Making sense of ‘study says’

Or: Why you’ll never say ‘within the margin of error’ again

Fred Vultee/Wayne State University

Why and wherefore

- We’re in an era in which news gets less writing time and less processing time
  - I don’t like that, but ...
- We’re in a political season –always a good time to get picky about survey research
  - But people will misuse research for ideological or commercial reasons at any time
- Selfish reason? I want the people on my side of the newsroom to own their stats as thoroughly as the sports department owns its stats
  - Margin of error is as easy to calculate as ERA, so ...

1.96 √(.25/n)

- Produces the maximum margin of sampling error at 95% confidence for a random sample of size ‘n’
- Wasn’t that painless?
- But let’s get back to this magic tool of awesome supernatural power after a detour into how studies say what they say – and what we can do about it

What we'll do

- Talk about what it means for a study to 'say'
  - Different sorts of quantitative studies
  - Different sorts of qualitative studies
- Some basics of statistical reasoning
  - Correlation vs. cause
- Statistical hygiene
  - Significance, error, randomness
  - Tests and results
- What kinds of tests say what they say
- What experts to call in and when
- And some tips for polling season

Different sets of priorities

Journalism and science have some similar goals, but also some radically different priorities

- News likes to be conclusive; research is incremental
  - No poll story was ever hurt by becoming duller
- News cuts through nuance; research is in the details
  - 'The science reporter is a specialist in journalism but a generalist in science'
- News is about stuff that happened; science starts with the assumption that nothing happened
- News needs to 'balance'; science doesn't
  - We have to take into account competing explanations, but we don't have to pretend flat-earthism is "science"

Assorted ways of knowing

- Intuition ('my gut reaction') can be wrong – what your gut says might be contrary to what's known.
  - Based on feelings, easily influenced by your mood
- Tradition: 'That's the way we do it around here'
  - Those people in Salem? *They weren't witches*
- Experience: 'I've been in journalism since ...'
  - Experience is selective and uncontrolled
  - It's also influenced by our prejudices
- Or empirical: Based on what we see and measure
  - Falsifiable and replicable
  - Generalizable – at least, in the aggregate

Follow the money?

Some people think the first and last thing you should do is see who paid for a poll (or experiment, or ...)

- Yes and no. More important...
- What kind of data did you gather, and how?
- What tests did you run on those data?
- What did those tests show?
  - Did you report the nonsignificant findings too?
- How well did you stay in bounds when interpreting?
  - 'Study says' vs. 'reporter says study says'
- Are you seeing something on its merits, or ...?

Step 1: Compared to what?

Is 'study says' the day's best piece of science?

- Three guesses on when this study appeared
  - The study focused on the 6,783 people who died in car crashes in the U.S. over the last 30 years on April 15, compared to the week preceding tax day and the week that comes after. The results found a nationwide average increase of 13 fatalities on tax day, which is a 6% increase.

Fatal car wrecks jump on tax day

CNNMoney

4 comments

By Aaron Smith @CNNMoney April 11, 2012: 9:43 AM ET

Recommend 56 Tweet 8 Share 1 Email Print

NEW YORK (CNNMoney) -- The odds of getting into a fatal crash increase by 6% on tax filing day, according to a study published Wednesday in the Journal of the American Medical Association.

- Capable science meets silly journalism
  - Calling a 6% increase a "jump" is pretty common
  - Here, it's a case where the percentage increase is irrelevant without the baseline number

Step 2: What kind of study?

- *Qualitative* research is suited for 'why' questions
  - Focus groups: Why did that ad make you feel the way you do about the candidate?
  - Participant observation: What do people get out of playing massive multiplayer online games?
- *Quantitative* research is about 'how many'; it counts stuff and draws statistical conclusions
  - Content analysis: What proportion of primary ads are negative compared with general election ads?
  - Experiment: Do people remember more details from stories with video than from stories without video?
  - Survey/CA: How likely are adolescents to start having sex at age 15 if they listen to naughty music?

‘How many’ ≠ ‘too many

“Don’t write a policy paper ...”

- A study that tells you ‘how many’ doesn’t tell you ‘too much’ or ‘too many’
- Studies don’t tell you there’s too much violence on TV or that campaign ads are too negative this year
- We talk about how much violence there is, and we might suggest steps that might work if you want to reduce TV violence, but – we report, you decide
- And public opinion isn’t really relevant

Correlation ≠ cause

'If you want proof, go to seminary'

- Quantitative research is about probability, not proof
- We want stuff to be compelling and useful, but not at the expense of accuracy
- Three steps needed to demonstrate causality:
  - Something has to change (sounds dumb, but ...)
  - The thing that changes has to come after the thing that changed it
  - Confounding explanations have to be ruled out
- Because research is probabilistic, be careful of saying any study applies to 'you,' or 'moms,' or 'kids'

‘False cause’ in real life

The developments [Sunday] came just a day after the embattled Thomas denied every one of Hill’s allegations. His testimony was so persuasive that a Washington Post-ABC News poll conducted Saturday night showed that 55 percent of the 515 people interviewed were inclined to believe him, not Hill.

--Knight-Ridder

- How many people sat around on a Saturday afternoon watching the Clarence Thomas hearings?
- How many of them reported any attitude change related to the testimony?

Step 3: Do we know it's real?

- 'Margin of error' is a wayward cousin of statistical significance
- Statistical significance is an arbitrary state that has nothing to do with importance
- It's the level of confidence at which we agree that our results reflect what we're observing or manipulating – not chance or error
- Conventionally, it's 95%

LOS ANGELES Eating red meat – any amount and any type – appears to significantly increase the risk of premature death, according to a long-range study that examined the eating habits and health of more than 110,000 adults for more than 20 years.

Statistical vs. practical

- The flip side of statistical significance is practical or clinical significance
- "Significance testing" is the process in which we compare averages, or proportions, or sequences to determine whether they're real enough to meet the arbitrary cutoff of 95%
  - If so, the difference is "significant" – even if small
- Significance tests are often reported with a measure of effect size, or how much of a difference the condition or treatment makes
  - Proportion of the outcome (LSAT score) that's associated with the treatment (taking a prep class), for example

Step 4: Now that we've caught it ...

Two broad categories of stuff are tested for significance

- *Correlational* research looks at the extent to which two (or more) variables are related
  - More spending on news = higher quality product
  - Higher degree of press freedom: lower chance of starting an armed conflict
- *Group comparison* research looks at differences in the value of a variable among groups
  - Are Romney v. Obama preferences different since Santorum dropped out?
  - Are kids who see violent cartoons more violent?
  - Does the tone of editorials change after a war begins?

Putting tests with methods

- Correlational questions
  - Correlation (either numbers or rank-orders)
  - Regression, which fits cases to a prediction line
- Group comparisons
  - Averages: T-test or analysis of variance
  - 'Nominal' data: chi-square
- Look for significance levels *and* effect sizes
- You can do several of these for free at home
  - Is laughing at the instructor's jokes associated with GPA?
    - Enter those two columns in an Excel sheet
    - Formulas=>more functions=>statistical=>correl
    - Enter the arrays and you get a correlation coefficient

laughs	grade
5	2.5
15	4
7	2.5
10	3.5
8	3.5
12	3.7
3	2
12	2
4	2.5
8	3.2

- Sen. Crook's office calls to complain that your coverage of him has gotten worse since that pesky indictment Jan. 31. Has it?
- Method: Chi-square
- http://vassarstats.net/
- Go to *Frequency data*, then *chi square, Cramer's V …*
- Select 3 rows and 2 columns, enter your data, press 'calculate'

	Pos	Neg
Dec	13	8
Jan	9	16
Feb	7	19

Select the number of rows:	2	3	4	5	3
Select the number of columns:	2	3	4	5	2

Data Entry

	B_1	B_2	B_3	B_4	B_5	Totals
A_1	13	8	-----	-----	-----	21
A_2	9	16	-----	-----	-----	25
A_3	7	19	-----	-----	-----	26
A_4	-----	-----	-----	-----	-----	-----
A_5	-----	-----	-----	-----	-----	-----
Totals	29	43	-----	-----	-----	72

Reset Calculate

Chi-Square	df	P
6.2	2	0.045

Cramer's V = 0.2934

No message for this analysis.

Step 5: What's being looked at?

A 'census' looks at everything in the population. A sample looks at a subset

- A census is often expensive and time-consuming, so we sample
- Correctly executed, a small sample can say a lot about a huge population
  - "900 randomly selected likely voters" can predict national results as accurately as statewide results
- Nonprobability samples call for much more caution
  - Self-selecting polls are worthless by definition. Shun them!

AP

Military Times poll shows a landslide of support for McCain -- who captures 68 percent of the military vote to Obama's 23 percent. | VIDEO

Within Margin of Error

FOX News poll has Obama up by 3 points

Days before Democratic Convention, Obama has slim 42 percent to 39 percent edge over McCain | FULL RESULTS *(pdf)*

McCain Out Front

FNC poll: McCain leading Obama 45% to 42%

Latest FOX poll shows shift among independents giving McCain slim lead after GOP convention | FULL RESULTS *(pdf)*

- FOX News Poll: Sarah — Just Plain Popular
- Report: McCain a Hero to Taxpayers
- Obama: 'Lipstick on a Pig' a 'Phony Controversy' | VIDEO
- McCain Camp Launches 'Lipstick' Ad
- FOX FORUM: Brits for Barack — Who Gives a Toss?

- At left, Fox uses a self-selecting sample from a nonrepresentative population. Result: Fake news
- At right, competent poll meets partisan journalism: Identical results are 'out front' for the good guy, 'within margin of error' for the bad guy

Sampling and error

- We can measure sampling error, but it's not the only one to be concerned about
- Cell phones have increased the chance of 'coverage error': something nonsystematic in how households are sampled
- 'Nonresponse' error comes about when there are nonsystematic ways in which people do or don't pick up the phone
  - People my age have the 'answer the ringing phone' habit
- 'Measurement' error comes about from interviewer errors or faulty question design

Roper and the Holocaust question

"Just three days before the dedication of the U.S. Holocaust Memorial Museum, a survey ... revealed that one in three Americans is open to the possibility that the Holocaust never occurred at all." (LA Times, 4/20/93)

"Does it seem possible or does it seem impossible to you that the Nazi extermination of the Jews never happened?"

65% Impossible it never happened
12% Don't know
22% Possible it never happened

"Does it seem possible to you that the Nazi extermination of the Jews never happened, *or do you feel certain that it happened*?

91% Certain it happened
8% Don't know
1% Possible it never happened

How about the ‘150 students’?

- Generalizing from students to a population on a single variable is a bad idea
  - Political party preferences, musical tastes, etc.
- In a ‘multivariate’ comparison, you’re on safer ground
- ‘Political interest’ is distributed differently among students than in the population at large
- But if you divide students on news use and measure their interest in a story, you’ll probably find that interest varies between heavy and light news users in about the same way it does among all adults

Experiment v. survey

- Those features are what make a study an experiment: comparing groups that get different levels of a treatment or get treatment vs. placebo
- Experiments make the best claims for causal relationships:
  - Outcome is measured after treatment, so they're in the right order
  - Highest degree of control over confounding factors
- That can come at the expense of 'validity':
  - People rarely watch 25 PSAs in a row at one sitting
  - More control can mean less realism

Paved with good intentions

- Here's a good study (content analysis with a three-stage survey) that tempted editors into sin:
- *"Teens whose iPods are full of music with raunchy sexual lyrics start having sex sooner than those who prefer other songs, a study found. Its influence on behavior appears to depend on how sex is portrayed, the researchers found."*
- ... please tell me your hed didn't say

Lyrics incite teens to sex or

Raunchy lyrics spur teen sex, study concludes

What do the authors say?

Our results suggest that the relationship between exposure and behavior <u>may be causal</u> ... however, <u>our correlational data do not allow us to make causal inferences with certainty.</u>

- Teens were asked which names on a list of 'top Billboard artists' they'd listened to, and how much, since school began
- Lyrics from those artists' most recent albums were coded for sexual content
- In the next interview, teens were asked about their sexual behavior

Chicken and egg?

Or two eggs from some antecedent chicken?

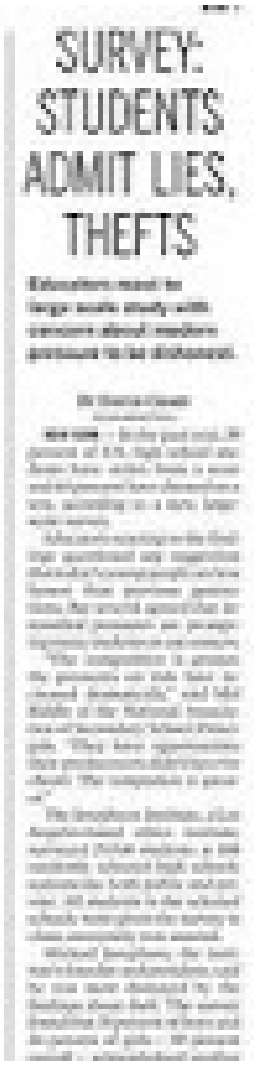
- So what went wrong to a perfectly good study relating sexual behavior and music exposure?
  - Ignoring the boundaries of causality
  - Featurizing: Just because you can say ‘iPod’ in the lede doesn’t mean you should
  - Talking down: ‘Raunchy’ may be the word everybody understands, but it misses the point of ‘sexual *and degrading*’
- The headlines are a category of sin all their own:

Teens are jumping to the jive

Step 6: Not just who pays …

… but who benefits

- People who sell their services benefit when they're on the front page!
  - People buy their services directly
  - People are impressed by their skillz and thus make them into authority figures
- *"Experts agree that dishonesty on surveys usually is an attempt to conceal misconduct."*
- No, they don't. They agree on social desirability bias


SURVEY: STUDENTS ADMIT LIES, THEFTS

But there's more!

Ever-morphing English nears its millionth word

- It's more insidious when people use our bona fides to make themselves look like 'experts'
- Here, a guy with no credentials uses a bean-counter in MS Word to convince CNN he's a 'language guru'
- What do linguists say about the professor-length in the Obama speech sentences?

- *"They were almost as long as the ones that President George W. Bush, that notorious pointy-headed intellectual, used in his 9/15/2005 speech to the nation about Hurricane Katrina."*

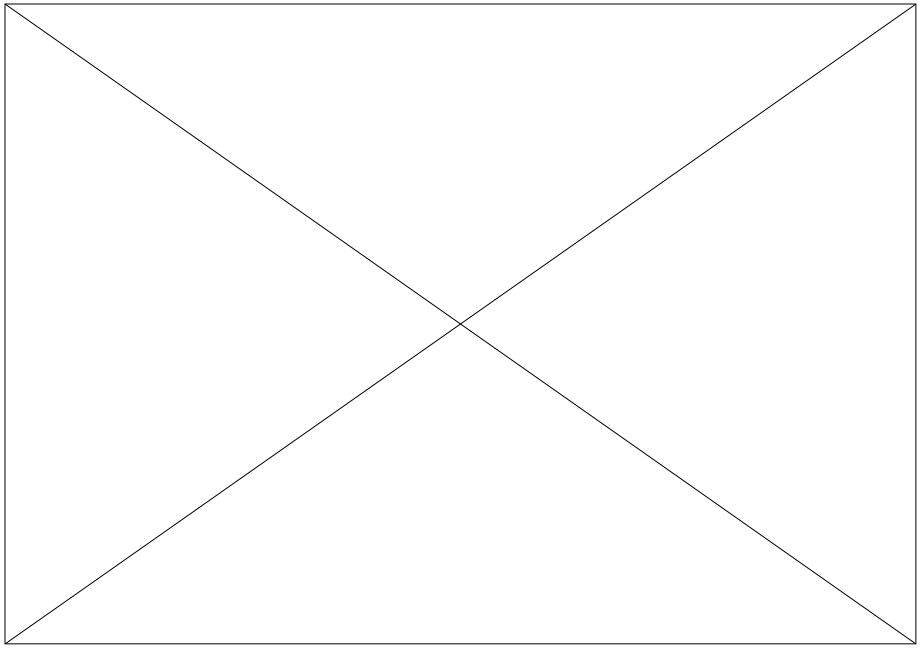
  Mark Liberman, Language Log

Back to polling

- People like to deceive journalists with survey research, but it's also one of the areas in which it's easiest to fool ourselves
- Polling is expensive; it's nice to get your money's worth
- News is about stuff that happened, rather than stuff that probably didn't
  - 'Clawed his way back into contention' sounds better than 'probably gained a little but maybe didn't'
- And good intentions too can fall for what a study ought to say, rather than what it says

"Before the advertisement aired, Gantt was leading in several polls, but his support plummeted as Helms' televised assault hit the airwaves. Helms won the election, 53% to 47%."

== Los Angeles Times, 2/21/95, "Affirmative action poised to become political divide"



- Well, not really. Gantt's support was pretty steady from three weeks before the ad to the election itself
- What did change was Helms' vote
  - And, of course, the undecideds
- Boring idea consistent with evidence and theory?
  - Helms voters/leaners were calling themselves undecided

Doesn’t the ‘margin of error’ help?

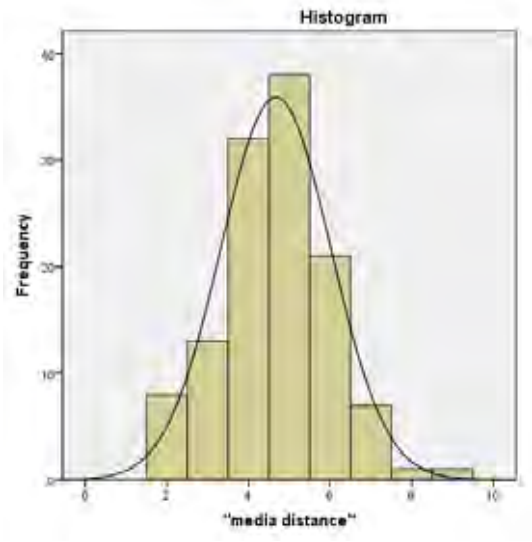
- Only if you use it right. So let’s get back to its roots with some sampling theory
- Random samples taken from a normally distributed population eventually form a normal distribution themselves
- In other words, if you ask a few dozen different samples of 900 registered voters whether they approve or disapprove of the president, the results will form a bell curve with the real population value in the middle

How it works

- The formula √((p)(1-p)/n) finds the 'standard error' for a sample of size n, where p and (1-p) – or 'q'– represent a proportion and ... all the leftovers
  - The √(.25/n) formula works because .25 is the largest value that the product pq can take: a 50-50 split
- About two-thirds of samples will fall with one standard error on either side of the population value
- Multiply the SE by 1.96 and you get an area that includes 95 percent of samples
- That's the "95 percent confidence level" that all poll stories should mention, without fail

Frequent margins of error

N	max error (at 95%)
100	+/- 9.8 points
200	+/- 6.9
300	+/- 5.7
400	+/- 4.9
500	+/- 4.4
800	+/- 3.5
1,000	+/- 3.1

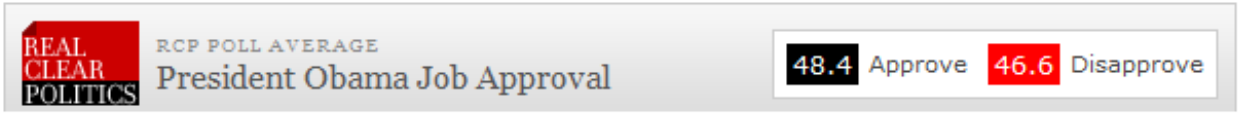


- You can see how the 95% curve will get a little tighter every time you add interviews – but with diminishing returns

President Obama Job Approval

Polling Data					
Poll	**Date**	**Sample**	**Approve**	**Disapprove**	**Spread**
RCP Average	**3/20 - 4/10**	--	**48.4**	**46.6**	**+1.8**
Rasmussen Reports	4/8 - 4/10	1500 LV	48	51	-3
Gallup	4/7 - 4/10	1500 A	45	45	Tie
ABC News/Wash Post	4/5 - 4/8	1103 A	50	45	+5
CNN/Opinion Research	3/24 - 3/25	1014 A	51	45	+6
McClatchy/Marist	3/20 - 3/22	846 RV	48	47	+1

See All President Obama Job Approval Polling Data



- Starting to see why any attempt to report a meaningful 'average' of polls is bogus?
  - Averaging different-sized samples is like averaging a guess and an estimate
  - Don't even start on averaging unlike populations
- You can, though, compare polls longitudinally

A made-up real-life example

- Remember, the margin of sampling error applies to both the proportions you report:

	3/24	3/31	4/7	4/14
Crook	41	40	42	44
Liar	43	44	42	40

- With 640 voters in each survey, we have a margin of sampling error (at 95% confidence) of 3.9 points
- What has happened in the past three weeks?
  - Crook surged into the lead!
  - Or nothing has changed. Both candidates are stuck on 42, and about two-thirds of samples will be within 2 of that

Sorry about that margin

- The margin of error is part of an estimate. It's not a magic tool that makes a race 'too close to call'
- Sampling error for subgroups is always larger than for the whole sample.

  Senate poll shows Crook leading Liar, 53-47

  – 1,200 adults (+/-2.8, 95%): Crook 51.2-55.8; Liar 44.2-49.8
  – Half of them men (+/- 4.0, 95%): Crook 49-57, Liar 43-51
- And never generalize from a poll to a population it doesn't represent (registered voters ≠ likely voters ≠ "all right-thinking Americans")

Some standard research hygiene

We can't stop people from doing bad things with numbers. But we can keep them from doing bad stuff to our readers with numbers.

- Does the story say what kind of study is involved?
  - Does it identify what was studied and how it was tested?
- Can you get to the published/presented work itself?
  - Google Scholar, good for abstracts
  - Anybody on the staff taking night classes? Library access!
- Confidence levels and confidence intervals are essential to reporting on public opinion. If 'no margin of error was reported,' don't trust the study

More safety tips

- Personalize/featurize with extreme caution. No study says what ‘you’ will do
- The dull answer is probably better than the lively one
- Look for the ‘limitations’ section. When the authors say they don’t claim a causal relationship, listen
- Not all experts are equal
  - Not all of them are even experts
  - Make sure they’re experts in what they’re talking about
- Political speech isn’t science. If the official comment on an empirical projection about climate change is an ad hominem argument – “why do you hate freedom?” – you may ignore it