In this module, you will learn:

- Why random sampling is of cardinal importance in political research
- Why samples that seem small can yield accurate information about much larger groups
- How to figure out the margin of error for the information in a sample
- How to use the normal curve to make inferences about the information in a sample

Inferential Statistics

- Refers to a set of procedures for deciding how closely a relationship we observe in a sample corresponds to the unobserved relationship in the population from which the sample was drawn.

Population

- A **population** may be defined generically as the universe of cases the researcher wants to describe.
- A characteristic of a population, such as the mean level of support for the Democratic Party (as measured by the Democratic feeling thermometer) is a *population parameter*.
- Population mean is symbolized by μ (“mew”)
- Ordinarily, we cannot directly observe population parameters.

Sample

- A **sample** is a number of cases or observations drawn from a population.
- A characteristic of a sample, such as the mean level of support for the Democratic Party, is a *sample statistic*.
- Sample mean is symbolized by $\bar{x}$ ("x bar")

Example: Student pollsters

- Student polling group wants to estimate the mean Democratic feeling thermometer rating (μ) in the student population.
- They plan to take a sample of size *n* and calculate the mean Democratic feeling thermometer rating ($\bar{x}$) of the sample.
- How accurately will $\bar{x}$ estimate μ?

3 factors determine how accurately $\bar{x}$ estimates μ

- The sampling procedure used
- The size of the sample (*n*)
- The amount of variation (σ) in the population parameter being estimated

Sampling procedure: Random

- In taking a **random sample**, the researcher ensures that every member of the population has an equal chance of being chosen for the sample.
  - In a student population (*N*) of 20,000, each student has a 1/20,000 chance of being chosen.
- A random sample eliminates selection bias, a source systematic error.
- A random sample introduces random sampling error.

Random sampling error

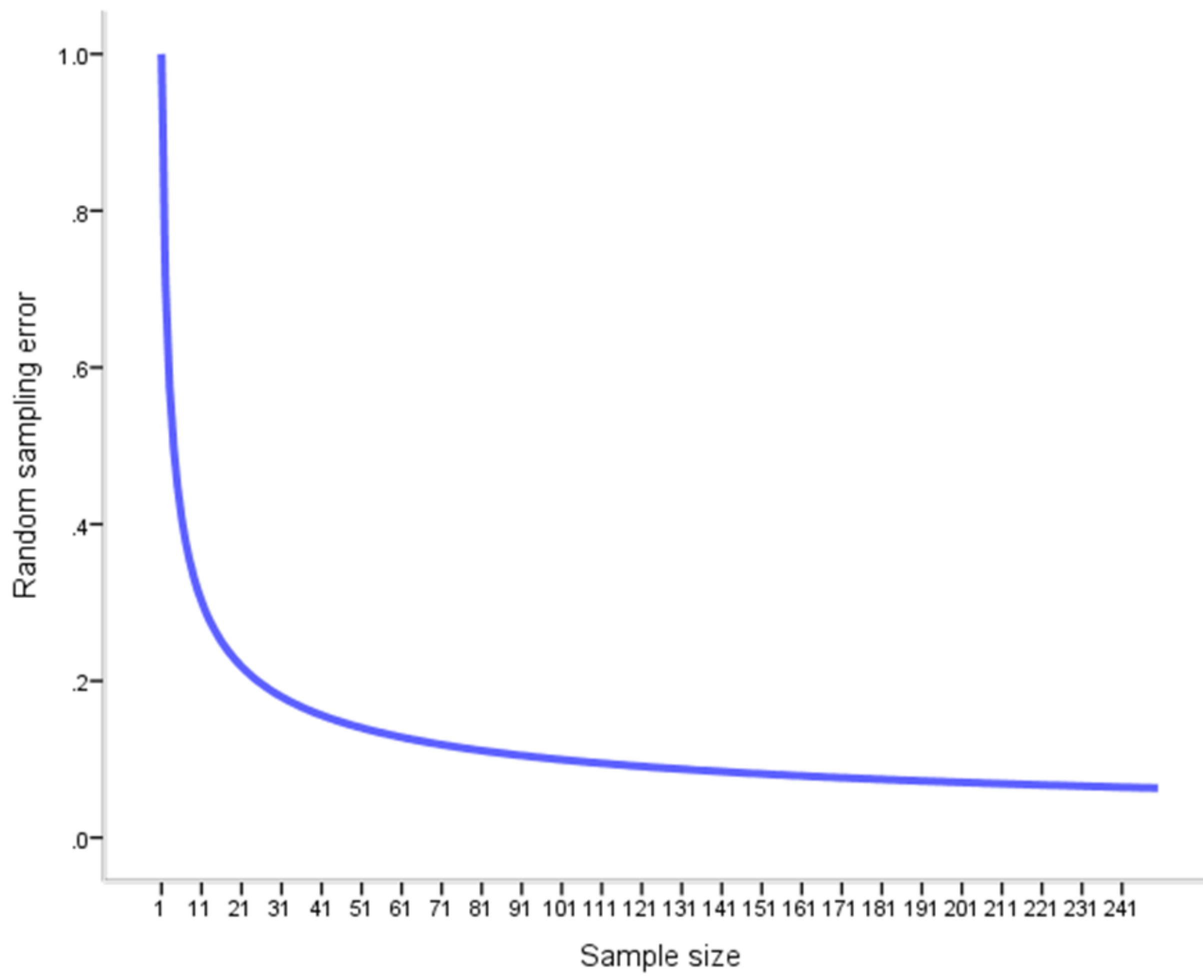
- Random sampling error is defined as the extent to which a sample statistic differs, *by chance,* from a population parameter.
- Population parameter = Sample statistic + Random sampling error
- $\mu = \bar{x}$ + Random sampling error

How much random sampling error is contained in a sample statistic?

- Depends on the sample size (n)
- Depends on the amount of variation in the population parameter (σ)
- Random sampling error = $\sigma / \sqrt{n}$

Sample size and Random sampling error

- Random sampling error declines as a function of the inverse of the square root of sample size (n).
- Because of this, we do not need huge samples to obtain accurate estimates of population parameters.


Random sampling error
1.0
.8
.6
.4
.2
.0
1 11 21 31 41 51 61 71 81 91 101 111 121 131 141 151 161 171 181 191 201 211 221 231 241
Sample size

Variation and Random sampling error

- As variation in the population parameter increases, random sampling error increases
- Assume that the pollsters' sample size (*n*) is equal to 100.
- How does the amount of variation in Democratic ratings in the population affect random sampling error?

Population A and Population B

- Both populations have the same population mean
  - In both A and B, $\mu = 58$
- But A has more variation, a larger value of σ, than B
- So, random sampling error is larger in A than in B: $\sigma_A/\sqrt{n_A} > \sigma_B/\sqrt{n_B}$

Illustrating $\sigma/\sqrt{n}$

- Population A: $\mu = 58$ and $\sigma \approx 25$
- Population B: $\mu = 58$ and $\sigma \approx 18$
- From each population take:
  - Ten random samples of n=25
  - Ten random samples of n=100
  - Ten random samples of n=400
- Calculate for each sample
- How accurately will the values of estimate μ?

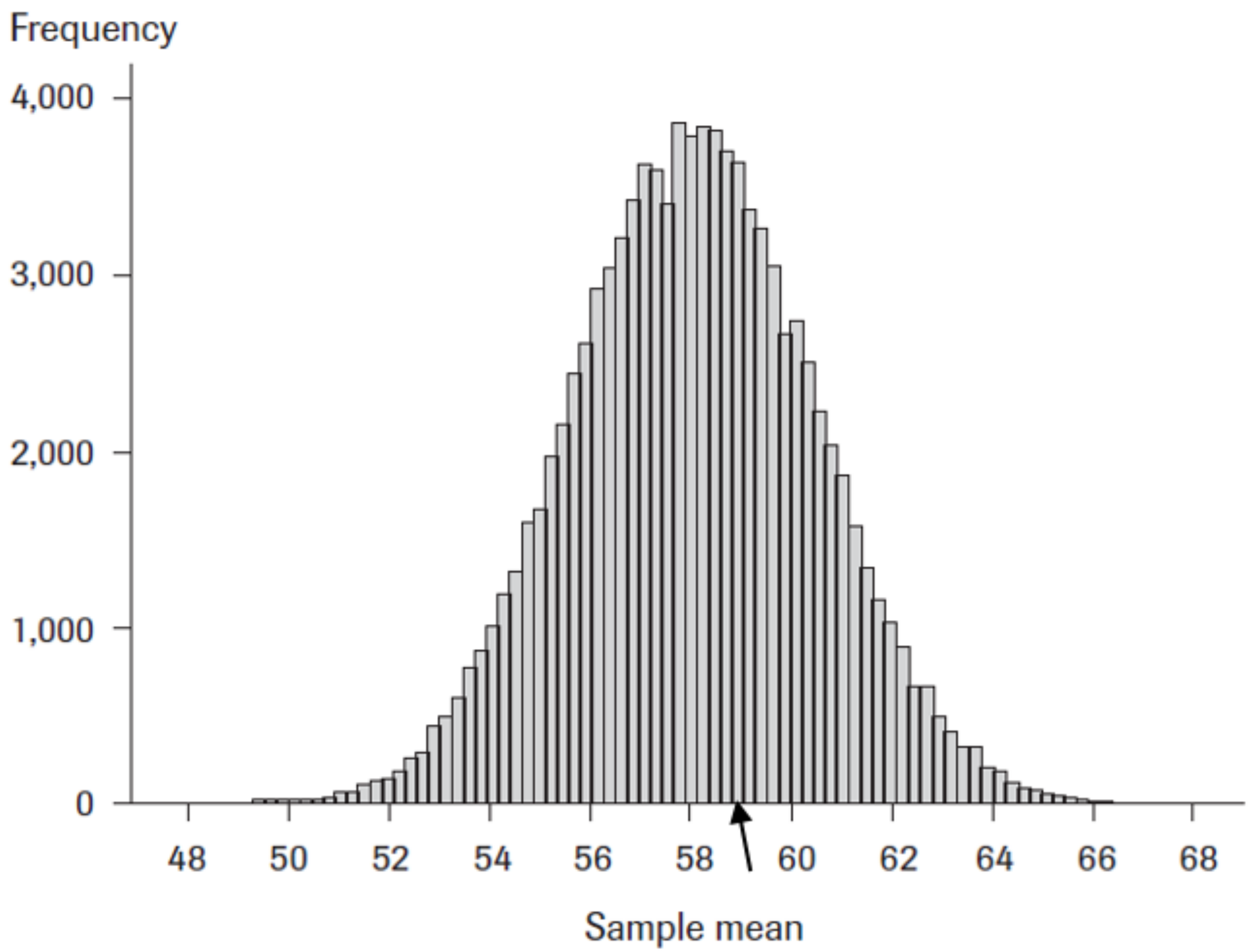
Standard error

- In estimating a population parameter, we ordinarily do not use the term "random sampling error."
- We use the term **standard error**.
- The two terms are synonymous
  - Random sampling error = $\sigma / \sqrt{n}$
  - Standard error of a sample mean = $\sigma / \sqrt{n}$

The Central Limit Theorem [1]

- Earlier we took only ten samples of n=100 from the student population
- Imagine taking 100,000 samples of n=100 and recording $\bar{x}$ for each sample
- What would the distribution of the 100,000 sample means look like?

Figure 6-3 Distribution of Means from 100,000 Random Samples



Note: Displayed data are means from 100,000 samples of $n = 100$. Population parameters: $\mu = 58$ and $\sigma = 24.8$.

The Central Limit Theorem [2]

- If we were to take an infinite number of samples of size n from a population of N members, the means of these samples would be normally distributed.
- The distribution of sample means would have a mean equal to the true population mean and have random sampling error equal to $\sigma/\sqrt{n}$.

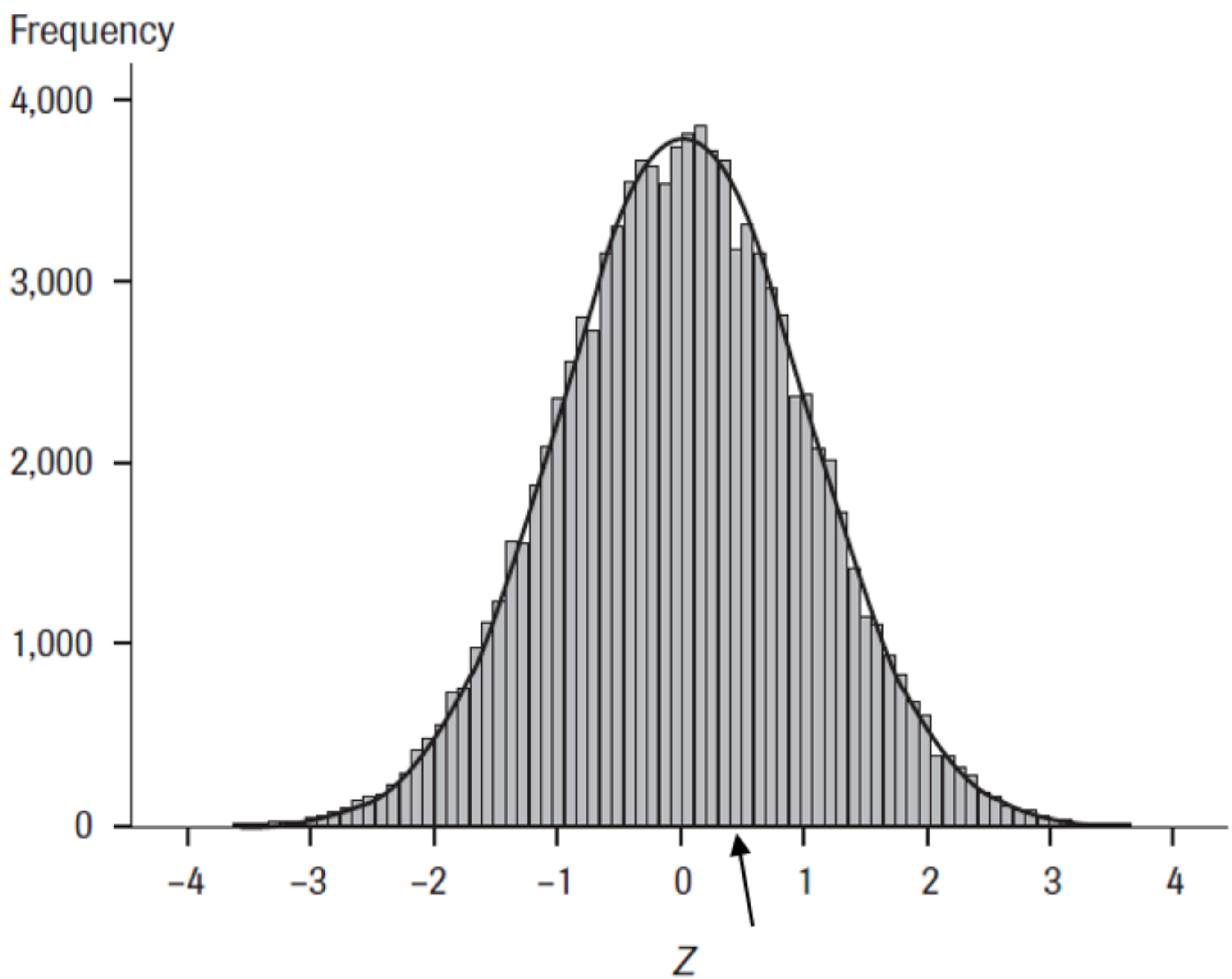
The Central Limit Theorem [3]

- Therefore, most random samples of n = 100 drawn from a population where μ=58 and σ=24.8 will yield means that are equal to 58 ± 24.8/√100 or 58 ± 2.5 or so: Between 55.5 and 60.5.
  - The student pollsters' mean of 59 falls in this high-probability interval.

The Normal Distribution [1]

- The normal distribution allows us to make precise inferences about the percentage of sample means that will fall within any given number of standard errors of the true population mean.
- Consider the standardized transformation of the distribution of 100,000 sample means.

Figure 6-4 Raw Values Converted to Z Scores



The Normal Distribution [2]

- The **normal distribution** is a distribution used to describe interval-level variables.
- To use the normal distribution, we first need to standardize each of the 100,000 sample means
- **Standardization** occurs when the numbers in a distribution are converted into standard units of deviation from the mean of the distribution.

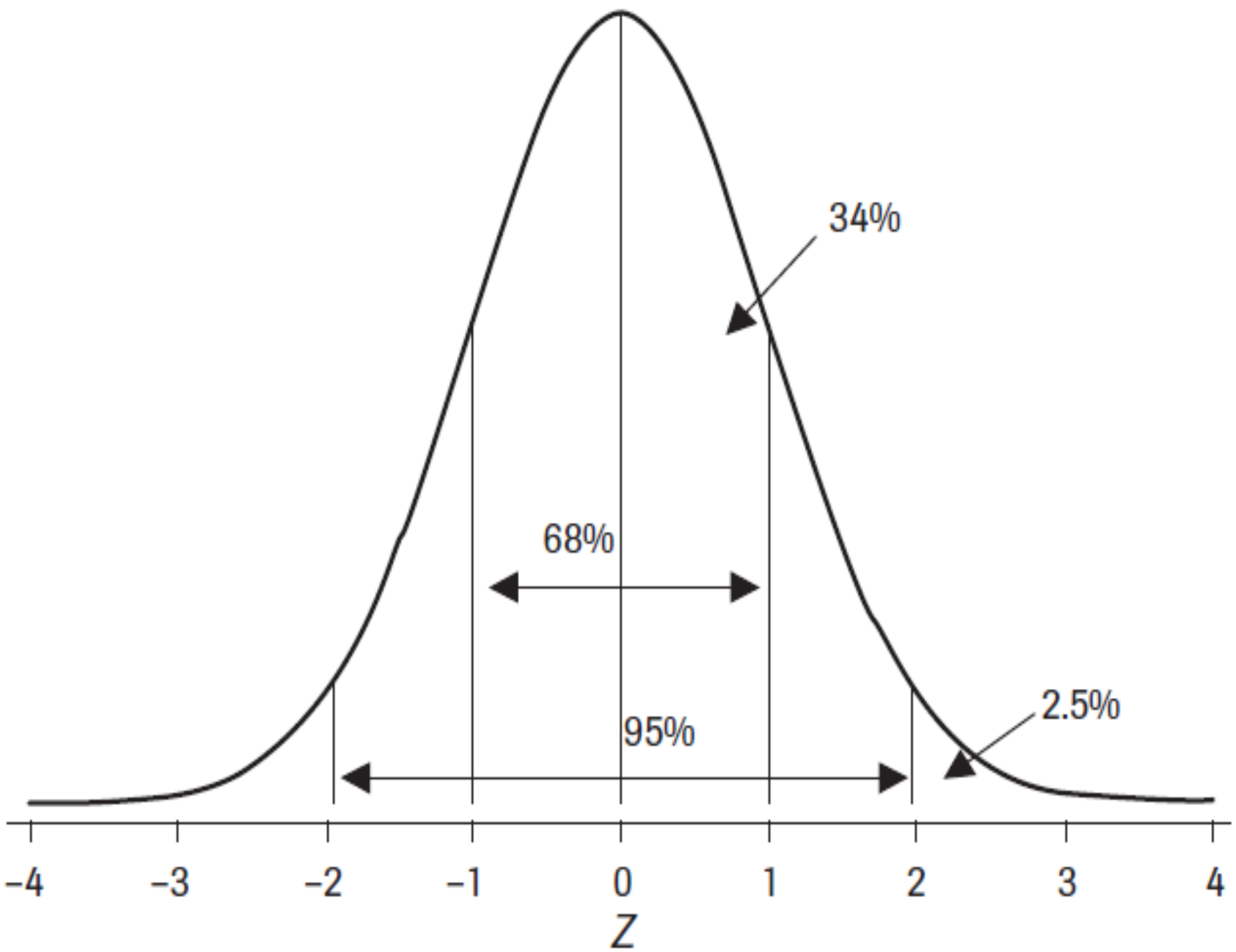
Z-scores

- To standardize the distribution of sample means, we would subtract the population mean from the sample mean, and then divide by the standard error
- Z = ($\bar{x}$ - μ) / (standard error)
- Z = (59 – 58) / 2.48 ≈ .40
  – Pollsters' mean lies .40 standard errors above the population mean.

The inferential power of Z

- Once a distribution is standardized, we can use the normal curve to determine the **probability** of observing any given sample mean, by chance.
- A **probability** is defined as the likelihood of the occurrence of an event or set of events.

Figure 6-5 Areas under the Normal Curve



Areas under the curve

- 68 percent of all possible sample means will fall in the interval between 1 standard error below the population mean (Z=-1) and 1 standard error above the population mean (Z=+1)
- 95 percent will fall between Z=-1.96 and Z=+1.96
- 5 percent of the time we will obtain sample means that are wide of the mark: 2.5 percent in the region below Z=-1.96 and 2.5 percent in the region about Z=+1.96.

Inference using the normal curve [1]

- The student pollsters do not know the population mean, but they do have a sample mean ($\bar{x}$ =59) and a standard error (≈ 2.5).
- Using the central limit theorem, they know that there is a 95% probability that μ lies between $\bar{x}$ ± 1.96(2.5):
  - 59 – 1.96 (2.5) = **54.1** at the low end, and
  - 59 + 1.96 (2.5) = **63.9** at the high end.

Inference using the normal curve [2]

- The 95 percent confidence interval or 95% CI defines the boundaries of plausible hypothetical claims and implausible hypothetical claims.
- All hypothetical values of μ that fall within the 95% CI are considered plausible and are not rejected.
- All hypothetical values of μ that fall outside the 95% CI are considered implausible and are rejected.

Applying the 95% CI

- Suppose someone claims that the true value of μ is equal to 66.
- Because 66 lies above the upper confidence boundary (63.9), we know that the probability is less than .025 that the true mean is 66.
- Reject the claim.