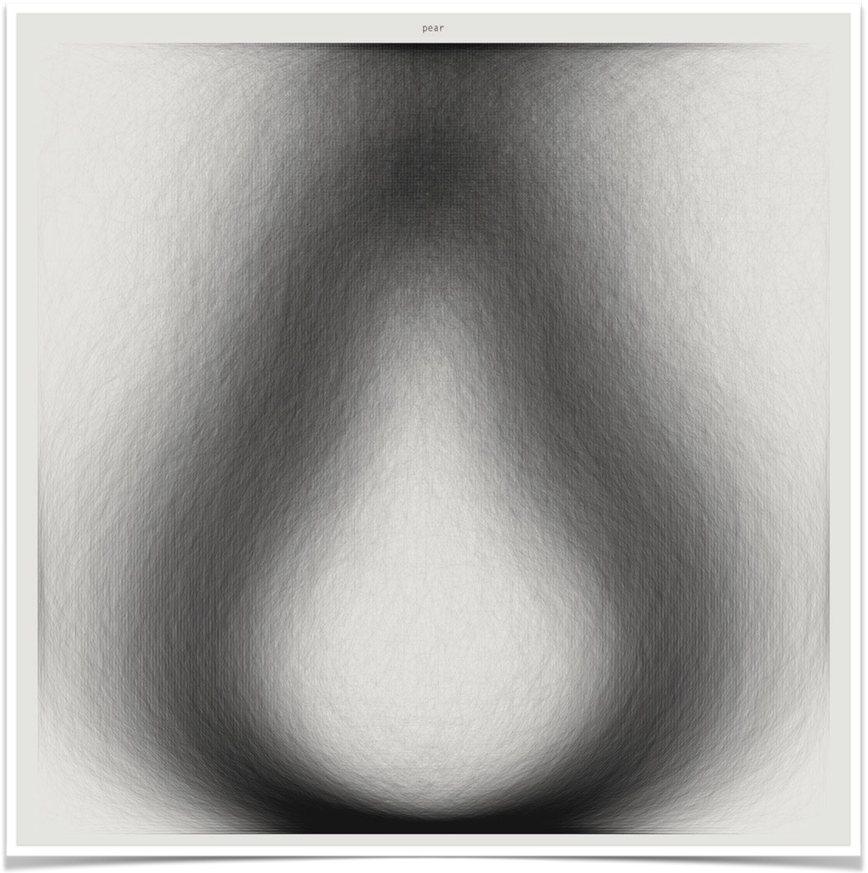
IT UNIVERSITY OF COPENHAGEN

Reflections in Data Science

BSREDAS1KU-20201

2020-03-03



Pear,
Overlap of drawings from people all around the world
formafluens.io

Exercise 4 - Plan

- Hack your way to scientific glory (10 minutes)
- P-Hacking. (20 minutes)
- Multiple Comparisons Bias and Bonferroni’s correction. (20 minutes)

Hack your way to scientific glory

IF YOU
TORTURE DATA
LONG ENOUGH
THEY WILL
CONFESS.

-RONALD COASE

https://projects.fivethirtyeight.com/p-hacking/

- Please record the configurations you found more interesting.

Hack your way to scientific glory

- The data can be narrowed and expanded (p-hacked) to make either hypothesis appear correct.

- Answering even a simple scientific question requires lots of choices that can shape the results.

- This doesn’t mean that science is unreliable. It just means that it’s more challenging than we sometimes give it credit for.

P-hacking

- A study made a hypothesis that green dice are loaded and that they will yield a six outcome more often than usual dice would.

- They took a green die, and rolled it 1000 times.

- A six appeared 188 times, which is rather high. One would expect that the number of a six outcome to be closer to 1,000/6 = 167.

P-hacking

- It seems that the hypothesis about green dice may have some truth in it.

- To prove it the scientists ran an experiment and computed the p-value.

- You are asked to evaluate their experiment and explain if you agree with their analyses and why.

https://github.com/tocunha/reflectionsdatascience/blob/master/exercise4-02-03-2020/Green-dice-p-value-hacking.ipynb

Multi comparison bias

- Effectively it's just the notion that the more tests you run, the more likely you are to get false positives
- Things that look like they confirm your hypothesis, but are really just random chance.
- If you don't correct for this at some point you're very likely to accept hypotheses that aren't based on any real relationships.

Bonferroni’s principle

- If you look in your data for too many things at the same time, you will see things that look interesting.

Bonferroni's principle

- You can expect events to occur, even if the data is completely random.

- The number of occurrences of these events will grow as the size of the data grows.

Bonferroni's principle

- Calculate the expected number of occurrences of the events you are looking for, on the assumption that data is random.

- If this number is significantly larger than the number of real instances you hope to find, then you must expect almost anything you find to be bogus.

Bonferroni's principle

The original p value

$$\textit{Bonferroni-corrected p value} = \frac{\alpha}{n}$$

The number of tests performed

Multi comparison bias

https://github.com/tocunha/reflectionsdatascience/blob/master/exercise4-02-03-2020/multiple_comparisons.ipynb