Video 4: p-hacking

wbg231

December 2022

introduction

- vedio link
- p values are everywhere in science
- they are often a request for publication because they show that what we see is not due to random flucuations
- but they are not the only metric that should be taken into account
- a small p value neither implies practical signefgence nor casual effects

practical signefgence

- we want to evaluate the cure rete of two really expensive drugs with a lot of side effects.
- in group 1A, where people do not recive the drug 30 out of 100 people recover
- in group 2A, where people recive drug A 52 out of 100 recover
- in group 1B, where people do not recive the drug 30,000 out of 100,000 people recover
- in group 2B, where people recive drug A 30650 out of 100.000 recover
- this seems to suggest that drug two is a lot less effecive than drug 1.
- lets apply a two sample z test
- null there is no difrent between the control and treatment group we assume all data are iid bern with cure rate parameter θ
- test stat difrence in cure reate between treatment and control groups
- under null the test stat $\tilde{t} \sim \mathcal{N}(0, \theta(1-\theta)(\frac{1}{n_{\text{treatment}}} + \frac{1}{n_{\text{control}}}))$

- notice that our variance depends a lot on the number of people in the study
- for drug one test stat is $t_data = .2$ and this yields a very small p-value
- for drug two our test stat is 0.007 but with a much smaller test sttat, so we have the same p value in the first trial
- this is intresting as they have the same signefgence level despite the fact that drug one makes much more of a difference in actual cure rate

what does this mean

- both ressults equally unlikely under the null
- both increase the cure rate
- but they increase the cure rete by difrent amount
- so how do we quantify this difference?
- we can do a confidince interval for the difference in cure rate

confidince interval

- let the true control cure rate be θ_c
- number of cured control subjects $\tilde{k}_c$
- this is the number of cured paitenets in the control group is distrbuted as a binomal with parameters n_C, θ_c
- we can apply a gaussian aproximation to the binomial with mean $n_C\theta_c$ and varinace $n_c(\theta_c)(1-\theta_c)$
- thus our observed control cure rate $\frac{\tilde{k}_C}{n_c} \sim \mathcal{N}(\theta_c, \frac{\theta_c(1-\theta_c)}{n_c})$
- we can do the same thing for the treatment rate
- so we can think of the difference between the cure rate in the treatment and cure rate as distrobution as $\sim \mathcal{N}((\theta_t - \theta_c), \frac{\theta_c(1-theat_c)}{n_c} + \frac{\theta_t(1-\theta_t)}{n_T})$
- so we can build a gaussian confidince inverval around a guassian rv $\tilde{a}$ using mean μ, σ^2
- as $\tilde{\mathcal{I}}_{1-\alpha} := [\tilde{a} - c_\alpha\sigma, \tilde{a} + c_\alpha\sigma]$ where c_α is some constnat
- we see that the ci for drug 1 is between 8 adn 35 percent
- drug 2 on the other hand as a difference between .25 and 1.05%
- so these two have the same p-value, but only drug 1 has a real effect

statstical vs practical signefgence

- so more or less, if there is a test with a tone of power it can pick an effect that is so small that it likely does not matter in reality

covid example

- an over powerd study can make almost any effect size signefgence so it is important to think if the difference is practically significant

publication bias

- we would expect the null to hold. ie that pizza does not curve covic -19
- so we will expect to see p -value unfiormally dstributed between 0 and 1
- so we would expect 5% of studies to find a false postive
- but if we do a lot of tests say 100 studies, then we would still excpet 5%
- we wouldd excpect to see 5 false postives, and 95 true negatives
- this is not a big deal if all of these ressults were published
- but it is much easier to publish a false posrive than a true negative
- so when this is done inentially this is called p-hacking
- part of the issue with this is that it may not be repoducable later
- so this is a reprodability issue.
- there are a lot of ways to fudge your stats in order to get statstical significant (this is unethical)
- it is an incentive problem

does p hacking happen in practice

- the short awnser is yes
- if we look at the distrobution of p values on pub med,
- there is evedince that there are more studies published with lower p values