Why You Need to Ask How in Order to Know What

Georges Monette
York University
georges@yorku.ca
DILBERT

TOUR OF ACCOUNTING

OVER HERE
WE HAVE OUR
RANDOM NUMBER
GENERATOR.

NINE NINE
NINE NINE
NINE NINE

ARE YOU
SURE
THAT'S
RANDOM?

THAT'S THE
PROBLEM
WITH RAND-
OMNESS:
YOU CAN
NEVER BE
SURE.
Which is more random: 

\[ \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} = 10^{-6} \]

or

\[ \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} = 10^{-6} \]
Which is more random:

9 9 9 9 9 9 or 6 5 3 5 8 9

The expression ‘random number’ is misleading:

It’s the process generating the numbers that is random, not the numbers themselves.
It’s a random (number generator) not a (random number) generator
Which numbers are more random:

9 9 9 9 9 9

Or

6 5 3 5 8 9

3.14159265358979
Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

– H. G. Wells (ca 1903)
The United States of Big Data

By Rachael King

Even the U.S. government is getting in on so-called big data.

Today, the Obama administration said it would make $200 million in new research and development investments to glean insights from large and complex collections of digital data. The government said it also would try to expand the number of workers who are able to sift through huge piles of digital data to find scientific or business insights.

"After Eisenhower, you couldn't win an election without radio. After JFK, you couldn't win an election without television. After Obama, you couldn't win an election without social networking. I predict that in 2012, you won't be able to win an election without big data."
Some things we can do when we get our hands on big data:
Increasing cigarette consumption by 1,000 per year corresponds to an increase in life expectancy of 7.2 years.
Increasing cigarette consumption by 1,000 per year corresponds to an increase in life expectancy of 7.2 years.

All it takes is 3 cigarettes a day to add 8 years to your life!
Increasing cigarette consumption by 1,000 per year corresponds to an increase in life expectancy of 7.2 years.

All it takes is 3 cigarettes a day to add 8 years to your life!

Is there anything wrong with our reasoning?
What do we know?

$Y = \text{Life Expectancy}$

$X = \text{Cigarette Consumption}$

In our data $X$ and $Y$ have a positive association (correlation)
Experiment

1) $X \rightarrow Y$

2) $Y \rightarrow X$

3) $Z \rightarrow Y$

4) Chance

5) Selection

$X \rightarrow Y$
Correlation is not causation

Is there a solution?
The ideal solution: 

R. A. Fisher (~1920)

Experiment

You need to know how levels of X were assigned to subjects. The choice to smoke or not should not be up to the subject. It must be determined by a random process.
Randomized experiments are an ideal way to discover whether X causes Y.

But there are many problems with randomized experiments:
- too costly
- too risky
- too long
- observational data already on hand so we should use it
- won't give an answer until it's too late
- experimental situation not realistic

Essentially randomized experiments can only be used for relatively few selected questions.

Often, all that's available is non-experimental (observational) data.
Is smoking harmful?

Early 1950s – Richard Doll established a strong correlation between smoking and the risk of lung cancer. Evidence mounted but it was based on observational data.
Dangers of Cigarette-smoking

Sir,—In the *Journal* of July 20 (p. 158) Dr. Robert N. C. McCurdy writes, “Fisher’s criticism (*Journal*, July 6, p. 43) . . . would not be so unfair if he had specified what alternative explanations of the facts still await exclusion.” I had hoped to be brief. A few days later the B.B.C. gave me the opportunity of putting forward examples of the two classes of alternative theories which any statistical association, observed without the precautions of a definite experiment, always allows—namely, (1) that the supposed effect is really the cause, or in this case that incipient cancer, or a pre-cancerous condition with chronic inflammation, is a factor in inducing the smoking of cigarettes, or (2) that cigarette-smoking and lung cancer, though not mutually causative, are both influenced by a common cause, in this case the individual genotype.
In the Preface to a collection of articles:

If, indeed, the statistical departments engaged in university teaching, were performing their appropriate task, of clarifying and confirming, in the future research workers who come within their influence, an understanding of the art of examining observational data, the fallacious conclusions drawn, from a simple association, about the danger of cigarettes, could scarcely have been made the basis of a terrifying propaganda.

For this reason I have thought that the fallacies must be attacked at both of two distinct levels; as an experimental scientist, and as a mathematical statistician. The lecture on *The Nature of Probability* was to a non-mathematical audience, on the general question of the validity of inferences from facts available on lung cancer.
The challenge:
How can good citizens develop a balanced view of evidence.

Why?

It’s critical in order to understand the nature of most scientific controversies, many political ones, and to make informed choices in everyday living.
Climate change

Personal health:
Does sunscreen cause cancer?
Does eating red mean cause heart attacks?

Effects of social policies

Side effects of drugs: are they ‘caused’ by the drug or by the factors that lead patients to take the drug?
We can’t escape the need to develop a balanced view of causal inference from observational data.

*For example*: clinical trials are used to assess the effectiveness of drugs but possible rare side-effects need to be monitored with observational data once the drug is on the market.
Current textbooks:

Most stop at ‘correlation is not causation’

A few go a bit further:


Statistical Ideas and Methods by Utts and Heckard (2006)
Footnote:

Another important ‘how’:

Interpreting surveys: what the figures mean depends critically on how the sample was selected.