New_New about Mann-Whitney

Zeki, Kaiyan

June 2018

1 Introduction

Suppose $n_1 < n_2$, and N is pubic.

1. Release $\hat{n}_1 = n_1 + \text{Lap}(1/\epsilon_1)$ to the public.

2. Calculate c such that with $1-\delta$ probability, $|n_1 - \hat{n}_1| \leq c$.

3. Choose $n_1^* \in [\hat{n}_1 - c, \hat{n}_1 + c]$ such that n_1^* leads to the smallest critical value (lowest power to ensure accuracy under all cases) and the largest change of U (the global sensitivity to ensure privacy under all cases).

4. Use $\Delta U = \max\{n_1^*, N - n_1^*\}$ as the sensitivity and release $\hat{U} = U + \Delta U$. The release is ϵ_2-δ -DP.

5. The researcher/public can find the null distribution with $\hat{n}_1$ and $N - \hat{n}_1$ to get the critical value they want an compare with the released $\hat{U}$.

2 Calculation of c

PDF of Laplace(b) = $f(x) = \frac{1}{2b} e^{\frac{|x-\mu|}{b}}$
With given δ, we want

$$1 - \delta = \int_{n_1-c}^{n_1+c} f(x)dx$$

.

The calculation starts as follows, with $b = 1/\epsilon_2$ and $\mu = n_1$.

$$\begin{aligned}\int_{n_1-c}^{n_1+c} f(x)dx &= 2 \cdot \int_{n_1-c}^{n_1} f(x)dx \\ &= 2 \cdot \int_{n_1-c}^{n_1} \frac{1}{2b} e^{\frac{|x-\mu|}{b}} dx \\ &= 2 \cdot \left(\frac{1}{2} e^{\frac{x-n_1}{1/\epsilon_2}}\right)\Big|_{n_1-c}^{n_1} \\ &= e^0 - e^{\frac{n_1-c-n_1}{1/\epsilon_2}} \\ &= 1 - e^{-c\epsilon_2}\end{aligned}$$

Thus,

$$\begin{aligned}1 - \delta &= 1 - e^{-c\epsilon_2} \\ \delta &= e^{-c\epsilon_2} \\ c &= -\frac{ln\delta}{\epsilon_2}\end{aligned}$$

3 Choice of n_1^*

Claim: For all choice of n_1^*, the larger $|n_1^* - n_2^*|$ is, the larger the sensitivity of U and the smaller the critical values are.
Pf/ By previous proof, $\Delta U = \text{Max}\{n_1, n_2\}$ and in this case is $\Delta U = \text{Max}\{n_1^*, n_2^*\}$.

Suppose $n_2^* > n_1^*$, then $\Delta U = n_2^*$, and $|n_1^* - n_2^*| = n_2^* - n_1^* = n_2^* - (N - n_2^*) = 2n_2^* - N$. Thus the larger $|n_1^* - n_2^*|$ is, the larger n_2^* is, and the larger ΔU is.
Suppose $n_2^* < n_1^*$, then $\Delta U = n_1^*$, and $|n_1^* - n_2^*| = n_1^* - n_2^* = n_1^* - (N - n_1^*) = 2n_1^* - N$. Thus the larger $|n_1^* - n_2^*|$ is, the larger n_1^* is, and the larger ΔU is.
In both cases, the larger $|n_1^* - n_2^*|$ is, the larger n_1^* is, and the larger ΔU is.

Let the significance level $= \alpha$, then for n_1,$n_2 > 20$, the critical values for U,

$$Z' = Z \cdot \sigma + \mu = Z \cdot \sqrt{\frac{(n_1 + n_2 + 1)n_1 n_2}{16}} + \frac{n_1 n_2}{2}.$$

As $n_1 n_2 = \frac{(n_1+n_2)^2 - (n_1-n_2)^2}{4}$, the larger $|n_1 - n_2|$, the smaller $n_1 n_2$, thus the smaller the critical values.

By supposition $n_1 < n_2$ and as $n_1^* \in [\hat{n}_1 - c, \hat{n}_1 + c]$, we choose $n_1^* = \hat{n}_1 - c$ which leads to largest difference between n_1^* and $n_2^* = N - n_1^*$. This is the worst case for both ΔU and critical values.

4 Verification of ΔU

With $n_1^* = \hat{n}_1 - c$, We can find the range for n_1^* and n_2^* to determine ΔU.

Consider n_1^*. As $|n_1 - \hat{n}_1| \leq c$, $\hat{n}_1 \in [n_1 - c, n_1 + c]$. Then it follows that $n_1^* \in [n_1 - 2c, n_1]$. Then $n_2^* = N - n_1^* \in [N - n_1, N - n_1 + 2c] = [n_2, n_2 + 2c]$. As $n_1 < n_2$, $n_2^* > n_1^*$. Thus $\Delta U = n_2^* = N - n_1^*$.

Since $n_2^* \in [n_2, n_2 + 2c]$, $n_2^* > n_2$. Thus the sensitivity is always larger than the true sensitivity, and therefore will not break the differential privacy.

5 Algortihm

With N is pubic. The release is $\epsilon_1 + \epsilon_2$ -DP.

1. WLOG, suppose $n_1 < n_2$,and let release $\hat{n}_1 = n_1 + \text{Lap}(1/\epsilon_1)$ to the public.

2. Let $c = -\frac{ln\delta}{\epsilon_2}$ so that with $1 - \delta$ probability, $|n_1 - \hat{n}_1| \leq c$.

3. Let $n_1^* = \hat{n}_1 - c$.

4. Let $\Delta U = N - n_1^*$ and release $\hat{U} = U + \Delta U$. The release is ϵ_2-δ -DP.

5. The researcher/public can find the null distribution with $\hat{n}_1$ and $N - \hat{n}_1$ to get the critical value they want an compare with the released $\hat{U}$.