36-450: Advanced Statistical Theory II	**Winter 2025**

Lecture 0: February 03

Lecturer: Matey Neykov *Scribe: Tuorui Peng*

0.1 Example 2: Multinomial Testing

Motivation: We are curious that: given a lottery with d balls, is the lottery fair? That is, is the probability of each ball being drawn equal to $1/d$?

0.1.1 Problem Statement

We have the distribution family $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ for which $\mathbb{P}_\theta$ is supported on $[d] := \{1, 2, \ldots, d\}$, and satisfies

$$\Theta = \Big\{\theta :\, p_\theta(i) \geq 0, \quad \sum_{i=1}^{d} p_\theta(i) = 1\Big\}$$

And we consider the uniformality test, i.e. the parameter of interest is

$$\{\theta_0\} = \Theta_0 = \Big\{\theta :\, p_\theta(i) = 1/d, \quad \forall i \in [d]\Big\}$$

w.r.t. the corresponding alternative. The rejection region we consider takes the form of ℓ_1 norm, i.e. our testing problem $\hat{\psi}_n$

$$H_0 : p_\theta = p_{\theta_0} = \mathrm{Unif}[d] \longleftrightarrow H_a : p_\theta \neq p_{\theta_0}$$

in the sense that we can control the probability of error

$$\mathbb{P}_0\left(\hat{\psi}_n = 1\right) + \sup_{p_\theta \in H_1} \mathbb{P}_\theta\left(\hat{\psi}_n = 0\right) \leq \varepsilon$$

for which, note that we have the relation between probability of error and the total variation distance d_{TV}, it suffices to control the total variation distance, which would leads to the following form of rejection region represented by ℓ_1 norm:

$$\text{Rejection Region} = \Big\{\theta :\, \left\|p_\theta - p_{\theta_0}\right\|_1 > \epsilon\Big\}$$

Goal: We are curious about the (asymptotic) behaviour of the critical value ϵ.

0.1.2 Challenge

Compared with the previous example of mean hypothesis testing, in which we can access an unbiased estimator (up to a constant) to the $\|y\|_2^2$, here an unbiased estimator to the $\|p_\theta - p_{\theta_0}\|_1$ is intractable. Thus we consider using other related norm to bound it.

0.2 Upper Bound Side

Denote our data $X = \{X_i\}_{i=1}^n$, $X_i = \{X_{i1}, X_{i2}, \ldots, X_{id}\}$, $X_i \in \{\hat{e}_1, \ldots, \hat{e}_d\}$ where $\hat{e}_j$ is the j-th canonical basis vector in $\mathbb{R}^d$. Then we have the following estimator for $\|p_\theta - p_{\theta_0}\|_2^2$:

Lemma 0.1 *With* $\underset{n\times d}{X}$ *being the data defined above and* $p_{\theta_0} = \mathrm{Unif}[d]$*, we have*

$$\mathbb{E}\left[\binom{n}{2}^{-1}\sum_{i\neq j} X_i' X_j - \frac{1}{d}\right] = \|p_\theta - p_{\theta_0}\|_2^2 .$$

Proof: Note that

$$\mathbb{E}_\theta\left[X_i X_j\right] = \delta_{ij} + (1-\delta_{ij})\sum_{k=1}^d p_\theta(k)^2$$

we have

$$\begin{aligned}
\mathbb{E}\left[\binom{n}{2}^{-1}\sum_{i\neq j} X_i' X_j - \frac{1}{d}\right] &= \binom{n}{2}^{-1}\sum_{i\neq j}\mathbb{E}_\theta\left[X_i X_j\right] - \frac{1}{d} \\
&= \sum_{k=1}^d p_\theta(k)^2 - \frac{1}{d} \\
&= \sum_{k=1}^d \left(p_\theta(k) - \frac{1}{d}\right)^2 \\
&= \|p_\theta - p_{\theta_0}\|_2^2 .
\end{aligned}$$

∎