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://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:

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

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://http://cnx.org/content/m17052/1.5/>.

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

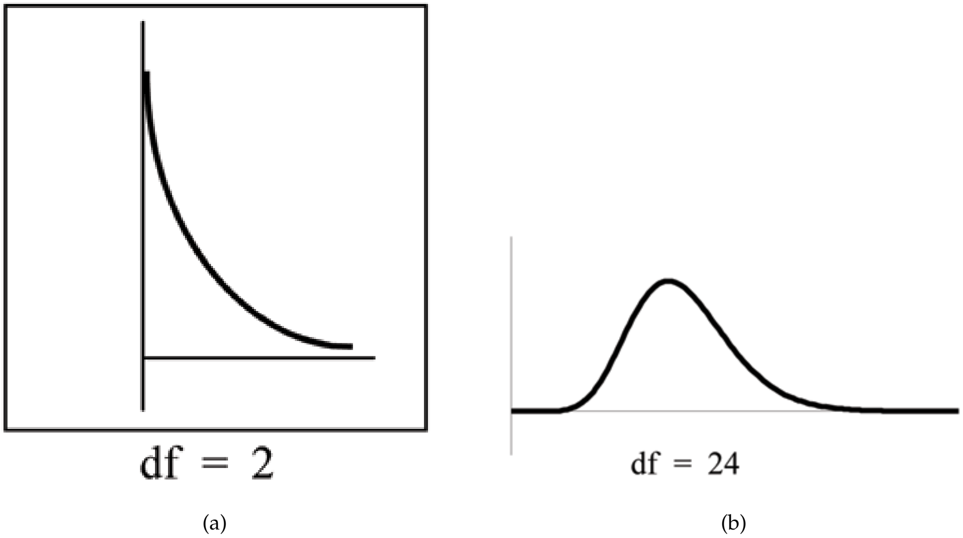


Figure 11.1

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.

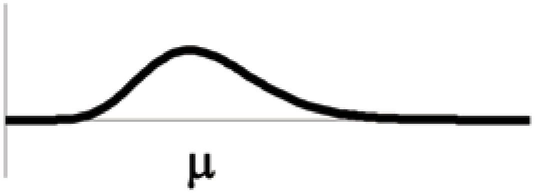


Figure 11.2

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://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_0: 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-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.

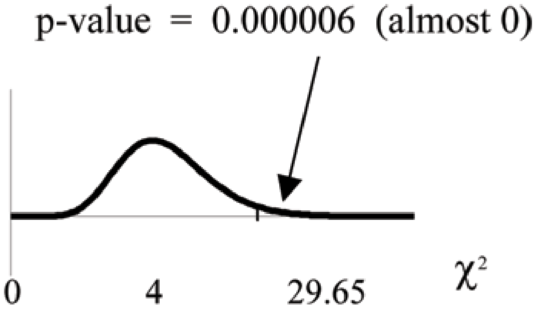
H_a: The "number of televisions" distribution of far western United States families is different from the "number of televisions" distribution of the American population.

Distribution for the test: χ^2_4 where $df =$ (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-value $= P\left(\chi^2 > 29.65\right) = 0.000006$.

Compare α and the p-value:

- $\alpha = 0.01$
- p-value $= 0.000006$

So, $\alpha >$ p-value.

Make a decision: Since $\alpha >$ p-value, reject H_0.

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:χ2cdf` 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.

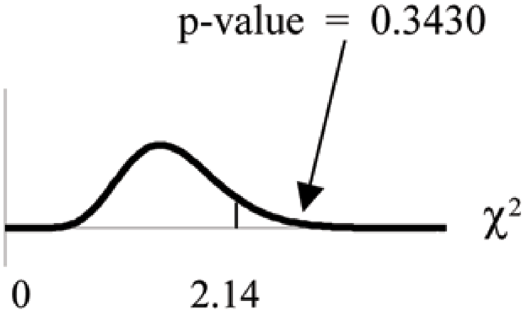
H_o: The coins are fair.

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-value $= P\left(\chi^2 > 2.14\right) = 0.3430$

Compare α and the p-value:

- $\alpha = 0.05$
- p-value $= 0.3430$

So, $\alpha <$ p-value.

Make a decision: Since $\alpha <$ p-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://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):

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

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

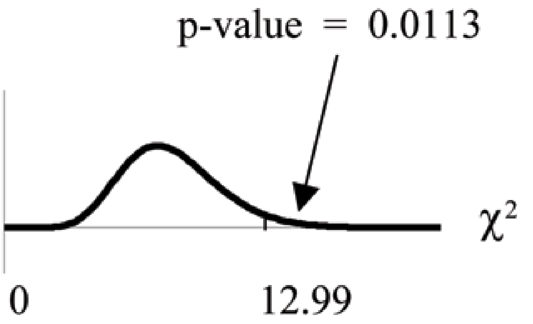
$E = \frac{\text{(row total)(column total)}}{\text{total number surveyed}} = \frac{255 \cdot 298}{839} = 90.57$

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-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-value = 0.0113. $\alpha >$ p-value.

Make a decision: Since $\alpha >$ p-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.

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

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

a. $E = \frac{\text{(row total)(column total)}}{\text{total surveyed}} =$

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://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 $\text{df} = n - 1$.

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:

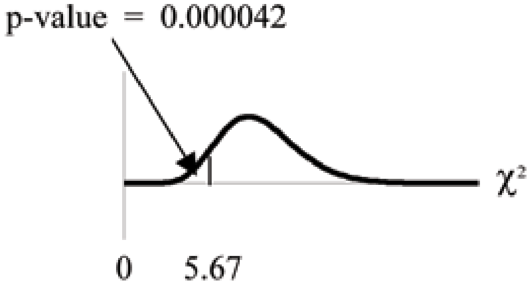
- n = the number of customers sampled
- df $= n - 1 = 25 - 1 = 24$

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-value $= P\left(\chi^2 < 5.67\right) = 0.000042$

Compare α **and the p-value:** $\alpha = 0.05$ p-value $= 0.000042$ $\alpha >$ p-value

Make a decision: Since $\alpha >$ p-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-value = 0.000042.

11.7 Summary of Formulas[7]

Rule 11.1: The Chi-square Probability Distribution
$\mu = \text{df}$ and $\sigma = \sqrt{2 \cdot \text{df}}$
Rule 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 $\underset{n}{\Sigma}\frac{(O-E)^2}{E}$, 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.

Rule 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 $\underset{(i \cdot j)}{\Sigma}\frac{(O-E)^2}{E}$ 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 $E = \frac{\text{(row total)(column total)}}{\text{total surveyed}}$.

Rule 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://http://cnx.org/content/m17058/1.5/>.