STAT-510		2017 Fall
	Lecture 1: Inference	
Instructor:	Scribe: Yutong Dai	Last Modified: 2017-10-12

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1.1 Concepts review

- **Sample Space**:

1.1.1 Statistical Models

A statistical model consists of three components, a random object X, its corresponding space $\mathcal{X}$ and an entire family of $\mathcal{P}$ of probabilty distribution. We denote it as $(X, \mathcal{X}, \mathcal{P})$.

Remark Often, the families are parametrized by a finite-dimensional parameter θ, i.e.,

$$\mathcal{P} = \{P_\theta | \theta \in \Theta \subset R^k\}$$

The Θ is called the parameter space.

1.1.2 Approch to inference

Task: Make inferences about θ based on observing the data $X = x$.

- **Frequenist**: The frequentist approach assumes that the parameter θ is fixed but unknown (that is, we know only that $\theta \in \Theta$).
  - An inference is an action, which is a function

    $$\delta : \mathcal{X} \to \mathcal{A}$$

    for some action space $\mathcal{A}$.
  - e.g. If one whish to estimate θ, then $\delta(x)$ would be the estimate and $\mathcal{A} = \Theta$.
- **Bayesian**: Given your prior distribution $\pi(\theta)$ on Θ, which may be your subjective distribution, the Bayes approach tells you how to update your opinion upon observing $X = x$. The update is of course the posterior and the posterior $f_{\Theta|X}(\theta|x)$ is the inference (at least all inferences are derived from it).

1.2 Point Estimation

Task: We assume a statistical model with parameter space Θ , and suppose we wish to estimate some function g of θ,

$$g : \Theta \to R^q$$

In the following, we try to find estimator of the parameter θ according to different principles.

Definition of estimator: Formally, an estimator is a function $\delta(x)$,

$$\delta : \mathcal{X} \to \mathcal{A}$$

where $\mathcal{A}$ is some space, presumably the space of $g(\theta)$, but not always. The estimator can be any function of x, but cannot depend on an unknown parameter.

Remark The estimate is the reliazation form of estimator. That is, suppose $X_1, ..., X_n \overset{iid}{\sim} X$. Then the $\delta(X_1, ..., X_n)$ is the estimator and $\delta(X_1 = x_1, ..., X_n = x_n)$ is the estimate.

If $\delta(x)$ is an estimator of $g(\theta)$, we woulde denote $\delta(x)$ as $\hat{g(\theta)}$.

1.2.1 Plug-in methods

Under a parametric model $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$ (or a non-parametric $\mathcal{P} = \{P_F | F \in \mathcal{F}\}$), any real-valued characteristic g of a particular member P_θ (or P_F) can be written as a mapping from the parameter-space Θ (or $\mathcal{F}$) to the R^n. If your derive a estimator $\hat{\theta}$ of θ, it's natural to use $g(\hat{\theta})$ to estimate $g(\theta)$. This method for constructing estimates is commonly reffered to as the plug-in method.

Example 1(parametric) Suppose $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then $P(X \leq 10) = \Phi((10 - \mu)/\sigma)$. We simply plug in the mean estimator $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and the variance estimator $S^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$ to obtain the estimator

$$\delta(x) = \Phi(\frac{10 - \bar{x}}{s}).$$

Example 2(non-parametric) Suppose $\mathcal{F}$ is a large class of distribution functions and we are interested in estimating *functionals* on $\mathcal{F}$,

$$\theta : \mathcal{F} \to R^n.$$

For example,the mean $\theta(F) = EX = \int x dF(x)$ is a appropriate functionals on $\mathcal{F}$. The $\theta(F)$ is then the population parameter. We would like to use emperical distribution function $\hat{F}_n(x) = \frac{1}{n}\sum_i \mathbf{I}(X_i \leq x)$ to estimate $F(x)$. Hence, the plug-in estimator of the parameter $\theta(F)$ would be $\int x d\hat{F}_n(x)$.

Remark $E[\hat{F}_n(x)] = 1/n \sum_{i=1}^n E[\mathbf{I}(X_i \leq x)] = F(x)$. Therefore, $\hat{F}_n(x)$ is an unbiased estimation.

Theorem 1.1 (Glivenko–Cantelli) $\sup_{x \in R} ||\hat{F}_n(x) - F(x)||$ *converges to 0 almost surely, that is,*

$$P[\lim_{n \to +\infty} \sup_{x \in R} ||\hat{F}_n(x) - F(x)|| = 0\ ] = 1.$$

Remark Application of Glivenko–Cantelli theorem.

$$||\hat{\theta}(F) - \theta(F)|| \triangleq || \int x d\hat{F}_n(x) - \int x dF(x)|| \leq \int ||x|| (d\hat{F}_n(x) - dF(x)) \to 0 \quad (n \to +\infty)$$

1.2.2 Methods of Moment

Suppose $X_1, ..., X_n \overset{iid}{\sim} f(x;\theta)$ and $\theta \in R^k$. The idea of moment estimation is simply regarding the sample moment equals to the population moment. And we can establish k equations to estimate the θ,

$$\frac{1}{n}\sum_{i=1}^{n} X_i^j = EX^j, j = 1, 2, ..., k.$$

The problems come along with the moment method is,

- The estimated $\hat{\theta}$ may not in the parameter space.
  - e.g. Use moment method to estimate the parameter of Bernoulli(n, p).
- The k equations may be overdetermined.
  - e.g. Use moment method to estimate the parameter of Possion(λ).

1.2.3 Maximum Likelihood Method

If $x_i \overset{iid}{\sim} f(x;\theta), i = 1, ..., n$, then the likelihood function is defined as $L(X;\theta) = \prod_i f(x_i;\theta)$, where $X = (x_1, ..., x_n)$. The estimator of the Maximum Likelihood Method is

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta\in\Theta} L(X;\theta)$$

We usually consider the log-likelihood function, which takes the form of $\ell(x;\theta) = \log L(X;\theta)$.

The procedures for finding the $\hat{\theta}_{\text{MLE}}$ are as follows.

1. Find all $\hat{\theta}$ that satisfy
$$\frac{\partial \ell(X;\theta)}{\partial\theta}|_{\theta=\hat{\theta}} = 0$$
2. Check the Hessian matrix at those $\hat{\theta}$. To be specific, define
$$H(\hat{\theta}) = \frac{\partial^2 \ell(X;\theta)}{\partial\theta^2}|_{\theta=\hat{\theta}}$$
. If $H(\hat{\theta}) < 0$ and $\Theta = R^k$, where k is the dimension of the parameter space, then $\hat{\theta}_{\text{MLE}} = \hat{\theta}$.
3. If $\Theta \subsetneq R^k$, then we have to compare the $\hat{\theta}$ with the boundaries.

Proposition 1.2 (Invariance of the MLEs) *If $\hat{\theta}_{MLE}$ is the maximum likelihood estimation of θ, then $g(\hat{\theta}_{MLE})$ is the maximum likelihood estimation of $\eta = g(\theta)$, where $g(.)$ is a function.*

Proof: If $g(.)$ is injective, just follow the defintion of $L(X;\theta)$, we have

$$L^*(X;\eta) = \prod_{i=1}^{n} f(x_i; g^{-1}(\eta)) = L(X; g^{-1}(\eta)).$$

Furthermore, we have

$$\max_{\eta} L^*(X;\eta) = \max_{\eta} L(X; g^{-1}(\eta)) = \max_{\theta} L(X;\theta).$$

Therefore, the maximum of $L^*(X;\eta)$ is attained at $\hat{\eta} = g(\hat{\theta}_{\text{MLE}})$, which means the MLE estimator of $g(\theta)$ is $g(\hat{\theta}_{\text{MLE}})$.

If $g(.)$ is not injective, we have to define another version of likelihood function called induced likelihood function, which takes form of

$$L^*(X;\eta) = \sup_{\{\theta: g(\theta)=\eta\}} L(X;\theta).$$

By definition, we know the maxima of $L^*(X;\eta)$ coincides with that of $L(X;\theta)$. Notice that, if $\hat{\eta}$ maximize $L^*(X;\eta)$, then

$$\begin{aligned} L^*(X;\hat{\eta}) &= \sup_{\eta} \sup_{\{\theta: g(\theta)=\eta\}} L(X;\theta) && (1.1)\\ &\overset{(1)}{=} \sup_{\theta} L(X;\theta) && (1.2)\\ &= L(X;\hat{\theta}_{\text{MLE}}) && (1.3) \end{aligned}$$

where (1) holds due two L^* and L have the same maxima. Furthermore

$$\begin{aligned} L(X;\hat{\theta}_{\text{MLE}}) &= \sup_{\{\theta: g(\theta)=g(\hat{\theta}_{\text{MLE}})\}} L(X;\theta) && (1.4)\\ &= L^*(X;\hat{\eta}) && (1.5) \end{aligned}$$

Therefore, the MLE estimator of $g(\theta)$ is $g(\hat{\theta}_{\text{MLE}})$.

■

1.2.3.1 Profile Likelihood Method

If we partition the parameter θ into two parts $\theta = (\lambda, \phi)$. λ is the parameter we are interested in and ϕ is nuisance. To find θ_{MLE}, we can follow two steps:

1. Fix the λ and find $\hat{\phi}(\lambda) = \arg\max_{\phi\in\Phi} \ell(X;\lambda,\phi)$.
2. Find $\hat{\lambda} = \arg\max_{\lambda\in\Lambda} \ell(X;\lambda,\hat{\phi}(\lambda))$.

Then we have following claim.

Claim 1.3 $\hat{\theta}_{MLE} = (\hat{\lambda}, \hat{\phi}(\hat{\lambda}))$.

Proof: By definition, $\forall \phi \in \Phi, \lambda \in \Lambda$, we have following inequalities

$$\ell(X;\lambda,\phi) \le \ell(X;\lambda,\hat{\phi}(\lambda)) \le \ell(X;\hat{\lambda},\hat{\phi}(\hat{\lambda})).$$

Hence, $\ell(X;\lambda,\phi) \le \ell(X;\hat{\lambda},\hat{\phi}(\hat{\lambda}))$, which concludes the claim. ■

1.2.3.2 When Maxium Likelihood fails?

When the MLE can not be solved analyticaly and must be found by the numerical method, the maxium likelihood estimator might be unstable. For example, consider $X_i \overset{iid}{\sim}$ Binomial(k,p). The MLE estimator of k is unstable, since the likelihood function is very flat in the neighborhood of its maximum.

1.2.4 Bayes Estimators

$$\pi(\theta|x) = f(x|\theta)\pi(\theta)/m(x)$$

where the $m(x)$ is the marginal distribution of sample X obtained by $\int f(x|\theta)\pi(\theta)d\theta$ and $f(x|\theta)$ is the sampling distribution.

Remark Sampling distribution is the distribution of the statistic. For example, suppose $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$. The sampling distribution of $\bar{X}_n \sim N(\mu, \sigma^2/n)$ and the sampling distribution of $(X_1, X_2, ..., X_n)$ is the likelihood.

Example 1: (Discrete Case) $X_i \overset{iid}{\sim}$ Bernoulli(p) then $Y = \sum_i^n X_i \sim$ Binomial(n, p). Assume the the prior distribution

$$\pi(p) = Beta(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}.$$

Since sampling distribution $f(y|\theta) = C_n^y p^y (1-p)^{n-y}$, we can find the $m(y)$ and furthermore the $\pi(p|y) = Beta(\alpha+y, \beta+n-y)$. Thus $\hat{p}_{\text{Bayes}} = E(y|\theta) = \frac{\alpha+y}{\alpha+\beta+n}$.

Example 2: (Continuous Case) Suppose $X_i \sim N(\theta, \sigma^2)$ (σ^2 is known) and prior distribution of θ is $N(\mu, \tau^2)$. Then

$$\theta|\bar{x}_n \sim N(\frac{\tau^2}{\sigma^2/n+\tau^2}\bar{x}_n + \frac{\sigma^2/n}{\sigma^2/n+\tau^2}\mu, \frac{\tau^2\sigma^2/n}{\sigma^2/n+\tau^2})$$

1.3 Methods of Evaluating Estimators

The methods discussed in the previous section have outlined resonable techniques for finding point estimators of parameters. In this section, we discuss some basic criteria for evaluating estimators and examine several estimators against these criteria.

1.3.1 Mean square Error

Definition 1.4 *The mean square error(MSE) of an estimator $\hat{\theta} = T(X)$ is defined as*

$$\begin{aligned} MSE(\hat{\theta}) &= E_{X\sim f(x;\theta)}(\hat{\theta}-\theta)^2 && (1.6)\\ &= Var(\hat{\theta}) + (E_{X\sim f(x;\theta)}(\hat{\theta})-\theta)^2 && (1.7)\\ &= Var(\hat{\theta}) + Bias^2(\hat{\theta}) && (1.8) \end{aligned}$$

One problem with MSE is that it penalizes equally for overestimation and underestimation, which is not reasonable for paramters that have natural bounds. For example, the variance σ^2 in $N(\mu, \sigma^2)$, has a restriction that $\sigma^2 > 0$. Genearlly, when the estimation problem is not symmetric, using MSE criteria might bring some problems.