Fisher's Method

Katie Fitzgerald and Rrita Zejnullahi

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_j is the treatment effect for study j:

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

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

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

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

where under H_0, the p_j^*'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 $k = 64$). 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 tell you *which* study did in fact have a true non-zero effect. Even still, it provides no information about the size or direction of that non-zero 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_j has no bearing on the 63 other θ_j's, the presumed goal would be to detect if there are *any* false negatives among the replicate studies.

The true distribution of Fisher's test statistic under the alternative hypothesis is unknown, and therefore the power cannot be calculated exactly. The asymptotic distribution of X^2 can be shown to be approximately normal, but this approximation is not valid for the sample sizes in the OSC data.[1] We therefore 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 treatment effects on the standardized scale of Cohen's d, defined as $\delta_j = \frac{\theta_j}{\sigma_j}$, where θ_j is the mean difference between the treatment and control groups in study j, and σ_j is the known and equal variance among the treatment and control populations. Assuming equal sample sizes in the treatment and control groups within study j (that is, let

[1] See Appendix B

$n_j^t = n_j^c = n_j$), δ_j is estimated by $d_j \sim N(\delta_j, \frac{2}{n_j})$.[2] Under this framework, Fisher's method can be represented in Cohen's d as testing the hypotheses

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

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

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

We first consider the power of Fisher's method when $n_j = 76/2 = 38$ for all j, 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.[3] 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 need 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 has just 70% power to reject.

Table 1: Power of Fisher's method given median sample size ($n_j = 38$) 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.[4][5] 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

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

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

[4]This provides a "best-case scenario" because large n_j's lead to smaller p_j's, which in turn result in a larger test statistic X^2 and greater likelihood of rejecting H_0 (i.e. higher power).

[5]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_j among the studies for which the power is at most 99.99% to detect the given δ_j. The largest sample sizes used for $\delta_j = 0.2; 0.5$; and 0.8 were $n_j = 745; 159$; and 100 respectively. See Appendix A.3 for further discussion.

# of non-null effects	$\delta = 0.2$	$\delta = 0.5$	$\delta = 0.8$
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

Appendices

Appendix A: Fisher's method power simulations

A.1 Framework and assumptions

Assume each study j is testing the presence of some treatment effect θ_j, where $j = 1, ..., k$, and k is the number of studies. Assume equal sample sizes for the treatment and control groups ($n_{tj} = n_{cj} = n_j$). 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 $Y_{jl}^t \sim N(\mu_j + \theta_j, \sigma_j^2)$ and $Y_{jl}^c \sim N(\mu_j, \sigma_j^2)$, where the σ_j^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_j is the difference in means between the treatment and control groups, and its estimate $T_j = \overline{Y_{j\cdot}^t} - \overline{Y_{j\cdot}^c}$ has variance $v_j = \frac{\sigma^2}{n_{cj}} + \frac{\sigma^2}{n_{tj}} = \frac{2\sigma^2}{n_j}$. For simplicity of interpretation but without loss of generality, we work with the standardized scale of Cohen's d, defined as $\delta_j = \frac{\theta_j}{\sigma}$. Note then that δ_j is estimated by $d_j = \frac{T_j}{\sigma} = \frac{T_j}{\sqrt{\frac{v_j}{2/n_j}}} = \frac{T_j}{\sqrt{v_j}}\sqrt{\frac{2}{n_j}}$. Since under H_0, $T_j \sim N(\theta_j, v_j) \Rightarrow \frac{T_j}{\sqrt{v_j}} \sim N(\frac{\theta_j}{\sqrt{v_j}}, 1)$ and $\sqrt{\frac{2}{n_j}}$ is a constant, then we have $Var(d_j) = Var(\frac{T_j}{\sqrt{v_j}}\sqrt{\frac{2}{n_j}}) = \frac{2}{n_j}$. Therefore, $d_j \sim N(\delta_j, \frac{2}{n_j})$.

A.2 Power simulation logic and code

Let there be m false negatives (i.e. m true non-zero effects) 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_j from a $N(\delta_j, \frac{2}{n_j})$ distribution, where $\delta_j \neq 0$ and compute its p-value. We continue drawing $d_j's$ 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 $X^2 = -2\sum_{j=1}^{k} ln(\frac{p_j - 0.05}{0.95})$. We run this procedure N times and calculate the simulated power of Fisher's method under these conditions to be $\sum_{q=1}^{N} I_{\{X_q^2 > 155.4047\}}/N$, where I is the indicator function and $\chi_{128}^2 = 155.4047$ is the critical value for Fisher's test with $k = 64$ studies. We let N=100,000. The code is given below.

```
power_sims<-function(N,M,delta,n){
  ##############################################################################################
  # TAKES: N; number of simulations
  #        M; vector of number of non-null effects (e.g. M=c(1, 2, 3, 4, 5, 32, 64))
  #        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)
}
```

A.3 "Best-case scenario"" power simulations

As noted in footnote 4 of the text, the largest sample sizes were dropped out of necessity in the "best-case scenario" power simulations presented in Table 2. For example, the largest study had $n_j = 384351.5$, which is powered at 100% to detect even a small $\delta = 0.2$. Therefore, it would have been impossible for this study to result in a p-value greater than 0.05 if there was a true non-null effect. Table 3 presents the power of the OSC replicate studies to detect a given δ using the true sample sizes for the 64 studies. For each δ, we began the power simulations with the largest n for which the power was at most 0.9999. The largest sample sizes used for $\delta_j = 0.2; 0.5;$ and 0.8 therefore were $n_j = 745; 159;$ and 100 respectively. Table 4 presents the sample size vectors used for the "best-case scenario" power simulations.

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

n	$\delta = 0.2$	$\delta = 0.5$	$\delta = 0.8$
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

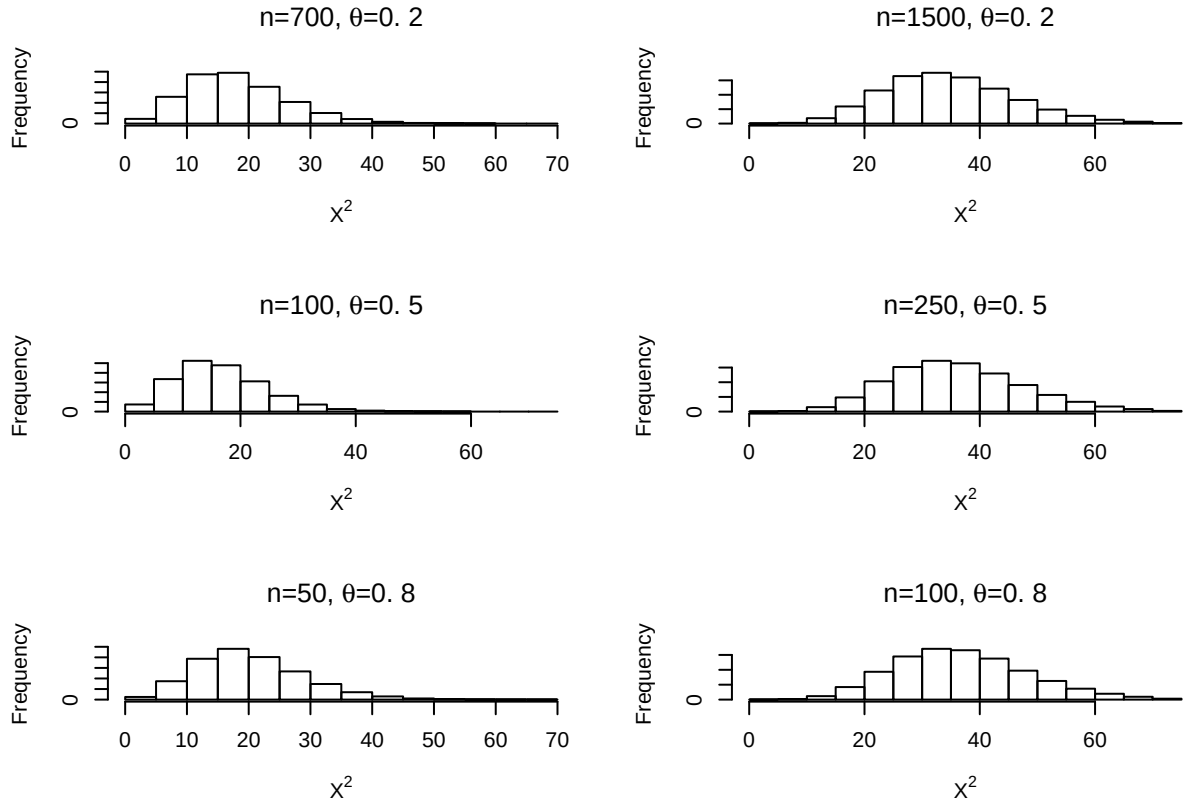
Appendix B: Fisher's method asymptotic results

Under the Neyman and Pearson hypothesis testing framework, the significance level follows a uniform distribution on $[0, 1]$ when the null hypothesis holds, however, the exact distribution under the alternative hypothesis is unknown. As a consequence of this, we rely on long known results from asymptotic theory. Lambert and Hall have shown that, given the test statistic is asymptotically normal, the one sided P-value follows a lognormal distribution with mean $-nc(\theta)$ and variance $n\tau^2(\theta)$ (1982). The parameter $c(\theta)$ is defined as half the Bahadur slope, given by $-\frac{1}{n}lim_{n\to\infty}logP_n = c(\theta)$, and is the exponential rate at which the significance level converges to zero under the alternative hypothesis. In addition, observe that the variance of the standardized P-value is $\frac{\tau^2(\theta)}{n}$. We can approximate the two sided P-value by doubling the one sided one, which implies that $P_n \sim AlogN[-\frac{1}{2}nc(\theta), \frac{1}{4}n\tau^2(\theta)]$, and thus $logP_n \sim AN[-\frac{1}{2}nc(\theta), \frac{1}{4}n\tau^2(\theta)]$. Multiplying $logP_n$ by -2, we obtain $-2logP_n \sim AN[nc(\theta), n\tau^2(\theta)]$. Note that $-2logP_{n_j}$, $j = 1, ..., k$, are asymptotically independent random variables following identical distributions, therefore under H_a

$$X_n^2 = -2\sum_{j=1}^{k} logP_{n_j} \sim AN(\sum_{j=1}^{k} n_j c_j(\theta), \sum_{j=1}^{k} n_j \tau_j^2(\theta))$$

.

Examining the behavior of X_n^2 using simulated data, we find that the normal approximation is valid if the within study sample size n_j is at least 1500, 250, and 100 for θ equal to 0.2, 0.5, and 0.8, respectively. Note that θ in this case can be interpreted as Cohen's d because σ is assumed to be 1. When the within study sample sizes are smaller than indicated previously, the distribution of the test statistic is skewed to the right, and therefore the normal approximation is not valid.



Now consider the two sample shift problem as an example. Let $Y_{11}, Y_{21}, ..., Y_{n1}$ denote a sample of i.i.d. observations from a normal distribution with mean μ and standard deviation 1. Let $Y_{12}, Y_{22}, ..., Y_{n2}$ denote a second sample, independent of the first, with i.i.d. observations from a normal $(\mu + \theta, 1)$. For simplicity, suppose that n_1 and n_2 are equal. We test $H_0 : \theta = 0$ versus $H_a : \theta \neq 0$ by using $\frac{\sqrt{n}(\overline{Y_2} - \overline{Y_1})}{\sqrt{2}}$. According to Lambert and Hall, $c(\theta) = \frac{1}{2}\lambda\overline{\lambda}\theta^2$ and $\tau^2(\theta) = \lambda\overline{\lambda}\theta^2$, where λ denotes the fractional sample size, and $\overline{\lambda} = 1 - \lambda$. Since we are assuming n_1 is equal to n_2, both λ and $\overline{\lambda}$ are $\frac{1}{2}$. Then, if all k studies are testing the same hypothesis, it follows that X_n^2 is asymptotically normal with grand mean $\sum_{j=1}^{k} \frac{1}{2} n_j \lambda_j \overline{\lambda_j} \theta_j^2$ and grand variance $\sum_{j=1}^{k} n_j \lambda_j \overline{\lambda_j} \theta_j^2$. Consequently, the asymptotic power is:

$$
\begin{aligned}
Power &= Pr(reject\ H_0 \mid H_a\ is\ true) \\
&= 1 - Pr(fail\ to\ reject\ H_0 \mid H_a\ is\ true) \\
&= 1 - Pr(|X_n^2| \ < \ \chi_{2k}^2 \mid H_a\ is\ true) \\
&= 1 - \Phi\Big(\frac{\chi_{2k}^2 - \sum_{j=1}^{k} \frac{1}{2} n_j \lambda_j \overline{\lambda}_j \theta_j^2}{\sum_{j=1}^{k} n_j \lambda_j \overline{\lambda}_j \theta_j^2}\Big) + \Phi\Big(\frac{-\chi_{2k}^2 - \sum_{j=1}^{k} \frac{1}{2} n_j \lambda_j \overline{\lambda}_j \theta_j^2}{\sum_{j=1}^{k} n_j \lambda_j \overline{\lambda}_j \theta_j^2}\Big). \\
&= 1 - \Phi\Big(\frac{\chi_{2k}^2}{\sum_{j=1}^{k} n_j \lambda_j \overline{\lambda}_j \theta_j^2} - \frac{1}{2}\Big) + \Phi\Big(\frac{-\chi_{2k}^2}{\sum_{j=1}^{k} n_j \lambda_j \overline{\lambda}_j \theta_j^2} - \frac{1}{2}\Big) \\
&= 1
\end{aligned}
$$

The following plots of the power function reveal that asymptotic power approaches 1 when the total within

study sample size is greater than 600, 100, and 50 for a fixed θ of 0.2, 0.5, and 0.8, respectively.