Estimating Distribution Parameters from Data

M. Degonish

February 21, 2023

1 Sample Mean and Sample Variance

Given that we take a *sample* of n observations from some population, we can estimate the mean and variance of the population by using the taken sample. The calculations of the sample mean $\bar{x}$ and sample variance s are shown below:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

2 Using Sample Mean and Sample Variance to Estimate Parameters of a Normal Distribution

The normal (or Gaussian) distribution is the most frequently used statistical model. Its probability density function is:

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

where μ and σ are location and scale parameters, respectively, of the distribution.

In many problems μ and σ are not known. We can use sample statistics to estimate the parameters. To estimate the population mean μ and population variance σ^2:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

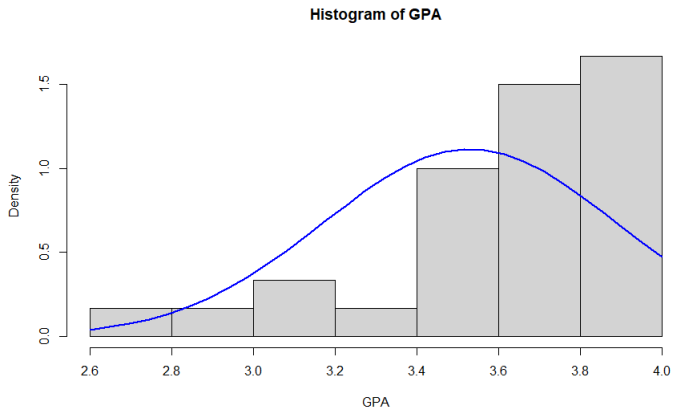
$$or\ s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

As an example, take the following highschool GPA data:

$$\vec{x} =$$

$$\begin{bmatrix} 3.9 & 3.8 & 3.0 & 3.9 & 3.6 \\ 3.2 & 3.8 & 3.2 & 3.8 & 3.5 \\ 3.8 & 3.9 & 3.8 & 4.0 & 3.3 \\ 3.7 & 3.9 & 3.5 & 3.9 & 4.0 \\ 3.5 & 3.8 & 3.9 & 3.5 & 3.8 \\ 4.0 & 3.5 & 2.6 & 3.8 & 4.0 \end{bmatrix}$$

The obtained parameters for the normal distribution can be compared to a histogram of the empirical data. Although parameters can be determined that best fit the data, as apparent from the histogram, the data is left-skewed and most likely the population is not normally distributed.



3 Using Sample Mean and Sample Variance to Estimate Parameters of a Lognormal Distribution

The lognormal distribution is the model for a random variable whose logarithm follows the normal distribution with parameters μ and σ. Thus, we can replace x in the normal distribution with *ln(x)*.

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

where μ and σ are *shape* and *scale* parameters, respectively, of the distribution.

In problems where μ and σ are not known, we can use sample statistics (just as we did for the normal distribution) to estimate the parameters. To estimate the population mean μ and population variance σ^2:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} ln(x_i)$$

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n}(ln(x_i) - \bar{x})^2}{n-1}$$

$$or\ s = \sqrt{\frac{\sum_{i=1}^{n}(ln(x_i) - \bar{x})^2}{n-1}}$$

As an example, take the following set of offsite power recovery estimates (in hours) from the nuclear industry:

$$\vec{x} = \begin{bmatrix} 2.33 & 0.67 & 4.62 & 0.95 & 0.17 & 0.25 \\ 0.5 & 0.5 & 4.0 & 0.7 & 1.5 & 0.67 \\ 1.82 & 6.67 & 6.45 & 7.32 & 12.6 & 2.95 \\ 33.88 & 5.43 & 27.6 & 27.6 & 3.27 & 0.42 \end{bmatrix}$$

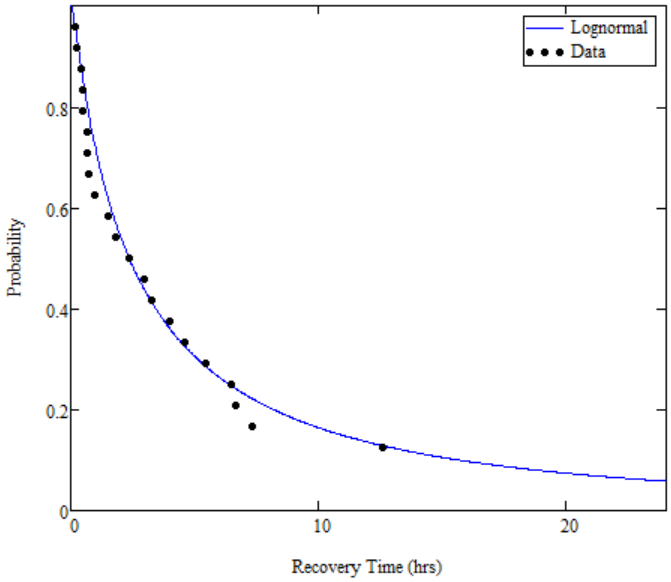
The parameters of the lognormal distribution can be estimate for this data set as follows:

$$\hat{\mu} = \frac{1}{24}\sum_{i=1}^{24} ln(x_i) = 0.855$$

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n}(ln(x_i) - \bar{x})^2}{n-1} = 2.29$$

$$s = \sqrt{s^2} = 1.51$$

The obtained parameters for the lognormal distribution can be compared to the empirical data, but to plot the empirical data an additional step must occur. Expected values of the observed data must be generated. To do this, the complement of the empirical distribution function (namely the reliability function) is generated for the data. the data points of the realiability function as well as the lognormal distribution are then plotted together.



3.1 Frequentist and Bayesian Confidence Intervals

Given the calculation of the lognormal shape and scale parameters, confidence intervals can be obtained either using the typical frequentist analysis, or a Bayesian approach.

Frequentist Approach

In the frequentist approach, to generate a confidence interval for the parameters (well, specifically for the shape parameter as this is typically the value of interest), first a significance level (α) must be selected. By default this value is typically set to $\alpha = 0.05$ which would correspond to a 95% confidence interval. The confidence interval can then be estimated by:

$$95\%\ CI = \bar{x} \pm t_{\frac{\alpha}{2},DOF}\frac{s}{\sqrt{n}}$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation, n is the number of data points, and DOF is equal to the number of data points minus 1.
By using a *t-table*, we can fill in the necessary variables to generate the confidence interval:

$$95\%\ CI = 0.855 \pm 1.714\frac{1.51}{\sqrt{23}} = 0.855 \pm 0.5396 = (0.315, 1.395)$$

These values can be used as the corresponding 5th and 95th percentile values to support uncertainty analysis of the fitted shape parameter.

Bayesian Approach

The Bayesian approach uses the winBUGS software to generate the estimates of the 5th and 95th percentile values for the parameters. In this approach we can also generate 5th and 95th percentile values for the scale parameter. The script used for this approach is taken from [2] and modifieid to meet the parameters for this example. The most notable change to the script is the data input, as well as the initialization of the shape and scale parameters. The script converges better if good estimates of these parameters are provided. The initial estimates of the parameters are dispersed around the MLE values found in Section 3.0.

The image above shows the script, the sampled nodes (for the shape and scale parameters), and output plots of the simulation which show convergence in the script. From the simulation, it can be seen that the mean value of the shape parameter is $\hat{\mu} = 0.855$ and $\hat{\sigma} = 1.561$. These parameters agree well with the MLE values found in Section 3.0. Furthermore, We can compare the 5th and 95th percentile estimates for the shape parameter $\hat{\mu}$. We can see that the Bayesian confidence interval for the shape parameter is (0.325, 1.376) which is in good agreement with the frequentist confidence interval. Additionally, we obtain a Bayesian confidence interval for the scale parameter which is (1.22, 2.115). To obtain a similar interval using the frequentist method, some assumptions regarding the distribution of the standard deviatino would need to be made; therefore, the Bayesian approach provides for a more straightforward approach to generating this estimate.

4 References

1. NUREG/CR-6823, *Handbook of Parameter Estimation for Probabilistic Risk Assessment*, U.S. Nuclear Regulatory Commission, September 2003.

2. D. Kelly and C. Smith, *Bayesian Inference for Probabilistic Risk Assessment - A Practitioner's Guidebook*, Springer 2011.