Fisher's Method

Katie Fitzgerald

06/20/2018

Fisher's test

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

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

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

Because the OSC conducted Fisher's test only among the 64 replication studies that did not have significant results and thus $p_i \geq 0.05$ for all i, they used the following transformation of Fisher's test statistic

$$X^2 = -2\sum_{i=1}^{k} ln(p_i^*) = -2\sum_{i=1}^{k} ln(\frac{p_i - 0.05}{0.95}),$$

where under H_0, the p_i^*'s follow a uniform distribution on [0,1] and thus $X^2 \sim \chi^2_{2k}$, where k is the number of studies, in this case 64. Low p-values give a larger test statistic, leading to a rejection of H_0.

We believe that Fisher's test is not well suited to assess replication. In the event that you do reject H_0, this result only tells you that at least one study was a false negative, but it does not till you *which* study did in fact have a true effect. Even still, it provides no information about the size or direction of that true effect and whether or not it replicates the original finding. Presumably, a finding of "no false negatives" would be most informative in assessing replication in this scenario, but this can never be validly concluded from Fisher's method since that would require concluding the 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 OSC was able to reject their null hypothesis ($X^2 = 155.83, p = 0.048$), we think that in general Fisher's method is underpowered to answer the question at hand. We hypothesize that it requires many and possibly large non-null effects in order to skew the distribution of p-values enough to reject Fisher's null hypothesis. In the scenario where a researcher is combining the results of k studies of the same treatment to test if an overall treatment effect $\Theta = 0$, 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 turn to simulations to investigate the power of Fisher's method in the OSC scenario.

For simplicity of interpretation but without loss of generality, we will work with the standardized scale of Cohen's d, defined as $\delta_i = \frac{\theta_i}{\sigma_i}$, where θ_i is the mean difference between the treatment and control groups in study i and σ_i is the known and equal variance among the treatment and control populations. Assuming equal sample sizes in the treatment and control groups within study i (that is, let $n_i^t = n_i^c = n_i$), δ_i is estimated by $d_i \sim N(\delta_i, \frac{2}{n_i})$.[1] Under this framework, Fisher's method can be represented in Cohen's d as testing the hypotheses

$$H_0\text{: } \delta_1 = \delta_2 = ... = \delta_{64} = 0$$

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

[1] See Appendix A.1 for proof and discussion of simplifying assumptions.

and the 64 p-values to be summed in Fisher's test statistic can be calculated as $p_i = 2(1 - \Phi(\frac{|d_i|}{\sqrt{2/n_i}}))$.

We first consider the power of Fisher's method when $n_i = 76/2 = 38$ for all i, 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 in Table 1.[2] Note that 0.2, 0.5, and 0.8 correspond to small, medium, and large effects sizes on the scale of Cohen's d. As shown in the last column of Table 1, we find that there needs to be 10 large effects in order to achieve close to the standard 80% power (power=0.7743). 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, Fisher's method only has 70% power to reject.

Table 1: Power of Fisher's method given median sample size ($n_i = 38$) for varying δ and true # of non-null effects

# 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 there is just one non-null effect we let it be from the largest study; if there are two non-null effects we let them be from the two largest studies, etc.[3] As shown in the first row of Table 2, 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).

Table 2: Power of Fisher's method given large sample sizes for varying δ and true # of non-null effects (i.e. "Best-case scenario")

# 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

[2] See Appendix A.2 for code and details on how the power simulations were conducted.

[3] Note 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, and therefore we begin the simulations with the largest n_i among the studies for which the power is at most 99.99% to detect the given δ_i. The largest sample sizes used for $\delta_i = 0.2; 0.5$; and 0.8 were $n_i = 745; 159$; and 100 respectively. See Table 4 in Appendix A.2 for the full sample size vectors that were used to conduct these "best-case scenario" power simulations.

Asymptotic power

Appendices

Appendix A: Fisher's method power simulations

A.1 Framework and assumptions

Assume each study i is testing the presence of some treatment effect θ_i, where $i = 1, ..., k$ and k is the number of studies. Assume 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 ($i = 1, ..., k$, and $j = 1, ...n_i$), and assume $Y_{ij}^t \sim N(\mu_i + \theta_i, \sigma_i^2)$ and $Y_{ij}^c \sim N(\mu_i, \sigma_i^2)$, where the σ_i^2's are 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 the OSC 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 $T_i = \overline{Y_{i\cdot}^t} - \overline{Y_{i\cdot}^c}$ 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 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, $d_i \sim N(\delta_i, \frac{2}{n_i})$.

A.2 Power simulation logic and code

Let there be m false negatives among the 64 studies, $m = 1, ..., 64$, that is m true non-null effects. 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 OSC restriction 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 let N=100,000.

Table 3: Power of OSC replicate studies to detect δ given n

n	$\delta = 0.2$	$\delta = 0.5$	$\delta = 0.8$
384351.5	1	1	1
745	0.9713	1	1
573	0.923	1	1
159	0.4299	0.9938	1
152	0.4144	0.9918	1
140	0.3873	0.9869	1
135	0.3758	0.9841	1
131.5	0.3677	0.9819	1
125.5	0.3538	0.9773	1
113	0.3242	0.9639	1
111	0.3194	0.9612	1
100	0.293	0.9424	0.9999
...	...	...	...

Table 4: Sample size vectors used in "best case scenario" power simulations for given δ's

	$\delta = 0.2$	$\delta = 0.5$	$\delta = 0.8$
1	745	159	100
2	745	159	100
3	573	159	100
4	159	159	100
5	152	152	100
6	140	140	100
7	135	135	100
8	131.5	131.5	100
9	125.5	125.5	100
10	113	113	100
11	111	111	100
12	100	100	100
13	88.5	88.5	88.5
...	...	...	...
64	4	4	4

```
power_sims<-function(N,M,delta,n){
  ##############################################################################################
  # TAKES: N; number of simulations
  #        M; vector of number of non-null effects
  #        delta; effect size under alternative hypothesis, on scale of cohen's d
  #        n; vector of treatment/control sample size across studies (total sample size/2)
  # RETURNS: power of Fisher's test to reject
  # Assumes 2-sided p-values, throws away p-values<=0.05 to match OSC methods
  ##############################################################################################

  T<-c() #empty list to store Fisher's test statistic
  Power<-matrix() #empty matrix to store results

  for(k in 1:length(M)){

    for (i in 1:N){

      #print(i) # can uncomment to show progress for lengthy simulations

      p0<-runif(64 - M[k], 0.05, 1) #draws p-values for the true null effects

      p1<-c() #create list to store p-values drawn for non-null effects

      for (j in 1:M[k]){

        p1[j]<-0

        while (p1[j] <= 0.05) { #throw away p-values<=0.05
          p1[j]<-2 * (1 - pnorm(abs(rnorm(1, delta, sqrt(2 / n[j]))) / sqrt(2 / n[j])))
        }

        #print(n[j]) # can uncomment to show progress for lengthy simulations
        }
```

```
      #test statistic for Fisher's method, with transformation for truncating p-values
      T[i]<--2 * sum(log((p0 - 0.05)/0.95)) - 2 * sum(log((p1 - 0.05) / 0.95))

      }

    Power[k]<-sum(T > 155.4047) / N

    }

    return(Power)
}
```