Course 02323 & 02402 Introduction to Statistics

Lecture 4: Confidence intervals

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Overview

Overview

Example - Height of 10 students:

Sample, $n = 10$:

168	161	167	179	184	166	198	187	191	179

Example - Height of 10 students:

Sample, $n = 10$:

168	161	167	179	184	166	198	187	191	179

Sample mean and standard deviation:

$$\bar{x} = 178$$
$$s = 12.21$$

Example - Height of 10 students:

Sample, $n = 10$:

168	161	167	179	184	166	198	187	191	179

Sample mean and standard deviation:

$$\bar{x} = 178$$
$$s = 12.21$$

Estimates for the population mean and standard deviation:

$$\hat{\mu} = 178$$
$$\hat{\sigma} = 12.21$$

Example - Height of 10 students:

Sample, $n = 10$:

168	161	167	179	184	166	198	187	191	179

Sample mean and standard deviation:

$$\bar{x} = 178$$
$$s = 12.21$$

Estimates for the population mean and standard deviation:

$$\hat{\mu} = 178$$
$$\hat{\sigma} = 12.21$$

But how reliable are these estimates?

Example - Height of 10 students:

Sample, $n = 10$:

168	161	167	179	184	166	198	187	191	179

Sample mean and standard deviation:

$$\bar{x} = 178$$
$$s = 12.21$$

Estimates for the population mean and standard deviation:

$$\hat{\mu} = 178$$
$$\hat{\sigma} = 12.21$$

But how reliable are these estimates?

NEW: **Confidence Intervals:**

$$\hat{\mu} = 178 \; [169.3; 186.7]$$
$$\hat{\sigma} = 12.21 \; [8.4; 22.3]$$

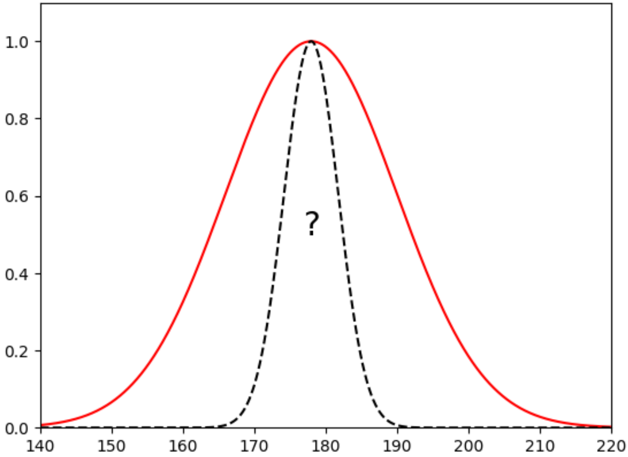
Python: (Empirical) Distribution of the sample mean

- Go to today's Python notebook in VS Code
  - "Simulation: Distribution of the sample mean"

Visual Studio Code

Overview

(Theoretical) distribution of sample mean



What is the theoretical distribution of the sample mean?

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

Now the sample mean, $\bar{X} = \frac{1}{n}(X_i + ... + X_n)$, itself is a **random variable** (with its own distribution).

The distribution of $\bar{X}$

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

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

The mean and variance follow from the rules of calculation

The mean of $\bar{X}$:

$$\mathsf{E}[\bar{X}] = \mathsf{E}[\frac{1}{n}X_i + \ldots + \frac{1}{n}X_n] = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu = \mu$$

The mean and variance follow from the rules of calculation

The mean of $\bar{X}$:

$$\mathrm{E}[\bar{X}] = \mathrm{E}[\frac{1}{n}X_i + \ldots + \frac{1}{n}X_n] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu = \mu$$

The variance of $\bar{X}$:

$$\mathrm{V}[\bar{X}] = \mathrm{V}[\frac{1}{n}X_i + \ldots + \frac{1}{n}X_n] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{V}[X_i] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

The mean and variance follow from the rules of calculation

The mean of $\bar{X}$:

$$\mathsf{E}[\bar{X}] = \mathsf{E}[\frac{1}{n}X_i + \ldots + \frac{1}{n}X_n] = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu = \mu$$

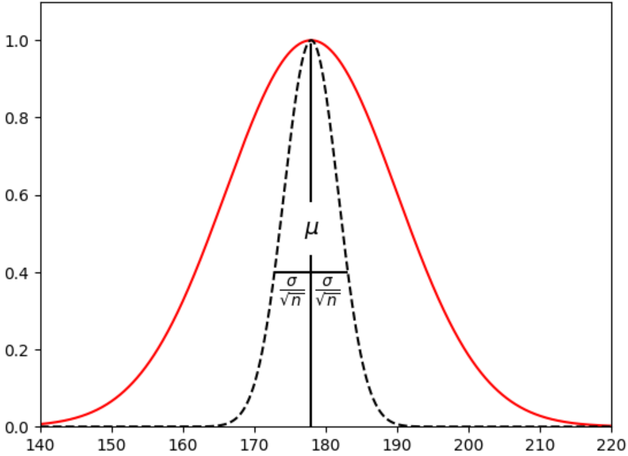
The variance of $\bar{X}$:

$$\mathsf{V}[\bar{X}] = \mathsf{V}[\frac{1}{n}X_i + \ldots + \frac{1}{n}X_n] = \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{V}[X_i] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

Normality of $\bar{X}$:

A linear combination of normally distributed random variables will also be normally distributed (with mean and variance as given above).

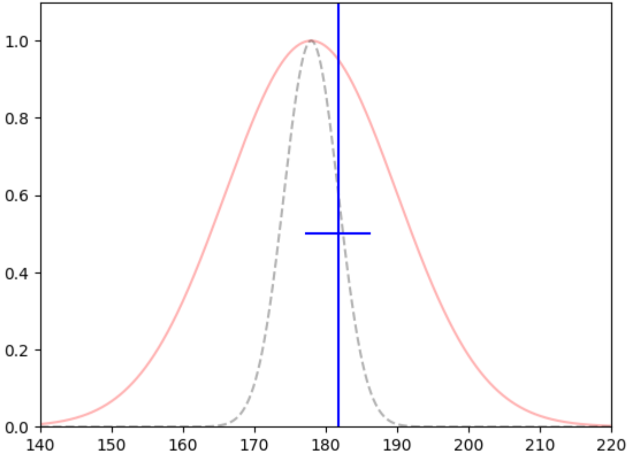
(Theoretical) distribution for the mean



Distribution of sample mean

In the individual case

We have now estimated $\bar{x}$ and s^2 from our sample data.
But we do not know the "true" μ and σ^2.
What do we do then?



How large is the error $(\bar{X} - \mu)$?

Expected average error $E[\bar{X} - \mu]$

$$\mathrm{E}[\bar{X} - \mu] = \mu - \mu = 0$$

How large is the error $(\bar{X} - \mu)$?

Expected average error $E[\bar{X} - \mu]$

$$\mathsf{E}[\bar{X} - \mu] = \mu - \mu = 0$$

Variance of the error $\mathsf{V}[\bar{X} - \mu]$

$$\mathsf{V}[\bar{X} - \mu] = \mathsf{V}[\bar{X}] = \frac{\sigma^2}{n}$$

Standardized version of the above, Theorem 3.4

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

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

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

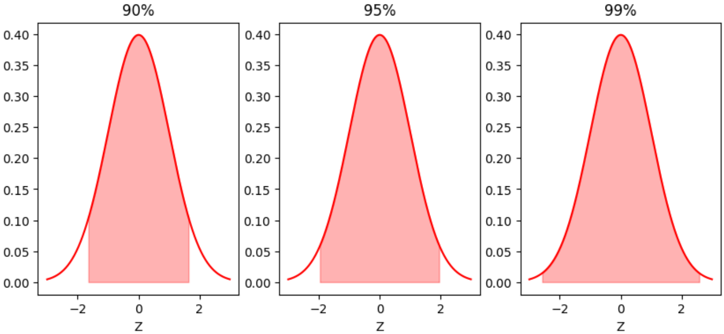
That is, the standardized sample mean Z follows a standard normal distribution.

Kahoot!

(x2)

Z-intervals

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



You need to choose a **"significance level"** (α).
Then, based on the interval boundaries of Z, you can calculate the interval boundaries for $\bar{X} - \mu$ (board).

Practical problem (and solution)

How do we use the results from the previous slides to say something about μ ...

... when the 'true', unknown, population standard deviation σ enters into all the formulas?

Practical problem (and solution)

How do we use the results from the previous slides to say something about μ ...

... when the 'true', unknown, population standard deviation σ enters into all the formulas?

Obvious solution:

Use the estimate s instead of σ in formulas.

Practical problem (and solution)

How do we use the results from the previous slides to say something about μ ...

... when the 'true', unknown, population standard deviation σ enters into all the formulas?

Obvious solution:

Use the estimate s instead of σ in formulas.

BUT:

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

Theorem 3.5, a more applicable extension of the above

Instead we consider the random variable: $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$

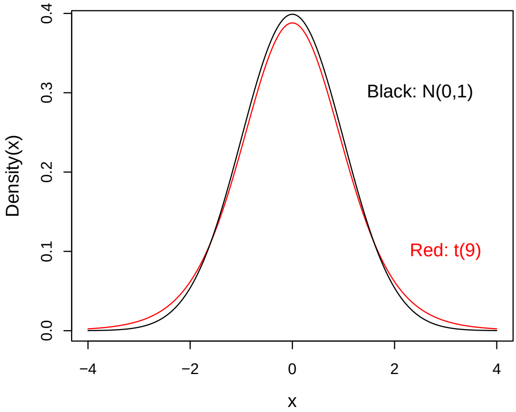
The t-distribution takes the uncertainty of s into account:

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

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

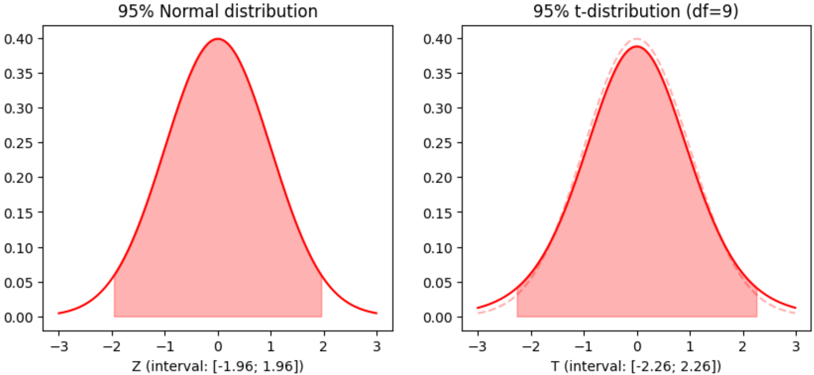
where $t(n-1)$ is the t-distribution with $n-1$ degrees of freedom.

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



https://en.wikipedia.org/wiki/Student's_t-distribution

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

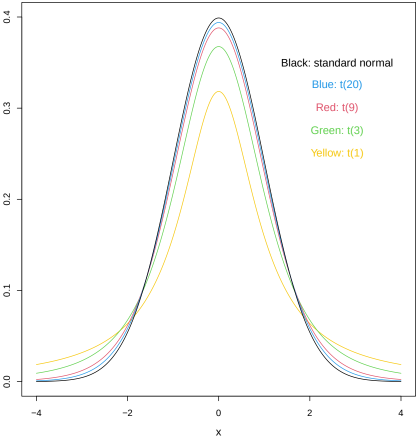


The t-distribution resembles the normal distribution.

In the t-distribution, the probability mass is spread a bit wider.

The width of the t-distribution depends on the number of degrees of freedom (df).

Different t-distributions:



With the increase in sample size or df, (1) the shape of the distribution becomes narrower (2) Tails Become Thinner (3) t-distribution approaches Normal distribution (4) Critical t-values approach Z-values, (5) Confidence intervals become narrower

Python: t-distribution

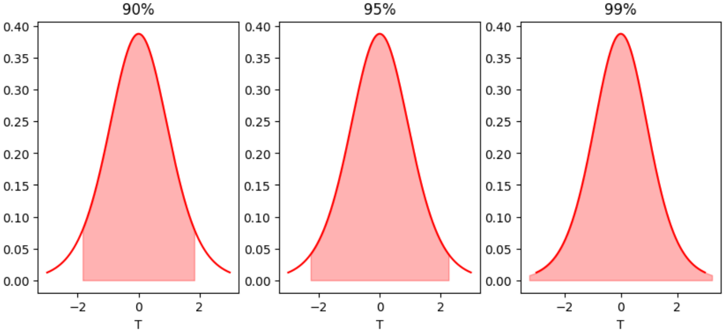
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.t.html

- Go to today's Python notebook in VS Code
  - "t-distribution"

Visual Studio Code

T-intervals

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



Again: We need to choose a **"significance level"** (α), typically, $0.01, 0.05, 0.01$. After that, based on the interval boundaries of Z, you can calculate the interval boundaries for $\bar{X} - \mu$ (on the board).

Overview

Method 3.9: One-sample Confidence Interval (CI) for μ

Use the correct t-distribution to construct the confidence interval:

For a sample $x_1, \ldots, x_n$ 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.

Method 3.9: One-sample Confidence Interval (CI) for μ

Use the correct t-distribution to construct the confidence interval:

For a sample $x_1, \ldots, x_n$ 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:

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

'Repeated sampling' interpretation

If $\alpha = 0.05$:

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

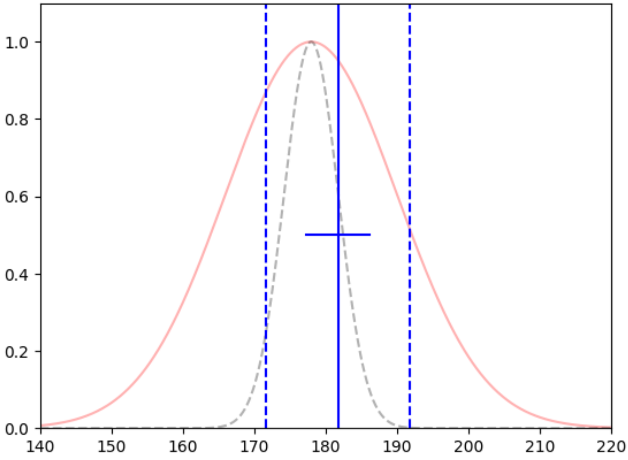
In the long run, we capture the true value in 95% of cases:

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

Kahoot!
(x4)

In the individual case

We have now estimated $\bar{x}$ and s^2 from our sample data.
We do not know the "true" μ and σ^2.
But we can provide a confidence interval for $\hat{\mu}$.



Height Example

- Go to today's Python notebook in VS Code
  - Example: find correct quantiles in t(9)-distribution
  - 95% interval estimate for students' height sample data
  - Visualize the confidence interval

Overview

Python: (Empirical) Distribution of the sample variance

We have talked a lot about the sample **mean**, but what about the sample **variance**?

- Go to today's Python notebook in VS Code
  - "Simulation: Distribution of the sample variance"
- Note
  - The variance is always positive (squared unit) and does not follow a normal distribution (variance is a measure, not a random variable).
  - The distribution of variance is not symmetrical.

Distribution of the sample variance, Theorem 2.81

Assume that $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables, $X_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$.

The sample variance (variance estimate) follows a χ^2 distribution:

Let

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

It holds that:

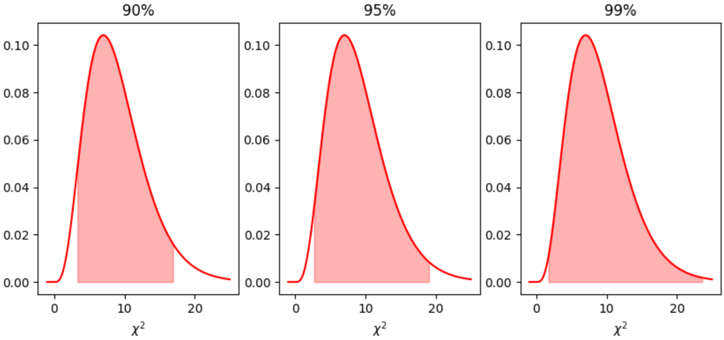
$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

follows a χ^2 distribution with $n-1$ degrees of freedom.

χ^2 is always positive and the χ^2 distribution is not symmetric.

χ^2 distribution

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



Again: We need to choose a **"significance level"** (α).
Then, We can calculate interval bounds for σ^2 from the interval bounds on χ^2 (on the board).

Method 3.19: Confidence intervals for variance and standard deviation

Let $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$ be independent (and identically distributed).

Variance:

A $100(1-\alpha)\%$ confidence interval for the variance σ^2 is given by:

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

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

Method 3.19: Confidence intervals for variance and standard deviation

Let $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$ be independent (and identically distributed).

Variance:

A $100(1-\alpha)\%$ confidence interval for the variance σ^2 is given by:

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

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

Standard deviation:

A $100(1-\alpha)\%$ confidence interval for the standard deviation σ is:

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

Python: Confidence interval for the variance estimate

Kahoot!

- Go to today's Python notebook in VS Code
  - "Example: Variance of student heights"

Notice that the confidence interval for the variance and standard deviation are not symmetric.

Visual Studio Code

Overview

Theorem 3.14: The Central Limit Theorem (CLT)

The distribution of the sample mean approaches a normal distribution as the sample size n becomes sufficiently large, regardless of the shape of the population distribution.

Let $\bar{X}$ be the average of a randomly drawn sample of size n taken from a population with mean μ and variance σ^2. Then, the distribution of

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

approaches the standard normal distribution, $N(0, 1^2)$, as $n \to \infty$.

Theorem 3.14: The Central Limit Theorem (CLT)

The distribution of the sample mean approaches a normal distribution as the sample size n becomes sufficiently large, regardless of the shape of the population distribution.

Let $\bar{X}$ be the average of a randomly drawn sample of size n taken from a population with mean μ and variance σ^2. Then, the distribution of

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

approaches the standard normal distribution, $N(0, 1^2)$, as $n \to \infty$.

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

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

English: *Central Limit Theorem* (CLT)

Python: CLT in action

- Go to today's Python notebook in VS Code
  - "CLT in action"

Visual Studio Code

Consequence of the Central Limit Theorem:

The confidence interval for μ also applies to non-normal data:

Confidence intervals for the mean can be calculated based on the t-distribution in almost all situations, as long as n is "large enough." Why?: Assumptions for the t-test:(1) observations are independent (2) for small samples, the underlying population from which the sample is drawn should be approximately normally distributed. For larger samples, t-test applicable for CLT.

Consequence of the Central Limit Theorem:

The confidence interval for μ also applies to non-normal data:

Confidence intervals for the mean can be calculated based on the t-distribution in almost all situations, as long as n is "large enough." Why?: Assumptions for the t-test:(1) observations are independent (2) for small samples, the underlying population from which the sample is drawn should be approximately normally distributed. For larger samples, t-test applicable for CLT.

When is n "large enough"?

Difficult to give a precise answer, BUT:

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

(Note: CLT applies to **means** – not variance)

Overview

Example: Exam Question from 2016

Exercise IX

A course at a university is offered each semester typically with more than 300 students taking the exam. Examination results for 280 students who have passed the course at the previous exam is given in the table below. For example, the tables shows that 24 students got the grade 12. The distribution of the 280 grades is considered in the next 4 questions.

Grade	02	4	7	10	12	In total
Count	22	78	84	72	24	280

The data (grades) can be loaded into R by:

```
grades = rep(x=c(2,4,7,10,12), times=c(22,78,84,72,24))
```

Question IX.1 (12)

Use the central limit theorem to determine a 95% confidence interval for the mean grade based on the students who have passed the exam. (It is important in this question that the grades are perceived numerically, eg. 02 corresponds to the number 2, etc.).

Python: Exam Question from 2016

- Go to today's Python notebook in VS Code
  - "Example: Exam question from 2016"

Visual Studio Code

Example: Tablet Production

Tablet Production:

In the production of tablets, an active ingredient is mixed with a powder, after which the mixture is formed into tablets. The goal is to produce a homogeneous mixture so that the strength of the tablets is consistent.

We consider a mixture of the active ingredient and filler powder, from which we aim to produce a large number of tablets.

We want the concentration of the active ingredient in the tablets to be 1 mg/g with the least possible variation. A random sample of 20 measurements is taken, each corresponding to what becomes one tablet. In each sample, the concentration of the active ingredient (in mg/g) is measured. Furthermore, we assume that our measurements follow a normal distribution.

Now, we want to estimate the variance of the concentration of the active ingredient and provide a confidence interval for this estimate.

From the sample of 20 measurements (representing 20 tablets), we find that the average concentration is 1.01 mg/g, with a sample standard deviation of 0.07 mg/g.

Python: Confidence Interval for Variance Estimate

Kahoot!

(x5)

- Go to today's Python notebook in VS Code
  - "Example: Production of tablets"

Visual Studio Code

Overview

The formal framework for *statistical inference*

From Chapter 1 of the book:

- An *observational unit* is the single entity/level about which information is sought (e.g. a person) (**Observation Unit**)
- 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**)

The formal framework for *statistical inference*

From Chapter 1 of the book:

- An *observational unit* is the single entity/level about which information is sought (e.g. a person) (**Observation Unit**)
- 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**)

Terminology and concepts:

- μ and σ are *parameters* that describe the population
- $\hat{\mu} = \bar{x}$ is the *estimate* for μ (specific outcome value)
- $\bar{X}$ is the *estimator* for μ (now seen as a random variable)
- The concept of *statistic* is a common term for both ($\bar{x}$ and $\bar{X}$)

The formal framework for *statistical inference* - Example

From Chapter 1 of the book: Modified height example

We measure the height of 10 random individuals in Denmark.

The formal framework for *statistical inference* - Example

From Chapter 1 of the book: Modified height example

We measure the height of 10 random individuals in Denmark.

The sample:

The 10 observations: $x_1, \ldots, x_{10}$.

The formal framework for *statistical inference* - Example

From Chapter 1 of the book: Modified height example

We measure the height of 10 random individuals in Denmark.

The sample:

The 10 observations: $x_1, \ldots, x_{10}$.

The population:

The heights of all people in Denmark.

The formal framework for *statistical inference* - Example

From Chapter 1 of the book: Modified height example

We measure the height of 10 random individuals in Denmark.

The sample:

The 10 observations: $x_1, \ldots, x_{10}$.

The population:

The heights of all people in Denmark.

The observational unit:

One individual.

Statistical inference: Learning from data

Learning from data:

We want to deduce the parameter values for the underlying population.

Statistical inference: Learning from data

Learning from data:

We want to deduce the parameter values for the underlying population.

Important in this regard:

The sample must meaningfully be *representative* of a well-defined population.

Statistical inference: Learning from data

Learning from data:

We want to deduce the parameter values for the underlying population.

Important in this regard:

The sample must meaningfully be *representative* of a well-defined population.

How do you ensure this?

For example, by ensuring that the sample is fully *randomly selected*.

Random sampling

Definition 3.12:

- A random sample from an (infinite) population: The random variables $X_1, X_2, \ldots, X_n$ constitute a random sample of size n from the infinite population if:

    1. All the random variables have the same distribution
    2. The n random variables are independent

Random sampling

Definition 3.12:

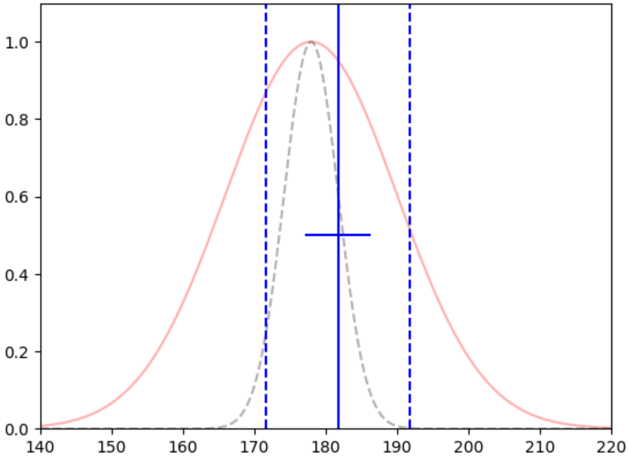
- A random sample from an (infinite) population: The random variables $X_1, X_2, ..., X_n$ constitute a random sample of size n from the infinite population if:

  1. All the random variables have the same distribution
  2. The n random variables are independent

What does this mean?

1. All observations must come from the same population
2. They must NOT share information with each other (e.g. if you had sampled entire families instead of individuals)

Statistical inference: Learning from data



Agenda

1. Introduction to confidence intervals
2. Distribution of the sample mean
    - The t-distribution
3. Confidence interval (CI) for μ
    - Example: Heights
4. Confidence interval for variance and standard deviation
5. The Central Limit Theorem (CLT)
6. Examples
7. Summary of Statistical Inference