IC105: Probability and Statistics **2021-22-M**

Lecture 22: Statistical Inference

Instructor: Dr. Kuldeep Kumar Kataria *Scribe:*

22.1. Statistical Inference

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 althouh 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: distribution 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).

$\mathscr{H}$: set of all possible values of unknown parameter θ. $\mathscr{H}$ is called parameter space.

Basic Parametric Statistical Inference Problem:

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

Non Parametric Statistical Inference:

Here we know nothing about the d.f F (except perhaps that F is absolutely continous or discrete)

<u>Goal</u> : To make inference about unknown d.f F.

In this course we only concentrate on parametric inference.

$\underline{Data\ Collection}$:

The statistician can observe n independent observations (say $x_1, \ldots, x_n$) 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, $i = 1, 2, \ldots, n$, where $X_1, \ldots, X_n$ are independent r.v's with common d.f F.

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

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

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

Sample Space:

The space of possible values of $(X_1, \ldots, X_n)$ is called the sample space and denoted by $\mathscr{X}$.

Random Sample:

Let X be a r.v with d.f F and let $X_1, \ldots, X_n$ be a collection of independent and identically distributed (i.i.d) r.v's with common d.f F. Then the collection $X_1, \ldots, X_n$ is known as a random sample of size n from d.f F, or the corresponding population.

Statistics:

Let $T : \mathscr{X} \to \mathbb{R}^k$ be a Borel function. Then the r.v $T(X_1, \ldots, X_n)$ is called a (sample) statistic provided it is not a function of any unknown parameters, *i.e.* T only depends on sample $(X_1, \ldots, X_n)$.

Example:

Let $X_1, \ldots, X_n$ be a random sample (r.s) from $N(\mu, \sigma^2)$.

$\mathscr{H} = \{(\mu, \sigma) : \mu \in \mathbb{R}, \sigma > 0\} = \mathbb{R} \times (0, \infty)$ is unknown. Then $T_1(\underline{X}) = \sum(X_i - \bar{X})^2$, $T_2 = \sum X_i$ are statistics but $T_3(\underline{X}) = \sum(X_i - \mu)^2$ is not statistics.

In this statistical inference we study two main topics:

(i) Point Estimation, Interval Estimation

(ii) Testing of Hypothesis.

Point Estimation

$\underline{Estimator}$:

Any function of the random sample which is used to estimate the unknown value of the given parametric function $g(\theta)$ is called an estimator.

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

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

Let $X_1, \ldots, X_n$ be a random sample from a population described by family $\mathcal{F} = \{f_{\underline{\theta}}(\cdot) : \underline{\theta} \in \mathscr{H}\}$ of pdf's/pmf's where for each $\underline{\theta} \in \mathscr{H}$ form of $f_{\underline{\theta}}(\cdot)$ is known but $\underline{\theta} \in \mathscr{H}$ is unknown.

Here knowledge of unknown $\underline{\theta} \in \mathscr{H}$ yields knowledge of unknown $\underline{\theta} \in \mathscr{H}$ yields knowledge of the entire population. Moreover $\underline{\theta}$ itself may represent an important characterstic of the population (such as population mean, variance *etc.*) and there may be direct interest in obtaing a point estimate of $\underline{\theta}$. Sometimes there may be interest in obtaing point estimate of $g(\theta)$, a given function of $\underline{\theta}$.

<u>Goal</u>: Based on a random sample $X_1, \ldots, X_n$ from the population, find a good point estimate $g(\underline{\theta})$.

Definition 22.1. *A point estimator of $g(\theta)$ is a function $W(\underline{X})$ of the random sample $\underline{X} = (X_1, \ldots, X_n)$.*

$\underline{Note}$:

(i) An estimator $W(\underline{X})$ is a random variable whereas an estimate is an observed value of the estimator based on an observed sample.

Different Method of Finding Estimator:

(i) $\underline{Method\ of\ moment\ estimator\ (MME)}$:

Let $X_1,\ldots,X_n$ be a random sample from a population with distribution $F_{\underline{\theta}}$, $\underline{\theta}=(\theta_1,\ldots,\theta_k)\in\mathscr{H}$.

Consider k noncentral moments

$$\begin{cases}\mu_1' = E(X_1) = g_1(\underline{\theta})\\ \mu_2' = E(X_1^2) = g_2(\underline{\theta})\\ \vdots \qquad\qquad \vdots\\ \mu_k' = E(X_1^k) = g_k(\underline{\theta})\end{cases}. \tag{22.1}$$

Assume the system of equation (22.1) have solution and solving for $\theta_1,\ldots,\theta_k$ we get

$$\begin{cases}\theta_1 = h_1(\mu_1',\ldots,\mu_k')\\ \vdots \quad \vdots\\ \theta_k = h_k(\mu_1',\ldots,\mu_k')\end{cases}.$$

Define the 1^{st} k noncentral sample moments

$$\alpha_1 = \frac{1}{n}\sum_{i=1}^n X_i,\ \ \alpha_2 = \frac{1}{n}\sum_{i=1}^n X_i^2,\ldots,\alpha_k = \frac{1}{n}\sum_{i=1}^n X_i^k.$$

In method of moments we estimate k^{th} population moment by k^{th} sample moment *i.e.* $\hat{\mu}_j=\alpha_j,\ \ j=1,2,\ldots,k.$

Thus the method of moment estimators (MME) of $\theta_1,\ldots,\theta_k$ are

$$\hat{\theta}_1 = h_1(\alpha_1,\ldots,\alpha_k),\ldots,\hat{\theta}_k = h_k(\alpha_1,\ldots,\alpha_k).$$

Example:

$X_1,\ldots,X_n \overset{iid}{\sim} N(\mu,\sigma^2)$. Find MME of $\underline{\theta}=(\mu,\sigma^2)$ and $(\mu+\sigma)=\psi(\underline{\theta})$.

$\underline{Solution}$:

$\mu_1' = E(X_1)=\mu,\ \ \mu_2' = E(X_1^2)=\sigma^2+\mu^2$. So $\mu=\mu_1'$ and $\sigma^2=\mu_2'-{\mu_1'}^2$.

So $\hat{\mu}_{MME}=\bar{X},\ \hat{\sigma}^2_{MME} = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^2$. Thus,

$$\hat{\mu}_{MME}=\bar{X},\ \ \ \hat{\sigma}^2_{MME} = \frac{1}{n}\sum(X_i-\bar{X})^2.$$

MME of $\psi(\underline{\theta})=\mu+\sigma$ is $\bar{X}+\sqrt{\frac{1}{n}\sum(X_i-\bar{X})^2}$.

Example:

$X_1,\ldots,X_n \overset{iid}{\sim} Bin(m,p),\ \underline{\theta}=(m,p)\in\mathscr{H},\ \mathscr{H}=\{1,2,\ldots\}\times(0,1).$

$$\mu_1' = E(X_1) = mp,\ \ \mu_2' = E(X_1^2) = mp(1-p)+m^2p^2.$$

Now we have

$$mp(1-p) = \mu_2'-{\mu_1'}^2 \implies (1-p) = \frac{\mu_2'-{\mu_1'}^2}{\mu_1'} \implies p = 1-\frac{\mu_2'-{\mu_1'}^2}{\mu_1'} \text{ and } m=\frac{\mu_1'}{p}.$$

$$\hat{p}_{MME} = 1-\frac{\frac{1}{n}\sum_{i=1}^n X_i^2-\bar{X}^2}{\bar{X}} = 1-\frac{\frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^2}{\bar{X}},\ \hat{m}_{MME}=\frac{\bar{X}}{\hat{p}_{MME}} = \frac{\bar{X}^2}{1-\frac{1}{n}\sum_{i=1}^n(X_i-\bar{X})^2}.$$

Remark 22.2. *(i) Some times for k-dimensional parameter we may have to consider more than k equations.*

Example:

$X_1, \ldots, X_n \overset{iid}{\sim} U[-\theta, \theta]$, $\theta > 0$. $\mu_1' = E(X_1) = 0$, $\mu_2' = E(X_1^2) = \frac{\theta^2}{3} \implies \theta = \sqrt{3\mu_2'}$,

$\hat{\theta}_{MME} = \sqrt{\frac{3}{n}\sum_{i=1}^{n} X_i^2}$.

Here the 1^{st} moment did not give any solution so we are using 2^{nd} moment.

(ii) MME may not exist.

(ii) MME is not unique.

Method of Maximum Likliehood Estimation:

For a given observed sample point $\underline{x} = (x_1, \ldots, x_n)$, define

$$L_{\underline{x}}(\underline{\theta}) = \prod_{i=1}^{n} f(x_i|\underline{\theta}), \ \ \underline{\theta} \in \mathscr{H}$$

as a function of $\underline{\theta} = (\theta_1, \ldots, \theta_k)$.

$L_{\underline{x}}(\underline{\theta})$: The probability that the observed sample point $\underline{x}$ came from population represented by pdf/pmf $f(\underline{x}|\underline{\theta})$, $\underline{\theta} \in \mathscr{H}$, *i.e.* the likelihood of observing r.s. $(x_1 \ldots x_n)$ from $f(\cdot|\underline{\theta})$, $\theta \in \mathscr{H}$.

Definition 22.3. *(1) For a given sample point $\underline{x}$, the function $L_{\underline{x}}(\underline{\theta})$ as a function of $\theta \in \mathscr{H}$ is called the likelihood function.*

It means sense to find $\underline{\hat{\theta}}$ that maximizes $L_{\underline{x}}(\underline{\theta})$ for a given sample point $\underline{x}$, as the corresponding population (represented by pdf/pmf) is most likely to have yielded the observed sample $\underline{x}$.

Definition 22.4. *For each $\underline{x} \in \mathfrak{X}$, let $\hat{\theta} = \hat{\theta}(\underline{x})$ be such that*

$$L_{\underline{x}}(\hat{\theta}) = \sup_{\underline{\theta} \in \mathscr{H}} L_x(\underline{\theta}).$$

Then a maximum likelihood estimator (MLE) of the parameter $\underline{\theta} = (\theta_1, \ldots, \theta_k)$, based on random sample $\underline{X}$ is $\hat{\theta}(\underline{X})$.

Finding MLE:

$\hat{\theta}$: MLE if $\quad L_{\underline{x}}(\hat{\theta}) = \sup_{\underline{\theta} \in \mathscr{H}} L_{\underline{x}}(\underline{\theta})$.

Define, $\ l_{\underline{x}}(\underline{\theta}) = \log L_{\underline{x}}(\underline{\theta})$, $\underline{\theta} \in \mathscr{H}$. This is called log-likelihood function.

MLE maximizes $L_{\underline{x}}(\underline{\theta})$ or equivalently $l_{\underline{x}}(\underline{\theta})$. If $l_{\underline{x}}(\underline{\theta})$ is differentiable and maximum at the interior of $\mathscr{H}$ then MLE $\hat{\theta}(\underline{x})$ satisfies

$$\frac{\partial}{\partial \theta_i} l_{\underline{x}}(\underline{\theta})\Big|_{\theta = \hat{\theta}(\underline{x})} = 0, \ \ i = 1, 2, \ldots k.$$

Examples:

(1) Let $X \sim Bin(n, p)$, $0 \le p \le 1$, $\theta = p \in [0, 1]$. Here n is known.

$$L_x(p) = f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}, \ \ x = 0, 1, \ldots, n$$

$$\log L_x(p) = l_x(p) = \log\binom{n}{x} + x\log p + (n-x)\log(1-p)$$

$$\frac{\partial l_x}{\partial p} = \frac{x}{p} - \frac{n-x}{1-p} = \frac{x-np}{p(1-p)} \begin{cases} <0, & \text{if } p > \frac{x}{n} \\ >0, & \text{if } p < \frac{x}{n} \end{cases}.$$

So,

$$l_x(p) \begin{cases} \uparrow, & \text{if } p < \frac{x}{n} \\ \downarrow, & \text{if } p > \frac{x}{n} \end{cases}.$$

So, $\hat{p}_{ML} = \frac{X}{n}$.

(2) Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{P}(\lambda),\ \lambda > 0$ and $\lambda \le \lambda_0$.

$$L_{\underline{x}}(\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod_{i=1}^{n}(x_i!)}$$

$$l_{\underline{x}}(\lambda) = \log L_{\underline{x}}(\lambda) = -n\lambda + \sum x_i \log\lambda - \log\Big\{\prod_{i=1}^{n}(x_i!)\Big\}$$

$$\frac{\partial l_x(\lambda)}{\partial\lambda} = -n + \frac{\sum x_i}{\lambda} = \frac{\sum x_i - n\lambda}{\lambda} \begin{cases} >0, & \text{if } \lambda < \bar{x} \\ <0, & \text{if } \lambda > \bar{x} \end{cases}.$$

So,

$$l_x(\lambda) \begin{cases} \uparrow, & \text{if } \lambda < \bar{x} \\ \downarrow, & \text{if } \lambda > \bar{x} \end{cases}.$$

Given $\lambda \le \lambda_0$ *i.e.* $\mathscr{H} = (0, \lambda_0)$.

$$\hat{\lambda}_{ML} = \begin{cases} \bar{x}, & \text{if } \bar{x} \le \lambda_0 \\ \lambda_0, & \text{if } \bar{x} > \lambda_0 \end{cases}.$$

This is restricted MLE.

Remark 22.5. *(1) MLE is not unique.*

Example:

$X_1, \ldots, X_n \overset{iid}{\sim} U[\theta - a, \theta + a],\ a \in \mathbb{R},\ a > 0$ where a is a known constant.

The likelihood function

$$L_{\underline{x}}(\theta) = \begin{cases} \left(\dfrac{1}{2a}\right)^n, & \text{if } \theta - a \le x_{(1)} < x_{(2)} < \cdots < x_{(n)} \le \theta + a \\ 0, & \text{otherwise.} \end{cases}$$

So $L_{\underline{x}}(\theta)$ is maximum when $\theta - a \le x_{(1)}$ and $x_{(n)} \le \theta + a$ *i.e.* $x_{(n)} - a \le \theta \le x_{(1)} + a$. So any value of θ between $x_{(n)} - a$ to $x_{(1)} + a$ is MLE of θ. We may choose the midpoint *i.e.* $\dfrac{x_{(1)} - x_{(n)}}{2}$ as the MLE.

Example:

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $\mu \in \mathbb{R},\ \sigma > 0$.

$$L_{\underline{x}}(\theta) = \prod_{i=1}^{n}\left[\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}\right] = \frac{1}{\sigma^n(2\pi)^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2},\ \mu \in \mathbb{R},\ \sigma > 0,\ x_i \in \mathbb{R},\ i = 1(1)n.$$

$$\begin{aligned}
l_{\underline{x}}(\theta) &= \log L_{\underline{x}}(\underline{\theta}) = -n\log\sigma - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum(x_i-\mu)^2\\
\frac{\partial l_{\underline{x}}(\theta)}{\partial\mu} &= \frac{1}{\sigma^2}\sum(x_i-\mu),\quad \frac{\partial l_{\underline{x}}(\theta)}{\partial\mu} = 0 \implies \mu = \bar{x}\\
\frac{\partial l}{\partial\sigma^2} &= 0 \implies \sigma^2 = \frac{1}{n}\sum(x_i-\mu)^2.
\end{aligned}$$

So $\hat{\mu}_{ML} = \bar{X}$ and $\sigma^2_{ML} = \frac{1}{n}\sum(X_1-\bar{X})^2$. Consider the MLE of μ and σ^2 when $\mu > 0$. In this case $\mu \in (0,\infty)$. We have

$$\frac{\partial l_x}{\partial\mu} = \frac{n(\bar{x}-\mu)}{\sigma}\begin{cases} >0, & \text{if } \mu < \bar{x}\\ <0, & \text{if } \mu > \bar{x}\end{cases}.$$

So,

$$\hat{\mu}_{RML} = \begin{cases}\bar{X}, & \text{if } \bar{X} > 0\\ 0, & \text{if } \bar{X} \le 0\end{cases} = \max\{\bar{X}, 0\}.$$

$$\hat{\sigma}^2_{RML} = \frac{1}{n}\sum(X_1-\hat{\mu}_{RML})^2 = \begin{cases}\frac{1}{n}\sum(X_1-\bar{X})^2, & \text{if } \bar{X} > 0\\ \frac{1}{n}\sum X_1^2, & \text{if } \bar{X}\le 0\end{cases}.$$

$\underline{Exercise}$:

(i) Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Exp(\mu,\sigma)$. Find MLE of μ and σ.

(ii) Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Gamma(r,\lambda)$. Find MLE of r and λ.

Example:

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} U(0,\theta)$. Then we have

$$L_{\underline{x}}(\theta) = \begin{cases}\frac{1}{\theta^n}, & 0 < x_i < \theta,\ i = 1(1)n\\ 0, & \text{otherwise}\end{cases} = \begin{cases}\frac{1}{\theta^n}, & 0 < x_{(1)} < x_{(2)} < \cdots < x_{(n)} < \theta\\ 0, & \text{otherwise}\end{cases}$$

$L_{\underline{x}}(\theta)$ is the decreasing function of θ. So $L_{\underline{x}}(\theta)$ attains its maximum when θ is minimum $\implies \hat{\theta}_{ML} = X_{(n)}$.

$\underline{Exercise}$:

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} U(\theta, \theta+1)$. Find MLE of θ.

Remark 22.6. *(3) Finding MLE requires maximization of likelihood function which is sometimes difficult and may require numerical optimization techniques.*

Remark 22.7. *(4) MLE is sometimes sensitive to data, a slightly different data may produce a vastly different MLE.*

Invariance Property of MLE:

Let $\hat{\theta} = \theta(\underline{X})$ be a MLE of $\hat{\theta}$ then the of MLE of $\psi(\theta)$ is $\psi(\hat{\theta})$.

$\underline{Exercise}$:

(1) Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Bin(n,\theta)$. Then find a MLE of $\theta(1-\theta)$.

(2) Let $X_1, X_2, \ldots, X_m \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$. The two sample are independent. Then find the distribution of

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{(m-1)S_1^2}{\sigma_1^2} + \frac{(n-1)S_2^2}{\sigma_2^2}}} \sqrt{\frac{m+n-2}{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}.$$

(3) Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} U(\theta_1, \theta_2)$. Find MLE of θ_1 and θ_2.

Efficiency of Estimators:

Let $g(\theta)$ be a parametric function and $\delta(\underline{X})$ be an estimator.

$$\text{Mean absolue Error} : E|\delta(\underline{X}) - g(\theta)|,$$
$$\text{Mean squared Error} : E(\delta(\underline{X}) - g(\theta))^2$$

Definition 22.8. *We say that estimator δ_1 is better (more efficient) than δ_2 if*

$$MSE(\delta_1) \leq MSE(\delta_2) \;\; \forall \theta \in \mathscr{H}.$$

If $E(\delta(\underline{X})) = g(\theta)$ then $MSE(\delta) = E(\delta(\underline{X}) - g(\theta))^2 = \mathrm{Var}(\delta(\underline{X}))$.

Unbaised Estimator:

A estimator $\delta(\underline{X})$ is said to be an unbaised estimator of g(θ) if

$$E((\delta(\underline{X})) = g(\theta) \;\; \forall \theta \in \mathscr{H}.$$

Example: $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$, $\delta_1(\underline{X}) = \bar{X}$.

$E(\delta_1(\underline{X})) = \mu \implies \delta_1(\underline{X})$ is unbaised for μ.