

EXPERIMENTAL DESIGN

Based on Coursera's "The Data Scientist's Toolbox"

FORMULATE YOUR QUESTION

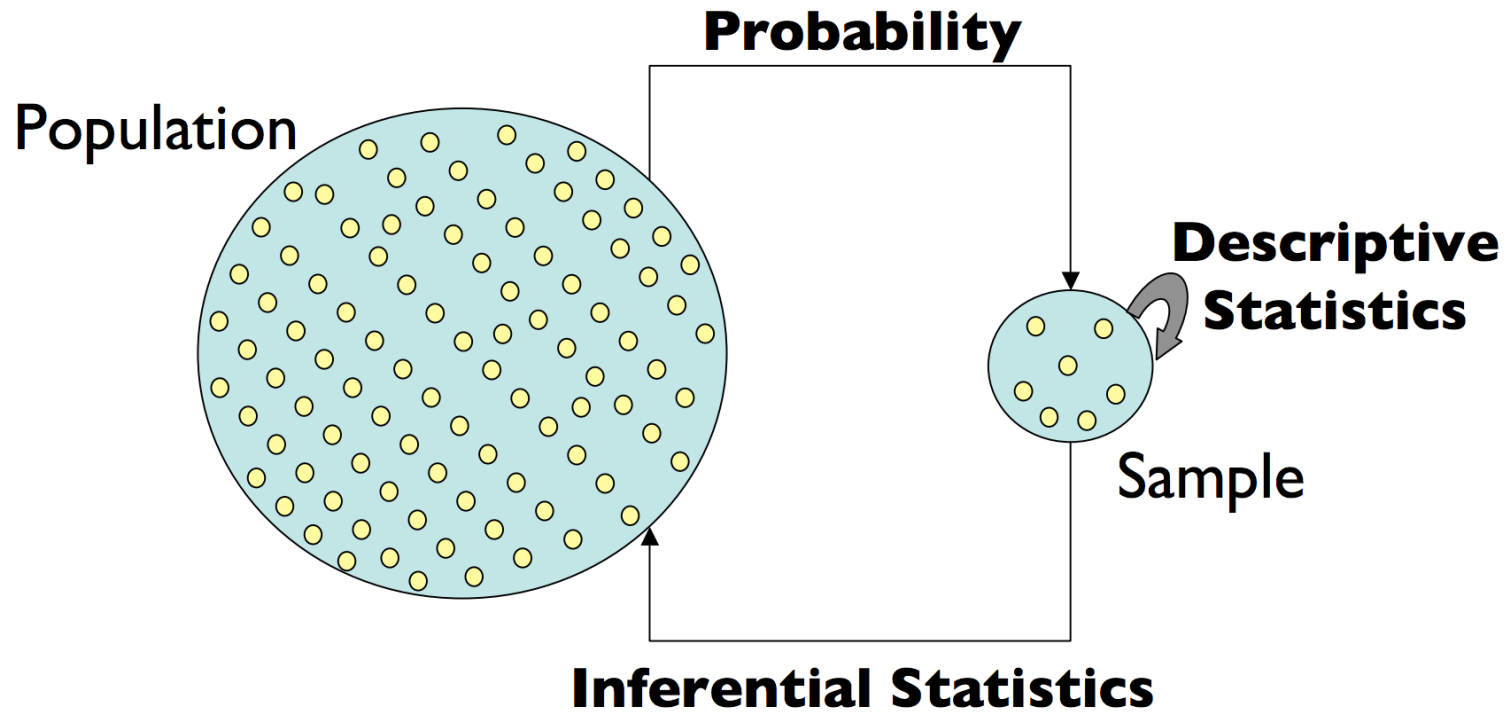
Question: Does changing the text on your website improve donations?

Experiment:

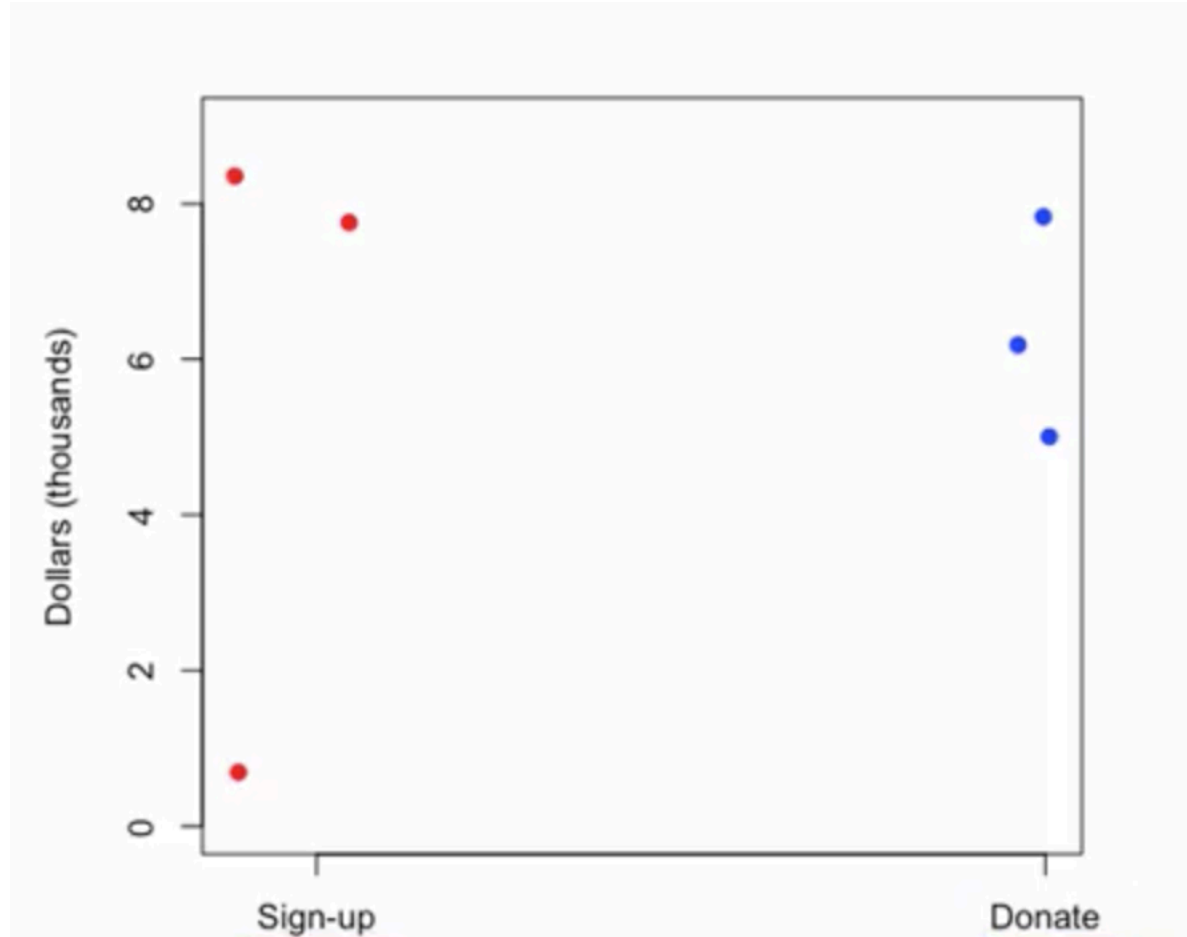
- ❖ Randomly show visitors different versions of the website (“learn more” vs “sign up”)
- ❖ Measure how much they donate
- ❖ Determine which performs better

A/B Testing: http://www.wired.com/2012/04/ff_abtesting

STATISTICAL INFERENCE



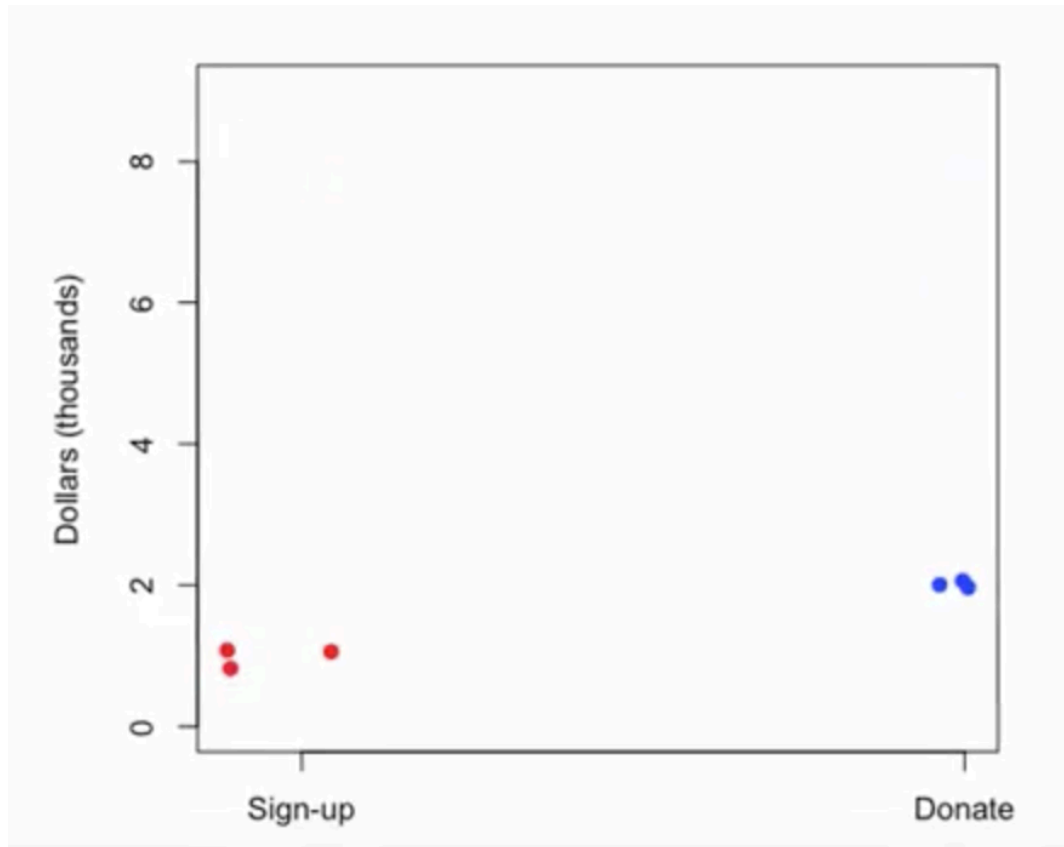
VARIABILITY



Large Variability:

- ❖ Large amount of variability
- ❖ Hard to draw conclusions

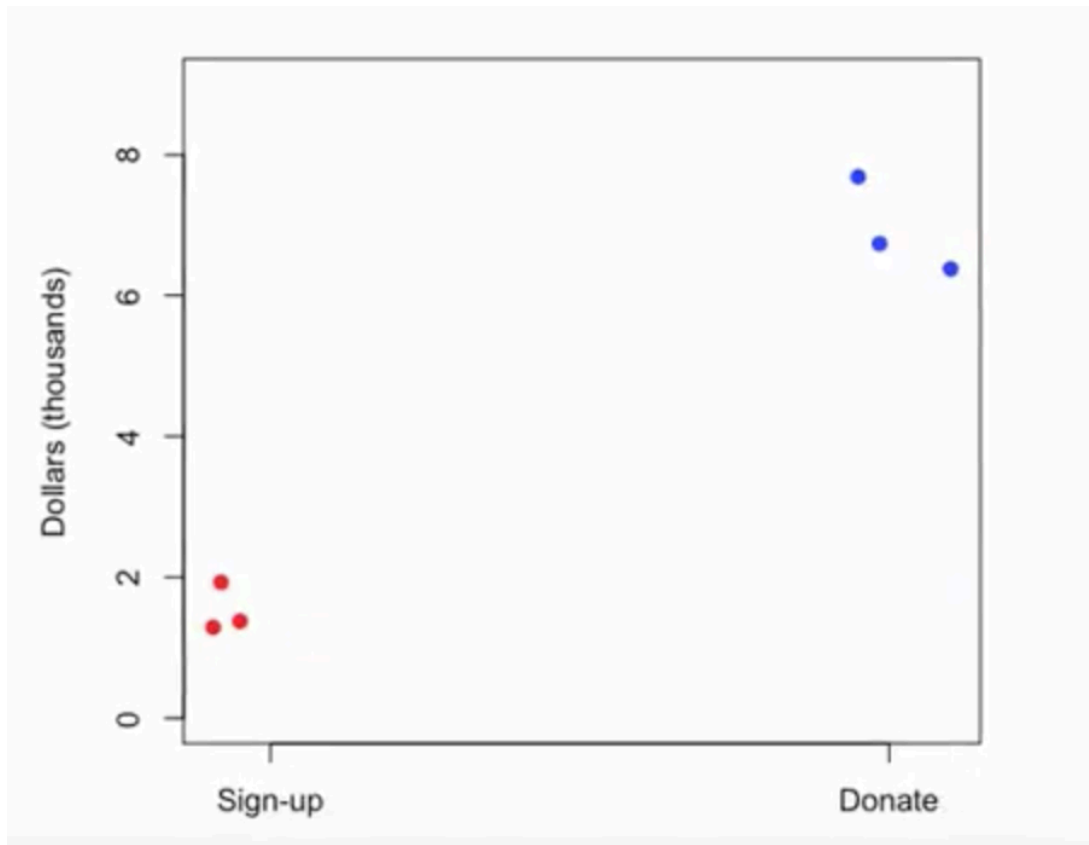
VARIABILITY



Small Variability:

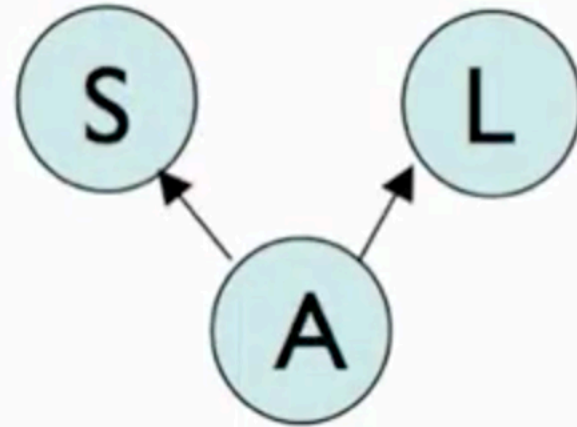
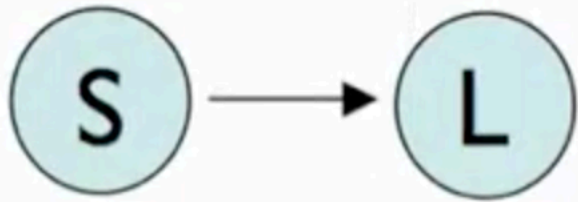
- ❖ Consistent data points
- ❖ May not have a lot to gain by switching to “donate” version but it is better

VARIABILITY



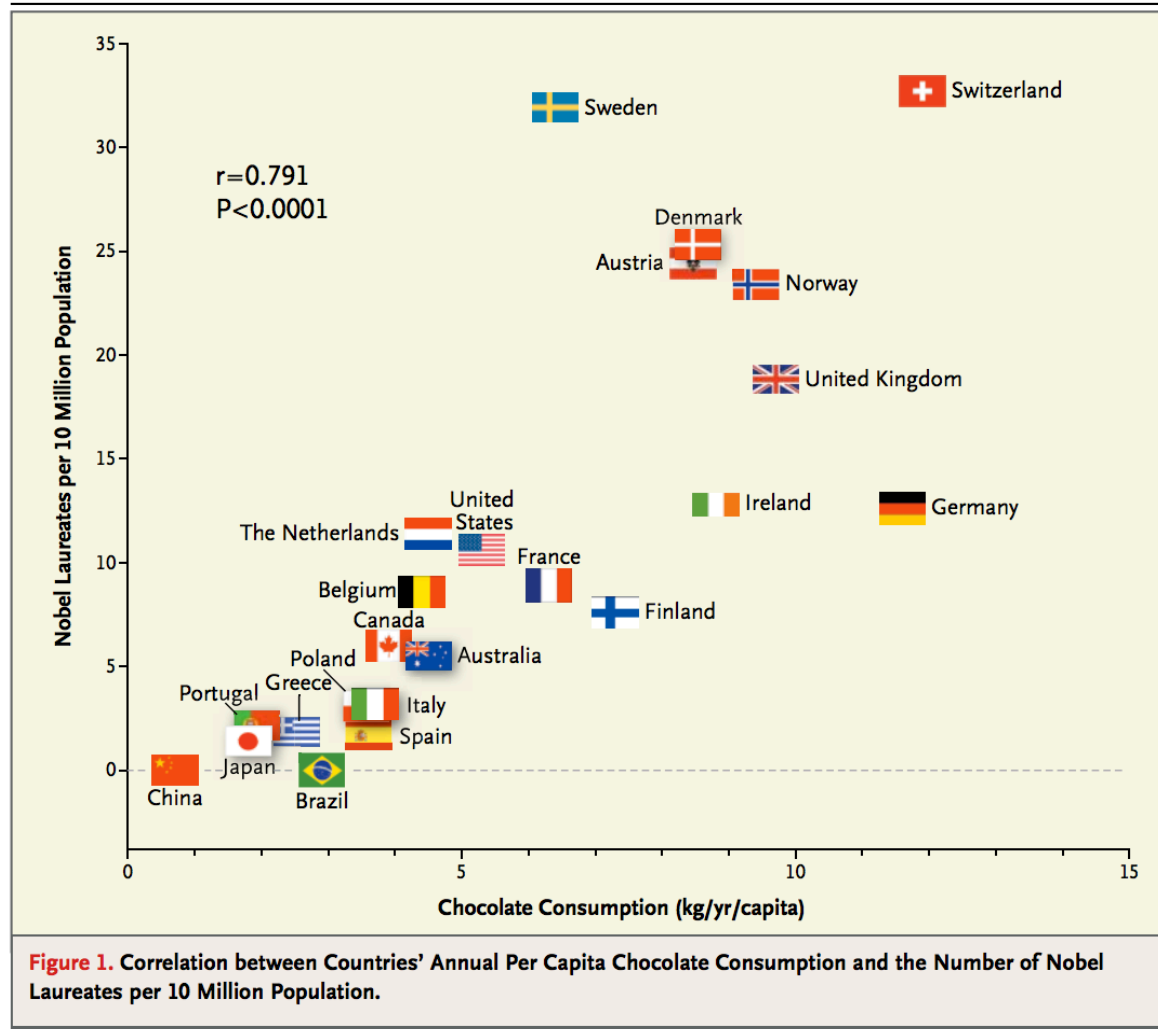
Small Variability
within cases,
large difference
between cases –
clear “winner”

CONFOUNDING FACTORS



Shoe size, length and age

CORRELATION IS NOT CAUSATION!



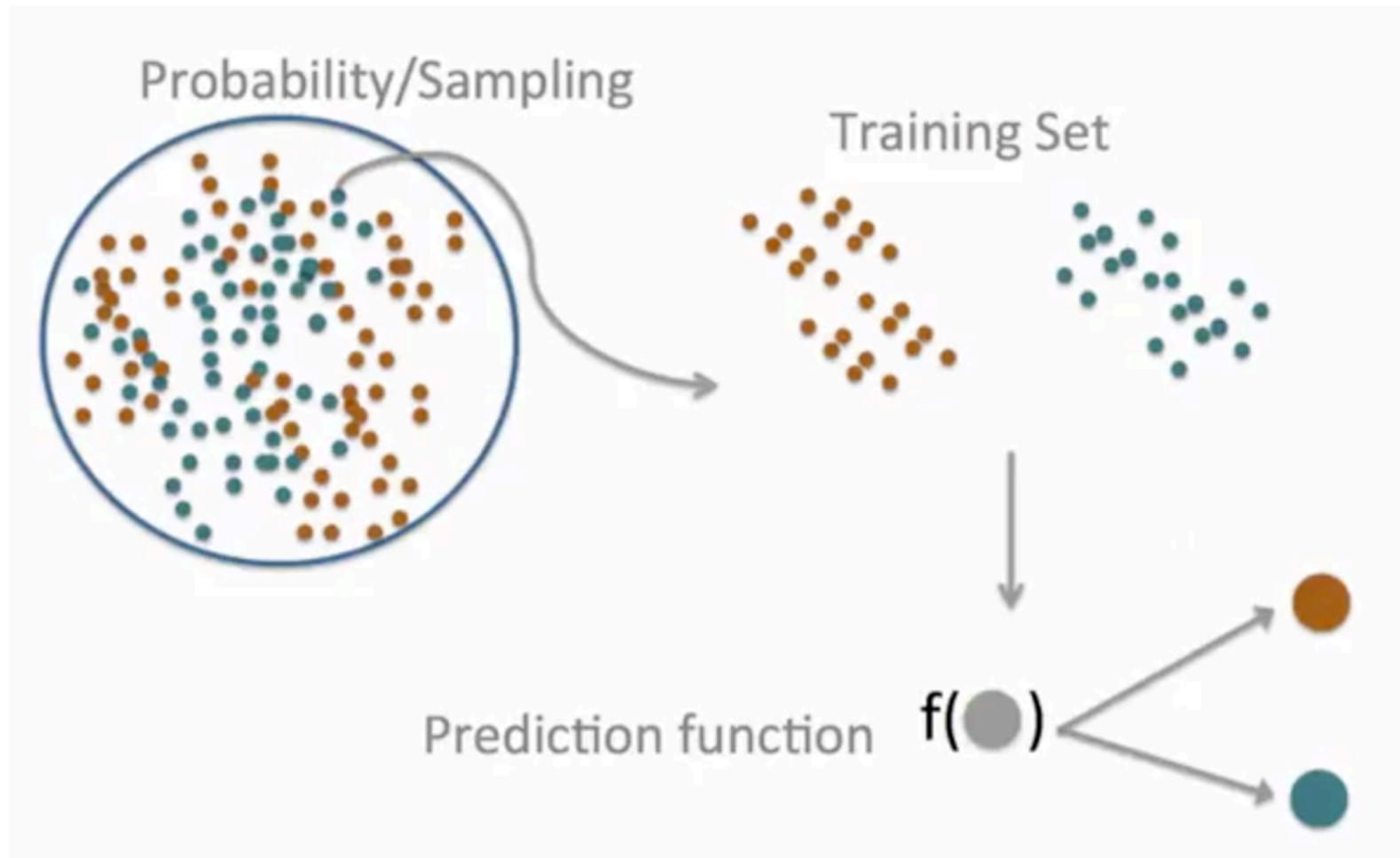
[http://www.biostat.jhsph.edu/courses/bio621/misc/Chocolate%20consumption%20cognitive%20function%20and%20nobel%20laureates%20\(NEJM\).pdf](http://www.biostat.jhsph.edu/courses/bio621/misc/Chocolate%20consumption%20cognitive%20function%20and%20nobel%20laureates%20(NEJM).pdf)

RANDOMIZATION AND BLOCKING

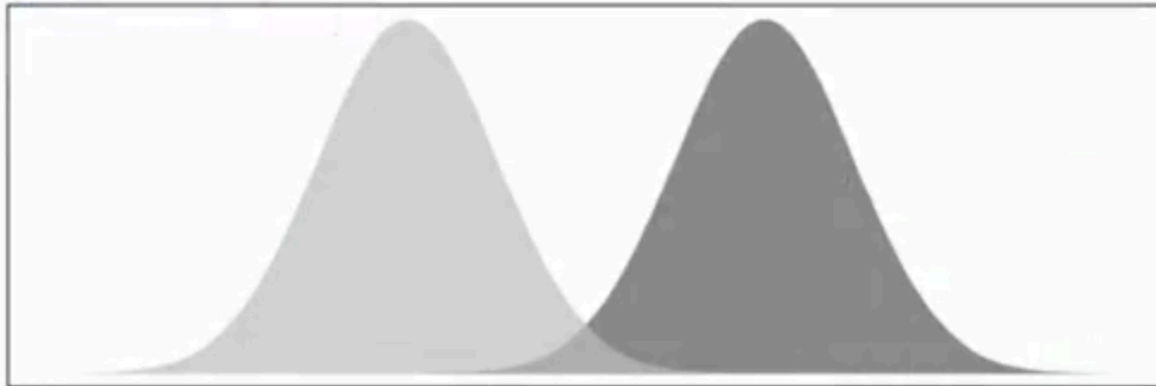
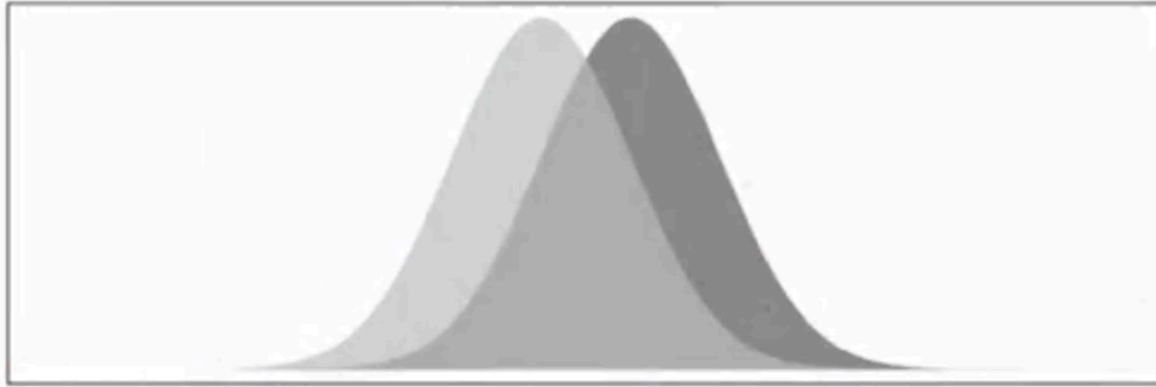
In order to try to avoid spurious correlations:

- ❖ Fix variables (always have the text say the same thing)
- ❖ Stratify variables (if you have two colors and two phrases, use options equally in combination)
- ❖ Randomize variables

PREDICTION



PREDICTION VS INFERENCE



Different means, so we can infer things about the two populations. However, if you choose a point, you'll have a hard time predicting whether that point is part of dark or light

DATA DREDGING

<http://xkcd.com/882/>

