Dr Isabella Gollini

Probability and Statistics

12 Chi-Square Tests of Goodness of Fit

12.1 Example — throwing a die

Suppose that we throw a die repeatedly to see if it appears to be fair, i.e., in the long run all six faces occur equally often. Table 1 shows the frequency distribution of the score obtained at each throw from a sequence of 120 throws.

score	1	2	3	4	5	6	total
observed frequency	16	14	18	28	24	20	120

Table 1: Frequency distribution of scores from 120 throws of a die

If the die is unbiased, so that each face is equally likely to occur, then on average we would expect the frequency of each score to be 20. On the other hand, because of the sampling variation, we would not expect to see a frequency of exactly 20 for each score. But the question that arises is whether the discrepancy between the observed frequencies and expected frequencies is large enough to suggest that the die is biased.

In terms of hypothesis testing, we are testing the null hypothesis that the die is unbiased against the alternative hypothesis that it is biased. We assume that the results of the throws are independent of each other, so that we have a random sample of 120 throws of the die. It turns out that in such circumstances we may use the chi-square test statistic

$$X^2 = \sum_{r=1}^{6} \frac{(O_r - E_r)^2}{E_r} \ ,$$

where O_r represents the observed frequency of the score r and E_r represents the expected frequency of the score r under the null hypothesis. In this case $E_r = 20$ $(1 \le r \le 6)$. From the formula, we can see that the test statistic is an overall measure of the discrepancies between the observed and expected frequencies, so that the larger the value of X^2 the stronger appears to be the evidence against the null hypothesis.

The calculation of the test statistic is laid out in Table 2, from which the value $X^2 = 6.8$ is obtained.

score, r	1	2	3	4	5	6	total
observed frequency, O_r	16	14	18	28	24	20	120
expected frequency, E_r	20	20	20	20	20	20	120
$(O_r - E_r)^2/E_r$	0.8	1.8	0.2	3.2	0.8	0	6.8

Table 2: Calculation of the chi-square statistic

Under the null hypothesis, the test statistic X^2 has approximately the χ^2_5 distribution, where the degrees of freedom, 5, are one less than the number of "cells" in the table of observed frequencies, i.e., one less than the number of possible outcomes of each throw. As we shall only reject the null hypothesis if the test statistic is large enough, we use a one-tail test. Using Table 7 of the χ^2 distribution function with 5 degrees of freedom and some interpolation, we find that the p-value corresponding to our calculated value 6.8 of the test statistic is given by

$$p = 1 - F(6.8) = 1 - 0.764 = 0.236 \; .$$

Since $p > 0.2$, this is not significant even at the 20% significance level, and hence not at the 5% level, so our data provide no strong evidence to reject the null hypothesis that the die is unbiased.

Alternatively, from Table 8 we see that $\chi^2_5(20) = 7.289$. As our calculated $X^2 = 6.8 < \chi^2_5(20)$, we do not reject the null hypothesis at the 20% significance level.

In R, you can use the function `chisq.test` for carrying out the chi-square test. Type `?chisq.test` if you want to access help files. The first argument you have to specify is a vector containing the observed values. It is also possible to specify the argument `p` that is a vector containing the probabilities under the null hypothesis. If nothing is specified for `p` R assumes equal probability for all the outcomes.

In the present case we construct a vector called `count` containing the observed counts:

```
count <- c(16, 14, 18, 28, 24, 20)
names(count) <- 1:6
count

##  1  2  3  4  5  6
## 16 14 18 28 24 20
```

and we perform the test by using the function `chisq.test`. We don't have to specify the value of `p` as we are testing the null hypothesis of equal proportions.

```
test <- chisq.test(count)
test

##
##  Chi-squared test for given probabilities
##
## data:  count
## X-squared = 6.8, df = 5, p-value = 0.2359
```

The output confirms the calculations above.

12.2 A more general scenario

12.2.1 Categorical data

Up to now we have considered *numerical data*, where the outcome of an experiment is associated with a number, and we have distinguished between discrete and continuous data. But now we suppose that with the outcome of a statistical experiment (or "trial") there is associated not a numerical value but a category. For example, if the observation is of eye colour in a human population, any individual might be classified as having either "blue", "brown" or "green" eye colour.

In the example of Section 12.1, the faces of the die might have been painted different colours, or they might have had different symbols painted on them, instead of the numbers 1 to 6. The analysis of Section 12.1 did not depend upon the fact that the six faces were associated with the numbers 1, 2, ..., 6, respectively, only on the fact that the faces represented the six possible outcomes of each throw.

Data in which each observation falls into one of k, say, categories, is referred to as *categorical data.* If a sample of such data is taken then the results may be summarized by the counts of the numbers of occurrences, i.e., the observed frequencies, of each of the k categories.

12.2.2 The multinomial distribution

Label the k categories by $r = 1, 2, \ldots, k$. In any trial let p_r denote the probability that the outcome belongs to category r, or, if we are sampling from a large population, let p_r be the proportion of individuals in the population who belong to category r. It follows that (p_r) $(r = 1, 2, \ldots, k)$ is a probability distribution, so that $\sum_{r=1}^{k} p_r = 1$.

- A special case is that of equal probabilities, or equal proportions,

  $$p_r = \frac{1}{k} \qquad (r = 1, 2, \ldots, k),$$

  as in the example of Section 12.1, where $k = 6$.

Suppose that n independent trials are carried out, or a random sample of size n is taken from a large population. Let O_r denote the number of occurrences of category r $(r = 1, 2, \ldots, k)$, i.e., the observed frequency. It follows that $\sum_{r=1}^{k} O_r = n$.

The probability of observing a given set of such values O_r is given by the *multinomial distribution*,

$$\Pr(O_1, O_2, \ldots, O_k) = \frac{n!}{O_1! O_2! \ldots O_k!} \, p_1^{O_1} p_2^{O_2} \ldots p_k^{O_k} \; .$$

- The multinomial distribution is a generalization of the binomial distribution, which corresponds to the special case $k = 2$ with categories "success" and "failure." We may, in this special case, identify O_1 with the number of successes and O_2 with the number of failures.

We shall not make detailed use of the multinomial distribution, but we shall want to have expressions for the expected frequencies E_r for each category,

$$E_r = E(O_r) \qquad (r = 1, 2, \ldots, k).$$

As for the binomial distribution, we have

$$E_r = np_r \qquad (r = 1, 2, \dots, k). \tag{1}$$

It follows that $\sum_{r=1}^{k} E_r = n \sum_{r=1}^{k} p_r = n$.

12.2.3 The chi-square statistic

Suppose that we have a sample of size n of categorical data, with k categories, and that we are testing the null hypothesis H_0 that we have a random sample with the probabilities of the different categories specified by a particular probability distribution (p_r) $(r = 1, 2, \dots, k)$. The alternative hypothesis H_1 is that we do not have a random sample from the given probability distribution (p_r).

Given the sample data, we have, or we can calculate, the observed frequencies O_r. In addition, using Equation (1), we can calculate the expected frequencies E_r under H_0. To test H_0 we use the test statistic

$$X^2 = \sum_{r=1}^{k} \frac{(O_r - E_r)^2}{E_r} \ , \tag{2}$$

which, under H_0, has approximately the χ^2_{k-1} distribution. We shall reject H_0 if the test statistic is large enough, so we carry out a one-tail test and reject H_0 at the $100\alpha\%$ significance level if $X^2 \geq \chi^2_{k-1}(100\alpha)$.

- The $k - 1$ degrees of freedom, one less than the number of categories or "cells," k, may be interpreted in the following way. There are $k - 1$ frequencies that can be entered into the table more or less arbitrarily, but the kth frequency, whether observed or expected, is then determined by the constraint that the sum of the frequencies must be n.

- In the case $k = 2$, this test procedure is equivalent to the test for a proportion of Section 11.3, with a two-sided alternative.

For the chi-square approximation to be adequate, the expected frequencies E_r should not be too small. The requirement that $E_r \geq 5$ for all r is often suggested, although if one or two of the E_r lie just below this limit then this is acceptable.

If the E_r do turn out to be too small, we may remedy the situation by amalgamating some of the categories, i.e., by amalgamating some of the cells in the table of observed frequencies.

12.3 An example from genetics

In the plant-breeding experiments reported by Gregor Mendel in 1865, the following set of data were obtained on hybrid plants of a certain variety. The plants that were raised yielded seeds of four types. In all, there were 556 seeds, and of these there were:

- 315 round and yellow,
- 101 wrinkled and yellow,
- 108 round and green,
- 32 wrinkled and green.

According to a genetic model, the seeds of the four types should be in the ratio 9:3:3:1. Are the data obtained consistent with the model?

There are four categories of seed, which we may label 1, 2, 3, 4. According to the model, the probabilities that a randomly chosen seed lies in each of the categories are 9/16, 3/16, 3/16, 1/16, respectively. We test the null hypothesis that the data obtained are a random sample with the probabilities of the categories as predicted by the genetic model. The calculations are shown in Table 3.

Type of seed, r	p_r	O_r	$E_r = 556p_r$	$(O_r - E_r)^2/E_r$
1. round and yellow	9/16	315	312.75	0.016
2. wrinkled and yellow	3/16	101	104.25	0.101
3. round and green	3/16	108	104.25	0.135
4. wrinkled and green	1/16	32	34.75	0.218
sum	1	556	556.00	0.470

Table 3: Mendel data

The calculated value of the chi-square test statistic is 0.470 with 3 degrees of freedom. Using Table 7 of the χ^2 distribution function with 3 degrees of freedom and some interpolation, we find that the corresponding p-value is given by

$$p = 1 - F(0.470) = 1 - 0.075 = 0.925 .$$

This p-value is large and nowhere near significant. There is no evidence whatsoever to reject the null hypothesis. The fit of the data to the model is so good that it has been suggested that the data were fixed, or selected to support the model.

Using `R`, the names of the four seed types have been entered in a vector named `seedtype` and the corresponding observed frequencies have been entered in a vector named `freq`. The probabilities specified by the model have been entered in decimal form in a column named `prob`.

```
seedtype <- c("round and yellow", "wrinkled and yellow",
              "round and green", "wrinkled and green")

freq <- c(315, 101, 108, 32)
names(freq) <- seedtype

prob <- c(0.5625, 0.1875, 0.1875, 0.0625)
sum(prob)

## [1] 1

cbind(freq, prob)

##                        freq   prob
## round and yellow     315 0.5625
## wrinkled and yellow  101 0.1875
## round and green      108 0.1875
## wrinkled and green    32 0.0625
```

We make use of the function `chisq.test`, and we specify the value of `p`.

```
test <- chisq.test(freq,
                   p = prob)
test

##
##  Chi-squared test for given probabilities
##
## data:  freq
## X-squared = 0.47002, df = 3, p-value = 0.9254
```

The output confirms the calculations of Table 3 and results in the same p-value, 0.925, as above.

Extra Example

A random sample of 100 student absences yielded the following data:

Day	Monday	Tuesday	Wednesday	Thursday	Friday	total
observed frequency	27	19	13	15	26	100

Test the hypothesis that an absence is equally likely to occur on any of the 5 days. (Use $\alpha = 0.05$).