STATISTICS FOR DATA SCIENCE

Confidence Intervals for Large Samples

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

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Confidence Intervals for Large Samples

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

STATISTICS FOR DATA SCIENCE

Topics to be covered...

- Confidence Intervals
- Confidence Intervals for population mean of large samples
- Confidence Levels
- Confidence Co-efficient
- Probability Vs Confidence
- One sided confidence intervals
- Confidence Intervals for Proportions

STATISTICS FOR DATA SCIENCE

Confidence Intervals

- Its an interval estimate for a population parameter and is how much uncertainty there is with any particular statistic.
- Based on sample data and provides a range of plausible values for a parameter.
- Confidence Interval differs from sample to sample (taken from same population).
- Confidence intervals are intrinsically connected to confidence levels.

STATISTICS FOR DATA SCIENCE

Confidence Intervals

- Confidence intervals are often used with a margin of error.
- It tells you how confident you can be that the results from a poll or survey reflect what you would expect to find if it were possible to survey the entire population.
- For example: we may be 95% confident that μ (population parameter) lies in the interval (-0.2, 3.1). Here, being 95% is the confidence level associated with the Confidence Interval (-0.2, 3.1).

STATISTICS FOR DATA SCIENCE

Confidence Intervals

- Confidence intervals are often used with a margin of error.
- It tells you how confident you can be that the results from a poll or survey reflect what you would expect to find if it were possible to survey the entire population.
- For example: we may be 95% confident that μ (population parameter) lies in the interval (-0.2, 3.1). Here, being 95% is the confidence level associated with the Confidence Interval (-0.2, 3.1).

Confidence Intervals VS Confidence Levels

- Confidence levels are expressed as a percentage(for example, a 95% confidence level).
- It means that should you repeat an experiment or survey over and over again, 95 percent of the time your results will match the results you get from a population (in other words, your statistics would be sound!).
- Confidence intervals are your results, usually numbers.

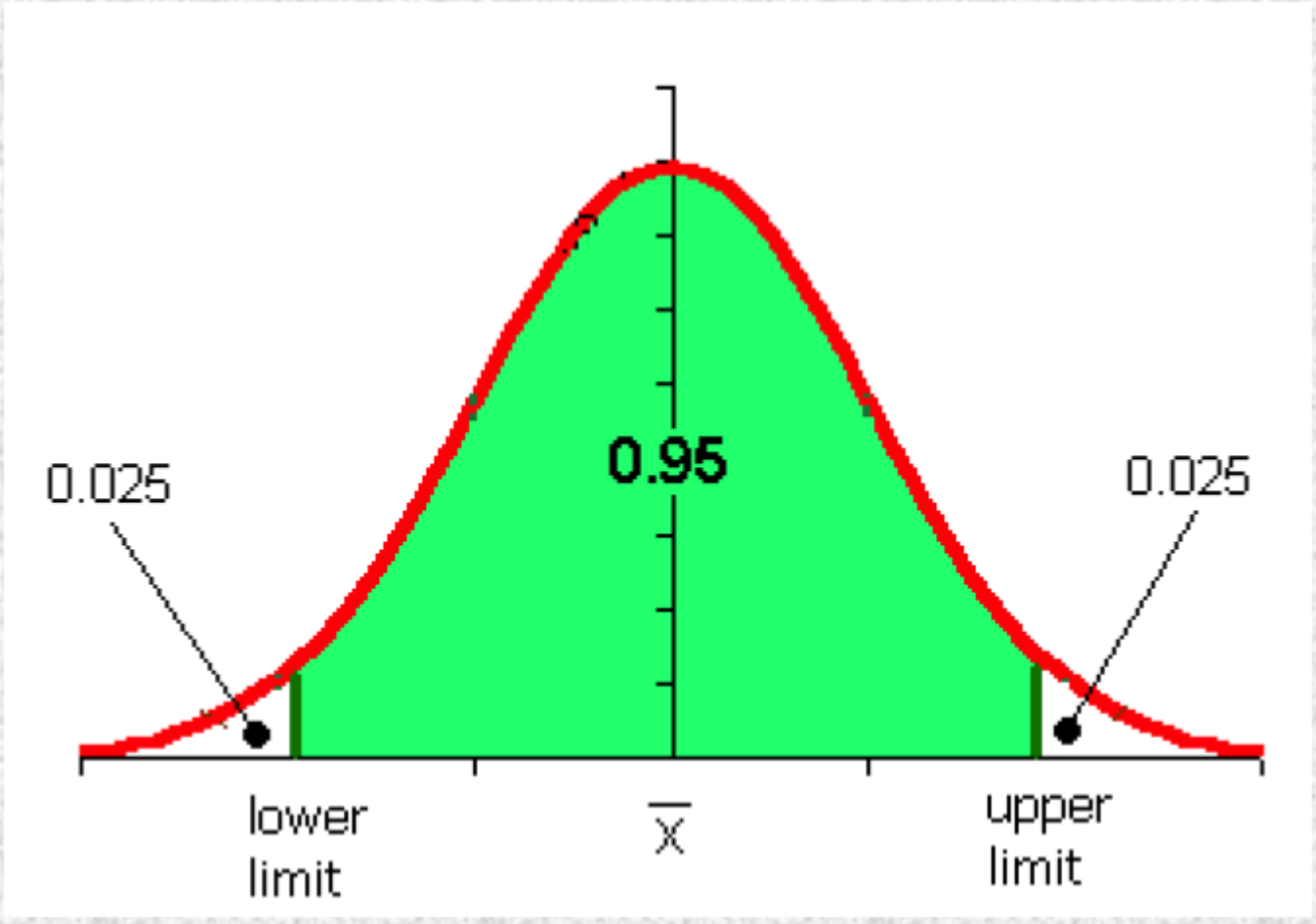
STATISTICS FOR DATA SCIENCE

Confidence Intervals VS Confidence Levels

- For example, you survey a group of pet owners to see how many cans of dog food they purchase a year.
- You test your statistics at the 99 percent confidence level and get a confidence interval of (200,300). That means you think they buy between 200 and 300 cans a year. You're super confident (99% is a very high level!) that your results are sound, statistically.

STATISTICS FOR DATA SCIENCE

95% Confidence Interval



Text-Book-Chapter 5.1- Page.No:323

Confidence Co-efficient

- The confidence coefficient is the confidence level stated as a proportion, rather than as a percentage.
- For example, if you had a confidence level of 99%, the confidence coefficient would be .99.
- In general, the higher the coefficient, the more certain you are that your results are accurate.

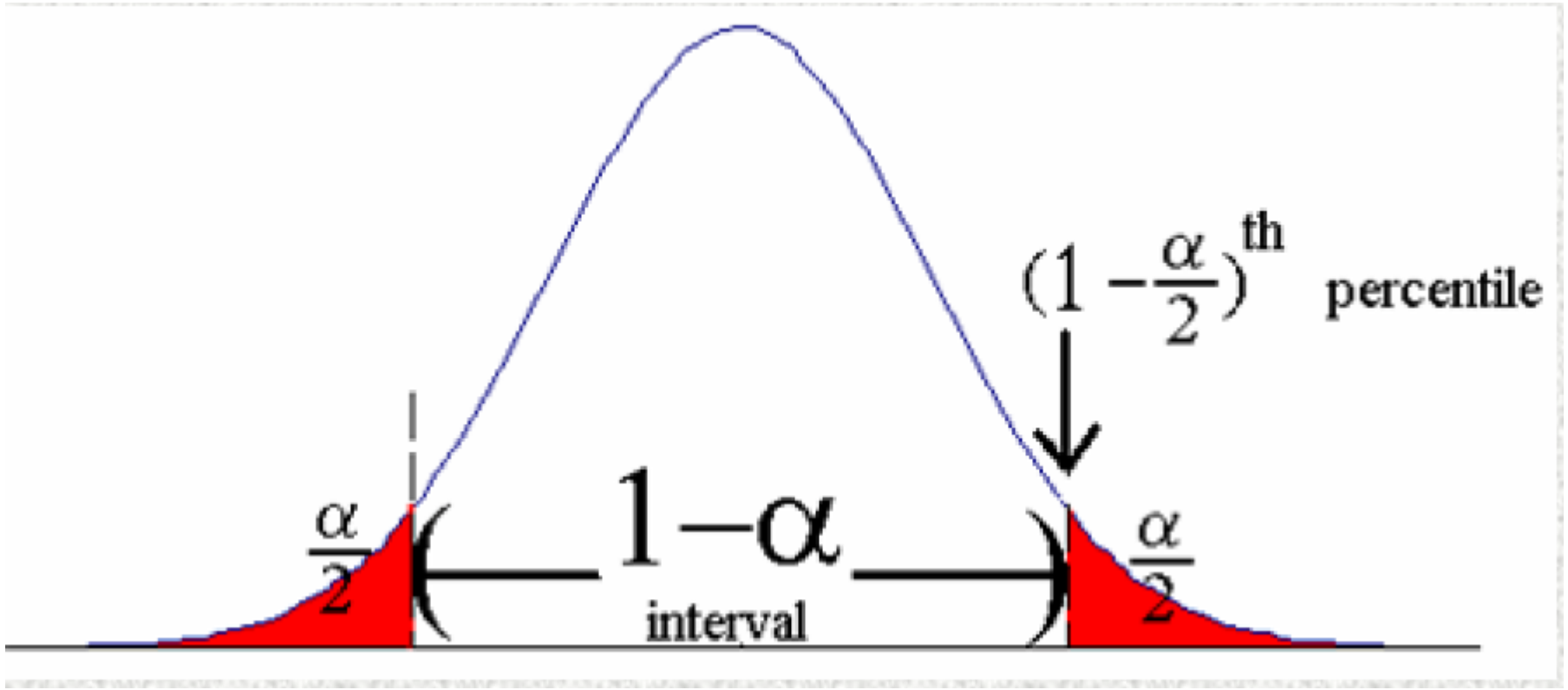
Confidence Co-efficient

The following table lists confidence coefficients and the equivalent confidence levels.

Confidence coefficient (1 – α)	Confidence level (1 – α * 100%)
0.90	90%
0.95	95%
0.99	99%

STATISTICS FOR DATA SCIENCE

Confidence Co-efficient



STATISTICS FOR DATA SCIENCE

Confidence Intervals for population mean of large samples

Let $X_1, \ldots, X_n$ be a *large* ($n > 30$) random sample from a population with mean μ and standard deviation σ, so that $\overline{X}$ is approximately normal. Then a level $100(1-\alpha)\%$ confidence interval for μ is

$$\overline{X} \pm z_{\alpha/2}\sigma_{\overline{X}} \qquad (5.1)$$

where $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. When the value of σ is unknown, it can be replaced with the sample standard deviation s.

In particular,

- $\overline{X} \pm \frac{s}{\sqrt{n}}$ is a 68% confidence interval for μ.
- $\overline{X} \pm 1.645\frac{s}{\sqrt{n}}$ is a 90% confidence interval for μ.
- $\overline{X} \pm 1.96\frac{s}{\sqrt{n}}$ is a 95% confidence interval for μ.
- $\overline{X} \pm 2.58\frac{s}{\sqrt{n}}$ is a 99% confidence interval for μ.
- $\overline{X} \pm 3\frac{s}{\sqrt{n}}$ is a 99.7% confidence interval for μ.

Text-Book-Chapter 5.1- Page.No:326

STATISTICS FOR DATA SCIENCE

Confidence Intervals for population mean of large samples

point estimate ± Margin of error

Confidence Interval for μ will be of the form:

X_bar ± Margin of error

Margin of error:

$$\pm 1.96 \frac{s}{\sqrt{n}}$$

STATISTICS FOR DATA SCIENCE

Confidence Intervals for population mean of large samples

A $(1-\alpha)$ 100% Confidence Interval for μ is given by:

$$X\_bar \pm Z_{\alpha/2}\,(\sigma/\sqrt{n})$$

where, the quantity $Z_{\alpha/2}\,(\sigma/\sqrt{n})$ is the Margin of error.

Note:

$(\sigma/\sqrt{n})$ is the standard deviation of Sampling distribution of sample mean (X_bar).

Text-Book-Chapter 5.1- Page.No:326

Example1

Find the value of $z_{\alpha/2}$ to use to construct a confidence interval with level:

a) 95%

b) 98%

c) 99%

d) 80%

Text-Book-Chapter 5.1- Exercise 1-Page.No:335

Solution

a)95% : X_bar ± 1.96(σ/√n)

b)98% : X_bar ± 2.33 (σ/√n)

c)99% : X_bar ± 2.57(σ/√n) or

X_bar ± 2.58(σ/√n)

d)80% : X_bar ± 1.28 (σ/√n)

Example2

Find the levels of confidence intervals that have the following values of $z_{\alpha/2}$:

a) $z_{\alpha/2} = 2.17$

b) $z_{\alpha/2} = 3.28$

Examples

a) $z_{\alpha/2} = 2.17$

$$P(-z_{0.015} < Z < z_{0.015}) = 1 - 0.03$$

Hence the confidence level is 97%

b) $z_{\alpha/2} = 3.28$

$$P(-z_{0.0005} < Z < z_{0.0005}) = 1 - 0.001$$

Hence the confidence level is 99.9%

Interpreting confidence intervals

- To interpret the confidence interval of the mean, you must assume that:
- All the values were **independently and randomly sampled** from a population whose values are distributed according to a **Gaussian(Normal) distribution.**
- A confidence interval is calculated from one given sample. It either covers or misses the true parameter. Since the true parameter is unknown, you'll never know which one is true.

Interpreting confidence intervals

- If independent samples are taken repeatedly from the same population, and a confidence interval calculated for each sample, then a certain percentage (**confidence level**) of the intervals will include the unknown population parameter.

- The **confidence level** associated with a confidence interval is the success rate of the confidence interval.

STATISTICS FOR DATA SCIENCE

Probability Vs Confidence

- It is correct to say that there is a 95% chance that the confidence interval you calculated contains the true population mean.
- It is not quite correct to say that there is a 95% chance that the population mean lies within the interval.
- The population mean has one value.
- In contrast, the confidence interval you compute depends on the data you happened to collect.

STATISTICS FOR DATA SCIENCE

Example

A 90% confidence interval for the mean diameter (in cm) of steel rods manufactured on a certain extrusion machine is computed to be (14.73, 14.91). True or false: The probability that the mean diameter of rods manufactured by this process is between 14.73 and 14.91 is 90%.

Solution

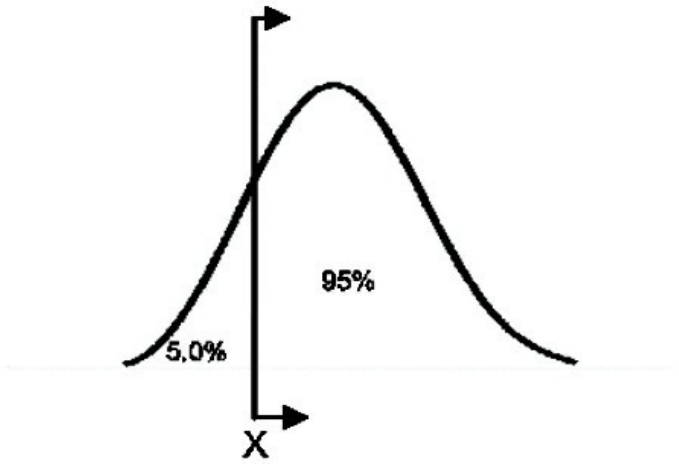
False. A specific confidence interval is given. The mean is either in the interval or it isn't. We are 90% confident that the population mean is between 14.73 and 14.91. The term *probability* is inappropriate.

STATISTICS FOR DATA SCIENCE

One-Sided Confidence Intervals

1)An upper one-sided bound defines a point that a certain percentage of the population is less than.

For example, if X is a 95% upper one-sided bound, this would indicate that 95% of the population is less than X.

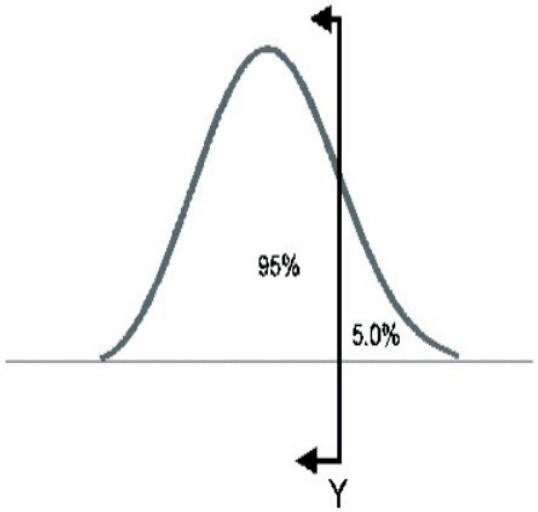


Text-Book-Chapter 5.1- Page.No:331

One-Sided Confidence Intervals

2)A lower one-sided bound defines a point that a specified percentage of the population is greater than.

For example, If X is a 95% lower one-sided bound, this would indicate that 95% of the population is greater than X.



STATISTICS FOR DATA SCIENCE

One-Sided Confidence Intervals

Let $X_1, \ldots, X_n$ be a *large* ($n > 30$) random sample from a population with mean μ and standard deviation σ, so that $\overline{X}$ is approximately normal. Then level $100(1-\alpha)\%$ lower confidence bound for μ is

$$\overline{X} - z_\alpha \sigma_{\overline{X}} \tag{5.2}$$

and level $100(1-\alpha)\%$ upper confidence bound for μ is

$$\overline{X} + z_\alpha \sigma_{\overline{X}} \tag{5.3}$$

where $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. When the value of σ is unknown, it can be replaced with the sample standard deviation s.

Text-Book-Chapter 5.1- Page.No:332

STATISTICS FOR DATA SCIENCE

One-Sided Confidence Intervals

In particular,

- $\overline{X} + 1.28\frac{s}{\sqrt{n}}$ is a 90% upper confidence bound for μ.
- $\overline{X} + 1.645\frac{s}{\sqrt{n}}$ is a 95% upper confidence bound for μ.
- $\overline{X} + 2.33\frac{s}{\sqrt{n}}$ is a 99% upper confidence bound for μ.

The corresponding lower bounds are found by replacing the "+" with "−."

Example1

In a sample of 80 ten-penny nails, the average weight was 1.56g and the standard deviation was 0.1g.

a) Find a 90% upper confidence bound for the mean weight.

b) Find a 80% lower confidence bound for the mean weight.

c) Someone says that the mean weight is less than 1.585g. With what level of confidence can this statement be made?

Solution

a) 90% upper confidence bound for the mean weight.

= 1.5743

b) Find a 80% lower confidence bound for the mean weight.

= 1.551

c) Someone says that the mean weight is less than 1.585g. With what level of confidence can this statement be made?

Hence we can make the statement with 98.75% confidence.

Example2

One step in the manufacture of a certain metal clamp involves the drilling of four holes. In a sample of 150 clams, the average time needed to complete this step was 72 seconds and the standard deviation was 10 seconds.

An efficiency expert says that the mean time is greater than 70 seconds. With what level of confidence can this statement be made?

Solution

Given:

mean = 72, sigma = 10, n = 150

Mean > 70

That means the lower confidence bound = 70

mean – lower_bound = 72 - 70 = 2

=> - z * (10/sqrt(150)) = 2

=> z = - 2.449 = - 2.45

=> Area to right of -2.45 = Area to the left of 2.45 = 0.9929

Hence we can make the statement with 99.29% confidence.

Confidence Intervals for Proportions

Let p represents the proportion in the population and it is given by:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

STATISTICS FOR DATA SCIENCE

Confidence Intervals for Proportions

Summary

Let X be the number of successes in n independent Bernoulli trials with success probability p, so that $X \sim \text{Bin}(n, p)$.

Define $\tilde{n} = n + 4$, and $\tilde{p} = \dfrac{X + 2}{\tilde{n}}$. Then a level $100(1 - \alpha)\%$ confidence interval for p is

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1 - \tilde{p})}{\tilde{n}}} \tag{5.5}$$

If the lower limit is less than 0, replace it with 0. If the upper limit is greater than 1, replace it with 1.

Text-Book-Chapter 5.2- Page.No:339

STATISTICS FOR DATA SCIENCE

One-sided Confidence Intervals for Proportions

Summary

Let X be the number of successes in n independent Bernoulli trials with success probability p, so that $X \sim \text{Bin}(n, p)$.

Define $\tilde{n} = n + 4$, and $\tilde{p} = \dfrac{X + 2}{\tilde{n}}$. Then a level $100(1 - \alpha)\%$ lower confidence bound for p is

$$\tilde{p} - z_\alpha \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{\tilde{n}}} \tag{5.6}$$

and level $100(1 - \alpha)\%$ upper confidence bound for p is

$$\tilde{p} + z_\alpha \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{\tilde{n}}} \tag{5.7}$$

If the lower bound is less than 0, replace it with 0. If the upper bound is greater than 1, replace it with 1.

Example

A random sample of n = 50 males showed a mean average daily intake of dairy products equal to 756 grams with a standard deviation of 35 grams. Find a 95% confidence interval for the population average μ?

Solution

$$\bar{x} \pm 1.96\frac{s}{\sqrt{n}} \Rightarrow 756 \pm 1.96\frac{35}{\sqrt{50}} \Rightarrow 756 \pm 9.70$$

$$\text{or } 746.30 < \mu < 765.70 \text{grams.}$$

Example

Of a random sample of n = 150 college students, 104 of the students said that they had played on a soccer team during their K-12 years. Estimate the proportion of college students who played soccer in their youth with a 90% confidence interval.

Solution

$$\hat{p} \pm 1.645\sqrt{\frac{\hat{p}\hat{q}}{n}} \quad \Rightarrow \frac{104}{150} \pm 1.645\sqrt{\frac{.69(.31)}{150}}$$

$$\Rightarrow \quad .69 \pm .06 \quad \text{or} \quad .63 < p < .75.$$

THANK YOU

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

Department of Computer Science and Engineering