Observational Studies

- subjects self-assign to treatment/control

- can show association between a factor and a response

- may be hidden confounding factors

"association is not causation"

eg cleaning with bleach + childhood infections.

"controlling for certain factors"

eg smoking mold in house

to try to compare groups that are as similar as possible

- may still be unrecognized confounding factors
Simpson's Paradox

Relationship between percentages in subgroups can be reversed when the subgroups are combined.

Sex bias in UC Berkeley graduate admissions

- 8,442 men applied 42% admitted
- 4,321 women applied 35% admitted

Break the applications down by major to see where the sex bias is.

- The subgroups may be more homogeneous.
<table>
<thead>
<tr>
<th>major</th>
<th># applicants</th>
<th>% admit</th>
<th># applicants</th>
<th>% admit</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>825</td>
<td>62</td>
<td>108</td>
<td>82</td>
</tr>
<tr>
<td>B</td>
<td>560</td>
<td>62</td>
<td>25</td>
<td>68</td>
</tr>
<tr>
<td>C</td>
<td>325</td>
<td>37</td>
<td>593</td>
<td>34</td>
</tr>
<tr>
<td>D</td>
<td>417</td>
<td>33</td>
<td>375</td>
<td>35</td>
</tr>
<tr>
<td>E</td>
<td>191</td>
<td>28</td>
<td>391</td>
<td>24</td>
</tr>
<tr>
<td>F</td>
<td>372</td>
<td>6</td>
<td>341</td>
<td>7</td>
</tr>
</tbody>
</table>

For most majors, women are accepted at a greater rate than the men.

- Some majors are easy to get accepted into (e.g., A, B)
- Over 50% of men applied to these majors.
other majors are hard to get in to
- women overwhelmingly (>90%) applied to these.

confounding, of sex + major when group all the students together.

---

Sample Surveys:
- how to design a sample survey
  (how to recognize a carefully designed survey)
  - bias

population - everything of interest (voters/door on campus/etc)
sample - what we've collected data about.

→ How to design the survey that generates the sample to get a representative sample that avoids bias.

→ How to use that sample to make inference about the parameters of the population
Selection bias.

Ideally, we would like a "simple random sample" - every element of the population has equal chance to be in the sample.

Selection bias is when the people you select for the sample are not representative of the whole population.

E.g., back row of the lecture is probably not representative of the whole class in terms of freshman/sophomore/junior/senior.

1936 Election

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roosevelt 43%</td>
<td>62%</td>
</tr>
<tr>
<td>Landon 57%</td>
<td>38%</td>
</tr>
</tbody>
</table>

Based on 2.4 million responses.

Why were they so badly wrong?
Selection bias.

- how did they choose the participants?

  - mailed surveys to 10M people,
    chosen from telephone directories
    club membership lists.

1936 - telephones were not common amongst poor people.

- during the Great Depression poor people were likely to vote differently than rich people.

large sample, but biased towards one part of the population.

Taking a large sample with a biased procedure does not improve the results.
Non-response bias.

2.4M responses out of 10M.
- Are the people who responded ever representative of the people who were polled (never mind representative of the population).
- Probably not.
- Middle class people are more likely to respond.

Hite Report on Female Sexuality

98% of respondents were dissatisfied with their marriage.
75% had extra-marital affairs.
but only 4% of surveys returned.
- Biased towards unhappy people.

Follow up: 93% satisfied.
70% affairs.
'1 in 2' admits to plagiarism

Michael Stothard
Chief News Editor

Half of Cambridge students have committed plagiarism as defined by the university, according to a Varsity survey. 49 per cent of students admitted that they had plagiarised work, although this differed radically between subjects and colleges. Ironically, students of the Law faculty plagiarised the most out of any subject, with 65 per cent of them breaking the university rules. The second highest was the Archaeology and Anthropology department with 59 per cent.

"It is a depressing set of statistics," said Robert Foley, a Professor in Biological Anthropology at King's College. "The college at the bottom of the Tompkins Table. St Edmund's, has the highest proportion of plagiarising students, with 89 per cent admitting to breaking the university rules. Selwyn, at the top of the Tompkins table, has the fewest number of plagiarising students.

"It stands to reason that those students who are performing less well will resort to more underhand means to get by," said a member of the University Council, the principal executive and policy making body of the university. "It is perhaps not surprising that 50 per cent of students said that the university is doing enough to punish plagiarists. "We can see why students, a great number of whom are frequently breaking the rules to their own benefit, would be keen to uphold the impression that the system is working," said a member of the General Board, the body responsible for education policy at the University.

"Sometimes when I am really fed up," said Land Economy student at Pembroke. "Google the essay title, copy and paste everything on to a blank word document and dig up the entire text. They usually end up being the best essays." 100 per cent of Land Economy and Natural Sciences students plagiarized, but the results should be taken lightly because less than five per cent of the student population admitted to plagiarism.

49%
Percentage of survey respondents who admitted to some form of plagiarism

62%
Percentage of respondents studying Land Economy or Natural Sciences who admitted to plagiarism

Cambridge application levels hit record high of 15,000

Vicky Woolley
News Reporter

The University has announced that the number of applications for undergraduate courses has reached a record high this year. For the past four years, around 13,500 students have applied to study at Cambridge, whereas this year the figure is set to be closer to 15,000. This equates to an increase of around 12 per cent, compared with an average rise of only 6.5 per cent across all UK universities.

Some subjects have fared better than others. Computer Science has seen a 53 per cent rise in applications, with the next biggest rise being in Philosophy which is up by 45 per cent. Theology and ASNC saw applications drop by 18.5 and 10 per cent respectively.

This rise in applicant numbers comes at the end of a year in which many changes have been made to the admissions process. The requirement for a foreign language GCSE has been removed because of fears it discriminated against applicants from state schools. In addition, foreign language teaching has declined rapidly in recent years. The Cambridge Application Form was also abolished from this year, which students had to fill in alongside their UCAS application, was felt to be off-putting for some students.

Continued on page 3

49%
Percentage of survey respondents who admitted to some form of plagiarism

62%
Percentage of respondents studying Land Economy or Natural Sciences who admitted to plagiarism

Richard Eyre, former Director of the National Theatre, on the ADC

Continued on page 3
How to choose a good sample?

- **simple random sampling**.

  from the list of the entire population
  - draw one element at random
  - delete that element from the list
  - repeat until have desired sample size.

  -> allows "easy" inference for population parameters & their margins of error

  -> often pretend we're doing this as it makes the math easier.

**Quota Sampling** / **Stratified Sampling**

- try to match the sample to the population

  specify # women / # men
  age distribution
  race distribution
  income distribution
  etc.
If the final choice of who to interview is left to the interviewer, the sample may still be biased.

Eg: if interview people in downtown SC at 11am on a Tuesday, mothers of small children are likely to be over-represented, even though the sample matches the population in terms of age/income/race etc.

Cluster Sampling.

A simple random sample will result in a geographically diverse set of locations - hard to travel to all the different places.
The person to be interviewed in each household is determined in advance. (e.g. youngest female over age of 18). - removes selection bias.

- allows for smaller errors with smaller samples. (Sample of a few 1000 gives error of ±3%).
Telephone Surveys.

generate phone numbers randomly.

- cell phones have made this less useful
  - large parts of the population only have cell phones, and
  can't be called for opinion polls.

levance thing nothing has changed.
Looking at Data:

- Graphical representations that can aid in understanding the distribution of data sets.

  - Range of values
  - Most likely values
  - Symmetry
  - Uni/multi modality
  - Tail behaviour
  - Outliers
Pie chart - don't do it

Humans are good at judging linear measures (distances) not angles or areas.