PES UNIVERSITY ONLINE

STATISTICS FOR DATA SCIENCE

Factors affecting margin of errors

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Factors affecting margin of errors

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Confidence Intervals

Suppose we want to estimate an actual population mean μ. As you know, we can only obtain $\bar{x}$, the mean of a sample randomly selected from the population of interest. We can use $\bar{x}$ to find a range of values:

Lower value<population mean μ<Upper value

that we can be really confident contains the population mean μ. The range of values is called a "**confidence interval**."

Example

Should using a hand-held cell phone while driving be illegal?

For example, a newspaper report (ABC News poll, May 16-20, 2001) was concerned whether or not U.S. adults thought using a hand-held cell phone while driving should be illegal. Of the 1,027 U.S. adults randomly selected for participation in the poll, 69% thought that it should be illegal. The reporter claimed that the poll's "**margin of error**" was 3%. Therefore, the confidence interval for the (unknown) population proportion *p* is 69% ± 3%. That is, we can be really confident that between 66% and 72% of all U.S. adults think using a hand-held cell phone while driving a car should be illegal.

General form of Confidence Intervals

Sample estimate±margin of error

The lower limit is obtained by:

the lower limit L of the interval=estimate−margin of error

The upper limit is obtained by:

the upper limit U of the interval=estimate+margin of error

Once we've obtained the interval, we can claim that we are really confident that the value of the population parameter is somewhere between the value of **L** and the value of **U**.

t-interval

To be more specific about their use, let's consider a specific interval, namely the "***t*-interval for a population mean *μ*.**“

(1-α)100% t-interval for the population mean μ

If we are interested in estimating a population mean μ, it is very likely that we would use the *t*-interval for a population mean μ.

t-Interval for a Population Mean

Formula for the confidence interval is:

Sample mean ± (t-multiplier×standard error)

and you might recall that the formula for the confidence interval in notation is:

$\bar{x} \pm t\alpha/2\ ,n-1(s/\sqrt{n})$

Observations

- "***t*-multiplier**," which we denote as $t\alpha/2,n-1$, depends on the sample size through *n* - 1 (called the "**degrees of freedom**") and the confidence level $(1-\alpha)\times100$ through $\alpha2$.
- "**standard error**," which is sn, quantifies how much the sample means x¯ vary from sample to sample. That is, the standard error is just another name for the estimated standard deviation of all the possible sample means.

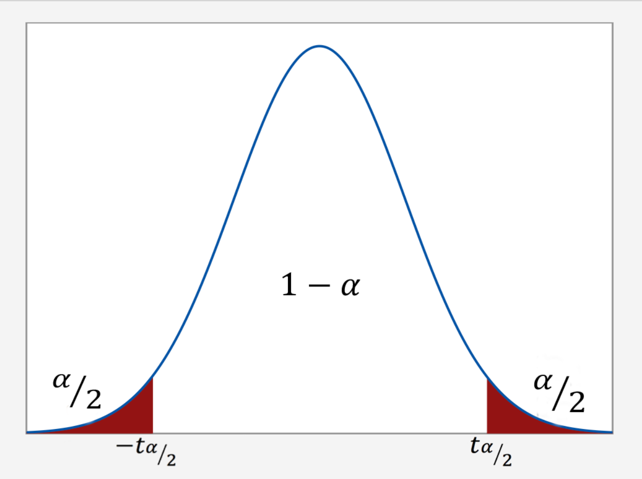
Observations

- The quantity to the right of the ± sign, *i.e.*, "***t*-multiplier** × **standard error**," is just a more specific form of the margin of error. That is, the margin of error in estimating a population mean μ is calculated by multiplying the *t*-multiplier by the standard error of the sample mean.

- The formula is only appropriate if a certain assumption is met, namely that the data are normally distributed.

How is the t-multiplier determined?

- We put the confidence level 1−α in the center of the *t*-distribution.
- Then, since the entire probability represented by the curve must equal 1, a probability of α must be shared equally among the two "tails" of the distribution.
- That is, the probability of the left tail is α2 and the probability of the right tail is α2. If we add up the probabilities of the various parts (α2+1−α+α2), we get 1.



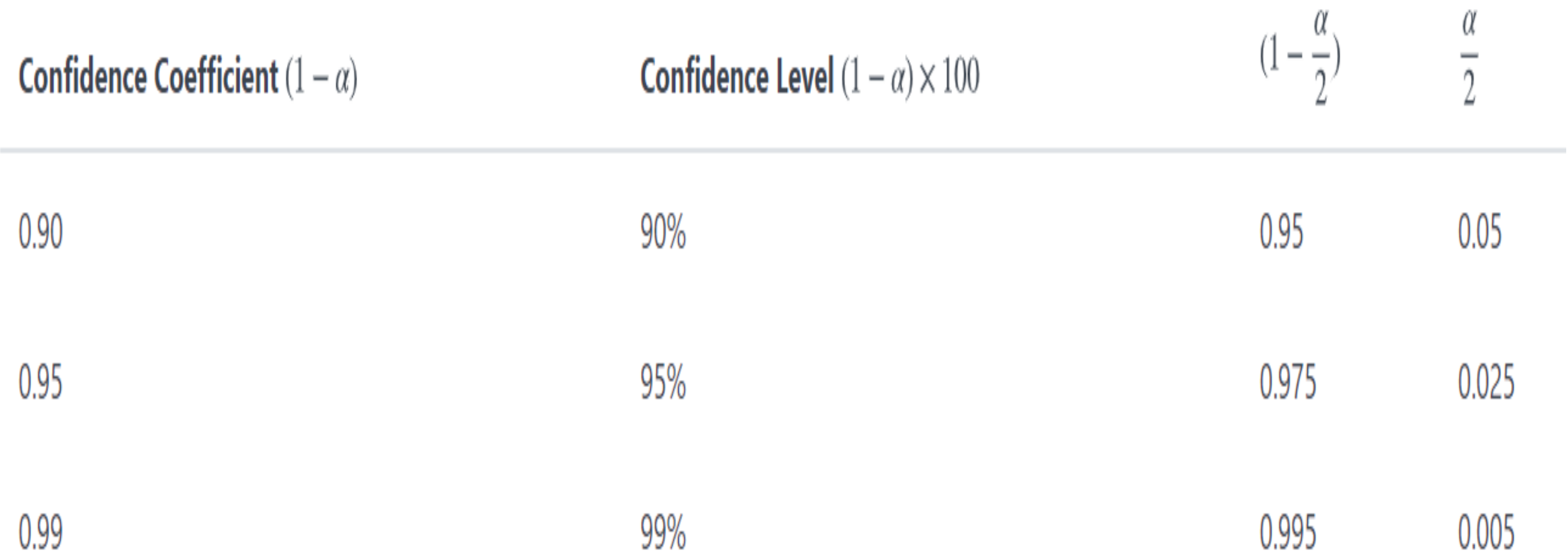
Observations

- This is why confidence levels are typically very high. The most common confidence levels are 90%, 95% and 99%.

- The following table contains a summary of the values of α2 corresponding to these common confidence levels.

- The"**confidence coefficient**" is merely the confidence level reported as a proportion rather than as a percentage.

Confidence Co-efficient table



Confidence Coefficient $(1-\alpha)$	Confidence Level $(1-\alpha) \times 100$	$(1-\frac{\alpha}{2})$	$\frac{\alpha}{2}$
0.90	90%	0.95	0.05
0.95	95%	0.975	0.025
0.99	99%	0.995	0.005

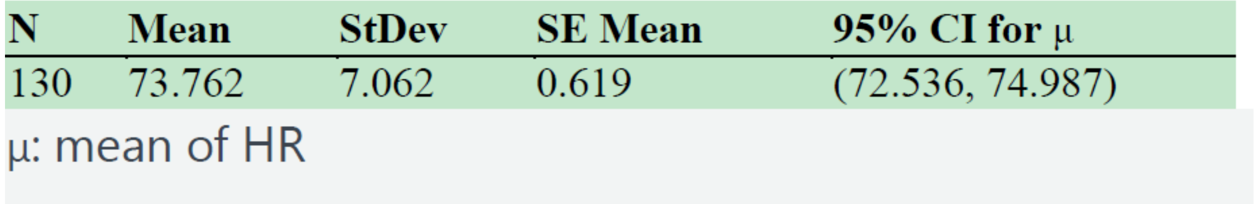
Example

Let's take an example of researchers who are interested in the average heart rate of male college students. Assume a random sample of 130 male college students were taken for the study.

Output of a one-sample *t*-interval output using this data.

One-Sample T: Heart Rate

Descriptive Statistics:



N	Mean	StDev	SE Mean	95% CI for μ
130	73.762	7.062	0.619	(72.536, 74.987)

μ: mean of HR

Factors Affecting the Width of the t-interval for the Mean μ

- We are 95% confident that the average GPA of all college students is between 1.0 and 4.0.

- We are 95% confident that the average GPA of all college students is between 2.7 and 2.9.

Factors Affecting the Width of the t-interval

To find the width of the confidence interval:

Width = Upper Limit - Lower Limit

What factors affect the width of the confidence interval?

We can examine this question by using the formula for the confidence interval and seeing what would happen should one of the elements of the formula be allowed to vary.

$\bar{x} \pm t\alpha/2,n-1(s/\sqrt{n})$

What is the width of the *t*-interval for the mean? If you subtract the lower limit from the upper limit, you get:

Width $=2 \times t\alpha/2,n-1(s/\sqrt{n})$

Observations

- As the sample mean increases, the length stays the same. That is, the sample mean plays no role in the width of the interval.

- As the sample standard deviation *s* decreases, the width of the interval decreases. Since *s* is an estimate of how much the data vary naturally,

Observations

We have little control over *s* other than making sure that we make our measurements as carefully as possible.

- As we decrease the confidence level, the *t*-multiplier decreases, and hence the width of the interval decreases. In practice, we wouldn't want to set the confidence level below 90%.
- As we increase the sample size, the width of the interval decreases. This is the factor that we have the most flexibility in changing, the only limitation being our time and financial constraints.

PES UNIVERSITY ONLINE

THANK YOU

Prof D.Uma
Prof S. Suganthi
Prof J. Silviya Nancy

Computer Science and Engineering

umaprabha@pes.edu
+91 99 7251 5335