Lecture 5: Monte Carlo simulations

IST5573

統計方法 Statistical methods

2016/10/12

Monte Carlo simulation

- Computers can be used to generate **pseudo-random numbers**.
- These pseudorandom numbers can be used to **imitate random variables** from the real world.
  - This permits us to examine properties of random variables using a computer instead of theoretical or analytical derivations.
  - We can also create simulated data to test out ideas or competing methods, without actually having to perform laboratory experiments.

- Simulations can also be used to check theoretical or analytical results. Also, many of the theoretical results we use in statistics are based on **asymptotic**s: they hold when **the sample size goes to infinity**.
- In practice, we never have an infinite number of samples so we may want to know how well the theory works with our actual sample size.
- Sometimes we can answer this question analytically, but not always. Simulations are extremely useful in these cases.

- As an example, let's use a Monte Carlo simulation to compare the CLT to the t-distribution approximation for different sample sizes.
  - RMD_example 5.1

Parametric simulations for the observations

- What we did in RMD_example 5.1 was a type of Monte Carlo simulation **when we had access to population data** and generated samples at random.
- In practice, we do not have access to the entire population.
  - The reason for using the approach here was for educational purposes.

- When we want to use Monte Carlo simulations in practice, it is much more typical to **assume a parametric distribution** and generate a population from this, which is called a **parametric simulation**.
  - This means that we take parameters estimated from the real data (here the mean and the standard deviation), and plug these into a model (here the normal distribution).
  - RMD_example 5.2

Parametric vs. nonparametric simulations

- **Parametric simulations** assumed that the random sample was from some known distribution, and we then generated random numbers from this distribution to simulate characteristics of estimators of interest.
- In reality, the distribution that the random sample belongs to is either unknown or too complicated to generate data from.
- We can consider methods that allow us to generate the underlying distribution using the **observed sample** (i.e., **nonparametric (Monte Carlo) simulations**).

What is the bootstrap?

- Bootstrap methods are a class of nonparametric simulations that estimate the distribution of a population by **resampling**.
- Resampling methods treat an observed sample as a finite population, and random samples are generated (resampled) from it to estimate population characteristics and make inferences about the sampled population.
- Bootstrap methods are often used when the distribution of the target population is not specified; the sample is the only information available.

(Nonparametric) bootstrap

- Suppose the observed sample is $x_1, x_2, \cdots, x_M$, θ is the population characteristics we want to estimate (e.g., the population standard deviation σ_X), and we use $\hat{\theta}$ (a random variable) to estimate it (e.g., the sample standard deviation s_X)

1. For each bootstrap replicate, indexed by $b = 1, \cdots, B$:
   a) Draw M values with **replacement** from the set $\{x_1, x_2, \cdots, x_M\}$
   b) Compute $\hat{\theta}$ with values sampled in a), denoted as $\hat{\theta}^{(b)}$
2. The bootstrap estimate can be obtained via $\hat{\theta}^{(1)}, \cdots, \hat{\theta}^{(B)}$

Bootstrap estimate of variance

- The variance of your estimator (i.e., $\mathrm{var}(\hat{\theta})$) (e.g., the variance of the sample standard deviation $\mathrm{var}(s_X)$) can be used to evaluate how accurate your estimator.
- However, this variance is typically difficulty to obtain theoretically. The bootstrap can help here.
- The bootstrap estimate of **variance** of $\hat{\theta}$

$$\widehat{\mathrm{var}}(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^{(b)} - \bar{\hat{\theta}}\right)^2$$

where $\bar{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}$

- The bootstrap estimate of **standard error** of $\hat{\theta}$ (i.e., $\sqrt{\text{var}(\hat{\theta})}$) is the square root of the variance estimate.

Bootstrap estimate of confidence interval

- For the bootstrap estimate of 95% confidence interval for θ (we can use percentages other than 95%), one can use

the upper 2.5 percentage point and the upper 97.5 percentage point of $\hat{\theta}^{(1)}, \cdots, \hat{\theta}^{(B)}$

- RMD_example 5.3

Permutation tests

- Suppose we have a situation in which none of the standard mathematical statistical approximations apply.
- In practice, we do not have access to all values in the population so we can't perform a simulation as done above.
- Permutation tests can be useful in these scenarios.

- For example, we have computed a summary statistic, the difference in mean, to determine if the observed difference was significant.
- In previous sections, we showed parametric approaches (CLT) that can help.
- **Permutation tests** take advantage of the fact that **if we randomly shuffle the cases and control labels, then the null is true.**
  - Here is how we generate a null distribution by shuffling the data 1,000 times: RMD_example 5.4

Notes

- Keep in mind that there is no theoretical guarantee that the null distribution estimated from permutations approximates the actual null distribution.
  - For example, if there is a real difference between the populations, some of the permutations will be unbalanced and will contain some samples that explain this difference.
- This is why permutations result in conservative p-values. For this reason, when we have **few samples**, we **cannot do permutations**.