Fisher's Method

Katie Fitzgerald and Rrita Zejnullahi

08/31/2018

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. Because only 3 of the original studies found null effects, when using the "sign and significance" metric to assess replication, the OSC took this to indicate that there was widespread failure in replication. In general, a finding of a null effect indicates that either the true effect is in fact zero or that it is non-zero but the test was underpowered to detect it. Hoping to determine which of these two scenarios is true of their replication studies, the OSC applied Fisher's method to the set of 64 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. That is, they wish to rule out the possibility that being under-powered was the only reason for failure to find significant results in the replication studies. Their hypothesis for Fisher's test could be formalized as follows, where θ_j is the treatment effect for study j:

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

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

Note that since $p_j \geq 0.05$ for all j in this scenario, the following transformation of Fisher's test statistic was used

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

where $X^2 \sim \chi^2_{2k}$ under H_0 and 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 distinguish between low power and the existence of false negatives. 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.

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 as within study sample sizes go to infinity, but this approximation is only valid for impractical sample sizes and detectable effect sizes.[1] We therefore turn to simulations to investigate the power of Fisher's method.

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 equal variance among the treatment and control populations in study j. Assuming equal sample sizes in the treatment and control groups within study j (that is, let $n_j^t = n_j^c = n_j$), δ_j is estimated by $d_j \sim N(\delta_j, \frac{2}{n_j} + \frac{\delta_j^2}{4n_j})$.[2] Under this framework, Fisher's method can be represented in Cohen's d as testing the hypotheses

[1] See Appendix B

[2] Hedges & Olkin (1985).

$$H_0\text{: } \delta_1 = \delta_2 = ... = \delta_{64} = 0$$
$$H_a\text{: at least one } \delta_j \neq 0, \quad j = 1, ..., 64,$$

and the 64 p-values to be summed in Fisher's test statistic can be calculated as $p_j = 2(1 - \Phi(\frac{|\delta_j|}{\sqrt{\frac{2}{n_j} + \frac{\delta_j^2}{4n_j}}}))$.

In order to consider the power of Fisher's test under a "best-case scenario" in the OSC 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] This will provide the highest possible power of Fisher's method given these data.[4] As shown in the first row of Table 1, even when the study with the largest 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).

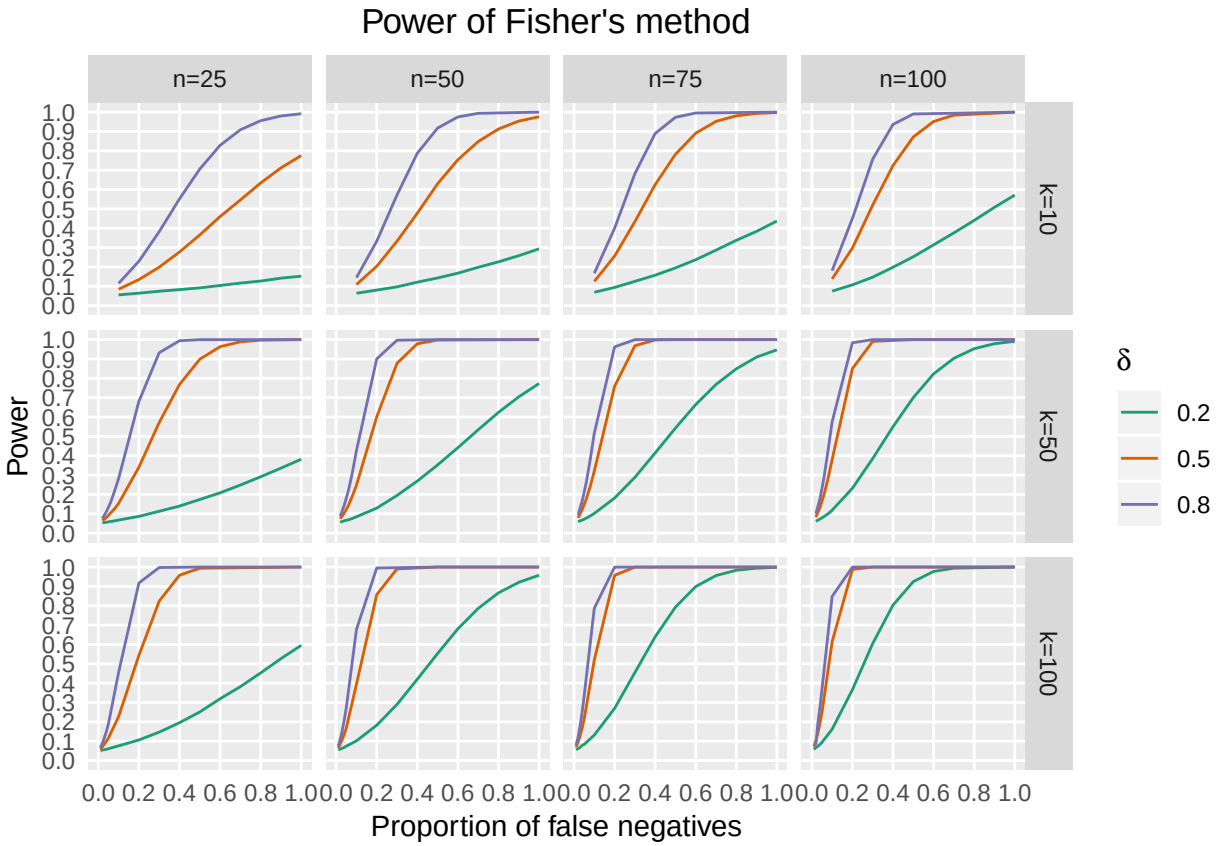
Table 1: Power of Fisher's method given large sample sizes for varying δ and true # of non-null effects (i.e. "Best-case scenario")

# 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

We now consider the power of Fisher's method under more general circumstances. Figure 1 plots the power of Fisher's method against the proportion of studies with false negatives, for varying numbers of total studies (k), effect sizes (δ), and within-group within-study sample sizes (n).

[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_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.2 for further discussion.

[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).



For example, in the first pane of the grid, the green curve indicates that when there are 10 studies each having within-group sample sizes of 25, even when all 10 studies have true small effects, the power of Fisher's method is only around 15%. Note the general pattern: the power of Fisher's method increases as within-group sample size increases (left to right in each row), as the number of studies increases (top to bottom in each column), and as the magnitude of the true effect size increases (indicated by the color of the curve).

As demonstrated by Figure 1, Fisher's method has low power and thus a large Type II error rate under many common scenarios in social science research, and we do not recommend it as a valid metric for assessing replication. Failing to reject H_0 is often likely an artifact of the poor properties of Fisher's method and is not valid evidence to conclude there are no false negatives. The test is only adequately powered when there are many effects, potentially large in magnitude and coming from large studies. 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. Furthermore, in the event that H_0 is rejected, it tells the researcher nothing about *which* study has a non-zero effect or if the size or direction of that effect is consistent with the original study.

Appendix A: Fisher's method power simulations

A.1 Power simulation logic and code

Let there be m false negatives (i.e. m true non-zero effects) among the k studies, $m = 1, ..., k$. 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 $N(\delta_j, \frac{2}{n_j} + \frac{\delta_j^2}{4n_j})$ 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 $k-m$ 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 $\sum_{q=1}^{N} I_{\{X_q^2 > \chi_{2k,\alpha}^2\}}/N$, where I is the indicator function and $\chi_{2k,\alpha}^2$ is the critical value for Fisher's test with k studies and level α. We let N=100,000. The code is given below.

```
power_sims<-function(N,M,delta,n,k){
  ##############################################################################################
  # 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)
  #        k; number of studies
  # RETURNS: 5 column dataframe of results [Power N k M delta]
  # 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(l in 1:length(M)){
    for (i in 1:N){
      p0<-runif(k - M[l], 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[l]){
        p1[j]<-0
        while (p1[j] <= 0.05) { #throw away p-values<=0.05
          z[j] <- rnorm(1,delta, sqrt(2/n[j] + delta^2/(4*n[j])))
          p1[j] <- 2*(1-pnorm(abs(z[j])/sqrt(2/n[j] + delta^2/(4*n[j]))))
        }
      }
      #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[l]<-sum(T > qchisq(0.95, 2*k) )/N
    print(power[l])
    if(power[l]>0.99) break
  }
  data <- as.data.frame(cbind(power = power,
                              n = n[1:length(power)],
                              k = rep(k,length(power)),
                              M = M[1:length(power)],
                              delta = rep(delta,length(power))))
  return(data)
}
```

A.2 "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 1. 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 small effect. Table 3 presents the power of the 12 largest OSC replicate studies (among the subset of 64) to detect a given δ. For example, the twelfth largest OSC study had a within-group sample size of 100 and had 29% power to detect a small effect, 94% power to detect a medium effect, and >99% power to detect a large effect. 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 3 presents the resulting sample size vectors used for the "best-case scenario" power simulations.

Table 2: Power of 12 largest OSC replicate studies to detect δ given n

Study	n	$\delta = 0.2$	$\delta = 0.5$	$\delta = 0.8$
1	384351.5	1.0000	1.0000	1.0000
2	745.0	0.9713	1.0000	1.0000
3	573.0	0.9230	1.0000	1.0000
4	159.0	0.4299	0.9938	1.0000
5	152.0	0.4144	0.9918	1.0000
6	140.0	0.3873	0.9869	1.0000
7	135.0	0.3758	0.9841	1.0000
8	131.5	0.3677	0.9819	1.0000
9	125.5	0.3538	0.9773	1.0000
10	113.0	0.3242	0.9639	1.0000
11	111.0	0.3194	0.9612	1.0000
12	100.0	0.2930	0.9424	0.9999

Table 3: 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 $n\tau^2(\theta)$, where n is the within group sample size (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 $\frac{\tau^2(\theta)}{2n}$. 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 independent random variables asymptotically following Normal distributions, therefore under H_a

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

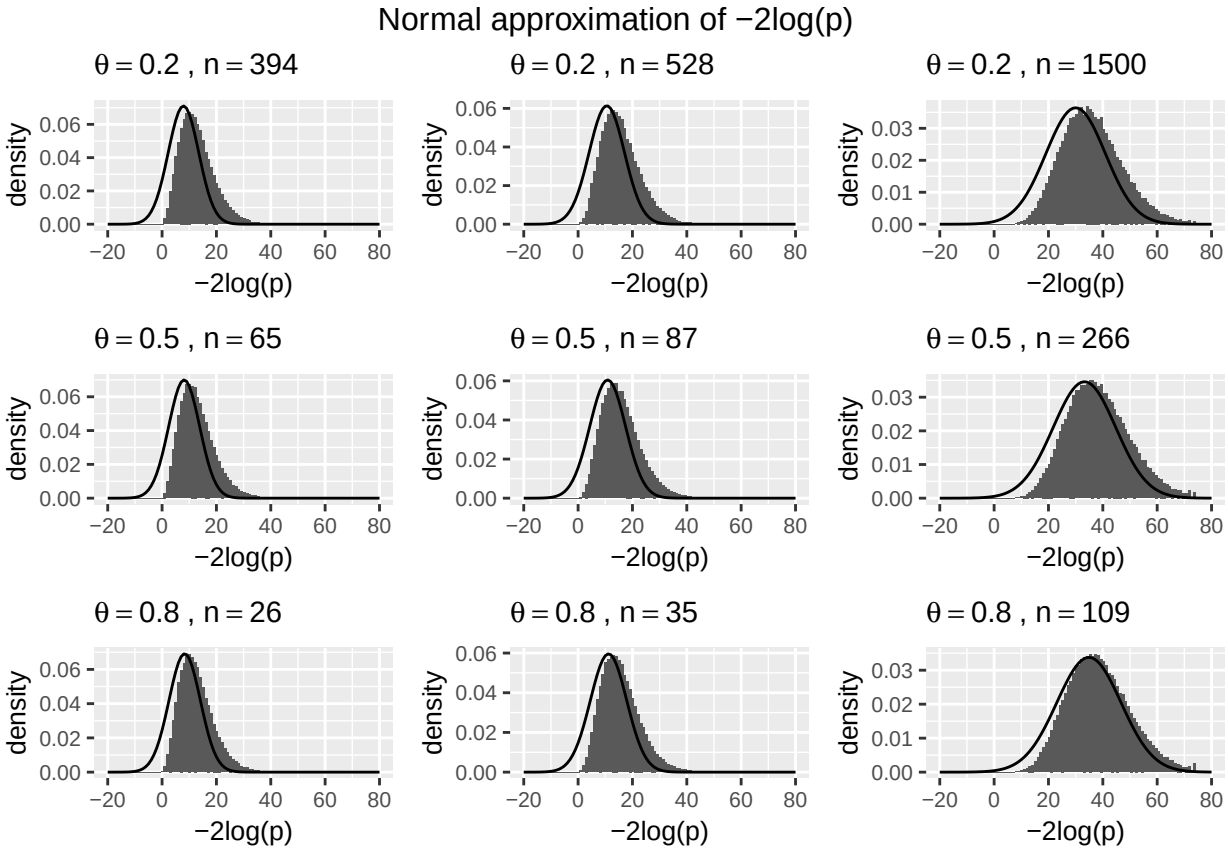
.

Now consider the two sample shift problem as an example. Let $Y_{11}, Y_{21}, ..., Y_{n_1 1}$ denote a sample of i.i.d. observations from a normal distribution with mean μ and standard deviation 1. Let $Y_{12}, Y_{22}, ..., Y_{n_2 2}$ denote a second sample, independent of the first, with i.i.d. observations from a normal $(\mu + \theta, 1)$. For simplicity, suppose that $n_1 = n_2 = n$. We test $H_0 : \theta = 0$ versus $H_a : \theta \neq 0$ using the test statistic $\frac{\sqrt{n}(\overline{Y_2} - \overline{Y_1})}{\sqrt{2}}$. 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 $n_1 = n_2$, both λ and $\overline{\lambda}$ are $\frac{1}{2}$. Therefore, in the two-sample shift problem we have

$$-2logP_n \sim AN(\frac{1}{4}n\theta^2, \frac{1}{2}n\theta^2).$$

Note that θ in this case can be interpreted as Cohen's d because σ is assumed to be 1.

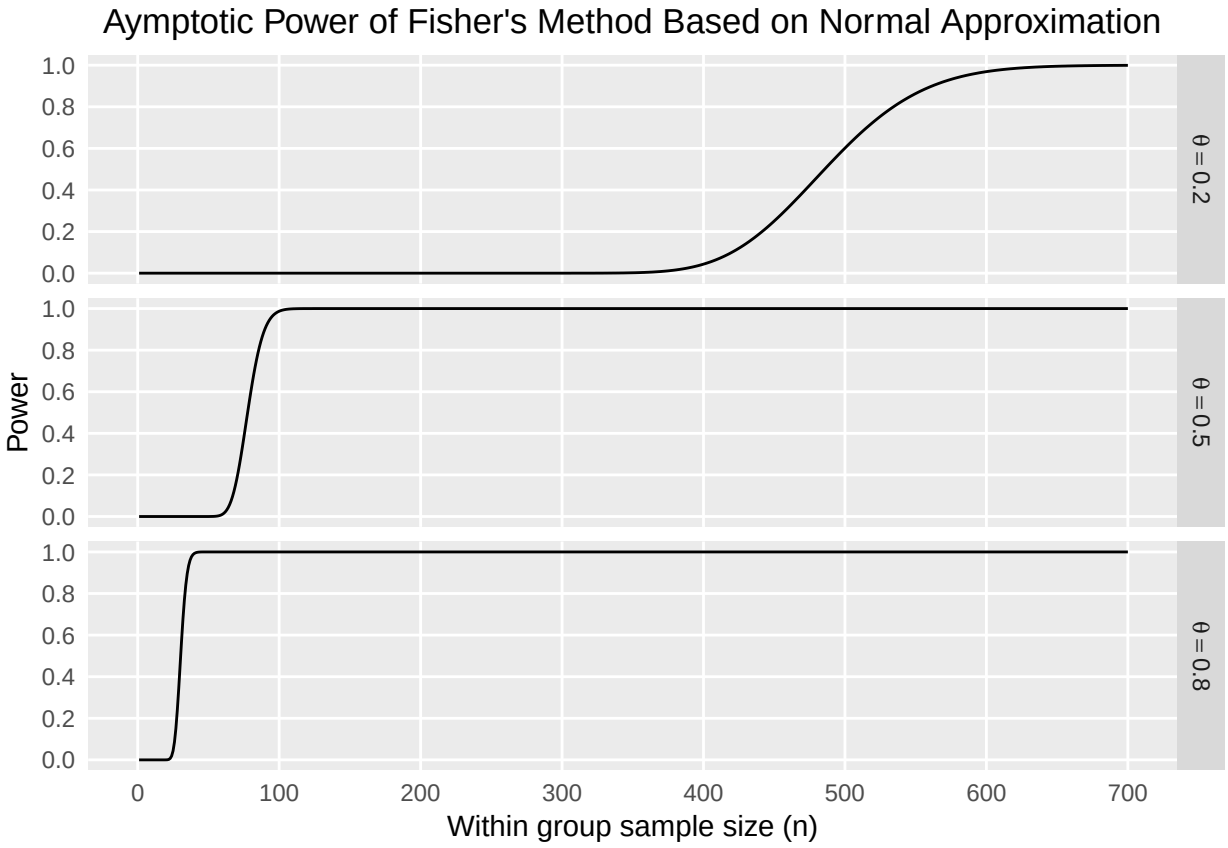
Examining the behavior of $-2log(p)$ using simulated data, we find that the normal approximation is only valid for impractical within study sample sizes. The following figure shows histograms of simulated data and the corresponding normal approximations for various values of θ and n. Since the normal approximation is a function of n and θ, it can also be considered a function of the power of a single test to detect the specified θ. Values of n were chosen to achieve power of 0.80, 0.90, and 0.9999 for each θ. For example, when $\theta = 0.2$, a within-group sample size of $n = 394$ is needed to achieve 80% power, and the first plot in the grid shows that the normal appoximation is poor in this scenario. Note that in most cases, the lower tail of the normal approximation includes negative values, whereas we know the true values of $-2log(p)$ can never be negative. Unless the within study sample sizes are very large, the distribution of $-2log(p)$ is skewed to the right, and therefore the normal approximation is not valid. Even in scenarios with sample sizes large enough the achieve 99.99% power (the third column in the figure), the normal distribution is only a rough approximation.



The asymptotic results can confirm that the power of Fisher's method approaches 1 as the within study sample sizes go to infinity, which is consistent with the power simulations presented in the body of the paper. To see this, note that 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 of Fisher's test is:

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

Shown for the the case when $k = 64$, the following plots of the power function reveal that asymptotic power approaches 1 when the total within study sample sizes are approximately 600, 100, and 40 for fixed θ's of 0.2, 0.5, and 0.8, respectively.


Aymptotic Power of Fisher's Method Based on Normal Approximation
Power
1.0
0.8
0.6
0.4
0.2
0.0
θ = 0.2
θ = 0.5
θ = 0.8
0
100
200
300
400
500
600
700
Within group sample size (n)