Section 13: What can go wrong

STOR 155.02, Spring ’21

updated 2021-04-27

What you will learn

- is in the title

Resources

- links in slides

Two lectures remain

- April 27, May 4

Last data/webassign homeworks are this week

May 6 class period is extended OH for final

Final project due May 11, 3p.m. sharp

Time to review: Common mistakes keep costing you points

Q5, hw8: Compare the sample s.d. of the mean and the s.d. derived from the CLT formula. Explain why your answers do or not not make sense together using concepts from class.

:-(

"It makes sense that they are different because the CLT uses the sample standard deviation in the calculation."

:-) **thanks to Kaylee Tackett**

"My answer for Q4 makes sense because the CLT [says] a sampling distribution of the sampling means approaches a normal distribution as the sample size gets larger, irregardless of the shape of the distribution. ... Our computations with the CLT is only valid on a size of data that is substantial enough to reflect our actual population."

We've done a bunch of examples showing the kinds of questions you can answer with hypothesis tests.

But those idealized situations don't always work out so well.

There are many cases in published research where conclusions drawn from hypothesis tests are nearly worthless.

Goal of this lecture

- get a feel for how to spot problems with hypothesis testing
- explore assumptions a bit more

Some examples give broader problems than just for hypothesis testing, such as bad data

'p-hacking'

538 p-hacking demo

Question: Is the the U.S. economy better when Democrats are in office, or worse?

Data

- who's in office: President, Senate, House, governors
- economy: GDP, inflation, employment, stock prices

Multiple difference of means tests you could do

group (Dem v Rep)	outcome
President	GDP
Senate	GDP
President	employment

etc.

p-hacking

- finding a group and outcome combo
- so that the difference of means test is statistically significant
- at some 'standard' level

easy when the **question is so poorly defined**

This is actually a problem...

"Women are more likely to wear red or pink when at peak fertility"

"Men with larger arms of higher socio-economic status more likely to oppose wealth distribution policies"

"Kids are more likely to eat fruit if it has Sesame Street stickers on it"

"People who perform 'power poses' subsequently have increased testosterone and cortisol levels"

(original 'power pose' research paper here)

```
_________________________
< y u click bait tho? >
-------------------------
    \   ^__^
     \  (oo)\_______
        (__)\       )\/\
            ||----w |
            ||     ||
```

Let's p-hack the affairs data

this is what you got in homework

p value for two-sided hypothesis test of some affairs in `young` (age <=32) vs not young groups.

```
## [1] 0.3713905
```

but we demand significance!

Looking for significant difference of `some_affairs` between people with and without children **within each young** group

young	children	mhat	shat	n
FALSE	no	0.2500000	0.4472136	16
FALSE	yes	0.2731959	0.4467535	194
TRUE	no	0.1483871	0.3566356	155
TRUE	yes	0.2966102	0.4577340	236

young	p
FALSE	0.8419298
TRUE	0.0003356

Why is this a problem?

You've gone in search of a conclusion your data will support

rather than asking whether your data support a pre-determined conclusion

> > There are many roads to statistical significance; if data are gathered with no preconceptions at all, **statistical significance can obviously be obtained even from pure noise** by the simple means of repeatedly performing comparisons ...
> >
> > Andrew Gelman, Eric Loken, in the American Scientist

Conclusions determined by p-hacking probably are not replicatable in a different study

which was part of what exposed power poses as fraud

and exposed Brian Wansink's masterclass in p-hacking

Garbage data

Bad data ⟶ bad conclusions

article: "Automated inference on criminality using face images"

data

Authors use ~1,100 'non-criminal' face pictures scraped from the internet and ~700 mug shots of convicted criminals.

claim

Authors "empirically establish the validity of automated face-induced inference on criminality ... discriminating structural features for predicting criminality have been found by machine learning." The algorithm found that **criminals have shorter distances between inner corners of the eyes** (d), smaller angles between nose and mouth corner, and more upper-lip curvature.

problem: wouldn't you frown in your mugshot, and smile in your office headshot?

Source: Calling bullshit

Moral: Don't be bad

At least try not to be bad. It's usually not so hard.

PollEv.com/brendanbrown849

poll closes at _________

five more minutes