TA Lecture 11 - Statistical Inference

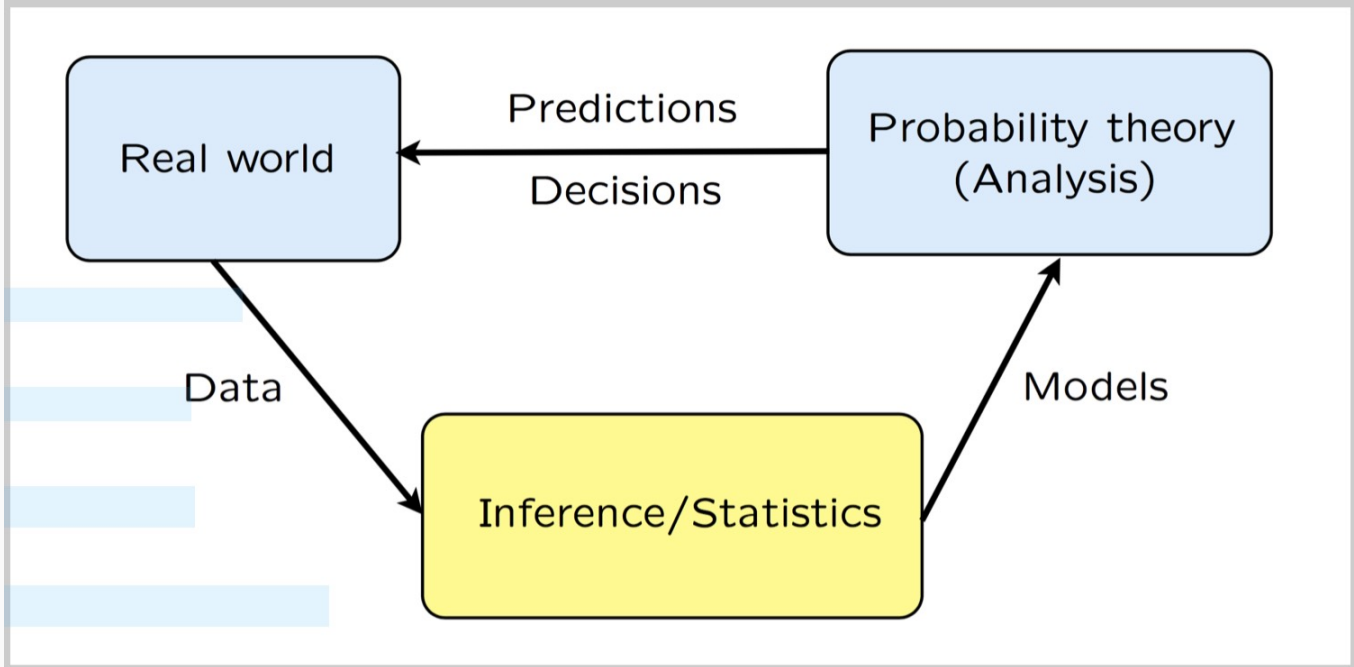
May 22 - 23

School of Information Science and Technology,
ShanghaiTech University

上海科技大学
ShanghaiTech University

Outline

Topic I: Statistical Inference



- Point Estimation
- Interval Estimation (Confidence Interval)
- Hypothesis Testing

Point Estimation

Point Estimation refers to providing a single "best guess" of some quantity of interest such as

- a parameter θ (possibly multi-dimensional) in a parametric model
- a CDF F
- a probability density function f
- a prediction for a future value Y of some random variable

Interval Estimation

Definition

An interval estimate of a real-valued parameter θ is any pair of functions, $L(x_1, \ldots, x_n)$ and $U(x_1, \ldots, x_n)$, of a sample that satisfy $L(\mathbf{x}) \leq U(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. When $\mathbf{X} = \mathbf{x}$ is observed, the inference $L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$ is made. The random interval $[L(\mathbf{X}), U(\mathbf{X})]$ is called an interval estimator.

Definition

A $1 - \alpha$ confidence interval for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ are functions of the data such that

$$P_\theta(\theta \in C_n) \geq 1 - \alpha, \forall \theta \in \Theta.$$

In words, (a, b) traps θ with probability $1 - \alpha$. We call $1 - \alpha$ the coverage of the confidence interval. If θ is a vector, then we use a confidence set instead of an interval.

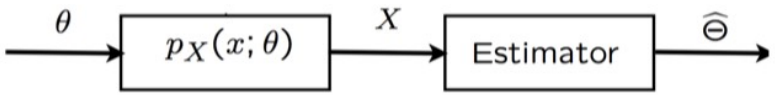
Hypothesis Testing

Start with a finite number of competing hypotheses and use the available data to decide which of them is true.

- Example 1: given a noisy picture, decide whether there is a person in the picture or not
- Example 2: given a noisy received signal, decide whether symbol 1 or 0 was sent by the transmitter
- Example 3: given a set of trials with three alternative medical treatments, decide which treatment is the most effective

Classical v.s. Bayesisan Inference

- Inference using the Bayes rule:
  unknown Θ and observation X are both random variables
  - Find $p_{\Theta|X}$
- Classical statistics: unknown constant θ



  - also for vectors X and θ: $p_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\ldots,\theta_m)$
  - $p_X(x;\theta)$ are NOT conditional probabilities; θ is NOT random
  - mathematically: many models, one for each possible value of θ

MLE

- Joint distribution of the vector of observations $X = (X_1, \ldots, X_n)$: PMF $P_X(x; \theta)$ (or PDF $f_X(x; \theta)$)
- θ: unknown (scalar or vector) parameter θ.
- We observe a particular value $x = (x_1, \ldots, x_n)$ of X, then a **maximum likelihood estimate** (MLE) is a value of the parameter that maximizes the numerical function $P_X(x_1, \ldots, x_n; \theta)$ (or $f_X(x_1, \ldots, x_n; \theta)$) over all θ:

$$\hat{\theta}_n = \arg\max_{\theta} P_X(x_1, \ldots, x_n; \theta)$$

$$\hat{\theta}_n = \arg\max_{\theta} f_X(x_1, \ldots, x_n; \theta)$$

MAP

- Given the observation value x, the MAP rule selects a value $\hat{\theta}$ that maximizes over θ the posterior distribution $p_{\Theta|X}(\theta|x)$ (if Θ is discrete) or $f_{\Theta|X}(\theta|x)$ (if Θ is continuous).
- Equivalently, it selects $\hat{\theta}$ that maximizes over θ:
  - $p_\Theta(\theta)p_{X|\Theta}(x|\theta)$: if both Θ and X are discrete
  - $p_\Theta(\theta)f_{X|\Theta}(x|\theta)$: if Θ is discrete and X is continuous
  - $f_\Theta(\theta)p_{X|\Theta}(x|\theta)$: if Θ is continuous and X is discrete
  - $f_\Theta(\theta)f_{X|\Theta}(x|\theta)$: if both Θ and X are continuous

- An estimator is a random variable of the form $\hat{\Theta} = g(X)$, for some function g. Different choices of g correspond to different estimators.
- An estimate is the value $\hat{\theta}$ of an estimator, as determined by the realized value x of the observation X.
- Once the value x of X is observed, the MAP estimator, sets the estimate $\hat{\theta}$ to a value that maximizes the posterior distribution over all possibly values of θ.

Topic II: Conditional Expectation

- Conditional expectation is a powerful tool for calculating expectations: first-step analysis
- Conditional expectation allows us to predict or estimate unknowns based on whatever evidence is currently available.
- Conditional Expectation given an event: $E(Y|A)$
- Conditional Expectation given a random variable: $E(Y|X)$

Conditional Expectation Given an Event

Definition

Let A be an event with positive probability. If Y is a discrete r.v., then the *conditional expectation of Y given A* is

$$E(Y|A) = \sum_y y \cdot P(Y = y|A) = \sum_y y \cdot P_{Y|A}(y),$$

where the sum is over the support of Y. If Y is a continuous r.v. with PDF f, then

$$E(Y|A) = \int_{-\infty}^{\infty} y \cdot f_{Y|A}(y)\, dy.$$

Conditional Expectation Given an R.V.

Definition

Let $g(x) = E(Y|X = x)$. Then the *conditional expectation of Y given X*, denoted $E(Y|X)$, is defined to be the random variable $g(X)$. In other words, if after doing the experiment X crystallizes into x, then $E(Y|X)$ crystallizes into $g(x)$.

- $E(Y|X)$ is a function of X, and it is a random variable.
- It makes sense to computer $E(E(Y|X))$ and $Var(E(Y|X))$.

Conditional Expectation Properties

Theorem

If X and Y are independent, then $E(Y|X) = E(Y)$.

Theorem

For any function h,

$$E\left(h(X)\,Y|X\right) = h(X)\,E\left(Y|X\right)$$

Theorem

$E(Y_1 + Y_2|X) = E(Y_1|X) + E(Y_2|X)$.

Conditional Expectation & Adam's Law

Theorem

For any r.v.s X and Y,

$$E(E(Y|X)) = E(Y).$$

Theorem

For any r.v.s X, Y, Z, we have

$$E(E(Y|X,Z)|Z) = E(Y|Z)$$
$$E(E(X|Z,Y)|Y) = E(X|Y)$$

Conditional Variance & Eve's Law

Definition

The *conditional variance of Y given X* is

$$\mathrm{Var}\left(Y|X\right) = E\left(\left(Y - E\left(Y|X\right)\right)^2 |X\right).$$

This is equivalent to

$$\mathrm{Var}\left(Y|X\right) = E\left(Y^2|X\right) - \left(E\left(Y|X\right)\right)^2.$$

Theorem

For any r.v.s X and Y,

$$\mathrm{Var}\left(Y\right) = E\left(\mathrm{Var}\left(Y|X\right)\right) + \mathrm{Var}\left(E\left(Y|X\right)\right).$$

The ordering of E's and Var's on the right-hand side spells EVVE, whence the name Eve's law. Eve's law is also known as the law of total variance or the variance decomposition formula.

LOTE

Theorem

Let $A_1, ..., A_n$ be a partition of a sample space, with $P(A_i) > 0$ for all i, and let Y be a random variable on this sample space. Then

$$E(Y) = \sum_{i=1}^{n} E(Y|A_i) P(A_i).$$

LOTUS

Definition

Let A be an event with positive probability and g is a function from $\mathbf{R}$ to $\mathbf{R}$. If Y is a discrete r.v., then the *conditional expectation of* $g(Y)$ *given* A is

$$E(g(Y)|A) = \sum_y \cdot P_{Y|A}(y),$$

where the sum is over the support of Y.
If Y is a continuous r.v. with PDF f, then

$$E(g(Y)|A) = \int_{-\infty}^{\infty} g(y) \cdot f_{Y|A}(y)\, dy.$$

Rethinking with Conditional Expectation I

You toss a fair coin repeatedly. What is the expected number of tosses until the pattern HT appears for the first time? What about the expected number of tosses until HH appears for the first time?

Rethinking with Conditional Expectation II

Suppose we have a stick of length 1 and break the stick at a point X chosen uniformly at random. Given that $X = x$, we then choose another breakpoint Y uniformly on the interval $[0, x]$. Find $E(Y|X)$, and its mean and variance.

Rethinking with Conditional Expectation III

A store receives N customers in a day, where N is an r.v. with finite mean and variance. Let X_j be the amount spent by the jth customer at the store. Assume that each X_j has mean μ and variance σ^2, and that N and all the X_j are independent of one another. Find the mean and variance of the random sum $X = \sum_{j=1}^{N} X_j$, which is the store's total revenue in a day, in terms of μ, σ^2, $E(N)$, and $\mathrm{Var}(N)$.

Topic III: Estimation

- Estimate Y from the observed value X
- Choose the estimator (inference function) $g(\cdot)$ to minimize the expected error $E(c(Y, g(X)))$
- $c(Y, \hat{Y})$ is the cost of guessing $\hat{Y}$ when the actually value is Y.
- When $c(Y, \hat{Y}) = ||Y - \hat{Y}||^2$, the best guess is called "the least square estimate (LSE)" estimate of Y given X.
- Further, if the function $g(\cdot)$ is restricted to be linear, i.e., of the form $a + bX$, it is called "the Linear Least Square Estimate (LLSE)" estimate of Y given X.
- Further, if the function $g(\cdot)$ can be arbitrary, it is called "the Minimum Mean Square Estimate (MMSE)" estimate of Y given X.

Linear Regression

An extremely widely used method for data analysis in statistics is *linear regression*. In its most basic form, the linear regression model uses a single explanatory variable X to predict a response variable Y, and it assumes that the conditional expectation of Y is *linear* in X:

$$E(Y|X) = a + bX.$$

(a) Show that an equivalent way to express this is to write

$$Y = a + bX + \epsilon,$$

where ϵ is an r.v. (called the *error*) with $E(\epsilon|X) = 0$.

(b) Solve for the constants a and b in terms of $E(X)$, $E(Y)$, $\mathrm{Cov}(X, Y)$, and $\mathrm{Var}(X)$.

LLSE

Theorem

The Linear Least Square Estimate (LLSE) of Y given X, denoted by $L[Y|X]$, is the linear function $a+bX$ that minimizes $E[(Y-a-bX)^2]$. In fact,

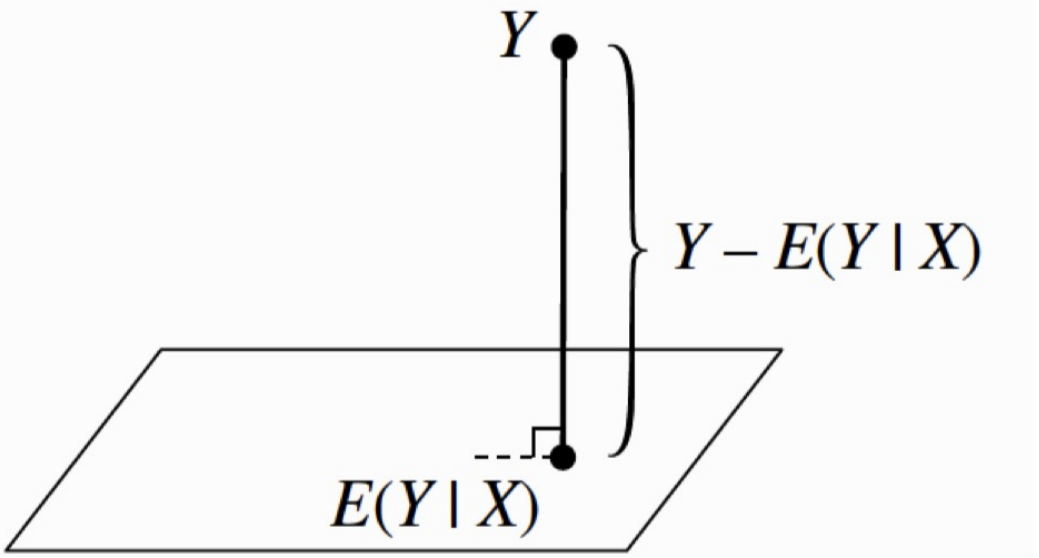
$$L[Y|X] = E(Y) + \frac{Cov(X,Y)}{Var(X)}(X - E(X))$$

MMSE

Theorem

The MMSE of Y given X is given by

$$g(X) = E[Y|X]$$



Theorem

For any function h, the r.v. $Y - E(Y|X)$ is uncorrelated with $h(X)$. Equivalently,

$$E((Y - E(Y|X))h(X)) = 0.$$

(This is equivalent since $E(Y - E(Y|X)) = 0$, by linearity and Adam's law.)

MMSE Properties

Theorem

(a) For any function $\phi(\cdot)$, one has

$$E[(Y - E[Y|X])\phi(X)] = 0$$

(b) Moreover, if the function $g(X)$ is such that

$$E[(Y - g(X))\phi(X)] = 0, \forall \phi(\cdot).$$

then $g(X) = E(Y|X)$

Theorem

Let X, Y be jointly Gaussian random variables. Then

$$E[Y|X] = L[Y|X] = E(Y) + \frac{Cov(X,Y)}{Var(X)}(X - E(X)).$$

Biased Coin: LLSE & MMSE

We wish to estimate the probability of landing heads, denoted by θ, of a biased coin. We model θ as the value of a random variable Θ with a known prior PDF $f_\Theta \sim \mathit{Unif}(0,1)$. We consider n independent tosses and let X be the number of heads observed. Find the MMSE $E(\Theta|X)$ and LLSE $L(\Theta|X)$.

Outline

Problem 3

Let $X_1 \sim \text{Expo}(\lambda_1), X_2 \sim \text{Expo}(\lambda_2)$, and $X_3 \sim \text{Expo}(\lambda_3)$ be independent.

(a) Find $E(X_1|X_1 > 2024)$.

(b) Find $E(X_1|X_1 < 1997)$.

(b) Find $E(X_1 + X_2 + X_3|X_1 > 1997, X_2 > 2014, X_3 > 2025)$ in terms of $\lambda_1, \lambda_2, \lambda_3$.

Problem 3 Solution

Problem 4

Let X and Y be two continuous random variables with joint PDF

$$f_{X,Y}(x,y) = \begin{cases} 6xy & \text{if } 0 \leq x \leq 1, 0 \leq y \leq \sqrt{x} \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the marginal distributions of X and Y. Are X and Y independent?

(b) Find $E[X|Y = y]$ and $\text{Var}[X|Y = y], \forall y \in [0, 1]$.

(c) Find $E[X|Y]$ and $\text{Var}[X|Y]$.

Problem 4 Solution

Problem 5

Let X be a discrete r.v. whose distinct possible values are $x_0, x_1, ...$, and let $p_k = P(X = x_k)$. The entropy of X is $H(X) = \sum_{k=0}^{\infty} p_k \log_2(1/p_k)$.

(a) Find $H(X)$ for $X \sim Geom(p)$

(b) Show that $P(X = Y) \geq 2^{-H(X)}$

Problem 5 Solution

Problem 6

Instead of predicting a single value for the parameter, we give an interval that is likely to contain the parameter: A $1-\delta$ confidence interval for a parameter p is an interval $[\hat{p}-\epsilon, \hat{p}+\epsilon]$ such that $\Pr(p \in [\hat{p}-\epsilon, \hat{p}+\epsilon]) \geq 1-\delta$. Now we toss a coin with probability p landing heads and probability $1-p$ landing tails. The parameter p is unknown and we need to estimate its value from experiment results. We toss such coin N times. Let $X_i = 1$ if the i th result is head, otherwise 0 . We estimate p by using

$$\hat{p} = \frac{X_1 + \ldots + X_N}{N}.$$

Problem 6 Continued

Find the $1-\delta$ confidence interval for p, then discuss the impacts of δ and N.

(a) Method 1: Adopt Chebyshev inequality to find the $1-\delta$ confidence interval for p, then discuss the impacts of δ and N.

(b) Method 2: Adopt Hoeffding bound to find the $1-\delta$ confidence interval for p, then discuss the impacts of δ and N.

(c) Discuss the pros and cons of the above two methods.

Problem 6 Solution

Theorem

Let X have mean μ and variance σ^2. Then for any $a > 0$,

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}.$$

Theorem

Let the random variables $X_1, X_2, \ldots, X_n$ be independent with $E(X_i) = \mu$, $a \leq X_i \leq b$ for each $i = 1, \ldots, n$, where a, b are constants. Then for any $\epsilon \geq 0$,

$$\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

Problem 6 Solution

Problem 6 Solution

Theorem

Let X have mean μ and variance σ^2. Then for any $a > 0$,

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}.$$

Chebyshev's inequality (see Cantelli's inequality for the one-side improvement):

- Pros: 1) sharp bound and cannot be improved in general (given no extra assumption). 2) can be improved with extra distributional information on polynomial moments.
- Cons: 1) requires the existence of moments until the second order. 2) quadratic convergence rate.

Problem 6 Solution

Theorem

Let the random variables $X_1, X_2, \ldots, X_n$ be independent with $E(X_i) = \mu$, $a \leq X_i \leq b$ for each $i = 1, \ldots, n$, where a, b are constants. Then for any $\epsilon \geq 0$,

$$\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

Hoeffding's inequality (see Theorem 2.8 and 2.9 of paper "old and new concentration inequalities" for the one-side improvement):

- Pros: 1) exponential convergence rate. 2) does not require assumption on moments.
- Cons: 1) works only for sub-Gaussian (e.g., bounded random variables). 2) in general not sharp when the variance is small (e.g., see popoviciu's inequality on variances and Bernstein's inequality).