Workshop 3

Bolun

01/27/2021

Review

- Let me preach a little bit (again)
- Population and sample
- Statistic inference: from sample to population
- Normal distribution
- Before we enter into the discussion, let's review the codes in small groups (and thank Zahra!).

Importance of the randomness

Group Discussion:

- Why randomness is important for inference?
- What kind of normative principle lies behind it?
- What if a sample is conentrate among certain group of people?

BTW: Beware! When people talk about inference/prediction, they might mean a lot of different things.

Sampling Distribution 1

What is sampling distribution: The probability distribution of a given statistics (mean, sd etc.).

Here using sampling distribution of the mean as an example.

Imagine that you have a given population of 1000 people, each have a given value of asset. You randomly selected 100 from them, and calculate the mean of this sample. Then, you repeat this process for almost infinite time, the probability distribution you come up with in the end is a sampling distribution of the mean.

Sampling Distribution 2

Let's simulating it in codes.

Standard Error 1

Standard error is the standard deviation of the sampling distribution.

Mathematically you can derive that

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Where σ is the standard deviation of the population, and n is the size of the sample.

In a lot of situation, if we know the kind of distribution the population belongs, we can infer the sampling distribution mathematically. In the case of mean estimate, the sampling distribution is a normal distribution.

Standard Error 2

But usually we do not know the standard deviation of the population as we did in the last simulation. Thus, we usually estimate the standard error using the standard deviation of the sample.

$$\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}}$$

or a old fashioned way

$$\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n-1}}$$

Normal Distribution and Student Distribution

Review: normal distribution and student distribution.

Two different situations:

1) When the sample size is large enough ($n > 30$), the sampling distribution is also a normal distribution.
2) When our sample size is small, the standard error estimate is not accurate enough. In this situation, we do not use the normal distribution for the purpose of statistic inference. Instead, we use student distribution.

Group Activities

Group Activity: Changing the codes and modify the size of the sample to a small number, see how the distribution changed.

Mean point estimate and confidential intervals 1

Group Activity:

- Take a large n size sample from the population we generated previously, using simulation to generate a distribution of (sample mean - population mean).
- Use normal distribution to calculate the 95% confidential intervals for the point estimate.
- Compare the results

Mean point estimate and confidential intervals 2

Group Activity:

- Take a small n size sample from the population we generated previously, using simulation to generate a distribution of (sample mean - population mean).
- Use student distribution to calculate the 95% confidential intervals for the point estimate.
- Compare to previous result, what do you find?

Ending: from estimation to hypothesis testing

- Similar logic
- Clarification about confidential intervals:
  - "A 95% confidence level does not mean that for a given realized interval there is a 95% probability that the population parameter lies within the interval (i.e., a 95% probability that the interval covers the population parameter)."
  - It's about the difference between the estimate and the true parameter.
- The sampling distribution is set differently: based on sample VS based on hypothesis
- Central Limit Theorem.