MA/CS 109 Lecture 13

Cautionary Tales

Midterm Thursday Oct 26
Random Samples

The Central Limit Theorem we stated requires that the sample be chosen using the Equally Likely Outcome probability model on the space of all possible samples. So each group of \( n \) people is as likely as any other to be chosen as the sample...
Random Samples

This is harder than it sounds...

If you have a list of everyone, then you can pull names out of a hat. But then you have to find everyone and they have to answer (and say what they think rather than what they think you want to hear).
Dewey Wins!

In the 1948 Presidential election, the polls (and many experts) predicted that Thomas Dewey would win...

A Chicago morning paper that had to start printing before the election results were final even printed that Dewey had won...

But there was no President Dewey.
DEWEY WINS!

Famous example—

One story goes that the polls were conducted by calling people on the phone—but in those days not everyone had a phone, so the Equally Likely Outcome model was not used to choose the sample (there was no chance for people without a phone to be in the sample).
Even worse today

While today everybody has a phone—many people do not have land lines with listed phone numbers.

(And even if pollsters get your number, do you answer? Caller ID has made life very difficult for pollsters.)
But some “polls” have even bigger problems...


“Syria After Assad Could Be Even Worse”

https://www.nytimes.com/roomfordebate/2012/02/06/is-assads-time-running-out/syria-after-assad-could-be-even-worse
The tragic civil war, which sadly continues today, had been going on for a year already then.

The article stated that a recent poll showed the 55% of Syrians wanted Assad to stay as president. This would certainly have an impact on my view of the situation...if true.

When quoting the poll, the article gave a link...
The Thread

...to another article in The Guardian (a British newspaper) which made the same statement, that 55% of Syrians wanted Assad to stay as president.

The Guardian article gave the source of the poll data—A poll conducted by the Doha Debates.
The Source

The Doha Debates gave a clear write up of its data...

And what do we see?

The poll had 1012 respondents...that’s not bad —the margin of error is $1/\sqrt{1012}$ or about 3%.
So...

So if 55% wanted Assad to stay as president, in the poll then with 95% confidence, between 52% and 58% of the population as a whole wanted Assad to stay as president.

But look more closely....
Plot thickens...

Only 211 of the 1012 people polled were from the Levant (the area including Syria)...

And only 46% of them were in Syria...so only 97 of people polled were in Syria. So the margin of error for the proportion of Syrians who wanted Assad to remain was $1/\sqrt{97} = 0.1$ or 10%.

So with 95% confidence we can only say that between 45% and 65% of Syrians want Assad to stay...
Plot thickens

So we are not 95% confident that a majority wanted Assad to stay...

But wait...how was the poll taken? How was the sample chosen?
Plot thickens

So we are not 95% confident that a majority wanted Assad to stay...

But wait...how was the poll taken? How was the sample chosen?

People from a “panel” were “invited to participate”...
The Crime!

From the panel of 220,000 people, only 1,012 “responded”...

This was not a random sample using the Equally likely outcomes model for people in Syria

Not even a random sample of the 220,000 panel members...

This was the worst of the worst...
This “so called” poll...

THIS WAS TALK RADIO!!
This “so called” poll...

THIS WAS TALK RADIO!!

I HATE TALK RADIO!!

From a relatively small group of people (listeners), a tiny proportion self select to participate by phoning in...
The evil of Talk Radio

When you listen to Talk Radio you have already selected yourself. Those who go to the trouble to call in are as far from randomly selected as possible—because of their strongly held beliefs that they want to share, they have taken the time and trouble to call in and wait to get on air...

(Note: Applies to Fox, CNN, NPR, and all the rest! The call in process selects extremes!)
Who is the criminal...

Well, the Doha Debates published their methods and the details...while you can complain that the press release wasn’t completely clear—even a little effort looking at the details reveals what the so called “poll” really is...and the details are available!
Who is the criminal...

The Guardian article writer gave the link to Doha...but either didn’t read the details (irresponsible) or misrepresented what the data said (because of the polling technique).

So lazy or evil...take your pick.
Who is the criminal

The New York Times article writer buried the information even deeper giving even more chance to mislead the public...

So more lazy or more evil...

(I heard about this case on a BBC podcast called “More or Less”...)
CHECK THE DETAILS! There is a lot of misleading data out there, but if you ask the right questions and dig for the answers, (and know a little about the proper way to do things), you can avoid being evil...
Moral

CHECK THE DETAILS! There is a lot of misleading data out there, but if you ask the right questions and dig for the answers, (and know a little about the proper way to do things), you can avoid being evil...

And TURN OFF TALK RADIO SHOWS!
Lurking Variables...

One more cautionary tale...and another use of the dangers of not using Equally likely outcomes model to choose your sample.

One of the great successes of statistics has been to link, or show correlation, between two different events...
Correlation

For example, we now know that smoking is dangerous…but this wasn’t always understood.

When I was young almost every old person I knew smoked…and they lived to be old—how could smoking be dangerous?
Correlation

The first step to showing smoking caused cancer was to show that there was a correlation between smoking and cancer. That is, that the cancer rate was higher among people who smoked than among those who didn’t smoke...

Of course, correlation does not prove causation —It could be that sick people started smoking because it made sick people feel better...
Establishing correlation

The processing establishing correlation between two events is also the first step in showing medical treatments are effective.

Here is a case study about a study of treatments of kidney stones...

(see https://en.wikipedia.org/wiki/Simpson's_paradox)
Kidney Stones

The traditional treatment, called Drug A for kidney stones has been around for years. A new treatment, Drug B, is developed, but it isn’t clear if it is better than Drug A, so a test was set up...

The results were as follows:
Drug A given to 350 patients and 273 recovered.
Drug B given to 350 patients and 289 recovered.
Drug tests

So 78% of those who received Drug A recovered while 83% of those who received Drug B recovered...

Looks like Drug B is better! (Of course, this is just a sample...we need a Central Limit Theorem result to give us a confidence interval, but turns out that with these sample sizes, we are 95% confident that Drug B is better)...
More Details

But then the study was looked at more carefully. The Doctors that assigned patients to the different drugs cared deeply about the health of their patients so they looked carefully at their conditions. Patients with large kidney stones were more often given Drug A (because the Doctors knew it worked pretty well) while those with small stones were given the experimental Drug B (wouldn’t be so bad if it didn’t work).
So if we include the division by size of stone

<table>
<thead>
<tr>
<th></th>
<th>Drug A</th>
<th>Drug B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>81 of 87 recovered</td>
<td>234 of 270 recovered</td>
</tr>
<tr>
<td></td>
<td>93% recovered</td>
<td>87% recovered</td>
</tr>
<tr>
<td>Large</td>
<td>192 of 263 recovered</td>
<td>55 of 80 recovered</td>
</tr>
<tr>
<td></td>
<td>73% recovered</td>
<td>69% recovered</td>
</tr>
</tbody>
</table>
More Details

So if we include the division by size of stone

<table>
<thead>
<tr>
<th></th>
<th>Drug A</th>
<th>Drug B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>81 of 87 recovered</td>
<td>234 of 270 recovered</td>
</tr>
<tr>
<td></td>
<td>93% recovered</td>
<td>87% recovered</td>
</tr>
<tr>
<td>Large</td>
<td>192 of 263 recovered</td>
<td>55 of 80 recovered</td>
</tr>
<tr>
<td></td>
<td>73% recovered</td>
<td>69% recovered</td>
</tr>
</tbody>
</table>

Wait!! Drug A is better in BOTH large and small stone cases! What happened?
Lurking Variables

Because patients with more difficult cases (less likely to be cured) tended to be assigned to Drug A, the overall results for Drug A were worse.

The size of the stone is a “Lurking Variable” of the original data...a variable that impacts the results.
Battle with Lurking Variables

How do we fight Lurking variables?

The problem with the data from the original study was that the Equally likely outcomes model wasn’t used to assign patients to the groups... Doctors assigned the large stone patients to Drug A. If patients has been assigned randomly to the drugs, the lurking variable wouldn’t impact the results...
Ethical problems

But the assignments were made with the best interests of the patients in mind...is it ethical to assign a patient to a group you don’t think is best for them just to do an experiment...Isn’t that really evil?

These are really hard issues...
Another example

Legal cases can also rest on similar issues.

Consider the case of a well known university (not BU) accused of bias in admission of graduate students...
The results of the study were quite clear

<table>
<thead>
<tr>
<th></th>
<th>Admitted</th>
<th>Denied</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>3714</td>
<td>4728</td>
<td>8442</td>
</tr>
<tr>
<td>Female</td>
<td>1512</td>
<td>2804</td>
<td>4321</td>
</tr>
</tbody>
</table>
So 44% of male applicants were admitted and only 35% of female applicants were admitted.

Is this bias? Does this sample represent (with, say 95% confidence) the behavior of the well known school?

The version of the Central Limit Theorem for these cases says yes.
More data

But there is a lurking variable here...

Let’s look at the data broken down by individual departments.
### Admissions by department

FROM https://en.wikipedia.org/wiki/Simpson%27s_paradox

<table>
<thead>
<tr>
<th>Dept</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Applicants</td>
<td>Admitted</td>
</tr>
<tr>
<td>A</td>
<td>825</td>
<td>62%</td>
</tr>
<tr>
<td>B</td>
<td>560</td>
<td>63%</td>
</tr>
<tr>
<td>C</td>
<td>325</td>
<td>37%</td>
</tr>
<tr>
<td>D</td>
<td>417</td>
<td>33%</td>
</tr>
<tr>
<td>E</td>
<td>191</td>
<td>28%</td>
</tr>
<tr>
<td>F</td>
<td>373</td>
<td>6%</td>
</tr>
</tbody>
</table>
Different Story

Four of the departments had actually admitted a higher proportion of women than men.

Is there still bias? Well, maybe in some departments one way and others the other ways?

Students didn’t use Equally likely outcomes model when choosing what department to apply to.
Think Harder

So maybe the University admissions were not bias...maybe the bias came much earlier when men and women were encouraged to pursue different fields because of stereotypes...

What this could point to is a systemic problem that can’t be solved by the time students start applying for graduate school...
Big picture...

Remember—

In mathematics, we do not except a statement as fact until it has a proof—until we know WHY it must be true. Examples, experiments, observations are evidence for what we might guess is true, but until there is a proof, there is only a conjecture, not a theorem.
But...

But, to give proofs, we must understand what we are saying. We must have precise formulations and definitions. We must have a “model”.

In some cases (e.g., arithmetic with integers or Euclidean geometry) the models are well accepted.
Life in a Model

But even Euclidean geometry doesn’t model the curved space-time of general relativity...

So our theorems are correct. If they don’t answer the original question, then it is the “model” we have to work on...we have to cycle back and be more careful setting up our precise description of reality.
Template for Doing Mathematics

Problem

| Model-------------------Repeat-----------------------------|

| Modify                |

Examples/Conjectures

| Model                |

Proof---Did we answer the question?---No

Yes—Fame + $$$
Review

Why do you take a course?
Review

Why do you take a course?

To change the way you look at the world! To change (permanently) your brain and how it works—which neurons and connected to which and how strongly they are connected...

Pretty scary.
Two types of changes

Every class has 2 types of topics—one type is the very specific facts or skills. You will remember these only if you continue to review them. They are presented in order to help make clear the bigger picture and the major themes.

The exam questions should reflect both types, but emphasize the major themes...
Example:

a) Does the network below have an Euler circuit?

Justify your answer.

**No**, the degree of the left node is 5 which is odd, we have a theorem that says all networks with Euler circuits have even degree at all nodes.

b) Explain in a sentence or two how you can be sure your answer to part (a) is correct.

Every time you enter a node on an Euler circuit you must leave, so there must be an even number of edges at each node.
Or, just go for the big picture...

Explain in a few sentences the differences between an accepted “Theory” in science (e.g., Theory of Evolution, or Theory of Relativity) and an accepted “Theorem” in mathematics (e.g., the Central Limit Theorem).

“A theory has yet to be disproved but a theorem has been proven...”

Proofs are permanent justification...Proof applies to every case and explains why the statement must to be true..
The details ARE useful to remember...

a) Below is the daily close price of a stock and for each day the Log of the price...using the graph paper provided, is the price increasing exponentially?

<table>
<thead>
<tr>
<th>Day</th>
<th>Price</th>
<th>Log Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>55</td>
<td>5.8</td>
</tr>
<tr>
<td>1</td>
<td>73</td>
<td>6.2</td>
</tr>
<tr>
<td>2</td>
<td>87</td>
<td>6.4</td>
</tr>
<tr>
<td>3</td>
<td>113</td>
<td>6.8</td>
</tr>
<tr>
<td>4</td>
<td>160</td>
<td>7.3</td>
</tr>
<tr>
<td>5</td>
<td>244</td>
<td>7.9</td>
</tr>
<tr>
<td>6</td>
<td>395</td>
<td>8.6</td>
</tr>
</tbody>
</table>

From the graph...

Faster than exponential growth...
But never forget the big picture...

b) Based JUST ON THIS INFORMATION, do you think an investor should buy shares of this stock? (Justify your answer.)

The shape of exponentially growing data continues grow faster and faster as it gets larger...so this speed of growth can’t go on forever!!

Absolutely can’t buy and hold this stock because this rate of growth is non sustainable!!.
Hard problems of both types...

a) Below is the graph of a population model (blue graph) where the units of population are millions of individuals. Suppose we start harvesting $H=1/2$ (that is $\frac{1}{2}$ million) per time period. What is the new model graph?
b) What does your new model predict for the long-term behavior of the population? (Answer in a couple of sentences.)
Population decreases every year to extinction...

c) Constant rate harvesting is a simple way to put harvesting into a model—but is it really realistic? In what situation might constant rate harvesting be difficult? (Answer in a couple of sentences.)
Harvesting at constant rate is difficult when they are rare...
Remember....

Reading the answers to sample problems and homework is like watching somebody else exercise. Easy to do, but it doesn’t make you stronger.

Talking about problems with other people, sharing ideas and helping each other is both useful and fun...but remember, on the test, you will be by yourself, so make sure you understand.