How to Lie with Statistics

or maybe... not get lied to...

Grinnell College

Spring 2025

Review

We have done **a lot** with statistics in only a few months. Let us very briefly review some of the things we have seen.

What is Statistics?

Statistics is the science and art of collecting and using data to learn about things

Statistics is about **variation**

- world is full of data
- these data exhibit variation (they aren't all the same)
- noticing, displaying, and quantifying this variation helps us learn
- end goal is to explain variation (why are things different?)

Why do we need statistics?

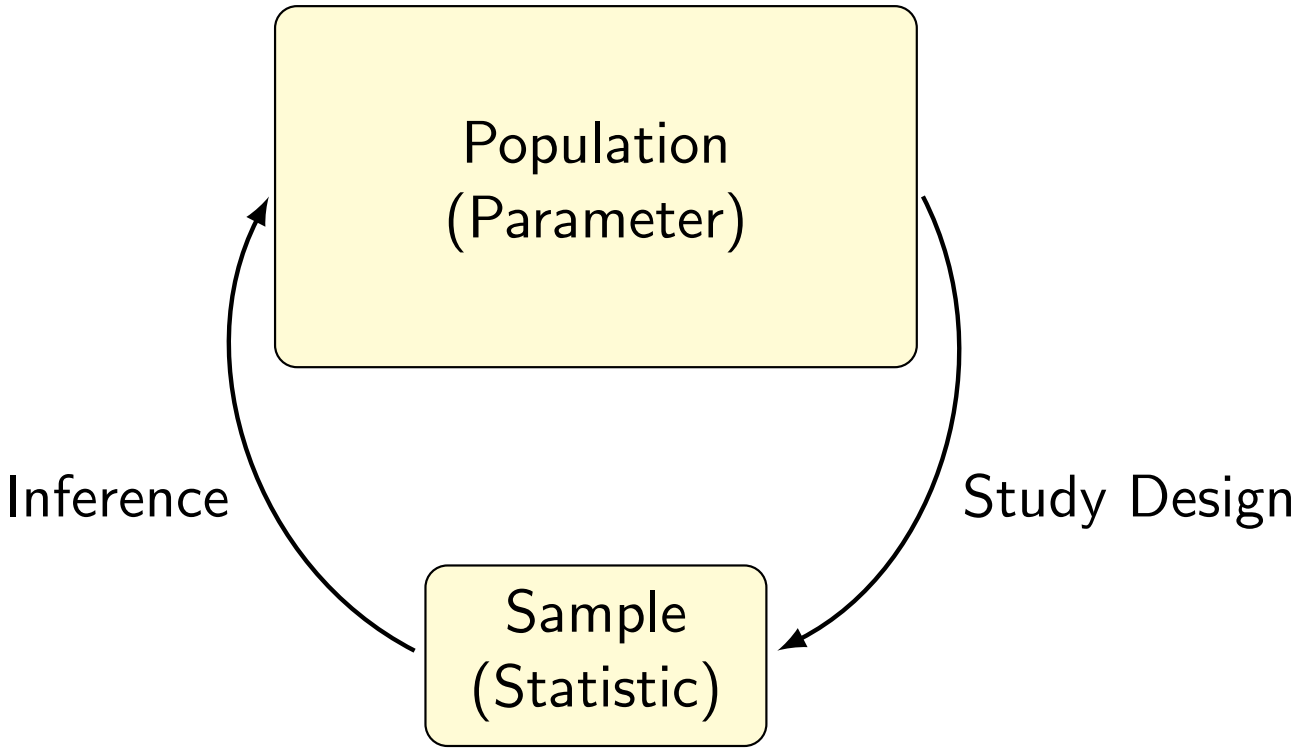
Human beings are great at identifying patterns

- Cognitive biases
- Poor intuition of uncertainty and randomness

Statistics gives us a framework for answering questions about the world using data (scientific method)

1. Construct a hypothesis
2. Collect data
3. Consider evidence
4. Draw conclusions

The Statistical Framework



We are often trying to learn about a population using limited information

- how data is collected affects the conclusions we can draw
- will our estimates be perfect? No! Quantify error due to randomness

Statistics: Science or Art?

In the beginning of this course I said Statistics was both a science *and* an art of data analysis. What do I mean by this?

Science

- Methodology
- Math Theory
- Simulations
- Predictions

Art

- Data Analysis Choice
- Communication
- Decision Making

How to Lie with Statistics

The majority of the remainder of these slides are adaptations of examples taken from the book "How to Lie with Statistics" by Darrell Huff and published in 1954.

The Well-Chosen Average

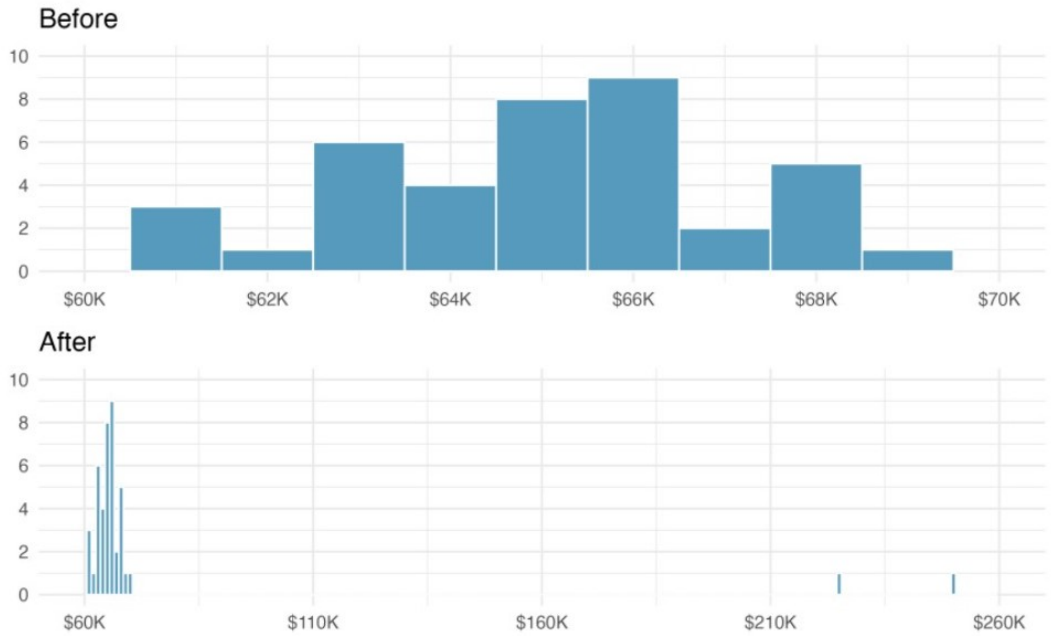
We have spent a lot of time working with averages and medians. Averages are not always good measures to use to describe a 'typical value' for our population!

- skews and outliers can make the mean really big or small, so that it is not a good measure of 'typical'
- mean only really works with "nice data"
- if you cannot see the distribution, which should you trust more: mean or median?

NOTE: We have used means over and over in CIs and HTs. Ask yourself when performing a test, "Does the mean make sense, or would median be better?"

The Well-Chosen Average

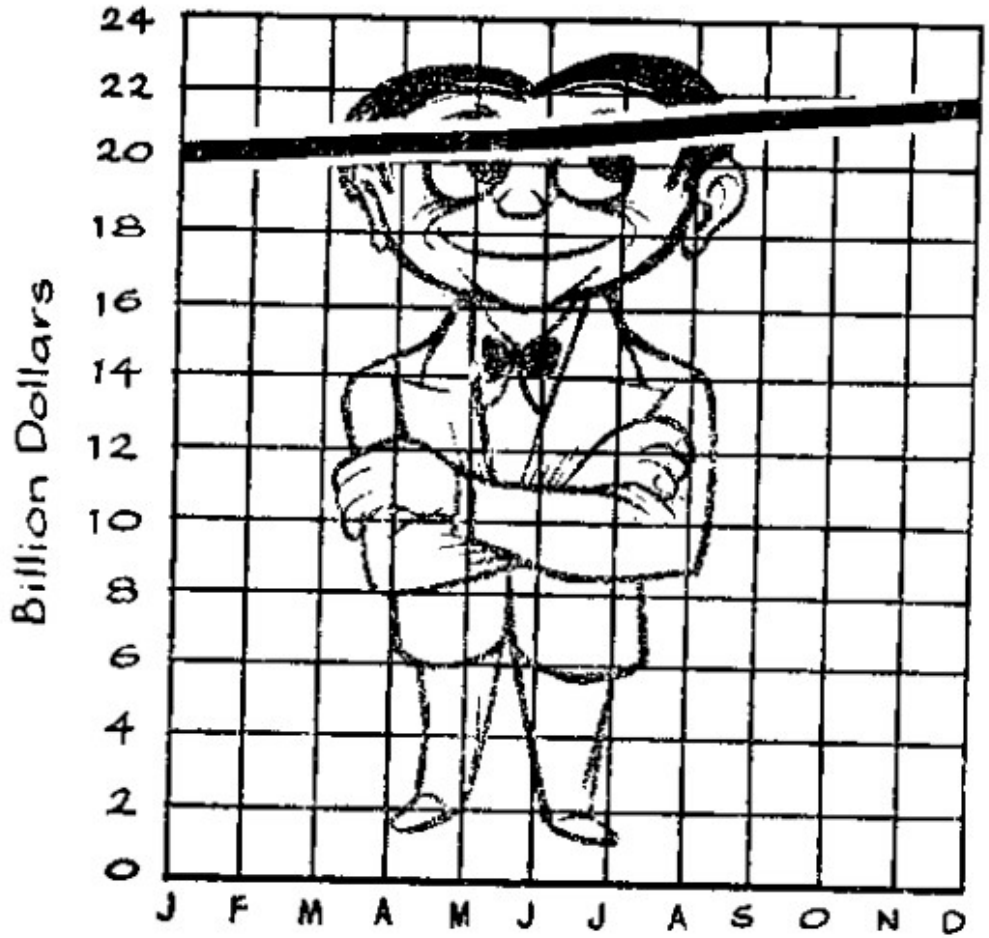
Income data for a sample of customers for a coffee shop



	n	Min	Q1	Median	Mean	Max	SD
Before	40	$60,679	$60,818	$65,238	$65,089	$69,885	$2,122
After	42	$60,679	$60,838	$65,352	$73,299	$250,000	$37,321

Is the mean still a good measure of 'typical value' for incomes?

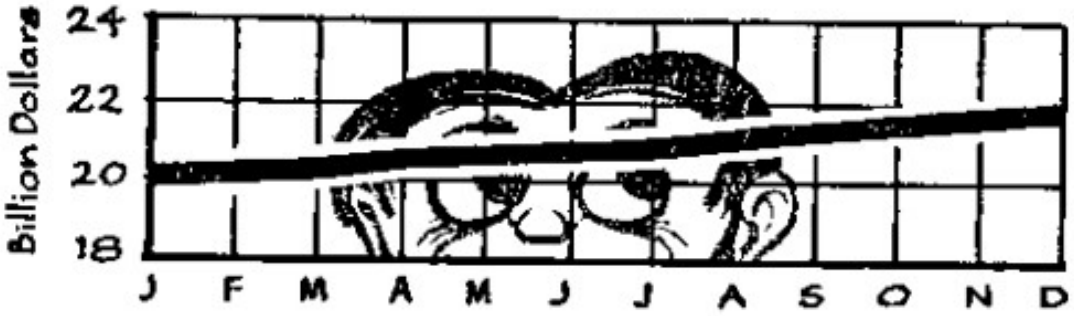
The Gee-Whiz Graph



A decent graph. We get an idea of how big a 10% increase in the variable is. BORING! I want to add some schmaltz to this thing

The Gee-Whiz Graph

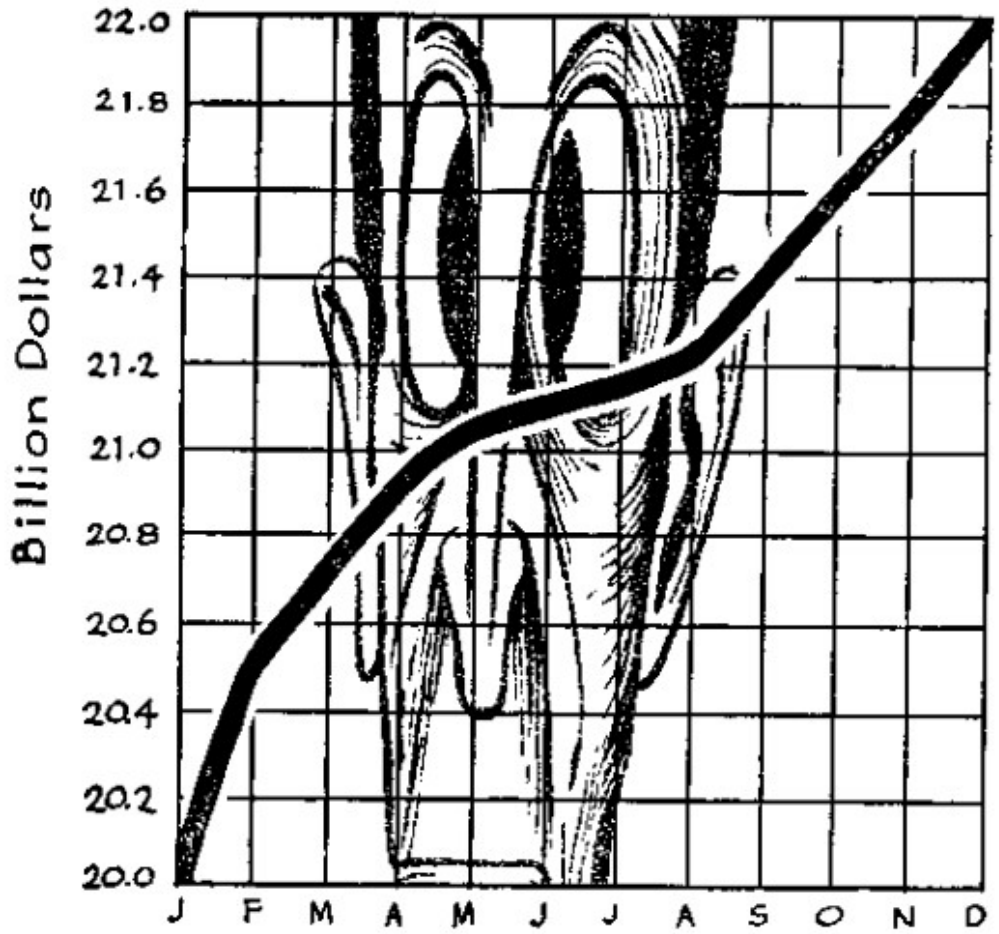
Chicanery: use of trickery to achieve a political, financial, or legal purpose



By chopping off the bottom of the y-axis I can make the increase look like a much bigger deal than it actually is. Still kind of boring though. I can make it better.

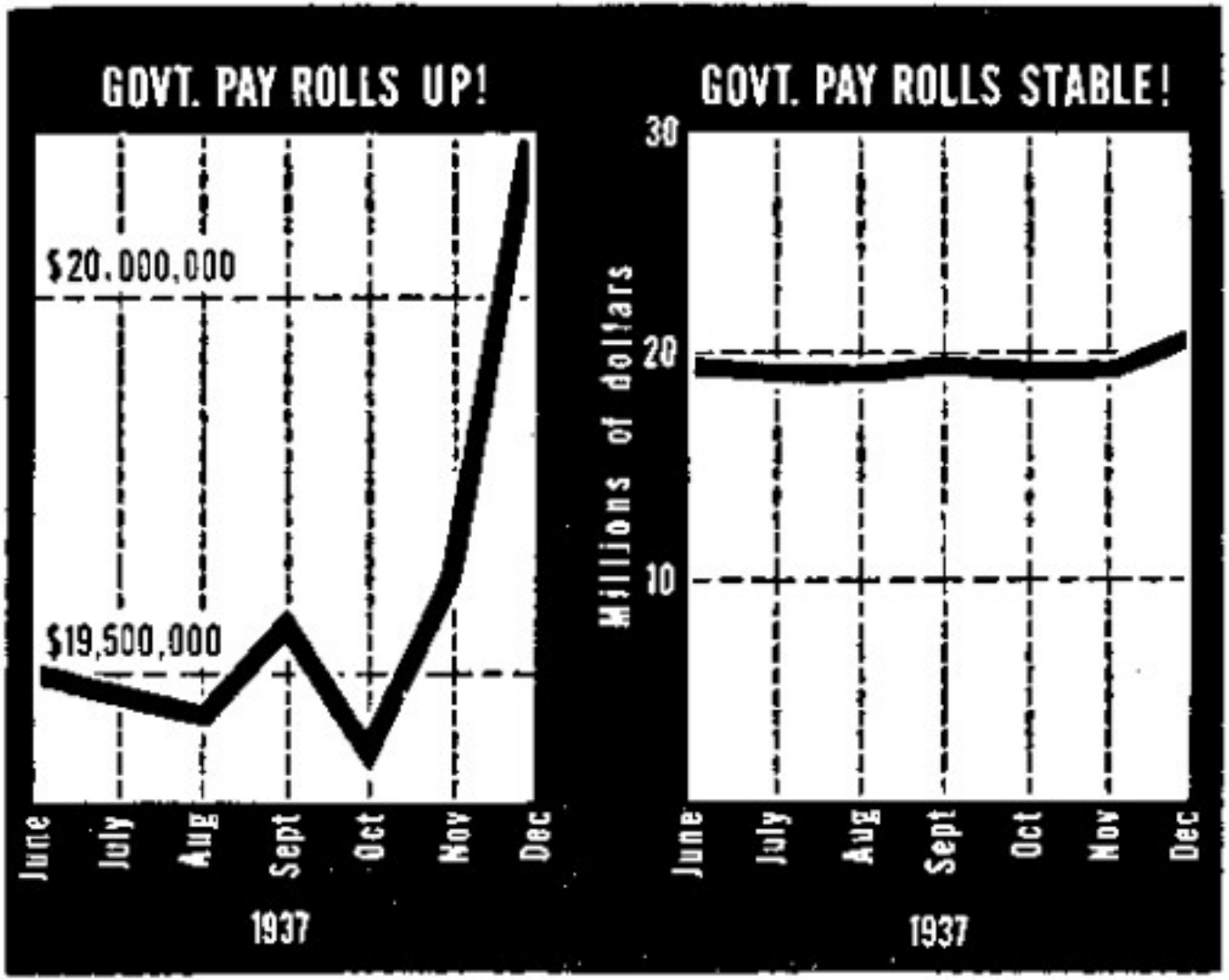
The Gee-Whiz Graph

Chicanery: use of trickery to achieve a political, financial, or legal purpose

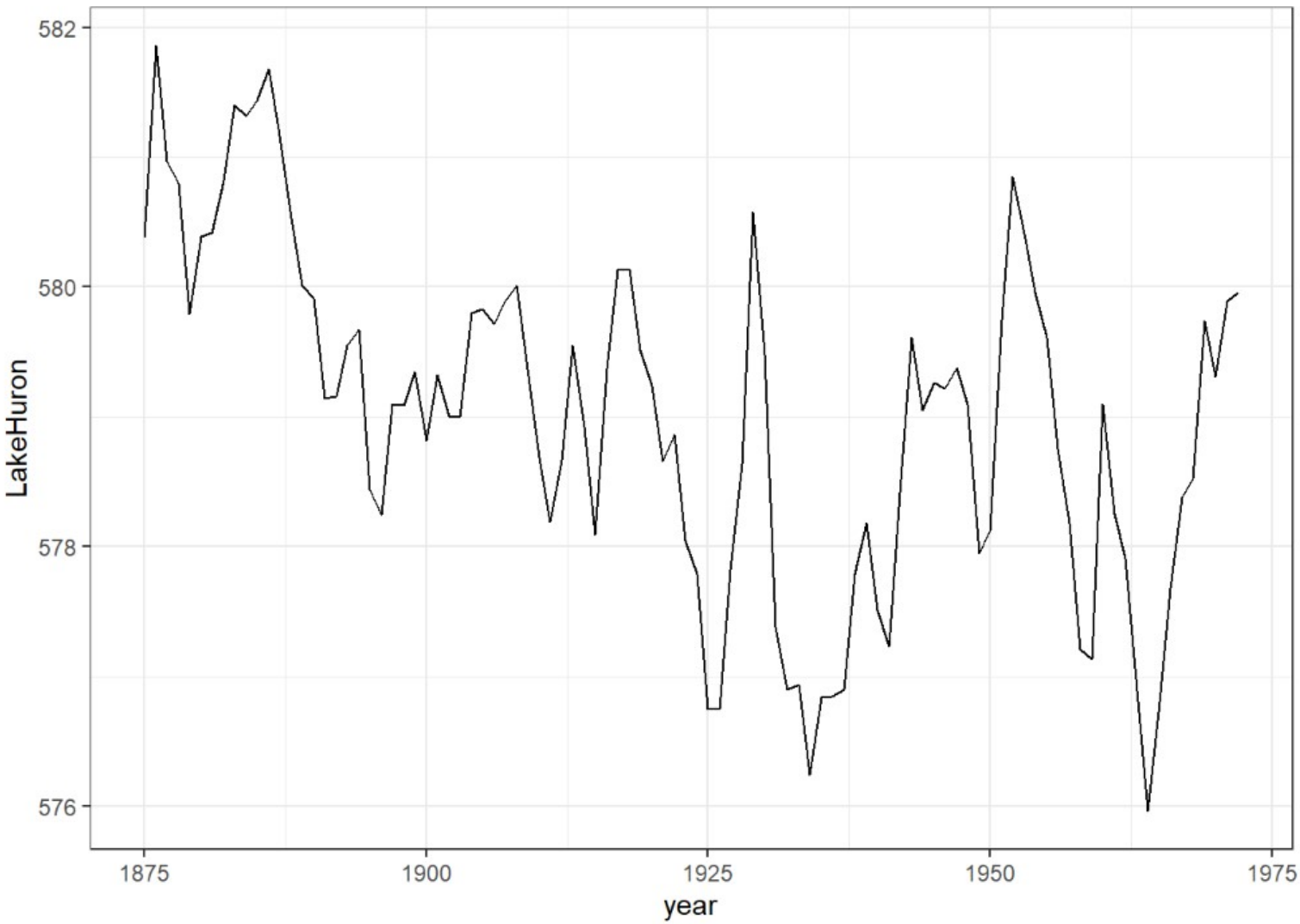


Stretching the y-axis a bit really makes that increase look huge!

The Gee-Whiz Graph

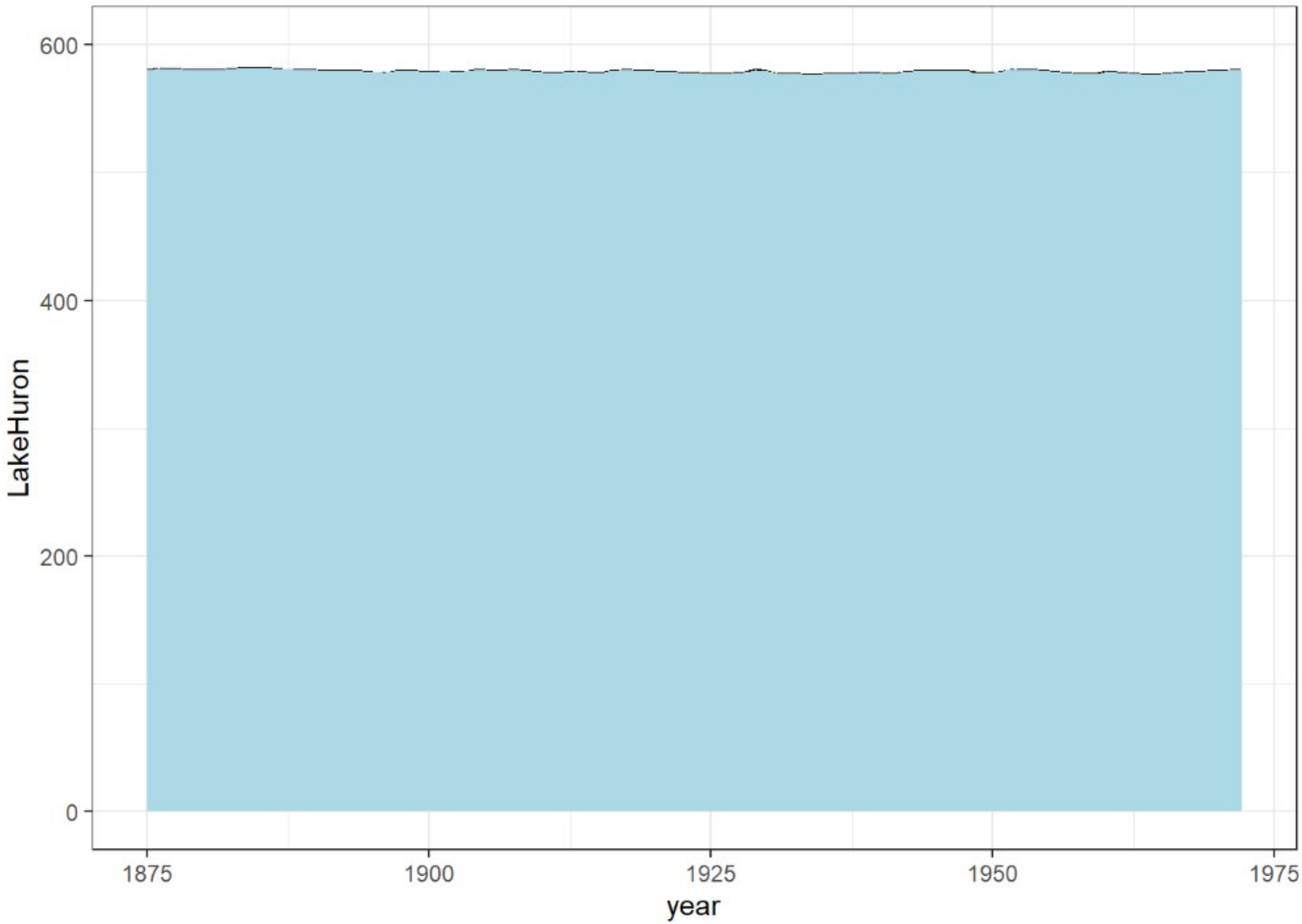


Lake Huron Example



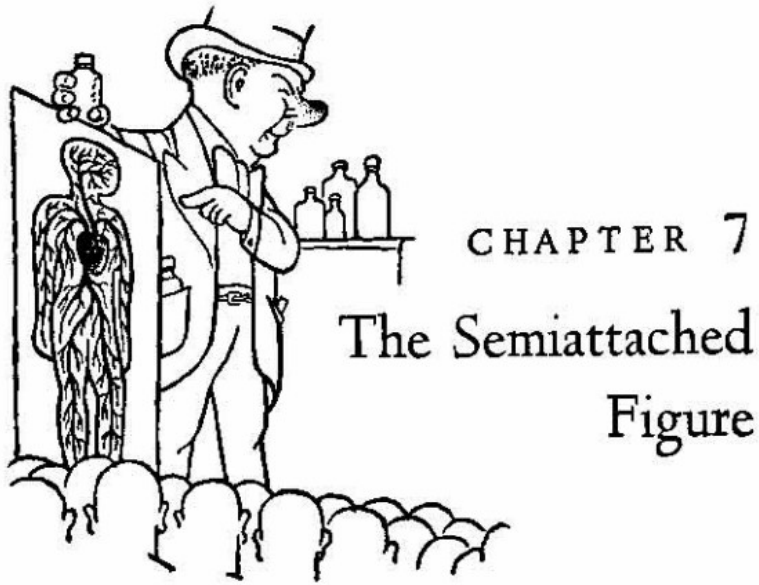
"Lake Huron levels trending down, water in short supply!"

Lake Huron Example



"Water levels stable, no need for alarm!"

The Semi-attached Figure



"If you can't prove what you want to prove, demonstrate something else and pretend they are the same thing."

A semi-attached figure is a piece of evidence presented as proof of something, even if there really isn't a connection between them at all.

The Semi-attached Figure

"27% of a large sample of eminent physicians smoke Throaties – more than any other brand."

"The only answer to a figure so irrelevant is 'So what?' With all proper respect toward the medical profession, do doctors know any more about tobacco brands than you do? [...] Of course they don't, and your doctor would be the first to say so. Yet the '27%' somehow manages to sound as if it meant something."

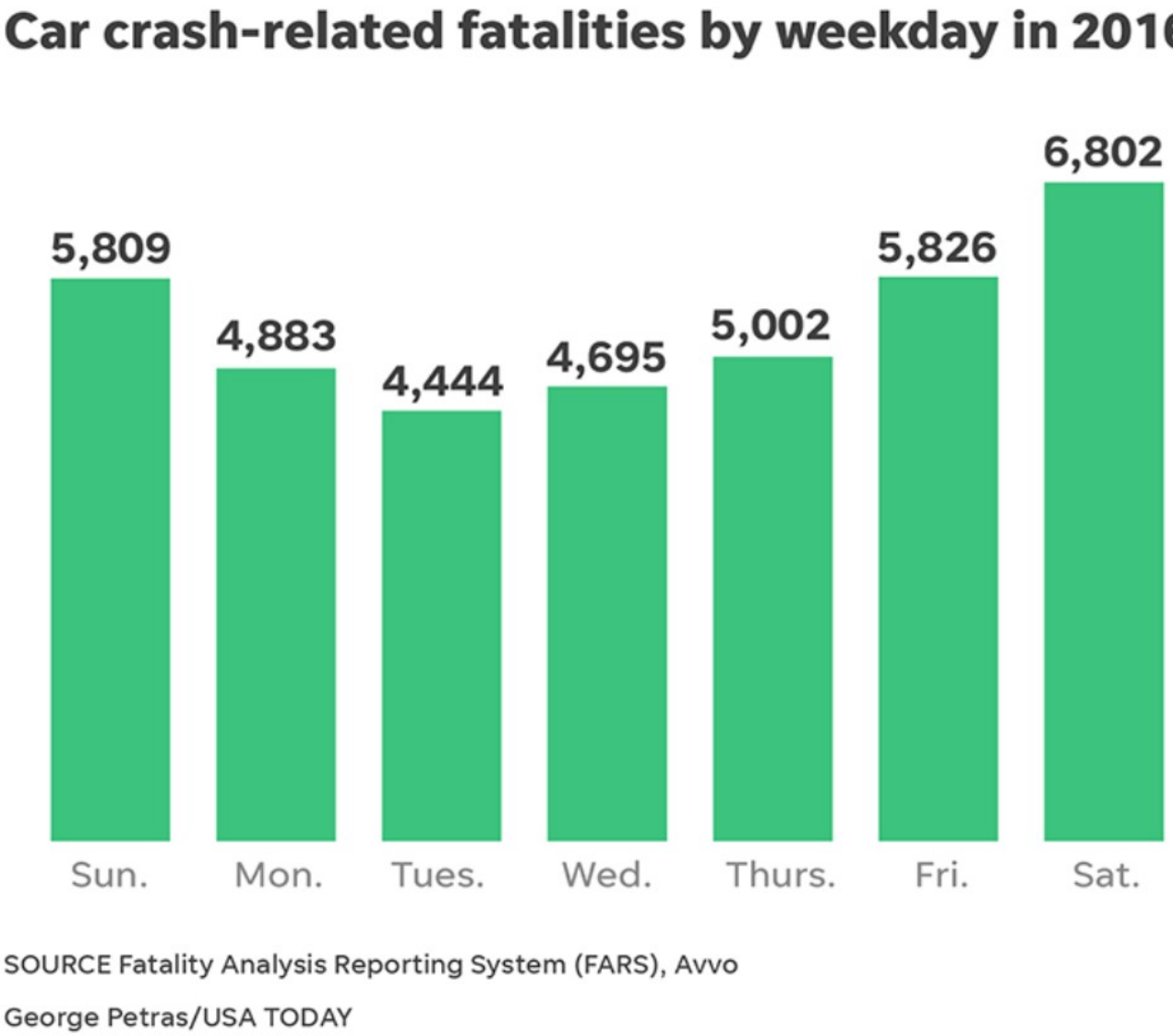
The Semi-attached Figure

"If you can't prove what you want to prove, demonstrate something else and pretend they are the same thing."

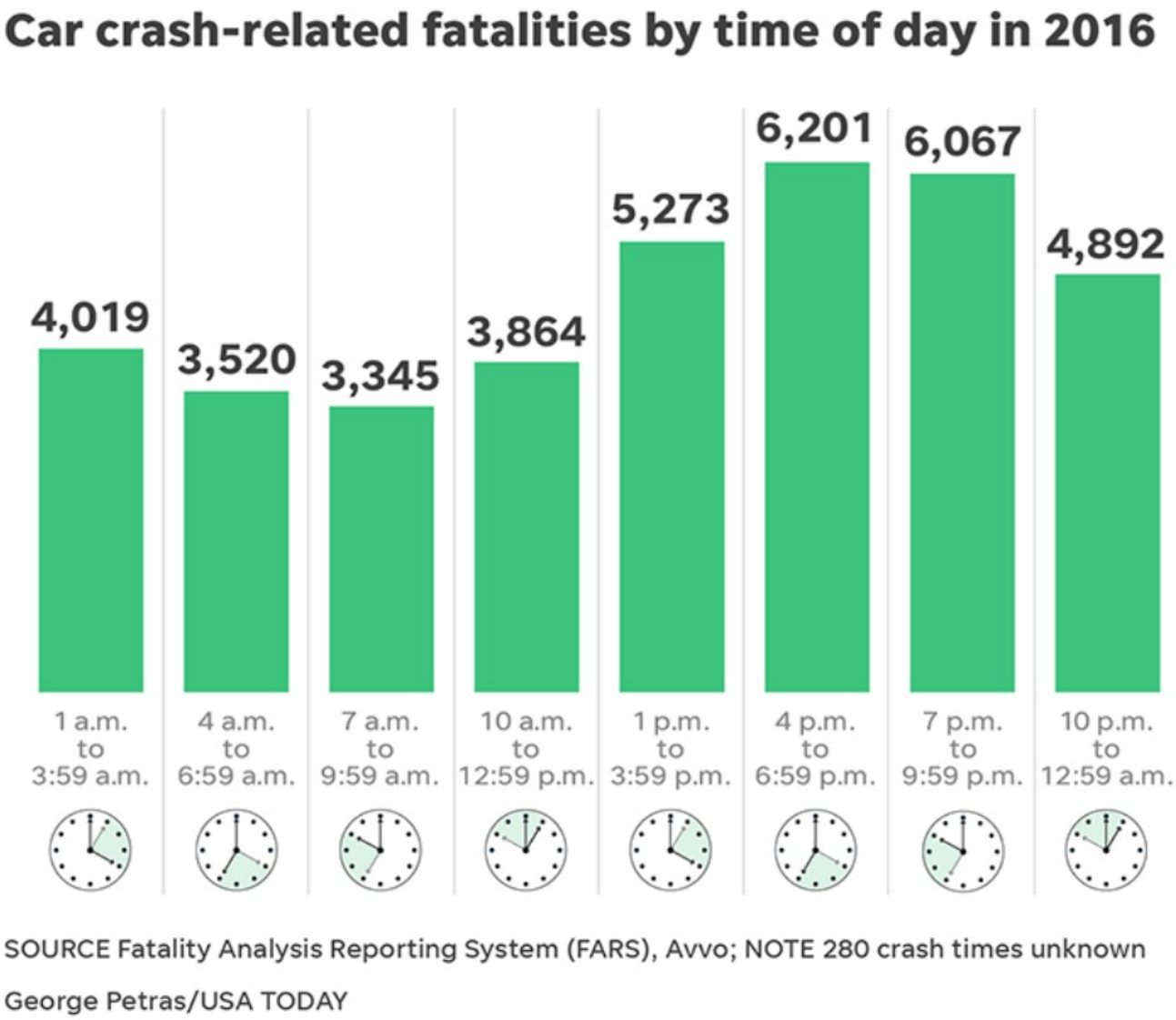
The Semi-attached Figure

Source: https://www.newscentermaine.com/article/news/nation-world/saturday-is-most-dangerous-day-of-the-week-to-drive-new-study-says/507-558703422

The Semi-attached Figure



The Semi-attached Figure



Conditional Probabilities

One thing touched in here in the last few examples is that the type of probability used *really* matters as to what you can do with it. When we want to compare rates that things happen we NEED to use conditional probabilities.

P(a car crash is on Saturday) is not the same thing as P(a car crashing *given* that it is Saturday)! We have not adjusted for how many cars are out and about on Saturday! In fact we are actually using the wrong direction of conditions here.

$18\% = \text{P(Saturday } \textit{given} \text{ car fatality)} \neq \text{P(fatality } \textit{given} \text{ Saturday)}.$

Semi-attached Figure – Navy

"The death rate in the Navy during the Spanish-American War (1898) was 9/1000. For civilians in New York City during the same time period the death rate was 16/1000. Navy recruiters later used these figures to show that it was safer to be in the Navy than out of it."

Issues:

- Groups are not comparable, navy is mainly young healthy males, NYC includes infants, old, and ill
- Need to explicitly compare outcomes for young healthy males (obviously will find being in a war zone is more dangerous than not)
- "These figures do not at all prove that men meeting Navy standards will live longer in the Navy than out." Recruiters' conclusion is completely worthless.

Post Hoc Rides Again

"Somebody once went to a good deal of trouble to find out if cigarette smokers make lower college grades than non-smokers. It turned out that they did. This pleased a good many people and they have been making much of it ever since. The road to good grades, it would appear, lies in giving up smoking; and to carry the conclusion one reasonable step further, smoking makes dull minds.

"This particular study was, I believe, properly done; sample big enough and honestly and carefully chosen, correlation having a high significance, and so on.

"The fallacy is an ancient one which, however, has a powerful tendency to crop up in statistical material [...]. It is the one that says that if B follows A, then A has caused B."

Correlation $\neq$ Causation

We can have a large correlation value between 2 variables. This does not mean one variable is causing a change in another.

"An unwarranted assumption is being made that since smoking and low grades go together, smoking causes low grades. Couldn't it just as well be the other way around? Perhaps low marks drive students not to drink but to tobacco. When it comes right down to it, this conclusion is about as likely as the other and just as well support by the evidence."

Lurking Variables

Lurking Variable: a third variable that explains the relationship between two variables with high correlation

"It seems a good deal more probably however, that neither of these things has produced the other, but both are a product of some third factor. Can it be that the sociable sort of fellow who takes his books less than seriously is also likely to smoke more?"

Statisticulate

"Misinforming people by the use of statistical material might be called statistictal manipulation; in a word (though not a very good one, statisticulation."

"The title of this book and some of the things in it might seem to imply that all such operations are the product of intent to deceive. [...] Possibly more important to keep in mind is that the distortion of statistical data and its manipulation to an end are not always the work of professional statisticians. What comes full of virtue from the statistician's desk may find itself twisted, exaggerated, oversimplified, and distorted-through-selection by salesman, public-relations expert, journalist, or advertising copywriter."

How to Talk Back to a Statistic

Ok. I spent a lot of time just yelling things at you about ways in which people can mislead us (whether due to intention or incompetence).

"I'll face up to the serious purpose that I like to think lurks just beneath the surface of this book: explaining how to look a phony statistic in the eye and face it down; and no less important, how to recognize sound and usable data in that wilderness of fraud to which the previous chapters have been largely devoted."

Five simple questions:

1. Who says so?
2. How do they know?
3. What's missing?
4. Did somebody change the subject?
5. Does it make sense?

Source:

"How to Lie with Statistics" – Darrell Huff (1954)

PDF link (https://www.horace.org/blog/wp-content/uploads/2012/05/How-to-Lie-With-Statistics-1954-Huff.pdf)