Chapter 5: Estimation

This chapter is concerned with estimating parameters from data. Some examples:

(i) We may wish to estimate the proportion of defective components produced by a particular manufacturer by examining a random sample of such components.

(ii) A quality control engineer may wish to estimate the mean lifetime of a particular type of light bulb produced by a particular production line, by testing a sample of these light bulbs.

(iii) A child health research team may wish to estimate the mean weight of one-year old girls in a particular ethnic group by considering a sample of such one-year olds.

1 Definitions

a) **Statistic:** Given data $X_1, ..., X_n$, any function of the data is called a **statistic**.

b) **Estimator:** Suppose we have data $X_1, ..., X_n$ and we are interested in fitting a mathematical model to it (for example, a Normal distribution in (iii) above). To do this, we must estimate the parameters in the model (μ and σ^2 for a Normal distribution). Any statistic which is used to estimate the parameters is called an **estimator**.

E.g. $\bar{X}$ is an estimator of the population mean μ.
s^2 is an estimator of the population variance σ^2.

Estimators are themselves random variables, as their values change depending on the data.

c) **Estimate:** The particular value of an estimator, obtained from a single set of data, is called an **estimate**.

2 Estimating the mean of a random variable

Suppose we have an independent random sample $X_1, ..., X_n$ from a distribution with mean μ and variance σ^2. Then the sample mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is an estimator of μ. We have seen previously that

$$E(\bar{X}) =$$

– we say that $\bar{X}$ is **unbiased** for μ, i.e. correct on average – and

$$\mathrm{Var}(\bar{X}) =$$

In addition, we have seen from the Central Limit Theorem that, for large n, the distribution of $\bar{X}$ is approximately . This theoretical distribution of $\bar{X}$ is called the **sampling distribution** of $\bar{X}$.

The **mean square error** of the estimator $\bar{X}$ is

$$E[(\bar{X} - \mu)^2] =$$

and its square root (i.e. the standard deviation of $\bar{X}$) is called the **standard error** of the estimator $\bar{X}$. The smaller the standard error (i.e. the smaller the mean square error), the more precise the estimator is. For a Normal distribution, the sample median is also an unbiased estimator of μ, but it has a larger standard error than $\bar{X}$ and is therefore not so good as an estimator for μ. The unbiased estimator with the smallest possible variance (and hence the smallest possible standard error) is called the **minimum variance unbiased estimator (MVUE)** (if it exists).

3 Confidence intervals for the mean

A sample mean $\bar{x}$ gives us a **point estimate** of the true mean μ. It is often important to attach a **confidence interval (CI)** to such an estimate, i.e. an interval of which we are reasonably confident that it will contain the true value of μ - this is also called **interval estimation**.

If the sample size is large enough for the Central Limit Theorem to apply, or if the data are Normally distributed, then

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

and

$$P\left(\bar{X}-1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

The interval

$$\left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}}\,,\,\bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right)$$

is said to be a 95% CI for μ.

Similarly, a 99% CI for μ is

$$\left(\bar{x}-2.58\frac{\sigma}{\sqrt{n}}\,,\,\bar{x}+2.58\frac{\sigma}{\sqrt{n}}\right).$$

More generally, a R% CI for μ is given by

$$\left(\bar{x}-z\cdot\frac{\sigma}{\sqrt{n}}\,,\,\bar{x}+z\cdot\frac{\sigma}{\sqrt{n}}\right)$$

where z is a constant such that

$$P(-z < Z < z) = R\%$$

for a standard Normal random variable Z (the value of z can be found from the table of the percentage points of the Normal distribution).

Note that the higher the confidence level required, the wider the interval. Note also that as the sample size n increases, the CI becomes narrower.

The interpretation of, for example, a 95% CI is: if we were to collect a large number of data sets from the same distribution, and for each data set we compute a 95% CI for μ, then in about 95% of cases the CI would contain the true value of μ, whilst in about 5% of cases the true value of μ would lie *outside* the CI.

Example (1): A sample of four rods is taken from a process that produces rods of about 30cm in length. The average length of the four rods in the sample is 29.575cm. Assuming that the lengths are Normally distributed with standard deviation 0.500cm, obtain 90%, 95% and 99% confidence intervals for the mean length of rods produced by this process (the 'process mean'). How large a sample should the engineer take in order to obtain a 95% confidence interval that is no longer than 0.200 cm?

Note that the CI's on the previous page depend on the (true) variance σ^2. In practice, we often don't know what σ^2 is, so we have to estimate it by

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

Now s^2 is also an estimator and hence is random. Some allowance must be made for this. If the sample size is large enough for the Central Limit Theorem to apply, or if the data are Normally distributed, then the statistic

$$\frac{\bar{X} - \mu}{s/\sqrt{n}}$$

has a **t-distribution with (n-1) degrees of freedom**. This distribution can be found in tables (for example, the New Cambridge Elementary Statistical Tables). In this case, a $R\%$ CI for μ is given by

$$\left(\bar{x} - t \cdot \frac{s}{\sqrt{n}}\,,\, \bar{x} + t \cdot \frac{s}{\sqrt{n}}\right)$$

where t is a constant such that

$$P(-t < T < t) = R\%$$

for a random variable T which has a t-distribution with $n-1$ degrees of freedom (the value of t can be found from the table of the percentage points of the t-distribution). This CI is always wider than the corresponding CI when σ^2 is known, reflecting the additional uncertainty present. Again, as the sample size n increases, the CI becomes narrower.

Example (2): A computer network is about to be updated to cope with growing demand. To assess the current demand, the network is monitored at hourly intervals during working hours (9am-5pm) over a period of a week (Mon-Fri). The mean number of requests for connection per hour is found to be 43.2, and the sample variance is found to be 50.9. Calculate a 95% and a 99% CI for the true mean hourly demand.

4 Estimating the variance

The sample variance

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

is an unbiased estimator of σ^2, i.e. $E(s^2) = \sigma^2$ (this is why we use $n-1$ instead of n in the denominator). Its sampling distribution is such that

$$\frac{(n-1)s^2}{\sigma^2}$$

follows a **chi-squared (χ^2) distribution with (n-1) degrees of freedom** (abbreviated to χ^2_{n-1}). This can be used to obtain confidence intervals for σ^2 if necessary.