Statistical Inference (Introduction) ①

Statistical Inference Problem: We seek information about some numerical characteristic(s) of a collection of elements called population.

For reasons of time or cost it may not be possible to study each individual element of the population (which although is the best thing to do).

Goal to draw conclusions (or make inferences) about the unknown characteristics of the population on the basis of information on characteristic(s) of a suitable selected sample:

$\underline{X}$: a random sample (or random vector) describing the characteristics of a population under study

F: dist^n function (d.f.) of X.

Parametric Statistical Inference: Here the r.v.s $\underline{X}$ has a d.f. $F \equiv F_\theta$ with a known functional form (except perhaps for the unknown parameter θ, which may be vector).

Θ : set of all possible values of unknown parameter θ. Θ is called parameter space.

Basic Parametric Statistical Inference Problem :

To decide, on the basis of a suitably selected sample, which member or members of the family $\{F_\theta : \theta \in \Theta\}$ can represent the d.f. of X.

Nonparametric Statistical Inference : Here we know nothing about the d.f. F (expect perhaps that F is absolutely continuous or discrete)

Goal : To make inferences about unknown d.f. F.

In this course we only concentrate on parametric inference

Data collection : The statistician can observe n independent observations (say $x_1, \dots, x_n$) on r.v's X that describes the population under study.

Here each x_i can be regarded as the value assumed by a random variable X_i, $i = 1, 2, \dots, n$, where $X_1, \dots, X_n$ are independent r.v's with common d.f. F.

So the values of $(x_1, \dots, x_n)$ are the values assumed by $(X_1, \dots, X_n)$.

$X_1, X_2, \dots, X_n$: a sample of size n taken from a population with d.f. F.

$(x_1, \dots, x_n)$: realization of the sample $(X_1, \dots, X_n)$

<u>Sample Space</u> : The space of possible values of $(X_1, \dots, X_n)$ is called the sample space and denoted by $\mathcal{X}$.

<u>Random Sample</u> : Let X be a r.v. with d.f. F and let $X_1, \dots, X_n$ be a collection of independent and identically distributed (i.i.d.) rvs with common d.f. F. Then the collection $X_1, \dots, X_n$ is known as a random sample of size n from d.f. F, or the corresponding population.

<u>Statistics</u> : Let $T : \mathcal{X} \to \mathbb{R}^K$ be Borel fun. Then the r.v. $T(X_1, \dots, X_n)$ is called a (sample) statistics provided it is not a fun of any unknown parameters, i.e. T only depends on sample $(X_1, \dots, X_n)$.

<u>Example</u>: Let $X_1, \ldots, X_n$ be a random sample (r.s.) from $N(\mu, \sigma^2)$.

$$\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma > 0\} = \mathbb{R} \times (0, \infty) \text{ is unknown}$$

Then

$$T_1(\underset{\sim}{X}) = \sum (X_i - \bar{X})^2, \quad T_2 = \sum X_i \quad \text{are statistics}$$

But $T_3(\underset{\sim}{X}) = \sum (X_i - \mu)^2$ is not statistics.

In this statistical inference we study

Two main topics: (I) Point Estimation, Interval Estimation

(II) Testing of Hypothesis

<u>Point Estimation</u>:

<u>Estimator</u>: Any fun of the random sample which is used to estimate the unknown value of the given parametric fun $g(\theta)$ is called an estimator.

If $\underline{X} = (X_1, \ldots, X_n)$ is a random sample from a population with prob distⁿ F_θ, a funⁿ $T(\underline{X})$ used for estimating $g(\theta)$ is known as estimator.

Let $\underline{x} = (x_1, \ldots, x_n)$ be a realization of $\underline{X}$. Then $T(\underline{x})$ ~~$d(\underline{x})$~~ is called an estimate.