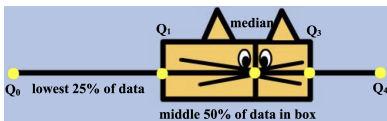


- probability: likeliness/chance: 0 to 1 (or 0 to 100%)
- algebra of probability, like probability of sharing a birthday
 - if independent then probabilities multiply
 - event will occur = 1 – probability it won't
- expected value can give us an idea of likely outcomes, e.g.
 - $\frac{1}{16}$ # coin tossers for HHHH
 - $.5 \cdot 85 + .05 \cdot 100 + .3 \cdot 75 + .15 \cdot 95$ for grade
- decision matrix, law of large numbers
- Benford's law: first digit in many real-life data sets > 500 approximates a logarithmic trend where 1 occurs about 30% of the time while 9 less than 5%
- respect for persons, max benefits while min risks, justice
- central tendency measures
 - average or mean: sum all the numbers and divide by how many
 - median: middle of ordered data—number and place



- boxplot
- 7 most popular in data

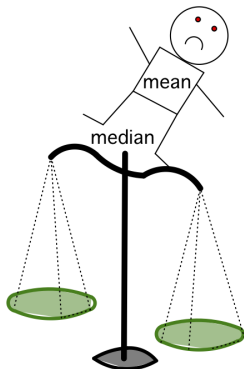
Here is a data set that measures population growth rates in the US from 1910–1919:

2.1	1.56	1.56	1.96	1.92	1.44	1.4	1.27	-.06	1.26
1910	1911	1912	1913	1914	1915	1916	1917	1918	1919

How would 1918 impact the mean/average?

Use a scale balancing idea

- a) drag the mean down from the median
- b) drag the mean up from the median
- c) would not impact the mean



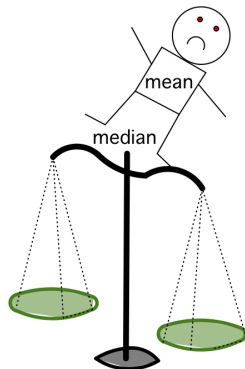
Here is a data set that measures population growth rates in the US from 1910–1919:

2.1	1.56	1.56	1.96	1.92	1.44	1.4	1.27	-.06	1.26
1910	1911	1912	1913	1914	1915	1916	1917	1918	1919

How would 1918 impact the mean/average?

Use a scale balancing idea

- a) drag the mean down from the median
- b) drag the mean up from the median
- c) would not impact the mean



What happened in 1918?

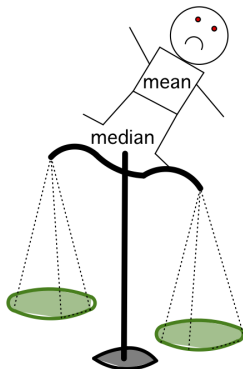
Here is a data set that measures population growth rates in the US from 1910–1919:

2.1	1.56	1.56	1.96	1.92	1.44	1.4	1.27	-.06	1.26
1910	1911	1912	1913	1914	1915	1916	1917	1918	1919

How would 1918 impact the mean/average?

Use a scale balancing idea

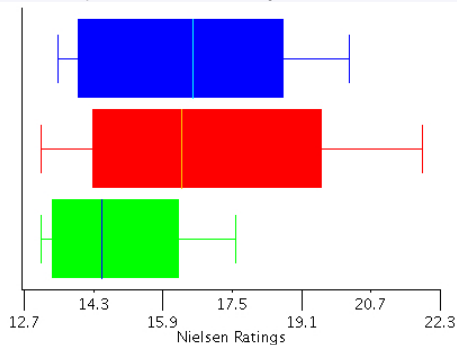
- a) drag the mean down from the median
- b) drag the mean up from the median
- c) would not impact the mean



What happened in 1918?

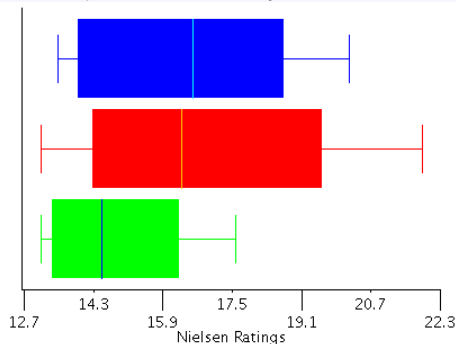
median: 1.56. mean: 1.48.

Here is Nielsen ratings (roughly represent the percentage of households tuned in). Use the boxplots to award “best network”



- a) ABC (top boxplot)
- b) CBS (middle)
- c) NBC (bottom)
- d) There should be more than 1 winner
- e) other

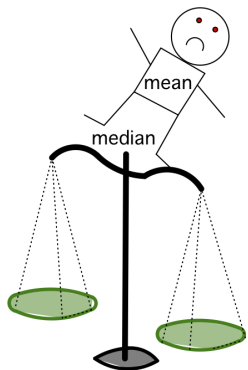
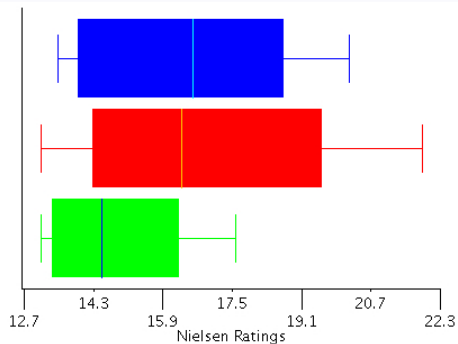
Here is Nielsen ratings (roughly represent the percentage of households tuned in). Use the boxplots to award “best network”



- a) ABC (top boxplot)
- b) CBS (middle)
- c) NBC (bottom)
- d) There should be more than 1 winner
- e) other

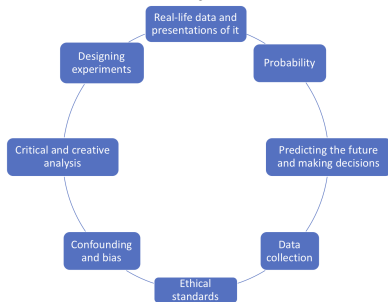
Discuss how to spin the statistics positively for each network: Here's good news, we are the best network because... In one of them, it may be challenging to say something positive but truthful, but think creatively!

Which is true for CBS, the middle boxplot?



- a) The mean is probably higher than the median
- b) The mean is probably lower than the median
- c) The mean is probably about the same as the median

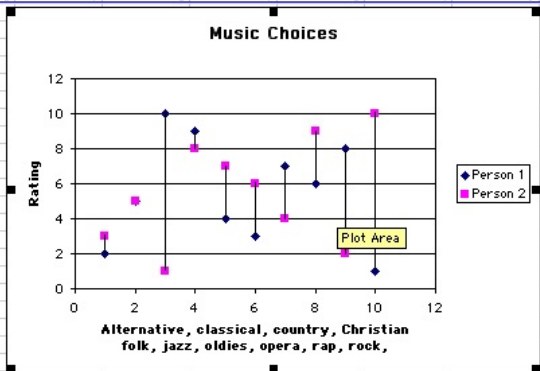
- collecting data: reproducibility, consensus, and random sampling if possible
- presenting data: entire data set versus numerical or visual snapshots of it
- expected value: weighted probabilities for decisions
- mean and median: central tendencies
- box plots: comparisons
- **regressions**: correlations
- **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

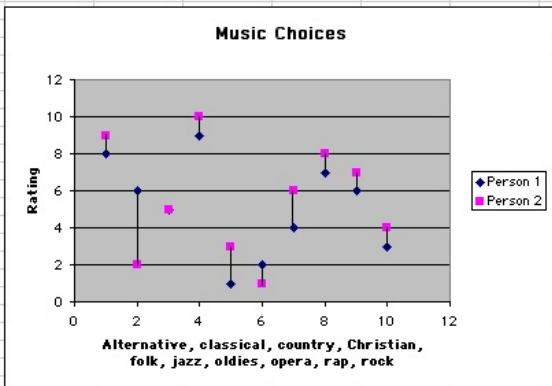
Music Compatibility

	A	B	C	D	E	F	G	H	I	J
1	alternative	classical	country	christian	folk	jazz	oldies	opera	rap	rock
2	2	5	10	9	4	3	7	6	8	1
3	3	5	1	8	7	6	4	9	2	10



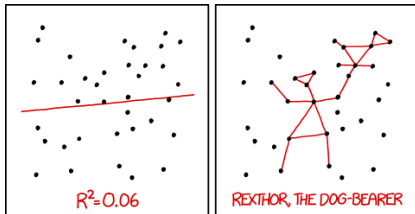
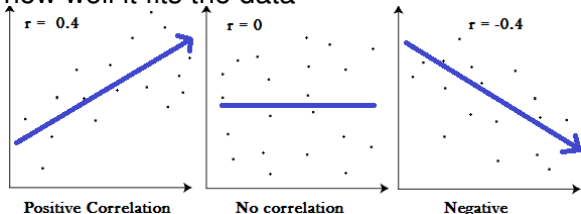
Music Compatibility

	A	B	C	D	E	F	G	H	I	J		
1	alternative	classical	country	christian	folk	jazz	oldies	opera	rap	rock		
2		8		5		1		4		6	3	
3		9	2	5	10		3	1	6	8	7	4
4												
5												
6												
7												
8												
9												
10												
11												
12												
13												
14												
15												
16												
17												
18												
19												
20												
21												
22												
23												
24												
25												



Inferences and Regression

- Karl Pearson statistics, eugenics
- correlation coefficient r gives sign of slope of best fit line and a measure of how well it fits the data



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

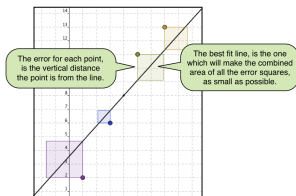
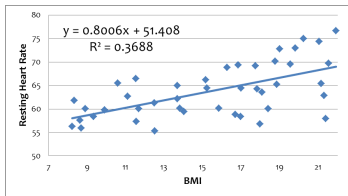
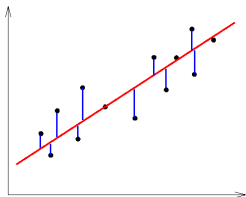
Picture citations: 1. Public domain

2. <https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/correlation-coefficient-formula/>

3. <https://xkcd.com/1725/>

Strength of the Relationship: r^2 percent

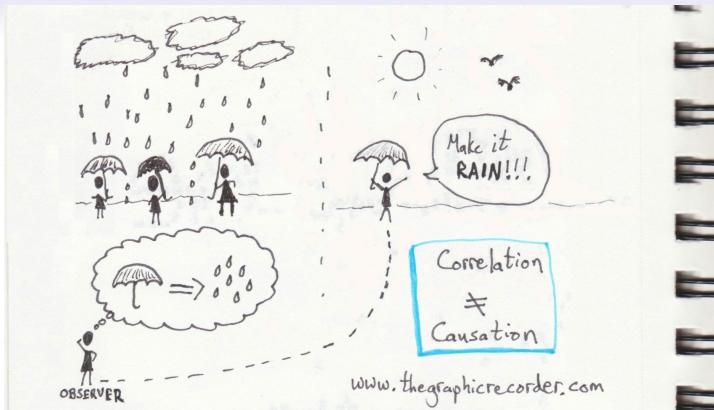
- 0 to 10% **no**
 - 10% to 25% **weak**
 - 25% to 65% **moderate**
 - above 65% **strong**
- NOT a probability for correct nor a likelihood of on the line
 - measures the y-values distances via sum of squares as variation in the dependent variable explained by linearity



Picture citations:

1. <http://cs.wellesley.edu/~cs199/lectures/35-correlation-regression.html>
2. <http://www2.nau.edu/mat114-c/ch3a.php>
3. <http://math.maine121.org/welcome/chapter-5/>

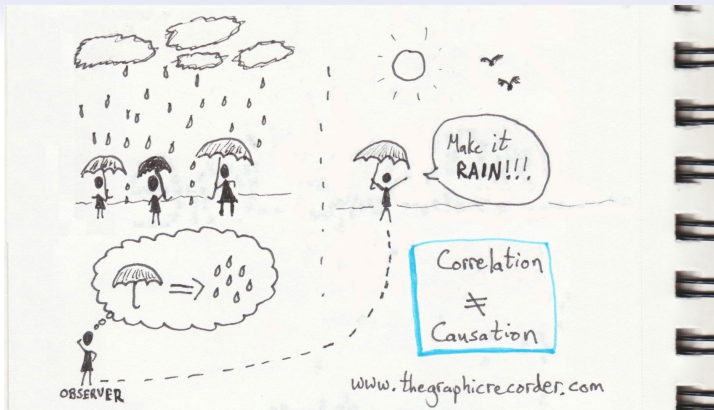
Correlation versus Causation



Plausible? Analyze the data. Look for random sampling, consensus, and confounding variables.

- Does hand length predict forearm?

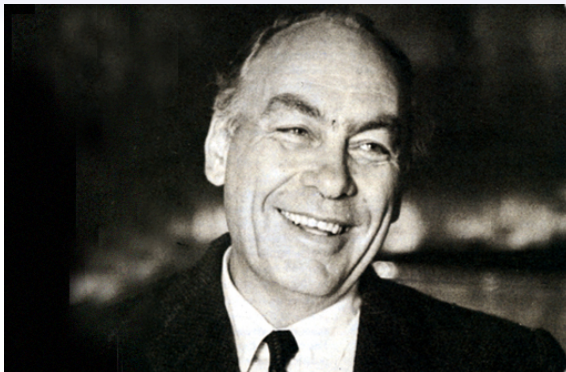
Correlation versus Causation



Plausible? Analyze the data. Look for random sampling, consensus, and confounding variables.

- Does hand length predict forearm?
 $\sim 1.618 \pm .4$: (1.218, 2.018)
- Do hours without sleep the night before an exam predict midterm errors?

George Gallup and the 1936 Election



The Literary Digest

NEW YORK

OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the

Republican National Committee purchased THE LITERARY DIGEST? And all types and varieties, including: "Have the Jews purchased

returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.' "We never make any claims before election but we respectfully refer you to the opinion of one of the most trusted citizens

Picture credit: 1. <http://www.thegalluphouse.com/georgegallupbiography.html>

2. <http://www.unz.com/print/LiteraryDigest-1936oct31:4/>

Nate Silver and the 2016 Election



- “Data-driven predictions can succeed-and they can fail. It is when we deny our role in the process that the odds of failure rise. Before we demand more of our data, we need to demand more of ourselves.”
- **sabermetrics, elections, FiveThirtyEight**

Picture credit: Danielle Levitt

<https://www.out.com/out-exclusives/out100-2013/2013/11/05/nate-silver-espn-statistician-predict>

Nate Silver on the 2016 Presidential Election

Post-election view: Should never, ever trust polls again, as they proved Secretary of State Clinton almost certain to win

Nate Silver on the 2016 Presidential Election

Post-election view: Should never, ever trust polls again, as they proved Secretary of State Clinton almost certain to win

Really: Polls showed an uncertain and volatile race with Clinton as a modest favorite and Trump (who was elected as president) with a chance

- myth there was a catastrophic failure for the polls. Trump outperformed his national polls by only 1 to 2 percentage points. He beat his polls by only 2 to 3 percentage points in the average swing state.

Nate Silver on the 2016 Presidential Election

Post-election view: Should never, ever trust polls again, as they proved Secretary of State Clinton almost certain to win

Really: Polls showed an uncertain and volatile race with Clinton as a modest favorite and Trump (who was elected as president) with a chance

- myth there was a catastrophic failure for the polls. Trump outperformed his national polls by only 1 to 2 percentage points. He beat his polls by only 2 to 3 percentage points in the average swing state.
- there were individual pollsters that had some explaining to do, especially in Michigan, Wisconsin and Pennsylvania, where Trump beat his polls by a larger amount.

Nate Silver on the 2016 Presidential Election

Post-election view: Should never, ever trust polls again, as they proved Secretary of State Clinton almost certain to win

Really: Polls showed an uncertain and volatile race with Clinton as a modest favorite and Trump (who was elected as president) with a chance

- myth there was a catastrophic failure for the polls. Trump outperformed his national polls by only 1 to 2 percentage points. He beat his polls by only 2 to 3 percentage points in the average swing state.
- there were individual pollsters that had some explaining to do, especially in Michigan, Wisconsin and Pennsylvania, where Trump beat his polls by a larger amount.
- not some sort of massive outlier; the polls were pretty much as accurate as they'd been, on average, since 1968
- FiveThirtyEight: Trump a 3 in 10 chance of winning the Electoral College. Others: 1 in 100

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.

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.

Gallup Positive Events for Investors Nov 7, 2018—<https://news.gallup.com/poll/244268/positive-events-investors-buying-home-getting-married.aspx>

[//news.gallup.com/poll/244268/positive-events-investors-buying-home-getting-married.aspx](https://news.gallup.com/poll/244268/positive-events-investors-buying-home-getting-married.aspx)

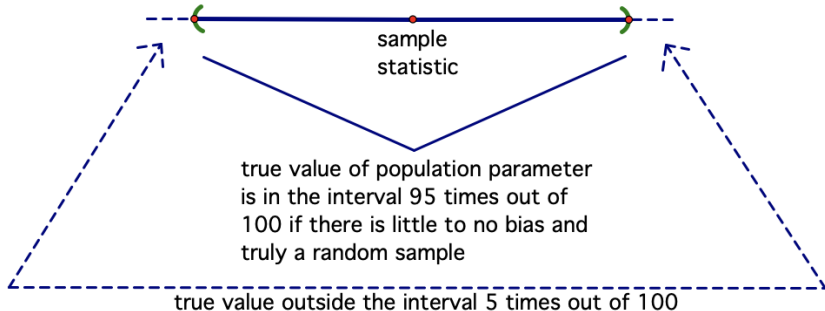
- Likelihood of the sample outcome—no way to know which intervals contain the true percentage and which don't

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.

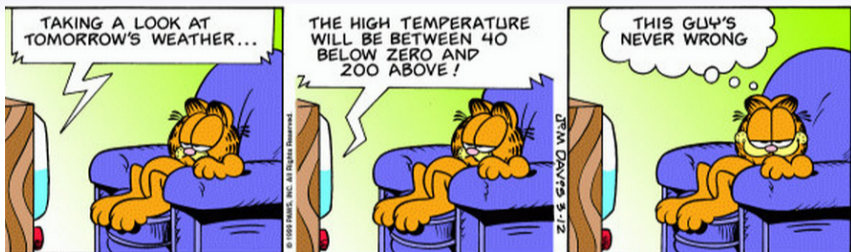
-margin of error=lower boundary

+margin of error=upper boundary



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

Margin of Error

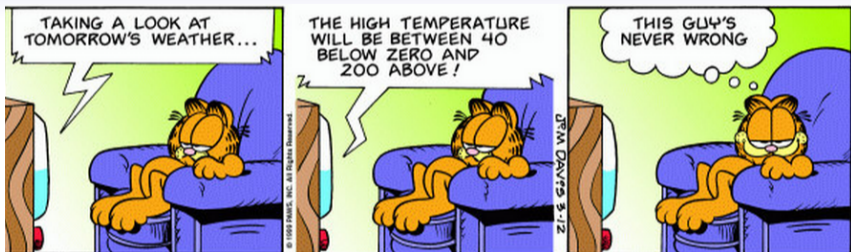


Garfield by Jim Davis <https://garfield.com/comic/1999/03/12>

- **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**).

A survey reports a margin of error of 3% at the 95% confidence interval. Approximately how many people were surveyed?

Margin of Error



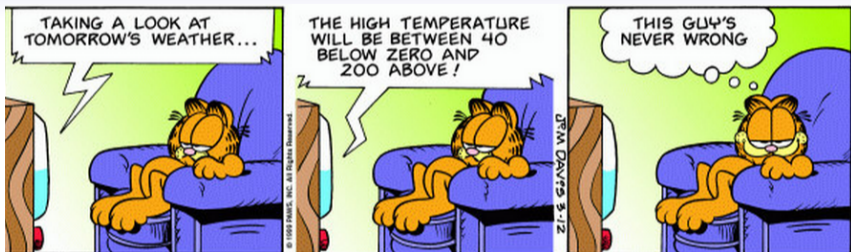
Garfield by Jim Davis <https://garfield.com/comic/1999/03/12>

- **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**).

A survey reports a margin of error of 3% at the 95% confidence interval. Approximately how many people were surveyed?

$$\frac{1}{\sqrt{n}} = .03$$

Margin of Error



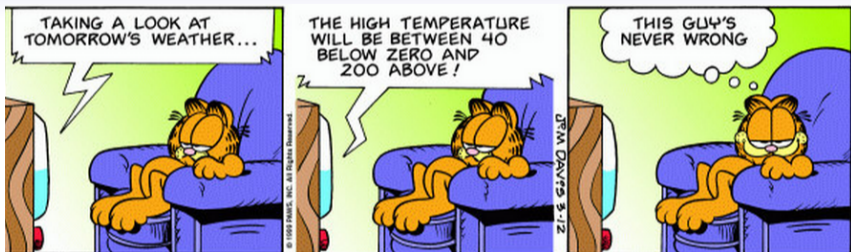
Garfield by Jim Davis <https://garfield.com/comic/1999/03/12>

- **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**).

A survey reports a margin of error of 3% at the 95% confidence interval. Approximately how many people were surveyed?

$$\frac{1}{\sqrt{n}} = .03 \text{ so } \frac{1}{.03} = \sqrt{n} \text{ and so}$$

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**).

A survey reports a margin of error of 3% at the 95% confidence interval. Approximately how many people were surveyed?

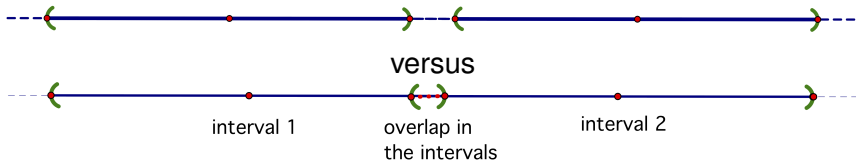
$$\frac{1}{\sqrt{n}} = .03 \text{ so } \frac{1}{.03} = \sqrt{n} \text{ and so } \left(\frac{1}{.03}\right)^2 = n$$

Gallup Polls

SURVEY METHODS

Results for this Gallup World poll are based on an aggregation of telephone and in person interviews conducted with 5,011 individuals aged 15 or older residing in Algeria, Libya, Egypt, Morocco and Tunisia in 2016 and 5,030 in 2017. The interviews were conducted with between 1,000 and 1,016 individuals in each country, each year. For results based on an aggregation of adults residing in these countries, the margin of sampling error is ± 2 percentage points at the 95% confidence level. All reported margins of sampling error include computed design effects for weighting.

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 World poll are based on an aggregation of telephone and in person interviews conducted with 5,011 individuals aged 15 or older residing in Algeria, Libya, Egypt, Morocco and Tunisia in 2016 and 5,030 in 2017. The interviews were conducted with between 1,000 and 1,016 individuals in each country, each year. For results based on an aggregation of adults residing in these countries, the margin of sampling error is ± 2 percentage points at the 95% confidence level. All reported margins of sampling error include computed design effects for weighting.

conservative margin of error: $\frac{1}{\sqrt{5011}} \sim .014 \sim 1.4\%$

Gallup uses 2%

Statistically Accurate Claim?

We check for overlaps in the intervals in order to evaluate the statistical validity of headlines and statements in polls

“Desire to Migrate Rises in North Africa”

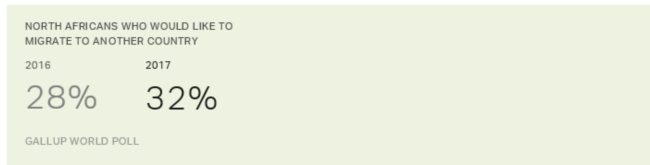
2017 lower boundary: $32 - 2 = 30\%$

2016 upper boundary: $28 + 2 = 30\%$

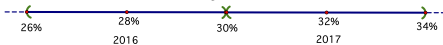
WORLD APRIL 24, 2018

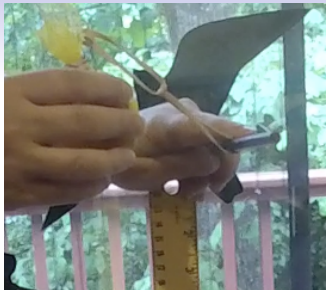
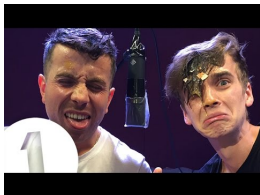
Desire to Migrate Rises in North Africa

BY IMAN BERRACHED AND RJ REINHART



it could have stayed the same!





<https://i.ytimg.com/vi/vcb6CmUyxn0/hqdefault.jpg>

