Statistical Inference Note

May 28, 2023

Notation

1. χ - sample space
2. X - random variable
3. $F_X(x)$ - cdf of X
4. $f_X(x)$ - pdf of X
5. $X = (X_1, \ldots, X_n)$ - X is a random sample of size n

Chapter 5: Properties of a Random Sample

Definition 1 (Random Sample) *: The random variables $X_1, \ldots, X_n$ are called a random sample of size n from the population $f(x)$ if $X_1, \ldots, X_n$ are mutually independent variables and the marginal pdf or pmf of each X_i is the same $f(x)$. $\{X_i\}$ are called iid rv with pdf or pmf $f(x)$.*

Definition 2 (Statistics) *: Let $X_1, \ldots, X_n$ be **a** random sample of size n from a population and let $T(x_1, \ldots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \ldots, X_n)$. Then the random variable or random vector $Y = T(X_1, \ldots, X_n)$ is called a **statistics**. The probability distribution of a statistic Y is called the sampling distribution of Y.*

Definition 3 (Sample mean and variance)

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \textit{(Sample mean)}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad \textit{(Sample variance)}$$

$\bar{x}, s^2$ denote to observed values of $\overline{X}, S^2$

Result 1 *Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then*

1. $\mathrm{E}\bar{X} = \mu$
2. $\mathrm{Var}\bar{X} = \frac{\sigma^2}{n}$
3. $\mathrm{E}S^2 = \sigma^2$

Remark: The statistics $\bar{X}$ is unbiased estimator of μ. S^2 is unbiased estimator of σ^2 due to the $n-1$ denominator. If we use n as denominator, $\mathrm{E}S^2$ would be $\frac{n-1}{n}\sigma^2$.

Chapter 6: Principles of Data Reduction

Remark: Any statistics $T(X)$ defines a form of data reduction or data summary. When we use only the observed value of the statistics $T(X)$ rather than the entire observed sample x, we will treat two samples as equal if $T(x) = T(y)$. Therefore data reduction in terms of a particular statistic can be though of as a partition of the sample space. The image $\mathcal{T} = \{t : t = T(x) for some x \in \chi\}$ partition χ into $A_t = \{x|T(x) = t\}$

The Sufficiency Principle of data reduction

Definition 4 (Sufficient Statistics) *A statistic $T(X)$ is a sufficient statistics for θ if the conditional distribution of the sample X given the value of $T(X)$ does no depend on θ. In other word, $T(X)$ captures all the information about θ contained in the sample X. Knowing X does not provide more information about θ.*

Definition 5 (Sufficiency Principle) *If $T(X)$ is sufficient statistics for θ, then any inference about θ should depend on sample X only through $T(X)$. If x, y are two sample points such that $T(x) = T(y)$, the inference about θ should be the same whether $X = x$ or $X = y$ is observed.*

Result 2 (Factorization Theorem) *Let $f(x|\theta)$ denote the joint pdf or pmf of a sample X. A statistic $T(X)$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(x)$ such that for all sample points x and all parameter points θ.*

$$f(x|\theta) = g(T(x)|\theta)h(x)$$

The Likelihood Principle

Definition 6 *Let $f(x|\theta)$ be joint pdf/pmf of a sample $X = (X_1, \ldots, X_n)$, the likelihood function is defined as*

$$L(\theta|x) = f(x|\theta)$$

LIKELIHOOD PRINCIPLE (Informal)

If x and y are two sample points such that $L(\theta|x)$ is proportional to $L(\theta|y)$, that is there exist a constant $C(x, y)$ (only depends on (x, y)) such that

$$L(\theta|y) = C(x, y)L(\theta|x)$$

Then the conclusions drawn from x and y should be identical.

Chapter 7: Point Estimation

Motivation: we want to find a good estimator for θ or $\tau\theta$ using samples from a pdf $p(x|\theta)$ since θ yields knowledge of the entire population.

Definition 7 (Point Estimator) *A point estimator is any function $W(X_1 \ldots X_n)$ of a sample; that is any statistic is a point estimator.*

Remark: When a sample is taken, estimator is a function of the rv $X_1, \ldots, X_n$ while an estimate is a function of realized values $X_1, \ldots, X_n$.

Method of Finding Estimator

0.0.1 Method of Moments

Result 3 (Method of Moments) *Let $X_1, \ldots, X_n$ be a sample from population $f(x|\theta_1, \ldots, \theta_n)$. The method of moments estimators are found by equating the first k sample moments ($m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$) to the corresponding k population moments ($\mu_k' = EX^k$)*

Example: Suppose $X_1, \ldots, X_n$ are iid from $f(x|\theta, \sigma^2)$, we have sample moment $m_1 = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}$, $m_2 = \frac{1}{n}\sum_{i=1}^n X_i^2$ and population moment $\mu_1' = \mathbf{E}X = \theta$, $\mu_2' = \mathbf{E}X^2 = \theta^2 + \sigma^2$.

Then we have

$$\theta = \bar{X} \qquad \rightarrow \theta = \frac{1}{n}\sum_{i=1}^n X_i$$

$$\theta^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 \qquad \rightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \theta^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2$$

0.0.2 Maximum Likelihood Estimator

Definition 8 *(Likelihood function) If $X_1, \ldots, X_n$ is an iid sample from a population $f(x|\theta_1, \ldots, \theta_n)$, the likelihood function is defined by*

$$L(\theta|x) = L(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n) = \prod_{i=1}^n f(x_i|\theta_1, \ldots, \theta_k)$$

Definition 9 *For each sample point x, let $\hat{\theta}(x)$ be a parameter value at which $L(\theta|x)$ attains its maximum as a function of θ with x held fixed. A maximum likelihood estimator(MLE) of the parameter θ based on a sample X is $\hat{\theta}(X)$*

If the likelihood function is differentiable wrt θ_i, the candidate extrema are $\frac{\partial}{\partial \theta_i} L(\theta|x) = 0$ (Extrema can occur on boundary, we need to check those as well).

Remark: The drawbacks are:

1. The problem of actually finding the global maximum and verifying it.
2. Numerical sensitivity; how sensitive is the estimate to the change in data. This can occur when the likelihood function is very flat in the neighborhood of its maximum or when there is no finite maximum. When using numerical methods, spend some time investigating the stability of the solution.

Result 4 (invariance property of MLE) *If $\hat{\theta}$ is the MLE of θ, then for any function τ, $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$.*

Bayes Estimators

In the Bayesian approach, θ is not thought to be fixed but a quantity whose variation can be described by a probability distribution (**the prior distribution**).

Definition 10 (Prior distribution) *is a subjective distribution based on experimenter's belief and is formulated before the data are seen. Denoted as $\pi(\theta)$*

Definition 11 (Posterior distribution) *is the updated prior distribution with the information after a sample taken from the population. Denoted as $\pi(\theta|x)$.*

The prior distribution and posterior distribution are related by Bayesian rule:

$$\pi(\theta|x) = \frac{f(x, \theta)}{m(x)} = \frac{f(x|\theta)\pi(\theta)}{m(x)}$$

where $m(x) = \int f(x|\theta)\pi(\theta)d\theta$

Remark: Posterior distribution is a conditional distribution based on observing the sample. It can be used to make statements about θ. E.g. the mean of posterior distribution can be used as a point estimate of θ.

Definition 12 (Conjugate family) *Let $\mathcal{F}$ denote the class of pdf/pmf $f(x|\theta)$ (indexed by θ). A class $\prod$ of prior distribution is called the conjugate family for $\mathcal{F}$ if the posterior distribution is in $\prod$ for all $f \in \mathcal{F}$, all priors in $\prod$ and all $x \in \mathcal{X}$*

Methods of Evaluating Estimators

Mean Squared Error

Definition 13 *The bias of a point estimator W of a parameter θ is the difference between the expected value of W and θ; $\text{Bias}_\theta W = \mathbf{E}_\theta W - \theta$. An estimator whose ibas is equal to 0 is called unbiased and satisfies $\mathbf{E}_\theta W = \theta$ for all θ.*

Definition 14 *The mean square error (MSE) of an estimator W of a parameter θ is the function of θ defined by*

$$\mathbf{E}_{x|\theta}(W-\theta)^2 = \text{Var}_\theta W + (\mathbf{E}_\theta W - \theta)^2 = \text{Var}_\theta W + (\text{Bias}_\theta W)^2$$

Remark: Be aware that controlling bias does not guarantee that MSE is controlled. A biased estimator can have a lower MSE due to low variance.

Remark: MSE is a reasonable criterion for location parameters but not for scale parameters. In scale case, 0 is a natural lower bound, so the estimation problem is not symmetric. MSE tends to be forgiving of underestimation.