Mathematical Statistics 2B

Practical 2: Fitting Distributions

Department of Statistics - University of Johannesburg

1. Introduction

In practical statistics we observe data through proper sampling methodologies in order to estimate the behaviour of the population based on the sampled information. This means that we need to know something about the distribution of our variables of interest. We know that observed data consist of random variability, and therefore the empirical distribution will never conform to a perfect theoretical distribution. But if we can find a theoretical distribution that adequately describes the empirical distribution, we can make use of the properties of the theoretical distribution to calculate probabilities and do inference. To do this, we must be able to fit a theoretical distribution to the empirical data and assess the fit. There are several methods used for assessing distribution fit: formal and informal methods. We should always use all available information to determine whether a distribution is a good fit and not rely on a single method.

Formal methods

Formal methods use some form of calculated value to assess distribution fit. Some are relative fit measures and others are absolute goodness-of-fit. Relative fit measures only tell whether one distribution is better than another by ranking several different fits, but do not indicate how well it fits. Information criterion can be used to rank multiple distributions in a relative way, for example Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC). There are many goodness-of-fit measures, but none is without limitations. This means no goodness-of-fit statistic can be used with absolute certainty. Chi-squared goodness-of-fit is used for discrete distributions. The Kolmogorov-Smirnov, Anderson-Darling and Chi-Squared goodness-of-fit measures are used for continuous distributions.

Informal methods

Informal methods make use of graphical representation and are more subjective than formal methods. As such the distribution fit is assessed through qualitative methods. The typical graphs include a histogram of the data with a curve of the fitted distribution, the empirical cumulative distribution of the data along with the cumulative distribution of the fitted distribution, the quantiles of the distribution vs. the quantiles of the data (Q-Q plot), and the probability of the distribution vs. the empirical probability distribution of the data (P-P plot).

2. Chi-Squared Goodness-Of-Fit

The Chi-squared goodness-of-fit hypothesis test is used to test if the distribution of a categorical variable differs significantly from what is expected. Depending on the nature of the problem statement, what is expected could be either an equal distribution across the levels of the variable (i.e., equal proportions) or an unequal distribution (i.e., the proportions are equal to some known distribution). The chi-squared goodness-of-fit test depends on an adequate sample size for the approximations to be valid, and a minimum requirement is that the expected frequency of each of the k levels of the variable must be at least 5.

The null and alternative hypotheses are:

H_0: Distribution follows a known pattern

H_1: At least 1 proportion is different from the known pattern

The level of significance (α) and degrees of freedom are used to find the critical value from the χ^2 table, i.e., the rejection region. If the values of the p parameters of interest are assumed or known in advance, there are $k - 1$ degrees of freedom. If the values of the p parameters of interest must first be estimated from the data, there are $k - p - 1$ degrees of freedom.

The test statistic is:

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

where

$O_i = n_i =$ observed frequencies for category I

$E_i = n\pi_i =$ expected frequencies for category i

$\pi_i =$ probability associated with category i of the theoretical distribution, i.e., under the null hypothesis

Exercise 1

Suppose that we observed the following number of children in n = 100 families:

Number of children	0	1	2	3	4+
Frequency	19	26	29	13	13

Are these data consistent with a Poisson distribution? The raw data are given in the *child.csv* data file.

The procedure in R to fit a discrete probability distribution, in this case the Poisson distribution, using the Chi-squared goodness-of-fit is as follows:

1) If the parameter is not assumed to be equal to a specific value, estimate the parameter (in this case λ) using the average of the variable. The maximum likelihood estimate of λ is just the sample mean (you will learn more about this in the theory lectures).
2) Calculate the Poisson probabilities for each of the values of the variable using the *dpois()* function. Adapt this function for the required discrete distribution. Note: for the chi-squared test all probabilities must add up to 1, so the probability of the last category must be calculated using the complement rule.
3) Perform a chi-squared test using the *chisq.test()* function.

Exercise 2

A doctor at a diabetes clinic wants to know how well newly diagnosed insulin-dependent patients manage the difficult task of regulating their blood sugar levels. The doctor measured the blood sugar levels (BSL) of n = 121 patients when they visited the clinic for their first examination after their initial diagnosis. The first question the doctor wants to answer is whether the BSL levels are normally distributed, as this will determine if he should use parametric or non-parametric statistics tests (we will do more assumption checking in the next practical). The data are given in the *BSL.csv* data file.

The procedure in R to fit a continuous probability distribution, in this case the normal distribution, using the Chi-squared goodness-of-fit is as follows:

1) If the parameters are not assumed to be equal to specific values, estimate the parameters (in this case μ and σ^2) using the sample mean $\bar{x}$ and sample variance s^2. These are the maximum likelihood estimates of the parameters and we will discuss this in the theory lectures.
2) Standardise each observed data value to create z-scores.

$$z = \frac{x - \mu}{\sigma}$$

3) Group the z-scores into classes, called bins, all of equal class width. There must be sufficient sample size in each bin for the test to be valid. For this example, create bins for the majority of the standard normal distribution in increments of 0.5, using the *cut()* function. Because the expected frequencies in the tail areas are small, combine cells to conform to the requirement that the expected frequencies must be at least 5. Also, combine the observed frequencies cells at the extremes if these cells have a frequency of zero.
4) For each bin/class interval (a, b), calculate the $P\left(a<Z<b\right)=P\left(Z<b\right)-P\left(Z<a\right)$ using the *pnorm()* function.
5) Perform a chi-squared test using the *chisq.test()* function.

3. Empirical Cumulative Distribution Function

The empirical cumulative distribution function (ECDF) is a step function that displays all the observed data values (x-axis) of a continuous random variable from lowest to highest against their percentiles (y-axis). The function assigns the probability 1/n to each value in the sample and computes the proportion of the sample at or below that value.

Exercise 3

1) Plot the ecdf of the standardized BSL data using the *ecdf()* function.
2) Add the curve of the standard normal CDF to the plot using the function:

```
curve(pnorm, from = −2, to = 2,add = T)
```

3) Comment on the results.

4. Accessing Built-In Datasets in R

I have given you a few datasets for this practical session and the previous session, but you can access all the built-in datasets in R using the following code in the command line:

```
library(help = "datasets")
```

This will show a list of the dataset names and a short description. You can then access a specific dataset, say the Cars dataset, in the help function, which will give more detail about the dataset, such as the number of observations, the variables, etc.

You can use these datasets to fit distributions. You can also simulate data from some distribution and then test the fit.