# Understanding scientific evidence: Storytelling with data

Albert Sorribas

Ester Vilaprinyo

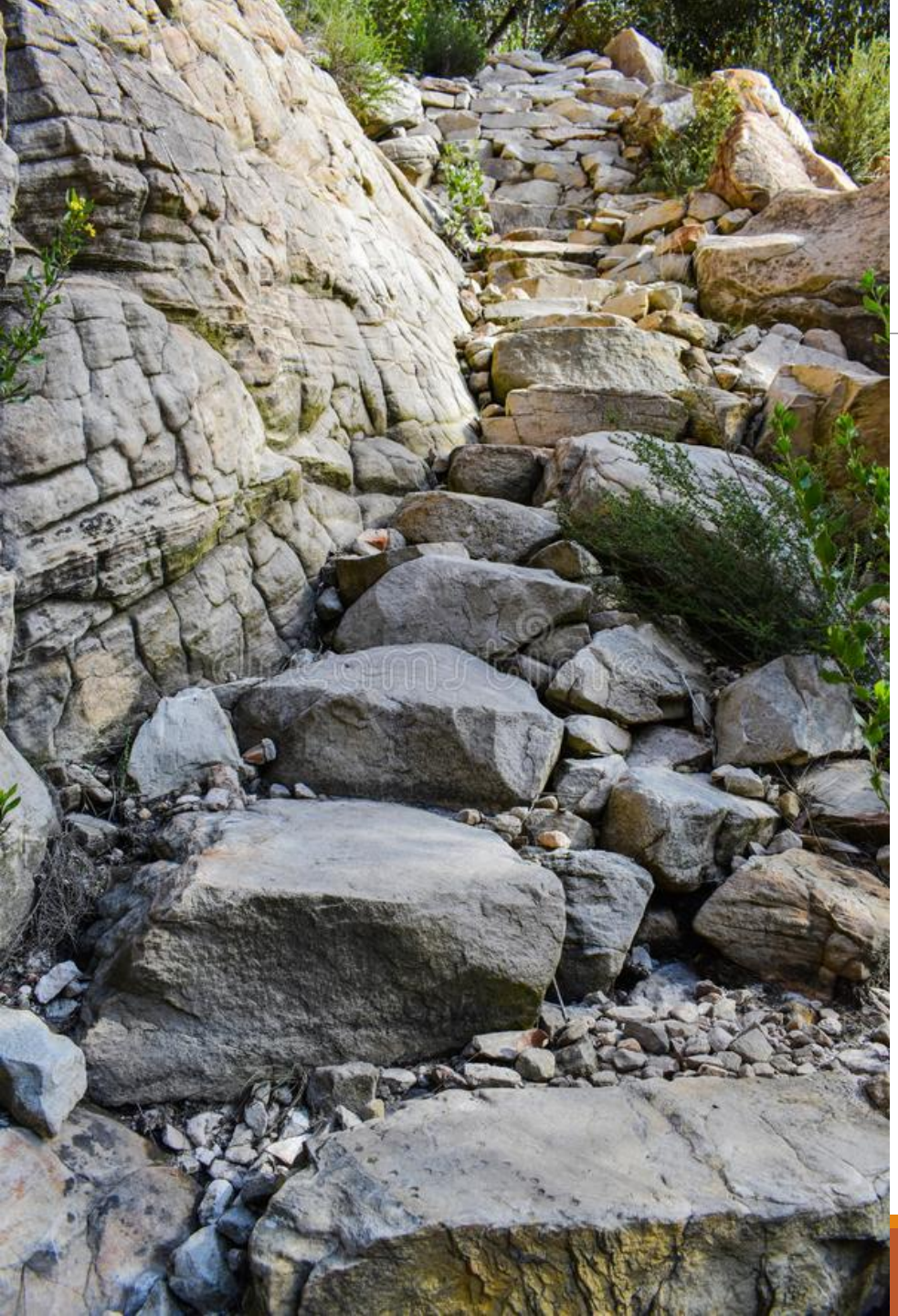
Departament de Ciències Mèdiques Bàsiques
Lab.4.1 Edifici Biomedicina II
Institut de Recerca Biomèdica de Lleida


# Why are we here?

Telling stories is one important aspect of being humans. With stories, we **share information, ideas, predictions**, etc.

Stories can be invented, and they can be very instructive and imaginative. **In science, stories should be told around evidence** (theory, data, and analysis).

**Evidence is the basis of scientific method** and it results from a **careful analysis** of appropriate data.



# From the idea to the conclusions: a rocky path to follow

Which is the context of your study?

Have you identified the main question(s)?

Can you obtain the right data?

Have you planned the analysis?

Is your study plan correct? Is there any alternative?

Do you understand the analysis?

Is there enough evidence for sound conclusions?

# Storytelling (with data):

Avoid inventing, go for evidence!

Which is the main question?

Do I have the right data?

Where to look for right data?

Analyze the data

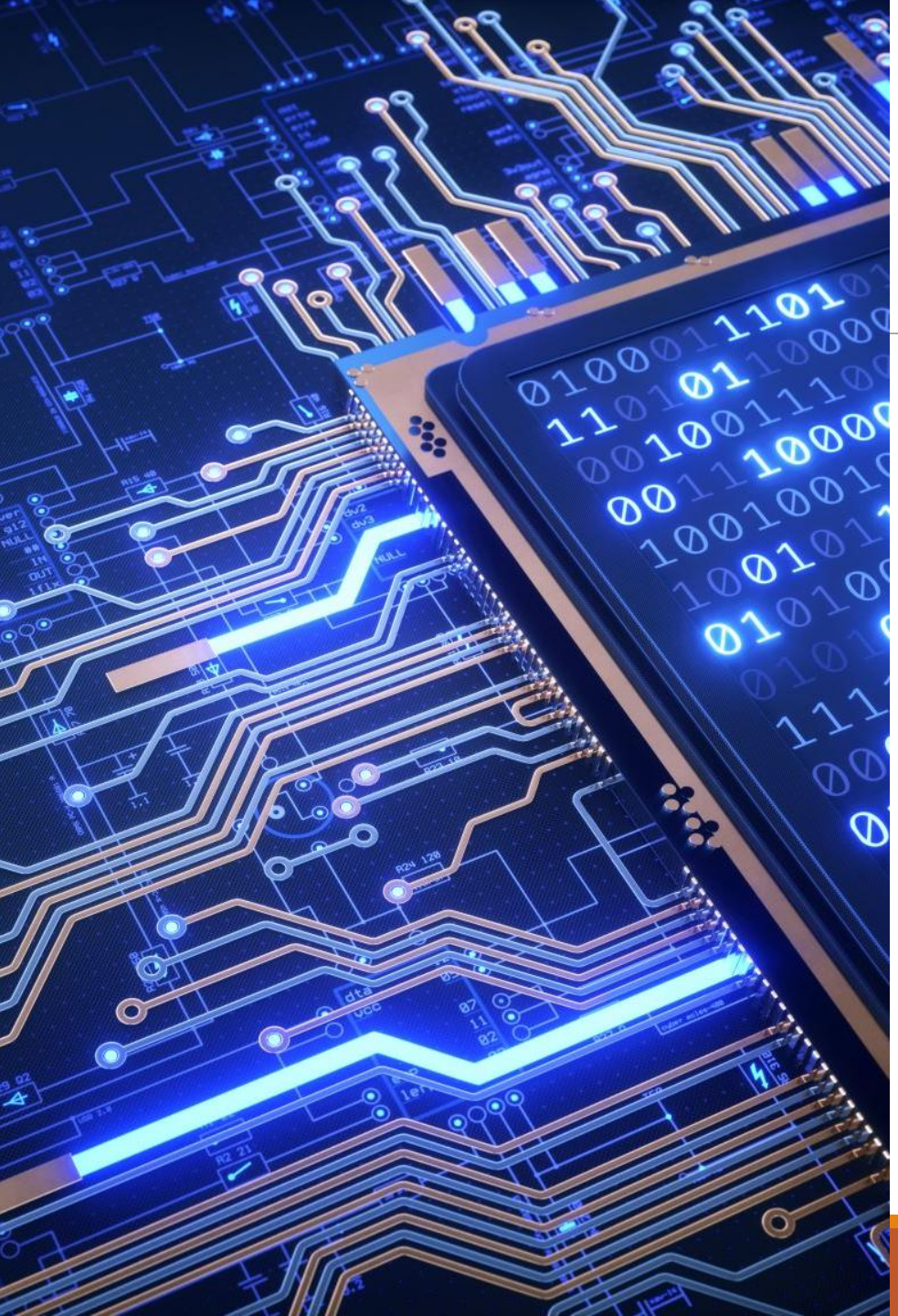
Data must show (not just tell) the story

Visualize data and results

Look for impact and good communication

Define a great story

Build a story around your data and results



# Elements you need for telling a story with data

Understand how to go from raw data to clue ideas

It is fundamental to know the basic tools

- Means, standard deviation, confidence intervals, boxplots, barplots, clusters, inphografics, tables, heat maps, p values, etc.

Computer applications are paramount: R

**Understand how to evaluate the evidence**

Read a lot, discuss ideas, be curious, don’t be afraid of errors, be open minded, ….

# What do we need to know?

Organize data

Visualize data

Analyze data

Communicate conclusions

| | | Teoria |
|---|---|---|
| 1 | From research goals to data: Study Designs | 2 |
| 2 | Clues from Looking at Data: Descriptive statistics | 1 |
| 3 | Understanding probability: Bayes' Rule and clinical diagnostic. Probability Distributions: reference intervals in | 8 |
| 4 | About risk factors: analyzing frequencies. Understanding risk ratio and odds ratios. | 3 |
| 5 | Statistical thinking: confidence intervals. Interpretation and limitations. | 5 |
| 6 | Statistical modelling: linear regresion. | 4 |
| 7 | Statistical modelling: experimental dessign. | |
| 8 | Statistical modelling: logistic regression. | 4 |
| 9 | Statistical modelling: survival analysis. | 3 |
| | | **30** |

**Pràctiques**

| Sessió | Hores | |
|---|---|---|
| 1 | 2 | Intro R |
| 2 | 2 | Tidyverse |
| 3 | 2 | ggplot |
| 4 | 2 | compareGroups |
| 5 | 2 | Descriptiva cuantitativa |
| 6 | 2 | Conf intervals |
| 7 | 2 | Case study EDA |
| 8 | 2 | Regressió |
| 9 | 2 | Regressió |
| 10 | 2 | ANOVA |
| 11 | 2 | Logistica |
| 12 | 2 | Case study |
| 13 | 2 | Case study |
| 14 | 2 | Case study |
| 15 | 2 | Case study |
| | **30** | |

# ¿Qué sabéis de la estadística?

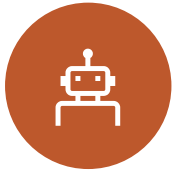
¿Cómo definiríais el concepto de estadística?

¿Porqué se estudia en este grado?

¿Se puede mentir con la estadística?

¿Porqué es importante conocer los conceptos básicos de la estadística?

Los gráficos de datos, ¿son estadística?



¿Qué significa la media? ¿Sabéis lo que es la varianza?

# ¿Qué es una muestra?

Población y muestra

¿Cómo se obtiene un buena muestra?

Encuestas, estudios observacionales, y estudios experimentales

# Recoger información no es tan sencillo

¿Qué vamos a medir?

¿Cómo vamos a medir?

¿Cómo validaremos los datos?

¿Cómo usaremos los datos?

Cuidado: Garbage in, garbage out!!!

# Analizar los datos

Hoja de cálculo

Software estadístico

Análisis descriptivo
Ajuste de modelos
Estimación de efectos

El programa R

# Paso a paso

- Hipótesis de interés
- ¿Qué queremos caracterizar?
  - Una probabilidad (enfermar, morir…)
  - El valor de una determinada propiedad (años de supervivencia, peso,…)
  - Clasificar (sanos/enfermos, buena evolución, …)
- ¿Qué factores debemos tener en cuenta?
  - Edad, sexo, antecedentes familiares, dieta, etc.
- Criterios de inclusión en el estudio
- Enmascaramiento y control de sesgos

# Del problema a la solución

**Es fundamental conocer bien el problema**

Antecedentes

Marco conceptual

Hipótesis de trabajo

**Diseño del estudio**

No todo vale

El diseño debe de ser adecuado

Garantizar la calidad de la información

**Análisis**

Conocer los conceptos básicos

Utilizar la técnica adecuada

# Procedimiento general

En general, los métodos estadísticos tienen las siguientes características comunes:

- Se define un modelo que relacione la variable(s) de interés con variables que se consideran explicativas de su variación.
  - El riesgo de sufrir un infarto en función de la edad, sexo, BMI, hábitos alimenticios, ejercicio, etc.
  - Disminución del colesterol en función de la medicación que se toma.
  - Diferencia de opinión en diversos grupos de interés en función de sus características socio-económicas.
- Se ajusta el modelo a los datos disponibles. La manera de realizar este ajuste depende del problema y de las características de la variable.
- Se evalúa si las predicciones del modelo basadas en hipótesis concretes sobre sus parámetros son compatibles con los datos

# Consideraciones

**Una hipótesis no se demuestra**, en todo caso, en función de la evidencia de las observaciones, puede ser rechazada.

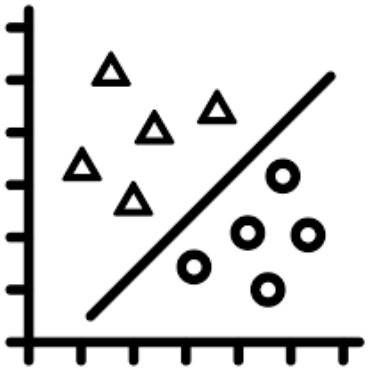
Los métodos estadísticos no demuestran nada; proporcionan una evaluación acerca de la compatibilidad de hipótesis y observaciones.

Debemos evitar la tortura de los datos!

Es más importante dedicar tiempo al diseño de un estudio que a su análisis. Un mal diseño conduce a conclusiones que no son científicamente defendibles.

Es muy importante verificar la selección de las muestras, los procedimientos de medida y la corrección de los datos recogidos.

# Understanding statistical models




Sample

Exploratory data analysis

Study design
- Goals
- Hypothesis
- Sampling plan

Statistical model
- Assumptions
- Functional form
- Parameters
- Fitting procedure

Fitted model
- Does the model fit the data?
- Confidence intervals for parameters
- Goodness-of-fit
- Possible issues

Change/Adapt/New design?

No

Yes

Predictions
Conclusions
Generalization of results

# Why do we need statistical models?

Distribution of a characteristic in the population

Predict the value of a characteristic

- Identify the best predictors
- Which factors are important to consider in the predictions?
- Which is the probability of death for an ICU patient?

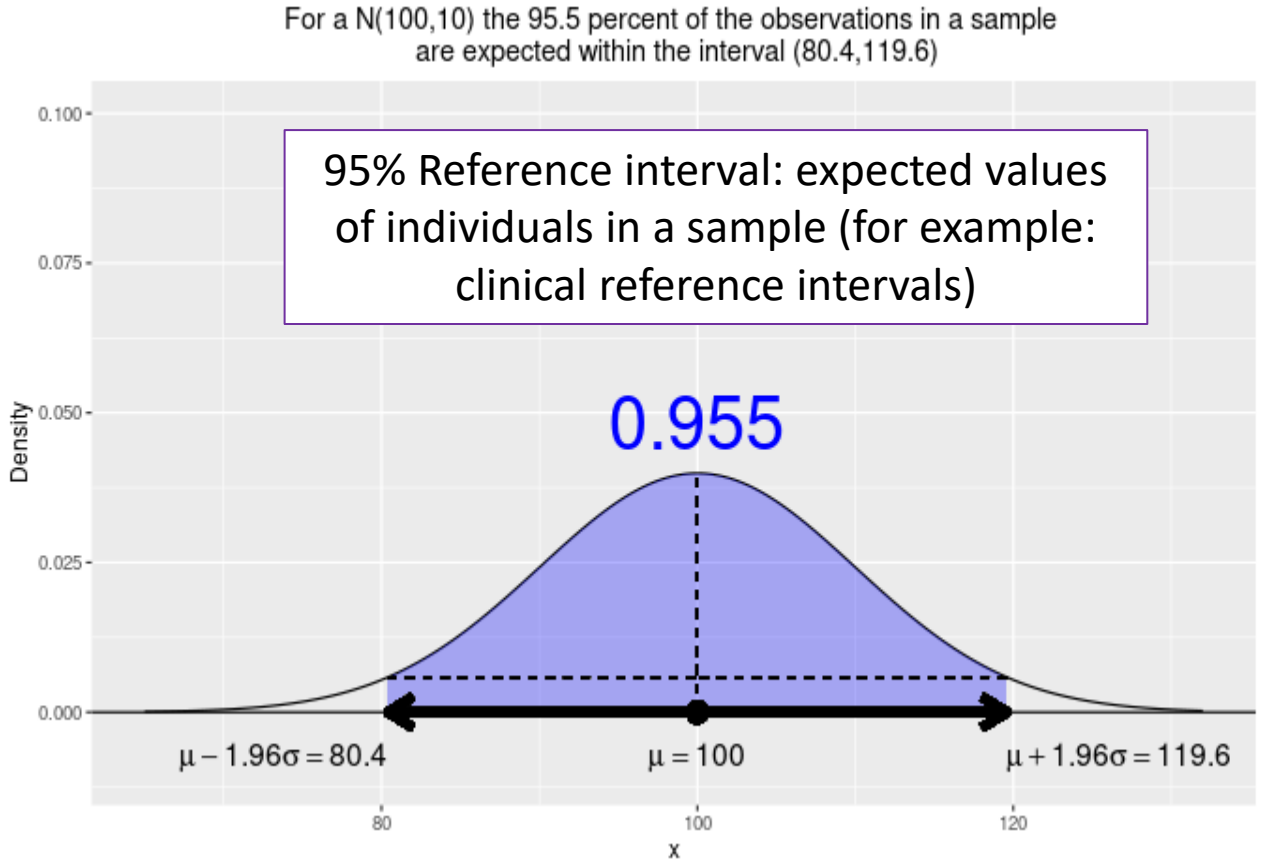
Evaluate treatments

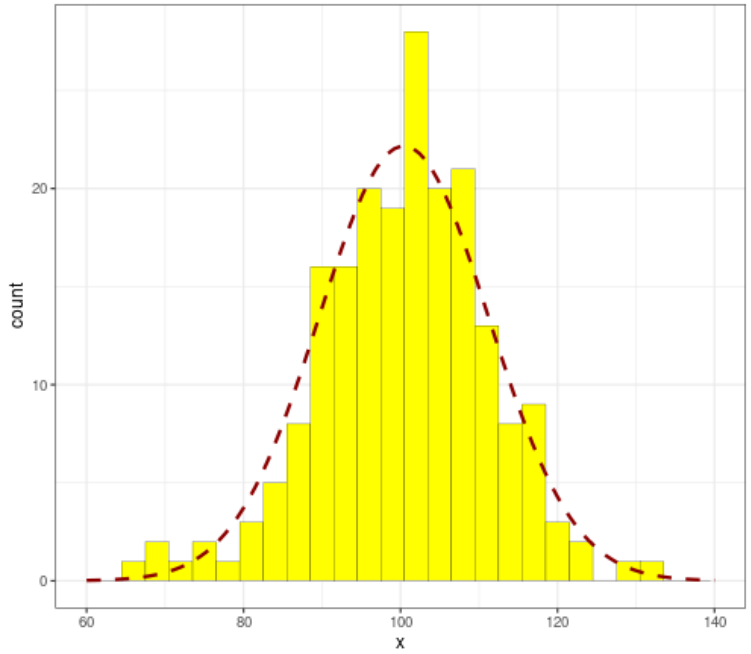
- Is the treatment effective?
- Which is the best design for a given purpose?
- How do we evaluate survival?

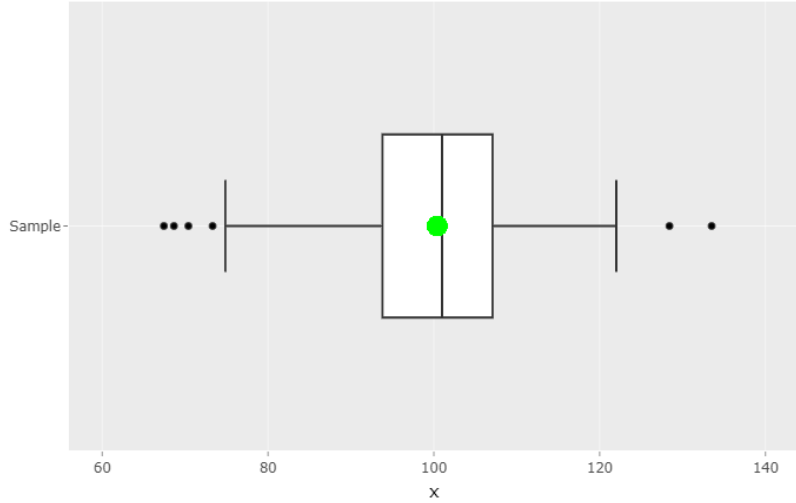
Classification

- Identify risk groups

# Normal distribution

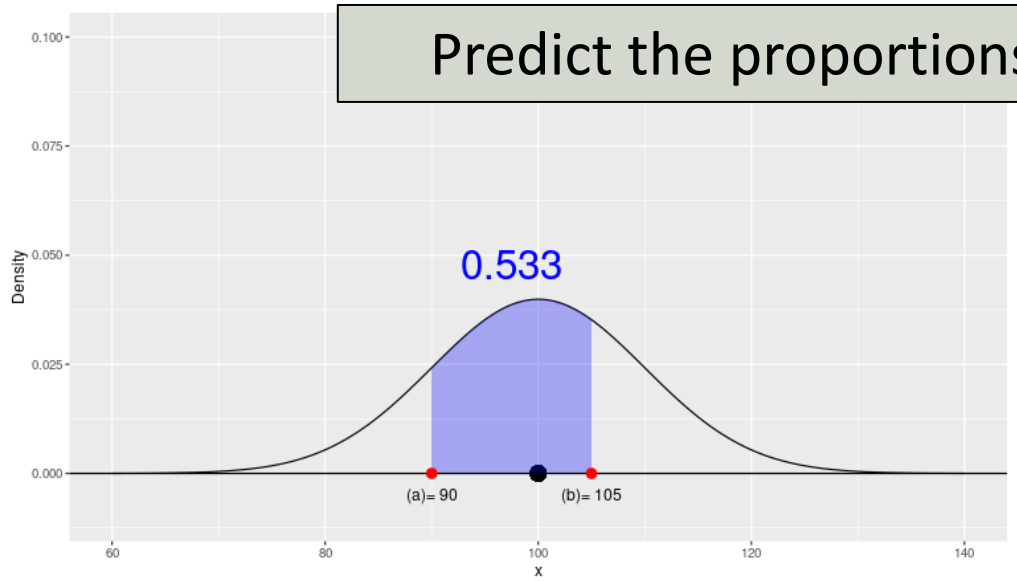


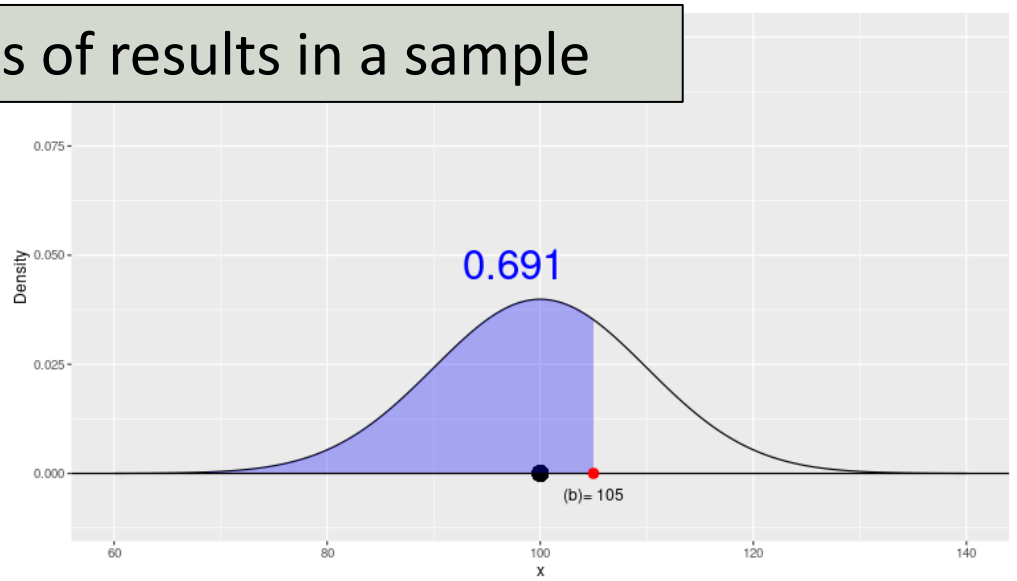



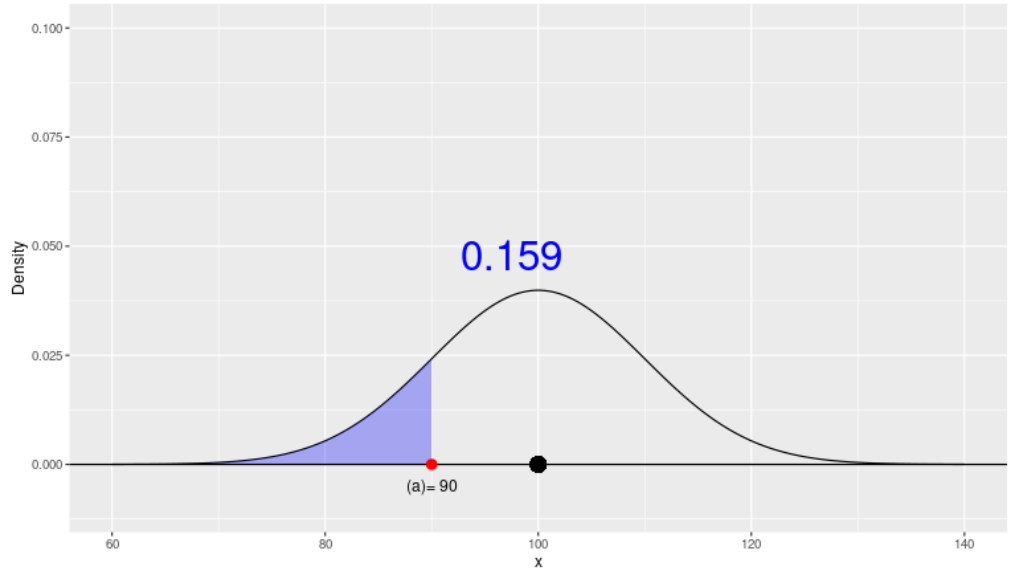



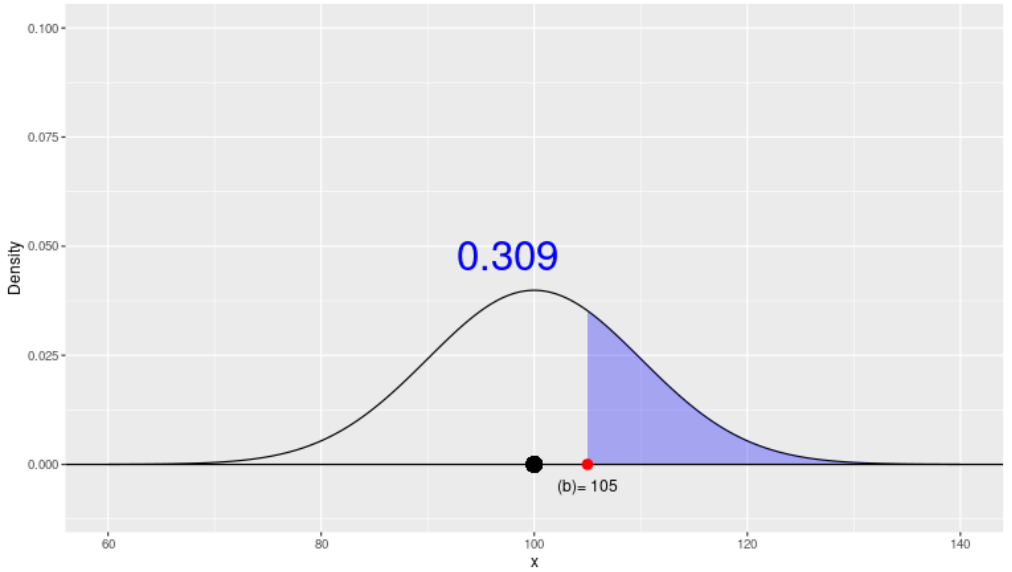
# N(100,10)

## Predict the proportions of results in a sample



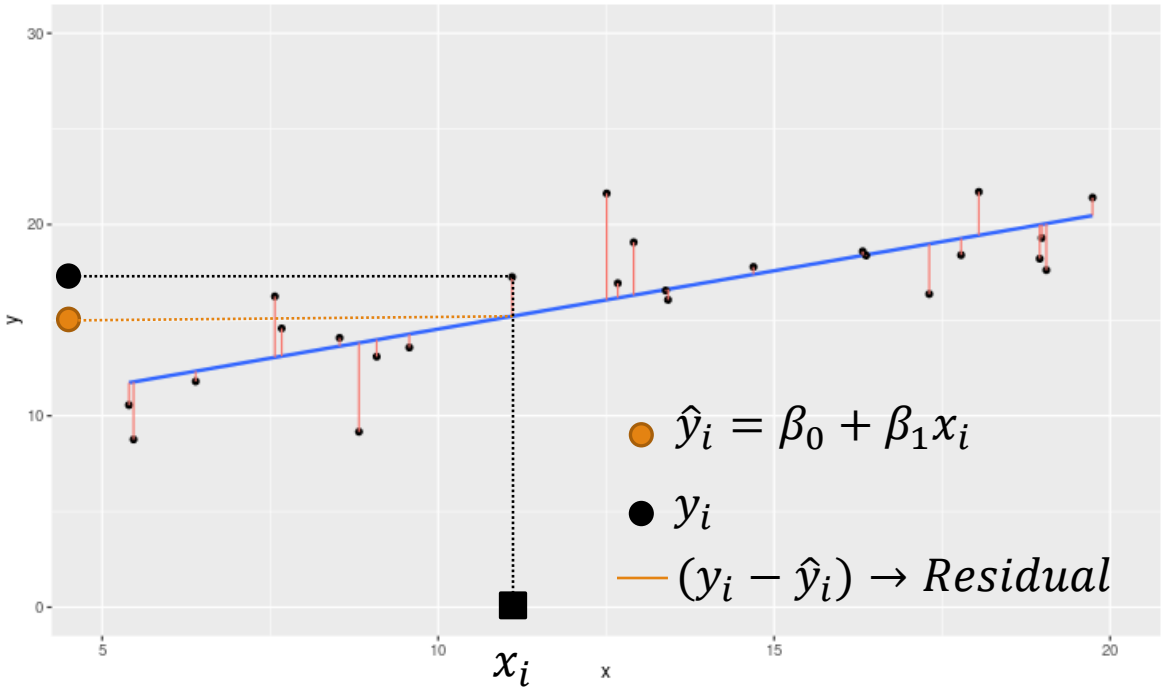



# Linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

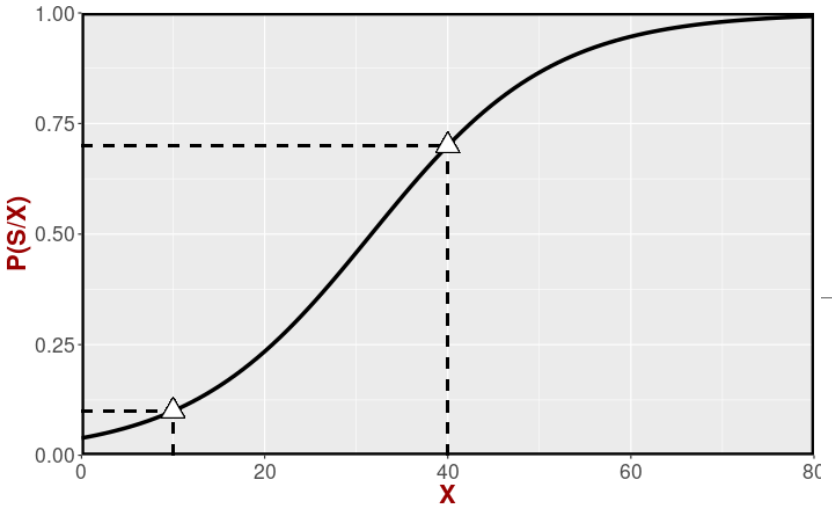



Least squares criterion

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

$$(y_i - \hat{y}_i) \rightarrow Residual$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \rightarrow Minimim$$




# Logistic model: predicting a probability

The logistic regression is a model for the **odds** of an event at a given value of a predictive variable. The odds of an event of probability $p$ are defined as: $p/(1-p)$

The (univariate) logistic model states that the logit is a linear function of the predictor:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times x$$

Thus, in the logistic regression, the probability of an event (S) as a function of the value of a variable X is modeled as:

$$P(S|X=x) = \frac{1}{1+e^{-(\beta_0+\beta_1\times x)}}$$

This model can be extended for several predictor variables:

$$P(S|X=\{x_1,x_2,..,x_n\}) = \frac{1}{1+e^{-(\beta_0+\beta_1\times x_1+\beta_2\times x_2+..+\beta_n\times x_n)}}$$

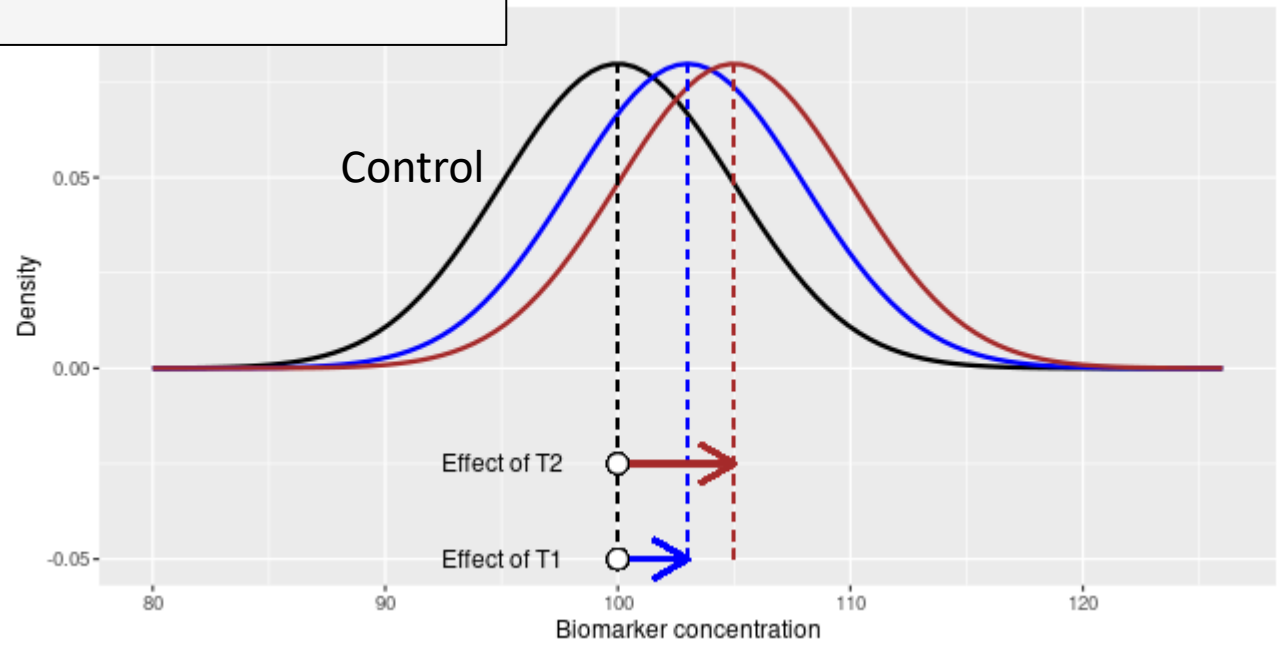
## Model for one fixed factor

In a design with on fixed factor we define several experimental groups (factor levels) and measure a variable in different subjects within each experimental unit. This is the case of one Control group treated with a placebo compared to three treatments. In this case the factor is the treatment, with four levels (placebo and three treatments) A linear model for one fixed factor is defined as:

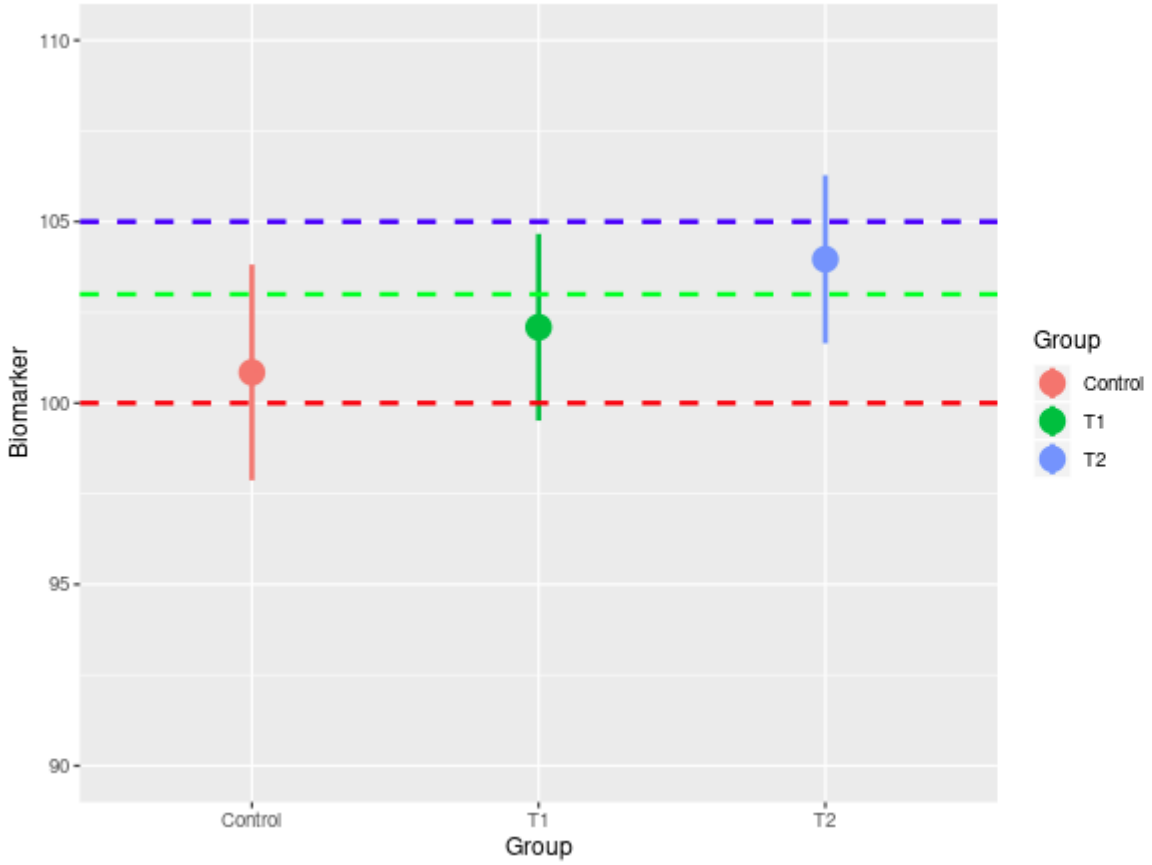
$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

Where $y_{ij}$ is the $j^{th}$ observation in the $i^{th}$ group (level). $\mu$ is the mean of the results if there is no effect of the factor. $\alpha_i$ is the effect (additive) of the level $i$. Thus, the expected mean in the $i^{th}$ group is $\mu_i = \mu + \alpha_i$ Finally, $\epsilon_{ij}$ indicates the random variation around the mean. It is assumed that the random variation follows a $N(0, \sigma)$ distribution.

# Evaluate the effect of two treatments (T1 and T2) vs. control



# Evaluate the effect of two treatments (T1 and T2) vs. control




| Group | Mean | SD | n | Median | low | upper |
|---|---|---|---|---|---|---|
| Control | 100.71 | 6.01 | 15 | 100.49 | 97.38 | 104.04 |
| T1 | 102.69 | 4.88 | 15 | 101.19 | 99.98 | 105.39 |
| T2 | 104.47 | 4.33 | 15 | 103.86 | 102.07 | 106.87 |

**Range of biomarker axes**

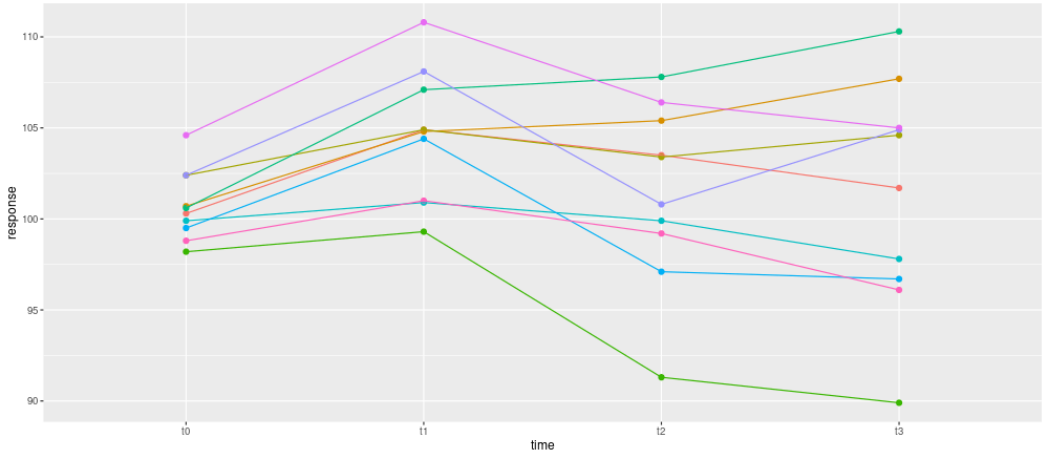
50 90 110 200

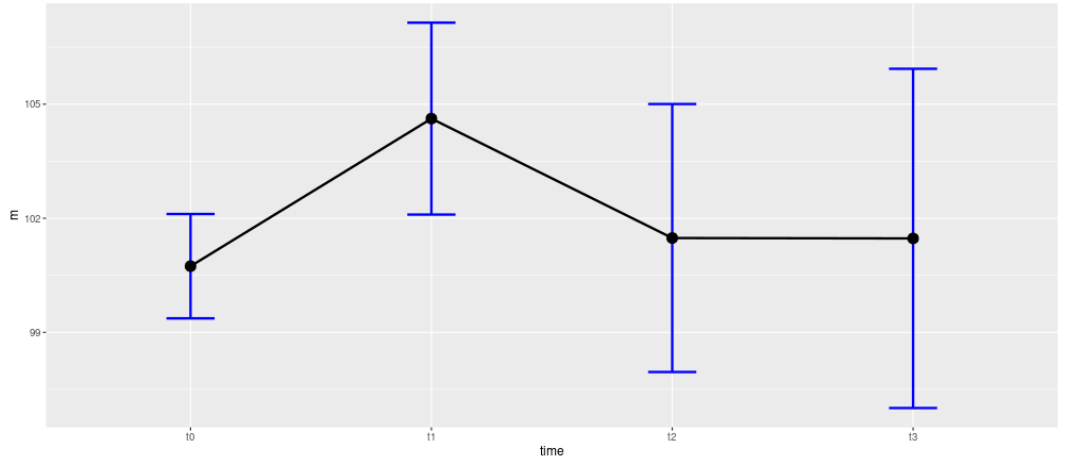
50 65 80 95 110 125 140 155 170 185 200

```
Analysis of Variance Table

Response: Biomarker
           Df  Sum Sq Mean Sq F value Pr(>F)
Group       2  106.11  53.054  2.0215 0.1451
Residuals 42 1102.30  26.245
```

# Repeated measures:
## the importance of design




This is not a fixed factor model (with time and subject as a factors)
We should consider the effect of each subject
Observations of each subject are not independent in different times

Here the previous model does not apply.

# Assessing survival

The probability of survival is modeled in a non-parametric way (Kaplan-Meier)

Cox-regression can be used when covariates are to be considered (age, sex, etc.)

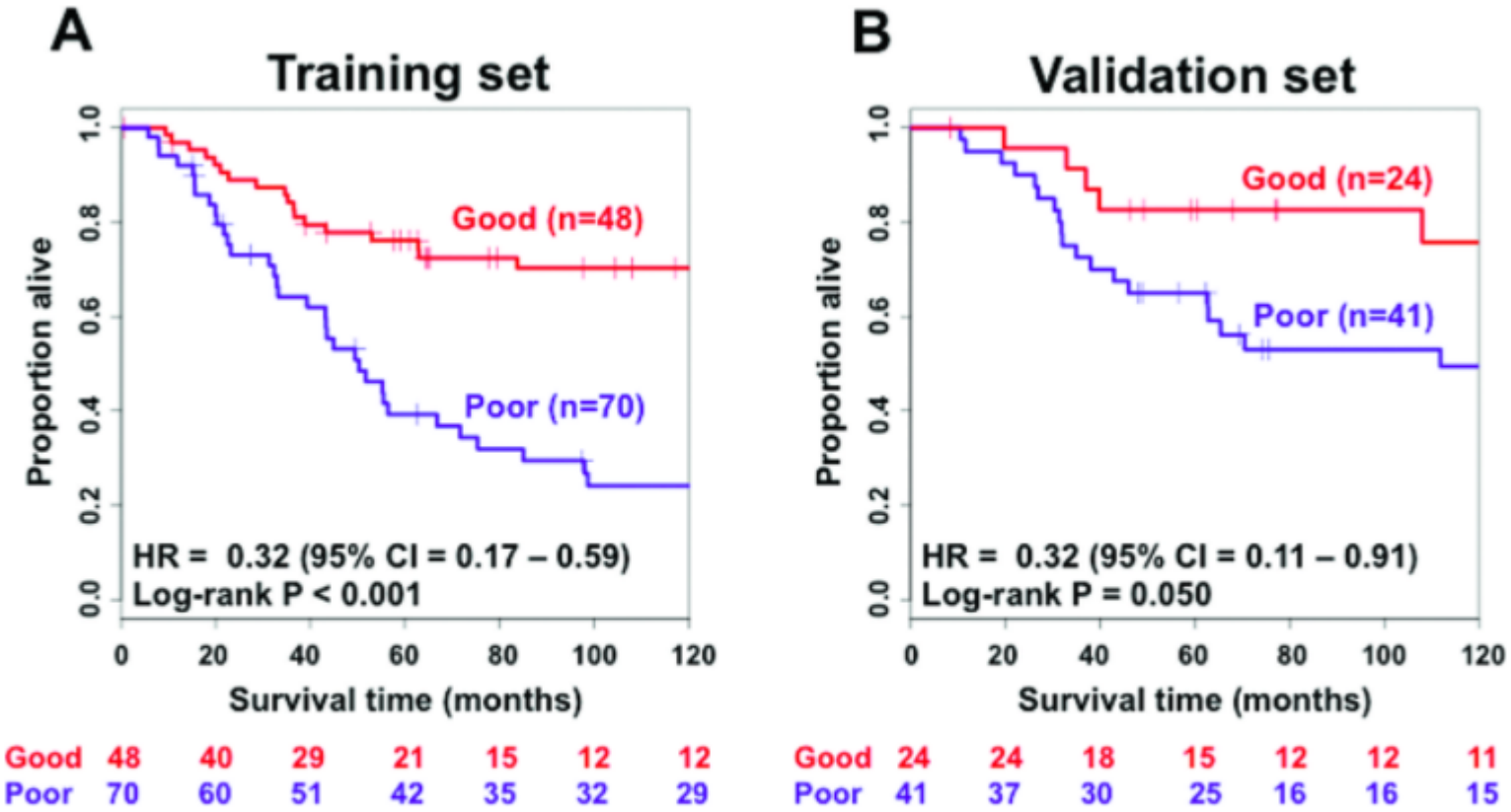


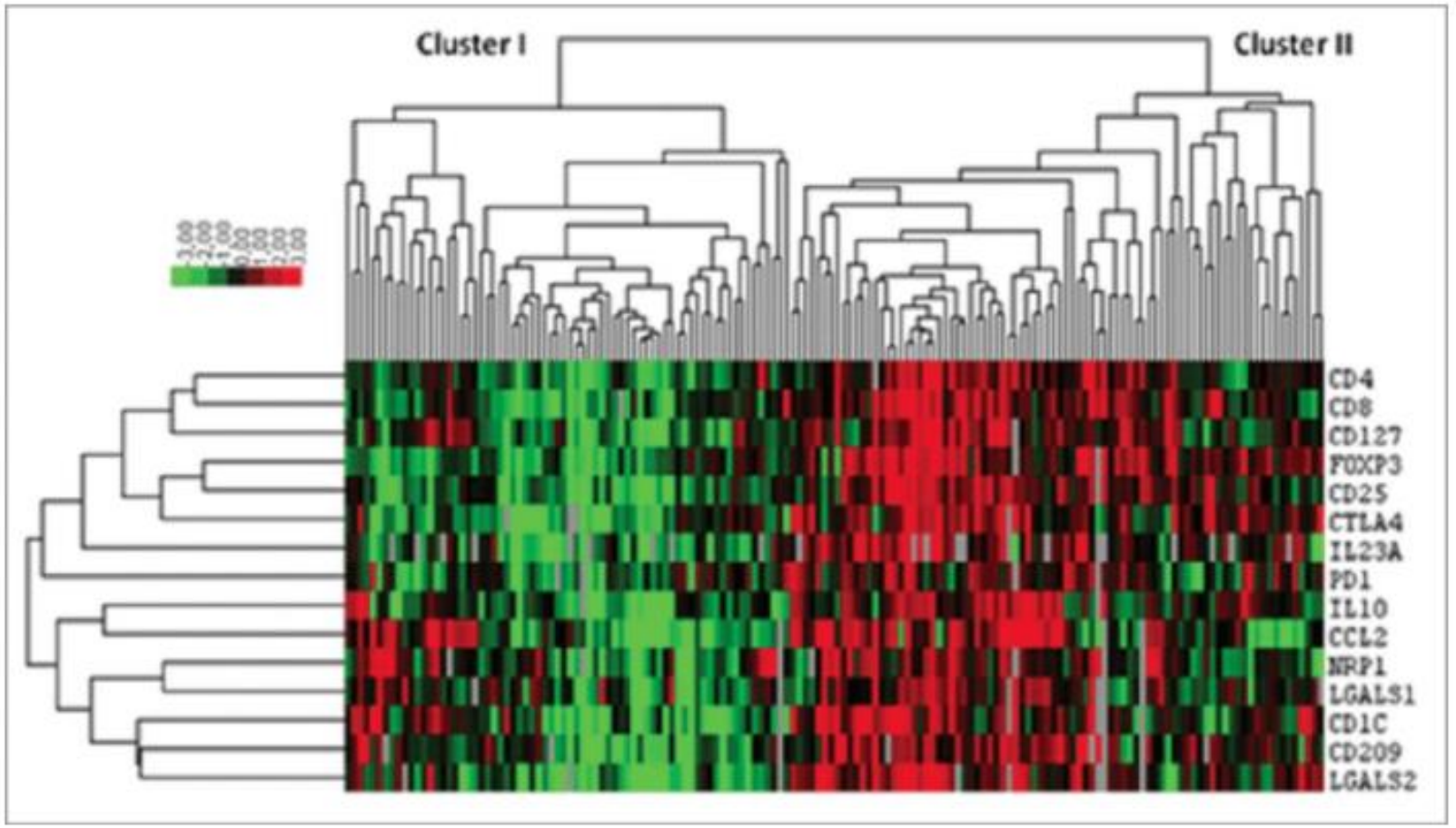

**Figure 2** Kaplan-Meier analysis showing that the four-biomarker signature is associated with survival in triple negative breast cancer. (A) Kaplan-Meier curve of the four-biomarker signature in the training set. (B) Kaplan-Meier curve of the four-biomarker signature in the validation set. CI = confidence interval; HR = hazard ratio. HR and 95% CI were estimated by multivariate Cox regression with age at diagnosis, grade, the presence of nodes, and menopausal status included as covariates. The *p*-value was obtained by the log-rank test of the Kaplan-Meier curve.

# Cluster analysis

Define a distance between subjects (i.e. gene expression)

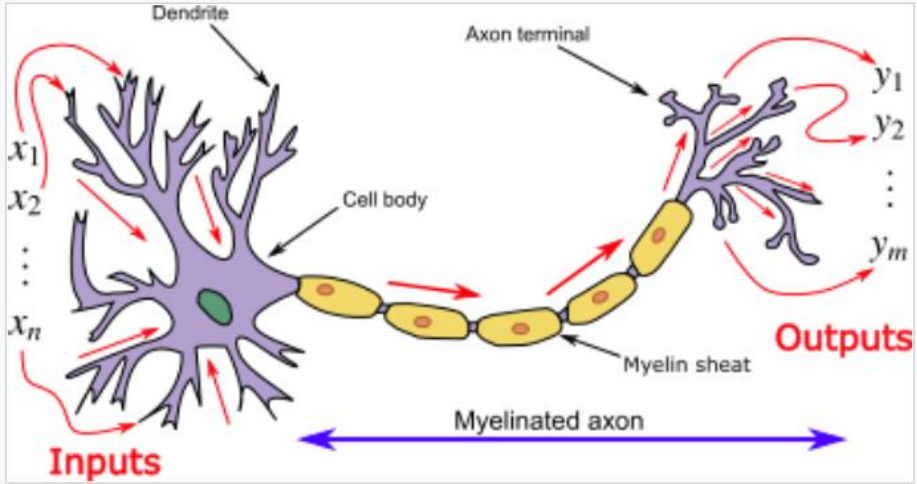
Find the most similar subjects

Move on by aggregation until no more groups can be identified.

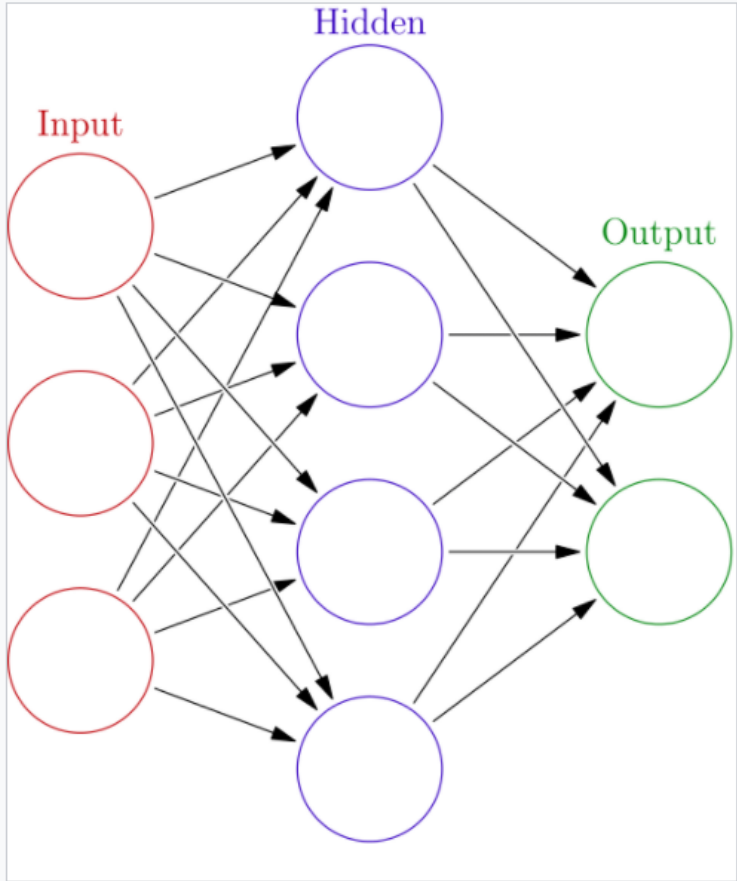



Hierarchical cluster based on selected gene expression. Patients in the original cohort were clustered into a hierarchical tree based on the expression of immune related genes. The clustering separated the patients into two distinct groups. Red indicates high expression and green indicates low expression levels.

# Reaching the edge:
## computational models



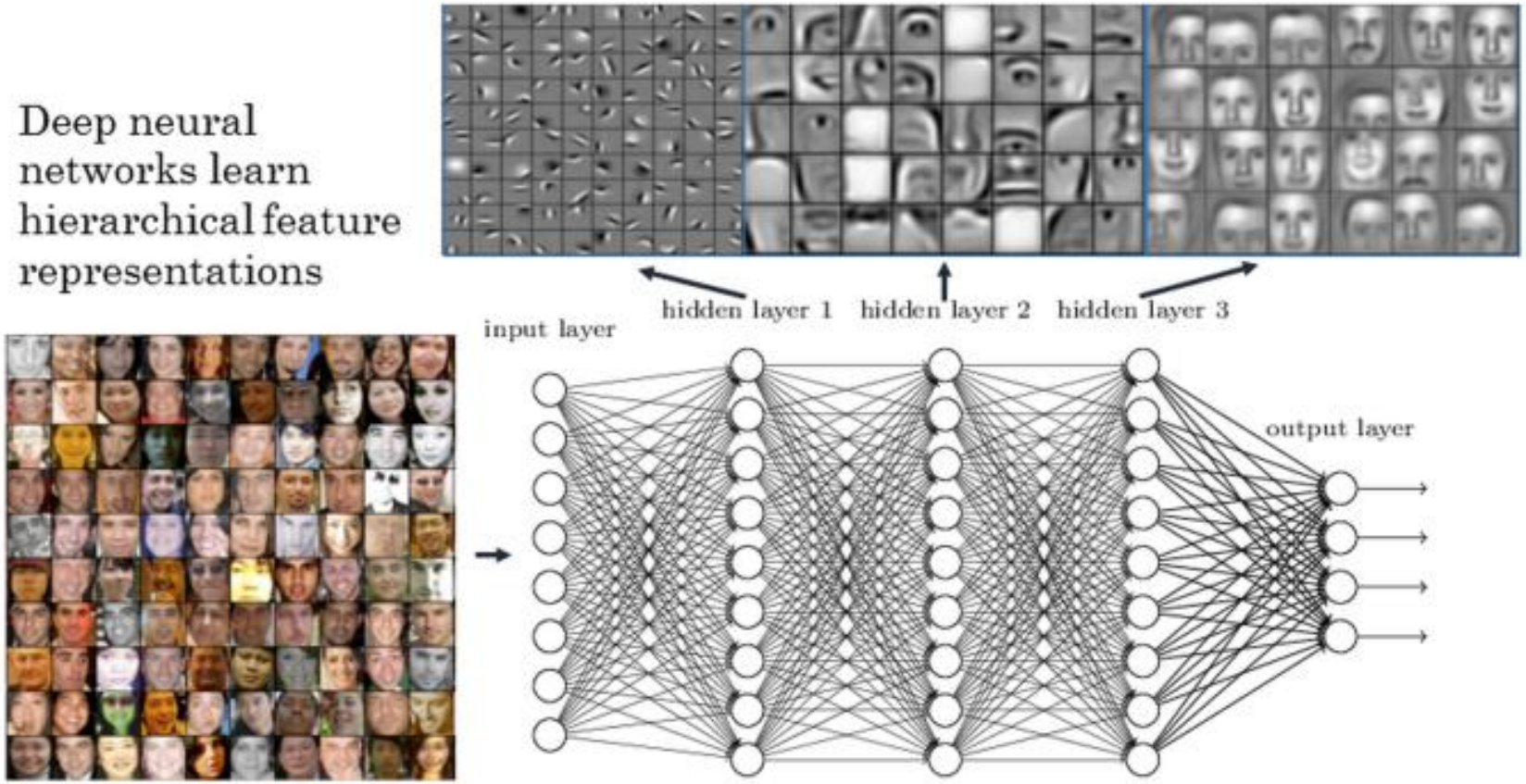

Neuron and myelinated axon, with signal flow from inputs at dendrites to outputs at axon terminals




An artificial neural network is an interconnected group of nodes, inspired by a simplification of neurons in a brain. Here, each circular node represents an artificial neuron and an arrow represents a connection from the output of one artificial neuron to the input of another.

The model is trained by examples
Interactions evolve their values until an accurate output is reached
This is an iterative and time-consuming process

# Deep learning



# Take home messages

**Working hypotheses and study design are the starting point.**

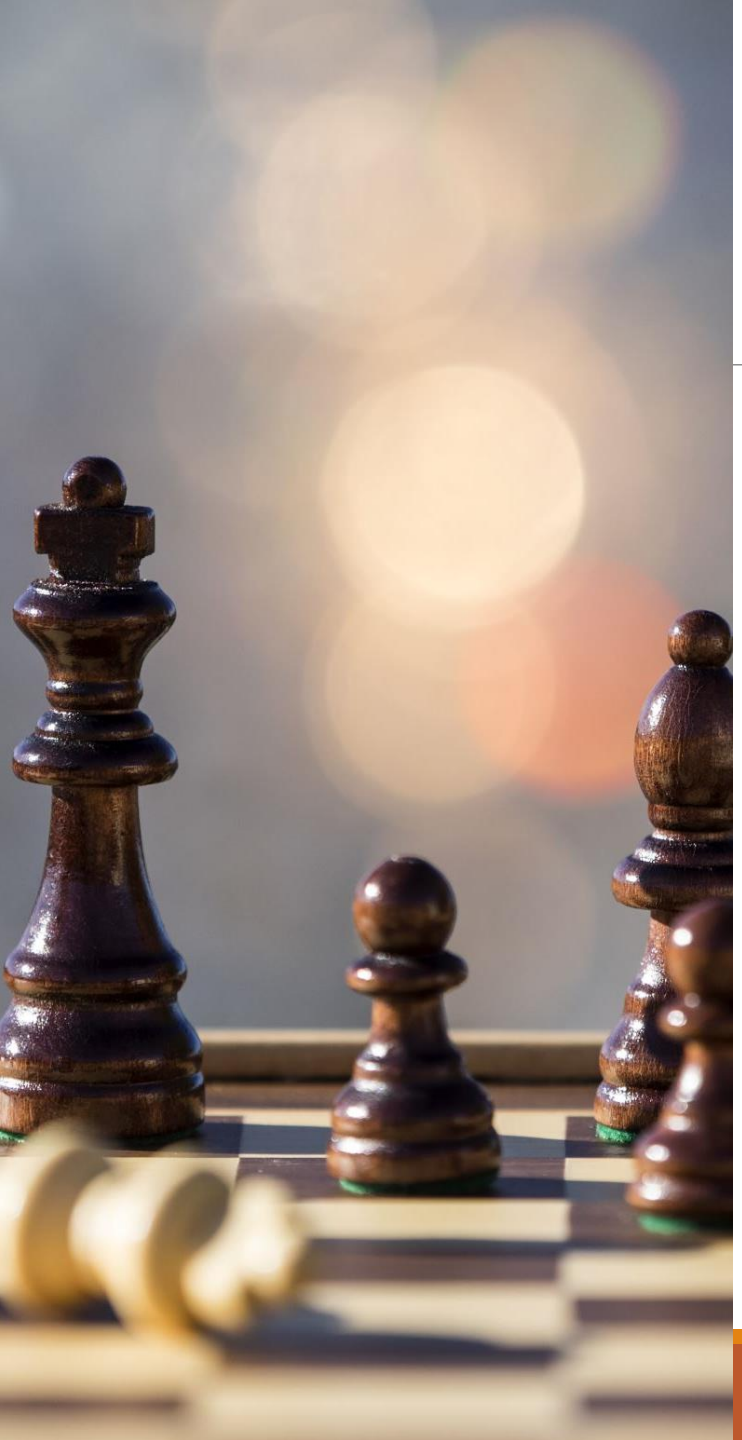
Each model (method) has its own **requisites for validity**. Assessment of such requisites is extremely important for getting valid conclusions.

**Quality of data** (sampling and validation) is very important. Remember: **garbage in, garbage out!**

**Almost any model can be fitted to data**. It is important to assess if the fit is appropriate. Narrow confidence intervals for the model parameters provide a clue.

**The value of a model is its capacity for generalization**. Models may learn a set of data and fail to produce good predictions for a new set of the same problem.

**For complex problems, alternative models may be considered.** Model comparison and a careful selection is an important, and sometimes difficult, step.



# Final

- El avance de la ciencia se basa en la evidencia.
- La estadística es necesaria para evaluar hasta qué punto disponemos de suficientes evidencias para defender una conclusión.
- Es importante conocer los conceptos fundamentales del método estadístico para poder entender y comunicar los resultados de un estudio.
- El dominio de las técnicas estadísticas es un reto complejo, pero no hay escusa para no entender los conceptos en que se basan.