Data collection, Analysis and Inference

Subject Code: CPE-RPE,

May 2021,
SRM Univeristy-AP, Andhrapradesh

Lecture- 9: Large Sample Tests and Small Sample Tests: χ^2-test

Aim:

- To interpret the χ^2 (Chi-square) probability distribution as the sample size changes.
- Conduct and interpret chi-square tests for
  - goodness-of-fit hypothesis.
  - independence hypothesis.
  - single variance hypothesis.
  - homogeneity hypothesis.

- Have you ever wondered
  - if lottery numbers were evenly distributed or if some numbers occurred with a greater frequency?
  - 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 using chi-square distribution.

Three major applications of the chi-square distribution

1. The goodness-of-fit test, which determines if data fit a particular distribution, such as in the lottery example.

2. The test of independence, which determines if events are independent, such as in the movie example

3. the test of a single variance, which tests variability, such as in the coffee example.

The Chi-Square Distribution

- For the χ^2 distribution,

  the population mean is $\mu = df$ and

  the population standard deviation is $\sigma = \sqrt{(2df)}$.

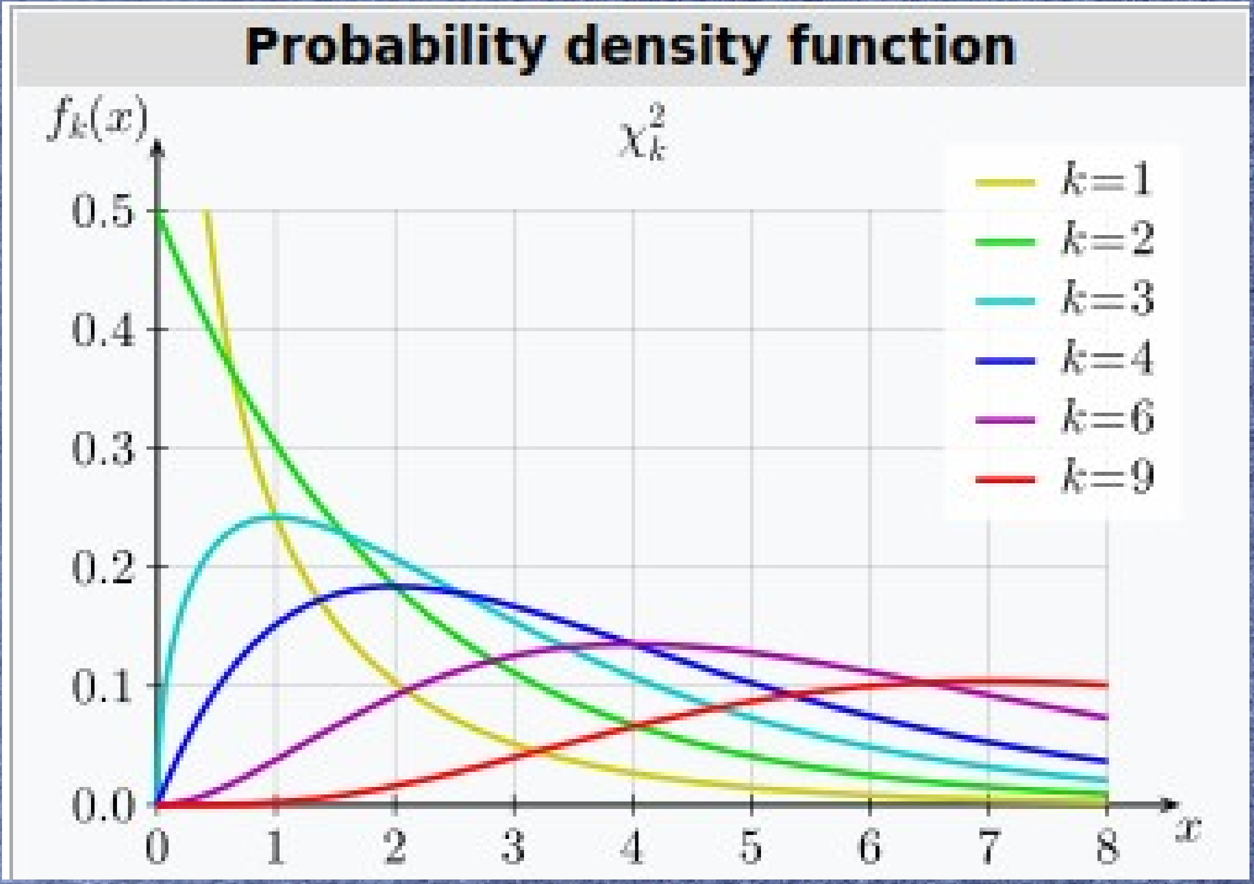
- Notation : $\chi \sim \chi^2_{df}$

  where df = n-1=degrees of freedom which depends on how chi-square is being used.

Caution *: (The degrees of freedom for the three major uses are each calculated differently.)*

Probability density function

$$f(x;\,k) = \begin{cases} \dfrac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\,\Gamma\left(\frac{k}{2}\right)}, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$



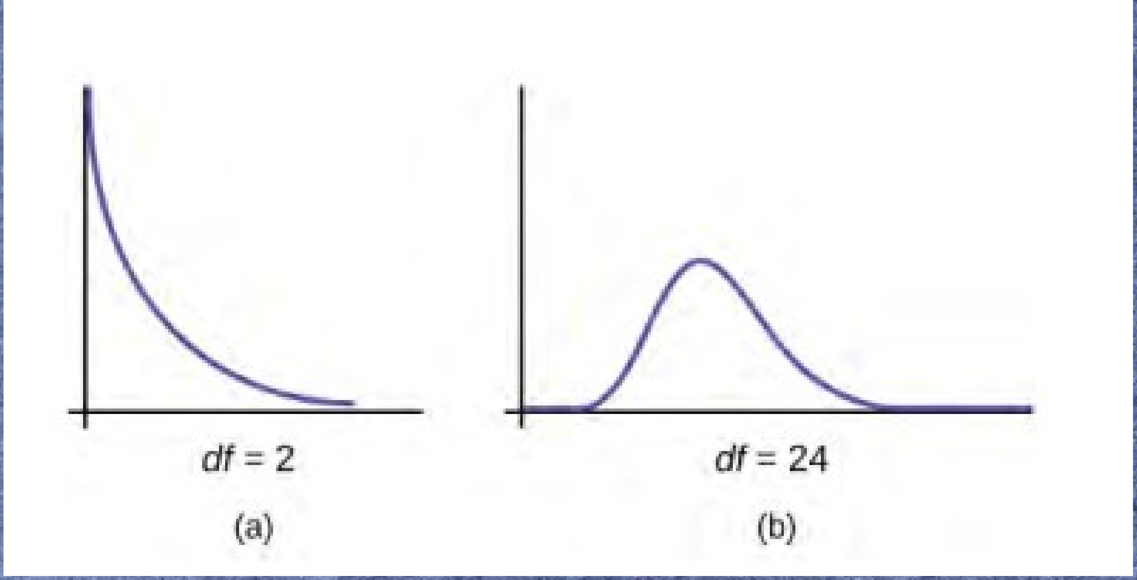
Probability density function

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

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

1. The curve is non symmetrical and skewed to the right.

2. There is a different chi-square curve for each df.



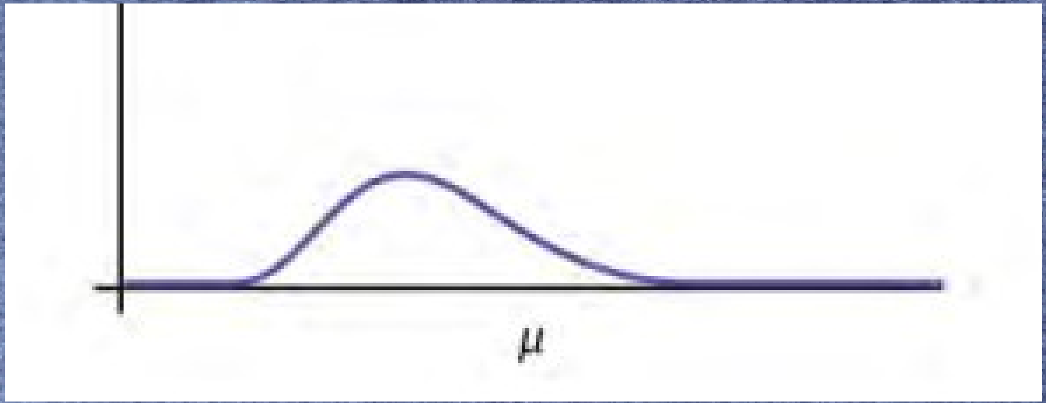
3. The test statistic for any test is always greater than or equal to zero.

4. When df > 90, the chi-square curve approximates the normal distribution. For $X \sim \chi^2_{1,000}$

the mean, $\mu = df = 1{,}000$

and the standard deviation, $\sigma = \sqrt{2(1{,}000)} = 44.7$. Therefore, $X \sim N(1{,}000, 44.7)$, approximately.

5. The mean, μ, is located just to the right of the peak.



Goodness-of-Fit Test

- 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 !
- **The null and the alternative hypotheses for this test may be written *<u>in sentences</u> or may be stated as equations or inequalities.***

Test-statistic (χ^{2-}score)

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

$$\sum_{k} \frac{(O-E)^{2}}{E}$$

where:

- O = observed values (data)
- E = expected values (from theory)
- k = 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(model) were true.

Right- tailed / Right-skewed

The number of degrees of freedom is

df = (number of categories – 1).

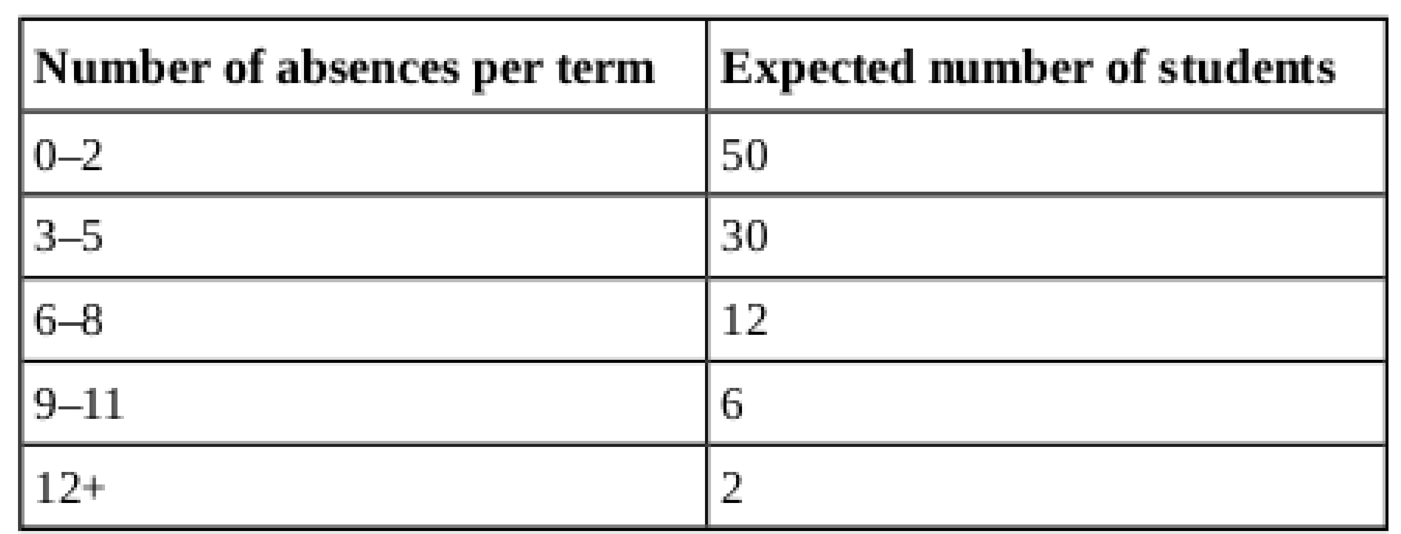
The goodness-of-fit test is almost always right-tailed.

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.

- *Note: The expected value for each cell needs to be **at least five** in order for you to use this test.*

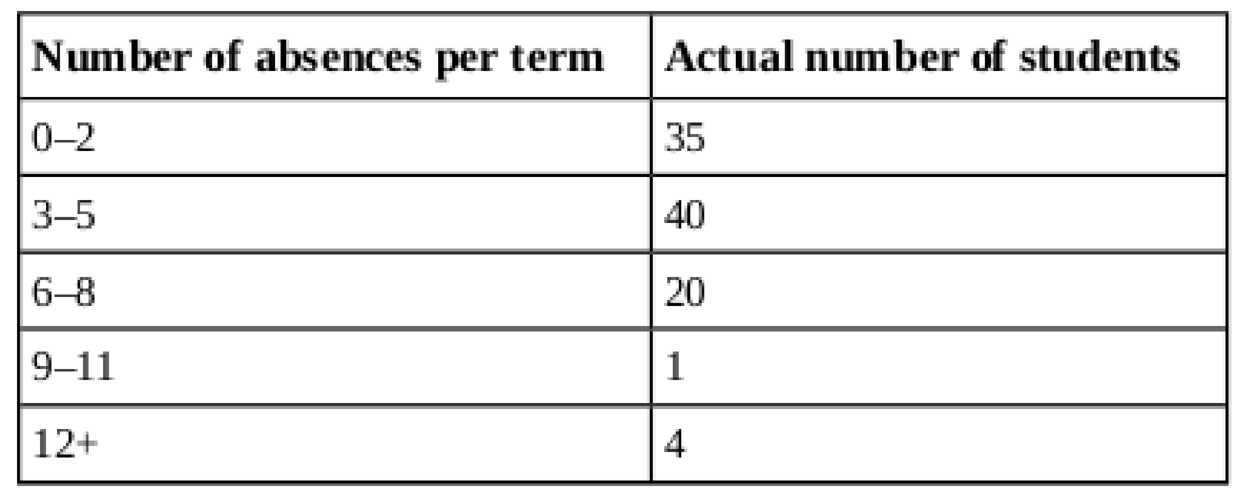
Faculty perception : *Absenteeism of college students from math classes is a major concern to math instructors because missing class appears to increase the drop rate.*

- Suppose that a study was done to determine if the actual student absenteeism rate follows faculty perception. The faculty **expected** that a group of 100 students would miss class according to



Number of absences per term	Expected number of students
0–2	50
3–5	30
6–8	12
9–11	6
12+	2

Hence, a random survey across all mathematics courses was then done to determine the actual number (observed) of absences in a course.



Number of absences per term	Actual number of students
0–2	35
3–5	40
6–8	20
9–11	1
12+	4

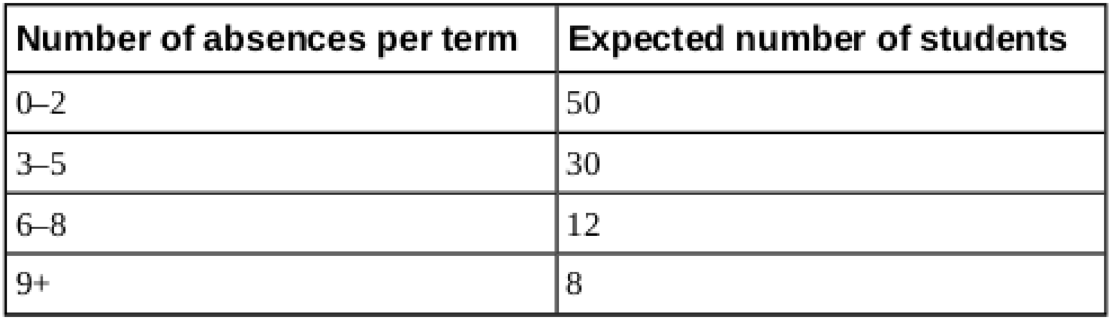
To conduct a goodness-of-fit test for the hypotheses.

H_0 : Student absenteeism fits faculty perception.

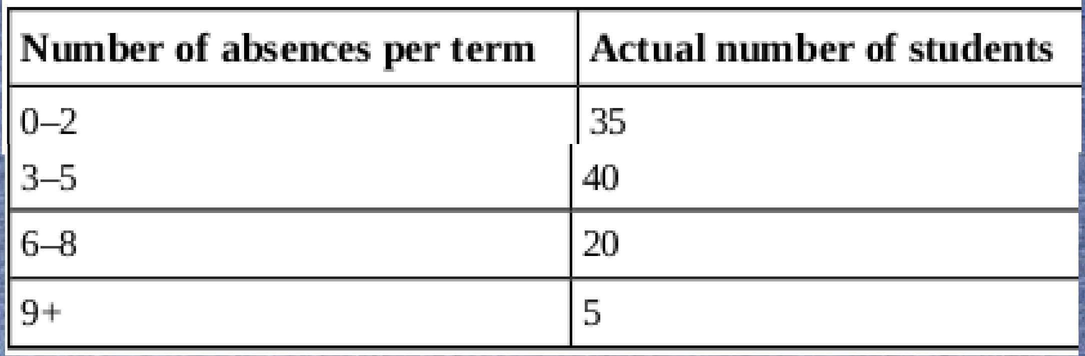
H_a : Student absenteeism does not fit faculty perception.

We can not use the information as it appears in the charts to conduct the goodness-of-fit test (why?)

- Absences for the "12+" entry is less than five !
- Remedy : Combine that group with the "9–11" group



Number of absences per term	Expected number of students
0–2	50
3–5	30
6–8	12
9+	8

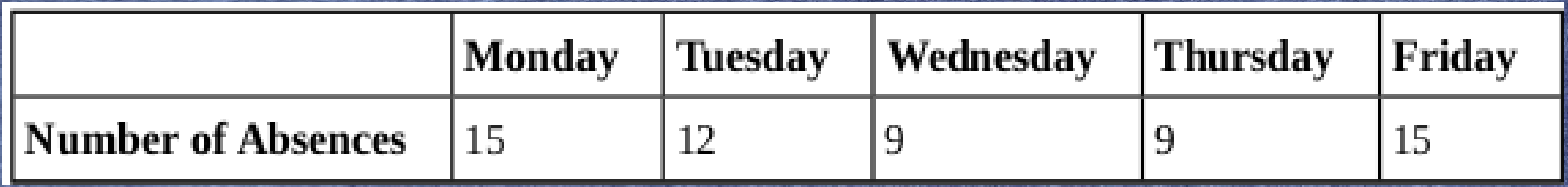


Number of absences per term	Actual number of students
0–2	35
3–5	40
6–8	20
9+	5

There are four "cells" or categories in each of the new tables.
df = number of cells – 1 = 4 – 1 = 3

Example : Employers 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***.

Suppose a random sample of 60 managers were asked on which day of the week they had the highest number of employee absences. The results were distributed as in the Table.



	Monday	Tuesday	Wednesday	Thursday	Friday
Number of Absences	15	12	9	9	15

For the population of employees, do the days for the highest number of absences occur with equal frequencies during a five-day work week? Test at a 5% significance level.

- Solution: The null and alternative 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 60 absent days there would be 12 absences on each week day. These numbers are the expected (E) values. The values in the table are the observed (O) values or data.

Calculate the χ^2 test statistic by hand. Make a chart with the following headings and fill in the columns:

- Expected (E) values (12, 12, 12, 12, 12)
- Observed (O) values (15, 12, 9, 9, 15)
- (O – E), • $(O – E)^2$ • $(O – E)^2/E$

Now sum the last column.

The sum is $\{(12\text{-}15)^2+(12\text{-}12)^2+(12\text{-}9)^2+(12\text{-}9)^2+(12\text{-}15)^{\wedge 2}\}/12=36/12=3$.

- This is the χ^2 test statistic.

To find the p-value, calculate $P(\chi^2 > 3)$.

- This test is right-tailed. (*Use a computer or calculator to find the p-value*. You should get p-value = 0.5578.)

(The dfs are the number of cells – 1 = 5 – 1 = 4)

- The decision is not to 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.

The test of independence

A test of independence determines whether two factors are independent or not.

- If A and B are independent then

$$P(A \text{ AND } B) = P(A)P(B)$$

In a test of independence,

the contingency table (row-column-wise data) consists of two factors, the null hypothesis states that the factors are independent and the alternative hypothesis states that they are not independent (dependent).

Computing the expected number (E) from the independence relation

- The test of independence is always right-tailed because of the calculation of the test statistic.
- The number of degrees of freedom for the test of independence is:

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

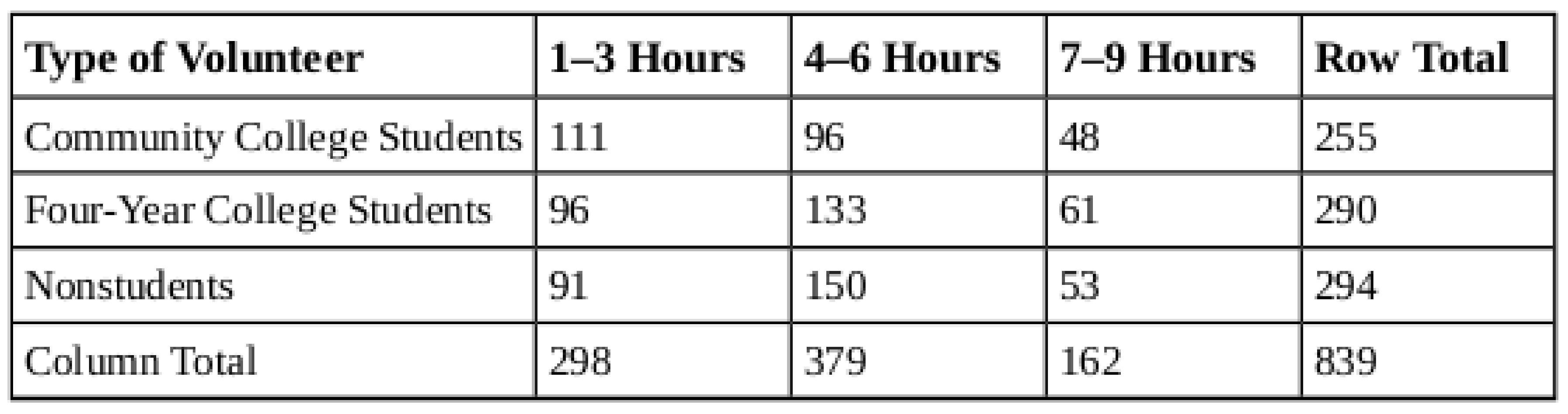
- The following formula calculates the expected number (E):

E =(row total)(column total)/total number surveyed

Volunteer groups vs Hours per week

- In a volunteer group, *adults 21 and older* volunteer from 1-9 hours each week to spend time with *a disabled senior citizen*. The program recruits among community college students, four-year college students, and non-students.
- Contingency Table is a sample of the adult volunteers and the number of hours they volunteer per week. The table contains observed (O) values (data).

Is the number of hours volunteered independent of the type of volunteer?



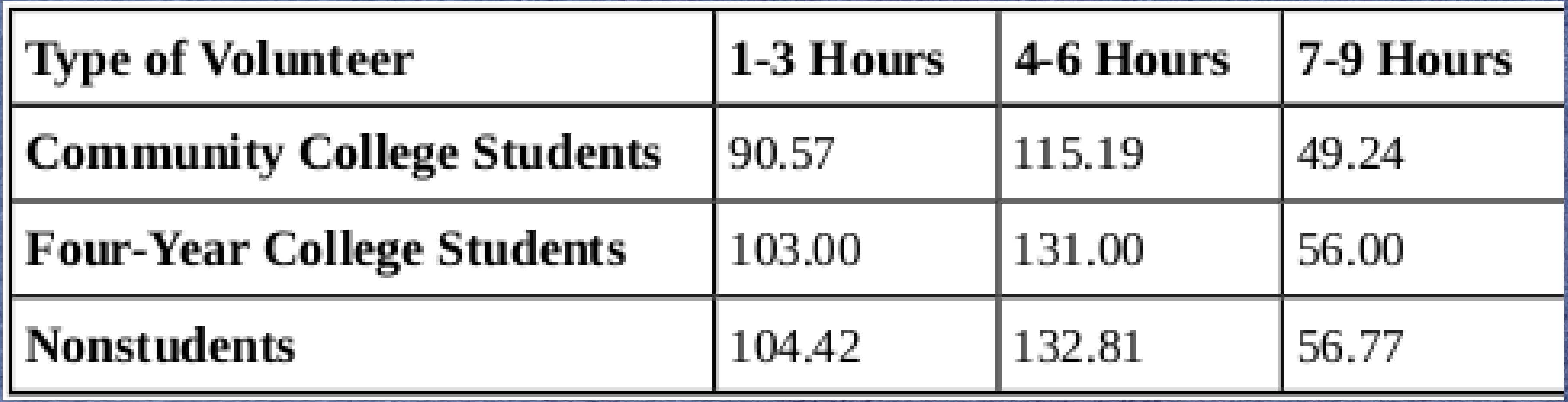
Type of Volunteer	1–3 Hours	4–6 Hours	7–9 Hours	Row Total
Community College Students	111	96	48	255
Four-Year College Students	96	133	61	290
Nonstudents	91	150	53	294
Column Total	298	379	162	839

- The two factors are number of hours volunteered and type of volunteer.
- This test is always right-tailed.

H_0 : The number of hours volunteered is independent of the type of volunteer.

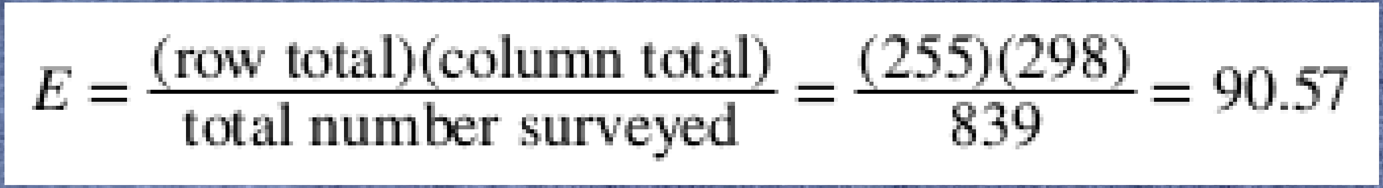
H_a : The number of hours volunteered is dependent on the type of volunteer.

The expected result are computed as



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

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

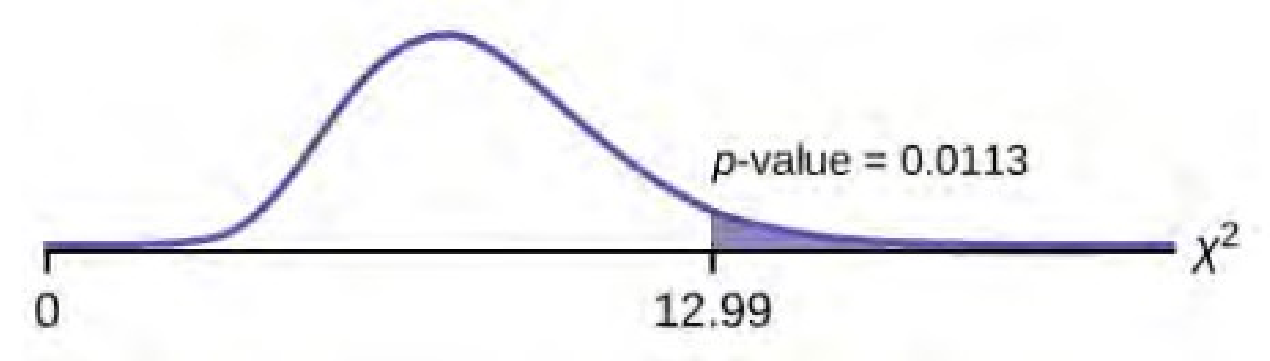


$$E = \frac{(\text{row total})(\text{column total})}{\text{total number surveyed}} = \frac{(255)(298)}{839} = 90.57$$

Calculated the test statistic: $\chi^2 = 12.99$

- Distribution for the test: $\chi^2_{df=4}$

- df= (3 columns – 1)(3 rows – 1) = (2)(2) = 4



p-value=P(χ^2 > 12.99) = 0.0113

- Assume **β** = 0.05. then p-value=**0.0113** < **β**

Hence, reject H_0 , this means that the factors are *not independent.*

<u>Conclusion</u>: *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.*

Test of a Single Variance

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.

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

Here s is the random variable in this test.

The number of degrees of freedom is df = n – 1

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

The test statistic is: $\chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$

where:

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

Example :

Suppose a math instructor believes that the standard deviation for his final exam is five points. One of his best students thinks otherwise. The student claims that the standard deviation is more than five points.

If the student were to conduct a hypothesis test, what would the null and alternative hypotheses be?

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

- $H_0 : \sigma^2 = 5^2$
- $H_a : \sigma^2 > 5^2$

- We need the sample variance s^2 and sample size n to compute the chi-square statistic:

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2}$$

- Make an inference with 5% level of significance, **β**

Test for Homogeneity

- The goodness–of–fit test can be used to decide whether a population fits a given distribution, but it will not suffice to decide whether two populations follow the same unknown distribution (Homogeneity).
- A different test, called the test for homogeneity, can be used to draw a conclusion about whether two populations have the same distribution.
- To calculate the test statistic for a test for homogeneity, follow the same procedure as with the test of independence.
- But with the Degrees of Freedom

  df = number of columns - 1

- END