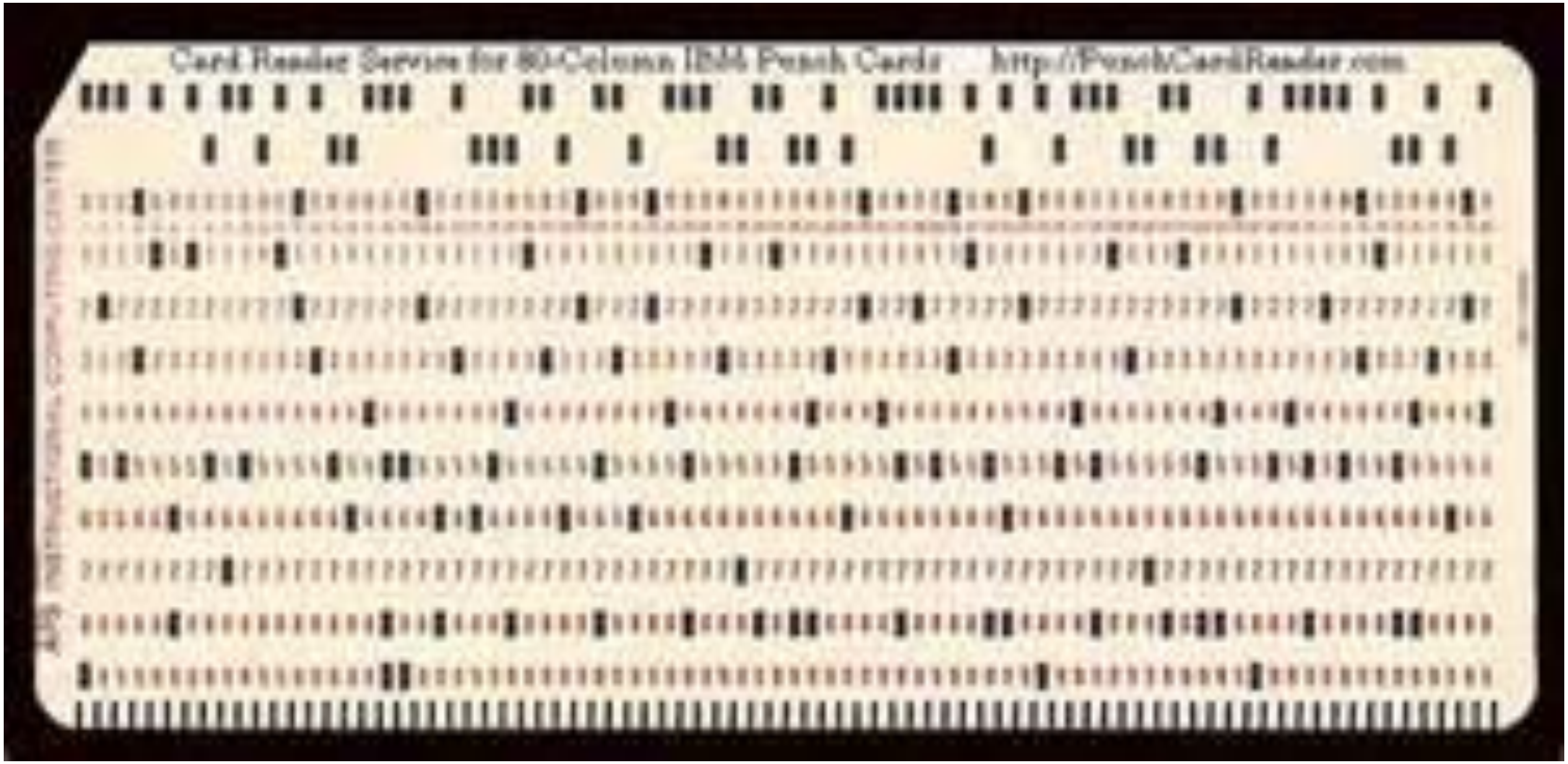
Computational Capacity and Statistical Inference: A Never Ending Interaction

Finbarr Sloane

EHR/DRL

“Studies in Crop Variation I (1921)”

- It has been estimated that Sir Ronald A. Fisher spent about 185 hours to generate a single complicated table in this now classic paper (Salsburg, 2001).
- The paper includes fifteen other tables of similar complexity as well as four large complicated graphs.
- Physical time estimate: ???


Card Reader Service for 80-Column IBM Punch Cards
http://PunchCardReader.com

Central Goals

- R. A. Fisher was able to provide statistical theory that took full advantage of the computational capacity of the 1920's.

- We can see from the presentations that new methods are evolving to computational capacity.

- Our goal today is to do the same for the 2020's for STEM education (i.e., STEM learning).

Data Models

- Statisticians in applied research consider data modeling as the template for statistical analysis.
- The applied researcher can formulate a reasonably good parametric class of models for a complex mechanism derived by nature.
- Once published, the model tends to take on a life of its own.

Small Data with Big Problems

- Re-analyzing the Stanford data on Chronic Hepatitis;
- Between 1975 and 1980:
  - 155 patients observed with Chronic Hepatitis;
  - 33 patients died;
  - 19 variables measured (medical histories, biopsies, liver function tests, etc.);

Bootstrapping

- The basic idea of bootstrapping is that inference to a population from sample data (sample to population) can be modeled by resampling the sample data and performing the inference on the resampled data.
- As the population is unknown, the true error in a sample statistic against the population value is unknowable.
- In bootstrap resamples, the population is now the sample, and this is known: hence the quality of inferences from resample to data is measurable.

Stanford’s MS Stepwise Logistic Regression

- The original Stanford Medical School analysis concluded that the important variables were numbers 6,12,14, & 19.
- Diaconis and Efron (1983) drew 500 bootstrap samples from the original data set and used a similar procedure, including logistic regression, to isolate the important variables in each bootstrapped data set.

D&E Results

- “Of the four variables originally selected not one was selected in more than the 60 percent of samples”.

- The variables selected in the original analysis cannot be taken as a good model of reality.

Problems of Current Data Modeling Techniques

- Bickel, Ritov and Stoker (2001) show that goodness of fit statistics have very little power unless the direction of alternative is precisely specified;
  - Omnibus goodness of fit tests have little power and will not reject until the lack of fit is extreme.

Data Modeling problems continued:

- Residual analyses are not sensitive in more than a few dimensions,
  - Misleading conclusions may follow.
- A multiplicity of data models – data will often point to more than one model. These competing models may give different pictures of the relation between the predictor and the response variables.
- A lack of predictive accuracy (needs cross validation).

Why are the D&E Results Important?

- These practices are still quite common (e.g., Regression in Biology):
  - A simple linear regression method for quantitative loci linkage analysis with censored observations. **Genetics, 173:** 1735-1745 (2006).
  - Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. **Nature Communications 5:** 1-10 (2014).

Five Inferences from the Presentations

- With any model fit to data, the information extracted is as much about the model as it is about nature (this is likely more true at the beginning of the modeling process).
- The better the model emulates nature, the more reliable our information.
- The error rate in predicting future outcomes should be a prime criterion as to how good the emulation is.

Inferences (continued)

- The most accurate current prediction algorithms can be applied to very high dimensional data, but they are also complex

- A complex predictor can yield a wealth of “interpretable” scientific information about the prediction mechanism and the data

Data Modelers

- Model Formulation:
  - Prediction and or Classification
  - What happens when you have more variables than data points?
  - Linearity Vs non-linearity
  - Original forms of data: numerical, text, video
  - Understanding events to create meta-data
  - Populations Vs Samples
  - Data Visualization
  - Model interpretation

Questions

- Model Estimation (examples):
    - Choosing from many competing models?
    - Sensitivity of models to changes in the data?
- Model Validation: Estimated predictive capacity?
    - – Predictive accuracy