video 2: multiple testing

wbg231

December 2022

introduction

- vedio link
- the whole idea is to avoid false postives

cluth

- a player is clutch if they play better when it matters
- data 3 points shooting in nba
- clutch time: 4th quarter of close games
- conjecture: player shoots better in the clutch
- null: players shoots the same
- test stat: 3's made in clutch
- so lets set up a hypothesis test
- under the null the percentage of making a clutch 3 is the season 3 point percentage
- the test stat under the ull is a binomal with parameter n adn θ_{season}
- item so our p value is $pv(t_{data}) = P(\tilde{t} \geq t_{data}) = \Sigma_{i=t_{data}}^{n} \binom{n}{i} \theta^i (1-\theta)^{n-i}$
- we can check this for each player and see how well this holds
- there are a few players that have low p-values on the first half of the season
- does this convince you?
- no we may want to test our conlusions on held out data.
- so we test those same players on the second half of the season and those p values no longer hold
- so what is going on?

what is going on

- the liklyhood of a single player overpreforming by chance is quite low
- but we are looking at all players in the nba so that is 146 total players
- so the liklyhood that a few of them overpreformed by chance is much higher than α

p -value distrobtion

- the p -vale distrotuon under the null is uniform in zero and 1 (this aprxemently holds for discrete as well)
- so the distrobution of p values for a single player is distrbuted unfiormally between zero and one
- and thus the liklyhood of a false postive in that case is α
- but in our example we are doing many hypothsis tests. there are over 146 players in our data set
- so how many false postives are we likely to see?
- aprxemently a fraction equal to α
- item so in other words we would expect 5% of the total players to be false postives

multiple testig

- spose we preform k indpednt hyptohsis test with signgence level α
- the probability of a false postive for any test is α
- $P(\geq 1\text{false postive}) = 1 - P(\text{no false postives}) = 1 - (1-\alpha)^k$ (given the tests are independt)
- if k=100 and $\alpha = .05$ then the liklyhood of at least 1 false postive is 99%
- so what can we do? we could lower α

chalange

- we want to find a value α such that we keep the liklyhood of any valuse postives bellow α while doing k indepdnint hypothsis tests at the same time
- $P(\text{false postive}) = P(\cup_{i=1}^{k}(\text{false postive in test i})$

union bound

- for events $A_1..A_k$
- $P(\cup_{i=1}^k A_i) \leq \Sigma_{i=1}^k P(A_i)$

- ## bonferroni's correction

- how to set the p value threshold τ so that $P(\text{false postive}) \leq \alpha$
- $P(\text{false postives}) = P(\cup_{i=1}^k \text{false postive in test i }) \leq \Sigma_{i=1}^k P(\text{false postve in test i }) \leq k\tau = \alpha$
- so in other words we reject the null if p value $\leq \tau := \frac{\alpha}{k}$
- this garuntees that $P(\text{false postve}) \leq \alpha$

back to clutch example

- we are testing 146 players
- so if set our α to $\alpha_{mt} = \frac{\alpha}{146}$

example 2

- goal evalute impact of a single player on team preformance
- stat $t_d ata := m_{with} - m_{without}$ that is the mean number of points with our with out the player
- our data is nba games between 2012 and 2018
- we see that some players that do not play that much have a very high test stat
- the issue is that players who did not play that much may have a lot of noise

hypothsis test

- so we can approach this as a permuation test
- we do a montecarlo subset of the permutations and estiamte the p value using a permuation test
- doing this we see small p-values.

- are we convinced? no there are so many players so it is really easy to get false postives
- so we can apply bonferronis correction and see how that affects things
- this gives us lebron james is the only real sigfnget test
- so if we sort by p-value the list starts makign a lot of sense
- so by ordering the p-values we can see players that overall have strong impact on there games
- this is also weighted by how much evedine we have in favor of the evedince of the players

p value distrobtion

- bonferronis correction bascailly zooms into the uniform p values to such a point that it is unlikely that there will be players that are sigfnget due to noise alone.
- but this does natrually reduce power
- so we see kevin durant, is not listed as statistically signefgence with bonferronis correction
- so we want to think about how this impacts our power.
- so testing many hypothesis at the same time and being very strict about our number of false postives forces us to incur some false negatives
- there is a trade off and sometimes it is better to allow for some more false postives so that you can still avoid a lot of false negatives.