Chapter 5

Estimators

Definition. A random sample is a sample that satisfies two conditions:

- Every object has an equal probability of being selected
- The objects are selected independently.

Consider a random sample $x_1, ..., x_n$. Notice that - although any given sample observation x_i takes a specific numerical value - in each sample it is still a random variable, since if the sampling process were repeated it will then take a different numerical value. Therefore each of these observations are ex-ante identical and independent random variables.

5.1. Definition and properties

Let $x_1, ..., x_n$ be a random sample of X. Let τ be a population parameter that characterizes X and let T be a measure designed to estimate τ. Since each observations in a sample is ex-ante an iid random variable, also T is random variable.

Estimators. Let T be a statistics to estimate τ. Then T is an estimator of τ. It is a random variable and depends on the sample data. Its distribution is called *sampling distribution*.

Unbiasedness. An estimator T of τ is said to be unbiased if

$$E[T] = \tau$$

The bias of an estimator T is defined as the difference between its expected value and value of the population parameter, i.e.

$$\text{bias}[T] = E[T] - \tau$$

If $E[T] > \tau$, we say that T is biased upward. If $E[T] < \tau$, we say that T is biased downward.

Efficiency. Let T_1 and T_2 be two two unbiased estimators of τ. Let $Var[T_1] < VAR[T_2]$. Then T_1 is said to be more efficient than T_2.

Mean square error. A combined measure of bias and inefficiency is given by the mean squared error, defined as follows:

$$\begin{aligned}
\mathrm{MSE}[T] = \mathrm{E}[(T-\tau)^2] = \mathrm{E}[(T-\mathrm{E}[T]+\mathrm{E}[T]-\tau)^2] = \\
\mathrm{E}[(T-\mathrm{E}[T])^2 + (\mathrm{E}[T]-\tau)^2 + 2(T-\mathrm{E}[T])(\mathrm{E}[T]-\tau)] = \\
\mathrm{E}[(T-\mathrm{E}[T])^2] + \mathrm{E}[(\mathrm{E}[T]-\tau)^2] + 2\mathrm{E}[(T-\mathrm{E}[T])(\mathrm{E}[T]-\tau)]] = \\
\mathrm{E}[(T-\mathrm{E}[T])^2] + \mathrm{E}[(\mathrm{E}[T]-\tau)^2] + 2\mathrm{E}[(T\mathrm{E}[T] - T\tau - (\mathrm{E}[T])^2 + \mathrm{E}[T]\tau)] = \\
\mathrm{E}[(T-\mathrm{E}[T])^2] + \mathrm{E}[(\mathrm{E}[T]-\tau)^2] = \\
\mathrm{VAR}(T) + \mathrm{E}[\mathrm{BIAS}(T)^2] = \mathrm{VAR}(T) + \mathrm{BIAS}(T)^2
\end{aligned}$$

Among the unbiased estimators, the one with the minimum MSE has the smallest variance, i.e. is the most efficient.

Asyptotical unbiasedness. Let $\{T_n\}$ be a sequence of the same estimator T of τ, constructed over samples with different sizes n. Then T is said to be asyptotically unbiased if

$$\lim_{n\to\infty} E[T_n] = \tau$$

i.e. if the bias tends to zero as the sample size n increases.

Consistency. Let $\{T_n\}$ be a sequence of the same estimator T of τ, constructed over samples with different sizes n. Then T is said to be a consistent estimator of τ if the probability of deviations of T_n from τ decreases as n increases, i.e.

$$\lim_{n\to\infty} P[|T_n - \tau| > \epsilon] = 0$$

for any $\epsilon > 0$. Notice that if an estimator is asymptotically unbiased, then it is consistent if

$$\lim_{n\to\infty} VAR[T_n] = 0$$

Notice that any asympotically unbiased estimator can still display positive variance as n increases if it inconsistent.

5.1.1 Example: The sample mean

Consider a series of continuous iid random variable X. Assume μ the expected value of X and σ^2 be the variance. Let $\bar{x}$ be the sample mean of X from a sample of size n.

The expected value of $\bar{x}$ is equal to:

$$\mathrm{E}[\bar{x}] = \mathrm{E}\left[\frac{\sum_{i=1}^{n} x_i}{n}\right] = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n} x_i\right] =$$
$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[x_i] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[x_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu = \mu$$

This implies that the $\bar{x}$ is an unbiased estimator of μ, i.e.

$$\mathrm{BIAS}[\bar{x}] = \mathrm{E}[\bar{x}] - \mu = 0$$

The variance of $\bar{x}$ is equal to:

$$\mathrm{VAR}[\bar{x}] = \mathrm{VAR}\left[\frac{\sum_{i=1}^{n} x_i}{n}\right] = \frac{1}{n^2}\mathrm{VAR}\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{VAR}[x_i] = \frac{\sigma^2}{n}$$

Notice that if X where not iid, then:

$$\mathrm{VAR}[\bar{x}] = \mathrm{VAR}\left[\frac{\sum_{i=1}^{n} x_i}{n}\right] = \frac{1}{n^2}\mathrm{VAR}\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\left[\sum_{i=1}^{n}\mathrm{VAR}[x_i] + 2\sum_{i=1}^{n}\sum_{j=i} COV[x_i, x_j]\right]$$

Let s^2 be the sample variance of X from a sample of size n. The expected value of s^2 is equal to

$$\mathrm{E}[s^2] = \mathrm{E}\left[\frac{\sum_{i=1}^{n}[x_i - \mathrm{E}[\bar{x}]]^2}{n-1}\right] = \sigma^2$$

The variance of s^2 is equal to

$$\mathrm{VAR}[s^2] = \mathrm{VAR}\left[\frac{\sum_{i=1}^{n}[x_i - \mathrm{E}[\bar{x}]]^2}{n-1}\right] = \frac{2\sigma^2}{n-1}$$

5.1.2 Theorem: The central limit theorem

Informally, the central limit theorem states that the sample mean of a random sample of n observations drawn from a population with any probability distribution will be approximately normally distributed, if n is large.

Theorem: Let $X_1, X_2, ...$ be i.i.d. random variables obtained by sampling from an arbitrary population. Let $S_n = X_1 + X_2 + ...X_n$. Denote by $\mathrm{E}[S_n]$ and $\mathrm{VAR}[S_n]$

expected value and variance of S_n. Let

$$z_n = \frac{S_n - \mathrm{E}[S_n]}{\sqrt{\mathrm{VAR}[S_n]}}$$

Then, z_n converge in distribution to a standard normal distribution for n that increases.

Corollary The random variable $\frac{S_n}{n}$ converges to a normal distribution with expected value $\frac{1}{n}\mathrm{E}[S_n]$ and variance $\frac{1}{n^2}\mathrm{VAR}[S_n]$.

5.1.3 Theorem: The law of large number

Informally, the law of large number states that, for a given a random sample of size n taken from a population mean, the sample mean will *approach* the population mean as n increases, regardless of the underlying probability distribution of the data.

Theorem: Let $X_1, X_2, ...$ be i.i.d. random variables obtained by sampling from an arbitrary population. Let $\bar{x}_n$ be the sample mean for a sample with size n. Then, for any positive number $\epsilon > 0$,

$$\lim_{n \to \infty} \mathrm{P}\left(|\bar{x}_n - \mu| > \epsilon\right) = 0$$

5.2. Confidence Interval

Informally, a confidence interval (CI) indicates a range of values that's likely to encompass the population value. The probability that the confidence interval encompasses the true value is called the confidence level of the CI.
To construct a confidence interval for a population parameter τ at confidence level $\alpha\%$, we have to find the interval where τ will falls into with probability $1-\alpha$. Practically, we identify a sample statistic that we cab use to estimate a population parameter, say $T \sim N(\tau, \sigma_T^2)$.

Case 1 $VAR[T]$ is a known parameter.

Then we compute

$$P(a \leq T \leq b) = 1 - \alpha$$

To do so, we standardize T and obtain:

$$P\left(\frac{a-\tau}{\sigma_T} \leq Z \leq \frac{b-\tau}{\sigma_T}\right) = 1 - \alpha$$

or equivalently,

$$P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$

where $z_{\frac{\alpha}{2}}$ is called critical value of z statistics corresponding to α for a two-tail confidence interval. $z_{\frac{\alpha}{2}}$ can be computed using the statistical tables for a standard normal random variable. Therefore, we get:

$$\begin{aligned} P\left(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}\right) &= 1 - \alpha \\ P\left(-z_{\frac{\alpha}{2}} \leq \frac{T-\tau}{\sigma_T} \leq z_{\frac{\alpha}{2}}\right) &= 1 - \alpha \\ P\left(-z_{\frac{\alpha}{2}}\sigma_T \leq T - \tau \leq z_{\frac{\alpha}{2}}\sigma_T\right) &= 1 - \alpha \\ P\left(T - z_{\frac{\alpha}{2}}\sigma_T \leq \tau \leq T + z_{\frac{\alpha}{2}}\sigma_T\right) &= 1 - \alpha \end{aligned}$$

The quantity $z_{\frac{\alpha}{2}}\sigma_T$ is called margin of error.

Case 2 σ_T^2 is a unknown parameter. In this case, we standardize T using an estimator of σ_T^2, i.e. the sample variance s_T^2 and obtain:

$$P\left(\frac{a-\tau}{s_T} \leq Z \leq \frac{b-\tau}{s_T}\right) = 1 - \alpha$$

or equivalently,

$$P(-t_{n-1,\frac{\alpha}{2}} \leq t_{n-1} \leq t_{n-1,\frac{\alpha}{2}}) = 1 - \alpha$$

where t_{n-1} is now a t-student random variable with $n-1$ degrees of freedom, and $t_{n-1,\frac{\alpha}{2}}$ are critical values of t associated to α for a two-tail confidence interval. $t_{n-1,\frac{\alpha}{2}}$ have to be computed using now the statistical tables for a t-student random variable. Therefore in this case, we get:

$$\begin{aligned} P(-t_{n-1,\frac{\alpha}{2}} \leq t_{n-1} \leq t_{n-1,\frac{\alpha}{2}}) &= 1 - \alpha \\ P\left(T - t_{n-1,\frac{\alpha}{2}}s_T \leq \tau \leq T + t_{n-1,\frac{\alpha}{2}} \leq t_{n-1}s_T\right) &= 1 - \alpha \end{aligned}$$

Property. As n increases, t_{n-1} converges to z. This means that for values $n > 30$, one can use values from the statistical table for z.

5.3. Exercises

Exercise 1 Suppose you are a drink producer and your drinks are sold in 250ml bottles. Due to some imperfections in the production process, the actual volume in each bottle varies. The production process has a mean of 250ml and a standard deviation of 20ml. A consumer watchdog takes a sample of 30 of your bottles and measures their content precisely. You will get bad press if the sample mean is below 245ml. What is the probability that you will get bad press?

Exercise 2 A random sample of 16 bags of a chemical were tested to estimate the mean impurity content. It is known that the impurity content is distributed normally. The sample mean impurity content was 20.4 grams, and the sample standard deviation was 6.4 grams. Find the 95% confidence interval for the population mean.

Exercise 3 An auditor takes a random sample of 400 invoices relating to the activities of a company in a particular year. The sample mean of the invoices is 250GBP and the sample standard deviation is 64GBP. Find a 95% confidence interval for the population mean of the company invoices in the same year.