# Understanding Statistics in the News

Kimball Martin

The University of Oklahoma
Norman, Oklahoma

July 2021

# I. Truth and logic

**(Logical) statements:** declarative sentences which are objectively true or objectively false

Which of these are statements?

1. It's hot today.
2. It is freezing now.
3. There is a 20% chance of rain today.
4. The weather forecast said there is a 20% chance of rain today in Norman.
5. Is today Wednesday?
6. *Gone with the wind* is the most popular movie of all time.
7. My name is Mike.
8. My favorite food is pizza.
9. $x = 7$.

## More on Statements

- statements can be *relative* (e.g., contain "it", "here" or "now") or contain *variables*
- all terms in statements must be *well-defined*
- often terms are defined *implicitly*
- statements, both true and false, may be *verifiable* or *unverifiable* (e.g., "The butler did it.")

— In practice, statements require interpretation and for accurate communication, we need to use a common (shared) interpretation.

— Many statements are implicitly approximations (e.g., "my dog weights 50 lbs") or opinions

— Precisely defining things is often hard

— It's often useful to work with concepts that are not well-defined (hot/cold, good/evil, ...)

— Statistics and statements in the news/media/etc are often misinterpreted or false

# Understanding truths

1. **Concrete truth** (facts)

Reality $\rightarrow$ direct observer $\rightarrow \cdots \rightarrow$ reporter $\rightarrow \cdots \rightarrow$ you

- Many opportunities for errors in this "telephone game"
- *Daily Mirror* reports AC Omonia's fans wear hats made from shoes based on Wiki page[1]

2. **Abstract truth** (ideas)

- $1 + 2 = 3$
- "that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness"

Both "formal" and "informal" abstract truths come from observation + logical thinking/arguments

[1] Source: *Virtual Unreality*, Charles Seife

# Thinking logically

A good argument is like a good recipe or directions

Tips from D.Q. McInerny's *Being Logical*, Chapter I:

- Be attentive
- Get the facts straight
  Basic types of facts: things (White House) and events (Lincoln's asassination)
- Ideas: objects and origins (cats, centaurs)
- Match ideas to facts (don't just live in ideas)
- Match words to ideas
- Effective communication
  — make definitions/assumptions clear (be on same page)
  — clearly connect facts and ideas
- Avoid ambiguous or evasive language
  — vague words: good, evil, love, democracy, fair, ...

# In practice...

- There's neither time nor resources to verify everything yourself, so we need to rely on others
- The real world is complicated—there's much that we can't know for sure, and we have to take *probabilistic* and *heuristic* approaches to understanding the world
- Even verified truths often don't fit into quick soundbites, so are not easy to communicate accurately
- Two logical people with access to the same information may arrive at very different conclusions
- Quality sources and expertise matters, but they aren't infallible

— Distinguish what you know to be true from what you think is true (biases)
— Prioritize what's important to verify for yourself
— Find reliable sources, but understand their limits, and learn how to corroborate evidence

# Logic quiz

Analyze the following:[2]

1. Every dog has 3 heads. Collies are dogs. Therefore, collies have 3 heads.
2. Pierre Poseur was an All-American football player. Pierre Poseur earned his first million before he was 30. Pierre Poseur is handsome and has a winning smile. Therefore, Pierre Poseur should be elected governor.
3. Every squirrel is a mammal. Every chipmunk is a mammal. Therefore, every chipmunk is a squirrel.
4. No men are daughters. No waitresses are men. Therefore, no waitresses are daughters.
5. No Oklahomans are Californians. Every Normanite is an Oklahoman. Therefore, no Normanites are Californians.

[2] Adapted from *Being Logical*, D.Q. McInerny

# II. Evaluating information

1. Do sanity checks
   — consistency, internally and externally
2. Look for common fallacies/tricks
   — survivorship bias
   — misinterpreting percentages
   — correlation does not imply causation
   — etc.
3. Consider the source
   — credibility
   — biases
   — motives
4. Compare with other sources/perspectives
   — corroborating evidence
   — omitted crucial points

# Sanity checks

- Internal consistency: do the claims made support each other or contradict each other? or are they completely independent of each other?
  — 100% of adults will be obese by 2048. 100% of black men will be obese by 2095. (*How not to be wrong*, Chapter 3: Everyone is obese)
- Do different approaches lead to the same conclusions?
  — 200 dead in Madrid = 1300 dead Americans = 600 dead New Yorkers (*How not to be wrong*, Chapter 4: How much is that in dead Americans?)

# Basic misconceptions

- Not all relations are linear (*How not to be wrong*, Chapter 1: Less like Sweden)
- Looking at proportions but ignoring sample sizes (Law of Strong Numbers)
  — brain cancer is 2–3x more likely in South Dakota vs North Dakota (*How not to be wrong*, Chapter 4: How much is that in dead Americans?)
- Misinterpreting percentages/sample populations
  — efficacy rates of COVID-19 vaccines:
  Pfizer 91%, Moderna 94%, Johnson & Johnson 72%
- Improbable things happen a lot, and don't indicate significance (*How not to be wrong*, Chapter 6: Baltimore stock broker and the bible code)

# Significance

**(Statistical) significance:** detect *some* difference between experiment and control (thresholds depend on experiment)

Improbable $\neq$ significant

- License plates (AJS 317)
- Baltimore stockbroker (*How not to be wrong*, Chapter 6)
- Bible code (*How not to be wrong*, Chapter 6)

Signficance $\neq$ risk ratio $\neq$ importance

- Significance does not mean there is a strong effect
- Just reporting risk ratios may be misleading

— 1995: certain brands of oral contraceptives double risk of veinous thrombosis (*How not to be wrong*, Chapter 7)
— infant fatality rate for in-home day care is 7x that for day care centers (*How not to be wrong*, Chapter 7)

# Correlation

**(Statistical) correlation:** a measure of how similarly 2 quantities vary with each other, or with respect to time

- Positively correlated: age and height in children, COVID hospitalizations and COVID deaths (with time delay)
- No correlation: completely independent
- Negatively correlated: HDL (good) cholesterol levels and heart attack rates

Correlation does not imply causation

- Correlations can happen by chance (improbable things happen a lot, "spurious correlations", bible code, ...)
- Correlations may be indirectly related (same root cause, e.g., ice cream sales and drowning deaths)
- Even if there is a causal relationship, statistics don't tell you which way the cause goes (e.g., violence and TV watching in children)

Reference: https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

# Hypothesis testing

1. Clearly formulate hypothesis and design *before* experiment (avoid reading into low probability outcomes)
2. As much as possible, test with a large, random (unbiased) sample
3. Account for all experiments (avoid cherry picking, data-dredging—insert xkcd comic)
4. Test for effect as well as significance (and report confidence intervals not just $p$-values)

— Report results and analysis as honestly and accurately as possible
— Experiments should be repeated by other researchers
(the scientific method is progressive, not absolute)