Statistical Inference I
A Note on Interval Estimation

Kiranmoy Chatterjee*

Let us consider a population which is characterized by some parameters such as mean(*for* location), variance(*for* scale), skewness or kurtosis(*for* shape). In statistical analysis, one of the major aim is to make inference about the population that means about its unknown parameter(s). In *point estimation* chapter, we estimate a parameter, say θ, by a specific value calculated from the given sample data. Now in *Interval Estimation* chapter, we will give an interval, based on the given sample data, with a strong belief that unknown value of the parameter θ lies in that interval.

Let $\underline{X}$ =$(X_1, X_2, \ldots, X_n)$ be a set of random variables of size n drawn from a population with density f_θ and $\underline{x}$ =$(x_1, x_2, \ldots, x_n)$ is a realization of $\underline{X}$. Consider, $T_1(\underline{X})$ and $T_2(\underline{X})$ are two statistics (based on $\underline{X}$) satisfying $T_1(\underline{X}) \leq T_2(\underline{X})$ for all $\underline{X} \in \mathcal{X}$. Suppose that on seeing the data $\underline{X} = \underline{x}$, we make the inference $T_1(\underline{x}) \leq \theta \leq T_2(\underline{x})$. Now, this calculated interval has fixed endpoints for fixed data points $\underline{x}$, where θ might be in between (or not). Thus this event has probability either 0 or 1. Since, θ is fixed and $T_1(\underline{X})$ and $T_2(\underline{X})$ are two random variables (due to randomness of $\underline{X}$), so $T_1(\underline{x}) \leq \theta \leq T_2(\underline{x})$ is just an random event. As much as the $Prob_\theta(T_1(\underline{X}) \leq \theta \leq T_2(\underline{X}))$ is being higher, our confidence on the inference that $\theta \in [T_1(\underline{X}), T_2(\underline{X})]$ increases. In practice,

$$Prob_\theta(T_1(\underline{X}) \leq \theta \leq T_2(\underline{X})) = 1 - \alpha,$$

where α does not depend on θ and in practice, α has to be set by the experimenter. Then the random interval $[T_1(\underline{X}), T_2(\underline{X})]$ is called a $100(1-\alpha)\%$ *confidence interval* for θ. However, if we repeat the procedure of sampling and compute the confidence interval $[T_1(\underline{x}), T_2(\underline{x})]$ based on sample $\underline{x}$ each time, then our confidence interval will contain the true θ $100(1-\alpha)\%$ of the time. Typically α is 0.05 or 0.01, so that the probability the interval contains θ is close to 1. Then by giving up precision in our assertion about the value of θ, we gain confidence that our assertion is correct.

Remark: One may choose $[T_1(\underline{X}), T_2(\underline{X})]$ such that α is exactly 1, but that interval will be useless as it would be too wide.

Definition: Interval Estimator

Consider a pair of functions of random variables, $T_1(\underline{X})$ and $T_2(\underline{X})$ that satisfy $T_1(\underline{X}) \leq T_2(\underline{X})$ for all $\underline{X} \in \mathcal{X}$. If $\underline{X} = \underline{x}$ is observed, the inference that $T_1(\underline{x}) \leq \theta \leq T_2(\underline{x})$ is made on a real-valued parameter θ. This interval $[T_1(\underline{x}), T_2(\underline{x})]$ is called *interval estimate* of θ. The associated random interval $[T_1(\underline{X}), T_2(\underline{X})]$ is called an *interval estimator* of θ.

*Department of Statistics, Bidhannagar College, Salt Lake, Kolkata-700064; E-mail: *kiranmoy07@gmail.com*
Target Students: B.Sc.(Hons.) in Statistics

Definition: Confidence Interval & Confidence Coefficient
For an *interval estimator* $[T_1(\underline{X}), T_2(\underline{X})]$ of a real-valued parameter θ, a coefficient $\gamma \in (0,1)$ is considered such that

$$Prob_\theta(T_1(\underline{X}) \leq \theta \leq T_2(\underline{X})) \geq \gamma.$$

Here the γ said to be *confidence coefficient* of $[T_1(\underline{X}), T_2(\underline{X})]$ and $[T_1(\underline{X}), T_2(\underline{X})]$ is said to be $100\gamma\%$ *confidence interval.*

Interpretation of Confidence Interval: $[T_1(\underline{X}), T_2(\underline{X})]$ is a $100\gamma\%$ confidence interval. This means "probability that the interval $[T_1(\underline{X}), T_2(\underline{X})]$ contains the true θ is at least γ".

OR

$[T_1(\underline{X}), T_2(\underline{X})]$ is a $100\gamma\%$ confidence interval. This means "if we repeat the sampling strategy 100 times and calculate the interval based on $T_1(\underline{X})$ and $T_2(\underline{X})$ each time, then θ will lie in the interval $[T_1(\underline{X}), T_2(\underline{X})]$ at least $100\gamma\%$ times.

Example: If $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$ independently, with μ and σ^2 both are unknown and we are interested in finding a confidence interval for μ at $100(1-\alpha)\%$ confidence. Now,

$$\frac{\sqrt{n}(\overline{X}-\mu)}{\sqrt{S_{xx}/(n-1)}} \sim t_{n-1},$$

where t_{n-1} denotes the *Students' t*-distribution with $(n-1)$ degrees of freedom and $\overline{X} = n^{-1}\sum_{i=1}^n X_i$, $S_{xx} = \sum_{i=1}^n (X_i - \overline{X})^2$. So if a and b are such that

$$Prob(a \leq \frac{\sqrt{n}(\overline{X}-\mu)}{\sqrt{S_{xx}/(n-1)}} \leq b) = 1-\alpha,$$

which can be rewritten as

$$Prob(\overline{X} - b\sqrt{S_{xx}/n(n-1)}\mu) \leq \mu \leq \overline{X} - a\sqrt{S_{xx}/n(n-1)}) = 1-\alpha.$$

Again the choice of a and b is not unique, but it is natural to try to make the length of the confidence interval as small as possible. The symmetry of the t-distribution implies that we should choose a and b symmetrically about 0. In practice, $a = -t_{\alpha/2;(n-1)}$ and $b = t_{\alpha/2;(n-1)}$.

References

[1] Casella, G. and Berger, R. L. (2002). *Statistical Inference*, 2nd Edition. Cengage Learning India Private Ltd., New Delhi, India.

[2] Lecture Notes on Statistics by Prof. Richard Weber, Statistical Laboratory, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WB, UK.

[3] Gun, A. M., Gupta, M. K. and Dasgupta, B. (2008). *Fundamental of Statistics, Volume One*, 8th Edition. The World Press Private Ltd., Kolkata, India.