

Topic 13. Chisquare Tests

Cheng Peng

Contents

1	Introduction	1
2	Chi-square Test of Goodness-of-fit	1
2.1	A Motivational Example	2
2.2	Chi-squares Distribution	2
2.3	Formulation of Chi-square Test of Goodness-of-fit	4
3	Chi-square Test of Independence	6
3.1	Independence of Two Categorical Variables	7
3.2	Expected Table Under Independence Assumption (H_0)	7
3.3	Formulation of Chi-squares Test of Independence	8
4	Use of Technology	11
4.1	Goodness-of-fit Chi-square	11
4.2	Chi-square Test of Independence	11
5	Practice Exercises	12

1 Introduction

We have discussed the relationship between two numerical variables using linear correlation coefficient and linear regression. We now explore the relationship between two categorical variables. The idea is to make an assumption (hypothesis) about the variable (s) and use the assumption to construct a frequency table (called the expected table). At the same time, we tabulated the data to obtain an observed table. The discrepancy between the expected table and the observed table can be used to make the inference about the relationship between categorical variables.

2 Chi-square Test of Goodness-of-fit

A goodness-of-fit test of a distribution is a testing procedure that justifies whether the null hypothesis that specified distribution is correct based on sample information.

For a single categorical variable, the null hypothesis should specify the cell probabilities. In other words, if the category has k categories, then (p_1, p_2, \dots, p_k) must be specified in the null hypothesis.

2.1 A Motivational Example

As a special case, we look at the following example of testing the proportion problem.

Example 1. We want to justify a claim that about 30% of WCU students are STEM majors. That is, we test the following hypotheses.

$$H_0 : p = 0.3 \quad v.s. \quad H_a : p \neq 0.3.$$

We take a random sample of 100 students and record the majors and found that 33 of them claimed a major in STEM. This means 67 of them are non-STEM majors. We have introduced a procedure to test the above hypotheses with the test statistic

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

that compares the claimed proportion with the sample proportion.

Note that the proportion of STEM majors contains the number of majors (frequencies) in both STEM and non-STEM disciplines. we can think about using (observed)sample frequencies and null (expected) frequencies to define the test statistic.

- Under H_0 , we would **expect** to have 30 STEM majors and 70 non-STEM majors.
- We **observed** 33 STEM majors and 67 STEM majors in the random sample.

The above observed and expected number of STEM and non-STEM majors are summarized in the following table.

	Observed Values (from the sample)	Expected Value (under Ho)
STEM	$O_1 = 33$	$E_1 = 30$
non-STEM	$O_2 = 67$	$E_2 = 70$

In fact, a test statistic that measures the “distance” between the observed and expected frequency tables and has a χ^2 (chi-square) distribution is defined below

$$TS = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \rightarrow \chi_1^2.$$

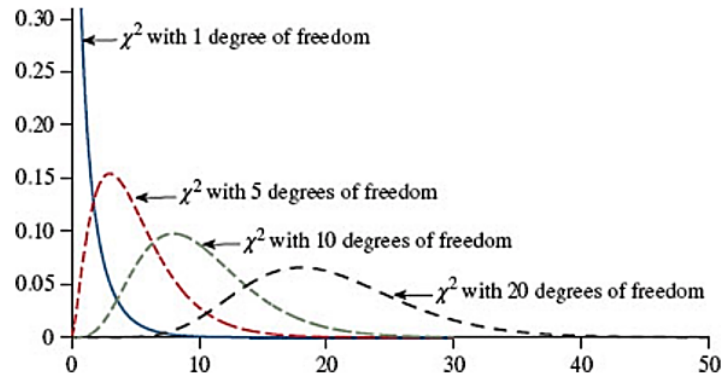
The value of the test statistic in this example

$$TS = \frac{(33 - 30)^2}{30} + \frac{(67 - 70)^2}{70} = 9/30 + 49/70 = 3/10 + 7/10 = 1$$

With the above value of test statistic, we can make a statistical decision on H_0 based on a given significance level.

2.2 Chi-squares Distribution

The chi-square distribution is used to characterize the positive random variable. Unlike normal and t distributions that have symmetric density curves, the chi-square distributions (dependent on the degrees of freedom) have skewed density curves.



We can find the critical value of chi-square distribution from the chi-square table that is available on the course web page. The structure of the chi-square table is similar to the t-table.

d.f	0.9950	0.9900	0.9750	0.9500	0.9000	0.1000	0.0500	0.0250	0.0100	0.0050
1	0.0000	0.0002	0.0010	0.0039	0.0158	2.705	3.841	5.024	6.635	7.879
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.605	5.992	7.378	9.210	10.597
3	0.0717	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348	11.345	12.838
4	0.2070	0.2971	0.4844	0.7107	1.0626	7.779	9.488	11.143	13.277	14.860
5	0.4117	0.5543	0.8312	1.1455	1.6103	9.236	11.070	12.832	15.086	16.750
6	0.6757	0.8721	1.2373	1.6354	2.2041	10.645	12.592	14.449	16.812	18.548
7	0.9893	1.2390	1.6899	2.1697	2.8331	12.017	14.067	16.013	18.475	20.278
8	1.3444	1.6465	2.1797	2.7326	3.4895	13.362	15.507	17.535	20.090	21.955
9	1.7349	2.0879	2.7004	3.3251	4.1682	14.684	16.919	19.023	21.666	23.589
10	2.1559	2.5582	3.2470	3.9403	4.8652	15.987	18.307	20.483	23.209	25.188
11	2.6032	3.0535	3.8157	4.5748	5.5778	17.275	19.675	21.920	24.725	26.757

The possible degrees of freedom are listed in the first column, the possible right-tail areas are listed in the top row, and the critical values are listed in the main body of the table.

The steps for finding the critical values are the same as those we followed for finding t- critical values.

Example 2. Find the critical value of the chi-square distribution with 5 degrees of freedom with significance level 0.05.

Chisquare value

Possible significance level: right-tailed areas

d.f	0.9950	0.9900	0.9750	0.9500	0.9000	0.1000	0.0500	0.0250	0.0100	0.0050
1	0.0000	0.0002	0.0010	0.0039	0.0158	2.705	3.841	5.024	6.635	7.879
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.605	5.992	7.378	9.210	10.597
3	0.0717	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348	11.345	12.838
4	0.2070	0.2971	0.4844	0.7107	1.0636	7.779	9.488	11.143	13.277	14.860
5	0.4117	0.5543	0.8312	1.1455	1.6103	9.236	11.070	12.832	15.086	16.750
6	0.6757	0.8721	1.2373	1.6354	2.2041	10.645	12.592	14.449	16.812	18.548
7	0.9893	1.2390	1.6899	2.1673	2.8331	12.017	14.067	16.013	18.475	20.278
8	1.3444	1.6465	2.1797	2.7326	3.4895	13.362	15.507	17.535	20.090	21.955
9	1.7349	2.0879	2.7004	3.3251	4.1682	14.684	16.919	19.023	21.666	23.589
10	2.1559	2.5582	3.2470	3.9403	4.8652	15.987	18.307	20.483	23.209	25.188
11	2.6032	3.0535	3.8157	4.5748	5.5778	17.275	19.675	21.920	24.725	26.757

Possible degrees of freedom

The above figure shows how to find the critical value, denoted by $CV = \chi_{5,0.05}^2 = 11.071$. The first subscript denotes 5 degrees of freedom and the second subscript is the significance level of 0.05.

2.3 Formulation of Chi-square Test of Goodness-of-fit

Let k be the number of categories of categorical variable Y . The category labels are C_1, C_2, \dots, C_k . Let $P_1 = Pr(C_1), P_2 = Pr(C_2), \dots, P_k = Pr(C_k)$. The null hypothesis claims that the categorical follows a specific distribution, and the alternative hypothesis claims that the categorical distribution does NOT follow the distribution specified in the null hypothesis. That is,

$$H_0 : P_1 = p_1, P_2 = p_2, \dots, P_k = p_k \quad v.s. \quad H_a : \text{the distribution in } H_0 \text{ is not correct.}$$

The N is the sample size. We can then calculate the **expected cell frequency** of each category using formulas: $E_1 = N \times p_1, E_2 = N \times p_2, \dots, E_k = N \times p_k$. The **observed cell frequency**, denoted by O_i (for $i = 1, 2, \dots, k$), of each category is obtained from the data set. The **expected** and **observed** frequencies are summarized in the following table.

	Category 1	Category 1	...	Category (k-1)	Category k
Observed	O_1	O_2	...	O_{k-1}	O_k
Expected	E_1	E_2	...	E_{k-1}	E_k

The chi-square statistic is

$$G^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \dots + \frac{(O_k - E_k)^2}{E_k} \rightarrow \chi_{k-1}^2$$

A small G^2 indicates a lack of evidence for rejecting the null hypothesis. This implies that **the Pearson chi-square test of goodness is always a right-tailed test**. The degrees of freedom are always $(k - 1)$ if the categorical factor variable has k levels.

Example 3. A gambler wants to test a die to determine whether it is fair. The gambler rolls a die that has six possible outcomes: 1, 2, 3, 4, 5, and 6; and the die is fair if each of these outcomes is equally likely. The gambler rolls the die 60 times and counts the number of times each number comes up. These counts, which are called the observed frequencies, are

Outcome	1	2	3	4	5	6
Observed	12	7	14	15	4	8

Solution: The null hypothesis is that the six-sided die is fair. That is equivalent to

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6 \quad v.s. \quad H_a : \text{The die is NOT fair.}$$

Based on the observed frequency table, the size of the sample is 60. Using the *cell probabilities* in H_0 , we have **expected frequencies** of the 6 categories to be equal to 10. We summarize the **expected** and **observed** frequencies in the following table.

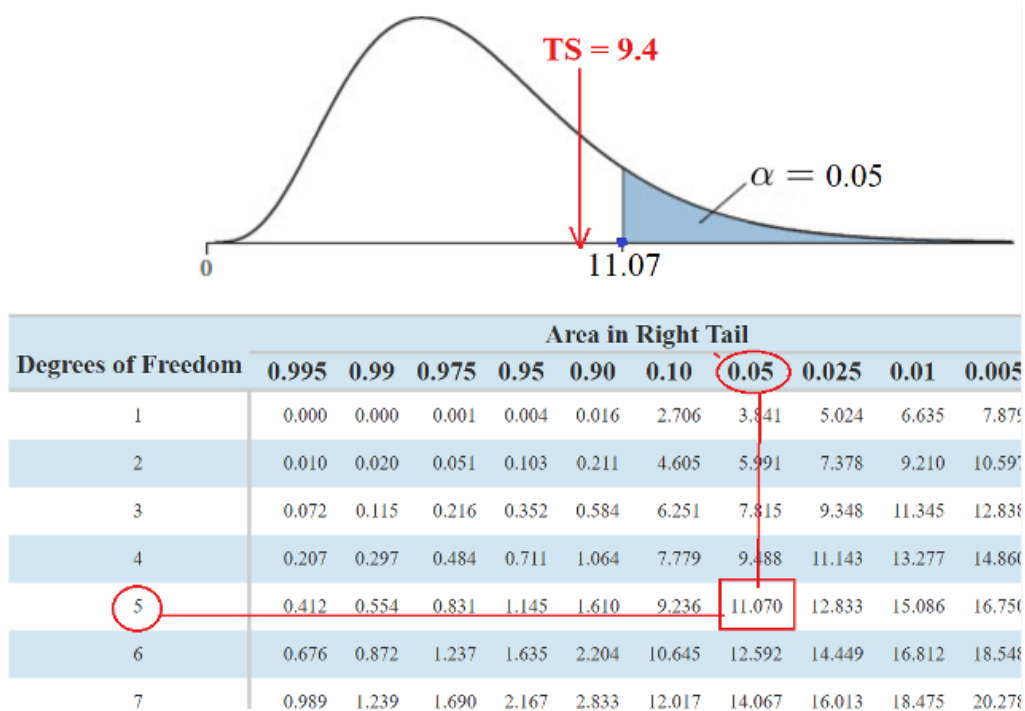
Outcome	1	2	3	4	5	6	Total
Observed	12	7	14	15	4	8	N=60
Expected	10	10	10	10	10	10	

$Np_i = E_i \Rightarrow 60 \times \frac{1}{6} \quad 60 \times \frac{1}{6} \quad 60 \times \frac{1}{6} \quad 60 \times \frac{1}{6} \quad 60 \times \frac{1}{6} \quad 60 \times \frac{1}{6}$

The test statistic for testing H_0 is defined by

$$G^2 = \frac{(12-10)^2}{10} + \frac{(7-10)^2}{10} + \frac{(14-10)^2}{10} + \frac{(15-10)^2}{10} + \frac{(4-10)^2}{10} + \frac{(8-10)^2}{10} = (4+9+16+25+36+4)/10 = 9.4$$

Since the categorical variable (number of dots on the faces of the 6-sided die) has 6 categories. The test statistic has 5 degrees of freedom. The critical value based on the significance level of 0.05 is given in the following figure



The test statistic is NOT in the rejection region. We fail to reject the null hypothesis. The die is a fair die.

Example 4. Grade distribution: A statistics teacher claims that, on average, 20% of her students get a grade of A, 35% get a B, 25% get a C, 10% get a D, and 10% get an F. The grades of a random sample of 100 students were recorded. The following table presents the sample frequencies of each grade.

Grade	A	B	C	D	F
Observed	29	42	20	5	4

Solution: Based on the given information. $N = 100$. The null hypothesis and the alternative hypothesis are

$$H_0 : p_A = 0.2, p_B = 0.35, p_C = 0.25, p_D = 0.1, p_F = 0.1 \quad v.s. H_a : \text{The claimed distribution is wrong.}$$

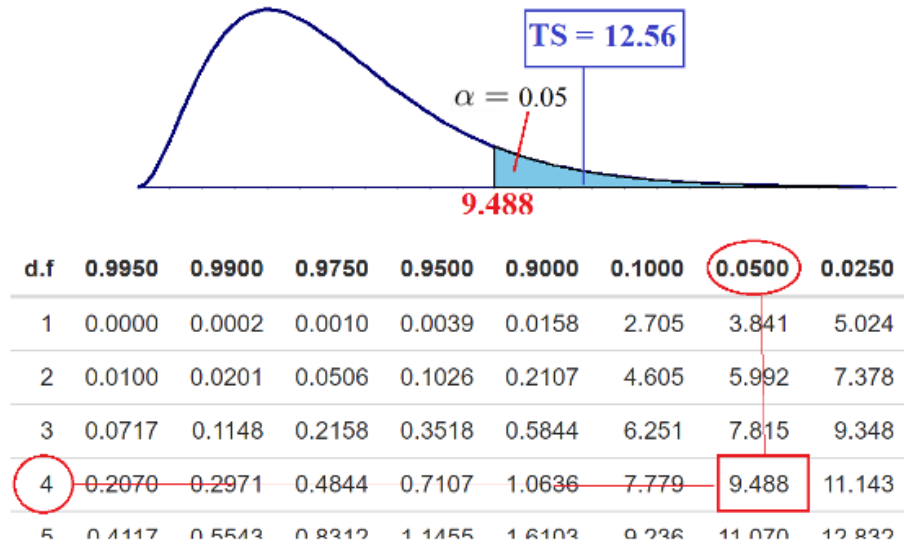
Under the null hypothesis, the expected table is given by

Grade	A	B	C	D	F
Expected	20	35	25	10	10

The test statistic has 4 degrees of freedom and the value is calculated as follows

$$G^2 = \frac{(29 - 20)^2}{20} + \frac{(42 - 35)^2}{35} + \frac{(20 - 25)^2}{25} + \frac{(5 - 10)^2}{10} + \frac{(4 - 10)^2}{10} = 12.56$$

The critical value with a significance level of 0.05 is given by



Since the test statistic is inside the rejection region, we reject the null hypothesis. That is, the sample does not support the claimed grade distribution in H_0 .

Remark: In the chi-square goodness-of-fit test, the crucial step is to find the *expected** frequency table under the null hypothesis.

Here is another example that is different from the above example

3 Chi-square Test of Independence

Let X and Y be two categorical variables with k and m categories respectively. Their relationship between X and Y is characterized by their joint distribution (table). For simplicity, we use the following two special

categorical to explain the ideas of statistical testing of independence.

3.1 Independence of Two Categorical Variables

We use the following example to illustrate **independence** and **dependence** between two categorical variables.

Example 5. *Joint probabilities and contingency tables.* Let X = political preference (Democrat vs Republican) and Y = gender (Male and Female). Let's **assume** their joint distribution to be of the following *contingency table*.

	Democrat	Republican	Row total
Male	$p_{11} = 0.3$	$p_{12} = 0.2$	$p_{1*} = 0.50$
Female	$p_{21} = 0.3$	$p_{22} = 0.3$	$p_{2*} = 0.50$
Column Total	$p_{*1} = 0.50$	$p_{*2} = 0.50$	$p_{**} = 1.00$

Column marginal probabilities

Row marginal probabilities

The cell numbers are joint probabilities. For example, $p_{12} = 0.2 = 20\%$ says 20% of the *study population* are male republicans. The row and column totals represent the percentage of male/female and democrats/republicans in the study population. Any observed data table is **governed** by the above joint distribution table.

Definition Two categorical variables are **independent** if and only if their joint probabilities are equal to the product of their corresponding marginal probabilities.

With this definition, we can see that X and Y with joint distribution specified in the above table (in Example 5) are NOT independent since $p_{11} = 0.3 \neq 0.5 \times 0.5 = 0.25$.

Example 6. We consider two variables X = preference of hair color (Blonde and Brunette) and Y = gender (Male and Female). Assume the joint distribution of the two variables is given by

	Male	Female	total
Blonde	0.18	0.27	0.45
Brunette	0.22	0.33	0.55
total	0.40	0.60	1.00

Based on the definition of independence. The preference for hair color is independent of gender. Since **all** joint probabilities are equal to the product of their corresponding marginal probabilities.

$$0.45 \times 0.40 = 0.18, 0.45 \times 0.60 = 0.27, 0.55 \times 0.40 = 0.22, \text{ and } 0.55 \times 0.60 = 0.33.$$

3.2 Expected Table Under Independence Assumption (H_0)

We construct the **expected table** under the **null hypothesis of independence** and the **observed contingency table**. For ease of interpretation, we use an example to illustrate the steps for obtaining the expected table.

Example 7. Consider the potential dependence between the attendance (good vs poor) and course grade (pass vs fail). We take 50 students from a population and obtain the following observed table.

	Pass	Fail	total
Good	25	2	27
Poor	8	15	23
total	33	17	50

Question: Whether the attendance is independent of class performance?

H_0 : attendance is independent of the performance

versus

H_a : attendance is dependent of the performance

To obtain the expected table, we follow the next few steps.

1. Estimate the marginal probabilities

	Pass	Fail	Marginal Probability
Good			0.54
Poor			0.46
Marginal Probability	0.66	0.34	1.00

where marginal probabilities are calculated by $\Pr(\text{Good}) = 27/50 = 0.54$, $\Pr(\text{Poor}) = 23/50 = 0.46$, $\Pr(\text{Pass}) = 33/50 = 0.66$, $\Pr(\text{Fail}) = 17/50 = 0.34$.

2. Estimate the joint probability under the null hypothesis of independence

	Pass	Fail	Marginal Probability
Good	0.3564	0.1836	0.54
Poor	0.3036	0.1564	0.46
Marginal Probability	0.66	0.34	1.00

where joint probabilities under the independence assumption (H_0) are calculated by taking the product of the corresponding marginal probabilities. For example, $0.54 \times 0.66 = 0.3564$.

3. Calculate Expected Table

The expected frequencies are calculated in the following table (with detailed steps).

	Pass	Fail	Row Total
Good	$0.3564 \times 50 = 17.82$	$0.1836 \times 50 = 9.18$	27
Poor	$0.3036 \times 50 = 15.18$	$0.1564 \times 50 = 7.82$	23
Column Total	33	17	Sample size = 50

Remark. For categorical variables with **more than two categories**, the expected table can be found using **** the same 3 steps**** as those used in the above example.

3.3 Formulation of Chi-squares Test of Independence

The test statistic used to test the independence of two categorical variables is the same as that used in the goodness-of-fit test. That is the standardized “distance” between the observed and the expected table (under H_0).

Assume that the two categorical variables have k and m categories respectively, then the resulting test statistic has a chi-square distribution with $(k - 1) \times (m - 1)$ degrees of freedom.

Example 8. [Continuation of **Example 7**]. Test whether **attendance** and **class performance**.

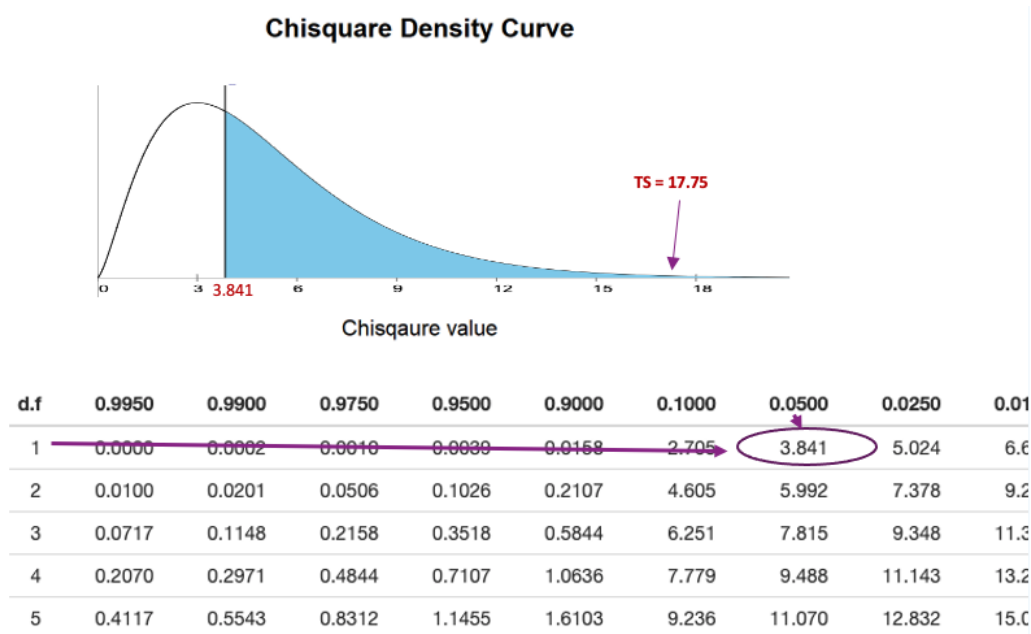
Solution: We have found the expected table under H_0 in **Example 7**, we put the observed and expected tables in the following.

Observed Table				Expected Table			
	Pass	Fail	total		Pass	Fail	total
Good	25	2	27	Good	17.82	9.18	27
Poor	8	15	23	Poor	15.18	7.82	23
total	33	17	50	total	33	17	50

The test statistic is given by

$$TS = \frac{(25 - 18.82)^2}{17.82} + \frac{(2 - 9.18)^2}{9.18} + \frac{(8 - 15.18)^2}{15.18} + \frac{(15 - 7.82)^2}{7.82} = 17.75$$

The test statistic has a chi-square distribution with $(2 - 1) \times (2 - 1) = 1$ degrees of freedom. The critical value at the significance level of 0.05 is found in the following figure.



Since the test statistic is inside the rejection region, we reject the null hypothesis that attendance and class performance are independent.

Example 9. Do some college majors require more studying than others? The National Survey of Student Engagement asked a number of college freshmen what their major was and how many hours per week they spent studying, on average. A sample of 1000 of these students was chosen, and the numbers of students in each category are tabulated in the following two-way contingency table.

Hours Studying Per Week	Major				Total
	Humanities	Social Science	Business	Engineering	
0-10	68	106	131	40	345
11-20	119	103	127	81	430
More Than 20	70	52	51	52	225
Total	257	261	309	173	1000

Solution: The null and alternative hypotheses are given by

Ho: studying time is INDEPENDENT on majors
versus

Ha: studying time is DEPENDENT on majors.

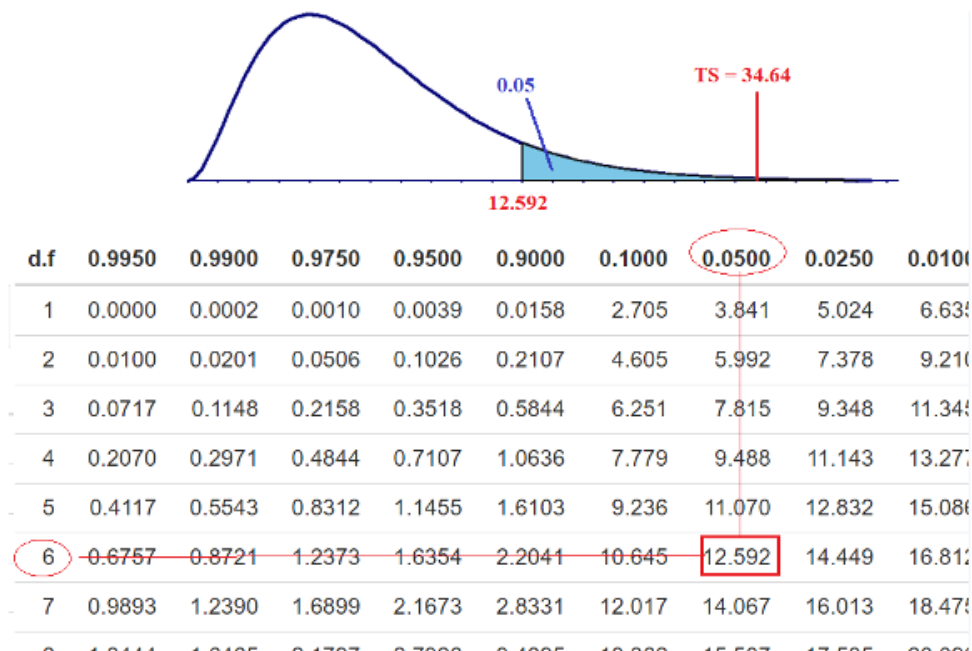
Under the null hypothesis, we obtained the expected table using the same steps in Example 7 in the following.

	Humanities	Soc Sci	Business	Engineering	Total
0-10	88.7	90.0	106.6	59.7	345
11-20	110.5	112.2	132.87	74.4	430
>20	57.8	58.7	69.5	38.9	225
total	457	261	309	173	1000

The test statistic is given by

$$TS = \frac{(68 - 88.7)^2}{88.7} + \frac{(106 - 90.0)^2}{90.0} + \dots + \frac{(52 - 38.9)^2}{38.9} \approx 34.64$$

The critical value and rejection region based on significance level 0.05 is given by



Conclusion: Since the test statistic is inside the rejection region, we reject the null hypothesis and conclude that the studying time is dependent on the majors.

4 Use of Technology

Two apps were created for the two chi-square tests of goodness-of-fit and independence respectively.

4.1 Goodness-of-fit Chi-square

The app is at: <https://chpeng.shinyapps.io/chisq-gof/>. The following screenshot illustrates the use of this app using Example 03 in this note.


IntroStatsApps: Chi-squared χ^2 Goodness-of-fit Test

Instructions:

The table in the panel 2 is editable. You double click the cell to modify the default value.

1. Type the cell counts in the first column. All values must be **non-negative integers**.

2. Type the cell probabilities in the null hypothesis (Ho). Please make sure the cell probabilities in the right column **must add up to 1** and **each individual cell value is between 0 and 1**.



Report bugs to C. Peng

Input Shelf Table:

	Observed.value	Prob.in.Ho
1	12	0.1666667
2	7	0.1666667
3	14	0.1666667
4	15	0.1666667
5	4	0.1666667
6	8	0.1666667
7	0	0
8	0	0
9	0	0
10	0	0

Summarized Input:

Observed vs Expected Counts

Observed.Value	Expected.Value
12.00	10.00
7.00	10.00
14.00	10.00
15.00	10.00
4.00	10.00
8.00	10.00

χ^2 Test Results

Null Hypothesis
Ho: The data follows the designated distribution.

Test Statistic (TS):
 $\chi^2 = 9.4$.

Calculation of TS:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(12-10)^2}{10} + \frac{(7-10)^2}{10} + \frac{(14-10)^2}{10} + \frac{(15-10)^2}{10} + \frac{(4-10)^2}{10} + \frac{(8-10)^2}{10} = 9.4$$

P-value:
The above χ^2 test statistic with 5 degrees of freedom yields a p-value = 0.0941.

4.2 Chi-square Test of Independence

The app is at <https://chpeng.shinyapps.io/chisq-independence/>. We use this app with Example 09.


IntroStatsApps: Chi-squared (χ^2) Independent Test

How many rows?
☐ 2 ☒ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10 ☐ 11 ☐ 12

How many columns?
☐ 2 ☐ 3 ☒ 4 ☐ 5

Enter Counts Here

	1	2	3	4
1	68	106	131	40
2	119	103	127	81
3	70	52	51	52



Report bugs to C. Peng

HYPOTHESES

Ho: Row and Column are independent.
Ha: Row and Column are dependent.

OBSERVED & EXPECTED COUNTS

	Observed	Expected	Col 1	Col 2	Col 3	Col 4	Total
Row 1	68	106	131	40			345
Row 2	119	103	127	81			430
Row 3	70	52	51	52			225
Total	257	261	309	173			1000

TEST STATISTIC

$\chi^2 = 34.64$

CALCULATION

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(68-88.7)^2}{88.7} + \frac{(106-90)^2}{90} + \frac{(131-106.6)^2}{106.6} + \frac{(40-59.7)^2}{59.7} + \frac{(119-110.5)^2}{110.5} + \frac{(103-112.2)^2}{112.2} + \frac{(127-132.9)^2}{132.9} + \frac{(81-74.4)^2}{74.4} + \frac{(70-57.8)^2}{57.8} + \frac{(52-58.7)^2}{58.7} + \frac{(51-69.5)^2}{69.5} + \frac{(52-38.9)^2}{38.9}$$

NULL DISTRIBUTION OF TEST STATISTIC

χ^2 distribution with df = 6

P-VALUE

p-value < 0.0001

5 Practice Exercises

You can use the apps to do the following exercises.

1. College Sports

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

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

We specifically want to compare the distribution of responses in the sample to the distribution reported the previous year (i.e., 60%, 25%, 15% reporting no, sporadic and regular exercise, respectively). Whether the data supports the above distribution at a significance level of 0.05.

2. Political Affiliation and Opinion

The following table based on the sample will be used to explore the relationship between Party Affiliation and Opinion on Tax Reform.

	favor	indifferent	opposed	total
democrat	138	83	64	285
republican	64	67	84	215
total	202	150	148	500

Find the expected counts for all of the cells.

3. Tire Quality

The operations manager of a company that manufactures tires wants to determine whether there are any differences in the quality of work among the three daily shifts. She randomly selects 496 tires and carefully inspects them. Each tire is either classified as perfect, satisfactory, or defective, and the shift that produced it is also recorded. The two categorical variables of interest are the shift and condition of the tire produced. The data can be summarized by the accompanying two-way table. Does the data provide sufficient evidence at the 5% significance level to infer that there are differences in quality among the three shifts?

	Perfect	Satisfactory	Defective	Total
Shift 1	106	124	1	231
Shift 2	67	85	1	153
Shift 3	37	72	3	112
Total	210	281	5	496

4. Condiment preference and gender

A food services manager for a baseball park wants to know if there is a relationship between gender (male or female) and the preferred condiment on a hot dog. The following table summarizes the results. Test the

hypothesis with a significance level of 10%.

		Condiment			
		Ketchup	Mustard	Relish	Total
Gender	Male	15	23	10	48
	Female	25	19	8	52
	Total	40	42	18	100