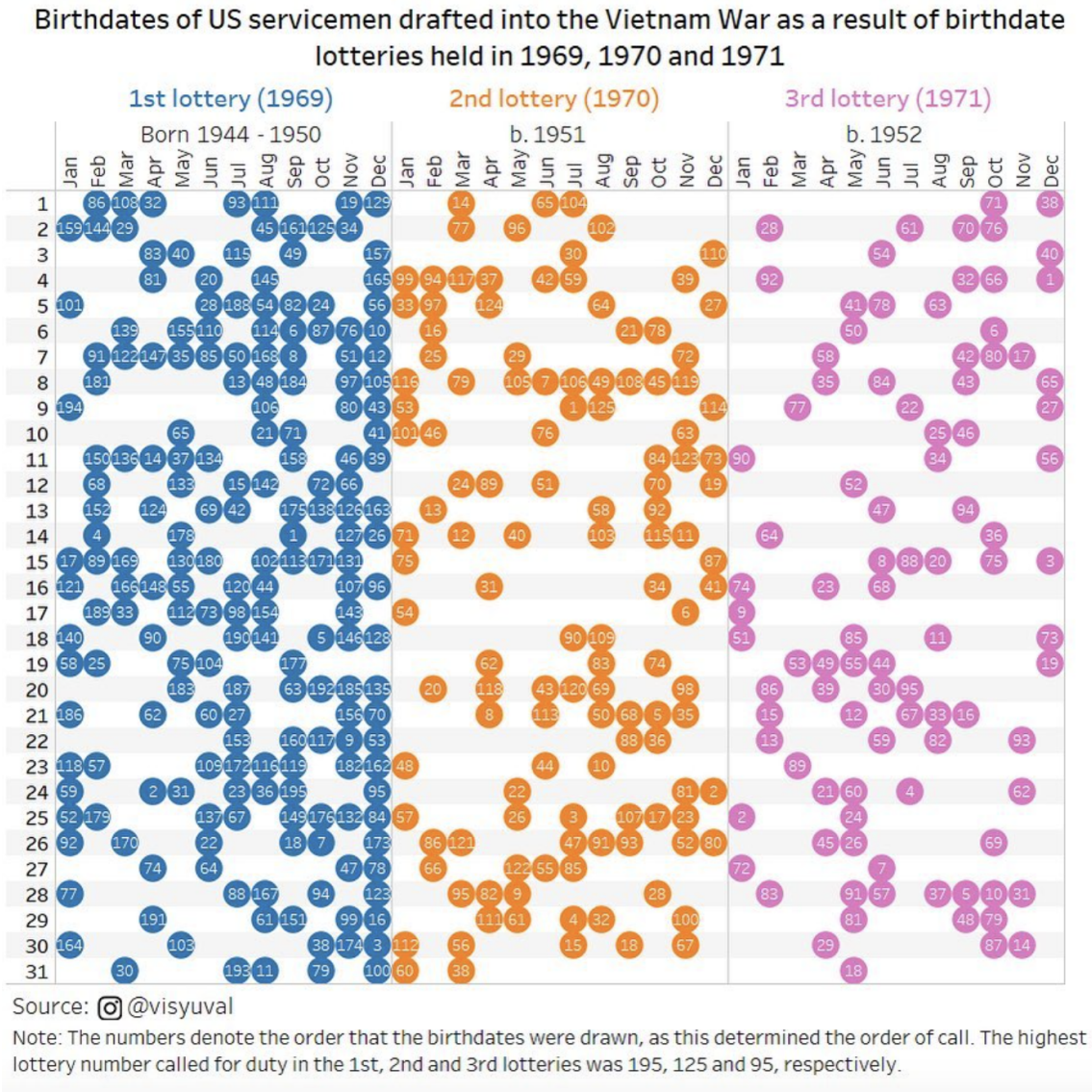
# Sampling

Data Analysis for Journalism and Political Communication (Fall 2024)

Prof. Bell

# 1970 Vietnam War Draft

# 1970 Vietnam War Draft



# 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**

# Sampling

- There are many ways to derive an estimate from a sample, but recall that “garbage in = garbage out”: no amount of statistical wizardry can compensate for bad data

# Sampling

- There are many ways to derive an estimate from a sample, but recall that “garbage in = garbage out”: no amount of statistical wizardry can compensate for bad data
- The key to any data analysis project is a quality sample, which is determined by two elements:

# Sampling

- There are many ways to derive an estimate from a sample, but recall that “garbage in = garbage out”: no amount of statistical wizardry can compensate for bad data
- The key to any data analysis project is a quality sample, which is determined by two elements:
  1. A **random sample** of the population

# Sampling

- There are many ways to derive an estimate from a sample, but recall that "garbage in = garbage out": no amount of statistical wizardry can compensate for bad data
- The key to any data analysis project is a quality sample, which is determined by two elements:
  1. A **random sample** of the population

## Definition

The probability of any given unit being drawn from the population is uniform (the same)

# Sampling

- There are many ways to derive an estimate from a sample, but recall that “garbage in = garbage out”: no amount of statistical wizardry can compensate for bad data
- The key to any data analysis project 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!)

# Margin of Error

- There is a mathematical formula to estimate the confidence interval for a continuous mean, but more frequently we estimate a proportion

# Margin of Error

- There is a mathematical formula to estimate the confidence interval for a continuous mean, but more frequently we estimate a proportion
- The confidence interval for a proportion is also called the **margin of error (MOE)**

# Margin of Error

- There is a mathematical formula to estimate the confidence interval for a continuous mean, but more frequently we estimate a proportion
- The confidence interval for a proportion 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

# Sample Size and the Margin of Error

- E.g. Washington Post-University of Maryland poll, December 14-18, 2023

# Sample Size and the Margin of Error

- E.g. Washington Post-University of Maryland poll, December 14-18, 2023

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

# Sample Size and the Margin of Error

- E.g. Washington Post-University of Maryland poll, December 14-18, 2023

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

$$1.96 * \sqrt{.33 * (1 - .33)/1024} = .029$$

# Sample Size and the Margin of Error

- E.g. Washington Post-University of Maryland poll, December 14-18, 2023

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

  $$1.96 * \sqrt{.33 * (1 - .33)/1024} = .029$$

- We report the estimate with the MOE, i.e., 33 +/- 2.9%.

# Sample Size and the Margin of Error

- E.g. Washington Post-University of Maryland poll, December 14-18, 2023

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

$$1.96 * \sqrt{.33 * (1 - .33)/1024} = .029$$

- We report the estimate with the MOE, i.e., 33 +/- 2.9%.
- This means that there is a 5% chance that the true population proportion is outside of about 30-36%.

# Sample Size and the Margin of Error

- E.g. Washington Post-University of Maryland poll, December 14-18, 2023

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

$$1.96 * \sqrt{.33 * (1 - .33)/1024} = .029$$

- We report the estimate with the MOE, i.e., 33 +/- 2.9%.
- This means that there is a 5% chance that the true population proportion is outside of about 30-36%. *(Note: this does not mean that every value in the MOE is equally likely!)*

# Sample Size and the Margin of Error

- E.g. Washington Post-University of Maryland poll, December 14-18, 2023

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

$$1.96 * \sqrt{.33 * (1 - .33)/1024} = .029$$

- We report the estimate with the MOE, i.e., 33 +/- 2.9%.
- This means that there is a 5% chance that the true population proportion is outside of about 30-36%. *(Note: this does not mean that every value in the MOE is equally likely!)*
- Typically, pollsters will use a proportion ($p$) of .5 to calculate an MOE for the entire poll, rather than individual questions

# Sample Size and the Margin of Error

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

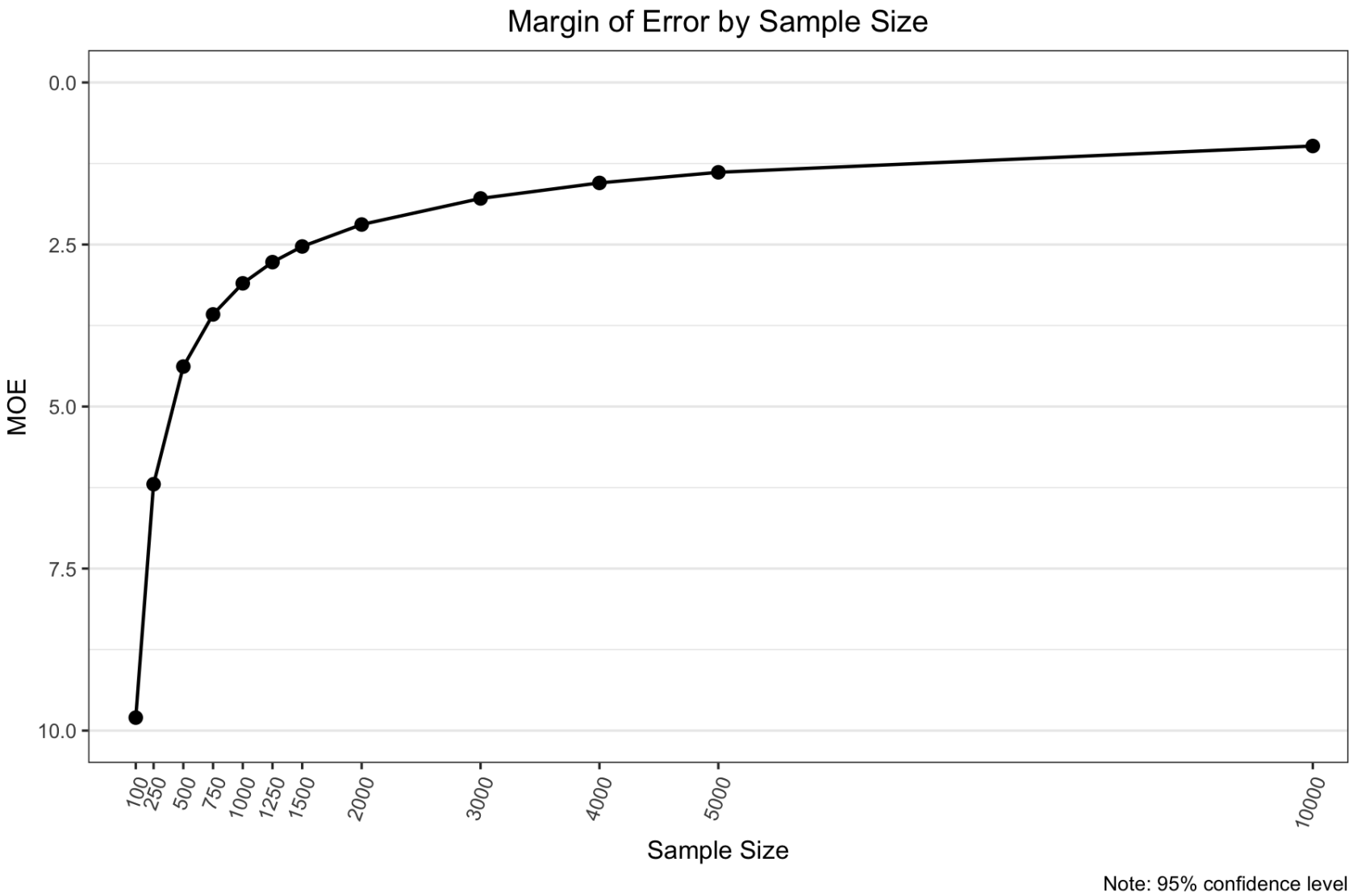
# Sample Size and the Margin of Error

- 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

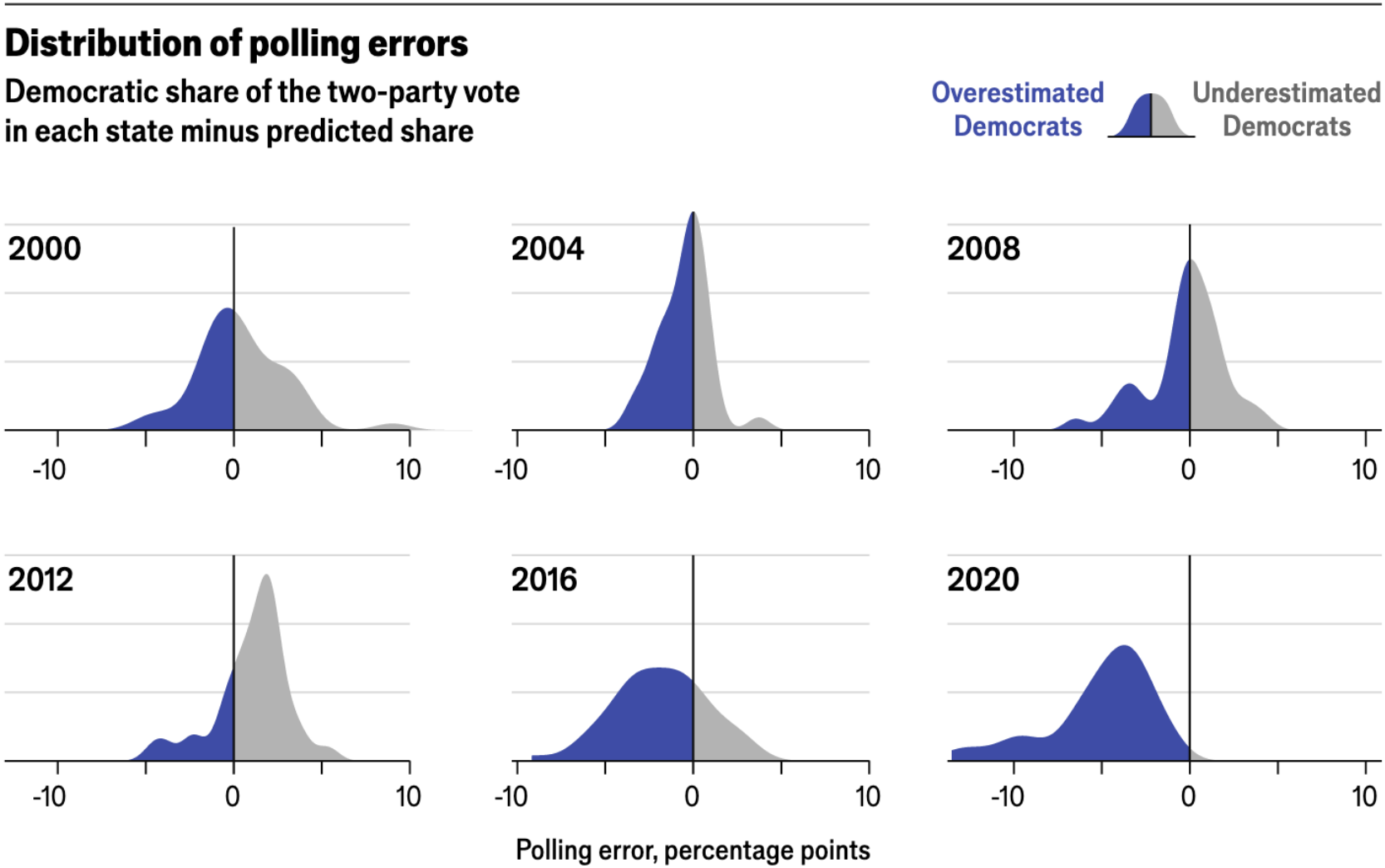
# Sample Size and the Margin of Error



# 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

# Getting a Random Sample

How would you generate a random sample of American voters?

# Getting a Random Sample

- **Systematic Sampling**: Select every $n^{th}$ unit of a population (e.g. exit polls)

# Getting a Random Sample

- **Systematic Sampling**: Select every $n^{th}$ unit of a population (e.g. exit polls)
- **List-based Sampling**: Randomly select units from an existing list of the population (e.g. registered voter lists)

# Getting a Random Sample

- **Systematic Sampling**: Select every $n^{th}$ unit of a population (e.g. exit polls)
- **List-based Sampling**: Randomly select units from an existing list of the population (e.g. registered voter lists)
- **Address-based Sampling (ABS)**: Randomly select households from a list of addresses provided by the U.S. Postal Service

# Getting a Random Sample

- **Systematic Sampling**: Select every $n^{th}$ unit of a population (e.g. exit polls)
- **List-based Sampling**: Randomly select units from an existing list of the population (e.g. registered voter lists)
- **Address-based Sampling (ABS)**: Randomly select households from a list of addresses provided by the U.S. Postal Service
- **Random-digit Dialing (RDD)**: Randomly select area codes, and then random digits are added to the end to create 10-digit phone numbers

# Getting a Random Sample

- **Systematic Sampling**: Select every $n^{th}$ unit of a population (e.g. exit polls)
- **List-based Sampling**: Randomly select units from an existing list of the population (e.g. registered voter lists)
- **Address-based Sampling (ABS)**: Randomly select households from a list of addresses provided by the U.S. Postal Service
- **Random-digit Dialing (RDD)**: Randomly select area codes, and then random digits are added to the end to create 10-digit phone numbers
- **Non-probability/Quota Sampling**: Pseudo-randomly selecting, from an opt-in pool of respondents, a sample that approximates the make-up of the general population