Fisher's Method

Katie Fitzgerald

05/03/2018

Fisher's test

Among the 100 replication studies conducted by the Open Science Collaboration (OSC), 64 of them found null effects, defined as having a p-value greater than 0.05. The OSC authors applied Fisher's method to this set of non-significant p-values to test the hypothesis that these 64 replication studies did in fact have "no evidential value." That is, they tested whether the null hypothesis of zero effect held for each study and that there were no false negatives among them. Therefore, their hypothesis for the Fisher's test could be formalized as follows, where θ_i is the treatment effect for study i:

$$H_0\text{: } \theta_1 = \theta_2 = ... = \theta_{64} = 0$$

$$H_a\text{: at least one } \theta_i \neq 0, \quad i = 1, ..., 64$$

Often, Fisher's method is used to combine the results of several independent tests on the same overall hypothesis that one $\Theta = 0$. For example, if researchers are interested in whether one particular treatment has a true effect, they could conduct k studies and pool the results to give an indication if the true $\Theta = 0$. In that scenario, if the null hypothesis for each of the k studies is true (i.e. H_0 above holds), the p-values follow a uniform distribution on [0,1]. The test statistic for Fisher's method is

$$X^2 = -2\sum_{i=0}^{k} ln(p_i),$$

which follows a χ^2 distribution with $2k$ degrees of freedom when H_0 is true. Low p-values give a larger test statistic, leading to a rejection that $\Theta = 0$.

In the OSC scenario, however, the authors are conducting Fisher's test only among the 64 replication studies that did not have significant results. By design then, the p-values have a U[0.05,1] distribution. The OSC makes an appropriate transformation, $p_i^* = \frac{p_i - 0.05}{0.95}$, so that $p_i^* \sim U[0,1]$ and Fisher's method can be applied. Similarly, the transformed test statistic is

$$X^2 = -2\sum_{i=0}^{k} ln(p_i^*) = -2\sum_{i=0}^{k} ln(\frac{p_i - 0.05}{0.95}),$$

which follows a χ^2 distribution with $2k$ degrees of freedom under H_0.

Among the 64 non-significant replicate p-values, the OSC finds a test statistic of $X^2 = 155.83$, which falls just beyond the critical value $\chi^2_{128} = 155.40$ and therefore is significant with $p = 0.048$. The authors acknowledge this suggests there is at least one replication finding that could be a false negative, but they conclude "nonetheless, the wide distribution of p-values suggests against insufficient power as the only explanation of failures to replicate." This statement seems to indicate the authors are looking for evidence that there are no false negatives among the non-significant replication studies, which is their null hypothesis.

While it is never advised to conduct a test in order to conclude the null hypothesis, this switched framework is especially problematic when the test is underpowered to reject H_0 because the Type II error rate will be large. Even though the authors were able to reject their null hypothesis, we think that in general Fisher's method might be underpowered to answer the question at hand, so we want to explore this more rigorously. The intuition is that if there is just one $\theta_i \neq 0$, and the remaining $k - 1$ studies have p-values from a U[0,1] distribution, the one non-null effect will need to be very large to skew the distribution enough to look differently from U[0,1]. Or alternatively, there would need to be several non-null effects to skew the

distribution to a detectable degree. In the scenario where a researcher is combining the results of k studies of the same treatment to determine if an overall $\Theta = 0$, this type of conservative test may be appropriate. In the OSC scenario, however, since the 64 studies are not testing the same treatment effect, and one θ_i has no bearing on the 63 other $\theta_i's$, the presumed goal would be to detect if there are *any* false negatives among the replicate studies. We will use power simulations to determine if the intuition holds that it would take many and possibly large non-null effects for Fisher's test to reject H_0.

We make a few simplifying assumptions about the framework of these 64 studies in order to proceed. Assume each study i is testing the presence of some treatment effect θ_i and has equal sample sizes for the treatment and control groups ($n_{ti} = n_{ci} = n_i$). Let Y_{ij}^t and Y_{ij}^c be the observations from the treatment and control groups respectively ($j = 1, ...n_i$), and assume $Y_{ij}^t \sim N(\mu_i + \theta_i, \sigma^2)$ and $Y_{ij}^c \sim N(\mu_i, \sigma^2)$, where σ^2 is known. Note this is not an unreasonbale assumption since this framework is required for both the t-test and the ANOVA test, and the majority of the tests in this subset of 64 replicate studies are of these two types (89%).

Note then, θ_i is the difference in means between the treatment and control groups, and its estimate $T_i = \overline{Y_{i.}^t} - \overline{Y_{i.}^c}$ has variance $v_i = \frac{\sigma^2}{n_{ci}} + \frac{\sigma^2}{n_{ti}} = \frac{2\sigma^2}{n_i}$. For simplicity of interpretation but without loss of generality, we will work with the standardized scale of Cohen's d, defined as $\delta_i = \frac{\theta_i}{\sigma}$. Note then that δ_i is estimated by $d_i = \frac{T_i}{\sigma} = \frac{T_i}{\sqrt{\frac{v_i}{2/n_i}}} = \frac{T_i}{\sqrt{v_i}}\sqrt{\frac{2}{n_i}}$. Since under H_0, $T_i \sim N(\theta_i, v_i) \Rightarrow \frac{T_i}{\sqrt{v_i}} \sim N(\frac{\theta_i}{\sqrt{v_i}}, 1)$ and $\sqrt{\frac{2}{n_i}}$ is a constant, then we have $Var(d_i) = Var(\frac{T_i}{\sqrt{v_i}}\sqrt{\frac{2}{n_i}}) = \frac{2}{n_i}$. Therefore, $d_i \sim N(\delta_i, \frac{2}{n_i})$.

The hypotheses of each of the 64 studies can be represented on the scale of Cohen's d as follows:

$$H_{0i} : \delta_i = 0$$

$$H_{ai} : \delta_i \neq 0,$$

and the p-value of this test can therefore be calculated as $2(1 - \Phi(\frac{|d_i|}{\sqrt{2/n_i}}))$.

In our simulations, we will assume that there are m false negatives among the 64 studies, $m = 1, ...64$. Therefore we must draw m p-values from a distribution consistent with the alternative hypothesis. That is, we draw a random variable d_i from a $N(\delta_i, \frac{2}{n_i})$ distribution, where $\delta_i \neq 0$ and compute its p-value. We continue drawing $d_i's$ until we obtain m p-values greater than 0.05 (due to the restriction in the OSC scenario of only considering replicate studies with non-significant results). We will draw the remaining $64 - m$ p-values from a U[0.05,1] distribution and then calculate Fisher's test statistic $X^2 = -2\sum_{i=0}^{k} ln(\frac{p_i - 0.05}{0.95})$. We run this procedure N times and calculate the simulated power of Fisher's method under these conditions to be $\sum_{l=0}^{N} I_{\{X_l^2 > 155.4047\}}/N$, where I is the indicator function and $\chi^2_{128} = 155.4047$ is the critical value for Fisher's test with $k = 64$ studies.

We first consider the case when $n_i = 76/2 = 38$, because 76 is the median sample size in the OSC dataset, and we are assuming equal sample sizes in the treatment and control groups. The results of the power simulations for median sample size are given below. Note that 0.2, 0.5, and 0.8 correspond to small, medium, and large effects sizes on the scale of Cohen's d. We find that there needs to be 10 large effects $\delta_i = 0.8$ in order to achieve close to the standard 80% power (power=0.77426). If there is only one large effect, Fisher's test only has approximately 8% power to reject. Even when all 64 studies have small effects, there is only 70% power.

# of non-null effects	$\delta = 0.2$	$\delta = 0.5$	$\delta = 0.8$
1	0.0540	0.0669	0.0805
2	0.0573	0.0876	0.1214
3	0.0608	0.1157	0.1753
4	0.0666	0.1456	0.2452
5	0.0717	0.1841	0.3250
10	0.1008	0.4332	0.7743
32	0.3068	0.9971	1.0000
64	0.6976	1.0000	1.0000

In order to consider the power of Fisher's test under a "best-case scenario" in this dataset, we sort the 64 sample sizes and let the non-null effects be from the studies with the largest sample sizes first. That is, if $m = 1$ we let the non-null effect be from the largest study, if $m = 2$ we let the non-null effects be from the two largest studies, etc. Note, however that because we are working with a set of replicate studies which found a p-value greater than 0.05, pairing large effects with very large sample sizes is not realistic. For example, the largest study has $n_i = 768703/2 = 384351.5$ which has 100% power to detect a large effect of $\delta_i = 0.8$ (see tableXX in Appendix). In other words, it would not have been possible to obtain a p-value less than 0.05 with this sample size and true $\delta_i = 0.8$. We therefore take the largest n_i among the studies for which the power is at most 99.99% to detect the given δ_i. TableXX gives the largest sample sizes used for each δ.

δ	n	Power of OSC replicate to detect
0.2	745	0.9713
0.5	159	0.9938
0.8	100	0.9999

See TableXX in the Appendix for the full sample size vectors that were used to conduct the "best-case scenario" power simulations.

As shown in TableXX, even when a study with a very large sample size has a large effect, Fisher's method has less than 10% power to detect it. When the non-null effects come from the studies with the largest sample sizes, about half of the studies need to have $\delta = 0.2$ in order to achieve approximately 80% power (power=0.8340)

# of non-null effects	$\delta = 0.2$	$\delta = 0.5$	$\delta = 0.8$
1	0.0823	0.0870	0.0933
2	0.1291	0.1411	0.1614
3	0.1872	0.2153	0.2552
4	0.2169	0.3113	0.3760
5	0.2506	0.4167	0.5124
10	0.4264	0.8742	0.9625
32	0.8340	1.0000	1.0000
64	0.9341	1.0000	1.0000

Asymptotic power

If all 64 have 0.2, need n=278 (total sample size) for 80% power, n=45 for d=0.5, n=18 for d=0.8.

Appendix

n	d=0.2	d=0.5	d=0.8
384351.5	1.0000	1.0000	1.0000
745.0	0.9713	1.0000	1.0000
573.0	0.9230	1.0000	1.0000
159.0	0.4299	0.9938	1.0000
152.0	0.4144	0.9918	1.0000
140.0	0.3873	0.9869	1.0000
135.0	0.3758	0.9841	1.0000
131.5	0.3677	0.9819	1.0000
125.5	0.3538	0.9773	1.0000

n	d=0.2	d=0.5	d=0.8
113.0	0.3242	0.9639	1.0000
111.0	0.3194	0.9612	1.0000
100.0	0.2930	0.9424	0.9999
88.5	0.2650	0.9140	0.9996
83.0	0.2515	0.8964	0.9993
79.0	0.2417	0.8815	0.9989
76.5	0.2355	0.8713	0.9986
74.0	0.2294	0.8602	0.9982
72.0	0.2244	0.8508	0.9977
70.5	0.2207	0.8434	0.9974
70.0	0.2195	0.8409	0.9972
67.5	0.2133	0.8276	0.9964
63.0	0.2022	0.8013	0.9943
62.5	0.2009	0.7982	0.9940
59.0	0.1923	0.7751	0.9915
56.5	0.1861	0.7573	0.9890
54.0	0.1799	0.7383	0.9860
52.5	0.1762	0.7263	0.9838
45.5	0.1590	0.6645	0.9683
44.0	0.1553	0.6500	0.9635
41.5	0.1491	0.6246	0.9539
39.0	0.1430	0.5979	0.9421
38.0	0.1406	0.5869	0.9366
37.5	0.1393	0.5813	0.9337
37.5	0.1393	0.5813	0.9337
36.0	0.1357	0.5641	0.9242
35.5	0.1344	0.5583	0.9208
35.5	0.1344	0.5583	0.9208
35.0	0.1332	0.5524	0.9172
35.0	0.1332	0.5524	0.9172
33.0	0.1283	0.5283	0.9014
29.0	0.1186	0.4777	0.8613
25.5	0.1102	0.4308	0.8150
25.5	0.1102	0.4308	0.8150
25.0	0.1089	0.4239	0.8074
24.0	0.1065	0.4100	0.7914
24.0	0.1065	0.4100	0.7914
23.5	0.1053	0.4029	0.7830
21.5	0.1005	0.3744	0.7463
16.5	0.0886	0.3005	0.6323
16.0	0.0874	0.2930	0.6190
16.0	0.0874	0.2930	0.6190
15.0	0.0850	0.2778	0.5913
14.5	0.0838	0.2702	0.5770
12.0	0.0779	0.2318	0.4999
11.0	0.0756	0.2164	0.4667
10.5	0.0744	0.2087	0.4496
10.0	0.0732	0.2009	0.4321
9.5	0.0720	0.1932	0.4144
9.0	0.0709	0.1855	0.3964
8.0	0.0685	0.1701	0.3596
7.5	0.0673	0.1624	0.3408

n	d=0.2	d=0.5	d=0.8
6.0	0.0639	0.1393	0.2833
4.0	0.0592	0.1089	0.2047
4.0	0.0592	0.1089	0.2047

d=0.2	d=0.5	d=0.8
745.0	159.0	100.0
745.0	159.0	100.0
573.0	159.0	100.0
159.0	159.0	100.0
152.0	152.0	100.0
140.0	140.0	100.0
135.0	135.0	100.0
131.5	131.5	100.0
125.5	125.5	100.0
113.0	113.0	100.0
111.0	111.0	100.0
100.0	100.0	100.0
88.5	88.5	88.5
83.0	83.0	83.0
79.0	79.0	79.0
76.5	76.5	76.5
74.0	74.0	74.0
72.0	72.0	72.0
70.5	70.5	70.5
70.0	70.0	70.0
67.5	67.5	67.5
63.0	63.0	63.0
62.5	62.5	62.5
59.0	59.0	59.0
56.5	56.5	56.5
54.0	54.0	54.0
52.5	52.5	52.5
45.5	45.5	45.5
44.0	44.0	44.0
41.5	41.5	41.5
39.0	39.0	39.0
38.0	38.0	38.0
37.5	37.5	37.5
37.5	37.5	37.5
36.0	36.0	36.0
35.5	35.5	35.5
35.5	35.5	35.5
35.0	35.0	35.0
35.0	35.0	35.0
33.0	33.0	33.0
29.0	29.0	29.0
25.5	25.5	25.5
25.5	25.5	25.5
25.0	25.0	25.0
24.0	24.0	24.0
24.0	24.0	24.0

d=0.2	d=0.5	d=0.8
23.5	23.5	23.5
21.5	21.5	21.5
16.5	16.5	16.5
16.0	16.0	16.0
16.0	16.0	16.0
15.0	15.0	15.0
14.5	14.5	14.5
12.0	12.0	12.0
11.0	11.0	11.0
10.5	10.5	10.5
10.0	10.0	10.0
9.5	9.5	9.5
9.0	9.0	9.0
8.0	8.0	8.0
7.5	7.5	7.5
6.0	6.0	6.0
4.0	4.0	4.0
4.0	4.0	4.0