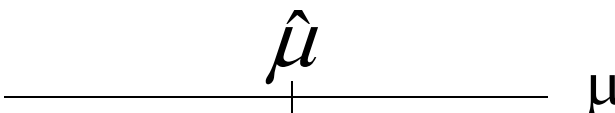


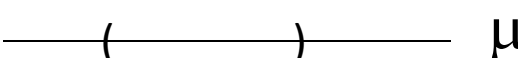
LECTURE 9

POINT ESTIMATION METHODS

STATISTICAL INFERENCE

- Determining certain unknown properties of a probability distribution on the basis of a sample (usually, a r.s.) obtained from that distribution

Point Estimation: $\hat{\mu} = 5$ 

Interval Estimation: $3 \leq \mu \leq 8$ 

Hypothesis Testing:
 $H_0 : \mu = 5$
 $H_1 : \mu \neq 5$

- **Parameter Space** (Ω or Θ): The set of all possible values of an unknown parameter, θ ; $\theta \in \Omega$.
- **Statistic**: A function of rvs (usually a sample rvs in an estimation) which does not contain any unknown parameters.

$$\bar{X}, S^2, etc$$

- **Estimator** of an unknown parameter θ : A statistic $\hat{\theta}$ used for estimating θ .

$$\hat{\theta} : estimator = U(X_1, X_2, \dots, X_n)$$

$$\bar{X} : Estimator$$



An observed value

$\bar{x} : Estimate$: A particular value of an estimator

POINT ESTIMATION

- We have a sample $\mathbf{x} = (x_1, \dots, x_n)$ from a population
- The population contains an unknown parameter θ
- The functional forms of the distributional functions may be known or unknown, but they depend on the unknown θ .
- Denote generally by $f(x; \theta)$ the probability density or mass function of the distribution
- A *point estimate* of θ is a function of the sample values

$$\hat{\theta} = \hat{\theta}(x_1, \dots, x_n) = \hat{\theta}(\mathbf{x})$$

such that its values should be close to the unknown θ .

STANDARD POINT ESTIMATES

- The sample mean \bar{x} is a point estimate of the population mean μ

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \hat{\mu}(x_1, \dots, x_n)$$

- The sample variance s^2 is a point estimate of the population variance σ^2

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \hat{\sigma}^2(x_1, \dots, x_n)$$

- The sample proportion p of a specific event (a specific value or range of values) is a point estimate of the corresponding population proportion π

$$p = \frac{\#\{x_i : \text{event is satisfied}\}}{n} = \hat{\pi}(x_1, \dots, x_n)$$

POINT ESTIMATION METHODS

- It is often the case that we are interested in finding values of some parameters of the system. Then we design an experiment and get some observations (x_1, \dots, x_n) . We want to use these observations and estimate the parameters of the system.
- The result of the estimation is a function of observation $T(x_1, \dots, x_n)$. A function of the observations is called statistic. It is a random variable and in many cases we want to find its distribution also.
- Maximum Likelihood Method and Method of Moments are most popular techniques to estimate parameters using observations or experimental data.

MAXIMUM LIKELIHOOD ESTIMATION (MLE) METHOD

MLE is a method of fitting statistical models to observed data. Let us assume that we know that our random sample points came from a population with the distribution with parameter(s) - θ . We do not know θ . If we would know it, then we could write the probability distribution of a single observation $f(x | \theta)$. Here $f(x | \theta)$ is the conditional distribution of the observed random variable if the parameter(s) would be known. If we observe n independent sample points from the same population then the joint conditional probability distribution of all observations can be written:

$$f(x_1, x_2, \dots, x_n | \theta) = \prod_{i=1}^n f(x_i | \theta)$$

We could write the product of the individual probability distributions because the observations are independent.

We could interpret $f(x_1, x_2, \dots, x_n | \theta)$ as the probability of observing given sample points if we would know the parameter θ . If we would vary the parameter(s) we would get different values for the probability f . Since f is the probability distribution, parameters are fixed and observation varies. For a given set of observations we define likelihood proportional to the conditional probability distribution.

$$L(x_1, x_2, \dots, x_n | \theta) \propto f(x_1, x_2, \dots, x_n | \theta)$$

When we talk about conditional probability distribution of the observations given parameter(s) then we assume that parameters are fixed and observations vary. When we talk about likelihood then observations are fixed parameters vary. That is the major difference between likelihood and conditional probability distribution. Sometimes to emphasize that parameters vary and observations are fixed, likelihood is written as:

$$L(\theta | x_1, x_2, \dots, x_n)$$

Principle of maximum likelihood states that the best parameters are those that maximize probability of observing current values of the observations. Maximum likelihood chooses parameters that satisfy:

$$L(x_1, x_2, \dots, x_n | \hat{\theta}) \geq L(x_1, x_2, \dots, x_n | \theta)$$

Purpose of the maximum likelihood is to maximize the likelihood function and estimate parameters. If the derivatives of the likelihood function exist then it can be done using:

$$\frac{dL(x_1, x_2, \dots, x_n | \theta)}{d\theta} = 0$$

Solution of this equation will give possible values for maximum likelihood estimator.

Usually instead of likelihood its logarithm is maximized. Since log is strictly monotonically increasing function, derivative of the likelihood and derivative of the log of likelihood will have exactly same roots. If we use the fact that observations are independent then the joint probability distributions of all observations is equal to the product of the individual probabilities.

$$l(x_1, x_2, \dots, x_n | \theta) = \ln(L(x_1, x_2, \dots, x_n | \theta)) = \sum_{i=1}^n \ln(f(x_i | \theta))$$

Usually working with sums is easier than working with products.

Example

Let us assume that we carry out a Bernoulli experiment. Possible outcomes of the trials are success or failure. Probability of success is π and probability of failure is $1 - \pi$. We do not know the value of π . Let us assume we have n trials and k of them are successes and $n - k$ of them are failures. Values of random variables in our trials can be either 0 (failure) or 1 (success). Let us denote observations as $y = (y_1, y_2, \dots, y_n)$. Probability of the observation y_i at the i th trial is:

Since individual trials are independent we can write for n trials:

$$f(y_i | \pi) = \pi^{y_i} (1 - \pi)^{1 - y_i} \quad L(y_1, y_2, \dots, y_n | \pi) = \prod_{i=1}^n \pi^{y_i} (1 - \pi)^{1 - y_i}$$

log of this function is:

$$l(y_1, y_2, \dots, y_n | \pi) = \sum_{i=1}^n (y_i \ln(\pi) + (1 - y_i) \ln(1 - \pi))$$

Equating the first derivative of the likelihood w.r.t unknown parameter to zero we get:

$$\frac{dl}{d\pi} = \frac{\sum_{i=1}^n y_i}{\pi} - \frac{\sum_{i=1}^n (1 - y_i)}{1 - \pi} = 0 \Leftrightarrow \hat{\pi} = \frac{\sum_{i=1}^n y_i}{n} = \frac{k}{n}$$

The ML estimator for the parameter is equal to the fraction of successes.

Example

Let us assume that the sample points came from the population with normal distribution with unknown mean and variance. Let us assume that we have n observations, $y=(y_1, y_2, \dots, y_n)$. We want to estimate the population mean and variance. Then log likelihood function will have the form:

$$l(y_1, y_2, \dots, y_n | \mu, \sigma^2) = \sum_{i=1}^n \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mu)^2}{2\sigma^2}}\right) = -n \ln(\sqrt{2\pi}) - \frac{n}{2} \ln(\sigma^2) - \sum_{i=1}^n \frac{(y_i - \mu)^2}{2\sigma^2}$$

If we get derivative of this function w.r.t mean value and variance then we can write:

$$\frac{dl}{d\mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu) = 0 \Leftrightarrow \hat{\mu} = \frac{\sum_{i=1}^n y_i}{n} \Leftrightarrow \hat{\mu} = \bar{y}$$

$$\frac{dl}{d(\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{n}{2\sigma^4} \sum_{i=1}^n (y_i - \mu)^2 = 0 \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mu})^2$$

ADVANTAGES OF MLE

- Often yields good estimates, especially for large sample size.
- Asymptotic distribution of MLE is Normal.
- Most widely used estimation technique.
- Usually they are consistent estimators. [will define consistency later]

DISADVANTAGES OF MLE

- Requires that the pdf or pmf is known except the value of parameters.
- MLE may not exist or may not be unique.
- MLE may not be obtained explicitly (numerical or search methods may be required.). It is sensitive to the choice of starting values when using numerical estimation.
- MLEs can be heavily biased for small samples.

METHOD OF MOMENTS ESTIMATION (MME)

- Let X_1, X_2, \dots, X_n be a r.s. from a population with pmf or pdf $f(x; \theta_1, \theta_2, \dots, \theta_k)$. The MMEs are found by equating the first k population moments to corresponding sample moments and solving the resulting system of equations.

Population Moments

$$\mu_k = E[X^k]$$

Sample Moments

$$M_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

Example

- Suppose you have 10 data about x
0.3, 4, 5, 1, 1.3, 6.5, 0.85, 2.5, 4.56, 3.14

$$E(X) = \frac{1}{n} \sum_{i=1}^n x_i$$

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

- After calculation, mean = 2.915, var = 4.2981

Example

- Suppose we want to fit with uniform,

Uniform $f_X(x) = \frac{1}{b-a} \quad a < x < b$ a, b

$$E(X) = (a + b)/2$$

$$\text{Var}(X) = \frac{1}{12} (b - a)^2$$

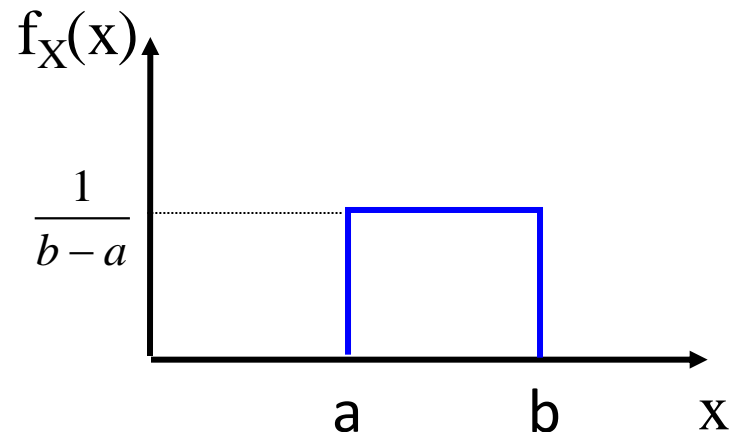
- Now

$$E(X) = \frac{1}{2} (a + b) = 2.915$$

$$\text{Var}(X) = \frac{1}{12} (b - a)^2 = 4.2981$$

- Solving,

$$b = 6.5142, a = -0.684$$



METHOD OF MOMENTS ESTIMATION (MME)

$$\mu_1 = M_1$$

$$\mu_2 = M_2$$

$$\mu_3 = M_3$$

so on...

$$E(X) = \frac{1}{n} \sum_{i=1}^n X_i \quad E(X^2) = \frac{1}{n} \sum_{i=1}^n X_i^2 \quad E(X^3) = \frac{1}{n} \sum_{i=1}^n X_i^3$$

Continue this until there are enough equations to solve for the unknown parameters.

DRAWBACKS OF MMES

- Although sometimes parameters are positive valued, MMES can be negative.
- If moments does not exist, we cannot find MMES.