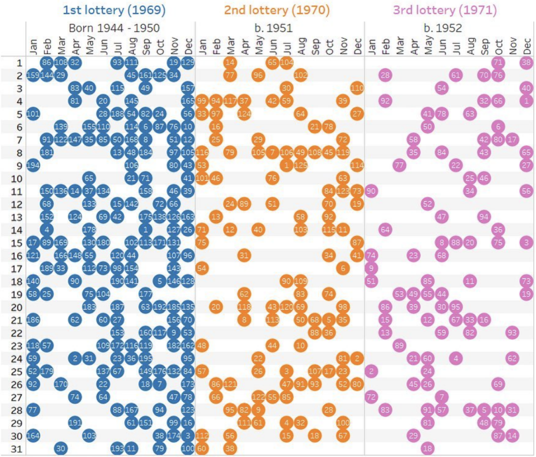
Sampling

Data Analysis for Journalism and Political Communication (Spring 2026)

Prof. Bell

1970 Vietnam War Draft

1970 Vietnam War Draft



Birthdates of US servicemen drafted into the Vietnam War as a result of birthdate lotteries held in 1969, 1970 and 1971

Source: @visyuval

Note: The numbers denote the order that the birthdates were drawn, as this determined the order of call. The highest lottery number called for duty in the 1st, 2nd and 3rd lotteries was 195, 125 and 95, respectively.

DEFINITIONS

- The group we are interested in studying is known as the **population**

Definitions

- The group we are interested in studying is known as the **population**
- Often, we are not able to count every unit in the population, so we take a **sample**

DEFINITIONS

- The group we are interested in studying is known as the **population**
- Often, we are not able to count every unit in the population, so we take a **sample**
- Our best guess about the population based on our sample is the **estimate**

Definitions

- The group we are interested in studying is known as the **population**
- Often, we are not able to count every unit in the population, so we take a **sample**
- Our best guess about the population based on our sample is the **estimate**
- The key to a good estimate is a quality sample, which is determined by two elements:
  1. A **random sample** of the population
  2. The **sample size** is sufficiently large

In-class exercise

SAMPLE SIZE

- How many units should you sample from the population?

Sample Size

- How many units should you sample from the population? **It depends on your desired level of certainty.**

Sample Size

- How many units should you sample from the population? **It depends on your desired level of certainty.**
- The most common level of certainty is 95% (the inverse of a **p-value** of .05, meaning that there is a 5% chance we are committing Type I error)

Sample Size

- How many units should you sample from the population? **It depends on your desired level of certainty.**
- The most common level of certainty is 95% (the inverse of a **p-value** of .05, meaning that there is a 5% chance we are committing Type I error)
- In other words, there is a 5% chance that the true population value is outside of the **confidence interval**

Sample Size

- How many units should you sample from the population? **It depends on your desired level of certainty.**
- The most common level of certainty is 95% (the inverse of a **p-value** of .05, meaning that there is a 5% chance we are committing Type I error)
- In other words, there is a 5% chance that the true population value is outside of the **confidence interval**
- If we re-sampled the population 100 times, 95 of our estimates would fall within the confidence interval (let's see this in action!)

Google Colab Example

Margin of Error

- The confidence interval for a proportion[1] is also called the **margin of error (MOE)**

[1] There is a different formula for continuous variables.

Margin of Error

- The confidence interval for a proportion[1] is also called the **margin of error (MOE)**
- The 95% MOE is calculated as:

$$1.96 * \sqrt{p * (1 - p)/n}$$

where p is the proportion and n is the sample size

[1] There is a different formula for continuous variables.

Margin of Error

- The confidence interval for a proportion[1] is also called the **margin of error (MOE)**
- The 95% MOE is calculated as:

  $$1.96 * \sqrt{p * (1 - p)/n}$$

  where p is the proportion and n is the sample size
- Typically, pollsters will use a proportion (p) of .5 to calculate an MOE for the entire poll:

  $$1.96 * \sqrt{.5 * (1 - .5)/1000} = .031$$

[1] There is a different formula for continuous variables.

Margin of Error

- The confidence interval for a proportion[1] is also called the **margin of error (MOE)**
- The 95% MOE is calculated as:

  $$1.96 * \sqrt{p * (1 - p)/n}$$

  where p is the proportion and n is the sample size
- Typically, pollsters will use a proportion (p) of .5 to calculate an MOE for the entire poll:

  $$1.96 * \sqrt{.5 * (1 - .5)/1000} = .031$$

- We report the estimate with the MOE, e.g., 45 +/- 3.1%.

[1] There is a different formula for continuous variables.

Sample Size and the Margin of Error

- Mathematically, the MOE gets smaller as the sample size increases

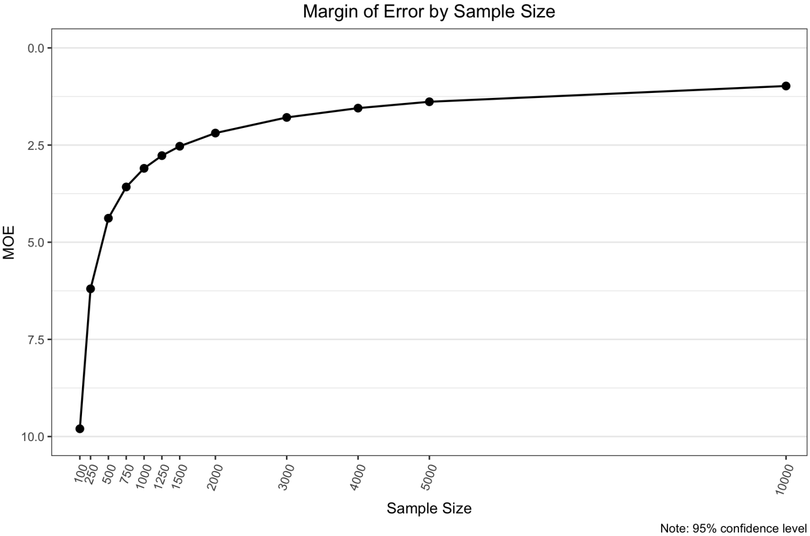
Sample Size and the Margin of Error

- Mathematically, the MOE gets smaller as the sample size increases
- Intuitively, the larger our sample size, the less likely we are to draw an “outlier” sample and the more likely we are to get a representative sample

Sample Size and the Margin of Error

- Mathematically, the MOE gets smaller as the sample size increases
- Intuitively, the larger our sample size, the less likely we are to draw an "outlier" sample and the more likely we are to get a representative sample
- But the marginal improvement in the MOE from adding units to the sample decreases as the sample size grows

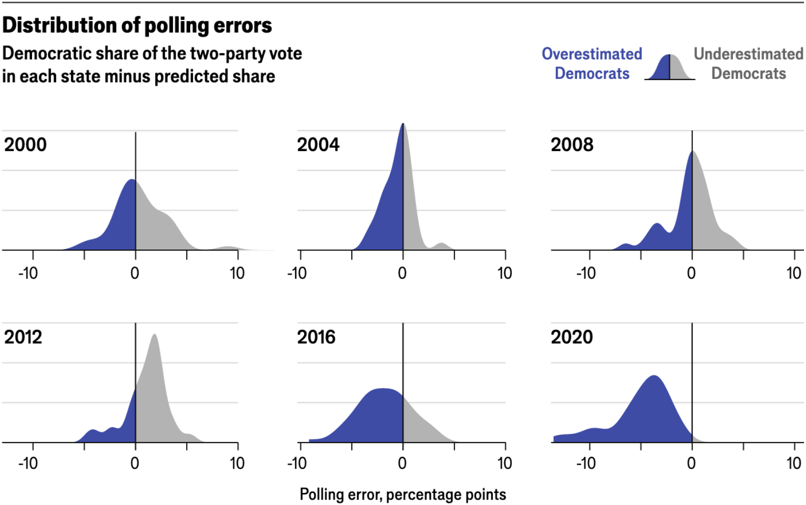
Sample Size and the Margin of Error



Sample Size and the Margin of Error

- Mathematically, the MOE gets smaller as the sample size increases
- Intuitively, the larger our sample size, the less likely we are to draw an "outlier" sample and the more likely we are to get a representative sample
- But the marginal improvement in the MOE from adding units to the sample decreases as the sample size grows
- Remember that the MOE only takes into account the sample size, not the potential for selection bias

Sample Size and the Margin of Error



Source: The Economist