MTH410

Quantitative Business Analysis

Module 7: Statistical Analyses

In Module 7, we will continue our examination of statistical inference. However, we will now consider statistical inference when the comparison between two or more population means or proportions is important. This type of analysis is used frequently to evaluate data between groups for comparison and decision-making purposes.

Learning Outcomes

1. Develop tests about two means for independent samples.
2. Develop tests about two means for dependent samples (matched pairs).
3. Develop tests about the differences between two proportions.
4. Test for goodness of fit.
5. Test for variable independence.

For Your Success & Readings

The analysis of two or more populations adds complexity to statistical calculations and analysis. This type of analysis is common due to the value of these types of calculations. As you read the chapters and work through the problems, consider ways that you could use this analysis of information at your workplace or in other organizations.

This week's discussion question asks you to locate the results of a recent survey that shows at least two variables in a newspaper, magazine, or internet article. Outline the survey data so that your peers can understand the variables and results, and then identify at least one key formula from this module that you could use to evaluate the data. Provide a brief explanation of why you selected the formula you did and why it matters. Also, explain what the formula is, where it is in the textbook, and clearly define your **parameters.** Be sure to support your statements with logic and argument, citing any sources referenced. Post your initial response early and check back often to continue the discussion. Be sure to respond to your peers' and instructor's posts, as well.

Required

- Sections 10.1, 10.4, 10.6, 11.1, 11.3, 11.4, in *Introductory Business Statistics*

Recommended

- Taylor, C. (2018a, June 27). Example of two sample t test and confidence interval. *ThoughtCo*. Retrieved from **https://www.thoughtco.com/sample-t-test-confidence-interval-example-4022456** (https://www.thoughtco.com/sample-t-test-confidence-interval-example-4022456)
- Taylor, C. (2018b, Dec. 4). Example of a chi-square goodness of fit test. *ThoughtCo*. Retrieved from **https://www.thoughtco.com/chi-square-goodness-of-fit-test-example-3126382** (https://www.thoughtco.com/chi-square-goodness-of-fit-test-example-3126382)

1. Independent and Dependent (Matched Pairs) Samples, Tests About Two Means

We have discussed how to develop interval estimates and how to conduct hypothesis tests for situations involving a single population mean or proportion. What if we wanted to compare the means of two populations? For example, what if we wanted to learn if there was a difference in mean post-graduate earnings between men and women?

Independent and Dependent (Matched Pairs) Samples

When making inferences about the difference between two population means, one must first determine whether the two sets of samples are **matched pairs**. Two samples form a **matched pair** if each member of the first sample corresponds to a member of the other sample. Two samples are **independent** if the samples from the first population are not related to the samples from the second population.

Suppose a new diet has been invented. Five different people are weighed before the diet. Then each of those same persons is weighed after the diet. The before and after weights in lbs. after six months are listed here for the five people:

before	175	189	235	167	289
after	150	180	190	148	250

This set of data consists of matched pairs because each of the five sets of weights is matched (before, after). Matched pairs are often referred to as **dependent samples**.

Consider these five pairs of sales associates who use two different methods to sell cars. Each method was tried for a week:

	Associate 1	Associate 2	Associate 3	Associate 4	Associate 5
Method 1 Sales in Thousands of Dollars	75	58	89	90	92
Method 2 Sales in Thousands of Dollars	60	79	100	90	98

The set is a matched pair because each pair of data is related (Method 1, Method 2).

Test About Differences in Means with Independent Samples

As mentioned above, two samples are independent if they are not paired in any way.

We may develop hypothesis tests for the difference of two population means, μ_1 and μ_2, from two independent populations:

δ_0 is a hypothesized difference between the population means:

left-tailed	**right-tailed**	**two-tailed**
H_0: $\mu_1 - \mu_2 \geq \delta_0$ H_a: $\mu_1 - \mu_2 < \delta_0$	H_0: $\mu_1 - \mu_2 \leq \delta_0$ H_a: $\mu_1 - \mu_2 > \delta_0$	H_0: $\mu_1 - \mu_2 = \delta_0$ H_a: $\mu_1 - \mu_2 \neq \delta_0$

If δ_0=0, the above hypotheses can be rewritten as follows:

left-tailed	**right-tailed**	**two-tailed**
H_0: $\mu_1 \geq \mu_2$ H_a: $\mu_1 < \mu_2$	H_0: $\mu_1 \leq \mu_2$ H_a: $\mu_1 > \mu_2$	H_0: $\mu_1 = \mu_a$ H_a: $\mu_1 \neq \mu_2$

The **point estimate for the difference of two population means**, , from independent samples is:

The **standard error** is:

s_1, $n_1$and s_2, n_2 are the sample standard deviations and sample sizes of populations 1 and 2, respectively.

The **test statistic** uses the t-distribution:

df =

The degrees of freedom, *df*, are:

In the above formula for *df*, always **round down** decimal answers. For example, if the degrees of freedom above is computed to be 39.7, then round down to obtain *df=39*.

The above formula for the degrees of freedom is often used in statistical software packages. Because the above formula is tedious to compute, there is an **alternative, less accurate, formula for the degrees of freedom:**

It is common to use the above simpler formula for the degrees of freedom if both sample sizes are more than 30.

Note: In some other textbooks, the degrees of freedom are the smaller of n_1-*1* and n_2-*1*.

EXAMPLE

Let $\mu_1$and μ_2 represent the population mean post-graduate earnings of men and women, respectively. Suppose we want to determine if the mean post-graduate earning of men is less than the mean post-graduate earning of women. Test the following at the 0.01 significance level:

H_0: $\mu_1 \geq \mu_2$
H_a: $\mu_1 < \mu_2$

Suppose sample means of $80,000 and $80,500 were computed for males and females with sample sizes 85 and 81, respectively. Suppose the respective sample standard deviations are $1,000 and $1,200. The male and female earnings are **independent** since there is no mention that each male is related in some way to the set of females, and vice versa.

Click "Solution" to check your thinking.

Solution

– – – – 2.91

The test statistic is:

The degrees of freedom are:

***p*-value Method:**

The *p*-value is:

To find *P(t<-2.91)* with *df=166*, we used this Excel command

=T.DIST(-2.91, 166, TRUE)

Since the *p*-value of 0.002055 is less than the level of significance (0.002055< 0.01), we can reject H_0. There is strong enough evidence at the 0.01 significance level to conclude the mean post-graduate earning of men is less than the mean post-graduate earning of women.

Critical Value/Rejection Region Method:

The critical value with *df=166* is:

We used Excel to find the above critical t-value. For example, to find

use

=T.INV(0.01, 166)

Since the test statistic is less than a critical value (-2.91< -2.35), we can reject H_0. There is strong enough evidence at the 0.01 significance level to conclude the mean post-graduate earning of men is less than the mean post-graduate earning of women.

1.1. Distinguishing Between Dependent (Paired) and Independent Samples

When considering analysis of paired data, it is important to distinguish between dependent(paired) and independent samples.

For example, this set of data is not a matched pair because they are two independent sets of data:

Weekly shopping amount spent by female, in dollars	58	76	58	92	89
Weekly shopping amount spent by male, in dollars	55	79	88	60	59

However, the following set on married couples is a matched pair. One is pairing the wife with the husband. These are matched pairs because the amount spent by one might be affected by the spouse. For example, one spouse might have to spend less due to the other spouse spending more:

Weekly shopping amount spent by wife, in dollars	58	76	58	92	89
Weekly shopping amount spent by husband, in dollars	55	79	88	60	59

The following summary represents independent data on credit card debt, in dollars:

	New York	Maine
Sample size	358	280
Sample mean	$3,058	$2,500
Sample standard deviation	$300	$350

The two sets of data are independent because there is no pairing of the debt among the two states. However, the following are paired data on the stock prices of five companies:

	Company A	Company B	Company C	Company D	Company E
Stock price at 9 a.m.	$5.50	$39.76	$42.11	$22.52	$70.30
Stock price at 4 p.m.	$5.00	$32.58	$48.50	$18.95	$65.49

The data are dependent because for each company, the stock price at 9 a.m. is paired with the one for 4 p.m.

1.2. More on Tests About Two Means for Independent Samples

Recall that two samples are independent if the elements of the first sample are not paired with those of the second sample.

Two fashion retailers, Clothes Mode and Fashion Max, offer separate in-store credit cards to their customers. Let $\mu_1$and μ_2 represent the mean credit card debt of Clothes Mode and Fashion Max card holders, respectively. A business analyst claims that there is a difference in mean credit card debt:

H_0: $\mu_1 = \mu_2$
H_a: $\mu_1 \neq \mu_2$

Suppose sample means of $600 and $650 were computed with sample sizes 50 and 60 for Clothes Mode and Fashion Max card holders, respectively. Suppose the respective sample standard deviations are $100 and $120. Develop a hypothesis test at the 0.01 level of significance to determine if the population mean wages differ.

Click “Solution” to check your thinking.

Solution

The two samples are independent because customers of Clothes Mode and Fashion Max are not paired in any way. The test statistic is:

The degrees of freedom are:

***p*-value Method:**

Since this is a two-tailed test, the *p*-value is twice the left-tail probability:

We can use Excel to find the above probability. The following provides the cumulative probability of the t-distribution:

=T.DIST(-2.384, 108, TRUE)

P(t< -2.384)= 0.009435

Since the *p*-value of 0.019 is greater than the level of significance (0.019 > 0.01), fail to reject H_0. There is not strong enough evidence at the 0.01 level of significance to conclude the mean credit card debt of Clothes Mode and Fashion Max card holders differ.

Critical Value/Rejection Region Method:

The critical values with 108 degrees of freedom are

We used Excel to find the above critical t-values. The following command provides t-values for left tail areas:

=T.INV(probability, degrees of freedom)

For example, to find the t-value

use

=T.INV(0.005, 108)

We obtained a t-value of -*2.62*.

Since the test statistic of -2.384 is between the two critical values (-2.62<-2.384<2.62), fail to reject H_0. There is not strong enough evidence at the 0.01 level of significance to conclude the mean credit card debt of Clothes Mode and Fashion Max card holders differ.

2. Tests About Two Means of Dependent Samples (Matched Pairs)

Tests About Differences in Matched or Paired (Dependent) Samples

Recall that two samples form a **matched pair** if each member of the first sample corresponds to a member of the other sample. If the data consist of matched pairs, let be the population mean difference in values between the two populations.

Let be the sample mean difference in values between the two populations:

n **is the number of pairs**, not the total number of samples. *d* is the difference between each pair. The sample standard deviation of the differences is:

The test statistic for matched samples uses the t-distribution with $df = n - 1$ degrees of freedom:

EXAMPLE

Recall the five pairs of sales associates who use two different methods for a week to sell cars. Let represent the population mean difference between Method 1 and Method 2. Suppose we are testing whether there is a difference between the mean sales of Method 1 and Method 2 at a 0.05 significance level:

Let *d* denote the difference between Method 1 and Method 2, for each sales associate:

	Associate 1	Associate 2	Associate 3	Associate 4	Associate 5	Sum:
Method 1 Sales in Thousands of Dollars	75	58	89	90	92	
Method 2 Sales in Thousands of Dollars	60	79	100	90	98	
Difference, *d*	15	-21	-11	0	-6	-23

4.6

The sample standard deviation of the differences is that of the d values:

d	15	-21	-11	0	-6	Sum:
	19.6	-16.4	-6.4	4.6	-1.4	
=	384.16	268.96	40.96	21.16	1.96	717.20

= 13.39

$= \quad = \quad = \quad = -0.77$

The test statistic is:

The *p*-value with degrees of freedom *df = n - 1 = 5 - 1 = 4* is:

We can use Excel to find *P(t< -0.77)*:

=T.DIST(-0.77, 4, TRUE)

Since the *p*-value is greater than the level of significance (0.484252 > 0.05), fail to reject H_0. There is insufficient evidence to conclude that there is a difference between the mean sales of Method 1 and Method 2 at a 0.05 significance level.

2.1. More on Tests About Two Means for Dependent (Paired, Matched) Samples—*p*-value Method

Recall that two samples are dependent if the elements of the first sample are paired with those of the second sample.

EXAMPLE

A real estate agent wants to determine if an appraiser (Appraiser 1) underestimates the values of homes compared to a second appraiser who has been hired (Appraiser 2). Five homes are randomly chosen and appraised by both appraisers.

	House 1	House 2	House 2	House 3	House 4	Sum
Appraiser 1	$35,000	$56,000	$78,000	$57,000	$47,000	
Appraiser 2	$43,000	$55,000	$85,000	$62,000	$56,000	
Difference, d	-8,000	1,000	-7,000	-5,000	-9,000	-28,000

The hypotheses are intended to determine if the mean difference is less than 0 at the 0.05 level of significance:

5000

The sample standard deviation of the differences is that of the *d* values. It can be found this way:

d	-8,000	1,000	-7,000	-5,000	-9,000	Sum:
	-2400	6600	-1400	600	-3400	
	5,760,000	43,560,000	1,960,000	360,000	11,560,000	63,200,000

= = = = 3,974.92

– – – – −3.15

The test statistic is:

The degrees of freedom are based on the number of pairs: *df = n - 1 = 5 - 1 = 4*.

***p*-value Method:**

Since this is a left-tailed test, the *p*-value is:

We can use Excel to find $P(t< -3.15)$:

=T.DIST(-3.15, 4, TRUE)

Since the *p*-value is less than the level of significance ($0.017<0.05$), reject the null hypothesis. At the 0.05 level of significance, there is strong enough evidence to conclude that Appraiser 1 underestimates home prices relative to Appraiser 2.

2.2. More on Tests About Two Means for Dependent (Paired, Matched) Samples—Critical Value/Rejection Region Method

In this Section, we consider the critical value/rejection region method.

EXAMPLE

A real estate agent wants to determine if an appraiser (Appraiser 1) underestimates the values of homes compared to a second appraiser who has hired (Appraiser 2). Five homes are randomly chosen and appraised by both appraisers.

	House 1	House 2	House 2	House 3	House 4	Sum
Appraiser 1	$35,000	$56,000	$78,000	$57,000	$47,000	
Appraiser 2	$43,000	$55,000	$85,000	$62,000	$56,000	
Difference, d	-8,000	1,000	-7,000	-5,000	-9,000	-28,000

The hypotheses are intended to determine if the mean difference is less than 0 at the 0.05 level of significance:

The sample standard deviation of the differences is that of the d valuesIt can be found this way:

d	-8,000	1,000	-7,000	-5,000	-9,000	Sum:
	-2400	6600	-1400	600	-3400	
=	5,760,000	43,560,000	1,960,000	360,000	11,560,000	63,200,000

= 3,974.92

= = = = −3.15

The test statistic is:

The degrees of freedom are based on the number of pairs: $df = n - 1 = 5 - 1 = 4$.

Critical Value/Rejection Region Method:

This is a left-tailed test. Thus, the critical values with 4 degrees of freedom are

We can use Excel to find the above critical t-values. The following command provides t-values:

To find the t-value, we used this Excel command

=T.INV(0.05, 4)

We obtained a t-value of *-2.13*.

Since the test statistic is less than the critical t-value (-3.15 < -2.13), reject the null hypothesis. At the 0.05 level of significance, there is strong enough evidence to conclude that Appraiser 1 underestimates home prices relative to Appraiser 2.

3. Tests of the Difference Between Two Proportions, Chi-Square Distribution

Tests of the Difference Between Two Proportions

Given two populations A and B, we may also make interferences regarding the difference of two population proportions:

Here are the three types of hypotheses involving the difference of proport. δ_0 ions is a hypothesized difference between the population proportions:

left-tailed	**right-tailed**	**two-tailed**

In many applications, δ_0=0 and the above hypotheses can be rewritten as follows:

left-tailed	**right-tailed**	**two-tailed**

The sample proportions are

Here, is the number of samples that have a specified characteristic out of a sample of from population A. Also, is the number of samples that have a specified characteristic out of a sample of from population B.

The point estimate of the difference between the two population proportions is the corresponding difference between the two sample proportions:

There are three **assumptions** regarding populations A and B:

1. The two samples from A and B are independent (not matched pairs) random samples.
2.
3. Each sample size, and , cannot be more than 10% of the population.

Part 2 above says that both the number of successes and the number of failures should be at least 5.

Under the assumption that is true as an equality, , then the formula for the standard deviation (standard error) of becomes

However, we do not know the value of p. To fix that, we have the following estimator of p. **Weighted (Pooled) Proportion of and :**

If one substitutes for p_c in the standard error formula, the **test statistic (*z*-score) formula** is:

In most examples, .

EXAMPLE

Suppose one is interested in the spending habits of males versus females on soda. Let and be the population proportion of males and females who buy soda at a certain grocery store, respectively. Let and be the sample proportion of males and females who buy soda at a certain grocery store, respectively.

For this grocery store, suppose we want to test whether the proportion of male soda buyers differs from the proportion of female soda buyers at the 0.05 level of significance:

Suppose the following data are obtained:

Males	Females
Sample size = 100	Sample size = 200
Number who bought soda	Number who bought soda

The three assumptions regarding both populations are satisfied:

1. The male and female samples are independent (not matched pairs) random samples.
2. and
3. Each sample size, and , does not appear to be more than 10% of the population of all store customers.

The pooled average is

The test statistic is:

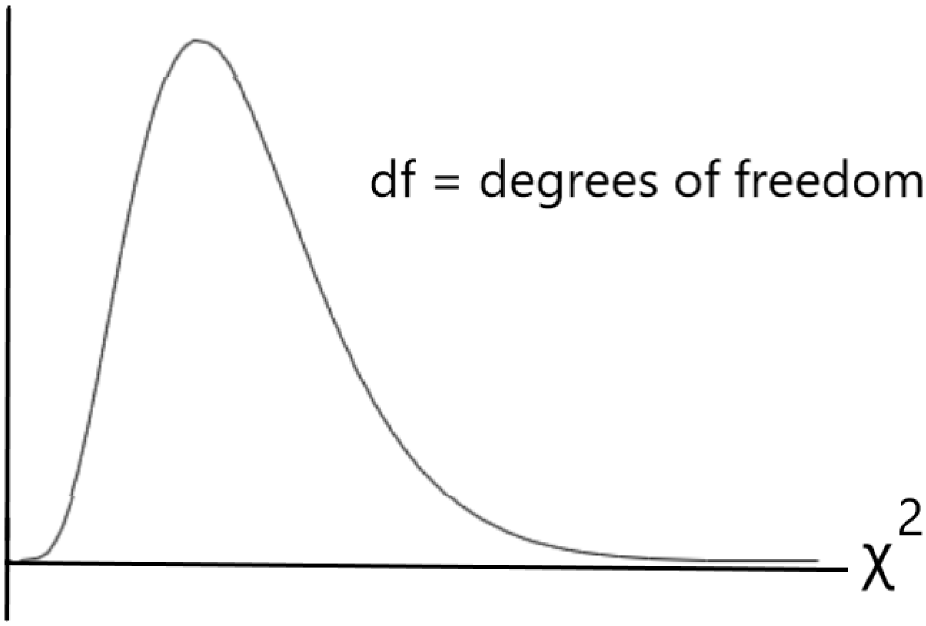
The *p*-value is

The probability, *P(Z < -0.36)*, is found by Appendix A on page 607. Appendix A only has positive z values:

Since the *p*-value is greater than the level of significance, (0.7188 > 0.05) we would fail to reject Thus, there is insufficient evidence at the 0.05 level of significance to conclude that there is a difference between the proportion of males and females who buy soda at the grocery store.

Chi-Square Distribution

The chi-square and t-distribution are similar in that they both have degrees of freedom. Each degree of freedom provides a different chi-square distribution. See Figure 11.2 for the different chi-square distribution graphs and degrees of freedom. Notice that the chi-square distribution is skewed to the right. Hence, the mean is to the right of the peak.



As with the t-distribution, the chi-square distribution approximates the normal distribution when the degrees of freedom are large enough. For the chi-square distribution, if the degrees of freedom are 90 or more, we can assume that the chi-square distribution approximates the normal distribution.

For example, the following ^ value corresponding to 12 degrees of freedom is found by going to Appendix Table A13 on page 610:

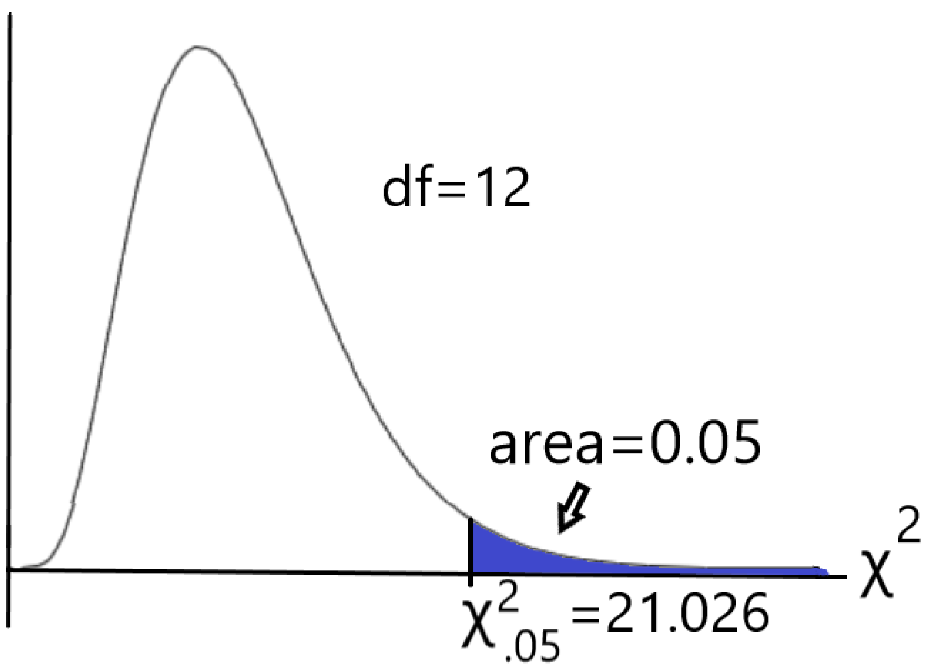


Table A13 corresponds to ^ values whose area is to the right.

Alternatively, we may use the Excel command to find the chi-square value corresponding to a right-tail probability:

=CHISQ.INV.RT(probability, degrees of freedom)

In the above example, we would use

=CHISQ.INV.RT(0.05,12)

to get 21.02607.

3.1. More on Inferences About the Difference Between Two Proportions—*p*-value Method

Instead of testing whether proportions differ, we can test whether one proportion is greater than another.

EXAMPLE

A banker claims that a larger proportion of Bank A clients with good credit (credit score of 700 or more) get accepted for a mortgage loan than those from Bank B. Let and be the population proportion of Bank A and Bank B clients who get approved for a mortgage loan, respectively. Let and be the sample proportion of Bank A and Bank B clients who get approved for a mortgage loan, respectively.

The banker wants to test whether a larger proportion of Bank A clients with good credit get accepted for a mortgage loan than those from Bank B at the 0.1 level of significance:

Suppose the following data is obtained:

Bank A	Bank B
Sample size = 50	Sample size = 70
Number who were approved	Number who were approved

The three assumptions regarding both populations are satisfied:

1. The Bank A and Bank B mortgage applicants are independent (not matched pairs) random samples.
2. and
3. Each sample size, and , does not appear to be more than 10% of the population of all mortgage applicants from either Bank A or Bank B.

The pooled average is

The test statistic is:

_p_-value Method:

The *p*-value is

The probability, *P(Z>1.48)*, is found in Appendix A on page 607. Appendix A only has positive z values:

Since the *p*-value is less than the level of significance, (0.0694 < 0.10), we would reject . Thus, there is enough evidence at the 0.1 level of significance to conclude that a larger proportion of Bank A clients with good credit get accepted for a mortgage loan than those from Bank B.

3.2. More on Inferences About the Difference Between Two Proportions—Critical Value/Rejection Region Method

In this section, we consider the critical value/rejection region method.

EXAMPLE

A banker claims that a larger proportion of Bank A clients with good credit (credit score of 700 or more) get accepted for a mortgage loan than those from Bank B. Let and be the population proportion of Bank A and Bank B clients who get approved for a mortgage loan, respectively. Let and be the sample proportion of Bank A and Bank B clients who get approved for a mortgage loan, respectively.

The banker wants to test whether a larger proportion of Bank A clients with good credit get accepted for a mortgage loan that those from Bank B at the 0.1 level of significance:

Suppose the following data are obtained:

Bank A	Bank B
Sample size = 50	Sample size = 70
Number who were approved	Number who were approved

The three assumptions regarding both populations are satisfied:

1. The Bank A and Bank B mortgage applicants are independent (not matched pairs) random samples.
2.
   and
3. Each sample size, and , does not appear to be more than 10% of the population of all mortgage applicants from either Bank A or Bank B.

The pooled average is

The test statistic is:

Critical Value/Rejection Region Method

The critical value is

The *z*-value was found from Appendix Table A11 on page 607. We subtracted 0.1 from 0.5 to get 0.4. Then the probability of 0.4 is between the table values of 0.3997 and 0.4015. The value of 1.285 is then obtained by computing the average of *z*-values 1.28 and 1.29.

Since the test statistic is greater than the critical value (1.48> 1.285) reject . Thus, there is enough evidence at the 0.1 level of significance to conclude that a larger proportion of Bank A clients with good credit get accepted for a mortgage loan than those from Bank B.

4. Goodness of Fit and Tests of Independence

Goodness of Fit Tests

A **chi-square goodness-of-fit test** determines whether a frequency distribution fits an expected distribution.

: The frequency distribution fits the expected distribution
:   is not true

Alternatively, we can state a goodness of fit test as proportions. We would like to determine whether the population proportions are equal to given values b_1, b_2, ..., bk (where the b_i's add up to 1):

: $p_1 = b_1, p_2 = b_2, p_3 = b_3, \ldots, p_k = b_k$
:   is not true

The test statistic for goodness of fit compares the sample of observed results with the expected results under the assumption that the null hypothesis is true. The formula is this sum of *k* terms:

O= observed frequency for a given category
E = corresponding expected frequency for the category
k = the number of categories

One uses the distribution with k-1 degrees of freedom, provided that the expected frequencies are 5 or more for each category. A goodness of fit test is always a **right-tailed hypothesis test**.

EXAMPLE

Suppose that, historically, 25%, 35%, and 40% of the people prefer Malls A, B, and C, respectively. A study of 598 people was done to see if this result is still accurate. The expected frequencies are thus

Expected Frequencies			
Prefer Mall A	Prefer Mall B	Prefer Mall C	Total
0.25·598=149.5	0.35·598=209.3	0.4·598=239.2	568

The following results were obtained:

Observed Frequencies			
Prefer Mall A	Prefer Mall B	Prefer Mall C	Total
144	198	226	568

Suppose one is testing at the 0.1 level of significance:

H_0: The mall preference distribution fits the expected distribution
H_a: The mall preference distribution does not fit the expected distribution

or equivalently as proportions

H_0: p_1 = 0.25, p_2 = 0.35, p_3 = .40
H_a: H_0 is not true

Here, are the population proportion of people who prefer malls A, B, and C, respectively. The 3 terms of the sum are [illegible]

p-**value Method**:

The *p*-value is (with $k - 1 = 3 - 1 = 2$ degrees of freedom) is the right-tail area: *P*(> 0.037) = 0.964158

The Excel command to find the above right tail area with 2 degrees of freedom is

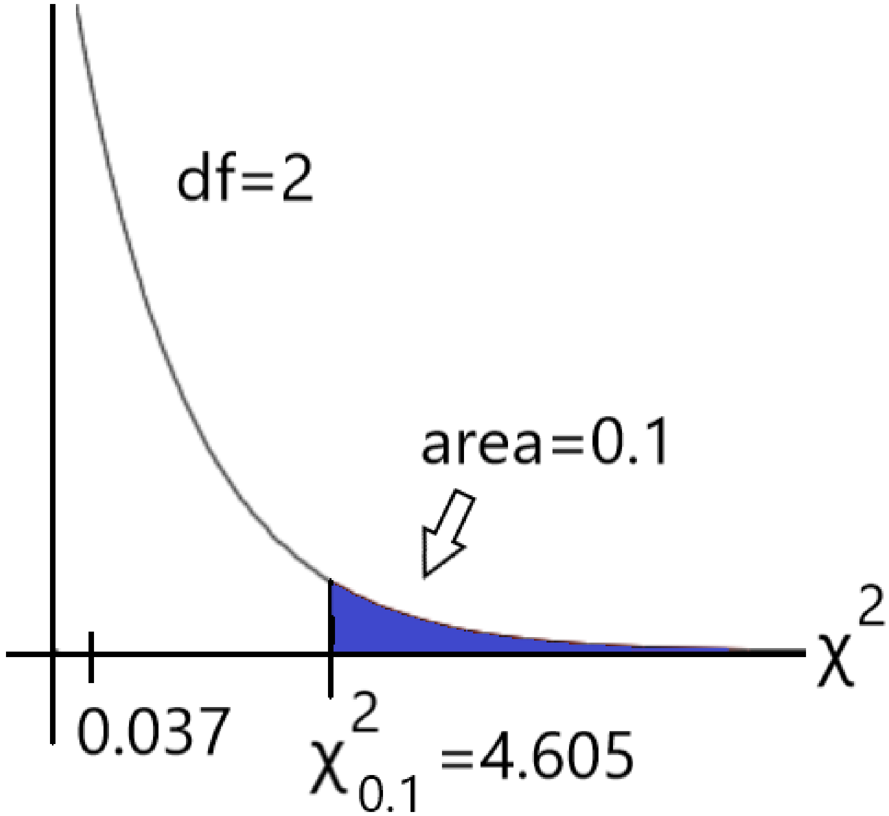
=CHISQ.DIST.RT(0.037 ,2)

Since the *p*-value is not less than the level of significance (0.964158 > 0.1), we would fail to reject H_0. We cannot conclude that the mall preference distribution differs from the historical (expected) ones.

Critical Value/Rejection Region Method:

The critical value with 2 degrees of freedom is

The above value corresponding to 3 degrees of freedom was found by going to Appendix Table A13 on page 610. Table A13 corresponds to values whose area is to the right.



Since the test statistic is less than the critical value (0.037<4.605), we would fail to reject H_0. We cannot conclude that the mall preference distribution differs from the historical (expected) ones.

Tests of Independence

The **test of independence** is intended to identify whether variables in a given sample have a relationship.

For example, suppose males and females were sampled as to which of the three local malls they prefer:

	Prefer Mall A	Prefer Mall B	Prefer Mall C	Total
Male	38	68	78	184
Female	42	76	82	200
Total	80	144	160	384

Suppose one is interested if gender and mall preference are independent of each other. The hypotheses are:

H_0: Gender and mall preference are independent of each other.
H_a: Gender and mall preference are dependent on each other.

In general, if one makes a contingency table (as above), the hypotheses for tests of independence are written as:

H_0: The rows are independent of the columns and vice versa.
H_a: The rows are dependent of the columns or vice versa.

The test statistic is like the one for goodness of fit tests. The formula is this sum of *i·j terms:*

O= observed frequency for a given entry in the table
E = corresponding expected frequency for entry in the table
i = the number of rows in the table
j = the number of columns in the table

A test of independence is always a **right-tailed hypothesis test**.

One uses a distribution with *(i - 1)· (j - 1)* degrees of freedom, provided that the expected frequencies are 5 or more for each cell.

Note: The number of expected results must be 5 or more for us to use this test.

Example: Recall the previous example:

Suppose males and females were sampled as to which of the three local malls they prefer:

	Prefer Mall A	Prefer Mall B	Prefer Mall C	Total
Male	38	68	78	184
Female	42	76	82	200
Total	80	144	160	384

At the .05 level of significance, test the hypotheses of whether gender is independent of mall preference:

H_0: Gender and mall preference are independent of each other
H_a: Gender and mall preference are dependent on each other

Calculate the 2·3=6 expected frequencies, E. For example, E for the row 1 column 2 entry is

The expected frequencies, E, are

	Prefer Mall A	Prefer Mall B	Prefer Mall C	Total
Male	38.3	69	76.7	184
Female	41.7	75	83.3	200
Total	80	144	160	384

The 6 terms of the sum are

The p-value with is

The Excel command to find the above right tail area with 2 degrees of freedom is

=CHISQ.DIST.RT(0.073 ,2)

Since the p-value is not less than the level of significance (0.964158 > 0.05), we would fail to reject H_0 we would fail to reject that gender and mall preference are independent of each other at the 0.05 significance level.

Note: Here is an explanation why the formula for E is the following:

If row 1 and column 2 are independent, then by the multiplication rule for independent events

Thus, we expect approximately .179685 of 388 people to be Male and Prefer Mall B:

So, to test for goodness of fit and test of independence, we would follow these similar steps:

State the null and alternative hypotheses.

Select a random sample and record the observed frequencies.

Determine the expected frequency in each category by multiplying the category probability by the sample size.

Compute the test statistic.

Accept or reject the H_0 and the H_a with consideration of the level of significance.

4.1. More on Goodness of Fit Tests

A goodness of fit test is used to determine if the observed frequency distribution fits the expected frequencies. The following is an example to determine if different sizes of cereal boxes of one type of cereal are sold with equal frequencies.

A manufacturer expects to sell small, medium, large, and family sized Sugar Rush cereals with equal frequencies at any Global Grocer grocery store over a period of 2 months. A Global Grocer store was randomly chosen, and the number of Sugar Rush cereals sold was recorded for a period of 2 months. A total of 800 Sugar Rush cereal boxes were sold over a period of 2 months:

	Actual (Observed) Number Sold	Expected Number sold
small	180	200
medium	220	200
large	179	200
family	221	200
Total	800	800

Suppose the manufacturer is testing the following hypotheses at the 0.05 level of significance:

H_0: The cereal size sales distribution fits the expected distribution
H_a: The cereal size sales distribution does not fit the expected distribution

The test statistic is a ˆ distribution value with *(k - 1)=4-1=3* degrees of freedom:

***p*-value Method**:

The *p*-value with 3 degrees of freedom is

P(ˆ > 8.41) = 0.038256

The Excel command to find the above ˆ right tail area with 2 degrees of freedom is

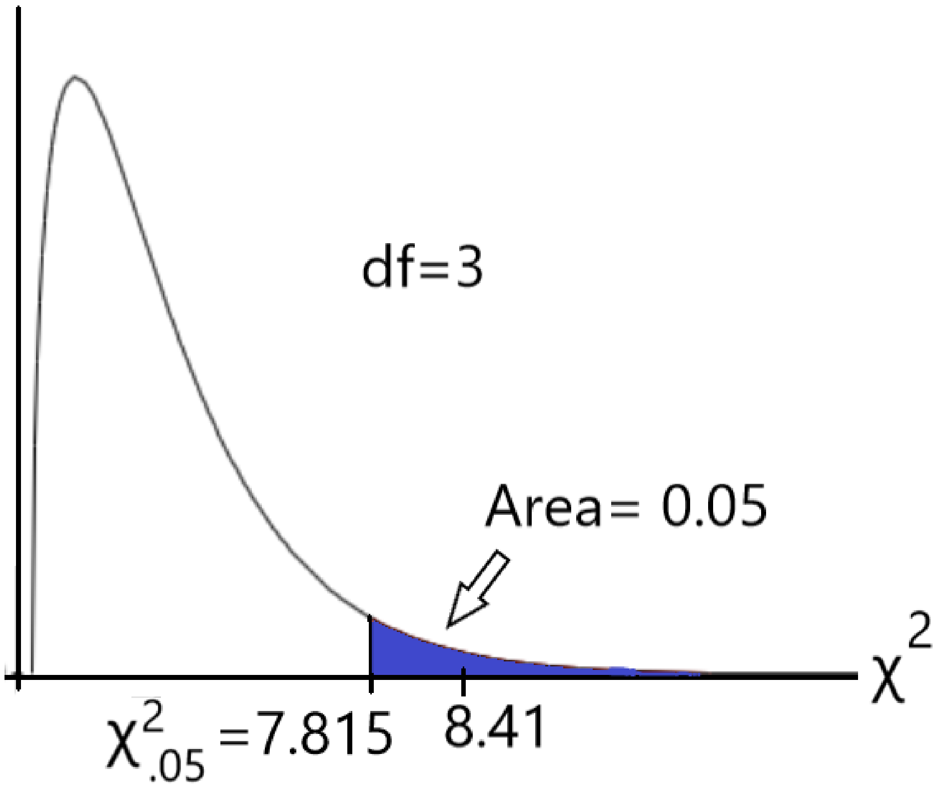
=CHISQ.DIST.RT(8.41 ,3)

Since the *p*-value is less than the level of significance (0.038256 < 0.05), we would reject H_0. At the 0.05 level of significance, there is strong enough evidence to conclude that the cereal size sales distribution does not fit the expected distribution.

Critical Value/Rejection Region Method:

The critical value with 2 degrees of freedom is

The above value corresponding to 3 degrees of freedom was found by going to Appendix Table A13 on page 610. Table A13 corresponds to values whose area is to the right.



Since the test statistic is greater than the critical value (8.41>7.815) we would reject H_0. At the 0.05 level of significance, there is strong enough evidence to conclude that the cereal size sales distribution does not fit the expected distribution.

4.2. More on Tests of Independence

A test of independence helps determine if the rows and columns are independent of each other.

EXAMPLE

A consumer researcher wants to determine if the brand of cell phone and whether the client had some type of college education are independent of each other.

	Brand A	Brand B	Brand C	Total
Number with some college education	28	55	76	159
Number without any college education	32	73	58	163
Total	60	128	134	322

At the 0.1 level of significance, test the hypotheses of whether a client had some college education is independent of cell phone brand:

H_0: Some type of college education and cell phone preference are independent of each other
H_a: Some type of college education and cell phone preference are dependent on each other

Calculate the 2·3=6 expected frequencies, E. For example, E for the row 2 column 1 entry are

The 6 terms of the sum are

p-value Method:

The p-value with is

$P($ > 5.166) = 0.075547

The Excel command to find the above right tail area with 2 degrees of freedom is

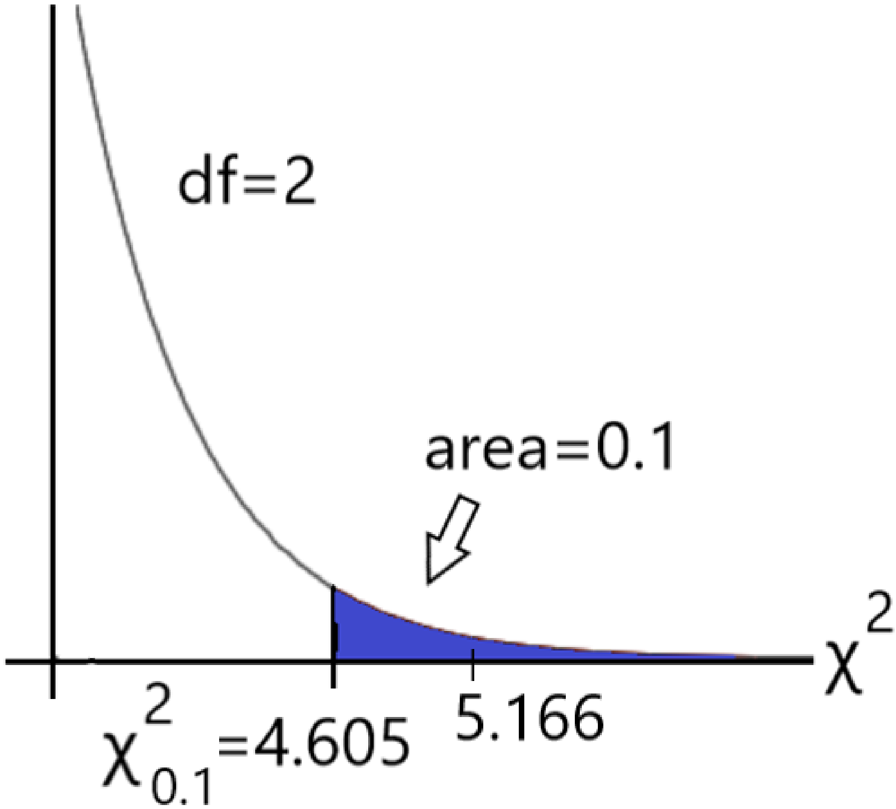
=CHISQ.DIST.RT(5.166 ,2)

Since the p-value is less than the level of significance (.075547<0.1), we would reject . Some type of college education and cell phone preference are dependent on each other, at the 0.1 level of significance.

Critical Value/Rejection Region Method:

The critical value with 2 degrees of freedom is

The above value corresponding to 2 degrees of freedom was found by going to Appendix Table A13 on page 610. Table A13 corresponds to values whose area is to the right.



Since the test statistic is greater than the critical value (5.166 > 4.605), we would reject the null hypothesis. Some type of college education and cell phone preference are dependent on each other, at the 0.1 level of significance.

5. Summary

We first determined how to distinguish between independent and dependent (paired) samples. Dependent samples are paired in some way. For example, the weight of a person before the diet is paired with the weight of that person after the diet. We covered hypothesis tests for the mean involving either independent or dependent (paired) samples. We then covered two applications of the chi-square distribution. Both use similar test statistic formulas. The first is a goodness of fit test. Such a test is used to determine if observed frequencies follow the expected frequencies in a two-way table. Then we covered a test of independence between the rows and columns in a two-way table.

Here is the list of the objectives that we have covered and are part of the Mastery Exercises in Knewton Alta:

- Determine the p-value for a hypothesis test for the difference between two means (population standard deviations unknown)
- Make a conclusion and interpret the results for testing the difference between two means (population standard deviation unknown) using the p-value Approach
- Identify dependent samples versus independent samples
- Determine the critical value(s) for a hypothesis test for the mean of the differences for the paired data in order to define rejection region(s)
- Calculate the test statistic (t-value) and degrees of freedom for a hypothesis test for the differences of paired data (dependent samples)
- Determine the p-value for a hypothesis test for the mean of the differences for the paired data
- Make a conclusion and interpret the results for testing the difference between means for paired data (dependent samples) using the p-value Approach
- Determine the critical value(s) for a hypothesis test to test the difference between two population proportions in order to define rejection region(s)
- Compute the value of the test statistic (z-value) for a hypothesis test to test the difference between two population proportions
- Determine the p-value for a hypothesis test to test the difference between two population proportions
- Understand the properties of the chi-square distribution
- Compute the value of the test statistic using the expected frequencies for a chi-square goodness-of-fit test
- Conduct and interpret a chi-square goodness-of-fit test
- Compute the value of the test statistic using the expected frequencies for a chi-square independence test

Conduct and interpret a test of independence with the chi-square distribution

Check Your Understanding

Check your understanding of the contents of this module by answering the following questions. If you get the answers wrong, consider reviewing the primary and secondary pages and trying again.

Embedded Media Content! Please use a browser to view this content.

References

None