Applied Statistics

Lectures 17+18

David Barton & Sabine Hauert

Department of Engineering Mathematics

Outline

- Lies, damn lies, and statistics
- Fallacies, bias, and mis-representation

Lies, damn lies, and statistics. . .

If you can't prove what you want to prove, demonstrate something else and pretend they are the same thing. In the daze that follows the collision of statistics with the human mind, hardly anyone will notice the difference. Darrell Huff, How to Lie with Statistics, 1954

Lies, damn lies, and statistics. . .

Statistics, assuming theyve been done correctly, can't lie.

But the mathematical component is never the whole story, when statistics interact with the real world a process of deciding which statistics to look at, interpretation of the statistics and representation begins. . .

Types of errors

- Fallacies
  - An error in reasoning
- Bias
  - A partial perspective, either intentional or not, which taints the outlook of the whole
- Misleading Representation
  - The statistics may be correct, their representation may not be.

The Gamblers' fallacy

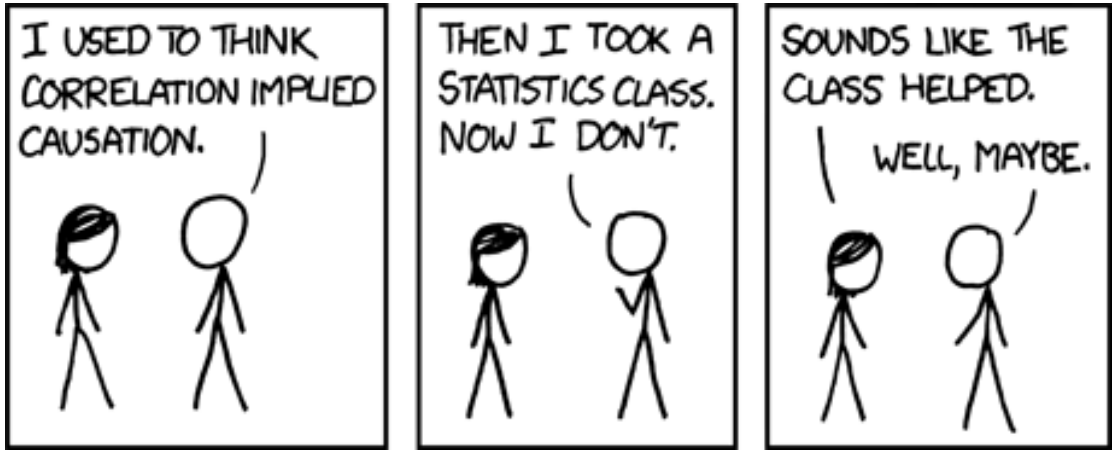
In a roulette game in Monte Carlo, during 1913, the ball landed in the black 26 times in a row.

After this gamblers lost millions disproportionately betting red on the same wheel, in the assumption that red must turn up eventually.

This fallacy is based on the assumption that the universe somehow evens out, that the number of head thrown will be the same as the number of tails thrown.

The causation is back to front, the random nature of the event drives the distribution, the likely distribution doesnt drive the events.

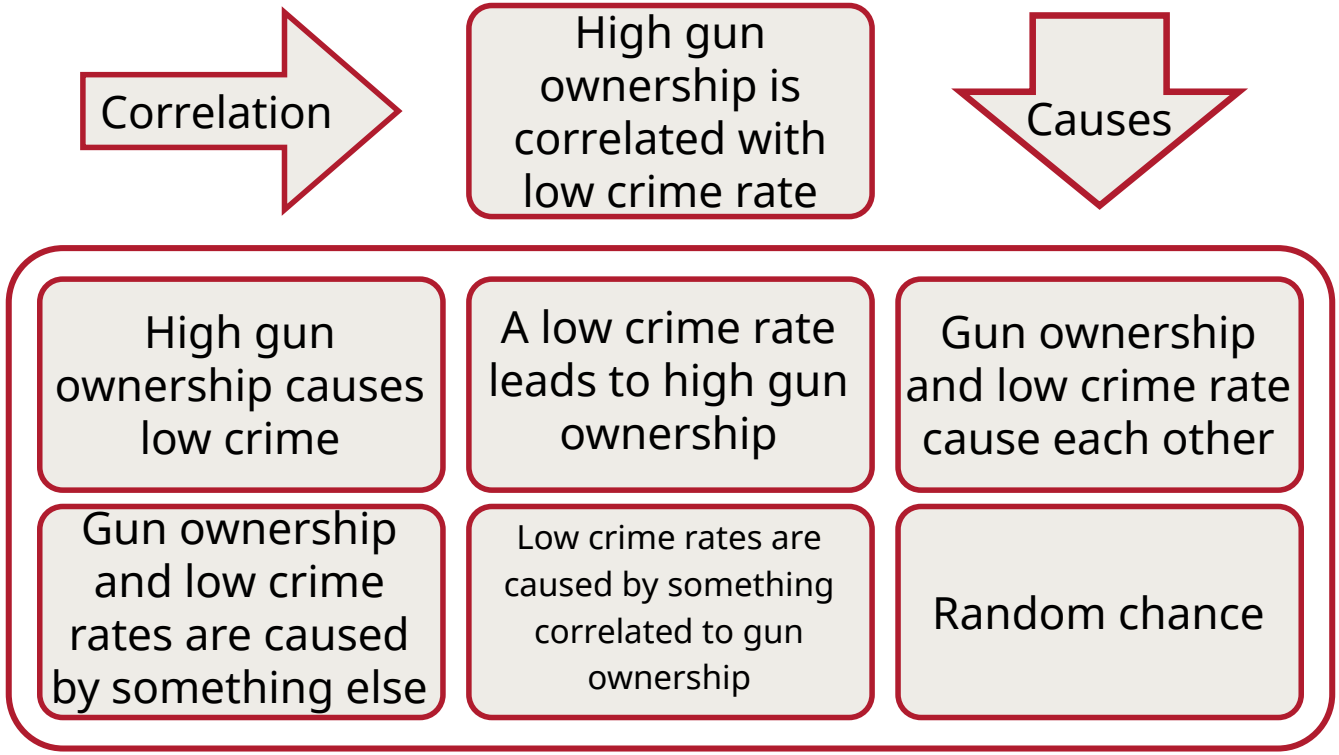
Correlation is not causation



XKCD 552

"Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'."

Correlation is not causation



Unpicking the chain of causation

If is difficult to determine which variable is causing the other

- Without extra information it is often impossible!

Easy(!) to check in experiments; alter one variable and observe the effect

- Less easy in sociological studies where it is impossible to re-run an entire society.
- Also problematic where there might be ethical concerns — is your intervention causing more traffic accidents?

Regression to the mean

> *Regression to the mean is the phenomenon that if a variable is extreme on its first measurement, it will tend to be closer to the average on its second measurement. (Wikipedia)*

If you are playing higher-or-lower with a pack of cards and you draw a Jack, you are almost certainly better off guessing that the next card drawn will be lower.

Seems obvious but there are many situations where people get confused. . .

The Madden Curse

American football players have often had a worse season than the previous after they have featured on the front cover of the Madden series of video games!

Regression to the mean?!

Other instances

While speed cameras do have an impact, their benefits may be overstated because of regression-to-the-mean in the same way as the Madden Curse. . .

The conjunction fallacy

> *Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.*
>
> *Tversky and Kahneman*

Which is more probable?

1. Linda is a bank clerk.
2. Linda is a bank clerk and is active in the feminist movement.

The conjunction fallacy

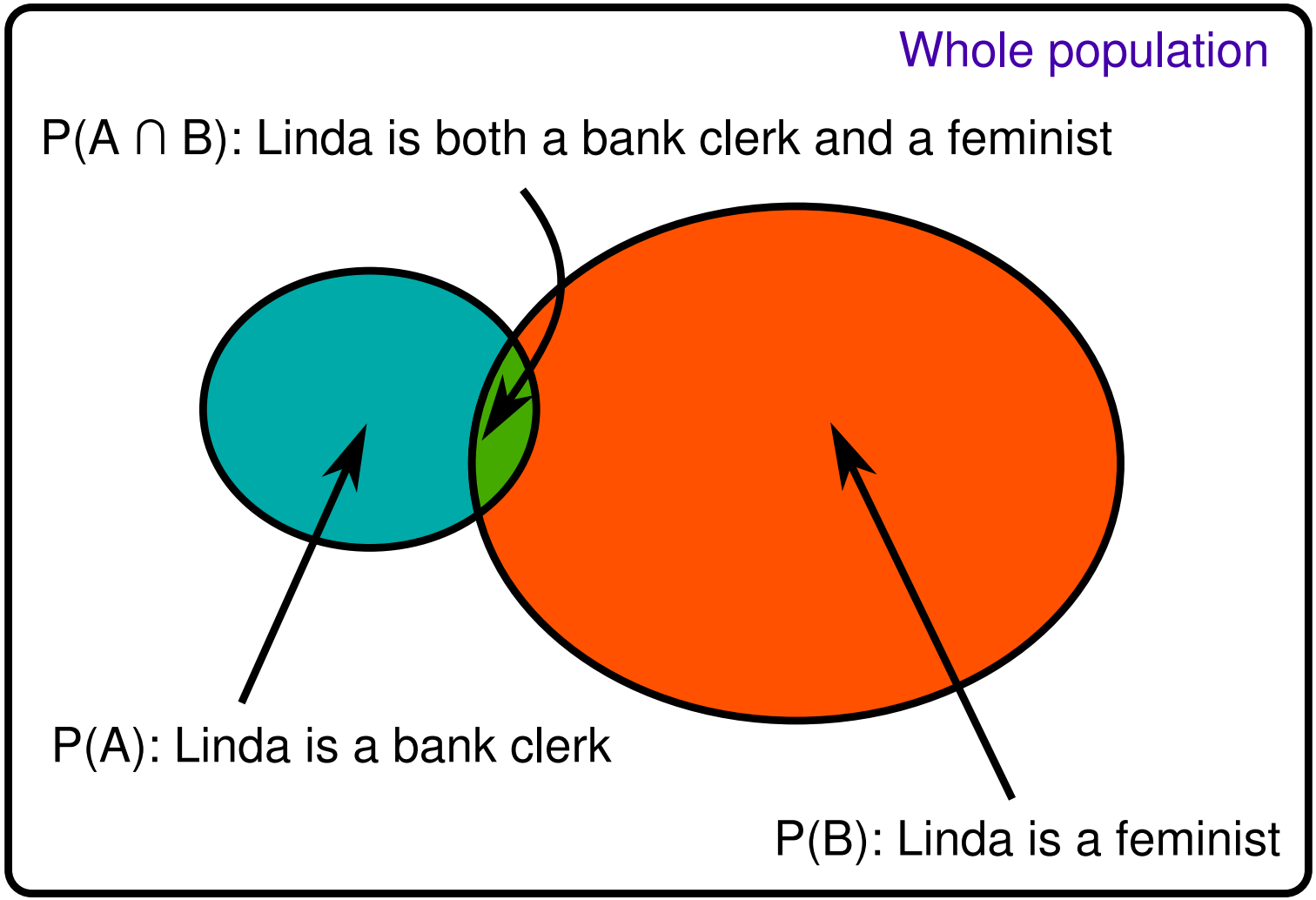
The conjunction fallacy occurs when you assume that specific conditions are more likely than a general one.

$\mathrm{P}(A)$	Probability that Linda is a bank clerk.
$\mathrm{P}(B)$	Probability that Linda is a feminist.
$\mathrm{P}(A \cap B)$	Probability that Linda is both a bank clerk and a feminist.

In general we have that $\mathrm{P}(A \cap B) \leqslant \mathrm{P}(A)$ and $\mathrm{P}(A \cap B) \leqslant \mathrm{P}(B)$. At best we have that $\mathrm{P}(A) = \mathrm{P}(A \cap B) = \mathrm{P}(B)$, that is, A and B are perfectly correlated.

If A and B are independent, then $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$. Remember that probabilities are in the range $[0, 1]$!

The conjunction fallacy



Base rate fallacy

If a medical test for a particular disease is 95% accurate and I am tested positive, what is the probability that I have the disease?
Not 95%!

The fallacy comes about because there are actually two random variables involved.

1. A random variable that determines whether I have the disease or not (ignoring any testing)
2. A random variable that determines whether the test is correct or not.

The 95% claim is (presumably) a marginal probability whereas we really need the conditional probability

$$\mathrm{P}(\text{have the disease} \mid \text{positive test})$$

Base rate fallacy explained

Consider a disease that has a rate of occurrence of 1 in 100. Assume that the test has equal false-positive and false-negative rates. This gives the joint distribution

	Present	Not present	Total
Test positive	0.0095	0.0495	0.059
Test negative	0.0005	0.9405	0.941
Total	0.01	0.99	1.00

Base rate fallacy explained

The calculation is thus

$$\begin{aligned} \mathrm{P}(\text{disease present} \mid \text{test positive}) &= \frac{\mathrm{P}(\text{disease present and test postive})}{\mathrm{P}(\text{test positive})} \\ &= \frac{0.0095}{0.059} \approx 0.16 \end{aligned}$$

Hence for a test with an accuracy of 95% for a disease which occurs in 1% of people means you only have a 16% chance of actually having the disease if you test positive for it!

Be careful that you've identified all the (random) variables!

(Also see the prosecutors' fallacy.)

The Ludic fallacy

> *"The misuse of games to model real-life situations."*

Identified by Nassim Nicholas Taleb in his 2007 book The Black Swan.

The real world is a complicated system, there are many underlying systems and relationships which are not apparent.

- Information may be missing, and it may not be known which information (if any) is missing (known unknowns, unknown unknowns. . .)
- Small unknown variations could have a large impact (different from chaos theory)
- Any model or theory can only be based on what has been observed. That which hasnt been observed will probably not be modelled.

An example

A coin is flipped 99 times, coming up heads each time.

A scientist, and a confidence trickster are each asked what is the most likely outcome, heads or tails?

- The scientist, assuming that the coin is perfectly predicted by the model in their head and also being aware of the Gamblers Fallacy, says heads or tails are both equally likely.

- The confidence trickster thinks something is up, and calls heads.

Summary of fallacies

- Gamblers' fallacy
- Correlation is not causation
- Regression to the mean
- Conjunction fallacy
- Base rate fallacy
  - Prosecutors' fallacy
- Ludic fallacy

Each of these particular types of thinking can blind a person to the underlying information in the numbers!

BIAS

The 1936 American Presidential Election In 1936 a postal survey was conducted to predict the next president of the USA. The survey was comprised of readers of the American Literary Digest magazine, with additional responses from registered car and phone owners.
The survey predicted Alf Landon, the Republican candidate, would easily win. The actual election was an easy victory for Franklin Roosevelt.
What Happened?

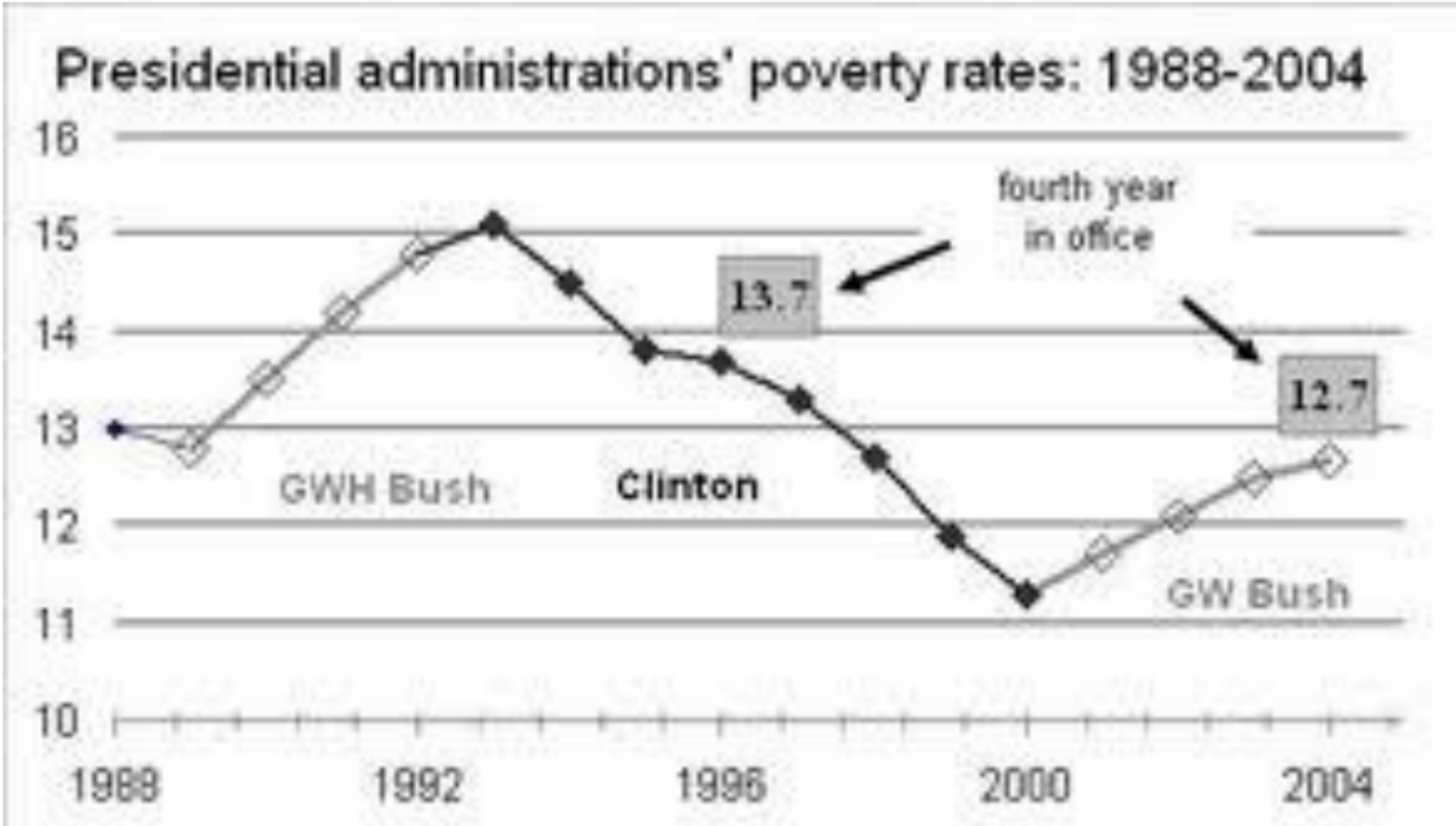
Sample Bias

The people surveyed were not randomly chosen and were not a statistically representative sample of the American Population
They were disproportionately rich, when compared to the average voter, and more likely to vote Republican

Cherry Picking

- A data set is rarely fully complete
  - Sometimes through experimental limitations
  - Sometimes through data being lost or unavailable
  - Sometimes it is deliberately missing
- If the data set is large enough, and the data loss is not systematic then this shouldnt matter
  An Example: In 2004 under Bush, the poverty level in America was 12.7%, under Clinton in 1996, the level was at 13.7%. From this we can conclude the Poverty reduced under the Bush Administration

"Bush Reduced Poverty"



Publication Bias

There is a natural tendency to prefer a positive result

- We want to hear what cures cancer, we dont want a list of medicinal failures. This is not necessarily good science. A positive study is no more likely to be methodologically sound than a negative study.
- Despite this, a positive study is more than three times as likely to be published.

File Drawer Effect

The previous example was publisher bias, but there is also researcher bias.
A researcher conducts a study with a statistically insignificant result.
Rather than attempt to publish, the result is filed away and forgotten.
A piece of research is lost

But more importantly

Out of 20 random trials, one will be deemed statistically significant at the P=0.05 level.
Example: 20 researchers trial the promising new drug Pheelbettaphasta.
19 researchers show no improvement in the patients at the P=0.05 level.
They dont publish
1 researcher detects an improvement.
They publish, and a systematic bias has been introduced, unintended, into the research.

How to combat this

Pre-registration of medical trials before results

- If the trial is pre-registered then it will be impossible to hide negative results, willingly or unwillingly

Better and less biased meta-reviews

- In meta-reviews the instinct is to look for high impact journal publications, by its nature this will overlook the negative results
- By looking for research that is more methodologically sound, regardless of the result, we get a better indication of the research

Frame Title

One effect of computers has been the development of Data Science, a discipline aimed at distilling large data sets into manageable and understandable sections.
However it has also led to Data Dredging, the practice of sifting through millions of data points and hundreds of variables, searching for any connection.

A Toy Example

100 data points are created, each with 50 separate variables.
This is the equivalent of a survey for 100 people with 50 questions.
In the example, the survey results are randomly generated, there is no relationship, by design.

Results

Although most relationship between variables are not significant some are.
Data dredging would pick up these spurious relationships as real.
Almost paradoxically, increasing the survey depth increases the chance of false relationships in the data.

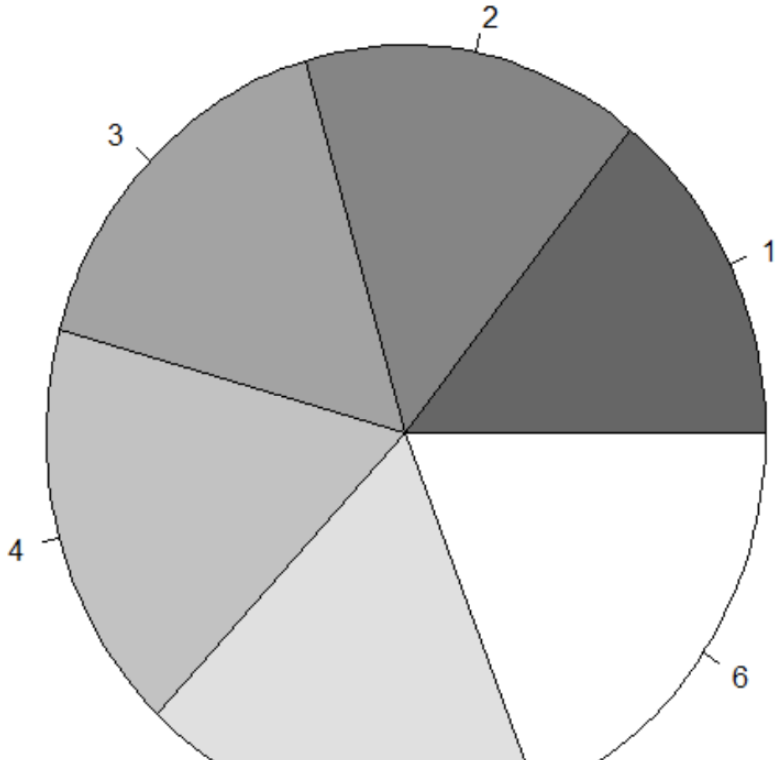
Misleading Representations

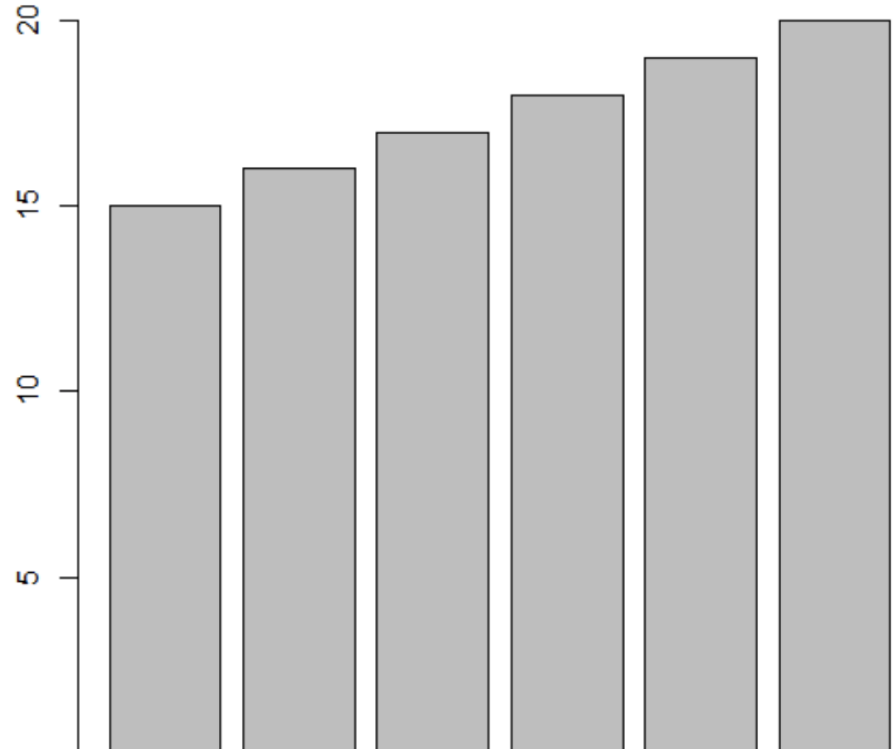
Relative: An intervention reduces the probability of an accident by 50%.
Absolute: An intervention reduces the number of accidents by 1 per month.
Which is better? It depends on the absolute incidence rate for the accident but 50% sounds better!

Technically Correct... but Misleading

If You Really Want to Obscure Your Data

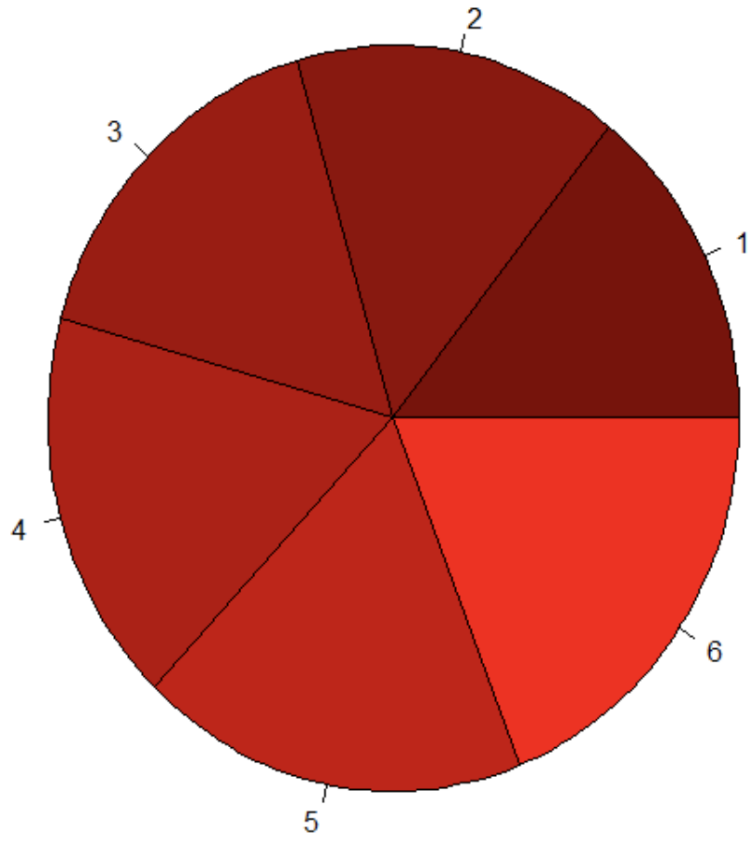
Pie charts can... minimise differences





If You Really Want to Obscure Your Data

Emphasise differences



Maury: Please put all my chips on red 21.
Dealer: Are you sure you want to do that? Red 21 just came up in the last spin.
Maury: I didnt know that! Thank you! Put it on black 15 instead. I cant believe I almost made that mistake!

Based on a survey of 1000 American homeowners, 99% of those surveyed have two or more automobiles worth on average $100,000 each. Therefore, Americans are very wealthy.

My headache went away because thats what headaches eventually do – they are a temporary disruption in the normal function of a brain.

My political candidate gives 10% of his income to the needy, goes to church every Sunday, and volunteers one day a week at a homeless shelter. Therefore, he is honest and morally straight.

While jogging around the neighborhood, you are more likely to get bitten by someones pet dog, than by any member of the canine species.

There are two people:
Dr. John, who is regarded as a man of science and logical thinking.
Fat Tony, who is regarded as a man who lives by his wits.
A third party asks them, ”assume a fair coin is flipped 99 times, and each time it comes up heads. What are the odds that the 100th flip would also come up heads?” Dr. John says that the odds are not affected by the previous outcomes so the odds must still be 50/50. Fat Tony says that the odds of the coin coming up heads 99 times in a row are so low (less than 1 in 6.33 1029) that the initial assumption that the coin had a 50/50 chance of coming up heads is most likely incorrect.

https://www.logicallyfallacious.com/tools/lp/Bo/LogicalFallacies

Real World Implications

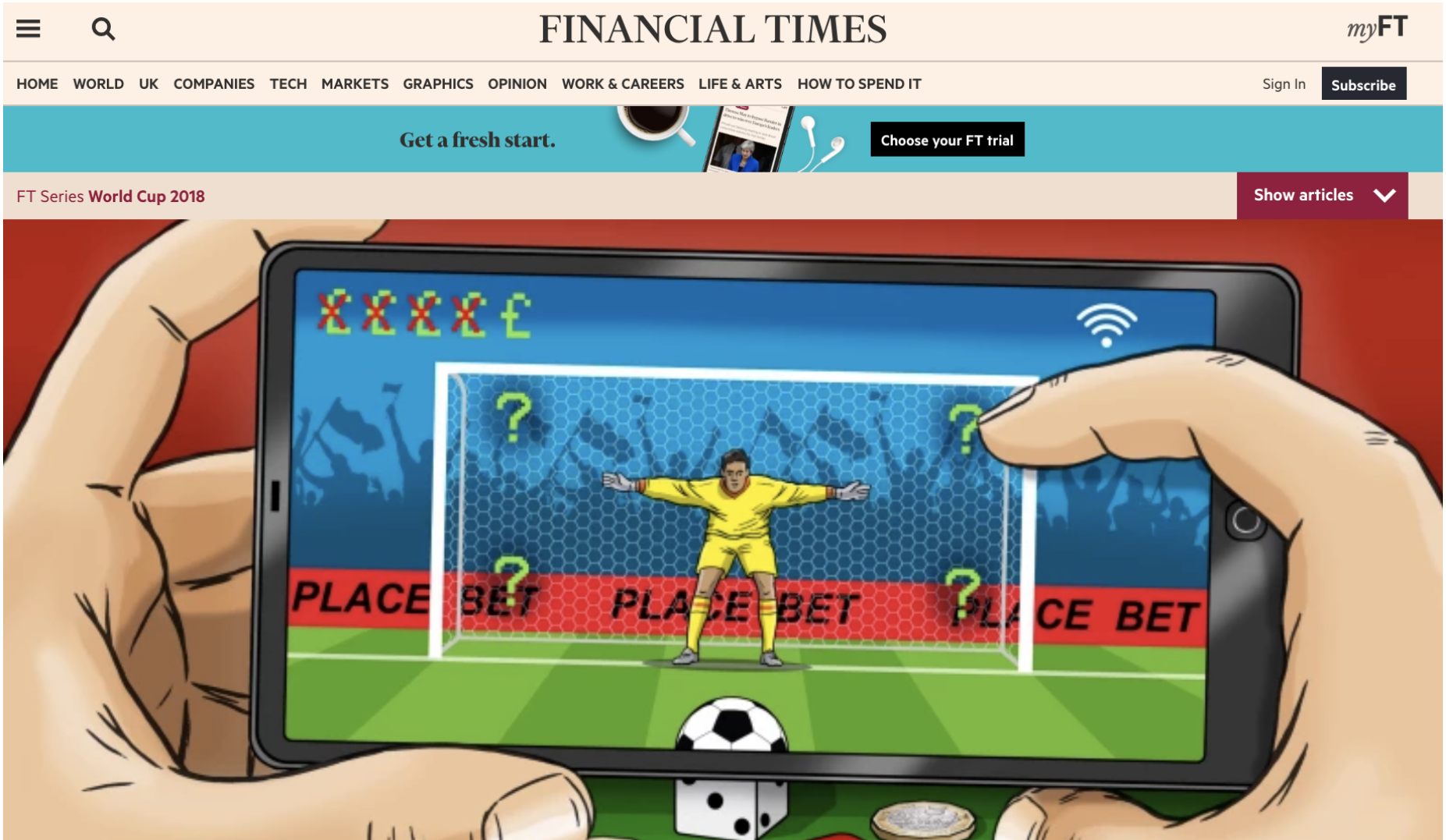
32,026 views | Feb 2, 2017, 08:50pm

Lies, Damned Lies And Statistics: How Bad Statistics Are Feeding Fake News

Kalev Leetaru Contributor ⓘ

I write about the broad intersection of data and society.

Real World Implications



Real World Implications

The New York Times

How Data Failed Us in Calling an Election

Real World Implications

Bloomberg

Economics

China Government Advisers Worry About AI Taking Over Jobs

Bloomberg News

6 March 2019, 08:10 GMT