TOPICS IN TWO-SAMPLE TESTING

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Nelson C. Ray

201?

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Susan P. Holmes) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Persi W. Diaconis)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Bradley Efron)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Jerome H. Friedman)

Approved for the University Committee on Graduate Studies.

Contents

Chapter 1

Stein's method

In this chapter we present an introduction to Stein's method of exchangeable pairs which we use to prove the core theoretical result of this thesis: a rate of convergence bound for the randomization distribution.

1.1 Introduction

Stein's method provides a means of bounding the distance between two probability distributions in a given probability metric. When applied with the normal distribution as the target, this results in central limit type theorems. Several flavors of Stein's method (e.g. the method of exchangeable pairs) proceed via auxiliary randomization. We reproduce Stein's proof of the Hoeffding combinatorial central limit theorem (HCCLT) with explicit calculation of various constants. It will be instructive to follow the proof of the HCCLT because our proof proceeds in a similar fashion but with the following generalizations: an approximate contraction property, less cancellation of terms due to separate estimation of various denominators, and non-unit variance of an r.v. in the exchangeable pair.

1.2 Hoeffding combinatorial CLT

Theorem 1.1. *Let $\{a_{ij}\}_{i,j}$ be an $n \times n$ matrix of real-valued entries that is row- and column-centered and scaled such that the sums of the squares of its elements equals $n-1$:*

$$\sum_{j=1}^{n} a_{ij} = 0 \tag{1.1}$$

$$\sum_{i=1}^{n} a_{ij} = 0 \tag{1.2}$$

$$\sum_{i=1,j=1}^{n} a_{ij}^2 = n-1 \tag{1.3}$$

Let Π be a random permutation of $\{1,\dots,n\}$ drawn uniformly at random from the set of all permutations:

$$P(\Pi = \pi) = \frac{1}{n!}. \tag{1.4}$$

Define

$$W = \sum_{i=1}^{n} a_{i\Pi(i)} \tag{1.5}$$

to be the sum of a random diagonal. Then

$$|P(W \le w) - \Phi(w)| \le \frac{C}{\sqrt{n}} \left[\sqrt{\sum_{i,j=1}^{n} a_{ij}^4} + \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^3} \right]. \tag{1.6}$$

Proof. In order to construct our exchangeable pair, we introduce the ordered pair of random variables (I,J) independent of Π that represents a uniformly at random draw from the set of all non-null transpositions:

$$P(I = i, J = j) = \frac{1}{n(n-1)} \quad i,j \in \{1,\dots,n\}, i \ne j. \tag{1.7}$$

Define the random permutation Π' by

$$\Pi'(i) = \Pi \circ (I, J) = \begin{cases} \Pi(J) & i = I \\ \Pi(I) & i = J \\ \Pi(i) & \text{else.} \end{cases} \tag{1.8}$$

We construct our exchangeable pair by defining

$$W' = \sum_{i=1}^{n} a_{i\Pi'(i)} = W - a_{I\Pi(I)} + a_{I\Pi(J)} - a_{J\Pi(J)} + a_{J\Pi(I)}. \tag{1.9}$$

We now verify the contraction property:

$$\begin{aligned}
\mathbb{E}[W - W'|\Pi] &= \mathbb{E}[a_{I\Pi(I)} - a_{I\Pi(J)} + a_{J\Pi(J)} - a_{J\Pi(I)}|\Pi] && (1.10)\\
&= \frac{2}{n}\sum_{i=1}^{n} a_{i\Pi(i)} - \frac{2}{n}\frac{1}{n-1}\sum_{i,j=1,i\neq j}^{n} a_{i\Pi(j)} && (1.11)\\
&= \frac{2}{n}W - \frac{2}{n}\frac{1}{n-1}\left[\sum_{i,j=1}^{n} a_{i\Pi(j)} - \sum_{i}^{n} a_{i\Pi(i)}\right] && (1.12)\\
&= \frac{2}{n}W + \frac{2}{n}\frac{1}{n-1}W - \frac{2}{n}\frac{1}{n-1}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i\Pi(j)}\right] && (1.13)\\
&= \frac{2}{n}W\left(1 + \frac{1}{n-1}\right) - 0 && (1.14)\\
&= \frac{2}{n-1}W && (1.15)
\end{aligned}$$

This satisfies our contraction property with

$$\lambda = \frac{2}{n-1}. \tag{1.16}$$

To bound the variance component, compute

$$
\begin{aligned}
\mathbb{E}[(W-W')^2|\Pi] &= \mathbb{E}[(a_{I\Pi(I)} - a_{I\Pi(J)} + a_{J\Pi(J)} - a_{J\Pi(I)})^2|\Pi] \\
&= \mathbb{E}[a_{I\Pi(I)}^2 + a_{J\Pi(J)}^2 + a_{I\Pi(J)}^2 + a_{J\Pi(I)}^2 \\
&\quad - 2a_{I\Pi(I)}a_{I\Pi(J)} - 2a_{J\Pi(J)}a_{J\Pi(I)} - 2a_{I\Pi(I)}a_{J\Pi(I)} - 2a_{J\Pi(J)}a_{I\Pi(J)} \\
&\quad + 2a_{I\Pi(I)}a_{J\Pi(J)} + 2a_{I\Pi(J)}a_{J\Pi(I)}|\Pi] \\
&= \frac{2}{n}\sum_{i=1}^{n} a_{i\Pi(i)}^2 + \frac{2}{n}\frac{1}{n-1}\sum_{i,j=1,i\neq j}^{n} a_{i\Pi(j)}^2 \\
&\quad - \frac{4}{n}\frac{1}{n-1}\sum_{i,j=1,i\neq j}^{n} a_{i\Pi(i)}a_{i\Pi(j)} - \frac{4}{n}\frac{1}{n-1}\sum_{i,j=1,i\neq j}^{n} a_{i\Pi(i)}a_{j\Pi(i)} \\
&\quad + \frac{2}{n}\frac{1}{n-1}\sum_{i,j=1,i\neq j}^{n} a_{i\Pi(i)}a_{j\Pi(j)} + \frac{2}{n}\frac{1}{n-1}\sum_{i,j=1,i\neq j}^{n} a_{i\Pi(j)}a_{j\Pi(i)} \\
&= \frac{2}{n}\sum_{i=1}^{n} a_{i\Pi(i)}^2 + \frac{2}{n}\frac{1}{n-1}\left(\sum_{i,j=1}^{n} a_{i\Pi(j)}^2 - \sum_{i=1}^{n} a_{i\Pi(i)}^2\right) \\
&\quad - \frac{4}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left(a_{i\Pi(i)}\sum_{j=1}^{n}\left(a_{i\Pi(j)} + a_{j\Pi(i)}\right) - 2a_{i\Pi(i)}^2\right) \\
&\quad + \frac{2}{n}\frac{1}{n-1}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\left(a_{i\Pi(i)}a_{j\Pi(j)} + a_{i\Pi(j)}a_{j\Pi(i)}\right) - 2\sum_{i=1}^{n} a_{i\Pi(i)}^2\right) \\
&= \frac{2}{n}\left(1 - \frac{1}{n-1}\right)\sum_{i=1}^{n} a_{i\Pi(i)}^2 + \frac{2}{n} \\
&\quad + \frac{8}{n}\frac{1}{n-1}\sum_{i=1}^{n} a_{i\Pi(i)}^2 \\
&\quad + \frac{2}{n}\frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(a_{i\Pi(i)}a_{j\Pi(j)} + a_{i\Pi(j)}a_{j\Pi(i)}\right) - \frac{4}{n}\frac{1}{n-1}\sum_{i=1}^{n} a_{i\Pi(i)}^2 \\
&= \frac{2}{n} + \frac{2}{n-1}\sum_{i=1}^{n} a_{i\Pi(i)}^2 + \frac{2}{n}\frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{i\Pi(i)}a_{j\Pi(j)} + a_{i\Pi(j)}a_{j\Pi(i)})
\end{aligned}
\tag{1.17}
$$

Theorem 1.2 (The c_r-inequality). *Let $r > 0$. Suppose that $\mathbb{E}|X|^r < \infty$ and $\mathbb{E}|Y|^r <$*

∞. *Then*

$$\mathbb{E}|X+Y|^r < c_r(\mathbb{E}|X|^r + \mathbb{E}|Y|^r), \tag{1.18}$$

where $c_r = 1$ *when* $r \leq 1$ *and* $c_r = 2^{r-1}$ *when* $r \geq 1$.

Corollary 1.3. *Suppose that* $\mathrm{Var}(X) < \infty$ *and* $\mathrm{Var}(Y) < \infty$. *Then*

$$\mathrm{Var}(X+Y) < 2(\mathrm{Var}(X) + \mathrm{Var}(Y)). \tag{1.19}$$

Proof. This follows immediately from applying Theorem 1.2 to the centered random variables $X' = X - \mathbb{E}[X]$ and $Y' = Y - \mathbb{E}[Y]$. □

□