MA26620: Practical 8

Practical 8: Discrete Inference – Poisson and Chi-square

Today's practical is mainly composed of pen-and-paper (and stats tables) exercises which are not a million miles away from the kinds of questions that may appear in the final exam for this module. They use the methods we've met in the lectures, in particular calculating confidence intervals and performing hypothesis tests for quantities relating to discrete distributions. The notes on the module webpages (https://stats.vellender.com) should be useful, as of course should your lecture notes.

1 Chi-square test

In this week's lecture, we looked at the chi-square test. Let's recap...

Recall the setup in the example we saw: 50 digits (0,1,2,3,...,8,9) were supposedly generated randomly. We wanted to test whether each of the $k = 10$ possibilities were equally likely (our null hypothesis will be that they are).

Are the differences between the observed numbers and expected numbers large enough to be incompatible with the theory? To answer this, we calculate the *Pearson chi-squared statistic*:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

Doing this with the data from the lecture (the recording's on the webpages), we obtained $\chi^2 = 10.8$. We then wanted to assess whether 10.8 is large enough to be plausible if all digits are truly equally likely. To do this, we used the following result:

For a large sample size and assuming the null hypothesis is true, then χ^2 has an approximate chi-squared distribution with $k - 1 - p$ degrees of freedom, where p is the number of parameters that have been estimated from the data. That is:

$$\chi^2 \sim \chi^2_{(k-1-p)}.$$

In our example, $k = 10$ (there were 10 different digits). The value of p is zero in our example, because in order to arrive at our expected numbers of observations for each digit, we didn't need to use our observed number to estimate any parameters. Consequently, $\chi^2 \sim \chi^2_{(9)}$, and from $\chi^2_{(9)}$ tables, we see that the probability of this distribution exceeding 10.8 is over 0.3. Thus this value of the observed chi-squared statistic is not unexpectedly large (not significant at the 10% level), so there's very little evidence indeed to suggest that each digit is not equally likely.

2 Exercises

1. The Highways Department schedules a road for resurfacing if the rate of potholes exceeds 6 per 100m. Assuming that potholes occur randomly on roads at a constant expected rate, would they resurface a road which has 40 potholes in 500m? Formulate this problem as a test of hypotheses involving the Poisson distribution i.e. state

   (i) what variable has the Poisson distribution, defining its parameter;

   (ii) the two hypotheses and

   (iii) the distribution when H_0 is true. Use tables to evaluate the p-value and state your conclusion.

2. Police records for the idyllic country town of Much-Thuggery-on-the-Binge, show that over a 24 week period, 32 serious assaults took place outside Mad Mick's Boozerama. Use a suitable Normal approximation to calculate a 95% confidence interval for λ, the weekly rate of assaults. State clearly what distribution you are assuming for the observation.

3. A group of rats, one-by-one, proceed down a ramp to one of three doors. Test the hypothesis that the rats have no preference concerning the choice of a door and therefore that

   $$H_0 : p_1 = p_2 = p_3 = 1/3,$$

   where p_i is the probability that a rat will choose door i for $i = 1, 2$, or 3, given that in $n = 90$ trials, the observed frequencies were $n_1 = 23$, $n_2 = 36$, and $n_3 = 31$.

4. (Slightly trickier) The number of accidents Y per week at a junction was checked for $n = 75$ weeks, with the results as shown below:

y	Frequency
0	48
1	18
2	9
3	0

   Test the hypothesis at the 95% level that the random variable Y has a Poisson distribution, assuming the observations to be independent.

   Solution *(fill in the gaps)*: The null hypothesis H_0 states that Y has the _____ distribution, which has probability mass function:

   $$P(Y = y) = \_\_\_\_\_, \qquad y = 0, 1, 2, \ldots$$

   Because λ is unknown, we must estimate it as $\hat{\lambda} =$ _____.

   We have, for the given data, three cells with non-zero observations – the cells defined by $Y = 0$, $Y = 1$, $Y \geq 2$ (together, these categories cover all possible events so their probabilities sum to 1). Under H_0, the probabilities for these cells (in terms of λ) are:

   $$p_0 = P(Y = 0) = \_\_\_\_\_; \quad p_1 = P(Y = 1) = \_\_\_\_\_; \quad p_2 = P(Y \geq 2) = \_\_\_\_\_.$$

   These probabilities are estimated by replacing λ with $\hat{\lambda}$, which gives:

   $$\hat{p}_0 = \_\_\_\_\_; \qquad \hat{p}_1 = \_\_\_\_\_; \qquad \hat{p}_2 = \_\_\_\_\_.$$

   The expected cell frequencies are therefore:

   $$\widehat{\mathbb{E}[n_0]} = n\hat{p}_0 = \_\_\_\_\_; \qquad \widehat{\mathbb{E}[n_1]} = n\hat{p}_1 = \_\_\_\_\_; \qquad \widehat{\mathbb{E}[n_2]} = n\hat{p}_2 = \_\_\_\_\_.$$

   (Note that if we had instead defined $\hat{p}_3$ to be $P(Y = 3)$, then $\widehat{\mathbb{E}[n_3]} = n\hat{p}_3 =$ _____, which is less than 5. This justifies combining the events as we have done.)

   Thus, the χ^2 statistic is given by

   $$\chi^2 = \sum_{i=0}^{2} \frac{[n_i - \widehat{\mathbb{E}[n_i]}]^2}{\widehat{\mathbb{E}[n_i]}},$$

   which has an approximate χ^2 distribution with _____ degree(s) of freedom. On computing the χ^2 statistic, we find $\chi^2 =$ _____ . Because _____ (from the tables), we *[do/do not]* reject H_0. The data *[does/does not]* not present sufficient evidence to contradict our hypothesis that Y possesses a Poisson distribution.