Sampling and Uncertainty

Gov 51: Data Analysis and Politics

Scott Cunningham

Harvard University

Week 4, Tuesday
February 18, 2026

Where We Are

Last week we built three tools for describing data:

Mean | **Variance** | **Covariance**

These describe *data we already have*. But political science usually needs something harder:

> What can a **sample** tell us about
> a **population** we can't observe?

Today we cross from description to **inference**.

What We'll Learn Today

1. **Why polls work**—random sampling and the logic of inference
2. **Probability distributions**—the Bernoulli and normal distributions, and why we need them
3. **The standard error**—a formula that tells us how much samples vary
4. **Confidence intervals**—quantifying our uncertainty about the truth

By the end of today, you'll understand what "margin of error ±3%" actually means.

Wrapping Up: Correlation from Last Thursday

We ended last week with **correlation**—standardized covariance:

$$r_{xy} = \frac{\mathrm{Cov}(x, y)}{s_x \cdot s_y}$$

Key takeaways to carry forward:

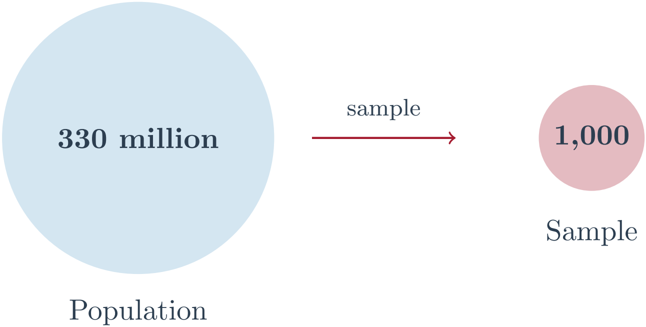
- r is unitless and ranges from -1 to $+1$
- It measures **linear** association only (Anscombe's quartet!)
- Correlation $\neq$ causation
- Covariance explained *why* weighted and unweighted means diverge

These tools describe patterns in data. Now we ask: can we trust those patterns?

Why Do Polls Work?

The Miracle of Polling

Claim: You can learn what 330 million Americans think by asking just 1,000.



This seems impossible. Why does it work?

The Key: Random Sampling

If every person has an **equal chance** of being selected, the sample will look like the population.

Intuition: Imagine a giant jar of marbles—60% blue, 40% red.

- Shake well and grab 100 marbles blindfolded
- You'll get *approximately* 60 blue, 40 red
- Not exactly—but close!

The same logic applies to polling voters.

What Could Go Wrong?

Random sampling fails when selection isn't truly random:

1. **Non-response bias**: People who answer phones differ from those who don't
2. **Coverage bias**: Your sampling frame misses some groups
   - 1936: *Literary Digest* polled car/phone owners → missed poor voters
3. **Social desirability bias**: People lie about unpopular views
4. **Likely voter screens**: Who will actually vote?

These are why polls can be wrong—not sampling error.

Today's Focus: Quantifying Uncertainty

Even with perfect random sampling, samples vary.

Question: How much variation should we expect?

Answer: It depends on sample size—and we can calculate it exactly.

52% 48% 51% 49% 50%

Different samples give different answers

Distributions and the Standard Error

Probability Was Born from Gambling and Astronomy

- **1650s**: Pascal and Fermat invented probability theory to settle gambling disputes—how should you split the pot in an interrupted dice game?
- **1700s–1800s**: Gauss, Laplace, and Legendre discovered that measurement errors in astronomy follow a bell-shaped curve—the **normal distribution**
- **Key insight**: The same mathematics that predicted planetary orbits now predicts polling errors

Two centuries of math, distilled into the formulas we'll use today.

Sampling Variability

Imagine the true population support for a candidate is 50%.

If we take many samples of size $n = 1000$:

- Sample 1 might show 51%
- Sample 2 might show 49%
- Sample 3 might show 48%
- And so on...

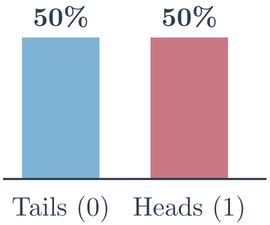
The **spread** of these sample estimates is called the **sampling distribution**.

Its standard deviation is the **standard error**.

A Distribution Describes How Outcomes Spread Out

A **probability distribution** tells you which values are possible and how likely each one is.

Example: Flip a fair coin. Two outcomes, equally likely.

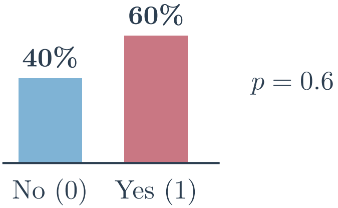


Every distribution has two key numbers: the **mean** (center) and the **variance** (spread).

Each Survey Response Is a Coin Flip

A **Bernoulli trial**: one yes/no outcome with probability p.
Let X be the result of asking **one** voter "Do you support Candidate A?"

$$X = \begin{cases} 1 & \text{if yes (with probability } p) \\ 0 & \text{if no (with probability } 1-p) \end{cases}$$



The true proportion $p = 0.6$ is **fixed**—it never changes. But you don't know what any *single* voter will say. That per-person uncertainty is what we measure.

The Bernoulli Mean and Variance

For a single Bernoulli draw X with probability p:

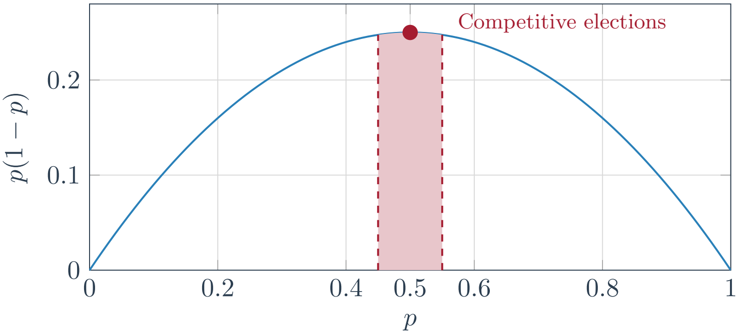
$$\text{Mean: } E[X] = p \qquad \text{Variance: } \mathrm{Var}(X) = p(1-p)$$

What does this variance measure? Not uncertainty about p—that's a constant. It measures the spread in **individual responses**: each person is either 0 or 1, and you can't predict which.

- If $p = 1$ (everyone says yes): no uncertainty, $\mathrm{Var} = 0$
- If $p = 0.5$ (coin flip): maximum uncertainty, $\mathrm{Var} = 0.25$
- If $p = 0.6$: $\mathrm{Var} = 0.6 \times 0.4 = 0.24$ (nearly maximal)

Competitive Elections Mean Maximum Uncertainty

Variance $p(1-p)$ is a parabola—maximized at $p = 0.5$, zero at the extremes:



Most U.S. elections fall in the 45–55% range—right at the peak. Competitive democracies produce the **hardest** polling problem: maximum variance, maximum uncertainty.

The Sample Proportion Averages Many Bernoulli Trials

Each person in our sample gives us one Bernoulli draw. The sample proportion averages them:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

When you average n independent Bernoulli trials:

- Mean of $\hat{p}$: still p (unbiased—centers on the truth)
- Variance of $\hat{p}$: $\frac{p(1-p)}{n}$ (shrinks with n!)

$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n} \quad \Rightarrow \quad \text{SE}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

But **how** does it shrink? Let's simulate it.

Why $1/n$ and Not $1/(n-1)$?

Last week we used $n-1$ when **estimating** variance from data: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. That correction exists because using $\bar{X}$ instead of the true mean costs one degree of freedom.

Here is different: we **already know** the variance of each draw is $p(1-p)$. We aren't estimating it—it's a property of the Bernoulli distribution.

The $1/n$ comes from a probability rule about averages:

If $X_1, \ldots, X_n$ are independent with variance σ^2, then $\mathrm{Var}(\bar{X}) = \dfrac{\sigma^2}{n}$

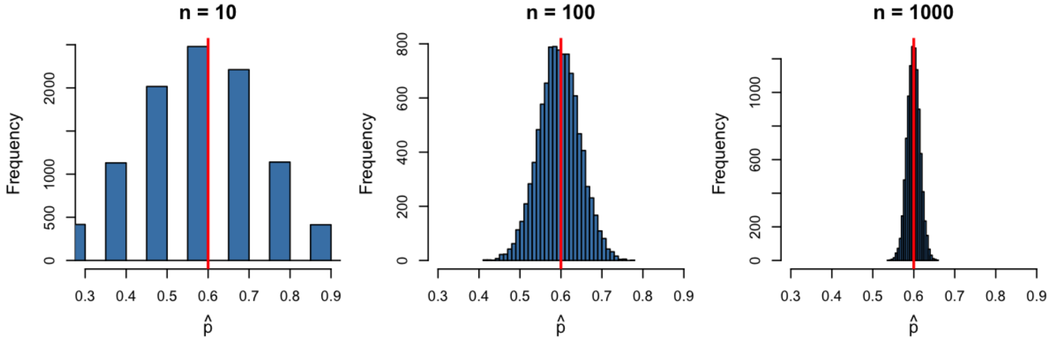
Known variance $\div$ sample size $=$ no correction needed.

Larger Samples Produce Tighter Estimates

```r
set.seed(51); p_true <- 0.6
par(mfrow = c(1, 3))
for (n in c(10, 100, 1000)) {
  p_hats <- replicate(10000,
                      mean(rbinom(n, 1, p_true)))
  hist(p_hats, breaks = 30,
       main = paste("n =", n),
       xlab = expression(hat(p)),
       col = "steelblue", xlim = c(0.3, 0.9))
  abline(v = p_true, col = "red", lwd = 2)
}
```

Run this code—what do the three histograms look like?

The Sampling Distribution Narrows with n

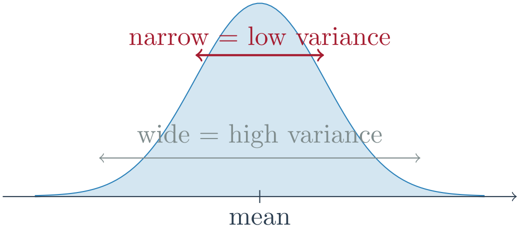


$n = 10$: wide and lumpy $\quad n = 100$: tighter $\quad n = 1000$: very narrow

The spread of these histograms **is** the standard error.

The Normal Distribution

The **normal distribution** (also called the "bell curve") is a symmetric, continuous distribution defined by two numbers: its mean and its variance.



- ▹ Symmetric around the mean—equally likely to be above or below
- ▹ Most values cluster near the center; extreme values are rare
- ▹ Completely described by just **mean** and **variance**

Regression to the Mean

Francis Galton (1886) measured heights of parents and children. He noticed: exceptionally tall parents had children who were tall—but *not as tall*. Short parents had children who were short—but *not as short*.

Why? Extreme values are rare in a normal distribution. If you're in the far tail, most of the probability mass is closer to the center. The next observation is likely to be less extreme.

- A team that wins 75% of games this season will probably win fewer next season
- A student who scores 99th percentile on one test will likely score lower on the next
- A poll showing a candidate at 62% will probably show closer to 50% next time

Extreme outcomes don't persist—the bell curve pulls everything back toward the center.

Large Samples Produce Bell-Shaped Distributions

The **Central Limit Theorem** (CLT):

> No matter what the original distribution looks like, the sampling distribution of $\hat{p}$ becomes approximately **normal** for large n.

- Each individual response is Bernoulli—just 0 or 1, not bell-shaped at all
- But the *average* of many responses forms a smooth bell curve
- Look back at the simulation: the $n = 1000$ histogram already looks normal

This is why we can use the normal distribution's properties—like the **1.96 rule**—to build confidence intervals.

This is one of the most important results in all of statistics.

The Standard Error Formula

For a proportion (like support for a candidate):

$$\text{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Where:

- $\hat{p}$ = sample proportion (e.g., 0.52 = 52%)
- n = sample size

Key insight: SE shrinks as n grows—but slowly (square root).

Example: A Typical Poll

A poll of $n = 1,000$ voters finds 52% support for Candidate A.

$$\text{SE} = \sqrt{\frac{0.52 \times 0.48}{1000}} = \sqrt{\frac{0.2496}{1000}} = \sqrt{0.0002496} \approx 0.016$$

So SE $\approx$ 1.6 percentage points.

Interpretation: If we repeated this poll many times, the results would typically vary by about 1.6 points.

The Margin of Error

The **margin of error** (MOE) you see in news reports is:

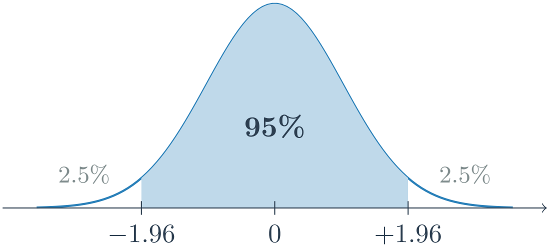
$$\boxed{\text{MOE} \quad = \quad 1.96 \times \text{SE} \quad \approx \quad 2 \times \text{SE}}$$

For our example: MOE $= 2 \times 1.6 = 3.2$ percentage points.

So the poll would report: **"52% $\pm$ 3 points"**

The "$\pm$ 3" is the margin of error.

Why 1.96? The 95% Confidence Level



95% of a normal distribution falls within ±1.96 standard deviations.

So 95% of samples will fall within ±1.96 × SE of the truth.

The Confidence Interval

A **95% confidence interval** is:

$$\text{CI} \quad = \quad \hat{p} \pm 1.96 \times \text{SE}$$

For our poll: $0.52 \pm 0.032 = [0.488, 0.552]$ or **[48.8%, 55.2%]**

Interpretation (careful!):

- Correct: If we repeated this procedure many times, 95% of intervals would contain the true value
- Incorrect: There's a 95% chance the true value is in this interval

The true value is fixed—either it's in the interval or it isn't.

How Sample Size Affects Uncertainty

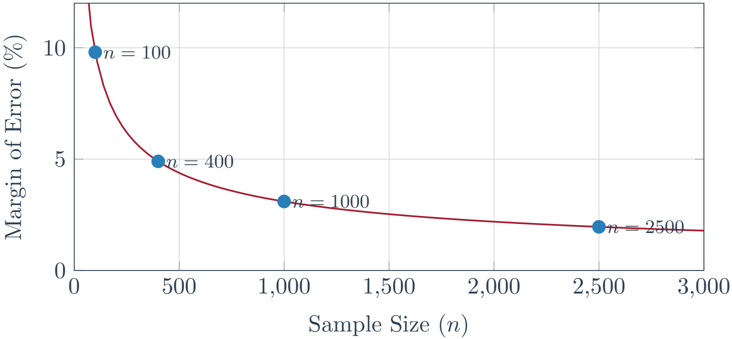
The Square Root Rule

Sample Size (n)	Standard Error	Margin of Error
100	5.0%	±10%
400	2.5%	±5%
1,000	1.6%	±3%
2,500	1.0%	±2%
10,000	0.5%	±1%

Pattern: To halve the MOE, you need $4\times$ the sample size.

This is why most polls use 1,000–1,500 respondents.
Beyond that, improvements are expensive and small.

Visualizing Sample Size Effects



Diminishing returns: the curve flattens as n grows.

Population Size Doesn't Matter (Much)

Surprising fact: The margin of error depends on sample size, *not* population size.

- A poll of 1,000 Americans (pop: 330 million) has MOE $\approx$ 3%
- A poll of 1,000 Bostonians (pop: 700,000) has MOE $\approx$ 3%
- A poll of 1,000 Harvard students (pop: 25,000) has MOE $\approx$ 3%

The absolute number sampled matters, not the fraction.

(There's a small "finite population correction" but it rarely matters.)

Real Data: The 2008 Election

Case Study: Obama vs. McCain

We have 1,333 state-level polls from the 2008 presidential election.

Data: `polls08.csv`

- 50 states + DC
- Multiple pollsters per state
- Poll dates from June to November 2008

Question: How much did polls vary within states?

And how close were they to the actual results?

Loading the Data

```
library(tidyverse)
polls08 <- read_csv("polls08.csv")

head(polls08, 4)
## # A tibble: 4 x 5
##   state Pollster          Obama McCain middate
##   <chr> <chr>             <dbl>  <dbl> <date>
## 1 AL    SurveyUSA-2          36     61 2008-10-27
## 2 AL    Capital Survey-2     34     54 2008-10-15
## 3 AL    SurveyUSA-2          35     62 2008-10-08
## 4 AL    Capital Survey-2     35     55 2008-10-06
```

Variation Within a Swing State

```
# Focus on Florida
florida <- polls08 %>%
  filter(state == "FL")

# How many polls?
nrow(florida)
## [1] 73

# What's the range of Obama support?
range(florida$Obama)
## [1] 44 53
```

73 polls, with Obama support ranging from 44% to 53%—a 9-point spread!

Some of this is sampling error. Some is real change over time.

Why Do Polls Disagree?

When two polls show different results, it could be:

1. **Sampling error**: Random variation (expected!)
2. **Different timing**: Opinion changed between polls
3. **Different methods**:
   - Phone vs. online
   - Likely voter vs. registered voter screens
   - Question wording
4. **House effects**: Some pollsters consistently lean D or R
   - Not intentional—different methods produce different systematic errors

This is why we aggregate polls—to average out the noise.

Comparing Polls to Results

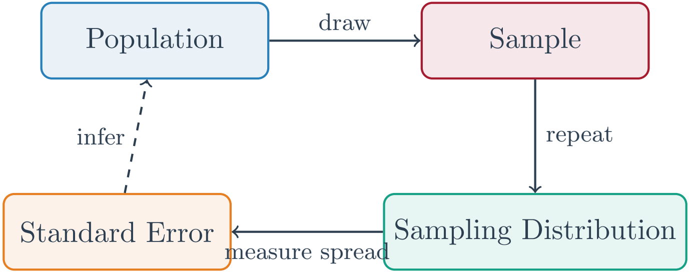
```
# Load actual results
pres08 <- read_csv("pres08.csv")

# Florida actual result
pres08 %>% filter(state == "FL")
## # A tibble: 1 x 5
##   state.name state Obama McCain    EV
##   <chr>      <chr> <dbl>  <dbl> <dbl>
## 1 Florida    FL       51     48    27
```

Obama won Florida with 51%.

The final polls showed 48-53%, so most were close—but some missed badly.

Key Concepts

The Frequentist Framework



The standard error tells us how much samples vary, which lets us make inferences about the population.

Vocabulary Summary

Population	The entire group we want to learn about
Sample	A subset we actually observe
Parameter	True value in the population (unknown, fixed)
Estimate	Our best guess from the sample
Standard Error	How much estimates vary across samples
Margin of Error	$\approx 2\times$ SE (for 95% confidence)
Confidence Interval	Range that contains the truth 95% of the time

What We Learned Today

1. **Random sampling** is why polls work
2. **Standard error** quantifies sampling uncertainty: $\text{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
3. **Margin of error** $\approx 2 \times \text{SE}$
4. **Sample size matters**: SE shrinks with $\sqrt{n}$
5. Real polls vary—some due to sampling, some due to methods

Thursday: How do we combine multiple polls to get better estimates?

**Every poll has uncertainty.
The margin of error
tells you how much.**

Questions?