etc.); number of decimal places; user-defined missing values; and measurement level (scale, ordinal, nominal). Custom variable attributes can also be created. For further information on variable attributes and their default settings, go to the **Help menu** in SPSS and select **Topics**. Then, type "variable view" in the **Index tab** and select **Customizing Variable View**.

Some researchers choose to develop and maintain a data dictionary for their variables in a word processing program, especially if they need to document further information specific to their study. The format is typically a table containing the variable name, description, and categorical value labels, as well as other important information such as the source of the data (e.g., existing data from a school district's database versus data obtained from questionnaires administered as part of the study) and when the data were obtained (month and year). Creating and maintaining a data dictionary is not only helpful when conducting analysis for the current study, it is also beneficial if the researcher needs to use, or refer to, the data file in the future.

The file is available on the website and is called *module1_disability_standard.sav*. The naming convention for all data files in the book is to begin with the word "module" and its number (e.g., module 1) followed by a descriptive word or two about the file. The descriptors are separated by an underscore, _. For example, the file *module1_disability_standard.sav* indicates that the file is for use with module 1, it contains disability information, and it's used to illustrate the standard approach to conduct a one-sample chi-square test. Although there are no specific rules with regard to naming files, it is important to create your own naming conventions for data files that are used within a research project. Keep the file names as short as possible, and be diligent about sticking to the conventions you created.

It is also useful to write a brief description of each data file. This is possible in SPSS by selecting **Data File Comments** under the **Utilities menu**. The dialog box in Figure 1.1 will open, and comments can be typed in the **Comments** box. After **OK** is clicked, the comments will be saved with the file. At any point, comments can be viewed by opening the dialog box or displaying them in the output file by selecting **Display comments in output**.

Figure 1.1 Utilities → Data File Comments



Contains the county and primary disability type for all 165 students whose parent/guardian gave permission to participate in the interview portion of the evaluation study.

C. Planning and Decision Making

Working with the data available to him, Dr. Porter decided to check the representativeness of his sample in two ways. ****Dr. Porter(1)** First, he wanted to know how well the sample he obtained using the cluster method reflected the population with regard to student distribution by county. Using the frequencies in Tables 1.1 and 1.2, he will need to create a data file and then conduct a one-sample chi-square test. ****Dr. Porter(2)** This type of statistical test allows for examining whether the proportions of the sample students in each county are similar to, or significantly different from, the hypothesized proportions. In his situation, the hypothesized proportions are the population proportions of students in each county.

Dr. Porter also wanted to determine whether the group of students whose parents provided permission for interviews was representative of all students in the sample in terms of disability type. The participation rate for the interview collection data was quite high (87%), but he wanted to have statistical evidence that would support representativeness. Therefore, Dr. Porter will conduct another one-sample chi-square test using the *module1_disability_standard.sav* data file.

In order to conduct a one-sample chi-square test in SPSS, data must be represented in one of two ways. The most common is the standard method, which requires that the number of cases in the data file is equal to the number of subjects. In other words, there is a row for each student and a variable that contains a value or code for each student. The second way to conduct a one-sample chi-square test does not require a student-level data file. Rather, only the total frequencies in each category of the variable are necessary. This approach is called the weighted cases method. The data file is structured so that the total number of rows equals the total number of categories in the variable of interest. Typically, the file contains two variables. One variable lists the codes for each category (e.g., 1, 2, 3), and the other variable indicates the observed total frequencies for each category.

**1 Dr. Porter says: "I had access to certain variables for my sample but not the entire population. For example, I knew the disability types for all sample students but not for all students in the population. In addition, some of my data were in aggregate form (e.g., the frequencies in Table 1.1 for the population), and other data were at the student level (e.g., the data file containing the primary disability type for sample students). Therefore, the data available to me and [their] format guided my planning and approach to evaluating the representativeness of the sample."

**2 Dr. Porter says: "Some people refer to this statistical test as a chi-square goodness-of-fit test because it reflects whether empirical distributions match (or 'fit') theoretical distributions. I prefer to call it the one-sample chi-square test because it examines frequencies/proportions for one variable. There is another type of chi-square test available to examine frequencies/proportions across categories for two variables. It is called the two-sample chi-square test or the chi-square test of association because it evaluates whether the two variables are correlated or associated with each other."

To summarize, Dr. Porter will use two types of chi-square analysis. The first analysis involves the weighted cases method to determine the extent to which his sample (*n* = 190) generalizes to the population (*N* = 463) in terms of the distribution of students across the three counties. The second chi-square analysis uses the standard method to determine the extent to which the distribution of disability types for the 165 students whose parents/guardians provided permission for interviews generalizes to the distribution of disability types for all 190 students in the sample. ****Dr. Porter(3)**

**3 Dr. Porter says: "For the first analysis, I could use the frequencies in the county categories to create a file that has the same number of rows as students (*n* = 190) so that the standard method for the chi-square test could be conducted. However, I chose to use the weighted cases method because it is more efficient. It is not necessary to spend time entering data for 190 students into a data file, and data entry error is also reduced."

D. Using SPSS to Address Issues and Prepare Data △

First Chi-Square Analysis—Weighted Approach

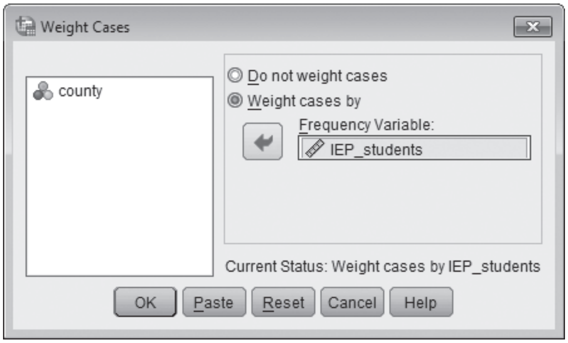
To examine how well the sample generalizes to the population in terms of distribution of students across the three counties, Dr. Porter used the weighted cases approach to conducting the one-sample chi-square test.

His first step was to create a file with 3 cases (rows) and 2 variables. The file is on the website and is called *module1_county_weighted.sav*. The variables are *county* and *IEP_students*. ****SPSS Tip 1** Values for *county* represent the county numerical codes. Values for *IEP_students* represent the frequency of students in each county in the sample obtained from Table 1.2. In order to conduct the chi-square test using the weighted cases approach, SPSS requires that the frequencies for each category be listed in the data file from lowest to highest. Therefore, Dr. Porter's file contains *county* values of 1, 3, and 2, and *IEP_students* values of 31, 44, and 115. Not following this rule will lead to an inaccurate chi-square result.

SPSS Tip 1: The *county* variable is not absolutely necessary in the file because SPSS needs only the frequency variable (*IEP_students*) to calculate the chi-square test. However, placing *county* in the file will make it easier for Dr. Porter to associate the frequencies with their corresponding county codes.

Now, Dr. Porter needs to weight the three cases in his file so that SPSS will know that, for example, 31 is not a value for one student, but instead represents 31 students in County 1. Under the **Data menu**, Dr. Porter selected **Weight Cases** and obtained the SPSS dialog box shown in Figure 1.2. He

Figure 1.2 Data → Weight Cases



Weight Cases

county

○ Do not weight cases
● Weight cases by

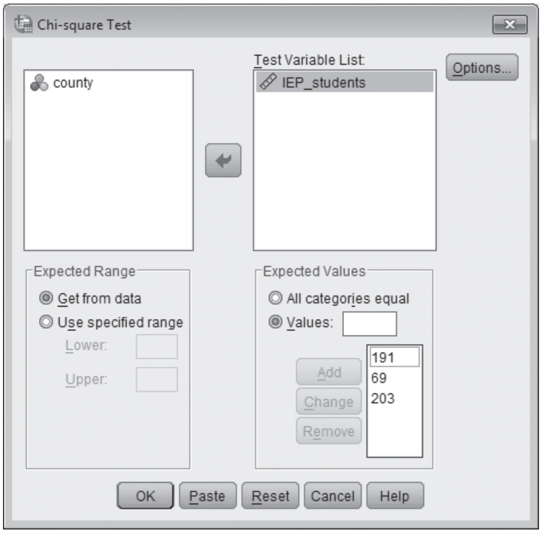
Frequency Variable: IEP_students

Current Status: Weight cases by IEP_students

selected the **Weight cases by** button and placed the variable containing the frequencies of students (*IEP_students*) in the **Frequency Variable** box and clicked **OK**. This essentially tells SPSS that there are actually 31 students in County 1, 44 students in County 3, and 115 students in County 2.

Dr. Porter is now ready to conduct the one-sample chi-square test. Under the **Analyze menu**, he selected **Nonparametric Tests**, **Legacy Dialogs**, and **Chi-square** in order to get the dialog box shown in Figure 1.3. The frequency variable of interest, *IEP_students*, is placed in the **Test**

Figure 1.3 Analyze → Nonparametric Tests → Legacy Dialogs → Chi-square



Chi-square Test

county

Test Variable List: IEP_students

Expected Range
● Get from data
○ Use specified range
Lower:
Upper:

Expected Values
○ All categories equal
● Values:

191
69
203

Variable List box. The *county* variable is not placed in the box. Recall that the only purpose of this variable is to help Dr. Porter remember which frequencies are associated with each county.

In the **Expected Values** box, the population frequencies for each county (from Table 1.1) must be entered by selecting the **Values** button, typing in the frequency, and clicking **Add.** If an error is made, the **Remove** or **Change** buttons can be used. It is extremely important to remember that the order of the counties was from the smallest sample (observed frequency) to the largest sample (observed frequency) (i.e., counties 1, 3, and 2). Therefore, the population frequencies must be entered in the same order. ****SPSS Tip 2**

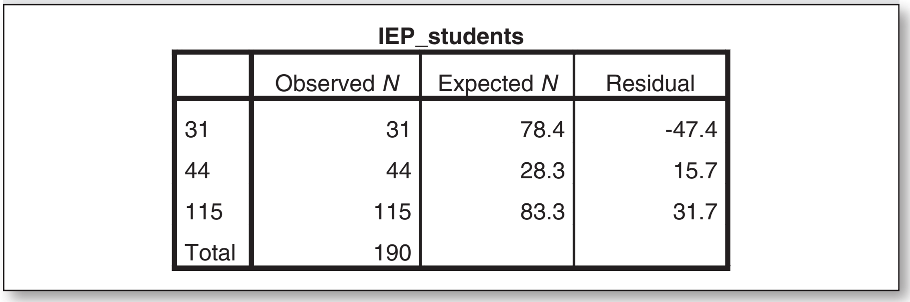
SPSS Tip 2: The expected values could also be equal. For example, suppose a researcher needed to sample an equal number of participants from each category of a variable. To examine whether his final sample statistically had the same frequency of participants across categories, he would simply select "All categories equal" and SPSS would calculate the Expected Values to be the total number of cases divided by the total number of categories.

There is also an **Expected Range** box. This allows you to select only some of the categories to include in the chi-square analysis. Because Dr. Porter's analysis incorporated all three categories of his county variable, he did not need to use this box. He used the default option (**Get from data**), which means that all categories will be used in the analysis. ****SPSS Tip 3**

SPSS Tip 3: An example of when the Expected Range box might be helpful is if an ethnicity variable contains five categories but the analysis only focuses on the first three categories because they contain the majority of total cases. In this situation, you would select "Use specified range," then type "1" for the Lower code in the range of the ethnicity variable and "3" for the Upper code. Note that the codes for the variable would need to be consecutively ordered in the data file for this technique to work.

After he clicked **OK**, the output showed two tables. In Table 1.4, the

Table 1.4 SPSS Output Indicating Observed and Expected Values for Students by County



IEP_students

	Observed *N*	Expected *N*	Residual
31	31	78.4	-47.4
44	44	28.3	15.7
115	115	83.3	31.7
Total	190		

Observed N values show the sample frequencies of students with IEPs in County 1, 3, and 2. The next column indicates the expected values of the sample based on the population frequencies of students by county. ****SPSS Tip 4**

SPSS Tip 4: The Expected *N* is calculated using the proportion of the student population in each county. For example, 191 out of 463 students in the population were in County 1 (proportion = .4125). If the sample proportion was equal to the population proportion, then .4125 of the total of 190 sample students (expected *N* = 78.4) in the sample would be in County 1. However, there were only 31 sample students (observed *N*) in County 1, which leads to a residual of -47.4. The actual sample proportion was .1632 (31 divided by 190), which was much smaller than the population proportion. Expected *N* and residual values for Counties 3 and 2 in Table 1.4 are calculated in the same way.

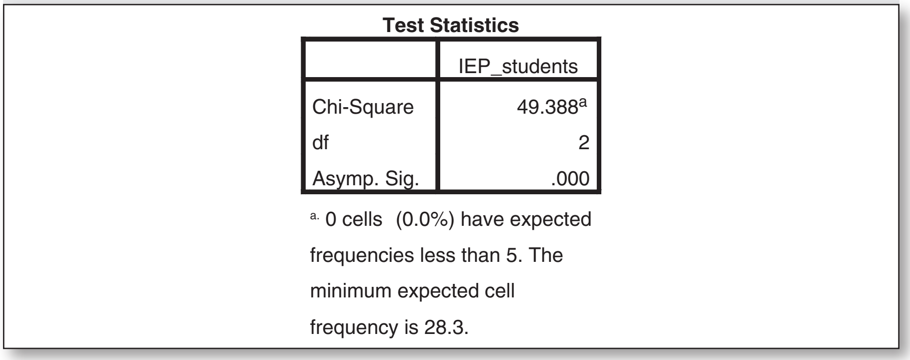
The second table of output presents the statistical results from the chi-square test. ****SPSS Tip 5** Table 1.5 shows a chi-square value of 49.388 with a *p* value that is less than .001. There are two degrees of freedom for this test because three categories were examined (*df* = number of categories minus 1). The null hypothesis that the proportions in the sample are equal to the proportions in the population is rejected. Therefore, the sample proportions are significantly different from the population proportions across the three counties. In other words, the distribution of students across the three counties for the sample is not the same as the distribution in the population. This conclusion brings into question the generalizability of the results for the evaluation study.

SPSS Tip 5: If you did not follow the SPSS rule to order the frequencies from lowest to highest in the data file, you will not obtain these results. The incorrect order of 31, 115, 44 (representing Counties 1, 2, and 3, respectively) will produce a chi-square result that is incorrect [χ^2 (*df* = 2) = 312.562, *p* < .001].

Second Chi-Square Analysis—Standard Approach

Dr. Porter's next analysis was to examine how well the distribution of disability types for students with permission to be interviewed reflects the distribution of disability types for all sample students. The majority of parents/guardians allowed their child to be interviewed (87%). Based on the aggregated data that was available to Dr. Porter, he knew that of the 190 total students in the sample, the primary disability was documented as learning for 67% of them, mental for 19%, and emotional for the remaining 14% of students. Of the 165 students in the sample who had permission to be interviewed, percentages were somewhat similar (64% for learning, 24% for mental, and 12% for emotional disabilities). However, Dr. Porter wanted to have statistical

Table 1.5 SPSS Output Indicating the Chi-Square Test Statistic for Students by County



Test Statistics

	IEP_students
Chi-Square	49.388[a]
df	2
Asymp. Sig.	.000

[a] 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 28.3.

evidence to support his conclusion, so he conducted the one-sample chi-square test.

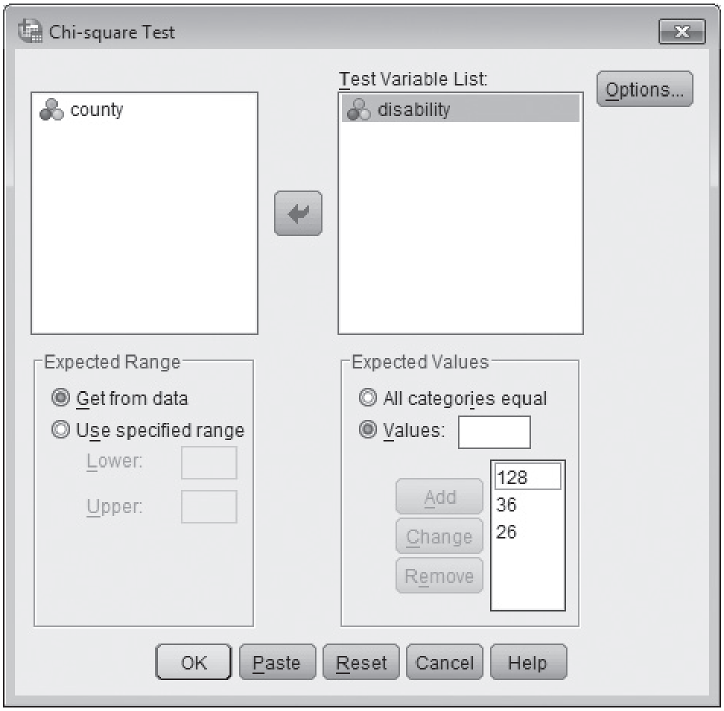
As mentioned earlier, Dr. Porter had a data file (*module1_disability_standard.sav*) containing the disability type for each student whose parent(s) agreed to participation. As described in Table 1.3, the codes of 1, 2, and 3 for *disability* represented learning, mental, and emotional disabilities. Because the file contains the same number of cases (rows) as students, he can use the standard approach when conducting the chi-square analysis. That is, he does not need to weight cases by the *disability* variable.

He begins by opening the same dialog box shown in Figure 1.4 under the **Analyze menu** as in the first analysis. The *disability* variable is moved to the **Test Variable List** box. The *county* variable is not the variable of interest in this analysis. Then, Dr. Porter selected the **Values** button in the **Expected Values** box in order to provide SPSS with the expected frequencies in each disability type across all 190 students in the sample. He typed in the value of 128 for learning disability, the first category in the *disability* variable, and clicked the **Add** button. He entered the expected values for the mental and emotional categories in the same manner (36 and 26, respectively). ****SPSS Tip 6**

Once again, two tables of output are produced. Table 1.6 indicates that the observed and expected values for each category are similar, and the residuals are small. Results from the statistical test

SPSS Tip 6: Unlike the weighted cases approach to conducting the chi-square test, the standard approach requires that the expected values match the numerical order of the categorical codes for the variable (1, 2, and 3 in this study, which represent learning, mental, and emotional disabilities). The expected values should *not* be ordered from smallest to largest frequency.

Figure 1.4 Analyze → Nonparametric Tests → Legacy Dialogs → Chi-square



Chi-square Test

county

Test Variable List:

disability

Expected Range

● Get from data

○ Use specified range

Lower:

Upper:

Expected Values

○ All categories equal

● Values:

128
36
26

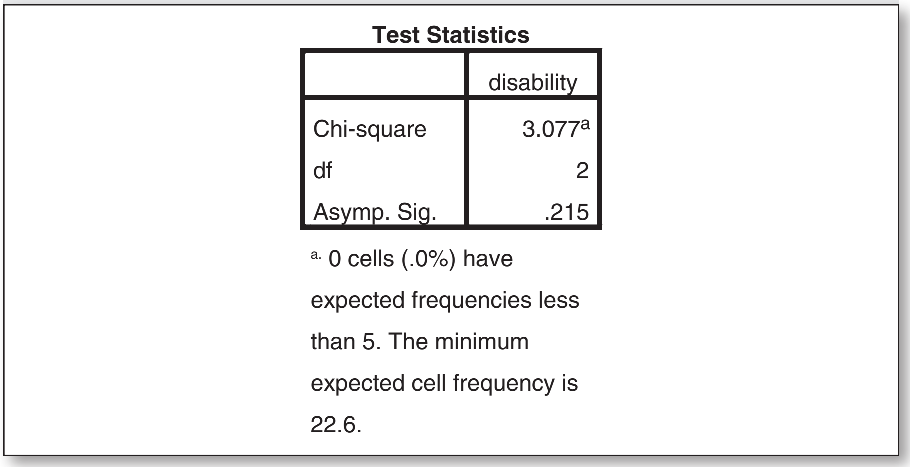
Table 1.6 SPSS Output Indicating Observed and Expected Values for Students by Disability

disability

	Observed *N*	Expected *N*	Residual
1 learning disability	105	111.2	-6.2
2 mental	40	31.3	8.7
3 emotional	20	22.6	-2.6
Total	165		

shown in Table 1.7 confirm Dr. Porter's initial thoughts about the data. The *p* value of .215 indicates that the null hypothesis of equality of the two sets of proportions is not rejected. Therefore, the distribution of disability type

Table 1.7 SPSS Output Indicating the Chi-Square Test Statistic for Students by Disability



Test Statistics

	disability
Chi-square	3.077[a]
df	2
Asymp. Sig.	.215

[a] 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 22.6.

for students who were permitted to participate in interviews is similar to the distribution for all students in the sample. This outcome indicates that the group of students to be interviewed generalizes well to all students in the sample with respect to the type of primary disability.

E. Reflection and Additional Decision Making

Dr. Porter was pleased that the group of students participating in the interview portion of the evaluation study was representative of the sample in terms of disability type. However, he was interested in further exploring data for the first chi-square analysis. He wondered if County 1 was the reason the sample distribution of students by county was not similar to the population distribution. The absolute value of its residual (observed minus expected value) was higher than the residuals for Counties 2 and 3. He decided to conduct three follow-up chi-square tests using the weighted approach, one test for each possible pair of counties, using the same procedures as described in Section D. A summary of results is presented in Table 1.8. The chi-square test results were obtained from the SPSS output. The population and sample proportions were calculated as described in SPSS Tip 4.

The first two tests included County 1 data. Results showed statistically significant chi-square values, and the sample proportions were quite different from the population proportions. Conversely, the third test of Counties

Table 1.8 Summary of Chi-Square Results for Each Unique Pair of Counties With Sample and Population Proportions

<table>
<tr><th colspan="3">Test 1: [χ^2 (<i>df</i> = 1) = 39.710, <i>p</i> &lt; .001]</th></tr>
<tr><th>County</th><th>Sample Proportions</th><th>Population Proportions</th></tr>
<tr><td>1</td><td>.41</td><td>.73</td></tr>
<tr><td>3</td><td>.59</td><td>.27</td></tr>
<tr><th colspan="3">Test 2: [χ^2 (<i>df</i> = 1) = 43.388, <i>p</i> &lt; .001]</th></tr>
<tr><th>County</th><th>Sample Proportions</th><th>Population Proportions</th></tr>
<tr><td>1</td><td>.21</td><td>.48</td></tr>
<tr><td>2</td><td>.79</td><td>.52</td></tr>
<tr><th colspan="3">Test 3: [χ^2 (<i>df</i> = 1) = .446, <i>p</i> = .504]</th></tr>
<tr><th>County</th><th>Sample Proportions</th><th>Population Proportions</th></tr>
<tr><td>3</td><td>.28</td><td>.25</td></tr>
<tr><td>2</td><td>.72</td><td>.75</td></tr>
</table>

2 and 3 did not produce a significant chi-square result, and the sample versus the population proportions were similar. Thus, County 1 appears to be the reason for lack of representation of the sample to the population. Only two schools were in the county. One school had a large population of students (N = 160), the other had a small population (N = 31). The large school was originally selected to be in the sample, but it declined participation in the study. The school with the small population agreed to participate. Because there was such a large difference in the number of students between the two schools in the county, regardless of which school was selected, the values would have greatly impacted the overall chi-square statistic. Unfortunately, this is one of the drawbacks of the cluster sampling method when there are only a small number of clusters/schools in each area. There is nothing Dr. Porter can do to alter his data, but at least his analysis allows him to identify the issue and acknowledge it when writing reports for his evaluation study.

F. Writing It Up

Dr. Porter wrote the following paragraphs to describe his sampling procedure, his approach to investigating the sample's representativeness, and the findings regarding the sample.

"A cluster sampling approach was used to randomly select approximately half of the schools in each county. The administrative office in each of the selected schools was contacted by phone and letter to request their participation in the evaluation study. Five of the seven schools agreed to participate, two schools declined. In these two cases, another school from the same county was randomly selected to be in the study. Both schools in the second round agreed to participate.

In order to determine how well the sample (n = 190) generalized to the population (n = 463) in terms of distribution of students across the three counties, a one-sample chi-square test was conducted. Results indicated that the sample proportions of students by county were significantly different from the population proportions by county (χ^2 (df = 2) = 49.388, p < .001). Overall, the sample did not represent the population. Upon further examination of the data using follow-up chi-square tests on each pair of counties, County 1 appeared to be the reason for the lack of generalization from sample to population. There were only two schools in the county. The school that participated was smaller than the school that was originally selected and declined participation (31 versus 160 students). Thus, the sample proportion in County 1 was quite different than the population proportion. The two follow-up chi-square tests that included this county produced significant results (p < .001). The result from the follow-up chi-square test for Counties 2 and 3 was not significant (p = .504). Sample proportions of students in these two counties were similar to the population proportions, indicating that the sample represented the population for Counties 2 and 3.

Of the 190 total students in the sample, 87% of the parents/guardians allowed their children to participate in the interview portion of data collection. To check whether the distribution of disability types for the 165 students to be interviewed generalized to the distribution of disability types for all students in the sample, another one-sample chi-square test was conducted. The statistical evidence indicates that the students to be interviewed are representative of all sample students in terms of disability type (χ^2 (df = 2) = 3.077, p = .215). In the full sample, the primary disability was documented to be learning for 67% of the students, mental for 19%, and emotional for 14%. Of the students to be interviewed, the percentages were 64%, 24%, and 12%, respectively."

Reflective Questions △

- Which of the two approaches to conducting a one-sample chi-square test would you use if you had a file that contained all the necessary data by individual participants? What are the processes for carrying out this approach?

- Which of the two approaches to conducting a one-sample chi-square test would you use if you had only aggregated data who represented sample and population frequencies for the variables of interest? What are the processes for carrying out this approach?

Extensions

- Using other research examples, discuss how well the sample generalizes to a targeted population. If possible, create the necessary data files and conduct a one-sample chi-square analysis to determine the generalizability of the sample.
- Find research articles that mention a sample's generalizability, then discuss how the authors determined if results from their sample generalized to the target population.
- Discuss the unique aspects of the cluster sampling approach in comparison to other methods for obtaining a sample. Compare the advantages and disadvantages of using this approach.

Note

[1] Throughout the rest of this module, the use of the word "students" refers to students with IEPs that have a learning, mental, or emotional disability documented as their primary disability.