STATISTICS FOR DATA SCIENCE

Confidence Intervals for Small Samples

Prof. Uma D

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Confidence Intervals for Small Samples

D. Uma

Topics to be covered...

- Confidence Intervals for population mean of small samples
- Student's t Distribution
- Confidence Intervals using t Distribution
- Student's t Distribution Is Appropriate?
- One-Sided CI for Small Samples

$n = 15 \Rightarrow$ small sample

t-distribution

When population SD (σ) is given with $n=15$

Normal distribution

σ known $\Rightarrow$ z-distribution

Confidence Intervals

- If the sample size is small, standard deviation (s) of the sample may not be close to σ (population standard deviation). Hence $\bar{X}$ (sample_mean) may not be approximately normal.

$\bar{x} \nsim N$

- However, if the population from which the sample is drawn is known to be approximately normal (can be confirmed using normal probability plot).

$\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim$

sample

Pop $\sim N$

Confidence Intervals

- It turns out that we can still use the quantity.
- $(\bar{X} - \mu) / (s/\sqrt{n})$,

  but since s is not necessarily close to σ , the quantity will not have a normal distribution.
- Instead it has Student's t distribution with n − 1 degrees of freedom, denoted as t_{n-1} .

Large sample $Z_{\alpha/2}$ (2-sided)

Z_{α} (1-sided)

$df = n-1$.

sample size n

$n = 7$

1st ⇒ 7

2nd ⇒ 6

3rd = 5

i = n-1

$df = n-1$

Small sample $t_{n-1, \alpha/2}$ (2-sided)

$t_{n-1, \alpha}$ (1-sided)

t - Distribution

- The t distribution is a theoretical probability distribution.
- It is symmetrical, bell-shaped, and similar to the standard normal curve.
- It differs from the standard normal curve, however, in that it has an additional parameter, called **degrees of freedom**, which changes its shape.

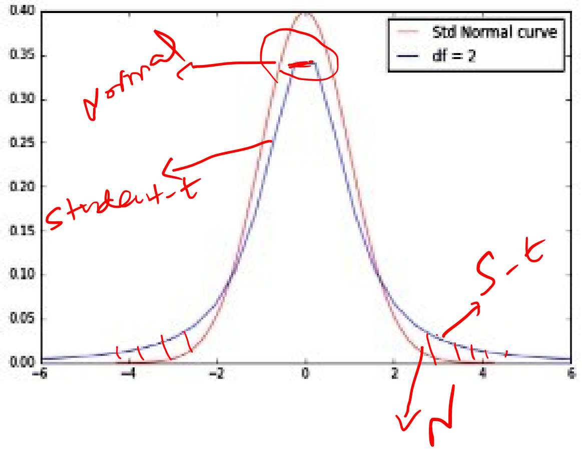
df = sample size – 1

- Setting the value of df defines a particular member of the family of t distributions. (df > 0 => Sample Size > 1)

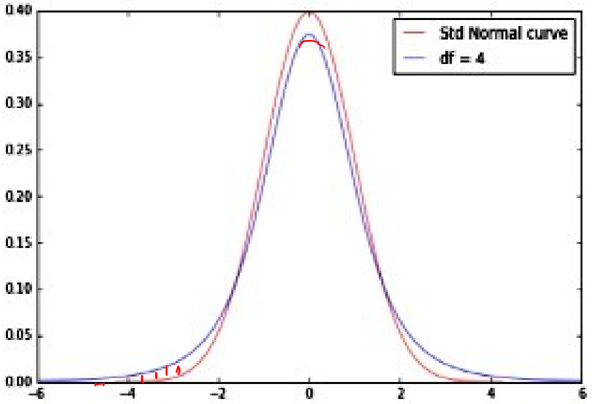
STATISTICS FOR DATA SCIENCE

Students t Distribution

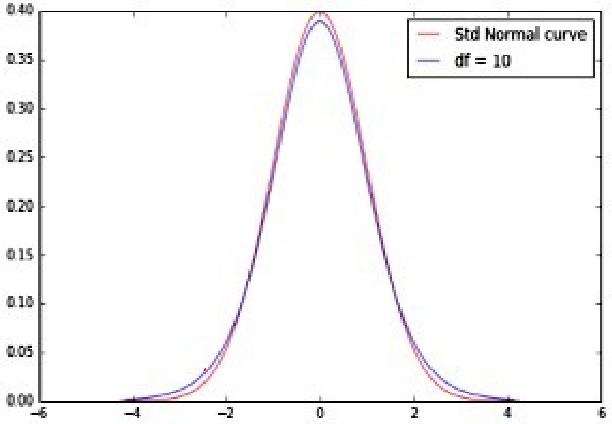
1) df = 2



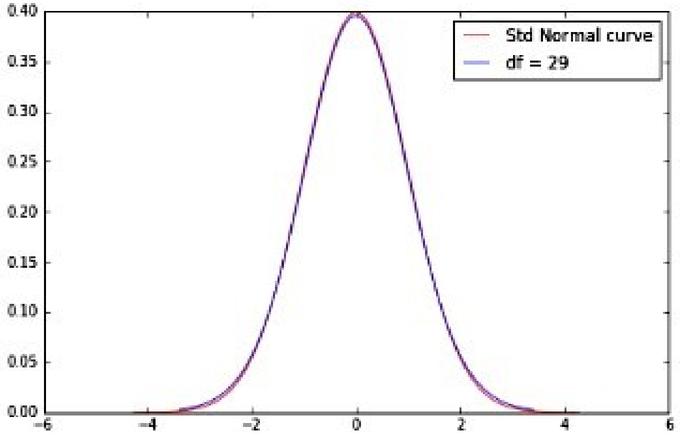
2) df =4



3) df = 10

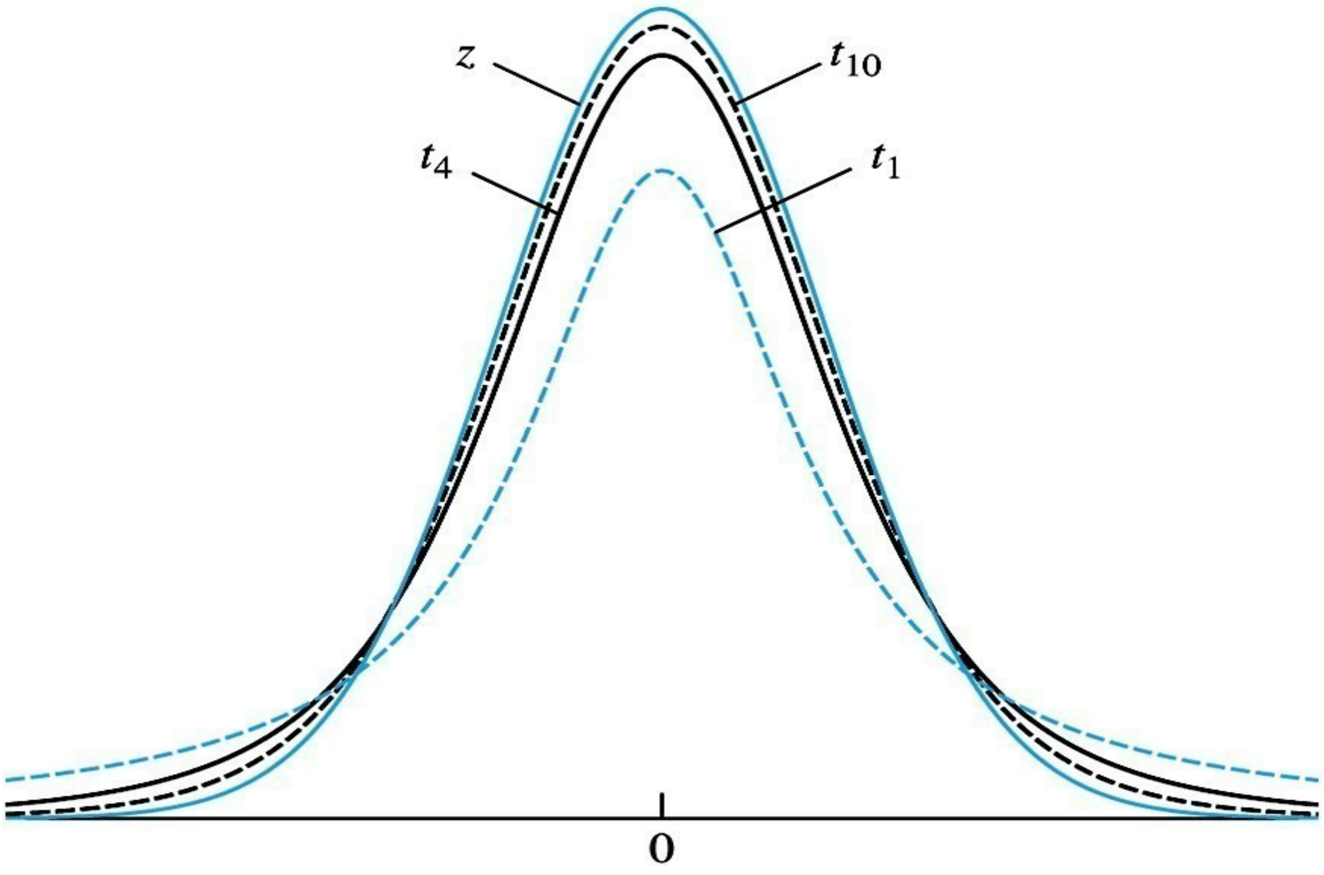


4) df = 30



PDF for Students t curve

Note that the smaller the distribution function, the flatter the shape of the distribution, resulting in greater area in the tails of the distribution.



Relationship to the normal curve

- As the df increase, the t distribution approaches the standard normal distribution (μ=0.0, σ=1.0).
- The standard normal curve is a special case of the t distribution when df= infinity.
- For practical purposes, the t distribution approaches the standard normal distribution relatively quickly, such that when df=30 the two are almost identical.

Using t table

- **We use t table to find probabilites associated with t distribution.**
- **Row headings** – denotes degree of freedom
- **Column headings** – denotes the area to the right(probabilities)
- The value in particular row and column specifies the t-score where,

value

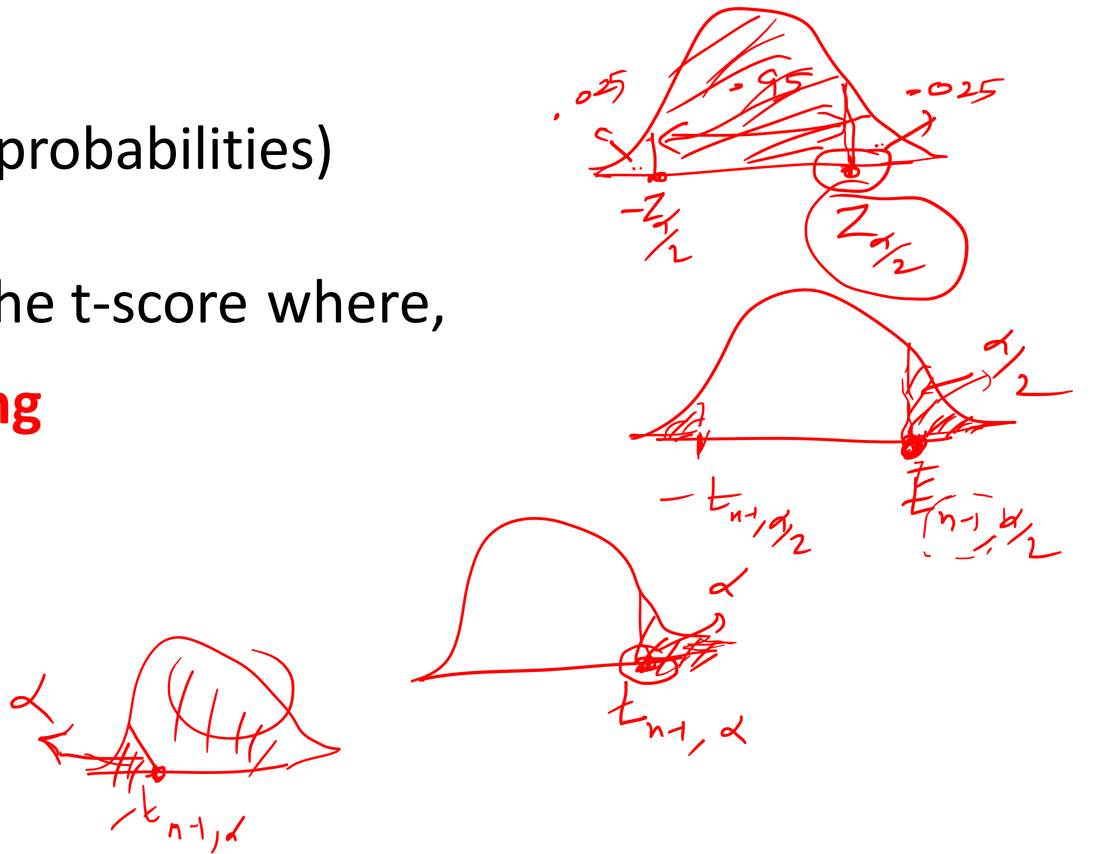
P(t > t-tscore) = col_heading

Student t table

Area

	0.1	0.025	.05
df 1		t value	
2			
3			

.05 .025 .01



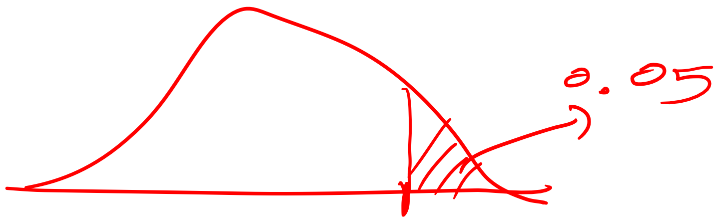
Examples

1) A random sample of size 10 is drawn from a normal distribution with mean 4.

$n = 10$ $\quad df = n-1 = 9$

a) Find P(t >1.833)

$P(t > t\text{-score}) = 0.05$



1.833 $\quad t_9$

$1.383 < 1.5 < 1.833$

$P(t > 1.833) = 0.05$

$P(t > 1.383) = 0.1$

$P(t$

b) Find P(t > 1.5)

$P(t > 1.5) = 0.05 < P(t > 1.5) < 0.1$

Solution

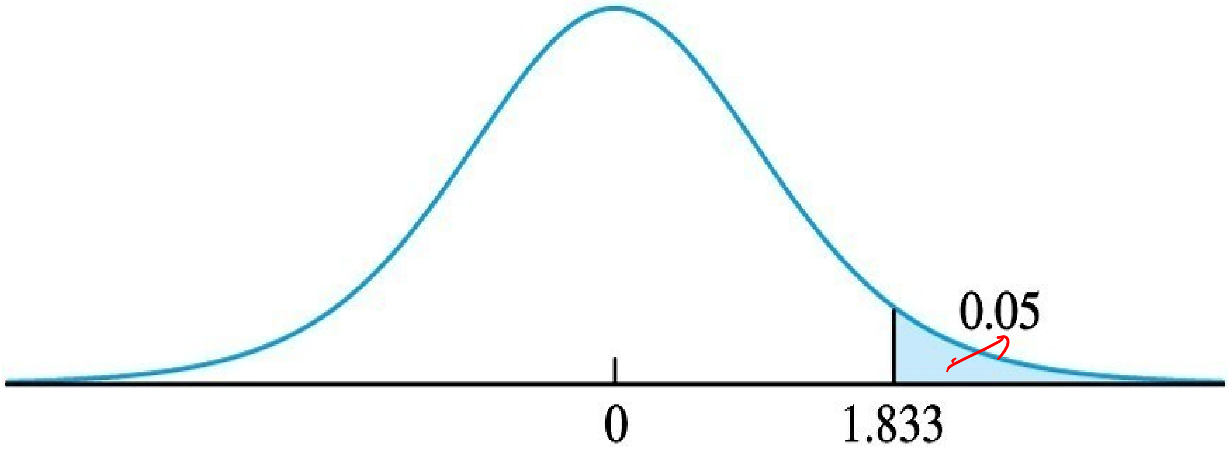
a) Find P(t >1.833)

df = 9 (row_heading)

t-score = 1.833

corresponding col_heading = 0.05

P(t >1.833) = 0.05



Solution

b) Find P(t > 1.5)

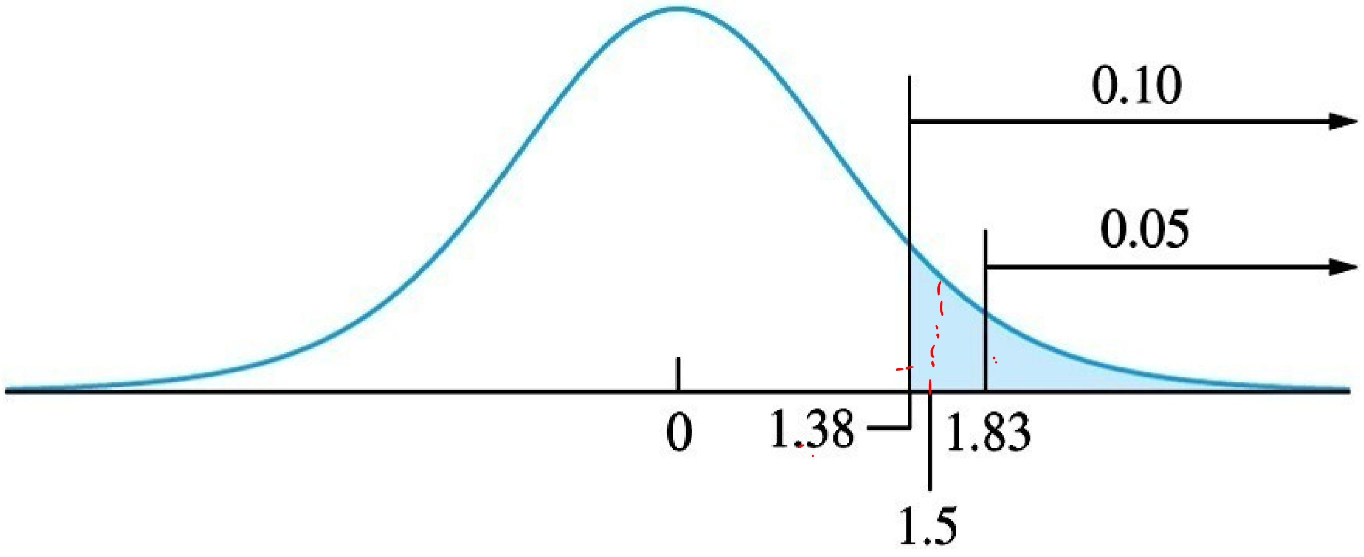
df = 9 (row_heading)

t-score = 1.5 [does not correspond to any of the values in that row]

but we do have t-scores 1.383, 1.833 corresponding to upper tail probabilties 0.10 and 0.05 respectively. That is,

P(t >1.383) = 0.10 and P(t >1.833) = 0.05

Since 1.383 < 1 .5 < 1.833 => 0.05 < P(t > 1.5) < 0.10



Student's t Distribution is Appropriate when

- Sample size is small (**$n < 30$**)
- Sample comes from a population that is **approximately normal**.
- In many cases, we must examine the sample for normality, by constructing a box plot or normal probability plot.

Sample   Approx. Normaly

- Unfortunately, when the sample size is small, departures from normality may be hard to detect.

- If these plots do not reveal a strong asymmetry or any outliers, then in most cases the Student's *t* distribution will be reliable.

Confidence Interval for Small Samples using t distribution:

For μ: (Two-Sided)

$(1-\alpha)*100\%$ CI is given by

$$\bar{X} \pm \underbrace{\boxed{\underbrace{\boxed{t_{n-1,\alpha/2}}}_{\text{t-multiplier}} * \boxed{\frac{s}{\sqrt{n}}}}}_{\text{MOE}}$$

Standard Error

s - Sample SD

$\frac{s}{\sqrt{n}} \Rightarrow$ Standard Error

For μ: (one-sided)

UB: $\left(-\infty, \underbrace{\bar{X} + t_{n-1,\alpha} * \frac{s}{\sqrt{n}}}_{UB}\right) \Rightarrow$ Lower Interval

LB: $\left(\bar{X} - t_{n-1,\alpha} * \frac{s}{\sqrt{n}}, +\infty\right) \Rightarrow$ Upper Interval

One-Sided Confidence Intervals for small samples

We can generate a (1 – a) 100% Upper Confidence bound for μ as:

$$X\_bar + \mathbf{t}_{n-1,\alpha} * s/sqrt(n)$$

We can generate a (1 – a) 100% Lower Confidence bound for μ as:

$$X\_bar - \mathbf{t}_{n-1,\alpha} * s/sqrt(n)$$

Example1

Find the value of $t_{n-1,\ \alpha/2}$ needed to construct a two-sided confidence interval of the given level with the given sample size:

a) 90% with sample size 12

b) 95% with sample size 7

$t_{n-1, \alpha/2}$

$n = 12$

$df = 11$

$CL = 90\%$

$1-\alpha = .90$

$\alpha = .10$

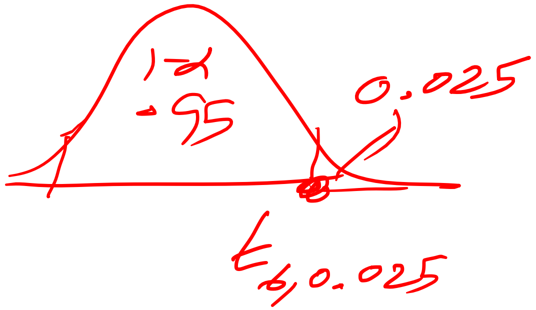
$\alpha/2 = .05$

a) $t_{df,\ \alpha/2} = \boxed{t_{11,\ .05} = 1.796}$

$1-\alpha = .95$

$\alpha = .05$

$\alpha/2 = .025$



b) $n = 7$; $df = 6$

$t_{6,\ 0.025} = \boxed{2.447}$

Solution

a) **90% with sample size 12**

df = 11

alpha = 0.10 => alpha/2 = 0.05

=> in t table : row_heading = 11, col_heading = 0.05 => $\mathbf{t}_{11,\,0.05}$ = 1.796

b) **95% with sample size 7**

df = 6

alpha = 0.05 => alpha/2 = 0.025

=> in t table : row_heading = 6, col_heading = 0.025 => $\mathbf{t}_{6,\,0.025}$ = 2.447

Example2

Find the level of two-sided confidence interval that is based on the given value of t n -1 , α/2 and the given sample size:

$t_{n-1,\alpha/2}$

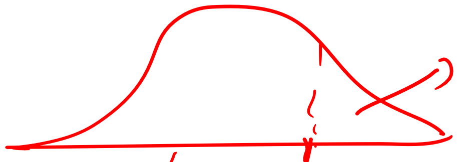
a) t = 5.841, sample size = 4

$t = 5.841 \quad df = 3 \quad \alpha/2 =$

$\alpha/2 = 0.005$

$\Rightarrow \alpha = 2(0.005)$

$\alpha = 0.01$



$t_{3,\alpha/2} = 5.841$

b) t = 1.746, sample size = 17

$CL = (1-\alpha) * 100\%$

$= (1-0.01) * 100\%$

$= (0.99) \cancel{*100}\%$

$CL = 99\%$

CL= ? 90%

CL = ?

Example2

Find the level of two-sided confidence interval that is based on the given value of t n -1 , α/2 and the given sample size:

a) t = 5.841, sample size = 4

CL = ?

df = 4 - 1 = 3

$\alpha/2 = 0.005 \Rightarrow \alpha = 2(0.005)$

$\alpha = 0.01$

$\therefore 1 - \alpha = 0.99 \Rightarrow \therefore CL = 99\%$

b) t = 1.746, sample size = 17

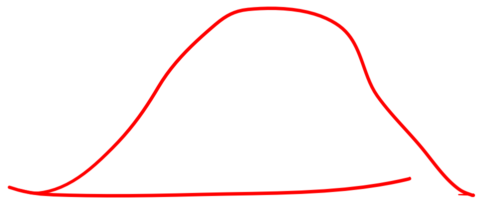
df = 16

$\alpha/2 = .05$

$\Rightarrow \alpha = 2(.05) = 0.1$

$\therefore 1 - \alpha = 0.9$

$\therefore CL = 90\%$



Example2

Following represents the measurements of the nominal shear strength (in kN) for a sample of 15 pre-stressed concrete beams:

IQR

580 400 428 825 850 875 920 550 575 750 636 360 590 735 950

n=15 No outliers

a) Is it appropriate to use the Student's t statistic to construct a 99% confidence interval for the mean shear strength? CI for μ

CL=99%

b) b) If so, construct the confidence interval. If not, explain why not.

Example2

Following represents the measurements of the nominal shear strength (in kN) for a sample of 15 pre-stressed concrete beams:

580 400 428 825 850 875 920 550 575 750 636 360 590 735 950

a) **Is it appropriate to use the Student's t statistic to construct a 99% confidence interval for the mean shear strength?**

Yes.

Since there are no outliers in the data set, Student's t statistic can be used to construct 99% CI.

Example2

Following represents the measurements of the nominal shear strength (in kN) for a sample of 15 pre-stressed concrete beams:

580 400 428 825 850 875 920 550 575 750 636 360 590 735 950 $\Rightarrow \bar{x} = 668.27$

$\Rightarrow s = 192.087$

b) If so, construct the confidence interval. If not, explain why not.

$n = 15$ $\quad CL = 99\%$

CI for μ: $\bar{x} \pm t_{n-1, \alpha/2} * \frac{s}{\sqrt{n}}$

$1 - \alpha = .99$

$\alpha = .01$

$\alpha/2 = .005$

$668.27 \pm 2.977 * \frac{192.087}{\sqrt{15}}$

$t_{14, .005} = 2.977$

$668.27 \pm$

$(520.62, 815.92)$

Example2

Following represents the measurements of the nominal shear strength (in kN) for a sample of 15 pre-stressed concrete beams:

580 400 428 825 850 875 920 550 575 750 636 360 590 735 950 ⇒ Dataset

b) If so, construct the confidence interval. If not, explain why not.

CL = 99%

$1-\alpha = 0.99$

$\alpha = 0.01$

$\alpha/2 = 0.005$

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

$n = 15$

$df = 14$

$\bar{x} \pm t_{14, 0.005} * \frac{s}{\sqrt{15}}$

sample mean $\bar{x} = 668.27$

Sample S.D $s = 192.089$

∴ 99% CI for μ is $668.27 \pm 2.977 * \frac{192.089}{\sqrt{15}}$

$(520.62, 815.92)$

Example 3

PES UNIVERSITY ONLINE

The table below shows data on a subsample of n=10 participants in the 7th examination of the Framingham Offspring Study.

Characteristic	n	Sample Mean	Standard Deviation (s)
Systolic Blood Pressure	10	121.2	11.1
Diastolic Blood Pressure	10	71.3	7.2
Total Serum Cholesterol	10	202.3	37.7
Weight	10	176.0	33.0
Height	10	67.175	4.205
Body Mass Index	10	27.26	3.10

Suppose we compute a 95% confidence interval for the true systolic blood pressure using data in the subsample.

Source: sphweb.bumc.bu.edu

Example 3 Solution

Given CL = 95% $\quad n = 10 \quad df = 9$

SBP $\quad \bar{x} = 121.2 \quad 11.1$

To find CI for true Systolic Blood Pressure (μ)

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

$$121.2 \pm 2.262 * \frac{11.1}{\sqrt{10}}$$

$$\underset{\bar{x}}{121.2} \pm 7.94$$

$$(113.3, 129.1)$$

Source: sphweb.bumc.bu.edu

Interpretation

Interpretation: Based on this sample of size n=10, our best estimate of the true mean systolic blood pressure in the population is 121.2.

μ

Based on this sample, we are 95% confident that the true systolic blood pressure in the population is between 113.3 and 129.1.

μ

Note that the margin of error is larger here primarily due to the small sample size.

Use z,Not t, if σ is known

If it is known that the sample indeed was drawn from a **normal population**, also the **standard deviation of the population is known**, use z not t distribution to find out the confidence interval irrespective of the sample size.

Summary

Let $X_1, \ldots, X_n$ be a random sample (of any size) from a *normal* population with mean μ. If the standard deviation σ is known, then a level $100(1-\alpha)\%$ confidence interval for μ is

$$\overline{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{5.12}$$

σ

THANK YOU

D. Uma
Computer Science and Engineering

umaprabha@pes.edu

+91 99 7251 5335