Causality, Dependency, Correlation, and Designed Experiments

Arnaud Legrand and Jean-Marc Vincent

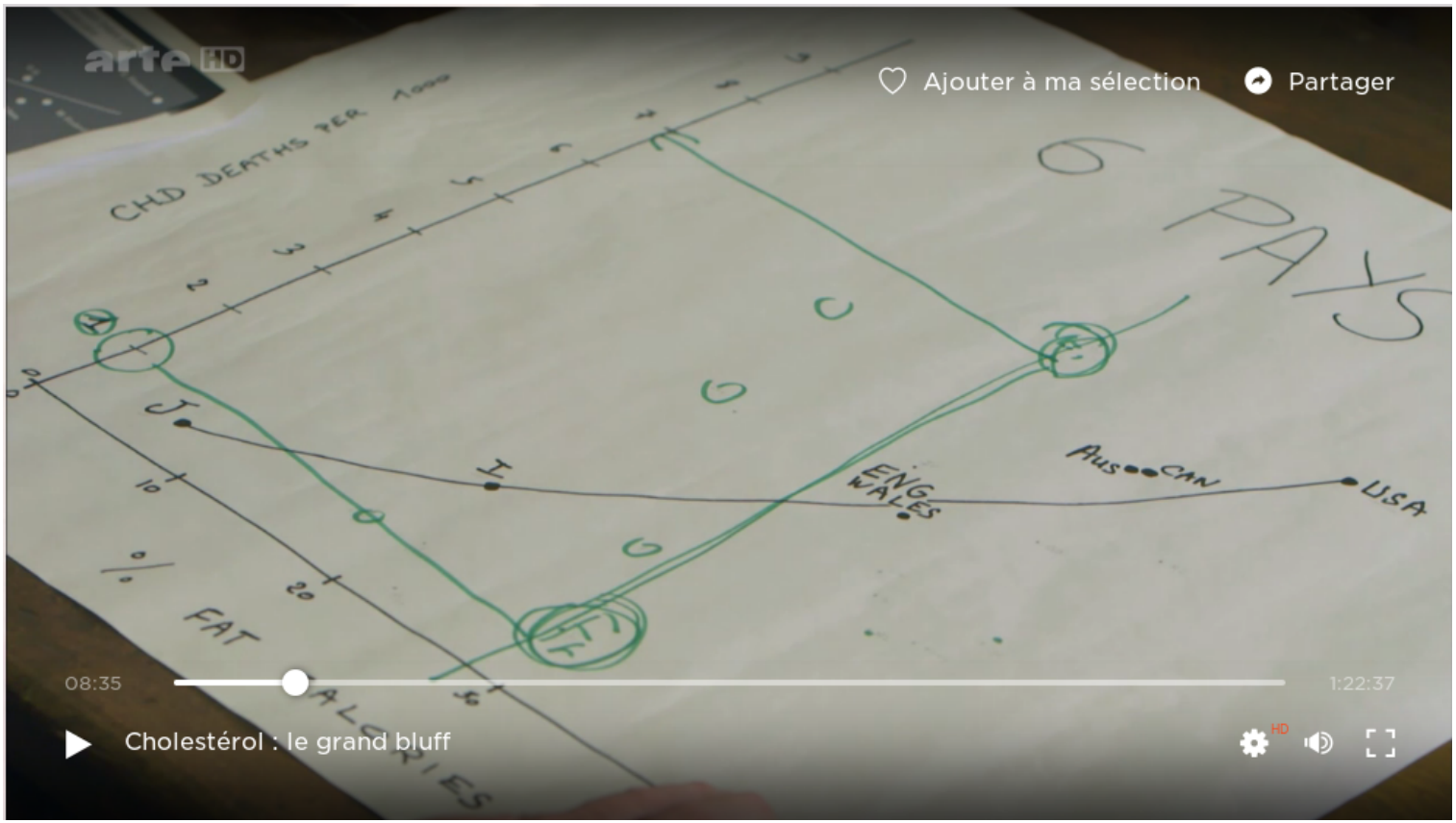
Scientific Methodology and Performance Evaluation

A vivid debate: Cholesterol and Statins

Cholesterol: le grand bluff (Arte, 18/10/2016 @ 20h50)

A vivid debate: Cholesterol and Statins

Cholesterol: le grand bluff (Arte, 18/10/2016 @ 20h50)



"Careful" selection of data and influence from the industry ☹

But that's not what I want to illustrate now... Even if data hadn't been removed, could we really conclude something from such data?

Outline

Correlation and Causation

Let me illustrate this inference story with a few examples.

It may be the case that two random variables X and Y are dependent

- E.g., Let's pick a student at random and measure its *DrinkingHabit* and its *TestScore*
  - In general, the more a student drinks the more his test goes down ☺

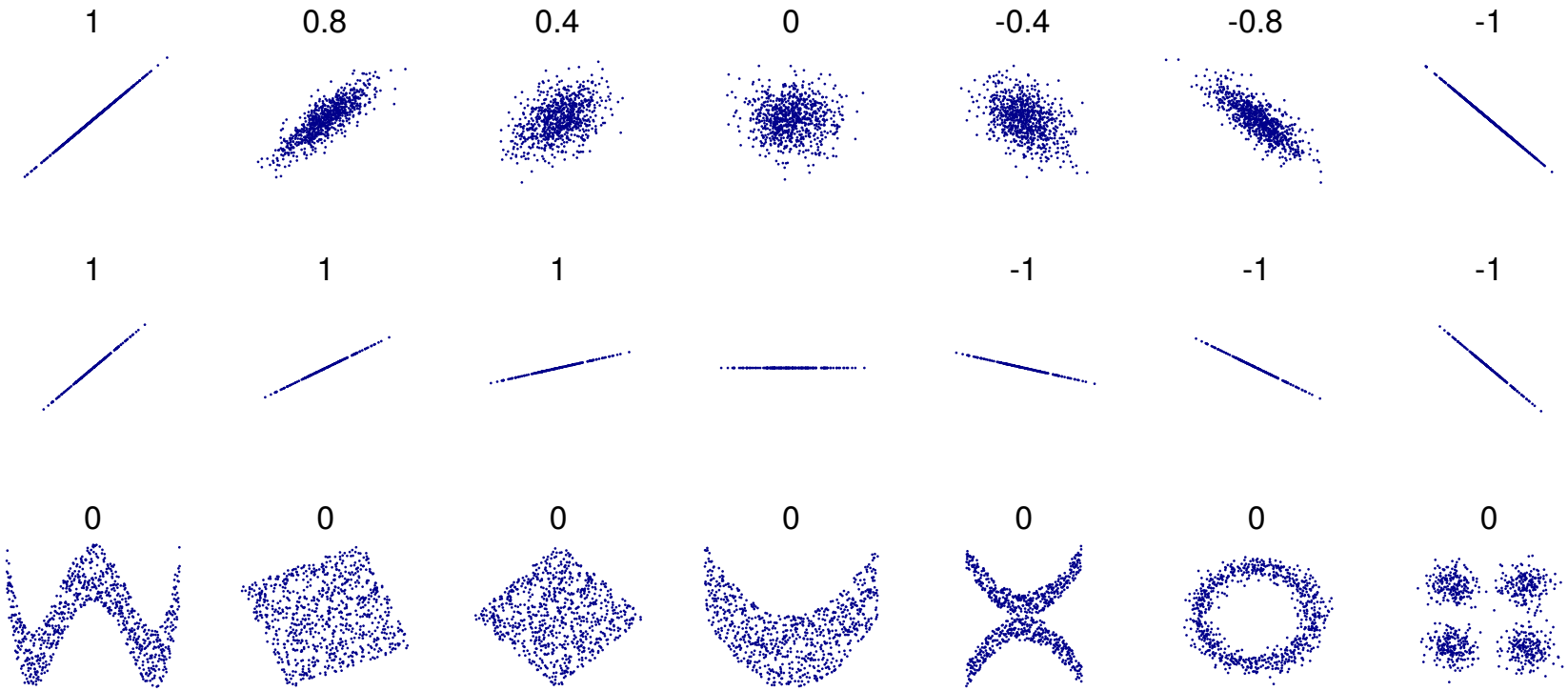
The correlation of two variables X and Y is defined as:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\text{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

- The correlation is symmetrical ($\text{corr}(X, Y) = \text{corr}(Y, X)$)
- The correlation is in $[-1, 1]$
- $\text{corr}(Y, X) = 1$ or $-1 \Rightarrow$ perfectly linear relationship
- X independent of $Y \Rightarrow \text{corr}(X, Y) = 0$
- Y grows when X grows $\Rightarrow \text{corr}(X, Y) > 0$

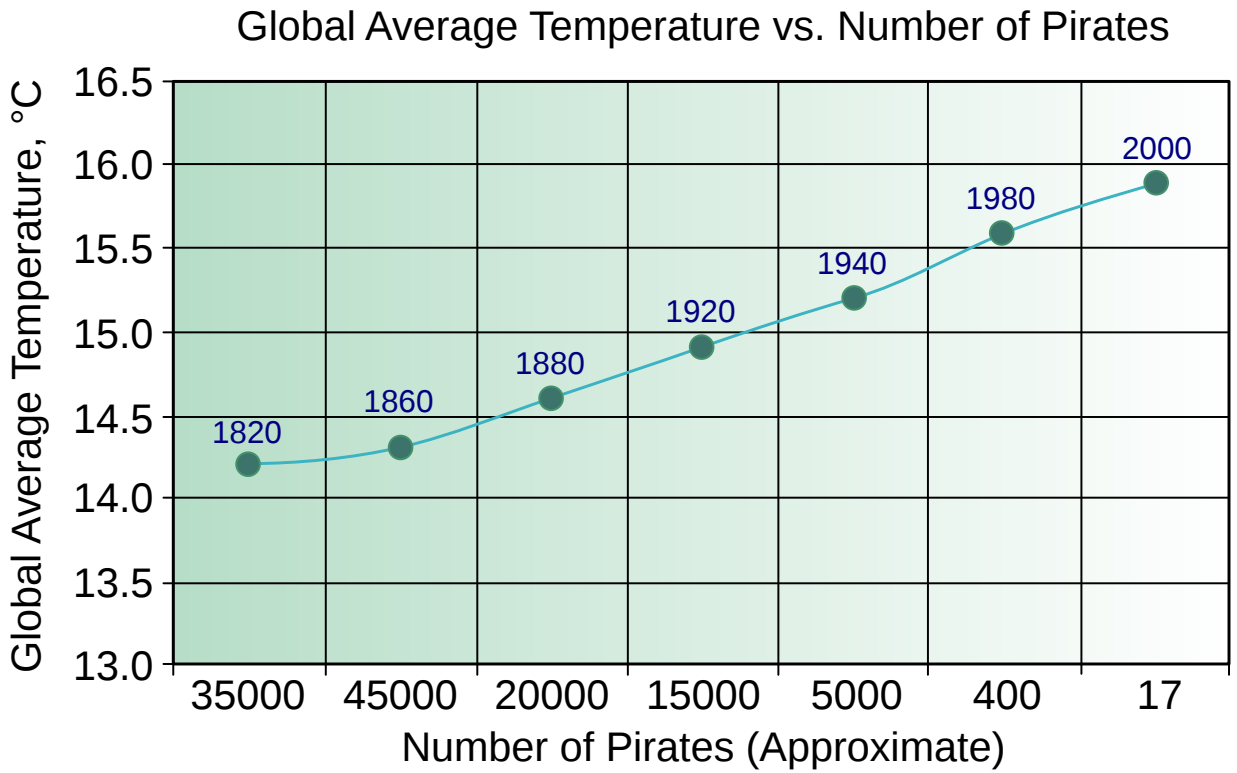
It is thus very tempting to use sample correlation as a way of knowing whether some variables are dependant

Scatter plot and correlation



Non-linear relations or hidden variables are not be well trapped by correlation

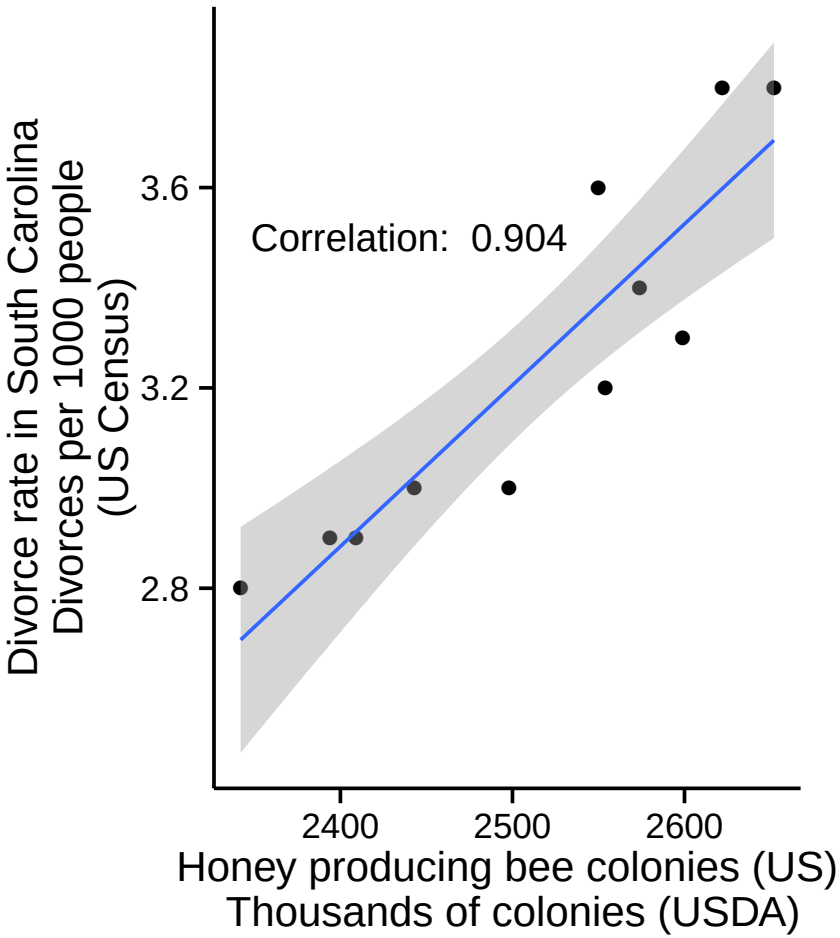
Correlation does not imply Causation

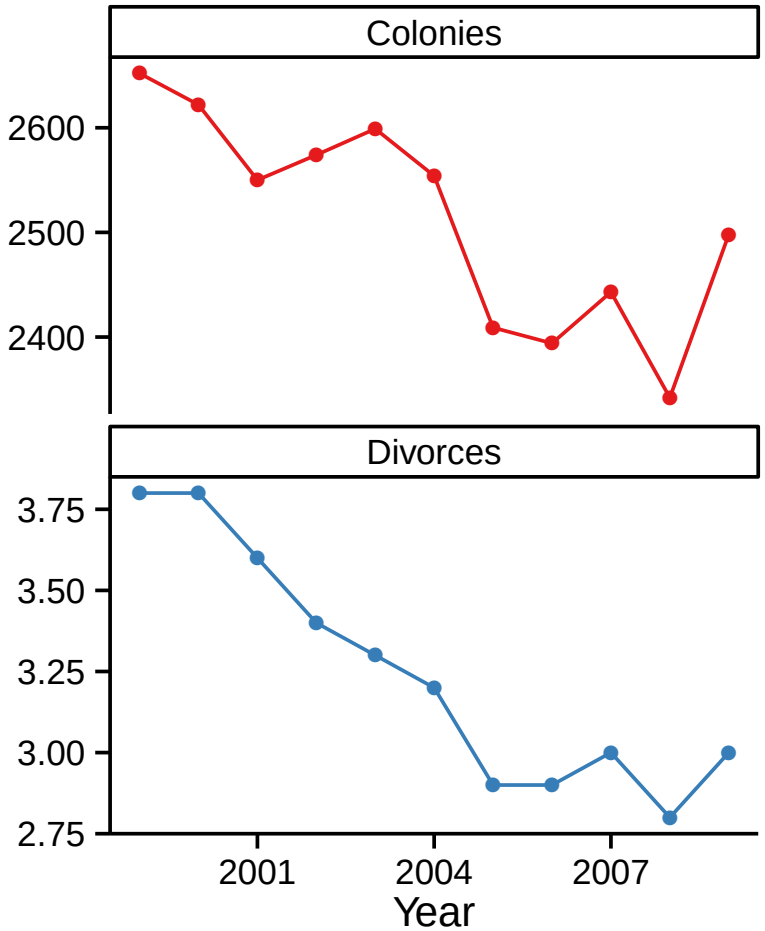


Mikhail Ryazanov (talk) - PiratesVsTemp.svg.
Licensed under CC BY-SA 3.0 via Wikimedia Commons

- 2 variables can be strongly correlated to a third one (e.g., year)
- Btw, what is wrong with this figure? 😉

Observational vs. Experimental Data Illustration





Source: *Spurious correlations*. For the good of the US society, we should try to get rid of honey bees 😉

The Deluge of Spurious Correlations in Big Data

The Deluge of Spurious Correlations in Big Data, by C. Calude and G. Longo, Foundations of ScienceMarch 2016)

Is Data science is the end of science ?

- Powerful algorithms can now explore huge databases and find therein correlations and regularities.
- Properly defining "meaning" or "content" of such correlations is very difficult. But do we need to ?

Ergodic Theory

- Almost every trajectory (even deterministic and chaotic) will eventually iterate in a similar way
- So regularity is expected but it does not mean that prediction can be done.

Ramsey Theory

- Any sufficiently long string contains an arithmetic progression
  - $\underline{0}$, 1, 1, 0, $\underline{0}$, 1, 1, 0, $\underline{0}$
  - 0, 1, $\underline{1}$, 0, 0, $\underline{1}$, 1, 0, $\underline{1}$
- Similar result for n ary relations

Simpson's Paradox

UC Berkeley admission figures in fall 1973.

Men		Women	
Applicants	Admitted	Applicants	Admitted
8442	44%	4321	35%

Simpson's Paradox

UC Berkeley admission figures in fall 1973.

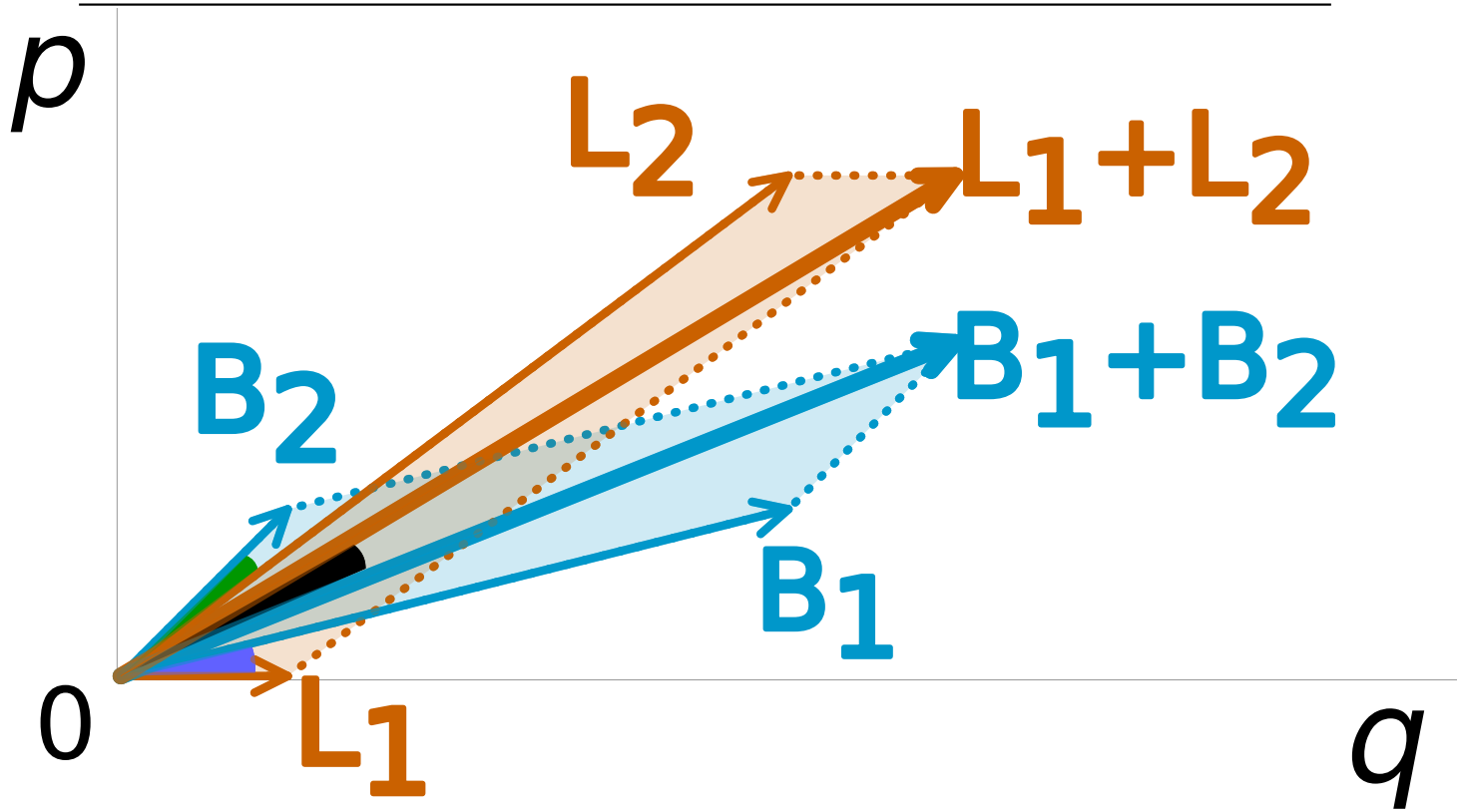
Men		Women	
Applicants	Admitted	Applicants	Admitted
8442	44%	4321	35%

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

Simpson's Paradox

UC Berkeley admission figures in fall 1973.

Men		Women	
Applicants	Admitted	Applicants	Admitted

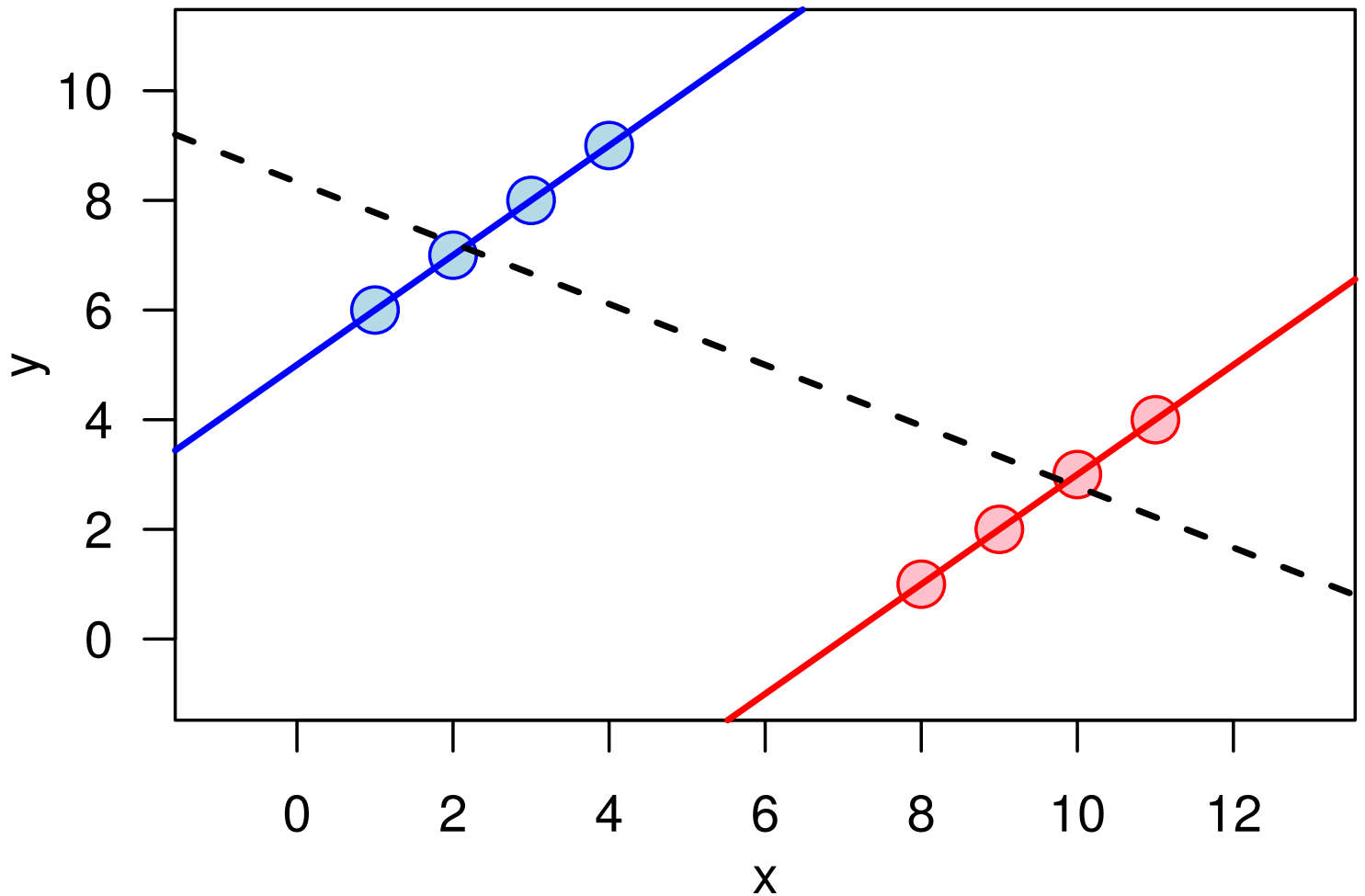


$L_1 < B_1$ and $L_2 < B_2$, yet $L_1 + L_2 > B_1 + B_2$

Simpson's Paradox

UC Berkeley admission figures in fall 1973.

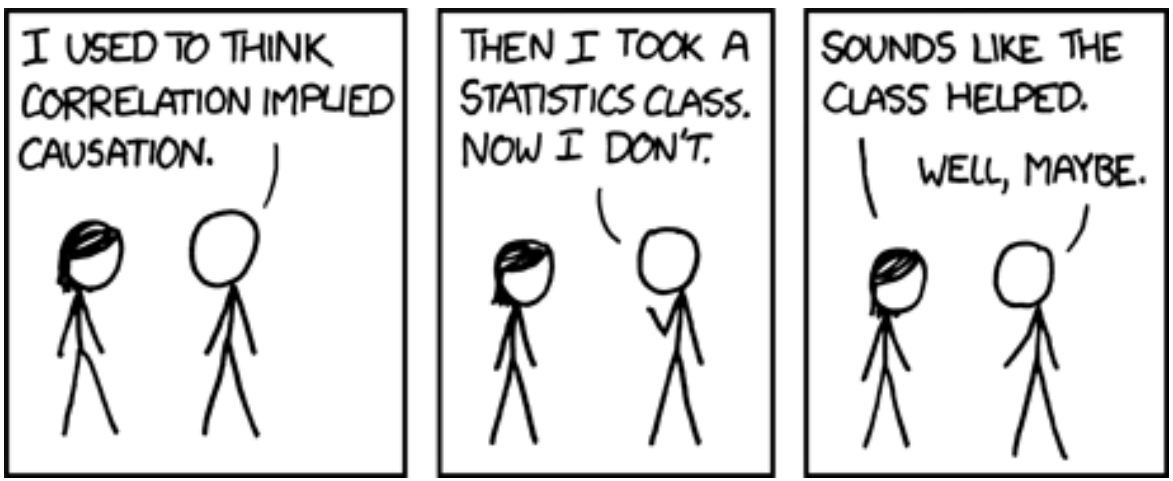
Men		Women	
Applicants	Admitted	Applicants	Admitted



Correlation does not imply Causation

For any two correlated events, A and B, the following relationships are possible:

- A causes B (direct causation) 
- A causes B and B causes A (bidirectional or cyclic causation) 
- A causes C which causes B (indirect causation) 
- B causes A; (reverse causation) 
- A and B are consequences of a common cause, but do not cause each other 
- There is no connection between A and B; it is a "coincidence" 
  - But designed experiments can help you ruling this option out



Outline

Experimental data vs. Observational data

You need a good blend of observation, theory and experiments

- Many scientific experiments appear to be carried out with no hypothesis in mind at all, but simply to see what happens.
- This may be OK in the early stages (inductive reasoning) but drawing conclusions on such observations is difficult (large number of equally plausible explanations; without testable prediction no experimental ingenuity; . . .).

Experimental data vs. Observational data

You need a good blend of observation, theory and experiments

- Many scientific experiments appear to be carried out with no hypothesis in mind at all, but simply to see what happens.
- This may be OK in the early stages (inductive reasoning) but drawing conclusions on such observations is difficult (large number of equally plausible explanations; without testable prediction no experimental ingenuity; . . .).

erence Essential steps:

1. Formulate a clear hypothesis
2. Devise an acceptable test

erence It would be silly to disregard all observational data that do not come from designed experiments. Often, they are the only information we have (e.g. the trace of a system).
But we need to keep the limitations of such data in mind. It is possible to use it to derive hypothesis but not to test hypothesis (i.e., claim facts).

Experimental Design

There are two key concepts:

replication and randomization

You replicate to increase reliability. You randomize to reduce bias.

If you replicate thoroughly and randomize properly, you will not go far wrong.

Experimental Design

There are two key concepts:

replication and randomization

You replicate to increase reliability. You randomize to reduce bias.

If you replicate thoroughly and randomize properly, you will not go far wrong.

> *It doesn't matter if you cannot do your own advanced statistical analysis. If you designed your experiments properly, you may be able to find somebody to help you with the statistics.*
>
> *If your experiments is not properly designed, then no matter how good you are at statistics, you experimental effort will have been wasted.*

No amount of high-powered statistical analysis can turn a bad experiment into a good one.

Other important concepts:

- Pseudo-replication
- Experimental vs. observational data

Replication vs. Pseudo-replication

Measuring the same configuration several times is not replication. It's pseudo-replication and is generally biased

Instead, test other configurations (with a good randomization)

In case of pseudo-replication, here is what you can do:

- average away the pseudo-replication and carry out your statistical analysis on the means
- carry out separate analysis for each time period
- use proper time series analysis

Outline

Linux Users Got Bigger Ding Dong 😉

The world famous Kinsey institutes for Sex Studies have proved that the average Linux user has a bigger penis than the average Windows PC user.
The study, carried out over a 6 month period showed that just using Linux for six months caused an average growth of 1 cm in the overall girth of a man's penis.
Scientist at first theorize that since the average Linux user spends more time in front of his computer than a windows user, that perhaps radiation from the monitor is responsible for the increase is size.

– https://forums.pcbsd.org/thread-4392.html

(Heavily inspired from Richard Monvoisin's post.)

What would such a study look like ?

1. Measure the size of the penis of sample of linux users
   - representative ?
   - number of samples ?
2. Sum these measurements and divide by the number by the number of samples
3. Conduct a similar study with Windows and Mac OS X users.
   - Same number of samples as before ?
4. Conclude

Bias #1: Uncertainty

No information about the standard error (variability).
Let's imagine they gathered the following data (in cm):

- Windows: 10, 10, 10, 10, 10 $\rightsquigarrow$ 10 on average
- Linux: 8, 9, 9, 9, 40 $\rightsquigarrow$ 15 on average 😉

If I repeat the experiment, will I get the same results ? similar results ?
What are the odds ?

Bias #1: Uncertainty

No information about the standard error (variability).
Let's imagine they gathered the following data (in cm):

- Windows: 10, 10, 10, 10, 10 $\rightsquigarrow$ 10 on average
- Linux: 8, 9, 9, 9, 40 $\rightsquigarrow$ 15 on average 😉

If I repeat the experiment, will I get the same results ? similar results ?
What are the odds ?

Handle "outliers", confidence intervals

No information about the protocol:

- volunteer users / rewarded / random sampling ?
- room temperature ?

Bias #2: Does such a computation make any sense ?

What does this even mean ?

- Is the average of penises representative of the "average penis"?
- Can we transpose relations between populations to individuals ?
- The average human has one breast and one testicle. . . 🙂
  - By the way how did they handle female linux users ?
- Anyway, "The bigger the better"?

Similar disturbing fact:

- High child mortality rate is corelated with the number of doctors
- Can we conclude that we should decrease the number of doctors ?

Bias #3: The stork effect

- Maybe men with a larger penis tend to use linux rather than other OS.

Bias #3: The stork effect

- Maybe men with a larger penis tend to use linux rather than other OS.
- A "better" explanation: Linux makes you look cool, hence the linux users were mostly teenagers in full growth... ☺
- Maybe linux users were easier to find at University than in companies, hence they belong to a different population

The Stork effect:

- Cities that host storks tend to have a higher birth rate.
- Stork probably bring babies ;)
- Or Cities that host storks are more likely found in rural environment where birth rate is higher for socio-economical reasons...

Citing Sources and Reproducible Research

On 10 October 2006, the number of sites that relayed this information has exploded...
But although there exists a Kinsey Institute, there has never been any such news nor data that would support such a study...

- Just imagine what it is like now that we have twitter 😉

Outline

Select the problem to study

Clearly define the kind of system to study, the kind of phenomenon to observe (state, evolution of state through time), the kind of study to conduct (descriptive, exploratory, prediction, hypothesis testing, ...)

This is quite important as the set of experiments to perform will be completely different when you are:

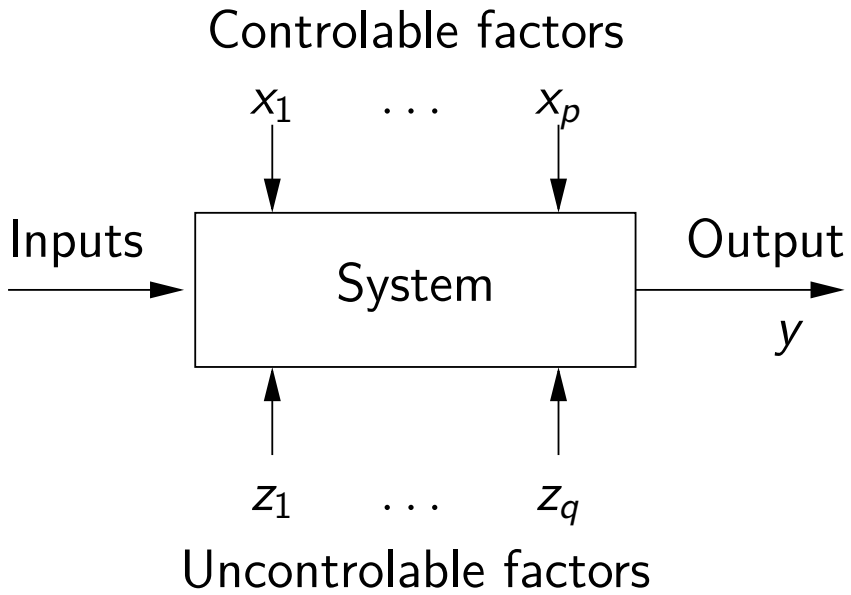
- studying the stabilization of a peer-to-peer algorithm under a high churn
- trying to compare various scheduling algorithms or code versions
- modeling the response time of a server under a workload close to the server saturation
- ...

This step will help you to determine which kind of experiment design you should use.

Determine the set of relevant *factors* and *responses*

The system under study is generally modeled though a black-box model:

- some output variable/response(y)
- some inputs are fully unknown
- some input variables ($x_1, \ldots, x_p$) are controllable
- whereas some others ($z_1, \ldots, z_q$) are uncontrollable



Typical controllable variables could be:

- the heuristic used (e.g., FIFO, HEFT, . . .)
- one of their parameters (e.g., replication factor, a threshold, . . .)
- the size of the platform
- the degree of heterogeneity
- the version of the compiler

Uncontrollable variables could be:

- temperature, humidity, moon phase, road surface conditions
- someone using the machine and interfering with the experiment

You can organize them in a dogbone diagram

You should carefully record all the factors you can think of

Typical case studies

The typical case studies defined in the first step could include:

- Determining which variables are most influential on the response y (factorial designs, screening designs, analysis of variance)
  - Allows to distinguish between primary factors whose influence on the response should be modeled and secondary factors whose impact should be averaged
  - Allows to determine whether some factors interact in the response
- Devise an analytical model of the response y as a function of the primary factors x (regression, lhs designs)
- Fit a an analytical model (regression, response surface methodology, optimal designs)
  - Can then be used to determine where to set the primary factors x so that response y is always close to a desired value or is minimized/maximized
- Determining where to set the primary factors x so that variability in response y is small i.e., so that the effect of uncontrollable variables $z_1, \ldots, z_q$ is minimized (robust designs, Taguchi designs)

General Workflow