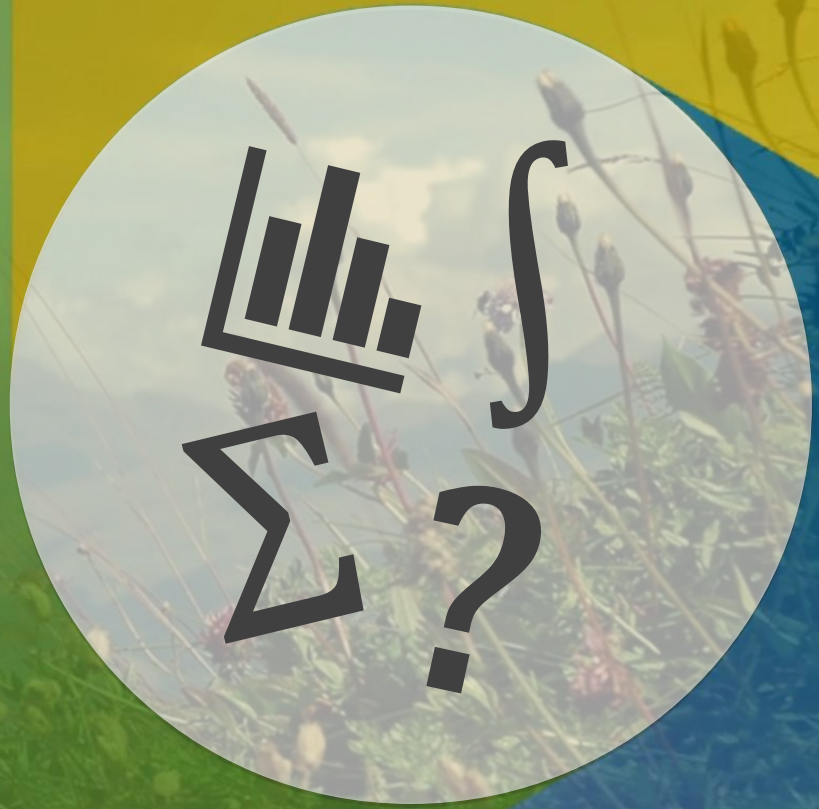
Introduction to Bayesian Statistics



Part 1

Statistical modeling

Benjamin Rosenbaum

iDiv 2025

About me

- Reseacher & Statistical Consultant
- Quantitative Ecologist
- Started out as a mathematician
- Main research interests:
  - Statistical methods for process-based models
  - Population & community dynamics
  - Species interactions, functional responses

New course!

2025 RELAUNCH

Deployed to **github-pages** by
benjamin-rosenbaum via pages-build-deployment #10

`master`

Target audience

- You have used **R** before

- You have just a little bit of stats experience

Or:

- You have some stats experience,
  but not in Bayesian stats

Or:

- You have some Bayesian stats experience,
  but not with the *brms* package

Image by storyset on Freepik

Course goals

- Building blocks of statistics modeling

  Think in terms of models, not tests

- Revision of classical models:

  Learn something useful even if you want to stick to frequentist stats.

- Basic understanding of Bayesian statistics

- Write code with the *brms* package

- Interpret model output & statistical inference

→ Analyze your own datasets

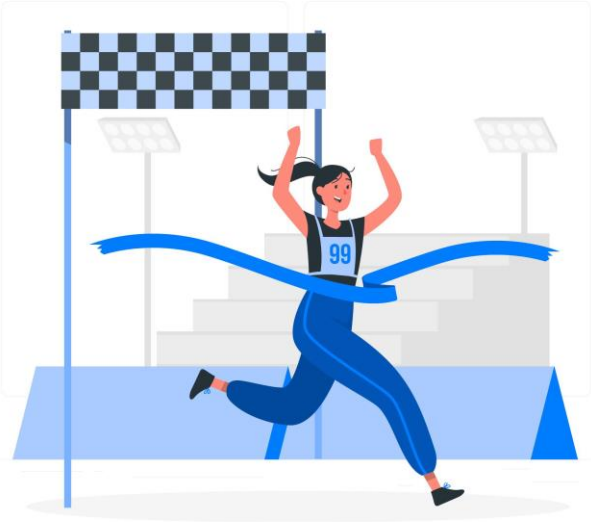


Image by storyset on Freepik

Contents

1. Statistical modeling
2. Bayesian principles
3. Prior and posterior distributions
4. Linear models
5. Generalized linear models
6. Mixed effects models
7. Stan introduction
8. Conclusions + questions

→ Every lesson includes a **lecture** and a **practical** part.
There are small exercises for self-study, solutions in github

This lecture

Review: probability distributions

What is a statistical model?

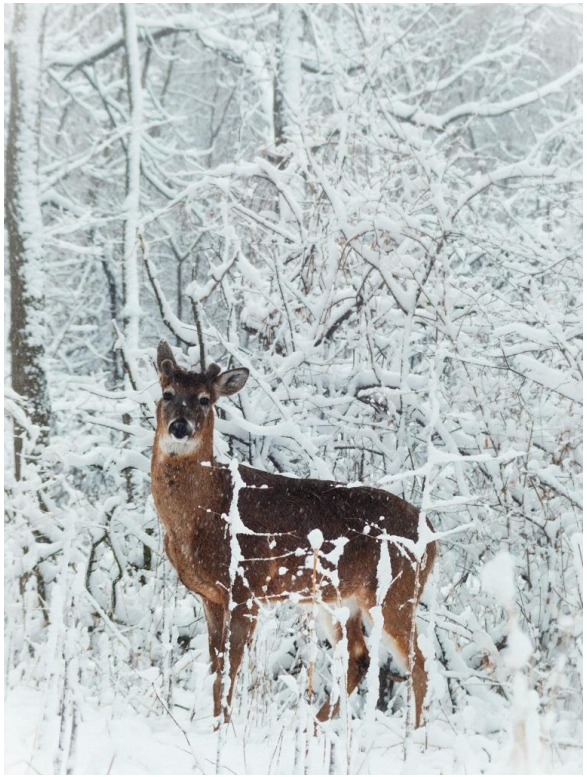
Probability and the likelihood function

Maximum likelihood estimation
(as preparation for Bayesian statistics)

Review: Probability distributions

Discrete distribution

- **Example:** number of individuals from a population of $N = 10$ that survive the winter
- y **discrete** and **bounded** variable with outcomes 0, 1, 2, …, 10
- Average survival probability $\theta = 0.6$ (60%)
- Binomial distribution: $y \sim \text{Binomial}(N, \theta)$



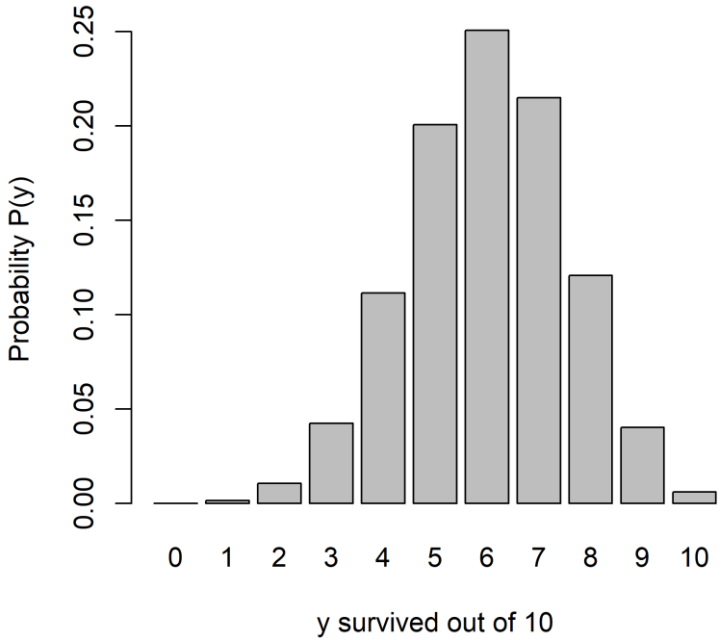
Discrete distribution

- Binomial distribution: $y \sim \text{Binomial}(N, \theta)$
- Probability function $P(y|\theta) = \binom{N}{y} \theta^y (1-\theta)^{N-y}$
  calcutates **probability** of each possible outcome for a fixed set of parameters ($N = 10, \theta = 0.6$)
- No need to memorize the equation. Use R:

```
> p = dbinom(y,size=10,prob=0.6)
```

- Draw random samples from this distribution

```
> y = rbinom(1,size=10,prob=0.6)
```



Discrete distribution

- Probabilities always sum up to 1:

  $P(y = 0) + P(y = 1) + \cdots + P(y = 10) = 1$

- Mean $\mu = N \cdot p = 0.6 \cdot 10 = 6$

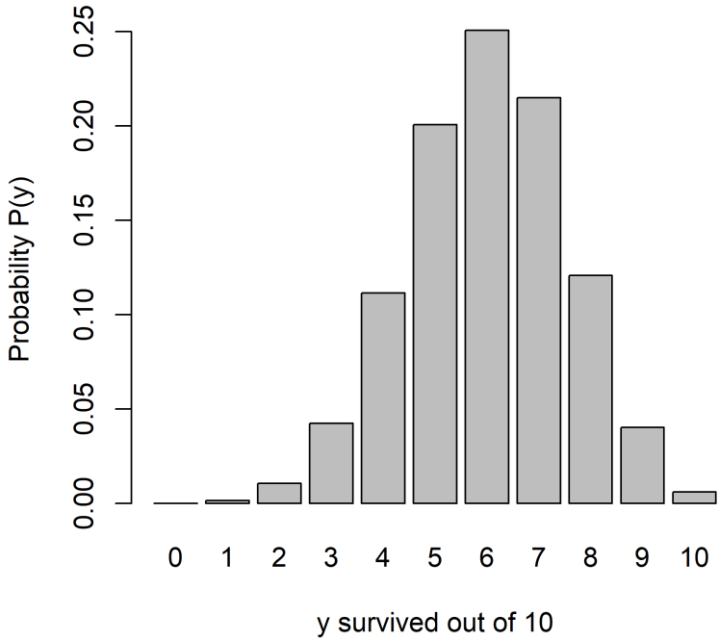
  (average outcome if experiment is repeated often)

- Compute probabilities of **events**, for example

  $P(y = 6) = 0.251$

  $P(y \geq 6) = P(y = 6) + \cdots + P(y = 10) = 0.633$

  $P(4 \leq y \leq 8) = P(y = 4) + \cdots + P(y = 8) = 0.899$



Discrete distribution

- Probabilities always sum up to 1:

  $P(y = 0) + P(y = 1) + \cdots + P(y = 10) = 1$

- Mean $\mu = N \cdot p = 0.6 \cdot 10 = 6$

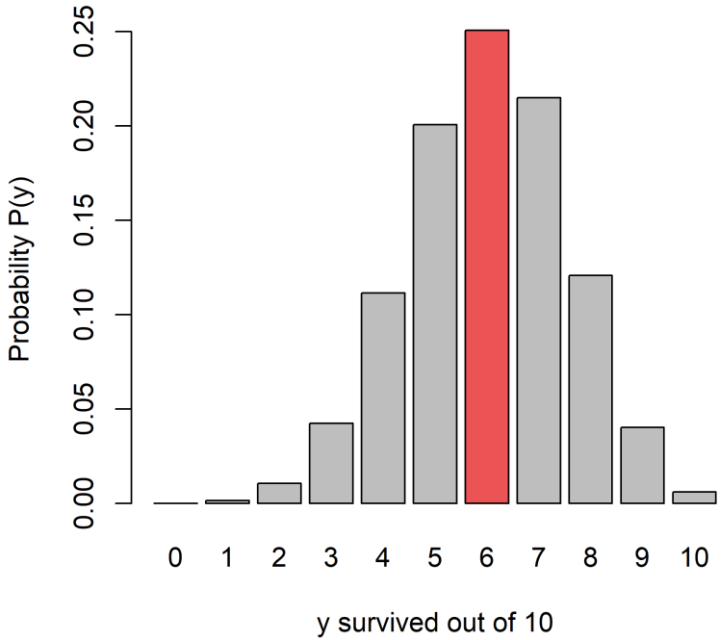
  (average outcome if experiment is repeated often)

- Compute probabilities of **events**, for example

  $P(\boldsymbol{y = 6}) = 0.251$

  $P(y \geq 6) = P(y = 6) + \cdots + P(y = 10) = \ 0.633$

  $P(4 \leq y \leq 8) = P(y = 4) + \cdots + P(y = 8) = 0.899$



Discrete distribution

- Probabilities always sum up to 1:

  $P(y = 0) + P(y = 1) + \cdots + P(y = 10) = 1$

- Mean $\mu = N \cdot p = 0.6 \cdot 10 = 6$

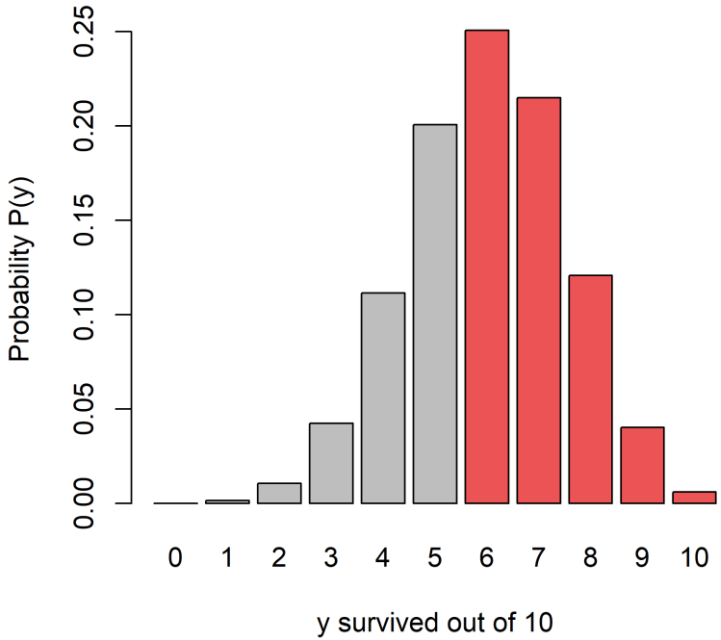
  (average outcome if experiment is repeated often)

- Compute probabilities of **events**, for example

  $P(y = 6) = 0.251$

  $P(\boldsymbol{y \geq 6}) = P(y = 6) + \cdots + P(y = 10) = 0.633$

  $P(4 \leq y \leq 8) = P(y = 4) + \cdots + P(y = 8) = 0.899$



Discrete distribution

- Probabilities always sum up to 1:

  $P(y = 0) + P(y = 1) + \cdots + P(y = 10) = 1$

- Mean $\mu = N \cdot p = 0.6 \cdot 10 = 6$

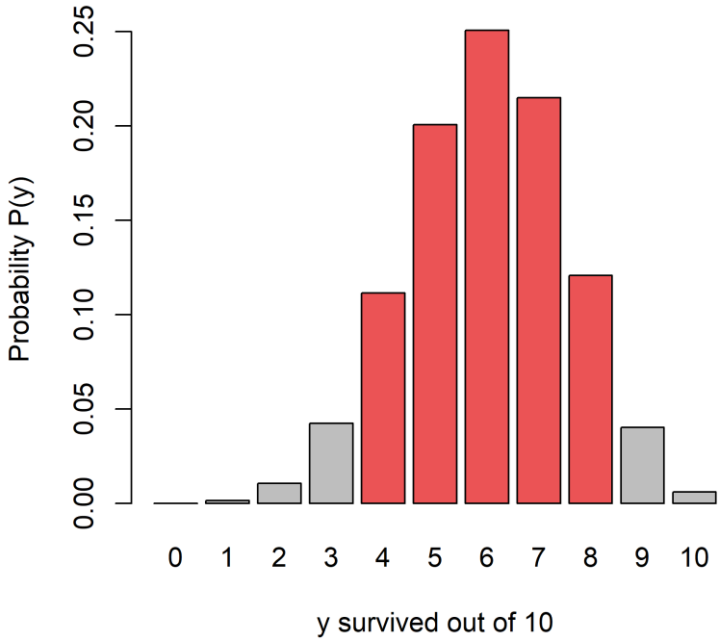
  (average outcome if experiment is repeated often)

- Compute probabilities of **events**, for example

  $P(y = 6) = 0.251$

  $P(y \geq 6) = P(y = 6) + \cdots + P(y = 10) = 0.633$

  $P(\mathbf{4 \leq y \leq 8}) = P(y = 4) + \cdots + P(y = 8) = 0.899$



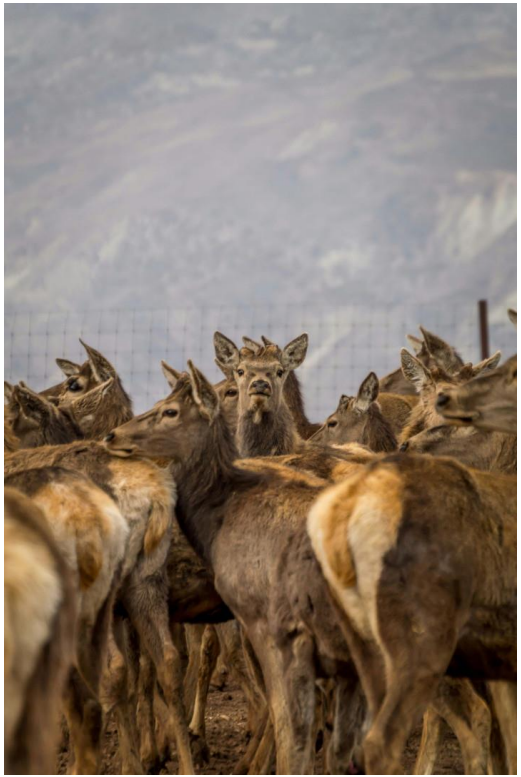
Continuous distribution

- **Example:** body mass of adult deer
- y can take any value (continuous)
- Average body mass $\mu = 100\ [kg]$
- Standard deviation $\sigma = 10$ (spread)
- Normal distribution: $y \sim \text{Normal}(\mu, \sigma)$

random variable

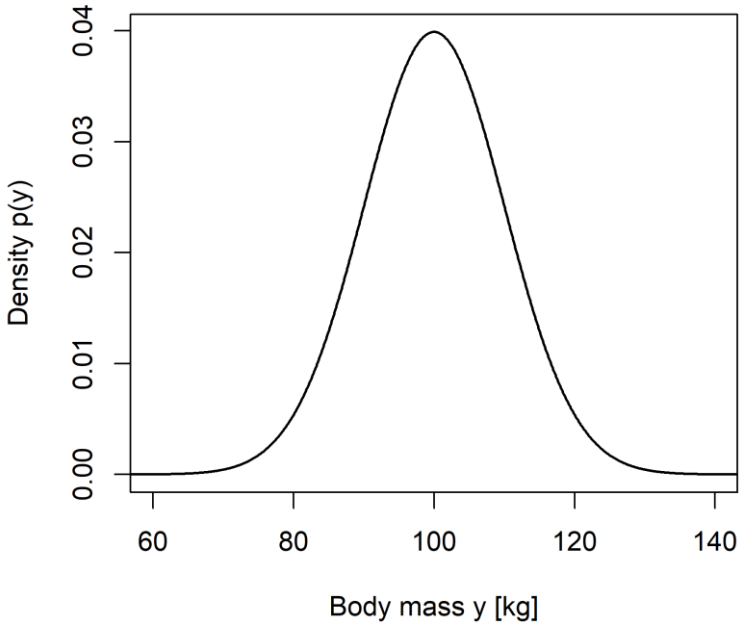
„distributed as"

parameters:
mean μ
standard deviation σ



Continuous distribution

- Normal distribution: $y \sim \text{Normal}(\mu, \sigma)$
- $p(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$ is the **probability density function** of each possible outcome y for a fixed set of parameters ($\mu = 100, \sigma = 10$)
- Mean μ and standard deviation σ (average outcome if experiment is repeated often)



Continuous distribution

- Normal distribution: $y \sim \text{Normal}(\mu, \sigma)$
- $p(y = 95.0|\mu, \sigma)$ is **not** the probability for $y = 95.0$

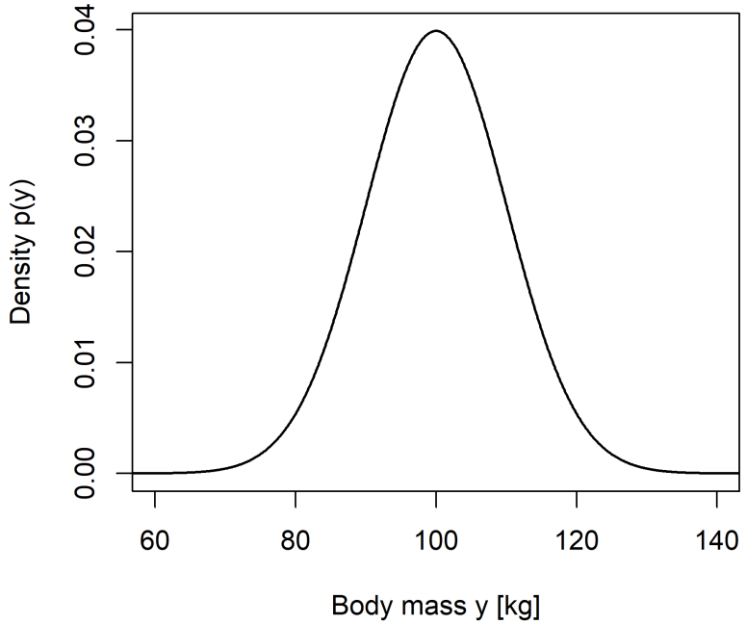
  For continuous distributions, prob. of an exact value is zero!

  (see next slide)
- No need to memorize the equation. Use R:

```
> p = dnorm(y,mean=100,sd=10)
```

- Draw random samples from this distribution

```
> y = rnorm(1,mean=100,sd=10)
```



Continuous distribution

- Probabilities always integrate to 1 (area under the curve):

  $\int p(y|\mu,\sigma)dy = 1$ for any μ, σ

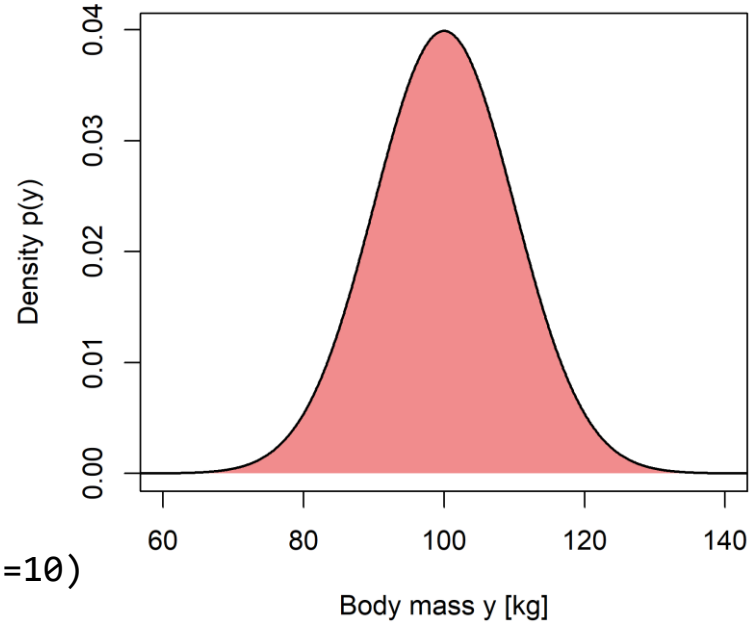
- Compute probabilities of an **interval**, for example

  $P(y \leq 110) = \int_{-\infty}^{110} p(y|100,10)dy = 0.841$

  ```
  > pnorm(110,mean=100,sd=10)
  ```

- $P(90 \leq y \leq 110) = \int_{90}^{110} p(y|100,10)dy = 0.682$

  ```
  > pnorm(110,mean=100,sd=10)-pnorm(90,mean=100,sd=10)
  ```



Continuous distribution

- Probabilities always integrate to 1 (area under the curve):

  $\int p(y|\mu,\sigma)dy = 1$ for any μ,σ

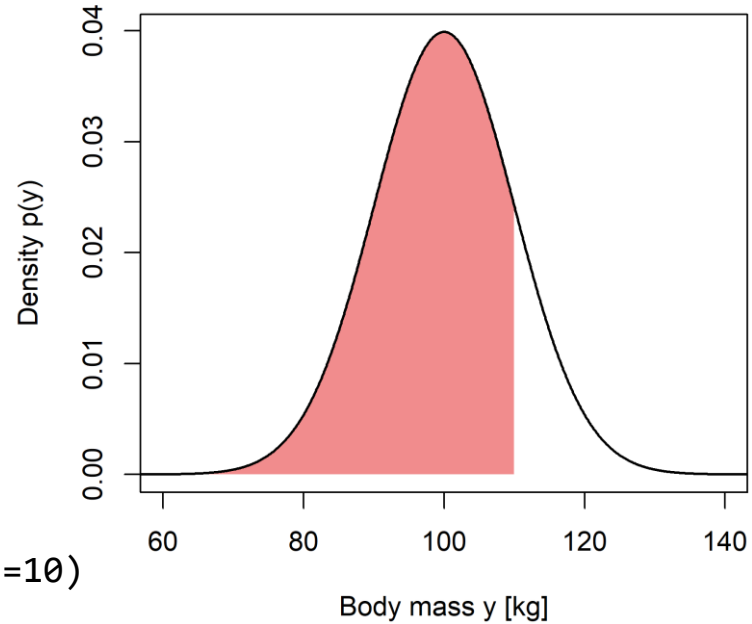
- Compute probabilities of an **interval**, for example

  $P(y \leq 110) = \int_{-\infty}^{110} p(y|100,10)dy = 0.841$

  ```
  > pnorm(110,mean=100,sd=10)
  ```

- $P(90 \leq y \leq 110) = \int_{90}^{110} p(y|100,10)dy = 0.682$

  ```
  > pnorm(110,mean=100,sd=10)-pnorm(90,mean=100,sd=10)
  ```



Continuous distribution

- Probabilities always integrate to 1 (area under the curve):

  $\int p(y|\mu,\sigma)dy = 1$ for any μ,σ

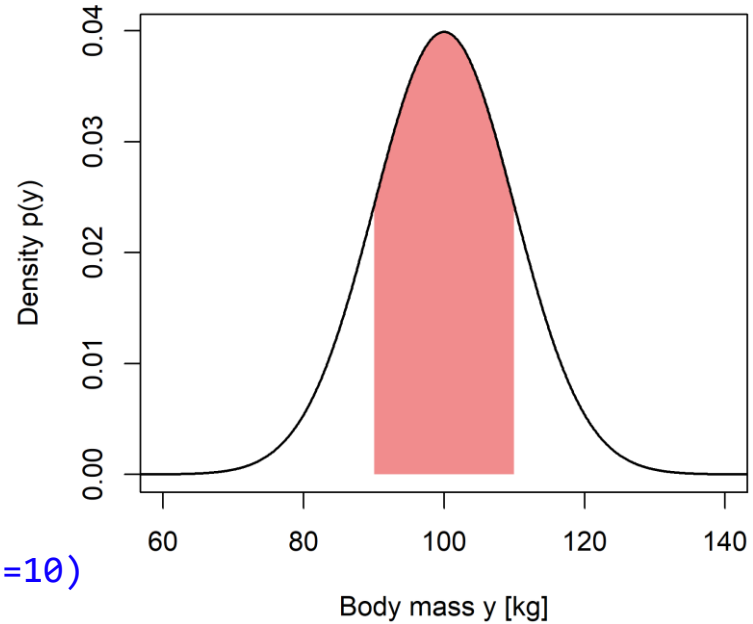
- Compute probabilities of an **interval**, for example

  $P(y \leq 110) = \int_{-\infty}^{110} p(y|100,10)dy = 0.841$

  ```
  > pnorm(110,mean=100,sd=10)
  ```

- $P(90 \leq y \leq 110) = \int_{90}^{110} p(y|100,10)dy = 0.682$

  ```
  > pnorm(110,mean=100,sd=10)-pnorm(90,mean=100,sd=10)
  ```



Summary: distributions

A random variable y has a distribution

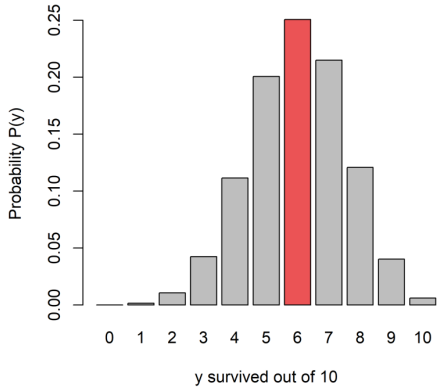
→ It has a known function to calculate probabilities

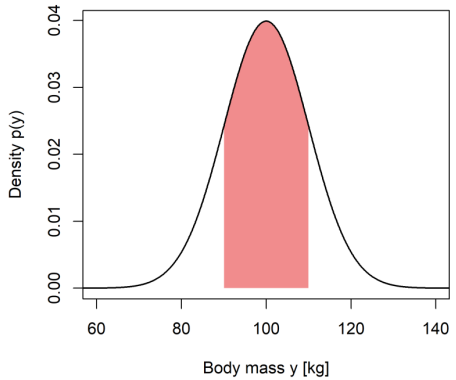
Discrete variable: „probability mass function"

- Assings actual probability to every single value
- Probabilities sum up to 1

Continuous variable: „probability density function"

- Assings probability density to every single value
- Actual probability of an interval = integral over density
- Density integrates to 1



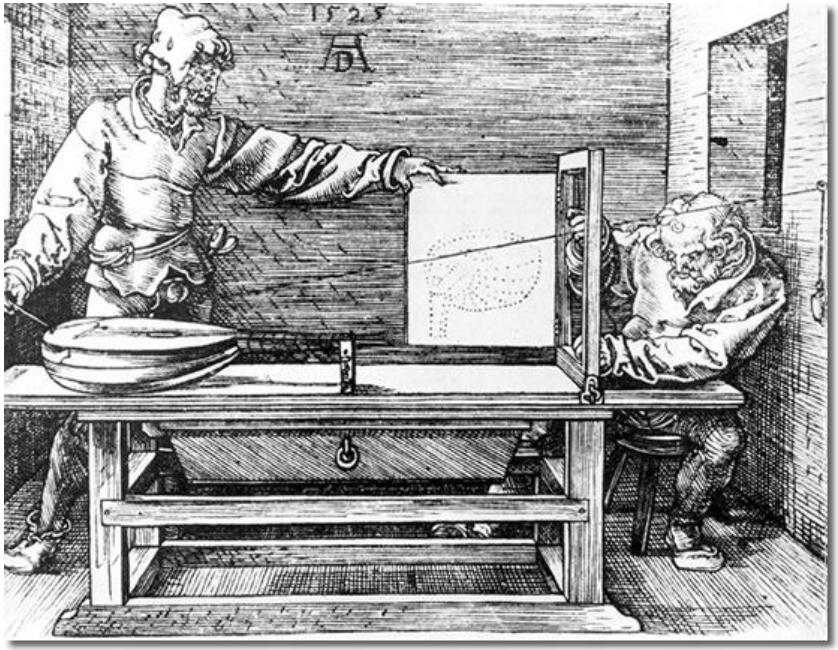


Statistical modeling

Why we need models

- Nature is complex. We need to simplify!
- Models are (mathematical) **abstractions** from nature.
- Explain **patterns** observed in nature (associations, trends, differences, …)
- Make **quantitative** statements.

→ Models can make sense out of your data!



Prediction and inference

Model

Does model describe data well?
→ **prediction**

Statements about processes / hypotheses
→ **inference**

Data / Nature

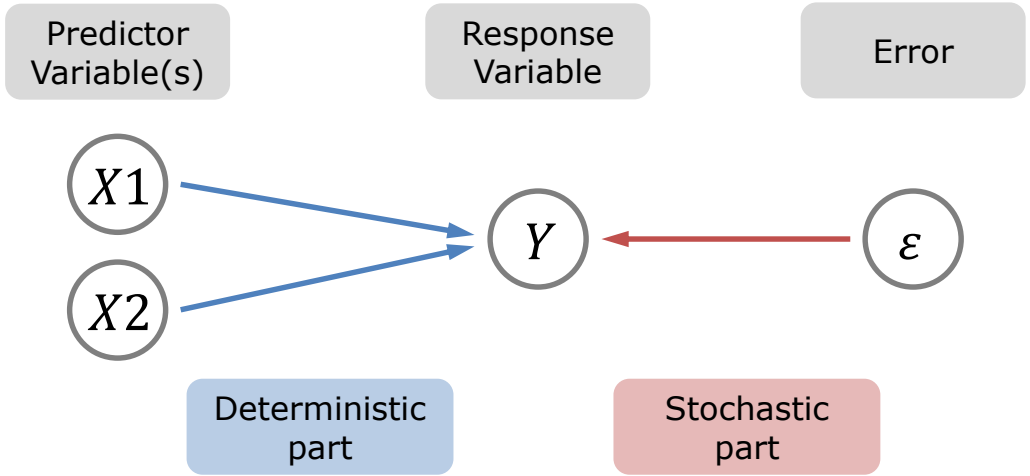
- Bring model **predictions** in correspondance with observed data

  Model fitting: estimate model parameters

  Model selection: choose between different models

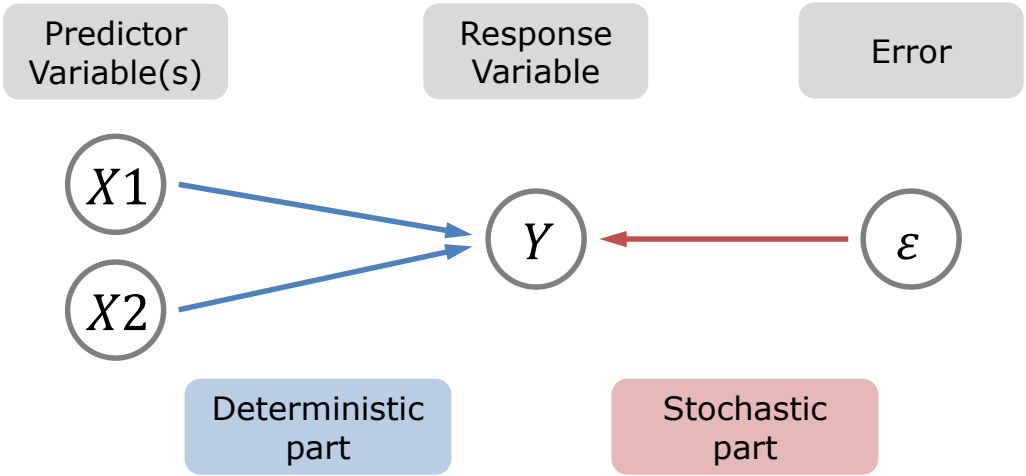
- **Inference:** What does the data tell me about the model (e.g. positive trend)?

Statistical model: building blocks



- Model the process that generates the data:
- We want to learn the association of a **single response** variable Y with **one or more predictor** variables $X1$, $X2$, …
- Predictors can be **categorical** (factor, e.g. „warm" vs „cold" treatment) or **continuous** (e.g. exact temperature values $11.0°C, 13.9°C, 12.1°C, …$)

Statistical model: building blocks



- Deterministic part:
  Prediction model, e.g. mean regression line
- Stochastic part:
  The prediction model cannot explain response perfectly, include random error
- Deterministic and stochastic parts both have **parameters** (e.g. effect sizes)

Example: linear regression

Example: linear relationship between age x and body mass y of sea turtles

Deterministic part: $\mu(x) = a + b \cdot x$ ← Probably a simplification!

Stochastic part: $y \sim \text{Normal}(\mu, \sigma)$ ← Connects the det. model to the data

Parameters:

a intercept

b slope

σ standard deviation

Example: linear regression

Example: linear relationship between age x and body mass y of sea turtles

Deterministic part: $\mu(x) = a + b \cdot x$

Probably a simplification!

Stochastic part: $y \sim \text{Normal}(\mu, \sigma)$

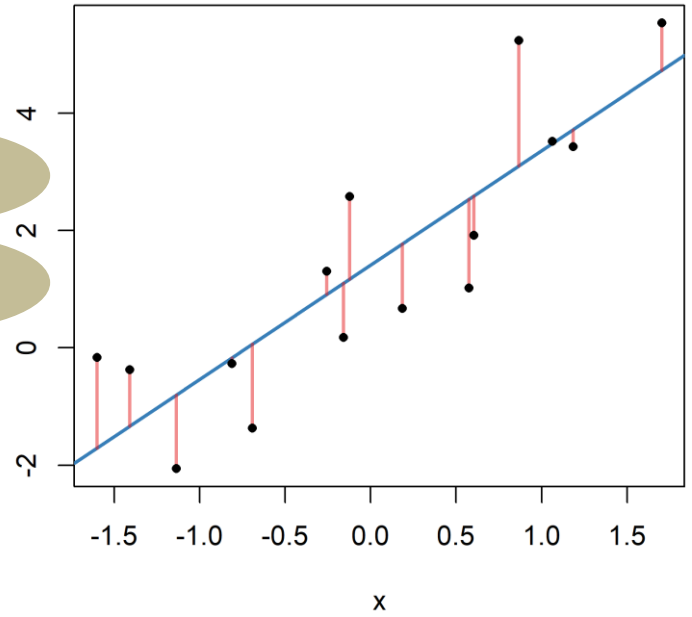
Connects the det. model to the data

Parameters:

a intercept

b slope

σ standard deviation



Example: linear regression

Data: independent observations
$(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$

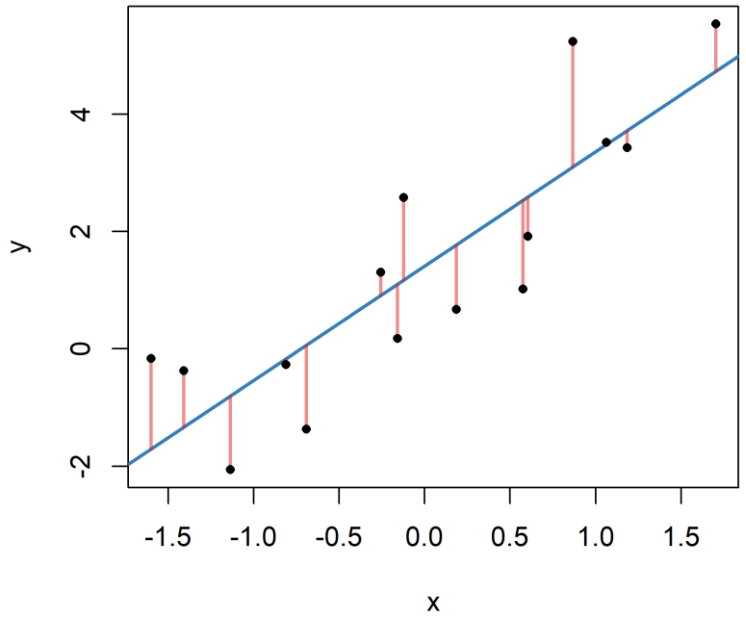
Deterministic part: $\mu_i = a + b \cdot x_i$

Stochastic part: $y_i \sim \text{Normal}(\mu_i, \sigma)$

Can be rewritten: $y_i = \mu_i + \varepsilon_i$

$\varepsilon_i \sim \text{Normal}(0, \sigma)$

ε_i **residuals** (difference between pred. and obs.)



Example: linear regression

Question:

Do datapoints $y_1 \ldots y_n$ need to come from a joint normal distribution?

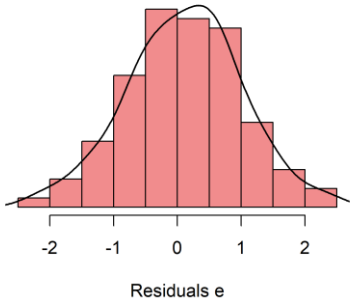
Answer:

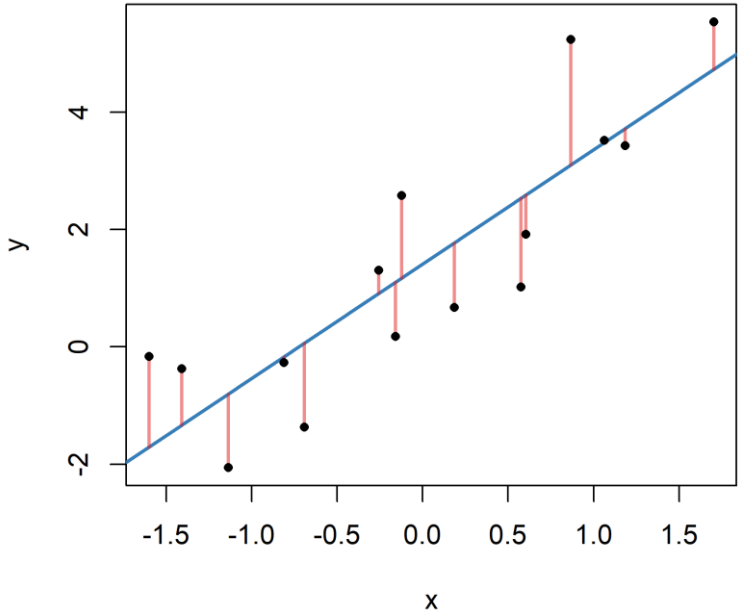
No, assumption not about the response values y_i !!!

Response y_i has shifting mean: μ_i

Assumption is about the **residuals** ε_i,

they have a joint zero mean and joint sdev σ





Assumptions in linear regression

1. Independent observations.

   Systematic differences in y are because of x !

2. Trend of y follows (linear) prediction model

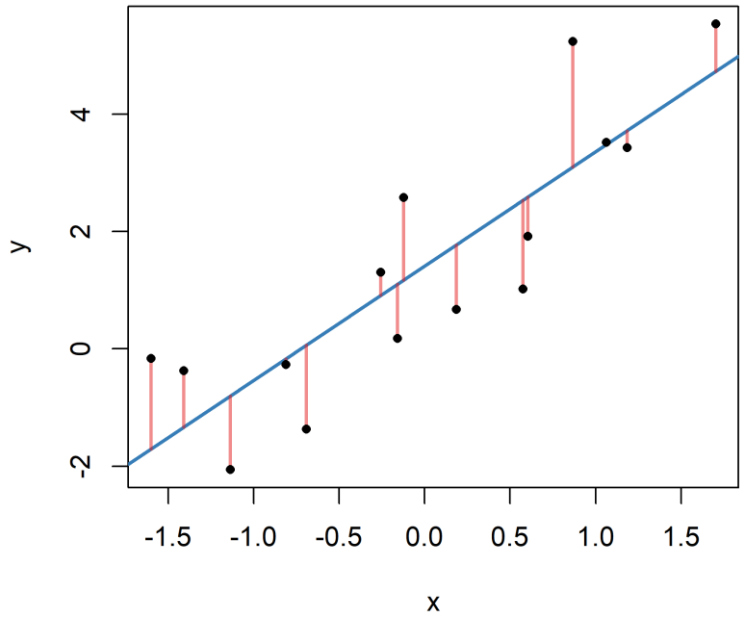
   $\mu(x) = a + b \cdot x$

3. Residuals follow normal distribution

   $\varepsilon \sim \text{Normal}(0, \sigma)$

4. Constant variance (standard deviation)

   across whole range of x



Assumptions in linear regression

1. Independent observations.

   Systematic differences in y are because of x !

2. Trend of y follows (linear) prediction model

   $\mu(x) = a + b \cdot x$

3. Residuals follow normal distribution

   $\varepsilon \sim \text{Normal}(0, \sigma)$

4. Constant variance (standard deviation)

   across whole range of x

Beyond linear models

Mixed effects / hierarchical models can account for grouping factors like „plot“

Generalized linear models, or even nonlinear models allow a wide range of trends

Choose other residual distributions to model y (e.g. Poisson for count)

Other distributions with non-constant variance available (e.g. for overdispersion)

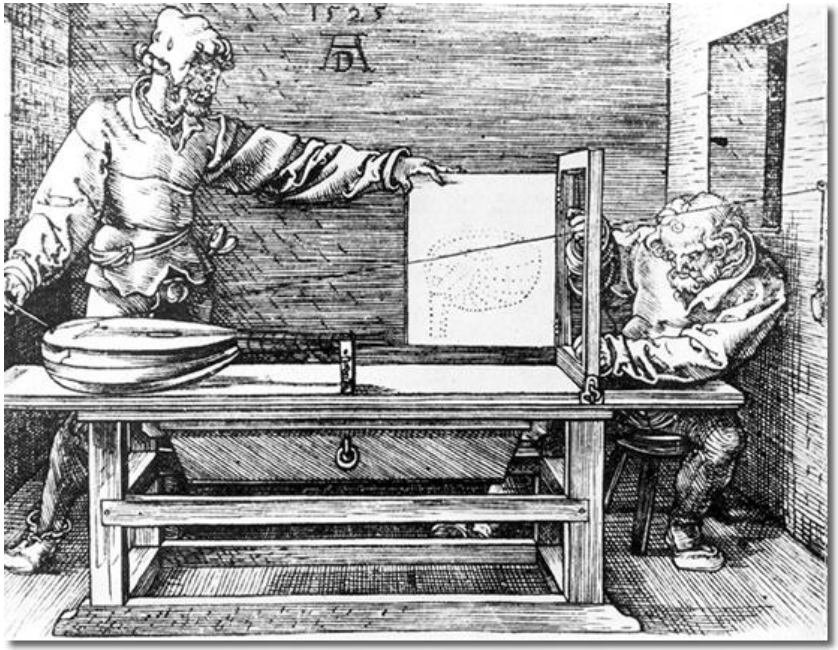
Statistical modeling

There is no such thing as a „Bayesian model"!

Statistical model:
- Deterministic part
- Stochastic part
- Model assumptions

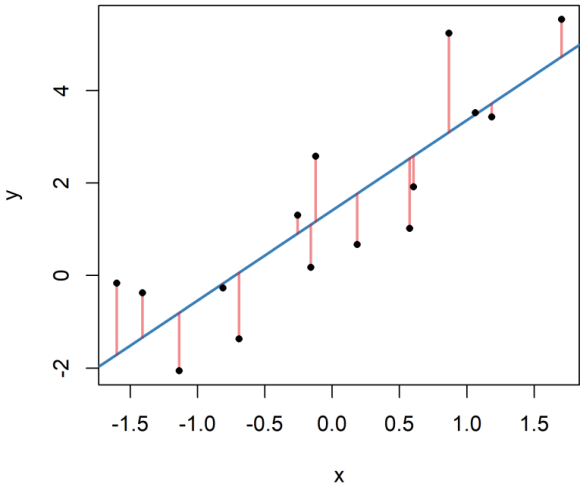
2 approaches to model fitting / parameter estimation / statements about hypotheses:
- Frequentist statistics
- Bayesian statistics

They are different in the way model parameters are computed and how their **uncertainty** is treated.

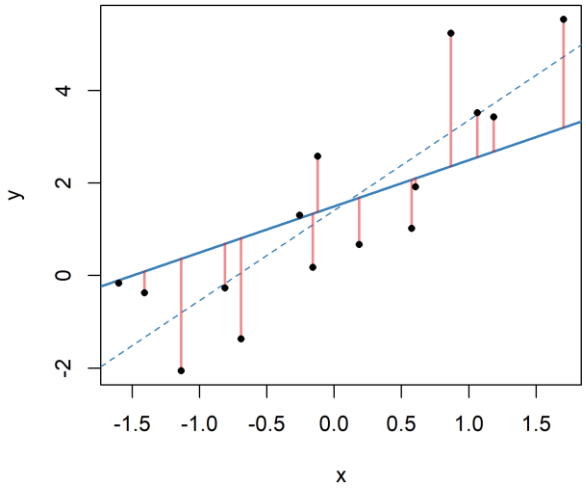


Maximum likelihood estimation

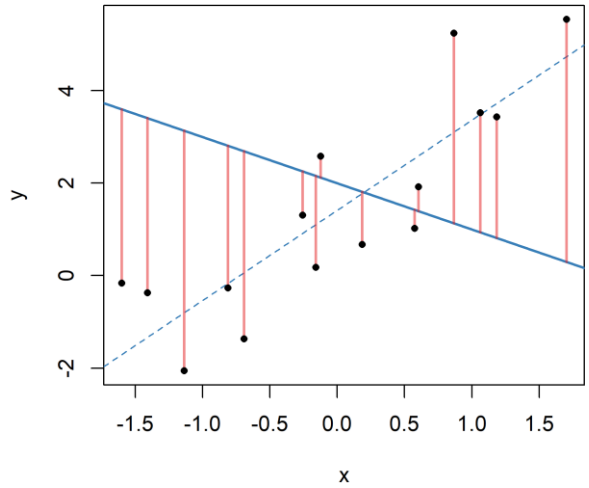
How to estimate parameters?



Best model fit:
$a = 1.41 \quad b = 1.94$



Worse fit:
$a = 1.5 \quad b = 1.0$

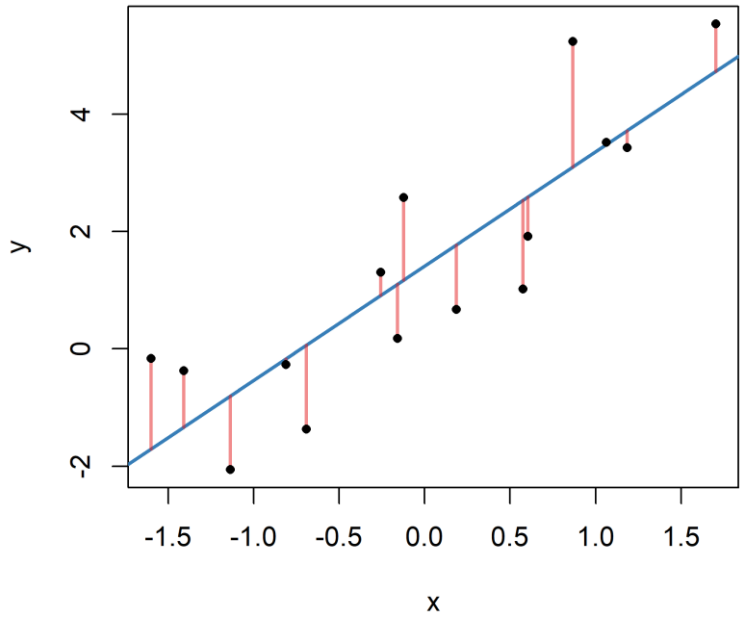


Really bad fit:
$a = 2.0 \quad b = -1.0$

How to estimate parameters?

- Ordinary **least-squares**
- Find intercept a and slope b that

  minimize $\sum_{i=1}^{n}(y_i - \mu_i)^2$ (sum of squares)
- Works perfectly for linear models
- Formulas for intercept and slope(s) available!



- But what about other models (GLM, LMM, …)?
- Other measure of model fit?
- Stochastic part of the model → Probability distribution of datapoints

The likelihood function

Example: survival rate

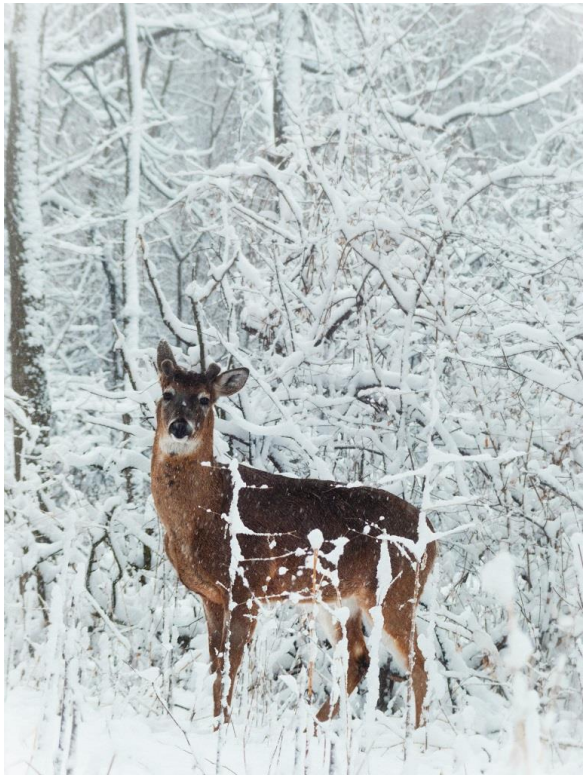
Statistical model: deterministic part: $\mu = \theta$

stochastic part: $y \sim \text{Binomial}(N, \theta)$

Probability: data unknown, parameters given
- The average survival rate is $\theta = 0.6$
- How many of the 10 individuals will survive the winter?

Likelihood: parameters unknown, data given
- Last winter, 6 out of 10 individuals survived
- What is the average survival rate?

→ Likelihood is the **reverse** of probability !



The likelihood function

Model

Prediction

Probability
$P(\boldsymbol{y}|\theta)$

Likelihood
$L(\boldsymbol{\theta}|y) = P(y|\boldsymbol{\theta})$

Inference

Data

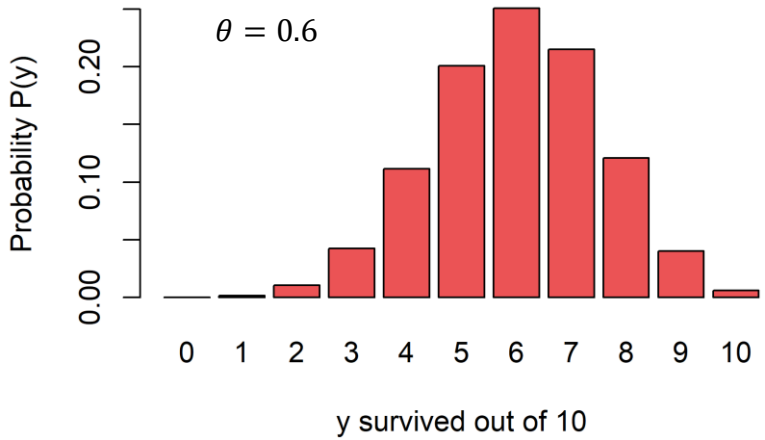
The likelihood function

Probability is function for unknown data

unknown given

$$P(\boldsymbol{y}|\theta) = \binom{N}{\boldsymbol{y}} \theta^{\boldsymbol{y}} (1-\theta)^{N-\boldsymbol{y}}$$

$$= \binom{10}{\boldsymbol{y}} 0.6^{\boldsymbol{y}} (1-0.6)^{10-\boldsymbol{y}}$$

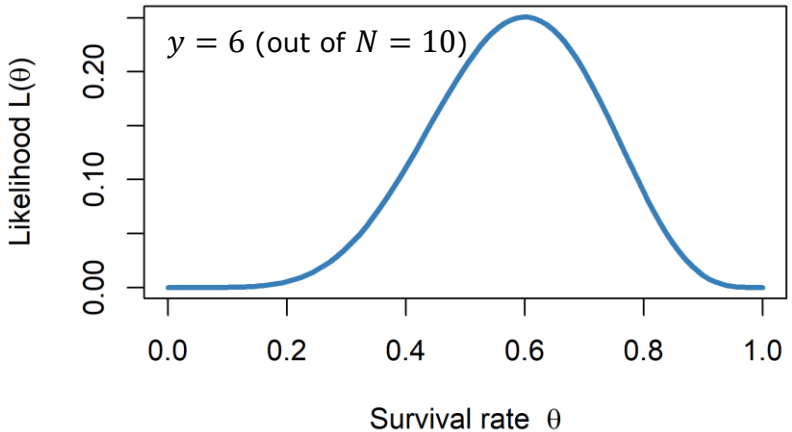


Likelihood is function for unknown parameters

unknown given

$$L(\boldsymbol{\theta}|y) = \binom{N}{y} \boldsymbol{\theta}^{y} (1-\boldsymbol{\theta})^{N-y}$$

$$= \binom{10}{6} \boldsymbol{\theta}^{6} (1-\boldsymbol{\theta})^{10-6}$$



Maximum likelihood estimation

Given:
Data y and statistical model

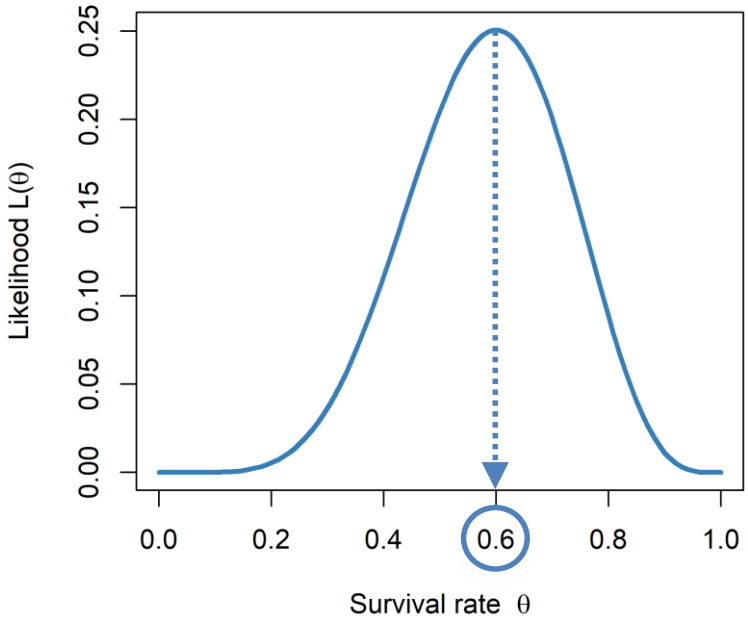
→ Defines likelihood function $L(\theta|y) = p(y|\theta)$

How likely did a parameter value θ produce the observed data?

Find the value for which the likelihood is highest!

→ We get a **point estimate** θ^*

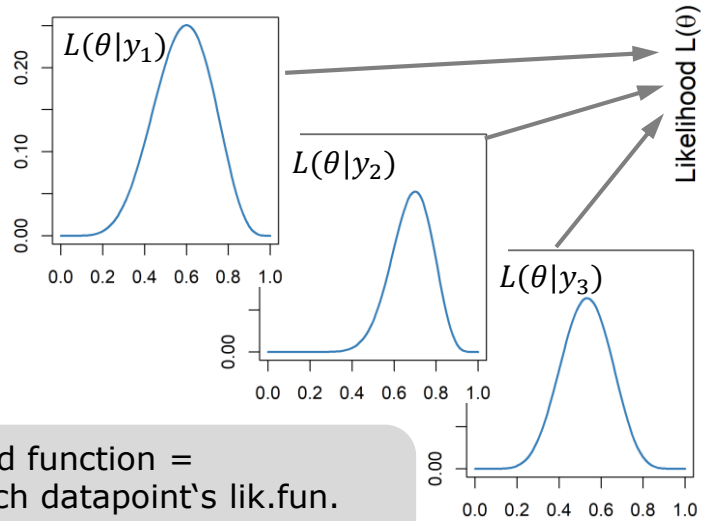
„Maximum likelihood estimate"

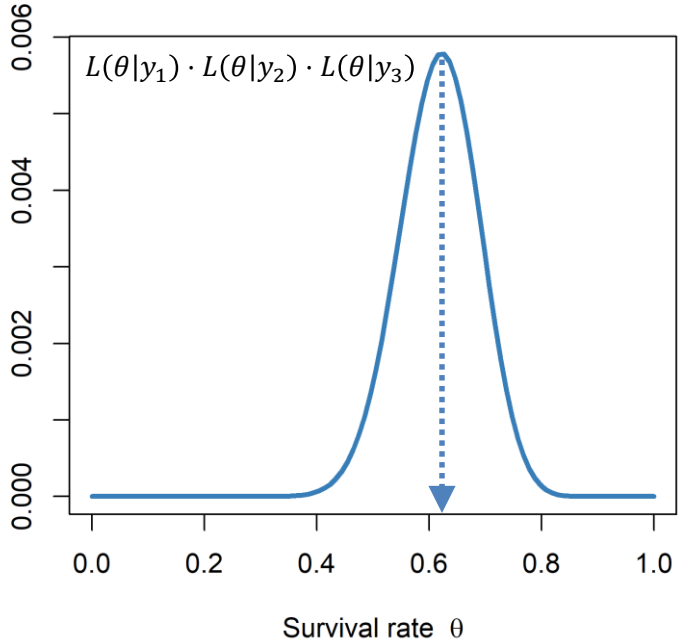


Maximum likelihood estimation

Now: **multiple observations**

Survival: $y_1 = 6/10,\ y_2 = 14/20,\ y_3 = 8/15$





Joint likelihood function =
product of each datapoint's lik.fun.

$$L(\theta|y) = L(\theta|y_1) \cdot L(\theta|y_2) \cdot L(\theta|y_3)$$

Maximum likelihood estimation

Example: linear regression

Deterministic part: $\mu(x) = \boldsymbol{a} + \boldsymbol{b} \cdot x$

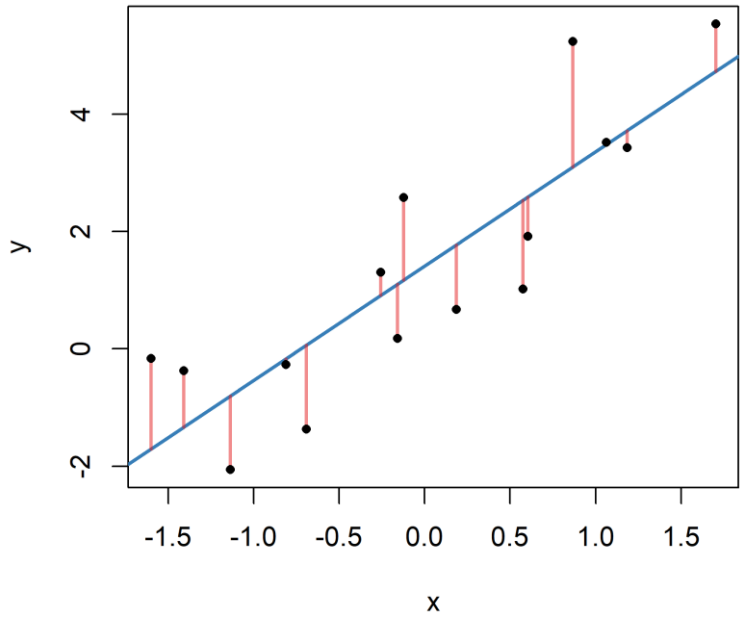
Stochastic part: $y \sim \text{Normal}(\mu, \boldsymbol{\sigma})$

3 parameters: intercept a, slope b, sdev σ

$$L(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\sigma}|y) = p(y|a, b, \sigma)$$
$$= p(y_1|a, b, \sigma) \cdot \ldots \cdot p(y_n|a, b, \sigma)$$

Now it's getting more complicated:

Find $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\sigma}$ that maximizes $L(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\sigma}|y)$

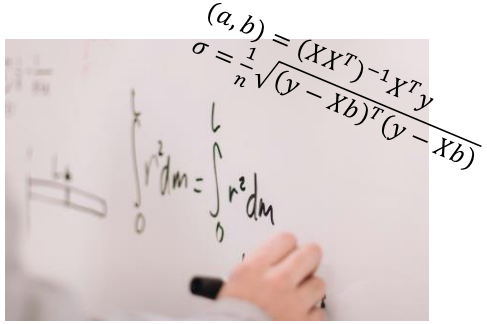


Maximum likelihood estimation

1) Analytical solution: find a mathematical formula for θ

→ Works for linear models with normal distribution

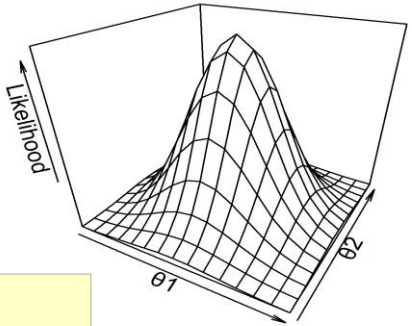
→ But too complicated for most applications



2) Brute force (e.g. grid)

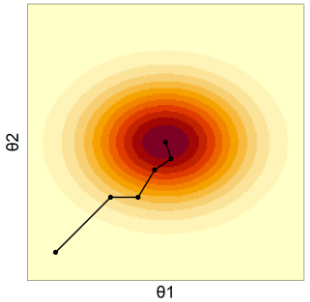
→ Effort grows exponentially with number of parameters

→ Too expensive for most applications



3) Numerical optimization

→ Iterative algorithm that tries to improve $L(\theta|y)$ in every step until no further improvement is possible



Beyond point estimates ?

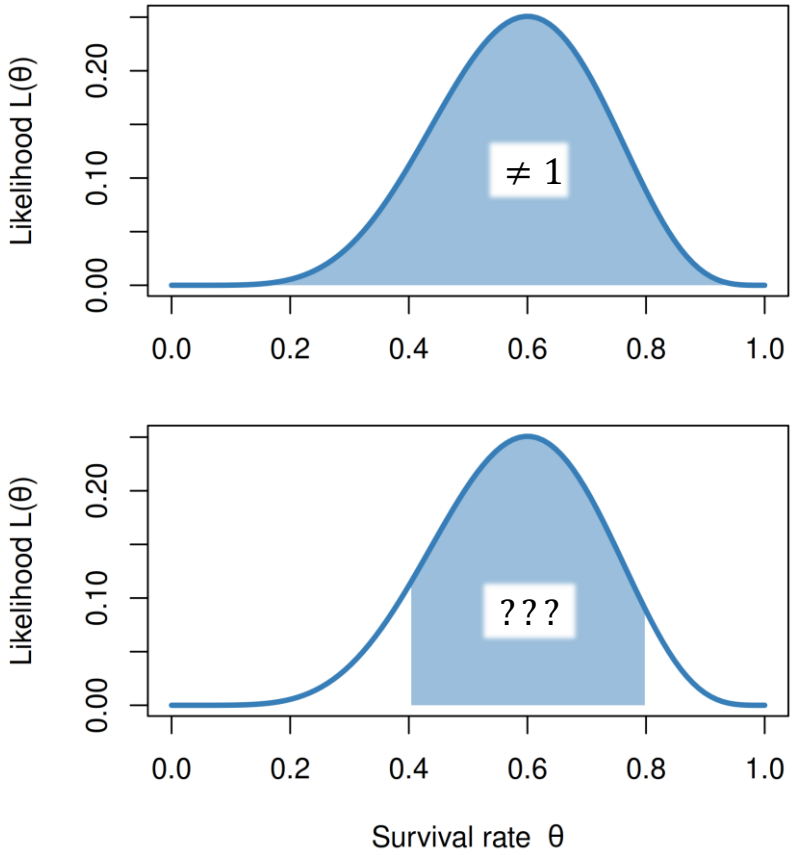
Why can't we use the likelihood for probability statements on the parameters ?

$L(\boldsymbol{\theta}|y)$ is not a probability density function for parameters $\boldsymbol{\theta}$!

$\int L(\boldsymbol{\theta}|y) \neq 1$ (area under the curve)

E.g. $\int_{0.4}^{0.8} L(\boldsymbol{\theta}|y)$ is a meaningless value.
It does **not** describe $P(0.4 < \boldsymbol{\theta} < 0.8)$!

But likelihood tells us that, e.g., survival rate of 0.3 is less likely than 0.5. Can we use that?



Beyond point estimates ?

$L_{new}(\boldsymbol{\theta}|y) = \frac{L(\boldsymbol{\theta}|y)}{c}$ scale by constant $c = \int L(\theta|y)$

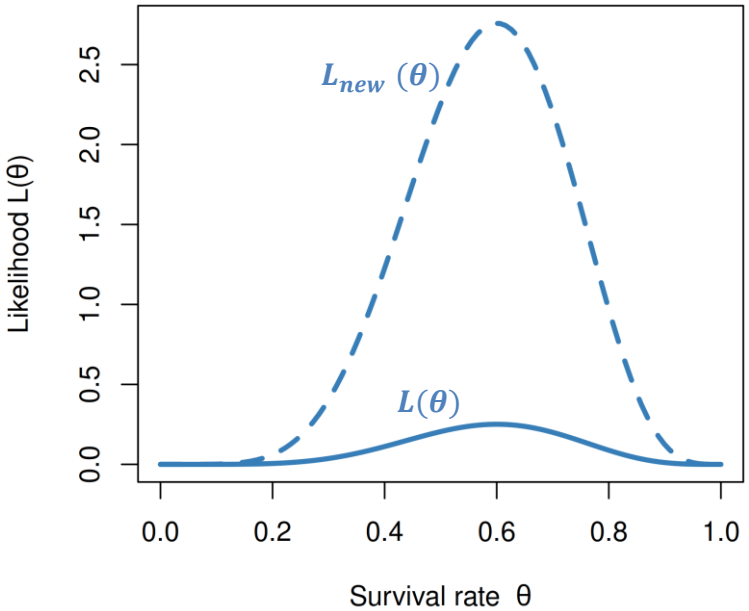
$\int L_{new}(\boldsymbol{\theta}|y) = 1$ (area under the curve)

Probability statements would be possible!

E.g. $\int_{0.4}^{0.8} L_{new}(\boldsymbol{\theta}|y) = P(0.4 < \boldsymbol{\theta} < 0.8)$

But we arrived at the same problem:
Can't compute the integral $c = \int L(\boldsymbol{\theta}|y)$

It's not practical. Solution in next lecture!

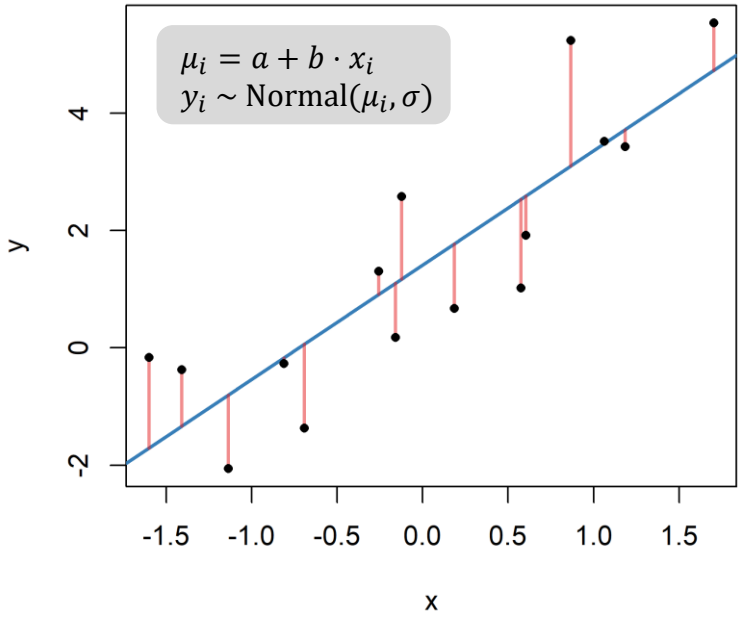


Summary

Summary MLE

- Every statistical model has a likelihood function, defined by distribution of the stochastic part, that connects deterministic part to data (prob of the data, given a fixed parameter)
- Find model parameters such that observed data is most likely
- Maximum likelihood estimation → point estimates
- Does not allow probability statements about the model parameters $P(\theta|y)$

→ The frequentist „short cut“:
Null hypothesis significance testing (NHST)



Further reading

Fieberg, J. (2024). Statistics 4 Ecologists. https://statistics4ecologists-v2.netlify.app/ [Chapters 1,9,10]

Inchausti, P. (2023). Statistical Modeling With R: a dual frequentist and Bayesian approach for life scientists. *Oxford University Press*. [Chapter 3]

Kery, M. & Kellner, F. (2024): Applied Statistical Modelling for Ecologists. *Elsevier.* [Chapters 1,2]

Essington, T. (2021). Introduction to Quantitative Ecology. *Oxford University Press*. [Chapter 8]

Warton, D. (2022). Eco-Stats: Data Analysis in Ecology. *Springer (Methods in Statistical Ecology)*. [Chapter 1]