Chapter 5: Estimation

(Ott & Longnecker Sections: 5.8)

Duzhe Wang

https://dzwang91.github.io/stat371/

Part 6

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Summary and motivation

- General form of CI: estimate $\pm$ multiplier $*$ estimated SE of the estimator

Population Distribution	$X \sim N(\mu, \sigma^2)$		$X \nsim N(\mu, \sigma^2)$	
subcase	σ is known	σ is unknown	n is large (like n>30)	n is small
CI	$\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$	$\bar{X} \pm t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$	$\bar{X} \pm z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$	Bootstrap???

- How do we make a CI when the population is not normal and the sample size is small?

Example

Story

Secondhand smoke is of great health concern, especially for children. The level of exposure could be determined by measuring the urinary concentration of cotanine.

Primary Research Question

The Child Protective Services (CPS) in a city need to know the mean cotanine level for children in foster care.

$$\mu = ?$$

Sampling

15 children were selected randomly, and their urinary concentration of cotanine was measured.

$$n = 15 < 30$$

Observed Data

29, 30, 53, 75, 34, 21, 12, 58, 117, 119, 115, 134, 253, 289, 287

$$\bar{x} = 108.4$$

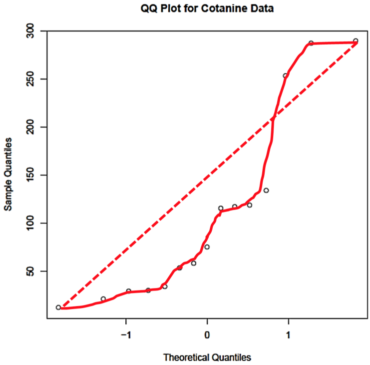
$$s = 95.6$$

Example

- Question 1: does the sample come from a normal population?

Example

- Question 1: does the sample come from a normal population?



It looks pretty bad, and with the small sample size we may not be able to rely on the CLT as an accurate approximation to the distribution of the sample mean.

Example

- Question 2: what's the challenge for us to make a CI of μ in this example?

Example

- Question 2: what's the challenge for us to make a CI of μ in this example?
  Without assuming normality of the population, the quantity
  $$t = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$$
  will not be a T-distribution.
- Question 3: how do we get the distribution of the statistic t?

Example

- Question 2: what's the challenge for us to make a CI of μ in this example?
  Without assuming normality of the population, the quantity
  $$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$
  will not be a T-distribution.
- Question 3: how do we get the distribution of the statistic t?
  By simulations! And we call it bootstrapping.

The bootstrap method

Given the original data set: $x_1, x_2, ..., x_n$.

1. Compute the sample mean $\bar{x}$ and sample standard deviation s of the original data.

The bootstrap method

Given the original data set: $x_1, x_2, ..., x_n$.

1. Compute the sample mean $\bar{x}$ and sample standard deviation s of the original data.
2. Draw n data points from the original data set with replacement. Call these new observations x_1^*, x_2^*, ..., x_n^*. (This is like treating your original sample as a population, and sampling iid from this new population!) (What is the mean of this new population if we assume $x_1, ..., x_n$ is uniformly distributed?)

The bootstrap method

Given the original data set: $x_1, x_2, ..., x_n$.

1. Compute the sample mean $\bar{x}$ and sample standard deviation s of the original data.
2. Draw n data points from the original data set with replacement. Call these new observations x_1^*, x_2^*, ..., x_n^*. (This is like treating your original sample as a population, and sampling iid from this new population!) (What is the mean of this new population if we assume $x_1, ..., x_n$ is uniformly distributed?)
3. Compute the mean and standard deviation of the resampled data. Call them $\bar{x}^*$ and s^*.

The bootstrap method

Given the original data set: $x_1, x_2, ..., x_n$.

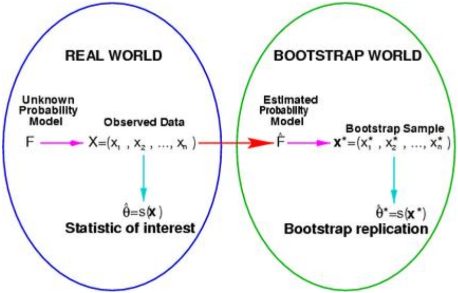
1. Compute the sample mean $\bar{x}$ and sample standard deviation s of the original data.
2. Draw n data points from the original data set with replacement. Call these new observations x_1^*, x_2^*, ..., x_n^*. (This is like treating your original sample as a population, and sampling iid from this new population!) (What is the mean of this new population if we assume $x_1, ..., x_n$ is uniformly distributed?)
3. Compute the mean and standard deviation of the resampled data. Call them $\bar{x}^*$ and s^*.
4. Compute the realization of statistic $\hat{t} = \frac{\bar{x}^* - \bar{x}}{\frac{s^*}{\sqrt{n}}}$. (If we treat our original sample like a population, $\bar{x}$ plays the role of the population mean!)

The bootstrap method

Given the original data set: $x_1, x_2, ..., x_n$.

1. Compute the sample mean $\bar{x}$ and sample standard deviation s of the original data.
2. Draw n data points from the original data set with replacement. Call these new observations x_1^*, x_2^*, ..., x_n^*. (This is like treating your original sample as a population, and sampling iid from this new population!) (What is the mean of this new population if we assume $x_1, ..., x_n$ is uniformly distributed?)
3. Compute the mean and standard deviation of the resampled data. Call them $\bar{x}^*$ and s^*.
4. Compute the realization of statistic $\hat{t} = \frac{\bar{x}^* - \bar{x}}{\frac{s^*}{\sqrt{n}}}$. (If we treat our original sample like a population, $\bar{x}$ plays the role of the population mean!)
5. Repeat steps 2-4 a large number of times(say 1000 times), and compute $\hat{t}$ from each one. Put these values of $\hat{t}$ in order and throw them into a histogram. This is an approximation to the true sampling distribution of t!

The bootstrap method

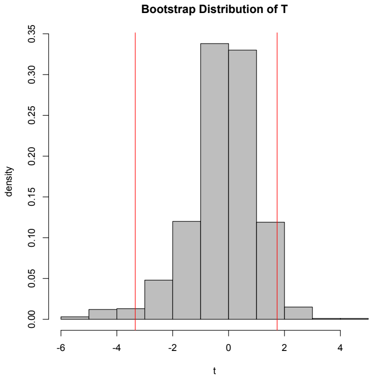


The bootstrap method

- How do we make a CI using the bootstrap method?
  - After step 1-5, find the $\alpha/2$ and $1-\alpha/2$ critical values of the approximate sampling distribution you've generated with all these $\hat{t}$. Call these critical values $\hat{t}_{(\alpha/2)}$ and $\hat{t}_{(1-\alpha/2)}$. (What is $\alpha/2$ critical value?)
  - An approximate $100(1-\alpha)\%$ CI for μ is now:
$$(\bar{x} - \hat{t}_{(\alpha/2)}\frac{s}{\sqrt{n}}, \bar{x} - \hat{t}_{(1-\alpha/2)}\frac{s}{\sqrt{n}}).$$

Example

- For the secondhand smoke data, we find $\bar{x} = 108.4$ and $s = 95.6$. Bootstrapping 1000 times yields the following approximate distribution of t:



Example

- You can see that this distribution is not very symmetric, and thus quite unlike a t or normal.
- Let's compute a 95% confidence interval. Using R, we have $\hat{t}_{(1-0.05/2)} = -3.34$ and $\hat{t}_{0.025} = 1.74$. Thus the 95% CI is:

$$(108.4 - 1.74\frac{95.6}{\sqrt{15}}, 108.4 - (-3.34)\frac{95.6}{\sqrt{15}}) = (65.53, 190.88).$$

- See relevant R code from course website.

What's the next?

We'll start hypothesis testing in the next lecture.