Math 408 - Mathematical Statistics

Lecture 38. Fundamental Concepts of Statistical Inference: an Overview

May 3, 2013

Agenda

- Probability Theory
- Survey Sampling
- Fundamental Concepts of Statistical Inference

Statistical Inference

Statistical inference is the process of using data to infer the distribution that generates the data. The basic statistical inference problem is the following:

Basic Problem

We observe $X_1, \ldots, X_n \sim \pi$. We want to estimate π or some features of π such as its mean.

Definition

A **statistical model** is a set of distributions or a set of densities $\mathcal{F}$.

- A **parametric model** is a set $\mathcal{F}$ that can be parameterized by a finite number of parameters.
- A **nonparametric model** is a set $\mathcal{F}$ that cannot be parameterized by a finite set of parameters.

Point Estimation, Confidence Intervals, Hypothesis Testing

Given a parametric model, $\mathcal{F} = \{\pi(x|\theta), \ \ \theta \in \Theta\}$, the problem of inference is then to estimate the parameter θ from the data.

Almost all problems in statistical inference can be identified as being one of three types: **point estimates**, **confidence intervals**, and **hypothesis testing**.

- Point Estimation refers to providing a single "best guess."
  Suppose $X_1, \ldots, X_n \sim \pi(x|\theta)$, where $\pi(x|\theta) \in \mathcal{F}$.
  A point estimator $\hat{\theta}_n$ of a parameter θ is some function of $X_1, \ldots, X_n$:

$$\hat{\theta}_n = f(X_1, \ldots, X_n)$$

- A $100(1-\alpha)\%$ Confidence Interval for a parameter θ is a random interval $I_n = (a, b)$ where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ such that

$$\mathbb{P}(\theta \in I_n) = 1 - \alpha$$

- In Hypothesis Testing, we start with some default theory, called a null hypothesis, and we ask if the data provide sufficient evidence to reject the theory. If not, we accept the null hypothesis.

Method of Moments

Suppose that $X_1, \ldots, X_n \sim \pi(x|\theta)$ where $\theta \in \Theta$, and we want to estimate θ based on the data $X_1, \ldots, X_n$.

Method of Moments

- *Let* $\mu_j(\theta) = \mathbb{E}_\theta[X^j]$ *be the* j^{th} *moment of a probability distribution* $\pi(x|\theta)$
- *Let* $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^n X_i^j$ *be the* j^{th} *sample moment*
  (LLN: $\hat{\mu}_j \xrightarrow{\mathbb{P}} \mu_j(\theta)$*, when* $n \to \infty$*)*
- *Suppose that the parameter* θ *has* k *components,* $\theta = (\theta_1, \ldots, \theta_k)$

The **method of moments estimator** $\hat{\theta}$ *is defined to be the value of* θ *such that*

$$\begin{cases} \mu_1(\theta) = \hat{\mu}_1 \\ \mu_2(\theta) = \hat{\mu}_2 \\ \qquad \ldots\ldots \\ \mu_k(\theta) = \hat{\mu}_k \end{cases} \tag{1}$$

- System (1) is a system of k equations with k unknowns: $\theta_1, \ldots, \theta_k$
- The solution of this system $\hat{\theta}$ is the MoM estimate of the parameter θ.

Consistency of the MoM estimator

Definition

Let $\hat{\theta}_n$ be an estimate of a parameter θ based on a sample of size n. Then $\hat{\theta}_n$ is **consistent** if

$$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$$

Theorem

The method of moments estimate is consistent.

The Likelihood Function

The most common method for estimating parameters in a parametric model is the method of maximum likelihood.

Suppose $X_1, \dots, X_n$ are i.i.d. from $\pi(x|\theta)$.

Definition

The **likelihood function** is defined by

$$\boxed{\mathcal{L}(\theta) = \prod_{i=1}^{n} \pi(X_i|\theta)}$$

Important Remark:

- The likelihood function is just the joint density of the data, except that we treat it as a function of the parameter θ.

Maximum Likelihood Estimate

Definition

The **maximum likelihood estimate** (MLE) of θ, denoted $\hat{\theta}_{\mathrm{MLE}}$, is the value of θ that maximizes the likelihood $\mathcal{L}(\theta)$

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta)$$

$\hat{\theta}_{\mathrm{MLE}}$ makes the observed data $X_1, \ldots, X_n$ “most probable” or “most likely”

Important Remark:

Rather than maximizing the likelihood itself, it is often easier to maximize its natural logarithm (which is equivalent since the log is a monotonic function). The log-likelihood is

$$\boxed{l(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log \pi(X_i|\theta)}$$

Properties of MLE

- MLE is consistent:
$$\hat{\theta}_{\mathrm{MLE}} \xrightarrow{\mathbb{P}} \theta_0$$
where θ_0 denotes the true value of θ.
- MLE is equivariant:
if $\hat{\theta}_{\mathrm{MLE}}$ is the MLE of $\theta \quad \Rightarrow \quad f(\hat{\theta}_{\mathrm{MLE}})$ is the MLE of $f(\theta)$.
- MLE is asymptotically optimal: among all well behaved estimators, the MLE has the smallest variance, at least for large sample sizes n.
- MLE is asymptotically Normal:
$$\hat{\theta}_{\mathrm{MLE}} \to \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$$
where
$$I(\theta) \stackrel{\mathrm{def}}{=} \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log \pi(X|\theta)\right)^2\right] = \int \left(\frac{\partial}{\partial\theta}\log \pi(x|\theta)\right)^2 \pi(x|\theta)dx$$
  - $I(\theta)$ is called Fisher Information.
- MLE is asymptotically unbiased:
$$\lim_{n\to\infty} \mathbb{E}[\hat{\theta}_{\mathrm{MLE}}] = \theta_0$$

Confidence Intervals from MLEs

Recall that

Definition

A $100(1-\alpha)\%$ **confidence interval** for a parameter θ is a random interval calculated from the sample,

$$X_1, \ldots, X_n \sim \pi(x|\theta)$$

which contains θ with probability $1-\alpha$.

There are three methods for constructing confidence intervals using MLEs $\hat{\theta}_{\mathrm{MLE}}$:

- Exact Method
- Approximate Method
- Bootstrap Method

Exact Method

Exact Method provides exact confidence intervals.

- Example: $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$

$$\mu: \quad \hat{\mu}_{\mathrm{MLE}} \pm \frac{1}{\sqrt{n-1}} \hat{\sigma}^2_{\mathrm{MLE}} t_{n-1}(\alpha/2)$$

$$\sigma^2: \quad \left(\frac{n\hat{\sigma}^2_{\mathrm{MLE}}}{\chi^2_{n-1}(\frac{\alpha}{2})}, \frac{n\hat{\sigma}^2_{\mathrm{MLE}}}{\chi^2_{n-1}(1-\frac{\alpha}{2})} \right)$$

These result is based of the following facts:

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \sim t_{n-1}$$

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

Remark:

The main drawback of the exact method is that in practice the sampling distributions — like t_{n-1} and χ^2_{n-1} in our example — are not known.

Approximate Method

One of the most important properties of MLE is that it is asymptotically normal:

$$\hat{\theta}_{\mathrm{MLE}} \to \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right), \qquad \text{as } n \to \infty$$

where $I(\theta_0)$ is Fisher information

$$I(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log \pi(X|\theta) \right)^2 \right]$$

Since the true value θ_0 is unknown, we will use $I(\hat{\theta}_{\mathrm{MLE}})$ instead of $I(\theta_0)$:

Result

An **approximate** $100(1-\alpha)\%$ *confidence interval for* θ_0 *is*

$$\hat{\theta}_{\mathrm{MLE}} \pm \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta}_{\mathrm{MLE}})}}$$

where z_α *is the point beyond which the standard normal distribution has probability* α.

Measure of Efficiency: Mean Squared Error

In most estimation problems, there are many possible estimates $\hat{\theta}$ of θ.
For example, the MoM estimate $\hat{\theta}_{\mathrm{MoM}}$ or the MLE estimate $\hat{\theta}_{\mathrm{MLE}}$.

Question: How would we choose which estimate to use?

Qualitatively, it is reasonable to choose that estimate whose distribution is most highly concentrated about the true parameter value θ_0. To make this idea work, we need to define a quantitative measure of such concentration.

Definition

The **mean squared error** of $\hat{\theta}$ as an estimate of θ_0 is

$$\boxed{\mathrm{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta_0)^2]}$$

- The mean squared error can be also written as follows:

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{V}[\hat{\theta}] + \underbrace{(\mathbb{E}(\hat{\theta}) - \theta_0)^2}_{\text{squared bias}}$$

- If $\hat{\theta}$ is unbiased, then $\mathrm{MSE}(\hat{\theta}) = \mathbb{V}[\hat{\theta}]$.

Cramer-Rao Inequality

Let $X_1, \ldots, X_n$ be i.i.d. from $\pi(x|\theta)$. Let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be any unbiased estimate of a parameter θ whose true values is θ_0. Then, under smoothness assumptions on $\pi(x|\theta)$,

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{V}[\hat{\theta}] \geq \frac{1}{nI(\theta_0)}$$

Important Remarks:

- $\hat{\theta}$ can't have arbitrary small MSE
- The Cramer-Rao inequality gives a lower bound on the variance of any unbiased estimate.

Definition

An unbiased estimate whose variance achieves this lower bound is said to be **efficient**.

Recall that MLE is asymptotically Normal: $\hat{\theta}_{\mathrm{MLE}} \to \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$

- Therefore, MLE is asymptotically efficient
- However, for a finite sample size n, MLE may not be efficient

Hypothesis Testing: General Framework

Suppose that we partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1 and that we wish to test

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

We call H_0 the **null hypothesis** and H_1 the **alternative hypothesis**.

Let X be data and let $\mathcal{X}$ be the range of X. We test a hypothesis by finding an appropriate subset of outcomes $\mathcal{R} \subset \mathcal{X}$ called the **rejection region**. If $X \in \mathcal{R}$ we reject the null hypothesis, otherwise, we do not reject the null hypothesis:

$$X \in \mathcal{R} \Rightarrow \text{ reject } H_0$$

$$X \notin \mathcal{R} \Rightarrow \text{ accept } H_0$$

Usually the rejection region $\mathcal{R}$ is of the form

$$\mathcal{R} = \{x \in \mathcal{X} : T(x) < c\}$$

where T is a **test statistic** and c is a **critical value**.
The main problem in hypothesis testing is

to find an appropriate test statistic T and an appropriate cutoff value c

Main Definitions

In hypothesis testing, there are two types of errors we can make:

- Rejecting H_0 when H_0 is true is called a **type I error**
- Accepting H_0 when H_1 is true is called a **type II error**

Definition

- The probability of a type I error is called the **significance level** of the test and is denoted by α

$$\alpha = \mathbb{P}(\text{type I error}) = \mathbb{P}(\text{Reject } H_0 | H_0)$$

- The probability of a type II error is is denoted by β

$$\beta = \mathbb{P}(\text{type II error}) = \mathbb{P}(\text{Accept } H_0 | H_1)$$

- $(1-\beta)$ is called the **power** of the test

$$\textbf{power} = 1 - \beta = 1 - \mathbb{P}(\text{Accept } H_0 | H_1) = \mathbb{P}(\text{Reject } H_0 | H_1)$$

Thus, the **power** of the test is the probability of rejecting H_0 when it is false.

Neyman-Pearson Lemma

Definition

- A hypothesis of the form $\theta = \theta_0$ is called a **simple hypothesis**.
- A hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is called a **composite hypothesis**.

The Neyman-Pearson Lemma shows that the test that is based on the likelihood ratio is optimal for simple hypotheses:

Neyman-Pearson Lemma

Suppose that H_0 and H_1 are simple hypotheses, $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$. Suppose that the **likelihood ratio test** *that rejects H_0 whenever the likelihood ratio is less than c,*

$$\textit{Reject } H_0 \quad \Leftrightarrow \quad \frac{\mathcal{L}(Data|\theta_0)}{\mathcal{L}(Data|\theta_1)} < c$$

has significance level α_{LR}. Then any other test for which the significance level $\alpha \leq \alpha_{\mathrm{LR}}$ has power less than or equal to that of the likelihood ratio test

$$1 - \beta \leq 1 - \beta_{LR}$$

Generalized Likelihood Ratio Test

Let $X = (X_1, \ldots, X_n)$ be data and let $\pi(x|\theta)$ be the joint density of the data. The likelihood function is then

$$\mathcal{L}(\theta) = \pi(X|\theta)$$

Suppose we we wish to test

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

where Θ_0 and Θ_1 are two disjoint sets of the parameter space Θ, $\Theta = \Theta_0 \sqcup \Theta_1$.

- Based on the data, a measure of relative plausibility of the hypotheses is the ratio of their likelihoods.
- If the hypotheses are composite, each likelihood is evaluated at that value of θ that maximizes it.

This yields the **generalized likelihood ratio**:

$$\boxed{\Lambda^* = \frac{\max_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\max_{\theta \in \Theta_1} \mathcal{L}(\theta)}}$$

Small values of Λ^* tend to discredit H_0.

Generalized Likelihood Ratio Test

For technical reasons, it is preferable to use the following statistic instead of Λ^*:

$$\boxed{\Lambda = \frac{\max_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\max_{\theta \in \Theta} \mathcal{L}(\theta)}}$$

- Λ is called the likelihood ratio statistic.
- Note that

  $$\Lambda = \min\{\Lambda^*, 1\}$$

  Thus, small values of Λ^* correspond to small values of Λ.

The rejection region $\mathcal{R}$ for a generalized likelihood test has the following form:

$$\text{reject } H_0 \quad \Leftrightarrow \quad X \in \mathcal{R} = \{X : \Lambda(X) < \lambda\}$$

The threshold λ is chosen so that

$$\mathbb{P}(\Lambda(X) < \lambda | H_0) = \alpha,$$

where α is the desired significance level of the test.

Distribution of $\Lambda(X)$

In order for the generalized likelihood ratio test to have the significance level α, the threshold λ must be chosen so that

$$\mathbb{P}(\Lambda(X) < \lambda | H_0) = \alpha$$

If the distribution of $\Lambda(X)$ under H_0 is known, then we can determine λ.
Generally, the distribution of Λ is not of a simple form, but in many situations the following theorem provides the basis for an approximation of the distribution.

Theorem

Under smoothness conditions on $\pi(x|\theta)$, the null distribution of $-2\log\Lambda(X)$ (i.e. distribution under H_0) tends to a χ^2_d as the sample size $n \to \infty$, where

$$d = \dim\Theta - \dim\Theta_0,$$

where $\dim\Theta$ *and* $\dim\Theta_0$ *are the numbers of free parameters in Θ and Θ_0.*

Summarizing Data: Empirical CDF

Suppose that $x_1, \ldots, x_n$ is a batch of numbers.

Remark: We use the word

- "sample" when $X_1, \ldots, X_n$ is a collection of random variables.
- "batch" when $x_1, \ldots, x_n$ are fixed numbers (data, realization of sample).

Definition

The **empirical cumulative distribution function** (eCDF) is defined as

$$\boxed{F_n(x) = \frac{1}{n}(\# x_i \leq x)}$$

Denote the ordered batch of numbers by $x_{(1)}, \ldots, x_{(n)}$.

- If $x < x_{(1)}$, then $F_n(x) = 0$
- If $x_{(1)} \leq x < x_{(2)}$, then $F_n(x) = 1/n$
- If $x_{(k)} \leq x < x_{(k+1)}$, then $F_n(x) = k/n$

The eCDF is the "data analogue" of the CDF of a random variable

Summarizing Data: Quantile-Quantile Plots

Quantile-Quantile (Q-Q) plots are used for comparing two probability distributions.

Suppose that X is a continuous random variable with a strictly increasing CDF F.

Definition

The p^{th} **quantile** of F is that value x_p such that

$$F(x_p) = p \quad \text{or} \quad \boxed{x_p = F^{-1}(p)}$$

Suppose we want to compare two CDF: F and G.

Definition

The **theoretical Q-Q plot** is the graph of the quantiles of a the CDF F, $x_p = F^{-1}(p)$, versus the corresponding quantiles of the CDF G, $y_p = G^{-1}(p)$, that is the graph $[F^{-1}(p), G^{-1}(p)]$ for $p \in (0, 1)$.

- If the two CDFs are identical, the theoretical Q-Q plot will be the line $y = x$.

Summarizing Data: Empirical Q-Q plots

In practice, a typical scenario is the following:

- $F(x) = F_0(x)$ is a specified CDF (e.g. normal) which is a theoretical model for data $X_1, \ldots, X_n$.
- $G(x)$ is the empirical CDF for $x_1, \ldots, x_n$, a realization of $X_1, \ldots, X_n$ (actually observed data).
- We want to compare the model $F(x)$ with the observation $G(x)$.

Let $x_{(1)}, \ldots, x_{(n)}$ be the ordered batch. Then

Definition

The **empirical Q-Q plot** is the plot of $F_0^{-1}(i/n)$ on the horizonal axis versus $G^{-1}(i/n) = x_{(i)}$ on the vertical axis, for $i = 1, \ldots, n$.

Remarks:

- The quantities $p_i = i/n$ are called plotting positions

Summarizing Data: Measures of Location and Dispersion

- Measures of Location
  - Arithmetic Mean: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (sensitive to outliers)
  - Median: the middle value of the ordered batch values $\tilde{x} = Q_2$
  - Trimmed Mean:

$$\overline{x}_\alpha = \frac{x_{([n\alpha]+1)} + \ldots + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$

  - M estimate: $y^* = \arg\min_{y\in\mathbb{R}} \sum_{i=1}^{n} \Psi(x_i, y)$
    - ⋆ if $\Psi(x_i, y) = (x_i - y)^2$, then $y^* = \overline{x}$
    - ⋆ it $\Psi(x_i, y) = |x_i - y|$, then $y^* = \tilde{x}$
- Measures of Dispersion
  - Sample Standard Deviation (sensitive to outliers):

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

  - Interquartile Range: $\mathrm{IQR} = Q_3 - Q_1$
  - Median Absolute Deviation: $\mathrm{MAD} =$ median of the numbers $|x_i - \tilde{x}|$

Thank you for attention and good luck on the final!