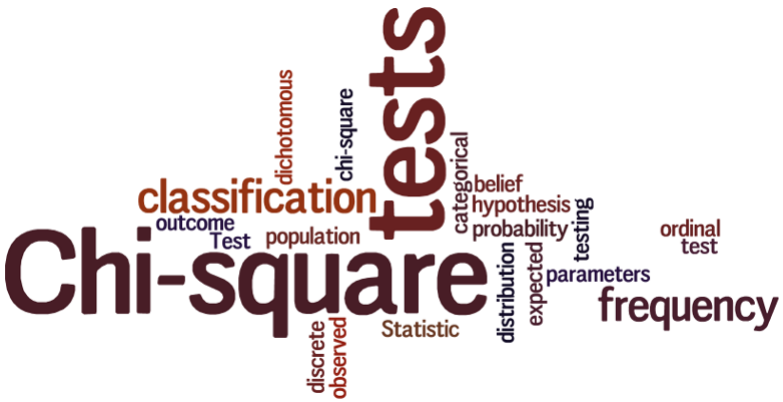
Hypothesis Testing - Chi Squared Test

Author:

Lisa Sullivan, PhD

Professor of Biostatistics

Boston University School of Public Health



Introduction

This module will continue the discussion of hypothesis testing, where a specific statement or hypothesis is generated about a population parameter, and sample statistics are used to assess the likelihood that the hypothesis is true. The hypothesis is based on available information and the investigator's belief about the population parameters. The specific tests considered here are called chi-square tests and are appropriate when the outcome is discrete (dichotomous, ordinal or categorical). For example, in some clinical trials the outcome is a classification such as hypertensive, pre-hypertensive or normotensive. We could use the same classification in an observational study such as the Framingham Heart Study to compare men and women in terms of their blood pressure status - again using the classification of hypertensive, pre-hypertensive or normotensive status.

The technique to analyze a discrete outcome uses what is called a chi-square test. Specifically, the test statistic follows a chi-square probability distribution. We will consider chi-square tests here with one, two and more than two independent comparison groups.

Learning Objectives

After completing this module, the student will be able to:

1. Perform chi-square tests by hand

2. Appropriately interpret results of chi-square tests
3. Identify the appropriate hypothesis testing procedure based on type of outcome variable and number of samples

Boston University School of Public Health

Tests with One Sample, Discrete Outcome

Here we consider hypothesis testing with a discrete outcome variable in a single population. Discrete variables are variables that take on more than two distinct responses or categories and the responses can be ordered or unordered (i.e., the outcome can be ordinal or categorical). The procedure we describe here can be used for dichotomous (exactly 2 response options), ordinal or categorical discrete outcomes and the objective is to compare the distribution of responses, or the proportions of participants in each response category, to a known distribution. The known distribution is derived from another study or report and it is again important in setting up the hypotheses that the comparator distribution specified in the null hypothesis is a fair comparison. The comparator is sometimes called an external or a historical control.

In one sample tests for a discrete outcome, we set up our hypotheses against an appropriate comparator. We select a sample and compute descriptive statistics on the sample data. Specifically, we compute the sample size (n) and the proportions of participants in each response

category ($\hat{p}_1$, $\hat{p}_2$, ... $\hat{p}_k$) where k represents the number of response categories. We then determine the appropriate test statistic for the hypothesis test. The formula for the test statistic is given below.

Test Statistic for Testing H_0: $p_1 = p_{10}$, $p_2 = p_{20}$, ..., $p_k = p_{k0}$

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

We find the critical value in a table of probabilities for the chi-square distribution with degrees of freedom (df) = k-1. In the test statistic, O = observed frequency and E=expected frequency in each of the response categories. The observed frequencies are those observed in the sample and the expected frequencies are computed as described below. χ^2 (chi-square) is another probability distribution and ranges from 0 to ∞. The test above statistic formula above is appropriate for large samples, defined as expected frequencies of at least 5 in each of the response categories.

When we conduct a χ^2 test, we compare the observed frequencies in each response category to the frequencies we would expect if the null hypothesis were true. These expected frequencies are determined by allocating the sample to the response categories according to the distribution specified in H_0. This is done by multiplying the observed sample size (n) by the proportions specified in the null hypothesis (p_{10} , p_{20} , ..., p_{k0}). To ensure that the sample size is appropriate for the use of the test statistic above, we need to ensure that the following: min(np_{10} , $n p_{20}$, ..., $n p_{k0}$) ≥ 5.

The test of hypothesis with a discrete outcome measured in a single sample, where the goal is to assess whether the distribution of responses follows a known distribution, is called the χ^2 goodness-of-fit test. As the name indicates, the idea is to assess whether the pattern or distribution of responses in the

sample "fits" a specified population (external or historical) distribution. In the next example we illustrate the test. As we work through the example, we provide additional details related to the use of this new test statistic.

Example:

A University conducted a survey of its recent graduates to collect demographic and health information for future planning purposes as well as to assess students' satisfaction with their undergraduate experiences. The survey revealed that a substantial proportion of students were not engaging in regular exercise, many felt their nutrition was poor and a substantial number were smoking. In response to a question on regular exercise, 60% of all graduates reported getting no regular exercise, 25% reported exercising sporadically and 15% reported exercising regularly as undergraduates. The next year the University launched a health promotion campaign on campus in an attempt to increase health behaviors among undergraduates. The program included modules on exercise, nutrition and smoking cessation. To evaluate the impact of the program, the University again surveyed graduates and asked the same questions. The survey was completed by 470 graduates and the following data were collected on the exercise question:

	No Regular Exercise	Sporadic Exercise	Regular Exercise	Total
Number of Students	255	125	90	470

Based on the data, is there evidence of a shift in the distribution of responses to the exercise question following the implementation of the health promotion campaign on campus? Run the test at a 5% level of significance.

In this example, we have one sample and a discrete (ordinal) outcome variable (with three response options). We specifically want to compare the distribution of responses in the sample to the distribution reported the previous year (i.e., 60%, 25%, 15% reporting no, sporadic and regular exercise, respectively). We now run the test using the five-step approach.

- **Step 1.** Set up hypotheses and determine level of significance.

The null hypothesis again represents the "no change" or "no difference" situation. If the health promotion campaign has no impact then we expect the distribution of responses to the exercise question to be the same as that measured prior to the implementation of the program.

H_0: p_1=0.60, p_2=0.25, p_3=0.15, or equivalently H_0: Distribution of responses is 0.60, 0.25, 0.15

H_1: H_0 is false. α =0.05

Notice that the research hypothesis is written in words rather than in symbols. The research hypothesis as stated captures any difference in the distribution of responses from that specified in the null hypothesis. We do not specify a specific alternative distribution, instead we are testing whether the sample data "fit" the distribution in H_0 or not. With the χ^2 goodness-of-fit test there is no upper or lower tailed version of the test.

- **Step 2.** Select the appropriate test statistic.

The test statistic is:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

We must first assess whether the sample size is adequate. Specifically, we need to check min(np_0, np_1, ..., n p_k) $\geq$ 5. The sample size here is n=470 and the proportions specified in the null hypothesis are 0.60, 0.25 and 0.15. Thus, min(470(0.65), 470(0.25), 470(0.15))=min(282, 117.5, 70.5)=70.5. The sample size is more than adequate so the formula can be used.

- **Step 3.** Set up decision rule.

The decision rule for the χ^2 test depends on the level of significance and the degrees of freedom, defined as degrees of freedom (df) = k-1 (where k is the number of response categories). If the null hypothesis is true, the observed and expected frequencies will be close in value and the χ^2 statistic will be close to zero. If the null hypothesis is false, then the χ^2 statistic will be large. Critical values can be found in a table of probabilities for the χ^2 distribution. Here we have df=k-1=3-1=2 and a 5% level of significance. The appropriate critical value is 5.99, and the decision rule is as follows: Reject H_0 if $\chi^2 \geq$ 5.99.

- **Step 4.** Compute the test statistic.

We now compute the expected frequencies using the sample size and the proportions specified in the null hypothesis. We then substitute the sample data (observed frequencies) and the expected frequencies into the formula for the test statistic identified in Step 2. The computations can be organized as follows.

	No Regular Exercise	Sporadic Exercise	Regular Exercise	Total
Observed Frequencies (O)	255	125	90	470
Expected Frequencies (E)	470(0.60) =282	470(0.25) =117.5	470(0.15) =70.5	470

Notice that the expected frequencies are taken to one decimal place and that the sum of the observed frequencies is equal to the sum of the expected frequencies. The test statistic is computed as follows:

$$\chi^2 = \frac{(256-282)^2}{282} + \frac{(125-117.5)^2}{117.5} + \frac{(90-70.5)^2}{70.5}$$

$$\chi^2 = 2.59 + 0.48 + 5.39 = 8.46$$

- **Step 5.** Conclusion.

We reject H_0 because 8.46 $\geq$ 5.99. We have statistically significant evidence at α=0.05 to show that H_0 is false, or that the distribution of responses is not 0.60, 0.25, 0.15. The p-value is p < 0.005.

In the χ^2 goodness-of-fit test, we conclude that either the distribution specified in H_0 is false (when we reject H_0) or that we do not have sufficient evidence to show that the distribution specified in H_0 is false (when we fail to reject H_0). Here, we reject H_0 and concluded that the distribution of responses to the exercise question following the implementation of the health promotion campaign was not the same as the distribution prior. The test itself does not provide details of how the distribution has shifted. A comparison of the observed and expected frequencies will provide some insight into the shift (when the null hypothesis is rejected). Does it appear that the health promotion campaign was effective?

Consider the following:

	No Regular Exercise	Sporadic Exercise	Regular Exercise	Total
Observed Frequencies (O)	255	125	90	470
Expected Frequencies (E)	282	117.5	70.5	470

If the null hypothesis were true (i.e., no change from the prior year) we would have expected more students to fall in the "No Regular Exercise" category and fewer in the "Regular Exercise" categories. In the sample, 255/470 = 54% reported no regular exercise and 90/470=19% reported regular exercise. Thus, there is a shift toward more regular exercise following the implementation of the health promotion campaign. There is evidence of a statistical difference, is this a meaningful difference? Is there room for improvement?

Example:

The National Center for Health Statistics (NCHS) provided data on the distribution of weight (in categories) among Americans in 2002. The distribution was based on specific values of body mass index (BMI) computed as weight in kilograms over height in meters squared. Underweight was defined as BMI< 18.5, Normal weight as BMI between 18.5 and 24.9, overweight as BMI between 25 and 29.9 and obese as BMI of 30 or greater. Americans in 2002 were distributed as follows: 2% Underweight, 39% Normal Weight, 36% Overweight, and 23% Obese. Suppose we want to assess whether the distribution of BMI is different in the Framingham Offspring sample. Using data from the n=3,326 participants who attended the seventh examination of the Offspring in the Framingham Heart Study we created the BMI categories as defined and observed the following:

	Underweight BMI<18.5	Normal Weight BMI 18.5-24.9	Overweight BMI 25.0-29.9	Obese BMI $\geq$ 30	Total
# of Participants	20	932	1374	1000	3326

- **Step 1.** Set up hypotheses and determine level of significance.

  H_0: $p_1=0.02$, $p_2=0.39$, $p_3=0.36$, $p_4=0.23$ or equivalently

  H_0: Distribution of responses is 0.02, 0.39, 0.36, 0.23

  H_1: H_0 is false. $\alpha=0.05$

- **Step 2.** Select the appropriate test statistic.

The formula for the test statistic is:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

We must assess whether the sample size is adequate. Specifically, we need to check min(np_0, np_1, ..., n p_k) $\geq$ 5. The sample size here is n=3,326 and the proportions specified in the null hypothesis are 0.02, 0.39, 0.36 and 0.23. Thus, min(3326(0.02), 3326(0.39), 3326(0.36), 3326(0.23))=min(66.5, 1297.1, 1197.4, 765.0)=66.5. The sample size is more than adequate, so the formula can be used.

- **Step 3.** Set up decision rule.

Here we have df=k-1=4-1=3 and a 5% level of significance. The appropriate critical value is 7.81 and the decision rule is as follows: Reject H_0 if $\chi^2 \geq 7.81$.

- **Step 4.** Compute the test statistic.

We now compute the expected frequencies using the sample size and the proportions specified in the null hypothesis. We then substitute the sample data (observed frequencies) into the formula for the test statistic identified in Step 2. We organize the computations in the following table.

	Underweight BMI<18.5	Normal BMI 18.5-24.9	Overweight BMI 25.0-29.9	Obese BMI $\geq$ 30	Total
Observed Frequencies (O)	20	932	1374	1000	3326
Expected Frequencies (E)	66.5	1297.1	1197.4	765.0	3326

The test statistic is computed as follows:

$$\chi^2 = \frac{(20-66.5)^2}{66.5} + \frac{(932-1297.4)^2}{1297.4} + \frac{(1374-1197.4)^2}{1197.4} + \frac{(1000-765.0)^2}{765.0}$$

$$\chi^2 = 32.52 + 102.77 + 26.05 + 72.19 = 233.53$$

- **Step 5.** Conclusion.

We reject H_0 because 233.53 $\geq$ 7.81. We have statistically significant evidence at α=0.05 to show that H_0 is false or that the distribution of BMI in Framingham is different from the national data reported in 2002, p < 0.005.

Again, the χ^2 goodness-of-fit test allows us to assess whether the distribution of responses "fits" a specified distribution. Here we show that the distribution of BMI in the Framingham Offspring Study is different from the national distribution. To understand the nature of the difference we can

compare observed and expected frequencies or observed and expected proportions (or percentages). The frequencies are large because of the large sample size, the observed percentages of patients in the Framingham sample are as follows: 0.6% underweight, 28% normal weight, 41% overweight and 30% obese. In the Framingham Offspring sample there are higher percentages of overweight and obese persons (41% and 30% in Framingham as compared to 36% and 23% in the national data), and lower proportions of underweight and normal weight persons (0.6% and 28% in Framingham as compared to 2% and 39% in the national data). Are these meaningful differences?

In the module on hypothesis testing for means and proportions, we discussed hypothesis testing applications with a dichotomous outcome variable in a single population. We presented a test using a test statistic Z to test whether an observed (sample) proportion differed significantly from a historical or external comparator. The chi-square goodness-of-fit test can also be used with a dichotomous outcome and the results are mathematically equivalent.

In the prior module, we considered the following example. Here we show the equivalence to the chi-square goodness-of-fit test.

Example:

The NCHS report indicated that in 2002, 75% of children aged 2 to 17 saw a dentist in the past year. An investigator wants to assess whether use of dental services is similar in children living in the city of Boston. A sample of 125 children aged 2 to 17 living in Boston are surveyed and 64 reported seeing a dentist over the past 12 months. Is there a significant difference in use of dental services between children living in Boston and the national data?

We presented the following approach to the test using a Z statistic.

- **Step 1.** Set up hypotheses and determine level of significance

  H_0: p = 0.75

  H_1: p ≠ 0.75 α=0.05

- **Step 2.** Select the appropriate test statistic.

We must first check that the sample size is adequate. Specifically, we need to check min(np_0, $n(1-p_0)$) = min(125(0.75), 125(1-0.75))=min(94, 31)=31. The sample size is more than adequate so the following formula can be used

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- **Step 3.** Set up decision rule.

This is a two-tailed test, using a Z statistic and a 5% level of significance. Reject H_0 if $Z \leq -1.960$ or if $Z \geq 1.960$.

- **Step 4.** Compute the test statistic.

We now substitute the sample data into the formula for the test statistic identified in Step 2. The sample proportion is:

$$\hat{p} = \frac{64}{125} = 0.512$$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.512 - 0.75}{\sqrt{\frac{0.75(1-0.75)}{125}}} = -6.15$$

- **Step 5.** Conclusion.

We reject H_0 because -6.15 ≤ -1.960. We have statistically significant evidence at a =0.05 to show that there is a statistically significant difference in the use of dental service by children living in Boston as compared to the national data. (p < 0.0001).

We now conduct the same test using the chi-square goodness-of-fit test. First, we summarize our sample data as follows:

	Saw a Dentist in Past 12 Months	Did Not See a Dentist in Past 12 Months	Total
# of Participants	64	61	125

- **Step 1.** Set up hypotheses and determine level of significance.

  H_0: p_1=0.75, p_2=0.25 or equivalently H_0: Distribution of responses is 0.75, 0.25

  H_1: H_0 is false. α=0.05

- **Step 2.** Select the appropriate test statistic.

The formula for the test statistic is:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

We must assess whether the sample size is adequate. Specifically, we need to check min(np_0, np_1, ...,np_k>) ≥ 5. The sample size here is n=125 and the proportions specified in the null hypothesis are 0.75, 0.25. Thus, min(125(0.75), 125(0.25))=min(93.75, 31.25)=31.25. The sample size is more than adequate so the formula can be used.

- **Step 3.** Set up decision rule.

Here we have df=k-1=2-1=1 and a 5% level of significance. The appropriate critical value is 3.84, and the decision rule is as follows: Reject H_0 if $\chi^2 \geq 3.84$. (Note that $1.96^2 = 3.84$, where 1.96 was the critical value used in the Z test for proportions shown above.)

- **Step 4.** Compute the test statistic.

We now compute the expected frequencies using the sample size and the proportions specified in the null hypothesis. We then substitute the sample data (observed frequencies) into the formula for the test statistic identified in Step 2. We organize the computations in the following table.

	Saw a Dentist in Past 12 Months	Did Not See a Dentist in Past 12 Months	Total
Observed Frequencies (O)	64	61	125
Expected Frequencies (E)	93.75	31.25	125

The test statistic is computed as follows:

$$\chi^2 = \frac{(64-93.75)^2}{93.75} + \frac{(61-31,25)^2}{31.35}$$

$$\chi^2 = 9.44 + 28.32 = 37.8$$

(Note that $(-6.15)^2 = 37.8$, where -6.15 was the value of the Z statistic in the test for proportions shown above.)

- **Step 5.** Conclusion.

We reject H_0 because $37.8 \geq 3.84$. We have statistically significant evidence at $\alpha=0.05$ to show that there is a statistically significant difference in the use of dental service by children living in Boston as compared to the national data. ($p < 0.0001$). This is the same conclusion we reached when we conducted the test using the Z test above. With a dichotomous outcome, $Z^2 = \chi^2$! In statistics, there are often several approaches that can be used to test hypotheses.

Tests for Two or More Independent Samples, Discrete Outcome

Here we extend that application of the chi-square test to the case with two or more independent comparison groups. Specifically, the outcome of interest is discrete with two or more responses and the responses can be ordered or unordered (i.e., the outcome can be dichotomous, ordinal or categorical). We now consider the situation where there are two or more independent comparison groups and the goal of the analysis is to compare the distribution of responses to the discrete outcome variable among several independent comparison groups.

The test is called the χ^2 test of independence and the null hypothesis is that there is no difference in the distribution of responses to the outcome across comparison groups. This is often stated as follows: The outcome variable and the grouping variable (e.g., the comparison treatments or comparison groups) are independent (hence the name of the test). Independence here implies homogeneity in the distribution of the outcome among comparison groups.

The null hypothesis in the χ^2 test of independence is often stated in words as: ***H_0: The distribution of the outcome is independent of the groups. The alternative or research hypothesis is that there is a difference in the distribution of responses to the outcome variable among the comparison groups*** (i.e., that the distribution of responses "depends" on the group). In order to test the hypothesis, we measure the discrete outcome variable in each participant in each comparison group. The data of interest are the observed frequencies (or number of participants in each response category in each group). The formula for the test statistic for the χ^2 test of independence is given below.

Test Statistic for Testing H_0: Distribution of outcome is independent of groups

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

and we find the critical value in a table of probabilities for the chi-square distribution with df=(r-1)*(c-1).

Here O = observed frequency, E=expected frequency in each of the response categories in each group, r = the number of rows in the two-way table and c = the number of columns in the two-way table. r and c correspond to the number of comparison groups and the number of response options in the outcome (see below for more details). The observed frequencies are the sample data and the expected frequencies are computed as described below. The test statistic is appropriate for large samples, defined as expected frequencies of at least 5 in each of the response categories in each group.

The data for the χ^2 test of independence are organized in a two-way table. The outcome and grouping variable are shown in the rows and columns of the table. The sample table below illustrates the data layout. The table entries (blank below) are the numbers of participants in each group responding to each response category of the outcome variable.

Table - Possible outcomes are are listed in the columns; The groups being compared are listed in rows.

	Outcome Variable				
Grouping Variable	**Response Option 1**	**Response Option 2**	**...**	**Response Option c**	**Row Totals**
Group 1					
Group 2					
. . .					
Group r					
Column Totals					N

In the table above, the grouping variable is shown in the rows of the table; r denotes the number of independent groups. The outcome variable is shown in the columns of the table; c denotes the number of response options in the outcome variable. Each combination of a row (group) and column (response) is called a cell of the table. The table has r*c cells and is sometimes called an r x c ("r by c") table. For example, if there are 4 groups and 5 categories in the outcome variable, the data are organized in a 4 X 5 table. The row and column totals are shown along the right-hand margin and the bottom of the table, respectively. The total sample size, N, can be computed by summing the row totals or the column totals. Similar to ANOVA, N does not refer to a

population size here but rather to the total sample size in the analysis. The sample data can be organized into a table like the above. The numbers of participants within each group who select each response option are shown in the cells of the table and these are the observed frequencies used in the test statistic.

The test statistic for the χ^2 test of independence involves comparing observed (sample data) and expected frequencies in each cell of the table. The expected frequencies are computed assuming that the null hypothesis is true. The null hypothesis states that the two variables (the grouping variable and the outcome) are independent. The definition of independence is as follows:

Two events, A and B, are independent if P(A|B) = P(A), or equivalently, if P(A and B) = P(A) P(B).

The second statement indicates that if two events, A and B, are independent then the probability of their intersection can be computed by multiplying the probability of each individual event. To conduct the χ^2 test of independence, we need to compute expected frequencies in each cell of the table. Expected frequencies are computed by assuming that the grouping variable and outcome are independent (i.e., under the null hypothesis). Thus, if the null hypothesis is true, using the definition of independence:

P(Group 1 and Response Option 1) = P(Group 1) P(Response Option 1).

The above states that the probability that an individual is in Group 1 and their outcome is Response Option 1 is computed by multiplying the probability that person is in Group 1 by the probability that a person is in Response Option 1. To conduct the χ^2 test of independence, we need expected frequencies and not expected probabilities. To convert the above probability to a frequency, we multiply by N. Consider the following small example.

	Response 1	Response 2	Response 3	Total
Group 1	10	8	7	25
Group 2	22	15	13	50
Group 3	30	28	17	75
Total	62	51	37	150

The data shown above are measured in a sample of size N=150. The frequencies in the cells of the table are the observed frequencies. If Group and Response are independent, then we can compute the probability that a person in the sample is in Group 1 and Response category 1 using:

P(Group 1 and Response 1) = P(Group 1) P(Response 1),

P(Group 1 and Response 1) = (25/150) (62/150) = 0.069.

Thus if Group and Response are independent we would expect 6.9% of the sample to be in the top left cell of the table (Group 1 and Response 1). The expected frequency is 150(0.069) = 10.4. We could do the same for Group 2 and Response 1:

P(Group 2 and Response 1) = P(Group 2) P(Response 1),

P(Group 2 and Response 1) = (50/150) (62/150) = 0.138.

The expected frequency in Group 2 and Response 1 is 150(0.138) = 20.7.

Thus, the formula for determining the expected cell frequencies in the χ^2 test of independence is as follows:

Expected Cell Frequency = (Row Total * Column Total)/N.

The above computes the expected frequency in one step rather than computing the expected probability first and then converting to a frequency.

Example:

In a prior example we evaluated data from a survey of university graduates which assessed, among other things, how frequently they exercised. The survey was completed by 470 graduates. In the prior example we used the χ^2 goodness-of-fit test to assess whether there was a shift in the distribution of responses to the exercise question following the implementation of a health promotion campaign on campus. We specifically considered one sample (all students) and compared the observed distribution to the distribution of responses the prior year (a historical control). Suppose we now wish to assess whether there is a relationship between exercise on campus and students' living arrangements. As part of the same survey, graduates were asked where they lived their senior year. The response options were dormitory, on-campus apartment, off-campus apartment, and at home (i.e., commuted to and from the university). The data are shown below.

	No Regular Exercise	Sporadic Exercise	Regular Exercise	Total
Dormitory	32	30	28	90
On-Campus Apartment	74	64	42	180
Off-Campus Apartment	110	25	15	150
At Home	39	6	5	50
Total	255	125	90	470

Based on the data, is there a relationship between exercise and student's living arrangement? Do you think where a person lives affect their exercise status? Here we have four independent comparison groups (living arrangement) and a discrete (ordinal) outcome variable with three response options. We specifically want to test whether living arrangement and exercise are independent. We will run the test using the five-step approach.

- **Step 1.** Set up hypotheses and determine level of significance.

  H_0: Living arrangement and exercise are independent

  H_1: H_0 is false. $\alpha=0.05$

The null and research hypotheses are written in words rather than in symbols. The research hypothesis is that the grouping variable (living arrangement) and the outcome variable (exercise) are dependent or related.

- **Step 2.** Select the appropriate test statistic.

The formula for the test statistic is:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}.$$

The condition for appropriate use of the above test statistic is that each expected frequency is at least 5. In Step 4 we will compute the expected frequencies and we will ensure that the condition is met.

- **Step 3.** Set up decision rule.

The decision rule depends on the level of significance and the degrees of freedom, defined as df = (r-1)(c-1), where r and c are the numbers of rows and columns in the two-way data table. The row variable is the living arrangement and there are 4 arrangements considered, thus r=4. The column variable is exercise and 3 responses are considered, thus c=3. For this test, df=(4-1)(3-1)=3(2)=6. Again, with χ^2 tests there are no upper, lower or two-tailed tests. If the null hypothesis is true, the observed and expected frequencies will be close in value and the χ^2 statistic will be close to zero. If the null hypothesis is false, then the χ^2 statistic will be large. The rejection region for the χ^2 test of independence is always in the upper (right-hand) tail of the distribution. For df=6 and a 5% level of significance, the appropriate critical value is 12.59 and the decision rule is as follows: Reject H_0 if $c^2 \geq 12.59$.

- **Step 4.** Compute the test statistic.

We now compute the expected frequencies using the formula,

Expected Frequency = (Row Total * Column Total)/N.

The computations can be organized in a two-way table. The top number in each cell of the table is the observed frequency and the bottom number is the expected frequency. The expected frequencies are shown in parentheses.

	No Regular Exercise	Sporadic Exercise	Regular Exercise	Total
Dormitory	32 (48.8)	30 (23.9)	28 (17.2)	90
On-Campus Apartment	74 (97.7)	64 (47.9)	42 (34.5)	180
Off-Campus Apartment	110 (81.4)	25 (39.9)	15 (28.7)	150
At Home	39 (27.1)	6 (13.3)	5 (9.6)	50
Total	255	125	90	470

Notice that the expected frequencies are taken to one decimal place and that the sums of the observed frequencies are equal to the sums of the expected frequencies in each row and column of the table.

Recall in Step 2 a condition for the appropriate use of the test statistic was that each expected frequency is at least 5. This is true for this sample (the smallest expected frequency is 9.6) and therefore it is appropriate to use the test statistic.

The test statistic is computed as follows:

$$\chi^2 = \frac{(32-48.8)^2}{48.8} + \frac{(30-23.9)^2}{23.9} + \frac{(28-17.2)^2}{17.2} + \frac{(74-48.8)^2}{48.8} + \frac{(64-47.9)^2}{47.9} + \frac{(42-34.5)^2}{34.5} + \frac{(110-81.4)^2}{81.4} + \frac{(25-39.9)^2}{39.9} + \frac{(15-28.7)^2}{28.7} + \frac{(39-27.1)^2}{27.1} + \frac{(6-13.3)^2}{13.3} + \frac{(5.9.6)^2}{9.6}$$

$$\chi^2 = 5.78 + 1.58 + 6.78 + 5.75 + 5.41 + 1.63 + 10.05 + 5.58 + 6.54 + 5.23 + 4.01 + 2.20 = 60.5$$

- **Step 5.** Conclusion.

We reject H_0 because 60.5 $\geq$ 12.59. We have statistically significant evidence at a =0.05 to show that H_0 is false or that living arrangement and exercise are not independent (i.e., they are dependent or related), $p < 0.005$.

Again, the χ^2 test of independence is used to test whether the distribution of the outcome variable is similar across the comparison groups. Here we rejected H_0 and concluded that the distribution of exercise is not independent of living arrangement, or that there is a relationship between living arrangement and exercise. The test provides an overall assessment of statistical significance. When the null hypothesis is rejected, it is important to review the sample data to understand the nature of the relationship. Consider again the sample data.

	No Regular Exercise	Sporadic Exercise	Regular Exercise	Total
Dormitory	32	30	28	90
On-Campus Apartment	74	64	42	180
Off-Campus Apartment	110	25	15	150
At Home	39	6	5	50
Total	255	125	90	470

Because there are different numbers of students in each living situation, it makes the comparisons of exercise patterns difficult on the basis of the frequencies alone. The following table displays the percentages of students in each exercise category by living arrangement. The percentages sum to 100% in each row of the table. For comparison purposes, percentages are also shown for the total sample along the bottom row of the table.

	No Regular Exercise	Sporadic Exercise	Regular Exercise
Dormitory	36%	33%	31%
On-Campus Apartment	41%	36%	23%
Off-Campus Apartment	73%	17%	10%
At Home	78%	12%	10%
Total	54%	27%	19%

From the above, it is clear that higher percentages of students living in dormitories and in on-campus apartments reported regular exercise (31% and 23%) as compared to students living in off-campus apartments and at home (10% each).

Test Yourself

Pancreaticoduodenectomy (PD) is a procedure that is associated with considerable morbidity. A study was recently conducted on 553 patients who had a successful PD between January 2000 and December 2010 to determine whether their Surgical Apgar Score (SAS) is related to 30-day perioperative morbidity and mortality. The table below gives the number of patients experiencing no, minor, or major morbidity by SAS category.

Surgical Apgar Score	No morbidity	Minor morbidity	Major morbidity or mortality
0-4	21	20	16
5-6	135	71	35
7-10	158	62	35

Question: What would be an appropriate statistical test to examine whether there is an association between Surgical Apgar Score and patient outcome? Using 14.13 as the value of the test statistic for these data, carry out the appropriate test at a 5% level of significance. Show all parts of your test.

Answer

In the module on hypothesis testing for means and proportions, we discussed hypothesis testing applications with a dichotomous outcome variable and two independent comparison groups. We presented a test using a test statistic Z to test for equality of independent proportions. The chi-square test of independence can also be used with a dichotomous outcome and the results are mathematically equivalent.

In the prior module, we considered the following example. Here we show the equivalence to the chi-square test of independence.

Example:

A randomized trial is designed to evaluate the effectiveness of a newly developed pain reliever designed to reduce pain in patients following joint replacement surgery. The trial compares the new pain reliever to the pain reliever currently in use (called the standard of care). A total of 100 patients undergoing joint replacement surgery agreed to participate in the trial. Patients were randomly assigned to receive either the new pain reliever or the standard pain reliever following surgery and were blind to the treatment assignment. Before receiving the assigned treatment, patients were asked to rate their pain on a scale of 0-10 with higher scores indicative of more pain. Each patient was then given the assigned treatment and after 30 minutes was again asked to rate their pain on the same scale. The primary outcome was a reduction in pain of 3 or more scale points (defined by clinicians as a clinically meaningful reduction). The following data were observed in the trial.

Treatment Group	n	Number with Reduction of 3+ Points	Proportion with Reduction of 3+ Points
New Pain Reliever	50	23	0.46
Standard Pain Reliever	50	11	0.22

We tested whether there was a significant difference in the proportions of patients reporting a meaningful reduction (i.e., a reduction of 3 or more scale points) using a Z statistic, as follows.

- **Step 1.** Set up hypotheses and determine level of significance

  H_0: $p_1 = p_2$

  H_1: $p_1 \neq p_2$ $\alpha=0.05$

Here the new or experimental pain reliever is group 1 and the standard pain reliever is group 2.

- **Step 2.** Select the appropriate test statistic.

We must first check that the sample size is adequate. Specifically, we need to ensure that we have at least 5 successes and 5 failures in each comparison group or that:

$$\min\left(n_1\hat{p}_1, n_1(1-\hat{p}_1), n_2\hat{p}_2, n_2(1-\hat{p}_2)\right) \geq 5$$

In this example, we have

$$\min\left(50(0.46),\ 50(1-0.46),\ 50(0.22),\ 50(1-0.22)\right) = \min(23, 27, 11, 39) = 11$$

Therefore, the sample size is adequate, so the following formula can be used:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\left(1-\hat{p}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}.$$

- **Step 3.** Set up decision rule.

Reject H_0 if $Z \leq -1.960$ or if $Z \geq 1.960$.

- **Step 4.** Compute the test statistic.

We now substitute the sample data into the formula for the test statistic identified in Step 2. We first compute the overall proportion of successes:

$$\hat{p} = \frac{(x_1+x_2)}{(n_1+n_2)} = \frac{23+11}{50+50} = \frac{34}{100} = \mathbf{0.34}$$

We now substitute to compute the test statistic.

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\left(1-\hat{p}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.46-0.22}{\sqrt{0.34\left(1-0.34\right)\left(\frac{1}{50}+\frac{1}{50}\right)}} = \frac{0.24}{0.095} = \mathbf{2.53}$$

- **Step 5.** Conclusion.

We reject H_0 because $2.53 \geq 1.960$. We have statistically significant evidence at α=0.05 to show that there is a difference in the proportions of patients on the new pain reliever reporting a meaningful reduction (i.e., a reduction of 3 or more scale points) as compared to patients on the standard pain reliever.

We now conduct the same test using the chi-square test of independence.

- **Step 1.** Set up hypotheses and determine level of significance.

  H_0: Treatment and outcome (meaningful reduction in pain) are independent

  H_1: H_0 is false. α=0.05

- **Step 2.** Select the appropriate test statistic.

The formula for the test statistic is:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

The condition for appropriate use of the above test statistic is that each expected frequency is at least 5. In Step 4 we will compute the expected frequencies and we will ensure that the condition is met.

- **Step 3.** Set up decision rule.

For this test, df=(2-1)(2-1)=1. At a 5% level of significance, the appropriate critical value is 3.84 and the decision rule is as follows: Reject H0 if $\chi^2 \geq 3.84$. (Note that $1.96^2 = 3.84$, where 1.96 was the critical value used in the Z test for proportions shown above.)

- **Step 4.** Compute the test statistic.

We now compute the expected frequencies using:

$$\text{Expected frequency} = \frac{(\text{Row Total} \times \text{Column Total})}{N}$$

The computations can be organized in a two-way table. The top number in each cell of the table is the observed frequency and the bottom number is the expected frequency. The expected frequencies are shown in parentheses.

Treatment Group	# with Reduction of 3+ Points	# with Reduction of <3 Points	Total
New Pain Reliever	23 (17.0)	27 (33.0)	50
Standard Pain Reliever	11 (17.0)	39 (33.0)	50
Total	34	66	100

A condition for the appropriate use of the test statistic was that each expected frequency is at least 5. This is true for this sample (the smallest expected frequency is 22.0) and therefore it is appropriate to use the test statistic.

The test statistic is computed as follows:

$$\chi^2 = \frac{(23-17.0)^2}{17.0} + \frac{(27-33.0)^2}{33.0} + \frac{(11-17.0)^2}{17.0} + \frac{(39-33.0)^2}{33.0}$$

$$\chi^2 = 2.12 + 1.09 + 2.12 + 1.09 = 6.42$$

(Note that $(2.53)^2 = 6.4$, where 2.53 was the value of the Z statistic in the test for proportions shown above.)

- **Step 5.** Conclusion.

We reject H_0 because $6.42 \geq 3.84$. We have statistically significant evidence at α=0.05 to show that H_0 is false or that treatment and outcome are not independent (i.e., they are dependent or related). This is the same conclusion we reached when we conducted the test using the Z test above. With a dichotomous outcome and two independent comparison groups, $Z^2 = \chi^2$! Again, in statistics there are often several approaches that can be used to test hypotheses.

Chi-Squared Tests in R

The video below by Mike Marin demonstrates how to perform chi-squared tests in the R programming language.

Answer to Problem on Pancreaticoduodenectomy and Surgical Apgar Scores

We have 3 independent comparison groups (Surgical Apgar Score) and a categorical outcome variable (morbidity/mortality). We can run a Chi-Squared test of independence.

- **Step 1:**

  H_0: Apgar scores and patient outcome are independent of one another.

  H_A: Apgar scores and patient outcome are ***not*** independent.

- **Step 2:**

$\chi^2 = \Sigma \frac{(O-E)^2}{E} = 14.3$ (We were given the chi-squared value)

- **Step 3:**

$$\text{degrees of freedom} = (r-1)(c-1) = (3-1)(3-1) = 4$$

Therefore reject H_0 if $\chi^2 \geq 9.49$

- **Step 4:**

Chi-squared = 14.3

- **Step 5:**

Since 14.3 is greater than 9.49, we reject H_0.

There is an association between Apgar scores and patient outcome. The lowest Apgar score group (0 to 4) experienced the highest percentage of major morbidity or mortality (16 out of 57=28%) compared to the other Apgar score groups.