Chapter 11

The Chi-Square Distribution

11.1 The Chi-Square Distribution[1]

11.1.1 Student Learning Objectives

By the end of this chapter, the student should be able to:

- Interpret the chi-square probability distribution as the sample size changes.
- Conduct and interpret chi-square goodness-of-fit hypothesis tests.
- Conduct and interpret chi-square test of independence hypothesis tests.
- Conduct and interpret chi-square single variance hypothesis tests (optional).

11.1.2 Introduction

Have you ever wondered if lottery numbers were evenly distributed or if some numbers occurred with a greater frequency? How about if the types of movies people preferred were different across different age groups? What about if a coffee machine was dispensing approximately the same amount of coffee each time? You could answer these questions by conducting a hypothesis test.

You will now study a new distribution, one that is used to determine the answers to the above examples. This distribution is called the Chi-square distribution.

In this chapter, you will learn the three major applications of the Chi-square distribution:

- The goodness-of-fit test, which determines if data fit a particular distribution, such as with the lottery example
- The test of independence, which determines if events are independent, such as with the movie example
- The test of a single variance, which tests variability, such as with the coffee example

NOTE: Though the Chi-square calculations depend on calculators or computers for most of the calculations, there is a table available (see the Table of Contents **15. Tables**). TI-83+ and TI-84 calculator instructions are included in the text.

[1]This content is available online at <http://cnx.org/content/m17048/1.7/>.

11.1.3 Optional Collaborative Classroom Activity

Look in the sports section of a newspaper or on the Internet for some sports data (baseball averages, basketball scores, golf tournament scores, football odds, swimming times, etc.). Plot a histogram and a boxplot using your data. See if you can determine a probability distribution that your data fits. Have a discussion with the class about your choice.

11.2 Notation[2]

The notation for the chi-square distribution is:

$\chi^2 \sim \chi^2_{df}$

where $df =$ degrees of freedom depend on how chi-square is being used. (If you want to practice calculating chi-square probabilities then use $df = n - 1$. The degrees of freedom for the three major uses are each calculated differently.)

For the χ^2 distribution, the population mean is $\mu = df$ and the population standard deviation is $\sigma = \sqrt{2 \cdot df}$.

The random variable is shown as χ^2 but may be any upper case letter.

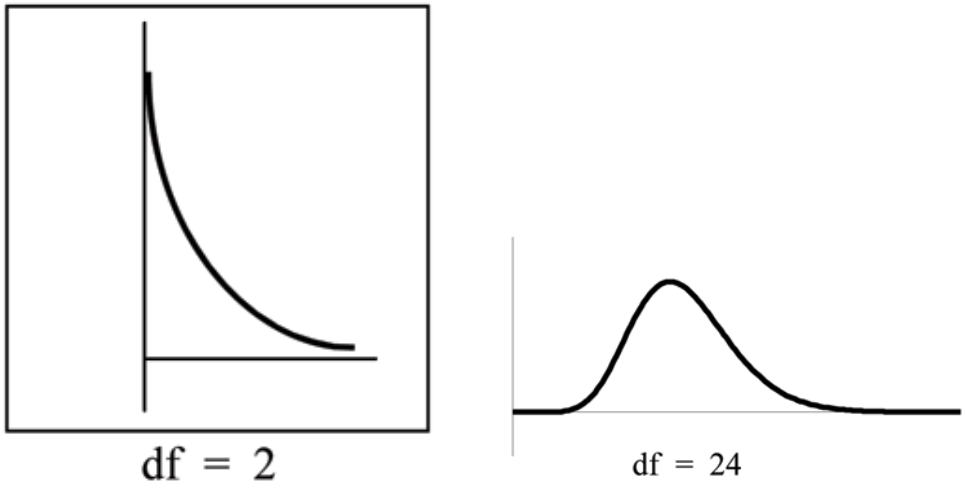
The random variable for a chi-square distribution with k degrees of freedom is the sum of k independent, squared standard normal variables.

$\chi^2 = (Z_1)^2 + (Z_2)^2 + ... + (Z_k)^2$

11.3 Facts About the Chi-Square Distribution[3]

1. The curve is nonsymmetrical and skewed to the right.
2. There is a different chi-square curve for each df.

[2]This content is available online at <http://cnx.org/content/m17052/1.5/>.
[3]This content is available online at <http://cnx.org/content/m17045/1.5/>.



(a) (b)

Figure 11.1

3. The test statistic for any test is always greater than or equal to zero.
4. When $df > 90$, the chi-square curve approximates the normal. For $X \sim \chi^2_{1000}$ the mean, $\mu = df = 1000$ and the standard deviation, $\sigma = \sqrt{2 \cdot 1000} = 44.7$. Therefore, $X \sim N\,(1000, 44.7)$, approximately.
5. The mean, μ, is located just to the right of the peak.

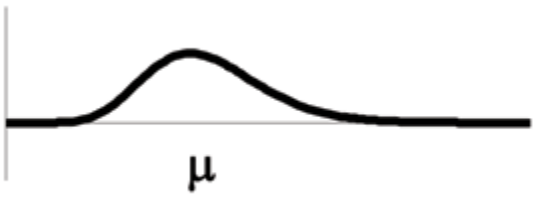


Figure 11.2

11.4 Goodness-of-Fit Test[4]

In this type of hypothesis test, you determine whether the data **"fit"** a particular distribution or not. For example, you may suspect your unknown data fit a binomial distribution. You use a chi-square test (meaning the distribution for the hypothesis test is chi-square) to determine if there is a fit or not. **The null and the alternate hypotheses for this test may be written in sentences or may be stated as equations or inequalities.**

[4]This content is available online at <http://cnx.org/content/m17192/1.7/>.

The test statistic for a goodness-of-fit test is:

$$\sum_{n} \frac{(O-E)^2}{E} \tag{11.1}$$

where:

- O = observed values (data)
- E = expected values (from theory)
- n = the number of different data cells or categories

The observed values are the data values and the expected values are the values you would expect to get if the null hypothesis were true. There are n terms of the form $\frac{(O-E)^2}{E}$.

The degrees of freedom are *df = (number of categories - 1)*.

The goodness-of-fit test is almost always right tailed. If the observed values and the corresponding expected values are not close to each other, then the test statistic can get very large and will be way out in the right tail of the chi-square curve.

Example 11.1
Absenteeism of college students from math classes is a major concern to math instructors because missing class appears to increase the drop rate. Three statistics instructors wondered whether the absentee rate was the **same** for every day of the school week. They took a sample of absent students from three of their statistics classes during one week of the term. The results of the survey appear in the table.

	Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
# of students absent	28	22	18	20	32

Table 11.1

Determine the null and alternate hypotheses needed to run a goodness-of-fit test.

Since the instructors wonder whether the absentee rate is the same for every school day, we could say in the null hypothesis that the data **"fit"** a uniform distribution.

H_o**:** The rate at which college students are absent from their statistics class fits a uniform distribution.

The alternate hypothesis is the opposite of the null hypothesis.

H_a**:** The rate at which college students are absent from their statistics class does not fit a uniform distribution.

Problem 1
How many students do you **expect** to be absent on any given school day?

Solution
The total number of students in the sample is 120. **If the null hypothesis were true,** you would divide 120 by 5 to get 24 absences expected per day. **The expected number is based on a true null hypothesis.**

Problem 2
What are the degrees of freedom (df)?

Solution
There are 5 days of the week or 5 "cells" or categories.

$df = no.\,cells - 1 = 5 - 1 = 4$

Example 11.2
Employers particularly want to know which days of the week employees are absent in a five day work week. Most employers would like to believe that employees are absent equally during the week. That is, the average number of times an employee is absent is the same on Monday, Tuesday, Wednesday, Thursday, or Friday. Suppose a sample of 20 absent days was taken and the days absent were distributed as follows:

Day of the Week Absent

	Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
Number of Absences	5	4	2	3	6

Table 11.2

Problem
For the population of employees, do the absent days occur with equal frequencies during a five day work week? Test at a 5% significance level.

Solution
The null and alternate hypotheses are:

- H_o: The absent days occur with equal frequencies, that is, they fit a uniform distribution.
- H_a: The absent days occur with unequal frequencies, that is, they do not fit a uniform distribution.

If the absent days occur with equal frequencies, then, out of 20 absent days, there would be 4 absences on Monday, 4 on Tuesday, 4 on Wednesday, 4 on Thursday, and 4 on Friday. These numbers are the **expected** (E) values. The values in the table are the **observed** (O) values or data.

This time, calculate the χ^2 test statistic by hand. Make a chart with the following headings:

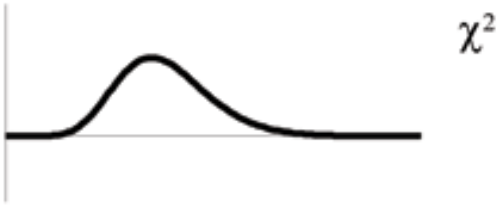
- Expected (E) values
- Observed (O) values
- $(O - E)$
- $(O - E)^2$
- $\frac{(O-E)^2}{E}$

Now add (sum) the last column. Verify that the sum is 2.5. This is the χ^2 test statistic.

To find the p-value, calculate $P\left(\chi^2 > 2.5\right)$. This test is right-tailed.

The dfs are the *number of cells* $- 1 = 4$.

Next, complete a graph like the one below with the proper labeling and shading. (You should shade the right tail. It will be a "large" right tail for this example because the p-value is "large.")



Use a computer or calculator to find the p-value. You should get $p\text{-}value = 0.6446$.

The decision is to not reject the null hypothesis.

Conclusion: At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the absent days do not occur with equal frequencies.

TI-83+ and TI-84: Press `2nd DISTR`. Arrow down to χ^2`cdf`. Press `ENTER`. Enter `(2.5,1E99,4)`. Rounded to 4 places, you should see 0.6446 which is the p-value.

NOTE: TI-83+ and some TI-84 calculators do not have a special program for the test statistic for the goodness-of-fit test. The next example (Example 11-3) has the calculator instructions. The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF`. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press `STAT TESTS` and `Chi2 GOF`. Enter the list names for the Observed list and the Expected list. Enter whatever else is asked and press `calculate` or `draw`. Make sure you clear any lists before you start. See below.

NOTE: **To Clear Lists in the calculators:** Go into `STAT EDIT` and arrow up to the list name area of the particular list. Press `CLEAR` and then arrow down. The list will be cleared. Or, you can press `STAT` and press 4 (for `ClrList`). Enter the list name and press `ENTER`.

Example 11.3
One study indicates that the number of televisions that American families have is distributed (this is the **given** distribution for the American population) as follows:

Number of Televisions	Percent
0	10
1	16
2	55
3	11
over 3	8

Table 11.3

The table contains expected (E) percents.

A random sample of 600 families in the far western United States resulted in the following data:

Number of Televisions	Frequency
0	66
1	119
2	340
3	60
over 3	15
	Total = 600

Table 11.4

The table contains observed (O) frequency values.

Problem

At the 1% significance level, does it appear that the distribution "number of televisions" of far western United States families is different from the distribution for the American population as a whole?

Solution

This problem asks you to test whether the far western United States families distribution fits the distribution of the American families. This test is always right-tailed.

The first table contains expected percentages. To get expected (E) frequencies, multiply the percentage by 600. The expected frequencies are:

Number of Televisions	Percent	Expected Frequency
0	10	$(0.10) \cdot (600) = 60$
1	16	$(0.16) \cdot (600) = 96$
2	55	$(0.55) \cdot (600) = 330$
3	11	$(0.11) \cdot (600) = 66$
over 3	8	$(0.08) \cdot (600) = 48$

Table 11.5

Therefore, the expected frequencies are 60, 96, 330, 66, and 48. In the TI calculators, you can let the calculator do the math. For example, instead of 60, enter .10*600.

H_o: The "number of televisions" distribution of far western United States families is the same as the "number of televisions" distribution of the American population.

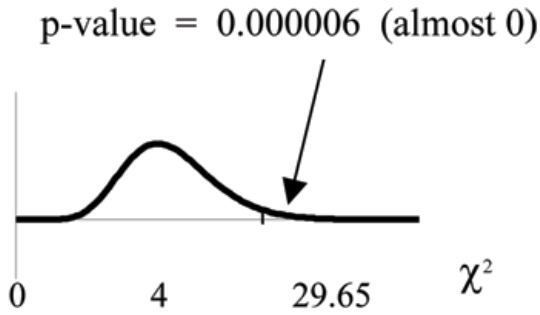
H_a: The "number of televisions" distribution of far western United States families is different from the "number of televisions" distribution of the American population.

Distribution for the test: χ^2_4 where $df = \textit{(the number of cells)} - 1 = 5 - 1 = 4$.

NOTE: $df \neq 600 - 1$

Calculate the test statistic: $\chi^2 = 29.65$

Graph:



Probability statement: $p\text{-}value = P\left(\chi^2 > 29.65\right) = 0.000006$.

Compare α and the p-value:

- $\alpha = 0.01$
- $p\text{-}value = 0.000006$

So, $\alpha > p\text{-}value$.

Make a decision: Since $\alpha > p\text{-}value$, reject H_o.

This means you reject the belief that the distribution for the far western states is the same as that of the American population as a whole.

Conclusion: At the 1% significance level, from the data, there is sufficient evidence to conclude that the "number of televisions" distribution for the far western United States is different from the "number of televisions" distribution for the American population as a whole.

NOTE: TI-83+ and some TI-84 calculators: Press `STAT` and `ENTER`. Make sure to clear lists `L1`, `L2`, and `L3` if they have data in them (see the note at the end of Example 11-2). Into `L1`, put the observed frequencies `66`, `119`, `349`, `60`, `15`. Into `L2`, put the expected frequencies `.10*600`, `.16*600`, `.55*600`, `.11*600`, `.08*600`. Arrow over to list `L3` and up to the name area `"L3"`. Enter `(L1-L2)^2/L2` and `ENTER`. Press `2nd QUIT`. Press `2nd LIST` and arrow over to `MATH`. Press `5`. You should see `"sum" (Enter L3)`. Rounded to 2 decimal places, you should see `29.65`. Press `2nd DISTR`. Press `7` or Arrow down to `7:`χ`2cdf` and press `ENTER`. Enter `(29.65,1E99,4)`. Rounded to 4 places, you should see `5.77E-6 = .000006` (rounded to 6 decimal places) which is the p-value.

Example 11.4
Suppose you flip two coins 100 times. The results are 20 HH, 27 HT, 30 TH, and 23 TT. Are the coins fair? Test at a 5% significance level.

Solution
This problem can be set up as a goodness-of-fit problem. The sample space for flipping two fair coins is {HH, HT, TH, TT}. Out of 100 flips, you would expect 25 HH, 25 HT, 25 TH, and 25 TT. This is the expected distribution. The question, "Are the coins fair?" is the same as saying, "Does the distribution of the coins (20 HH, 27 HT, 30 TH, 23 TT) fit the expected distribution?"

Random Variable: Let X = the number of heads in one flip of the two coins. X takes on the value 0, 1, 2. (There are 0, 1, or 2 heads in the flip of 2 coins.) Therefore, the **number of cells is 3**. Since X = the number of heads, the observed frequencies are 20 (for 2 heads), 57 (for 1 head), and 23 (for 0 heads or both tails). The expected frequencies are 25 (for 2 heads), 50 (for 1 head), and 25 (for 0 heads or both tails). This test is right-tailed.

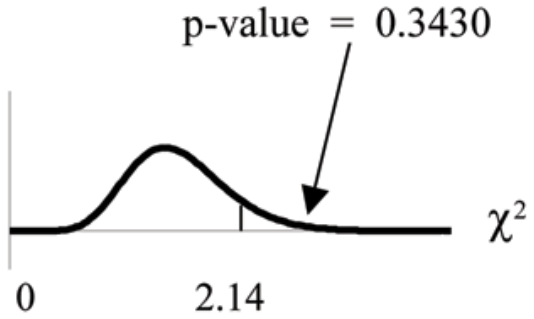
H_o: The coins are fair.

H_a: The coins are not fair.

Distribution for the test: χ^2_2 where $df = 3 - 1 = 2$.

Calculate the test statistic: $\chi^2 = 2.14$

Graph:



Probability statement: $p\text{-}value = P\left(\chi^2 > 2.14\right) = 0.3430$

Compare α and the p-value:

- $\alpha = 0.05$
- $p\text{-}value = 0.3430$

So, $\alpha < p\text{-}value$.

Make a decision: Since $\alpha < p\text{-}value$, do not reject H_o.

Conclusion: The coins are fair.

NOTE: TI-83+ and some TI- 84 calculators: Press `STAT` and `ENTER`. Make sure you clear lists `L1`, `L2`, and `L3` if they have data in them. Into `L1`, put the observed frequencies `20`, `57`, `23`. Into `L2`, put the expected frequencies `25`, `50`, `25`. Arrow over to list `L3` and up to the name area `"L3"`. Enter `(L1-L2)^2/L2` and `ENTER`. Press `2nd QUIT`. Press `2nd LIST` and arrow over to `MATH`. Press `5`. You should see `"sum".Enter L3`. Rounded to 2 decimal places, you should see `2.14`. Press `2nd DISTR`. Arrow down to `7:`χ`2cdf` (or press 7). Press `ENTER`. Enter `2.14,1E99,2)`. Rounded to 4 places, you should see `.3430` which is the p-value.

NOTE: For the newer TI-84 calculators, check `STAT TESTS` to see if you have `Chi2 GOF`. If you do, see the calculator instructions (a NOTE) before Example 11-3

11.5 Test of Independence[5]

Tests of independence involve using a **contingency table** of observed (data) values. You first saw a contingency table when you studied probability in the Probability Topics (Section 3.1) chapter.

The test statistic for a test of independence is similar to that of a goodness-of-fit test:

$$\sum_{(i \cdot j)} \frac{(O-E)^2}{E} \tag{11.2}$$

where:

- O = observed values
- E = expected values
- i = the number of rows in the table
- j = the number of columns in the table

There are $i \cdot j$ terms of the form $\frac{(O-E)^2}{E}$.

A test of independence determines whether two factors are independent or not. You first encountered the term independence in Chapter 3. As a review, consider the following example.

Example 11.5
Suppose A = a speeding violation in the last year and B = a car phone user. If A and B are independent then $P(A \ AND \ B) = P(A)\,P(B)$. $A \ AND \ B$ is the event that a driver received a speeding violation last year and is also a car phone user. Suppose, in a study of drivers who received speeding violations in the last year and who use car phones, that 755 people were surveyed. Out of the 755, 70 had a speeding violation and 685 did not; 305 were car phone users and 450 were not.

Let y = expected number of car phone users who received speeding violations.

If A and B are independent, then $P(A \ AND \ B) = P(A)\,P(B)$. By substitution,

$\frac{y}{755} = \frac{70}{755} \cdot \frac{305}{755}$

Solve for y : $y = \frac{70 \cdot 305}{755} = 28.3$

About 28 people from the sample are expected to be car phone users and to receive speeding violations.

In a test of independence, we state the null and alternate hypotheses in words. Since the contingency table consists of **two factors**, the null hypothesis states that the factors are **independent** and the alternate hypothesis states that they are **not independent (dependent)**. If we do a test of independence using the example above, then the null hypothesis is:

H_o: Being a car phone user and receiving a speeding violation are independent events.

If the null hypothesis were true, we would expect about 28 people to be car phone users and to receive a speeding violation.

The test of independence is always right-tailed because of the calculation of the test statistic. If the expected and observed values are not close together, then the test statistic is very large and way out in the right tail of the chi-square curve, like goodness-of-fit.

[5]This content is available online at <http://cnx.org/content/m17191/1.10/>.

The degrees of freedom for the test of independence are:

df = (number of columns - 1)(number of rows - 1)

The following formula calculates the **expected number** (E):

$E = \frac{(row\ total)(column\ total)}{total\ number\ surveyed}$

Example 11.6
In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four-year college students, and nonstudents. The following table is a **sample** of the adult volunteers and the number of hours they volunteer per week.

Number of Hours Worked Per Week by Volunteer Type (Observed)

Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours	Row Total
Community College Students	111	96	48	255
Four-Year College Students	96	133	61	290
Nonstudents	91	150	53	294
Column Total	298	379	162	839

Table 11.6: The table contains **observed (O)** values (data).

Problem
Are the number of hours volunteered **independent** of the type of volunteer?

Solution
The **observed table** and the question at the end of the problem, "Are the number of hours volunteered independent of the type of volunteer?" tell you this is a test of independence. The two factors are **number of hours volunteered** and **type of volunteer**. This test is always right-tailed.

H_o: The number of hours volunteered is **independent** of the type of volunteer.

H_a: The number of hours volunteered is **dependent** on the type of volunteer.

The expected table is:

Number of Hours Worked Per Week by Volunteer Type (Expected)

Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours
Community College Students	90.57	115.19	49.24
Four-Year College Students	103.00	131.00	56.00
Nonstudents	104.42	132.81	56.77

Table 11.7: The table contains **expected** (E) values (data).

For example, the calculation for the expected frequency for the top left cell is

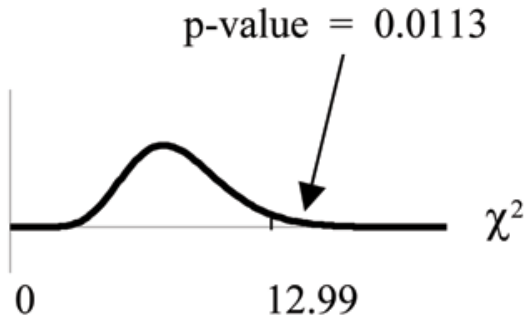
$E = \frac{(row\ total)(column\ total)}{total\ number\ surveyed} = \frac{255 \cdot 298}{839} = 90.57$

Calculate the test statistic: $\chi^2 = 12.99$ (calculator or computer)

Distribution for the test: χ^2_4

$df = (3\ columns - 1)\,(3\ rows - 1) = (2)\,(2) = 4$

Graph:



Probability statement: $p\text{-}value = P\left(\chi^2 > 12.99\right) = 0.0113$

Compare α and the *p-value*: Since no α is given, assume $\alpha = 0.05$. $p\text{-}value = 0.0113$. $\alpha > p\text{-}value$.

Make a decision: Since $\alpha > p\text{-}value$, reject H_o. This means that the factors are not independent.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that the number of hours volunteered and the type of volunteer are dependent on one another.

For the above example, if there had been another type of volunteer, teenagers, what would the degrees of freedom be?

NOTE: Calculator instructions follow.

TI-83+ and TI-84 calculator: Press the `MATRX` key and arrow over to `EDIT`. Press `1:[A]`. Press `3 ENTER 3 ENTER`. Enter the table values by row from Example 11-6. Press `ENTER` after each. Press `2nd QUIT`. Press `STAT` and arrow over to `TESTS`. Arrow down to `C:`χ`2-TEST`. Press `ENTER`. You should see `Observed:[A] and Expected:[B]`. Arrow down to `Calculate`. Press `ENTER`. The test statistic is 12.9909 and the *p-value* $= 0.0113$. Do the procedure a second time but arrow down to `Draw` instead of `calculate`.

Example 11.7
De Anza College is interested in the relationship between anxiety level and the need to succeed in school. A random sample of 400 students took a test that measured anxiety level and need to succeed in school. The table shows the results. De Anza College wants to know if anxiety level and need to succeed in school are independent events.

Need to Succeed in School vs. Anxiety Level

Need to Succeed in School	High Anxiety	Med-high Anxiety	Medium Anxiety	Med-low Anxiety	Low Anxiety	Row Total
High Need	35	42	53	15	10	155
Medium Need	18	48	63	33	31	193
Low Need	4	5	11	15	17	52
Column Total	57	95	127	63	58	400

Table 11.8

Problem 1
How many high anxiety level students are expected to have a high need to succeed in school?

Solution
The column total for a high anxiety level is 57. The row total for high need to succeed in school is 155. The sample size or total surveyed is 400.

$E = \frac{(row\ total)(column\ total)}{total\ surveyed} = \frac{155 \cdot 57}{400} = 22.09$

The expected number of students who have a high anxiety level and a high need to succeed in school is about 22.

Problem 2
If the two variables are independent, how many students do you expect to have a low need to succeed in school and a med-low level of anxiety?

Solution
The column total for a med-low anxiety level is 63. The row total for a low need to succeed in school is 52. The sample size or total surveyed is 400.

Problem 3 ***(Solution on p. 470.)***

a. $E = \frac{(row\ total)(column\ total)}{total\ surveyed}$ =
b. The expected number of students who have a med-low anxiety level and a low need to succeed in school is about:

11.6 Test of a Single Variance (Optional)[6]

A test of a single variance assumes that the underlying distribution is **normal**. The null and alternate hypotheses are stated in terms of the **population variance** (or population standard deviation). The test

[6]This content is available online at <http://cnx.org/content/m17059/1.6/>.

statistic is:

$$\frac{(n-1)\cdot s^2}{\sigma^2} \tag{11.3}$$

where:

- n = the total number of data
- s^2 = sample variance
- σ^2 = population variance

You may think of s as the random variable in this test. The degrees of freedom are $df = n - 1$.

A test of a single variance may be right-tailed, left-tailed, or two-tailed.

The following example will show you how to set up the null and alternate hypotheses. The null and alternate hypotheses contain statements about the population variance.

Example 11.8
Math instructors are not only interested in how their students do on exams, on average, but how the exam scores vary. To many instructors, the variance (or standard deviation) may be more important than the average.

Suppose a math instructor believes that the standard deviation for his final exam is 5 points. One of his best students thinks otherwise. The student claims that the standard deviation is more than 5 points. If the student were to conduct a hypothesis test, what would the null and alternate hypotheses be?

Solution
Even though we are given the population standard deviation, we can set the test up using the population variance as follows.

- H_o: $\sigma^2 = 5^2$
- H_a: $\sigma^2 > 5^2$

Example 11.9
With individual lines at its various windows, a post office finds that the standard deviation for normally distributed waiting times for customers on Friday afternoon is 7.2 minutes. The post office experiments with a single main waiting line and finds that for a random sample of 25 customers, the waiting times for customers have a standard deviation of 3.5 minutes.

With a significance level of 5%, test the claim that **a single line causes lower variation among waiting times (shorter waiting times) for customers**.

Solution
Since the claim is that a single line causes lower variation, this is a test of a single variance. The parameter is the population variance, σ^2, or the population standard deviation, σ.

Random Variable: The sample standard deviation, s, is the random variable. Let s = standard deviation for the waiting times.

- H_o: $\sigma^2 = 7.2^2$
- H_a: $\sigma^2 < 7.2^2$

The word **"lower"** tells you this is a left-tailed test.

Distribution for the test: χ^2_{24}, where:

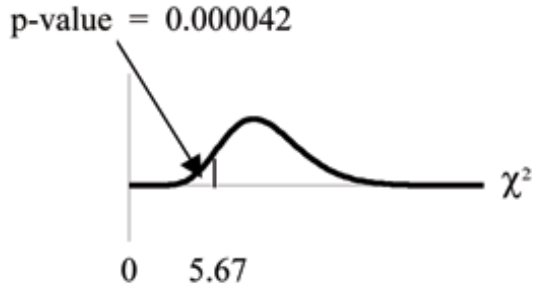
- n = the number of customers sampled
- $df = n - 1 = 25 - 1 = 24$

Calculate the test statistic:

$\chi^2 = \frac{(n-1)\cdot s^2}{\sigma^2} = \frac{(25-1)\cdot 3.5^2}{7.2^2} = 5.67$

where $n = 25$, $s = 3.5$, and $\sigma = 7.2$.

Graph:



Probability statement: $p\text{-}value = P\left(\chi^2 < 5.67\right) = 0.000042$

Compare α and the p-value: $\alpha = 0.05$ $p\text{-}value = 0.000042$ $\alpha > p\text{-}value$

Make a decision: Since $\alpha > p\text{-}value$, reject H_o.

This means that you reject $\sigma^2 = 7.2^2$. In other words, you do not think the variation in waiting times is 7.2 minutes, but lower.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that a single line causes a lower variation among the waiting times **or** with a single line, the customer waiting times vary less than 7.2 minutes.

TI-83+ and TI-84 calculators: In `2nd DISTR`, use `7:`χ`2cdf`. The syntax is `(lower, upper, df)` for the parameter list. For Example 11-9, χ`2cdf(-1E99,5.67,24)`. The $p\text{-}value = 0.000042$.

11.7 Summary of Formulas[7]

Formula 11.1: The Chi-square Probability Distribution
$\mu = df$ and $\sigma = \sqrt{2 \cdot df}$

Formula 11.2: Goodness-of-Fit Hypothesis Test

- Use goodness-of-fit to test whether a data set fits a particular probability distribution.
- The degrees of freedom are *number of cells or categories - 1.*
- The test statistic is $\sum_{n} \frac{(O-E)^2}{E}$, where O = observed values (data), E = expected values (from theory), and n = the number of different data cells or categories.
- The test is right-tailed.

Formula 11.3: Test of Independence

- Use the test of independence to test whether two factors are independent or not.
- The degrees of freedom are equal to *(number of columns - 1)(number of rows - 1).*
- The test statistic is $\sum_{(i \cdot j)} \frac{(O-E)^2}{E}$ where O = observed values, E = expected values, i = the number of rows in the table, and j = the number of columns in the table.
- The test is right-tailed.
- If the null hypothesis is true, the expected number $E = \frac{(row\ total)(column\ total)}{total\ surveyed}$.

Formula 11.4: Test of a Single Variance

- Use the test to determine variation.
- The degrees of freedom are the number of samples - 1.
- The test statistic is $\frac{(n-1) \cdot s^2}{\sigma^2}$, where n = the total number of data, s^2 = sample variance, and σ^2 = population variance.
- The test may be left, right, or two-tailed.

[7]This content is available online at <http://cnx.org/content/m17058/1.5/>.

11.8 Practice 1: Goodness-of-Fit Test[8]

11.8.1 Student Learning Outcomes

- The student will explore the properties of goodness-of-fit test data.

11.8.2 Given

The following data are real. The cumulative number of AIDS cases reported for Santa Clara County through December 31, 2003, is broken down by ethnicity as follows:

Ethnicity	**Number of Cases**
White	2032
Hispanic	897
African-American	372
Asian, Pacific Islander	168
Native American	20
	Total = 3489

Table 11.9

The percentage of each ethnic group in Santa Clara County is as follows:

Ethnicity	**Percentage of total county population**	**Number expected (round to 2 decimal places)**
White	47.79%	1667.39
Hispanic	24.15%	
African-American	3.55%	
Asian, Pacific Islander	24.21%	
Native American	0.29%	
	Total = 100%	

Table 11.10

11.8.3 Expected Results

If the ethnicity of AIDS victims followed the ethnicity of the total county population, fill in the expected number of cases per ethnic group.

[8]This content is available online at <http://cnx.org/content/m17054/1.8/>.

11.8.4 Goodness-of-Fit Test

Perform a goodness-of-fit test to determine whether the make-up of AIDS cases follows the ethnicity of the general population of Santa Clara County.

Exercise 11.8.1
H_o :

Exercise 11.8.2
H_a :

Exercise 11.8.3
Is this a right-tailed, left-tailed, or two-tailed test?

Exercise 11.8.4 ***(Solution on p. 470.)***
degrees of freedom =

Exercise 11.8.5 ***(Solution on p. 470.)***
Chi^2 test statistic =

Exercise 11.8.6 ***(Solution on p. 470.)***
p-value =

Exercise 11.8.7
Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.

Let $\alpha = 0.05$

Decision:

Reason for the Decision:

Conclusion (write out in complete sentences):

11.8.5 Discussion Question

Exercise 11.8.8
Does it appear that the pattern of AIDS cases in Santa Clara County corresponds to the distribution of ethnic groups in this county? Why or why not?

11.9 Practice 2: Contingency Tables[9]

11.9.1 Student Learning Outcomes

- The student will explore the properties of contingency tables.

Conduct a hypothesis test to determine if smoking level and ethnicity are independent.

11.9.2 Collect the Data

Copy the data provided in Probability Topics Practice 2: Calculating Probabilities into the table below.

Smoking Levels by Ethnicity (Observed)

Smoking Level Per Day	African American	Native Hawaiian	Latino	Japanese Americans	White	TOTALS
1-10						
11-20						
21-30						
31+						
TOTALS						

Table 11.11

11.9.3 Hypothesis

State the hypotheses.

- H_o :
- H_a :

11.9.4 Expected Values

Enter expected values in the above below. Round to two decimal places.

11.9.5 Analyze the Data

Calculate the following values:

Exercise 11.9.1 *(Solution on p. 470.)*
Degrees of freedom =

Exercise 11.9.2 *(Solution on p. 470.)*
Chi^2 test statistic =

Exercise 11.9.3 *(Solution on p. 470.)*
p-value =

Exercise 11.9.4 *(Solution on p. 470.)*
Is this a right-tailed, left-tailed, or two-tailed test? Explain why.

[9]This content is available online at <http://cnx.org/content/m17056/1.10/>.

11.9.6 Graph the Data

Exercise 11.9.5
Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.

11.9.7 Conclusions

State the decision and conclusion (in a complete sentence) for the following preconceived levels of α .

Exercise 11.9.6 ***(Solution on p. 470.)***
$\alpha = 0.05$

a. Decision:
b. Reason for the decision:
c. Conclusion (write out in a complete sentence):

Exercise 11.9.7
$\alpha = 0.01$

a. Decision:
b. Reason for the decision:
c. Conclusion (write out in a complete sentence):

11.10 Practice 3: Test of a Single Variance[10]

11.10.1 Student Learning Outcomes

- The student will explore the properties of data with a test of a single variance.

11.10.2 Given

Suppose an airline claims that its flights are consistently on time with an average delay of at most 15 minutes. It claims that the average delay is so consistent that the variance is no more than 150 minutes. Doubting the consistency part of the claim, a disgruntled traveler calculates the delays for his next 25 flights. The average delay for those 25 flights is 22 minutes with a standard deviation of 15 minutes.

11.10.3 Sample Variance

Exercise 11.10.1
Is the traveler disputing the claim about the average or about the variance?

Exercise 11.10.2 *(Solution on p. 470.)*
A sample standard deviation of 15 minutes is the same as a sample variance of _____ minutes.

Exercise 11.10.3
Is this a right-tailed, left-tailed, or two-tailed test?

11.10.4 Hypothesis Test

Perform a hypothesis test on the consistency part of the claim.

Exercise 11.10.4
H_o :

Exercise 11.10.5
H_a :

Exercise 11.10.6 *(Solution on p. 470.)*
Degrees of freedom =

Exercise 11.10.7 *(Solution on p. 470.)*
Chi^2 test statistic =

Exercise 11.10.8 *(Solution on p. 470.)*
p-value =

Exercise 11.10.9
Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade the p-value.

[10]This content is available online at <http://cnx.org/content/m17053/1.7/>.

Exercise 11.10.10
Let $\alpha = 0.05$

Decision:

Conclusion (write out in a complete sentence):

11.10.5 Discussion Questions

Exercise 11.10.11
How did you know to test the variance instead of the mean?

Exercise 11.10.12
If an additional test were done on the claim of the average delay, which distribution would you use?

Exercise 11.10.13
If an additional test was done on the claim of the average delay, but 45 flights were surveyed, which distribution would you use?

11.11 Homework[11]

Exercise 11.11.1

a. Explain why the "goodness of fit" test and the "test for independence" are generally right tailed tests.
b. If you did a left-tailed test, what would you be testing?

11.11.1 Word Problems

For each word problem, use a solution sheet to solve the hypothesis test problem. Go to The Table of Contents 14. Appendix for the solution sheet. Round expected frequency to two decimal places.

Exercise 11.11.2
A 6-sided die is rolled 120 times. Fill in the expected frequency column. Then, conduct a hypothesis test to determine if the die is fair. The data below are the result of the 120 rolls.

Face Value	Frequency	Expected Frequency
1	15	
2	29	
3	16	
4	15	
5	30	
6	15	

Table 11.12

Exercise 11.11.3 ***(Solution on p. 470.)***
The marital status distribution of the U.S. male population, age 15 and older, is as shown below. (*Source: U.S. Census Bureau, Current Population Reports*)

Marital Status	Percent	Expected Frequency
never married	31.3	
married	56.1	
widowed	2.5	
divorced/separated	10.1	

Table 11.13

Suppose that a random sample of 400 U.S. young adult males, 18 – 24 years old, yielded the following frequency distribution. We are interested in whether this age group of males fits the distribution of the U.S. adult population. Calculate the frequency one would expect when surveying 400 people. Fill in the above table, rounding to two decimal places.

[11]This content is available online at <http://cnx.org/content/m17028/1.10/>.

Marital Status	Frequency
never married	140
married	238
widowed	2
divorced/separated	20

Table 11.14

The next two questions refer to the following information: The real data below are from the California Reinvestment Committee and the California Economic Census. The data concern the percent of loans made by the Small Business Administration for Santa Clara County in recent years. (*Source: San Jose Mercury News*)

Ethnic Group	Percent of Loans	Percent of Population	Percent of Businesses Owned
Asian	22.48	16.79	12.17
Black	1.15	3.51	1.61
Latino	6.19	21.00	6.51
White	66.97	58.09	79.70

Table 11.15

Exercise 11.11.4
Perform a goodness-of-fit test to determine whether the percent of businesses owned in Santa Clara County fits the percent of the population, based on ethnicity.

Exercise 11.11.5 ***(Solution on p. 470.)***
Perform a goodness-of-fit test to determine whether the percent of loans fits the percent of the businesses owned in Santa Clara County, based on ethnicity.

Exercise 11.11.6
The City of South Lake Tahoe has an Asian population of 1419 people, out of a total population of 23,609 (*Source: U.S. Census Bureau, Census 2000*). Conduct a goodness of fit test to determine if the self-reported sub-groups of Asians are evenly distributed.

Race	Frequency	Expected Frequency
Asian Indian	131	
Chinese	118	
Filipino	1045	
Japanese	80	
Korean	12	
Vietnamese	9	
Other	24	

Table 11.16

Exercise 11.11.7 ***(Solution on p. 471.)***

Long Beach is a city in Los Angeles County (L.A.C). The population of Long Beach is 461,522; the population of L.A.C. is 9,519,338 (*Source: U.S. Census Bureau, Census 2000*). Conduct a goodness of fit test to determine if the racial demographics of Long Beach fit that of L.A.C.

Race	Percent, L.A.C.	Expected #, L.B.	Actual #, L.B.
American Indian and Alaska Native	0.8	3692	3,881
Asian	11.9		55,591
Black or African American	9.8		68,618
Native Hawaiian and Other Pacific Islander	0.3		5,605
White, including Hispanic/Latino	48.7		208,410
Other	23.5		95,107
Two or more races	5.0		24,310

Table 11.17

Exercise 11.11.8

UCLA conducted a survey of more than 263,000 college freshmen from 385 colleges in fall 2005. The results of student expected majors by gender were reported in *The Chronicle of Higher Education (2/2/06)*. Conduct a goodness of fit test to determine if the male distribution fits the female distribution.

Major	Women	Men
Arts & Humanities	14.0%	11.4%
Biological Sciences	8.4%	6.7%
Business	13.1%	22.7%
Education	13.0%	5.8%
Engineering	2.6%	15.6%
Physical Sciences	2.6%	3.6%
Professional	18.9%	9.3%
Social Sciences	13.0%	7.6%
Technical	0.4%	1.8%
Other	5.8%	8.2%
Undecided	8.0%	6.6%

Table 11.18

Exercise 11.11.9 ***(Solution on p. 471.)***

A recent debate about where in the United States skiers believe the skiing is best prompted the following survey. Test to see if the best ski area is independent of the level of the skier.

U.S. Ski Area	Beginner	Intermediate	Advanced
Tahoe	20	30	40
Utah	10	30	60
Colorado	10	40	50

Table 11.19

Exercise 11.11.10
Car manufacturers are interested in whether there is a relationship between the size of car an individual drives and the number of people in the driver's family (that is, whether car size and family size are independent). To test this, suppose that 800 car owners were randomly surveyed with the following results. Conduct a test for independence.

Family Size	Sub & Compact	Mid-size	Full-size	Van & Truck
1	20	35	40	35
2	20	50	70	80
3 - 4	20	50	100	90
5+	20	30	70	70

Table 11.20

Exercise 11.11.11 ***(Solution on p. 471.)***
College students may be interested in whether or not their majors have any effect on starting salaries after graduation. Suppose that 300 recent graduates were surveyed as to their majors in college and their starting salaries after graduation. Below are the data. Conduct a test for independence.

Major	$<$ $30,000	$30,000 - $39,999	$40,000 +
English	5	20	5
Engineering	10	30	60
Nursing	10	15	15
Business	10	20	30
Psychology	20	30	20

Table 11.21

Exercise 11.11.12
Some travel agents claim that honeymoon hot spots vary according to age of the bride and groom. Suppose that 280 East Coast recent brides were interviewed as to where they spent their honeymoons. The information is given below. Conduct a test for independence.

Location	20 - 29	30 - 39	40 - 49	50 and over
Niagara Falls	15	25	25	20
Poconos	15	25	25	10
Europe	10	25	15	5
Virgin Islands	20	25	15	5

Table 11.22

Exercise 11.11.13 ***(Solution on p. 471.)***

A manager of a sports club keeps information concerning the main sport in which members participate and their ages. To test whether there is a relationship between the age of a member and his or her choice of sport, 643 members of the sports club are randomly selected. Conduct a test for independence.

Sport	18 - 25	26 - 30	31 - 40	41 and over
racquetball	42	58	30	46
tennis	58	76	38	65
swimming	72	60	65	33

Table 11.23

Exercise 11.11.14

A major food manufacturer is concerned that the sales for its skinny French fries have been decreasing. As a part of a feasibility study, the company conducts research into the types of fries sold across the country to determine if the type of fries sold is independent of the area of the country. The results of the study are below. Conduct a test for independence.

Type of Fries	Northeast	South	Central	West
skinny fries	70	50	20	25
curly fries	100	60	15	30
steak fries	20	40	10	10

Table 11.24

Exercise 11.11.15 ***(Solution on p. 471.)***

According to Dan Lenard, an independent insurance agent in the Buffalo, N.Y. area, the following is a breakdown of the amount of life insurance purchased by males in the following age groups. He is interested in whether the age of the male and the amount of life insurance purchased are independent events. Conduct a test for independence.

Age of Males	None	$50,000 - $100,000	$100,001 - $150,000	$150,001 - $200,000	$200,000 +
20 - 29	40	15	40	0	5
30 - 39	35	5	20	20	10
40 - 49	20	0	30	0	30
50 +	40	30	15	15	10

Table 11.25

Exercise 11.11.16
Suppose that 600 thirty–year–olds were surveyed to determine whether or not there is a relationship between the level of education an individual has and salary. Conduct a test for independence.

Annual Salary	Not a high school grad.	High school graduate	College graduate	Masters or doctorate
< $30,000	15	25	10	5
$30,000 - $40,000	20	40	70	30
$40,000 - $50,000	10	20	40	55
$50,000 - $60,000	5	10	20	60
$60,000 +	0	5	10	150

Table 11.26

Exercise 11.11.17 ***(Solution on p. 471.)***
A plant manager is concerned her equipment may need recalibrating. It seems that the actual weight of the 15 oz. cereal boxes it fills has been fluctuating. The standard deviation should be at most $\frac{1}{2}$ oz. In order to determine if the machine needs to be recalibrated, 84 randomly selected boxes of cereal from the next day's production were weighed. The standard deviation of the 84 boxes was 0.54. Does the machine need to be recalibrated?

Exercise 11.11.18
Consumers may be interested in whether the cost of a particular calculator varies from store to store. Based on surveying 43 stores, which yielded a sample mean of $84 and a sample standard deviation of $12, test the claim that the standard deviation is greater than $15.

Exercise 11.11.19 ***(Solution on p. 471.)***
Isabella, an accomplished **Bay to Breakers** runner, claims that the standard deviation for her time to run the 7 ½ mile race is at most 3 minutes. To test her claim, Rupinder looks up 5 of her race times. They are 55 minutes, 61 minutes, 58 minutes, 63 minutes, and 57 minutes.

Exercise 11.11.20
Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. They are also interested in the variation of the number of babies. Suppose that an airline executive believes the average number of babies on flights is 6 with a variance of 9 at most. The airline conducts a survey. The results of the 18 flights surveyed give a sample average of 6.4 with a sample standard deviation of 3.9. Conduct a hypothesis test of the airline executive's belief.

Exercise 11.11.21 ***(Solution on p. 472.)***
According to the *U.S. Bureau of the Census, United Nations,* in 1994 the number of births per woman in China was 1.8. This fertility rate has been attributed to the law passed in 1979 restricting

births to one per woman. Suppose that a group of students studied whether or not the standard deviation of births per woman was greater than 0.75. They asked 50 women across China the number of births they had. Below are the results. Does the students' survey indicate that the standard deviation is greater than 0.75?

# of births	Frequency
0	5
1	30
2	10
3	5

Table 11.27

Exercise 11.11.22
According to an avid aquariest, the average number of fish in a 20–gallon tank is 10, with a standard deviation of 2. His friend, also an aquariest, does not believe that the standard deviation is 2. She counts the number of fish in 15 other 20–gallon tanks. Based on the results that follow, do you think that the standard deviation is different from 2? Data: 11; 10; 9; 10; 10; 11; 11; 10; 12; 9; 7; 9; 11; 10; 11

Exercise 11.11.23 ***(Solution on p. 472.)***
The manager of "Frenchies" is concerned that patrons are not consistently receiving the same amount of French fries with each order. The chef claims that the standard deviation for a 10–ounce order of fries is at most 1.5 oz., but the manager thinks that it may be higher. He randomly weighs 49 orders of fries, which yields: mean of 11 oz., standard deviation of 2 oz.

11.11.2 Try these true/false questions.

Exercise 11.11.24 ***(Solution on p. 472.)***
As the degrees of freedom increase, the graph of the chi-square distribution looks more and more symmetrical.

Exercise 11.11.25 ***(Solution on p. 472.)***
The standard deviation of the chi-square distribution is twice the mean.

Exercise 11.11.26 ***(Solution on p. 472.)***
The mean and the median of the chi-square distribution are the same if $df = 24$.

Exercise 11.11.27 ***(Solution on p. 472.)***
In a Goodness-of-Fit test, the expected values are the values we would expect if the null hypothesis were true.

Exercise 11.11.28 ***(Solution on p. 472.)***
In general, if the observed values and expected values of a Goodness-of-Fit test are not close together, then the test statistic can get very large and on a graph will be way out in the right tail.

Exercise 11.11.29 ***(Solution on p. 472.)***
The degrees of freedom for a Test for Independence are equal to the sample size minus 1.

Exercise 11.11.30 ***(Solution on p. 472.)***
Use a Goodness-of-Fit test to determine if high school principals believe that students are absent equally during the week or not.

Exercise 11.11.31 *(Solution on p. 472.)*
The Test for Independence uses tables of observed and expected data values.

Exercise 11.11.32 *(Solution on p. 472.)*
The test to use when determining if the college or university a student chooses to attend is related to his/her socioeconomic status is a Test for Independence.

Exercise 11.11.33 *(Solution on p. 472.)*
The test to use to determine if a coin is fair is a Goodness-of-Fit test.

Exercise 11.11.34 *(Solution on p. 472.)*
In a Test of Independence, the expected number is equal to the row total multiplied by the column total divided by the total surveyed.

Exercise 11.11.35 *(Solution on p. 472.)*
In a Goodness-of Fit test, if the p-value is 0.0113, in general, do not reject the null hypothesis.

Exercise 11.11.36 *(Solution on p. 472.)*
For a Chi-Square distribution with degrees of freedom of 17, the probability that a value is greater than 20 is 0.7258.

Exercise 11.11.37 *(Solution on p. 472.)*
If $df = 2$, the chi-square distribution has a shape that reminds us of the exponential.

11.12 Review[12]

The next two questions refer to the following real study:

A recent survey of U.S. teenage pregnancy was answered by 720 girls, age 12 - 19. 6% of the girls surveyed said they have been pregnant. (*Parade Magazine*) We are interested in the true proportion of U.S. girls, age 12 - 19, who have been pregnant.

Exercise 11.12.1 ***(Solution on p. 472.)***
Find the 95% confidence interval for the true proportion of U.S. girls, age 12 - 19, who have been pregnant.

Exercise 11.12.2 ***(Solution on p. 472.)***
The report also stated that the results of the survey are accurate to within $\pm$ 3.7% at the 95% confidence level. Suppose that a new study is to be done. It is desired to be accurate to within 2% of the 95% confidence level. What will happen to the minimum number that should be surveyed?

Exercise 11.12.3
Given: $X \sim Exp\left(\frac{1}{3}\right)$. Sketch the graph that depicts: $P(X > 1)$.

The next four questions refer to the following information:

Suppose that the time that owners keep their cars (purchased new) is normally distributed with a mean of 7 years and a standard deviation of 2 years. We are interested in how long an individual keeps his car (purchased new). Our population is people who buy their cars new.

Exercise 11.12.4 ***(Solution on p. 472.)***
60% of individuals keep their cars **at most** how many years?

Exercise 11.12.5 ***(Solution on p. 473.)***
Suppose that we randomly survey one person. Find the probability that person keeps his/her car **less than** 2.5 years.

Exercise 11.12.6 ***(Solution on p. 473.)***
If we are to pick individuals 10 at a time, find the distribution for the **average** car length ownership.

Exercise 11.12.7 ***(Solution on p. 473.)***
If we are to pick 10 individuals, find the probability that the **sum** of their ownership time is more than 55 years.

Exercise 11.12.8 ***(Solution on p. 473.)***
For which distribution is the median not equal to the mean?

A. Uniform
B. Exponential
C. Normal
D. Student-t

Exercise 11.12.9 ***(Solution on p. 473.)***
Compare the standard normal distribution to the student-t distribution, centered at 0. Explain which of the following are true and which are false.

a. As the number surveyed increases, the area to the left of -1 for the student-t distribution approaches the area for the standard normal distribution.
b. As the number surveyed increases, the area to the left of -1 for the standard normal distribution approaches the area for the student-t distribution.

[12]This content is available online at <http://cnx.org/content/m17057/1.8/>.

c. As the degrees of freedom decrease, the graph of the student-t distribution looks more like the graph of the standard normal distribution.
d. If the number surveyed is less than 30, the normal distribution should never be used.

The next five questions refer to the following information:

We are interested in the checking account balance of a twenty-year-old college student. We randomly survey 16 twenty-year-old college students. We obtain a sample mean of \$640 and a sample standard deviation of \$150. Let X = checking account balance of an individual twenty year old college student.

Exercise 11.12.10
Explain why we cannot determine the distribution of X.

Exercise 11.12.11 ***(Solution on p. 473.)***
If you were to create a confidence interval or perform a hypothesis test for the population average checking account balance of 20-year old college students, what distribution would you use?

Exercise 11.12.12 ***(Solution on p. 473.)***
Find the 95% confidence interval for the true average checking account balance of a twenty-year-old college student.

Exercise 11.12.13 ***(Solution on p. 473.)***
What type of data is the balance of the checking account considered to be?

Exercise 11.12.14 ***(Solution on p. 473.)***
What type of data is the number of 20 year olds considered to be?

Exercise 11.12.15 ***(Solution on p. 473.)***
On average, a busy emergency room gets a patient with a shotgun wound about once per week. We are interested in the number of patients with a shotgun wound the emergency room gets per 28 days.

a. Define the random variable X.
b. State the distribution for X.
c. Find the probability that the emergency room gets no patients with shotgun wounds in the next 28 days.

The next two questions refer to the following information:

The probability that a certain slot machine will pay back money when a quarter is inserted is 0.30 . Assume that each play of the slot machine is independent from each other. A person puts in 15 quarters for 15 plays.

Exercise 11.12.16 ***(Solution on p. 473.)***
Is the expected number of plays of the slot machine that will pay back money greater than, less than or the same as the median? Explain your answer.

Exercise 11.12.17 ***(Solution on p. 473.)***
Is it likely that exactly 8 of the 15 plays would pay back money? Justify your answer numerically.

Exercise 11.12.18 ***(Solution on p. 473.)***
A game is played with the following rules:

- it costs \$10 to enter
- a fair coin is tossed 4 times
- if you do not get 4 heads or 4 tails, you lose your \$10
- if you get 4 heads or 4 tails, you get back your \$10, plus \$30 more

Over the long run of playing this game, what are your expected earnings?

Exercise 11.12.19 ***(Solution on p. 473.)***

- The average grade on a math exam in Rachel's class was 74, with a standard deviation of 5. Rachel earned an 80.
- The average grade on a math exam in Becca's class was 47, with a standard deviation of 2. Becca earned a 51.
- The average grade on a math exam in Matt's class was 70, with a standard deviation of 8. Matt earned an 83.

Find whose score was the best, compared to his or her own class. Justify your answer numerically.

The next two questions refer to the following information:

70 compulsive gamblers were asked the number of days they go to casinos per week. The results are given in the following graph:

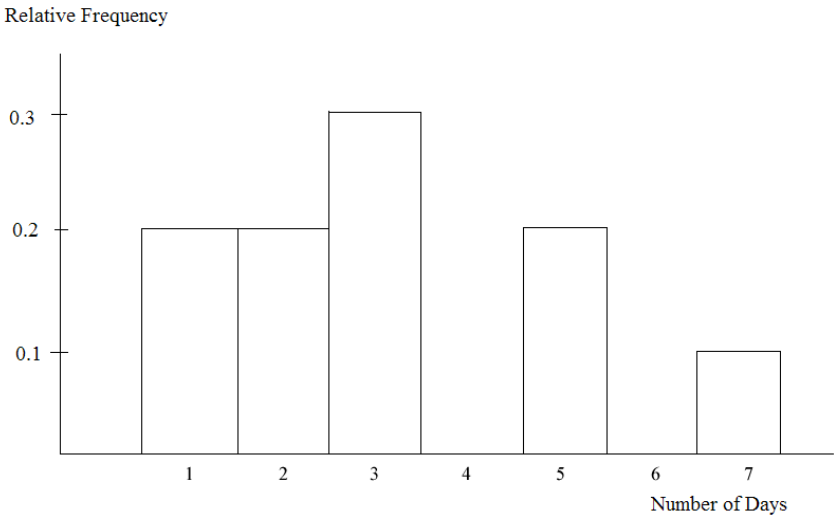


Figure 11.3

Exercise 11.12.20 ***(Solution on p. 473.)***

Find the number of responses that were "5".

Exercise 11.12.21 ***(Solution on p. 473.)***

Find the mean, standard deviation, all four quartiles and IQR.

Exercise 11.12.22 ***(Solution on p. 473.)***

Based upon research at De Anza College, it is believed that about 19% of the student population speaks a language other than English at home.

Suppose that a study was done this year to see if that percent has decreased. Ninety-eight students were randomly surveyed with the following results. Fourteen said that they speak a language other than English at home.

a. State an appropriate **null** hypothesis.
b. State an appropriate **alternate** hypothesis.

c. Define the Random Variable, P'.
d. Calculate the test statistic.
e. Calculate the p-value.
f. At the 5% level of decision, what is your decision about the null hypothesis?
g. What is the Type I error?
h. What is the Type II error?

Exercise 11.12.23
Assume that you are an emergency paramedic called in to rescue victims of an accident. You need to help a patient who is bleeding profusely. The patient is also considered to be a high risk for contracting AIDS. Assume that the null hypothesis is that the patient does **not** have the HIV virus. What is a Type I error?

Exercise 11.12.24 ***(Solution on p. 473.)***
It is often said that Californians are more casual than the rest of Americans. Suppose that a survey was done to see if the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals. Fifty of each was surveyed with the following results. 10 Californians wear jeans to work and 4 non-Californians wear jeans to work.

- C = Californian professional
- NC = non-Californian professional

a. State appropriate **null** and **alternate** hypotheses.
b. Define the Random Variable.
c. Calculate the test statistic and p-value.
d. At the 5% level of decision, do you accept or reject the null hypothesis?
e. What is the Type I error?
f. What is the Type II error?

The next two questions refer to the following information:

A group of Statistics students have developed a technique that they feel will lower their anxiety level on statistics exams. They measured their anxiety level at the start of the quarter and again at the end of the quarter. Recorded is the paired data in that order: (1000, 900); (1200, 1050); (600, 700); (1300, 1100); (1000, 900); (900, 900).

Exercise 11.12.25 ***(Solution on p. 474.)***
This is a test of (pick the best answer):

A. large samples, independent means
B. small samples, independent means
C. dependent means

Exercise 11.12.26 ***(Solution on p. 474.)***
State the distribution to use for the test.

11.13 Lab 1: Chi-Square Goodness-of-Fit[13]

Class Time:

Names:

11.13.1 Student Learning Outcome:

- The student will evaluate data collected to determine if they fit either the uniform or exponential distributions.

11.13.2 Collect the Data

Go to your local supermarket. Ask 30 people as they leave for the total amount on their grocery receipts. (Or, ask 3 cashiers for the last 10 amounts. Be sure to include the express lane, if it is open.)

1. Record the values.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Table 11.28

2. Construct a histogram of the data. Make 5 - 6 intervals. Sketch the graph using a ruler and pencil. Scale the axes.

[13]This content is available online at <http://cnx.org/content/m17049/1.8/>.

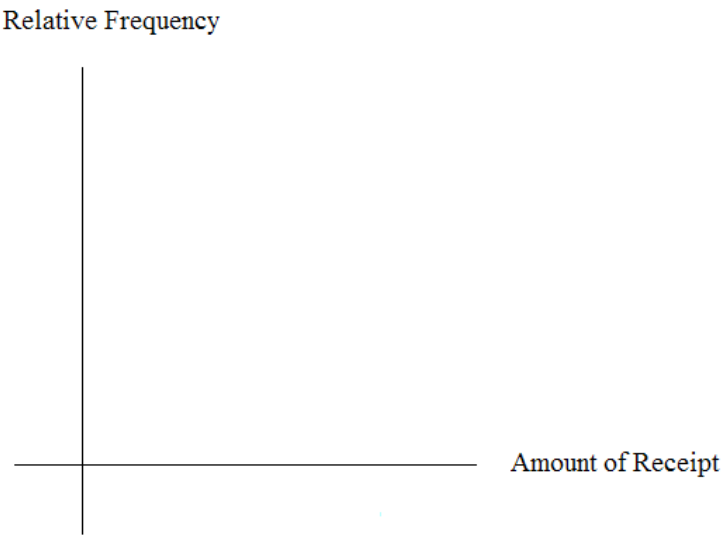


Figure 11.4

3. Calculate the following:
   a. $\overline{x} =$
   b. $s =$
   c. $s^2 =$

11.13.3 Uniform Distribution

Test to see if grocery receipts follow the uniform distribution.

1. Using your lowest and highest values, $X \sim U$ (_____,_____)
2. Divide the distribution above into fifths.
3. Calculate the following:
   a. Lowest value =
   b. 20th percentile =
   c. 40th percentile =
   d. 60th percentile =
   e. 80th percentile =
   f. Highest value =
4. For each fifth, count the observed number of receipts and record it. Then determine the expected number of receipts and record that.

Fifth	Observed	Expected
1st		
2nd		
3rd		
4th		
5th		

Table 11.29

5. H_o:
6. H_a:
7. What distribution should you use for a hypothesis test?
8. Why did you choose this distribution?
9. Calculate the test statistic.
10. Find the p-value.
11. Sketch a graph of the situation. Label and scale the x-axis. Shade the area corresponding to the p-value.

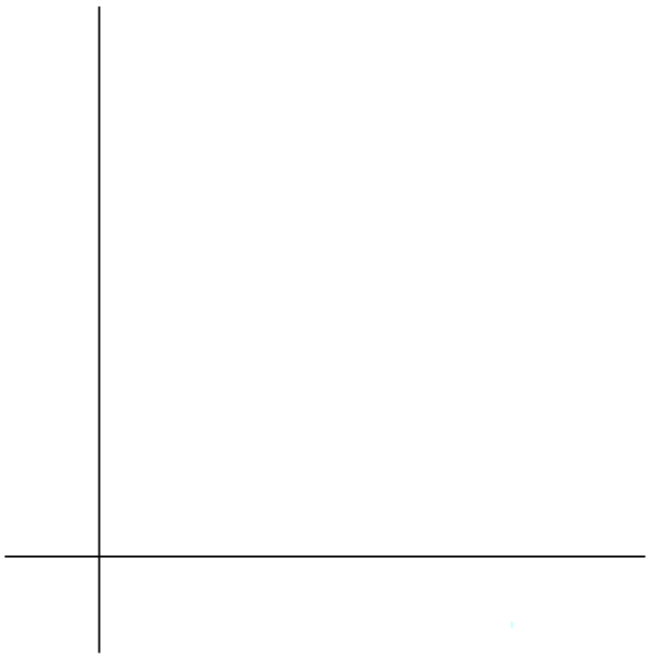


Figure 11.5

12. State your decision.
13. State your conclusion in a complete sentence.

11.13.4 Exponential Distribution

Test to see if grocery receipts follow the exponential distribution with decay parameter $\frac{1}{\bar{x}}$.

1. Using $\frac{1}{\bar{x}}$ as the decay parameter, $X \sim Exp$ (_____).
2. Calculate the following:
   a. Lowest value =
   b. First quartile =
   c. 37th percentile =
   d. Median =
   e. 63rd percentile =
   f. 3rd quartile =
   g. Highest value =
3. For each cell, count the observed number of receipts and record it. Then determine the expected number of receipts and record that.

Cell	Observed	Expected
1st		
2nd		
3rd		
4th		
5th		
6th		

Table 11.30

4. H_o
5. H_a
6. What distribution should you use for a hypothesis test?
7. Why did you choose this distribution?
8. Calculate the test statistic.
9. Find the p-value.
10. Sketch a graph of the situation. Label and scale the x-axis. Shade the area corresponding to the p-value.

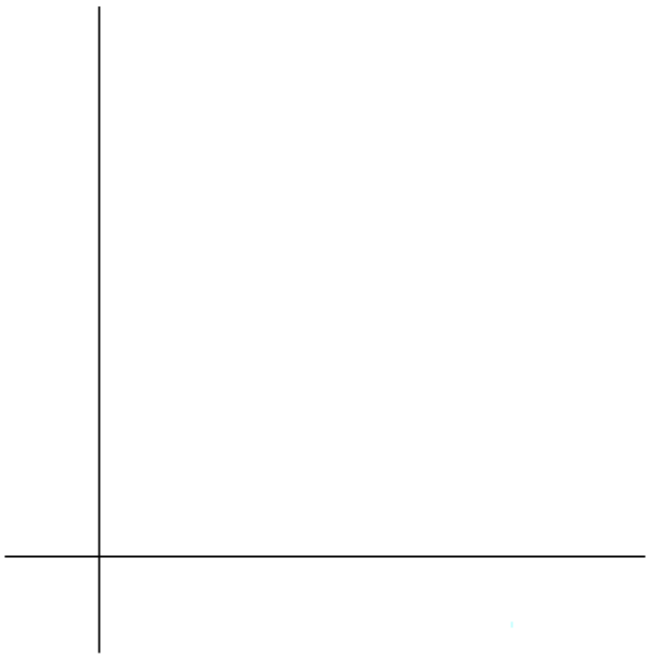


Figure 11.6

11. State your decision.
12. State your conclusion in a complete sentence.

11.13.5 Discussion Questions

1. Did your data fit either distribution? If so, which?
2. In general, do you think it's likely that data could fit more than one distribution? In complete sentences, explain why or why not.

11.14 Lab 2: Chi-Square Test for Independence[14]

Class Time:

Names:

11.14.1 Student Learning Outcome:

- The student will evaluate if there is a significant relationship between favorite type of snack and gender.

11.14.2 Collect the Data

1. Using your class as a sample, complete the following chart.

Favorite type of snack

	sweets (candy & baked goods)	ice cream	chips & pretzels	fruits & vegetables	Total
male					
female					
Total					

Table 11.31

2. Looking at the above chart, does it appear to you that there is dependence between gender and favorite type of snack food? Why or why not?

11.14.3 Hypothesis Test

Conduct a hypothesis test to determine if the factors are independent

1. H_o:
2. H_a:
3. What distribution should you use for a hypothesis test?
4. Why did you choose this distribution?
5. Calculate the test statistic.
6. Find the p-value.
7. Sketch a graph of the situation. Label and scale the x-axis. Shade the area corresponding to the p-value.

[14]This content is available online at <http://cnx.org/content/m17050/1.9/>.

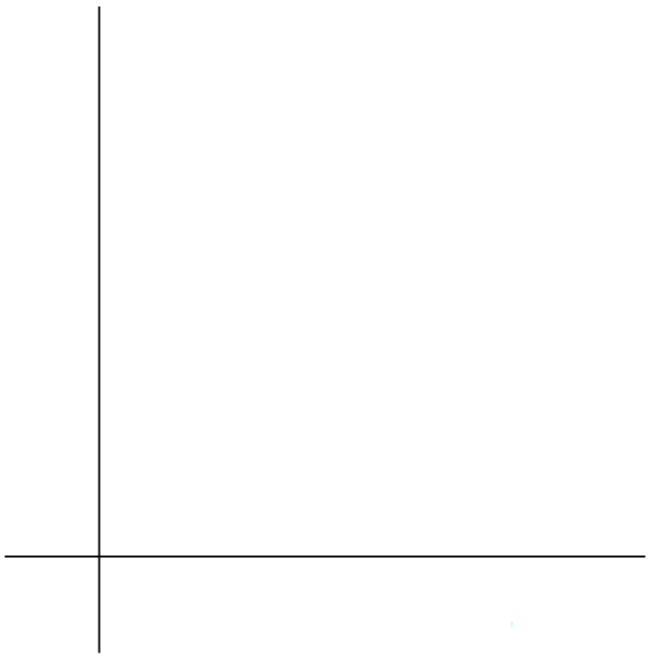


Figure 11.7

8. State your decision.
9. State your conclusion in a complete sentence.

11.14.4 Discussion Questions

1. Is the conclusion of your study the same as or different from your answer to (I2) above?
2. Why do you think that occurred?

Solutions to Exercises in Chapter 11

Solution to Example 11.7, Problem 3 (p. 441)

a. $E = \frac{(row\ total)(column\ total)}{total\ surveyed} = 8.19$
b. 8

Solutions to Practice 1: Goodness-of-Fit Test

Solution to Exercise 11.8.4 (p. 446)
degrees of freedom = 4
Solution to Exercise 11.8.5 (p. 446)
951.69
Solution to Exercise 11.8.6 (p. 446)
0

Solutions to Practice 2: Contingency Tables

Solution to Exercise 11.9.1 (p. 447)
12
Solution to Exercise 11.9.2 (p. 447)
10301.8
Solution to Exercise 11.9.3 (p. 447)
0
Solution to Exercise 11.9.4 (p. 447)
right
Solution to Exercise 11.9.6 (p. 448)

a. Reject the null hypothesis

Solutions to Practice 3: Test of a Single Variance

Solution to Exercise 11.10.2 (p. 449)
225
Solution to Exercise 11.10.6 (p. 449)
24
Solution to Exercise 11.10.7 (p. 449)
36
Solution to Exercise 11.10.8 (p. 449)
0.0549

Solutions to Homework

Solution to Exercise 11.11.3 (p. 451)

a. The data fits the distribution
b. The data does not fit the distribution
c. 3
e. 19.27
f. 0.0002
h. Decision: Reject Null; Conclusion: Data does not fit the distribution.

Solution to Exercise 11.11.5 (p. 452)

c. 3
e. 10.91
f. 0.0122
g. Decision: Reject null when $a = 0.05$; Conclusion: Percent of loans does not fit the distribution. Decision: Do not reject null when $a = 0.01$; Conclusion Percent of loans fits the distribution.

Solution to Exercise 11.11.7 (p. 453)

c. 6
e. 27,876
f. 0
h. Decision: Reject null; Conclusion: L.B. does not fit L.A.C.

Solution to Exercise 11.11.9 (p. 453)

c. 4
e. 10.53
f. 0.0324
h. Decision: Reject null; Conclusion: Best ski area and level of skier are not independent.

Solution to Exercise 11.11.11 (p. 454)

c. 8
e. 33.55
f. 0
h. Decision: Reject null; Conclusion: Major and starting salary are not independent events.

Solution to Exercise 11.11.13 (p. 455)

c. 6
e. 25.21
f. 0.0003
h. Decision: Reject null

Solution to Exercise 11.11.15 (p. 455)

c. 12
e. 125.74
f. 0
h. Decision: Reject null

Solution to Exercise 11.11.17 (p. 456)

c. 83
d. 96.81
e. 0.1426
g. Decision: Do not reject null; Conclusion: The standard deviation is at most 0.5 oz.
h. It does not need to be calibrated

Solution to Exercise 11.11.19 (p. 456)

c. 4
d. 4.52
e. 0.3402
g. Decision: Do not reject null.
h. No

Solution to Exercise 11.11.21 (p. 456)

c. 49
d. 54.37
e. 0.2774
g. Decision: Do not reject null; Conclusion: The standard deviation is at most 0.75.
h. No

Solution to Exercise 11.11.23 (p. 457)

a. $\sigma^2 \leq (1.5)^2$
c. 48
d. 85.33
e. 0.0007
g. Decision: Reject null.
h. Yes

Solution to Exercise 11.11.24 (p. 457)
True
Solution to Exercise 11.11.25 (p. 457)
False
Solution to Exercise 11.11.26 (p. 457)
False
Solution to Exercise 11.11.27 (p. 457)
True
Solution to Exercise 11.11.28 (p. 457)
True
Solution to Exercise 11.11.29 (p. 457)
False
Solution to Exercise 11.11.30 (p. 457)
True
Solution to Exercise 11.11.31 (p. 458)
True
Solution to Exercise 11.11.32 (p. 458)
True
Solution to Exercise 11.11.33 (p. 458)
True
Solution to Exercise 11.11.34 (p. 458)
True
Solution to Exercise 11.11.35 (p. 458)
False
Solution to Exercise 11.11.36 (p. 458)
False
Solution to Exercise 11.11.37 (p. 458)
True

Solutions to Review

Solution to Exercise 11.12.1 (p. 459)
$(0.0424, 0.0770)$
Solution to Exercise 11.12.2 (p. 459)
2401

Solution to Exercise 11.12.4 (p. 459)
7.5
Solution to Exercise 11.12.5 (p. 459)
0.0122
Solution to Exercise 11.12.6 (p. 459)
$N(7, 0.63)$
Solution to Exercise 11.12.7 (p. 459)
0.9911
Solution to Exercise 11.12.8 (p. 459)
B
Solution to Exercise 11.12.9 (p. 459)

a. True
b. False
c. False
d. False

Solution to Exercise 11.12.11 (p. 460)
student-t with $df = 15$
Solution to Exercise 11.12.12 (p. 460)
$(560.07, 719.93)$
Solution to Exercise 11.12.13 (p. 460)
quantitative - continuous
Solution to Exercise 11.12.14 (p. 460)
quantitative - discrete
Solution to Exercise 11.12.15 (p. 460)

b. $P(4)$
c. 0.0183

Solution to Exercise 11.12.16 (p. 460)
greater than
Solution to Exercise 11.12.17 (p. 460)
No; $P(X = 8) = 0.0348$
Solution to Exercise 11.12.18 (p. 460)
You will lose \$5
Solution to Exercise 11.12.19 (p. 461)
Becca
Solution to Exercise 11.12.20 (p. 461)
14
Solution to Exercise 11.12.21 (p. 461)

- . Mean = 3.2
- . Quartiles = 1.85, 2, 3, and 5
- . IQR = 3

Solution to Exercise 11.12.22 (p. 461)

d. $z = -1.19$
e. 0.1171
f. Do not reject the null

Solution to Exercise 11.12.24 (p. 462)

c. $z = 1.73$; $p = 0.0419$

d. Reject the null

Solution to Exercise 11.12.25 (p. 462)
C
Solution to Exercise 11.12.26 (p. 462)
t_5