Introduction:

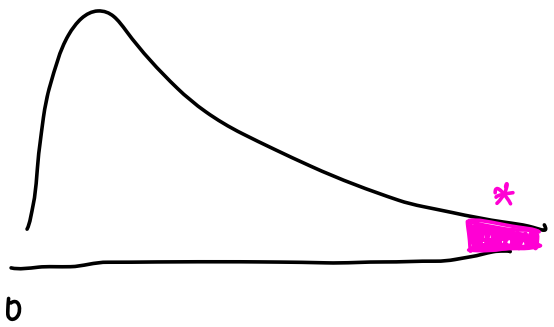
In this chapter, we would encounter another kind of tests, called chi-squared tests.

If you recall from last chapter, the logic of hypothesis testing, if, based on some advanced statistical theory, you know the distribution of your statistic under null hypothesis, then to do hypothesis testing, you first ASSUME that null hypothesis is true, and compute the probability of seeing current data or something more extreme (we call this p-value), if p-value is small, it means it is unlikely to see current data under null hypothesis, so we reject the null.

As a result, chi-squared tests are just another way to perform hypothesis testing. It is called chi-squared test since under null hypothesis, our test statistic would follow a chi-squared distribution.

properties of chi-squared distribution

graph



1. the chi-squared variable is nonnegative.
2. since it is nonnegative, the extreme value only happens on the positive tail.* Therefore, you would never see people discussing if it is one-sided or two-sided test.

3. you have to determine the degree of freedom.

Two kinds of chi-squared tests

1. chi-square goodness of fit test:

(a) The chi-square goodness of fit test is a statistical hypothesis test used to determine whether a variable is likely to come from a specific distribution or not.

(b) You can use the test when you have counts of values for a categorical variable.

(c) The null hypothesis is usually like:
variable A follows specific distribution
The alternative hypothesis is like:
variable A does not follow specific distribution

(d) The test statistic takes the form of

$$\frac{(\text{observed} - \text{Expected})^2}{\text{Expected}}$$

(e) the degree of freedom is (number of categories - 1)

(f) you reject the null hypothesis when your p-value is less than pre-determined significance level.

2. Chi-Square test of Independence:

(a) The chi-square test of Independence is a statistical hypothesis test used to determine whether two categorical variables are independent or not.

(b) You can use the test when you have counts of values for two categorical variables.

(c) The null hypothesis is usually like:
Two variables are independent.
The alternative hypothesis is like:
Two variables are not independent.

(d) The test statistic takes the form of the sum of $\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$ from each cell, which is (row total x column total) ÷ overall total

(e) the degree of freedom is
(number of columns - 1) X (number of rows - 1)

(f) You reject the null when your p-value is less than pre-determined significance level.

Exercises with Solutions

Exercise 1. – Goodness of Fit

We collect a random sample of ten bags. Each bag has 100 pieces of candy and five flavors. We want to test if the proportions of the five flavors in each bag are the same.
The following table shows the total counts of each flavor from 10 bags.

Flavor	Apple	Lime	Cherry	Orange	Grape
Number of pieces of candy	180	250	120	225	225

Set up the hypotheses and perform suitable test to make a conclusion, at 5% significance level.

Solution:

Step 1: We set up the hypotheses.

Null: Each flavor has the same proportion

Alternative: Each flavor does not have the same proportion.

Step 2: Set up the expected county table

under null hypothesis, each flavor has the same counts. We have 1,000 pieces of candy in total, thus, under null hypothesis, each flavor has $1,000 \div 5 = 200$ counts.

step 3: compute the statistic $\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$ for each cell.

Flavor	Apple	Lime	Cherry	Orange	Grape
Observed	180	250	120	225	225
Expected	200	200	200	200	200
observed - Expected	-20	50	-80	25	25
$(\text{observed} - \text{Expected})^2$	400	2500	6400	625	625
$\frac{(\text{observed} - \text{Expected})^2}{\text{Expected}}$	$\frac{400}{200}$	$\frac{2500}{200}$	$\frac{6400}{200}$	$\frac{625}{200}$	$\frac{625}{200}$

Finally, sum each term from the last row.

$$\frac{400}{200} + \frac{2500}{200} + \frac{6400}{200} + \frac{625}{200} + \frac{625}{200}$$

$$= 2 + 12.5 + 32 + 3.125 + 3.125$$

$$= 52.75$$

step 4: determine the degree of freedom is

(Number of flavors - 1) = (5 - 1) = 4

step 5: check the chi-square table, when significance level is 0.05 with 4 degrees of freedom, the critical value is 9.488, and as we know the extreme value

of chi-square is positive, $52.75 > 9.488$ more extreme, the p-value thus is less than 0.05, we would reject the null hypothesis.

Exercise 2 – Independence test

We have data for 600 people who saw a movie at our theater. For each person, we know the type of movie they saw and whether or not they bought snacks. The following table shows the data.

Type of movie	Action	Comedy	Family	Horror
Snacks	50	125	90	45
No Snacks	75	175	30	10

perform suitable test and make a conclusion at 5% significance level.

Solution:

It is important to calculate the row/column/overall total first.

Observed Table

counts of people having snacks and watch action

Type of movie	Action	Comedy	Family	Horror	row total
snacks	50	125	90	45	310
No snacks	75	175	30	10	290
column total	125	300	120	55	600

counts of people watching action

counts of people having snacks

Expected table

Type of movies	Action	Comedy	Family	Horror	row total
snacks					310
No snacks					290
column total	125	300	120	55	600

(Action, snack)'s Expected counts = $\frac{310 \times 125}{600} = 64.58$

see the arrows *

(Action, no snack)'s expected County = $\frac{290 \times 125}{600} = 60.42$

see the arrows *

With similar approach, we can fill in each cell of the Expected table

Type of movies	Action	Comedy	Family	Horror	row total
snacks	64.58	155	62	28.42	310
No snacks	60.42	145	58	26.58	290
column total	125	300	120	55	600

thus the test statistic is the sum of $\frac{(\text{observed}-\text{Expected})^2}{\text{Expected}}$

$$\frac{(50-64.58)^2}{64.58} + \frac{(75-60.42)^2}{60.42} + \frac{(125-155)^2}{155}$$

$$+ \frac{(175-145)^2}{145} + \frac{(90-62)^2}{62} + \frac{(30-58)^2}{58}$$

$$+ \frac{(45-28.42)^2}{28.42} + \frac{(10-26.58)^2}{26.58}$$

$$= 3.29 + 3.52 + 5.81 + 6.21 + 12.65 + 13.52 + 9.68 + 10.35$$

$$= 65.03$$

the degree of freedom is $(4-1)\times(2-1) = 3$

the critical value with 5% significance level and three degree of freedom is 7.815, 65.03 > 7.815, so our p-value would be less than 5%, we reject the null hypothesis.

Exercise 3 - Goodness of fit test

In a typical year, out of 365 days, about 250 days are weekdays, 105 days are weekends, and 10 days are U.S. Federal holidays (if a holiday falls on a weekend, it is 'observed' on a weekday). It was desired to know whether the distribution of births matched the distribution of days in a year. A simple random sample of 400 students at a large university were asked whether they were born on a weekday, weekend, or holiday.

(a) State null and alternative hypotheses relevant to the question. *The null is that the distribution of births matches the distribution of days in the year. The alternative is that the distribution of births does not match the distribution of days in the year.*

(b) Suppose that from the 400 students sampled, 302 were born on a weekday, 92 were born on a weekend, and 6 were born on a holiday. Perform a test of the stated hypotheses at the 5% significance level. Make sure to choose an appropriate test statistic, compute the test statistic and P-value, and make a conclusion in the context of the problem.

If we think of this as a box model, under the null, the box has three kinds of tickets, weekday ($250/365 = 68.5\%$ of the tickets), weekend ($105/365 = 28.8\%$ of the tickets) and holiday ($10/365 = 2.7\%$ of tickets). We can check this using a χ^2 GOF test. The expected counts are the population percentages times the sample size, 400, so they are 274.0, 115.2, and 10.8 for weekday, weekend, and holiday, respectively. The statistic is $\chi^2 = \frac{(302-274)^2}{274} + \frac{(92-115.2)^2}{115.2} + \frac{(6-10.8)^2}{10.8} = 2.86 + 4.67 + 2.13 = 9.66$, on $3 - 1 = 2$ df. Since the expected counts are all greater than 5, we can compare to a χ^2 table, and the P-value is a little less than 1%. So we reject the null and conclude the the distribution of birth days does not match the distribution in the year. It seems people are more likely to be born on weekdays than expected.

(c) Suggest a reason why the findings from above are not surprising. *C-Sections are more typically planned for a weekday*

Exercise 4 - Independence test

An experiment was conducted to compare three pesticides (call them A, B, and C) for use on alfalfa plants. The pesticides are designed to control aphids. 250 alfalfa plants were randomly assigned to each of the three treatments. Pesticide A had 94 plants, B had 84, and C had 72. After 8 weeks, the plants were observed, and the aphids on each plant were counted. If the plant had 10 or fewer aphids, it was designated as "successful control", and if it had more than 10 aphids, it was designated as "failed control". It was desired to know if the successful control percentages were equal for the three pesticides.

(a) State null and alternative hypotheses appropriate to the study question.

Solution:

Null Hypothesis: the category of pesticides and whether the control is successful or not is independent.

Alternative : the category of pesticides and whether the control is successful or not is not independent.

(b) Suppose that the number of successful control plants for pesticides A, B, C, were 45, 40, 38. Use an appropriate test to make a decision about the hypotheses using a significance level of 1%.

Solution:

You need to first identify this is an chi-square independence test.

Based on the question, we have the following observed county table.

Observed Table

	A	B	C	Total
success	45	40	38	123
failure	49	44	34	127
Total	94	84	72	250

Expected Table

	A	B	C	Total
success	46.25	41.33	35.42	123
failure	47.75	42.67	36.58	127
Total	94	84	72	250

the expected count for each cell is computed

$$\text{Expected} = \frac{\text{row total} \times \text{column total}}{\text{overall total}}$$

For example,

$$\text{Expected}(A, \text{success}) = \frac{94 \times 123}{250} = 46.25 \ *$$

$$\text{Expected}(C, \text{failure}) = \frac{72 \times 127}{250} = 36.58 \ *$$

After you obtain both tables, you compute the test statistic.

$$\frac{(45-46.25)^2}{46.25} + \frac{(49-47.75)^2}{47.75} + \frac{(40-41.33)^2}{41.33} + \frac{(44-42.67)^2}{42.67}$$
$$+ \frac{(38-35.42)^2}{35.42} + \frac{(34-36.58)^2}{36.58}$$

$$= 0.034 + 0.043 + 0.188 + 0.033 + 0.041 + 0.182 = 0.521$$

The degree of freedom is $(2-1) \times (3-1) = 2$

Check the table to know that the p-value is greater than 70%, so we do not reject the null hypothesis.