Chapter g07 – Univariate Estimation

1. Scope of the Chapter

This chapter deals with the estimation of unknown parameters of a univariate distribution.

2. Background

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.

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.

2.1. Normal distribution

For a single sample, $x_1 \ldots x_n$, from a Normal distribution with mean μ and variance $\sigma\tau^2$ the unbiased estimators with minimum variance are $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = \sum_{i=1}^{n}(x - \bar{x})^2/(n-1)$. From these two statistics confidence intervals for μ and σ^2 can easily be calculated. In the case of two samples $x_1 \ldots x_n$ and $y_1 \ldots y_n$ from Normal distributions with means μ_x and μ_y and variances σ_x^2 and σ_y^2 respectively the difference in means $\hat{\mu}_x - \hat{\mu}_y$ can be estimated by $\bar{x} - \bar{y}$. Further inferences about the difference depend on assumptions about σ_x^2 and σ_y^2. If it is assumed that $\sigma_x^2 = \sigma_y^2 = \sigma^2$ then a pooled estimate for σ^2 is computed from both samples and used to calculate confidence intervals and compute a t-statistic to test the hypothesis $\mu_x = \mu_y$. If $\sigma_x^2 \neq \sigma_y^2$ then it is known as the Behrens–Fisher problem and Satterthwaites, procedure is used.

2.2. Robust Estimation

While the mean and the sample variance represent the best estimates of the required assumptions hold they can be greatly affected by the presence of unusual or extreme values. Robust estimators are those which are less affected by unusual values. One type of robust estimator is the M-estimator, which is based on the approach used in maximum likelihood estimators.

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

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

and the log-likelihood is

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

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

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

For particular cases the above probability density function can be written as

$$f_{X_i}(x_i;\theta) = \frac{1}{\theta_2} g\left(\frac{x_i - \theta_1}{\theta_2}\right), \text{ for a suitable function } g,$$

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

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \hat{\theta}_1}{\hat{\sigma}}\right) = 0, \quad (4)$$

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

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

$$\sum_{i=1}^{n} \chi\left(\frac{x_i - \hat{\theta}_1}{\hat{\sigma}}\right) = \frac{n}{2}, \quad (5)$$

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

For the Normal distribution $\psi(z) = z$ and $\psi(z) = z^2/2$. 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, (5) can be replaced by

$$\sum_{i=1}^{n} \chi\left(\frac{x_i - \hat{\theta}_1}{\hat{\sigma}}\right) = (n-1)\beta, \quad (6)$$

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

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

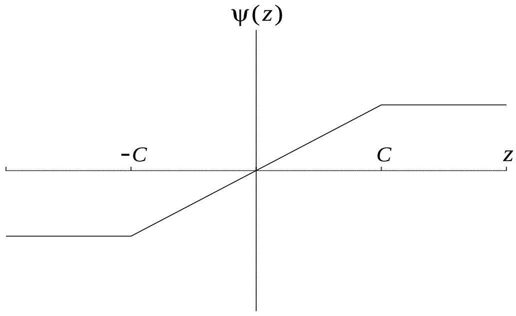

Figure 1

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

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

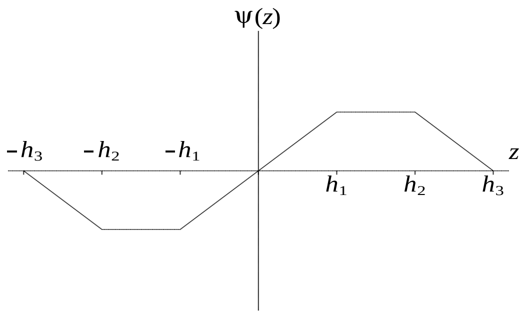

Figure 2

Usually a χ-function based on Huber's ψ-function is used: $\chi = \psi^2/2$. Estimators based on such bounded ψ-functions are M-estimators.

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).

2.3. 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)* Griffin (3rd Edition).

3. Available Functions

Compute two sample t-test and confidence interval	`g07cac`
Compute median and median absolute deviation	`g07dac`
Compute M-estimate of location and scale	`g07dbc`
Compute trimmed and Winsorized means	`g07ddc`