B04901098 電機三 陳立富

統計學報告

Approximation Theorems of Mathematical Statistics

Introduction

U-statistics, in which 'U' stands for 'unbiased,' generalize common notions of unbiased estimation such as the sample mean and the sample variance and thus play a vital role in estimation. In this chapter, it describes the basic description, martingale structure, and the projection of a U-statistic.

Basic Description

We first introduce the concept of statistical functional and use it to comprehend the definition of U-statistics and the symmetry of a 'kernel,' denoted by $h^*(x_1, x_2, \cdots, x_n)$

Let S be a set of cumulative distribution functions and let T denote a mapping from S into the real numbers R. Then T is called a statistical functional. The functional T can be regarded as $\theta(F)$, the parameter of interest.

Next, to see the relation between the U-statistics and the statistical functional T, we define what is called an expectation functional in the following formula.

$$T(F) = E_F[h(X_1, X_2, \cdots, X_a)]$$

The above formula means the expectation of $h(X_1, X_2, \cdots, X_a)$, where $X_1, X_2, \cdots, X_a$ is a simple random sample from the distribution function F. From the i.i.d. property of X_i, it is not hard to see that for any permutation π mapping $\{1, . . . , a\}$ onto itself, we have

$$E_F[h(X_1, X_2, \cdots, X_a)] = E_F[h(X_{\pi(1)}, X_{\pi(2)}, \cdots, X_{\pi(a)})]$$

Since there are a! such permutations, it may be reasonable to consider the following function

$$h^*(x_1, x_2, \cdots, x_a) = \frac{1}{a!}\sum_{all\ \pi} h(x_{\pi(1)}, x_{\pi(2)}, \cdots, x_{\pi(a)})$$

$$E_F[h(X_1, X_2, \cdots, X_a)] = E_F[h^*(X_1, X_2, \cdots, X_a)]$$

It is obvious that $h^*(x_1, x_2, \cdots, x_a)$ is symmetric in its arguments and call it the kernel function associated with $T(F)$.

Finally, we can use the above concepts to define the U-statistics as follow:

Let $h^*(x_1, x_2, \cdots, x_a)$ be the kernel function associated with an expectation functional $T(F)$, the U-statistic corresponding to this functional equals

$$U_n = \frac{1}{\binom{n}{a}}\sum_{c} h^*(X_{i_1}, X_{i_2}, \cdots, X_{i_a})$$

, where c denotes the combination of a distinct elements $\{i_1, i_2, \cdots, i_a\}$ from $\{1,2,\cdots, n\}$

Example

① $T(F)$= mean of F

Since $\mu = E[X]$, the kernel function is $h^*(x) = x$.

The corresponding U-statistic is $U_n = \frac{1}{n}\sum_{i=1}^{n} h^*(X_i) = \frac{1}{n}\sum_{i=1}^{n} X_i$

② $T(F)$= variance of F

$\sigma^2 = E[(X - E[X])^2] = E[\frac{1}{2}\{(X_1 - E[X_1])^2 + (X_2 - E[X_2])^2\}] = E[\frac{1}{2}\{(X_1 - E[X_1]) - (X_2 - E[X_2])\}^2] = E[\frac{1}{2}(X_1 - X_2)^2]$

The kernel function is $h^*(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2$.

The corresponding U-statistic is

$U_n = \frac{2}{n(n-1)}\sum_{1\le i<j\le n} h^*(X_1, X_2) = \frac{2}{n(n-1)}\sum_{1\le i<j\le n} \frac{1}{2}(X_i - X_j)^2$

Next we show that U_n is the conventional sample variance S_n^2.

$U_n = \frac{1}{n(n-1)}\sum_{1\le i<j\le n}(X_i - X_j)^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}(X_i - X_j)^2 = \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}[(X_i - \bar{X}_n + \bar{X}_n - X_j)^2]$

$= \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}[(X_i - \bar{X}_n)^2 + (X_j - \bar{X}_n)^2] = \frac{1}{n-1}\{\sum_{i=1}^{n}\frac{1}{2}(X_i - \bar{X}_n)^2 + \sum_{j=1}^{n}\frac{1}{2}(X_j - \bar{X}_n)^2\} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

$= S_n^2$

Property

We can use the Rao-Blackwell theorem to prove that

$$Var(U(X_1, X_2, \cdots, X_n)) \le Var(h^*(X_1, X_2, \cdots, X_a))$$

The key here lies in the definition of U-statistics. Since $U_n = \frac{1}{\binom{n}{a}}\sum_{c} h^*(X_{i_1}, X_{i_2}, \cdots, X_{i_a})$ is an average over the permutations π of $h(x_{\pi(1)}, x_{\pi(2)}, \cdots, x_{\pi(a)})$, we can write it as the result of conditioning on the order statistic.

That is, $U(X_1, X_2, \cdots, X_n) = E[h^*(X_1, X_2, \cdots, X_a) \mid X_{(1)}, X_{(2)}, \cdots, X_{(n)}]$

$Var(h^*(X_1, X_2, \cdots, X_a)) = Var(E[h^*(X_1, X_2, \cdots, X_a) \mid X_{(1)}, X_{(2)}, \cdots, X_{(n)}]) + E[Var(h^*(X_1, X_2, \cdots, X_a \mid X_{(1)}, X_{(2)}, \cdots, X_{(n)})]$

Therefore, we have $Var(U(X_1, X_2, \cdots, X_n)) \le Var(h^*(X_1, X_2, \cdots, X_a))$

Variance of U-Statistics

$U_n = \frac{1}{\binom{n}{a}}\sum_{c} h^*(X_{i_1}, X_{i_2}, \cdots, X_{i_a})$, where c denotes the combination of a distinct elements $\{i_1, i_2, \cdots, i_a\}$ from $\{1,2,\cdots, n\}$.

$Var(U_n) = \frac{1}{\binom{n}{a}^2}\sum_{S}\sum_{S'} Cov(h^*(X_S), h^*(X_{S'}))$, where S, S' range over the combinations of a distinct elements $\{i_1, i_2, \cdots, i_a\}$ from $\{1,2,\cdots, n\}$.

Hence, we can further write the variance as $\frac{1}{\binom{n}{a}^2}\sum_{c=1}^{a}\binom{n}{a}\binom{a}{c}\binom{n-a}{a-c}\zeta_c$, where c means the size of the intersection of S and S', [1] $\binom{n}{a}\binom{a}{c}\binom{n-a}{a-c}$ means the number of ways of choosing S and S' with $|S \cap S'| = c$, and ζ_c means $Cov(h^*(X_S), h^*(X_{S'}))$.

Next, we explain why it is suitable to denote $Cov(h^*(X_S), h^*(X_{S'}))$ by ζ_c. That is, for any S and S' with $|S \cap S'| = c$, $Cov(h^*(X_S), h^*(X_{S'}))$ is the same and thus denoted by ζ_c.

For $k = 1,2,\cdots,a$, let

$h_k^*(x_1, x_2, \cdots, x_k) = E[h^*(X_1, X_2, \cdots, X_a) \mid X_1 = x_1, \cdots, X_k = x_k] = E[h^*(x_1, x_2, \cdots, x_k, X_{k+1}, \cdots, X_a)]$We have that $h_a^* = h^*$.

It can be shown by the iterated expectation that
$E_F[h_k^*(X_1, X_2, \cdots, X_k)] = E_F[h^*(X_1, X_2, \cdots, X_a)] = T(F) = \theta(F)$

Given S={$i_1, i_2, \cdots, i_a$}and S'={$j_1, j_2, \cdots, j_a$} are two combinations of a distinct elements from $\{1,2,\cdots,n\}$ with c elements in common, then

$Cov(h^*(X_{i_1}, X_{i_2}, \cdots, X_{i_a}), h^*(X_{j_1}, X_{j_2}, \cdots, X_{j_a}))$
$= E[(h^*(X_1, X_2, \cdots, X_c, X_{c+1}, \cdots, X_a) - \theta)(h^*(X_1, X_2, \cdots, X_c, X'_{c+1}, \cdots, X'_a) - \theta)]$ by the symmetry of h^*

Conditioning on $X_1, X_2, \cdots, X_c$, the two terms in this expectation are independent, so taking the expectation of the conditional expectation, we have

$Cov(h^*(X_{i_1}, X_{i_2}, \cdots, X_{i_a}), h^*(X_{j_1}, X_{j_2}, \cdots, X_{j_a}))$=

$$= E[E[(h^*(X_1, X_2, \cdots, X_c, X_{c+1}, \cdots, X_a) - \theta)(h^*(X_1, X_2, \cdots, X_c, X'_{c+1}, \cdots, X'_a) - \theta) \mid X_1, X_2, \cdots X_c]]$$

$$= E[E[h^*(X_1, X_2, \cdots, X_c, X_{c+1}, \cdots, X_a) - \theta \mid X_1, X_2, \cdots X_c]E[h^*(X_1, X_2, \cdots, X_c, X'_{c+1}, \cdots, X'_a) - \theta \mid X_1, X_2, \cdots X_c]]$$
$$= E[(h_c^*(X_1, X_2, \cdots, X_c) - \theta)(h_c^*(X_1, X_2, \cdots, X_c) - \theta)]$$

$$= \zeta_c$$

Last but not least, let's recall that in the midterm, the first problem asks us to express $Var(S_n^2)$ in terms of the moments of F. After noticing that S_n^2 is a U-statistic, we could apply the formulas derived above to answer the question.

$$h^*(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2$$

$$\zeta_1 = Cov(h^*(X_1, X_2), h^*(X_1, X_3))$$

$$= E[(\frac{1}{2}(X_1 - X_2)^2 - \sigma^2)(\frac{1}{2}(X_1 - X_3)^2 - \sigma^2)] = E[(\frac{1}{2}(X_1 - \mu + \mu - X_2)^2 - \sigma^2)(\frac{1}{2}(X_1 - \mu + \mu - X_3)^2\sigma^2)]$$

$$= E[(\frac{1}{2}((X_1 - \mu)^2 + (X_2 - \mu)^2 - 2(X_1 - \mu)(X_2 - \mu) - 2\sigma^2)(\frac{1}{2}((X_1 - \mu)^2 + (X_3 - \mu)^2 - 2(X_1 - \mu)(X_3 - \mu) - 2\sigma^2)]$$

$$= \frac{1}{4}(\mu_4 + \sigma^4 - 2\sigma^4 + \sigma^4 + \sigma^4 - 2\sigma^4 - 2\sigma^4 - 2\sigma^4 + 4\sigma^4) = \frac{1}{4}(\mu_4 - \sigma^4), where\ \mu_4 = E[(X - \mu)^4]$$

---

[1] First choose S, then choose the intersection from S, then choose the non-intersection for the rest of S'

$\zeta_2 = Cov(h^*(X_1, X_2), h^*(X_1, X_2))$

$$= E[(\frac{1}{2}(X_1 - X_2)^2 - \sigma^2)(\frac{1}{2}(X_1 - X_2)^2 - \sigma^2)] = E[(\frac{1}{2}(X_1 - \mu + \mu - X_2)^2 - \sigma^2)(\frac{1}{2}(X_1 - \mu + \mu - X_2)^2\sigma^2)]$$

$$= E[(\frac{1}{2}((X_1 - \mu)^2 + (X_2 - \mu)^2 - 2(X_1 - \mu)(X_2 - \mu) - 2\sigma^2)(\frac{1}{2}((X_1 - \mu)^2 + (X_2 - \mu)^2 - 2(X_1 - \mu)(X_2 - \mu) - 2\sigma^2)]$$

$$= \frac{1}{4}(\mu_4 + \sigma^4 - 2\sigma^4 + \sigma^4 + \mu_4 - 2\sigma^4 + 4\sigma^4 - 2\sigma^4 - 2\sigma^4 + 4\sigma^4) = \frac{1}{2}(\mu_4 + \sigma^4)$$

Hence,

$$Var(S_n^2) = \frac{1}{\binom{n}{2}^2}\sum_{c=1}^{2}\binom{n}{2}\binom{2}{c}\binom{n-2}{2-c}\zeta_c = \frac{1}{\binom{n}{2}}\sum_{c=1}^{2}\binom{2}{c}\binom{n-2}{2-c}\zeta_c = \frac{2}{n(n-1)}[2(n-2)\frac{1}{4}(\mu_4 - \sigma^4) + \frac{1}{2}(\mu_4 + \sigma^4)]$$

$$= \frac{\mu_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)}$$

To obtain an expression of $Var(S_n^2)$ in terms of μ and the moments μ_2', μ_3', μ_4' about the zero, we can just substitute the following formulas into the above expression:

$$\sigma^2 = \mu_2' - \mu^2,\ \mu_4 = \mu_4' - 4\mu\mu_3' + 6\mu^2\mu_2' - 3\mu^4$$

After several calculations and simplification, we get the solution

$$Var(S_n^2) = \frac{1}{n}\mu_4' - \frac{4}{n}\mu\mu_3' - \frac{n-3}{n(n-1)}\mu_2'^2 + \frac{4(2n-3)}{n(n-1)}\mu^2\mu_2' - \frac{2(2n-3)}{n(n-1)}\mu^4$$

Reference

1. Serfling, R. J.(1980). Approximation Theorems of Mathematical Statistics. Chap5 5.1-5.3
2. http://personal.psu.edu/drh20/asymp/fall2006/lectures/
3. Peter L. Bartlett's lectures for the course of Theoretical Statistics
   https://www.stat.berkeley.edu/~bartlett/courses/2013spring-stat210b/