STAT575
Lrage Sample Theory

Pingbang Hu

January 16, 2024

Abstract

This is a graduate-level theoretical statistics course taught by Georgios Fellouris at University of Illinois Urbana-Champaign, aiming to provide an introduction to asymptotic analysis of various statistical methods, including weak convergence, Lindeberg-Feller CLT, asymptotic relative efficiency, etc.

We list some references of this course, although we will not follow any particular book page by page: *Asymnptotic Statistics* [Vaa98], *Asymptotic Theory of Statistics and Probability* [Das08], *A course in Large Sample Theory* [Fer17], *Approximation Theorems of Mathematical Statistics* [Ser09], and *Elements of Large-Sample Theory* [Leh04].

This course is taken in Spring 2024, and the date on the cover page is the last updated time.

Contents

Chapter 1

Introduction

Lecture 1: Introduction to Large Sample Theory

Say we first collect n data points $x_1, \ldots, x_n \in \mathbb{R}^d$, large sample theory concern with problem that when $n \to \infty$. We may treat x_i as a realization of random vector X_i on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. In this course, we will primarily consider the case that X_i's are i.i.d., i.e., independent and identically distributed from a distribution function (CDF) F such that 16 Jan. 9:30

$$X = (X^1, \ldots, X^d) \sim F(x_1, \ldots, x_d) \equiv \mathbb{P}(X^1 \leq x_1, \ldots, X^d \leq x_d)$$

for all $x \in \mathbb{R}$. If we have access to F, we can compute (PDF) $\mathbb{P}(X \in A)$ for all (measurable) $A \subseteq \mathbb{R}^d$ of interest. If we know this, we know every thing about the population. Hence, the goal is to compute this by collecting data x_i's, i.e., statistical inference problem.

1.1 Parametrized Approach

By postulate a family of CDFs $\{F_\theta, \theta \in \Theta\}$ where Θ is often a subset of $\mathbb{R}^m$ for some m (generally $\neq n$), and select the member of this family that is the "closet" or the "best fit" to the truth, i.e., F, based on the data. To emphasize that this depends on the data, we sometimes write the function we found as $\hat{\theta}_n(x_1, \ldots, x_n)$ so that $F_{\hat{\theta}_n(x_1, \ldots, x_n)}$ is my proxy for F.

Now, assume that the family is initially given, the problem is then how to select $\hat{\theta}_n$. Fisher suggested that we should look at the maximum likelihood estimator (MLE). The justification for MLE is not about finite n, but really about the asymptotic behavior when $n \to \infty$. Specifically, we have the following (informally).

> **Theorem 1.1.1** (Fisher). If $F \in \{F_\theta \colon \theta \in \Theta\}$, i.e., if $F = F_{\theta^*}$ for some $\theta^* \in \Theta$, then under certain conditions, $\hat{\theta}_n$ will be "close" to θ^* as $n \to \infty$. Under some other conditions, we can say that $\sqrt{n}(\hat{\theta}_n - \theta)$ is approximately Gaussian with variance is the "best possible" in some sense.

On the other hand, in the misspecified case, i.e., $F \notin \{F_\theta, \theta \in \Theta\}$, we can still compute the MLE. Another justification for MLE is that even in this case, $\hat{\theta}_n$ will still be "close" to θ^* such that F_{θ^*} is, in some sense, the "closest" to F among all possible F_θ.

1.2 Hypothesis Testing

We will also develop theory for hypothesis testing for some hypothesis we're interested in, e.g., whether the data we collect is really i.i.d., or whether our proposed family is reasonable enough.

Say now X_i's are scalar random variable with $\mathbb{E}[X] = \mu$, and we want to test the null hypothesis $H_0 \colon \mu = 0$.

> **Example.** Say we have a controlled group and a treatment group, and we observe $Z_1, \ldots, Z_n$, and $Y_1, \ldots, Y_n$, respectively, and compute $X_i = Z_i - Y_i$ for all i.

Let s_n to be the sample standard derivation, then we can compute

$$T_n = \frac{\overline{X}_n}{s_n/\sqrt{n}} \sim t_{n-1}$$

as long as X is Gaussian, i.e., the t-test for H_0. What if X is not an Gaussian? We will show that even if X is not Gaussian, this result is "approximately valid" when n is "large enough" as long as $\operatorname{Var}[X] < \infty$.

1.3 Sample Size

When we say n is "large enough", it really depends on how fast the underlying distribution will approach Gaussian as n grows. Hence, if we can say more about the underlying population, we can say more about when does n is "large enough"; otherwise the theory might be completely useless.

What if now $\operatorname{Var}[X]$ doesn't exit?

> **Example** (Heavy tail). When the population has a heavy tail distribution, then second moment may not exit.

We might reject H_0 when $\sum_{i=1}^{n} \mathbb{1}_{X_i>0}$ is large... Note that under H_0, $\sum_{i=1}^{n} \mathbb{1}_{X_i>0} \sim \operatorname{Bin}(n, 1/2)$. This test is valid even if expectation doesn't exist. We see that without saying anything about F, this test is valid even for $n = 3$ or 5. Now the question becomes how can we compare these different tests? This will also be addressed in this course.

Consider $R_{i,n}$ to be the rank of $|X_i|$, then consider

$$\sum_{i=1}^{n} \mathbb{1}_{X_i>0} R_{i,n}.$$

This is the so-called Wilcoxon's Rank-Sum test. As one can imagine, the closed form will be complicated; however, asymptotically, the above statics will follow Gaussian again, such that the rate of convergence doesn't depend on the underlying population.

Appendix

Bibliography

[Das08] Anirban DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer Science & Business Media, Feb. 6, 2008. 727 pp. ISBN: 978-0-387-75971-5. Google Books: sX4_AAAAQBAJ.

[Fer17] Thomas S. Ferguson. *A Course in Large Sample Theory*. Routledge, Sept. 6, 2017. 140 pp. ISBN: 978-1-351-47005-6. Google Books: clcODwAAQBAJ.

[Leh04] E. L. Lehmann. *Elements of Large-Sample Theory*. Springer Science & Business Media, Aug. 27, 2004. 640 pp. ISBN: 978-0-387-98595-4. Google Books: geIoxvgTXlEC.

[Ser09] Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, Sept. 25, 2009. 399 pp. ISBN: 978-0-470-31719-8. Google Books: enUouJ4EHzQC.

[Vaa98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 1998. ISBN: 978-0-521-78450-4. DOI: 10.1017/CBO9780511802256. URL: https://www.cambridge.org/core/books/asymptotic-statistics/A3C7DAD3F7E66A1FA60E9C8FE132EE1D (visited on 10/17/2023).