Confidence Interval

Illustrated through inference on one population mean or proportion

Motivation & simple random sample

Eg) We wish to estimate the average height of adult US males

➔ Take a random sample.

- "Simple" random sample: every subject in the population has the same chance to be selected.

Introduction to statistical inference on one population mean

For a "**random** sample" of size n: $X_1, X_2, \ldots, X_n$

<i> Point estimator $\overline{X}$ → sample mean $\left(= \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}\right)$

Other estimators: median, mode, trimmed mean, …

<ii> Confidence Interval (C.I.)

Eg) 95% C.I. for μ

99.9999% C.I. ('6-9' in the manufacture industry)

<iii> Hypothesis Test

Eg) $H_0 : \mu \le 5'6''$

$H_1 : \mu \succ 5'6''$

Point Estimator, C.I., Test $\Rightarrow$ Statistical Inference

- Draw some conclusion on the population (parameters of interest) based on a random sample.

1. The Exact Confidence Interval for μ when the population is normal & σ^2 is known

① Point estimator and confidence interval for μ

- When the population is normal and the population variance is known.
- Let $X_1, X_2, \ldots, X_n$ be a random sample for a normal population with mean μ and variance σ^2. That is, $X_i \overset{iid.}{\sim} N(\mu, \sigma^2), i = 1, \ldots, n$.
- For now, we assume that σ^2 is known.

<i> Point Estimator for μ : $\hat{\mu} = \overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

$E(\hat{\mu}) = E(\overline{X}) = \mu \Rightarrow \hat{\mu} = \overline{X}$ is an unbiased estimator of μ

- Intuitively, this means if you take "many" samples of size n from the population, then the mean of these samples means would be equal to μ if you take a large enough # of samples.
- $\overline{X}$ is also a maximum likelihood estimator (MLE) of μ.
- $\overline{X}$ is also a method of moment estimator (MOME) of μ.
- Other good properties too.

<ii> Confidence Interval for μ

- Intuitive approach (backwards derivation for the CI boundaries $C_1 and\ C_2$):

$$P(C_1 \leq \mu \leq C_2) = 0.95$$
$$P(-C_1 \geq -\mu \geq -C_2) = 0.95$$
$$P(\bar{X} - C_1 \geq \bar{X} - \mu \geq \bar{X} - C_2) = 0.95$$

$$P\left(\frac{\bar{X} - C_1}{\sigma/\sqrt{n}} \geq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{\bar{X} - C_2}{\sigma/\sqrt{n}}\right) = 0.95$$

Since we know

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

We can compute the expressions for $C_1 and\ C_2$.

However, one question is that there are MANY ways to choose the C's.

Later you will see that for pivotal quantity with symmetric pdfs, the symmetric CIs are the optimal – in that they have the shortest lengths for the given confidence level 100(1-α)%.

Now we present a general approach to derive the CI's.

General approach for deriving CI's : the <u>Pivotal Quantity</u> (P.Q.) approach

*Definition: A pivotal quantity is a function of the sample and the parameter of interest. Furthermore, its distribution is entirely known.

1. We start by looking at the point estimator of μ. $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

* Is $\overline{X}$ a pivotal quantity for μ?

→ $\overline{X}$ is not because μ is unknown.

* function of $\overline{X}$ and μ : $\overline{X} - \mu \sim N(0, \frac{\sigma^2}{n})$

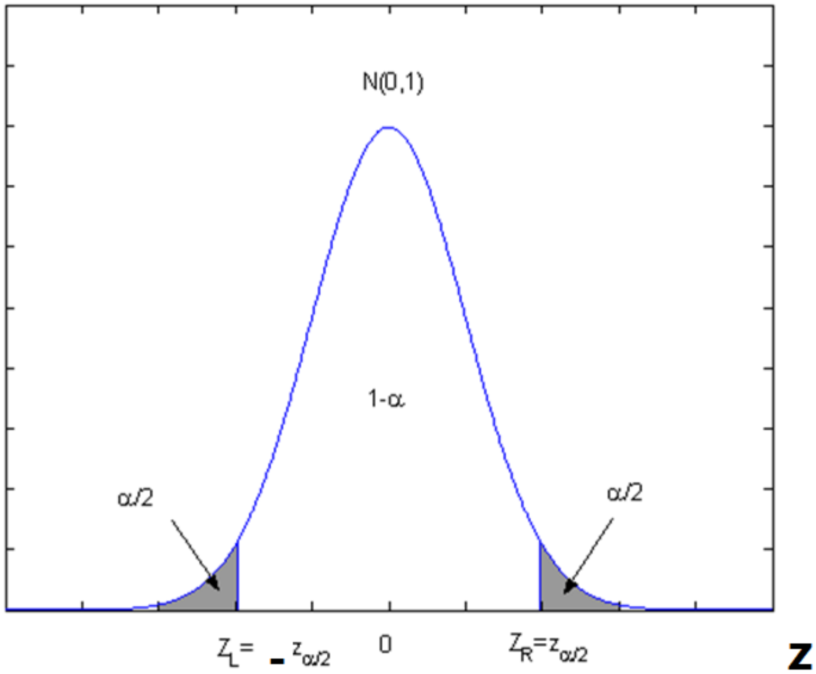
→ Yes, it is pivotal quantity.

* Another function of $\overline{X}$ and μ : $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

→ Yes, it is pivotal quantity.

So, Pivotal Quantity is not unique.

2. Now that we have found the pivotal quantity Z, we shall start the derivation for the <u>symmetrical CI's</u> for μ from the PDF of the pivotal quantity Z



100(1-α)% CI for μ, 0<α<1

(e.g. α=0.05 ⇒ 95% C.I.)

$$P(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}) = 1-\alpha$$

$$P(-Z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le Z_{\alpha/2}) = 1-\alpha$$

$$P(-Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \le \overline{X}-\mu \le Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$$P(-\overline{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \le -\mu \le -\overline{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$$P(\overline{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \ge \mu \ge \overline{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$$P(\overline{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

∴ the 100(1-α)% C.I. for μ is $[\overline{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}]$

***Note, some special values for α and the corresponding $Z_{\alpha/2}$ values are:**

1. The 95% CI, where $\alpha = 0.05$ and the corresponding $Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$

2. The 90% CI, where $\alpha = 0.1$ and the corresponding $Z_{\frac{\alpha}{2}} = Z_{0.05} = 1.645$

3. The 99% CI, where $\alpha = 0.01$ and the corresponding $Z_{\frac{\alpha}{2}} = Z_{0.005} = 2.575$

Example 1. A random sample of 400 adult US male was taken and the sample mean was found to be $\bar{X} = 5'7'' = 67\ inches$. Based on past studies, it is believed that the population distribution of all adult US male is normal and the standard deviation is 30 inches. Please construct a 95% confidence interval for the average height of all adult US male based on this sample.

Solution: The 95% CI for μ is

$$\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right] = \left[67 - 1.96\frac{30}{\sqrt{400}}, 67 + 1.96\frac{30}{\sqrt{400}}\right] \approx [64, 70]$$

That is, the estimated 95% confidence interval for the average height of all adult US male is [5'4", 5'10"].

… …

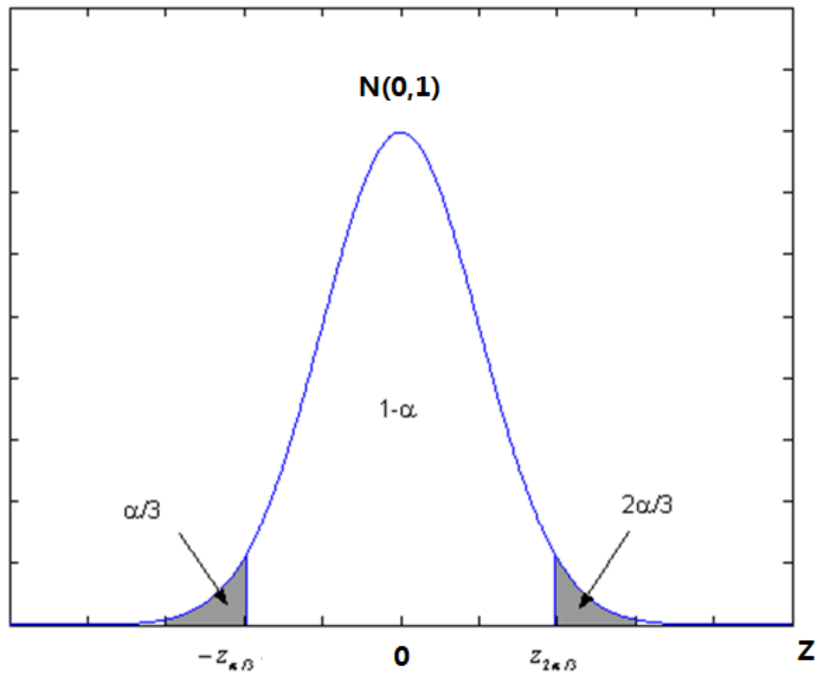
This means that we are 95% sure the population mean μ would lie between 5'4" and 5'10".

∴Recall the 100(1-α)% symmetric C.I. for μ is $[\bar{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}]$

Please note that this CI is symmetric around $\bar{X}$

The length of this CI is: $L_{sy} = 2 \cdot Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

Now we derive a non-symmetrical CI:



$$P(-Z_{\alpha/3} \le Z \le Z_{2/3\alpha}) = 1-\alpha$$

100(1-α)% C.I. for μ

$$\Rightarrow [\bar{X} - Z_{2/3\alpha} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{1/3\alpha} \cdot \frac{\sigma}{\sqrt{n}}]$$

Compare the lengths of the C.I.'s, one can prove theoretically that:

$$L = (Z_{\alpha/3} + Z_{2/3\alpha}) \cdot \frac{\sigma}{\sqrt{n}} > L_{sy} = 2 \cdot Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

You can try a few numerical values for α, and see for yourself. For example,

$$\boldsymbol{\alpha = 0.05}$$

HW: Please derive the 100(1-α)% symmetric C.I. for μ based on a random sample from a normal population with unknown variance

2. (Large Sample) Confidence interval for a population mean (*any population) or a population proportion p

<Theorem> Central Limit Theorem

$$\boxed{Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{n \to \infty} N(0,1)}$$

When n is large enough, we have

$$\boxed{Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1)}$$

That means Z follows approximately the normal (0,1) distribution.

Application #1. Inference on μ when the population distribution is unknown but the sample size is large

$$\boxed{Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \dot{\sim} N(0,1)}$$

By Slutsky's Theorem We can also obtain another pivotal quantity when σ is unknown by plugging the sample standard deviation S as follows:

$$\boxed{Z = \frac{\bar{X} - \mu}{S / \sqrt{n}} \dot{\sim} N(0,1)}$$

We subsequently obtain the $100(1-\alpha)\%$ C.I. using the second P.Q. for μ: $\boxed{\bar{X} \pm Z_{\alpha/2} \frac{S}{\sqrt{n}}}$

Application #2. Inference on one population proportion p when the population is Bernoulli(p) *** Let $X_i \overset{i.i.d.}{\sim} Bernoulli(p),\ i = 1, \cdots, n$, please find the 100(1-α)% CI for p.

Point estimator : $\hat{p} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ (ex. $n = 1000$, $\hat{p} = 0.6$)

Our goal: derive a 100(1-α)% C.I. for p

Thus for the Bernoulli population, we have:

$$\boxed{\mu = E(X) = p}$$

$$\boxed{\sigma^2 = Var(X) = p(1-p)}$$

Thus by the CLT we have:

$$\boxed{Z = \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)}$$

Furthermore, we have for this situation: $\bar{X} = \hat{p}$

Therefore we obtain the following pivotal quantity Z for p:

$$\boxed{Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)}$$

By Slustky's theorem, we can replace the population proportion in the denominator with the sample proportion and obtain another pivotal quantity for p:

$$\boxed{Z^* = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \dot{\sim} N(0,1)}$$

 # Thus the $100(1-\alpha)\%$ **(approximate, or large sample) C.I. for p based on the second pivotal quantity** Z^* is:

$$P(-z_{\alpha/2} \le Z^* \le z_{\alpha/2}) = 1-\alpha$$

$$P(-z_{\alpha/2} \le \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le z_{\alpha/2}) = 1-\alpha$$

$$P(-\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le -p \le -\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) = 1-\alpha$$

$$P(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) = 1-\alpha$$

=> The $100(1-\alpha)\%$ large sample C.I. for p is

$$\boxed{[\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]}.$$

CLT => n large usually means $n \ge 30$

special case for the inference on p based on a Bernoulli population. The sample size n is large means

Let $X = \sum_{i=1}^{n} X_i$, large sample means:

$n\hat{p} = X \ge 5$ (*Here X= total # of 'S'), and

$n(1-\hat{p}) = n - X \ge 5$ (*Here n-X= total # of 'F')

Example 2.

During one of the "beer wars" in the early 1980's, a taste test between Schlitz and Budweiser was the focus of a TV commercial. 100 people agreed to drink 2 unmarked mugs and indicate which of the two beers they liked better. 54 chose "Bud". Construct and interpret the corresponding 95% confidence interval for p - the proportion of beer drinkers who prefer Bud to Schlitz.

Solution.

<u>Confidence Interval for one population proportion (p) when the sample size is large</u>

Sample size : n $(n = 100)$

Sample proportion : $\hat{p} = \dfrac{\sum_{i=1}^{n} X_i}{n}$ $\left(\hat{p} = \dfrac{54}{100}\right)$

*** Recall we usually denote $X = \sum_{i=1}^{n} X_i$

"sample is large" means

- For one population mean, $n \geq 30$
- For one population proportion : $X \geq 5$ and $(n - X) \geq 5$

  $(X = 54 \geq 5\ ;\ n - X = 46 \geq 5)$

$n = 100,\ X = 54,$ 95% CI for p

From 95% confidence interval, $1 - \alpha = 0.95,\ \alpha = 0.05,\ \dfrac{\alpha}{2} = 0.025$

$$\hat{p} = \frac{54}{100} = 0.54\ ;\ Z_{0.025} = 1.96$$

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.54)(0.46)}{100}} = 0.049$$

$$Z_{0.025} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \times 0.049 = 0.096$$

$\therefore$ The 95% confidence interval for p is $[0.444, 0.636]$

If $n = 10000\ ;\ \hat{p} = 0.54$,

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.54)(0.46)}{10,000}} = 0.0049$$

$$Z_{0.025} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \times 0.0049 = 0.0096 \approx 0.01$$

$\therefore$ The 95% confidence interval for p is $[0.53, 0.55]$

3. The Exact Confidence Interval for μ when the population is normal & σ^2 is unknown

1. Point estimation : $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
2. $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
3. **Theorem.** Sampling from normal population
   a. $Z \sim N(0,1)$
   b. $W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$
   c. Z and W are independent.

Definition. $T = \frac{Z}{\sqrt{W/(n-1)}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

------ Derivation of CI, normal population, σ^2 is unknown ------

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ is not a pivotal quantity.

$\bar{X} - \mu \sim N(0, \frac{\sigma^2}{n})$ is not a pivotal quantity.

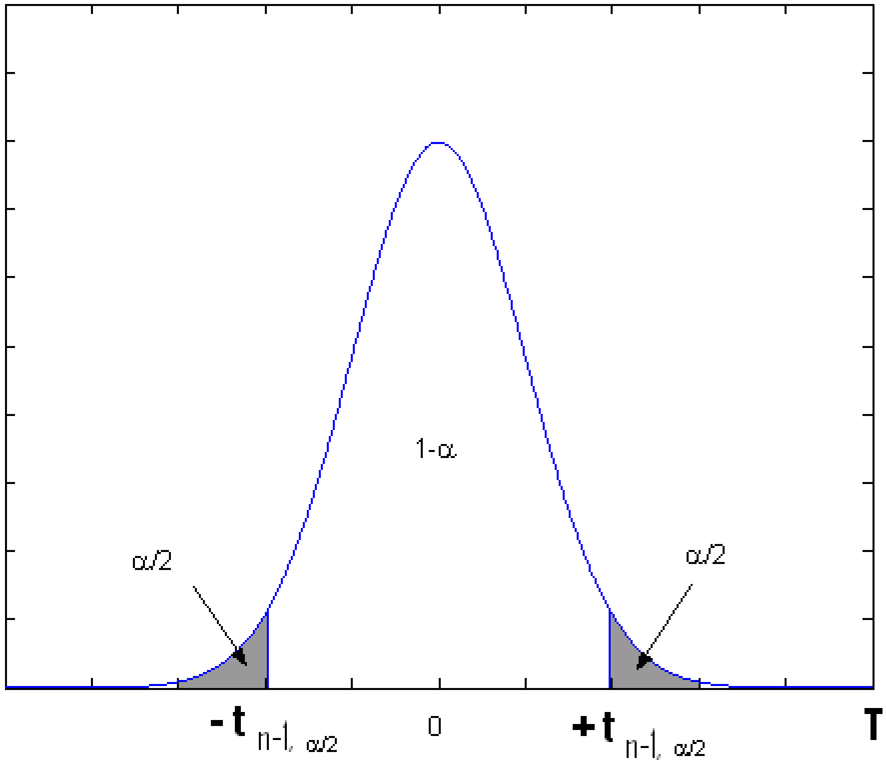
$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$ is not a pivotal quantity.

Remove σ !!!

Therefore $\boxed{T = \frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t_{n-1}}$ is a pivotal quantity.

Now we will use this pivotal quantity to derive the 100(1-α)% confidence interval for μ.

We start by plotting the pdf of the t-distribution with n-1 degrees of freedom as follows:



The above pdf plot corresponds to the following probability statement:

$$P(-t_{n-1,\alpha/2} \le T \le t_{n-1,\alpha/2}) = 1-\alpha$$

$$\Rightarrow P\left(-t_{n-1,\alpha/2} \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le t_{n-1,\alpha/2}\right) = 1-\alpha$$

$$\Rightarrow P\left(-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \le \bar{X}-\mu \le t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

$$\Rightarrow P\left(-\bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \le -\mu \le -\bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

$$\Rightarrow P\left(\bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \ge \mu \ge \bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

$$\Rightarrow P\left(\bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

=> Thus the $100(1-\alpha)\%$ C.I. for μ when σ^2 is unknown is

$$\boxed{[\bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}]}.$$

(*Please note that $t_{n-1,\alpha/2} \ge Z_{\alpha/2}$)

Example 3. In a psychological depth-perception test, a random sample of $n = 14$ airline pilots were asked to judge the distance between 2 markers at the other end of a laboratory. The data (in test) are

2.7, 2.4, 1.9, 2.4, 1.9, 2.3, 2.2, 2.5, 2.3, 1.8, 2.5, 2.0, 2.2, 2.6

Please construct a 95% CI for μ, the average distance.

Solution.

(Note: we can perform the Shapiro-Wilk test to examine whether the sample comes from a normal population or not. This test is not required in our class. Here we simply assume the population is normal. I will always give you such information in the exams.)

CI for μ, small sample, normal population, population variance unknown.

$n = 14,\ \overline{X} = 2.26,\ S = 0.28,\ \alpha = 0.05$

95% CI for μ is $\overline{X} \pm t_{n-1,\alpha/2} \cdot \frac{S}{\sqrt{n}} = 2.26 \pm 2.16 \cdot \frac{0.28}{\sqrt{14}}$

$\therefore [2.10, 2.42]$