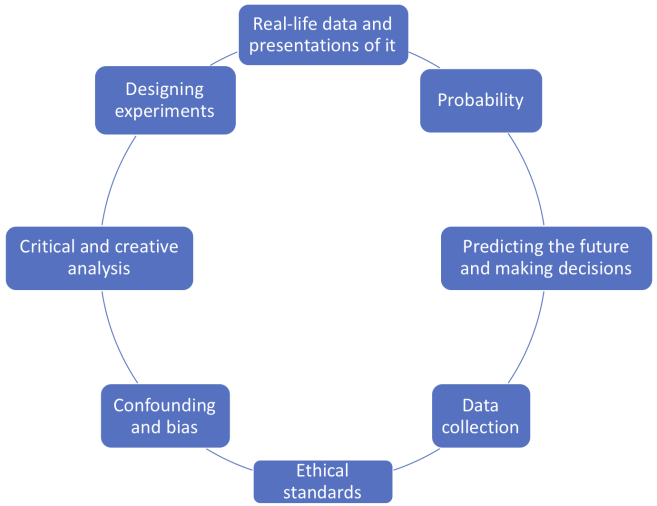
Case Studies

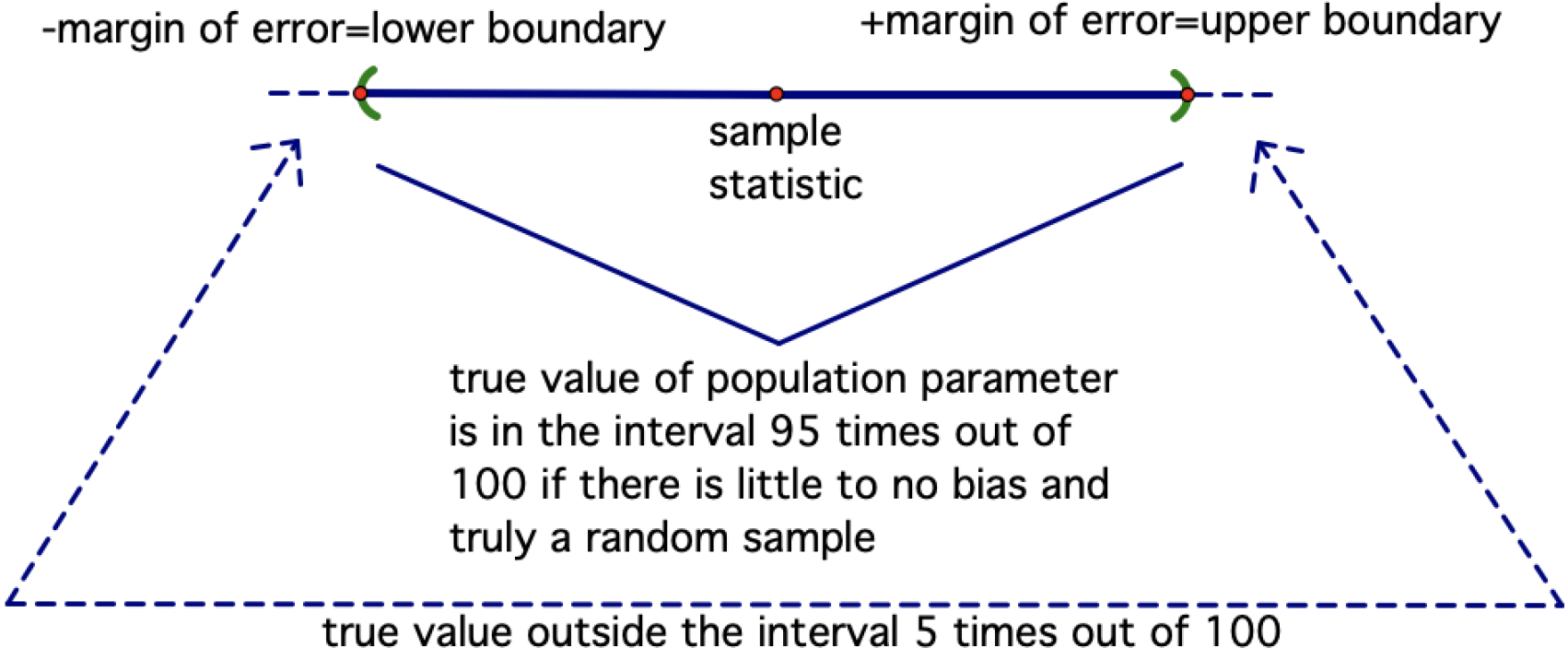
- collecting data: reproducibility, consensus, and random sampling
- presenting data: entire data set versus numerical or visual snapshots of it
- expected value: weighted probabilities for decisions
- mean and median: central tendencies
- box plots: comparisons
- regressions: correlations
- confidence intervals: uncertainty in even the best polls



all can be subject to bias and distortion, and are definitely subject to probability and random variations

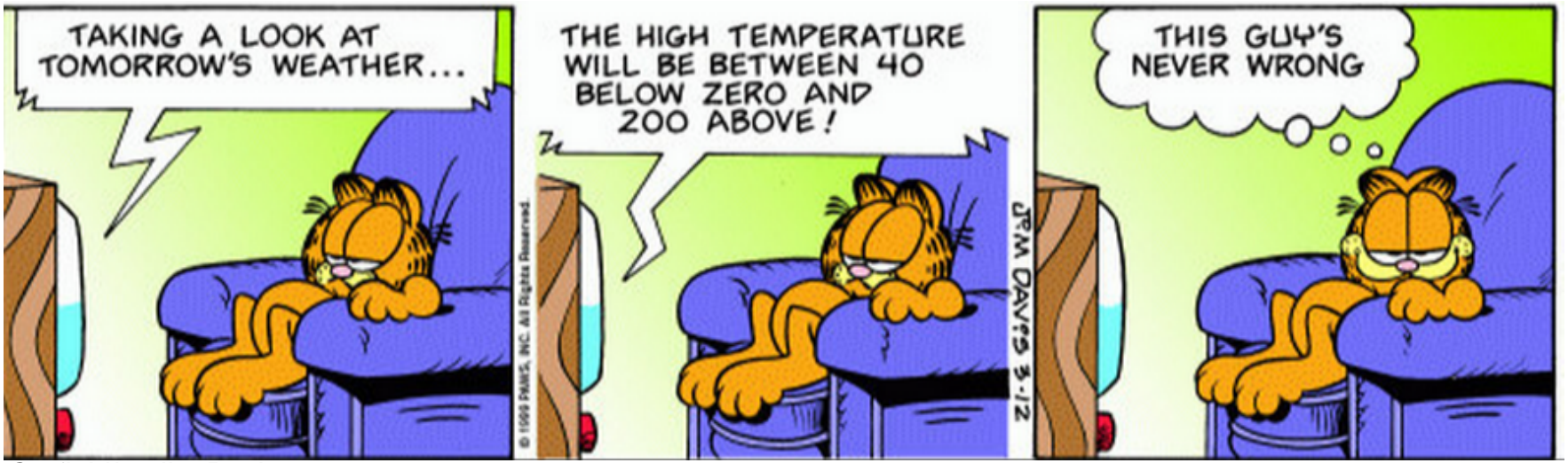
Confidence Levels

- If there is little to no bias and truly a random sample, then *x*% confidence interval is a numerical interval generated by a procedure that *x* times out of 100 will produce an interval that contains the true value for the entire population.



- Likelihood of the sample outcome—no way to know which intervals contain the true percentage and which don't

Margin of Error



Garfield by Jim Davis `https://garfield.com/comic/1999/03/12`

- margin of error gives a range the actual percentage is likely to be within if the sample size is large enough. Higher confidence level has a wider interval.
- For a 95% confidence interval, a sample of size n will have margin of error approximately $\frac{1}{\sqrt{n}}$ (conservative estimate).
- We check for overlaps in the intervals in order to evaluate the statistical validity of headlines and statements in polls

Gallup Polls

SURVEY METHODS

Results for this Gallup poll are based on telephone interviews conducted March 1-5, 2017, with a random sample of 1,018 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia. For results based on the total sample of national adults, the margin of sampling error is ±4 percentage points at the 95% confidence level. All reported margins of sampling error include computed design effects for weighting.

Each sample of national adults includes a minimum quota of 70% cellphone respondents and 30% landline respondents, with additional minimum quotas by time zone within region. Landline and cellular telephone numbers are selected using random-digit-dial methods.

SOCIAL & POLICY ISSUES MARCH 31, 2017

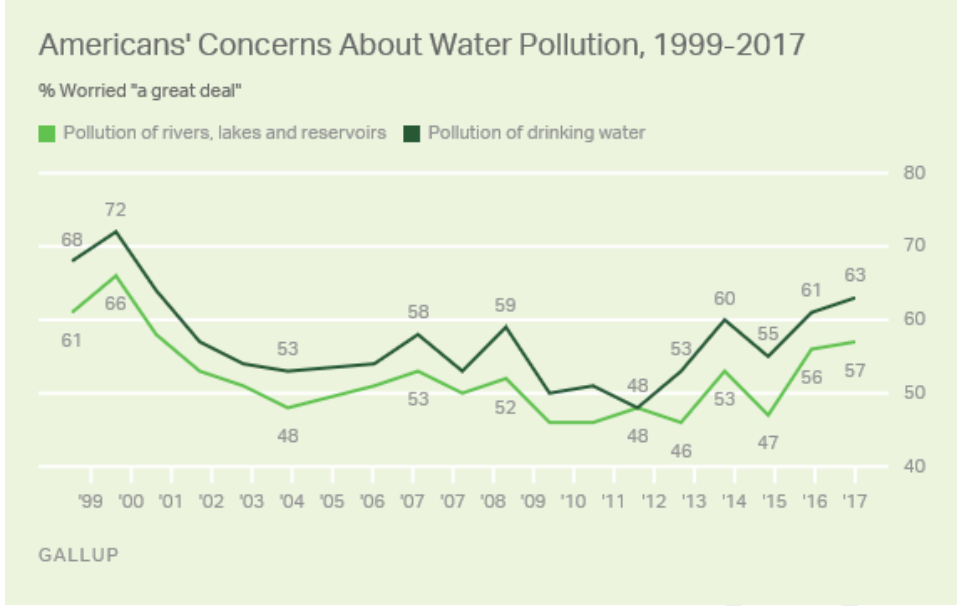
In U.S., Water Pollution Worries Highest Since 2001

BY **JUSTIN MCCARTHY**

AMERICANS WORRIED A GREAT DEAL
ABOUT POLLUTION OF DRINKING WATER

63%

GALLUP, MAR 1-5



Gallup Polls

SURVEY METHODS

Results for this Gallup poll are based on telephone interviews conducted March 1-5, 2017, with a random sample of 1,018 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia. For results based on the total sample of national adults, the margin of sampling error is ±4 percentage points at the 95% confidence level. All reported margins of sampling error include computed design effects for weighting.

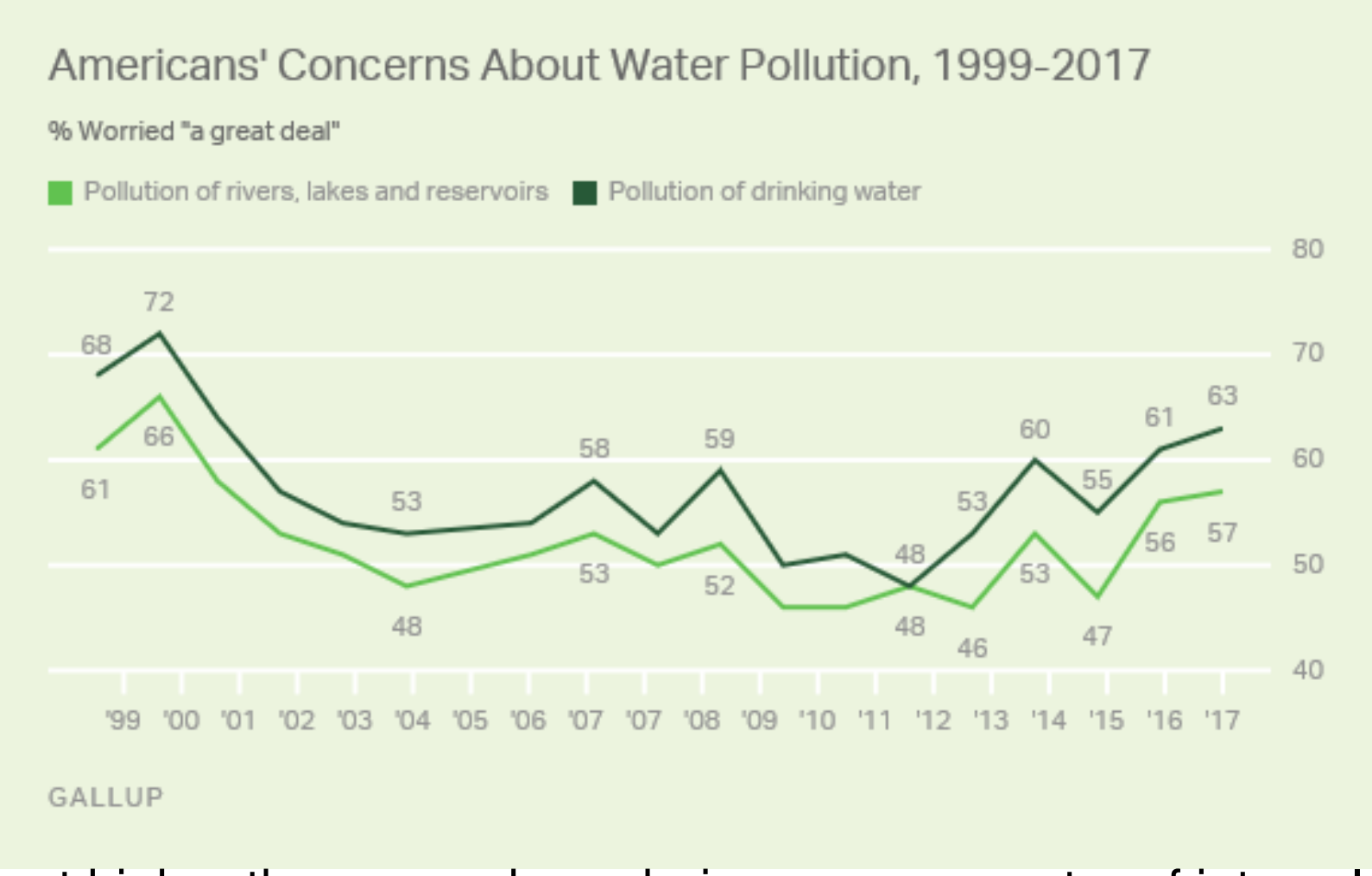
Each sample of national adults includes a minimum quota of 70% cellphone respondents and 30% landline respondents, with additional minimum quotas by time zone within region. Landline and cellular telephone numbers are selected using random-digit-dial methods.

conservative margin of error: $\frac{1}{\sqrt{1018}} \sim 0.03134 \sim 3.1\%$

Gallup uses 4%

Statistically Accurate Claim?

"In U.S., Water Pollution Worries Highest Since 2001"
lower boundary: $63 - 4 = 59\%$



not higher than upper boundaries, or even center of intervals

1. On April 4, 2017, Gallup published poll results on its web site under the headline, “Affordable Care Act Gains Majority Approval for First Time.”

If this was a simple random sample of the 1023 adults in 2017, what would the conservative 95% confidence interval margin of error be?

a) approximately 5%
b) approximately .03%
c) approximately 3.13%
d) other

1. On April 4, 2017, Gallup published poll results on its web site under the headline, “Affordable Care Act Gains Majority Approval for First Time.”

If this was a simple random sample of the 1023 adults in 2017, what would the conservative 95% confidence interval margin of error be?

a) approximately 5%
b) approximately .03%
c) approximately 3.13%
d) other

Gallup gives a 95% confident margin of error of plus or minus 3% for the 2012 poll, which had 48% of the sample “approved.” So the lower and upper boundaries for the confidence interval are 48% -3% = 45% to 51% = 48 % + 3%

2.
2012: 45% to 51% interval for the 95% confidence level
2017:

- First, compute the lower and upper boundaries for 2017 which had 55% of the sample “approved” and a margin of error plus or minus 4% for the 95% confidence level
- Second, was it likely a majority ($> 50\%$) in 2017?
- Third, could it have been a majority earlier—in 2012?

Is the headline “Affordable Care Act Gains Majority Approval for First Time” statistically accurate?

a) yes, it was likely a majority, and also a majority for the first time

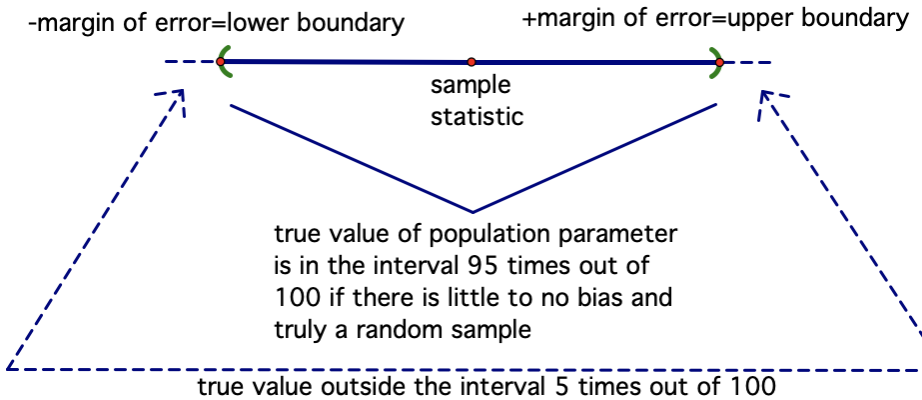
b) no, the headline isn’t statistically accurate when we take the full confidence interval into consideration

3. Assume little to no bias and truly a random sample. If a polling company conducted 100 such polls with a 95% confidence interval, then about how many of them are likely to include the true population percentage?

a) 95
b) 5
c) other

3. Assume little to no bias and truly a random sample. If a polling company conducted 100 such polls with a 95% confidence interval, then about how many of them are likely to include the true population percentage?

a) 95
b) 5
c) other



4. Is there any way to know which intervals from the 100 polls contain the true percentage and which ones don't?

a) yes
b) no
c) other

5. Gallup specifically targeted both landline and cellphone users in its polls. Are there any voices that are left out?

a) yes

b) no

5. Gallup specifically targeted both landline and cellphone users in its polls. Are there any voices that are left out?

a) yes

b) no

What percentage of US population has a phone?
How many US adults do not use the internet?

5. Gallup specifically targeted both landline and cellphone users in its polls. Are there any voices that are left out?

a) yes

b) no

What percentage of US population has a phone?
How many US adults do not use the internet?

6. How should we interpret the margin of error if the sample is very biased?

a) It is still valid as is

b) Garbage in garbage out, so the margin of error would not represent the entire population, although it could still be useful to interpret whatever biased sample it did represent.

7. For a simple random sample at the 95% confidence level, what sample size would be required to achieve a plus or minus 1% margin of error, using the conservative estimate?

a) 1
b) 100
c) 1000
d) 10000
e) other

7. For a simple random sample at the 95% confidence level, what sample size would be required to achieve a plus or minus 1% margin of error, using the conservative estimate?

a) 1
b) 100
c) 1000
d) 10000
e) other

8. In which of the following examples will the margin of error be the smallest? Assume each refers to a random sample that is not biased for a 95% confidence interval.

a) a sample of n = 400 from a population of 50,000
b) a sample of n = 1000 from a population of 10 million
c) a sample of n = 2500 from a population of 200 million
d) other

7. For a simple random sample at the 95% confidence level, what sample size would be required to achieve a plus or minus 1% margin of error, using the conservative estimate?

a) 1
b) 100
c) 1000
d) 10000
e) other

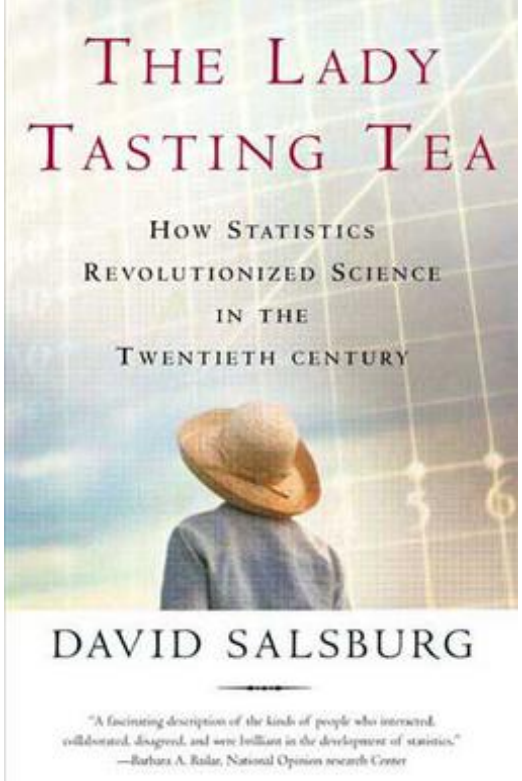
8. In which of the following examples will the margin of error be the smallest? Assume each refers to a random sample that is not biased for a 95% confidence interval.

a) a sample of n = 400 from a population of 50,000
b) a sample of n = 1000 from a population of 10 million
c) a sample of n = 2500 from a population of 200 million
d) other

Chicken Soup versus Literary Digest

9. What was the main point of Fisher's experiment on the Lady Tasting Tea from the homework readings?

a) sample size and random representative selection is what is important—not the percentage of the overall population
b) we can't assume that unusual data is incorrect
c) statistical significance can be obtained by deciding in advance the level of confidence we accept as persuasive and to collect data to make reasoned inferences



Front cover for *The Lady Tasting Tea* by David Salsburg

Readings: Deciding Personal & Public Policy

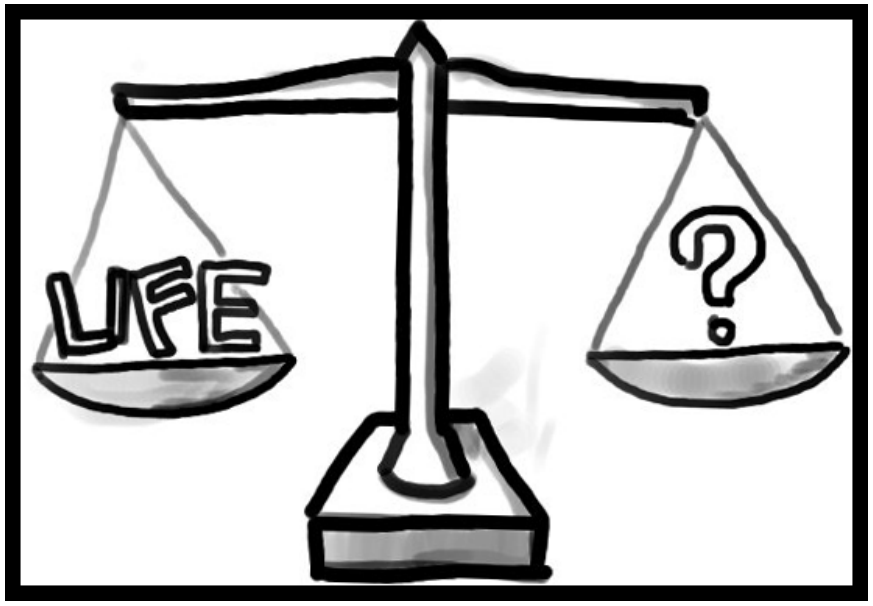


Image Credit: Linda Cai `http://cdn1.theodysseyonline.com/files/2015/07/20/635730278800703110204526443_price-of-life-by-linda-cai.png`

- What are the strongest arguments for each side? What makes the most sense from a probability argument?
- Should we vaccinate each citizen?
  Price to save a life: Cost per injection $\times$ number of shots to save just one life from the entire population

Readings: Deciding Personal & Public Policy

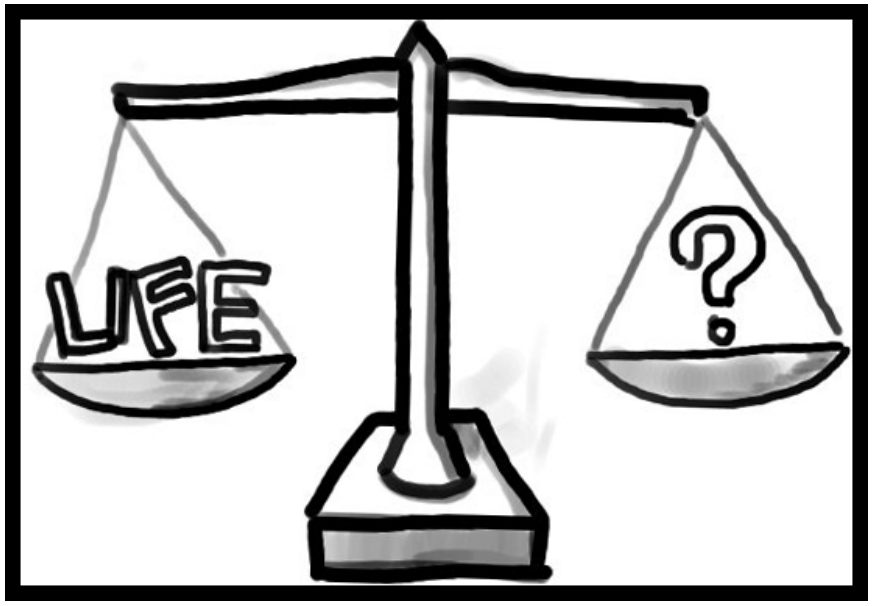


Image Credit: Linda Cai `http://cdn1.theodysseyonline.com/files/2015/07/20/635730278800703110204526443_price-of-life-by-linda-cai.png`

- What are the strongest arguments for each side? What makes the most sense from a probability argument?
- Should we vaccinate each citizen?
  Price to save a life: Cost per injection $\times$ number of shots to save just one life from the entire population
- If we had to choose between them, should we spend money to make airlines safer or cars safer?

Decision Matrix for Medical Testing

Combines probability with public policy

	Test +	Test -
Person is +		
Person is -		
Total		

Decision Matrix for Medical Testing

Combines probability with public policy

	Test +	Test -
Person is +		
Person is -		
Total		

	Test +	Test -
Person is +	true positive people	false negative people
Person is -	false positive people	true negative people
Total	total who test +	total who test -

ELISA: Mandatory HIV Testing in the US?

Sensitivity: probability the test correctly identifies someone who is HIV+ as positive = 95% $=$ .95
False Negative: The probability is $1 - .95 = .05$
Specificity: probability correctly identifies HIV- = 99% $=$ .99
False Positive: The probability is $1 - .99 = .01$

ELISA: Mandatory HIV Testing in the US?

Sensitivity: probability the test correctly identifies someone who is HIV+ as positive = 95% $=$ .95
False Negative: The probability is $1 - .95 = .05$
Specificity: probability correctly identifies HIV- = 99% $=$ .99
False Positive: The probability is $1 - .99 = .01$
US: $\approx 330,000,000$ https://www.census.gov/popclock/
probability of HIV+: 3/1000

ELISA: Mandatory HIV Testing in the US?

Sensitivity: probability the test correctly identifies someone who is HIV+ as positive = 95% $=$ .95
False Negative: The probability is $1 - .95 = .05$
Specificity: probability correctly identifies HIV- = 99% $=$ .99
False Positive: The probability is $1 - .99 = .01$
US: $\approx 330,000,000$ https://www.census.gov/popclock/
probability of HIV+: 3/1000 =.003
probability of HIV-:

ELISA: Mandatory HIV Testing in the US?

Sensitivity: probability the test correctly identifies someone who is HIV+ as positive = 95% $=$ .95
False Negative: The probability is $1 - .95 = .05$
Specificity: probability correctly identifies HIV- = 99% $=$ .99
False Positive: The probability is $1 - .99 = .01$
US: $\approx 330,000,000$ https://www.census.gov/popclock/
probability of HIV+: 3/1000 =.003
probability of HIV-: $1 - .003 = .997$

	Test +	Test -
Person is HIV+		

ELISA: Mandatory HIV Testing in the US?

Sensitivity: probability the test correctly identifies someone who is HIV+ as positive = 95% $=$ .95
False Negative: The probability is $1 - .95 = .05$
Specificity: probability correctly identifies HIV- = 99% $=$ .99
False Positive: The probability is $1 - .99 = .01$
US: $\approx 330,000,000$ https://www.census.gov/popclock/
probability of HIV+: 3/1000 =.003
probability of HIV-: $1 - .003 = .997$

	Test +	Test -
Person is HIV+	$330000000 \times .003$	

ELISA: Mandatory HIV Testing in the US?

Sensitivity: probability the test correctly identifies someone who is HIV+ as positive = 95% $=$ .95
False Negative: The probability is $1 - .95 = .05$
Specificity: probability correctly identifies HIV- = 99% $=$ .99
False Positive: The probability is $1 - .99 = .01$
US: $\approx 330,000,000$ https://www.census.gov/popclock/
probability of HIV+: 3/1000 =.003
probability of HIV-: $1 - .003 = .997$

	Test +	Test -
Person is HIV+	$330000000 \times .003 \times .95$	

ELISA: Mandatory HIV Testing in the US?

Sensitivity: probability the test correctly identifies someone who is HIV+ as positive = 95% $=$ .95
False Negative: The probability is $1 - .95 = .05$
Specificity: probability correctly identifies HIV- = 99% $=$ .99
False Positive: The probability is $1 - .99 = .01$
US: $\approx 330,000,000$ https://www.census.gov/popclock/
probability of HIV+: 3/1000 =.003
probability of HIV-: $1 - .003 = .997$

	Test +	Test -
Person is HIV+	$330000000 \times .003 \times .95$	$330000000 \times .003$

ELISA: Mandatory HIV Testing in the US?

Sensitivity: probability the test correctly identifies someone who is HIV+ as positive = 95% $=$ .95
False Negative: The probability is $1 - .95 = .05$
Specificity: probability correctly identifies HIV- = 99% $=$ .99
False Positive: The probability is $1 - .99 = .01$
US: $\approx 330,000,000$ https://www.census.gov/popclock/
probability of HIV+: 3/1000 =.003
probability of HIV-: $1 - .003 = .997$

	Test +	Test -
Person is HIV+	$330000000 \times .003 \times .95$ 940500	$330000000 \times .003 \times .05$ 49500
Person is HIV-		

ELISA: Mandatory HIV Testing in the US?

Sensitivity: probability the test correctly identifies someone who is HIV+ as positive = 95% $=$ .95
False Negative: The probability is $1 - .95 = .05$
Specificity: probability correctly identifies HIV- = 99% $=$ .99
False Positive: The probability is $1 - .99 = .01$
US: $\approx 330,000,000$ https://www.census.gov/popclock/
probability of HIV+: 3/1000 =.003
probability of HIV-: $1 - .003 = .997$

	Test +	Test -
Person is HIV+	$330000000 \times .003 \times .95$ 940500	$330000000 \times .003 \times .05$ 49500
Person is HIV-	$330000000 \times .997$	

ELISA: Mandatory HIV Testing in the US?

Sensitivity: probability the test correctly identifies someone who is HIV+ as positive = 95% $=$ .95
False Negative: The probability is $1 - .95 = .05$
Specificity: probability correctly identifies HIV- = 99% $=$ .99
False Positive: The probability is $1 - .99 = .01$
US: $\approx 330,000,000$ https://www.census.gov/popclock/
probability of HIV+: 3/1000 =.003
probability of HIV-: $1 - .003 = .997$

	Test +	Test -
Person is HIV+	$330000000 \times .003 \times .95$ 940500	$330000000 \times .003 \times .05$ 49500
Person is HIV-	$330000000 \times .997 \times .01$	

ELISA: Mandatory HIV Testing in the US?

Sensitivity: probability the test correctly identifies someone who is HIV+ as positive = 95% $=$ .95
False Negative: The probability is $1 - .95 = .05$
Specificity: probability correctly identifies HIV- = 99% $=$ .99
False Positive: The probability is $1 - .99 = .01$
US: $\approx 330,000,000$ https://www.census.gov/popclock/
probability of HIV+: 3/1000 =.003
probability of HIV-: $1 - .003 = .997$

	Test +	Test -
Person is HIV+	$330000000 \times .003 \times .95$ 940500	$330000000 \times .003 \times .05$ 49500
Person is HIV-	$330000000 \times .997 \times .01$	$330000000 \times .997$

ELISA: Mandatory HIV Testing in the US?

Sensitivity: probability the test correctly identifies someone who is HIV+ as positive = 95% $=$ .95
False Negative: The probability is $1 - .95 = .05$
Specificity: probability correctly identifies HIV- = 99% $=$ .99
False Positive: The probability is $1 - .99 = .01$
US: $\approx 330,000,000$ https://www.census.gov/popclock/
probability of HIV+: 3/1000 =.003
probability of HIV-: $1 - .003 = .997$

	Test +	Test -
Person is HIV+	$330000000 \times .003 \times .95$ 940500	$330000000 \times .003 \times .05$ 49500
Person is HIV-	$330000000 \times .997 \times .01$ 3290100	$330000000 \times .997 \times .99$ 325719900
Total		

ELISA: Mandatory HIV Testing in the US?

Sensitivity: probability the test correctly identifies someone who is HIV+ as positive = 95% $=$ .95
False Negative: The probability is $1 - .95 = .05$
Specificity: probability correctly identifies HIV- = 99% $=$ .99
False Positive: The probability is $1 - .99 = .01$
US: $\approx 330,000,000$ https://www.census.gov/popclock/
probability of HIV+: 3/1000 =.003
probability of HIV-: $1 - .003 = .997$

	Test +	Test -
Person is HIV+	$330000000 \times .003 \times .95$ 940500	$330000000 \times .003 \times .05$ 49500
Person is HIV-	$330000000 \times .997 \times .01$ 3290100	$330000000 \times .997 \times .99$ 325719900
Total	4230600	

ELISA: Mandatory HIV Testing in the US?

Sensitivity: probability the test correctly identifies someone who is HIV+ as positive = 95% = .95
False Negative: The probability is $1 - .95 = .05$
Specificity: probability correctly identifies HIV- = 99% = .99
False Positive: The probability is $1 - .99 = .01$
US: $\approx 330,000,000$ https://www.census.gov/popclock/
probability of HIV+: 3/1000 =.003
probability of HIV-: $1 - .003 = .997$

	Test +	Test -
Person is HIV+	$330000000 \times .003 \times .95$ 940500	$330000000 \times .003 \times .05$ 49500
Person is HIV-	$330000000 \times .997 \times .01$ 3290100	$330000000 \times .997 \times .99$ 325719900
Total	4230600	325769400

If you test positive your chance of having HIV: 940500/4230600 $\sim$ 22%

Connections to Geometry Segment

Russell Knightley. `http://www.rkm.com.au/VIRUS/HIV`

10. In *The Heart of Mathematics*, on pp. 663–664 you chose one to complete: amazing stats #15 or internet askew #18. Share as below:

If you selected #15:

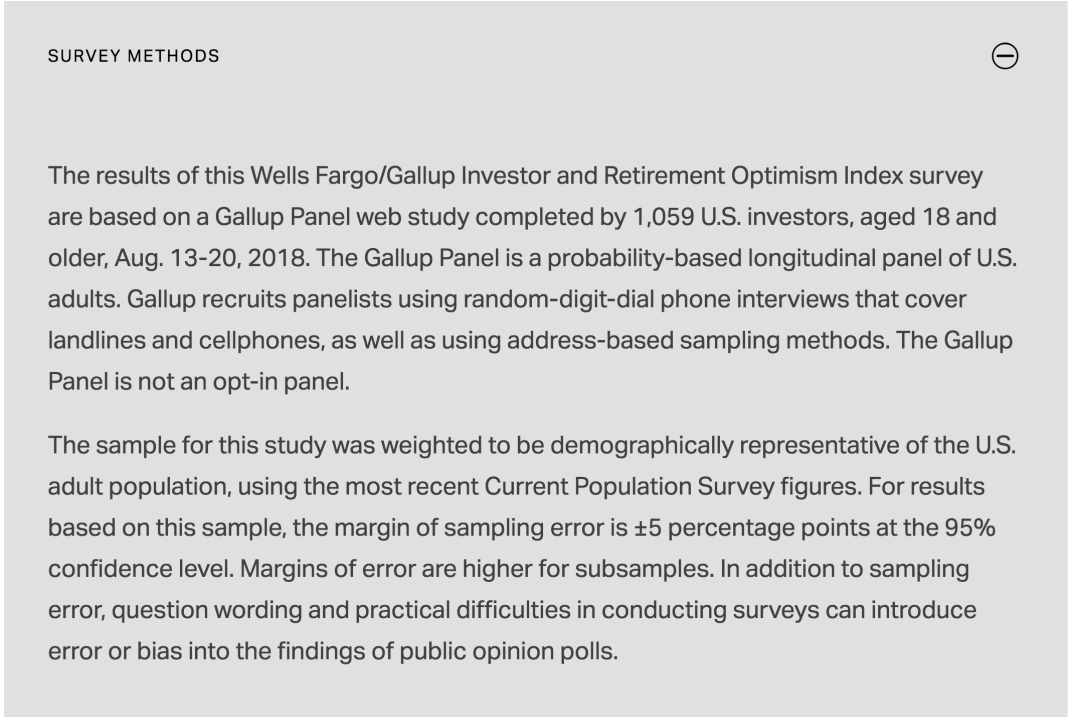
a) What is the question you asked?
b) What is the sample and the sample size?
c) What is the dubious conclusion?
d) How could one reduce bias in this instance?

If you selected #18:

a) What is the title or topic and provide the source
b) Summarize the graph
c) Is there distortion or bias? Explain.
d) How could one reduce bias in this instance?

https://news.gallup.com/home.aspx

Select a Gallup poll which also has a Survey Methods section that includes a margin of sampling error %



SURVEY METHODS

The results of this Wells Fargo/Gallup Investor and Retirement Optimism Index survey are based on a Gallup Panel web study completed by 1,059 U.S. investors, aged 18 and older, Aug. 13-20, 2018. The Gallup Panel is a probability-based longitudinal panel of U.S. adults. Gallup recruits panelists using random-digit-dial phone interviews that cover landlines and cellphones, as well as using address-based sampling methods. The Gallup Panel is not an opt-in panel.

The sample for this study was weighted to be demographically representative of the U.S. adult population, using the most recent Current Population Survey figures. For results based on this sample, the margin of sampling error is ±5 percentage points at the 95% confidence level. Margins of error are higher for subsamples. In addition to sampling error, question wording and practical difficulties in conducting surveys can introduce error or bias into the findings of public opinion polls.

Not all Gallup articles have this—articles that are tagged in Green with Report, Gallup Vault, Polling Matters, Gallup Blog, Gallup Podcast and more do not typically. Many articles that are tagged in Green with topic headers like Economy, Education, Politics, Social & Policy Issues, Well Being, and World are more likely to. May need to look around some...