1.017/1.010 Class 14
Estimation

Estimating Distributional Properties

We can use random samples to estimatethe unknown properties of random variable $\boldsymbol{x}$. These may be:

1. Distributional **parameters**, such as the lower and upper limits of a uniform distribution (a and b)
2. Other distributional properties such as the **mean**, **variance**, **90th percentile** value, etc.

Parametric statistics assumes that the form of the $\boldsymbol{x}$ distribution of $\boldsymbol{x}$ is known. **Nonparametric statistics** makes no assumptions about distribution of $\boldsymbol{x}$.

Estimators

An **Estimator** (or a **statistic**) $\hat{\boldsymbol{a}}(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N)$ is a function used to derive an **estimate** $\hat{\boldsymbol{a}}$ of the unknown distributional property a from a random sample $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$

$$\hat{\boldsymbol{a}} = \hat{\boldsymbol{a}}(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N)$$

Example:

Suppose we want to estimate from the random sample $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$ the mean of $F_x(x)$. A possible choice for the estimator is the sample mean $\boldsymbol{m}_x$. Then:

$$a = E[\boldsymbol{x}]$$

$$\hat{\boldsymbol{a}} = \hat{a}(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N) = \boldsymbol{m}_x = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i$$

How do we know if this is a “good” estimator?

Properties of Good Estimates

We would like the estimate $\hat{\boldsymbol{a}}$ to be as “close” as possible to the true value a.

Since $\hat{\boldsymbol{a}}$ is derived from a random sample (a set of random variables) it is also a random variable, defined by the probability distribution $F_{\hat{a}}(\hat{a})$. This distribution describes how much $\hat{\boldsymbol{a}}$ might deviate from a.

We can sometimes derive $F_{\hat{a}}(\hat{a})$ or we may use the Central Limit theorem to justify the assumption that $F_{\hat{a}}(\hat{a})$ is a **normal distribution** (if a is derived from a large random sample).

If $F_{\hat{a}}(\hat{a})$ is normal, it is completely defined by the mean $E[\hat{\boldsymbol{a}}]$ and variance $Var[\hat{\boldsymbol{a}}]$ of $\hat{\boldsymbol{a}}$.

Even if $F_{\hat{a}}(\hat{a})$ is not normal we can use the mean and variance of $\hat{\boldsymbol{a}}$ to measure the quality of the estimate. "Good" estimates should be **unbiased** and **consistent**:

1.**Unbiased**
The expected value of the estimate is the true property value:

$$E[\hat{\boldsymbol{a}}] = a$$

2. **Consistent**:
The variability of the estimate goes to zero as the sample size increases:

$$Var[\hat{\boldsymbol{a}}] \to 0 \quad as \quad N \to \infty$$

If the estimate is unbiased and consistent it will converge to the true value as the sample size increases.

In general, the mean and variance of an estimate are most readily estimated with **stochastic simulation** (see exercise below). In some cases they can be derived directly from the estimator function $\hat{a}$.

Example – Properties of the Sample mean:

Reconsider the sample mean $\boldsymbol{m}_x$ used to estimate the mean $E[\boldsymbol{x}]$ of $F_x(x)$:

$$\hat{\boldsymbol{a}} = \hat{a}(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N) = \boldsymbol{m}_x = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i$$

The mean and variance of this estimate are:

$$E[\hat{\boldsymbol{a}}] = E[\boldsymbol{m}_x] = E\left[\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i\right] = E[\boldsymbol{x}] = a$$

$$Var[\hat{\boldsymbol{a}}] = Var[\boldsymbol{m}_x] = Var\left[\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i\right] = \frac{1}{N}Var[\boldsymbol{x}]$$

So the sample mean is an **unbiased** and **consistent** estimator of the population mean.

For large samples the Central Limit Theorem implies that $\boldsymbol{m}_x$ is normally distributed. The distribution of $\boldsymbol{m}_x$ is centered on $E[\boldsymbol{x}]$ and gradually converges to $E[\boldsymbol{x}]$ as N increases.

Exercise: Comparing Alternative Estimates of Population Properties

Suppose that you have a sample of 10 observations of a random variable $\boldsymbol{x}$ that you believe to be exponentially distributed. Your objective is to estimate the 90 percentile value x_{90} of this variable. This value of the solution of the equation $F_x(x_{90}) = 0.9$ [i.e. $F_x^{-1}(0.9) = x_{90}$]

Propose at least two different methods for estimating x_{90} from the 10 observations.

Compare the performance of these alternative estimators with a stochastic simulation that performs the following steps:

1. Generate many (e.g. 1000) replicates, each consisting of 10 observations drawn from an exponential distribution. Select your own value of the distributional parameter $a = E[x]$.
2. For each replicate derive an estimate $\hat{x}_{90}$ of x_{90} from each of your two proposed estimators.
3. For each estimator compute the sample mean and variance of $\hat{x}_{90}$ over all replicates. Also, use the replicates to construct a histogram and a CDF plot (using MATLAB's `normplot` function) for each $\hat{x}_{90}$.
4. Determine whether your estimators are unbiased and consistent (check consistency by plotting the rerunning your simulation for a much larger number of observations).

Which of your estimators is better? Explain your reasoning.

Some relevant MATLAB functions: `exprnd, hist, normplot, mean,var`