Chapter 1

Inference for Distributions of Categorical Data

Is a hypothesized distribution valid for a categorical variable? Whether the distribution of a categorical variable differs for two or more populations or treatments? Is there an association between two categorical variables in a population? We will solve those three types of problems in this chapter by performing χ^2 tests.

1.1 Chi-square test for goodness of fit

Take a look at the following problem.

According to the Census Bureau, the distribution by ethnic background of the New York City population in a recent year was

Hispanic: 28%   Black: 24%   White: 35%   Asian: 12%   Others: 1%

The manager of a large housing complex in the city wonders whether the distribution by race of the complex's residents is consistent with the population distribution. To find out, she records data from a random sample of 800 residents. The table below displays the sample data.

Race:	**Hispanic**	**Black**	**White**	**Asian**	**Other**
Count:	212	202	270	94	22

Table 1.1: Observed residents of different races

Are these data significantly different from the citys distribution by race? Carry out an appropriate test at the $\alpha = 0.05$ level to support your answer.

- **The test statistic χ^2**

  The hypotheses for the above problem can be set as

  $\mathbf{H}_0$: The race distribution in the complex is the same as in the New York City.
  $\mathbf{H}_a$: The race distribution in the complex different from in the New York City.

  As the hypotheses tests we learned before, we need to calculate a P-value and compare it with the significance level α. Let's recall the definition of P-value.

  P-value is the probability, computed assuming H_0 is true, that the statistic(such as $\bar{x}$ and $\widehat{p}$) would be as extreme as or more extreme than the observed value, in the direction of H_a.

  Therefore, we have to find a statistic before we calculate the P-value. This statistic is χ^2.

  $$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

  χ^2 is also a *test statistic*, because it plays the same role as the *test statistic z* or *test statistic t* in calculating the P-value.

  The "Observed" in the formula means the observed value. Table 1.1 gives the observed values.

  The "Expected" in the formula means the expected value if $\mathbf{H}_0$ is true. In the above problem, the "Expected" are given below.

Race:	**Hispanic**	**Black**	**White**	**Asian**	**Other**
Expected:	$800 \times 28\% = 224$	$800 \times 24\% = 192$	$800 \times 35\% = 280$	$800 \times 12\% = 96$	$800 \times 1\% = 8$

Table 1.2: Expected number of residents of different races

Now let's plug in the values and calculate χ^2

$$\chi^2 = \frac{(212-224)^2}{224} + \frac{(202-192)^2}{192} + \frac{(270-280)^2}{280} + \frac{(94-96)^2}{96} + \frac{(22-8)^2}{8} = 26.1$$

Is a larger χ^2 or a smaller χ^2 in favour of $\mathbf{H}_a$?

- **χ^2 distribution**

We have to know the sampling distribution of χ^2 in order to calculate the P-value. χ^2 follows a χ^2 distribution with *degree of freedom* $df = n - 1$. n is the number of categories. In the above example, $n = 5$, $df = 5 - 1 = 4$.

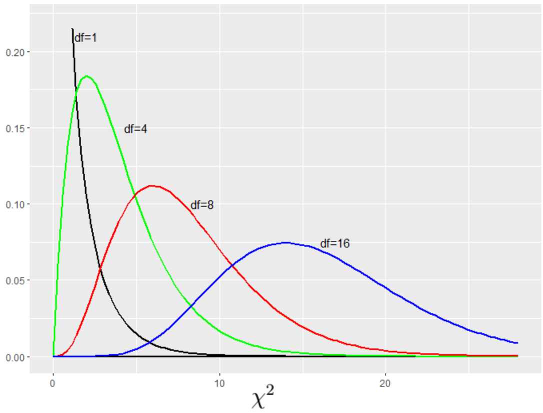

Figure 1.1: χ^2 distribution with different degree of freedom

From figure 1.1, we can see that the sampling distribution of χ^2 becomes more and more symmetric as the degree of freedom increase.

- **P-value**

According to the definition of the P-value, it should be the probability of χ^2 larger than the observed value of χ^2

$$\textbf{P-value} = \mathbf{P}(\chi^2 > 26.1) = 0.0003 \qquad df = 5 - 1 = 4$$

Clearly the P-value is smaller than the given significance level $\alpha = 0.05$. We reject $\mathbf{H}_0$, and have convincing evidence that the distribution of races in this complex is difference from that in the New York city.

- **Follow-up analysis**

If the sample data lead to a statistically significant result, we can conclude that there is convincing evidence that the distribution of the categorical variable is different from the stated one. By analysing the **components** of the χ^2, we can find out which one contributes the most.

In the above example, the *components* of χ^2 are

$$\frac{(212-224)^2}{224}, \frac{(202-192)^2}{192}, \frac{(270-280)^2}{280}, \frac{(94-96)^2}{96}, \frac{(22-8)^2}{8}$$

$\frac{(22-8)^2}{8} = 24.5$ contributes the most to $\chi^2 = 26.1$. The proportion of "other" ethnic groups is larger in this complex than in that in the New York city.

The general steps for the *Chi-square test for goodness of fit* is the same as all the other tests we learned before.

Mendelian inheritance

Biologists wish to mate pairs of fruit flies having genetic makeup RrCc, indicating that each has one dominant gene (R) and one recessive gene (r) for eye color, along with one dominant (C) and one recessive (c) gene for wing type. Each offspring will receive one gene for each of the two traits from each parent. The following table, known as a Punnett square, shows the possible combinations of genes received by the offspring:

	Parent 2 passes on:			
Parent 1 passes on:	**RC**	**Rc**	**rC**	**rc**
RC	RRCC(x)	RRCc(x)	RrCC(x)	RrCc(x)
Rc	RRCc(x)	RRcc(y)	RrCc(x)	Rrcc(y)
rC	RrCC(x)	RrCc(x)	rrCC(z)	rrCc(z)
rc	RrCc(x)	Rrcc(y)	rrCc(z)	rrcc(w)

Any offspring receiving an R gene will have red eyes, and any offspring receiving a C gene will have straight wings. So based on this Punnett square, the biologists predict a ratio of 9 red-eyed, straight-winged (x); 3 red-eyed, curly-winged (y); 3 white-eyed, straightwinged (z); 1 white-eyed, curly-winged (w) offspring.

To test their hypothesis about the distribution of offspring, the biologists mate a random sample of pairs of fruit flies. Of 200 offspring, 99 had red eyes and straight wings, 42 had red eyes and curly wings, 49 had white eyes and straight wings, and 10 had white eyes and curly wings. Do these data differ significantly from what the biologists have predicted? Carry out a test at the $\alpha = 0.01$ significance level.

1.2 Inference for two-way tables