Predictive social science

Soc Stats Reading Group

Alex Kindel

Princeton University

14 September 2017

Outline

1. Quantitative social research in historical perspective
2. How do we evaluate quantitative knowledge?
3. Challenges for a predictive social science

The social organization of quantitative social research

Quantitative social research in the 20th c. US relied on historically peculiar organizational infrastructure for producing knowledge about people:

- "Prominístrative" data collection and processing (Alonso & Starr 1987; Balogh 1991; Bouk 2016)
- Polling industry and public opinion research (Igo 2007)
- Federal funding for social science (Klausner & Lidz 1986; Geiger 1993)
  - NSF Division of Social Sciences (1958)
  - NDEA/HEA Title VI (1958, 1965)
- Social science research institutes, centers, bureaus (Lazarsfeld 1962; Barton 1979; Merchant 2015)
  - Data archives (e.g. ICPSR 1962)
  - Longitudinal survey research (e.g. WLS 1957; GSS 1972; ANES 1977)
- Formal ethical governance (Stark 2011)
  - Belmont Report (1978); Common Rule (1981)
  - IRBs + DHEW (now: OHRP in DHHS)

How do we evaluate quantitative knowledge?

As efforts to collect and synthesize numeric data about people expanded, how did social researchers evaluate the relation of statistical claims to what was really going on in the world?

In general, truth in formal axiomatic mathematics cannot be defined in its own terms (Tarski 1936)[1].

This goes beyond just getting the math right and making plausible assumptions: **what is the right conceptual language for evaluating statistical claims?**

[1] *Not* Gödel's (1931) work on incompleteness (cf. Collins 1984)

Mid-century social scientists sought relatively mechanical criteria

- Significance (Fisher 1935; also see Leahey 2005)
- Reliability (e.g. Cronbach 1951)
- Validity (e.g. Cronbach & Meehl 1955; Cronbach 1988)
- Consistency (e.g. Meehl 1978)
- Meta-analysis (Glass 1976, etc.)
- Causal inference (Rubin 1974, etc.; also see Holland 1986)

Much of this was developed in applied research settings (esp. applied psychology and education research)

Does any of this stuff even work?

By the 1980s, statistical inference was central to mainstream social research, but many influential scholars – many of them methodologists! – openly disavowed its signature practices (Meehl 1978; Duncan 1984; Collins 1984; Abbott 1988; etc.)

> *I am saying the whole business is so radically defective as to be scientifically almost pointless. . . . I suggest that when a reviewer tries to "make theoretical sense" out of such a table of favorable and adverse significant test results, what the reviewer is actually engaged in, willy-nilly or unwittingly, is meaningless substantive constructions on the properties of the statistical power function, and almost nothing else.*

– Paul Meehl (1978), "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology." *Journal of Consulting and Clinical Psychology* 46, p. 806-834.

Collins (1984), *Statistics vs. Words*

Statistics “implies a theoretical model of the external world”

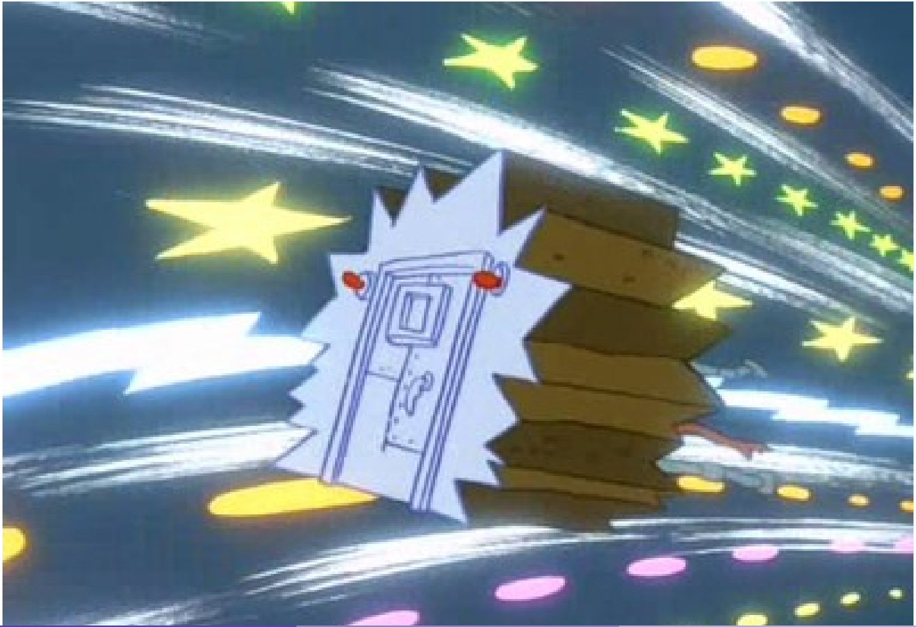
Consequently, statistical tests are a matter of “intellectual distrust” and don’t prove anything about the world:

> *The fact that we emphasize such tests in sociology, almost to the exclusion of all else, indicates the degree of competitiveness and institutionalized distrust in our intellectual community, rather than our scientific standing. (p. 339)*

Alternatives: historical sociology[2] or *consistency* with other theoretical claims

[2] For a critique of this section, see Sewell 2005.

30 years later...



Recently, more explicitly *social* criteria have become popular

- Replication (King 1995; Freese 2007)
- Changing incentives and/or reporting requirements (Simmons et al. 2011)
- Pre-registration (most prominently, Center for Open Science)
- Identify and call out "bad apples"
- Redefine significance (Benjamin et al. 2017)
- Multiple levels of confidence/evidence (e.g. DARPA 2017)
- Professional codes of ethics
- *Prediction*

Prediction

$\hat{\beta}$ vs. $\hat{y}$ (Breiman 2001)

Common task framework: lots of people try to predict the same outcome (Donoho 2015)

- Data withholding to assess out-of-sample performance
- Locally shared performance metric (e.g. everyone minimizes MSE)
- Competing models from multiple theoretical perspectives evaluated over same dataset

Predictive power and consistency

Using predictive criteria, particularly out-of-sample predictive criteria, is an exceedingly simple means to highlight the extent to which the theoretically informed models anticipate reality, and which among those models does a better job of it. (Cranmer & Desmarais 2017, p. 149)

Wherever possible, two or more nonredundant estimates of the same theoretical quantity should be made, because multiple approximations to a theoretical number are always more valuable . . . than a so-called exact test of significance, or even an exact setting of confidence intervals. (Meehl 1978, p. 829)

Five challenges for predictive social science

1. What kind of science is it?
2. What makes a good prediction task?
3. How do we interpret predictive results?
4. What do predictive limits mean?
5. How do we think about the ethics of prediction?

Challenge 1: What kind of science is it?

It's not obvious how prediction fits into what quantitative social science already does.

- Prediction as improving causal inference
- Predictive accuracy as a better performance goal for social science models
- Prediction as an intellectual and professional goal for social science generally
- Prediction as explanatory in its own right
  - Where does descriptive analysis fit in?

Challenge 2: What do we predict?

Given a task, predictions are relatively easy to aggregate (Breiman 1996; Donoho 2015)

In practice, this means that the task – i.e. the measurement and selection of outcomes – is even more important than before (Hofman, Sharma & Watts 2017)

Where and how do research questions come into common task predictive modeling?

Challenge 3: How do we interpret predictive results?

Aphoristically, predictive algorithms are “black boxes” or “opaque” or “uninterpretable”

This isn’t necessarily true, but we *do* lack institutionalized standards for theoretical interpretation (Cranmer & Desmarais 2017) – and whatever standards we choose are unlikely to look like a regression table

Challenge 4: What do predictive limits mean?

Are there intrinsic theoretical limits to how predictable classes of social phenomena are?

How do we know when we're scraping the ceiling?

Challenge 5: What is its ethical status?

The formal ethical organization of US social science – OHRP, the Common Rule, IRBs – presumes a different methodological paradigm (Stark 2007)

The kind of data that feed into predictive models have uncertain ethical status (Stevens 2014)

Ethical standards of industry are widely regarded as insufficient (e.g. Facebook “emotional contagion” study)

Works referenced

Abbott, Andrew. 1988. “Transcending General Linear Reality.” *Sociological Theory* 6(2): 169-186.

Alonso, William, and Paul Starr (eds.) 1989. *The Politics of Numbers.* Russell Sage Foundation.

Balogh, Brian. 1991. *Chain Reaction: Expert Debate and Public Participation in American Commercial Nuclear Power, 1945-1975*. Cambridge University Press.

Barton, Allen H. 1979. “Paul Lazarsfeld and Applied Social Research: Invention of the University Applied Social Research Institute.” *Social Science History* 3(3-4): 4-44.

Bouk, Dan. 2016. *How Our Days Became Numbered: Risk and the Rise of the Statistical Individual.* University of Chicago Press.

Benjamin, Daniel J. *et al.* 2017. “Redefine statistical significance.” *Nature Human Behavior*.

Breiman, Leo. 2001. “Statistical modeling: The two cultures.” *Statistical Science* 16(3): 199-231.

Collins, Randall. “Statistics Versus Words.” *Sociological Theory* 2: 329-362.

Cranmer, Skyler J. and Bruce A. Desmarais. 2017. “What Can We Learn from Predictive Modeling?” *Political Analysis* 25: 145-166.

Works referenced

Cronbach, Lee J. 1951. “Coefficient alpha and the internal structure of tests.” *Psychometrika* 16(3): 297-334.

Cronbach, Lee J. 1988. “Five perspectives on validity argument.” In H. Wainer & H. I. Braun (eds.), *Test Validity*, Routledge. Cronbach, Lee J., and Paul E. Meehl. 1955. “Construct validity in psychological tests.” *Psychological Bulletin* 52(4): 281-302.

Donoho, David. 2015. “50 years of Data Science.” *Princeton NJ, Tukey Centennial Workshop, 2015*.

Duncan, Otis D. 1984. *Notes on Social Measurement: Historical and Critical.* Russell Sage Foundation.

Fisher, R. A. 1935. *The Design of Experiments.* Oliver & Boyd, Edinburgh.

Freese, Jeremy. “Replication standards for quantitative social science: Why not sociology?” *Sociological Methods & Research* 36(2): 153-172.

Geiger, Roger L. 2008. *Research and relevant knowledge: American research universities since World War II.* Transaction Publishers.

Glass, Gene V. 1976. “Primary, secondary, and meta-analysis of research.” *Educational Researcher* 5(10): 3-8.

Hofman, Jake M., Amit Sharma, and Duncan J. Watts. 2017. “Prediction and explanation in social systems.” *Science* 355: 486-488.

Works referenced

Holland, Paul W. 1986. “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81(396): 945-960.

Igo, Sarah E. 2007. *The Averaged American: Surveys, Citizens, and the Making of a Mass Public.* Harvard University Press.

King, Gary. 1995. “Replication, Replication.” *PS: Political Science & Politics* 28(3): 444-452.

Klausner, Samuel Z., and Victor M. Lidz. 1986. *The nationalization of the social sciences.* University of Pennsylvania Press.

Lazarsfeld, Paul F. 1962. “The sociology of empirical social research.” *American Sociological Review* 27(6): 757-767.

Lazer, David *et al.* 2009. “Life in the network: the coming age of computational social science.” *Science* 323(5915): 721-723.

Leahey, Erin. 2005. “Alphas and asterisks: the development of statistical significance testing standards in sociology.” *Social Forces* 84(1): 1-24.

Meehl, Paul E. 1978. “Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology.” *Journal of Consulting and Clinical Psychology* 46: 806-834.

Works referenced

Merchant, Emily K. 2015. “Prediction and Control: Global Population, Population Science, and Population Politics in the Twentieth Century.” Doctoral dissertation, University of Michigan. Rubin, Donald B. 1974. “Estimating causal effects of treatments in randomized and nonrandomized studies.” *Journal of Educational Psychology* 66(5): 688-701.

Sewell Jr, William H. 2005. *Logics of History: Social Theory and Social Transformation.* University of Chicago Press.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. “False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant.” *Psychological Science* 22(11): 1359-1366.

Stark, Laura. 2011. *Behind Closed Doors: IRBs and the Making of Ethical Research.* University of Chicago Press.

Stevens, Mitchell L. 2014. “An ethically ambitious higher education data science.” *Research and Practice in Assessment* 9: 96-97.

Tarski, Alfred. 1936. “The Concept of Truth in Formalized Languages.” (J.H. Woodger, trans.)