Modeling of theoretical distributions to given data and the Chi squared test for goodness of fit

We are concerned here with testing a hypothesis that a given population follows a particular distribution. Up to now we have always assumed or been told that a particular variable follows a particular distribution. In this section we show how to test that assumption.

Thus we are testing a hypothesis that a variable X can be modelled as following a particular probability distribution. In order to do this, however, we need to be able to answer the question: what frequencies would we expect from X if X was distributed in the way described? We need to be able to construct a table of expected probabilities and frequencies. This involves no new theory and is best illustrated by example.

Example (1)

A discrete probability distribution

A ten-sided die is thought to be weighted in a way that makes it biased. The probability of throwing any even number is thought to be equal. For the odd numbers

> the probability of throwing a 1 is equal to the probability of throwing an even number
>
> the probability of throwing a 9 is five times that of throwing a 1
>
> the probabilities of throwing a1, 3, 5, 7 and 9 are in arithmetic progression

The die will be thrown 160 times; find the expected frequencies.

Answer

Let $P(2) = x$

Then

$P(2) = P(4) = P(6) = P(8) = P(10) = x$

Also

$P(1) = x$ and $P(9) = 5x$

and

$P(3) = 2x,\ P(5) = 3x,$ and $P(7) = 4x$

This generates the following discrete probability table:

X$_i$	1	2	3	4	5	6	7	8	9	10
P(X=x_i)	x	x	$2x$	x	$3x$	x	$4x$	x	$5x$	x

Since $\sum P(X = x_i) = 1$

Then $20x = 1$

$x = {}^1\!/_{20} = 0.05$

This will enable us to generate the probability distribution. Multiplying each probability by 20 will yield the expected frequencies.

X_i	1	2	3	4	5	6	7	8	9	10
P(X=X_i)	0.05	0.05	0.1	0.05	0.15	0.05	0.2	0.05	0.25	0.05
f_i	8	8	16	8	24	8	32	8	40	8

We will illustrate how expected frequencies can be obtained for other distributions later, but we shall first proceed to illustrate the testing of a hypothesis about a distribution by continuing this example.

Example (1) cont.

On throwing the die 160 times the following frequencies were observed.

X$_i$	1	2	3	4	5	6	7	8	9	10
f$_i$	6	11	18	10	32	5	29	8	34	7

Test the hypothesis at the 5% level that the die is biased as described.

The test will be conducted by comparing the expected frequencies that we have just calculated with the observed frequencies just stated.

We designate the observed frequencies O and the expected frequencies E. Then the test statistic is

$$\chi^2_{test} = \sum \frac{(O-E)^2}{E}$$

We square the difference between the observed and expected frequency and divide the result by the expected frequency; then we sum the whole lot.

Provided certain assumptions are fulfilled the statistic

$$\sum \frac{(O-E)^2}{E}$$

can be approximated by a probability distribution known as the χ^2-distribution (pronounced "kye squared"). Standardised values for the χ^2-distribution as functions of the critical significance levels are provided in tables.

In order to conduct a χ^2 test it is advisable to construct a contingency table. This is a table setting out the expected and observed frequencies side by side; from this the contributions of individual columns to the test statistic can be readily calculated

The contingency table for our example is:

X_i	1	2	3	4	5	6	7	8	9	10
f(observed)	6	11	18	10	32	5	29	8	34	7
f(expected)	8	8	16	8	24	8	32	8	40	8

Like other distributions (for example, the *t*-distribution) the χ^2-distribution is a function of the degrees of freedom. The degrees of freedom are the number of ways in which the contribution of $\sum (O-E)^2 / E$ can vary. The maximum degrees of freedom is therefore equal to the number of rows in the contingency table. From this number we subtract the number of constraints on the way the data has been constructed. There is always at least one constraint. This is because the last entry in a frequency table is determined by all the others – given the total frequency. In our current example, since the die is to be thrown 160 times and the sum of the first nine frequencies is153, this constrains the entry for the tenth frequency – it must be 160-153=7.

Constraints can arise in other ways. In particular, there is one other constraint that can arise in tests of goodness of fit. If the expected values are modelled around a parameter determined from the experimental data themselves this adds one more constraint to the contingency table. We will illustrate this in a further example. For the present, our current example has just one constraint, hence

degrees of freedom = v = 10-1 = 9

We are now in a position to complete the test.

x_i	1	2	3	4	5	6	7	8	9	10
f(observed)	6	11	18	10	32	5	29	8	34	7
f(expected)	8	8	16	8	24	8	32	8	40	8
$\frac{(O-E)^2}{E}$	$\frac{2^2}{8}$ $= 0.5$	$\frac{3^2}{8}$ $=1.125$	$\frac{4^2}{16}$ $=1$	$\frac{2^2}{8}$ $= 0.5$	$\frac{8^2}{24}$ $= 0.2667$	$\frac{3^2}{8}$ $=1.125$	$\frac{3^2}{32}$ $= 0.281$	0 $= 0$	$\frac{6^2}{40}$ $= 0.9$	$\frac{1^2}{8}$ $= 0.125$

$$\chi^2_{test} = 0.5+1.125+1+0.5+2.667+1.125+0.281+0+0.9+0.125$$
$$= 8.223$$

$\chi^2_{critical}$ (v=9, p=0.950) = 16.919

$\chi^2_{\text{test}} < \chi^2_{critical}$

$\therefore$ Accept H_0; reject H_1

The probability distribution for the die is as described.

Our next example illustrates the application of the test for goodness of fit and the determination of expected values for a Poisson distribution. It also illustrates the procedure for dealing with small expected frequencies.

Expected frequencies below 5 should not be used. Classes with lower frequencies should be combined to yield a class with frequency greater than 5.

Example (2)

The number of horror movies released per month in the USA is thought to follow a Poisson distribution. Releases of horror movies per month in the USA, X, were recorded by a fanatic over 100 months and she obtained the following observed frequencies.

x	0	1	2	3	4	5	6	7
f	15	25	34	16	7	2	1	0

Determine an appropriate Poisson distribution for this data and test the goodness of fit at the 5% significance level.

Solution

We will model this by $X \sim Po(\mu)$ where μ is estimated by $\bar{x}$, the mean of the distribution.

Then

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{185}{100} = 1.85$$

Hence $X \sim Po(1.85)$

H_0 : $X \sim Po(1.85)$

H_1 : X is not $Po(1.85)$

$$P(X = r) = \frac{\mu^r}{r!}e^{-\mu}$$

Expected frequencies are given by multiplying probabilities by the sample size. Hence

x_i	$P(X = x_i)$	f_i
0	$e^{-1.85} = 0.1572$	15.72
1	$1.85 \times e^{-1.85} = 0.2909$	29.09
2	$\frac{(1.85)^2}{2!} \times e^{-1.85} = 0.2691$	26.91
3	$\frac{(1.85)^3}{3!} \times e^{-1.85} = 0.1659$	16.59
4	$\frac{(1.85)^4}{4!} \times e^{-1.85} = 0.0767$	7.67

5	$\frac{(1.85)^5}{5!} \times e^{-1.85} = 0.0284$	2.84
6	$\frac{(1.85)^6}{6!} \times e^{-1.85} = 0.0088$	0.88
>7	$1 - \{P(1) + P(2) + \ldots + P(6)\}$ $= 1 - 0.9979 = 0.0030$	0.30

Expected frequencies less than 5 cannot be used, so we combined the classes for 4, 5, 6 and 7 to obtain the following contingency table, from which the contributions to the test statistic can be directly computed.

x	0	1	2	3	≥ 4
f(observed)	15	25	34	16	10
f(expected)	15.72	29.09	26.91	16.59	11.69
$\frac{(O-E)^2}{E}$	$\frac{0.72^2}{15.72}$ $= 0.0330$	$\frac{4.09^2}{29.09}$ $= 0.5750$	$\frac{7.09^2}{26.91}$ $= 1.8680$	$\frac{0.59^2}{16.59}$ $= 0.0210$	$\frac{0.69^2}{11.69}$ $= 0.0407$

$$\sum \frac{(O-E)^2}{E} = 0.0330 + 0.5750 + 1.8680 + 0.021 + 0.0407$$
$$= 2.5377$$

To calculate the critical value we need to know the degrees of freedom. The number of rows is 5.

There are two constraints:

1. Because frequencies in the last row are determined by the other frequencies.
2. Because the expected frequencies are calculated from a statistic determined from the observed frequencies.

$\therefore\ v = 5 - 2 = 3$

$\therefore\ \chi^2_{critical}(v = 3,\ p = 0.950) = 7.815$

$\therefore\ \chi^2_{test} < \chi^2_{critical}$

$\therefore$ Accept H_0, reject H_1

The distribution follows a Poisson distribution with parameter,
mean = 1.85 horror movies per month.

We now illustrate the application of the χ^2_{test} for goodness of fit to a hypothesis concerning a normal distribution.

Example (3)

The total weight of fish, in kilograms, caught by an experienced angler in a Russian lake in one day during the summer of 1964 is denoted by the random variable X. 90 anglers were sampled and the results obtained and summarised below:

x	<12	12-24	24-36	36-48	>48
Observed Frequency	10	14	30	23	13

It is thought that X is normally distributed by $N(30, 12^2)$. Calculate the expected frequencies for each of the five classes. Carry out a χ^2 goodness of fit test at the 5% level to test this hypothesis.

Answer

Let $X\sim N(30, 12^2)$

Total number of observations = 90

To determine the z value corresponding to the class boundaries:

For $x = 12$

$$z = \frac{30 - 12}{12} = 1.5$$

$P(X < 12) = P(X > 48) = 0.0668$

Expected frequency $= 0.0668 \times 90 = 6.012$

For $x = 24$

$$z = \frac{30 - 24}{12} = 0.5$$

$P(36 < X < 48) = P(12 < X < 24) = 0.3085 - 0.0668 = 0.2417$

Expected frequency $= 0.2417 \times 90 = 21.753$

$P(24 < X < 36) = (0.5 - 0.3085) \times 2 = 0.3830$

Expected frequency $= 0.3830 \times 90 = 34.47$

Determining the contingency table and contributions to the test statistic together:-

x	<12	12-24	24-36	36-48	>48
f(observed)	10	14	30	23	13
f(expected)	6.012	21.753	34.47	21.753	6.012
$\frac{(O-E)^2}{E}$	$\frac{(10-6.012)^2}{6.012}$ $= 2.645$	$\frac{(14-21.753)^2}{21.753}$ $= 2.763$	$\frac{(30-34.47)^2}{34.47}$ $= 0.580$	$\frac{(23-21.753)^2}{21.753}$ $= 0.072$	$\frac{(13-6.012)^2}{6.012}$ $= 8.122$

$$\chi^2_{test} = 2.645 + 2.763 + 0.580 + 0.072 + 8.122$$
$$= 14.182$$

There is one constraint - one cell is determined by the others; hence $v = 5 - 1 = 4$

$$\chi^2_{critical}\ (v = 4,\ p = 0.950) = 9.488$$

$$\chi^2_{test} > \chi^2_{critical}$$

$$\therefore \text{ reject } \mathrm{H}_0,\ \text{accept } \mathrm{H}_1$$

Fishing in Russian lakes is not normal!