STATISTICS FOR DATA SCIENCE

Confidence Intervals for Small Samples

D. Uma

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

- If the sample size is small, standard deviation (s) of the sample may not be close to σ (population standard deviation). Hence $\overline{X}$ (sample_mean) may not be aproximately normal.
- However, if the population from which the sample is drawn is known to be approximately normal (can be confirmed using normal probability plot).

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

  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} .

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.

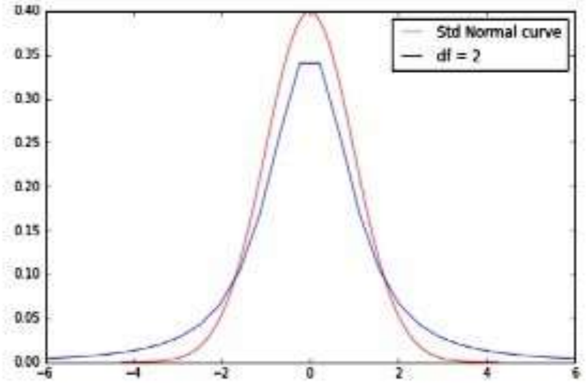
  df = sample size – 1

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

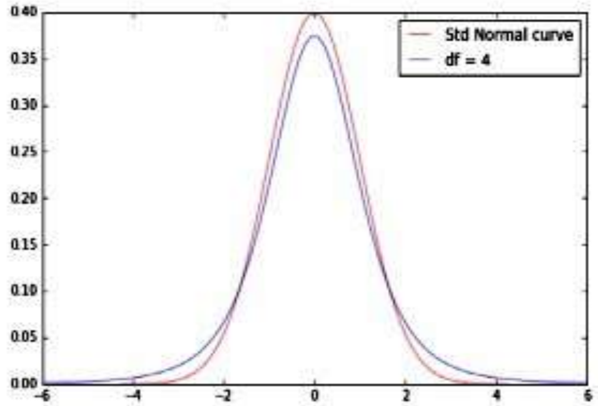
STATISTICS FOR DATA SCIENCE

Students t Distribution

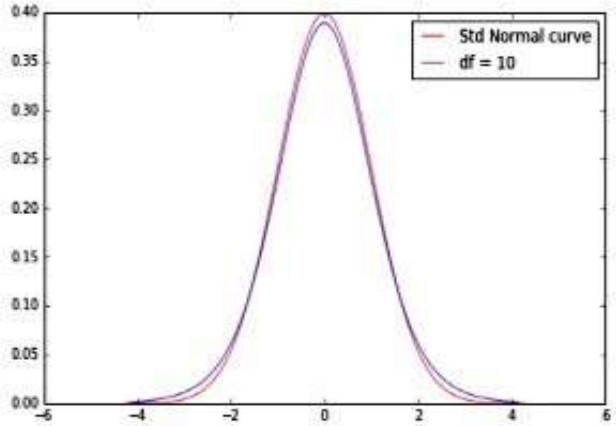
1) df = 2



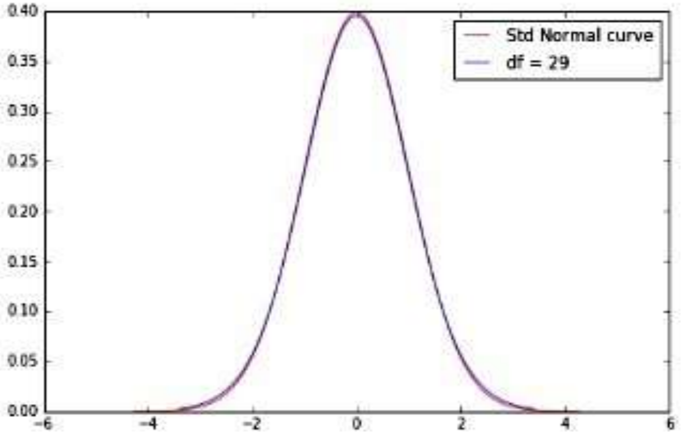
2) df =4



3) df = 10



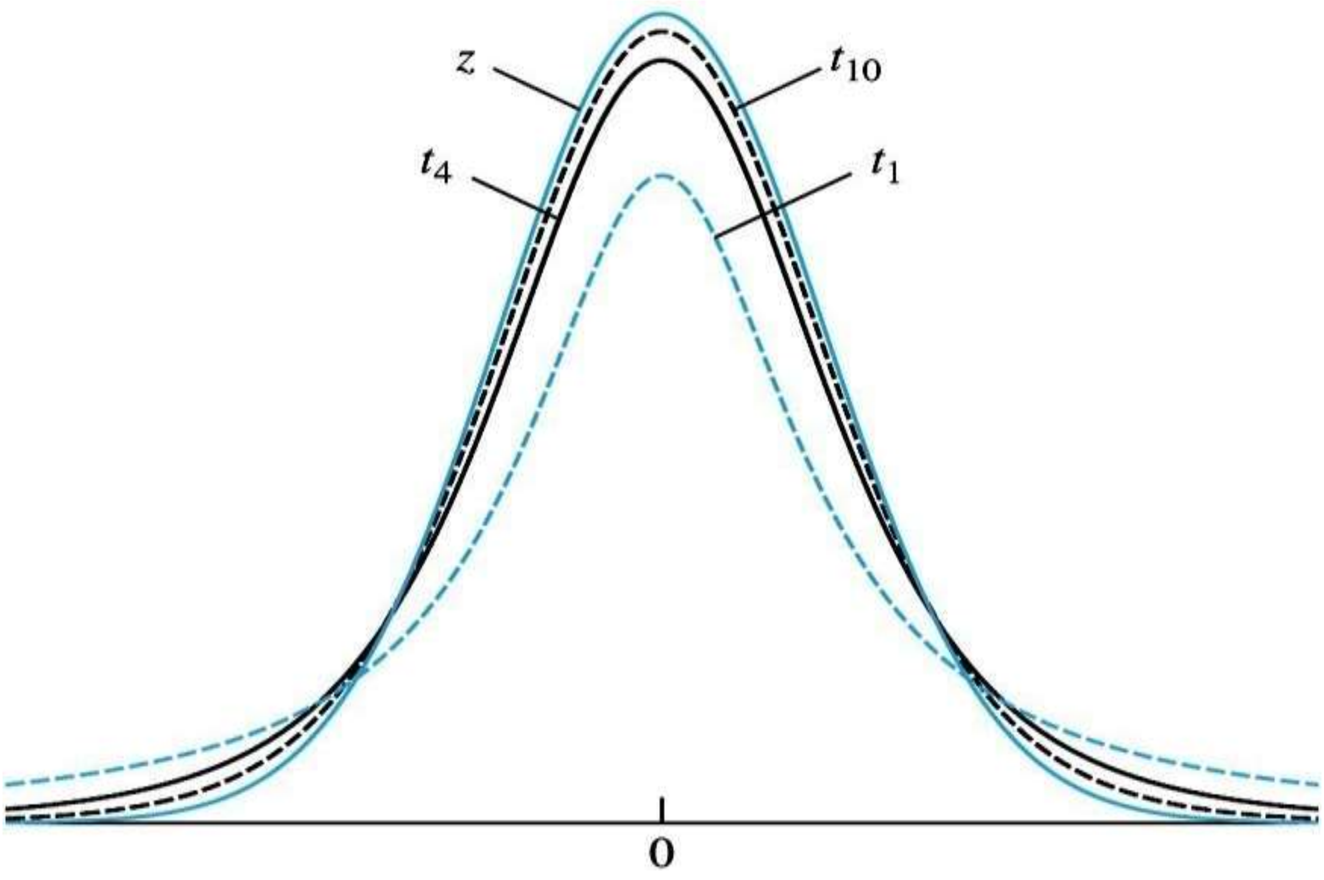
4) df = 30



STATISTICS FOR DATA SCIENCE

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,

$$P(t > t\text{-}tscore) = col\_heading$$

Examples

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

a) Find $P(t > 1.833)$

b) Find $P(t > 1.5)$

Solution

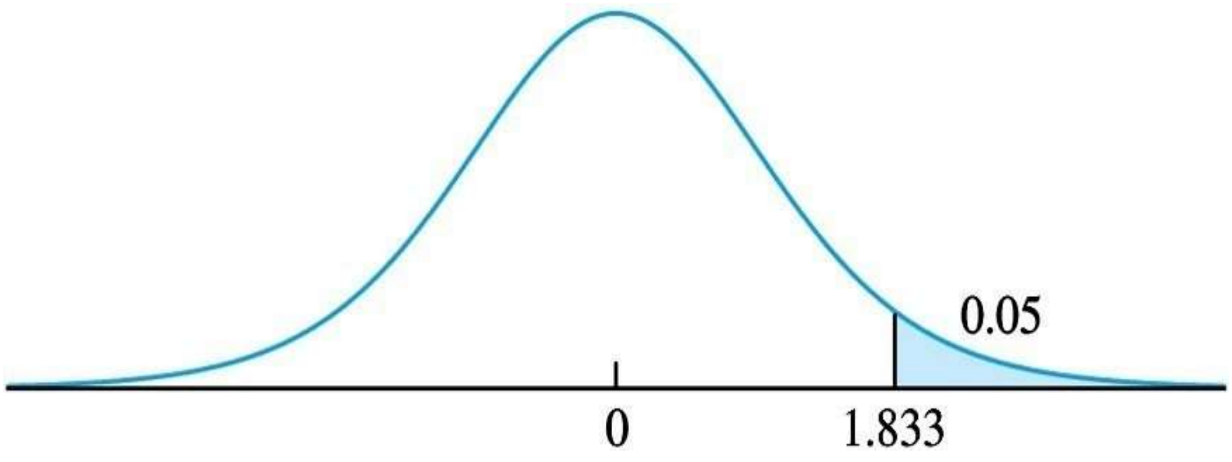
a) Find P(t >1.833)

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)

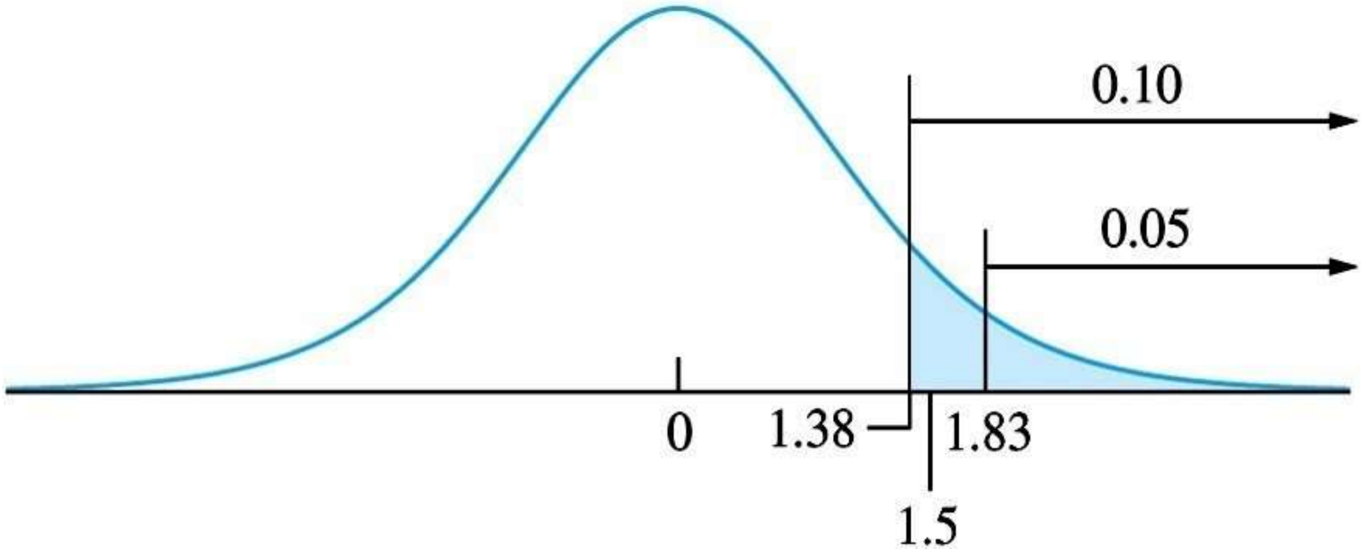
df = 9 (row_heading)

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



Confidence Interval for Small Samples using t distribution:

The quantity,

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

has a t distribution with n – 1 degrees of freedom.

We can generate a (1 – α) 100% Confidence Interval for μ as

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

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.
- 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.

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

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

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

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

THANK YOU

D. Uma

Computer Science and Engineering

umaprabha@pes.edu
+91 99 7251 5335