Statistics: Principles, Methods and R (II)

Gao Fengnan[1 2]

27.02.2017

[1]School of Data Science, Fudan University

[2]Shanghai Center for Mathematical Sciences

Overview

Basic Asymptotics Revisited

Bayesian Inference

Before we start

- **Lecturers**: Gao Fengnan
    - Office: N202 Zibin Building
    - Email:
    - Office Hour: 15–17, every Friday afternoon
- **Teaching Assistant**: He Siyuan
    - Email:

Basic Course Information

- Course Objective
  The course covers fundamental aspects of probability and statistical methods and principles. Data illustration using statistical software **R** constitutes an integral part throughout the course, therefore provides the hands-on experience in simulation and data analysis.
- Course Requirement
  - Understand key statistical concepts
  - Be able to program in **R**
  - Complete homework and project assignments
  - Pass the exams

- The topics covered in this course include but are not restricted to:
  EM-algorithm, robustness, Bayesian inference, importance sampling, linear regression, logistic regression, multivariate models, statistical decision theory, clustering, inference for independence, causal inference, graphical models, nonparametric kernel estimation

Basic Course Information—Continued

- Every Monday in the afternoon in HGX306
- The last 20–30 minutes every lecture might be used for solving problems
- Two important exams—the mid-term and final exam.
- Two quizzes, taking place approximately at a quarter and three quarters of the semester.
- For imperative reasons, I will be away for a week or two during the semester, the solutions include
  - finding someone to replace me, or
  - assigning that week to be the mid-term exam week
- A project assignment. Key aspects include
  - Working in teams of 2-3 people
  - A real-world data analysis problem
  - Program in **R**
- The final mark will be a weighted average of all the evaluations, subject to some proper rescaling.
- The evaluations consist of (in decreasing order in importance) final exam, mid-term exam, project and quizzes.

Course References

- Basic references
  - Pawitan, Yudi. In all likelihood: statistical modelling and inference using likelihood. Oxford University Press, 2001.
  - Wasserman, Larry. All of statistics: a concise course in statistical inference. Springer Science & Business Media, 2013.
  - Knight, Keith. Mathematical Statistics. Texts in Statistical Science Series. Boca Raton: Chapman & Hall/CRC Press, ©2000.
  - Wickham, Hadley. ggplot2: elegant graphics for data analysis. Springer, 2016.
- Advanced references
  - Tsybakov, A B. Introduction to Nonparametric Estimation. english ed. Springer Series in Statistics. New York: Springer, ©2009.
  - Van der Vaart, Aad W. Asymptotic statistics. Vol. 3. Cambridge university press, 2000.

In This Course

- Emphasis of the theoretical underpinnings and foundations of statistical inference.
- In the modeling/homework assignments, you will encounter other aspects of statistics, such as **gathering**, **description** and **summarization** of **data**

Basic Asymptotics Revisited

Recapitulation—Different Converges of Random Variables

Definition (Convergence in Distribution)

A sequence $X_1, X_2, \ldots$ of real-valued RV is said to *converge in distribution*, or *converge weakly*, or *converge in law* to a RV X if and only if (IIF)

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for every $x \in \mathbb{R}$ at which F is continuous. Here F_n and F are the distribution functions of RV X_n and X, respectively. If X_n converges to X in distribution, we write

$$X_n \rightsquigarrow X.$$

Definition (Convergence in Probability)

A sequence $X_1, X_2, \ldots$ of real-valued RV is said to *converge in probability* to the RV X IIF for any $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0.$$

If X_n converges to X in probability, we write

$$X_n \xrightarrow{P} X.$$

Definition (Convergence in Distribution)

A sequence $X_1, X_2, \ldots$ of real-valued RV is said to *converge in distribution*, or *converge weakly*, or *converge in law* to a RV X if and only if (IIF)

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for every $x \in \mathbb{R}$ at which F is continuous. Here F_n and F are the distribution functions of RV X_n and X, respectively. If X_n converges to X in distribution, we write

$$X_n \rightsquigarrow X.$$

Definition (Convergence in Probability)

A sequence $X_1, X_2, \ldots$ of real-valued RV is said to *converge in probability* to the RV X IIF for any $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0.$$

If X_n converges to X in probability, we write

$$X_n \xrightarrow{P} X.$$

Definition (Almost Sure Convergence)

A sequence $X_1, X_2, \ldots$ of real-valued RV is said to *converge almost surely* towards X IIF

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1.$$

If X_n converges to X almost surely, we write

$$X_n \xrightarrow{\text{a.s.}} X.$$

Definition (Convergence in Mean)

For some real number $r \geq 1$, $X_1, X_2, \ldots$ *converge in mean* towards X IIF

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^r) = 0,$$

If X_n converges to X in L^r, we write

$$X_n \xrightarrow{L^r} X.$$

Recapitulation—Relations of Different Convergence Modes

$$
\begin{tikzcd}
\text{Almost Sure Convergence} \arrow[dr] & & \text{Mean Convergence} \arrow[dl] \\
& \text{Convergence in Probability} \arrow[d] & \\
& \text{Convergence in Distribution} &
\end{tikzcd}
$$

Relations of Different Convergence Modes of Random Variables

Let X be a real-valued RV, and let $X_1, X_2, X_3, \ldots$ be an infinite sequence of IID copies of X. Let $\overline{X}_n = \left(\sum_{i=1}^n X_i\right)/n$ be the empirical averages of this sequence.

Theorem (Weak Law of Large Numbers)
Suppose that the first moment $\mathbb{E}[|X|]$ of X is finite. Then $\overline{X}_n$ converges in probability to $\mathbb{E}[X]$.

Theorem (Strong Law of Large Numbers)
Suppose that the first moment $\mathbb{E}[|X|]$ of X is finite. Then $\overline{X}_n$ converges almost surely to $\mathbb{E}[X]$.

Theorem (Lindeberg–Lévy CLT)
Suppose that the variance $\sigma^2 := \mathbb{E}[|X - \mathbb{E}[x]|^2]$ is finite and the expectation $\mathbb{E}[X]$ of X is μ. Then as $n \to \infty$, $\sqrt{n}(\overline{X}_n - \mu)$ converges in distribution to a normal law $N(0, \sigma^2)$

$$\sqrt{n}(\overline{X}_n - \mu) \rightsquigarrow N(0, \sigma^2).$$

Why are LLN and CLT important?

- Because the asymptotics enable us to do inference.

  What distinguishes statisticians from computer/data scientists are not estimation. Anyone may propose estimators and sometimes they are good, but only statisticians can do inference.

- Suppose we observe IID sequence $X_1, X_2, \ldots, X_n \sim N(\mu, 1)$ and would like to estimate μ, the maximum likelihood estimator (MLE) gives

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- But how to determine the quality of the

- By the strong LLN, $\hat{\mu}_n \xrightarrow{\text{a.s.}} \mu$
- How to construct a $(1-\alpha)$-confidence interval for μ?
- By the CLT, $\sqrt{n}(\hat{\mu}_n - \mu) \rightsquigarrow N(0,1)$, then

$$\mathbb{P}(\sqrt{n}|\hat{\mu}_n - \mu| > z_{\alpha/2}) \to \alpha.$$

- With approximately probability $1-\alpha$, $\mu \in (\hat{\mu}_n \pm z_{\alpha/2}/\sqrt{n})$

Theorem (Hoeffding's Inequality)

Let $X_1, \dots, X_n$ be independent RV*'s such that X_i takes value in $[a_i, b_i]$ almost surely for all $i \le n$. Let $S = \sum_{i=1}^n (X_i - \mathbb{E}[X_i])$. Then for every $t > 0$,*

$$\mathbb{P}(S \ge t) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Lemma (Hoeffding's lemma)

Let Y be a RV *with $\mathbb{E}[Y] = 0$, taking value in a bounded interval $[a, b]$. Then $\log \mathbb{E}[e^{\lambda Y}] \le \lambda^2 (b-a)^2/8$.*

Proof of both Hoeffding's lemma and inequality.

On the blackboard. $\square$

Application of Hoeffding's Inequality—Nonasymptotic Inference on Sample Mean

- Take X_i's to be IID RV's with value only from $[b, a]$
- Estimate the expecation $\mathbb{E}[X]$ with sample mean $\overline{X}_n$
- Hoeffding's inequaity says

$$\mathbb{P}(\sqrt{n}\,(\overline{X}_n - p) \geq t) \leq \exp\left(-\frac{2t^2}{(b-a)^2}\right).$$

Bayesian Inference

frequentist	Bayesian
Probability Refers to limiting relative frequencies. Probabilities are objective properties of the real world.	Probability describes degrees of belief, not limiting frequency.
Parameters are fixed, unknown constants.	We can make probability statements about parameters, even though they are fixed constants.
Statistical procedure should be designed to have well-defined long run frequency properties.	We make inferences about a parameter θ by producing a probability distribution for θ.

Frequentist v.s. Bayesian

The Bayesian Method

Bayesian inference is usually carried out in the following steps.

1. Choose a probability density $\pi(\theta)$—the *prior distribution*—to express our beliefs about a parameter θ before any data
2. Choose a statistical model $f(x|\theta)$ that reflects our belief about x given θ
3. After observing data $X_1, \ldots, X_n$, we update our beliefs and calculate the *posterior distribution* $\pi(\theta|X_1, \ldots, X_n)$

Recall Bayes' theorem

Theorem (Bayes' Theorem)

For two events A and B with $\mathbb{P}(B) \neq 0$

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Bayesian Procedure

Keep in mind that the parameter θ is random!

- Θ–the parameter, X—data
- Suppose θ only takes discrete values,

$$\mathbb{P}(\Theta = \theta | X = x) = \frac{\mathbb{P}(X = x, \Theta = \theta)}{\mathbb{P}(X = x)} = \frac{\mathbb{P}(X = x | \Theta = \theta)\mathbb{P}(\Theta = \theta)}{\sum_\theta \mathbb{P}(X = x | \Theta = \theta)\mathbb{P}(\Theta = \theta)}$$

- Suppose continuous θ, we use density function

$$\pi(\theta | x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}.$$

- Suppose n IID observations $X^{(n)} := \{X_1, \ldots, X_n\}$ and write non-random $x^{(n)} = \{x_1, \ldots, x_n\}$, then the likelihood function is

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = L_n(\theta).$$

Bayesian Procedure Continued

- We get

$$\pi(\theta|x^{(n)}) = \frac{f(x^{(n)}|\theta)\pi(\theta)}{\int f(x^n|\theta)\pi(\theta)d\theta} = \frac{L_n(\theta)\pi(\theta)}{c_n} \propto L_n(\theta)\pi(\theta)$$

  where $c_n = \int L_n(\theta)\pi(\theta)d\theta$ is called the normalizing constant.
- Posterior is proportional to Likelihood times Prior.
- With $L_n(\theta)\pi(\theta)$, c_n can always be recovered.
- Compare with normal distribution, the density is proportional to $\exp\left(-x^2/(2\sigma^2)\right)$, we can recover the full density by calculating the integral

$$\int \exp\left(-x^2/(2\sigma^2)\right)dx.$$

Example (Bernoulli Experiment)

Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, how to estimate p?

- The MLE gives $\hat{p}_n = \overline{X}_n$
- The Bayesian way—specify a prior π on p first—a density taking value on all possible p's
- We take uniform prior on $[0, 1]$, i.e., $\pi(p) = 1_{[0,1]}(p)$
- Any other possible prior for p?