Replicability Intro

The Scientific Method (Karl Popper)

Deductive methods of testing:

- Formulate hypothesis
- Collect data to test predictions
- Test hypothesis

What Has Changed? Big data and a new scientific paradigm

Collect data first $\Longrightarrow$ ask questions later

- Large data sets available prior to formulation of hypotheses.
- Need to adjust inference to reflect the fact that hypotheses generated by data snooping.
- Data snooping: Looking at the data to find some interesting effect, or testing many hypothesis on the data in order to find something interesting.

DS 101 and “replicability”

There are many reasons why studies may not replicate well.

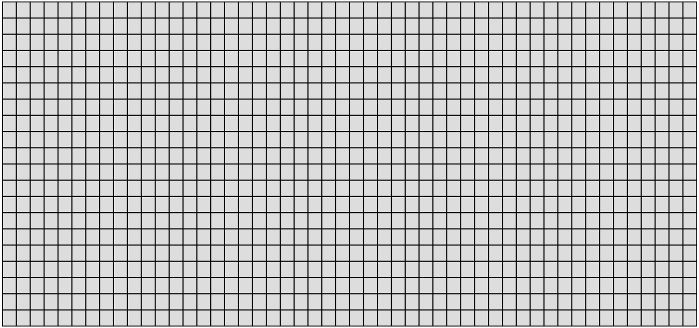
Our focus is:

- Multiple testing (look-everywhere effect)
- Winner’s curse (selection bias)

Intro: Testing Multiple Hypotheses

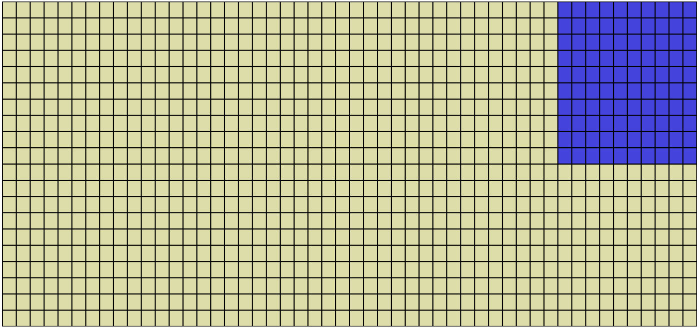
- Suppose we run a regression of an outcome Y on 1000 different predictors X. However, none of the 100 predictors are related to Y (they are independent, everything is just "pure noise").
- We then test $\hat{\beta} = 0$ vs. $\hat{\beta} \neq 0$ for all 1000 predictors.
- We reject each hypothesis if we observe a p-value that is less than 5%.
- Will we find significant results? How many?

Most discoveries may be false: Soric (1989)



1000 hypotheses to test

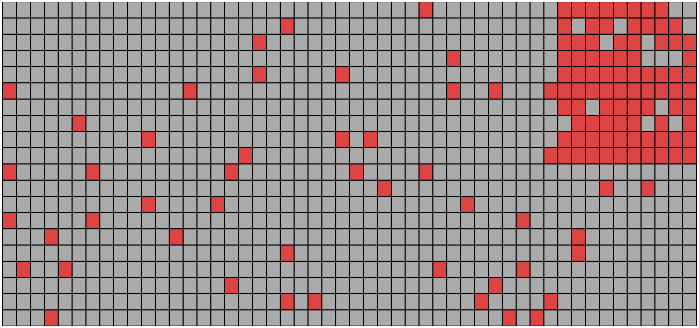
Most discoveries may be false: Soric (1989)



Nothing going on

Something going on

Most discoveries may be false: Soric (1989)

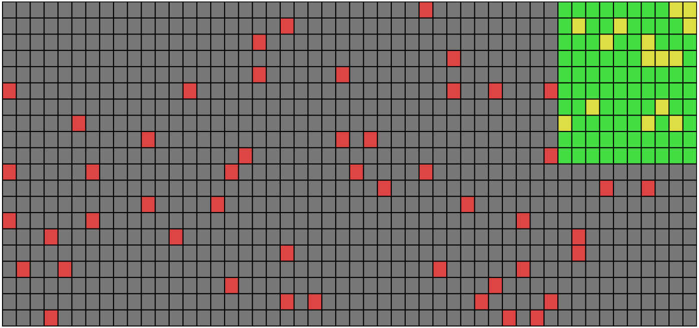


For each of the 1000 hypotheses, we make a decision: we make a decision: P(false positive)=0.05, P(false negative)=0.2.

The decisions we made are shown in the picture.

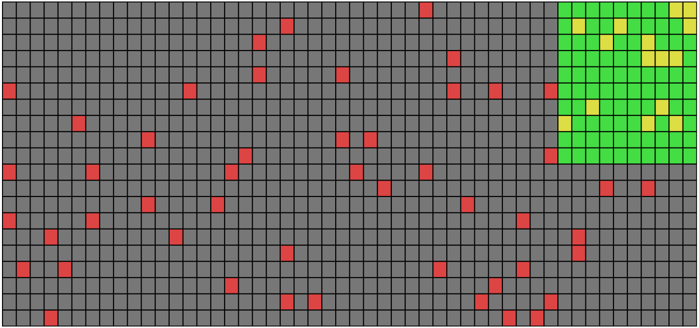
- Discovery
- Not a discovery

Most discoveries may be false: Soric (1989)



- We made 85 true discoveries
- We made 49 false discoveries
- Our *False Discovery Proportion* is 49/134=0.37.

Most discoveries may be false: Soric (1989)



Is this a problem?

Can we fix it?