Uniformity Testing with uniqueness

Ari

September 20, 2022

Abstract

1 Main Results

Let $Z_2 = \frac{1}{n}\sum_{j\in[k]} \mathbb{1}(N_j = 1)$ be the empirical avereage of the number of unique observations out of n i.i.d samples, where $N_j = \sum_{i=1}^{n} \mathbb{1}(x_i = i)$ counts the number of occurrences for all $j \in [k]$. The following algorithm is an optimal uniformity tester as long as $n \leq k$.

Theorem 1.1. *Let $\tau = (1-\frac{1}{k})^{n-1} - \frac{n\alpha^2}{8}$. As long as $n \geq O(\frac{\sqrt{k}}{\alpha^2})$*

$$\mathbb{P}\left[Z_2 \leq \tau \middle| d_{TV}(p, U_K) = 0\right] \geq \frac{2}{3} \tag{1}$$

and

$$\mathbb{P}\left[Z_2 \geq \tau \middle| d_{TV}(p, U_K) \geq \alpha\right] \geq \frac{2}{3} \tag{2}$$

Proof. The first thing to check is what the expectation of the statistic looks like:

$$\mathbb{E}[Z_2] = \frac{1}{n}\sum_{j\in[k]} p_j(1-p_j)^{n-1} \tag{3}$$

Now this does not look related to $||\vec{\boldsymbol{p}} - \vec{\boldsymbol{u_k}}||_1$ in anyway. However by the Lemma 2.1 described below we have

$$\mathbb{E}_{\vec{\boldsymbol{u_k}}}[Z_2] - \mathbb{E}_{\vec{\boldsymbol{p}}}[Z_2] \geq \frac{n}{16k} d_{TV}(\vec{\boldsymbol{p}}, \vec{\boldsymbol{u_k}})^2 \geq \frac{n}{16k}\alpha^2 := \Delta \tag{4}$$

Additionally, if $d_{TV}(\vec{\boldsymbol{p}}, \vec{\boldsymbol{u_k}})^2 = 0$ we have that $\mathbb{E}[Z_2] = \frac{1}{n}(1-\frac{1}{k})^{n-1}$. Thus from the picture below a reasonable test is to just set $\tau = \mathbb{E}_{\vec{\boldsymbol{u_k}}}[Z_2] - \Delta/2$. If the variance of Z_2 is much less than $\frac{n}{8k}\alpha^2$ we are would be done. So there is only one thing left to do is upper bound the variance for both cases. Then we just invoke Chebychev's inequality and the rest follows

$$\texttt{Var}[Z_2] = \frac{1}{n^2}\sum_{i,j\in[k]} \mathbb{E}[\mathbb{1}\{N_j = 1\}]\mathbb{E}[\mathbb{1}\{N_i = 1\}] - \mathbb{E}[Z_2]^2 \tag{5}$$

$$= \frac{1}{n}\sum_{j\in[k]} \frac{1}{n}\mathbb{E}[\mathbb{1}\{N_j = 1\}] + \frac{1}{n^2}\sum_{i\neq j} \mathbb{E}[\mathbb{1}\{N_j = 1\}]\mathbb{E}[\mathbb{1}\{N_i = 1\}] - \mathbb{E}[Z_2]^2 \tag{6}$$

$$= \frac{1}{n}\mathbb{E}[Z_2] - \frac{n}{n}\mathbb{E}[Z_2]^2 + \frac{1}{n^2}\sum_{i\neq j} n(n-1)p_i p_j(1-p_j-p-i)^{n-2} \tag{7}$$

$$= \mathbb{E}[Z_2]\frac{1}{n}(1-\mathbb{E}[Z_2]) + \frac{n-1}{n}\left(\sum_{i\neq j} p_i p_j(1-p_j-p-i)^{n-2} - \mathbb{E}[Z_2]^2\right) \tag{8}$$

$$\leq \frac{1}{n}(1-\mathbb{E}[Z_2]) + \left(\sum_{i\neq j} p_i p_j(1-p_j-p-i)^{n-2} - \mathbb{E}[Z_2]^2\right) \tag{9}$$

(7): That is the probability of selecting sampling two elements once and once only is $p_i p_j(1-p_i-p_j)$. There are n options for i and $n-1$ options for j.

Now when $\vec{\boldsymbol{p}} = \vec{\boldsymbol{u}}$, we have $\mathbb{E}[Z_2] = (1-\frac{1}{k})^{n-1}$, so if we plug this back into the above equation we get $\texttt{Var}Z_2 \leq \frac{3}{k}$. Thus in order for our tests to work out nicely, we want $O(\frac{1}{k}) \leq \Delta^2 = O(\frac{n^2\alpha^4}{k^2})$. Re-arranging and solving we get what we want. The explicit derivation is shown below

(10)

Now the variance for the far away case. ⟪ We use this magical lemma 2.2 to get this upper bound. I have no intuition for this. ⟫

$$\left(\sum_{i\neq j} p_i p_j(1-p_j-p-i)^{n-2} - \mathbb{E}[Z_2]^2\right) \tag{11}$$

$$= \sum_{i\neq j} p_i p_j(1-p_j-p-i)^{n-2} - \sum_{i,j} p_i(1-p_i)^{n-1}p_j(1-p_j)^{n-1} \tag{12}$$

$$\leq \sum_{i\neq j} p_i p_j\Big((1-p_j-p-i)^{n-2} - (1-p_i)^{n-1}(1-p_j)^{n-1}\Big) \tag{13}$$

$$\leq \frac{1}{n-1}\sum_{j\in[k]} p_j\Big(1-(1-p_j)^{n-1}\Big) \tag{14}$$

$$= \frac{1}{n-1}(1-\mathbb{E}[Z_2]) \tag{15}$$

(14): This is a direct application of Lemma 2.2 with $m = n-1$ and $x_i = p_i$
Plugging this back into (9) we get, for a general $\vec{\boldsymbol{p}}$

$$\texttt{Var}[Z_2] \leq \frac{1}{n}(1-\mathbb{E}[Z_2]) + \frac{1}{n-1}(1-\mathbb{E}[Z_2]) \tag{16}$$

$$\leq \frac{3}{n}(1-\mathbb{E}[Z_2]) \tag{17}$$

$$= \frac{3}{n}(1-\mathbb{E}_{\vec{\boldsymbol{u}_k}}[Z_2] + \mathbb{E}_{\vec{\boldsymbol{u}_k}}[Z_2] - \mathbb{E}[Z_2]) \tag{18}$$

$$= 3\Big(\frac{1}{k} + \frac{\Delta(p)}{n}\Big) \tag{19}$$

(17): $\frac{1}{n-1} \leq \frac{2}{n}$
The first term is the same as the uniform case, the second term is an equivalent statment.

$$\frac{3\Delta(p)}{n} \ll \Delta(p)^2 \tag{20}$$

$$\frac{3}{n} \ll \Delta(p) \qquad \leq \frac{n\alpha^2}{k} \tag{21}$$

$$\tag{22}$$

The algebra works out to need to $n \geq O\Big(\frac{\sqrt{k}}{\alpha}\Big)$

Ok we have the variance, and we have shown that it does not exceed the expectation gap by too much under our assumptions. Now we formalise all of this and apply standard chebychev. First the uniform case

$$\mathbb{P}\left[Z_2 \leq \tau \middle| d_{TV}(p, U_K) = 0\right] = \mathbb{P}\left[Z_2 \leq \mathbb{E}_{\vec{\boldsymbol{u_k}}}[Z_2] - \frac{\Delta}{2}\right] \tag{23}$$

$$\leq \frac{4\texttt{Var}[Z_2]}{\Delta^2} \tag{24}$$

$$\leq \frac{c_1 k}{n^2\alpha^4} \qquad \leq \frac{1}{3} \tag{25}$$

Now for the far away case

$$\mathbb{P}\left[Z_2 \geq \tau \middle| d_{TV}(p, U_K) \geq \alpha\right] = \mathbb{P}\left[Z_2 \geq \mathbb{E}_{\vec{\boldsymbol{u_k}}}[Z_2] - \frac{\Delta}{2}\right] \tag{26}$$

$$= \mathbb{P}\left[Z_2 \geq \mathbb{E}_{\vec{\boldsymbol{p}}}[Z_2] + \Delta(p) - \frac{\Delta}{2}\right] \tag{27}$$

$$= \mathbb{P}\left[Z_2 \geq \mathbb{E}_{\vec{\boldsymbol{p}}}[Z_2] + \frac{\Delta(p)}{2} + \frac{\Delta(p)}{2} - \frac{\Delta}{2}\right] \tag{28}$$

$$\leq \mathbb{P}\left[Z_2 \geq \mathbb{E}_{\vec{\boldsymbol{p}}}[Z_2] + \frac{\Delta(p)}{2}\right] \tag{29}$$

$$\leq \frac{4\texttt{Var}[Z_2]}{\Delta(p)^2} \tag{30}$$

$$\leq 12\Big(\frac{1}{k\Delta(p)^2} + \frac{1}{n\Delta(p)}\Big) \tag{31}$$

$$\leq 12\Big(\frac{1}{k\Delta^2} + \frac{1}{n\Delta}\Big) \tag{32}$$

$$\tag{33}$$

Plugging in $\Delta = \frac{n\alpha^2}{16k}$ we get what we want. □

2 Useful Lemmas

Lemma 2.1. *Let $\vec{\boldsymbol{X}} = (x_1, \ldots, x_n)$. If $n \leq k$ we have*

$$\mathbb{E}_{\vec{\boldsymbol{u_k}}}[Z_2] - \mathbb{E}_{\vec{\boldsymbol{p}}}[Z_2] \geq \frac{n}{16k} d_{TV}(\vec{\boldsymbol{p}}, \vec{\boldsymbol{u_k}})^2 \tag{34}$$

Proof. I have to admit the calculus tricks were not immediately intuitive to me. □

Lemma 2.2. *Fix $m \geq 1$, $k \in \mathbb{N}$. For any $x_1, \ldots, x_k$ where $(\sum_{i=1}^k x_i) = 1$, we have*

$$\frac{m\sum_{1\leq i<j\leq k} x_i x_j \Big((1-x_i-x_j)^{m-1} - (1-x_i)^m(1-x_j)^m\Big)}{\sum_{i=1}^k x_i\Big(1-(1-x_i)^m\Big)} \leq 1 \tag{35}$$

Proof. I do not have much intuition for why this is true. □

References