Class 3: Experiments in social choice
Outline

• Testing a simple hypothesis
  • Controlled Randomized Tests
• Measuring the effect of a variable
• Testing a more complex hypothesis
  • Group vs. individual behavior
• Testing and comparing behavioral models
• Field experiments
• Natural experiments
• Story time
Putting a theory to the test

“Testing a theory means **checking some predictions of the theory** and that is what we do in this article.”

Putting a theory to the test

Theory: People vote rationally (maximize expected utility)

(simple) setting: two candidates, Majority voting.
Each voter can vote (at a cost of 1) or abstain. Gets $v_i(\text{winner}) - \text{cost}$.

\[ v_i(A) = 20 - 15 = 5 \]
\[ v_i(B) = 20 - 5 = 15 \]

\[ u_i(\text{vote}) - u_i(\text{abstain}) = \begin{cases} \frac{1}{2} & \text{if } i \text{ is pivotal} \\ v_i(B) - v_i(A) & \text{if not pivotal} \end{cases} - 1 \]

"DiffVote"

Putting a theory to the test

• A voter should vote if and only if $\text{DiffVote}$ is positive
• Do they?

<table>
<thead>
<tr>
<th></th>
<th>% of participants who vote</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiffVote positive</td>
<td></td>
</tr>
<tr>
<td>(17% of votes)</td>
<td></td>
</tr>
<tr>
<td>DiffVote negative</td>
<td></td>
</tr>
<tr>
<td>(82% of votes)</td>
<td></td>
</tr>
</tbody>
</table>

Putting a theory to the test

• A voter should vote if and only if DiffVote is positive
• Do they?

<table>
<thead>
<tr>
<th>DiffVote positive (17% of votes)</th>
<th>% of participants who vote</th>
</tr>
</thead>
<tbody>
<tr>
<td>82%</td>
<td>82%</td>
</tr>
</tbody>
</table>

Putting a theory to the test

• A voter should vote if and only if DiffVote is positive
• Do they?

<table>
<thead>
<tr>
<th>% of participants who vote</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiffVote positive (17% of votes)</td>
</tr>
<tr>
<td>DiffVote negative (82% of votes)</td>
</tr>
</tbody>
</table>

“The poor performance ... may be due to the fact that subjects are not very good at predicting other voters' behavior, but their decision may be consistent with their perceptions.”

Putting a theory to the test

• A **consistent** voter should vote if and only if her **subjective DiffVote** is positive
• Do they?

<table>
<thead>
<tr>
<th>(subjective perception)</th>
<th>% of participants who vote</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiffVote positive (31% of votes)</td>
<td>72%</td>
</tr>
<tr>
<td>DiffVote negative (69% of votes)</td>
<td>76%</td>
</tr>
</tbody>
</table>

Can you think of an alternative explanation?

Measuring the effect of a variable

Theory: People vote rationally
Derived hypothesis I: “always vote for q in Scenario 3”

Derived hypothesis II: “score gap should not matter”
Measuring the effect of a variable.

Measuring the effect of a variable

(Leader-bias!)
Another example

Hypothesis: Candidates converge to the median voter’s position

With full information: Downs-Hotelling model*

More surprising: also true (theoretically!) with very limited information**

Does it hold in practice?

*Anthony Downs, "An Economic Theory of Democracy" (1957)

Voting-on-a-line experiment

- Several “voter subjects” with single-peak preferences are placed on a line. Their positions and preferences are private.

Voting-on-a-line experiment

• Several “voter subjects” with single-peak preferences are placed on a line. Their positions and preferences are *private*.

• Two “candidate subjects” A and B select positions (also privately)
  • It is only announced which candidate is Left and which is Right

Voting-on-a-line experiment

• Several “voter subjects” with single-peak preferences are placed on a line. Their positions and preferences are private.

• Two “candidate subjects” A and B select positions (also privately)
  • It is only announced which candidate is Left and which is Right

• All voters vote for A or B

Voting-on-a-line experiment

• Several “voter subjects” with single-peak preferences are placed on a line. Their positions and preferences are **private**.
• Two “candidate subjects” A and B select positions (also privately)
  • It is only announced which candidate is Left and which is Right
• All voters vote for A or B
• Winner’s position and margin are announced
• Repeat for up to one hour

Voting-on-a-line experiment

• Prediction: The two candidates will converge to the median voter position

Results

• The prediction is corroborated! 👍
  • But...

Group-level vs. Individual Behavior

- Many different dynamics may lead to the same outcome
- Outcome alone (e.g. convergence/equilibrium) does not tell us what voters did
- We can form and test explicit hypotheses
  - **H1**: voters vote as if they have full information
    - Consistent with ~82% of votes
    - Less on first trials
    - But not possible!
  - **H2**: voters form beliefs on candidates’ positions using regression on recent rounds
    - Consistent with ~85% of votes

Back to Individual Behavior

• We saw in second class several “theories” of strategic voting
  1. Truthful (i.e. non-strategic)
  2. Heuristic (say, 2-pragmatist, Laslier’s Leader Rule)
  3. Rational (say, Calculus of Voting)

• Which best describes voters’ behavior?
Comparing theories

- Setting: five candidates on a line
- Voters are placed at random *known* positions
- Vote simultaneously four times on consecutive days
- Voting rules: 1R Plurality, 2R Plurality, STV, Approval

Results

• STV: >90% votes consistent with truthful voting
• Approval: ~87% of votes consistent with the Leader Rule
• Impressive!
• No competing theories
• What about Plurality?
• No theory is consistent with the votes
• Why?
Voting experiments (one shot)

Vote only once.
Winner determined based on simulated votes.

Testing individual behavior

• Each participant voted in 20-40 different polls
• Experimental design: a single-player game
• For every heuristic/model:
  • Learn voter’s parameter(s) from samples
  • Predict remaining samples
  • Measure accuracy with 10-fold cross validation

• Which models are most predictive?
Always truthful (= 3-pragmatist)

Always leader (= 1-pragmatist)

2-pragmatist

KP (fit K)
Benchmark 1: machine learning with access to all data

Benchmark 2: machine learning with access to this subject’s data

(all these models discussed on Thursday!)
Voter types

(Frequency of truthful vote)
Voter types

(Frequency of dominated vote)
• Laslier, Myerson – repeated voting games
• Field experiments?
• Tal, Meir, Gal – lab experiments
Field experiments

• Theory: more information leads to more efficient outcomes
Welcome to the AAMAS 2024 program committee...
Please bid positively on at least 40 papers...

~1000 submissions

AAMAS 2024

**Paper Bidding**

**Explanation for Choices**

<table>
<thead>
<tr>
<th>Choice</th>
<th>Explanation</th>
<th>Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>I want to review this paper</td>
<td>11</td>
</tr>
<tr>
<td>maybe</td>
<td>I can review it</td>
<td>4</td>
</tr>
<tr>
<td>no</td>
<td>I prefer not to review it</td>
<td>2</td>
</tr>
<tr>
<td>conflict</td>
<td>I have a conflict of interest</td>
<td>0</td>
</tr>
</tbody>
</table>

**Submissions**

<table>
<thead>
<tr>
<th>#</th>
<th>Choice</th>
<th>Submission</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>yes maybe no conflict</td>
<td>(anonymous). <em>Designing an Adaptive Learning Module to Teach Software</em></td>
</tr>
<tr>
<td>5</td>
<td>yes maybe no conflict</td>
<td>(anonymous). <em>Something on game theory</em> (details)</td>
</tr>
<tr>
<td>6</td>
<td>yes maybe no conflict</td>
<td>(anonymous). <em>Al for dummies</em> (details)</td>
</tr>
</tbody>
</table>
Field experiments

• Theory: more information leads to more efficient outcomes

• Derived hypothesis:
  • Revealing dynamic information on demand will incentivize bidders to pick low-demand papers*

• Partition bidders in a conference into two groups:
  • Control group bids as usual
  • Treatment group see additional information on demand

• Preliminary results in lab experiments and a small workshop**

• A Large experiment at ECAI-2023


Results

• We average over all bids (papers selected by user) in each group
• Looks like hypothesis is true!

• But wait
  • Treatment group was only slightly larger
  • Number of bids is more than twice
  • Something is suspicious
These two bidders selected >1000 papers!

They also deselected them later.
Let’s re-do the analysis

• We now only average over all bids **that were not later removed** in each group

• Effect is gone 😞
Natural experiments

• Theory: geographical distance from the polling station affects turnout
• Hypothesis: moving the station farther will reduce turnout

• Option 1: analyze correlation between distance and turnout
  • Correlation ≠ causation!!!
• Option 2: Run a controlled experiment where we move some random polling stations and measure the effect
  • Good luck with that!

A *Natural experiment* takes advantage of some difference between populations that are expected to be otherwise similar
Natural experiment (cont.)

• Theory: geographical distance from the polling station affects turnout
• Hypothesis: moving the station farther will reduce turnout

• In October 2003, LA county consolidated some polling stations
  • These stations became farther from some of their voters
• First need to show that affected populations are similar to those unaffected
  • (mostly) uncorrelated with other factors that predict turnout
• Then measure difference in turnout between groups

What if there is correlation?

Doodle example

<table>
<thead>
<tr>
<th>3 participants</th>
<th>July 2014</th>
<th>Thu 3</th>
<th>Fri 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>John (Initiator)</td>
<td>![Checkmark](10:00 AM) ![Checkmark](11:00 AM) ![Checkmark](10:00 AM) ![Checkmark](10:00 AM)</td>
<td>![Checkmark](10:00 AM) ![Checkmark](11:00 AM) ![Checkmark](10:00 AM) ![Checkmark](10:00 AM)</td>
<td>![Checkmark](10:00 AM) ![Checkmark](11:00 AM) ![Checkmark](10:00 AM) ![Checkmark](10:00 AM)</td>
</tr>
<tr>
<td>Mary</td>
<td>![Checkmark](10:00 AM) ![Checkmark](11:00 AM) ![Checkmark](10:00 AM) ![Checkmark](10:00 AM)</td>
<td>![Checkmark](10:00 AM) ![Checkmark](11:00 AM) ![Checkmark](10:00 AM) ![Checkmark](10:00 AM)</td>
<td>![Checkmark](10:00 AM) ![Checkmark](11:00 AM) ![Checkmark](10:00 AM) ![Checkmark](10:00 AM)</td>
</tr>
<tr>
<td>Karl</td>
<td>![Checkmark](10:00 AM) ![Checkmark](11:00 AM) ![Checkmark](10:00 AM) ![Checkmark](10:00 AM)</td>
<td>![Checkmark](10:00 AM) ![Checkmark](11:00 AM) ![Checkmark](10:00 AM) ![Checkmark](10:00 AM)</td>
<td>![Checkmark](10:00 AM) ![Checkmark](11:00 AM) ![Checkmark](10:00 AM) ![Checkmark](10:00 AM)</td>
</tr>
</tbody>
</table>

| Your name      | ![Checkmark](2) ![Checkmark](2) ![Checkmark](3) ![Checkmark](1) |
How do people coordinate?

• Scheduling as a form of group coordination.

• Each participant balances her own interests with the group interest.

• Do people behave strategically?
  • Problem: We don’t know their preferences!

• Idea:
  • Compare behavior to a situation where there is no opportunity to strategize
(open) Doodle example

<table>
<thead>
<tr>
<th>3 participants</th>
<th>July 2014</th>
<th></th>
<th>Thu 3</th>
<th>Fri 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wed 2</td>
<td>10:00 AM</td>
<td>10:00 AM</td>
<td>10:00 AM</td>
</tr>
<tr>
<td>John (Initiator)</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Mary</td>
<td></td>
<td>✔</td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Karl</td>
<td></td>
<td></td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Your name</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2 2 3 1
Assumption 1: Populations on both conditions are similar
Assumption 2: People did not strategize on the “hidden” condition

We had no way of testing these assumptions 😞
Making conjectures

• In which condition more time slots get approved?
• How does availability behave as time passes?

• Try to guess the result *before* doing the analysis!
• Then check and compare
Availability over response positions

The availability of a voter is the fraction of slots that she approves. The response position of a voter is the order that she participates in.
Response curve at 11

Conditioned on a slot approved by \( x \) of the first 10 voters, what is the probability that the 11\textsuperscript{th} voter approves it?
Response curve at 11

Conditioned on a slot approved by $x$ of the first 10 voters, what is the probability that the 11th voter approves it?
Response curve at 11

Conditioned on a slot approved by $x$ of the first 10 voters, what is the probability that the 11\textsuperscript{th} voter approves it?
Response curve at 11

Conditioned on a slot approved by \( x \) of the first 10 voters, what is the probability that the 11\(^{\text{th}}\) voter approves it?
Stepping up: testing varying conditions


They collect data from various non-Plurality voting instances, estimate preferences under various assumptions, and run alternative voting rules.

Simulations:

Finally, there is the method of extrapolating from real outcomes obtained under one voting procedure the likely outcomes, ceteris paribus, that would be obtained under other voting procedures. Two difficulties are associated with this method. Firstly, a prerequisite for conducting extrapolations from an observed procedure to other procedures is that the voters' preference orderings among the candidates under the observed procedure are known. Since most
Recap: No silver bullets

Hypothesis: “Borda leads to better outcomes than 2-Approval and STV”

<table>
<thead>
<tr>
<th>Method</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>pros</td>
<td></td>
</tr>
<tr>
<td>cons</td>
<td></td>
</tr>
</tbody>
</table>

Recap: No silver bullets

Hypothesis: “Borda leads to better outcomes than 2-Approval and STV”

<table>
<thead>
<tr>
<th>Method</th>
<th>Simulations</th>
<th>Lab experiment</th>
<th>Survey</th>
<th>Extrapolation</th>
</tr>
</thead>
<tbody>
<tr>
<td>pros</td>
<td>Cheap, thorough</td>
<td>Control preferences</td>
<td>Real preferences</td>
<td>Real preferences</td>
</tr>
<tr>
<td>cons</td>
<td>Arbitrary assumptions on preferences and strategic behavior</td>
<td>Small, expensive, preferences arbitrary</td>
<td>Cannot trust hypothetical answers, cooperation</td>
<td>Lack of ranked data, Assume truthful vote</td>
</tr>
</tbody>
</table>

“... we must conclude that ... all four methods are complementary and should be employed”

(some) Considerations in experiment design

- What are the treatments?
  - Between / within subjects
- Which conditions to control?
  - Between / within subjects
- Who are the subjects?
- What is the interface?
- What information subject get?
  - Truthful / deceitful
- The order of conditions
- How to set incentives?
- How to explain and/or test understanding?
- IRB
- ...
Don't run experiments

Do you have a theory?

Does theory make concrete predictions?

Are those behavioral predictions?

Is there a simple experiment to test prediction?

Is there other available data to test predictions?

Did someone else run a similar experiment?

Do you want to run experiments?

OK, but be careful...