Introduction to Statistics I

Instructor: Jodin Morey moreyj@lemoyne.edu

Previous Lecture

- Proportion/Distribution (dist)
- Bar Charts & Dist's
- Dot plots & Dist's

Topic 3a: Drawing Conclusions

Vocab So Far:

- Observational Unit (obv-unit)
- Variable (var)
- Research Question (RQ)
- Distribution (dist)

Example: Elvis Presley died on August 16th, 1977.

Pretend it's 1989, when there was a conspiracy theory that Elvis faked his death.

RQ: What % of American adults believe Elvis Presley faked his death?

Obv-unit?/var/var type?

Sampling

What's best way to gather this data?

Listeners of 100+ radio stations called a 900 number ($2.50 per call) to voice their opinion of whether Elvis was really dead.

Will this **sampling method** answer original RQ?

——Generalizing...—-

Sample vs. Population

Population is collection of obv-units we're interested in.

Sample is subgroup of obv-units from whom data is collected.

Sample Size (denoted *n*) is # of obv-units in sample.

Statistic is # that summarizes our data. It's calculated from **sample**.

Parameter is # that summarizes the property of **entire population**.

Example: Call-in survey found 56% of callers thought Elvis was still alive.

Does this reflect views of population?

56% was calculated from people who called in, and so is a statistic.

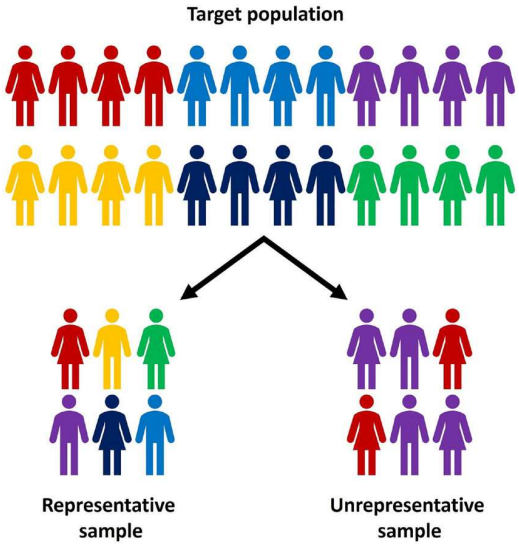
True % of Americans who believed Elvis was alive in 1989 is the paramater.

Parameter is unknown!

Ideally, statistic gives us knowledge about parameter.

Topic 3b: Drawing Conclusions

With our sample, can we draw a conclusion?



Sample Bias: For our statistic to be a good estimate of our parameter,
our sample must be **representative** of population.

If sample is gathered in a way that makes it *not* representative
(systematically overrepresents certain segments of population and underrepresents others),
then sampling method is **biased**.

What leads to biased samples?

- Sampling from only (nonrepresentative) part of the population

- Convenience Samples (lazy sampling)

- Voluntary Samples (self-selection)

- Allowing non-response

These result in members of the population having unequal chances of being in the sample.

Ideally you have a list of *all* obv-units in population!
This list (if it exists) is called the **sampling frame**.

Sampling frame: boys/girls

Having a **representative sample** is a necessary ingredient in drawing good conclusions.

Cause & Effect

Example RQ: Does our basketball team perform better (win more games)
at home games when there's a sellout crowd?
Obv-unit/var/var type?

Obv-unit: Home games

Var(s): Whether or not they won (cat/binary)
Whether or not there was a sellout crowd (cat/binary)

Poll:

bit.ly/introstatsdata

Poll: Sellout Crowds

Data: Oklahoma City Thunder, 2009

- **Sellout Crowds**: 3 wins, 15 losses

Wins $\frac{3}{18} = 0.167$. So, 16.7% win rate.

- **Non-Sellout Crowd**: 12 wins, 11 losses

Wins $\frac{12}{23} = 0.522$. So, 52.2% win rate.

——New Idea—

- **Explanatory Var**: Var we hypothesize is causing an effect.
- **Response Var**: Var we hypothesize is being affected.

Explanatory Var ————> Response Var

Back to Example: The data is against our hypothesis.

Thunder appears to lose more frequently in sellout crowds. Why?

Thunder was new in 2009. Maybe sellout crowds weren't coming to see *them*?

New Hypothesis

Sellout crowds happen more often when better teams play against the Thunder.

The Thunder (obviously) loses more frequently to good teams.

New Analysis: **Sellout/Good Opponent**

Do sellouts occur more frequently for good opponents?

Good opponents can be defined as teams that win more than ½ their games.

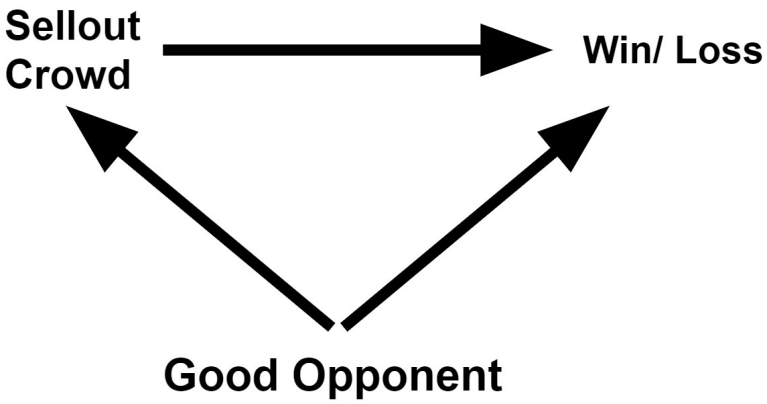
22 games were against good opponents. 13 games were sellouts.

$\frac{13}{22} = 0.59$ or 59%.

19 games were against bad opponents. 5 were sellouts.

$\frac{5}{19} = 0.26$ or 26%.

Conclusion: "Good opponent" is a **confounding var** in the relationship between sellouts & wins/losses (it affects both response & explanatory).



Var Types

Explanatory Var: Var we hypothesize is causing an effect.

Response Var: Var we hypothesize is being affected.

Lurking Var: Var we didn't measure that could affect **response** var.

Confounding Var: Var that could affect both **response** & **explanatory** vars.

Example: Childhood Obesity and Sleep.

A 2006 article found children who reported sleeping more hours a night were less likely to be obese than children who reported sleeping fewer hours.

Explanatory ——————————> Response ...

Sleep Obesity

Potential Lurking Vars? ...

Medication	Economic Standing	Self Esteem
Diet	Family History	Race
Income	Mental Health	Schooling
Exercise	Social Environment	Genetics
Nutrition	Quality of Sleep	Food Availability

—— Generalizing... —

Observational Study: A study in which data is collected and observed, but the vars are not controlled in any way by the researcher.

(eg: researcher doesn't randomly assign a treatment, or place subjects into different groups).

Observational studies cannot control for confounding variables.

Thus, observational studies can examine **relationships between vars**, but ...

❗ It can never conclude a **cause and effect relationship**!

When can you reasonably draw cause-and-effect connection between explanatory and response var?
When researcher actively and randomly imposes explanatory var on obv-units.

❗ Since we haven't the time to engage in these group activities in class today,
review these activities on your own or in study groups to ensure your understanding.

Activity 3-2

What did we learn?

- Sampling from a Population & Sampling Bias
- Explanatory/Response/Lurking/Compounding Vars
- Observational Studies & Cause/Effect