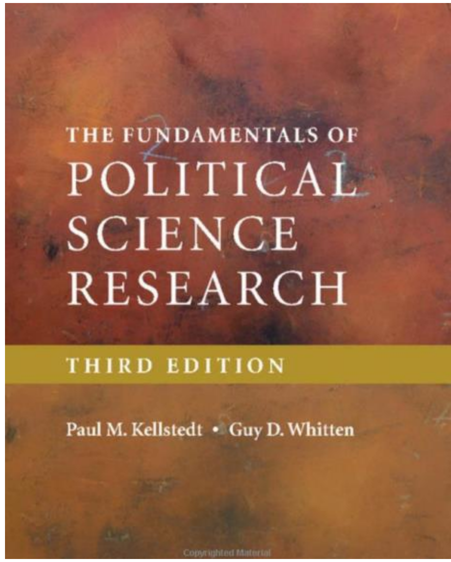


The Fundamentals of Political Science Research, 3rd edition

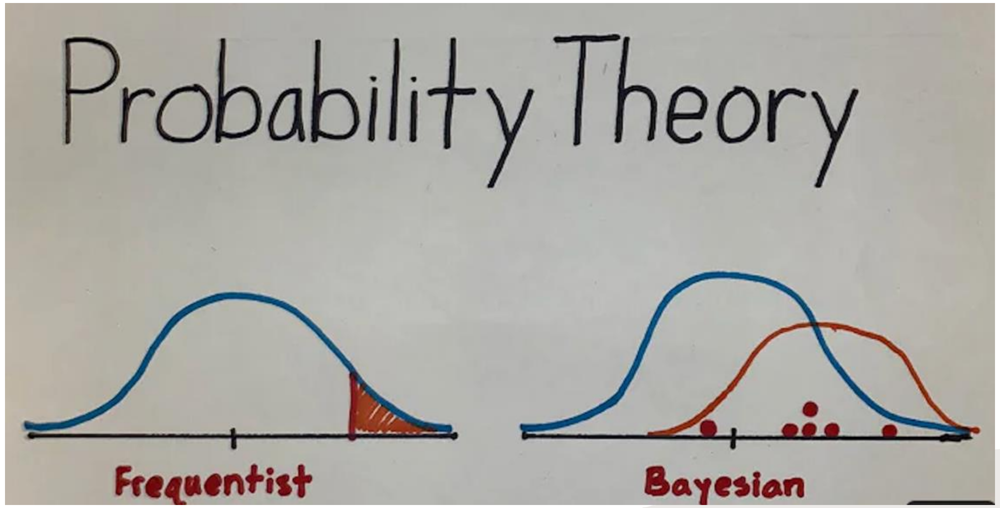
Chapter 7: Probability and Statistical Inference

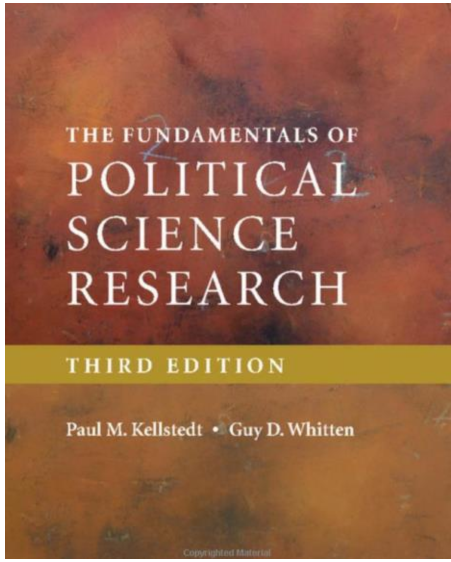
(Variable: Point Estimate and the Uncertainty)

Before we move on…
Bayesian Inference Recap

- Q: What do these equations and graphs tell us about Bayesian inference?
  - Q: how do you estimate a parameter?
  - Q: how do you build a credible interval?
- Q: The difference between Bayesian and Frequentist Inference?

$$\underbrace{\xi(\theta|x)}_{\text{posterior dist.}} \propto \underbrace{f(x|\theta)}_{\text{data/Likelihood}} \underbrace{\xi(\theta)}_{\text{prior dist.}}$$

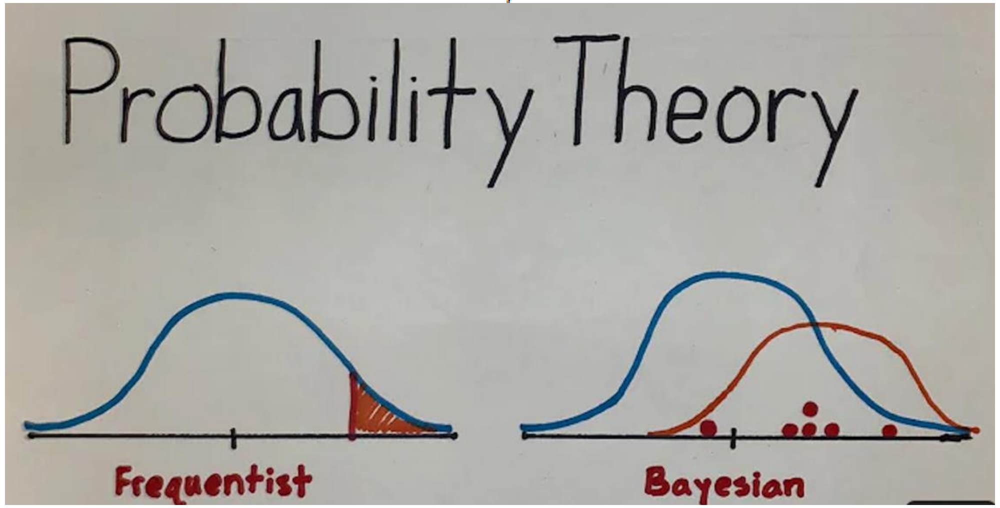




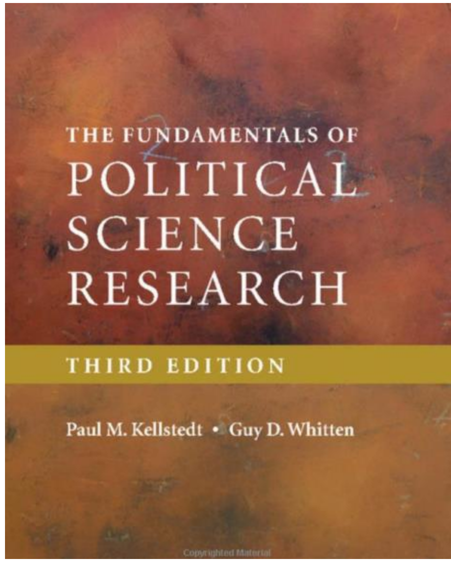
Before we move on…
Bayesian Inference Recap

- Q: What do these equations and graphs tell us about Bayesian inference?
  - Q: how do you estimate a parameter? MCMC (interaction between data and prior)
  - Q: how do you build a credible interval? MCMC
- Q: The difference between Bayesian and Frequentist Inference?

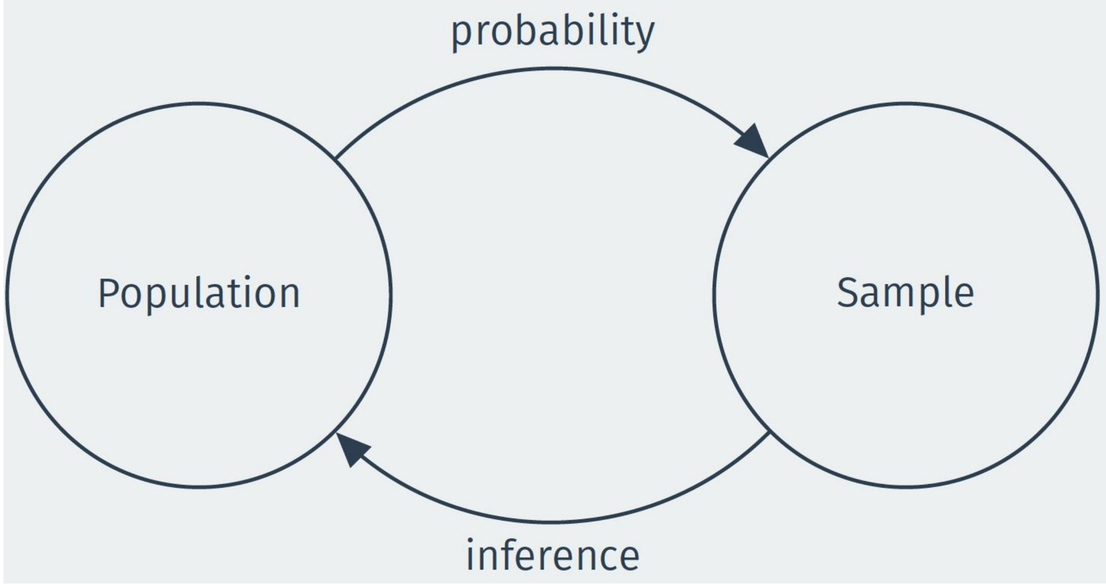
$$\underbrace{\xi(\theta|x)}_{\text{posterior dist.}} \propto \underbrace{f(x|\theta)}_{\text{data/Likelihood}} \underbrace{\xi(\theta)}_{\text{prior dist.}}$$



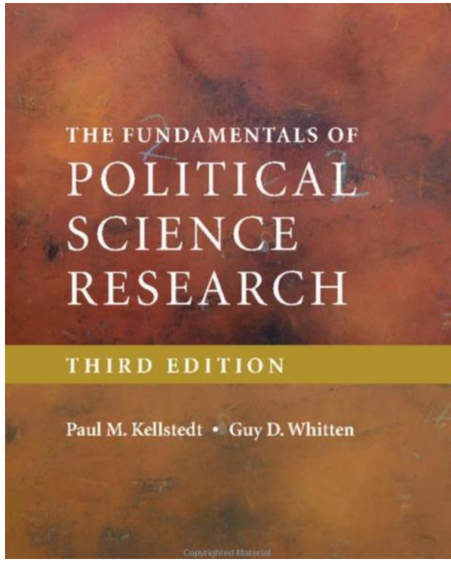
Populations versus samples



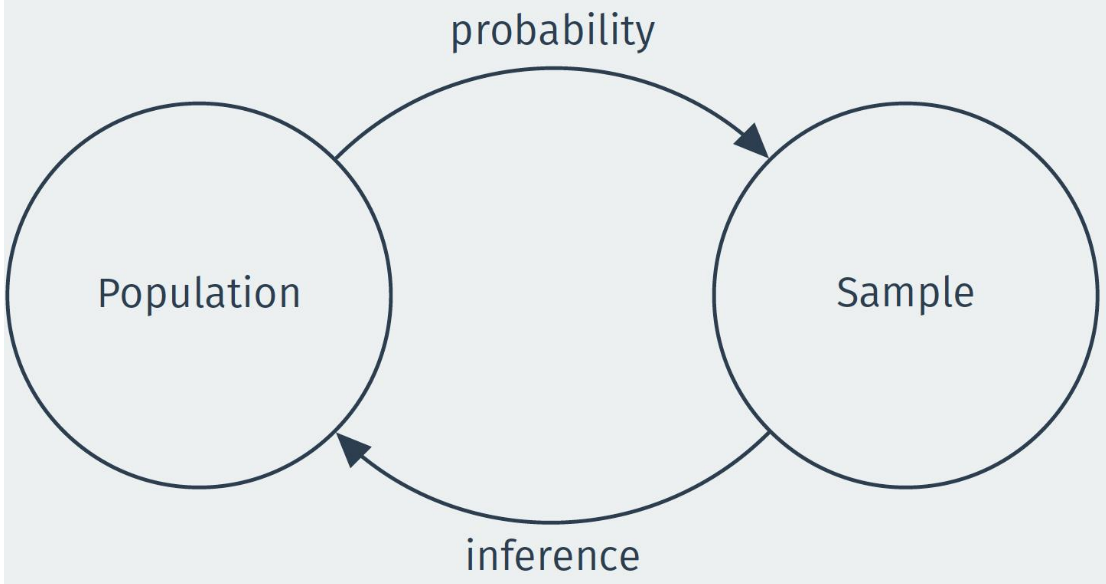
- We've learned how to describe your data:



Populations versus samples



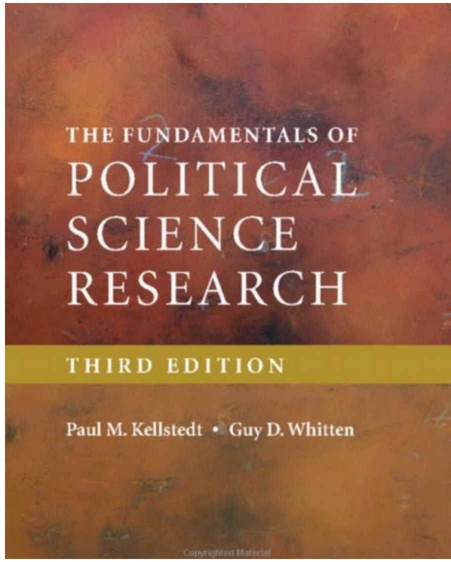
- We’ve learned how to describe your data:
  - Central tendency: mean
  - Dispersion: standard deviation
- What kind of data you have?
  - Population: data for every possible relevant case (e.g., civil war cases)
    - War is relatively rare events: we may have the population to study
  - Sample: a subset of population
    - Public opinion polls



Statistical inference (recap)

- **Sample → population**
- **Statistical inference** is the process of making probabilistic statements about a *population* characteristic based on our knowledge of *the sample characteristic*.
- In other words, there are things we know about with *certainty*—like the mean of some variable in our sample. But we care about the likely values of that variable in the entire population.
- Since we sometimes don't have data for an entire population, we need to use what we know to "infer" the likely range of values in the population.
- **Central Limit Theorem? (Core in the frequentist inference)**

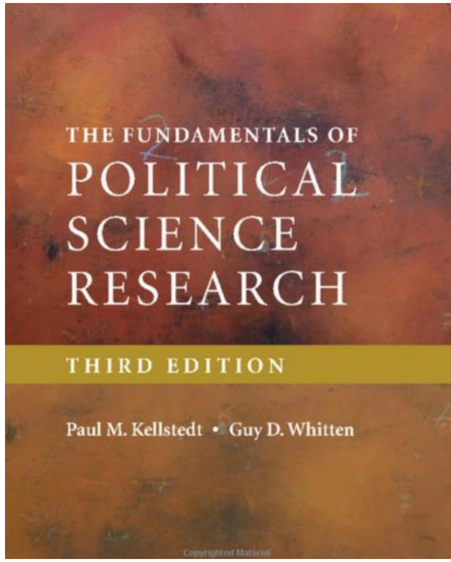
Central Limit Theorem



- The central limit theorem states that **if you have a population with mean μ and standard deviation σ and take sufficiently <u>large</u> <u>random samples</u> from the population with replacement, then the distribution of the sample means will be approximately normally distributed.**
- Normally distributed sample mean:

$$\text{Population mean } \mu = \bar{X}$$

$$\text{Population dispersion (standard error) } \sigma = \frac{sd}{\sqrt{n}}$$



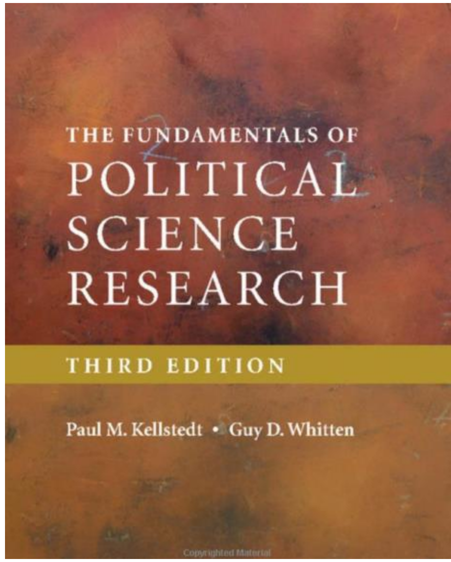
Central Limit Theorem

- The central limit theorem states that **if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.**
- Normally distributed sample mean:

$$\text{Population mean } \mu = \bar{X}$$

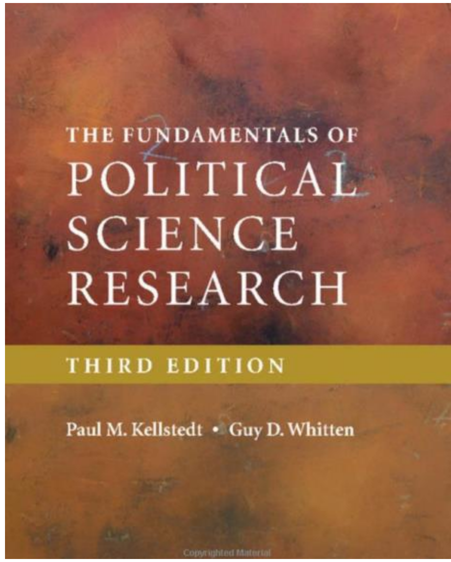
$$\text{Population dispersion (standard error) } \sigma = \frac{sd}{\sqrt{n}}$$

The normal distribution



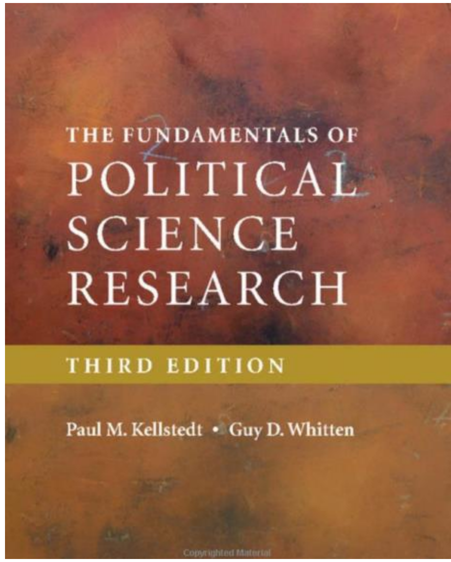
- The Central Limit Theorem will invoke a particular kind of distribution called the normal distribution, with which most of you are casually familiar. It's also called a bell-shaped distribution. But it has some unique features.
  - The distribution is **symmetrical**, so that the mode, median, and mean are all equal.

The normal distribution



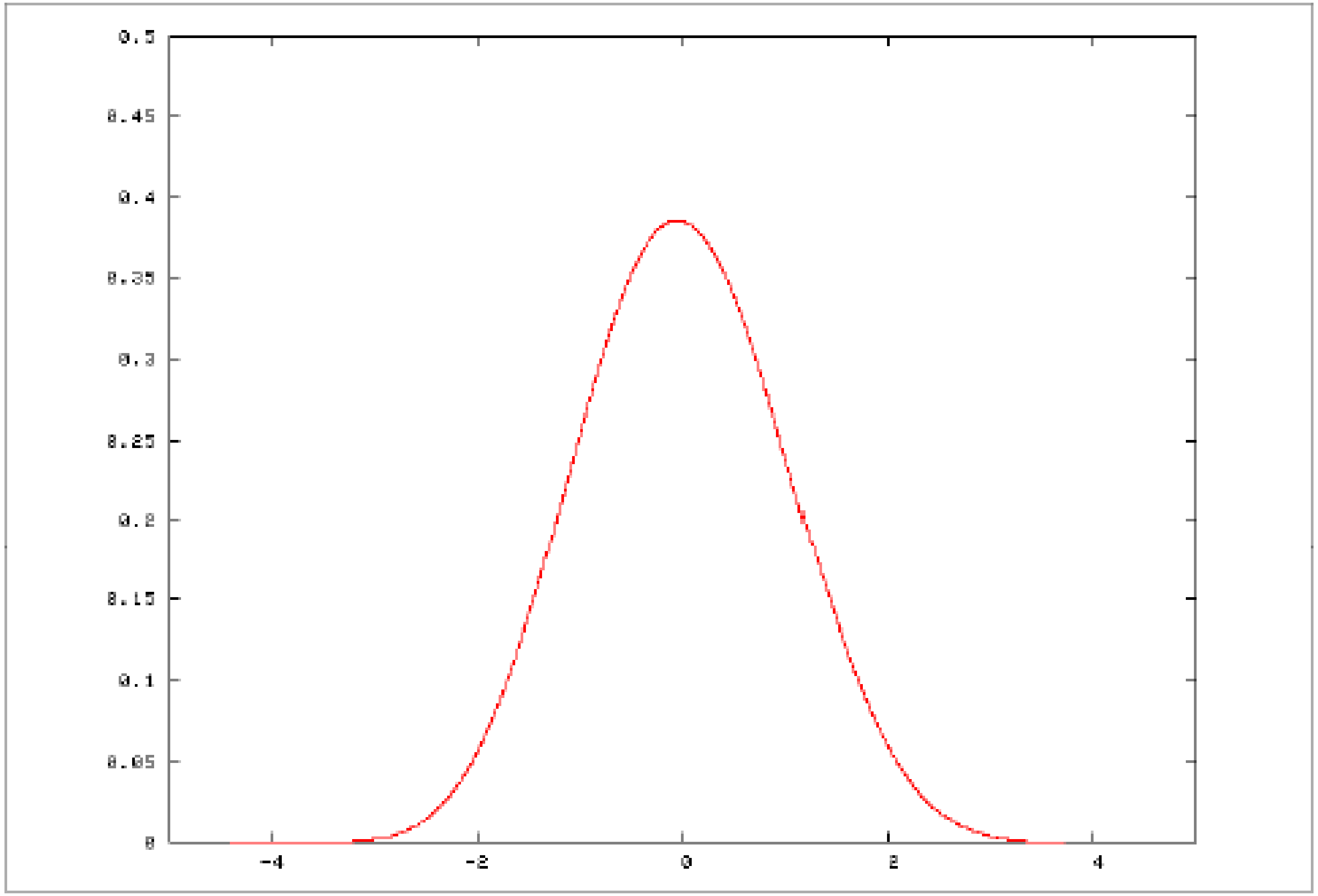
- The Central Limit Theorem will invoke a particular kind of distribution called the normal distribution, with which most of you are casually familiar. It's also called a bell-shaped distribution. But it has some unique features.
  - The distribution is **symmetrical**, so that the mode, median, and mean are all equal.
  - A particularly useful property: if a distribution is normally shaped, **we know a certain % of cases fall within a certain distance of the mean**.

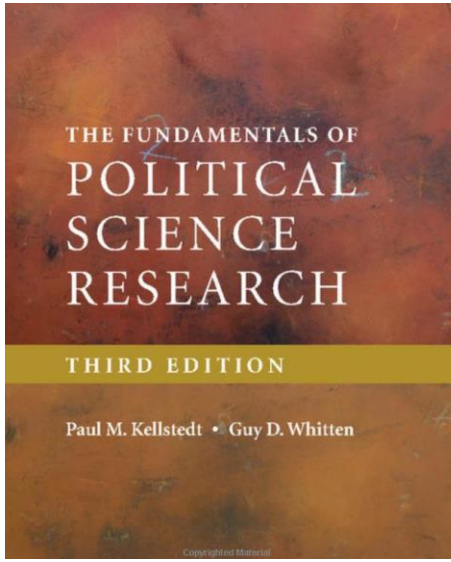
The normal distribution



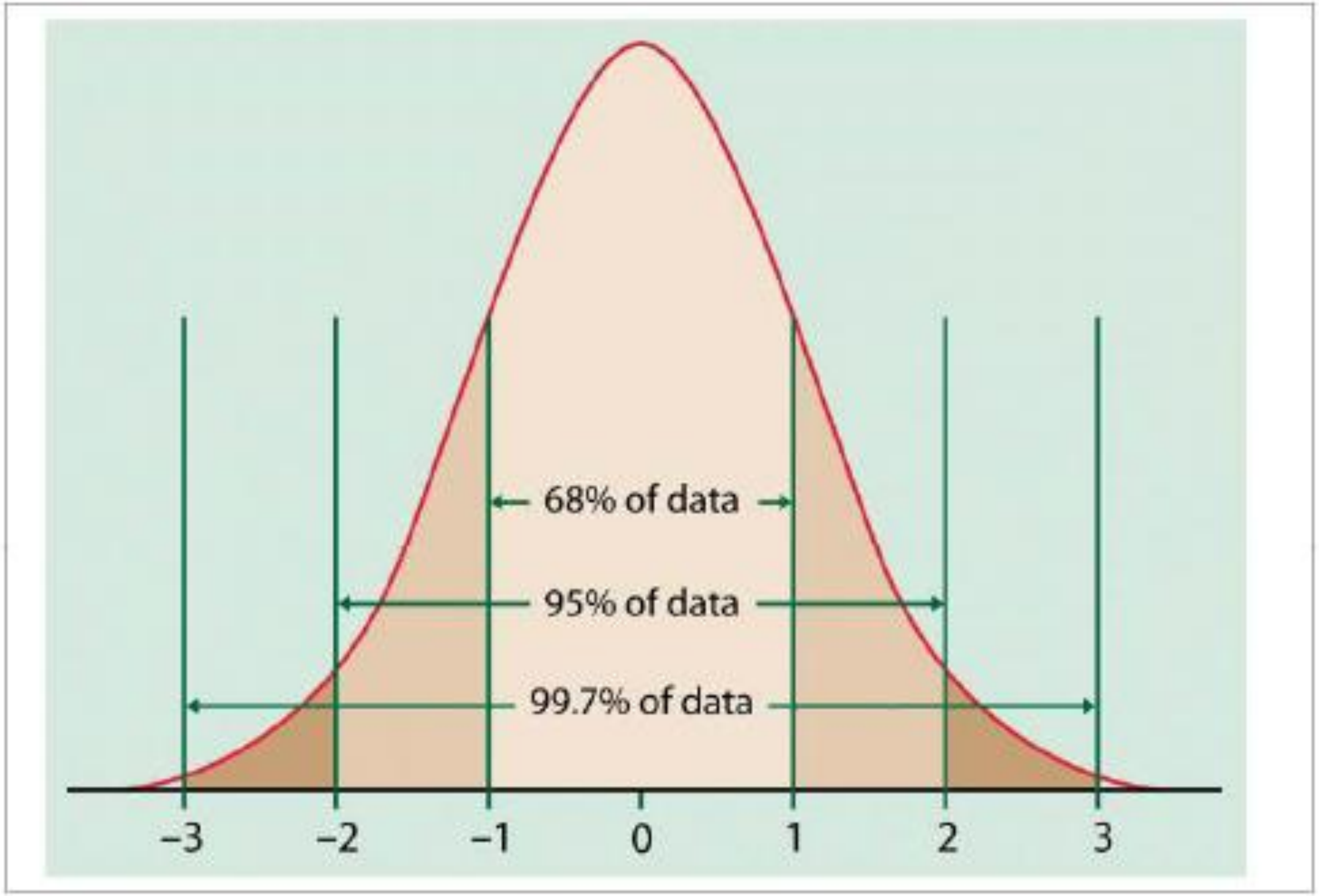
- The Central Limit Theorem will invoke a particular kind of distribution called the normal distribution, with which most of you are casually familiar. It's also called a bell-shaped distribution. But it has some unique features.
  - The distribution is **symmetrical**, so that the mode, median, and mean are all equal.
  - A particularly useful property: if a distribution is normally shaped, **we know a certain % of cases fall within a certain distance of the mean**
  - **→ helps build the concept of confidence intervals!**

It looks like this



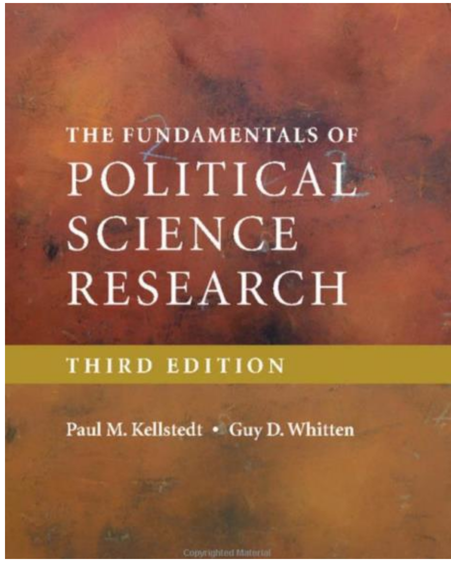


The 68 - 95 - 99 rule

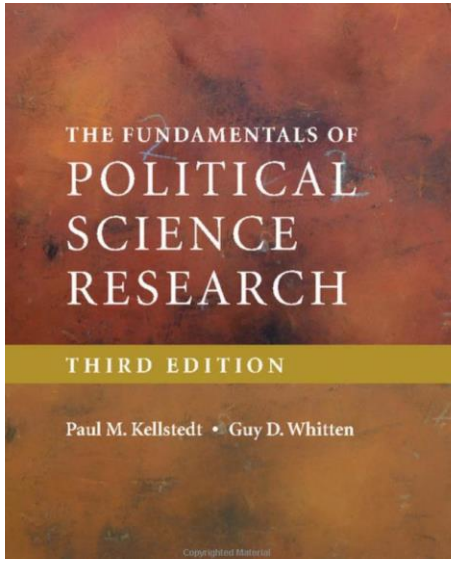


Density

Standard Deviation Away from Mean

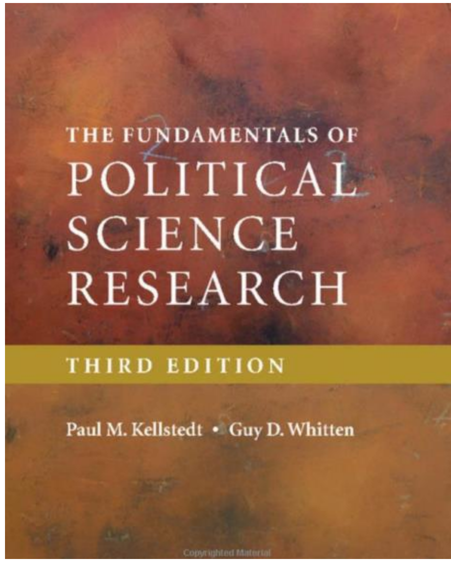


Normal Distribution Property

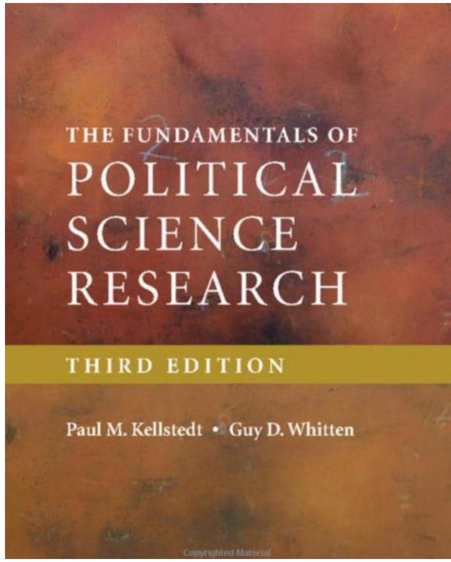


- The normal distribution is **symmetric, bell shaped**, and characterized by its **mean μ** and **standard deviation σ**.
- The probability range
  - 1 standard deviation → 68 % of data
  - 2 standard deviations (1.96 SD to be precise)→ 95 % of data
  - 3 standard deviations → 99% of the data

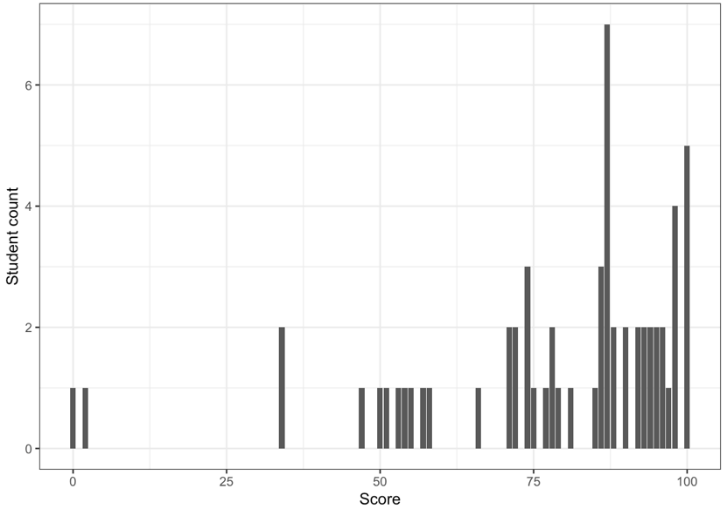
Are all distributions normal?



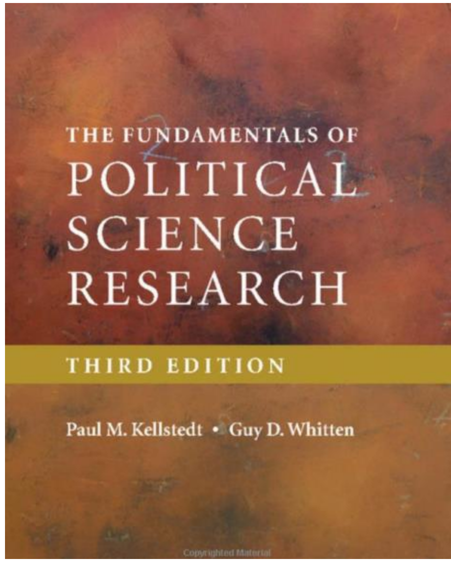
Are all distributions normal?



- NO!
- A frequency distribution is just a distribution of scores
  - like your scores on the midterm, or the distribution of income in Nebraska
- Most frequency distributions are not normally shaped.



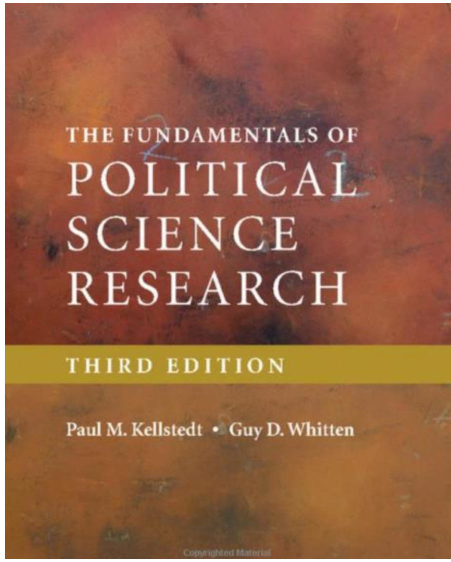
But…

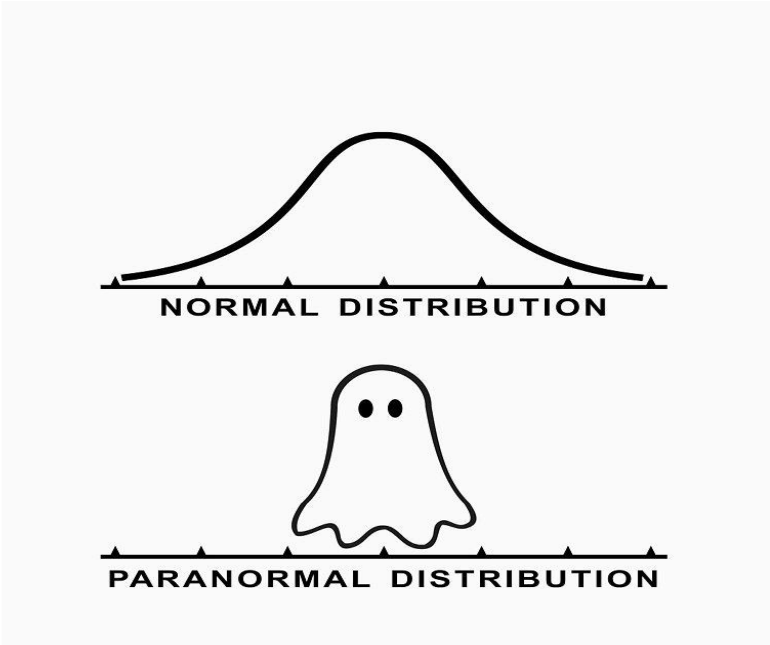


- Even if a frequency distribution is not normally shaped, if we imagine a (hypothetical) world in which we took an infinite *(N)* number of samples, and took **the mean** of each sample, and then plotted those means, then how would those plotted means be distributed?

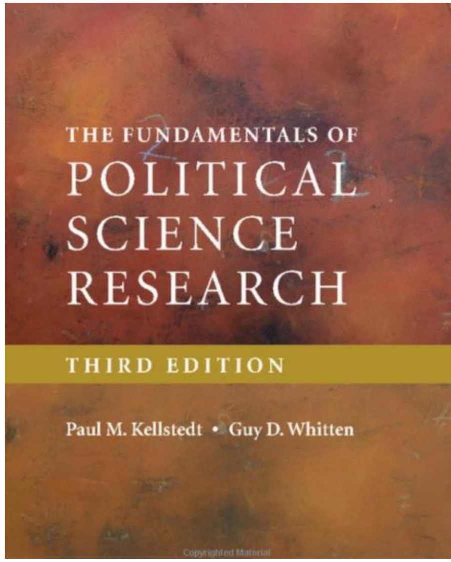
But…

- Even if a frequency distribution is not normally shaped, if we imagine a (hypothetical) world in which we took an <u>infinite *(N)* number of samples</u>, and took **the mean** of each sample, and then plotted those means, then how would those plotted means be distributed?



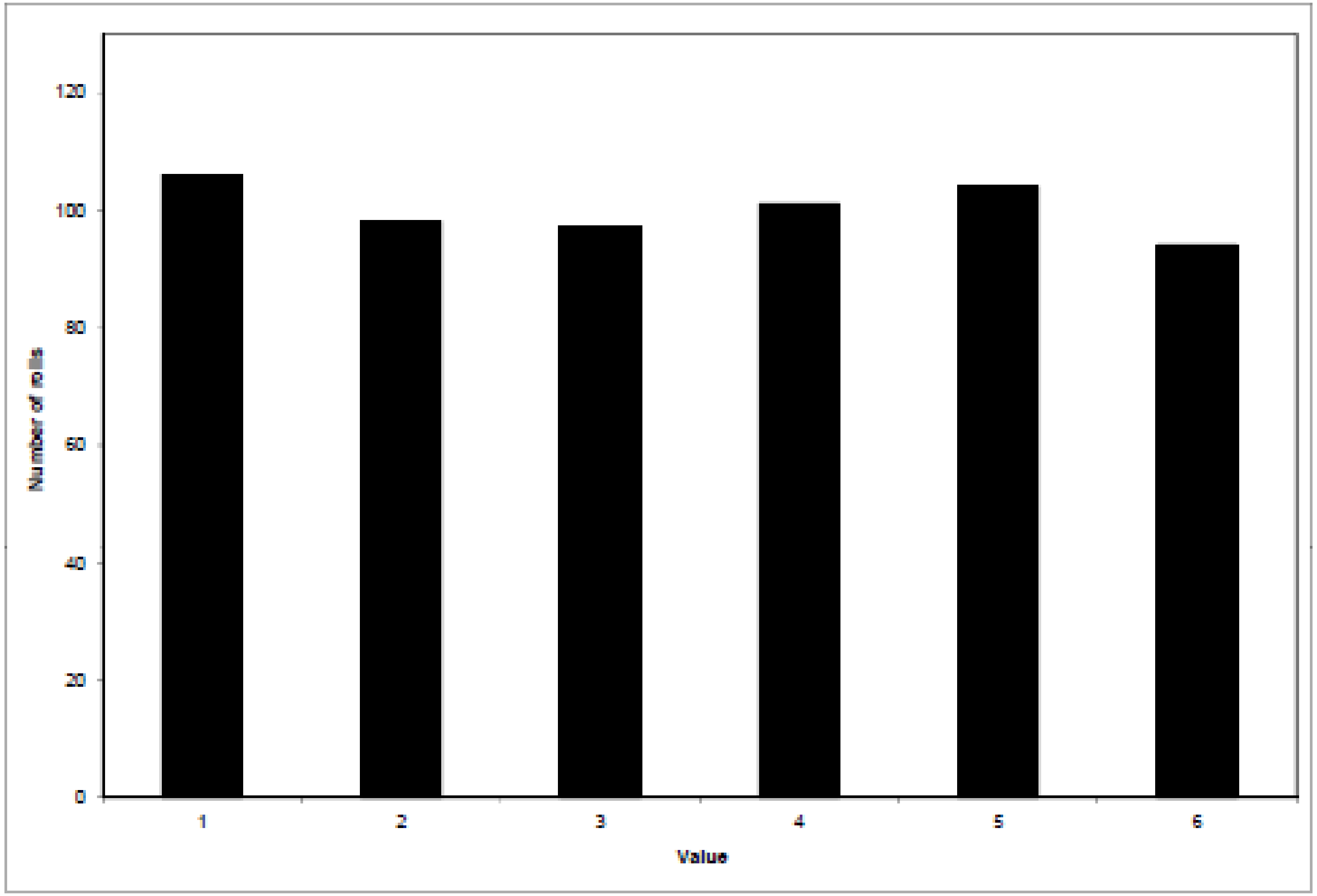


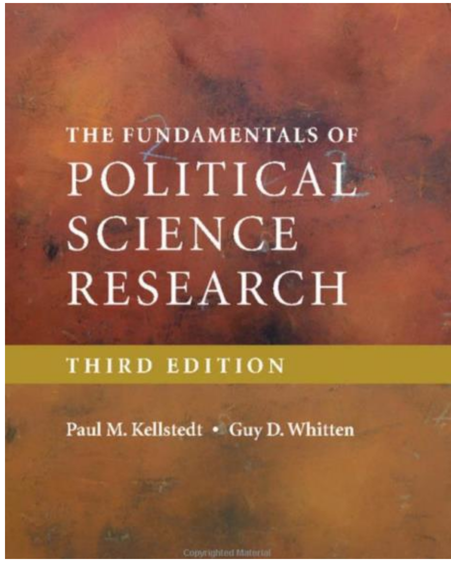
An example



- Imagine that we rolled a fair six-sided dice. It can come out as a 1, 2, 3, 4, 5 or 6 with equal probability, right?
- Let's say you rolled that dice 600 times (one sample/game). What would that distribution “look like"?

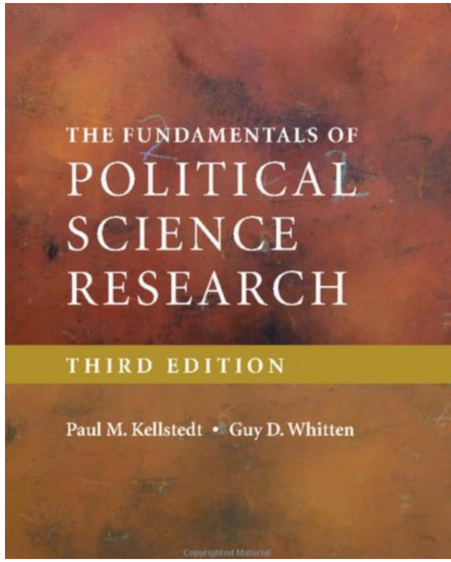
A uniform (not normal) distribution



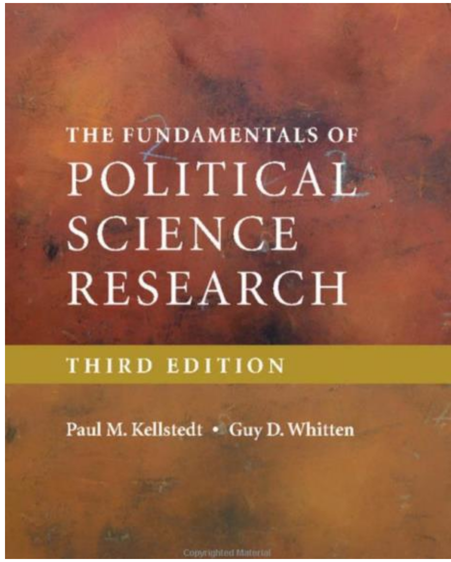


But that was not normal

- That's not normal, right?



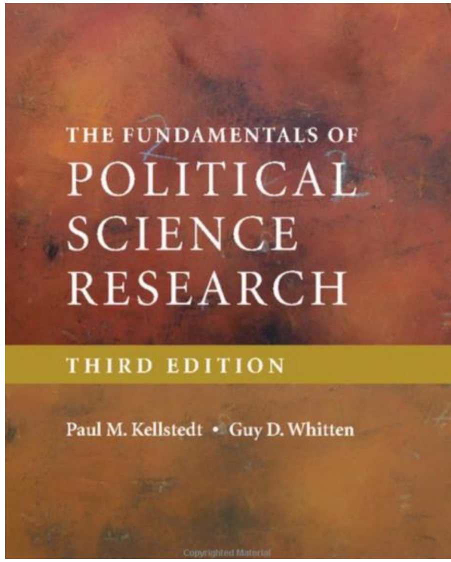
But that was not normal

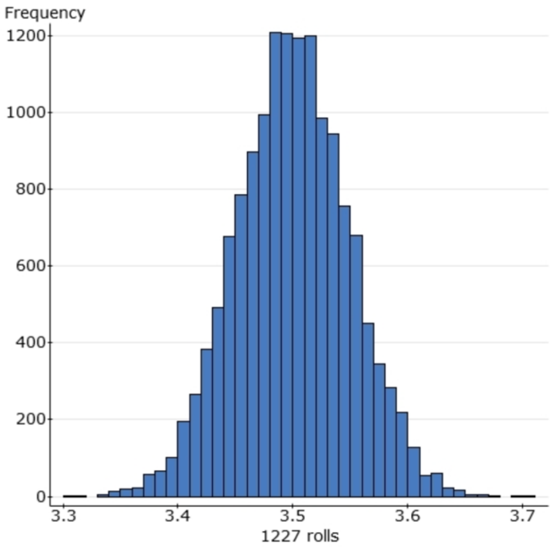


- That's not normal, right?
- Let's say we rolled that dice 600 times. What do you think the mean would be (about)?
  - Would it be exactly 3.5? Every time? No, of course not.
- But what would happen if we did that roll-it-600-times thing, say, a billion times *(N)*, then plotted the means?

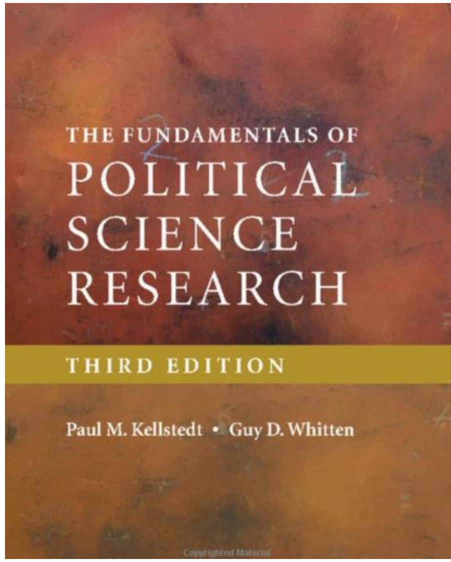
But that was not normal

- That's not normal, right?
- Let's say we rolled that dice 600 times. What do you think the mean would be (about)?
  - Would it be exactly 3.5? Every time? No, of course not.
- But what would happen if we did that roll-it-600-times thing, say, a billion times *(N)*, then plotted the means?





This is the Central Limit Theorem



- The **Central Limit Theorem** says that, no matter what the underlying shape of the frequency distribution (whether it's uniform, normal, or whatever), the hypothetical distribution of sample means--which is called a **sampling distribution**--will be normal, **with mean equal to the true population mean**, and **standard deviation** equal to

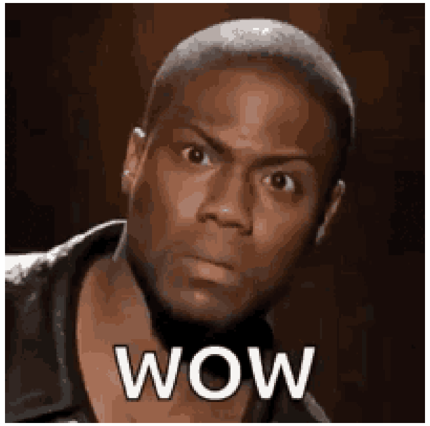
$$\sigma_{\bar{Y}} = \frac{s_Y}{\sqrt{n}}$$

$\sigma_{\bar{Y}}$ Standard error

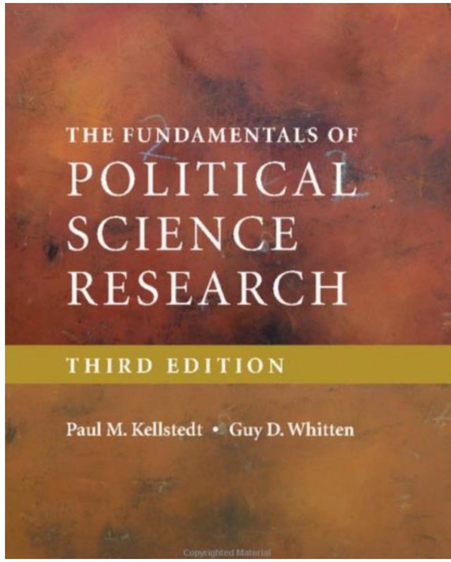
s_Y Sample standard deviation

$\sqrt{n}$ Sqrt of the sample size

The above is called the **standard error of the mean, or standard error**

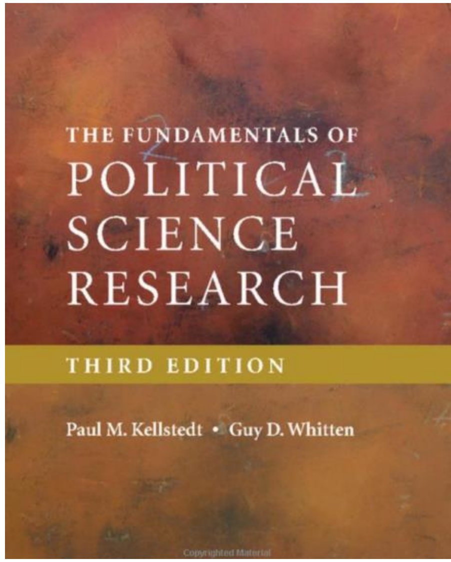


Normal Distribution and Confidence level



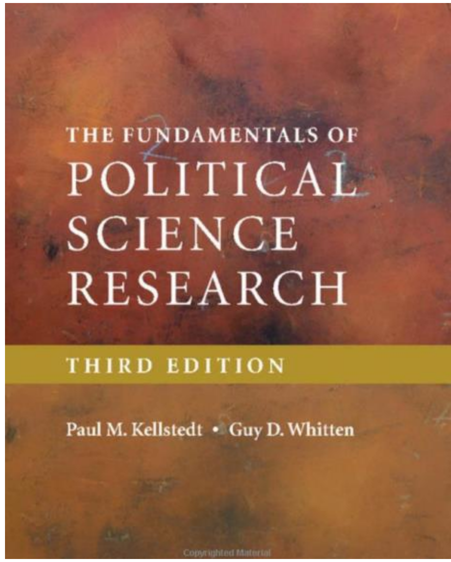
- So in the situation where we only have the sample mean, what we know is that **if we do this random sampling exercise *N* times, our sampling distribution will have a normal distribution**
- Key! With a normal distribution, we can invoke **the 68-95-99 rule** to create a *confidence level about the likely location of the population mean (true mean)*
  - For example, we have a sample mean $\mu = 3.47$ and $\sigma = 0.07$
  - By extending the coverage to 2 standard errors from the mean, $3.47 \pm (2 \times 0.07) = [3.33, 3.61]$
  - We can say that we are 95% confident that the true mean will be located in [3.33, 3.61]; 99% confident that the true mean will be in [3.26, 3.68]

Example: A polling result from September 2017



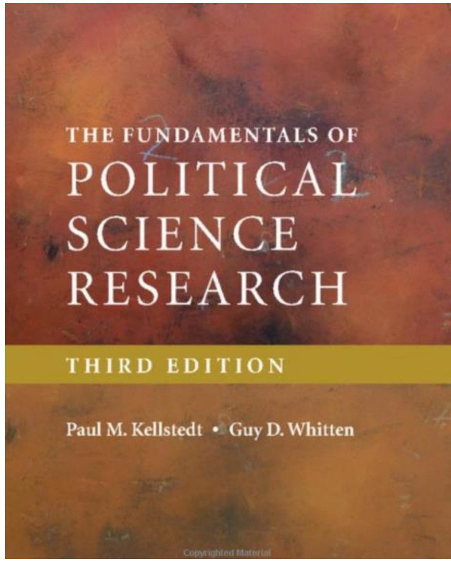
- Between September 14 and 18, 2017, NBC News and the *Wall Street Journal* sponsored a survey in which ***900 randomly selected*** adult American citizens were interviewed about their political beliefs. Among the questions they were asked was the following item intended to tap into a respondent's evaluation of the president's job performance:
- Q: In general, do you approve or disapprove of the job Donald Trump is doing as president?

The results



- In mid September, 2017, 43% of the sample approved of Trump's job performance, 52% disapproved, and 5% were unsure.
- We're only interested in the opinions of those 900 Americans who happened to be in the *sample insofar as they tell us something about the adult population as a whole*. But we can use these 900 responses to do precisely that, using the logic of the central limit theorem.

What we know with certainty about "*the sample*" *(900 Americans, 1 draw)*



Sample mean: We calculate our sample mean, as follows:

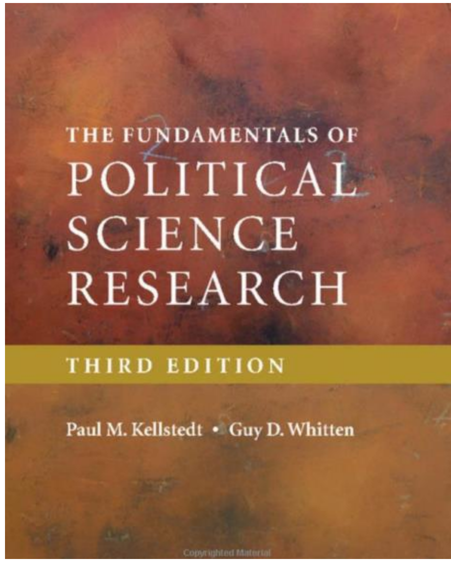
$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{\sum (387 \times 1) + (513 \times 0)}{900} = 0.43.$$

Sample SD: We calculate the sample standard deviation, in the following way:

$$s_Y = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}} = \sqrt{\frac{387(1-0.43)^2 + 513(0-0.43)^2}{900-1}}$$

$$= \sqrt{\frac{212.27}{899}} = 0.49.$$

What about the population as a whole?

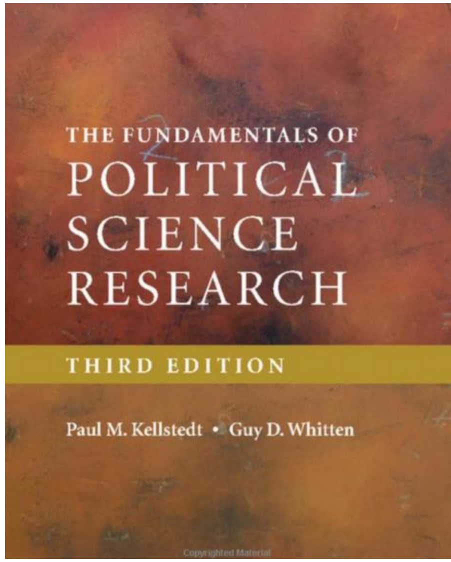


- Obviously, unlike the sample mean, the population mean cannot be known with certainty. But if we imagine that, instead of one sample of 900 respondents, we had an infinite number of samples of 900, then the central limit theorem tells us that those sample means would be distributed normally.
- Point estimate: Our best guess of the population mean, of course, is **0.43**, because it is our sample mean.
- Uncertainty: The standard error of the mean (SEM) is (based on the CLT):

$$\sigma_{\bar{Y}} = \frac{0.49}{\sqrt{900}} = 0.016,$$

which is our measure of "uncertainty about the population mean"

Creating a confidence interval

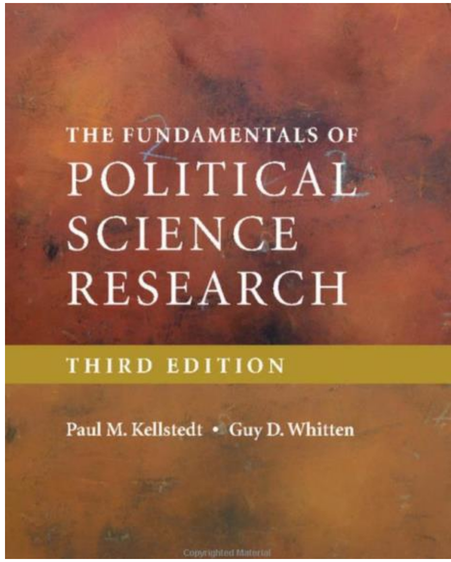


- If we use the rule of thumb and calculate the 95% confidence interval using two standard errors in either direction from the sample mean, we are left with the following interval:

$$\bar{Y} \pm 2 \times \sigma_{\bar{Y}} = 0.43 \pm (2 \times 0.016) = 0.43 \pm 0.032$$

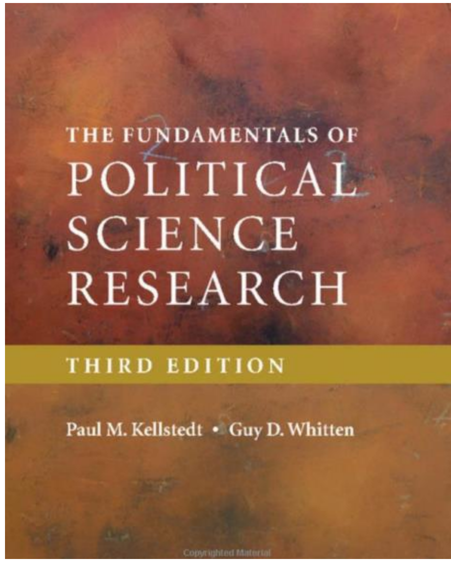
or between [0.398, 0.462], which translates into being 95% confident that the population value of Trump approval during September 14-18, 2017 was between 39.8% and 46.2%.

Those plus-or-minus figures



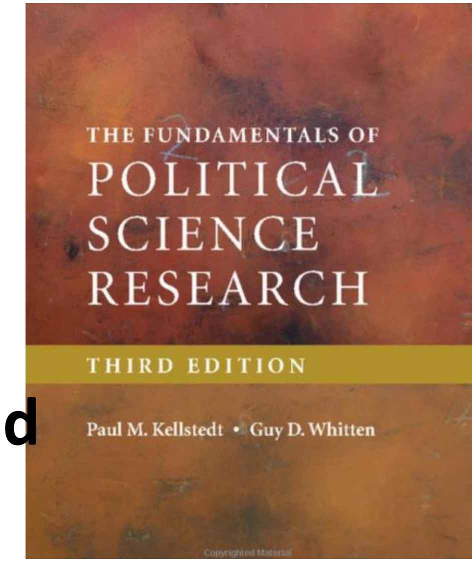
- This is where the “plus-or-minus" figures that we always see in public opinion polls come from.
- ***The best guess for the population mean value is the sample mean value, plus or minus 2 standard errors.***
- So the plus-or-minus figures we are accustomed to seeing are built, typically, on the 95% interval.

“Random” samples vs. samples of convenience



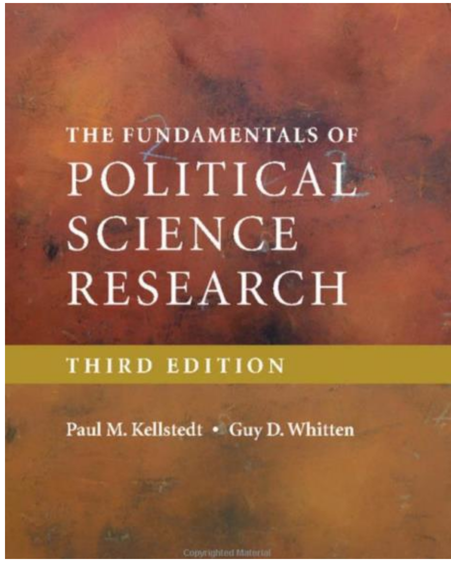
- The central limit theorem ***only*** applies to samples that are **selected randomly**. With a sample of convenience, by contrast, we cannot invoke the central limit theorem to construct a sampling distribution and create a confidence interval.
- What happens if we have a sample of convenience?

"Random" samples vs. samples of convenience

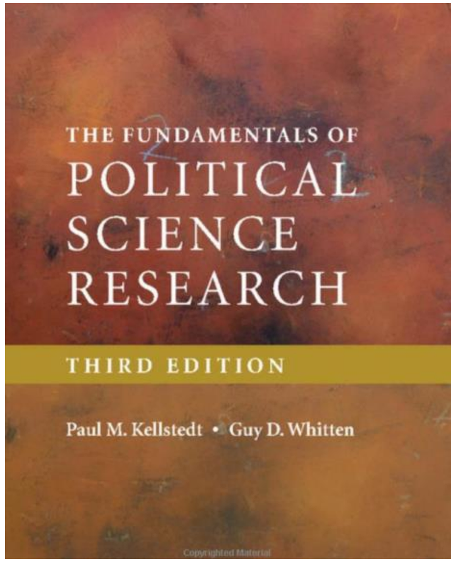


- The central limit theorem ***only*** applies to samples that are **selected randomly**. With a sample of convenience, by contrast, <u>we cannot invoke the central limit theorem</u> to construct a sampling distribution and create a confidence interval.
- What happens if we have a sample of convenience?
- A non-randomly selected **sample of convenience** does very little to help us build bridges between the sample and the population about which we want to learn. What do such "surveys" say about the population as a whole? <u>*Because their samples are clearly not random samples of the underlying population, the answer is "nothing."*</u>

How much does sample "size" matter?

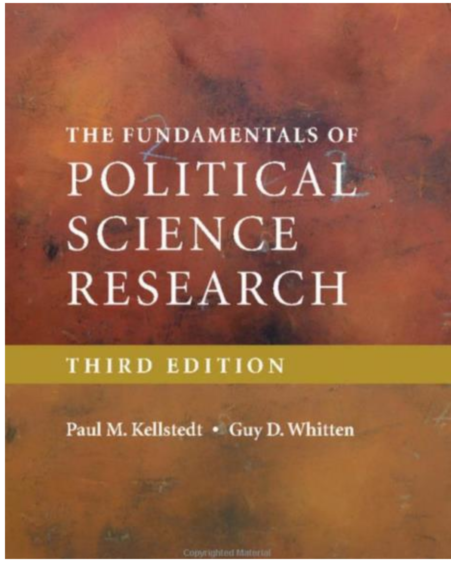


How much does sample "size" matter?



- As the formula for the confidence interval indicates, the smaller the standard errors, the “tighter" our resulting confidence intervals will be, and larger standard errors will produce “wider" confidence intervals.
- If we are interested in estimating population values, based on our samples, with as much precision as possible, then it is desirable to have tighter instead of wider confidence intervals.

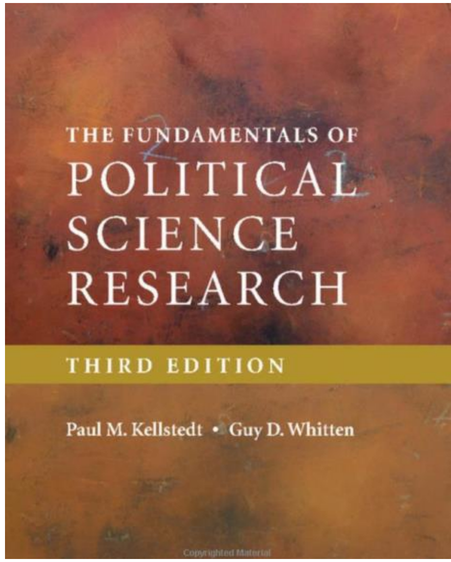
Some comparisons



- Instead of having our sample of 900, suppose we had 2,500 people. Then our standard errors would have been:

Consider the opposite case. If the sample were 400, then:

Some comparisons



- Instead of having our sample of 900, suppose we had 2,500 people. Then our standard errors would have been:

$$\sigma_{\bar{Y}} = \frac{0.49}{\sqrt{2500}} = 0.010$$

Consider the opposite case. If the sample were 400, then:

$$\sigma_{\bar{Y}} = \frac{0.49}{\sqrt{400}} = 0.024$$

which, when doubled to get our 95% confidence interval, would leave a plus-or-minus 0.048 (or nearly 4.8%) in each direction.