NAG Library Chapter Introduction

G07 – Univariate Estimation

Contents

1 Scope of the Chapter

This chapter deals with the estimation of unknown parameters of a univariate distribution. It includes both point and interval estimation using maximum likelihood and robust methods.

2 Background to the Problems

Statistical inference is concerned with the making of inferences about a **population** using the observed part of the population called a **sample**. The population can usually be described using a probability model which will be written in terms of some unknown **parameters**. For example, the hours of relief given by a drug may be assumed to follow a Normal distribution with mean μ and variance σ^2; it is then required to make inferences about the parameters, μ and σ^2, on the basis of an observed sample of relief times.

There are two main aspects of statistical inference: the **estimation** of the parameters and the **testing of hypotheses** about the parameters. In the example above, the values of the parameter σ^2 may be estimated and the hypothesis that $\mu \geq 3$ tested. This chapter is mainly concerned with estimation but the test of a hypothesis about a parameter is often closely linked to its estimation. Tests of hypotheses which are not linked closely to estimation are given in the chapter on nonparametric statistics (Chapter G08).

There are two types of estimation to be considered in this chapter: **point estimation** and **interval estimation**. Point estimation is when a single value is obtained as the best estimate of the parameter. However, as this estimate will be based on only one of a large number of possible samples, it can be seen that if a different sample were taken, a different estimate would be obtained. The distribution of the estimate across all the possible samples is known as the **sampling distribution**. The sampling distribution contains information on the performance of the estimator, and enables estimators to be compared. For example, a good estimator would have a sampling distribution with mean equal to the true value of the parameter; that is, it should be an **unbiased** estimator; also the variance of the sampling distribution should be as small as possible. When considering a parameter estimate it is important to consider its variability as measured by its variance, or more often the square root of the variance, the **standard error**.

The sampling distribution can be used to find interval estimates or confidence intervals for the parameter. A **confidence interval** is an interval calculated from the sample so that its distribution, as given by the sampling distribution, is such that it contains the true value of the parameter with a certain probability.

Estimates will be functions of the observed sample and these functions are known as **estimators**. It is usually more convenient for the estimator to be based on statistics from the sample rather than all the individuals observations. If these statistics contain all the relevant information then they are known as **sufficient statistics**. There are several ways of obtaining the estimators; these include least squares, the method of moments, and **maximum likelihood**. Least squares estimation requires no knowledge of the distributional form of the error apart from its mean and variance matrix, whereas the method of maximum likelihood is mainly applicable to situations in which the true distribution is known apart from the values of a finite number of unknown parameters. Note that under the assumption of Normality, the least squares estimation is equivalent to the maximum likelihood estimation. Least squares is often used in regression analysis as described in Chapter G02, and maximum likelihood is described below.

Estimators derived from least squares or maximum likelihood will often be greatly affected by the presence of extreme or unusual observations. Estimators that are designed to be less affected are known as **robust estimators**.

2.1 Maximum Likelihood Estimation

Let X_i be a univariate random variable with probability density function

$$f_{X_i}(x_i; \theta),$$

where θ is a vector of length p consisting of the unknown parameters. For example, a Normal distribution with mean θ_1 and standard deviation θ_2 has probability density function

$$\frac{1}{\sqrt{2\pi}\theta_2}\exp\left(-\frac{1}{2}\left(\frac{x_i-\theta_1}{\theta_2}\right)^2\right).$$

The likelihood for a sample of n independent observations is

$$\text{Like} = \prod_{i=1}^{n} f_{X_i}(x_i;\theta),$$

where x_i is the observed value of X_i. If each X_i has an identical distribution, this reduces to

$$\text{Like} = \prod_{i=1}^{n} f_X(x_i;\theta), \tag{1}$$

and the log-likelihood is

$$\log(\text{Like}) = L = \sum_{i=1}^{n} \log\left(f_X(x_i;\theta)\right). \tag{2}$$

The maximum likelihood estimates ($\hat{\theta}$) of θ are the values of θ that maximize (1) and (2). If the range of X is independent of the parameters, then $\hat{\theta}$ can usually be found as the solution to

$$\sum_{i=1}^{n} \frac{\partial}{\partial\hat{\theta}_j}\log\left(f_X\left(x_i;\hat{\theta}\right)\right) = \frac{\partial L}{\partial\hat{\theta}_j} = 0, \quad j = 1,2,\ldots,p. \tag{3}$$

Note that $\dfrac{\partial L}{\partial\theta_j}$ is known as the efficient score.

Maximum likelihood estimators possess several important properties.

(a) Maximum likelihood estimators are functions of the sufficient statistics.

(b) Maximum likelihood estimators are (under certain conditions) **consistent**. That is, the estimator converges in probability to the true value as the sample size increases. Note that for small samples the maximum likelihood estimator may be biased.

(c) For maximum likelihood estimators found as a solution to (3), subject to certain conditions, it follows that

$$E\left(\frac{\partial L}{\partial\theta}\right) = 0, \tag{4}$$

and

$$I(\theta) = -E\left(\frac{\partial^2 L}{\partial\theta^2}\right) = E\left(\left(\frac{\partial L}{\partial\theta}\right)^2\right), \tag{5}$$

and then that $\hat{\theta}$ is asymptotically Normal with mean vector θ_0 and variance-covariance matrix $I_{\theta_0}^{-1}$ where θ_0 denotes the true value of θ. The matrix I_θ is known as the information matrix and $I_{\theta_0}^{-1}$ is known as the Cramer–Rao lower bound for the variance of an estimator of θ.

For example, if we consider a sample, $x_1, x_2, \ldots, x_n$, of size n drawn from a Normal distribution with unknown mean μ and unknown variance σ_2 then we have

$$L = \log\left(\text{Like}\left(\mu,\sigma^2;x\right)\right) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2$$

and thus

$$\frac{\partial L}{\partial\mu} = \sum_{i=1}^{n}(x_i-\mu)/\sigma^2$$

and

$$\frac{\partial L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \sum_{i=1}^{n}(x_i - \mu)^2/2\sigma^4.$$

Then equating these two equations to zero and solving gives the maximum likelihood estimates

$$\hat{\mu} = \bar{x}$$

and

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

These maximum likelihood estimates are asymptotically Normal with mean vector a, where

$$a^{\mathrm{T}} = (\mu, \sigma^2),$$

and covariance matrix C. To obtain C we find the second derivatives of L with respect to μ and σ^2 as follows:

$$\begin{aligned}
\frac{\partial^2 L}{\partial \mu^2} &= -\frac{n}{\sigma^2} \\
\frac{\partial^2 L}{\partial (\sigma^2)^2} &= \frac{n}{2\sigma^4} - \sum_{i=1}^{n}(x_i - \mu)^2/\sigma^6 \\
\frac{\partial^2 L}{\partial \mu \partial \sigma^2} &= \frac{\partial^2 L}{\partial \sigma^2 \partial \mu} = -\frac{n(\bar{x} - \mu)}{\sigma^4}.
\end{aligned}$$

Then

$$C^{-1} = -E\begin{pmatrix} \dfrac{\partial^2 L}{\partial \mu^2} & \dfrac{\partial^2 L}{\partial \sigma^2 \partial \mu} \\ \dfrac{\partial^2 L}{\partial \mu \partial \sigma^2} & \dfrac{\partial^2 L}{\partial (\sigma^2)^2} \end{pmatrix} = \begin{pmatrix} n/\sigma^2 & 0 \\ 0 & n/2\sigma^4 \end{pmatrix}$$

so that

$$C = \begin{pmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{pmatrix}.$$

To obtain an estimate of C the matrix may be evaluated at the maximum likelihood estimates.

It may not always be possible to find maximum likelihood estimates in a convenient closed form, and in these cases iterative numerical methods, such as the Newton–Raphson procedure or the EM algorithm (expectation maximization), will be necessary to compute the maximum likelihood estimates. Their asymptotic variances and covariances may then be found by substituting the estimates into the second derivatives. Note that it may be difficult to find the expected value of the second derivatives required for the variance-covariance matrix and in these cases the observed value of the second derivatives is often used.

The use of maximum likelihood estimation allows the construction of generalized likelihood ratio tests. If $\lambda = 2(l_1 - l_2)$, where l_1 is the maximized log-likelihood function for a model 1 and l_2 is the maximized log-likelihood function for a model 2, then under the hypothesis that model 2 is correct, 2λ is asymptotically distributed as a χ^2 variable with $p - q$ degrees of freedom. Consider two models in which model 1 has p parameters and model 2 is a sub-model (nested model) of model 1 with $q < p$ parameters, that is model 1 has an extra $p - q$ parameters. This result provides a useful method for performing hypothesis tests on the parameters. Alternatively, tests exist based on the asymptotic Normality of the estimator and the efficient score; see page 315 of Cox and Hinkley (1974).

2.2 Confidence Intervals

Suppose we can find a function, $t(x,\theta)$, whose distribution depends upon the sample x but not on the unknown parameter θ, and which is a monotonic (say decreasing) function in θ for each x, then we can find t_1 such that $P(t_1 \leq t(x,\theta)) = 1-\alpha$ no matter what θ happens to be. The function $t(x,\theta)$ is known as a pivotal quantity. Since the function is monotonic the statement that $t_1 \leq t(x,\theta)$ may be rewritten as $\theta \geq \theta_1(x)$ see Figure 1. The statistic $\theta_1(x)$ will vary from sample to sample and if we assert that $\theta \geq \theta_1(x)$ for any sample values which arise, we will be right in a proportion $1-\alpha$ of the cases, in the long run or on average. We call $\theta_1(x)$ a $1-\alpha$ upper confidence limit for θ.

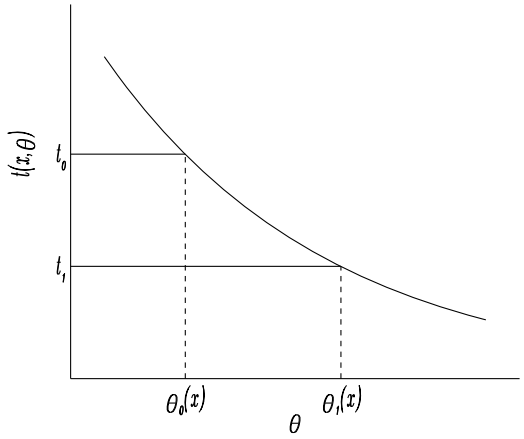

Figure 1

We have considered only an upper confidence limit. The above idea may be generalized to a two-sided confidence interval where two quantities, t_0 and t_1, are found such that for all θ, $\mathrm{P}(t_1 \leq t(x,\theta) \leq t_0) = 1-\alpha$. This interval may be rewritten as $\theta_0(x) \leq \theta \leq \theta_1(x)$. Thus if we assert that θ lies in the interval $[\theta_0(x), \theta_1(x)]$ we will be right on average in $1-\alpha$ proportion of the times under repeated sampling.

Hypothesis (significance) tests on the parameters may be used to find these confidence limits. For example, if we observe a value, k, from a binomial distribution, with known parameter n and unknown parameter p, then to find the lower confidence limit we find p_l such that the probability that the null hypothesis H_0: $p = p_l$ (against the one sided alternative that $p > p_l$) will be rejected, is less than or equal to $\alpha/2$. Thus for a binomial random variable, B, with parameters n and p_l we require that $P(B \geq k) \leq \alpha/2$. The upper confidence limit, p_u, can be constructed in a similar way.

For large samples the asymptotic Normality of the maximum likelihood estimates discussed above is used to construct confidence intervals for the unknown parameters.

2.3 Robust Estimation

For particular cases the probability density function can be written as

$$f_{X_i}(x_i;\theta) = \frac{1}{\theta_2} g\left(\frac{x_i - \theta_1}{\theta_2}\right)$$

for a suitable function g; then θ_1 is known as a location parameter and θ_2, usually written as σ, is known as a scale parameter. This is true of the Normal distribution.

If θ_1 is a location parameter, as described above, then equation (3) becomes

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \hat{\theta}_1}{\hat{\sigma}}\right) = 0, \tag{6}$$

where $\psi(z) = -\dfrac{d}{dz}\log\left(g(z)\right)$.

For the scale parameter σ (or σ^2) the equation is

$$\sum_{i=1}^{n} \chi\left(\frac{x_i - \hat{\theta}_1}{\hat{\sigma}}\right) = n/2, \tag{7}$$

where $\chi(z) = z\psi(z)/2$.

For the Normal distribution $\psi(z) = z$ and $\chi(z) = z^2/2$. Thus, the maximum likelihood estimates for θ_1 and σ^2 are the sample mean and variance with the n divisor respectively. As the latter is biased, (7) can be replaced by

$$\sum_{i=1}^{n} \chi\left(\frac{x_i - \hat{\theta}_1}{\hat{\sigma}}\right) = (n-1)\beta, \tag{8}$$

where β is a suitable constant, which for the Normal χ function is $\frac{1}{2}$.

The influence of an observation on the estimates depends on the form of the ψ and χ functions. For a discussion of influence, see Hampel *et al.* (1986) and Huber (1981). The influence of extreme values can be reduced by bounding the values of the ψ- and χ-functions. One suggestion due to Huber (1981) is

$$\psi(z) = \begin{cases} -C, & z < -C \\ z, & |z| \leq C \\ C, & z > C. \end{cases}$$

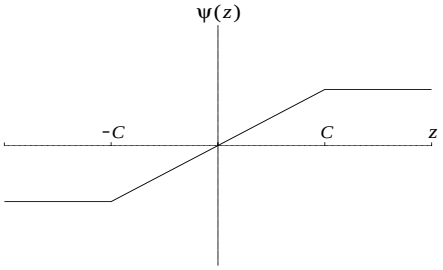

Figure 2

Redescending ψ-functions are often considered; these give zero values to $\psi(z)$ for large positive or negative values of z. Hampel *et al.* (1986) suggested

$$\psi(z) = \begin{cases} -\psi(-z) & \\ z, & 0 \leq z \leq h_1 \\ h_1, & h_1 \leq z \leq h_2 \\ h_1(h_3 - z)/(h_3 - h_2), & h_2 \leq z \leq h_3 \\ 0, & z > h_3. \end{cases}$$

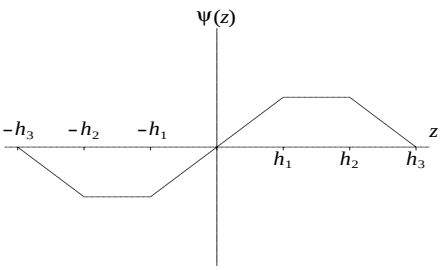

Figure 3

Usually a χ-function based on Huber's ψ-function is used: $\chi = \psi^2/2$. Estimators based on such bounded ψ-functions are known as M-estimators, and provide one type of robust estimator.

Other robust estimators for the location parameter are

(i) the sample median,

(ii) the trimmed mean, i.e., the mean calculated after the extreme values have been removed from the sample,

(iii) the winsorized mean, i.e., the mean calculated after the extreme values of the sample have been replaced by other more moderate values from the sample.

For the scale parameter, alternative estimators are

(i) the median absolute deviation scaled to produce an estimator which is unbiased in the case of data coming from a Normal distribution,

(ii) the winsorized variance, i.e., the variance calculated after the extreme values of the sample have been replaced by other more moderate values from the sample.

For a general discussion of robust estimation, see Hampel *et al.* (1986) and Huber (1981).

2.4 Robust Confidence Intervals

In Section 2.2 it was shown how tests of hypotheses can be used to find confidence intervals. That approach uses a parametric test that requires the assumption that the data used in the computation of the confidence has a known distribution. As an alternative, a more robust confidence interval can be found by replacing the parametric test by a nonparametric test. In the case of the confidence interval for the location parameter, a Wilcoxon test statistic can be used, and for the difference in location, computed from two samples, a Mann–Whitney test statistic can be used.

3 Recommendations on Choice and Use of Available Routines

Maximum Likelihood Estimation and Confidence Intervals

G07AAF provides a confidence interval for the parameter p of the binomial distribution.

G07ABF provides a confidence interval for the mean parameter of the Poisson distribution.

G07BBF provides maximum likelihood estimates and their standard errors for the parameters of the Normal distribution from grouped and/or censored data.

G07BEF provides maximum likelihood estimates and their standard errors for the parameters of the Weibull distribution from data which may be right-censored.

G07BFF provides maximum likelihood estimates and their standard errors for the parameters of the generalized Pareto distribution.

G07CAF provides a t-test statistic to test for a difference in means between two Normal populations, together with a confidence interval for the difference between the means.

Robust Estimation

G07DBF provides M-estimates for location and, optionally, scale using four common forms of the ψ-function.

G07DCF produces the M-estimates for location and, optionally, scale but for user-supplied ψ- and χ-functions.

G07DAF provides the sample median, median absolute deviation, and the scaled value of the median absolute deviation.

G07DDF provides the trimmed mean and winsorized mean together with estimates of their variance based on a winsorized variance.

Robust Internal Estimation

G07EAF produces a rank based confidence interval for locations.

G07EBF produces a rank based confidence interval for the difference in location between two populations.

Outlier Detection

This chapter provides two routines for identifying potential outlying values, G07GAF and G07GBF. Many of the model fitting routines, for examples those in Chapters G02 and G13 also return vectors of residuals which can also be used to aid in the identification of outlying values.

4 Functionality Index

2 sample t-test G07CAF

Confidence intervals for parameters,
- binomial distribution G07AAF
- Poisson distribution G07ABF

Maximum likelihood estimation of parameters,
- Normal distribution, grouped and/or censored data G07BBF
- Weibull distribution G07BEF

Outlier detection,
- Peirce,
  - raw data or single variance supplied G07GAF
  - two variances supplied G07GBF

Parameter estimates,
- generalized Pareto distribution G07BFF

Robust estimation,
- confidence intervals,
  - one sample G07EAF
  - two samples G07EBF
- median, median absolute deviation and robust standard deviation G07DAF
- M-estimates for location and scale parameters,
  - standard weight functions G07DBF
  - trimmed and winsorized means and estimates of their variance G07DDF
  - user-defined weight functions G07DCF

5 Auxiliary Routines Associated with Library Routine Arguments

None.

6 Routines Withdrawn or Scheduled for Withdrawal

None.

7 References

Cox D R and Hinkley D V (1974) *Theoretical Statistics* Chapman and Hall

Hampel F R, Ronchetti E M, Rousseeuw P J and Stahel W A (1986) *Robust Statistics. The Approach Based on Influence Functions* Wiley

Huber P J (1981) *Robust Statistics* Wiley

Kendall M G and Stuart A (1973) *The Advanced Theory of Statistics (Volume 2)* (3rd Edition) Griffin

Silvey S D (1975) *Statistical Inference* Chapman and Hall