GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
EXPERIMENTAL DESIGN

Experiments are designed around three central concepts:

Experiments are designed around three central concepts:

1. Causation: we want to be able to determine the causation or relationship between some dependent variables X and an independent variable y.

Experiments are designed around three central concepts:

1. Causation: we want to be able to determine the causation or relationship between some independent variables X and an dependent variable y.

2. Control: in order to determine above, we need some method of knowing the result when the independent variables are nil

Experiments are designed around three central concepts:

1. Causation: we want to be able to determine the causation or relationship between some independent variables X and an dependent variable y.

2. Control: in order to determine above, we need some method of knowing the result when the independent variables are nil

3. Variability: controlling variability in ease to detect reason behind change

Three most commonly used experimental design patterns

Three most commonly used experimental design patterns

Completely Random
+ effectively randomizes your sample completely, strives to eliminate bias
— how do you test randomness? How do you determine bias?

$$n\ samples = 1000$$

Three most commonly used experimental design patterns

Randomized Block
+ Similar to using a control, you now have a random sample blocked around a particular variable (or many)
+ limits variability!
— no longer truly random

n samples = 500 girls, 500 boys.
1000 all together

Three most commonly used experimental design patterns

Matched pairs
+ 1:1 relationship for each testing group = all variance eliminated
— literally impossible in the real world

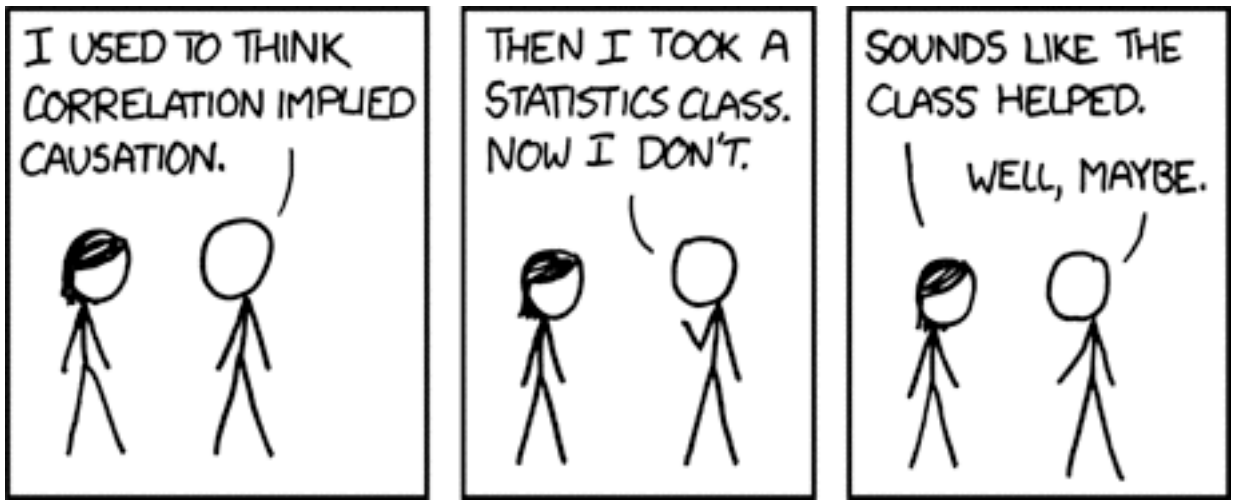
n samples = all the same on each end
1000 samples

How does this fit in with AB testing?

How does this fit in with AB testing?

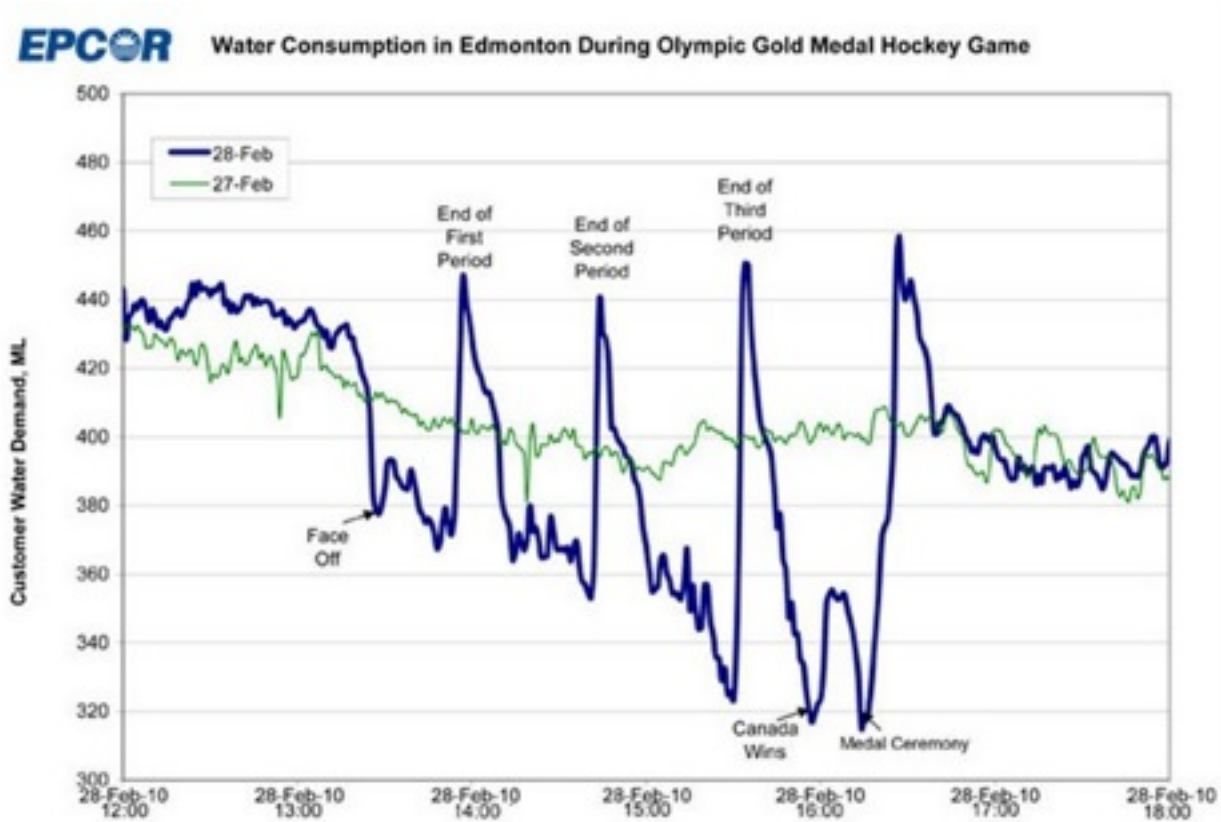
Sample: Testing how price effects conversion rate of a photography-related product on a web page.

As we start controlling for more features, how do you know each independent feature is causing a dependent feature?



http://xkcd.com/552/

DOES THIS CORRELATION SUGGESTION CAUSATION?



http://i.imgur.com/npieo.jpg

HOW ABOUT THIS ONE?

http://graphzoo.tumblr.com/post/84722935157/data-source-http-www-quandl-com

As we start controlling for more features, how do you know each independent feature is causing a dependent feature?

Short answer: We don’t

As we start controlling for more features, how do you know each independent feature is causing a dependent feature?

Short answer: We don't

Long answer: Start counting "correlation does not equal causation"s before going to bed

As we start controlling for more features, how do you know each independent feature is causing a dependent feature?

Short answer: We don't

Long answer: Start counting "correlation does not equal causation"s before going to bed

$42.94 answer:
http://bayes.cs.ucla.edu/BOOK-2K/ If you want to dive into the math.

As we start controlling for more features, how do you know each independent feature is causing a independent feature?

As we start controlling for more features, how do you know each independent feature is causing a independent feature?

We can test for this!

As we start controlling for more features, how do you know each independent feature is causing a independent feature?

We can test for this!

Contingency tables, chi squared (X^2) test (assumed independent)

Useful on categorical data

Useful on categorical data
requires to have at least 5 samples in each bucket, and that population is at least 10 times the sample size

Medicine Taken

Cold Length	yes	no	Total
1 -3 days	86	19	105
4 - 7 days	16	79	95
Total	102	98	200

null hypothesis (H0): The two variables are independent
alternative (H1): The two variables are dependent

null hypothesis (H0): The two variables (1, 2) are independent
alternative (H1): The two variables (1, 2) are dependent

		Medicine Taken		
Cold Length		yes	no	Total
	1 -3 days	86	19	105
	4 - 7 days	16	79	95
	Total	102	98	200

Degrees of freedom (DF)
amount of freedom within variables.
(var1 - 1) * (var2 - 1)

Medicine Taken

Cold Length	yes	no	Total
1 -3 days	86	19	105
4 - 7 days	16	79	95
Total	102	98	200

Is it significant?
chisquare: 81.9 for ?? DF

Medicine Taken

Cold Length	yes	no	Total
1 -3 days	86	19	105
4 - 7 days	16	79	95
Total	102	98	200

Is it significant?
chisquare: 81.9 for 1 DF
p-value: ~0

Medicine Taken

Cold Length	yes	no	Total
1 -3 days	86	19	105
4 - 7 days	16	79	95
Total	102	98	200

With the significantly low p value, Medicine taken and Cold Length are dependent on each other!

		Medicine Taken		
Cold Length		yes	no	Total
	1 -3 days	86	19	105
	4 - 7 days	16	79	95
	Total	102	98	200

qq plots: probability that two distributions are similar.

matches quantiles of expected vs quantiles of actual

great for checking normal distributions!

can be used for likelihood results given Bernoulli distribution

Cook's distance

measures validity of data (also continuous data, like qq plot)

How influential data points are, given if they were removed

Use cook's distance + qq plots to determine data validity