

# Course 02402 Introduction to Statistics

## Lecture 4: Confidence intervals

DTU Compute  
Technical University of Denmark  
2800 Lyngby – Denmark

## Overview

- 1 Intro and example
- 2 Distribution of the sample mean
  - The  $t$ -distribution
- 3 Confidence interval (CI) for  $\mu$ 
  - Example: Heights
- 4 The language of statistics and the formal framework
- 5 Non-normal data, the Central Limit Theorem (CLT)
- 6 Formal interpretation of the CI
- 7 CI for variance  $\sigma^2$  and standard deviation  $\sigma$

## Overview

- 1 Intro and example
- 2 Distribution of the sample mean
  - The  $t$ -distribution
- 3 Confidence interval (CI) for  $\mu$ 
  - Example: Heights
- 4 The language of statistics and the formal framework
- 5 Non-normal data, the Central Limit Theorem (CLT)
- 6 Formal interpretation of the CI
- 7 CI for variance  $\sigma^2$  and standard deviation  $\sigma$

## 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 for  $\mu$ :

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

NEW: Confidence interval for  $\sigma$ :

$$[8.4; 22.3]$$

## Overview

- 1 Intro and example
- 2 Distribution of the sample mean
  - The  $t$ -distribution
- 3 Confidence interval (CI) for  $\mu$ 
  - Example: Heights
- 4 The language of statistics and the formal framework
- 5 Non-normal data, the Central Limit Theorem (CLT)
- 6 Formal interpretation of the CI
- 7 CI for variance  $\sigma^2$  and standard deviation  $\sigma$

## (Empirical) distribution of sample mean

```
# 'True' mean and standard deviation
mu <- 178
sigma <- 12

# Sample size
n <- 10

# Simulate normal distributed  $X_i$  for  $n = 10$ 
x <- rnorm(n = n, mean = mu, sd = sigma)
x

# Empirical density
hist(x, prob = TRUE, col = 'blue')
# Compute sample mean
mean(x)

# Repeat the simulated sampling many times (100 samples)
mat <- replicate(100, rnorm(n = n, mean = mu, sd = sigma))

# Compute the sample mean for each sample
xbar <- apply(mat, 2, mean)
xbar

# See the distribution of the sample means
hist(xbar, prob = TRUE, col = 'blue')
# Empirical mean and variance of sample means
mean(xbar)
var(xbar)
```

## Theorem 3.3: Distribution of the sample mean of i.i.d. normal random variables

### The distribution of $\bar{X}$

Assume that  $X_1, \dots, X_n$  are independent and identically distributed (*i.i.d.*) normal 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, variance and 'normality' follow from 'rules':

The mean of  $\bar{X}$  (Theorem 2.56):

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

The variance of  $\bar{X}$  (Theorem 2.56):

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

The 'normality' of  $\bar{X}$  (Theorem 2.40):

By this theorem, the distribution of  $\bar{X}$  is a normal distribution with mean  $\mu$  and variance  $\sigma^2/n$  as specified above.

## Distribution of the error $\bar{X} - \mu$

The standard deviation of  $\bar{X}$ :

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

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

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

## Practical problem (and solution)

How do we use the results from the previous slides to say something about  $\mu$  ...  
... when the 'true', unknown, population standard deviation  $\sigma$  enters into all the formulas?

Obvious solution:

Use the estimate  $s$  instead of  $\sigma$  in formulas.

BUT:

Then, we need new theory! (There is also uncertainty linked to  $s$ .)

## Standardized version of the above, Theorem 3.4

Distribution of the standardized sample mean (or standardized error):

Assume that  $X_1, \dots, X_n$  are i.i.d. normal random variables,  $X_i \sim N(\mu, \sigma^2)$  for  $i = 1, \dots, n$ , 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.

## Theorem 3.5, a more applicable extension of the above

The  $t$ -distribution takes the uncertainty of  $s$  into account:

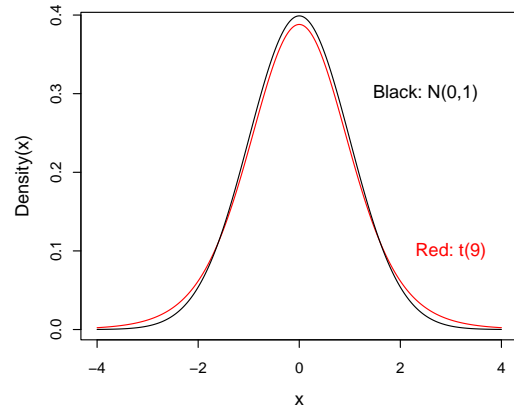
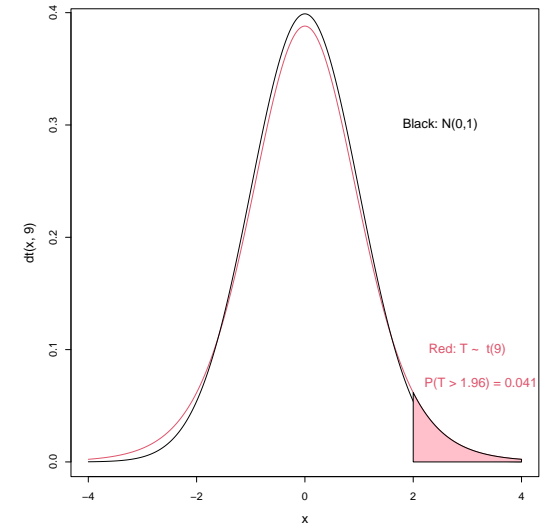
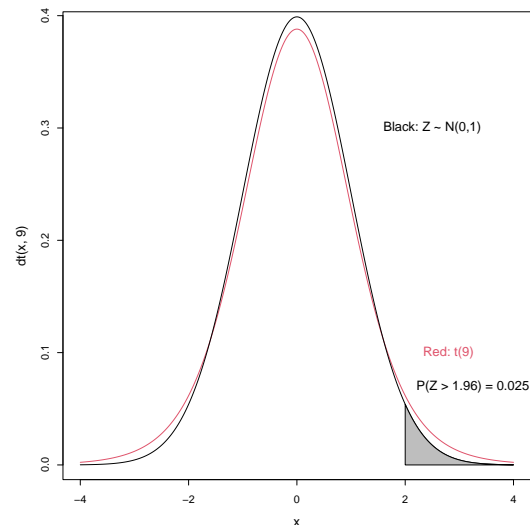
Assume that  $X_1, \dots, X_n$  are i.i.d. normal distributed random variables, where  $X_i \sim N(\mu, \sigma^2)$  for  $i = 1, \dots, n$ , then:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

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

The  $t$ -distribution with 9 degrees of freedom ( $n = 10$ )

```
x <- seq(-4, 4, by = 0.01)
plot(x, dt(x, df = 9), type = "l", col = "red", ylab = "Density(x)")
lines(x, dnorm(x), type = "l", col = "black")
text(2.5, 0.3, "Black: N(0,1)")
text(3, 0.1, "Red: t(9)", col = "red")
```

The  $t$ -distribution with 9 degrees of freedom and standard normal distributionThe  $t$ -distribution with 9 degrees of freedom and standard normal distribution

## Overview

- 1 Intro and example
- 2 Distribution of the sample mean
  - The  $t$ -distribution
- 3 Confidence interval (CI) for  $\mu$ 
  - Example: Heights
- 4 The language of statistics and the formal framework
- 5 Non-normal data, the Central Limit Theorem (CLT)
- 6 Formal interpretation of the CI
- 7 CI for variance  $\sigma^2$  and standard deviation  $\sigma$

## Method 3.9: One-sample Confidence Interval (CI) for $\mu$

Use the correct  $t$ -distribution to construct 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  $100(1 - \alpha/2)\%$  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:

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

## Example: Heights, 95% CI

```
# 0.975 quantile for the t(9) distribution (n = 10):
qt(0.975, df = 9)
```

Gives the result  $t_{0.975} = 2.26$ .

Now, we can recognize the already given result

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

which is

$$178 \pm 8.74 = [169.3, 186.7].$$

## Example: Heights, 99% CI

```
# 0.995 quantile for the t(9) distribution (n = 10):
qt(0.995, df = 9)
```

Gives the result  $t_{0.995} = 3.25$ .

In this case,

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

giving

$$178 \pm 12.55 = [165.5; 190.5]$$

## An R function for computing these CI (and more):

```
# Data
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)

# 99% CI for mu
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
```

## Overview

- 1 Intro and example
- 2 Distribution of the sample mean
  - The  $t$ -distribution
- 3 Confidence interval (CI) for  $\mu$ 
  - Example: Heights
- 4 **The language of statistics and the formal framework**
- 5 Non-normal data, the Central Limit Theorem (CLT)
- 6 Formal interpretation of the CI
- 7 CI for variance  $\sigma^2$  and standard deviation  $\sigma$

## The formal framework for *statistical inference*

From eNote, Chapter 1:

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

Language and concepts:

- $\mu$  and  $\sigma$  are parameters describing the population.
- $\bar{x}$  is the *estimate* of  $\mu$  (specific realization).
- $\bar{X}$  is the *estimator* of  $\mu$  (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. Example: Heights

We measured the heights of 10 randomly selected students.

The sample:

The 10 specific numbers (heights):  $x_1, \dots, x_{10}$ .

The population:

The heights for all people in Denmark.

Observational unit:

A person.

## Statistical inference = Learning from data

Learning from data:

Learning about parameters of distributions that describe populations.

Important:

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

How to ensure this:

For example, by making sure that the sample is taken completely at random.

## Random Sampling

### Definition 3.12:

- 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., shouldn't be related).

## Overview

- 1 Intro and example
- 2 Distribution of the sample mean
  - The  $t$ -distribution
- 3 Confidence interval (CI) for  $\mu$ 
  - Example: Heights
- 4 The language of statistics and the formal framework
- 5 Non-normal data, the Central Limit Theorem (CLT)
- 6 Formal interpretation of the CI
- 7 CI for variance  $\sigma^2$  and standard deviation  $\sigma$

## Theorem 3.14: The Central Limit Theorem (CLT)

"No matter the distribution of  $X_i$ ", the distribution of the mean of i.i.d. random variables approaches a normal distribution:

Let  $\bar{X}$  be the mean of a random sample of size  $n$  taken from a population with mean  $\mu$  and variance  $\sigma^2$ . 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 assume (approximately) that:

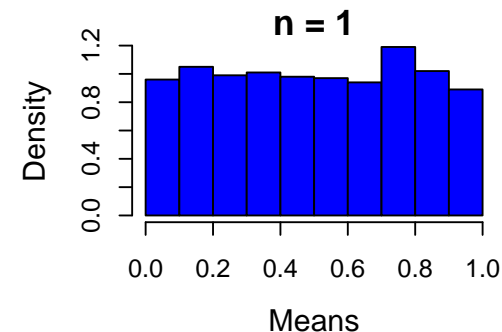
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$$

## CLT example: Mean of uniformly distributed observations

```
n <- 1 # Sample size
k <- 1000 # No. of samples (i.e. no. of means to be computed)

# Simulations from U(0,1)-distribution (k = 1000 samples, each of size n = 1)
u <- matrix(runif(k*n), ncol = n)

# Empirical density of means
hist(apply(u, 1, mean), col = "blue", main = "n = 1", xlab = "Means", prob = TRUE)
```

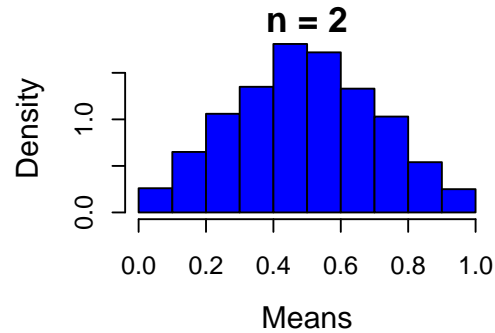


## CLT example: Mean of uniformly distributed observations

```
n <- 2 # Sample size
k <- 1000 # No. of samples (i.e. no. of means to be computed)

# Simulations from U(0,1)-distribution (k = 1000 samples, each of size n = 2)
u <- matrix(runif(k*n), ncol = n)

# Empirical density of means
hist(apply(u, 1, mean), col = "blue", main = "n = 2", xlab = "Means", xlim = c(0,1), prob = TRUE)
```

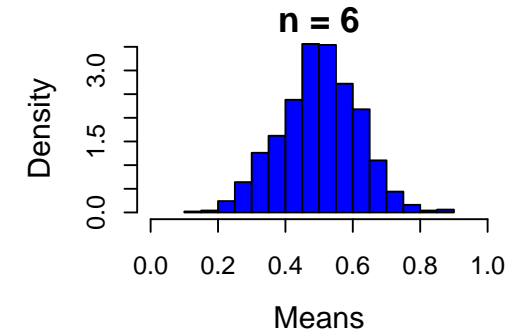


## CLT example: Mean of uniformly distributed observations

```
n <- 6 # Sample size
k <- 1000 # No. of samples (i.e. no. of means to be computed)

# Simulations from U(0,1)-distribution (k = 1000 samples, each of size n = 6)
u <- matrix(runif(k*n), ncol = n)

# Empirical density of means
hist(apply(u, 1, mean), col = "blue", main = "n = 6", xlab = "Means", xlim = c(0,1), prob = TRUE)
```

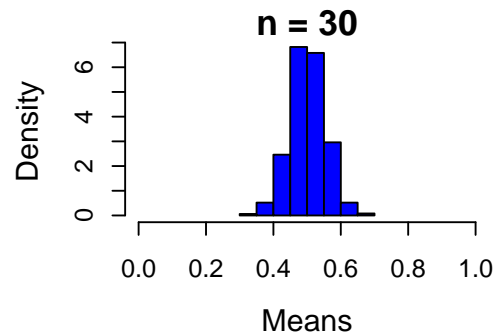


## CLT example: Mean of uniformly distributed observations

```
n <- 30 # Sample size
k <- 1000 # No. of samples (i.e. no. of means to be computed)

# Simulations from U(0,1)-distribution (k = 1000 samples, each of size n = 30)
u <- matrix(runif(k*n), ncol = n)

# Empirical density of means
hist(apply(u, 1, mean), col = "blue", main = "n = 30", xlab = "Means", xlim = c(0,1), prob = TRUE)
```



## Consequence of the 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.

When is  $n$  "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.



## Overview

- 1 Intro and example
- 2 Distribution of the sample mean
  - The  $t$ -distribution
- 3 Confidence interval (CI) for  $\mu$ 
  - Example: Heights
- 4 The language of statistics and the formal framework
- 5 Non-normal data, the Central Limit Theorem (CLT)
- 6 Formal interpretation of the CI**
- 7 CI for variance  $\sigma^2$  and standard deviation  $\sigma$

## Overview

- 1 Intro and example
- 2 Distribution of the sample mean
  - The  $t$ -distribution
- 3 Confidence interval (CI) for  $\mu$ 
  - Example: Heights
- 4 The language of statistics and the formal framework
- 5 Non-normal data, the Central Limit Theorem (CLT)
- 6 Formal interpretation of the CI
- 7 CI for variance  $\sigma^2$  and standard deviation  $\sigma$**

## 'Repeated sampling' interpretation

In the long run, we catch the true value in 95% of cases (95% CI):

The confidence interval will vary in both width ( $s$ ) and position ( $\bar{x}$ ) if the study is repeated.

More formally expressed (Theorem 3.5):

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

Which is equivalent to:

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

## 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

Assume i.i.d. normal distributed variables,  $X_i \sim N(\mu, \sigma^2)$  for  $i = 1, \dots, n$ .

Variance estimators behaves like a  $\chi^2$ -distribution:

Let

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

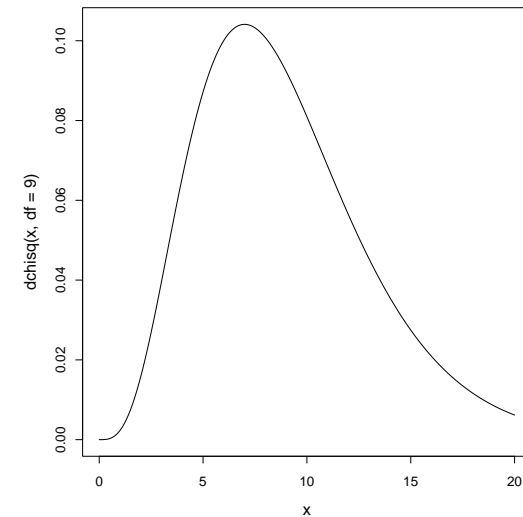
then:

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

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

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

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



## Method 3.19: Confidence interval for the variance and standard deviation

Assume i.i.d. normal distributed variables,  $X_i \sim N(\mu, \sigma^2)$  for  $i = 1, \dots, n$ .

The variance:

A  $100(1 - \alpha)\%$  confidence interval for the variance  $\sigma^2$  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  $\chi^2$ -distribution with  $\nu = n - 1$  degrees of freedom.

The standard deviation:

A  $100(1 - \alpha)\%$  confidence interval for the standard deviation  $\sigma$  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  $\chi^2$ -quantiles (19 degrees of freedom):

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

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

```
[1] 8.907 32.852
```

## Example

So the confidence interval for the variance  $\sigma^2$  becomes:

$$\left[ \frac{19 \cdot 0.07^2}{32.85}; \frac{19 \cdot 0.07^2}{8.907} \right] = [0.002834; 0.01045]$$

and the confidence interval for the standard deviation  $\sigma$  becomes:

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

## Example: Heights

Sample,  $n = 10$ :

168 161 167 179 184 166 198 187 191 179

Sample mean and standard deviation:

$$\begin{aligned} \bar{x} &= 178 \\ s &= 12.21 \end{aligned}$$

Estimate population mean and standard deviation:

$$\begin{aligned} \hat{\mu} &= 178 \\ \hat{\sigma} &= 12.21 \end{aligned}$$

NEW: Confidence interval,  $\mu$ :

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

NEW: Confidence interval,  $\sigma$ :

$$[8.4; 22.3]$$

## Example: Heights

We need the  $\chi^2$ -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  $\sigma$  becomes:

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

## Overview

- 1 Intro and example
- 2 Distribution of the sample mean
  - The  $t$ -distribution
- 3 Confidence interval (CI) for  $\mu$ 
  - Example: Heights
- 4 The language of statistics and the formal framework
- 5 Non-normal data, the Central Limit Theorem (CLT)
- 6 Formal interpretation of the CI
- 7 CI for variance  $\sigma^2$  and standard deviation  $\sigma$