MS-A0503 First course in probability and statistics

4B Confidence intervals

Jukka Kohonen

Deparment of mathematics and systems analysis
Aalto SCI

Academic year 2021–2022
Period III

Example. Coffee machine

Background: A coffee machine seems to serve random amounts.
Question: How much does it serve on average?
Study: 25 cups of coffee were taken and their volumes measured.

Data = observed volumes (centiliters):
$\vec{x} = (10.17, 11.23, 9.59, 8.94, 10.14, 9.66, 10.22, 9.59, 11.11, 9.94, 9.76, 9.92, 10.43, 10.05, 9.19, 10, 10.38, 10.02, 10.37, 9.93, 9.97, 10.24, 10.50, 9.38, 9.98)$

Average of these volumes $m(\vec{x}) = 10.0284$. (This we know exactly – assuming no measurement error.)

Question: Can we claim that the "true" average μ is near 10.0284? How near? How strongly can we claim this?

"True" average means the mean of that distribution from which the coffee volumes are drawn, randomly, whenever making a cup.

Why? Because the "true" average helps us understand what happens generally or in the future.

Contents

Data and stochastic model

Data set

From a data source, we have observed values $x_1, \ldots, x_n$. These we know. We want to infer where they came from.

Stochastic model

The possible values of data (that we might obtain when the data source generates n units) are modelled as random variables $X_1, \ldots, X_n$, which are independent, and each X_i has the same probability distribution $f(x)$.

Even if we do not know the distribution $f(x)$,
we think it "is" there and generates the data.

Modelling a "real" data source by a stochastic model

A stochastic model is a mathematical simplification of how the data source "really" works.

- The model may have parameters. If you fix their values, you get a probability distribution, such as $\text{Bin}(10, 0.5)$. Some of the parameters may be known and some unknown.
- A good model is reasonably accurate in telling what values can be generated, and with what probabilities.
- But a good model is also simple enough so that it is possible to calculate with it.
- Typically, we assume we can obtain many numbers from the data source, and that they are independent. (If not, we need a more complicated model.)

Examples – Observe mathematical similarity

Example (Coffee machine)

We assume that whenever the machine fills a cup, the volume is a random variable from an unknown distribution whose mean is μ.
The distribution is meant to model the results of the physical process in the coffee machine (determined by machine design, settings, and random details that we cannot predict exactly).

Example (Sampling from a population)

There are n Finns, and exactly k of them (proportion $p = k/n$) support building more nuclear power plants.
For practical reasons we pick a random Finn. Then his/her support for nuclear power is modeled by an indicator variable $X_1 \sim \text{Ber}(p)$.
The parameter p is a constant, but we do not know its value.
We may also pick more of these random Finns $X_2, X_3, \ldots$.

Lowercase and uppercase (one convention)

Data set $\vec{x} = (x_1, \ldots, x_n)$

- Contains the values that we observed/measured
- To obtain them, we need no stochastic modelling!
- Eg. $(x_1, x_2, x_3) = (10.17, 11.23, 9.59)$, from measuring the first 3 coffee servings.

Stochastic model $\vec{X} = (X_1, \ldots, X_n)$

- Contains random variables, following the distribution (stochastic model) by which we try to predict what the data source can generate
- To obtain this, we need no measurement data!
- Eg. (X_1, X_2, X_3), three independent normally distributed random variables, with mean 10 and standard deviation 5

Statistics of data sets and stochastic models

Recall the "descriptive" statistics from lecture 3B.

A statistic is a function $g : \mathbb{R}^n \to \mathbb{R}$. fi: tunnusluku

(Idea: "a rule that converts n observations into one number")

Example (Some well-known statistics)

- Average $m(\vec{x}) = \frac{1}{n} \sum_{i=1}^{n} x_i$
- Variance $\mathrm{var}(\vec{x}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - m(\vec{x}))^2$
- Standard deviation $\mathrm{sd}(\vec{x}) = \sqrt{\mathrm{var}(\vec{x})}$

If you apply such function to the random vector $\vec{X} = (X_1, \ldots, X_n)$, then you have a random number $g(X_1, \ldots, X_n)$. This is a transformation of a random variable (recall some lectures back).

If the X_i follow a stochastic model, then the laws of probability give you the distribution of $g(X_1, \ldots, X_n)$.

Average from a stochastic model

If data are coming from a stochastic model $\vec{X} = (X_1, \ldots, X_n)$ (note the randomness), such that each X_i has mean μ and standard deviation σ, then ...

... the average $m(\vec{X})$ is a random variable such that

$$\mathbb{E}[m(\vec{X})] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

$$\mathrm{SD}[m(\vec{X})] = \mathrm{SD}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathrm{SD}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sigma\sqrt{n} = \frac{\sigma}{\sqrt{n}}.$$

It seems that perhaps $m(\vec{X})$ is a good way to estimate μ.

Error of an estimate, and its distribution

Stochastic model: $X_1, \ldots, X_n$ independent random numbers with mean μ and standard deviation σ.

Suppose that we are using $m(\vec{X})$ as an estimator for μ. What is the error of the estimate? How is it distributed?

We already know $\mathbb{E}[m(\vec{X})] = \mu$ and $\mathrm{SD}[m(\vec{X})] = \frac{\sigma}{\sqrt{n}}$.

Then by linearity, the error $m(\vec{X}) - \mu$ has mean zero and standard deviation as above.

Let us go one step further. Divide the error by its standard deviation, to get standardized error

$$\frac{m(\vec{X}) - \mu}{\sigma/\sqrt{n}}.$$

By linearity, this quantity has mean=0 and standard deviation = 1.

Why is this useful? We might be able to calculate probabilities for the (standardized) error being small or large.

What about estimating other parameters than μ?

From data, one can calculate several different statistics (by different functions).

For example, if each data point X_i comes from some distribution f, then what is ...

- the distribution of $\max\{X_1, \ldots, X_n\}$?
- the distribution of $\mathrm{sd}\{X_1, \ldots, X_n\}$?

If we want to estimate the "true" standard deviation $\mathrm{SD}(X_i)$ by the observed statistic $\mathrm{sd}(\vec{x})$, we need to understand how the statistic is distributed $\rightarrow$ We need more tools from stochastics (e.g. MS-C1620).

But on this lecture we concentrate in one statistic (sample average), used to estimate one parameter (true mean).

Contents

Example. Coffee machine

The coffee machine is meant to serve 10.0 cl in each cup, on average. We measured the coffee volumes in 25 cups.

Observed volumes (cl):

$$\vec{x} = (10.17, 11.23, 9.59, 8.94, 10.14, 9.66, 10.22, 9.59, 11.11, 9.94, 9.76, 9.92, 10.43, 10.05, 9.19, 10, 10.38, 10.02, 10.37, 9.93, 9.97, 10.24, 10.5, 9.38, 9.98)$$

Observed average is $m(\vec{x}) = 10.03$. Let us try to calculate an interval that (hopefully) contains the true μ.

Point estimates and interval estimates

Let the unknown parameter be θ.

A point estimate for θ is some number $\hat{\theta}$ that is hopefully near the correct value: $\hat{\theta} \approx \theta$.

An interval estimate for θ is some interval $[a, b]$ that hopefully contains the correct value: $[a, b] \ni \theta$.

"Hopefully" and "near" must be defined somehow mathematically (there are different possibilities for this).

- On this lecture, we work with confidence intervals
- Next week another kind: Bayesian credible intervals

Both are interval estimates.

Point estimate for μ (normal model with known σ)

$\vec{x} = (10.17, 11.23, 9.59, 8.94, 10.14, 9.66, 10.22, 9.59, 11.11, 9.94, 9.76, 9.92, 10.43, 10.05, 9.19, 10, 10.38, 10.02, 10.37, 9.93, 9.97, 10.24, 10.5, 9.38, 9.98)$

Stochastic model: $X_1, \ldots, X_{25}$ independent and normally distributed with mean μ and known standard deviation $\sigma = 0.5$

Task: Estimate the parameter μ

Likelihood function

$$f(x_1, \ldots, x_n \mid \mu, \sigma) \;=\; \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$\Longrightarrow$ The maximum-likelihood estimate for μ is

$$m(\vec{x}) \;=\; \frac{1}{n} \sum_{i=1}^{n} x_i \;=\; 10.03$$

This is a point estimate, but how accurate?

Averages from normal model

Normal model: $X_1, \ldots, X_n$ independent, normally distributed with mean μ and standard deviation σ

By linearity of expectation, we know: The "standardized" error

$$\frac{m(\vec{X}) - \mu}{\sigma/\sqrt{n}}$$

has mean 0 and standard deviation 1.

Furthermore, because

- sum of independent normally dist. numbers is also normal, and
- shifted and scaled normal distribution is also normal,

the standardized error follows the standard normal distribution $N(0, 1)$.

Confidence interval for μ (normal model, known σ)

$\vec{x} = (10.17, 11.23, 9.59, 8.94, 10.14, 9.66, 10.22, 9.59, 11.11, 9.94, 9.76, 9.92, 10.43,$
$10.05, 9.19, 10, 10.38, 10.02, 10.37, 9.93, 9.97, 10.24, 10.5, 9.38, 9.98)$

Stochastic model: $X_1, \ldots, X_{25}$ independent and normal with mean μ and standard deviation $\sigma = 0.5$

$$\mathbb{P}(|m(\vec{X}) - \mu| \leq 0.2) = \mathbb{P}\left(\left|\frac{m(\vec{X}) - \mu}{\sigma/\sqrt{n}}\right| \leq \frac{0.2}{0.5/\sqrt{25}}\right) = \mathbb{P}(|Z| \leq 2) \approx 95\%.$$

Thus, we have relatively high probability (95%) that

$$\mu \in [m(\vec{X}) - 0.2, m(\vec{X}) + 0.2]$$

From the observed data $\vec{x}$ we can calculate, for μ,

- a point estimate $m(\vec{x}) = 10.03$
- a confidence interval $m(\vec{x}) \pm 0.2 = [9.83, 10.23]$

Can we now say that $[9.83, 10.23]$ contains μ with probability 95%?
Not quite. . .

Meaning of the confidence interval

$\vec{x} = (10.17, 11.23, 9.59, 8.94, 10.14, 9.66, 10.22, 9.59, 11.11, 9.94, 9.76, 9.92, 10.43, 10.05, 9.19, 10, 10.38, 10.02, 10.37, 9.93, 9.97, 10.24, 10.5, 9.38, 9.98)$

The interval

$$m(\vec{x}) \pm 0.2 \;=\; [9.83, 10.23]$$

is the confidence interval for μ, at confidence level 95%

From the stochastic model $\vec{X}$, we can get different actual data values; from different data, we will compute different confidence intervals.

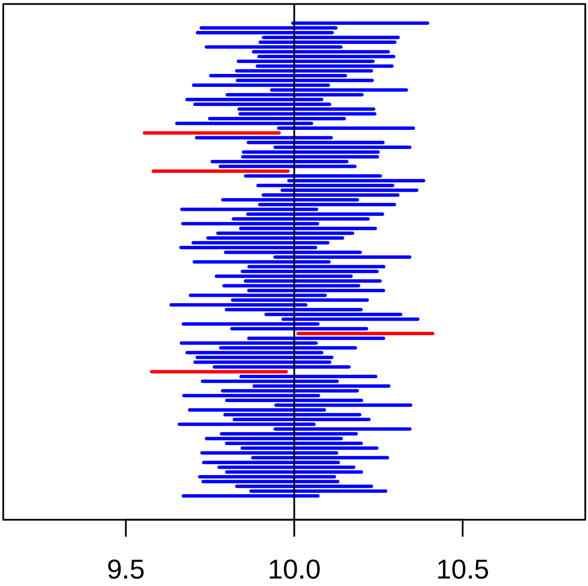
We have 95% probability for the event that our confidence interval will contain the μ:

$$\mathbb{P}(\mu \in [m(\vec{X}) - 0.2, m(\vec{X}) + 0.2]) \;=\; 95\%.$$

If you calculate <u>many</u> confidence intervals from data that come from such data sources, then you know that

- 95% of your confidence intervals <u>will</u> contain the unknown μ (but you do not know which ones)
- 5% of your confidence intervals <u>will not</u> contain the unknown μ (again you do not know which ones)

Confidence intervals, normal model ($\mu = 10, \sigma = 0.5$)



Confidence interval at 99% confidence (normal model)

Normal model: $X_1, X_2, \ldots$ independent and normal,
with unknown mean μ and known std.dev. σ

To determine the confidence interval:

1. Calculate $m(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$
2. Find a number $z > 0$, such that $\mathbb{P}(|Z| \leq z) = 1 - 2\Phi(-z) = 0.99$
   $\implies z = -\Phi^{-1}(\frac{1-0.99}{2}) \approx 2.58$
3. Let the confidence interval for μ be $m(\vec{x}) \pm z\frac{\sigma}{\sqrt{n}}$

Let's check we really now have 99% confidence.
For a random vector $\vec{X} = (X_1, \ldots, X_n)$ from the data source,

$$\mathbb{P}\Big(|m(\vec{X})-\mu| \leq z\frac{\sigma}{\sqrt{n}}\Big) = \mathbb{P}\left(\left|\frac{m(\vec{X})-\mu}{\sigma/\sqrt{n}}\right| \leq z\right) = \mathbb{P}(|Z| \leq z) = 99\%$$

Confidence intervals for μ in normal model: Summary

Normal model: $X_1, X_2, \ldots$ independent and normal
with unknown mean μ and known std.dev. σ

Maximum likelihood estimate for μ is $m(\vec{x})$ Confidence interval is
$m(\vec{x}) \pm z\frac{\sigma}{\sqrt{n}}$

- 95% confidence level, when $z = -\Phi^{-1}(\frac{1-0.95}{2}) \approx 1.96$
- 99% confidence level, when $z = -\Phi^{-1}(\frac{1-0.99}{2}) \approx 2.58$

Example. If $n = 25$, $\sigma = 0.5$, then intervals are:

$$m(\vec{x}) \pm z\frac{\sigma}{\sqrt{n}} = m(\vec{x}) \pm 0.196 \quad \text{(95\% level)}$$
$$m(\vec{x}) \pm z\frac{\sigma}{\sqrt{n}} = m(\vec{x}) \pm 0.258 \quad \text{(99\% level)}$$

Some practical problems:

- What if we do not know σ in advance?
- What if the data source it not a normal distribution?

CI for mean, normal model with unknown σ

Normal model: $X_1, X_2, \dots$ independent and normally distributed, with unknown mean μ and unknown mean σ

Let the confidence interval be $m(\vec{x}) \pm z \, \frac{\mathrm{sd}(\vec{x})}{\sqrt{n}}$, where $\mathrm{sd}(\vec{x})$ is the standard deviation of the sample.

Now for the random vector $\vec{X} = (X_1, \dots, X_n)$ from the model,

$$\mathbb{P}\Big(|m(\vec{X}) - \mu| \le z \frac{\mathrm{sd}(\vec{x})}{\sqrt{n}} \Big) \;=\; \mathbb{P}\left(\left| \frac{m(\vec{X}) - \mu}{\mathrm{sd}(\vec{X})/\sqrt{n}} \right| \le z \right) \;=\; ?$$

Trouble: $\frac{m(\vec{X})-\mu}{\mathrm{sd}(\vec{X})/\sqrt{n}}$ does not follow a normal distribution

Solution:

- If large data (n big), it is approximately normal
- If small data, instead of $\mathrm{sd}(\vec{x})$, use sample std.dev. $\mathrm{sd}_s(\vec{x})$ and take $z = -F^{-1}_{t,n-1}(\frac{1-0.99}{2})$ from the t distribution (this is the true distribution)

Contents

Estimating the mean of a general stochastic model

General model: $X_1, X_2, \ldots$ independent with unknown mean μ, from some distribution (e.g. uniform, exponential)

For μ we can use point estimate $m(\vec{x})$. It may not be maximum-likelihood, but it is unbiased. (Recall Ex. 4B)

By the CLT, $m(\vec{X})$ is approximately normal, so:
Determining an approximate confidence interval for μ:

1. From the data, calculate mean $m(\vec{x})$ and standard deviation $\mathrm{sd}(\vec{x})$
2. Determine number $z > 0$, such that
   $\mathbb{P}(|Z| \leq z) = 1 - 2\Phi(-z) = 0.99$
   $\implies z = -\Phi^{-1}(\frac{1-0.99}{2}) \approx 2.58$
3. Let the confidence interval be $m(\vec{x}) \pm z\frac{\mathrm{sd}(\vec{x})}{\sqrt{n}}$

For large data sets (n big) we have $\mathrm{sd}(\vec{X}) \approx \sigma$, and

$$\mathbb{P}\Big(|m(\vec{X}) - \mu| \leq z\frac{\mathrm{sd}(\vec{X})}{\sqrt{n}}\Big) \approx \mathbb{P}\left(\left|\frac{m(\vec{X}) - \mu}{\sigma/\sqrt{n}}\right| \leq z\right) \approx \mathbb{P}\left(|Z| \leq z\right) = 99\%.$$

Contents

Binary model for a data source

Binary model for a data source:
$X_1, X_2, \dots$ independent and $\{0, 1\}$-valued random variables with unknown mean p

The one parameter p determines fully the distribution of X_i:

$$\mathbb{E}(X_i) = 0 \cdot \mathbb{P}(X_i = 0) + 1 \cdot \mathbb{P}(X_i = 1) = \mathbb{P}(X_i = 1),$$

so X_i has distribution

$$f_p(k) = \begin{cases} 1 - p, & k = 0, \\ p, & k = 1, \\ 0, & \text{muuten.} \end{cases}$$

This is also called the Bernoulli distribution with parameter p, and denoted $\mathrm{Ber}(p)$, or $\mathrm{Bin}(1, p)$.

Example: Opinion poll

From the U.S. voting population, a random sample of $n = 2000$ persons were asked whether they are going to vote for Trump (0=No, 1=Yes).

The random variable $\vec{X} = (X_1, \ldots, X_{2000})$ roughly follows the binary model, with parameter p, where

$$p \;=\; \mathbb{E}(X_i) \;=\; \mathbb{P}(X_i = 1)$$

is the (unknown) proportion of Trump-voters in the population.

Task: Determine a point estimate and a 95% confidence interval for the proportion p.

From last lecture: The relative frequency of ones in the dataset, $\hat{p} = \hat{p}(\vec{x})$, is a maximum-likelihood estimate for p.

Confidence interval for binary model

Binary model for a data source:
$X_1, X_2, \ldots$ independent and $\{0,1\}$-valued random variables with unknown mean p

Because $p = \mathbb{E}(X_i)$, we are in fact estimating the mean, so let us apply the general method to our special case.

1. From data, calculate mean $m(\vec{x})$ and standard deviation $\mathrm{sd}(\vec{x})$
2. Determine number $z > 0$, such that
   $\mathbb{P}(|Z| \leq z) = 1 - 2\Phi(-z) = 0.95$
   $\implies z = -\Phi^{-1}(\frac{1-0.95}{2}) \approx 1.96$
3. Let the confidence interval be $m(\vec{x}) \pm z\frac{\mathrm{sd}(\vec{x})}{\sqrt{n}}$

On the next slide we will simplify the formula even further.

Confidence interval for binary model

The confidence interval for the parameter p is

$$m(\vec{x}) \pm z\frac{\mathrm{sd}(\vec{x})}{\sqrt{n}}$$

But observe that, for a dataset of zeros and ones only,

$$\text{Mean} \quad m(\vec{x}) \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \;=\; \frac{\#\{i : x_i = 1\}}{n} \;=\; \hat{p}$$

$$\text{Variance} \quad \mathrm{var}(\vec{x}) \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{p})^2 \;=\; \ldots \;=\; \hat{p}(1-\hat{p})$$

$$\text{Standard deviation} \quad \mathrm{sd}(\vec{x}) \;=\; \sqrt{\mathrm{var}(\vec{x})} \;=\; \sqrt{\hat{p}(1-\hat{p})}$$

Thus the confidence interval is simply

$$\hat{p} \pm z\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

where $\hat{p}$ is the relative frequency of ones in the sample.

Confidence interval for binary model — Summary

Binary model: $X_1, X_2, \ldots$ independent and $\{0,1\}$-valued random variables with unknown mean p

To find the (approximate) confidence interval for p (when n big):

1. From the data, compute the relative frequency of ones $\hat{p} = \hat{p}(\vec{x})$
2. Find a number $z > 0$, such that $\mathbb{P}(|Z| \leq z) = 1 - 2\Phi(-z) = 0.95$ $\implies z = -\Phi^{-1}(\frac{1-0.95}{2}) \approx 1.96$
3. Let the confidence interval be $\hat{p} \pm z\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$

Variant: "Conservative" confidence intervals

Binary model for a data source:
$X_1, X_2, \dots$ independent and $\{0,1\}$-valued random variables with unknown mean p

Sometimes we want to decide the length of the confidence interval before we have the data. Or we want to apply the same interval to several different estimates (e.g. different parties).

For "conservative" confidence intervals, replace $\sqrt{\hat{p}(1-\hat{p})}$ with

$$\max_{\hat{p}\in[0,1]} \sqrt{\hat{p}(1-\hat{p})} \;=\; \sqrt{\frac{1}{2}\left(1-\frac{1}{2}\right)} \;=\; 0.5.$$

To find a conservative confidence interval for p,

1. From data, find relative frequency of ones $\hat{p}$
2. Find a number $z > 0$, such that $\mathbb{P}(|Z| \le z) = 1 - 2\Phi(-z) = 0.95$ $\implies z = -\Phi^{-1}(\frac{1-0.95}{2}) \approx 1.96$
3. Let the confidence interval be $\hat{p} \pm z\frac{0.5}{\sqrt{n}}$

Conservative confidence intervals

Binary model: $X_1, X_2, \ldots$ independent and $\{0,1\}$-valued numbers with unknown mean p

The (approximate) conservative confidence interva for p is

$$\hat{p} \pm z \frac{0.5}{\sqrt{n}}.$$

- 95% confidence, when $z = -\Phi^{-1}(\frac{1-0.95}{2}) \approx 1.96$
- 99% confidence, when $z = -\Phi^{-1}(\frac{1-0.99}{2}) \approx 2.58$

Margin of error in opinion polls

Opinion polls commonly report margin of error (MOE), for example MOE=1.5%. This usually refers to half-length of the confidence interval.

Example. Point estimate $\hat{p} = 12.0\%$, margin of error 1.5%.
This means confidence interval is $[12.0-1.5,\ 12.0+1.5] = [10.5, 13.5]$.

Some points to consider ...

- Confidence level not always reported (most often 95%).
- The margin of error measures only the sampling error, caused by the random sampling. There may be other sources of error.
- Sometimes hard to remember what is the probability involved
- To calculate the length of a conservative, approximate CI at 95% confidence, $\hat{p}(\vec{x}) \pm 1.96 \times \frac{0.5}{\sqrt{n}}$, all we need to know is n:
  - $n = 1000 \implies \text{MOE} \approx 3\% \implies$ interval is $\hat{p}(\vec{x}) \pm 3\%$
  - $n = 2000 \implies \text{MOE} \approx 2\% \implies$ interval is $\hat{p}(\vec{x}) \pm 2\%$
  - $n = 9000 \implies \text{MOE} \approx 1\% \implies$ interval is $\hat{p}(\vec{x}) \pm 1\%$

Margin of error — What it tells

Remember that MOE measures only the “sampling error”, caused by the fact that we did not ask everyone, but only a random sample of the population.

There may be other sources of “error” between what we observe and what we want to know, e.g. ...

- we did the sampling wrong (not uniform from population) (1936 US presidential election: George Gallup’s 50 000 uniform sample vs. Literary Digest’s 2-million nonuniform)
- we measured what they say but we are trying to understand how they would vote now
- we measured the situation now but we are trying to know how they will vote 2 months later (population is changing)

The MOE of the sampling process says nothing about the probability such other “errors”.

Next lecture is about Bayesian inference. . .