Statistical Models

N. Torelli, L. Egidi, G. Di Credico

Spring 2020

University of Trieste

The concept of statistical model

Simulation from a statistical model

what the fuck is it?

The problems of statistical inference: an overview

The concept of statistical model

Intro: Aim of statistical inference

Statistics aims to **extract information from data**, and in particular on the process that generated the data.

Two intrinsic difficulties:

- It may be hard to infer what we wish to know from the data available;
- Most data contain some **random variability**: by replicating the data-gathering process several times we would obtain different data on each occasion. Noise ....

We search for conclusions drawn from a single data set that are **generally valid**, and not the result of random peculiarities of that data set.

like what is the random generating process

Role of statistical models

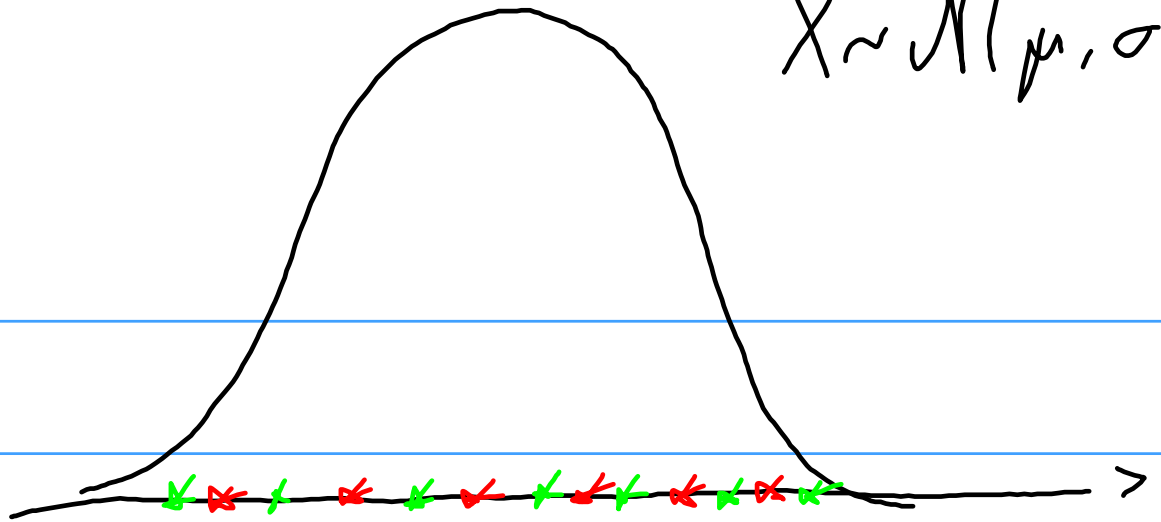
Statistics is able to draw conclusions from random data mainly though the use of **statistical models**.

A statistical model can be thought as a *mathematical cartoon* describing how our data might have been generated, if the unknown features of the data-generating process were actually known.

If the unknowns were known, a good model *can generate data* resembling the main features of observed data.

The purpose of **statistical inference** is to use the statistical model to go in the *reverse direction*: to infer the model unknowns that are consistent with the observed data.

$X \sim N(\mu, \sigma^2)$



set 1 set 2

I observe one ~~data~~ set and try to recover the features of the r. process I'm studying

Mathematical aspects

Notation:

- $\mathbf{y}$ random vector containing the observed data
- θ vector of parameters of unknown value

We assume that knowing the parameters would answer the question of interest about the process generating the data.

The model specifies how data akin to $\mathbf{y}$ may be simulated, implicitly defining the **distribution** of $\mathbf{y}$ and how it depends on θ.

Moreover, a statistical model may depend on some known parameters γ and some further data $\mathbf{x}$, treated as known and denoted as *covariates* or *predictor variables*.

Visually

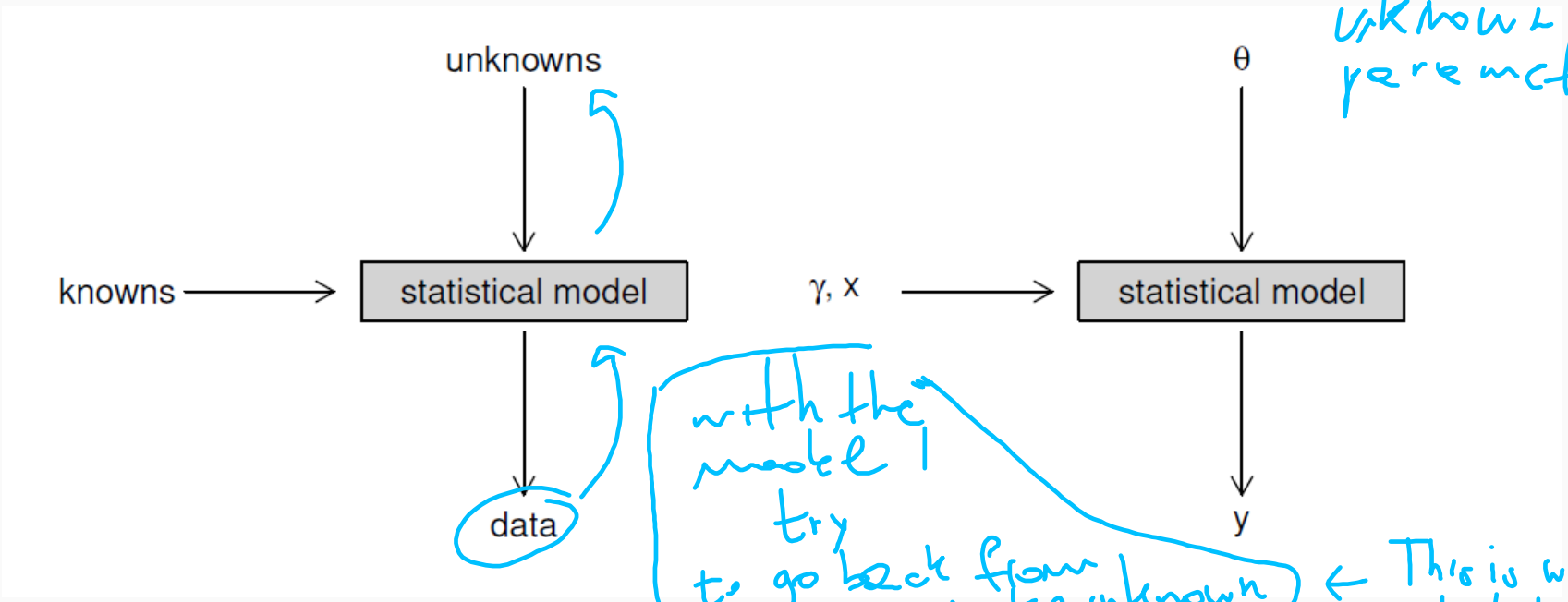


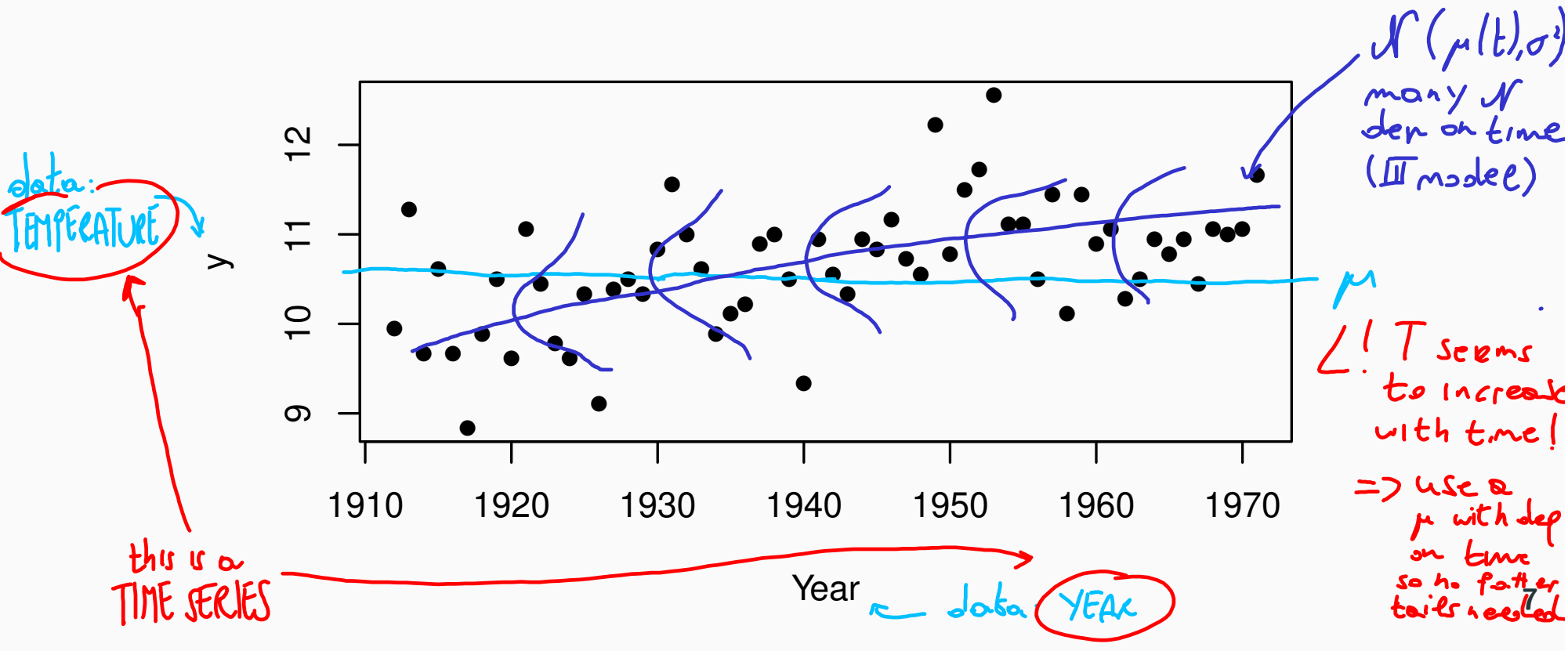
Figure 1: Taken from Wood's book, page 20

y has known distribution, but we don't know which one
→ from data we infer this distr

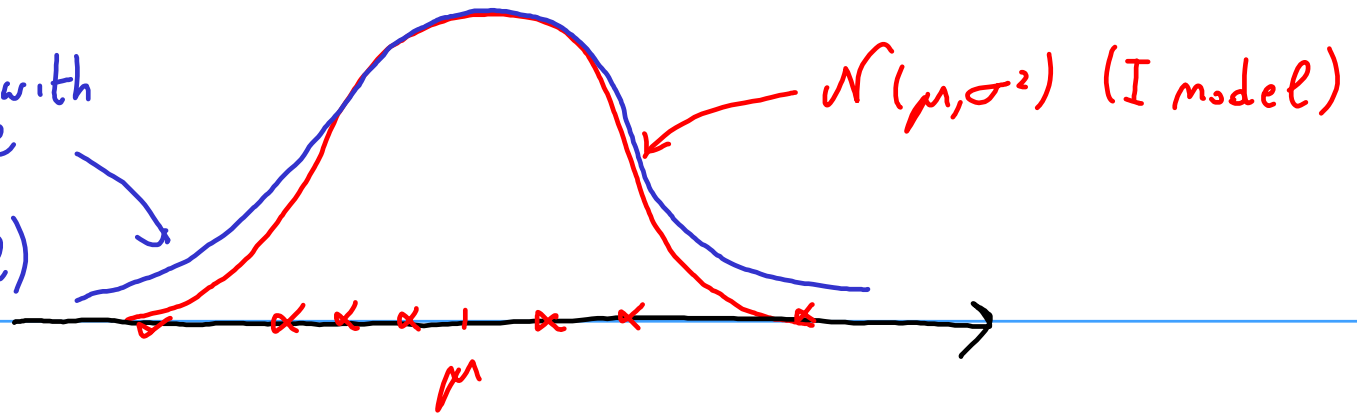
An example

Consider the following record of 60 mean annual temperatures in New Haven, expressed in $^\circ C$

```
y <- (nhtemp - 32) / 1.8
plot(1912:1971, y, pch = 16, xlab = "Year", ylab = "y")
```



μ may be a certain value but it's evident from data that they are not distributed for the majority around the estimated μ but spread across several values of y => fatter tails may describe better the situation



Not happy with data: many points are away from mean => Probably I need FATTER TAILS

II Model: we take μ fixed and study how data move around it → see correlation between data "today" and data "yesterday" → AUTOCORRELATION

$$\varepsilon_t = \phi \varepsilon_{t-1} + \delta i$$

MODEL FOR AUTOCORRELATED TIME SERIES

Example: Model 1

A first model simply assumes that the data are a random sample from a normal distribution namely they are the observation of i.i.d. r.v. from $\mathcal{N}(\mu, \sigma^2)$.

It follows that the distribution for the entire data vector is the product of the single contributions

$$f(\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sigma} \phi\{(y_i - \mu)/\sigma\} ,$$

We use INDEP VAR assumption to write tot distr as product of sngle distributions

where ϕ is the $\mathcal{N}(0,1)$ p.d.f.

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

Example: Model 2

A second model retains the random sample assumption, but replaces the normal distribution with a heavier-tailed t_5 distribution, assuming

$$\frac{Y_i - \mu}{\sigma} \sim t_5 \, .$$

The distribution of the data becomes

$$f(\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sigma} f_{t_5} \left\{ (y_i - \mu)/\sigma \right\} \, ,$$

where f_{t_5} is the t_5 p.d.f.

Example: Model 3

The third model relaxes the assumption of identical distribution, assuming a linear trend over time: after setting $t_i = \texttt{year}_i - 1911$, $i = 1, \ldots, 60$; we then take

$$y_i = \beta_0 + \beta_1 \, t_i + \varepsilon_i \,, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \,.$$

$\Rightarrow E(y_i) = \beta_0 + \beta_1 E(t_i) + \underbrace{E(\varepsilon_i)}_{=0} = \beta_0 + \beta_1 E(t_i) = \beta_0 + \beta_1 t_i$

The independence between observations still holds, so that

No assumed dependencies between y_i, y_{i+1}

$$f(\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sigma} \, \phi\{(y_i - \beta_0 - \beta_1 \, t_i)/\sigma\} \,.$$

Example: Model 4

The last model maintains the trend assumption, but also includes *autocorrelation* for the error term, meaning that we assume

$$y_i = \beta_0 + \beta_1 t_i + \varepsilon_i\,, \qquad \varepsilon_i = \rho\,\varepsilon_{i-1} + v_i\,,$$

with $v_i \sim \mathcal{N}(0, \sigma^2)$, and the autocorrelation $\rho \in (0, 1)$.

The model also requires to specify the distribution for the first observation, here taken as $Y_1 \sim \mathcal{N}\{\beta_0, \sigma^2/(1-\rho^2)\}$, so that all the variables in the sample have the same variance.

Example: Model 4 (cont'd.)

The model is an instance of a **linear regression model with autocorrelated errors**. The r.v. of the sample are not longer independent, yet the distribution of $\mathbf{Y}$ can be found with some algebra.

It is possible to verify that $\mathbf{Y}$ is multivariate normal, with mean vector given by the linear trend

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 \, t_i \, ,$$

and covariance matrix

$$\mathbf{\Sigma} = \frac{\sigma^2}{(1-\rho^2)} \begin{pmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{pmatrix} ,$$

so that $f(\mathbf{y}) = \phi_n(\mathbf{y}; \boldsymbol{\mu}, \mathbf{\Sigma})$, being ϕ_n the multivariate normal p.d.f.

Example: model parameters

It is useful to write down the vector parameters $\boldsymbol{\theta}$ for each of the four model specifications proposed:

- Model 1: $\boldsymbol{\theta} = (\mu, \sigma^2)$
- Model 2: $\boldsymbol{\theta} = (\mu, \sigma^2)$
- Model 3: $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$
- Model 4: $\boldsymbol{\theta} = (\beta_0, \beta_1, \rho, \sigma^2)$

Note that the meaning of each parameter depends on the chosen model: $\sigma^2 = \mathrm{var}(Y_i)$ in Model 1, but $\sigma^2 = 0.6\ \mathrm{var}(Y_i)$ in Model 2.

Simulation from a statistical model

Simulation from a statistical model

A decent model would allow to simulate data sets reproducing some of the features of the observed data, with better models providing more realistic results.

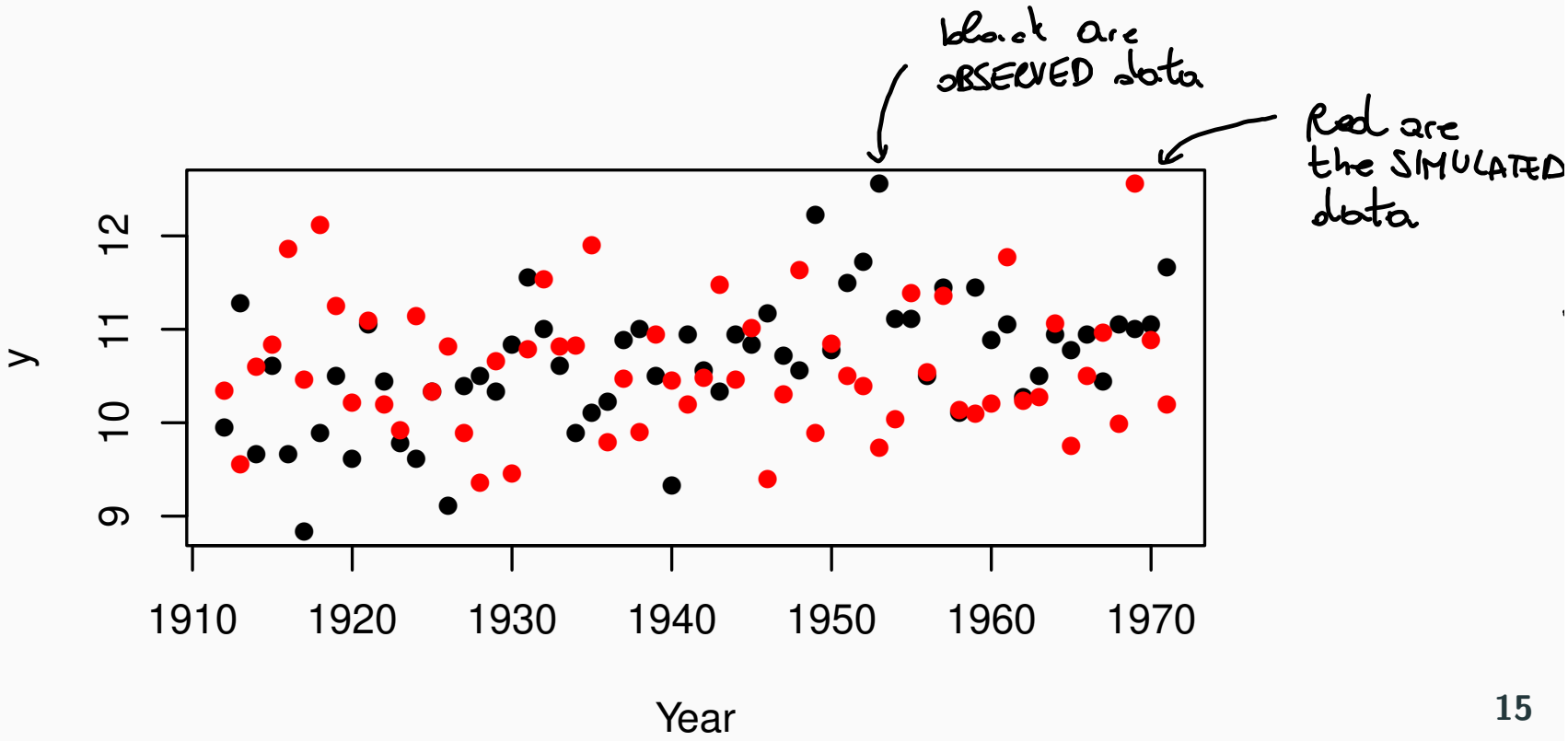
Simulation is an essential part of modern statistical inference. Its role is not only for the assessment of a candidate statistical model, but also to obtain **predictions** based on a chosen model.

Simulation requires that a value for the model parameters θ is chosen beforehand. This task is accomplished by **parameter estimation**, which would be illustrated later on.

Example: Model 1

For Model 1, the parameters μ and σ^2 are readily estimated by $\overline{y}$ and s^2. Then, a further dataset can be simulated using such values

```
set.seed(2018); ysim <- rnorm(length(y), m = mean(y), s = sd(y))
plot(1912:1971, y, pch = 16, xlab = "Year", ylab = "y")
points(1912:1971, ysim, col = 2, pch = 16)
```



Example: what should we look for?

In order to evaluate whether the simulated dataset is similar to the observed one, we should focus on some important features.

For example, climate changes over time may suggest that the temperature of a given year may be positively correlated with the temperature of the subsequent year, an example of *positive autocorrelation*.

We can quantify this point by computing the **sample autocorrelation**

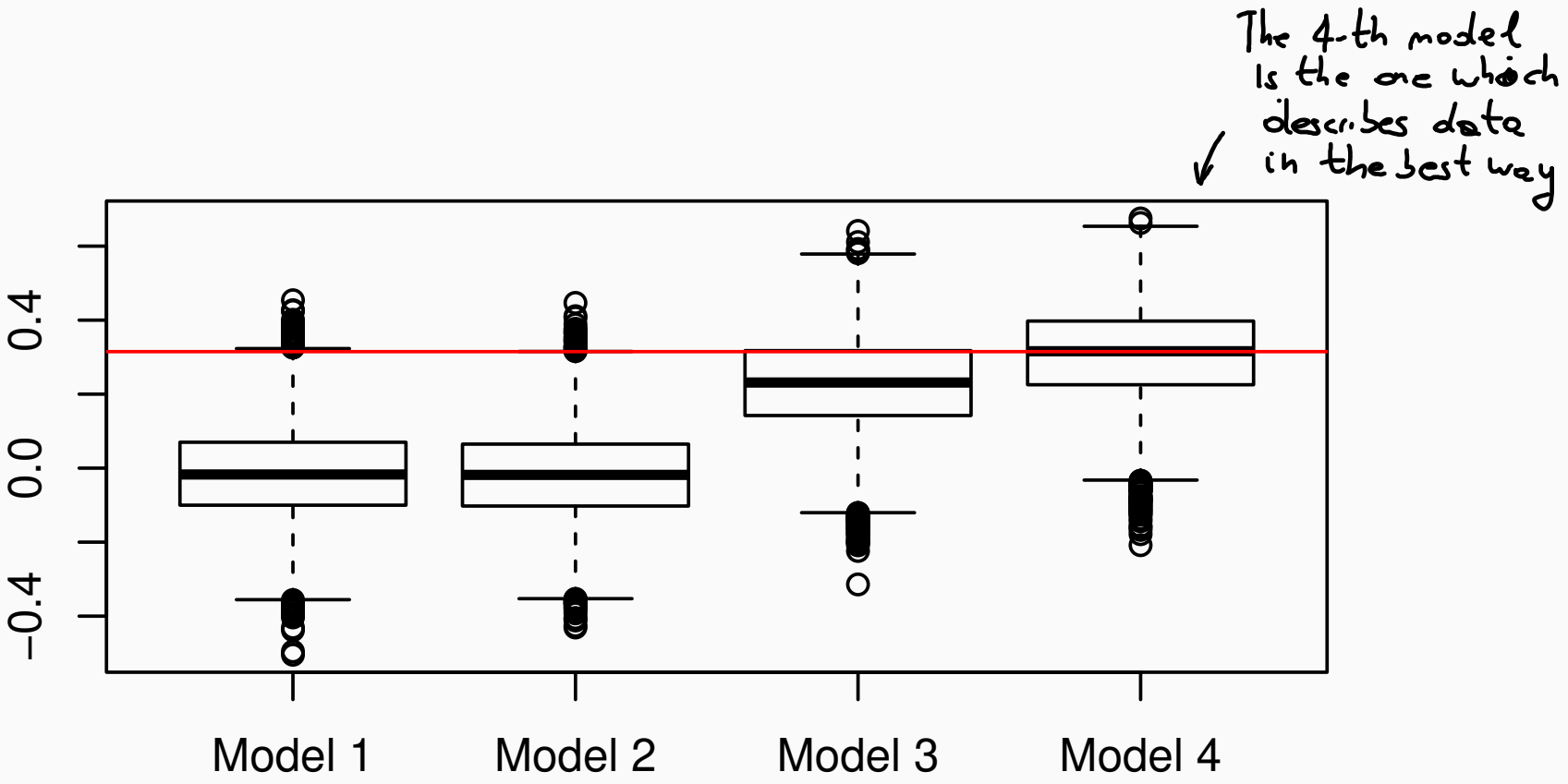
$$r_1 = \frac{\sum_{i=1}^{n-1}(y_i - \overline{y})\,(y_{i+1} - \overline{y})}{\sum_{i=1}^{n}(y_i - \overline{y})^2},$$

which is computed by the R function `acf`.

For the original data set $r_1 = 0.31$, whereas for the simulated data from Model 1 $r_1 = -0.12$. This is just a single data set, though.

Example: Simulated sample autocorrelation

We simulate 10,000 samples from each of the four models, and each time we compute the r_1 coefficient. The sample distributions obtained are displayed in the plot below. Model 4 is better at reproducing autocorrelation, as expected.



The problems of statistical inference: an overview

Inferential questions

Given a statistical model for data $\mathbf{y}$, with model parameters $\boldsymbol{\theta}$, there are some basic questions to ask (pasted from the CS book):

1. What values of $\boldsymbol{\theta}$ are most consistent with $\mathbf{y}$? [*Point estimation*]
2. What range of values of $\boldsymbol{\theta}$ are consistent with $\mathbf{y}$? [*Interval estimation*]
3. Is some prespecified restriction on $\boldsymbol{\theta}$ consistent with $\mathbf{y}$? [*Hypothesis testing*]
4. Is the model consistent with the data for any values of $\boldsymbol{\theta}$ at all? [*Model checking*]

Question 4 can be enlarged to include *which of several alternative models is most consistent with* $\mathbf{y}$? This is point of *model selection*, which partially overlaps with model checking.

The central issue is the acknowledgment of the **intrinsic uncertainty** inherent in trying to learn about $\boldsymbol{\theta}$.

A further question

For settings where some control over the data-gathering process is possible, a further question arises:

5. How might the data-gathering process be organized to produce data that enables answers to the preceding questions to be as accurate and precise as possible?

This is the core of *experimental and survey design methods*.

It represents an often neglected question, of central importance in many traditional fields where statistics is routinely applied (medical sciences, industrial research, biosciences . . .). It is also very relevant for business and web analytics, like in *A/B testing*.

Approaches to statistical inference

There two classes of methods providing an answer to questions 1-4, namely the **frequentist** and **Bayesian** approach.

They differ mainly for the role of model parameters θ, which are treated as fixed constants in the former approach and as r.v. in the latter one.

The difference may appear remarkable, and there has been controversy over the years about the merits of each approach.

Yet, from a *a practical perspective* the two approaches have much in common, and tend to give similar answers when properly applied, especially when compared to approaches that are not based on a statistical model.

Roadmap

In the rest of this course, a brief overview of classical frequentist methods for point estimation, interval estimation, hypothesis testing and design will be provided. The important idea of the bootstrap will be also illustrated.

Afterwards, the most important frequentist class of methods, given by **likelihood-based methods**, will be covered. This is rather comprehensive methodology, that provides also some tools for model selection.

Model checking will be illustrated with reference to some specific class of statistical models, such as **linear and generalized linear regression models**, whose theory will be covered in the course as well. We will skim over some important extensions, such as **nonparametric regression and mixed models**.

Some (limited) space will be devoted also to the main ideas of the Bayesian approach.

A first look at model diagnostics

Model diagnostics, a basic tool for model checking, it also has a role for simple models, like those of our illustrative example.

A basic tool is given by **quantile-quantile plots**, already briefly introduced, which can be used to verify whether the data **y** are consistent with an assumed model.

This is straightforward for i.i.d. models, like Model 1 and 2, where the fact that the assumed distribution for y_i depends on μ and σ is rather inconsequential.

↑
data spread in a way that does not reveal a manifest dependence on data

A first look at model diagnostics (cont'd.)

For more complex settings, such as Model 3 and 4, the general idea is as follows.

Assume that according to the fitted model the expected value and covariance matrix of $\mathbf{y}$ are $\mu_{\widehat{\theta}}$ and $\mathbf{\Sigma}_{\widehat{\theta}}$. ← values calculated by using the assumed model

Then the **standardized residuals** are

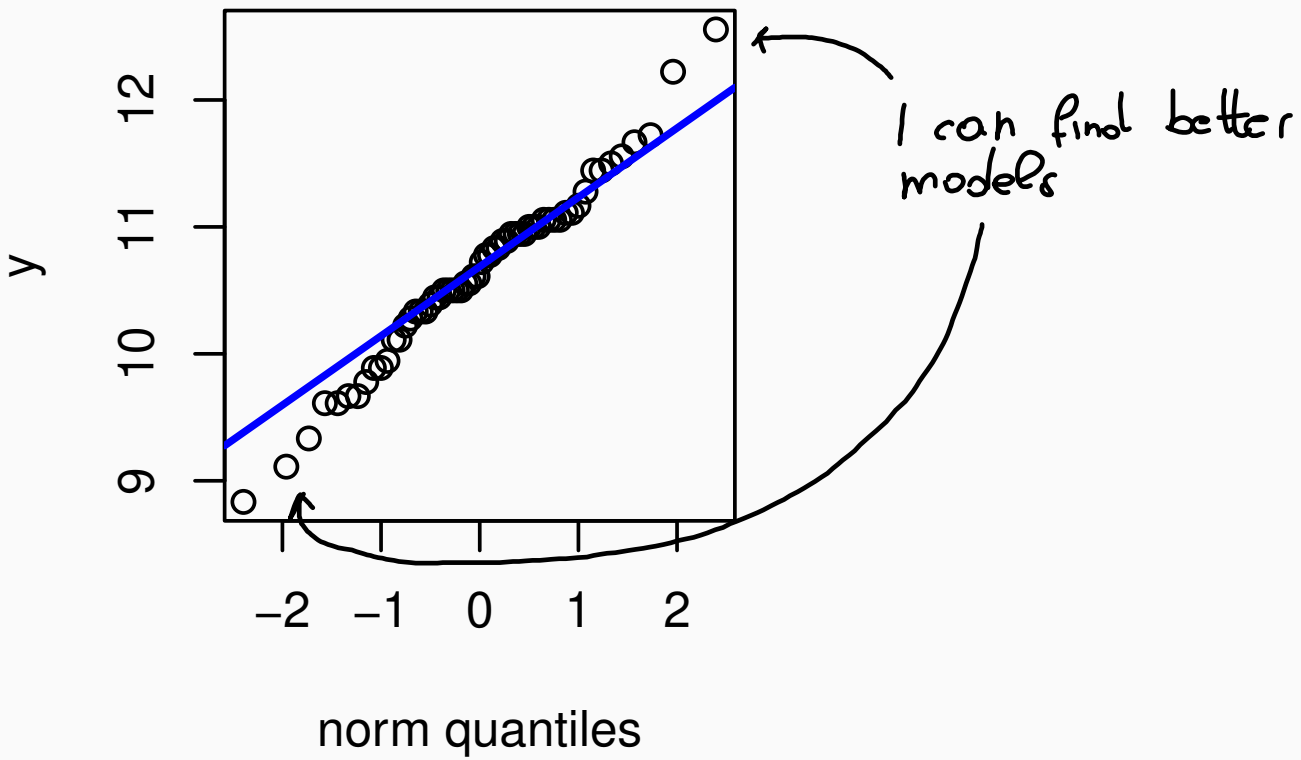
$$\widehat{\varepsilon} = \mathbf{\Sigma}_{\widehat{\theta}}^{-1/2} \left(\mathbf{y} - \mu_{\widehat{\theta}} \right),$$

where $\mathbf{\Sigma}_{\widehat{\theta}}^{-1/2}$ is any matrix *square root* of $\mathbf{\Sigma}_{\widehat{\theta}}^{-1}$, such as its Choleski factor.

If the model is correct, $\widehat{\varepsilon}$ should appear approximately independent, with zero mean and unit variance, and roughly normal if the model assumes normality.

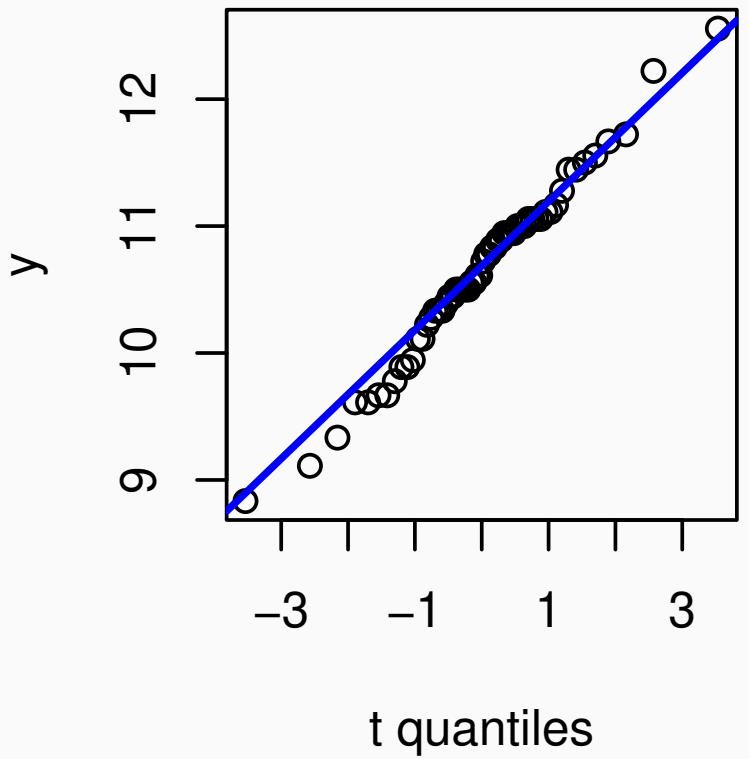
Example: model checking for Model 1

```
par(pty="s")
library(car)
qqPlot(y, dist="norm", envelope=FALSE, grid=FALSE, id=FALSE)
```



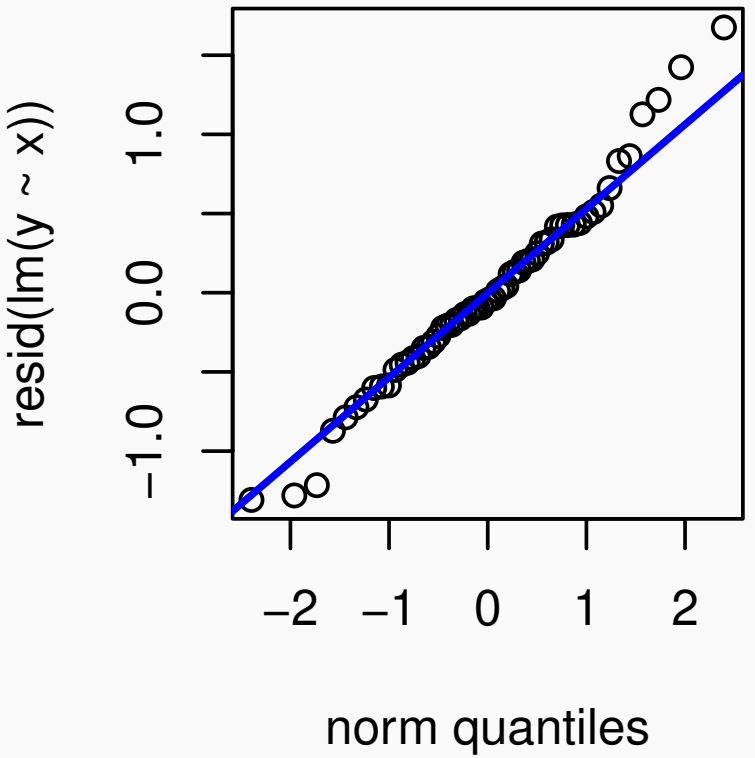
Example: model checking for Model 2

```
par(pty="s")
qqPlot(y, dist="t", df=5, envelope=FALSE, grid=FALSE, id=FALSE)
```



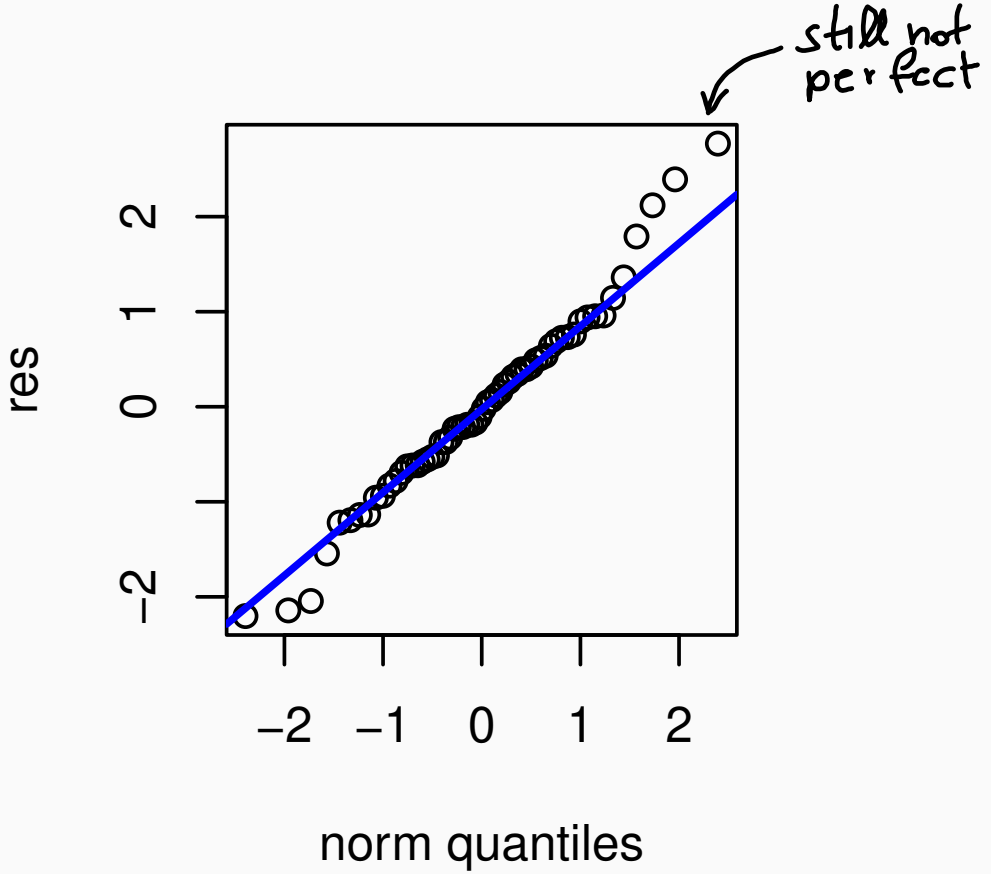
Example: model checking for Model 3

```
par(pty="s")
x <- 1912:1971-1911
qqPlot(resid(lm(y~x)), envelope=FALSE, grid=FALSE, id=FALSE)
```



Example: model checking for Model 4

```
par(pty="s")
qqPlot(res, dist="norm", envelope=FALSE, grid=FALSE, id=FALSE)
```



Example: winding up

A model should reproduce a statistical process using FEW parameters: too many pars leads me to OVERFITTING. THE SIMPLER THE MODEL, THE BETTER

The example shows that no model gives a perfect fit for this data set, a fact that we ought to accept in broad generality.

Model 3 and Model 4 both provide an acceptable fit, with the latter slightly better in reproducing some of the autocorrelation observed in data.

↖ because the control points lie better on a strait line. other models have central points not distributed as linearly as the one of the 4th model

More sophisticated models may give better results, but **simple models** conform to the **Occam's Razor principle**, that for statistical modelling argues in favor of *simple models for simple problems*, moving to more complex models when simple models are inappropriate.