FISHING EXPEDITIONS

Sparsh Sah

CONTEXT

In the following, assume that all experimental-design considerations are airtight:

- The study is double-blind,
- The participants are truly randomly selected,
- The control group achieved exactly $\hat{p} = p_0 = 0.50$,
- The Normal approximation to the Binomial is good enough (so we can use t-stats),
- Etc.

The point here is not to scrutinize experimental design, although that is a perfectly reasonable thing to scrutinize in real life. Also, I already alluded to this, but assume that without treatment every person has i.i.d.

$$p_0 := \Pr[\text{survive} \mid \text{no drug}] = 0.50.$$

We wish to investigate $p := \Pr[\text{survive} \mid \text{drug}]$. Note: When I use p herein, I mean a probability (proportion of survivors), not a p-value in the frequentist-inference framework.

THE EXPERIMENTS

- Alice gives her drug to 60 patients. She observes that exactly 30 survive. The t-stat here under the null that the drug has no effect (i.e. $p = p_0$) is, based on the Binomial distribution,

  $$t = \frac{S - Np_0}{\sqrt{Np_0(1-p_0)}} = \frac{30 - 60(0.50)}{\sqrt{60(0.50)(1-0.50)}} = 0,$$

  So Alice concludes there is no statistically-significant evidence that her drug works.

- Bob notices that in Alice's data, there were 40 women and 20 men. All 30 of the survivors were women. He calculates the t-stat for women as

  $$t_w = \frac{S_w - N_w p_0}{\sqrt{N_w p_0(1-p_0)}} = \frac{30 - 40(0.50)}{\sqrt{40(0.50)(1-0.50)}} = +3.16 \qquad > +1.96,$$

  and the t-stat for men as

  $$t_m = \frac{S_m - N_m p_0}{\sqrt{N_m p_0(1-p_0)}} = \frac{0 - 20(0.50)}{\sqrt{20(0.50)(1-0.50)}} = -4.47 \qquad < -1.96,$$

  Thereby concluding that actually, Alice's experiment yielded statistically-significant (at the $\alpha = 5\%$ level) evidence that her drug was very effective for women but actually actively increased mortality for men.

- Carol wants to verify Alice's result for herself. She conducts an identical experiment, but this time all 60 participants survive. She calculates the t-stat as

  $$t = \frac{S - Np_0}{\sqrt{Np_0(1-p_0)}} = \frac{60 - 60(0.50)}{\sqrt{60(0.50)(1-0.50)}} = +7.74 \qquad > +1.96,$$

  Thus concluding that this fresh experiment has yielded statistically-significant evidence that the drug works.

- Finally, Dave does a meta-analysis pooling Alice's and Carol's data. He calculates the pooled t-stat as

  $$t = \frac{S - Np_0}{\sqrt{Np_0(1-p_0)}} = \frac{(30+60) - (60+60)(0.50)}{\sqrt{(60+60)(0.50)(1-0.50)}} = +5.48 \qquad > +1.96,$$

  And therefore concludes that the drug works.

THE QUESTION

Which investigators' methods were valid from a frequentist-inference perspective? Answer: Only Alice and Dave.

- Bob's was a classic fishing expedition: Without preregistering his hypothesis, he was free to slice-and-dice his way to some arbitrary p-hacked conclusion. If he hadn't found an ostensibly-significant result among women, he could have split them up by zip code and tried each group like that, and if he still failed, he could split them up by the color of their socks and tried yet again.

- Carol's was a more insidious fishing expedition: She saw results she didn't like, and therefore just tried the experiment again. Even though her experiment in isolation seemed valid (indeed it was set up identically to Alice's), if her results had again turned up null, she could have just tried again and again and again until one flagged positive. Assume the drug has no effect. Because we're using the $\alpha = 5\%$ level, using the expected value of the geometric distribution (which models the number of trials until and including the trial where you hit a success), she'd expect to have to run just 20 trials before getting a success, despite the fact that the drug has no effect. Or from another perspective (using not the geometric but the Binomial distribution), if she committed to running 20 trials, she should expect one to flag positive just by random chance, again despite the fact that the drug has no effect.

- Why is Dave's meta-analysis valid despite being the union of one experiment deemed valid and one experiment deemed invalid? Because Dave isn't giving the drug a fresh second chance to work – He's essentially just collecting more data to add to the original data. This is perfectly valid and indeed encouraged! If you get a null result, the issue could simply be that you don't have enough statistical power – Your sample size is too small, so your standard errors are too wide. By going out and collecting new data to add to the original data, you're compressing the standard error, and getting closer to the truth. By the Central Limit Theorem, under some mild assumptions and given that we're using the MLE, as you converge to infinite sample size, the estimate $\hat{p}$ will converge to the ground truth p. If the drug has an effect, this value will be some $p > 0.50$, but if the drug truly has no effect i.e. $p = 0.50$, your $\hat{p}$ is going to get squeezed inexorably closer to 0.50. In fact, for this reason, it is even valid (I think) to say "I am so confident that my drug works that I'm going to keep adding new participants to my study for the rest of my life until I get a positive result" – If you're right, you'll eventually get your result with probability one, but if you're wrong, the more data you add in vain the more your $\hat{p}$ will keep getting squeezed closer and closer to 0.50.