Getting Ready for Big Data

Implications for intro stats

Bob Stine
Department of Statistics, Wharton
www-stat.wharton.upenn.edu/~stine
Change is upon us...

- Session topics
  - Shifting away from classical methods
  - Communication skills
  - Data visualization
  - Business analytics
  - Predictive analytics
  - Sports analytics
  - Analytics in curriculum

- Rather than discuss BA course, consider implications of ‘big data’ for intro courses
Big Data?

- **Examples**
  - Scanner data captured at retail transaction
  - Credit card, financial transactions
  - Health records and genetic testing
  - Social media, web visits

- **Characteristics**
  - Volume, variety, velocity, veracity…
  - Often not collected with stat in mind
Big Data?

• Examples
  • Scanner data captured at retail transaction
  • Credit card, financial transactions
  • Health records and genetic testing
  • Social media, web visits

• Characteristics
  • Volume, variety, velocity, veracity…
  • Often not collected with stat in mind

• Oops, we’re not in Kansas anymore
Big Data Changes Things

- Huge number of observations
  - All patient outcomes for a state in a year, all sales transactions, every web query…

→ ‘Everything’ seems statistically significant.
  p-values ≈ 1.0e-122
Big Data Changes Things

• Huge number of observations
  • All patient outcomes for a state in a year, all sales transactions, every web query…

  ➜ ‘Everything’ seems statistically significant.
  \[ p\text{-values} \approx 1.0\times 10^{\text{-122}} \]

• But…
  • Effect size
    Substantive versus statistical significance
  • Dependence
    Are those observations independent? Hurricane versus car insurance
    Behavior of credit markets, mortgages in 2008
Big Data Changes Things

• Data snooping, hypothesis discovery
  • Wide data sets offer many choices
  • Find important sales patterns
  • Beer and diapers

→ Model fits data very well
Big Data Changes Things

• Data snooping, hypothesis discovery
  • Wide data sets offer many choices
  • Find important sales patterns
  • Beer and diapers

  ➜ Model fits data very well

• Multiplicity
  • Look for items bought together in scanner data
    1000 items produces 500,000 pairs
  • Voter surveys include 1000s of questions related to preferences
Implications for Intro Stat

• Most students will have only one or maybe two semester exposure to statistics

• Promotional opportunity
  • Attract some to more majors
  • Provide practical knowledge for others

• Address issues for big data in this context
  • Dependence
  • Multiplicity
  • Effect size
  • Others

Zero-sum game
Getting Ready for Big Data

• Have a question to motivate, guide, control the modeling, statistical analysis
  • What question are we trying to answer?
  • Too easy to spend hours wandering in big data without a clear objective
Getting Ready for Big Data

- Have a question to motivate, guide, control the modeling, statistical analysis
  - What question are we trying to answer?
  - Too easy to spend hours wandering in big data without a clear objective

- Importance in intro courses
  - Why am I doing this? Who cares? Why does this matter?
  - Common metaphors ‘TST’, ‘MMMM’
Getting Ready for Big Data

• Data is happy to generate many, many hypotheses
  • Testing response to stimulus letters
  • Multiplicity (simultaneous inference)
Getting Ready for Big Data

• Data is happy to generate many, many hypotheses
  • Testing response to stimulus letters
  • Multiplicity (simultaneous inference)

• Importance in intro courses
  • Examples for regression models
    Stock market
  • Simple remedies are easy to teach
    (e.g. Bonferroni p-values)
Others have noticed...

xkcd

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!

WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).

THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.

SCIENTISTS!

[Panel 1: Character say, "JELLY BEANS CAUSE ACNE!"
Character holds a jar of jelly beans.

Panel 2: Character say, "WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05)."
Character sits at a desk with a stack of papers.

Panel 3: Character say, "THAT SETTLES THAT."
Character holds a chart with a bar graph.
]
Others have noticed...

xkcd
Others have noticed...

- Source of publication bias in journals
- Economist article
Getting Ready for Big Data

- 'Big Data' don’t always measure what you think they measure
  - Units, time lags, codebooks
  - Data preparation is key (95% rule)
  - Mailing list example is full of these problems
Getting Ready for Big Data

• ‘Big Data’ don’t always measure what you think they measure
  • Units, time lags, codebooks
  • Data preparation is key (95% rule)
  • Mailing list example is full of these problems

• Importance in intro courses
  • Give students data that is more realistic
    Missing values, vague definitions
  • Too much, too soon?
Getting Ready for Big Data

• Large data sets typically gathered as part of transaction processing, not for analysis
  • Repurposed accounting records
  • Justify that sparkling new data warehouse
Getting Ready for Big Data

• Large data sets typically gathered as part of transaction processing, not for analysis
  • Repurposed accounting records
  • Justify that sparkling new data warehouse

• Importance in intro courses
  • Always ask
    “What would be the ideal data to answer my question?”
  • Compare that to the data that you have
Getting Ready for Big Data

- Dependence often makes large data sets much smaller
  - Predicting credit behavior in US: dep customers
  - Repeated measurements (longitudinal)
Getting Ready for Big Data

• Dependence often makes large data sets much smaller
  • Predicting credit behavior in US: dep customers
  • Repeated measurements (longitudinal)

• Importance in intro courses
  • Carefully define assumption of independent observations
  • Divisor n is not number of cases, but ind cases
  • Relevant source of variation
  • Common examples: ‘lurking variable’
Getting Ready for Big Data

- Results may not generalize
  - On-line experiment on weekday not descriptive of weekend (Can imagine other factors)
  - Text model of one author not applicable to others
  - Transfer learning problem
Getting Ready for Big Data

• Results may not generalize
  • On-line experiment on weekday not descriptive of weekend (Can imagine other factors)
  • Text model of one author not applicable to others
  • Transfer learning problem

• Importance in intro courses
  • Sampling from what population?
  • Does same population exist? ‘Population drift’
  • Dynamics of election polls
Place for Classical Methods

- Surveys and sampling still make sense
  - Billions of credit card transactions each year
  - Do you need to see them all to track prices?
  - DoE analysis of prices for ethanol fuels

- Experimental design remains essential
  - Hard to beat that randomized experiment
  - Google ad response measurement
  - Trivial to do experiment

- Generalize?
Thanks!