Learning what works in populations for public health and public policy:
The role of statistics (and statisticians)

Elizabeth Stuart
www.biostat.jhsph.edu/~estuart

June 28, 2017
1. Evidence-based policymaking
2. Estimating population treatment effects
3. Experimental design options
   - Randomized trials in representative populations
   - Pragmatic randomized trials
4. Non-experimental design options
   - Comparative interrupted time series
   - Propensity score methods in representative samples
5. Combining experimental and population data
6. Conclusions
Outline

1. Evidence-based policymaking
2. Estimating population treatment effects
3. Experimental design options
   - Randomized trials in representative populations
   - Pragmatic randomized trials
4. Non-experimental design options
   - Comparative interrupted time series
   - Propensity score methods in representative samples
5. Combining experimental and population data
6. Conclusions
“Evidence-based policymaking”

• Past ten or so years has seen increased interest in “evidence-based policymaking”

• 2016: Bipartisan Act (Ryan/Murray) established the "Commission on Evidence-Based Policymaking" (CEP)

"The Act recognizes that better use of existing data may improve how government programs operate. The mission of the Commission is to develop a strategy for increasing the availability and use of data in order to build evidence about government programs, while protecting privacy and confidentiality. Through the course of the Commission’s work, members will study how data, research, and evaluation are currently used to build evidence, and how to strengthen the government’s evidence-building efforts."
Non-profits and foundations also getting in the game . . .

• Groups:
  • Evidence-based policymaking collaborative
  • Results First

• Groups/foundations involved: Arnold Foundation, Urban Institute, Brookings Institution, American Enterprise Institute, Pew Foundation, MacArthur Foundation, Heritage Foundation

• Companies like Mathematica Policy Research, Westat, etc. writing policy briefs and creating tools to help policymakers and decisionmakers understand and use data and evidence
Do statisticians have a seat at the table?

• Not so much: Many discussions and work dominated by economists and policy wonks

• Many Congressional committees have staff economists, but essentially no statisticians on Capitol Hill

• Evidence-based policymaking collaborative:
  • Staff and advisors mostly have public policy or economics degrees (and note few statisticians in policy schools . . . )

• Commission on Evidence-Based Policymaking
  • 2 statisticians (Robert Groves, Nancy Potok)
  • 7 economists
  • 6 other backgrounds (psychology, law, etc.)

• (National Academies and Committee on National Statistics are exceptions)
My premise and message for today:
Statisticians should have a greater role in this world
What do we have to offer?

• Skills at handling messy data and all of its complications
• More attention than many fields to things like measurement, sampling, how to handle non-random samples
• Ability to develop new methods when needed
• Inherent interest and expertise in rigorous study design and analysis
• Ability to bridge fields and make connections between methods
• Many of these topics become more, not less, important in the era of “big data"
What do we have to gain?

- Contribute to important policy questions
- Staying relevant in a world where the government Chief Data Scientist appointed with much fanfare, but not so the Chief Statistician
- Interesting statistical challenges
- Collaborations with individuals across many disciplines
Outline

1 Evidence-based policymaking

2 Estimating population treatment effects

3 Experimental design options
   - Randomized trials in representative populations
   - Pragmatic randomized trials

4 Non-experimental design options
   - Comparative interrupted time series
   - Propensity score methods in representative samples

5 Combining experimental and population data

6 Conclusions
Today: Estimating population treatment effects . . .

• Today I will talk about some of my own work, which broadly fits into this topic

• For public health and public policy purposes we often want to know about broad population effects of risk factors (e.g., lead exposure), interventions (e.g., flu shots), or policy changes (e.g., handgun control laws)

• Not “what works for whom" but rather “what is the best decision for the population"?
  • If we expose a population of people to this “treatment", what is the average outcome, as compared to if we don’t expose the population to the treatment?
Randomized trials: Give us “internal validity"

• In a randomized experiment we randomly assign who gets the intervention and who doesn’t

• Randomized experiments give us unbiased effect estimates in the sample at hand
  • Any difference in outcomes must be due to the intervention and not to any pre-existing differences between the groups

• This known as “internal validity"
But what about external validity?

• But increasing concerns about lack of external validity: That the results in the trial sample may not carry over to relevant target populations
  • External validity bias arises if the randomized trial sample and the population of interest differ on characteristics that also relate to treatment effects (Olsen, Orr, Bell, and Stuart, 2013)

• When interested in population treatment effects, external validity may be just as important as internal validity

• Framework for these trade-offs between internal and external validity in Imai, King, and Stuart (2010)
Two examples of lack of representativeness of trials

- Education research (Olsen, Bell, Orr, and Stuart, 2015)
  - School districts that participate in large-scale evaluations much larger and poorer than the average school district in potential target populations (Stuart et al., 2017)
  - In one empirical example, external validity bias on the order of 0.1 standard deviations

- Trials of drug abuse treatment (Susukida, Crum, Stuart, Ebnesajjad, and Mojtabai, 2015)
  - People participating in trials had more education and more likely to be employed full-time than are individuals seeking drug abuse treatment nationwide
Are there ways we can have studies with high internal and external validity?
Outline

1. Evidence-based policymaking
2. Estimating population treatment effects
3. Experimental design options
   - Randomized trials in representative populations
   - Pragmatic randomized trials
4. Non-experimental design options
   - Comparative interrupted time series
   - Propensity score methods in representative samples
5. Combining experimental and population data
6. Conclusions
1. Evidence-based policymaking
2. Estimating population treatment effects
3. Experimental design options
   - Randomized trials in representative populations
   - Pragmatic randomized trials
4. Non-experimental design options
   - Comparative interrupted time series
   - Propensity score methods in representative samples
5. Combining experimental and population data
6. Conclusions
Randomized trials in representative populations

- In some cases can conduct a randomized trial in a formally representative population
- This most common in evaluations of government programs, where sites can be mandated to participate
  - Head Start Impact Study
  - Upward Bound evaluation
- But in general difficult to carry out, and expensive
1. Evidence-based policymaking

2. Estimating population treatment effects

3. Experimental design options
   - Randomized trials in representative populations
   - Pragmatic randomized trials

4. Non-experimental design options
   - Comparative interrupted time series
   - Propensity score methods in representative samples

5. Combining experimental and population data

6. Conclusions
Pragmatic randomized trials

- Movements towards pragmatic clinical trials in many fields, especially mental health
  - CATIE trial of treatment for schizophrenia
  - STEP-BD trial of treatment for bipolar disorder
  - ADAPTABLE trial of high vs. low dose aspirin using PCORNet

- Main idea: conduct trial in large sample of people in range of "real-world" settings

- Great idea, but not always formally representative, and usually high-cost
But what if we can’t randomize?

Or at least not in the relevant population . . .
Outline

1. Evidence-based policymaking
2. Estimating population treatment effects
3. Experimental design options
   - Randomized trials in representative populations
   - Pragmatic randomized trials
4. Non-experimental design options
   - Comparative interrupted time series
   - Propensity score methods in representative samples
5. Combining experimental and population data
6. Conclusions
Outline

1. Evidence-based policymaking
2. Estimating population treatment effects
3. Experimental design options
   - Randomized trials in representative populations
   - Pragmatic randomized trials
4. Non-experimental design options
   - Comparative interrupted time series
   - Propensity score methods in representative samples
5. Combining experimental and population data
6. Conclusions
Option 1: Comparative interrupted time series

- In cases of abrupt policy changes, comparative interrupted time series methods can be used to estimate effects of that policy change.
- Essentially compare pre and post periods.
- Best if there are comparison sites without the change, to model trends over time.
- Example: Effects of federal mental health parity law on mental health service use (Stuart et al., 2017).
- Example: Gun control laws (Rudolph, Stuart, Vernick, and Webster, 2015).
January 5, 2016: President Obama announces new executive actions around gun control

One of these actions is around background checks. In justifying that, President Obama stated:

*In fact, we know that background checks make a difference. After Connecticut passed a law requiring background checks and gun safety courses, gun deaths decreased by 40 percent – 40 percent.*
CT permit to purchase handgun law

• Research question: How many homicides were prevented by Connecticut’s 1995 Permit-to-Purchase handgun law?

• Permit to purchase laws require individuals to get a permit to purchase a handgun, which involves a background check, fingerprints, safety training

• Use data on homicide rates over time (1984-2005) for CT and 39 states “at-risk” for implementing a PTP law in 1995 (i.e., without one)

• Fundamentally, compare trends and changes in CT with trends and changes in the other states

• Look at firearm homicides and nonfirearm homicides
Synthetic control approach

• Comparative interrupted time series uses the trends seen in the comparison states to help “guess” what would have happened in CT had it not passed the law

• This will generally work better if the comparison states are a good proxy for CT; more believable if the trends in the treatment and comparison sites were similar before the policy change

• “Synthetic control" method reweights comparison sites to make the pre-law trends as similar as possible to the site with the policy change

• Intuitively, the idea is that we might believe that their trends would have continued in similar ways if they were similar on pre-intervention factors

• Originally proposed by Abadie and Gardeazabal (2003), Abadie, Diamond, and Hainmueller (2010, 2013)
Benefits of the synthetic control approach

• Transparency: data-driven selection of comparison units
  • And can do this using just the pre-period data
  • Separation of “design" and “analysis" (Rubin, 2001)

• Safeguard against extrapolation: ensures similarity of intervention and comparison units
Results graphically

Firearm Homicide Rate

Year

Nonfirearm Homicide Rate

Year
Results

• Law associated with a significant 40% reduction in the firearm homicide rate for the 10 years following implementation (296 deaths prevented)

• Law associated with a nonsignificant 24 nonfirearm homicides prevented for the 10 years following implementation

• Comparative interrupted time series is often the best way to estimate the effects of interventions in broad populations
  • Allowed us to estimate the effects of this law across the state of CT
Methodological questions remain, though

• How to weight different variables in the weighting of comparison sites?

• How to use this approach when there are multiple intervention sites? Aggregate to one? Run on each and then aggregate? Something else?

• How to formally combine results that are complementary: e.g., similar laws in CT and MO, effects on homicides vs. suicides

• Statisticians can help answer these questions!

• And what if we don’t have a population change like this law change, or have individual-level data that we might want to take advantage of?
Outline

1. Evidence-based policymaking
2. Estimating population treatment effects
3. Experimental design options
   - Randomized trials in representative populations
   - Pragmatic randomized trials
4. Non-experimental design options
   - Comparative interrupted time series
   - Propensity score methods in representative samples
5. Combining experimental and population data
6. Conclusions
“Big data"

• There are more and more sources of data on broad populations of interest
  • Large-scale EHR systems, medical claims files, nationally representative surveys, etc.

• Could potentially use this data to estimate causal effects in those broad populations
  • e.g., effects of particular medical interventions or drugs, effects of environmental exposures

• But how to do so, given lack of randomization?
Propensity score methods as one approach

• Main problem in non-experimental studies is confounding: treatment and comparison individuals may be very different from one another on lots of factors

• Propensity scores commonly used as key design tool in such studies

• Goal is to replicate a randomized experiment as much as possible, by forming groups similar on the observed covariates

• (Relies on assumption that there are no unobserved differences between the groups once we condition on the observed ones; can also do sensitivity analysis regarding this assumption)
Effects of psychosocial therapy after suicide attempt

• Use Danish registry data (on population of people in Denmark) to compare outcomes of individuals who received psychosocial therapy after a suicide attempt to similar individuals who didn’t
  • Very large sample, allows long follow-up, extensive covariates available
• Suicide prevention clinics began operation in Denmark in 1992, now nationwide
• Erlangsen et al. (2014)
What do propensity scores do?

• The problem is that it is hard to find similar groups with respect to all covariates individually

• Propensity scores give a particular type of dimension reduction that allows matching on just the propensity score, not dealing with each covariate individually

• Propensity score methods attempt to replicate two features of randomized experiments
  • Create groups that look only randomly different from one another (at least on observed variables)
  • Don’t use outcome when setting up the design
Data

• Linked registers: Danish civil register, National registry of patients, Psychiatric central registry, and Registry of causes of death

• “Treatment group”: Users of suicide prevention centers after suicide attempt who received one or more psychotherapeutic treatment sessions

• “Comparison group”: Similar individuals who also had attempted suicide but who did not receive treatment from a suicide center after their suicide attempt. (Identified from hospital presentation).

• Ages 10+

• Follow-up from 1992 to 2011

• Treatment group: 5,678 people (42,893 person years); Comparison group: 58,281 people (544,602 person years)
Matching process

- For each user of the clinics, found 3 individuals with similar propensity scores (who also had the same values for “any psychiatric disorder" and “previous deliberate self-harm")
- Subjects selected to be similar on 31 observed covariates:
  - Demographics: Time period, gender, age, born in Denmark, civil status, educational level, SES, urban/rural, has children
  - Suicide attempt: Previous attempt, multiple repeats (3+), determined method
  - Psychiatric diagnoses: Mood disorders, anxiety, personality, PTSD, eating, drug abuse, alcohol abuse, schizophrenia, other, antidepressant treatment
  - Family history: Parents’ psychiatric disorder, parents’ suicidal behavior
Subjects look similar after the matching
Program was effective at reducing 10 year risk of . . .

- Repeat suicide attempts: OR=0.82; CI (0.75, 0.89)
- Death by suicide: OR=0.71; CI (0.56, 0.91)
- Death from any cause: OR=0.65; CI (0.57, 0.74)

- Findings robust to a potential unobserved confounder (Liu, Kuramoto, and Stuart, 2013)
Conclusions re propensity score methods

• Propensity score methods can be a useful tool for improving estimation of causal effects in non-experimental settings
• Can be used in large-scale population-based datasets
• Conclusions can then be drawn about the effects of exposures or interventions in that population
• But non-experimental studies will still have potential concerns about unobserved confounders that may bias the results
And again methods questions remain

- How to handle measurement error, which is especially common in EHR type datasets?
- How to combine propensity score methods with complex survey data? (This particularly relevant for population effects)
- How to ensure balance is obtained on the “right" variables?
- How to convey the key underlying assumptions and develop general sensitivity analysis approaches?
Outline

1. Evidence-based policymaking

2. Estimating population treatment effects

3. Experimental design options
   • Randomized trials in representative populations
   • Pragmatic randomized trials

4. Non-experimental design options
   • Comparative interrupted time series
   • Propensity score methods in representative samples

5. Combining experimental and population data

6. Conclusions
Generalizing trial results to target populations

• What about bridging experimental and non-experimental studies by using multiple data sources together?
• Recent work has looked at how to generalize results from randomized trials to target populations
• Particularly important to adjust for variables that differ between the trial and population and that moderate treatment effects
• Using data from the trial and information on characteristics of the population of interest, weight the trial sample to “look like” the population
• (Similar to propensity score or survey sampling weights; stuff statisticians know well!)
Assumptions

- Experiment was randomized
- "Sample ignorability for treatment effects": selection into the trial independent of impacts given the observed covariates
  - \((Y_k(1) - Y_k(0)) \perp S_k | X_k\)
  - For the same value of observed covariates, impacts the same across trial and population
  - No unmeasured variables related to selection into the trial and treatment effects
- "Overlap": all individuals in the population had a non-zero probability of participating in the trial

Analogous to strong ignorability/unconfoundedness of treatment assignment in non-experimental studies
Case study: The ACTG Trial (Cole & Stuart, 2010)

- Trial examined highly active antiretroviral (HAART) therapy for HIV compared to standard combination therapy
  - Intent-to-treat analysis: Hazard ratio of 0.51 (95% CI: 0.33, 0.77)
- Question: What would the effects of the treatment be if implemented nationwide?
- Trial and US population differ on variables that also moderate effects (age, race, sex)
Inverse probability of selection weighting

- Weight the trial subjects up to the population
- Each subject in trial receives weight \( w_i = \frac{1}{P(S_i=1|X)} \)
  - (Inverse of their probability of being in the trial)
- Use those weights when calculating means or running regressions
- Related to inverse probability of treatment weighting, Horvitz-Thompson estimation in surveys
### Estimated population effects

<table>
<thead>
<tr>
<th></th>
<th>Hazard ratio</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crude trial results</td>
<td>0.51</td>
<td>0.33, 0.77</td>
</tr>
<tr>
<td>Age weighted</td>
<td>0.68</td>
<td>0.39, 1.17</td>
</tr>
<tr>
<td>Sex weighted</td>
<td>0.53</td>
<td>0.34, 0.82</td>
</tr>
<tr>
<td>Race weighted</td>
<td>0.46</td>
<td>0.29, 0.72</td>
</tr>
<tr>
<td>Age-sex-race weighted</td>
<td>0.57</td>
<td>0.33, 1.00</td>
</tr>
</tbody>
</table>

- CI’s longer for weighted results
- Effects generally somewhat attenuated, except for weighting only by race
Everyone wants to assume that study results generalize

- But very few statistical methods exist
- At this point, lots of “hand waving,” qualitative statements
- Need more statistical methods to quantify and improve external validity
  - For both study design and study analysis
Methods questions remain here too!

• How to best assess treatment effect moderation, especially given limited power for such in randomized trials?
• What if there are unobserved effect moderators? (One solution; Nguyen et al., 2017)
• How to handle variables measured differently in the trial and population data?
• Do these weighting methods make sense, or should flexible regression models be used instead?
• How can we help people understand sample vs population effects?
Outline

1. Evidence-based policymaking
2. Estimating population treatment effects
3. Experimental design options
   - Randomized trials in representative populations
   - Pragmatic randomized trials
4. Non-experimental design options
   - Comparative interrupted time series
   - Propensity score methods in representative samples
5. Combining experimental and population data
6. Conclusions
We as statisticians have an opportunity, and obligation, to be engaged

• Statisticians can help provide a structure and framework for use of evidence and data in policymaking

• With core attention to quality and rigor of the studies (NOT the empirical results)
  • This particularly important in a polarized political climate

• Can also help teach methods, including when different methods are needed
  • May involve formal classroom teaching and more informal interactions
  • Pick up the phone when reporters call!
What shouldn’t we do?

• Just be seen as the people who always say “no” or that studies are bad; need to give solutions
• Provide overly complicated solutions that yield only minimal improvement
• Quibble with each other over minor differences in methods
• Be afraid to speak out when we see poor study designs or analyses
As inspiration

• Some great examples of statisticians engaged in these discussions
  • ASA Scientific and Public Affairs Advisory Committee
  • Nussbaum and Reiter editorial in The Hill
  • Steve Pierson at ASA (Science Policy and Advocacy section of ASA website)
  • Blogs (Simply Statistics, Andy Gelman, Frank Harrell, . . . )
  • stats.org helping journalists understand and use data

• And there seems to be demand for our insights and perspectives
  • Head of Scientific Review Office at NIH met with representatives from ASA
  • ASA statement on p-values viewed > 194,000 times
  • 10 simple rules for effective statistical practice paper viewed > 170,000 times
Conclusions

• Statisticians have a role to play in helping guide use of data and evidence in policymaking

• As one example, estimating population effects can be challenging, but is often what we need to make informed policy decisions

• Hopefully combinations of high-quality data, rigorous methods, and careful analysis (and us!) can help improve public health and public policy
For more information . . .

- http://www.biostat.jhsph.edu/~estuart/
- estuart@jhu.edu
- Funding thanks to NIH, NSF, IES, PCORI
- Brainstorming thanks to Frauke Kreuter, Avi Feller, and Steve Pierson
Selected references, with thanks to all my co-authors


