Class 22

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-10-25

Chi-Square Test for Independence in 2-Way Tables

1. Assume the 2 variables are ***independent***. (H_0: Independence)

Chi-Square Test for Independence in 2-Way Tables

1. Assume the 2 variables are ***independent***. (H_0: Independence)
2. Calculate the ***expected*** counts under the null hypothesis of independence.

Chi-Square Test for Independence in 2-Way Tables

1. Assume the 2 variables are ***independent***. (H_0: Independence)
2. Calculate the ***expected*** counts under the null hypothesis of independence.
3. Find the ***test statistic***.

Chi-Square Test for Independence in 2-Way Tables

1. Assume the 2 variables are ***independent***. (H_0: Independence)
2. Calculate the ***expected*** counts under the null hypothesis of independence.
3. Find the ***test statistic***.
4. Compute the ***degrees of freedom***.

Chi-Square Test for Independence in 2-Way Tables

1. Assume the 2 variables are ***independent***. (H_0: Independence)
2. Calculate the ***expected*** counts under the null hypothesis of independence.
3. Find the ***test statistic***.
4. Compute the ***degrees of freedom***.
5. Determine the probability of the ***observed*** counts under the null hypothesis.

Chi-Square Test for Independence in 2-Way Tables

1. Assume the 2 variables are ***independent***. (H_0: Independence)
2. Calculate the ***expected*** counts under the null hypothesis of independence.
3. Find the ***test statistic***.
4. Compute the ***degrees of freedom***.
5. Determine the probability of the ***observed*** counts under the null hypothesis.
6. If it is sufficiently unlikely to have gotten the ***observed*** data under the null hypothesis of independence, reject H_0 and accept H_A: Dependence.

Expected Counts

If we assume H_0: Independence, how can we estimate the data ***expected*** under the null hypothesis?

Expected Counts

If we assume H_0: Independence, how can we estimate the data ***expected*** under the null hypothesis?

The Multiplication Rule for Independent Events

The probability of event A ***and*** event B occurring is the product of the probability that A occurs and the probability that B occurs.

Proportions

Table 1: Proportion of Observations

	B	B'	
A	P(A + B)	P(A + B')	P(A)
A'	P(A' + B)	P(A' + B')	P(A')
	P(B)	P(B')	1

Calculating Expected Counts

Using proportions...

$$\text{Expected}_{\text{AandB}} = P(A) \times P(B) \times n$$

Calculating Expected Counts

Using proportions...

$$\text{Expected}_{\text{AandB}} = P(A) \times P(B) \times n$$

Using counts...

$$\text{Expected}_{\text{AandB}} = \frac{\text{count}(A) \times \text{count}(B)}{n}$$

The Test Statistic

The Chi-Square (χ^2) test statistic is the sum of the squared difference between observed and expected value divided by the expected value for all combinations of categories.

$$\chi^2_{df} = \sum_{i=1}^{k} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Degrees of Freedom

For a two-way table, the degrees of freedom for a χ^2 test statistic are...

$$\begin{aligned} df &= (n_{\text{rows}} - 1) \times (n_{\text{cols}} - 1) \\ &= (R - 1) \times (C - 1) \end{aligned}$$

...where R is the number of rows in the table and C is the number of columns.

P-Values in R

You can find p-values in R using the `pchisq()` function, which takes a test statistic (`q`) and the degrees of freedom (`df`) as parameters.

```
pchisq(2, df = 3, lower.tail = F)
```

```
[1] 0.5724067
```

For the χ^2 test, we always use the ***upper tail***, or the probability of seeing a result more extreme than what we observed.

Conditions

The χ^2 test is most appropriate for large sample sizes.

- $>$ 20% of expected counts are $>$5.

Conditions

The χ^2 test is most appropriate for large sample sizes.

- $>$ 20% of expected counts are $>$5.
- All individual expected counts are $>$1.

Conditions

The χ^2 test is most appropriate for large sample sizes.

- $>$ 20% of expected counts are $>$5.
- All individual expected counts are $>$1.
- In a 2x2 table, all 4 expected counts $>$5.

Conditions

The χ^2 test is most appropriate for large sample sizes.

- $>$ 20% of expected counts are $>$5.
- All individual expected counts are $>$1.
- In a 2x2 table, all 4 expected counts $>$5.

When sample sizes are small, use Fisher's Exact Test.

Limitations

- Neither the χ^2 test nor the Fisher's Exact Test will tell you the nature of the relationship between your 2 categorical variables.
- Additional tests are needed to determine if the outcomes are dependent on variable 1, variable 2, or both.

Sample Size Estimation for Proportions

You can use the margin of error calculation to estimate the sample size needed to detect a given difference in proportions.

$$\text{marginoferror} = Z^* \times \sqrt{\frac{p(1-p)}{n}}$$

Example: Sample Size

We want to know if people favor candidate 1 or candidate 2 (H_0: $P(C_1) = P(C_2)$), but it will be a very close race. If we want to find 52% for candidate 1 vs 48% for candidate 2, what size sample do we need?

Example: Sample Size

We want to know if people favor candidate 1 or candidate 2 (H_0: $P(C_1) = P(C_2)$), but it will be a very close race. If we want to find 52% for candidate 1 vs 48% for candidate 2, what size sample do we need?

What margin of error do we need?

a. 4%
b. 2%
c. 1%
d. 5%

Example: Sample Size

We want to know if people favor candidate 1 or candidate 2 (H_0: $P(C_1) = P(C_2)$), but it will be a very close race. If we want to find 52% for candidate 1 vs 48% for candidate 2, what size sample do we need?

Example: Sample Size

We want to know if people favor candidate 1 or candidate 2 (H_0: $P(C_1) = P(C_2)$), but it will be a very close race. If we want to find 52% for candidate 1 vs 48% for candidate 2, what size sample do we need?

What margin of error do we need?

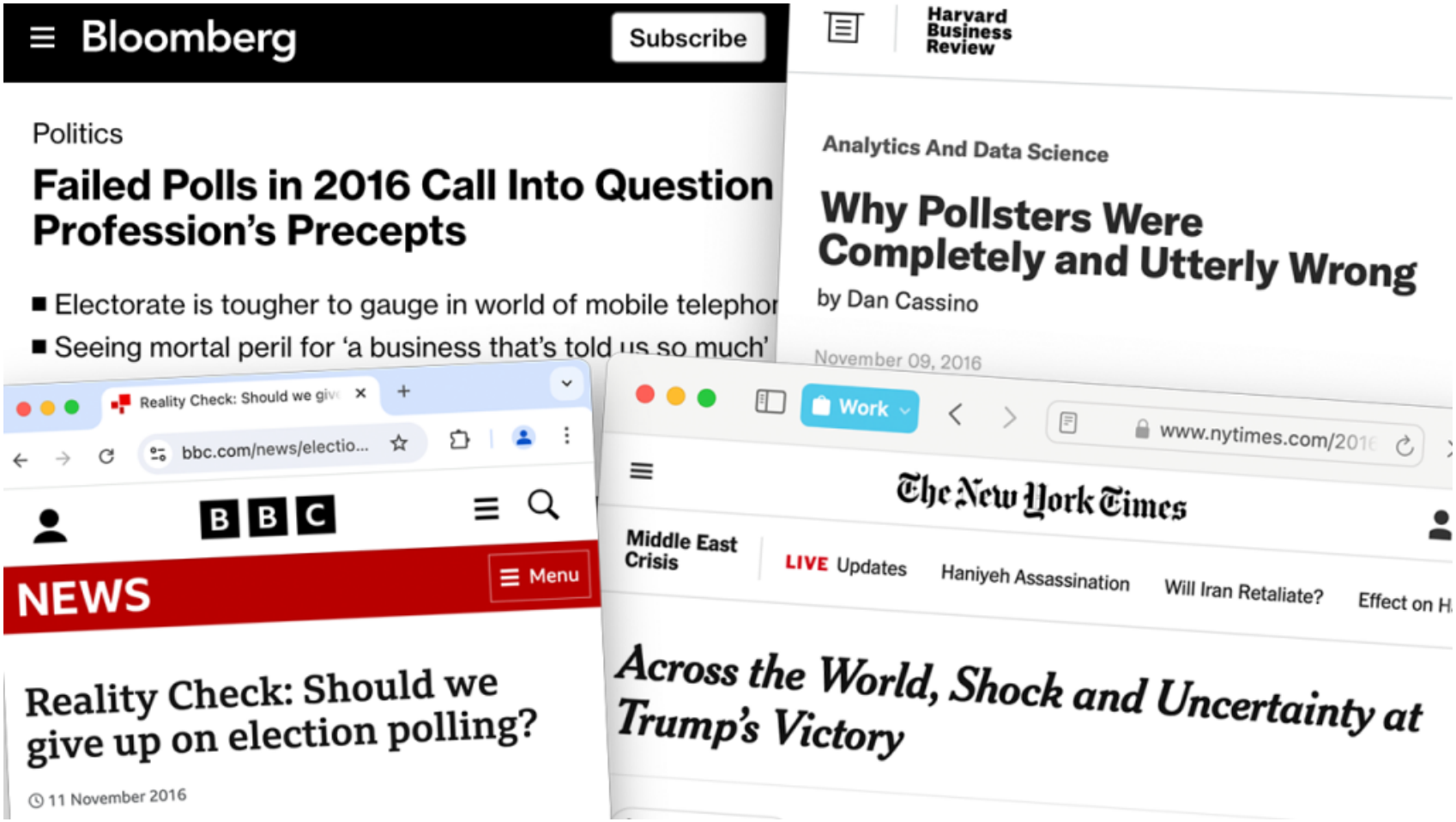
b. 2%

Example: Sample Size

If we need a margin of error of 2%, and we want to have 95% confidence in our results, we can solve for n to find the minimum sample size needed.

$$
\begin{aligned}
0.02 &= 1.96 \times \sqrt{\frac{0.52(1-0.52)}{n}} \\
0.01 &= \sqrt{\frac{0.52(1-0.52)}{n}} \\
0.0001 &= \frac{0.52(1-0.52)}{n} \\
0.0001n &= 0.250 \\
n &= 2397
\end{aligned}
$$

Polling History



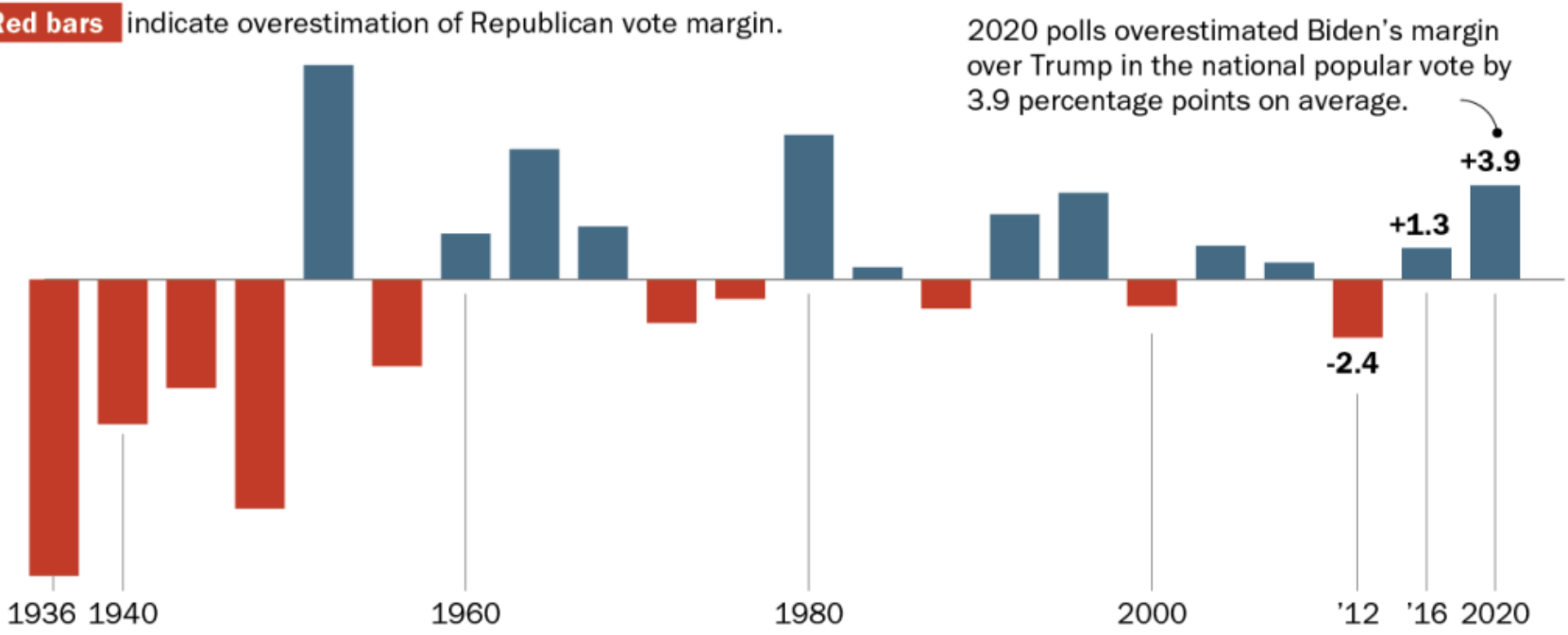
Polling Error

Polling errors in U.S. presidential elections

Bars represent average signed error

Blue bars indicate overestimation of Democratic vote margin.

Red bars indicate overestimation of Republican vote margin.



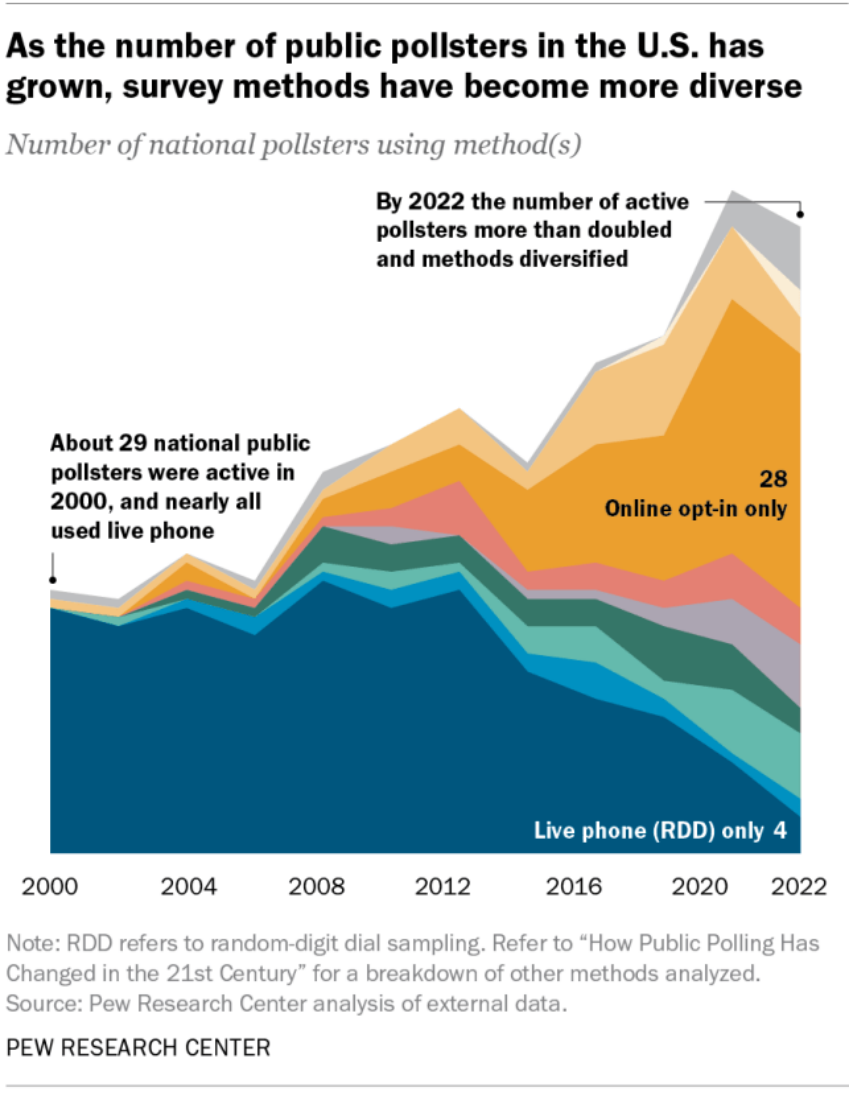
Note: The average signed error is the difference between the actual margin separating the candidates in the general election and the average margin in the polls. A negative error means that the Republican candidate's margin was overstated and a positive error means the Democratic candidate's margin was overstated.

Source: American Association for Public Opinion Research (AAPOR) Task Force on 2020 Pre-Election Polling: An Evaluation of the 2020 General Election Polls

PEW RESEARCH CENTER

Modern Polling

- Fewer live phone surveys
- More online opt-in surveys
- More active pollsters



Sampling

- Probabilitistic samples have less error than non-probabilistic samples
- Results coming out are only as good as the data going in
- Can you reliably, validly, generalizably describe the US when $n = 1000$? $n = 2000$?

Weighting

- Some demographics are hard to characterize in a pure random sample
- Some categories are oversampled while others are undersampled, not always intentionally
- Sampled data is adjusted to more closely match the study population characteristics

Likely Voters

- $\sim 1/3$ of eligible Americans do not vote in presidential elections
- Many people feel social pressure to say they'll vote even if it's unlikely
- The study population is not static over time!
- Popular vote does not necessarily win

Sources of Error

Sampling error is not the only kind of polling error

Error from ...	Error name	Reflected in margin of error
Excluding parts of the population	Noncoverage	No
Low response rates from certain groups	Nonresponse	No
People misunderstanding the question or misreporting their opinions	Measurement	No
Interviewing a sample rather than entire population	Sampling	Yes

PEW RESEARCH CENTER

Other Sources of Error

- Interviewer
- Responder
- Survey

Aggregators

- Combining results from multiple surveys may improve accuracy
- Are all polls created equal?
- If the underlying assumptions are faulty, more data won't improve the quality

Communicating Results

- Transparency
- Point estimates? Confidence intervals?
- Certainty

What to Look For

- Poll's sponsor and data collection firm
- Participant selection process
- Interview methods and dates
- Sample sizes, non-response rates
- Question phrasing
- Weighting methods