432 Class 27

https://thomaselove.github.io/432-2025

2025-04-24

## Agenda

* Building and Assessing Statistical Work in 2024
* Maintaining and Enhancing What You’ve Learned
  + [Statistical Problems to Document and To Avoid](https://discourse.datamethods.org/t/author-checklist/3407)
  + [A CHecklist for statistical Assessment of Medical Papers (the CHAMP statement): explanation and elaboration](https://bjsm.bmj.com/content/55/18/1009.2)
  + [Biostatistical Modeling Plan](https://www.fharrell.com/post/modplan) from Frank Harrell.
* and some surprises…

# On Fitting Regression Models

## On Fitting Regression Models

Some things are always true…

* Being honest about your findings is important.
* Identifying a stable phenomenon to study is crucial.
* Measuring that phenomenon well is crucial.
* Being transparent about your work is important. Describing your work so it can be replicated, and then actually replicating it, are good things. Sharing your data and your materials is a good idea.

## On Fitting Regression Models

Some things are always true…

* Having a large sample size (*n*) is helpful in fitting models.

but the impact of lots of other things changes depending on why you’re fitting a regression model.

## The Key Point

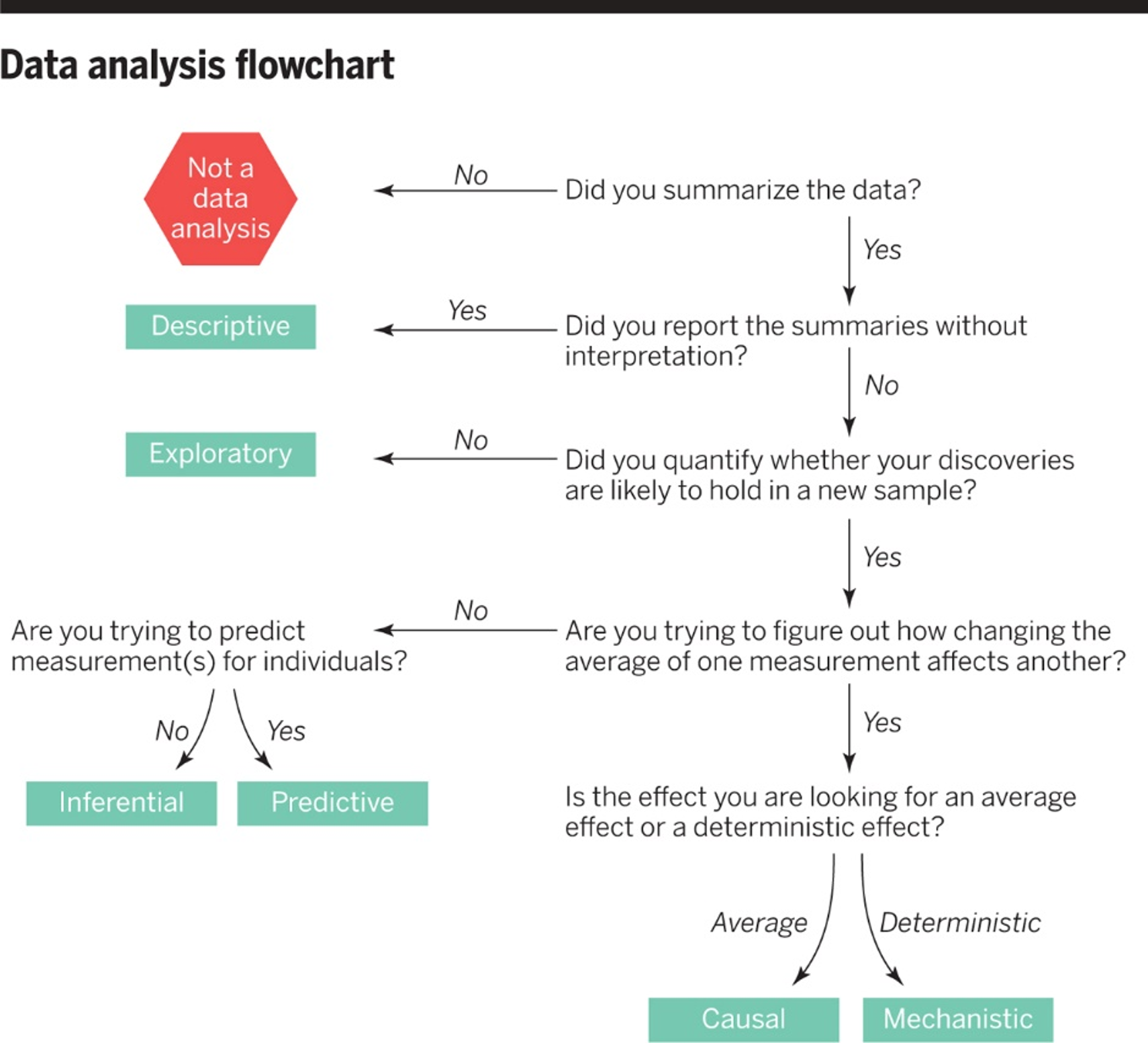
Decide what your research question is, and use it to help you think about what’s important in your modeling.

* Models that account for only a few of the possibly important dimensions of a problem don’t lead to causal conclusions, but can help screen out important from less important areas for future work.
* Model development strategies that work well with large sample sizes (n) and small numbers of predictors to consider don’t necessarily work well when the situation is reversed.

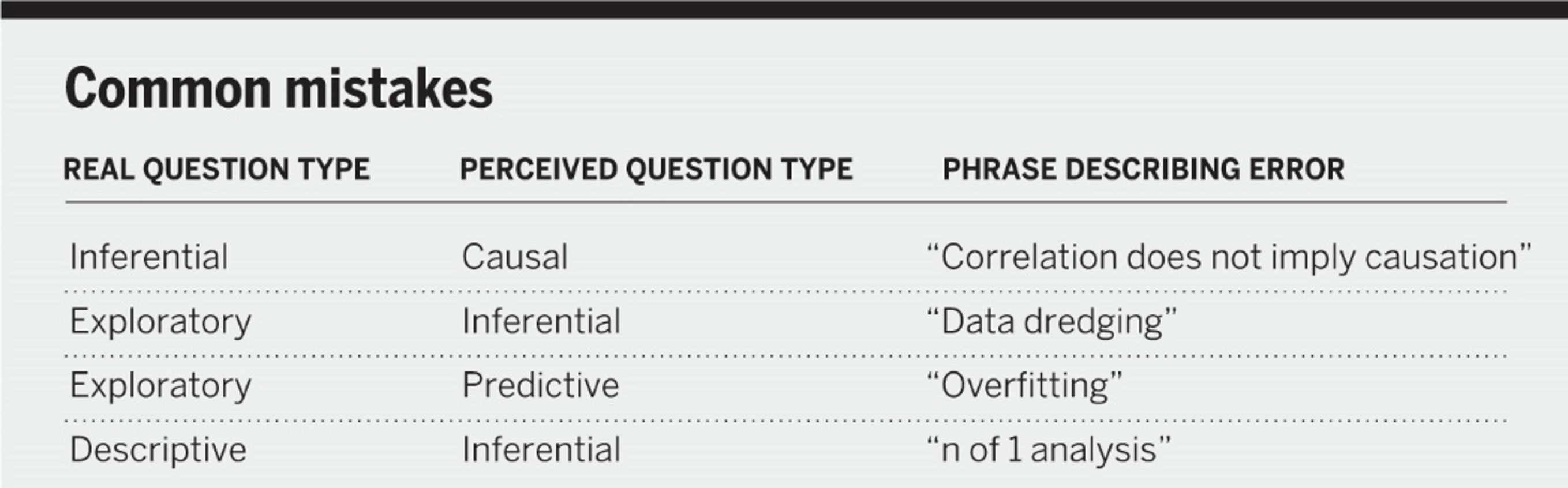
## The Key Point

Decide what your research question is, and use it to help you think about what’s important in your modeling.

* Most problems involve missing data, and problems in measurement. Getting those issues settled effectively is often overlooked.
* The sample size you need to fit a regression model changes depending on your aims.



## Leek JT and Peng RD “What is the Question” *Science* 2015-03-20



## On Fitting Models

What are some of the reasons you might fit a linear (or generalized linear) model?

1. All prediction, no explanations
2. All description, external validity is irrelevant
3. Clinical/Scientific Prediction with strong priors
4. Above all else, simplicity
5. Causal inference
6. Risk Adjustment

## On Fitting Models

What are some of the reasons you might fit a linear (or generalized linear) model?

1. (All prediction, no explanations) Because you want to make predictions about an outcome in new data based on some training data you have, but you’re happy to take those predictions as emerging from a mysterious magic “black box” that cannot be peered into without spoiling the surprise. You don’t care if your results are a little biased, so long as the predictions are strong. Parsimony doesn’t matter to you.

## On Fitting Models

1. (All prediction, no explanations)

* Some especially useful tools here include: variable (feature) selection through cross-validation, stepwise approaches, AIC and BIC, machine learning tools like regression trees, and other means of quickly searching through many possible models.
* Sample size is rarely a big issue here. The big problem is having more variables than you can possibly plot at once. You usually have enough data to partition into separate development and test samples.

## On Fitting Models

1. (All description, external validity is irrelevant) Because you want to describe, as accurately as possible, the nature of the associations you observe in the available data, but you don’t care much at all whether the conclusions you draw will hold up in new data.

* Confidence intervals for coefficients (slopes, mostly), and sometimes you’ll run the model on clinically relevant cutpoints rather than continuous predictors to see what’s happening more simply.

## On Fitting Models

1. (All description, external validity is irrelevant)

* Simple polynomial models can be appealing, and you’ll sometimes want to build this in the ANCOVA context, where you’re looking for the impact of specific pre-specified interactions.
* Residual plots play a big role here in deciding whether the model “fits” well enough, or identifying cases when it doesn’t.
* Cross-validation is useful, but not a big part of convincing people that the model is “right” or not.

## On Fitting Models

1. (Clinical prediction) You want to do an excellent job predicting an outcome in new data, but you have a lot of prior knowledge about the predictors under consideration, and want to use that information to help produce prediction rules as effectively as possible. You welcome the fact that most relationships are non-linear, but would like to be parsimonious if possible, as data are often expensive.

## On Fitting Models

1. (Clinical prediction)

* Some especially useful tools here include: scatterplot matrices (when the number of predictors is modest), cross-validation, assessments of discrimination and calibration, Spearman’s plot to point the way to non-linear terms that might be impactful if present, restricted cubic splines, polynomial functions, and graphical tools like nomograms
* Most stepwise tools aren’t helpful here. We try to not “peek” at the outcome-predictor relationships to maintain unbiased estimates of the relationship without extensive validation.

## On Fitting Models

1. (Above all else, simplicity) You want the problem to look like one in a statistics textbook, where everything is fit with the simplest possible model, where every term adds statistically significant predictive value, and where obtaining an unbiased estimate of the outcome is especially important. You still care a bit about what happens in new data, but you’re mostly concerned about parsimonious model development.

## On Fitting Models

1. (Above all else, simplicity)

* Some especially useful tools include stepwise and related approaches, and methods for pruning a set of predictors with clustering or principal components analysis. These models usually make the (often incorrect) assumption that relationships are linear.

## On Fitting Models

1. (Above all else, simplicity)

* Often this approach is used by people who are trying to pre-specify their entire model in advance, and want to be sure they can “explain” the result when they are done. That may not be a reasonable thing to hope for. It is important to set expectations in advance appropriately.

## On Fitting Models

1. (Causal inference) You want to identify whether a particular causal pathway you have pre-specified matches up well with what you see in new data.
2. (Risk Adjustment) You want to identify the impact of a particular exposure/predictor on an outcome, while controlling for the effects of a series of additional predictors. Perhaps you’ve done a randomized experiment / clinical trial, and want to identify whether particular results meet a standard for statistical (as well as clinical) significance. Power is very important.

## On Fitting Models

1. (Causal inference) or 6. (Risk Adjustment)

* Bias is very important, and you want to avoid it. Careful design of a comparison group (like I teach in 500) is a very good way to go about this work, but it’s also true that there’s a lot of epidemiology that goes into drawing causal conclusions, or even thinking hard about an association.
* Often the details of modeling take a back seat to the details of designing the study (and the comparison groups.)

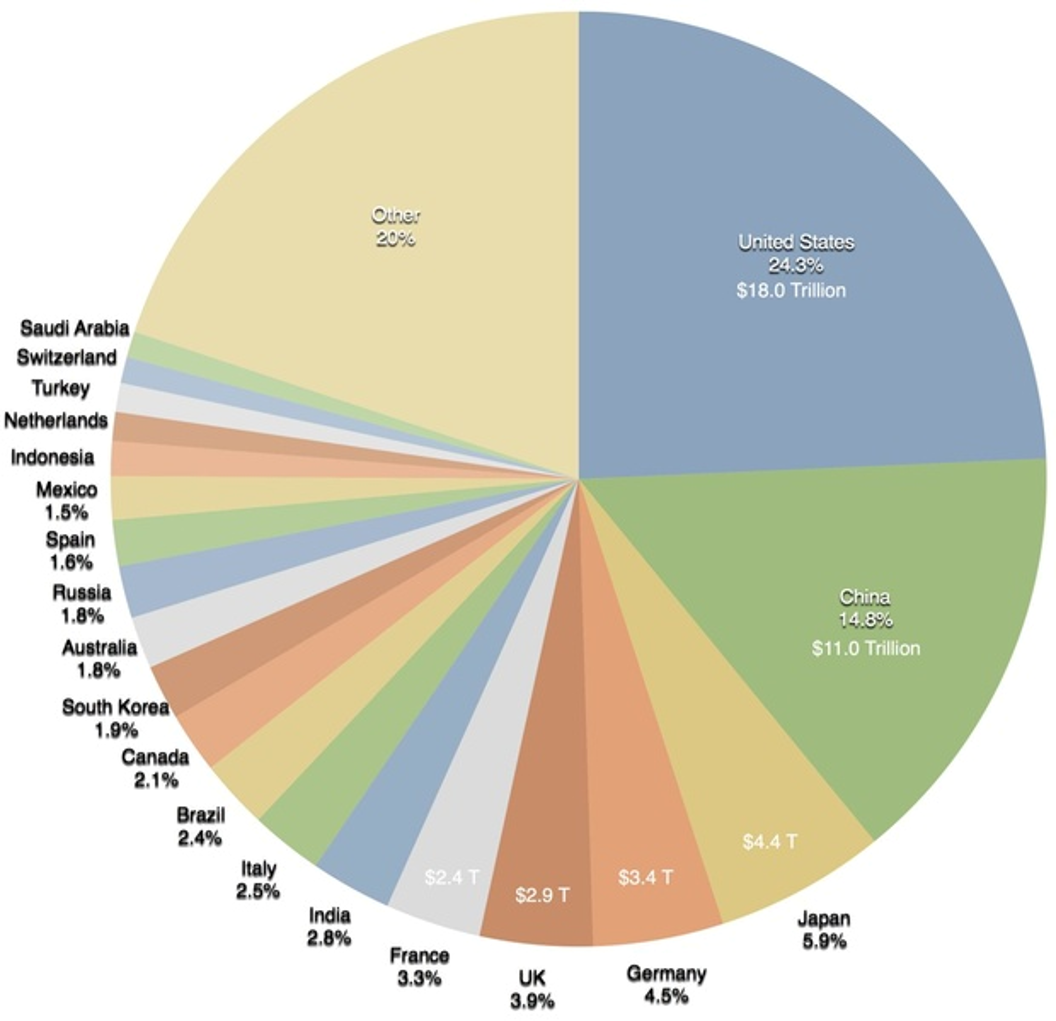
## from @EpiEllie (Dr. Ellie Murray)

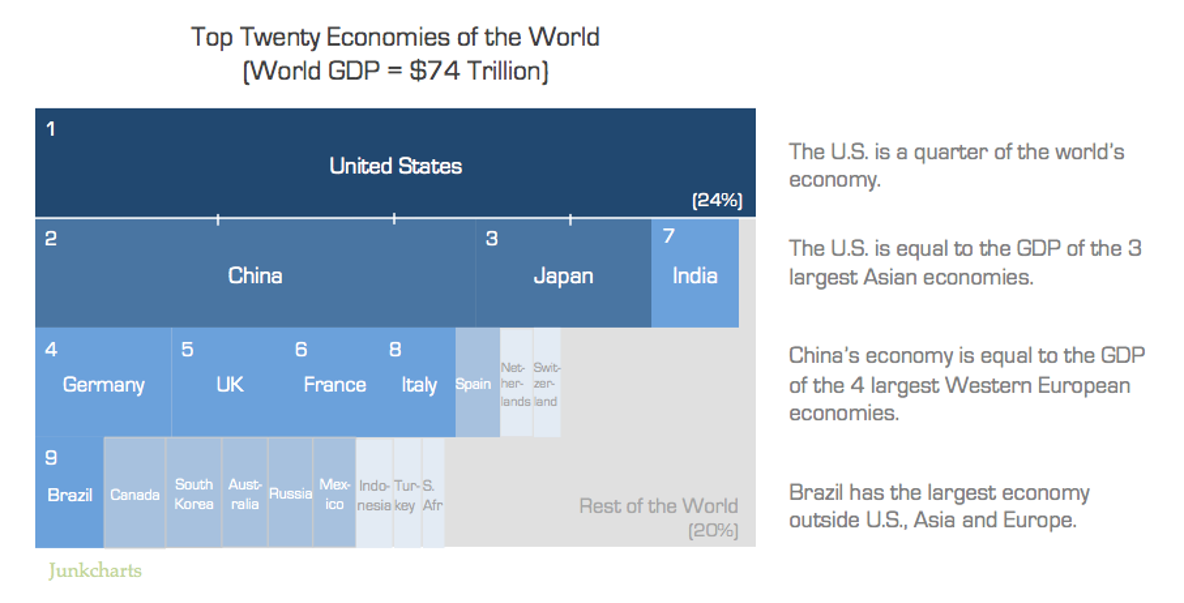


# Draw pictures, build models, and embrace uncertainty

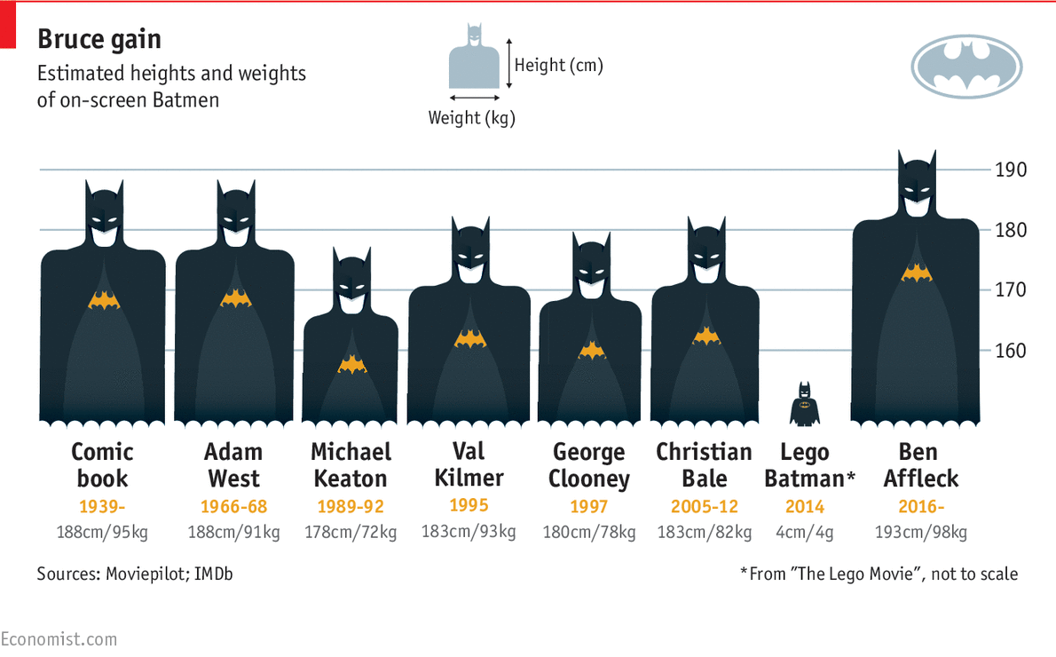
## Well, draw *good* pictures

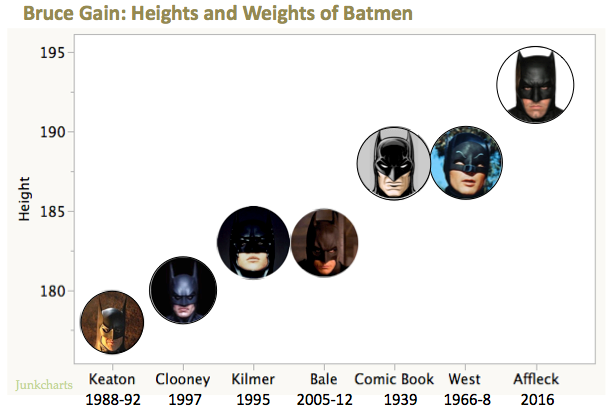












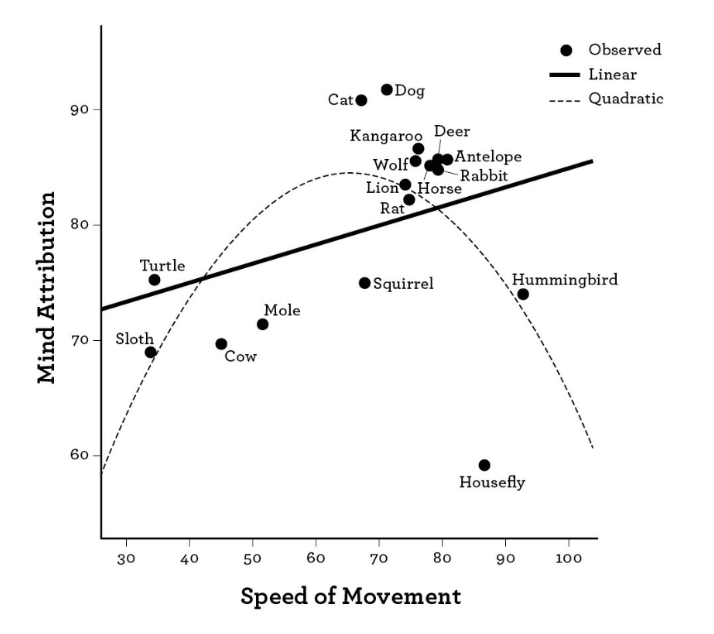
# Can we find bad data analysis out in the world?

## From Andrew Gelman 2023-04-23

[Hey - here’s some ridiculous evolutionary psychology for you, along with some really bad data analysis.](https://statmodeling.stat.columbia.edu/2023/04/23/hey-heres-some-ridiculous-evolutionary-psychology-for-you-along-with-some-really-bad-data-analysis/)

Jonathan Falk writes:

So I just started reading The Mind Club, which came to me highly recommended. I’m only in chapter 2. But look at this graph…

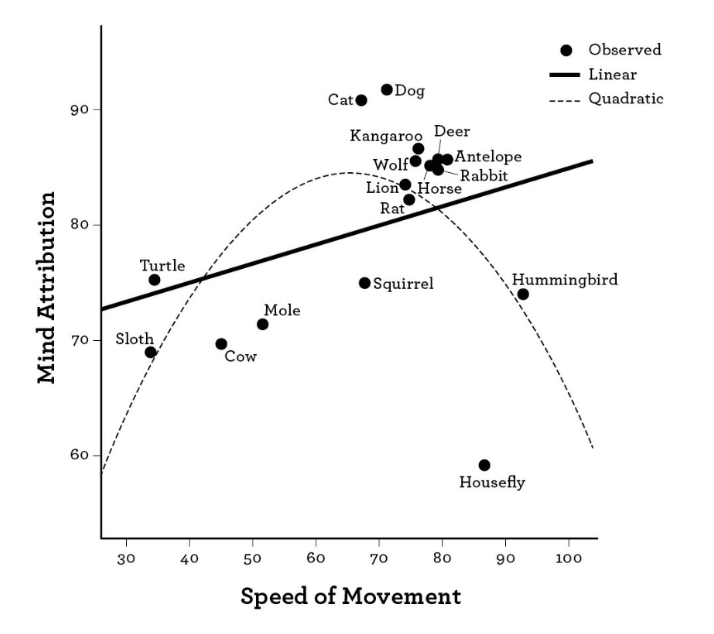


## More from Jonathan Falk …

So I just started reading The Mind Club, which came to me highly recommended. I’m only in chapter 2. But look at the above graph, which is used thusly…

As figure 5 reveals, there was a slight tendency for people to see more mind (rated consciousness and capacity for intention) in faster animals (shown by the solid sloped line)—it is better to be the hare than the tortoise. The more striking pattern in the graph is an inverted U shape (shown by the dotted curve), whereby both very slow and very fast animals are seen to have little mind, and human-speeded animals like dogs and cats are seen to have the most mind. This makes evolutionary sense, as potential predators and prey are all creatures moving at roughly our speed, and so it pays to understand their intentions and feelings. In the modern world we seldom have to worry about catching deer and evading wolves, but timescale anthropomorphism stays with us; in the dance of perceiving other minds, it pays to move at the same speed as everyone else.

* Wegner, Daniel M.; Gray, Kurt. The Mind Club (pp. 29-30). Penguin Publishing Group. Kindle Edition.



## Falk and Gelman…

That “inverted U shape” seems a bit housefly-dependent, wouldn’t you say? And how is the “slight tendency” less “striking” than this putative inverse U shape?

* Yeah, that quadratic curve is nuts. As is the entire theory.
* Also, what’s the scale of the x-axis on that graph? If a sloth’s speed is 35, the wolf should be more than 70, no? This seems like the psychology equivalent of that political science study that said that North Carolina was less democratic than North Korea.
* Falk sent me [the link to the article](https://scholar.harvard.edu/files/dwegner/files/morewedge_preston__wegner_2007.pdf), and it seems that the speed numbers are survey responses for “perceived speed of movement.” GIGO all around!

# Bayesian vs. Frequentist approaches

## Can you have confidence in a confidence interval?

<https://www.johndcook.com/blog/2023/04/23/confidence-interval/>

* Can you have confidence in a confidence interval? In practice, yes. In theory, no.
* If you have a 95% confidence interval for a parameter , can you be 95% sure that is in that interval? Sorta.
* The way nearly everyone interprets a frequentist confidence interval is not justified by frequentist theory. And yet it can be justified by saying if you were to treat it as a Bayesian credible interval, you’d get nearly the same result.

## An Example (from Cook) (part 1/3)

Suppose I want to know what percentage of artists are left handed and I survey 400 artists. I find that 127 of artists surveyed were southpaws. A 95% confidence interval, using the most common approach results in a confidence interval of (0.272, 0.363).

This comes from:

## An Example (from Cook) (part 2/3)

Suppose we redo our analysis using a Bayesian approach. Say we start with a uniform prior on .

* Then the posterior distribution on will have a beta(128, 264) distribution.

Looking at the density function, we can then say in clear conscience that there is a 94% posterior probability that is in the interval (0.272, 0.363).

## Some potential objections (part 3/3)

There are a couple predictable objections at this point. First, we didn’t get exactly 95%. No, we didn’t. But we got very close.

Second, the posterior probability depends on the prior probability. However, it doesn’t depend much on the prior.

Suppose you said “I’m pretty sure most people are right handed, maybe 9 out of 10, so I’m going to start with a beta(1, 9) prior.” If so, you would compute the probability of being in the interval (0.272, 0.373) to be 0.948.

## Can you have confidence in a confidence interval?

Often frequentist and Bayesian analyses reach approximately the same conclusions. A Bayesian can view frequentist techniques as convenient ways to produce approximately correct Bayesian results. And a frequentist can justify using a Bayesian procedure because the procedure has good frequentist properties.

* <https://www.johndcook.com/blog/2023/04/23/confidence-interval/>

# Is it clear what a “null result” actually means?



## Spread of Respiratory Viruses (Masking)

[Preprint by Miller, Tuia and Prasad (2023-04-17)](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4417585)

**Interpretation of Wide Confidence Intervals in Meta-Analytic Estimates: Is the ‘Absence of Evidence’ ‘Evidence of Absence’?**

An updated Cochrane review on physical interventions to slow the spread of respiratory viruses has sparked debate among researchers and in the media over the interpretation of the results, leading Cochrane’s editor-in-chief to issue a statement attempting to clarify comments made by the lead author.

## More from the preprint…

Among other topics, the review examined the effect of medical or surgical masks on the spread of respiratory viruses in the community and found a relative risk of 1.01 95% CI (0.72 - 1.42) after pooling 6 trials. The authors of the Cochrane review concluded, “Wearing masks in the community probably makes little or no difference to the outcome of laboratory‐confirmed influenza/SARS‐CoV‐2 compared to not wearing masks”, and the first author and senior reviewer Tom Jefferson was quoted by a news outlet saying, “There is just no evidence that they make any difference. Full stop.”

## More from the preprint…

In response, Cochrane editor-in-chief Karla Soares-Weiser, issued an unprecedented clarification, stating,

“Many commentators have claimed that a recently-updated Cochrane Review shows that ‘masks don’t work’, which is an inaccurate and misleading interpretation. It would be accurate to say that the review examined whether interventions to promote mask wearing help to slow the spread of respiratory viruses, and that the results were inconclusive.”

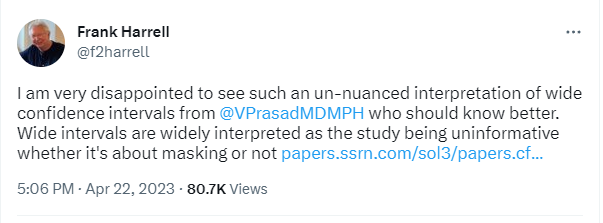
The editor went on to specifically criticize Jefferson, Soares-Weiser also said, though, that one of the lead authors of the review even more seriously misinterpreted its finding on masks by saying in an interview that it proved “there is just no evidence that they make any difference.” In fact, Soares-Weiser said, “that statement is not an accurate representation of what the review found.”

## Conclusion of the Preprint

We found that … the conclusions made by Jefferson and colleagues were not only appropriate, but in line with the standardized approach created by Cochrane. Further, Jefferson’s comment in the media about there being “no evidence that they make any difference” is consistent with their conclusion in the Cochrane review in which they stated, “Wearing masks in the community probably makes little or no difference to the outcome of laboratory‐confirmed influenza/SARS‐CoV‐2 compared to not wearing masks.”

We found no obvious difference between Jefferson’s review and other recent reviews that would justify the differential interpretation and treatment of this study or the unprecedented comments made over its findings. Clarifying comments of the editor-in-chief of Cochrane appear unjustified.

## Frank Harrell’s Reaction



## Frank Harrell’s Reaction



## Frank Harrell’s Reaction



# How I do Assessment of Projects

## George Cobb on Assessment (2004)

<http://www.rossmanchance.com/artist/proceedings/cobb.pdf>

(My remarks) address three concerns: fairness, grade inflation, and a third concern that for now I’ll simply label “Roger.” Each of the three concerns is linked to a corresponding attitude toward assessment.

## What am I doing in your Project B presentation?

1. I’m trying to **systematically** pay attention to you.
2. I’m trying to emphasize things you’ve done well and things you can fix.

 (from Gary Larson)

## What am I doing?

1. I’m trying to **systematically** pay attention to you.
2. I’m trying to emphasize things you’ve done well and things you can fix.
3. I’m trying to make it safe to screw up.
4. And I’m abandoning fairness in favor of assessing your work in your context.

## From Cobb…

For all students, both the more prepared and the less prepared, both the quicker learners and the slower learners, misdirected notions of fairness encourage a sense of competition, discourage helping others, and encourage students to judge themselves and their accomplishments by comparing themselves with others, rather than judging themselves by what works best for them as individuals. … Two different students taking the same course will inevitably get different things from it. We should embrace that inevitable difference, and try to see that each student gets as much as possible from our course, regardless of starting place.

## Project B involves comparing models…

* Modeling a count outcome
  + Compare a Poisson model to a ZIP model, for example
* Modeling a multi-categorical outcome
  + Compare an proportional odds logistic regression to a multinomial model
* Modeling using a Bayesian approach
  + Compare to what you’d get with the standard lm/ols or glm/lrm fit
* Modeling using a weighted linear regression
  + Compare to what you’d get with an unweighted model
* Modeling a time-to-event outcome
  + This is the one type of model where most of the things you demonstrated in Project A still hold, like a *nomogram* or *effects plots* or *ggplot(Predict)* results

## Want to impress me?

Instead of telling me what the value of a coefficient means in generic terms:

A one-unit increase in X is associated with an increase in Y of blah blah blah.

Talk about the impact of your predictors on your outcome using the actual context of the problem you’re studying. Be specific, not generic, to be more effective.

## Embrace Variation!

Show us graphs and tables that help us better understand how much we should “know” after your work about the relationships you observe in the data, and describe those things in terms of the **actual problem under study**.

## Statistics / Data Science is…

… a science, not a branch of mathematics, but uses mathematical models as essential tools.

* John Tukey

Statistics is an important tool in the data analysis/science toolbox. Statistics provides a coherent framework for thinking about random variation, and tools to partition data into signal and noise.

* Hadley Wickham

## Statistics / Data Science is …

… more than just p values and how you get to them.

In fact, forget about null hypothesis significance testing entirely and concentrate instead on:

* embracing variation,
* exploring data and building models and
* studying the size of effects more meaningfully,

even in the rare and unfortunate case where an important and binary decision “must” be made.

## Statistics / Data Science is …

… too important to be left to statisticians.