Statistical Inference (MA862)

Lecture Slides

Topic 2: Point Estimation

Statistical Inference

- In a typical statistical problem, our aim is to find information regarding numerical characteristic(s) of a collection of items/persons/products. This collection is called population.
- Suppose that we want to know the average height of Indian citizens.
  - Measure heights of all citizens
  - Find the average.
- However, it is a very costly (in terms of money and time) procedure.

Sample

- One approach to address these issues is to take a subset of the population based on which we try to find out the value of the numerical characteristic.
- Obviously, it will not be exact, and hence, it is an estimate.
- This subset is called a sample.
- The sample must be chosen such that it is a good representative of the population.
- There are different ways of selecting sample from a population.
- We will consider one such sample which is called *random sample*.

Modelling a Statistical Problem

- Different elements of a population may have different values of the numerical characteristic under study.
- Therefore, we will model it with a random variable and the uncertainty using a probability distribution.
- Let X be a random variable (either discrete or continuous random variable), which denotes the numerical characteristic under consideration.
- Our job is to find the probability distribution of X.
- Note that once the probability distribution is determined, the numerical summary (for example, mean, variance, median, etc.) of the distribution can be found.

Parametric and Non-parametric Inference

- There are two possibilities:
  - X has a CDF F with known functional form except perhaps some parameters. Here our aim is to (educated) guess value of the parameters. For example, in some case we may have $X \sim N(\mu, \sigma^2)$, where the functional form of the PDF is known, but the parameters μ and/or σ^2 may be unknown. In this case, we need to find value of the unknown parameters based on a sample. This is known as parametric inference.
  - X has a CDF F who's functional form is unknown. This is known as non-parametric inference.

Random Sample

Definition 1: The random variables $X_1, X_2, \ldots, X_n$ is said to be a random sample (RS) of size n from the population F if $X_1, X_2, \ldots, X_n$ are *i.i.d.* random variables with marginal CDF F. If F has a PMF/PDF f, we will write that $X_1, \ldots, X_n$ is a RS from the PMF/PDF f.

- The JCDF of a RS $X_1, \ldots, X_n$ from CDF F is

$$F(x_1, \ldots, x_n) = \prod_{i=1}^{n} F(x_i).$$

- The JPMF/JPDF of a RS $X_1, \ldots, X_n$ from PMF/PDF f is

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i).$$

Random Sample

- In the standard framework of parametric inference, we start with a data, say $(x_1, x_2, \ldots, x_n)$. Each x_i is an observation on the numerical characteristic under study.
- There are n observations and n is fixed, pre-assigned, and known positive integer.
- Our job is to identify (based on a data) the CDF (or equivalently PMF/PDF) of the RV X, which denote the numerical characteristic in the population.

Random Sample

- In practice, we have a data.
- How to model a data using RS?
- Notice that the first observation in the sample can be one of the member of the population.
- Thus, a particular observation is one of the realizations from the whole population.
- Therefore, it can be seen as a realization of a random variable X.
- Let X_i denote the ith observation for $i = 1, 2, \ldots, n$, where n is the sample size.
- Then, a meaningful assumption is that each X_i has same CDF F, as X_i is a copy of X.
- Now, if we can ensure that the observation are taken such a way that the value of one does not effect the others, then we can assume that $X_1, X_2, \ldots, X_n$ are independent.

Parametric Inference

- The functional form of the CDF/PMF/PDF of RV X is known.
- However, the CDF/PMF/PDF involves unknown but fixed real or vector valued parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$.
- If the value of $\boldsymbol{\theta}$ is known, the stochastic properties of the numerical characteristic is completely known.
- Therefore, our aim is to find the value of $\boldsymbol{\theta}$ or a function of $\boldsymbol{\theta}$.
- We assume that the possible values of $\boldsymbol{\theta}$ belong to a set Θ, which is called parametric space.
- $\boldsymbol{\theta}$ is a subset of $\mathbb{R}^n$.
- Here, $\boldsymbol{\theta}$ is an indexing or a labelling parameter. We say that $\boldsymbol{\theta}$ is an indexing parameter or a labelling parameter if the CDF/PMF/PDF is uniquely specified by $\boldsymbol{\theta}$, i.e., $F(x, \boldsymbol{\theta}_1) = F(x, \boldsymbol{\theta}_2)$ for all $x \in \mathbb{R}$ implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, where $F(\cdot, \boldsymbol{\theta})$ is the CDF of X.

Some Examples

Example 3:

- Suppose we want to find the probability of germination of seeds produced by a particular brand.
  - 100 seeds of a brand were planted one in each pot.
  - Let X_i equals one or zero according as the seed in the ith pot germinates or not.
  - The data consists of $(x_1, x_2, \ldots, x_{100})$, where each x_i is either one or zero.
  - The data is regarded as a realization of $(X_1, X_2, \ldots, X_{100})$, where the RVs are *i.i.d.* with $P(X_i = 1) = \theta = 1 - P(X_i = 0)$.
  - θ is the probability that a seed germinates.
  - The natural parametric space is $\Theta = [0, 1]$.
  - θ is an indexing parameter.

Some Examples

Example 4:

- Consider determination of gravitational constant g.
  - A standard way to estimate g is to use the pendulum experiment and use the formula

$$g = \frac{2\pi^2 l}{T^2},$$

    where l is the length of the pendulum and T is the time required for a fixed number of oscillations.
  - A variation is observed in the calculated values of g.
  - Let the repeated experiments are performed and the calculated values of g are $X_1, X_2, \ldots, X_n$.
  - Use the model $X_i = g + \epsilon_i$, where ϵ_i is the random error.
  - Assume $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
  - Then $X_i \overset{i.i.d.}{\sim} N(g, \sigma^2)$, and the parameter is $\boldsymbol{\theta} = \left(g, \sigma^2\right)$ with parametric space $\Theta = \mathbb{R} \times \mathbb{R}^+$.
  - $\boldsymbol{\theta}$ is an indexing parameter.

Some Examples

Example 5:

- Interested in estimating the average height of a large community of people.
    - Assume that $N(\mu, \sigma^2)$ is a plausible distribution.
    - As the average of heights of persons is always a positive real number, it is realistic to assume that $\mu > 0$.
    - Hence, a better choice of Θ is $\mathbb{R}^+ \times \mathbb{R}^+$.
    - Thus, we may need to choose the parametric space based on the background of the problem.

Some Examples

Example 6:

- Consider a series system with two components. A series system works if all its components work.
- Z: lifetimes of the first component.
- Y: lifetimes of the second component.
- $Z \sim Exp(\theta)$ and $Y \sim Exp(\lambda)$ (rates θ and λ)
- Y and Z are independent RVs.
- Z and Y are not observed.
- We observe $X = \min\{Z,\ Y\}$.
- $X \sim Exp(\theta + \lambda)$.
- $\alpha = \theta + \lambda$ is an indexing parameter.
- However, $(\theta,\ \lambda)$ is not an indexing parameter.

Statistic

Definition 2: Let $X_1, \ldots, X_n$ be a RS. Let $T(x_1, \ldots, x_n)$ be a real-valued function having domain that includes the sample space, χ^n, of $X_1, X_2, \ldots, X_n$. Then, the RV $\boldsymbol{Y} = T(X_1, \ldots, X_n)$ is called a statistic if it is not a function of unknown parameters.

Definition 3: In the context of estimation, a statistic is called a point estimator (or simply estimator). A realization of a point estimator is called an estimate.

Example 7: Let $X_1, \ldots, X_n$ be a RS from a $N(\mu, \sigma^2)$ distribution, where $\mu \in \mathbb{R}$ and $\sigma > 0$ are both unknown. Then $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ are examples of statistics. However, $\frac{\overline{X}-\mu}{\sigma}$ is not a statistic. Note that $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.