Asymptotics

Ye Wang
University of North Carolina at Chapel Hill

Mathematics and Statistics For Political Research
POLI783

Statistical inference

- For researchers, it is insufficient to know just an estimate.
- How confident are we that a new policy is effective and not harmful?
- After providing cash transfers to 1,000 families, 57 kids from these families perform worse at school.
- Is this a consequence of the policy or just randomness in the DGP?
- All estimates come with uncertainties.
- Estimates without measures of uncertainties surrounding them are not trustworthy.
- Ideally, we'd like to repeat the DGP (e.g., policy activation) for many times and record the estimate from each round.
- The distribution of estimates tells us whether an outcome is common or unlikely.
- This distribution is exactly the sampling distribution $F_{\hat{\tau}}(x)$.

Statistical inference

- Nevertheless, the sampling distribution is usually unknown to the researcher.
- Remember that we only have one estimate from one collected sample.
- It is unrealistic to repeat the DGP.
- Nevertheless, we can infer what the sampling distribution looks like with one dataset.
- This process is known as statistical inference.
- Inference will be more accurate when the sample size is larger.
- In practice, we often approximate the sampling distribution with the hypothetical one where N is infinite.
- This approach provides us with a convenient tool for statistical inference and is known as asymptotic analysis.

Variance estimation

- One measure we can use to gauge the uncertainty of our estimate is the estimator's variance.
- We do not know its value as it depends on the sampling distribution.
- Yet we can estimate it using an estimator, with the true variance being an estimand.
- For the sample average of an i.i.d. sample, $\bar{X}_N$, we know that its normalized variance is σ.
- An estimator for σ is the sample variance,

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(X_i - \bar{X}_N \right)^2 .$$

- We will show that the sample variance is unbiased for the true variance.

Sample variance

- First note that

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}_N\right)^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i^2 - 2X_i\bar{X}_N + \left(\bar{X}_N\right)^2\right)$$

$$= \frac{1}{N-1}\sum_{i=1}^{N} X_i^2 - \frac{N}{N-1}\left(\bar{X}_N\right)^2.$$

- Then, using the fact that $\mathbb{E}\left[X_i^2\right] = \mu^2 + \sigma^2$, we have

$$\mathbb{E}\left[s^2\right] = \frac{1}{N-1}\sum_{i=1}^{N}\mathbb{E}\left[X_i^2\right] - \frac{N}{N-1}\mathbb{E}\left[\left(\bar{X}_N\right)^2\right]$$

$$= \frac{N(\mu^2+\sigma^2)}{N-1} - \frac{1}{N(N-1)}\sum_{i=1}^{N}\mathbb{E}\left[X_i^2\right] - \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}\left[X_iX_j\right]$$

$$= \frac{N(\mu^2+\sigma^2)}{N-1} - \frac{\mu^2+\sigma^2}{N-1} - \frac{N(N-1)\mu^2}{N(N-1)} = \sigma^2.$$

Law of large numbers

- Is the sample variance a consistent estimator for the true variance?
- We first introduce the famous law of large numbers (LLN).
- If $(X_1, X_2, \dots, X_N)$ are i.i.d. with $\text{Var}[X_i] < \infty$, then

$$\bar{X}_N \xrightarrow{p} \mathbb{E}[X_i].$$

- This result follows directly from Chebyshev's inequality.
- Therefore, if $\text{Var}\left[X_i^2\right] \leq \mathbb{E}\left[X_i^4\right] < \infty$,

$$\frac{1}{N}\sum_{i=1}^{N} X_i^2 \xrightarrow{p} \mathbb{E}\left[X_i^2\right].$$

- Then, $\frac{1}{N-1}\sum_{i=1}^{N} X_i^2 = \frac{N}{N-1}\frac{1}{N}\sum_{i=1}^{N} X_i^2 \xrightarrow{p} \mathbb{E}\left[X_i^2\right]$ using the property of limits.

Continuous mapping theorem

- Using the LLN, we know that $\bar{X}_N \xrightarrow{p} \mathbb{E}[X_i]$.
- How about $\left(\bar{X}_N\right)^2$?
- Its convergence follows from the continuous mapping theorem (CMT).
- For a sequence of statistics, $\{\hat{\tau}_N\}$, if $\hat{\tau}_N \xrightarrow{p} \tau$, then

$$g\left(\hat{\tau}_N\right) \xrightarrow{p} g\left(\tau\right)$$

  for any continuous function $g()$.
- Since $\bar{X}_N \xrightarrow{p} \mathbb{E}[X_i]$, $\left(\bar{X}_N\right)^2 \xrightarrow{p} (\mathbb{E}[X_i])^2$.
- Therefore, $s^2 \xrightarrow{p} \mathbb{E}\left[X_i^2\right] - (\mathbb{E}[X_i])^2 = \sigma^2$.

St. Petersburg's paradox

- What would happen if the condition for LLN, $\text{Var}[X_i] < \infty$, is violated?
- Consider the following game: you will flip a fair coin until the first head appears.
- You receive \$$2^k$ if the first head appears on the k-th toss.
- E.g., you get $\$2^3 = \$8$ if the result is $\{T, T, H\}$.
- How much are you willing to pay to join the game?
- Let's compute the expectation of your payoff X:

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} 2^k * \frac{1}{2^k} = \infty.$$

- But are you willing to pay 1 million to play the game?

St. Petersburg's paradox

- What is the problem here?
- You cannot play any game forever, and what matters is how much you earn an average from playing it N times, i.e., $\bar{X}_N$.
- The return from the ith time of playing the game is X_i, which has the same distribution as X does.
- Let's compute X's variance:

$$\mathbb{E}\left[X^2\right] = \sum_{k=1}^{\infty}(2^k)^2 * \frac{1}{2^k} = \infty.$$

- Therefore, there is no guarantee that $\bar{X}_N \xrightarrow{p} \mathbb{E}[X_i] = \infty$.
- Using more advanced methods, we can show that $\bar{X}_N \approx \log_2 N$.
- If you play 100 rounds, the total return is roughly $100 * \log_2 100 = 100 * 6.64 = 664$.

Stochastic order

- In practice, we often want to know how fast an estimator converges to its limit when the sample size increases.
- If my sample size enlarges by 10 times, how much does the variance of $\hat{\tau}_N$ shrink?
- We can measure that with stochastic order, which compares the convergence rate of any $\hat{\tau}_N$ with that of a deterministic sequence.
- For a sequence $\{a_N\}$, we write $a_N = o(1)$ if $\lim_{N\to\infty} a_N = 0$.
- If $\lim_{N\to\infty} N^{\lambda} a_N = 0$, we write $a_N = o\left(N^{-\lambda}\right)$.
- If a_N is always bounded, we write $a_N = O(1)$.
- If $N^{\lambda} a_N$ is always bounded, we write $a_N = O\left(N^{-\lambda}\right)$.

Stochastic order

- We say $\hat{\tau}_N = o_p(1)$ when $\hat{\tau}_N \xrightarrow{p} 0$.
- If $N^{\lambda}\hat{\tau}_N = o_p(1)$, then we say $\hat{\tau}_N = o_p\left(N^{-\lambda}\right)$.
- $\hat{\tau}_N$ converges to zero faster than $1/N^{\lambda}$.
- E.g., $s^2 = o_p(1)$ and $\sqrt{N}s^2 = o_p(1)$, thus $s^2 = o_p(N^{-0.5})$.
- We say $\hat{\tau}_N = O_p(1)$ if it is bounded in probability: there exists $M > 0$ such that

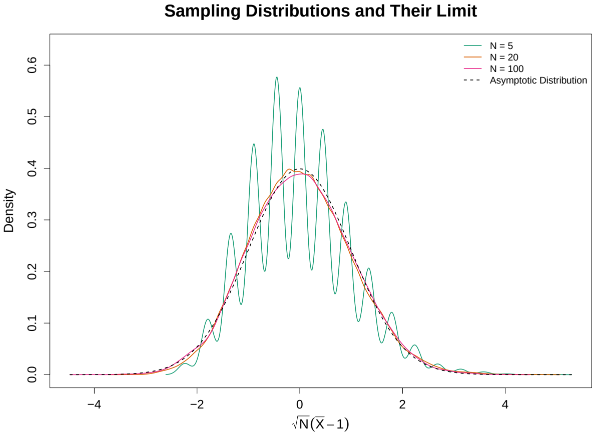
$$\mathbb{P}\left(|\hat{\tau}_N| > M\right) < \varepsilon$$

  for any $\varepsilon > 0$ when $N \to \infty$.
- In practice, it means that $\hat{\tau}_N$ converges to a fixed r.v.
- If $N^{\lambda}\hat{\tau}_N = O_p(1)$, then we say $\hat{\tau}_N = O_p\left(N^{-\lambda}\right)$.
- $\hat{\tau}_N$ converges to its limit is as fast as $1/N^{\lambda}$.
- E.g., $s^2 = O_p\left(N^{-1}\right)$ and $\bar{X}_N - \mu = O_p\left(N^{-0.5}\right)$.
- If $\hat{\tau}_N - \tau = O_p\left(N^{-0.5}\right)$, we say $\hat{\tau}_N$ is root-N consistent for τ.

The asymptotic distribution

- In many applications, we want to know more than the variance.
- As mentioned at the beginning, we need the sampling distribution to determine whether the estimate is driven by pure randomness.
- We usually do not know the sampling distribution under any sample size N.
- But we can learn the hypothetical distribution when $N = \infty$, which is known as the asymptotic distribution of $\hat{\tau}$.
- We first define the concept of convergence in distribution (weak convergence).
- We say $\hat{\tau}_N \xrightarrow{d} \tau$ if $F_{\hat{\tau}_N}(x) \rightarrow F_{\tau}(x)$ at any point x.
- For any x, $\{F_{\hat{\tau}_N}(x)\}$ can be seen as a sequence.
- The asymptotic distribution $F_{\tau}(\cdot)$ is the collection of limits of all the sequences.

The asymptotic distribution



Central limit theorem

- How do we know the functional form of the asymptotic distribution?
- If $(X_1, X_2, \ldots, X_N)$ are i.i.d. with $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \mathsf{Var}[X_i] < \infty$, then

$$\sqrt{N}\left(\bar{X}_N - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right).$$

- This result is known as the central limit theorem (CLT).
- Equivalently, $\bar{X}_N - \mu \xrightarrow{d} \mathcal{N}\left(0, \sigma^2/N\right)$.
- But we rescale the difference such that the asymptotic distribution remains the same.
- No matter what the population distribution is, the sampling distribution is approximately normal when N is large.

Central limit theorem

- The result holds for one particular estimator, the sample average.
- But most estimators in statistics can be written in the form of a sample average.
- E.g., the e.c.d.f. at x is a sample average of $\mathbf{1}\{X_i \leq x\}$.
- Any sample analogue is a sample average of $g(\mathbf{O}_i)$: $\hat{\tau}_N = \frac{1}{N}\sum_{i=1}^{N} g(\mathbf{O}_i)$.
- All these estimators are asymptotically normal under regularity conditions:
$$\sqrt{N}\left(\hat{\tau}_N - \tau\right) \xrightarrow{d} \mathcal{N}\left(0, N\sigma^2_{\hat{\tau}_N}\right).$$
- The CLT often holds even for non-i.i.d. data.
- But we may need additional constraints on higher-order moments of X_i or the covariance of any two observations.

Central limit theorem

- From the CLT, we know that

$$t_N = \frac{\hat{\tau}_N - \tau}{\sigma_{\hat{\tau}_N}} \xrightarrow{d} \mathcal{N}(0,1).$$

- Therefore, we can calculate the probability for t_N to be inside any set S using the c.d.f. of the standard normal distribution, e.g.,

$$\begin{aligned}\mathbb{P}\left(|\hat{\tau}_N - \tau| \leq c\right) =& \mathbb{P}\left(-\frac{c}{\sigma_{\hat{\tau}_N}} \leq t_N \leq \frac{c}{\sigma_{\hat{\tau}_N}}\right) \\ =& \Phi\left(\frac{c}{\sigma_{\hat{\tau}_N}}\right) - \Phi\left(-\frac{c}{\sigma_{\hat{\tau}_N}}\right) = 2\Phi\left(\frac{c}{\sigma_{\hat{\tau}_N}}\right) - 1.\end{aligned}$$

- In practice, we use $\hat{\sigma}_{\hat{\tau}_N}$ to approximate $\sigma_{\hat{\tau}_N}$.

Central limit theorem

- In Gerber and Green (2000), they estimated the effect of get-out-to-vote (GOTV) messages on turnout with $\hat{\tau}_N = 8.5\%$ and $\hat{\sigma}_{\hat{\tau}_N} = 2.6\%$.
- What is the probability that the deviation of the estimate from the true effect is larger than 8%?

$$\begin{aligned}\mathbb{P}\left(|\hat{\tau}_N - \tau| > 8\%\right) =& 1 - \mathbb{P}\left(|\hat{\tau}_N - \tau| \leq 8\%\right) \\ =& 2 - 2\Phi\left(\frac{8\%}{2.6\%}\right).\end{aligned}$$

```r
2 - 2*pnorm(0.08 / 0.026)
```

```
## [1] 0.002091493
```

- We can confidently claim that the probability for the true effect to be non-positive is lower than 1%.

Delta method

- The CLT suggests that $\frac{1}{N}\sum_{i=1}^{N} X_i^2$ is asymptotically normal.
- How about $\left(\bar{X}_N\right)^2$ and s^2?
- The p.d.f. for $g(X)$ is not just $g(f_X(x))$ in most cases.
- Nevertheless, if $g(\cdot)$ is continuously differentiable and $\hat{\tau}_N$ is asymptotically normal, then

$$\sqrt{N}\left(g\left(\hat{\tau}_N\right) - g\left(\tau\right)\right) \xrightarrow{d} \mathcal{N}\left(0, N\sigma^2_{\hat{\tau}_N}\left(g'\left(\tau\right)\right)^2\right).$$

- Using the first-order approximation, we know that

$$\sqrt{N}\left(g\left(\hat{\tau}_N\right) - g\left(\tau\right)\right) \approx \sqrt{N}g'(\tau)\left(\hat{\tau}_N - \tau\right).$$

- The previous result holds as $g'(\tau)$ is just a constant.
- This is known as the Delta method in statistics, and we can see

$$\sqrt{N}\left(\left(\bar{X}_N\right)^2 - \mu^2\right) \xrightarrow{d} \mathcal{N}\left(0, 4\mu^2\sigma^2\right).$$

Hoeffding's inequality (*)

- The asymptotic analysis is always an approximation.
- We approximate the sampling distribution with a normal distribution and the true variance with its estimate.
- Non-asymptotic analysis allows us to see the precision of this approximation with a given sample size
- Suppose $X_i \in [a_i, b_i]$ for any i with $\mathbb{E}[X_i] = \mu$, then

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right| \geq t\right) \leq 2\exp\left(-\frac{2N^2t^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\right).$$

- This is known as Hoeffding's inequality.
- Let X_i represent individual i's ideological score ranging from 1 to 7 ($a_i = 1$ and $b_i = 7$).
- How many observations do we need such that the probability $\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right| \geq 1$ is smaller than 0.05?
- We want to find N such that $2\exp(-2N/36) \leq 0.05$.
- The condition is that $N \geq 67$.

The finite-population perspective (*)

- So far, we have assumed that the data are i.i.d.
- Suppose we want to sample three individuals from the classroom with 15 people and measure their height.
- To ensure randomness, we draw the sample using lottery with three winning tickets.
- The probability for each person to be in the sample is $1/5$.
- But the probability for any two people to be included is $\frac{3}{15}\frac{2}{14} \neq \left(\frac{1}{5}\right)^2$.
- When the population is finite relative to the sample, data can hardly be seen as i.i.d.
- A more realistic view: heights are fixed values in the population, and randomness only comes from sampling.

The finite-population perspective (*)

- We define the sampling indicator:

$$R_i = \begin{cases} 1, & \mathbb{P}(R_i = 1) = 1/5, \\ 0, & \mathbb{P}(R_i = 0) = 4/5. \end{cases}$$

- The sample average estimator can be written as

$$\bar{X} = \frac{1}{3} \sum_{i=1}^{3} X_i = \frac{1}{15p} \sum_{i=1}^{15} R_i X_i,$$

  where $p = 1/5$ and each X_i is a fixed value.
- In this case, $\bar{X}$ is still unbiased for the population mean:

$$\mathbb{E}\left[\bar{X}\right] = \frac{1}{15p} \sum_{i=1}^{15} \mathbb{E}\left[R_i\right] X_i = \frac{1}{15} \sum_{i=1}^{15} X_i = \tau_{15}.$$

The finite-population perspective (*)

- For asymptotics, we consider a series of finite populations $\{\mathcal{X}_n\}$ with $|\mathcal{X}_n| = n$ and $n \to \infty$.
- The sample size N increases with n, and $\frac{N}{n} = r$.
- We can similarly show that $\mathbb{E}\left[\bar{X}_N\right] = \frac{1}{n}\sum_{i=1}^{n} X_i = \tau_n$ for any n and N.
- Note that $\text{Var}\left[R_i\right] = r(1-r)$ and $\text{Cov}\left[R_i, R_j\right] = rr^* - r^2$, where $r^* = \frac{N-1}{n-1}$.
- Therefore,

$$
\begin{aligned}
\text{Var}\left[\bar{X}_N\right] =& \frac{1}{n^2 r^2}\sum_{i=1}^{n}\text{Var}\left[R_i\right]X_i^2 + \frac{1}{n^2 r^2}\sum_{i=1}^{n}\sum_{j\neq i}\text{Cov}\left[R_i, R_j\right]X_i X_j \\
=& \frac{1-r}{n^2 r}\sum_{i=1}^{n} X_i^2 + \frac{r^*-r}{n^2 r}\sum_{i=1}^{n}\sum_{j\neq i} X_i X_j \\
=& \frac{1-r}{n(n-1)r}\sum_{i=1}^{n}(X_i - \tau_n)^2 \to 0.
\end{aligned}
$$

The finite-population perspective (*)

- Even though the variance takes a different form, $\bar{X}_N$ is still consistent for τ_n.
- Asymptotic normality requires that

$$\lim_{n\to\infty} \frac{\max_{1\leq i\leq n}(X_i - \tau_n)^2}{\sum_{i=1}^{n}(X_i - \tau_n)^2} = 0.$$

- Another result from Hoeffding.
- The finite-population perspective is useful when the sample and the population are similar in size.
- It has been increasingly popular for experimental analysis.
- E.g., we can compute the difference in average GDP per capita between red and blue states.
- What does the variance tell us?