Test Positive Fractions v. Test Positives

Suzanne M. Dufault

January 17, 2018

Contents

Abstract

Addressing reviewer's comment: Authors choose $T = \alpha_I - \alpha_C$. What would be the problem with choosing $T' = p_I - p_C$ where p_I is the average of p_{Dj} in the intervention clusters and p_C is the average of p_{Dj} in the control clusters? Under the null, $E(T')$ would also be null? I think authors' T is a better statistic, but it may still be influenced by differences in the incidence of non-arbovirus febrile illnesses between clusters, for example respiratory virus infections that occur at higher rates in more densely populated locations?

1 Process

1. I set up an example dataset for 10 clusters with variation between the clusters with respect to Cases and OFIs.
2. I then permuted 10 choose 5 (252) unique treatment allocations.
3. I computed the log(OR) for RR = 1 for each of the permuted allocations using 100 cases and 100 controls assigned to clusters according to their current proportions.

2 General Setup

Suppose there are $2m$ clusters with m randomly assigned to the intervention and the remainder untreated. How do we determine (via some sort of modified T-test) whether the intervention is working? The cases and controls are distributed as follows:

Clusters	Treatment	Case Distribution	Control Distribution
1	1	0.11	0.09
2	1	0.12	0.12
3	1	0.10	0.10
4	1	0.09	0.10
5	1	0.12	0.08
6	0	0.06	0.11
7	0	0.12	0.08
8	0	0.10	0.09
9	0	0.10	0.11
10	0	0.08	0.10

Table 1: Distribution of controls, from which we sample nD cases and nbarD = r x nD controls

If we want to sample $n_D = 1000$ cases and $n_{\bar{D}} = 1000$ controls, our data looks as follows.

Clusters	Treatment	Case Distribution	Control Distribution	Cases	Controls
1	1	0.1141	0.0944	114	94
2	1	0.1183	0.1224	118	122
3	1	0.0996	0.1012	100	101
4	1	0.0934	0.1031	93	103
5	1	0.1162	0.0848	116	85
6	0	0.0643	0.1079	64	108
7	0	0.1162	0.0829	116	83
8	0	0.0975	0.0915	98	92
9	0	0.0975	0.1118	98	112
10	0	0.0830	0.1002	83	100

Table 2: Case and control counts are rounded in order to be sensible. This leads to 1000 cases in this particular scenario and 1000 controls.

3 Test-Positives Only

The reviewer described ignoring the controls, such that we only consider the following case-specific data. Because treatment is randomized, this conceptually is a reasonable approach.

	Clusters	Treatment	Case Distribution	Cases
1	1	1	0.1141	114
2	2	1	0.1183	118
3	3	1	0.0996	100
4	4	1	0.0934	93
5	5	1	0.1162	116
6	6	0	0.0643	64
7	7	0	0.1162	116
8	8	0	0.0975	98
9	9	0	0.0975	98
10	10	0	0.0830	83

Table 3: Example data if we just consider the cases.

Let $j \in I$ be a cluster j in the intervention arm and $j \in C$ be a cluster j in the control arm. The proposed test statistic:

$$T = (n_D|j \in I) - (n_D|j \in C)$$

where $(n_D|j \in I)$ is equal to the total number of dengue cases in the intervention arm and $(n_D|j \in C)$ is equal to the total number of dengue cases in the control arm.

3.1 At the Null

Total

At the null, we expect half of all cases to fall in each arm $(n_D/2)$.

$$\begin{aligned} E[n_D|j \in C] &= E[n_D|j \in I] && (1)\\ &= \frac{1}{2}n_D && (2) \end{aligned}$$

Hence, at the null, $T = 0$

$$\begin{aligned} E[T|n_D] &= E[n_D|j \in I] - E[n_D|j \in C] && (3)\\ &= E[n_D|j \in I] - [n_D - E[n_D|j \in I]] && (4)\\ &= 2E[n_D|j \in I] - n_D && (5)\\ &= 2\frac{n_D}{2} - n_D && (6)\\ &= 0 && (7) \end{aligned}$$

3.1.1 Variance

The variance of T is estimated as follows:

$$\begin{aligned} Var(T) &= Var(2 \times [n_D|j \in I] - n_D) && (8)\\ &= 2^2 Var([n_D|j \in I]) && (9)\\ &= 2^2(2m)^2\left(\frac{2m-m}{2m}\right)\frac{s_x^2}{m} && (10)\\ &= 8ms_x^2 && (11) \end{aligned}$$

where $s_x^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{m-1}$ is the sample variance of the counts in either the control or intervention clusters.

3.2 Away from the Null

3.2.1 Recovering the RR

Away from the null: $n_D = \lambda \times (n_D|j \in I) + (n_D|j \in C)$. To recover lambda, we can simply take a ratio of the observed totals:

$$\frac{\lambda \times (n_D|j \in I)}{(n_D|j \in C)} = \lambda$$

3.2.2 Variance

$$T = \lambda \times (n_D|j \in I) - (n_D|j \in C) \tag{12}$$
$$= \lambda \times (n_D|j \in I) - (n_D - \lambda(n_D|j \in I)) \tag{13}$$
$$= 2 \times \lambda \times (n_D|j \in I) - n_D \tag{14}$$
$$\tag{15}$$

where $s_x^2 = \sum_{j \in I} \frac{(n_{D,j} - (\overline{n_D}|j \in I))^2}{m-1}$ is the sample variance. While it is currently expressed for the intervention clusters, we could alternatively choose to estimate the variance of the control clusters.

The standardized test statistic (under the null) is then $T/\sqrt{8ms_x^2}$. Since the arms both contain m clusters, we can average the two estimates of the variance of the total counts in each arm (the pooled variance estimators). Away from the null, we expect the variance to change by a multiple of λ^2. This will result in a decreased variance estimate (assuming $\lambda < 1$).

4 Simulation

4.1 Test-Positive Fraction Method Results