Chapter 5: Estimation
(Ott & Longnecker Sections: 5.8)

Duzhe Wang

https://dzwang91.github.io/stat324/

Part 5

Summary and motivation

- General form of CI:

$$L = \text{estimator} - \text{critical value} \times \text{standard error of the estimator},$$

$$U = \text{estimator} + \text{critical value} \times \text{standard error of the estimator}$$

- CI for population mean:

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

Summary and motivation

- General form of CI:

$$L = \text{estimator} - \text{critical value} \times \text{standard error of the estimator},$$

$$U = \text{estimator} + \text{critical value} \times \text{standard error of the estimator}$$

- CI for population mean:

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**

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

A motivating 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

$$\bar{x} = 108.4$$

$$s = 95.6$$

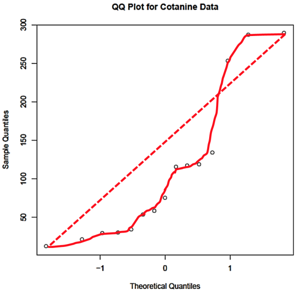
Our goal is to build a CI for μ.

A motivating example

Does the sample come from a normal distribution?

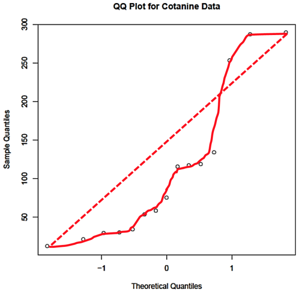
A motivating example

Does the sample come from a normal distribution?



A motivating example

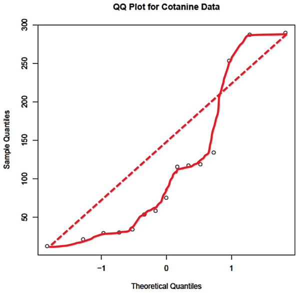
Does the sample come from a normal distribution?



- It looks pretty bad.

A motivating example

Does the sample come from a normal distribution?

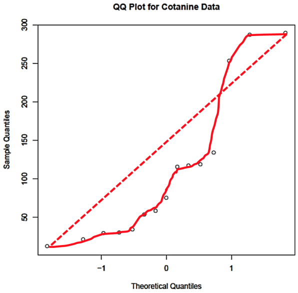


- It looks pretty bad.

Can we use CLT?

A motivating example

Does the sample come from a normal distribution?



- It looks pretty bad.

Can we use CLT?

- Too small sample size.

The essence of CI

- If $X_1, ..., X_n \sim N(\mu, \sigma^2)$ i.i.d., then

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}.$$

- Therefore,

$$\mathbb{P}(-t_{n-1,\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \leq t_{n-1,\frac{\alpha}{2}}) = 1 - \alpha.$$

- Therefore,

$$\mathbb{P}(\bar{X} - t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}}) = 1 - \alpha.$$

The essence of CI

What is the distribution of $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ when sample doesn't come from the Normal distribution?

The essence of CI

What is the distribution of $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ when sample doesn't come from the Normal distribution?

- Use simulations and get the empirical distribution!

The essence of CI

What is the distribution of $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ when sample doesn't come from the Normal distribution?

- Use simulations and get the empirical distribution!

How can we sample in each simulation?

The essence of CI

What is the distribution of $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ when sample doesn't come from the Normal distribution?

- Use simulations and get the empirical distribution!

How can we sample in each simulation?

- Bootstrap!

The bootstrap method

Given the original data set: $x_1, x_2, ..., x_n$. (n is small)

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

The bootstrap method

Given the original data set: $x_1, x_2, ..., x_n$. (n is small)

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

One run of bootstrap

2. Draw n data points from the original data set with replacement. Call these new observations x_1^*, x_2^*, ..., x_n^*.
3. Compute the mean and standard deviation of the resampled data. Call them $\bar{X}^*$ and S^*.
4. Compute the realization of statistic $\hat{t} = \frac{\bar{X}^* - \bar{X}}{\frac{S^*}{\sqrt{n}}}$.

The bootstrap method

Given the original data set: $x_1, x_2, ..., x_n$. (n is small)

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

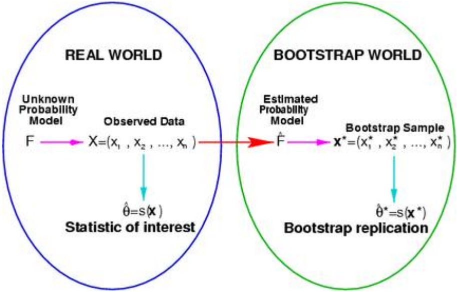
One run of bootstrap

2. Draw n data points from the original data set with replacement. Call these new observations x_1^*, x_2^*, ..., x_n^*.
3. Compute the mean and standard deviation of the resampled data. Call them $\bar{X}^*$ and S^*.
4. Compute the realization of statistic $\hat{t} = \frac{\bar{X}^* - \bar{X}}{\frac{S^*}{\sqrt{n}}}$.

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 $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$!

The bootstrap method



How do we make a CI using the histogram from bootstrap?

The bootstrap method

How do we make a CI using the histogram from bootstrap?

6. 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?)
7. An approximate $100(1 - \alpha)\%$ CI for μ is

$$(\bar{X} - \hat{t}_{(\alpha/2)}\frac{S}{\sqrt{n}}, \bar{X} - \hat{t}_{(1-\alpha/2)}\frac{S}{\sqrt{n}}).$$

Bootstrap in R

Step 1: Create the original data set

```r
data = c(29,30,53,75,34,21,12,58,117,119,
         115,134,253,289,287)
n=length(data) # sample size
```

Bootstrap in R

Step 2: Calculate the sample mean and sample standard deviation of the original data set

```r
xbar = mean(data)
s = sd(data)
```

Bootstrap in R

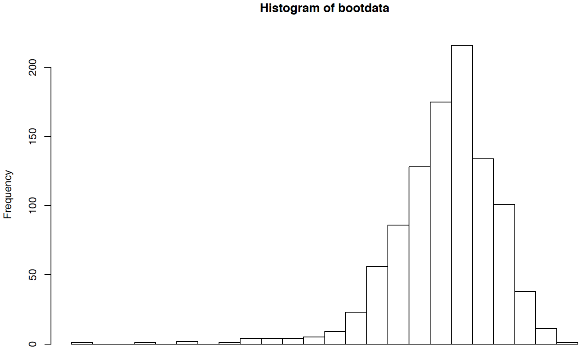
Step 3 (a): Build a bootstrap function

```
bootest=function(data,b) {
  bootstat=NULL
  truemean=mean(data)
  for(i in 1:b) {
    samp=sample(data, size = length(data), replace = T)
    bootmean=mean(samp)
    bootsd=sd(samp)
    bootstat[i]=(bootmean - truemean)/(bootsd/sqrt(n))
  }
  return(bootstat)
}
```

Bootstrap in R

Step 3 (b): Run boostrap b times

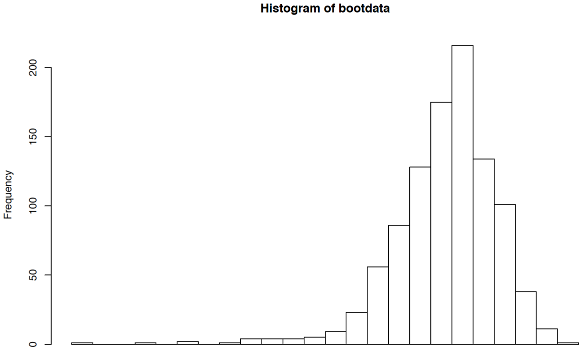
```
b = 1000        # number of bootstrap samples
bootdata=bootest(data,b)
hist(bootdata, breaks = 20)
```



Bootstrap in R

Step 3 (b): Run boostrap b times

```r
b = 1000        # number of bootstrap samples
bootdata=bootest(data,b)
hist(bootdata, breaks = 20)
```



This distribution is not very symmetric, and thus quite unlike a t or normal.

Bootstrap in R

Step 4: Calculate critical values

```
alpha=0.05    # 95% CI
```

$\hat{t}_{1-\alpha/2}$:

```
lower=quantile(bootdata, probs=alpha/2)
lower
```

```
##      2.5%
## -2.71855
```

$\hat{t}_{\alpha/2}$:

```
upper=quantile(bootdata, probs=1-alpha/2)
upper
```

```
##    97.5%
## 1.802621
```

Bootstrap in R

Step 5: Build a CI

```
CI_lower=xbar-upper*s/sqrt(n)
CI_upper=xbar-lower*s/sqrt(n)
print(c(CI_lower, CI_upper))
```

```
##     97.5%      2.5%
##  63.90435 175.50433
```

What's the next?

We'll start hypothesis testing in the next lecture.