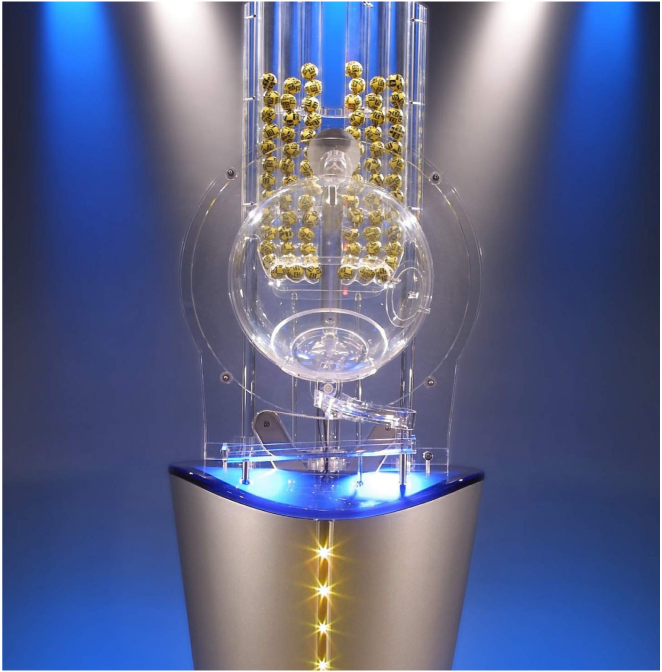
When Models Mislead

John Myles White

November 24, 2014

tl;dr: $A \to B \neq B \to A$

All of statistics starts with a data generating process



To build theories, we start with models and end up with data

In applied work, we try to go from data back to models

Sometimes we can reach conclusions safely

Often we can't

Elapsed time when executing sqrt(2.0)
10μs
10μs
10μs
10μs
10μs

How can moving backwards from data to models mislead?

Example 1: The Problem with P-Values

Reject a model if the model says observed data is unlikely

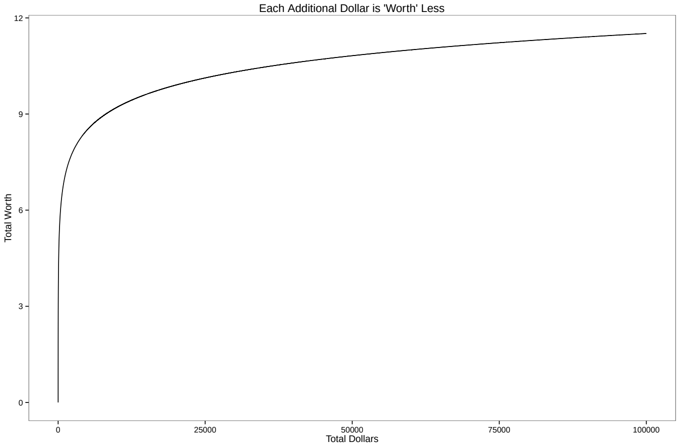
Misleads if the observed data is unlikely under other models

If a person is an American, then they are probably not a member of Congress.

Al Franken is a member of Congress.

Therefore, Al Franken is probably not an American.

Example 2: The Problem with Diminishing Marginal Utility


Each Additional Dollar is 'Worth' Less
Total Worth
12
9
6
3
0
0
25000
50000
75000
100000
Total Dollars

The St. Petersburg Lottery:

- If you get heads on your first turn, you win \$1.
- If you get tails, you take a second turn.
- If you get heads on your second turn, you win \$2.
- If you get tails, you take a third turn.
- etc. . .

How much would you pay to play the St. Petersburg Lottery?

The expected value of the St. Petersburg Lottery is infinite:

$$(\frac{1}{2})^1 2^0 + (\frac{1}{2})^2 2^1 + (\frac{1}{2})^3 2^2 + ... = \infty$$

People won't pay much to play the St. Petersburg Lottery

We'd expect this if the 'worth' of each dollar decreases

But we'd also predict some very strange behaviors

TABLE I

IF AVERSE TO 50-50 LOSE \$100 / GAIN g BETS FOR ALL WEALTH LEVELS, WILL TURN DOWN 50-50 LOSE L / GAIN G BETS; G'S ENTERED IN TABLE.

L	g = \$101	\$105	\$110	\$125
\$400	400	420	550	1,250
\$600	600	730	990	∞
\$800	800	1,050	2,090	∞
\$1,000	1,010	1,570	∞	∞
\$2,000	2,320	∞	∞	∞
\$4,000	5,750	∞	∞	∞
\$6,000	11,810	∞	∞	∞
\$8,000	34,940	∞	∞	∞
\$10,000	∞	∞	∞	∞
\$20,000	∞	∞	∞	∞

Example 3: The Problem with Delay Discounting

Would you rather have \$10 today or \$20 in one day?

Would you rather have \$10 today or \$40 in two days?

There are two popular models: exponential and hyperbolic

Exponential says you'll answer both questions the same way

Hyperbolic says you won't

People don't answer both questions the same way

Does that mean the hyperbolic model is right?

Learning from data is dangerous

If a model fits our data, that doesn't make it good

If a model doesn't fit our data, that doesn't make it bad

A model tells us exactly what kinds of data we can expect

Data tells us a bit about what kinds of models we can expect

References

- Basic Logic
- Urns
- 1 Dot per Person
- Benchmarking Fast Functions
- "The Earth is Round ($p < .05$)"
- The St. Petersburg Paradox
- Rabin's Calibration Theorem
- Exponential vs. Hyperbolic Discounting