Estimating Distribution Parameters using Coarse Data for Chi-Squared Goodness-of-Fit Test

Sahand Rabbani

Problem Statement

Given a coarse observation of data, we wish to test the hypothesis that said data represents samples of a certain distribution. The coarse data is incomplete in that we only know whether the observations fall within a range or bin, without knowing their actual value. Formally, we are given

n_i	Frequency of samples in bin i for $i \in \{1, 2, \ldots, k\}$
x_i^l	Lower bound on bin i
x_i^u	Upper bound on bin i
k	Number of bins

We note that the lowest bin has $x_1^l = -\infty$ and the highest bin has $x_k^u = \infty$. We have that n_i observations fall in the interval $s_i = [x_i^l, x_i^u)$ (except for the lowest bin, where the lower bound is open since it is $-\infty$). Also, we make the assumption that the bins do not overlap and that the union of all bins is the entire real line:

$$s_i \cap s_j = \emptyset \quad i \neq j \qquad\qquad \bigcup_{i=1}^{k} s_i = \mathbb{R}$$

We wish to test the hypothesis that said data represents samples of a m-parameter distribution with a density function

$$f(x; \theta_1, \theta_2, \ldots, \theta_m)$$

where θ_l for $l \in \{1, 2, \ldots, m\}$ are parameters of the distribution. We would like to use the chi-squared goodness-of-fit test; however, we do not know the parameters θ_l and must estimate them given the coarse data. The chi-squared test requires that the parameters be determined using maximum likelihood estimation, which uses fine data. Below, I propose two methods to determine the parameters when only coarse data is available:

1. Coarse maximum likelihood estimation
2. Minimum chi-squared statistic

Coarse Maximum Likelihood Estimation

This technique appeals to the spirit of maximum likelihood estimation while negotiating the problem of coarse data. Given some distribution $f(x; \theta_1, \theta_2, \ldots, \theta_m)$, we can calculate the probability the data, adhering to this distribution, gives the observed frequency profile. Assuming independence of samples, this probability is the likelihood function L:

$$L(n_i, x_i^l, x_i^u; \theta_1, \theta_2, \ldots, \theta_m) = \prod_{i=1}^{k} \left(\int_{x_i^l}^{x_i^u} f(x; \theta_1, \theta_2, \ldots, \theta_m) dx \right)^{n_i}$$

The coarse maximum likelihood estimation method selects the parameters θ_l that maximize this likelihood function or any monotonic transformation of this function, specifically, the log-likelihood function:

$$\Lambda(n_i, x_i^l, x_i^u; \theta_1, \theta_2, \ldots, \theta_m) = \ln L(n_i, x_i^l, x_i^u; \theta_1, \theta_2, \ldots, \theta_m) = \sum_{i=1}^{k} n_i \ln\left(\int_{x_i^l}^{x_i^u} f(x; \theta_1, \theta_2, \ldots, \theta_m)dx\right)$$

Our estimates of the parameters, denoted by $\hat{\theta}_l$, are the solution to

$$\max_{\theta_1, \theta_2, \ldots, \theta_m} \sum_{i=1}^{k} n_i \ln\left(\int_{x_i^l}^{x_i^u} f(x; \theta_1, \theta_2, \ldots, \theta_m)dx\right)$$

given

$$n_i, x_i^l, x_i^u, f(x)$$

We can solve the following system of m equations:

$$\frac{\partial \Lambda}{\partial \theta_l} = 0 \quad l \in \{1, 2, \ldots, m\}$$

Though evaluating these analytically may be difficult for distributions with many parameters, we can easily solve the problem in MATLAB using `fminsearch` by defining the objective function as the negative of the log-likelihood function.

Minimum Chi-Squared Statistic

This method appeals to the spirit of hypothesis testing in that it offers the most benefit of the doubt to the null hypothesis. The results of this method are most convincing in favor of the alternative hypothesis when the chi-squared statistic leaves a p-value in the upper tail less than the significance level. Here, we select the parameters θ_l that minimize the chi-squared statistic. That is, we have

$$\chi^2_{\text{ts}} = \sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i}$$

where

$$N = \sum_{i=1}^{k} n_i$$

and

$$p_i = \int_{x_i^l}^{x_i^u} f(x; \theta_1, \theta_2, \ldots, \theta_m)dx$$

Thus, we see that χ^2_{ts} is a function of the parameters θ_l, which we choose as the solution to

$$\min_{\theta_1, \theta_2, \ldots, \theta_m} \sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i}$$

We can solve the following system of m equations:

$$\frac{\partial \chi^2_{\text{ts}}}{\partial \theta_l} = 0 \quad l \in \{1, 2, \ldots, m\}$$

Again, evaluating these analytically may be difficult, but we can easily solve the problem in MATLAB using `fminsearch` by defining the objective function as the chi-squared statistic. To illustrate the facility of this procedure, we provide a simple example below.

Example

Consider empirical observations according to the following coarse histogram:

Range	< 66	66–68	68–70	70–72	72–74	> 74
Frequency	4	24	35	15	8	4

We wish to test the hypothesis that this random process is described by a normal distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The following MATLAB code estimates the parameters μ and σ using both of the methods discussed here:

```
% Observed statistics
ni  = [4 24 35 15 8 4];
xil = [-Inf 66 68 70 72 74];
xiu = [66 68 70 72 74 Inf];

% Initial guess: mean = 70, stdev = 1
X0  = [mean(xil(2:end)); 1];

Xcmle = fminsearch(@(X)loglikenorm(X(1),X(2),xiu,xil,ni),X0)
% COARSE MAXIMUM LIKELIHOOD ESTIMATION METHOD
% Xcmle =
%    69.2396  Mean
%     2.3103  Standard deviation

Xmcss = fminsearch(@(X)chisstat(X(1),X(2),xiu,xil,ni),X0)
% MINIMUM CHI-SQUARED STATISTIC METHOD
% Xmcss =
%    69.2569  Mean
%     2.3931  Standard deviation

% Negative log-likelihood function for null hypothesis
function L = loglikenorm(mu,sigma,xiu,xil,ni)
L = -sum( ni .* log( normcdf(xiu,mu,sigma)-normcdf(xil,mu,sigma) ) );

% Chi-squared statistic for null hypothesis
function chi2 = chisstat(mu,sigma,xiu,xil,ni)
ei   = sum(ni) * (normcdf(xiu,mu,sigma)-normcdf(xil,mu,sigma));
chi2 = sum( (ni - ei).^2 ./ ei );
```

These estimates are in fact reasonable. We also note that the two methods yield considerably different estimates and that the coarse maximum likelihood estimation method is more likely to lead to a rejection of the null hypothesis than the minimum chi-squared statistic, as the latter will always yield a smaller chi-squared statistic by its very construction.