

Course 02402 Introduction to Statistics Lecture 4:

Confidence interval for mean (and standard deviation)

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

Agenda

- 1 Intro and example
- 2 Distribution of sample mean
 - t -Distribution
- 3 Confidence interval for μ
 - Example
- 4 The language of statistics and the formal framework
- 5 Non-normal data, Central Limit Theorem (CLT)
- 6 A formal interpretation of the confidence interval
- 7 Confidence interval for variance and standard deviation

Oversigt

- 1 Intro and example
- 2 Distribution of sample mean
 - t -Distribution
- 3 Confidence interval for μ
 - Example
- 4 The language of statistics and the formal framework
- 5 Non-normal data, Central Limit Theorem (CLT)
- 6 A formal interpretation of the confidence interval
- 7 Confidence interval for variance and standard deviation

Example - heights:

Sample, $n = 10$:

168 161 167 179 184 166 198 187 191 179

Sample mean and standard deviation:

$$\bar{x} = 178$$

$$s = 12.21$$

Estimate population mean and standard deviation:

$$\hat{\mu} = 178$$

$$\hat{\sigma} = 12.21$$

NEW: Confidence interval, μ :

$$178 \pm 2.26 \cdot \frac{12.21}{\sqrt{10}} \Leftrightarrow [169.3; 186.7]$$

NEW: Confidence interval, σ :

$$[8.4; 22.3]$$

Oversigt

- 1 Intro and example
- 2 Distribution of sample mean
 - t -Distribution
- 3 Confidence interval for μ
 - Example
- 4 The language of statistics and the formal framework
- 5 Non-normal data, Central Limit Theorem (CLT)
- 6 A formal interpretation of the confidence interval
- 7 Confidence interval for variance and standard deviation

Let's simulate the key challenge of statistics!

```
## Mean
mu <- 178
## Standard deviation
sigma <- 12
## Sample size
n <- 10
## Simulate normally distributed  $X_i$ 
x <- rnorm(n=n, mean=mu, sd=sigma)
## See the realizations
x
## Empirical density
hist(x, prob=TRUE, col='blue')
## Find the sample mean
mean(x)
## Find the sample variance
var(x)
## Repeat the simulated sampling many times
mat <- replicate(100, rnorm(n=n, mean=mu, sd=sigma))
## Find the sample mean for each of them
xbar <- apply(mat, 2, mean)
## Now we have many realizations of the sample mean
xbar
## See their distribution
hist(xbar, prob=TRUE, col='blue')
## There mean
mean(xbar)
## and sample variance
var(xbar)
```

Theorem 3.2: The distribution of the mean of normal random variables

(Sample-) Distribution/ The (sampling) distribution for \bar{X}

Assume that X_1, \dots, X_n are independent and identically normally distributed random variables,
 $X_i \sim N(\mu, \sigma^2), i = 1, \dots, n$, then:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Mean and variance follow from 'rules': (Theorem 2.56)

The Mean of \bar{X}

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} n\mu = \mu$$

The variance of \bar{X}

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

We now know the distribution of the error we make:

(When using \bar{x} as an estimate of μ)

The standard deviation of \bar{X}

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

The standard deviation of $(\bar{X} - \mu)$

$$\sigma_{(\bar{X} - \mu)} = \frac{\sigma}{\sqrt{n}}$$

Standardized version of the same thing, Theorem 3.3:

Distribution for the standardized error we make:

Assume that X_1, \dots, X_n are independent and identically normally distributed random variables, $X_i \sim N(\mu, \sigma^2)$ where $i = 1, \dots, n$, then:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$$

That is, the standardized sample mean Z follows a standard normal distribution.

Practical problem in all this, so far:

How to transform this into a specific interval for μ ?

When the populations standard deviation σ is in all the formulas?

Obvious solution:

Use the estimate s in stead of σ in formulas!

BUT BUT:

The given theory then breaks down!!

Luckily:

We have an extended theory to handle it for us!!

Theorem 3.4: More applicable extension of the same stuff: (copy of Theorem 2.89)

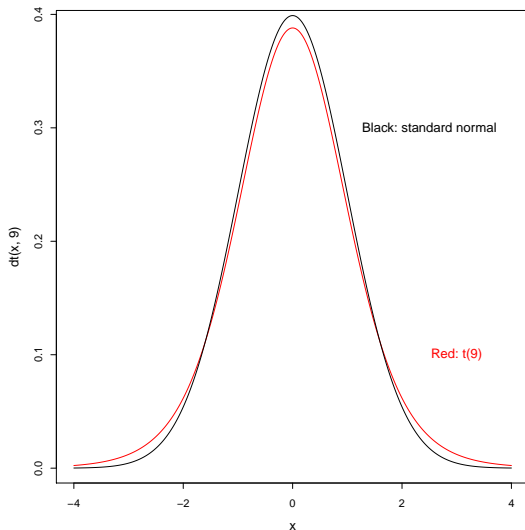
The t -Distribution takes the uncertainty of s into account:

Assume that X_1, \dots, X_n are independent and identically normally distributed random variables, where $X_i \sim N(\mu, \sigma^2)$ and $i = 1, \dots, n$, then:

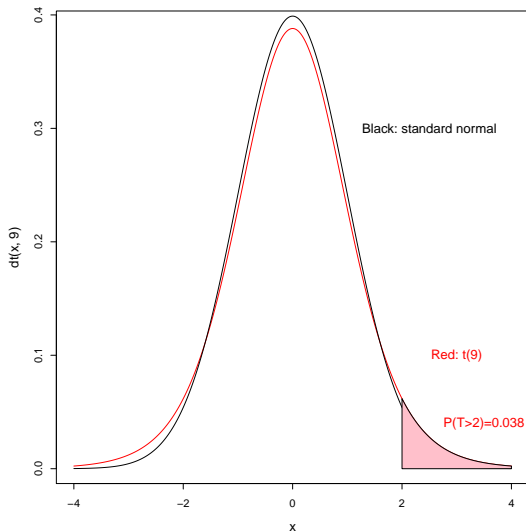
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t$$

where t is the t -distribution with $n - 1$ degrees of freedom.

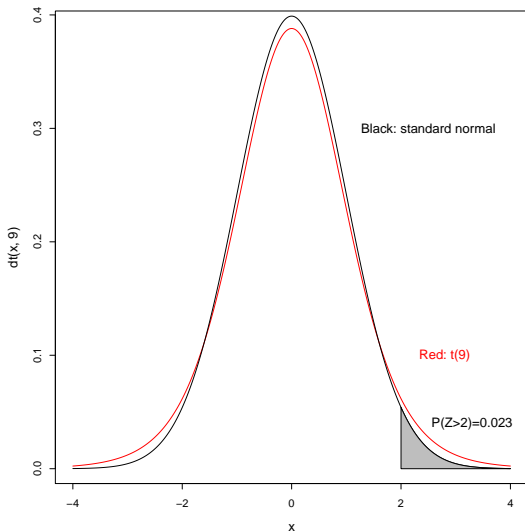
t -Distribution with 9 degrees of freedom ($n = 10$):



t -Distribution with 9 degrees of freedom and standard normal distribution:



t -Distribution with 9 degrees of freedom and standard normal distribution:



Oversight

- 1 Intro and example
- 2 Distribution of sample mean
 - t -Distribution
- 3 Confidence interval for μ
 - Example
- 4 The language of statistics and the formal framework
- 5 Non-normal data, Central Limit Theorem (CLT)
- 6 A formal interpretation of the confidence interval
- 7 Confidence interval for variance and standard deviation

Method box 3.8: One-sample Confidence interval for μ

Use the right t -distribution to make the confidence interval:
For a sample x_1, \dots, x_n the $100(1 - \alpha)\%$ confidence interval is given by:

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

where $t_{1-\alpha/2}$ is the $100(1 - \alpha)\%$ quantile from the t -distribution with $n - 1$ degrees of freedom.

Most commonly using $\alpha = 0.05$:

The most commonly used is the 95%-confidence interval:

$$\bar{x} \pm t_{0.975} \cdot \frac{s}{\sqrt{n}}$$

Student height Example

```
## The t-quantiles for n=10:  
qt(0.975,9)  
## [1] 2.262
```

and we can recognize the already given result:

$$178 \pm 2.26 \cdot \frac{12.21}{\sqrt{10}}$$

which is:

$$178 \pm 8.74 = [169.3; 186.7]$$

Student height example, 99% Confidence interval (CI)

```
qt(0.995, 9)
```

```
## [1] 3.25
```

$$178 \pm 3.25 \cdot \frac{12.21}{\sqrt{10}}$$

giving

$$178 \pm 12.55 = [165.4; 190.6]$$

There is an R-function, that can do it all
(and more than that):

```
x <- c(168,161,167,179,184,166,198,187,191,179)
t.test(x,conf.level=0.99)

##
##  One Sample t-test
##
## data:  x
## t = 46, df = 9, p-value = 5e-12
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  165.5 190.5
## sample estimates:
## mean of x
##      178
```

Oversight

- 1 Intro and example
- 2 Distribution of sample mean
 - t -Distribution
- 3 Confidence interval for μ
 - Example
- 4 The language of statistics and the formal framework
- 5 Non-normal data, Central Limit Theorem (CLT)
- 6 A formal interpretation of the confidence interval
- 7 Confidence interval for variance and standard deviation

The formal framework for *statistical inference*

From eNote, Chapter 1:

- An *observational unit* is the single entity/level about which information is sought (e.g. a person) (**Observations**)
- The *statistical population* consists of all possible “measurements” on each *observational unit* (**Population**)
- The *sample* from a statistical population is the actual set of data collected. (**Sample**)

Language and concepts:

- μ and σ are parameters describing the population
- \bar{x} is the *estimate* of μ (specific realization)
- \bar{X} is the *estimator* of μ (now seen as a random variable)
- The word '*statistic(s)*' is used for both

The formal framework for *statistical inference* - Example

From eNote, Chapter 1, heights example

We measure the heights of 10 randomly selected persons in Denmark

The sample:

The 10 specific numbers: x_1, \dots, x_{10}

The population:

The heights for all people in Denmark

Observational unit:

A person

Statistical inference = Learning from data

Learning from data:

Is learning about parameters of distributions that describe populations.

Important for this:

The sample must in a meaningful way represent some well defined population

How to ensure this:

F.ex. by making sure that the sample is taken completely at random

Random Sampling

Definition 3.11

- A random sample from an (infinite) population: A set of observations X_1, X_2, \dots, X_n constitutes a random sample of size n from the infinite population $f(x)$ if:
 - ① Each X_i is a random variable whose distribution is given by $f(x)$
 - ② These n random variables are independent

What does that mean????

- ① All observations must come from the same population
- ② They cannot share any information with each other (e.g. if we sampled entire families)

Oversight

- 1 Intro and example
- 2 Distribution of sample mean
 - t -Distribution
- 3 Confidence interval for μ
 - Example
- 4 The language of statistics and the formal framework
- 5 Non-normal data, Central Limit Theorem (CLT)
- 6 A formal interpretation of the confidence interval
- 7 Confidence interval for variance and standard deviation

Theorem 3.13: The Central Limit Theorem

No matter what, the distribution of the mean becomes a normal distribution:

Let \bar{X} be the mean of a random sample of size n taken from a population with mean μ and variance σ^2 , then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

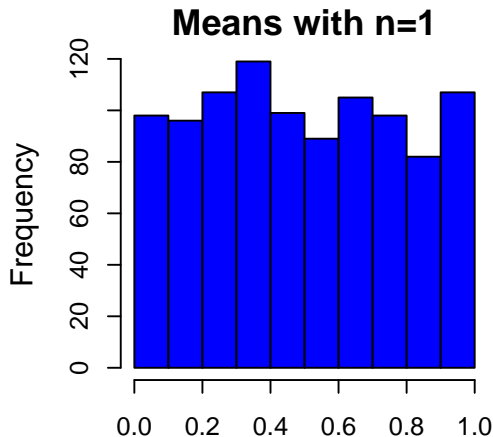
is a random variable whose distribution function approaches that of the standard normal distribution, $N(0, 1^2)$, as $n \rightarrow \infty$

Hence, if n is large enough, we can (approximately) assume:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$$

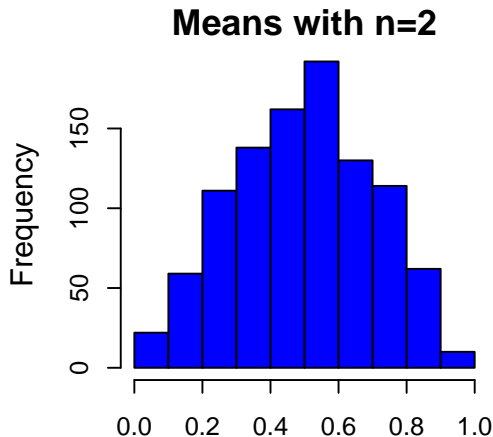
CLT in action - mean of uniformly distributed observations

```
n=1  
k=1000  
u=matrix(runif(k*n),ncol=n)  
hist(apply(u,1,mean), col = "blue", main = "Means with n=1", xlab = "")
```



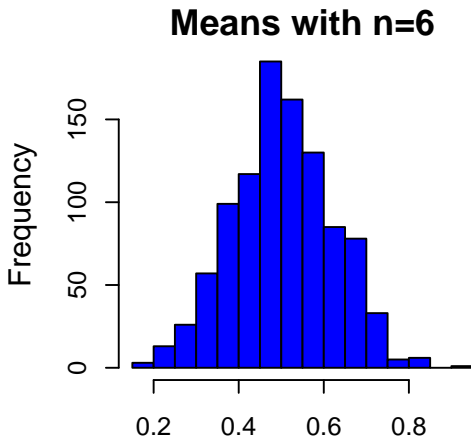
CLT in action - mean of uniformly distributed observations

```
n=2  
k=1000  
u=matrix(runif(k*n),ncol=n)  
hist(apply(u,1,mean), col = "blue", main = "Means with n=2", xlab = "")
```



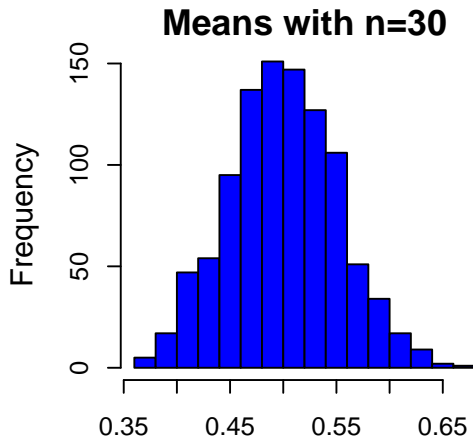
CLT in action - mean of uniformly distributed observations

```
n=6  
k=1000  
u=matrix(runif(k*n),ncol=n)  
hist(apply(u,1,mean), col = "blue", main = "Means with n=6", xlab = "")
```



CLT in action - mean of uniformly distributed observations

```
n=30  
k=1000  
u=matrix(runif(k*n),ncol=n)  
hist(apply(u,1,mean), col = "blue", main = "Means with n=30", xlab = "")
```



Consequence of CLT:

Our CI-method also works for non-normal data:

We can use the confidence-interval based on the t -distribution in basically any situation, as long as n is large enough.

What is "large enough"?

Actually difficult to say exactly, BUT:

- Rule of thumb: $n \geq 30$
- Even for smaller n the approach can be (almost) valid for non-normal data.

Oversight

- 1 Intro and example
- 2 Distribution of sample mean
 - t -Distribution
- 3 Confidence interval for μ
 - Example
- 4 The language of statistics and the formal framework
- 5 Non-normal data, Central Limit Theorem (CLT)
- 6 A formal interpretation of the confidence interval
- 7 Confidence interval for variance and standard deviation

'Repeated sampling' interpretation

In the long run we catch the true value in 95% of cases:
The confidence interval will vary in both width (s) and position (\bar{x}) if the study is repeated.

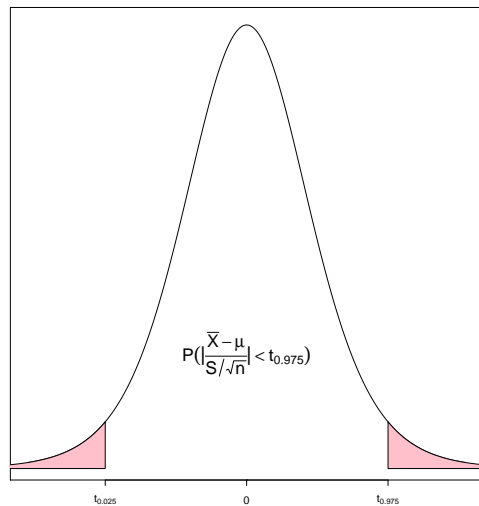
More formally expressed (Theorem 3.4 and 2.89):

$$P\left(\frac{|\bar{X} - \mu|}{S/\sqrt{n}} < t_{0.975}\right) = 0.95$$

Which is equivalent to:

$$P\left(\bar{X} - t_{0.975} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{0.975} \frac{S}{\sqrt{n}}\right) = 0.95$$

'Repeated sampling' interpretation



Oversight

- 1 Intro and example
- 2 Distribution of sample mean
 - t -Distribution
- 3 Confidence interval for μ
 - Example
- 4 The language of statistics and the formal framework
- 5 Non-normal data, Central Limit Theorem (CLT)
- 6 A formal interpretation of the confidence interval
- 7 Confidence interval for variance and standard deviation

Motivating Example: Production of tablets

In the production of tablets, an active matter is mixed with a powder and then the mixture is formed to tablets. It is important that the mixture is homogenous, so that each tablet has the same strength.

We consider a mixture (of the active matter and powder) from where a large amount of tablets is to be produced.

We seek to produce the mixtures (and the final tablets) so that the mean content of the active matter is 1 mg/g with the smallest variance as possible. A random sample is collected where the amount of active matter is measured. It is assumed that all the measurements follow a normal distribution with the unit mg/g.

The sampling distribution of the variance estimator (Theorem 2.81)

Variance estimators behaves like a χ^2 -distribution:

Let

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

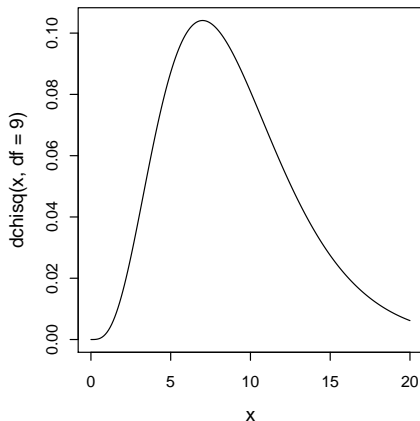
then:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

is a stochastic variable following the χ^2 -distribution with $\nu = n - 1$ degrees of freedom.

χ^2 -distribution with $\nu = 9$ degrees of freedom

```
x <- seq(0, 20, by = 0.1)
plot(x, dchisq(x, df = 9), type = "l")
```



Method 3.18: Confidence interval for sample variance and standard deviation

The variance:

A $100(1 - \alpha)\%$ confidence interval for a sample variance $\hat{\sigma}^2$ is:

$$\left[\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}; \frac{(n-1)s^2}{\chi_{\alpha/2}^2} \right]$$

where the quantiles come from a χ^2 -distribution with $\nu = n - 1$ degrees of freedom.

The standard deviation:

A $100(1 - \alpha)\%$ confidence interval for the sample standard deviation $\hat{\sigma}$ is:

$$\left[\sqrt{\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}}; \sqrt{\frac{(n-1)s^2}{\chi_{\alpha/2}^2}} \right]$$

Example

Data:

A random sample with $n = 20$ tablets is taken and from this we get:

$$\hat{\mu} = \bar{x} = 1.01, \hat{\sigma}^2 = s^2 = 0.07^2$$

95%-Confidence interval for the variance - we need the χ^2 -quantiles:

$$\chi_{0.025}^2 = 8.9065, \chi_{0.975}^2 = 32.8523$$

```
qchisq(c(0.025, 0.975), df = 19)
```

```
[1] 8.907 32.852
```

Example

So the confidence interval for the variance σ^2 becomes:

$$\left[\frac{19 \cdot 0.7^2}{32.85}; \frac{19 \cdot 0.7^2}{8.907} \right] = [0.002834; 0.01045]$$

and the confidence interval for the standard deviation σ becomes:

$$\left[\sqrt{0.002834}; \sqrt{0.01045} \right] = [0.053; 0.102]$$

Heights example

We need the χ^2 -quantiles with $\nu = 9$ degrees of freedom:

$$\chi_{0.025}^2 = 2.700389, \chi_{0.975}^2 = 19.022768$$

```
qchisq(c(0.025, 0.975), df = 9)
```

```
[1] 2.70 19.02
```

So the confidence interval for the height standard deviation σ becomes:

$$\left[\sqrt{\frac{9 \cdot 12.21^2}{19.022768}}; \sqrt{\frac{9 \cdot 12.21^2}{2.700389}} \right] = [8.4; 22.3]$$

Example - heights- recap:

Sample, $n = 10$:

168 161 167 179 184 166 198 187 191 179

Sample mean and standard deviation:

$$\bar{x} = 178$$

$$s = 12.21$$

Estimate population mean and standard deviation:

$$\hat{\mu} = 178$$

$$\hat{\sigma} = 12.21$$

NEW: Confidence interval, μ :

$$178 \pm 2.26 \cdot \frac{12.21}{\sqrt{10}} \Leftrightarrow [169.3; 186.7]$$

NEW: Confidence interval, σ :

$$[8.4; 22.3]$$

Agenda

- 1 Intro and example
- 2 Distribution of sample mean
 - t -Distribution
- 3 Confidence interval for μ
 - Example
- 4 The language of statistics and the formal framework
- 5 Non-normal data, Central Limit Theorem (CLT)
- 6 A formal interpretation of the confidence interval
- 7 Confidence interval for variance and standard deviation