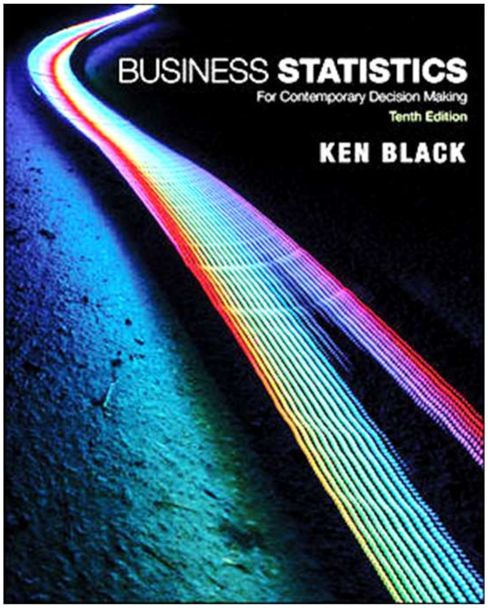
Lecture 12

Dr. Shafayat Abrar

Recap and Problems on Confidence Interval

1 Introduction to Statistics and Business Analytics
2 Visualizing Data with Charts and Graphs
3 Descriptive Statistics
4 Probability
5 Discrete Distributions
6 Continuous Distributions
7 Sampling and Sampling Distributions
8 Statistical Inference: Estimation for Single Populations
9 Statistical Inference: Hypothesis Testing for Single Populations
10 Statistical Inferences About Two Populations
11 Analysis of Variance and Design of Experiments
12 Simple Regression Analysis and Correlation
13 Multiple Regression Analysis
14 Building Multiple Regression Models
15 Time-Series Forecasting and Index Numbers
16 Analysis of Categorical Data
17 Nonparametric Statistics
18 Statistical Quality Control
19 Decision Analysis



An estimator is asymptotically normal if

$$\frac{\hat{\theta}_n - \theta}{\sigma_{\hat{\theta}_n}} \rightsquigarrow N(0,1) \quad \text{as} \quad n \to \infty$$

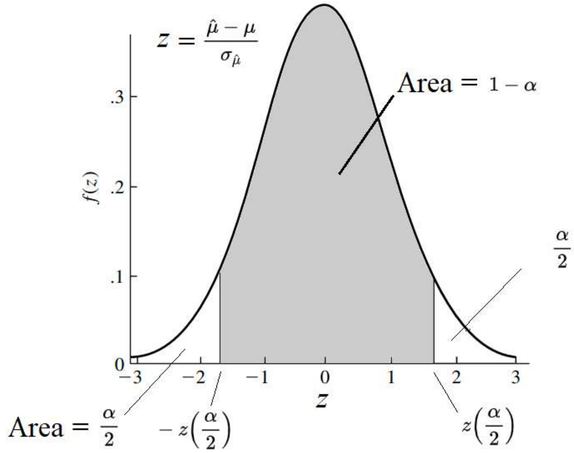
The estimation error in sample mean *scaled* by the standard error is thus asymptotically normal

$$\frac{\hat{\mu}_n - \mu}{\sigma_{\hat{\mu}_n}} = \frac{\tilde{\mu}_n}{\sigma_{\hat{\mu}_n}} \rightsquigarrow N(0,1) \quad \text{as} \quad n \to \infty$$

$$P\left(-z\left(\frac{\alpha}{2}\right) \le \frac{\hat{\mu} - \mu}{\sigma_{\hat{\mu}}} \le z\left(\frac{\alpha}{2}\right)\right) \approx 1 - \alpha$$

$$P\left(\hat{\mu} - z\left(\frac{\alpha}{2}\right)\sigma_{\hat{\mu}} \le \mu \le \hat{\mu} + z\left(\frac{\alpha}{2}\right)\sigma_{\hat{\mu}}\right) \approx 1 - \alpha$$

$$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$$



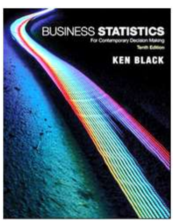
$100(1-\alpha)\%$ CONFIDENCE INTERVAL TO ESTIMATE μ : σ KNOWN

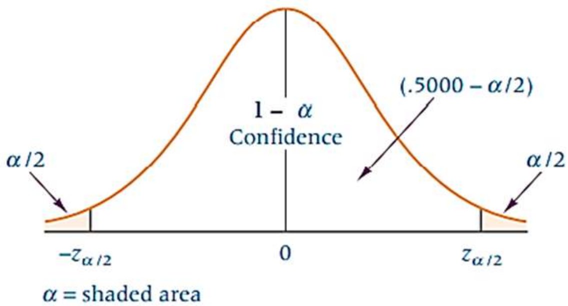
$$\hat{\mu} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

or

$$\hat{\mu} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \hat{\mu} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where
$\alpha =$ the area under the normal curve outside the confidence interval area
$\alpha/2 =$ the area in one end (tail) of the distribution outside the confidence interval





Example 1

A survey was taken of U.S. companies that do business with firms in India. One of the questions on the survey was: Approximately how many years has your company been trading with firms in India? A random sample of 44 responses to this question yielded a mean of 10.455 years. Suppose the population standard deviation for this question is 7.7 years. Using this information, construct a 90% confidence interval for the mean number of years that a company has been trading in India for the population of U.S. companies trading with firms in India.

Here, $n = 44$,
$\widehat{\mu} = \overline{X} = 10.455$,
and $\sigma = 7.7$,
$\mathrm{CI} = 90\%$,
$\alpha = 10\%$,
$\alpha' = \alpha/2 = 5\% = 0.05$
From one-tail table,
we get $z_{\alpha'} = 1.6449$

Critical Values, $P(Z \geq z_\alpha) = \alpha$

α	.10	.05	.025	.01	.005	.001	.0005	.0001
z_α	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	3.2905	3.7190

α	.00009	.00008	.00007	.00006	.00005	.00004	.00003	.00002	.00001
z_α	3.7455	3.7750	3.8082	3.8461	3.8906	3.9444	4.0128	4.1075	4.2649

The z distribution of $\bar{x}$ around μ contains .4500 of the area on each side of μ, or $\frac{1}{2}(90\%)$.
Table yields a z value of 1.645 for the area of .4500 (interpolating between .4495 and .4505).
The confidence interval is

$$\bar{x} - z\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z\frac{\sigma}{\sqrt{n}}$$

$$10.455 - 1.645\frac{7.7}{\sqrt{44}} \leq \mu \leq 10.455 + 1.645\frac{7.7}{\sqrt{44}}$$

$$10.455 - 1.910 \leq \mu \leq 10.455 + 1.910$$

$$8.545 \leq \mu \leq 12.365$$

The analyst is 90% confident that if a census of all U.S. companies trading with firms in India were taken at the time of this survey, the actual population mean number of years a company would have been trading with firms in India would be between 8.545 and 12.365.
The point estimate is 10.455 years.

Example 2

Use the following information to construct the confidence intervals specified to estimate μ.

a. 95% confidence for $\bar{x} = 25$, $\sigma = 3.5$, and $n = 60$

b. 98% confidence for $\bar{x} = 119.6$, $\sigma = 23.89$, and $n = 75$

c. 90% confidence for $\bar{x} = 3.419$, $\sigma = 0.974$, and $n = 32$

d. 80% confidence for $\bar{x} = 56.7$, $\sigma = 12.1$, $N = 500$, and $n = 47$

a. 95% confidence for $\bar{x} = 25$, $\sigma = 3.5$, and $n = 60$

Critical Values, $P(Z \geq z_\alpha) = \alpha$

α	.10	.05	.025	.01	.005	.001	.0005	.0001	
z_α	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	3.2905	3.7190	
α	.00009	.00008	.00007	.00006	.00005	.00004	.00003	.00002	.00001
z_α	3.7455	3.7750	3.8082	3.8461	3.8906	3.9444	4.0128	4.1075	4.2649

a) $\bar{x} = 25$ $\sigma = 3.5$ $n = 60$

95% Confidence $z_{.025} = 1.96$

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}} = 25 \pm 1.96\frac{3.5}{\sqrt{60}} = 25 \pm 0.89 = \mathbf{24.11 \leq \mu \leq 25.89}$$

b. 98% confidence for $\bar{x} = 119.6$, $\sigma = 23.89$, and $n = 75$

Critical Values, $P(Z \geq z_\alpha) = \alpha$

α	.10	.05	.025	.01	.005	.001	.0005	.0001	
z_α	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	3.2905	3.7190	
α	.00009	.00008	.00007	.00006	.00005	.00004	.00003	.00002	.00001
z_α	3.7455	3.7750	3.8082	3.8461	3.8906	3.9444	4.0128	4.1075	4.2649

b) $\bar{x} = 119.6$ $\quad\sigma = 23.89$ $\quad n = 75$

98% Confidence $\quad z_{.01} = 2.33$

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}} = 119.6 \pm 2.33\frac{23.89}{\sqrt{75}} = 119.6 \pm 6.43 = \mathbf{113.17 \leq \mu \leq 126.03}$$

c) $\bar{x} = 3.419$ $\quad\sigma = 0.974$ $\quad n = 32$

90% C.I. $\quad z_{.05} = 1.645$

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}} = 3.419 \pm 1.645\frac{0.974}{\sqrt{32}} = 3.419 \pm .283 = \mathbf{3.136 \leq \mu \leq 3.702}$$

d) $\bar{x} = 56.7$ $\quad\sigma = 12.1$ $\quad N = 500$ $\quad n = 47$

80% C.I. $\quad z_{.10} = 1.28$

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = 56.7 \pm 1.28\frac{12.1}{\sqrt{47}}\sqrt{\frac{500-47}{500-1}} =$$

$$56.7 \pm 2.15 = \mathbf{54.55 \leq \mu \leq 58.85}$$

Example 3

A community health association is interested in estimating the average number of maternity days women stay in the local hospital. A random sample is taken of 36 women who had babies in the hospital during the past year. The following numbers of maternity days each woman was in the hospital are rounded to the nearest day.

3	3	4	3	2	5	3	1	4	3
4	2	3	5	3	2	4	3	2	4
1	6	3	4	3	3	5	2	3	2
3	5	4	3	5	4				

Use these data and a population standard deviation of 1.17 to construct a 98% confidence interval to estimate the average maternity stay in the hospital for all women who have babies in this hospital.

Critical Values, $P(Z \geq z_\alpha) = \alpha$

α	.10	.05	.025	.01	.005	.001	.0005	.0001	
z_α	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	3.2905	3.7190	

α	.00009	.00008	.00007	.00006	.00005	.00004	.00003	.00002	.00001
z_α	3.7455	3.7750	3.8082	3.8461	3.8906	3.9444	4.0128	4.1075	4.2649

$n = 36$ $\quad \bar{x} = 3.306$ $\quad \sigma = 1.17$

98% C.I. $\quad z_{.01} = 2.33$

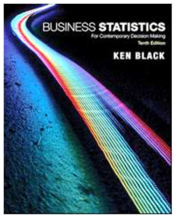
$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}} = 3.306 \pm 2.33\frac{1.17}{\sqrt{36}} = 3.306 \pm .454 = \mathbf{2.852 \leq \mu \leq 3.760}$$

As we briefly discussed in our first-week lectures that if the sample is taken from a finite population, a finite correction factor may be used to increase the accuracy of the solution. In the case of confidence interval estimation, the finite correction factor is used to reduce the width of the interval.

If the sample size is less than 5% of the population, the finite correction factor does not significantly alter the solution. The previous formula is modified to include the finite correction factor, the result is:

CONFIDENCE INTERVAL TO ESTIMATE μ USING THE FINITE CORRECTION FACTOR

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \le \mu \le \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$



Example 4

A study is conducted in a company that employs 800 engineers. A random sample of 50 of these engineers reveals that the average sample age is 34.30 years. Historically, the population standard deviation of the age of the company's engineers is approximately 8 years. Construct a 98% confidence interval to estimate the average age of all the engineers in this company.

This problem has a finite population. The sample size, 50, is greater than 5% of the population, so the finite correction factor may be helpful. In this case $N = 800$, $n = 50$, $\bar{x} = 34.30$, and $\sigma = 8$. The z value for a 98% confidence interval is 2.33 (.98 divided into two equal parts yields .4900; the z value is obtained from Table A.5 by using .4900). Substituting into formula 8.2 and solving for the confidence interval gives

$$34.30 - 2.33\frac{8}{\sqrt{50}}\sqrt{\frac{750}{799}} \le \mu \le 34.30 + 2.33\frac{8}{\sqrt{50}}\sqrt{\frac{750}{799}}$$

$$34.30 - 2.55 \le \mu \le 34.30 + 2.55$$

$$31.75 \le \mu \le 36.85$$

This problem has a finite population. The sample size, 50, is greater than 5% of the population, so the finite correction factor may be helpful. In this case $N = 800$, $n = 50$, $\bar{x} = 34.30$, and $\sigma = 8$. The z value for a 98% confidence interval is 2.33 (.98 divided into two equal parts yields .4900; the z value is obtained from Table A.5 by using .4900). Substituting into formula 8.2 and solving for the confidence interval gives

$$34.30 - 2.33\frac{8}{\sqrt{50}}\sqrt{\frac{750}{799}} \le \mu \le 34.30 + 2.33\frac{8}{\sqrt{50}}\sqrt{\frac{750}{799}}$$

$$34.30 - 2.55 \le \mu \le 34.30 + 2.55$$

$$31.75 \le \mu \le 36.85$$

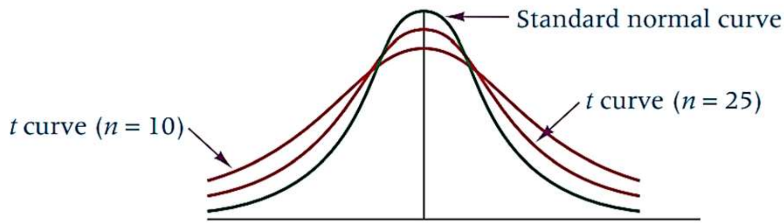
Without the finite correction factor, the result would have been

$$34.30 - 2.64 \le \mu \le 34.30 + 2.64$$
$$31.66 \le \mu \le 36.94$$

The finite correction factor takes into account the fact that the population is only 800 instead of being infinitely large. The sample, $n = 50$, is a greater proportion of the 800 than it would be of a larger population, and thus the width of the confidence interval is reduced when using the finite correction factor.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$



These are called the **sample mean** and the (**unbiased**) **sample variance**, respectively.

COROLLARY **B**

Let $\overline{X}$ and S^2 be as given at the beginning of this section. Then

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

We have found earlier that the maximum likelihood estimates of μ and σ^2 from an i.i.d. normal sample are

$$\widehat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

A confidence interval for μ is based on the fact that

$$\frac{\bar{X}-\mu}{\sqrt{S^2/n}} \sim t_{n-1} \qquad\qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

where t_{n-1} denotes the t distribution with $n-1$ degrees of freedom.

The **degrees of freedom** for the t statistic is $n-1$. The term degrees of freedom refers to the **number of independent observations** for a source of variation minus the number of **independent parameters** estimated in computing the variation.

In our case, one independent parameter, the population mean, μ, is being estimated by $\bar{x}$ in computing S. Thus, the degrees of freedom formula is n independent observations minus one independent parameter being estimated $(n-1)$. Because the degrees of freedom are computed differently for various t formulas, a degrees of freedom formula is given along with each t formula in the text.

Let $t_{n-1}(\alpha/2)$ denote that point beyond which the t distribution with $n-1$ degrees of freedom has probability $\alpha/2$.

Since the t distribution is symmetric about 0 , the probability to the left of $-t_{n-1}(\alpha/2)$ is also $\alpha/2$. Then, by definition,

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

The inequality can be manipulated to yield

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

According to this equation, the probability that μ lies in the interval

$$\bar{X} \pm St_{n-1}(\alpha/2)/\sqrt{n}$$

is $1-\alpha$.

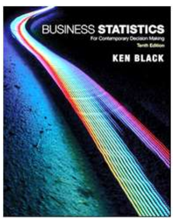
Note that this interval is random: The center is at the random point $\bar{X}$

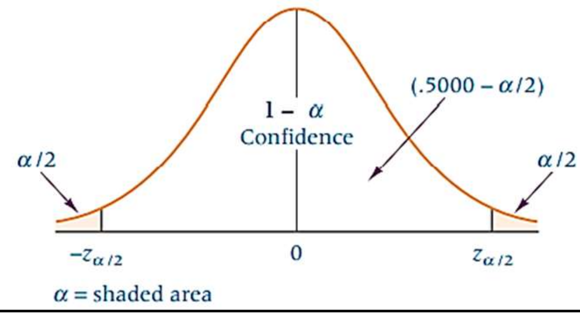
CONFIDENCE INTERVAL TO ESTIMATE μ: POPULATION STANDARD DEVIATION UNKNOWN AND THE POPULATION NORMALLY DISTRIBUTED

$$\bar{x} \pm t_{\alpha/2,\, n-1}\frac{s}{\sqrt{n}}$$

$$\bar{x} - t_{\alpha/2,\, n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2,\, n-1}\frac{s}{\sqrt{n}}$$

$$\mathrm{df} = n-1$$





Example 5

The owner of a large equipment rental company wants to make a rather quick estimate of the average number of days a piece of ditchdigging equipment is rented out per person per time. The company has records of all rentals, but the amount of time required to conduct an audit of *all* accounts would be prohibitive. The owner decides to take a random sample of rental invoices. Fourteen different rentals of ditchdiggers are selected randomly from the files, yielding the following data. She uses these data to construct a 99% confidence interval to estimate the average number of days that a ditchdigger is rented and assumes that the number of days per rental is normally distributed in the population.

3 1 3 2 5 1 2 1 4 2 1 3 1 1

```
>> mean([3 1 3 2 5 1 2 1 4 2 1 3 1 1])

ans =

    2.1429

>> sqrt(var([3 1 3 2 5 1 2 1 4 2 1 3 1 1],0))

ans =

    1.2924

>> sqrt(var([3 1 3 2 5 1 2 1 4 2 1 3 1 1],1))

ans =

    1.2454
```

S^2 **(Unbiased; it divides by N-1)**

$\widehat{\sigma^2}$ **(Biased; it divides by N)**

The sample mean is 2.14 and the sample standard deviation is 1.29.

The sample mean is 2.14 and the sample standard deviation is 1.29.

As $n = 14$, the $df = 13$.
The 99% level of confidence results in $\alpha/2 = .005$ area in each tail of the distribution.

The table t value is

$$t_{0.005,13} = 3.012$$

ν	.20	.10	.05	.025	.01	α .005
1	1.3764	3.0777	6.3138	12.7062	31.8205	63.6567
2	1.0607	1.8856	2.9200	4.3027	6.9646	9.9248
3	.9785	1.6377	2.3534	3.1824	4.5407	5.8409
4	.9410	1.5332	2.1318	2.7764	3.7469	4.6041
5	.9195	1.4759	2.0150	2.5706	3.3649	4.0321
6	.9057	1.4398	1.9432	2.4469	3.1427	3.7074
7	.8960	1.4149	1.8946	2.3646	2.9980	3.4995
8	.8889	1.3968	1.8595	2.3060	2.8965	3.3554
9	.8834	1.3830	1.8331	2.2622	2.8214	3.2498
10	.8791	1.3722	1.8125	2.2281	2.7638	3.1693
11	.8755	1.3634	1.7959	2.2010	2.7181	3.1058
12	.8726	1.3562	1.7823	2.1788	2.6810	3.0545
13	.8702	1.3502	1.7709	2.1604	2.6503	3.0123
14	.8681	1.3450	1.7613	2.1448	2.6245	2.9768

$$\bar{X} \pm t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}}$$

$$2.14 \pm 3.012\frac{1.29}{\sqrt{14}} = 2.14 \pm 1.04$$

$$1.10 \leq \mu \leq 3.18$$

The point estimate of the average length of time per rental is 2.14 days, with an error of ± 1.04. With a 99% level of confidence, the company's owner can estimate that the average length of time per rental is between 1.10 and 3.18 days. Combining this figure with variables such as frequency of rentals per year can help the owner estimate potential profit or loss per year for such a piece of equipment.

Example 6

Suppose the following data are selected randomly from a population of normally distributed values.

40	51	43	48	44	57	54
39	42	48	45	39	43	

Construct a 95% confidence interval to estimate the population mean.

```
>> d=[40   51   43   48   44   57   54  39   42   48   45   39   43]

d =

    40    51    43    48    44    57    54    39    42    48    45    39    43

>> [mean(d)  sqrt(var(d,0))  sqrt(var(d,1))]

ans =

   45.6154    5.6941    5.4707
```

$n = 13 \quad \bar{X} = 45.62 \quad S = 5.694 \quad \text{df } = 13 - 1 = 12$

95% Confidence Interval and $\alpha/2 = .025$

$$t_{0.025,12} = 2.179$$

$$\begin{aligned}\bar{x} \pm t\frac{s}{\sqrt{n}} &= 45.62 \pm 2.179\frac{5.694}{\sqrt{13}} \\ &= 45.62 \pm 3.44 \\ &= \mathbf{42.18 \le \mu \le 49.06}\end{aligned}$$

ν	.20	.10	.05	.025
1	1.3764	3.0777	6.3138	12.7062
2	1.0607	1.8856	2.9200	4.3027
3	.9785	1.6377	2.3534	3.1824
4	.9410	1.5332	2.1318	2.7764
5	.9195	1.4759	2.0150	2.5706
6	.9057	1.4398	1.9432	2.4469
7	.8960	1.4149	1.8946	2.3646
8	.8889	1.3968	1.8595	2.3060
9	.8834	1.3830	1.8331	2.2622
10	.8791	1.3722	1.8125	2.2281
11	.8755	1.3634	1.7959	2.2010
12	.8726	1.3562	1.7823	2.1788

Example 7

A valve manufacturer produces a butterfly valve composed of two semicircular plates on a common spindle that is used to permit flow in one direction only. The semicircular plates are supplied by a vendor with specifications that the plates be 2.37 millimeters thick and have a tensile strength of five pounds per millimeter. A random sample of 20 such plates is taken. Electronic calipers are used to measure the thickness of each plate; the measurements are given here. Assuming that the thicknesses of such plates are normally distributed, use the data to construct a 95% level of confidence for the population mean thickness of these plates. What is the point estimate? How much is the error of the interval?

2.4066	2.4579	2.6724	2.1228	2.3238
2.1328	2.0665	2.2738	2.2055	2.5267
2.5937	2.1994	2.5392	2.4359	2.2146
2.1933	2.4575	2.7956	2.3353	2.2699

```
>> d=[2.4066 2.4579 2.6724 2.1228 2.3238
2.1328 2.0665 2.2738 2.2055 2.5267
2.5937 2.1994 2.5392 2.4359 2.2146
2.1933 2.4575 2.7956 2.3353 2.2699]; d=d(:);
>> [mean(d) sqrt(var(d,0)) sqrt(var(d,1))]

ans =

    2.3612    0.1972    0.1922
```

$n = 20 \quad \text{df} = 19 \quad 95\%\text{CI} \quad t_{0.025,19} = 2.093(\text{table})$

$\bar{X} = 2.36116 \quad S = 0.19721$

$$2.36116 \pm 2.093\frac{0.1972}{\sqrt{20}} = 2.36116 \pm 0.0923$$

$$= \mathbf{2.26886 \le \mu \le 2.45346}$$

ν	.20	.10	.05	.025
1	1.3764	3.0777	6.3138	12.7062
2	1.0607	1.8856	2.9200	4.3027
3	.9785	1.6377	2.3534	3.1824
4	.9410	1.5332	2.1318	2.7764
19	.8610	1.3277	1.7291	$\boxed{2.0930}$

Point Estimate = $\mathbf{2.36116}$

Error = $\mathbf{0.0923}$ for 95 % CI