Mathematics for Data Science

Lecture 7: Chi-Squared Tests

Mohamad GHASSANY

EFREI Paris

The hypothesis-testing procedures that we have discussed in previous sections are designed for problems in which the population or probability distribution is known and the hypotheses involve the parameters of the distribution. Another kind of hypothesis is often encountered: We do not know the underlying distribution of the population, and we wish to test the hypothesis that a particular distribution will be satisfactory as a population model. For example, we might wish to test the hypothesis that the population is normal.

We have previously discussed a very useful graphical technique for this problem called **probability plotting** and illustrated how it was applied in the case of a normal distribution. In this section, we describe a formal **goodness-of-fit test** procedure based on the chi-square distribution.

Testing for Goodness of Fit

The test procedure requires a random sample of size n from the population whose probability distribution is unknown. These n observations are arranged in a frequency histogram, having k bins or class intervals. Let O_i be the observed frequency in the ith class interval. From the hypothesized probability distribution, we compute the expected frequency in the ith class interval, denoted E_i. The test statistic is

Goodness-of-Fit Test Statistic

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{9.47}$$

It can be shown that, if the population follows the hypothesized distribution, χ_0^2 has, approximately, a chi-square distribution with $k - p - 1$ degrees of freedom, when p represents the number of parameters of the hypothesized distribution estimated by sample statistics. This approximation improves as n increases. We should reject the null hypothesis that the population is the hypothesized distribution if the test statistic is too large. Therefore, the P-value would be the probability under the chi-square distribution with $k - p - 1$ degrees of freedom above the computed value of the test statistic χ_0^2 or $P = P(\chi_{k-p-1}^2 > \chi_0^2)$. For a fixed-level test, we would reject the hypothesis that the distribution of the population is the hypothesized distribution if the calculated value of the test statistic $\chi_0^2 > \chi_{\alpha,k-p-1}^2$.

EXAMPLE 9.12 | Printed Circuit Board Defects—Poisson Distribution

The number of defects in printed circuit boards is hypothesized to follow a Poisson distribution. A random sample of $n = 60$ printed circuit boards has been collected, and the following number of defects observed.

Number of Defects	Observed Frequency
0	32
1	15
2	9
3	4

The mean of the assumed Poisson distribution in this example is unknown and must be estimated from the sample data. The estimate of the mean number of defects per board is the sample average, that is, $(32 \cdot 0 + 15 \cdot 1 + 9 \cdot 2 + 4 \cdot 3)/60 = 0.75$. From the Poisson distribution with parameter 0.75, we may compute p_i, the theoretical, hypothesized probability associated with the ith class interval. Because each class interval corresponds to a particular number of defects, we may find the p_i as follows:

$$p_1 = P(X = 0) = \frac{e^{-0.75}(0.75)^0}{0!} = 0.472$$

$$p_2 = P(X = 1) = \frac{e^{-0.75}(0.75)^1}{1!} = 0.354$$

$$p_3 = P(X = 2) = \frac{e^{-0.75}(0.75)^2}{2!} = 0.133$$

$$p_4 = P(X \geq 3) = 1 - (p_1 + p_2 + p_3) = 0.041$$

The expected frequencies are computed by multiplying the sample size $n = 60$ times the probabilities p_i. That is, $E_i = np_i$. The expected frequencies follow:

Number of Defects	Probability	Expected Frequency
0	0.472	28.32
1	0.354	21.24
2	0.133	7.98
3 (or more)	0.041	2.46

Because the expected frequency in the last cell is less than 3, we combine the last two cells:

Number of Defects	Observed Frequency	Expected Frequency
0	32	28.32
1	15	21.24
2 (or more)	13	10.44

The seven-step hypothesis-testing procedure may now be applied, using $\alpha = 0.05$, as follows:

1. **Parameter of interest:** The variable of interest is the form of the distribution of defects in printed circuit boards.
2. **Null hypothesis:** H_0: The form of the distribution of defects is Poisson.
3. **Alternative hypothesis:** H_1: The form of the distribution of defects is not Poisson.
4. **Test statistic:** The test statistic is $\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$
5. **Reject H_0 if:** Because the mean of the Poisson distribution was estimated, the preceding chi-square statistic will have $k - p - 1 = 3 - 1 - 1 = 1$ degree of freedom. Consider whether the P-value is less than 0.05.
6. **Computations:**
$$\chi_0^2 = \frac{(32 - 28.32)^2}{28.32} + \frac{(15 - 21.24)^2}{21.24} + \frac{(13 - 10.44)^2}{10.44} = 2.94$$
7. **Conclusions:** We find from Appendix Table III that $\chi^2_{0.10,1} = 2.71$ and $\chi^2_{0.05,1} = 3.84$. Because $\chi_0^2 = 2.94$ lies between these values, we conclude that the P-value is between 0.05 and 0.10. Therefore, because the P-value exceeds 0.05, we are unable to reject the null hypothesis that the distribution of defects in printed circuit boards is Poisson. The exact P-value computed from software is 0.0864.

Contingency Table Tests

Many times the n elements of a sample from a population may be classified according to two different criteria. It is then of interest to know whether the two methods of classification are statistically **independent**; for example, we may consider the population of graduating engineers and may wish to determine whether starting salary is independent of academic disciplines. Assume that the first method of classification has r levels and that the second method has c levels. We will let O_{ij} be the observed frequency for level i of the first classification method and level j of the second classification method. The data would, in general, appear as shown in Table 9.2. Such a table is usually called an $r \times c$ **contingency table**.

We are interested in testing the hypothesis that the row-and-column methods of classification are independent. If we reject this hypothesis, we conclude some interaction exists between the two criteria of classification. The exact test procedures are difficult to obtain, but an approximate test statistic is valid for large n. Let p_{ij} be the probability that a randomly selected element falls in the ijth cell given that the two classifications are independent. Then $p_{ij} = u_i v_j$, where u_i is the probability that a randomly selected element falls in row class i and v_j is the probability that a randomly selected element falls in column class j. Now by assuming independence, the estimators of u_i and v_j are

$$\hat{u}_i = \frac{1}{n}\sum_{j=1}^{c} O_{ij} \qquad \hat{v}_j = \frac{1}{n}\sum_{i=1}^{r} O_{ij} \tag{9.48}$$

Therefore, the expected frequency of each cell is

$$E_{ij} = n\hat{u}_i\hat{v}_j = \frac{1}{n}\sum_{j=1}^{c} O_{ij}\sum_{i=1}^{r} O_{ij} \tag{9.49}$$

TABLE 9.2 An $r \times c$ Contingency Table

		Columns			
		1	2	⋯	c
	1	O_{11}	O_{12}	⋯	O_{1c}
Rows	2	O_{21}	O_{22}	⋯	O_{2c}
	⋮	⋮	⋮	⋮	⋮
	r	O_{r1}	O_{r2}	⋯	O_{rc}

Then, for large n, the statistic

$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{9.50}$$

has an approximate chi-square distribution with $(r-1)(c-1)$ degrees of freedom if the null hypothesis is true. We should reject the null hypothesis if the value of the test statistic χ_0^2 is too large. The P-value would be calculated as the probability beyond χ_0^2 on the $\chi^2_{(r-1)(c-1)}$ distribution, or $P = P(\chi^2_{(r-1)(c-1)} > \chi_0^2)$. For a fixed-level test, we would reject the hypothesis of independence if the observed value of the test statistic χ_0^2 exceeded $\chi^2_{\alpha,(r-1)(c-1)}$.

EXAMPLE 9.14 | Health Insurance Plan Preference

A company has to choose among three health insurance plans. Management wishes to know whether the preference for plans is independent of job classification and wants to use $\alpha = 0.05$. The opinions of a random sample of 500 employees are shown in Table 9.3.

TABLE 9.3 Observed Data for Example 9.14

	Health Insurance Plan			
Job Classification	1	2	3	Totals
Salaried workers	160	140	40	340
Hourly workers	40	60	60	160
Totals	200	200	100	500

To find the expected frequencies, we must first compute $\hat{u}_1 = (340/500) = 0.68$, $\hat{u}_2 = (160/500) = 0.32$, $\hat{v}_1 = (200/500) = 0.40$, $\hat{v}_2 = (200/500) = 0.40$, and $\hat{v}_3 = (100/500) = 0.20$. The expected frequencies may now be computed from Equation 9.49. For example, the expected number of salaried workers favoring health insurance plan 1 is

$$E_{11} = n\hat{u}_1\hat{v}_1 = 500(0.68(0.40)) = 136$$

The expected frequencies are shown in Table 9.4.

TABLE 9.4 Expected Frequencies for Example 9.14

	Health Insurance Plan			
Job Classification	1	2	3	Totals
Salaried workers	136	136	68	340
Hourly workers	64	64	32	160
Totals	200	200	100	500

The seven-step hypothesis-testing procedure may now be applied to this problem.

1. **Parameter of interest:** The variable of interest is employee preference among health insurance plans.
2. **Null hypothesis:** H_0: Preference is independent of salaried versus hourly job classification.
3. **Alternative hypothesis:** H_1: Preference is not independent of salaried versus hourly job classification.
4. **Test statistic:** The test statistic is

$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

5. **Reject H_0 if:** We will use a fixed-significance level test with $\alpha = 0.05$. Therefore, because $r = 2$ and $c = 3$, the degrees of freedom for chi-square are $(r-1)(c-1) = (1)(2) = 2$, and we would reject H_0 if $\chi_0^2 > \chi_{0.05,2}^2 = 5.99$.
6. **Computations:**

$$\begin{aligned}\chi_0^2 &= \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\ &= \frac{(160-136)^2}{136} + \frac{(140-136)^2}{136} + \frac{(40-68)^2}{68} \\ &\quad + \frac{(40-64)^2}{64} + \frac{(60-64)^2}{64} + \frac{(60-32)^2}{32} \\ &= 49.63\end{aligned}$$

7. **Conclusions:** Because $\chi_0^2 = 49.63 > \chi_{0.05,2}^2 = 5.99$, we reject the hypothesis of independence and conclude that the preference for health insurance plans is not independent of job classification. The P-value for $\chi_0^2 = 49.63$ is $P = 1.671 \times 10^{-11}$. (This value was computed by computer software.) Further analysis would be necessary to explore the nature of the association between these factors. It might be helpful to examine the table of observed minus expected frequencies.

Content is copied from book: Applied Statistics and Probability for Engineers