THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

SDS 384 11: Theoretical Statistics

Lecture 9: U Statistics cont.

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin

U Statistics

- We will see many interesting examples of U statistics.
- Interesting properties
  - Unbiased (done)
  - Reduces variance (done)
  - Concentration (via McDiarmid) (done)
  - Asymptotic variance
  - Asymptotic distribution

Variance of U statistic

- Consider a U Statistic of order r.

$$U = \frac{\sum_{\{i_1,\ldots,i_r\}\in\mathcal{I}_r} h(X_{i_1},\ldots,X_{i_r})}{\binom{n}{r}}$$

- Let $S, S' \in \mathcal{I}_r$.

$$\begin{aligned}\mathrm{var}(U) &= \frac{1}{\binom{n}{r}^2}\sum_{S,S'}\mathrm{cov}(h(X_S), h(X_{S'}))\\ &= \frac{1}{\binom{n}{r}^2}\sum_{c=0}^{r}\underbrace{\binom{n}{r}\binom{r}{c}\binom{n-r}{r-c}}_{Y_c}\xi_c,\end{aligned}$$

  - Assume that two subsets A, B have c elements in common.
  - Y_c is the number of ways to choose A, choose the intersection $A \cap B$ and then choose the rest of B, i.e. $B \setminus A$.
  - ξ_c will be defined now.

Variance of U statistic

- ξ_c is defined as $\text{cov}(h(X_S), h(X_{S'})$.
- Let $I := S \cap S'$ and $|I| = c$

$$\begin{aligned}\xi_c &:= \text{cov}(h(X_S), h(X_{S'})) \\ &= \text{cov}(h(X_I, X_{S\setminus I}), h(X_I, X_{S'\setminus I}))\end{aligned}$$

- $$\begin{aligned}&= \text{cov}(E[h(X_I, X_{S\setminus I}|X_I)], E[h(X_I, X_{S'\setminus I}|X_I)]) \\ &+ E[\text{cov}(h(X_I, X_{S\setminus I}), h(X_I, X_{S'\setminus I})|X_I)] \\ &= \text{var}(E[h(X_I, X_{S\setminus I}|X_I)]) \geq 0\end{aligned}$$

Variance of U statistic

$$
\begin{aligned}
\operatorname{var}(U) &= \frac{1}{\binom{n}{r}^2} \sum_{c=0}^{r} \underbrace{\binom{n}{r}\binom{r}{c}\binom{n-r}{r-c}}_{Y_c} \xi_c \\
&= \frac{1}{\binom{n}{r}} \sum_{c=0}^{r} \underbrace{\binom{r}{c}\binom{n-r}{r-c}}_{Y_c} \xi_c \\
&= \sum_{c=1}^{r} \frac{r!^2}{c!(r-c)!^2} \frac{\overbrace{(n-r)\dots(n-2r+c+1)}^{(n-r)^{r-c}}}{\underbrace{n(n-1)\dots(n-r+1)}_{n^r}} \xi_c \\
&= \frac{r^2}{n}\xi_1 + o(1/n)
\end{aligned}
$$

Variance of U statistic - example

Example

Let $h(x,y) = (x-y)^2/2$ and $\theta = \sigma^2$. The variance of the corresponding U statistics, aka the sample variance is given by $\frac{\mu_4 - \sigma^4}{n}$, where $\mu_4 := E[(X-\mu)^4]$.

- We will need ξ_1.
  $$\xi_1 := \mathrm{cov}(h(X_1, X_2), h(X_1, X_3))$$
  $$= \mathrm{cov}(E[h(X_1, X_2)|X_1], E[h(X_1, X_3)|X_1])$$
  - We have $E[h(X_1, X_2)|X_1] = E[(X_1 - X_2)^2|X_1]/2 = ((X_1 - \mu)^2 + \sigma^2)/2$
- So,
  $$\xi_1 := \frac{\mathrm{var}(X_1 - \mu)^2}{4} = \frac{E(X_1 - \mu)^4 - \sigma^4}{4} = \frac{\mu_4 - \sigma^4}{4}$$

Variance of U statistic-example

Example

Let $h(x, y) = xy$ and $\theta = \mu^2$. The variance of the corresponding U statistics, is given by $\frac{4\mu^2\sigma^2}{n}$.

- $E[h(X_1, X_2)|X_1] = \mu X_1$
- $\xi_1 := \text{var}(E[h(X_1, X_2)|X_1]) = \mu^2\sigma^2$

Normal Convergence of U statistics

Theorem

If $E[h^2] < \infty$, *we have*

$$\sqrt{n}(U - \theta) \xrightarrow{d} N(0, r^2 \xi_1).$$

- We will prove this using Hajek Projections.
- What happens when the limiting variance is zero?

Normal Convergence of U statistics-example

Example

Recall the U statistics associated with the Wilcoxon signed rank test. The kernel is $h(x, y) = 1(x + y > 0)$ and the parameter estimated is $\theta = P(X_1 + X_2 > 0)$. Under the null hypothesis that the underlying distribution is continuous and symmetric about 0, we have

$$\sqrt{n}(U - 1/2) \xrightarrow{d} N(0, 1/3)$$

- Under the null, $\theta = P(X_1 + X_2 > 0) = 1/2$

$$\begin{aligned}\xi_1 &= \mathrm{cov}(h(X_1, X_2), h(X_1, X_3)) = P(X_1 + X_2 > 0, X_1 + X_3 > 0) - \theta^2 \\ &= P(X_1 > -X_2, X_1 > -X_3) - 1/4 = P(X_1 > X_2, X_1 > X_3) - 1/4 \\ &= 1/3 - 1/4 = 1/12\end{aligned}$$

Convergence of U statistics-example

Example

Let $h(x, y) = xy$ and $\theta = \sigma^2$. Let $E[X^2] < \infty$. Then $\sqrt{n}(U - \mu^2) \xrightarrow{d} N(0, 4\xi_1)$, where $\xi_1 := \frac{\mu^2\sigma^2}{n}$.

- Say $\mu = 0$. Now what?
- This is called a degenerate U statistics.
- The variance of it is now $O(1/n^2)$, since $\xi_1 = 0$
- But is there a distributional convergence?

Convergence of U statistics-example

Example

Let $h(x, y) = xy$ and $\theta = \sigma^2$. Let $E[X^2] < \infty$. Then $\sqrt{n}(U - \mu^2) \xrightarrow{d} N(0, 4\xi_1)$, where $\xi_1 := \dfrac{\mu^2\sigma^2}{n}$.

$$
\begin{aligned}
U &= \frac{\sum_{i<j} X_i X_j}{\binom{n}{2}} = \frac{\sum_{i \neq j} X_i X_j}{n(n-1)} \\
&= \frac{(\sum_i X_i)^2 - \sum_i X_i^2}{n(n-1)} \\
&= \frac{(\sqrt{n}\bar{X}_n)^2 - \sum_i X_i^2/n}{n-1}
\end{aligned}
$$

$$
(n-1)U \xrightarrow{d} (Z^2 - 1)\sigma^2, \text{where } Z \sim N(0,1)
$$

Normal Convergence of U statistics-proof

Next time!