Simulation

zbjiao

October 2019

1 Simulation Study

(**This section is written based on the assumption that the methodology of all the methods have already been covered.**) In this section, we design simulation study to convey the idea that, of all the enrichment methods we've tried, permutation from raw data with t (**please replace this with a more proper term**) suits best for reasoning the casualty between vaccination and adverse effects (AE). Section 1.1 gives a detailed simulation setup. A summary of the simulation results is in Section 1.2. Finally, we gives a inference based on the simulation results in Section 1.3 and a toy simulation that favors KS in Section 1.4.

1.1 Simulation Setup

The idea is straightforward. We want to generate a pseudo table where records the vaccination and AE information for each observation. We artificially set one group of AEs that are enriched by the vaccine. Then, all methods raised in the article will be implemented to test if they can detect the enrichment.

To be very concise, we only specify two AE groups, the target AE group and the non-target AE group where there are 10 AEs in the target group and 40 AEs in another. The sample size, a.k.a the number of observations, is set at 100. For each observation, the probability of the vaccination follows a Bernoulli distribution with $p = 0.2$.

A logit model (**I am not sure whether the term logit model is appropriate**) is adopted to quantify the probability of one certain AE's emergence. Specifically, for AEs in the target group, we suppose that the emergence of the AE is directly influenced by the vaccine:

$$\frac{P(AE = 1)}{P(AE = 0)} = exp(\alpha + \beta I(\text{Vac})) \tag{1}$$

where $\alpha = -2$ and β is a tuning parameter range from 0 to 1 which can be interpreted as the one certain vaccine's "ability" to cause AE.
While for AEs in the non-target group, the emergence of AE is irrelevant to the

vaccine:

$$\frac{P(AE=1)}{P(AE=0)} = exp(\alpha) \tag{2}$$

In terms of the simulation strategy, we consider two scenarios. First, we permuted from the raw data for 1000 times. Every time by random sampling the Vac column, we obtain a new table. Then, all the computing will be conducted based on it. Second, we only permuted the RRs for 1000 times after we have obtained the true RRs for 50 AEs based on the raw table. The first strategy is denoted as 'out' and second as 'in'. For each strategy, two methods, KS and t, are respectively considered. Table 1 shows one typical simulated table.

Obs	Vac	AE_1	AE_2	...	AE_{50}
1	0	1	0	...	0
2	1	0	1	...	0
...	...	...	...	...	...
100	1	0	1	...	0

Table 1: The example of a randomly generated table where 1 and 0 in the AE columns are used to denote one certain AE's emergence or not respectively. 1 and 0 in the Vac column means the vaccination happened or not.

1.2 Results

Fig. 1 shows a summary of the simulation result. β takes value from 0 to 1. While each point is an average of 1000 times of experiments, while each experiment is 1000 times of permutation. Intuitively, as the β increases, we can get higher power scores for all 4 cases. A reasonable speculation is that they will all reach the power of 1 if β is large enough. Power unites at 0.05 when $\beta = 0$ since vaccination will make no difference in this situation.

Generally, t has a better performance than KS and "out" strategy outbid "in" strategy. t is also faster.

1.3 Inference

- For t method with fixed sample size, the difference between in and out is evident. Notion 'in' means that the permutation is conducted after the RR is calculated, which indicates that, suppose we have 10 AEs in the target group and 40 AEs not, the null distribution will be estimated with these 50 RRs.

  The theoretical value of RR for target AE is:

$$RR_1 = \frac{P(AE|\text{Vac})}{P(AE)} = \frac{\frac{e^{\beta-2}}{1+e^{\beta-2}}}{P(\text{Vac}) \times \frac{e^{\beta-2}}{1+e^{\beta-2}} + P(!\text{Vac}) \times \frac{e^{-2}}{1+e^{-2}}} \tag{3}$$

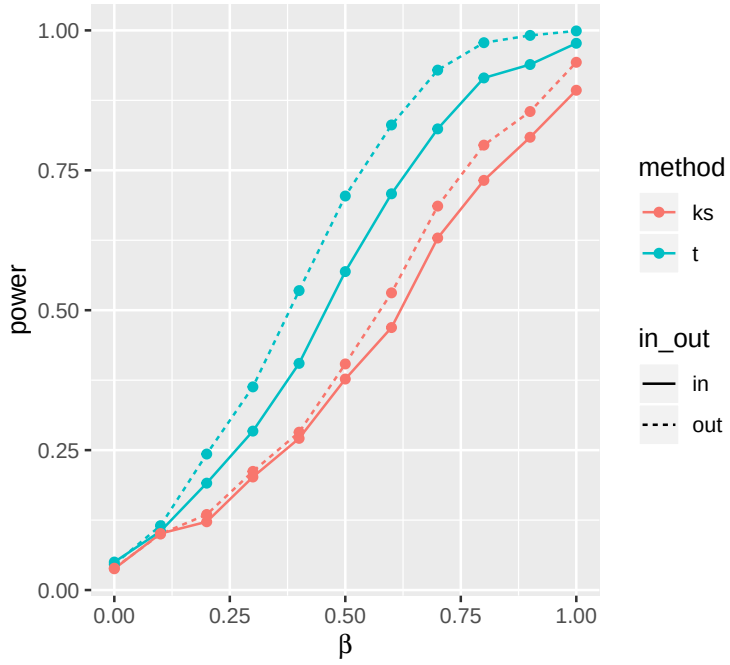

Figure 1: A line chart summary of the simulation result. Dashed red line: KS method with "out" strategy; Dashed blue line: t method with "out" strategy; Solid red line: KS method with "in" strategy; Solid blue line: t method with "out" strategy.

where $P(Vac) = 0.2, P(!Vac) = 0.8$.
For RR of non-target AE, the theoretical value should be 1:

$$RR_2 = \frac{P(AE|\text{Vac})}{P(AE)} = \frac{P(AE)}{P(AE)} = 1 \tag{4}$$

If we use 'in' method, the permutation will based on 10 $\hat{RR}_1$ with an expectation larger than 1 and 40 $\hat{RR}_2$ with an expectation of 1. So the null distribution is right biased. Hence, the null hypothesis will tend to be harder to be rejected when $\beta > 0$ which leads to a smaller power.

For 'out' method, the null distribution is certainly not biased. Since if we permute from the very beginning, the permuted $\hat{RR} = 1$. The permuted RR will have nothing to do with vaccine: $P(AE|Vac) = P(AE)$.

Basically, with the current setting of the simulation, permute from the very beginning certainly is a better choice.

An extreme case may help better illustrate the idea: suppose we have 10 AEs where 9 of them have $\beta < 0$ and one $beta = 0$. Then, under the

"in" strategy, the AE with $beta = 0$ will surely have $p < 0.05$. Since the null distribution is severely left biased due to the 9 AEs. As a result, the only neutral AE is considered as significant which is certainly not the case. This example shows why "in" strategy is flawed.

- For KS method with fixed sample size, most of the theory should be the same as t method, if we take $p > 0$ for ks method (we take $p = 1$ for this simulation). But, if we take p=0, in other words, we only consider the rank of each $\hat{RR}$, there shouldn't be any bias for 'in' method.

- For sample size, certainly, as we've derived in part1, as $n- > \infty$, all the $\hat{RR}$ will converge to the corresponding real value RR. Larger sample size could lead to smaller variance when doing permutation and higher power. See the line marked by small triangles and circles.

- For the difference of KS and t, one rational inference for currently t having a better performance simply is that the setting of the simulation is in favor of t. Noted that right now all AEs in the target group has the same probability of emerging. They exist in a more flat rather than a peaked way. This is generally not the pattern KS is good at. So, in the next section, we provide a simulation that favors KS.

1.4 Simulation that favors KS

Most of the setup is the same as Section 1.1, except for the βs in the target AE group are specified in different values. We consider half of the AEs in target AE group have same β_1, while the other half have the same β_2 where $\beta_1 = -\beta_2$. In other words, we consider conditions where the vaccine's effect on the target AE group is mixed with both positive and negative.

In order to find a scenario where KS can out-performs t, we need to build the simulation containing some of the $\beta < 0$s in the target group. Then, t method will be impaired since those β will lead to $RR < 1$ and cancel out the large RR when calculating the mean value of the target group. However, KS's performance will, on the contrary, be reinforced. Since it calculates the maximum step within ranked RRs and different settings of βs within a AE group increase the maximum step.

Fig. 2 shows a summary of the simulation that favors KS, where β_1 ranges from 0 to 2.

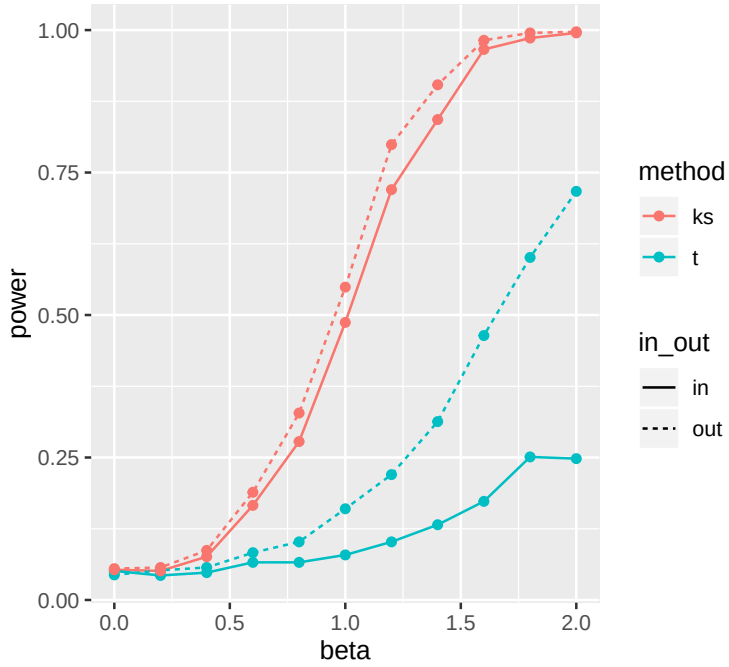

Figure 2: A line chart summary of the simulation that favors KS. The setting of line chart is the same as Fig. 1, except for β ranging from 0 to 2.