Quantitative Methods in Political Science: Sampling and Statistical Inference

Thomas Gschwend | Domantas Undzėnas | Muhammad Muhammad

Week 3 - 17 September 2025

Roadmap

- Understand and model stochastic processes
- Understand statistical inference
- Implement it mathematically and learn how to estimate it
  - OLS
  - Maximum Likelihood
- Implement it using software
  - R
  - Basic programming skills

Overview: Week 3

Statistical inference

    Basics

    Sampling

    Example

Confidence Intervals

    Construction of CIs: Normal approximation

    Construction of CIs: Bootstrapping
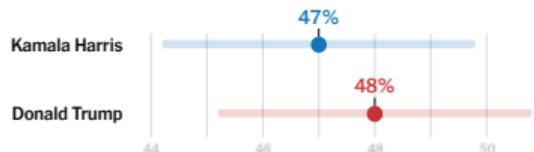
    Construction of CIs: Simulation

Testing Hypotheses

Statistical inference

If the 2024 presidential election were held today, who would you vote for if the candidates were **Kamala Harris** and **Donald Trump**?

Among likely voters. Shaded areas represent the margin of error.

The margin of error is plus or minus 2.8 points. • Based on a New York Times/Siena College poll of 1,695 voters nationwide conducted from Sept. 3 to 6. • By Lily Boyce and Andrew Park

"A national poll of likely voters by The New York Times and Siena College found Mr. Trump leading Ms. Harris, 48 percent to 47 percent, within the poll's three-percentage-point margin of [sampling] error …" *New York Times*, Sept 17, 2024.

The Goal of Statistical Inference

- Learn a quantity of interest about a particular group (population parameters).
  - Income of working age population in a country.
  - Reading ability of 4th graders in France.
  - Support for environmental policy in Germany.
- Often information on all members of the group (population) will not be available.
- We use sampling to collect a limited amount of information and use it to infer population properties (parameters).
- Since we have only information on a subset of the population we are uncertain about our inference (but there are other sources of uncertainty as well even if we observe the entire population).
- All inferences are inherently uncertain!

The goal of statistical inference is to estimate population parameters and summarize our uncertainty about these estimates.

- A parameter describes a feature of the population. The parameter is fixed at some value, and we will never be able to know it for sure.
- What we observe is a random sample, drawn from the population. A random sample is a proper subset of the population for which it is true that each member has an equal probability to be selected.
- From this sample, we can calculate a sample statistic of a population parameter. A sample statistic is a function that is applied on the observed sample. This function is called the estimator of the population parameter.
  - We can calculate the mean of a random sample. The mean is then a *sample statistic*, and the function that maps observations of the random sample to this sample statistic is the *estimator*.
- The population parameter is the estimand. The result of such a calculation is called an estimate.
  - Note: The estimator is a function, the estimate is a number!

The Principle of Sampling

- Probability sampling: Select from a population with size N a number of individuals, n (usually $n \ll N$), such that each individual has a non-zero probability of being chosen.
- Sources of variation across samples:
  - Sampling variability: Means and standard deviations of repeated samples will not be identical.
  - Sampling error: An estimate from a sample will not be identical to the value in the population.
- The sample size is positively related to the desired precision of the estimate.

Sampling model:

- We have: (1) a population, (2) a sample from this population, and (3) an estimate of a population parameter.
- How uncertain are we about that estimate?
- Alternatively, how precisely can we estimate the population parameter?
- In a different sample, our estimate would be slightly different. Hence, estimates vary over repeated samples.
- Applying an estimator on repeated samples yields a sampling distribution for this statistic.
- Calculating the spread of this sampling distribution yields a measure of uncertainty.

Example of Statistical Inference

- Let there be a country with 100,000 inhabitants.
- We want to know what the mean income of this country is.
- We sample 5000 individuals randomly from the population.
- The mean of obtained sample is 1400 ($= \hat{\theta}$) with a standard deviation of 2000 ($= \hat{\sigma}$).
- The standard error of the estimate of the population mean ($\hat{\theta}$) is: $\sigma_{pop}/\sqrt{n}$.
- For a large sample, the standard deviation ($\hat{\sigma}$) of a sample can be used as an approximation of the population standard deviation (σ_{pop}):

$$SE(\hat{\theta}) = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{\widehat{2000}}{\sqrt{5000}} = 28.3$$

- The mean income of our population is estimated to be 1400 ± 28.3

Digression: Derivation of Standard Error

- Let's assume that we have a random sample from a population and n random variables, $\theta_1, ..., \theta_n$ that come from the same population represented by the same distribution with mean μ and variance σ^2
- Such random variables are called _independently_ and _identically_ _distributed_ (_iid_)
- Hence, we know that $\text{Var}(\theta_i) = \sigma^2$ for all such random variables.
- Denote their mean as an estimate of the population mean by $\bar{\theta}$.
- Then, we get the sampling variance (i.e., variance of an estimator)

$$\text{Var}(\bar{\theta}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\theta_i\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n}\theta_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left(\theta_i\right)$$

$$= \frac{1}{n^2}n\sigma^2 = \frac{1}{n}\sigma^2$$

- The sampling variance equals the population variance divided by the sample size.
- Hence, the standard error of the mean is derived as $\text{SE}(\bar{\theta}) = \dfrac{\sigma}{\sqrt{n}}$

Confidence Intervals

Confidence Intervals

- Our estimate of a population feature (parameter) varies across repeated samples, thus generating a *sampling distribution.*

- Instead of a point estimate, we should better get an interval estimate – a range within which the true parameter lies with some level of certainty.

- We can construct confidence intervals using the standard error or the variance of our estimates.

- We call a confidence interval a q% confidence interval if it is constructed such that it contains the true parameter at least q% of the time *if we repeat the experiment a large number of times.*

- Check out this cool visualization.

- Note that this does not mean that there is a 95% probability for the population parameter to lie inside the interval!

1. Analytical
2. Bootstrapping (resampling)
3. Simulation (parametric)

1. Analytical

- We have a sample statistic $\hat{\theta}$ estimated for a parameter θ.
- If the sample is large enough, we can assume a normal sampling distribution with mean $\hat{\theta}$ and variance $Var(\hat{\theta})$
- We can then construct a 95% confidence interval using the quantiles from the standard normal distribution:

$$\hat{\theta} \pm 1.96\sqrt{Var(\hat{\theta})} = \hat{\theta} \pm 1.96 \cdot \frac{\hat{\sigma}}{\sqrt{N}}$$
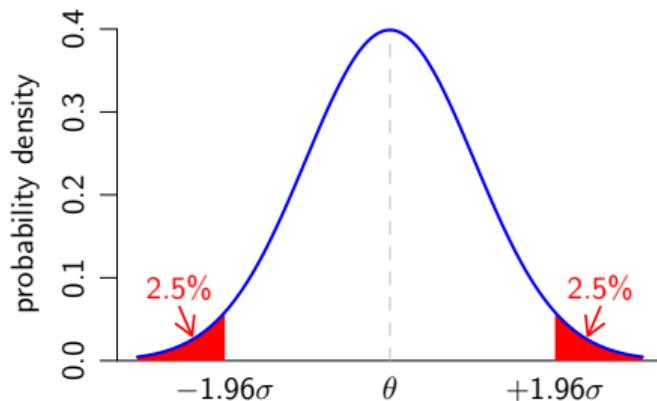
Example: Measurement With Normal Error

- Support for a new policy during the last month as measured by a polling firm ($N = 50$), as

  15 16 12 17 14 13 15 16 12 14 17 15 12 15 14 16 16 14 13 12 13 15 16 14 15

  11 13 13 16 15 17 14 12 15 14 13 16 17 14 15 16 14 13 14 13 15 17 11 14 15

  with an average value of $\hat{\theta} = 14.36$ and a standard deviation of $\hat{\sigma} = 1.61$.

- The standard error of the estimate $\hat{\theta}$ of the population mean θ is

$$\text{SE}(\hat{\theta}) = \frac{\hat{\sigma}}{\sqrt{N}} = \frac{1.61}{\sqrt{50}} = 0.228$$

- The 95% confidence interval ranges from

$$14.36 \pm 1.96 \times 0.228 = [13.91, 14.81]$$

-

2. Bootstrapping (resampling)
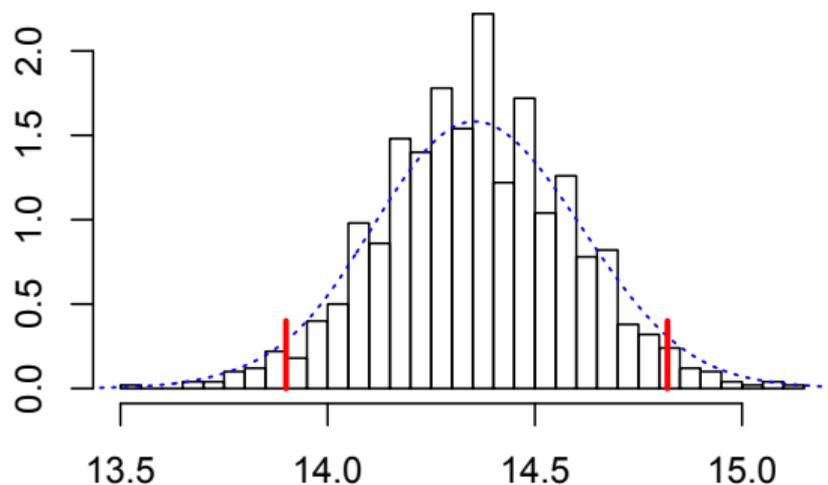
Bootstrapping Approach: Constructing CIs via Resampling

- What if the population does not appear normal and/or we have only a small sample?
- We do *not know* what the sampling distribution looks like, but we want an estimate of it.
- *Bootstrapping* estimates the sampling distribution of θ by *repeatedly sampling* (*with replacement*) from the original sample.
- In standard bootstrapping, the size of the bootstrap sample, n, is *identical* with the sample from the population. The variability comes from *sampling with replacement*.
  1. Take s samples of size n from your data.
  2. Calculate the quantity of interest ($\hat{\theta}_i$, e.g. the mean) for each of your s samples, which yields a vector of length s.
  3. A simple confidence interval for your quantity can be obtained by calculating quantiles (e.g., 2.5 and 97.5 percentiles for 95% CI) of this vector.

Just remember!
The population is to the sample as the sample is to the bootstrap sample.

- Bootstrapping $s = 1000$ samples of size $n = 50$ yields a mean of 14.36.
- A simple confidence interval can be obtained by just calculating the 2.5th and 97.5th quantile of the bootstrapped sampling distribution.
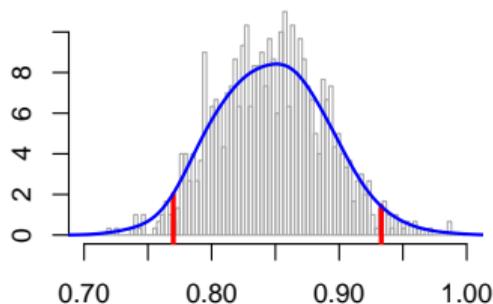- This yields a 95% confidence interval from 13.90 to 14.82.

3. Simulation (parametric)

- An alternative parametric approach sets up a population distribution and draws from it.
- This is an empirical / computational implementation of the idea of inference as repeated sampling.
- We obtain hypothetical repeated samples from an assumed population distribution.
- These simulations are done using a computer:
  1. Create a (normal) sampling distribution from the mean and standard error of your sample.
  2. Take s draws from that distribution $N(\hat{\theta}, \hat{\sigma}^2)$.
  3. Calculate your quantity of interest s times. Thus, we simulated its sampling distribution.
  4. Calculate summaries, such as means and standard errors, for the resulting vector of length s. To construct 95% CIs, you need to 2.5 and 97.5 percentiles.

Example: Ratio of Proportions

- Consider a sample of $n = 1000$ individuals, 500 of which are men and 500 are women.
- 55% of men vote left, while 65% of women vote left.
- *Quantity of interest*: The ratio for voting left for men compared to women is 0.846 ($\approx \frac{0.55}{0.65}$).
- Analytically, the SE for proportions is $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
- The standard error of these proportions for men is 0.0222 and for women is 0.0213.
- The confidence interval for this ratio obtained via simulation ranges from 0.765 to 0.937.

Mechanics:

- Take $s = 1000$ draws from normal distribution with $\hat{\mu}$ and $\hat{\sigma}^2$ like the data

$$p_m = \mathcal{N}(.55, .0222^2) = [0.56804, 0.53736, 0.53876, 0.55186, 0.55523, \dots]$$

$$p_w = \mathcal{N}(.65, .0213^2) = [0.63091, 0.63087, 0.61234, 0.64373, 0.68099, \dots]$$

- Calculate vector of 1000 ratios between men and women

$$r = p_m/p_w = [0.90036, 0.85178, 0.87984, 0.85728, 0.81533, \dots]$$

- The empirical confidence interval is the 2.5% and 97.5% quantile of the distribution of this vector of ratios (and ranges from 0.765 to 0.937).

Testing Hypotheses

- Usually, we want to learn about population parameters based on sample statistics. This requires us to use statistical inference.

- For example, is a population parameter different from zero? If you find a value different from zero in your sample, can you be sure that you did not select a sample where this is the case while in the population the value is zero?

- Formulate competing hypotheses:
  H_0: The population parameter is equal to zero (null hypothesis).
  H_A: The population parameter differs from zero (alternative hypothesis).

- Make a decision based on
  (a) Confidence intervals.
  (b) Test statistic.

(a) Decision based on Confidence Intervals

- Someone claims the mean income of a country is 1400. You take a random sample of size n=1000 and obtain a mean income of 1350, with a standard deviation of 750.
- What can we say about the "truth" of this claim?
- Construct two competing hypotheses:

  $H_0 : \mu = 1400$ and

  $H_A : \mu \neq 1400$
- Calculate the 95% confidence interval for the mean:

$$1350 \pm 1.96 \times \frac{\widehat{750}}{\sqrt{1000}} = [1303, 1396]$$

- The H_0 value of 1400 does not lie in this interval: the sample is very unlikely to come from a population with mean 1400.
- This allows us to reject H_0.
- Note that we can always only reject or fail to reject hypotheses, but we can never prove hypotheses.

(b) Decision based on a Test Statistic

- At a level of 95% confidence, the critical z values on a standard normal distribution are ± 1.96.
- How much is the sample mean (i.e., 1350) away from the hypothesized population mean (i.e., 1400) in repeated samples? Calculate the z-score of your realized sample:

$$z = \frac{\bar{x} - \mu_{pop}}{\sigma_{pop}/\sqrt{n}} = \frac{1350 - 1400}{\widehat{750}/\sqrt{1000}} = -2.1$$

- This is smaller than the critical value and you would reject H_0.