How I learned to stop visualizing and love statistics

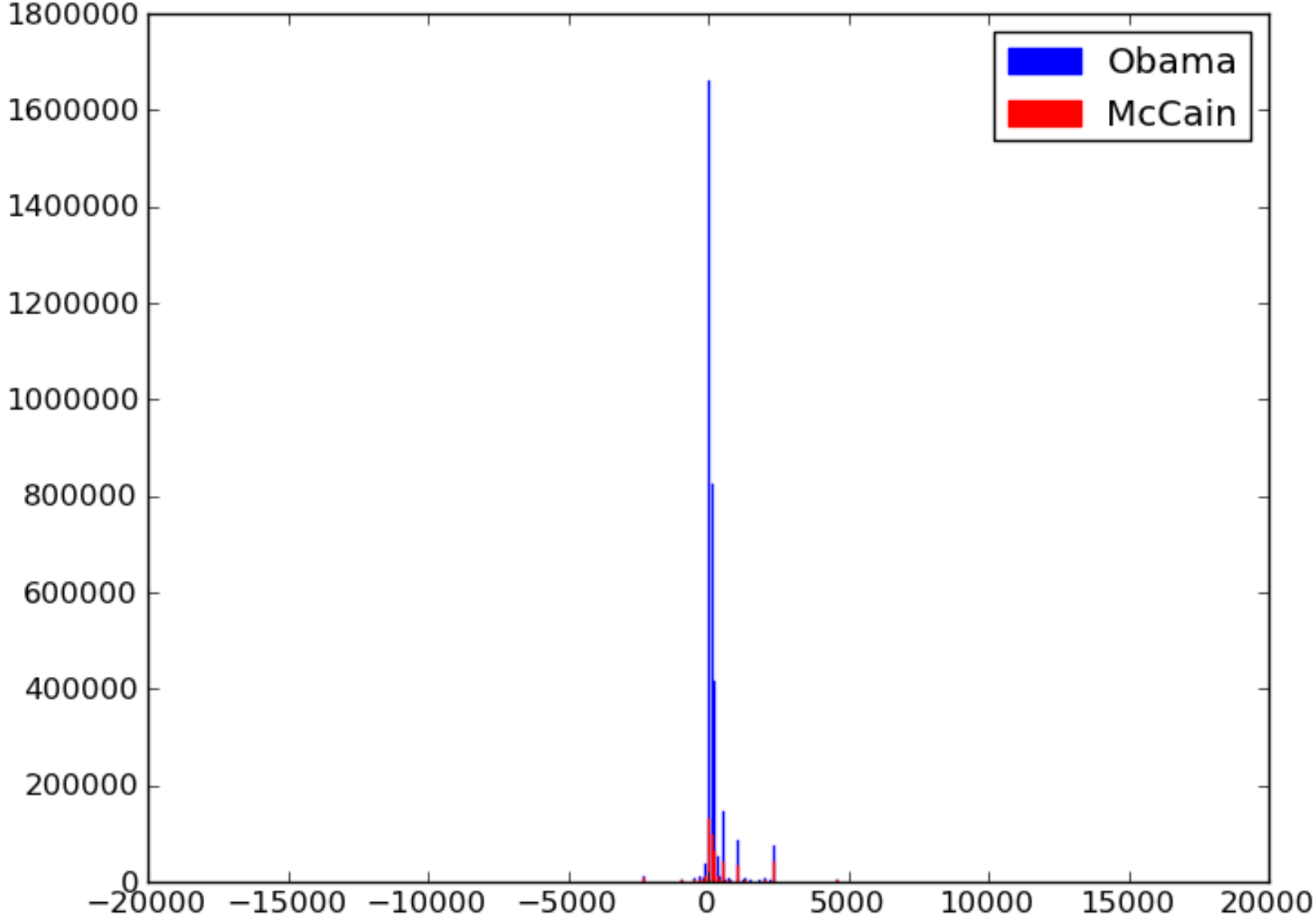
You have a hunch

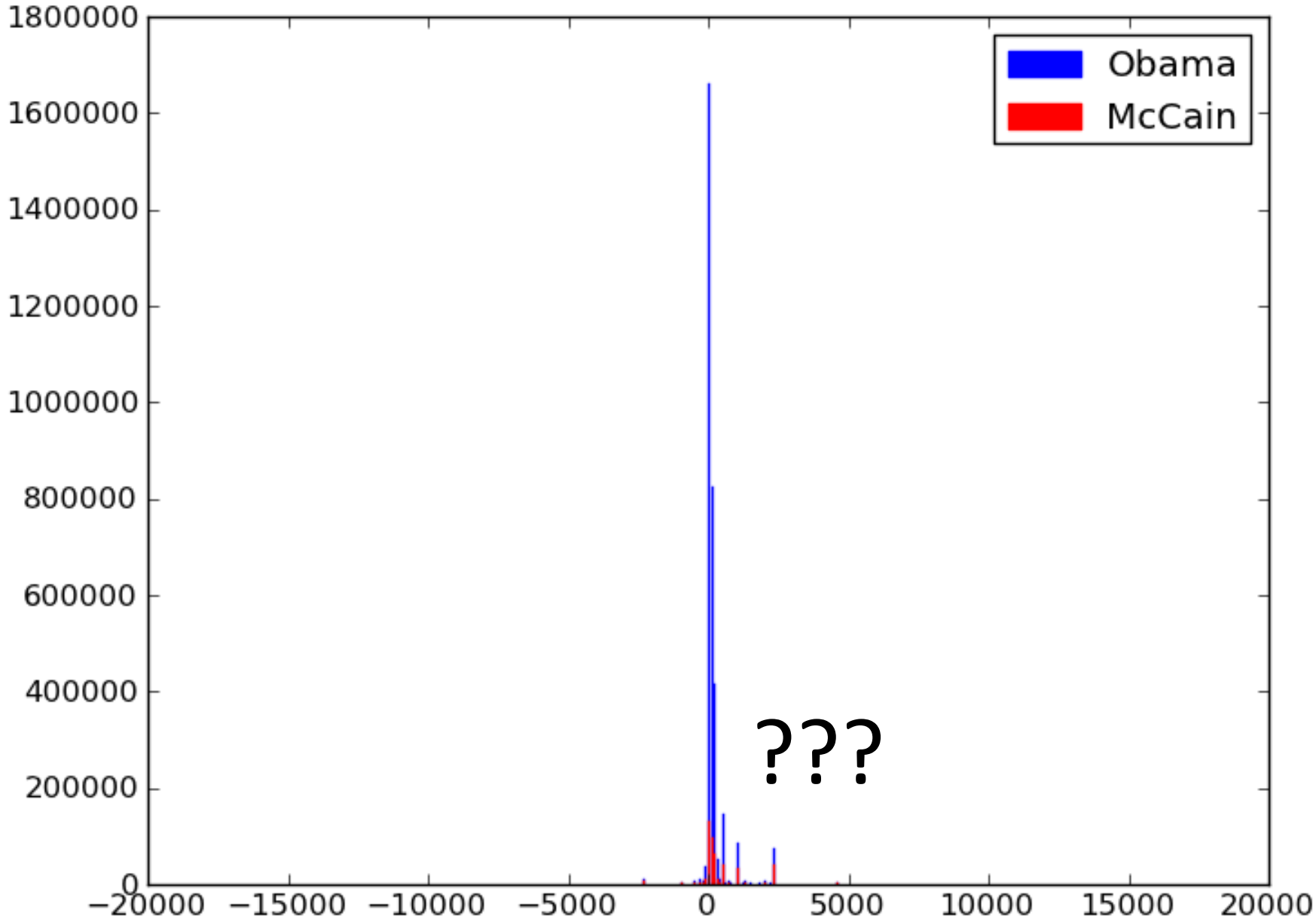
Visualizations → sanity check

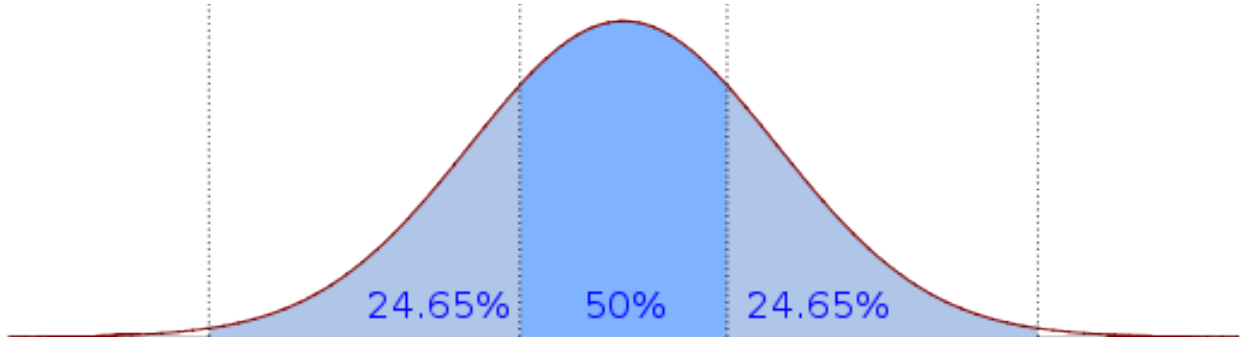
Statistics → quantify the hunch

(Visualizations → storytelling)

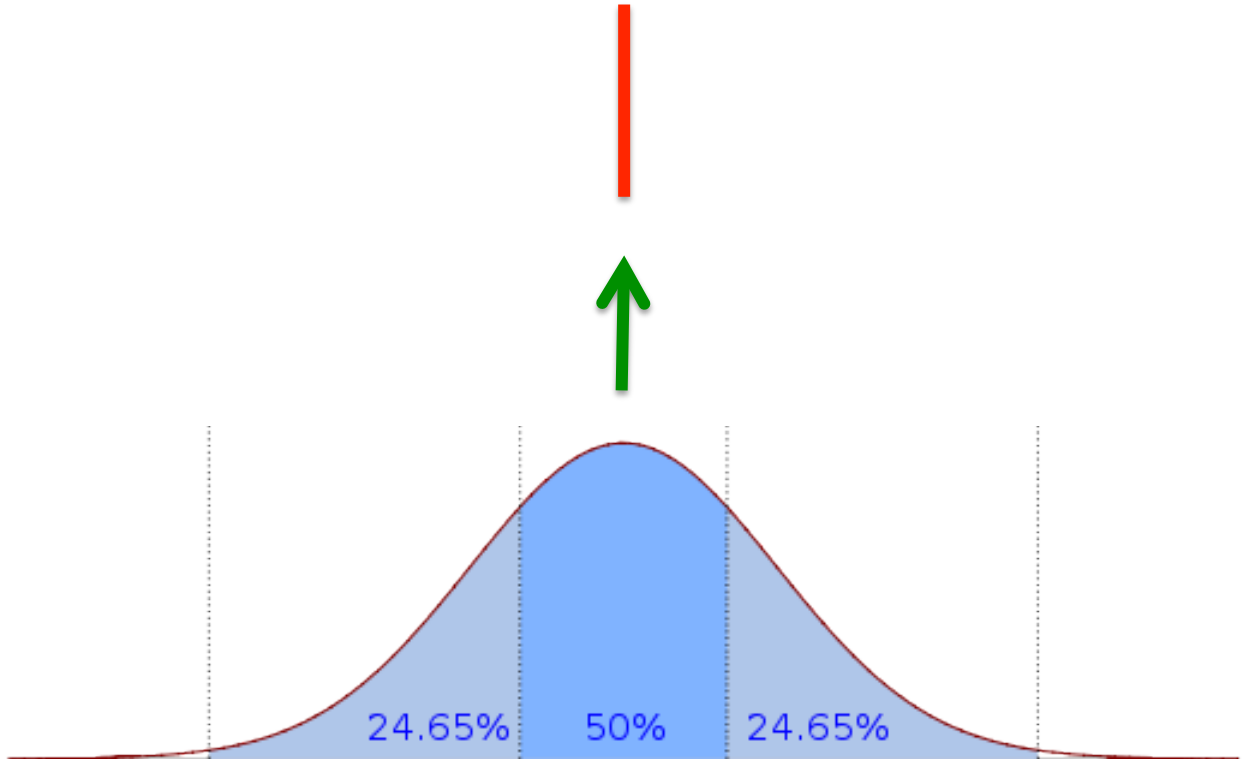
Someone says:
“Obama got more small campaign contributions than McCain”


Obama
McCain
1800000
1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
−20000
−15000
−10000
−5000
0
5000
10000
15000
20000

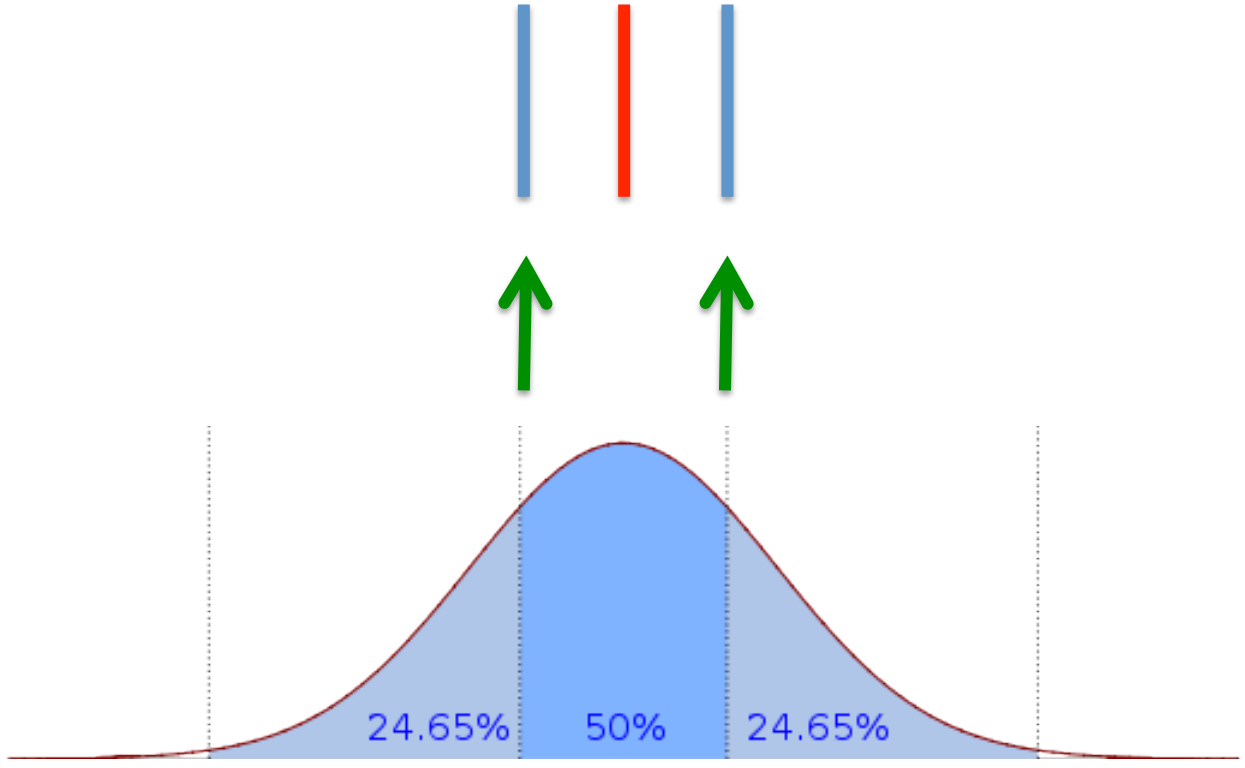

1800000
1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
−20000
−15000
−10000
−5000
0
5000
10000
15000
20000
Obama
McCain
???


24.65%
50%
24.65%

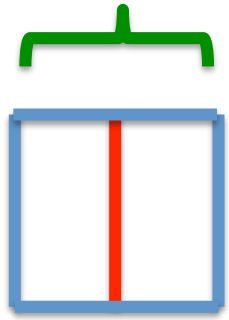
Median

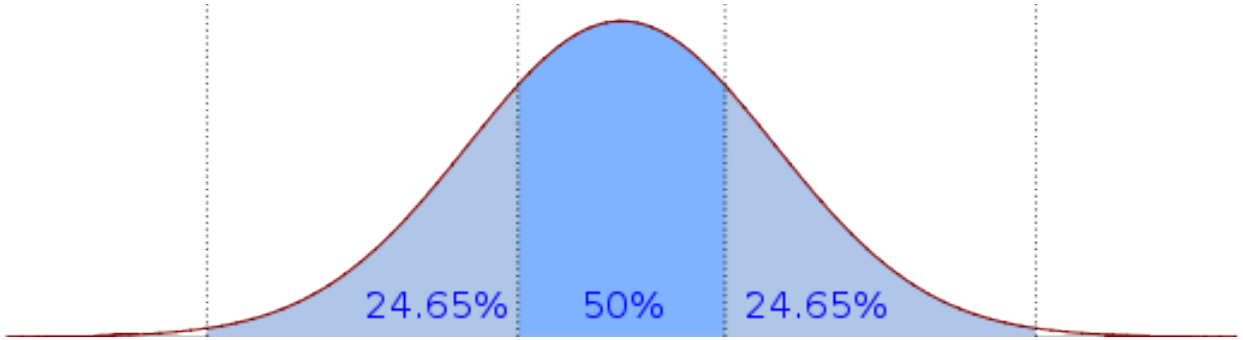


25% 75%

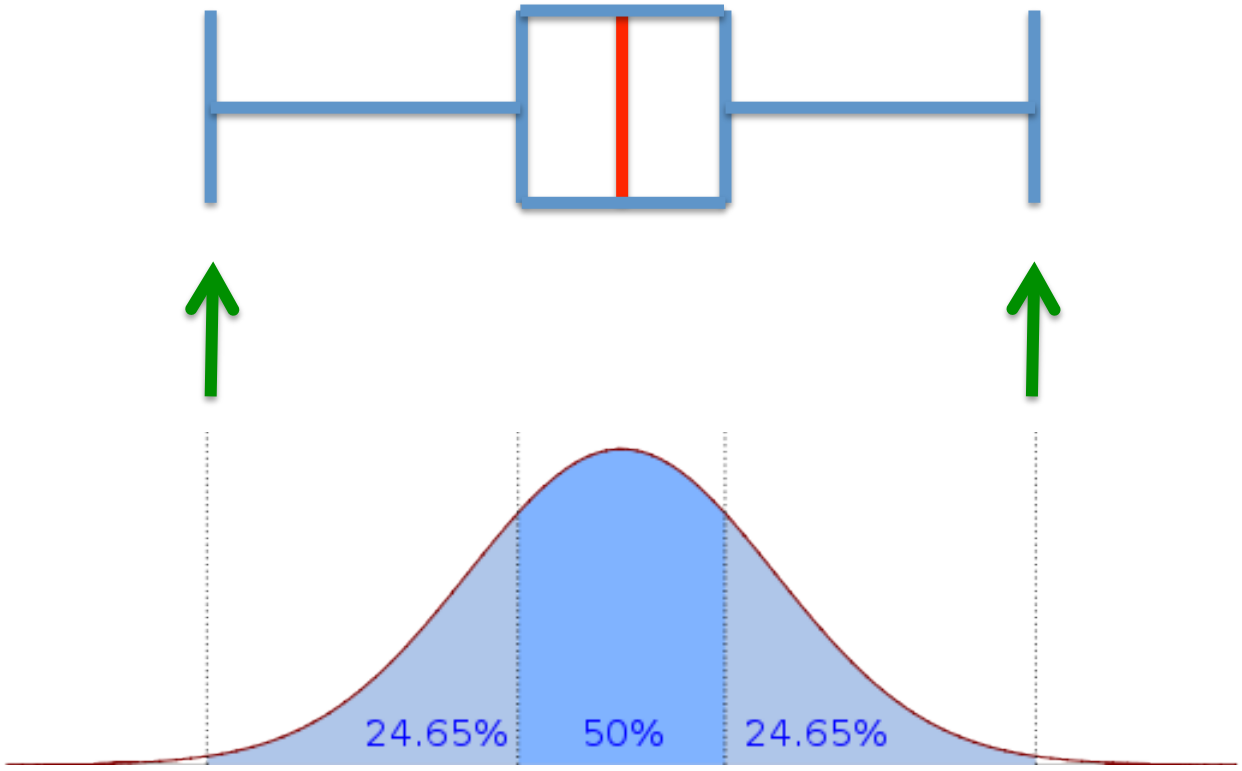


Inner Quartile Range

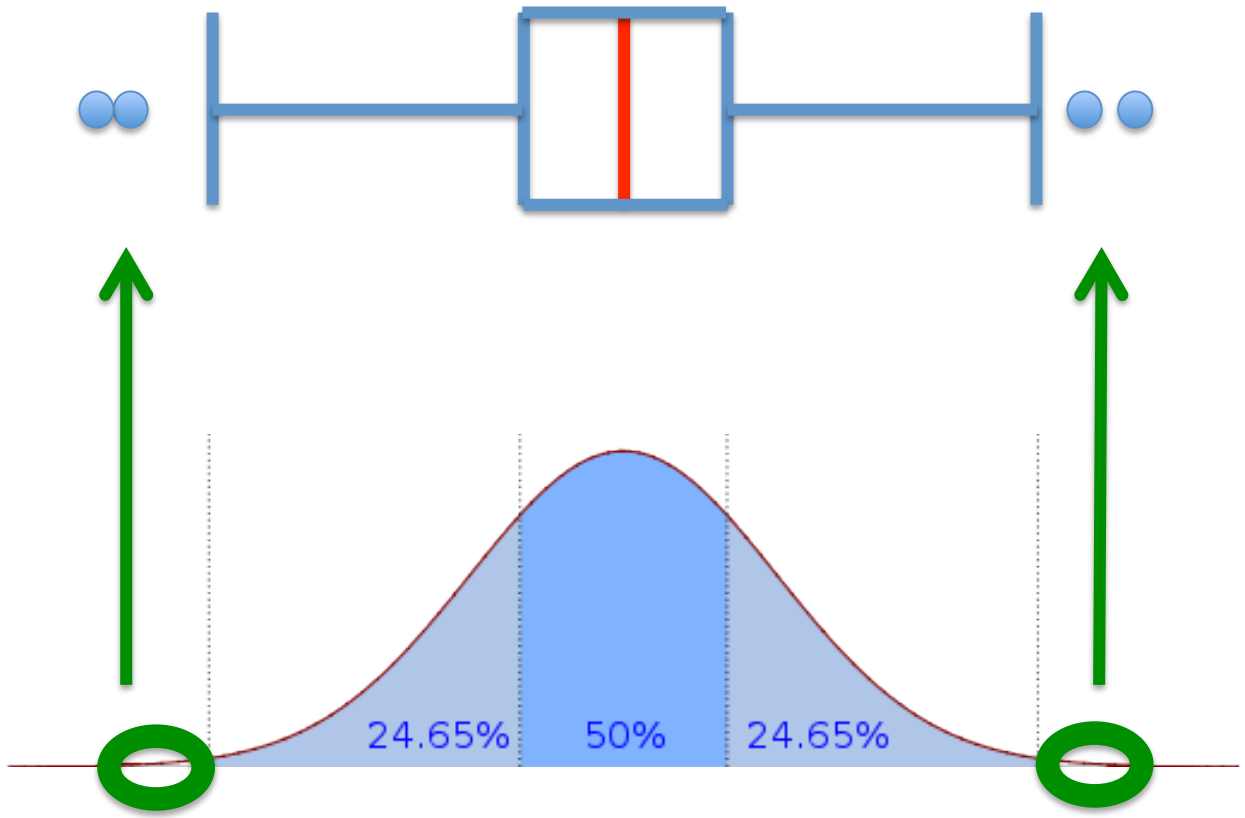




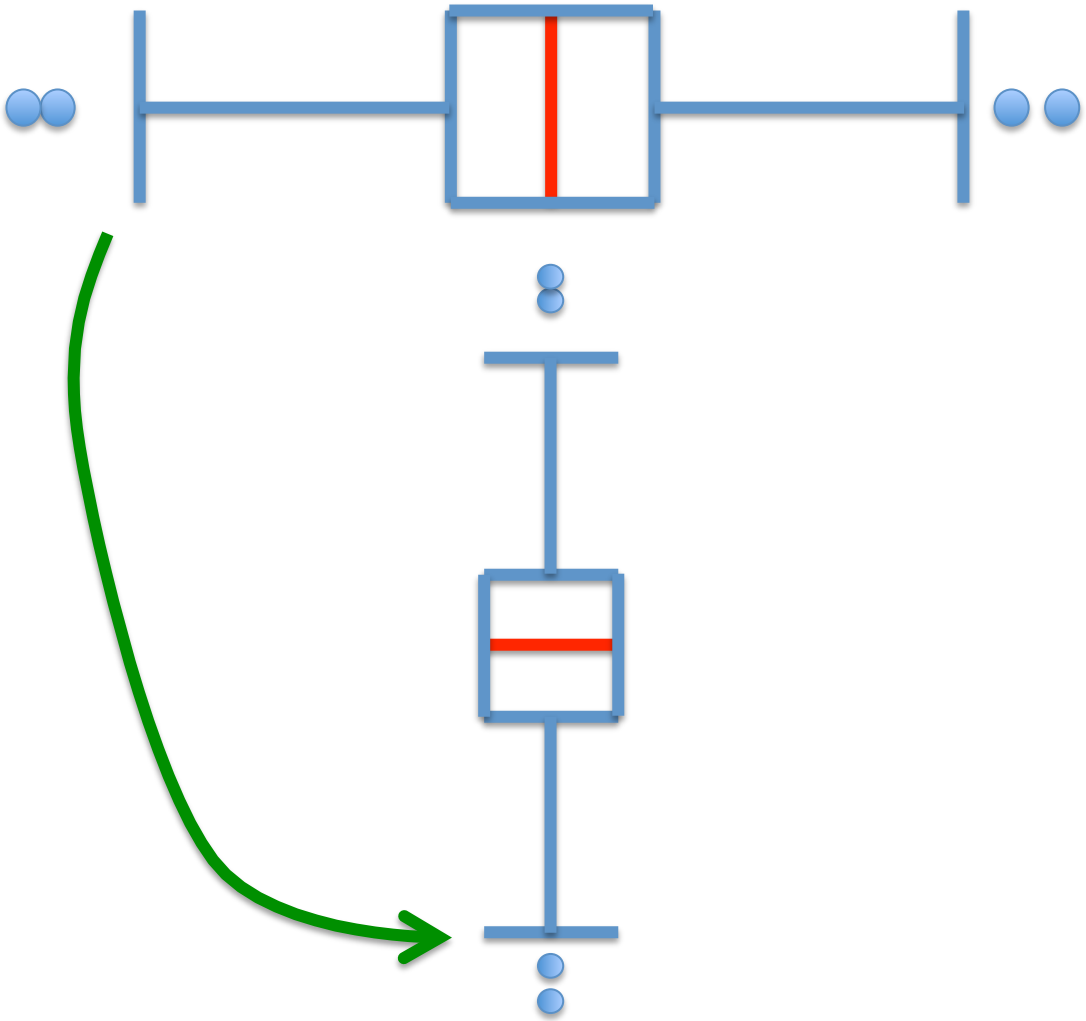
Whiskers / Extremes

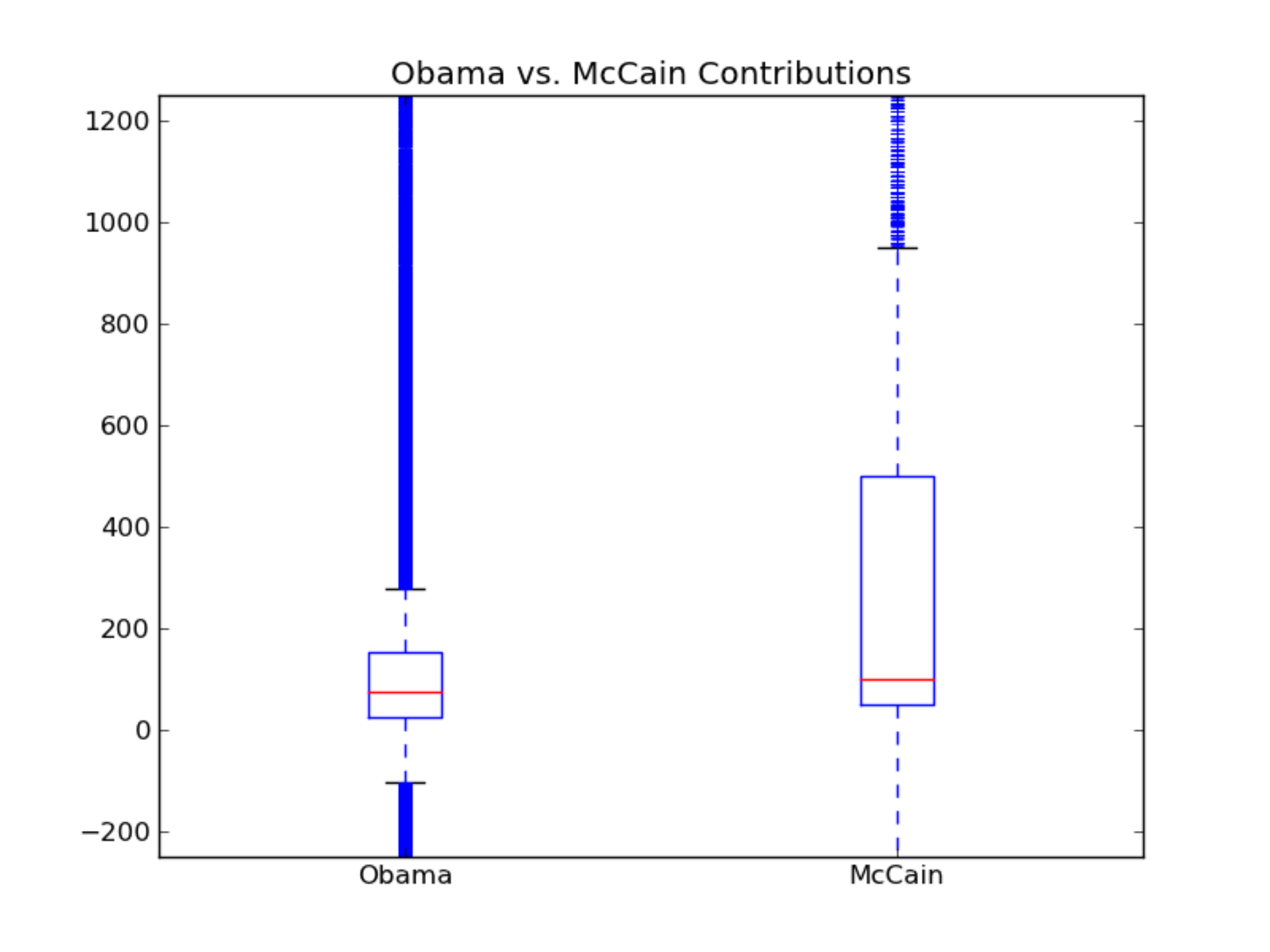


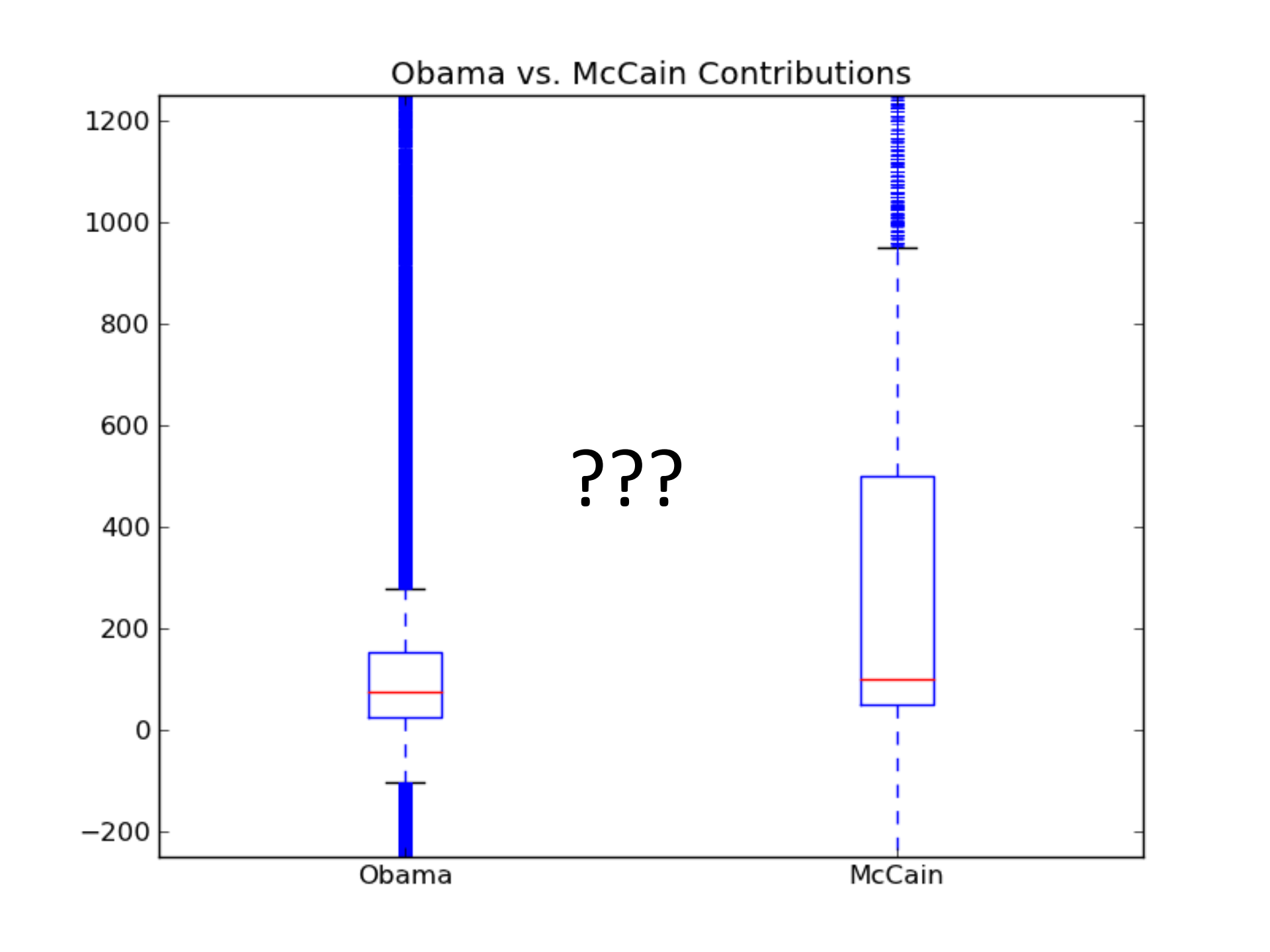
Outliers



Box-and-Whiskers Plot



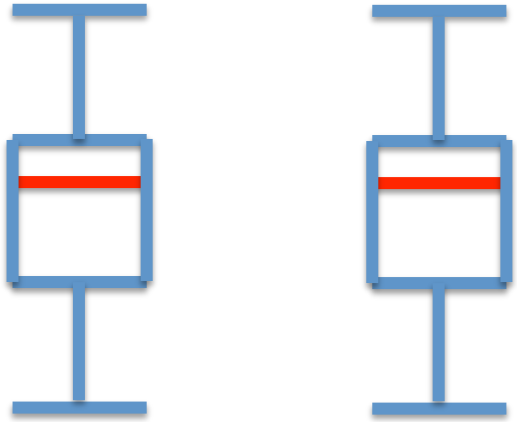

Obama vs. McCain Contributions
1200
1000
800
600
400
200
0
−200
Obama
McCain


Obama vs. McCain Contributions
1200
1000
800
600
400
200
0
−200
???
Obama
McCain

Are they actually different?

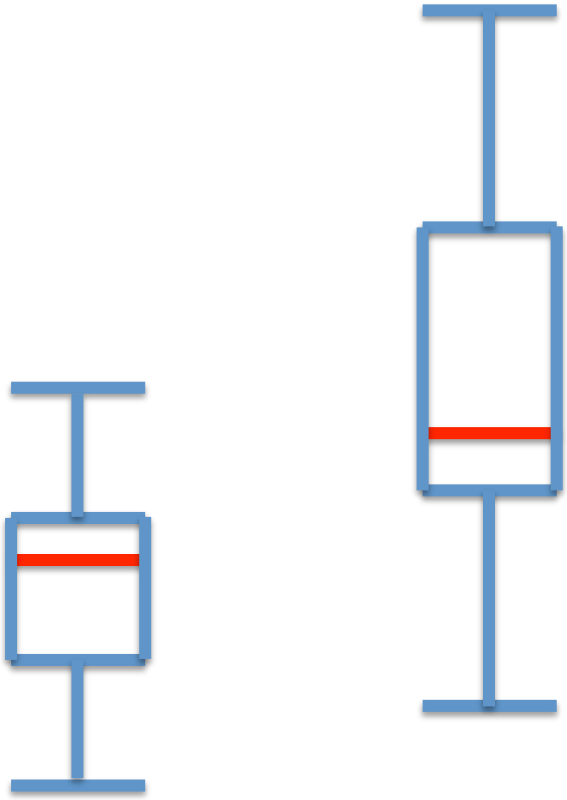
T-Test

Assume

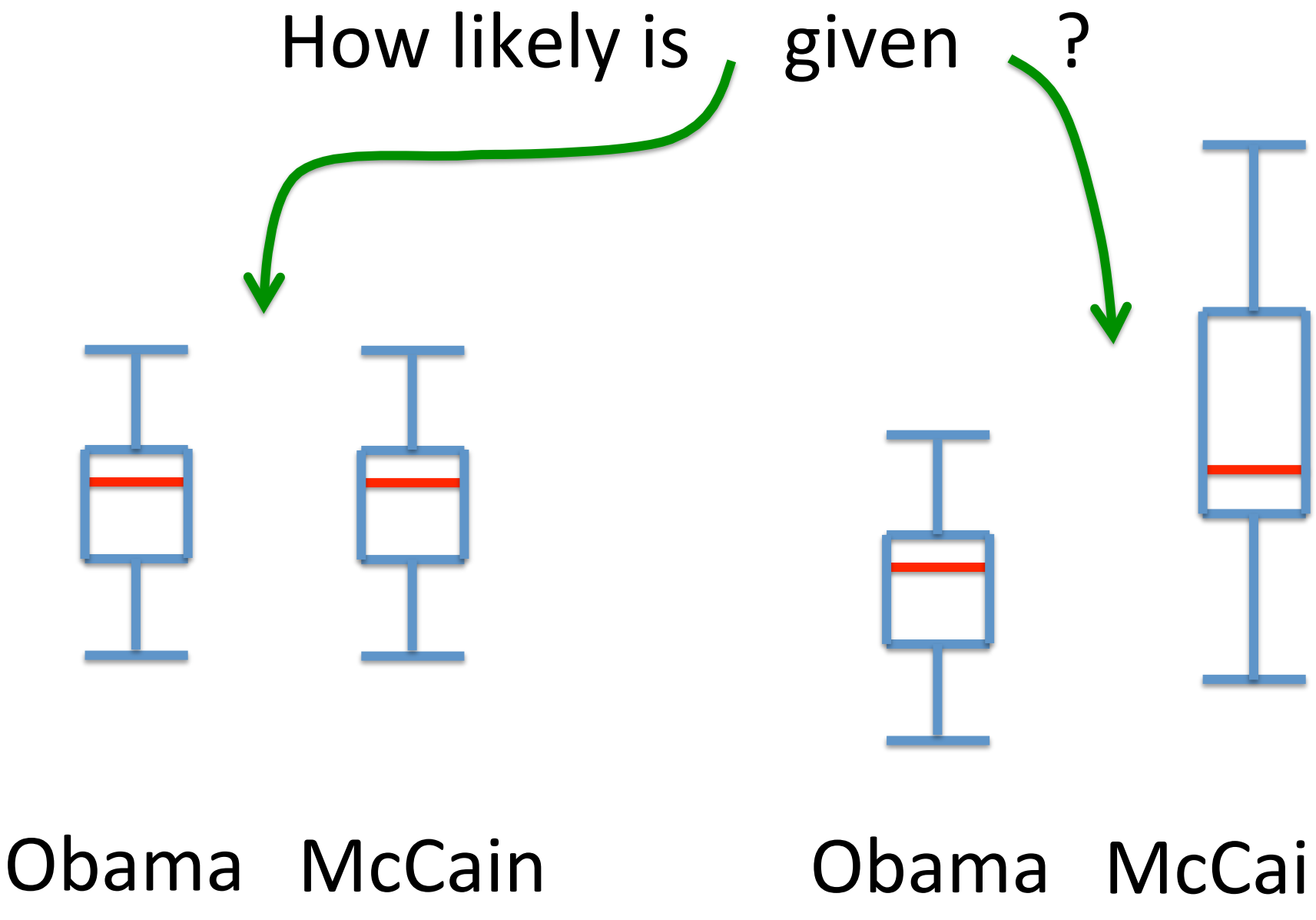


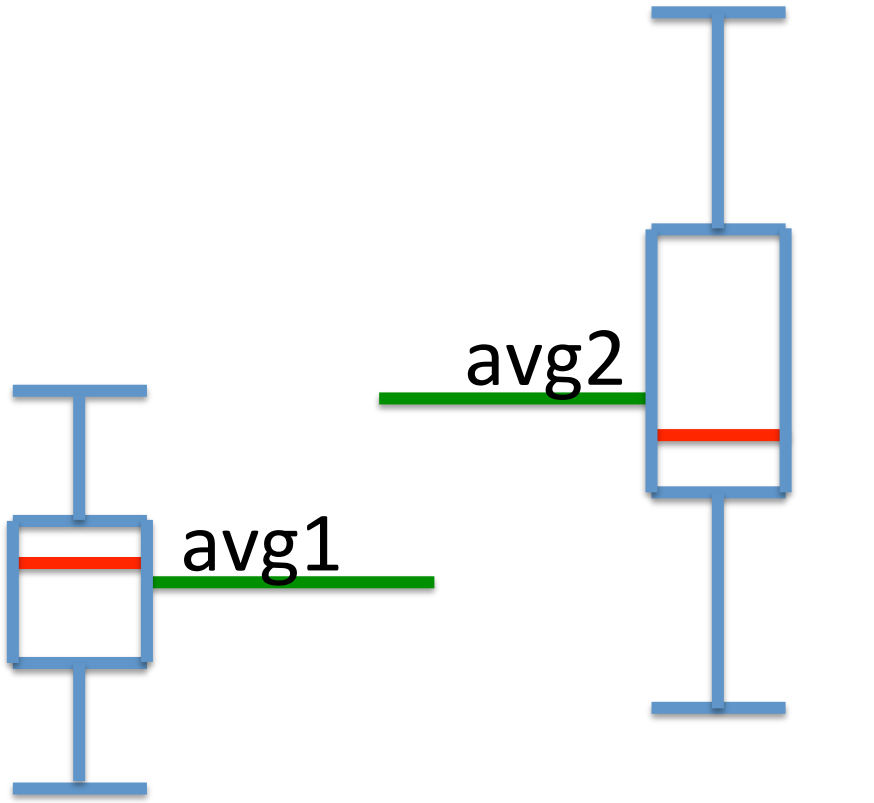
Obama McCain

Reality

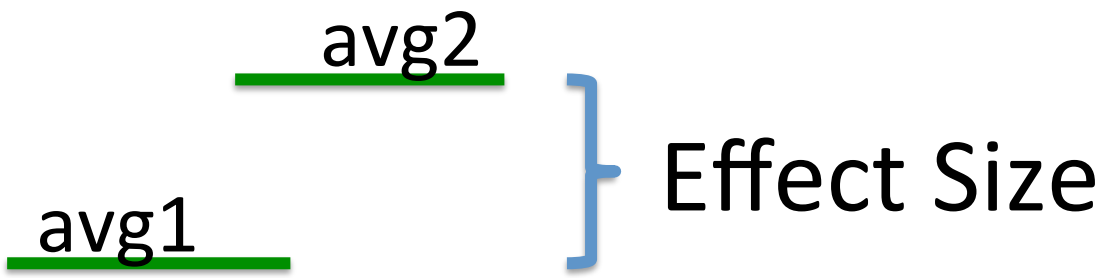


Obama McCain

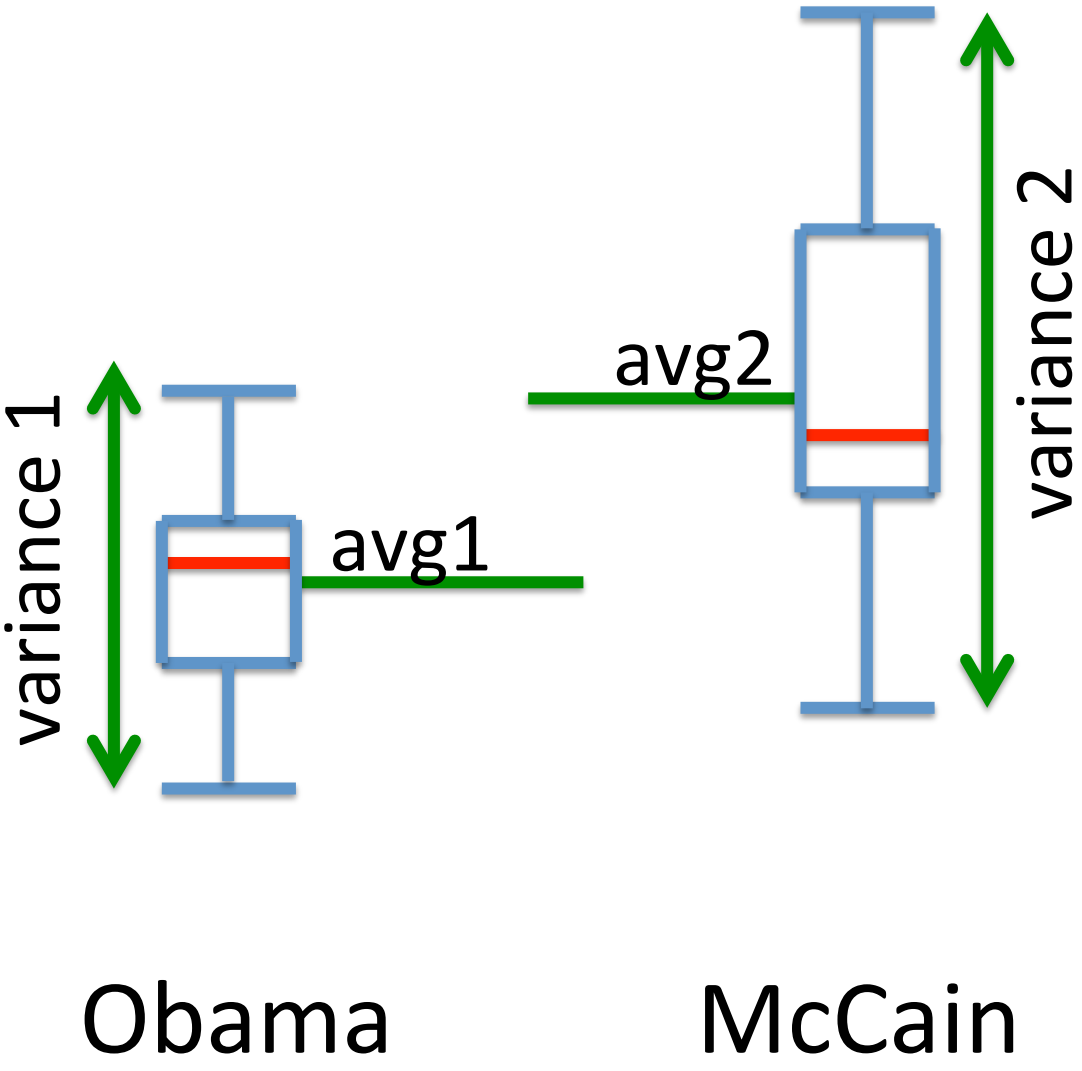

How likely is
given
?
Obama
McCain
Obama
McCain



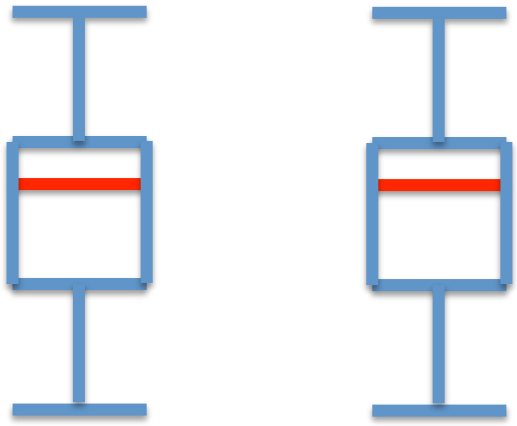
Obama McCain

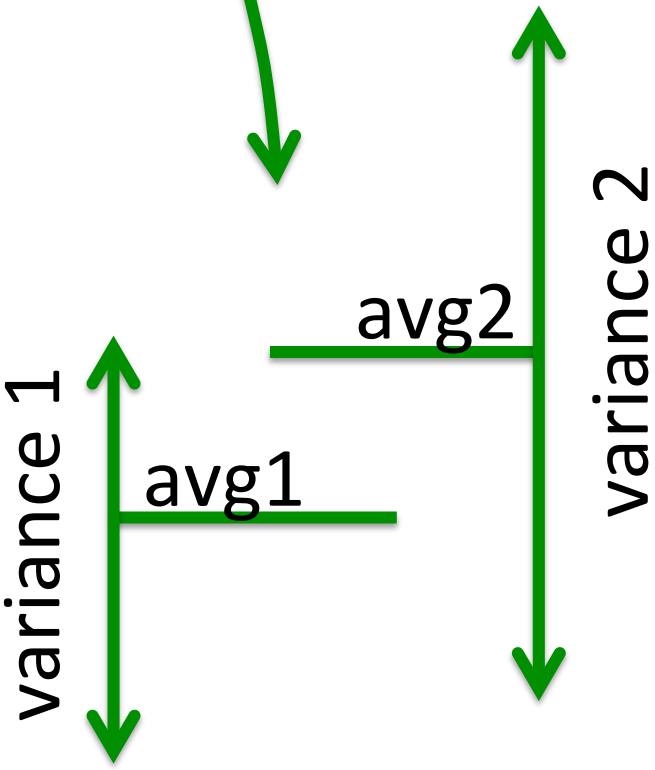


Obama McCain


variance 1
avg1
avg2
variance 2
Obama
McCain

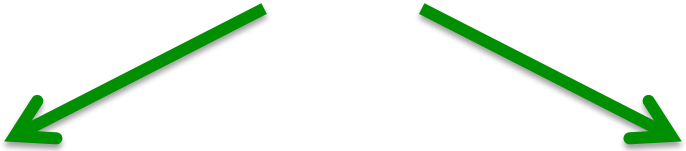
How likely is given ?





How likely are they equal
given avg/variance differences?

Probablility p



p is low

Obama, McCain are different

(significant)

p is high

Don't trust the difference

(not significant)

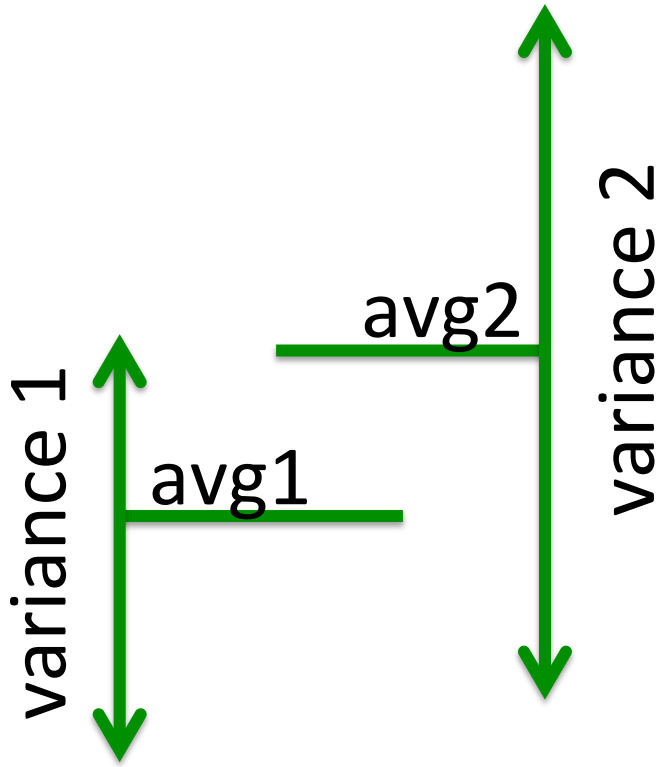
Significance is binary

- Pick a threshold: .01? .05?
- Is p > threshold, or ≤ threshold?

p ≤ .05? significant

p > .05? don't trust the difference

T-Test Signifiance



\+ # Samples
Obama: >1M
McCain: >1M

Correlation, Linear Regression

County Health Rankings

- Every county in USA
- Years of Potential Life Lost (YPLL): early morbidity
  - less is good
  - more is bad
- Median income, % population w/ diabetes,
  % population under 18, …

What is correlated with early death in a community?

Burgers

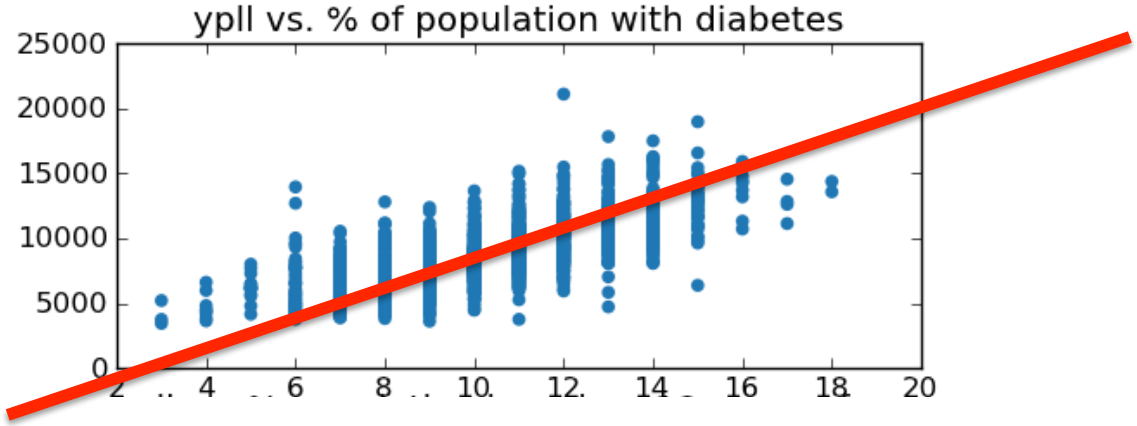
Sleep

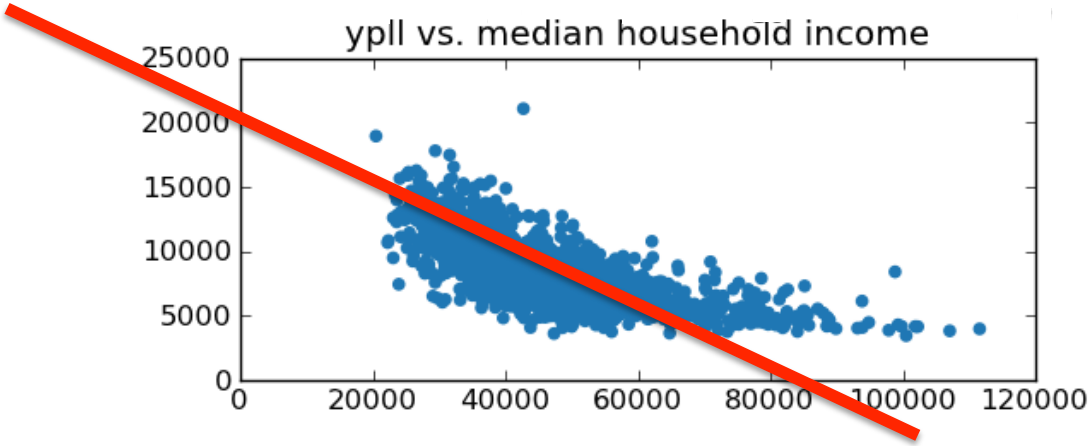
Education

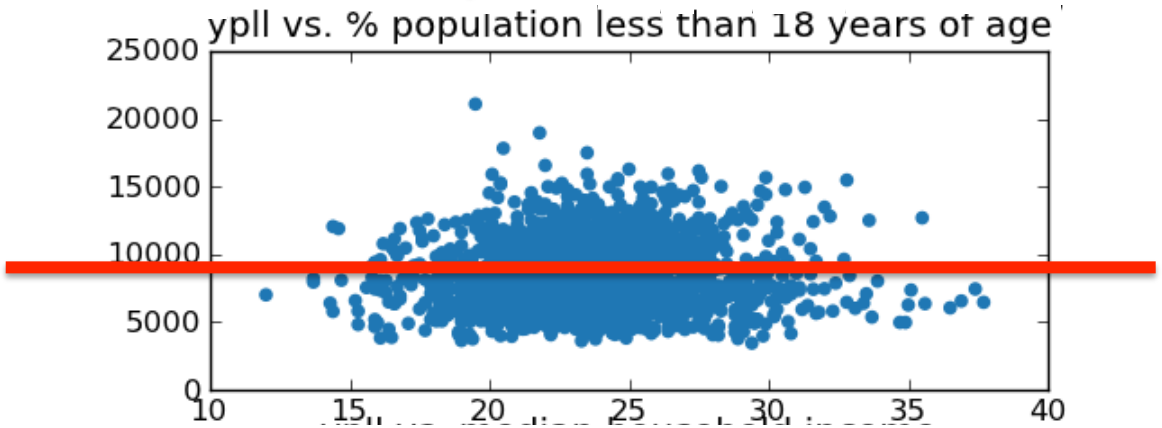
Exercise

Rappers

Your theory here


ypll vs. % of population with diabetes
25000
20000
15000
10000
5000
0
2
4
6
8
10
12
14
16
18
20

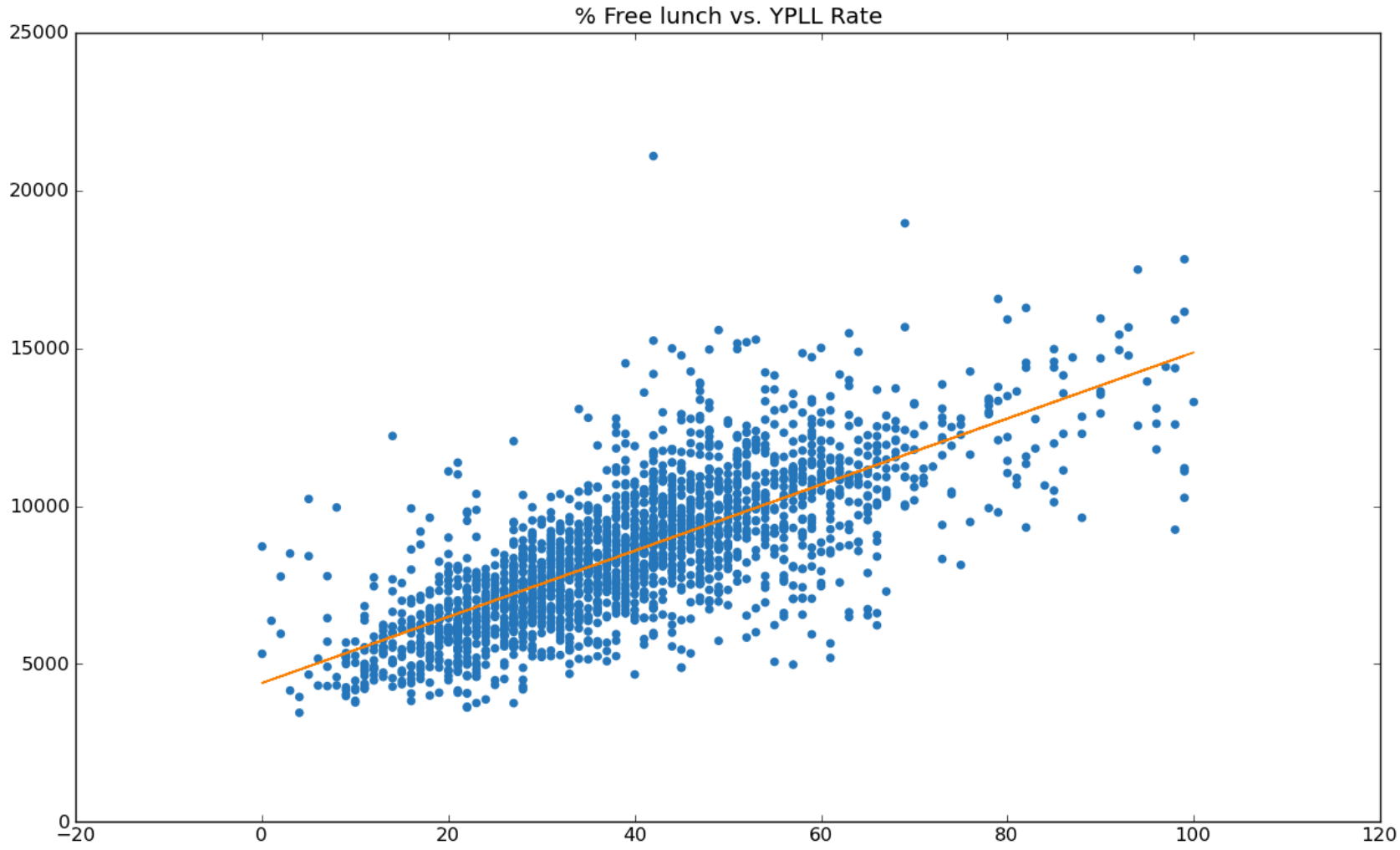

ypll vs. median household income
25000
20000
15000
10000
5000
0
0
20000
40000
60000
80000
100000
120000


ypll vs. % population less than 18 years of age
25000
20000
15000
10000
5000
0
10
15
20
25
30
35
40

Line coefficients: $y = mx + b$

Correlation amount: R^2 (0 to 1)

Significance: $p < .05$?


% Free lunch vs. YPLL Rate
25000
20000
15000
10000
5000
0
−20
0
20
40
60
80
100
120

Decrease amount of free lunch

→

Reduce early morbidity!


WARNING
ZOMBIES
AHEAD

Correlation != Causation

Correlation

Causal Hunch

Randomized Trial

T-Test!

Remember to **git pull**

http://dataiap.github.com/dataiap/day3/