Takeaways

1) Data neither lie nor tell the truth. Statistics interpret data and give it meaning.

2) Always put statistics in context. Is it representative of the whole world, a single country, or a city?

3) Correlation is not causation.

4) Garbage in, garbage out. If your data is ill and inappropriate in the context you explore it, your takeaways from the analysis will also be such.

5) We're all just humans, so always check potential biases in the selection, collection, processing, and presentation of data.

> It's easy to lie with statistics but hard to tell the truth without them.

> "An association between two variables is like a fingerprint at the scene of the crime. It points us in the right direction, but it's rarely enough to convict."

We use statistics in all aspects of our lives. It may not be as apparent, but when you play basketball you try to shoot from the spot that you have the best percentage of scoring from. In an election, you listen to the candidates and what they offer, as well as their and their parties' past behaviors when all combined lead to your voting decision.

So even though you may not be aware, your brain always collects this data and utilizes it in your day-to-day decision-making.

With various data sources (some are obvious, some are not; some are in CSV format, some are just your interaction with a friend), we are exposed to a myriad of data and it's in our hands to make use of them.

Some Concepts and their application in our lives

Descriptive Statistics and Deceptive Description

The difference between data presented as median or mean can be crucial—it could be the difference between profit and loss. Descriptive statistics help companies determine whether they are shipping many products with minor defects or a few with serious flaws, guiding them toward more effective solutions. However, these numbers can also be misleading.

Consider these scenarios:

- Verizon might cover more towns across the U.S., but if AT&T is more popular in major cities and states, AT&T could have a larger customer base and higher revenue.
- Or take a new drug that increases the life expectancy of cancer survivors by two weeks—is this enough to justify investment? What if the drug only benefits 40% of patients? The median life expectancy might not show this, but the mean would.

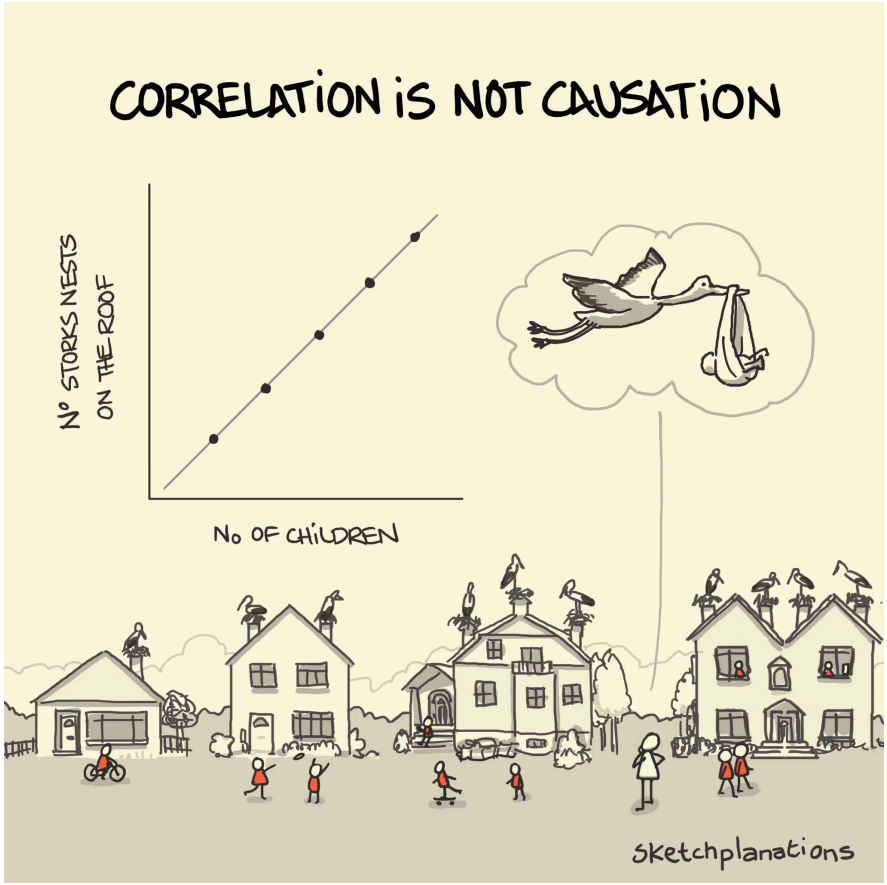
A more obvious example of deceptive statistics is in the U.S. News and World Report's university rankings. They heavily weigh "academic reputation," based on surveys from college professors. Harvard ranks high because professors think highly of it.

If U.S. News intends this score to reflect academic success, it's misleading. But if it aims to measure perceptions of academic success, then the statistic is accurate.

Correlation

Correlation does not necessarily mean causation. Changes in one variable may impact the other one, but not necessarily.

If people with more screen time on YouTube also have more screen time on Twitch, this could actually be due to them having social media addiction. It doesn't necessarily mean they watch Twitch more because they watch YouTube also more.



Basic Probability and Problems With It

Probability revolves around uncertainties. Coin toss, stock market, and slot machines all involve some uncertainty.

Because most things in our lives involve some certainty, a subdomain is born.

“Predictive analytics” analyzes data to decide the likeliness of an event.

However, just because something is highly unlikely, it doesn’t mean impossible– which leads me to my next point.

Many problems can occur if we forget **.000001% ≠impossible**.

This axiom is one of the reasons why models like VaR (Value at Risk) fall short sometimes.

1) The VaR saw the world as black and white and tried to predict if a stock would make or lose money, and it often disregarded “tail risk” — the potential chance of wild catastrophes.
2) It also didn’t consider that the financial markets are not predictable because uncertainties can and will happen.
    a) It can be argued that the false and confident precision of VaR added fuel to the fire of the 2008 housing market collapse.

Data Importance

Garbage in, garbage out :)

- Sample size large enough to represent the target population
- Randomly samples individuals of the population.

The Central Limit Theorem and Its Application in the Day-to-day Life

This theorem helps researchers pool surprisingly small amounts of data and still draw powerful conclusions and inferences about the broad population.

This is a really good video for the younger people reading:
https://www.youtube.com/watch?v=zeJD6dqJ5lo

Inference

Statistics don't prove anything with certainty. Analysts make estimates by applying common sense, good data and sound methodology. Experts use "statistical inference" to accept or reject a hypothesis. Inference measures the likelihood that something is true but can't be proven directly. Statistics can point a researcher in the right direction, but only testing and observation prove a hypothesis.

Polling

Polls – statistical analysis performed on survey data – can be important, valuable and potentially misleading. Capturing every response is impossible, so view poll results in context, like all statistical data.

Regression Analysis

Regression analysis is powerful when used properly. It can quantify the relationships between variables and outcomes. It can help identify and evaluate meaningful patterns, but only if the analyst uses the right questions and framework and determines the correct variables.

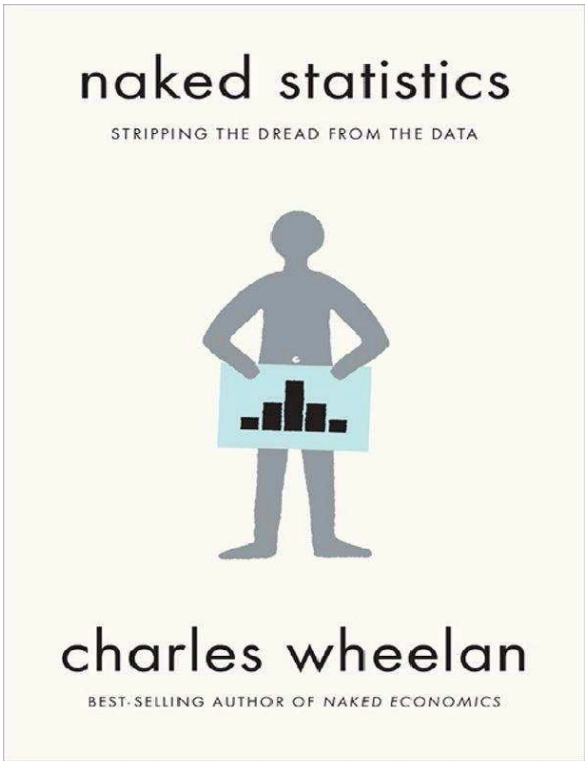
7 pitfalls of Regression analysis:

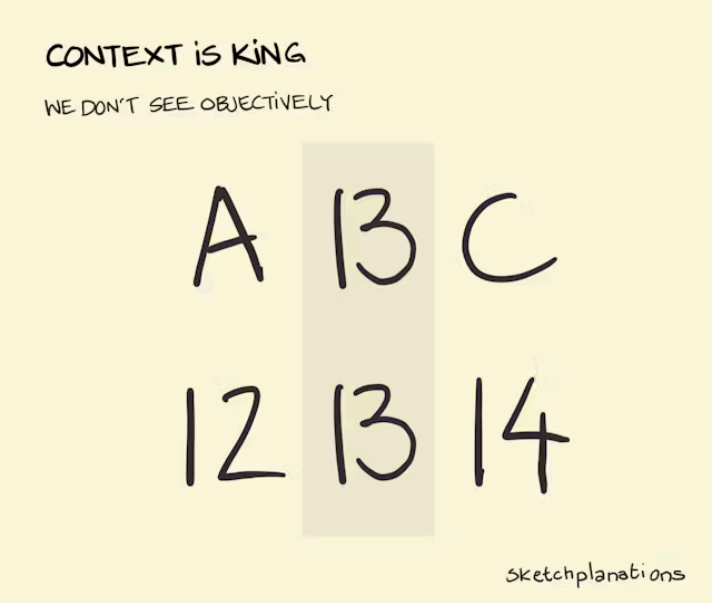
1) Using regression to analyse non-linear relationships (without special tweaks and adaptation)
2) Correlation ≠ causation (as always)
3) Reverse causality
    a) An association between A and B doesn't necessarily prove that A causes B, but it could be B causes A
4) Omitted variable bias
5) Highly correlated explanatory variables
    a) Just because 2 variables are strongly correlated, it doesn't mean that statistics can uncover the nature of their relationship with this information alone.
6) (Improperly) Extrapolating the data.
7) Data mining (too many variables)
    a) Could distort the outcomes.

Final Remarks

As I've pointed out numerous times in this documentation, context, induction, and explanations are crucial in statistical analysis. Data alone doesn't mean anything. Instead, it's how we interpret it that gives it some meaning.

Resources





sketchplanations.com (for the images)