1 Point Estimation

1.1 Statistical Inference

- The process of making educated guess and conclusions regarding a population, using a sample of that population is called **Statistical Inference**.
- Two important problems in statistical inference are **estimation of parameters** and **tests of hypothesis**
- Estimation can be of the form of **point estimation** and **interval estimation**.

1.2 Point Estimation

Main Task

- Assume that some characteristic of the elements in a population can be represented by a random variable X.
- Assume that $X_1, X_2, \ldots, X_n$ is a random sample from a density $f(x, \theta)$, where the form of the density is known but the parameter θ is unknown.
- The objective is to construct good estimators for θ or its function $\tau(\theta)$ on the basis of the observed sample values $x_1, x_2, \ldots, x_n$ of a random sample $X_1, X_2, \ldots, X_n$ from $f(x, \theta)$.

Definition: Statistic

Suppose $X_1, X_2, \ldots, X_n$ be n observable random variables. Then, a known function $T = g(X_1, X_2, \ldots, X_n)$ of observable random variables $X_1, X_2, \ldots, X_n$ is called a **statistic**. A statistic is always a random variable.

Definition: Estimator

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a density $f(x, \theta)$ and it is desired to estimate θ. Suppose $T = g(X_1, X_2, \ldots, X_n)$ is a *statistic* that can be used to determine and approximate value for θ. Then T is called an **estimator** for θ. An estimator is always a random variable.

Definition: Estimate

Suppose $T = g(X_1, X_2, \ldots, X_n)$ be an estimator for θ. Suppose that $x_1, x_2, \ldots, x_n$ is a set of observed values of the random variable $X_1, X_2, \ldots, X_n$. Then *the value* $t = g(x_1, x_2, \ldots, x_n)$ obtained by substituting the observed values in the estimator is called an **estimate** for θ.

- Therefore the **estimator** stands for the function of the sample, and the word **estimate** stands for the realized value of that function.
- *Notation:* An estimator of θ is denoted by $\hat{\theta}$. An estimate of θ is also denoted by $\hat{\theta}$. The difference between the two should be understood based on the context.

Parameter	Estimator: Using random sample $(X_1, X_2, \ldots, X_n)$	Estimate 1: Using observed sample $(1, 4, 2, 3, 4)$	Estimate 2: Using observed sample $(4, 2, 2, 6, 3)$
μ	$\hat{\mu} = \bar{X}$	$\hat{\mu} =$	$\hat{\mu} =$
σ^2	$\hat{\sigma^2} = S^2$	$\hat{\sigma^2} =$	$\hat{\sigma^2} =$

1.2.1 Methods of finding point estimators

- In some cases there will be an obvious or natural candidate for a point estimator of a particular parameter.
- For example, the sample mean is a good point estimator of the population mean
- However, in more complicated models we need a methodical way of estimating parameters.
- There are different methods of finding point estimators
  - Method of Moments
  - Maximum Likelihood Estimators (MLE)
  - Method of Least Squares
  - Bayes Estimators
  - The EM Algorithm
- However, these techniques do not carry any guarantees with them.
- The point estimators that they yeild still must be evaluated before their worth is established

1.2.1.1 Method of Moments

- Let $X_1, X_2, ... X_n$ be a random sample from a population with pdf or pmf $f(x;\theta)$, where $\theta = (\theta_1, \theta_2, ..., \theta_k)$ and $k \geq 1$.
- Sample moments m' (Sample moments m about 0) and population moments μ' are defined as follows

Sample moment	Population moment
$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\mu'_1 = E(X)$
$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$	$\mu'_2 = E(X^2)$
...	...
$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$	$\mu'_k = E(X^k)$

Each μ'_j is a function θ, i.e. $\mu'_j = \mu'_j(\theta_1, \theta_2, ..., \theta_k)$ for $j = 1, 2, ..., k$.

Method of Moments Estimators (MME)

We first equate the first k sample moments to the corresponding k population moments,

$$m_1 \approx \mu'_1,$$
$$m'_2 \approx \mu'_2,$$
$$...$$

$$m_k \approx \mu'_k,$$

Then we solve the resulting systems of simultaneous equations for $\hat{\theta_1}, \hat{\theta_2}, \ldots, \hat{\theta_k}$

Remarks on Method of Moments Estimators

- Very easy to compute
- Always give an estimator to start with
- Generally consistent (Since sample moments are consistent for population moments)
- Not necessarily the best or most efficient estimators

1.2.1.2 Maximum Likelihood Estimators (MLE)

Example

A company receives a certain number of orders per day, and this count appears to follow a Poisson distribution with a parameter λ.

The recorded number of orders over 10 randomly chosen days are: $12, 14, 15, 12, 13, 10, 11, 15, 10, 6$

Deriving the Joint Probability Expression

We can express the probability of observing this particular set of order counts, i.e.,$P(X_1 = 12,\ X_2 = 14, \ldots, X_{10} = 6)$ as a function of λ.

$$\text{Joint probability of the data} = P(X_1 = 12,\ X_2 = 14, \ldots, X_{10} = 6)$$
$$= \frac{e^{-\lambda}\lambda^{12}}{12!}\frac{e^{-\lambda}\lambda^{14}}{14!}\ldots\frac{e^{-\lambda}\lambda^{6}}{6!}$$

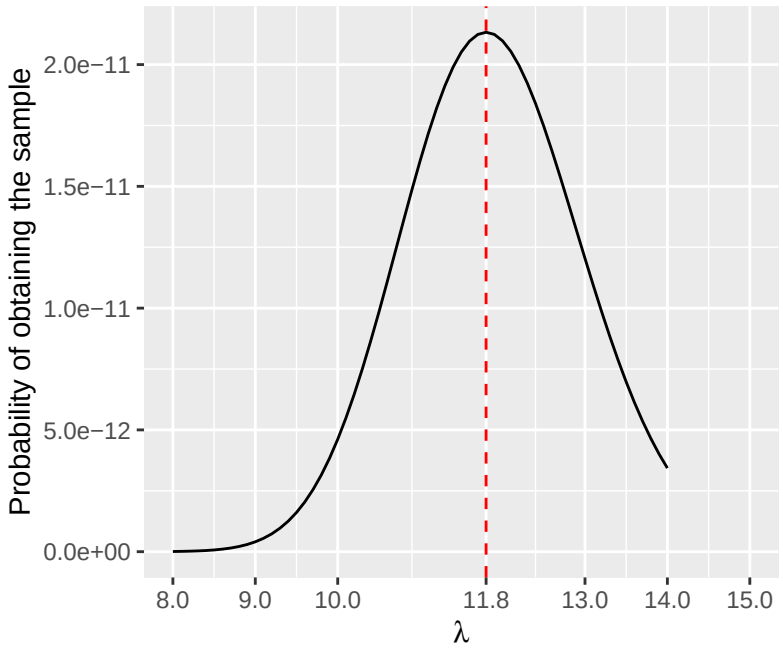

Figure 1.1: Probability of the sample is maximum when $\lambda = 11.8$

- By treating this probability of obtaining the sample as a function of λ, we obtain the **likelihood function**, which represents how likely different values of λ are given the observed data.
- The graphical representation indicates that the likelihood function reaches its maximum when $\lambda = 11.8$.
- Since these data have already occurred, it is very likely that the data have arisen from a Poisson distribution with $\lambda = 11.8$.
- This value is referred to as the **maximum likelihood estimate (MLE)** of λ.

Theory

- In order to define maximum-likelihood estimators, we shall first define the likelihood function.

Definition: Likelihood function

Let $x_1, x_2, \ldots, x_n$ be a set of observations of random variables $X_1, X_2, \ldots, X_n$ with the joint density of n random variables, say $f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n;\ \theta)$. This joint density function, which is considered to be a function of θ is called the **likelihood function of** θ for the set of observations (sample) $x_1, x_2, \ldots, x_n$.

In particular, if $x_1, x_2, \ldots, x_n$ is a random sample from the density $f(x;\theta)$, then the likelihood function is $f(x_1;\theta)f(x_2;\theta)\ldots f(x_n;\theta)$.

Notation

We use the notation $L(\theta;\ x_1, x_2, \ldots, x_n)$ for the likelihood function, in order to remind ourselves to think of the likelihood function as a function of θ.

- Likelihood function is seen as a function of θ rather than x
- Likelihood can be viewed as the degree of plausibility.
- An estimate of θ may be obtained by choosing the most plausible value, i.e., where the likelihood function is maximized.

Definition: Maximum Likelihood Estimator

Let $L(\theta) = L(\theta;\ x_1, x_2, \ldots, x_n)$ be the likelihood function of θ for the sample $x_1, x_2, \ldots, x_n$. Suppose $L(\theta)$ has its maximum when $\theta = \hat{\theta}$.

Then $\hat{\theta}$ is called the **Maximum likelihood estimate of** θ.

The corresponding estimator is called the **Maximum likelihood estimator of** θ.

- Many likelihood functions satisfy regularity conditions; so the maximum likelihood estimator is the solution of the equation

$$\frac{dL(\theta)}{d\theta} = 0$$

Log-likelihood function

Let

$$l(\theta) = ln[L(\theta)].$$

Then, $l(\theta)$ is called the **log-likelihood function**.

- Both $L(\theta)$ and $l(\theta)$ have their maxima at the same value of θ.
- It is sometimes easier to find the maximum of the logarithm of the likelihood and thereby simplify the calculations in finding the maximum likelihood estimate.

Invariance Property of MLE's

If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

1.2.2 Methods of evaluating estimators

- We discussed several methods of obtaining point estimators.
- It is possible that different methods of finding estimators will lead to same estimator or different estimators.
- In this section we discuss certain properties, which an estimator may or may not posses, that will guide us in deciding whether one estimator is better than another.

1.2.2.1 Unbiasedness

Definition: Unbiased estimator

An estimator $\hat{\theta}$ $(= t(X_1, X_2, ..., X_n))$ is defined to be an **unbiased estimator** of θ if and only if $E(\hat{\theta}) = \theta$

- The difference $E(\hat{\theta}) - \theta$ is called as the bias of $\hat{\theta}$ and denoted by $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$
- An estimator whose bias is equal to 0 is called **unbiased**.
- This means that, on average, the estimator correctly estimates θ, though it may fluctuate from sample to sample.

Asymptotic unbiasedness

An estimator $\hat{\theta}$ of a parameter θ is said to be *asymptotically unbiased* if its bias

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

satisfies

$$\lim_{n\to\infty} Bias(\hat{\theta}) = 0.$$

This means that as the sample size n increases, the bias of the estimator vanishes, making it approximately unbiased for large samples.

1.2.2.2 Consistency

Mean-Squared Error

- The *mean-squared error* is a measure of goodness or closeness of an estimator to the target.

Definition: Mean-squared Error (MSE)

The **mean-squared error** of an estimator $\hat{\theta}$ of θ is defined as $MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$

- The MSE measures the average squared difference between $\hat{\theta}$ and θ.
- The MSE is a function of θ and has the interpretation

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + \left[Bias(\hat{\theta})\right]^2$$

- Therefore the MSE incorporates two components, one measuring the variability of the estimator (*precision*) and the other measuring its bias (*accuracy*).
- Small value of MSE implies small combined variance and bias.
- If $\hat{\theta}$ is unbiased, then $MSE(\hat{\theta}) = Var(\hat{\theta})$
- The positive square root of MSE is known as the *root mean squared error* $RMSE(\hat{\theta}) = \sqrt{MSE(\hat{\theta})}$

Consistency

- Estimator $\hat{\theta}$ is said to be consistent for θ if $MSE(\hat{\theta})$ approaches zero as the sample size n approaches ∞.

$$lim_{n\to\infty} E\left[(\hat{\theta} - \theta)^2\right] = 0$$

- Mean-squared error consistency implies that the bias and the variance both approach to zero as n approaches ∞.

1.2.2.3 Sufficiency – Using all available information efficiently

Why do we need sufficiency?

An estimator should make full use of the **available data**. If we can summarize all relevant information into a single function of the data, we can reduce the problem to estimating from that summary.

Definition: Sufficient Statistic

A statistic $T(X)$ is a *sufficient Statistic for a parameter* θ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on θ.

In other words, once you know $T(X)$, **the sample data doesn't give any extra information about** θ.

Factorization Theorem (Easy way to check Sufficiency)

A statistic $T(X)$ is **sufficient** for θ if the likelihood function can be factored as:

$$L(\theta|X) = g(T(X), \theta)h(X)$$

where $g(T(X), \theta)$ depends on θ **only through** $T(X)$

Example

For a normal distribution $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$:, the statistic:

$$T = \frac{1}{n}\sum X_i$$

is sufficient for μ.

This means that all the useful information about μ is contained in the sample mean T.

These ideas are fundamental in statistics and machine learning for choosing the best estimator.