Lecture 18
Chi-square Goodness of Fit Tests

Chao Song

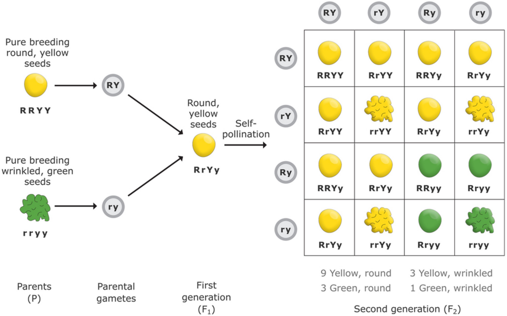
College of Ecology

Lanzhou University

November 25, 2024

A motivating example: Mendel's principle of inheritance

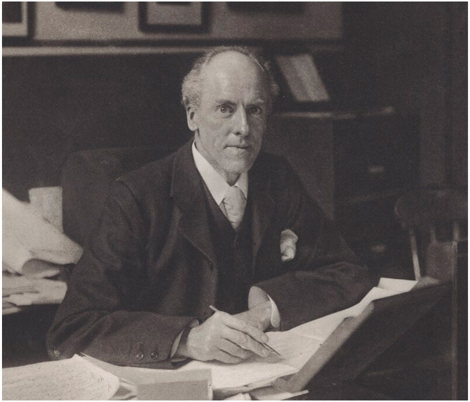
Mendel cross-bred plants with 2 or more traits and found that each trait was inherited independently of the other and produced its own 3:1 ratio. For example, a plant with round, yellow seeds crossed with a plant with wrinkled green seeds gives a ratio of 9:3:3:1. How do we test this theory with data?



Gregor Mendel (1822–1884) and the pea crossing experiment.

Goodness of fit test

Mendel's example represents a common problem we face in data analysis. We want to test how well the data fits a hypothesized distribution. British statistician Karl Pearson first developed the **goodness of fit tests**.



Karl Pearson (1857–1936)

Goodness of fit test

To introduce the goodness of fit test, we first start with a binomial case. Let Y_1 be the number of success in a binomial distribution with n trials and success probability p_1. Based on central limit theorem

$$Z = \frac{Y_1 - np_1}{\sqrt{np_1(1 - p_1)}}$$

is approximately $N(0, 1)$, particularly when $np_1 > 5$ and $n(1 - p_1) > 5$.

Given that the square of a standard normal distribution is chi-square distribution with 1 degree of freedom, $\chi^2(1)$, we have

$$Q = Z^2 = \frac{(Y_1 - np_1)^2}{np_1(1 - p_1)}$$

following a $\chi^2(1)$.

Goodness of fit test

We further rearrange the formula for Q_1

$$Q = \frac{(Y_1 - np_1)^2}{np_1(1 - p_1)} = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_1 - np_1)^2}{n(1 - p_1)}$$

Note here that

$$(Y_1 - np_1)^2 = (n - Y_1 - n + np_1)^2 = (Y_2 - np_2)^2$$

where $Y_2 = n - Y_1$ is the number of failures and $p_2 = 1 - p_1$ is the probability of failure. We thus rewrite Q_1 as

$$\begin{aligned} Q &= \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2} \\ &= \sum_{i=1}^{2} \frac{(Y_i - np_i)^2}{np_i} \sim \chi^2(1) \end{aligned}$$

Goodness of fit test

Let's examine the statistic Q carefully, we notice

- Y_i is the observed number of occurrence in a category;
- np_i is the expected number of occurrence in a category
- Q statistic thus measures the "closeness" of the observe numbers to the corresponding expected numbers. Large value of Q_1 indicates deviation of observation from the expectation.

Pearson generalized the case of 2 categories to k categories and constructed the Q statistic for multiple categories as

$$Q = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i} \sim \chi^2(k-1)$$

This test is referred to as the **chi-square goodness of fit test**.

Chi-square goodness of fit test

Let an experiment have k mutually exclusive and exhaustive outcomes. We would like to test whether probability in each category p_i is equal to a known number p_{i0}. We shall test the hypothesis

$$H_0: \; p_i = p_{i0}, \quad i = 1, 2, \ldots, k.$$

using the test statistic

$$Q = \sum_{i=1}^{k} \frac{(Y_i - np_{i0})^2}{np_{i0}}$$

We reject the null hypothesis if Q is large. Since $Q \sim \chi^2(k-1)$, we reject H_0 if $Q \geqslant \chi^2_\alpha(k-1)$. Alternatively, we can calculate the upper tail probability as the p-value and compare it to α.

Chi-square goodness of fit test

Data forensics: An ecologist measured plant height to the tenth of a centimeter, e.g., 10.3 cm. Plants should not have a preferred number in the first decimal place in their height. We thus expect equal chance of occurrence for all numbers. Suppose that we observed the following data. Are there any evidence that the data deviate from the expected equal occurrence?

Number	0	1	2	3	4	5	6	7	8	9
Frequency	11	10	14	11	14	6	5	6	9	14

The null hypothesis is H_0: $p_i = 0.1$, $ii = 0, 1, \ldots, 9$. The test statistic is

$$Q = \sum_{i=0}^{9} \frac{(Y_i - np_i)^2}{np_i} = 10.8$$

SInce $Q \sim \chi^2(9)$, $p = P(Q \geqslant 10.8) = 0.29$. At $\alpha = 0.05$, we do not observe significant deviation from H_0.

Chi-square goodness of fit test

The hypothesis we tested so far have been simple ones, i.e., completely specified cell probability. This is not always the case and it frequently happens that $p_{10}, p_{20}, \ldots, p_{k0}$ are functions of unknown parameters.

One way out of this difficulty is to estimate p_{i0} from the data and then carry out the computations with the use of this estimate. Typically, a maximum likelihood estimate is satisfactory. However, the Q_{k-1} statistic now follows $\chi^2(k-1-d)$, where d is the number of parameters estimated from the data.

Chi-square goodness of fit

Example: Let X denote the number of α particles emitted by barium–133 in one tenth of a second. Below listed are 50 observations of number of particles emitted in one tenth of a second. The experimenter is interested in determining whether X has a Poisson distribution.

7	4	3	6	4	4	5	3	5	3
5	5	3	2	5	4	3	3	7	6
6	5	3	11	9	6	7	4	5	4
7	3	2	8	6	7	4	1	9	8
4	8	9	3	9	7	7	9	3	10

Chi-square goodness of fit test

To test H_0: X is Poisson, we first need to estimate the mean of the distribution. Recall that the sample mean of a Poisson distribution is the maximum likelihood estimate of λ. Given $\overline{X} = 5.4$, we can calculate the probability of observing any number of particles using $\hat{\lambda} = 5.4$

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

The probability of each category is

Number	0	1	2	3	4	5
Probability	0.0045	0.024	0.066	0.119	0.160	0.173
Number	6	7	8	9	10	11
Frequency	0.156	0.120	0.081	0.049	0.026	0.013

Chi-square goodness of fit test

Recall that the theoretical foundation of the chi-square goodness of fit test is the normal approximation to multinomial distribution. The approximation works when the expected number of occurrence in any category is larger than 5. If the cell probability is too small, we usually combine categories.

Number	0–3	4	5	6	7	8–11
Probability	0.213	0.160	0.173	0.156	0.120	0.178
Expected	10.65	8.00	8.65	7.80	6.00	8.90
Observed:	13	9	6	5	7	10

The test statitstics

$$Q_5 = \frac{(13 - 10.65)^2}{10.65} + \cdots + \frac{(10 - 8.9)^2}{8.9} = 2.763$$

has $\chi^2(4)$ because we have 6 categories with 1 estimated parameter. $P(Q_5 \geqslant 2.763) = 0.598$. Thus H_0 is not rejected at $\alpha = 0.05$.

Chi-square goodness of fit test

Let us now consider the problem of testing a model for the distribution of a random variable W of the continuous type. In order to use the chi-square statistic, we must partition the set of possible values of W into k sets. One way this can be done is as follows:

- Partition the interval [0, 1] into k sets with points $b_1, b_2, \ldots, b_{k-1}$;
- Let $a_i = F^{-1}(b_i)$, where $F(x)$ is the CDF of the hypothesized distribution. Count number of values in each interval $(-\infty, a_1]$, $(a_2, a_3]$ ... (a_k, ∞); This is the observed frequency.
- The expected frequency in each category is $n(a_i - a_{i-1})$;
- Use chi-square goodness of fit test to test the null hypothesis.

Test for homogeneity

Suppose that each of two independent experiments can end in one of the k mutually exclusive and exhaustive categories. We observe data as follows

Category	1	2	3	4	5
Experiment 1	Y_{11}	Y_{21}	Y_{31}	Y_{41}	Y_{51}
Experiment 2	Y_{12}	Y_{22}	Y_{32}	Y_{42}	Y_{52}

Let p_{i1} be the probability of each categories in experiment 1 and p_{i2} be the probability of each category in experiment 2. We are interested in **testing for homogeneity** of category probabilities, i.e.,

$$H_0: \ p_{i1} = p_{i2}, \quad i = 1, 2, \ldots, k.$$

Test for homogeneity

From the chi-square goodness of fit test, we know for each experiment,

$$Q = \sum_{i=1}^{k} \frac{(Y_{ij} - n_j p_{ij})^2}{n_j p_{ij}} \sim \chi^2(k-1), \quad j = 1, 2.$$

Because the two experiment are independent and the sum of independent chi-square distributions is still a chi-square distribution, we have

$$Q = \sum_{j=1}^{2} \sum_{i=1}^{k} \frac{(Y_{ij} - n_j p_{ij})^2}{n_j p_{ij}} \sim \chi^2(2k-2)$$

Usually, p_{ij} are unknown.Under H_0: $p_{i1} = p_{i2}$, the maximum likelihood estimate for p_{ij} is $\hat{p}_{ij} = (Y_{i1} + Y_{i2})/(n_1 + n_2)$. We need to estimate $k-1$ parameters. Thus the test statistic

$$Q = \sum_{j=1}^{2} \sum_{i=1}^{k} \frac{(Y_{ij} - n_j \hat{p}_{ij})^2}{n_j \hat{p}_{ij}} \sim \chi^2(k-1)$$

Test for homogeneity

Example: To compare two methods of instructions, we applied each methods to 50 randomly selected students. At the end of the instruction, students took a test and got a grade. The data were recorded as follows:

Grade	A	B	C	D	F	Totals
Method I	8	13	16	10	3	50
Method II	4	9	14	16	7	50

We want to test if the two methods of instruction are equally effective.

Test for homogeneity

Under H_0 that the category probability is the same with the two methods, the estimated probabilities are

$$\frac{8+4}{50+50} = 0.12, 0.22, 0.30, 0.26, 0.10$$

and the corresponding expected frequency are calculated as $n_1 p_{i1}$ and $n_2 p_{i2}$.

The test statistic is then

$$Q = \sum_{j=1}^{2} \sum_{i=1}^{k} \frac{(Y_{ij} - n_j \hat{p}_{ij})^2}{n_j \hat{p}_{ij}} = 5.18$$

The test statistic follows $\chi^2(4)$ and thus $p(Q \geqslant 5.18) = 0.27$. Thus, we do not have evidence that the two methods differ in their effectiveness at $\alpha = 0.05$.

Contingency table

The test for homogeneity of 2 distributions can be applied to more than 2 distributions. Moreover, homogeneity of probabilities can also be thought of as independence of category probabilities and experiments. This leads us to a more general use of the chi-square test statistic, **contingency table**.

Suppose that a random experiment results in an outcome that can be classified by two different attributes. Assume that the first attribute is assigned to one of the k events, $A_1, A_2, \ldots, A_k$, and the second attributes falls into one of the h events, $B_1, B_2, \ldots, B_h$. Let $p_{ij} = P(A_i \cap B_j)$. The random experiment is repeated n independent times and Y_{ij} denotes the frequency of the event $A_i \cap B_j$. Commonly, we wish to test the hypothesis of the independence of the A and B attributes.

$$H_0 : \; P(A_i \cap B_j) = P(A_i)P(B_j)$$

Contingency table

Let $p_{i\cdot} = P(A_i)$ and $p_{\cdot j} = P(B_j)$, the hypothesis of independence can be formulated as

$$H_0: \; p_{ij} = p_{i\cdot} p_{\cdot j}, \quad i = 1, 2, \ldots, k, \; j = 1, 2, \ldots, h.$$

In practice, $p_{i\cdot}$ and $p_{\cdot j}$ are usually unknown and are estimated as

$$\hat{p}_{i\cdot} = \sum_{j=1}^{h} \frac{Y_{ij}}{n}, \quad \hat{p}_{\cdot j} = \sum_{i=1}^{k} \frac{Y_{ij}}{n}$$

Given that there are kh categories classified by the two attributes and we estimated $k - 1 + h - 1 = k + h - 2$ parameters, the test statistic

$$Q = \sum_{i=1}^{k} \sum_{j=1}^{h} \frac{(Y_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}} = \sum_{i=1}^{k} \sum_{j=1}^{h} \frac{(Y_{ij} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}}$$

follow a chi-square distribution with $kh - 1 - (k + h - 2) = (k - 1)(h - 1)$ degrees of freedom.

Contingency table

Example: A study was conducted to determine the media credibility for reporting news. Those surveyed were asked to give their education level and the most credible medium. Test whether media credibility differ across genders.

Gender	Newspaper	Television	Radio	Totals
Male	92	108	19	219
Female	97	81	32	210
Totals	189	189	51	429

Contingency table

We first calculate the marginal probabilities. The probability of male and female in the survey is $219/429 = 0.51$ and $210/429 = 0.49$. The probability of each medium type is $189/429$ =0.44,189/429=0.44, and $51/429 = 0.12$. The test statistic is

$$Q = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(Y_{ij} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}} = 7.12$$

The test statistic has $\chi^2(2)$ and thus $p = P(Q \geqslant 7.12) = 0.028$. Thus, at $\alpha = 0.05$, we reject the null hypothesis and conclude that credibility of medium differ across gender.