

# Introduction to Bayesian Statistics

## *Part 3* Prior & Posterior Distributions

Benjamin Rosenbaum

iDiv 2025



# This lecture

Short summary of last lecture

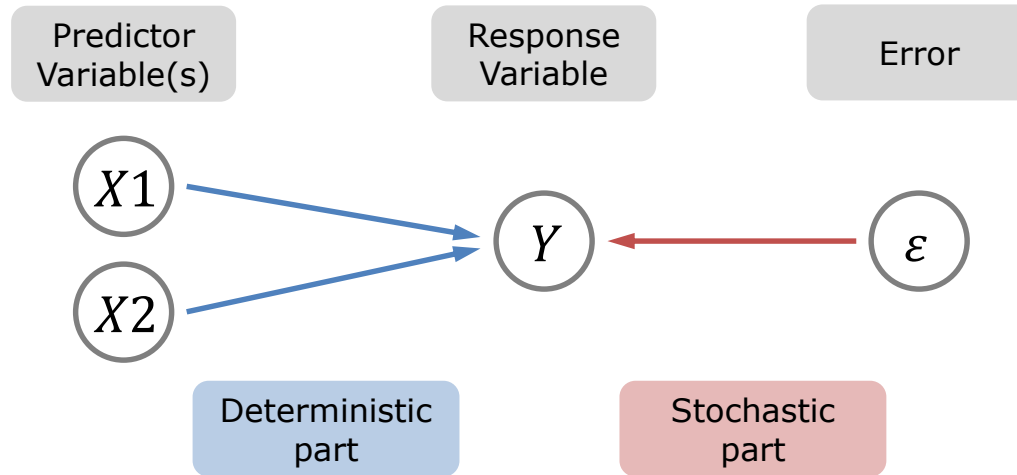
Some useful distributions

The prior distribution

The posterior distribution

Posterior predictions and model evaluation

## From last lecture



# From last lecture

## Bayes' rule

The diagram shows the Bayes' rule equation  $p(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{p(y)}$  inside a light beige rounded rectangle. Four arrows point to the components of the equation: 'Likelihood function  $L(\theta)$ ' points to  $p(y|\theta)$ , 'Prior distribution' points to  $p(\theta)$ , 'Posterior distribution' points to  $p(\theta|y)$ , and 'Normalization constant Independent of  $\theta$ ' points to  $p(y)$ .

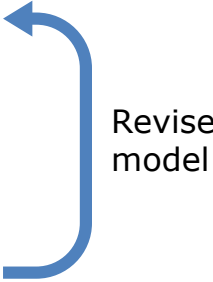
$$p(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{p(y)}$$

Update beliefs (prior)  
by gaining new information (data & likelihood)

Posterior distribution used for  
quantitative & direct statements on research questions

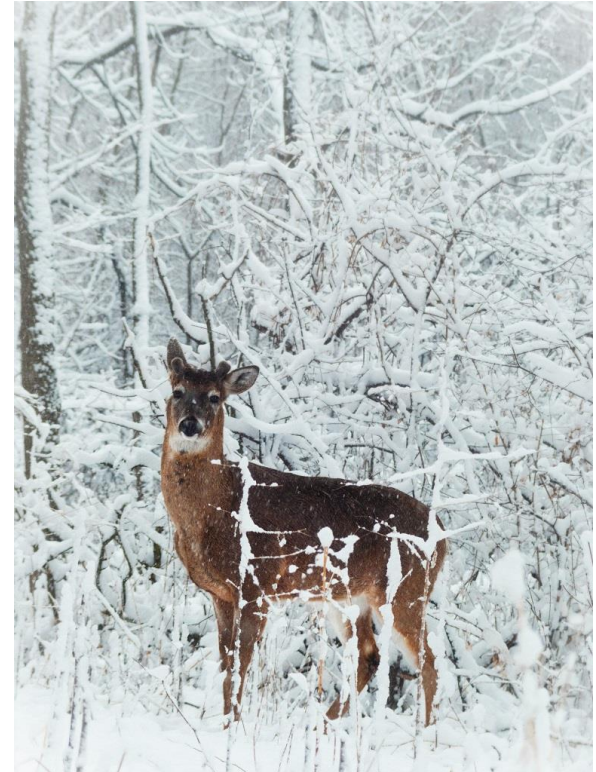
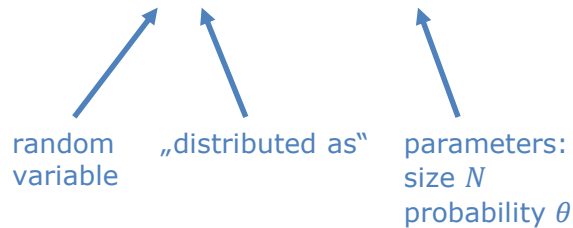
Don't have access to posterior distribution  
Approximate by MCMC sampling

# From last lecture

- 1) Research question (hypotheses)
  - 2) Data collection
  - 3) Statistical model
  - 4) Prior distribution choice
  - 5) Model fitting (MCMC)
  - 6) Evaluate model output
  - 7) Quantitative statements on hypotheses
- 
- ```
graph TD; 3((3)) --> 4((4)); 4 --> 5((5)); 5 --> 6((6)); 6 -- "Revise model" --> 3;
```

# From last lecture

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



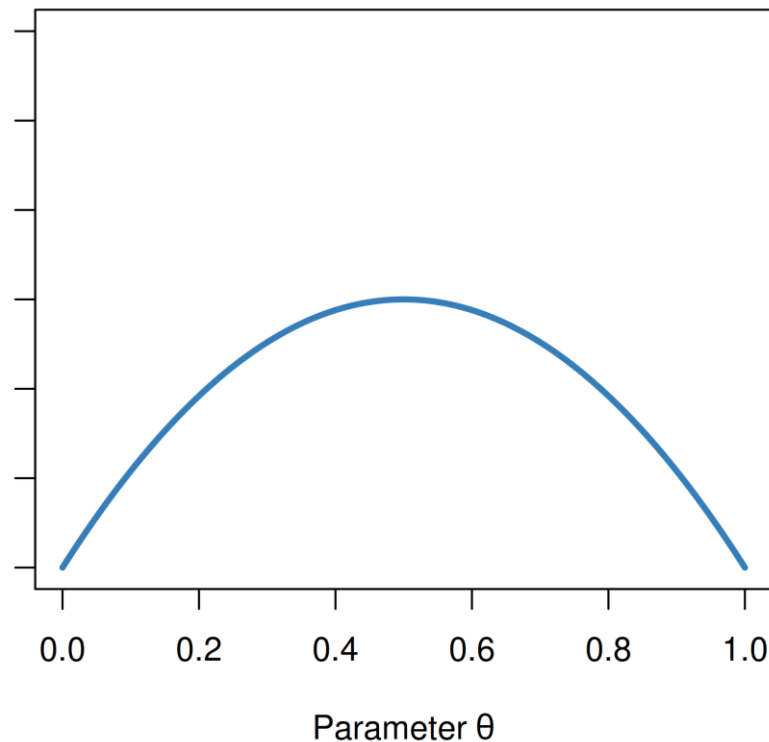
# From last lecture

## Prior distribution

Chosen by **you**

Density known over full parameter range

$$p(\theta) = \text{dbeta}(\theta \mid 2, 2)$$



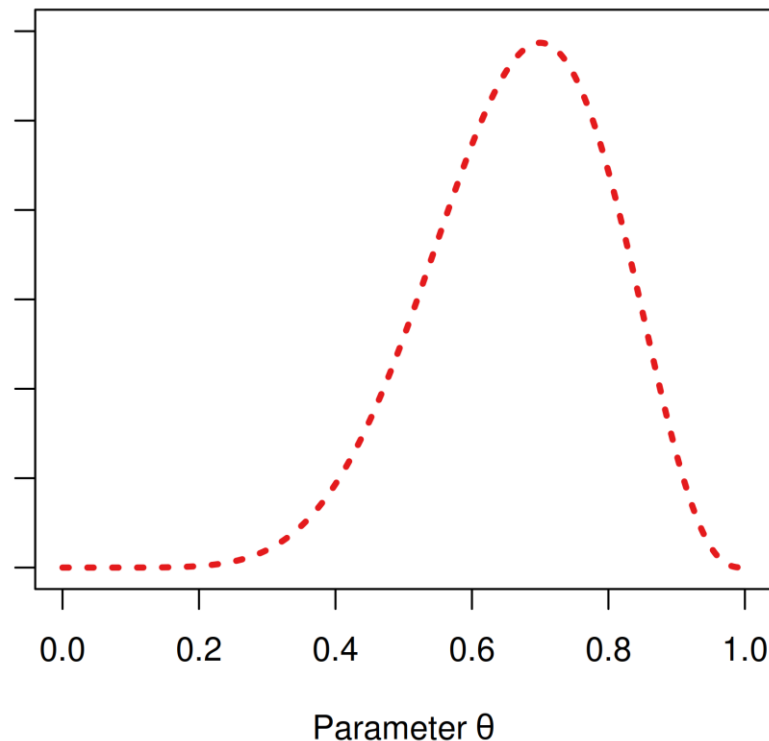
# From last lecture

## Likelihood function

Defined by **your** data and **your** statistical model  
(deterministic & stochastic part)

Can be computed for every single parameter value  
But values not known over full parameter range

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n p(y_i | \theta) \\ &= \prod_{i=1}^n \text{dBinom}(\text{survived}_i, \text{total}_i | \theta) \end{aligned}$$





# From last lecture

## Posterior distribution

Software **output**:  
approximation by MCMC sampling

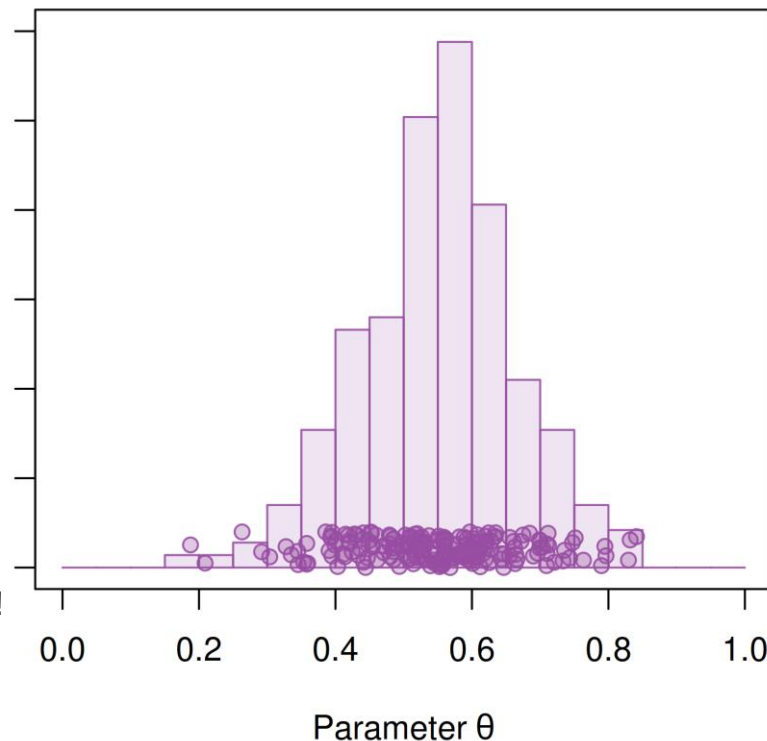
$$p(\theta|y) \sim p(\theta) \cdot L(\theta)$$

Represented by samples  $\theta_1, \theta_2, \theta_3, \dots, \theta_{1000}$  only!

Posterior density values  $p(\theta_1|y), p(\theta_2|y), \dots, p(\theta_{1000}|y)$   
not required, don't need to be saved

**Counting instead of integrating** for hypotheses testing!

Super easy once  
you have a  
posterior sample



# Distribution zoo app

<https://ben18785.shinyapps.io/distribution-zoo/>

Ben Lambert and Fergus Cooper

Last month: used by 203 people over 408 sessions in 33 countries

Since created: used by 19498 people over 36776 sessions in 144 countries

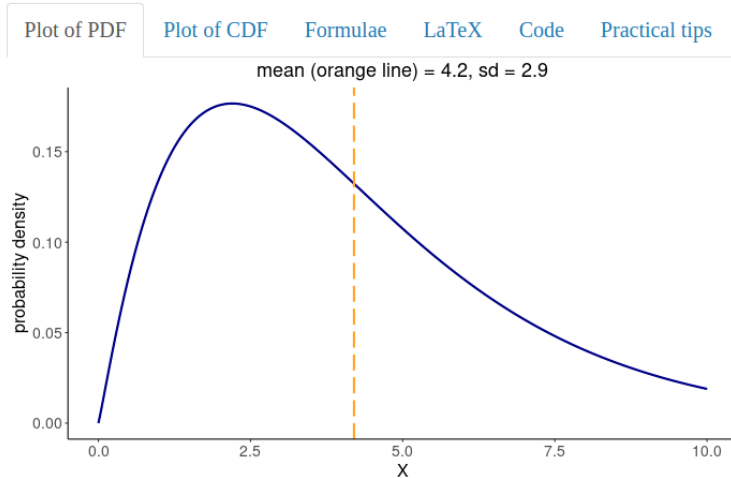
**Category of distribution**  
Continuous Univariate

**Distribution type:**  
Gamma

**Range**  
10 1,000

**Shape parameter**  
0.1 2.1 10

**Rate parameter**  
0.1 0.5 2



# Probability playground app

<https://www.acsu.buffalo.edu/~adamcunn/probability/probability.html>

The gamma distribution is a "waiting time" distribution. Suppose events occur independently and randomly with an average time between events of  $\beta$ . The waiting time until  $\alpha$  events have occurred is a  $\text{gamma}(\alpha, \beta)$  random variable.

The parameter  $\alpha$  is known as the shape parameter, and the parameter  $\beta$  is called the scale parameter. Increasing  $\alpha$  leads to a more "peaked" distribution, while increasing  $\beta$  increases the "spread" of the distribution.

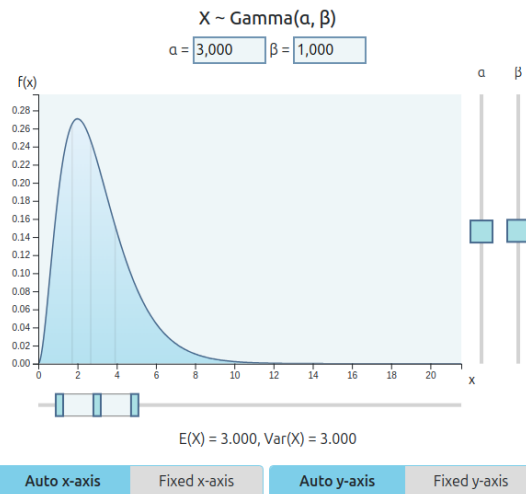
The function  $\Gamma(s)$  in the denominator of the pdf and cdf denotes the gamma function, while the function  $\gamma(s, x)$  in the cdf denotes the lower incomplete gamma function.

| Parameter | Range        | Description     |
|-----------|--------------|-----------------|
| $\alpha$  | $\alpha > 0$ | Shape parameter |
| $\beta$   | $\beta > 0$  | Scale parameter |

| Probability Density Function                                                           | Support             |
|----------------------------------------------------------------------------------------|---------------------|
| $f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$ | $0 \leq x < \infty$ |

| Mean          | Variance        |
|---------------|-----------------|
| $\alpha\beta$ | $\alpha\beta^2$ |

| Example                                                                                                                                   | $\alpha$ | $\beta$ |
|-------------------------------------------------------------------------------------------------------------------------------------------|----------|---------|
| A radioactive substance emits two alpha particles every second on average. Let $X$ be the waiting time for three particles to be emitted. | 3.000    | 0.5000  |
| Cars arrive at an intersection at an average rate of one every two minutes. Let $X$ be the waiting time until five cars have arrived.     | 5.000    | 2.000   |
| Garage door lightbulbs last five years on average and are replaced when they fail. Let $X$ be the time that a box of six bulbs lasts.     | 6.000    | 5.000   |



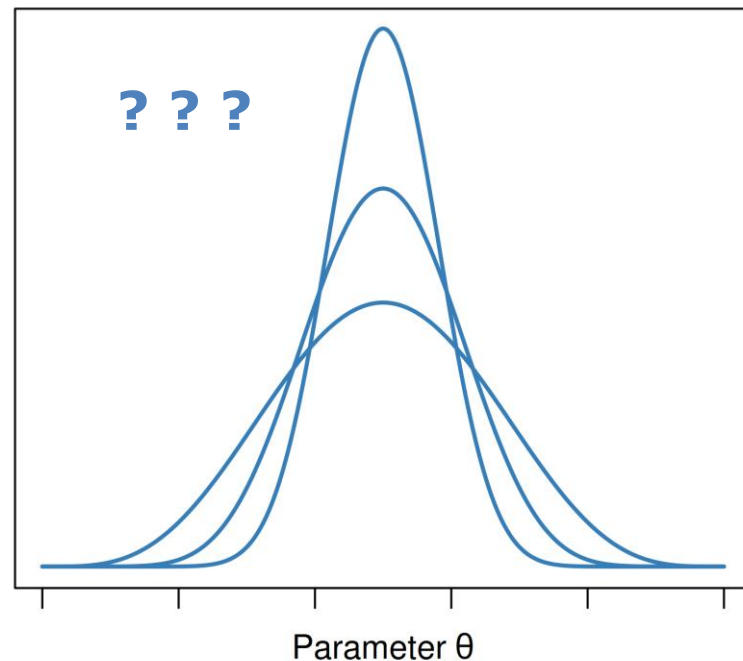
Note that the mean  $\alpha\beta$  is directly proportional to both  $\alpha$  and  $\beta$ . This is what we would intuitively expect - the mean time spent waiting for  $\alpha$  events to occur increases in proportion to both the number of events  $\alpha$  and the average time  $\beta$  between events.

The shape of the pdf depends on the parameter  $\alpha$ . For values of  $\alpha \leq 1$ , the pdf is strictly decreasing. For values of  $\alpha > 1$ , the pdf is unimodal.

## *Prior distributions*

# Prior information

- Priors represent belief about model parameters (for example effect size of an x-y association)
- *Traditional* viewpoint: before we see the data  $y$   
Data information is already contained in the likelihood!
- Use information from
  - General expectation / reasonable range
  - Previous experiments
  - Related studies in the literature
- *Modern* viewpoint:
  - Priors used for regularization
  - Prior predictive checks:  
Are predictions from the prior in the same range / magnitude as observed data?
- Priors are problem-specific



# Types of priors

## Flat / uninformative prior

- You know absolutely nothing about the parameter
- This is rarely the case

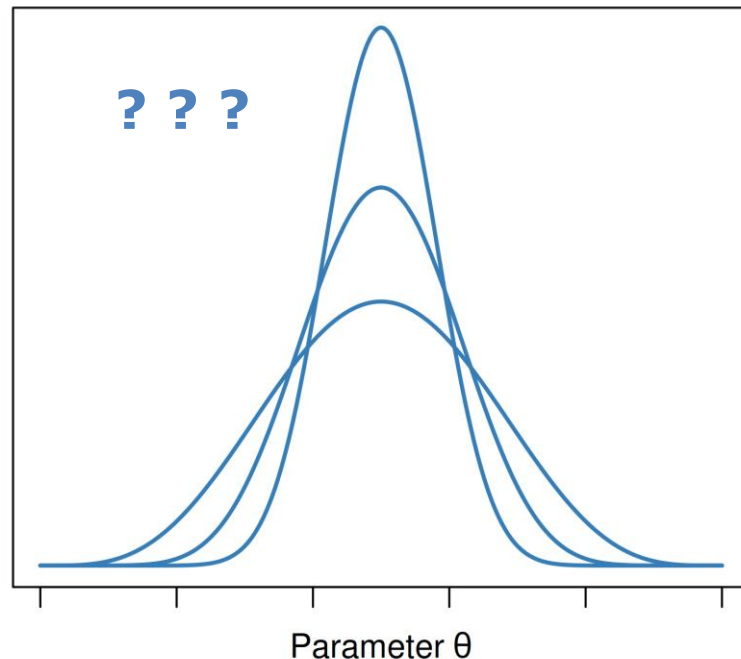
## Vague / weakly informative prior

- You have a vague idea
- For example about the order of magnitude, or sign

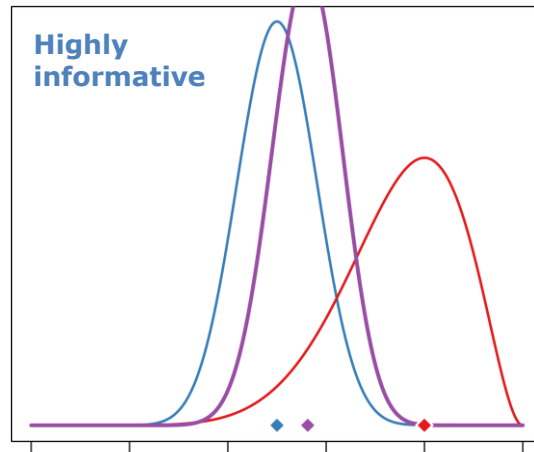
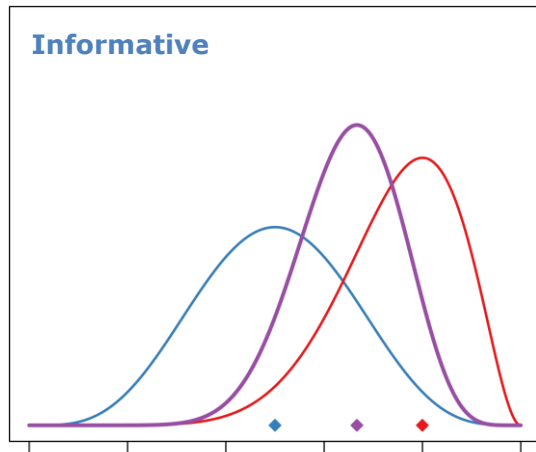
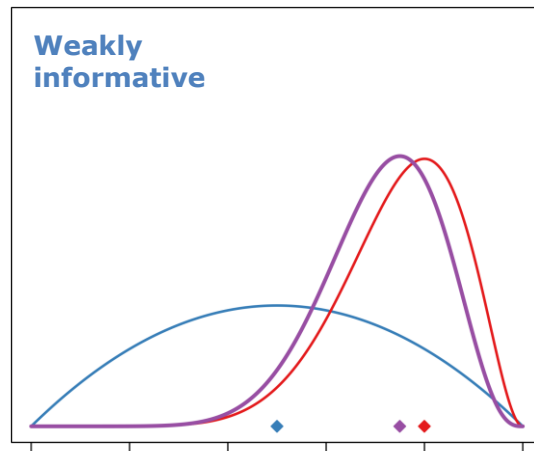
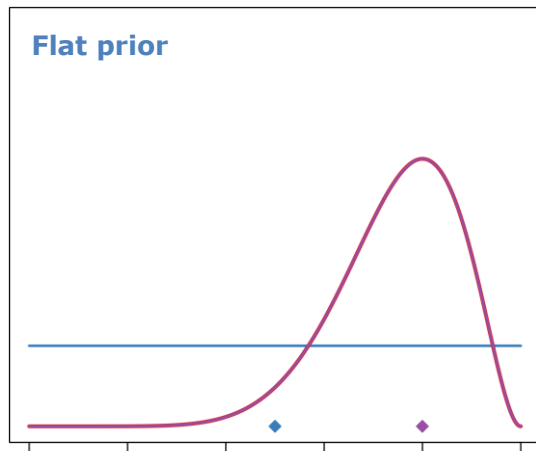
## Informative prior

- You have some idea about the parameter

→ There is no formal definition of these terms!



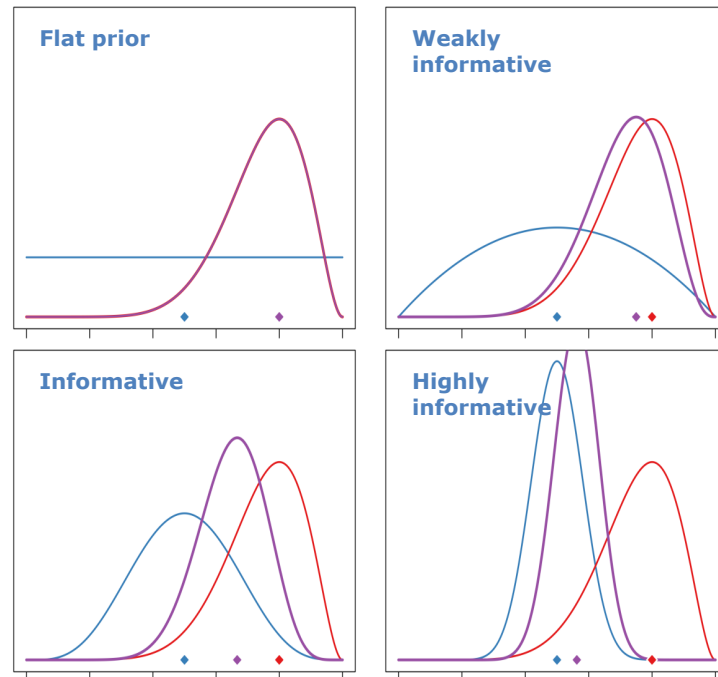
# Prior affects the posterior



- Example: survival rate  $\theta \in [0,1]$
- 1 Observation: 8/10 survived
- Binomial likelihood function
- Priors all beta distributions with mean = 0.5
- but different standard deviations (concentration around mean)

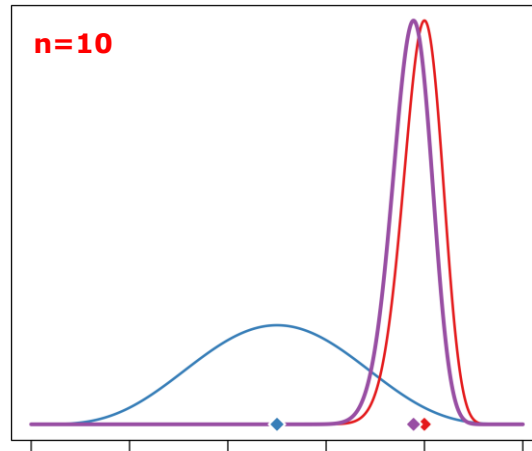
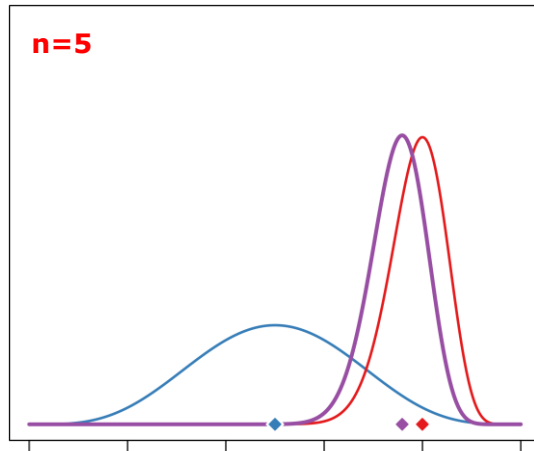
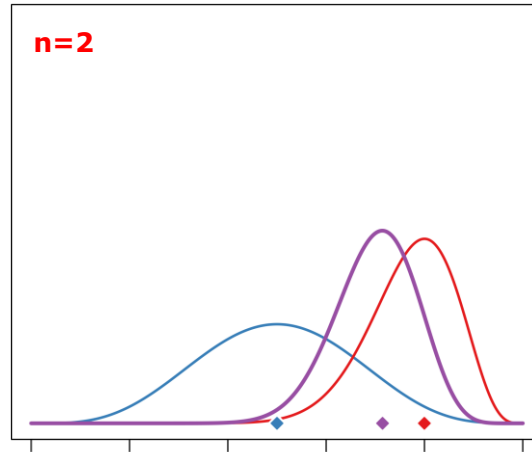
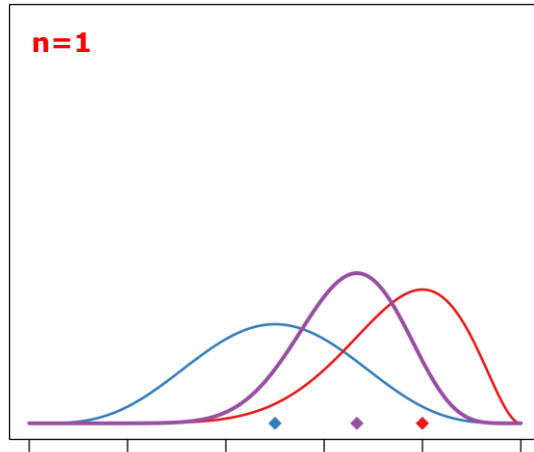
# Prior affects the posterior

- For flat priors, posterior is proportional to likelihood
- For any other prior, posterior is a compromise between prior and likelihood
- For weakly informative priors, even little data dominates the posterior
- More informative priors (lower sdev) draw the posterior mean further away from the maximum likelihood estimate (MLE) towards the prior mean





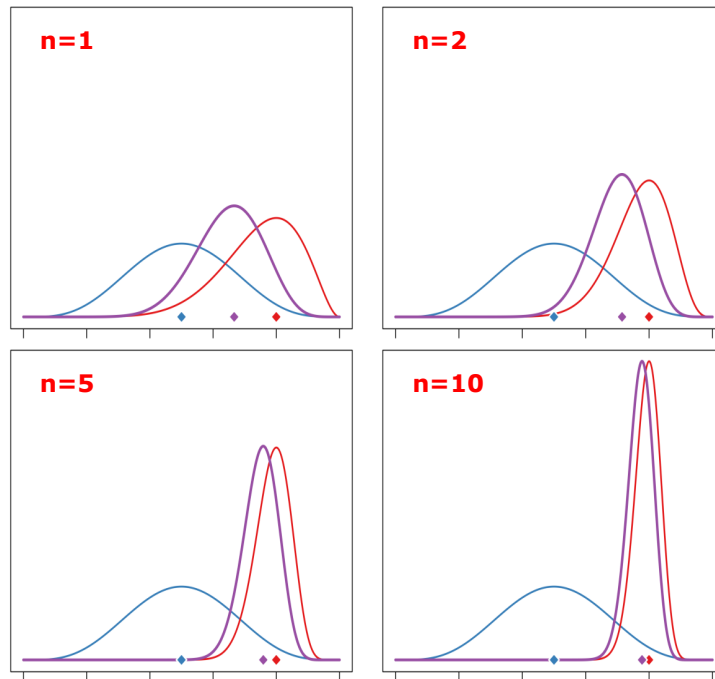
# Likelihood affects the posterior



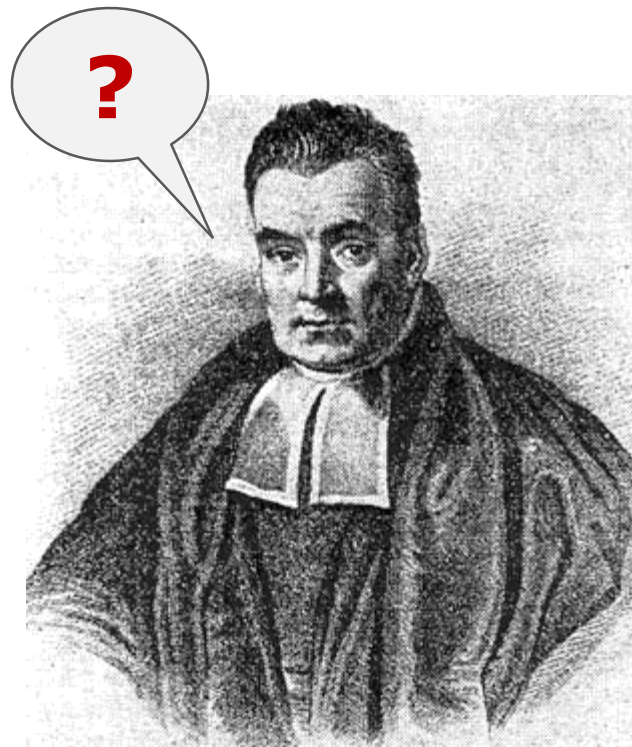
- Example: survival rate  $\theta \in [0,1]$
- „informative prior“ from last slide
- Different numbers of obs.  $n$
- Width of likelihood function decreases with  $n$   
(higher certainty)

# Likelihood affects the posterior

- For small datasets (little experimental evidence), the prior can dominate the posterior
- In large datasets, likelihood can dominate the posterior
- Number of observations decreases the width of the likelihood and therefore also posterior uncertainty (stronger experimental evidence)
- Number of observations draws posterior mean towards maximum likelihood estimate (MLE)



# Are priors subjective?



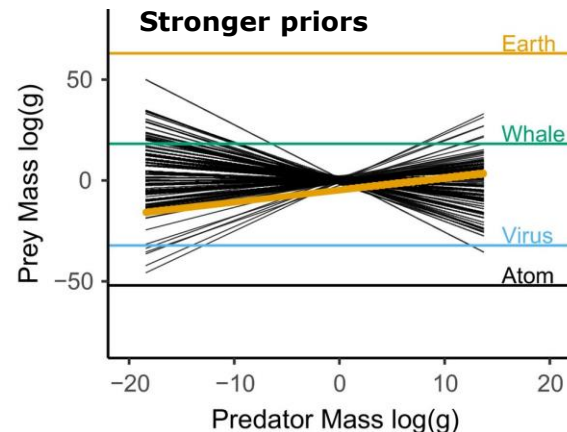
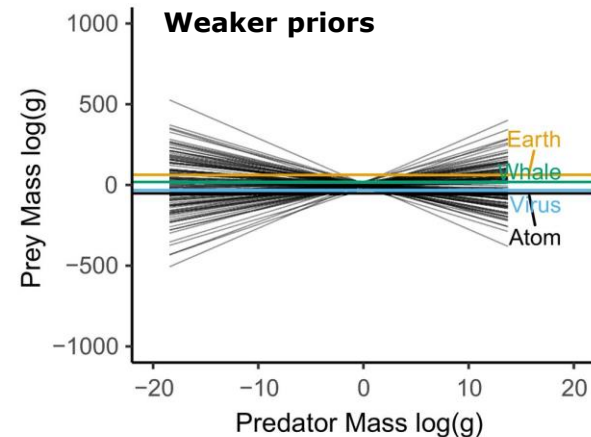
# Prior predictive checks

- Test if priors make sense
- Generate predictions with samples from prior distribution
- Compare them to the range of observed data
- Helpful when using data transformations

(GLMs use nonlinear link-functions, like  $\log$  or  $\text{logit}$  )

- *Traditional* viewpoint (old school):  
Priors should be chosen before even looking at the data
- *Modern* viewpoint: Prior predictive simulations are useful!  
E.g. McElreath: Statistical Rethinking (2020, 2nd ed.)

```
> prior_sim = brm(size~age, data=my_data, prior=my_prior,  
                  sample_prior="only")  
> pp_check(prior_sim)
```



# brms default priors

- brms automatically chooses priors for **intercepts** and **standard deviations**

- Based on the observed data

- Overriding intercept default prior must be handled carefully:

brms internally uses mean-centered predictors, which changes intercept

→ My advice: leave them unless you want to include specific information on these parameters

- brms default priors for effect sizes / regression **slopes** are flat priors!

→ Choose your own priors for them!

```
> prior_summary(fit2)
```

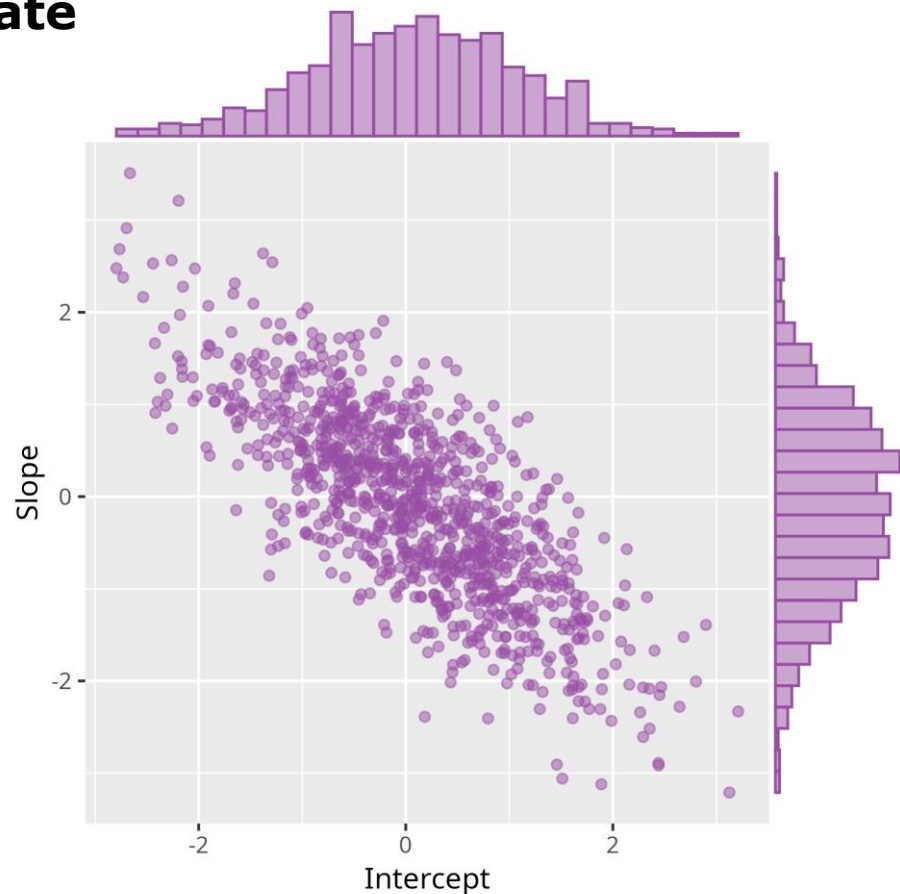
|  | prior                   | class     | coef | group | resp | dpar | nlpar | lb | ub | source       |
|--|-------------------------|-----------|------|-------|------|------|-------|----|----|--------------|
|  | (flat)                  |           | b    |       |      |      |       |    |    | default      |
|  | (flat)                  |           | b    | age   |      |      |       |    |    | (vectorized) |
|  | student_t(3, 110, 11.9) | Intercept |      |       |      |      |       |    |    | default      |
|  | student_t(3, 0, 11.9)   | sigma     |      |       |      |      |       | 0  |    | default      |

*Posterior distribution*

## The posterior sample is multivariate

|       | Intercept | Slope  |
|-------|-----------|--------|
| [1,]  | -0.524    | 0.638  |
| [2,]  | -0.703    | 0.473  |
| [3,]  | -0.200    | 0.942  |
| [4,]  | 0.610     | -0.357 |
| [5,]  | 0.928     | -0.581 |
| [6,]  | 0.807     | -0.941 |
| [7,]  | 1.905     | -1.924 |
| [8,]  | -0.136    | -0.079 |
| [9,]  | -0.799    | 0.403  |
| [10,] | 0.098     | -1.532 |

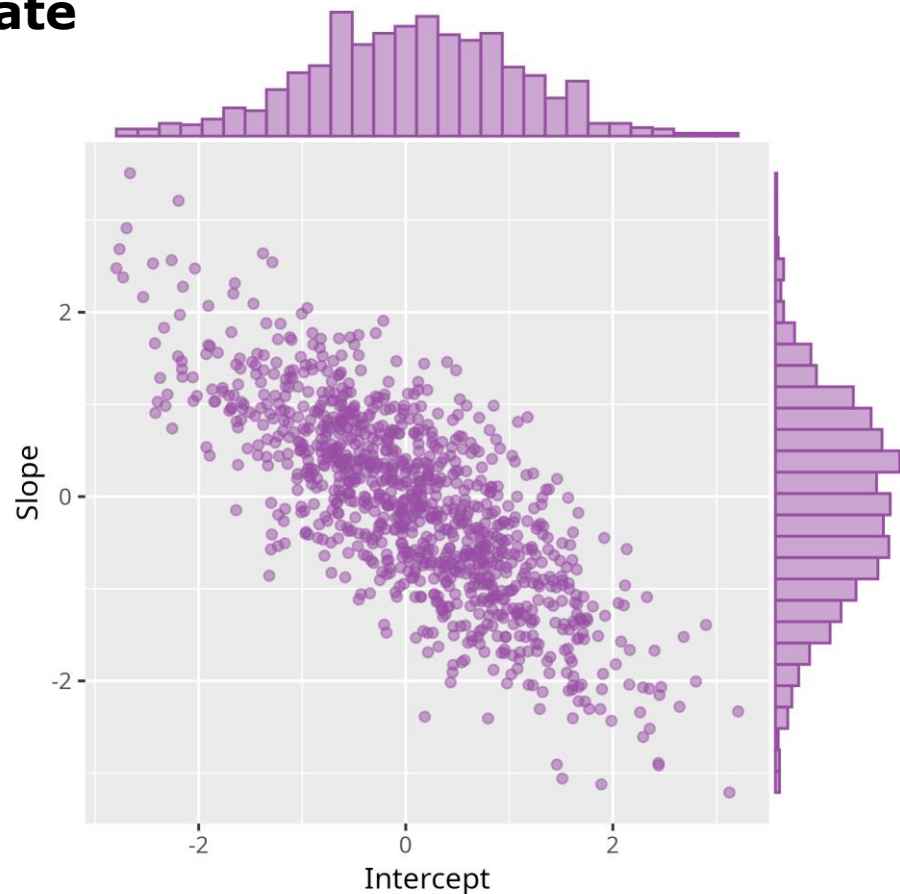
MCMC output is a matrix / dataframe !



## The posterior sample is multivariate

|       | Intercept | Slope  |
|-------|-----------|--------|
| [1,]  | -0.524    | 0.638  |
| [2,]  | -0.703    | 0.473  |
| [3,]  | -0.200    | 0.942  |
| [4,]  | 0.610     | -0.357 |
| [5,]  | 0.928     | -0.581 |
| [6,]  | 0.807     | -0.941 |
| [7,]  | 1.905     | -1.924 |
| [8,]  | -0.136    | -0.079 |
| [9,]  | -0.799    | 0.403  |
| [10,] | 0.098     | -1.532 |

Each **column** contains all samples of 1 parameter

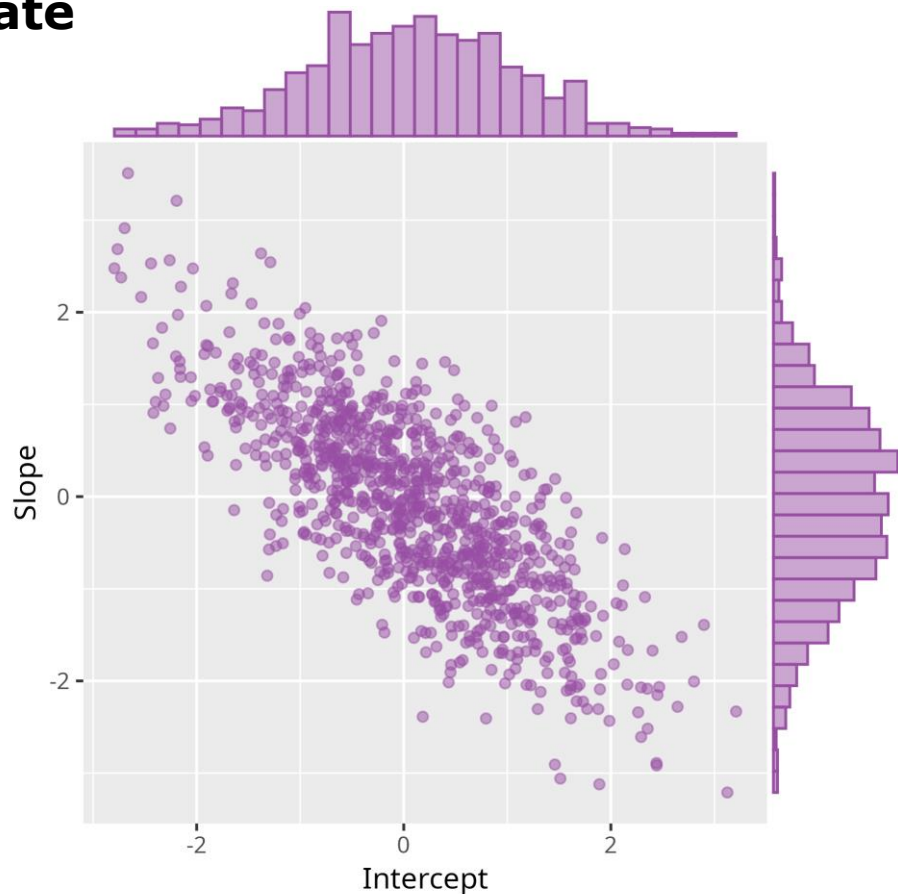




## The posterior sample is multivariate

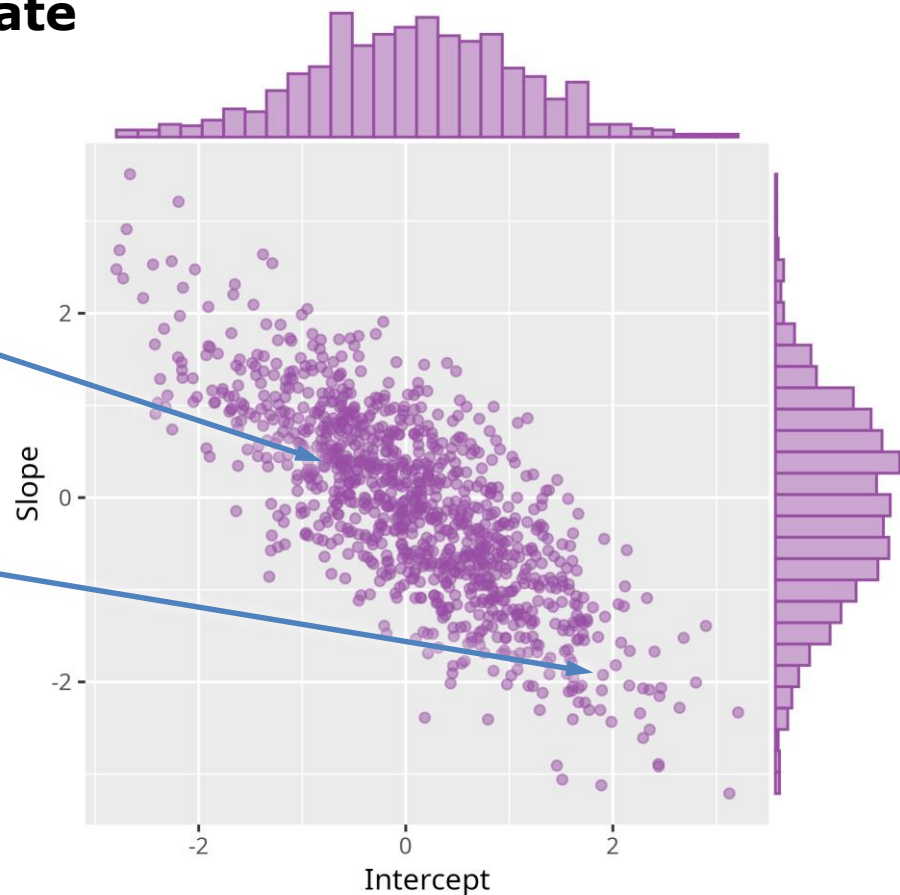
|       | Intercept | Slope  |
|-------|-----------|--------|
| [1,]  | -0.524    | 0.638  |
| [2,]  | -0.703    | 0.473  |
| [3,]  | -0.200    | 0.942  |
| [4,]  | 0.610     | -0.357 |
| [5,]  | 0.928     | -0.581 |
| [6,]  | 0.807     | -0.941 |
| [7,]  | 1.905     | -1.924 |
| [8,]  | -0.136    | -0.079 |
| [9,]  | -0.799    | 0.403  |
| [10,] | 0.098     | -1.532 |

Each **row** contains 1 sample of all parameters



## The posterior sample is multivariate

|       | Intercept | Slope  |
|-------|-----------|--------|
| [1,]  | -0.524    | 0.638  |
| [2,]  | -0.703    | 0.473  |
| [3,]  | -0.200    | 0.942  |
| [4,]  | 0.610     | -0.357 |
| [5,]  | 0.928     | -0.581 |
| [6,]  | 0.807     | -0.941 |
| [7,]  | 1.905     | -1.924 |
| [8,]  | -0.136    | -0.079 |
| [9,]  | -0.799    | 0.403  |
| [10,] | 0.098     | -1.532 |

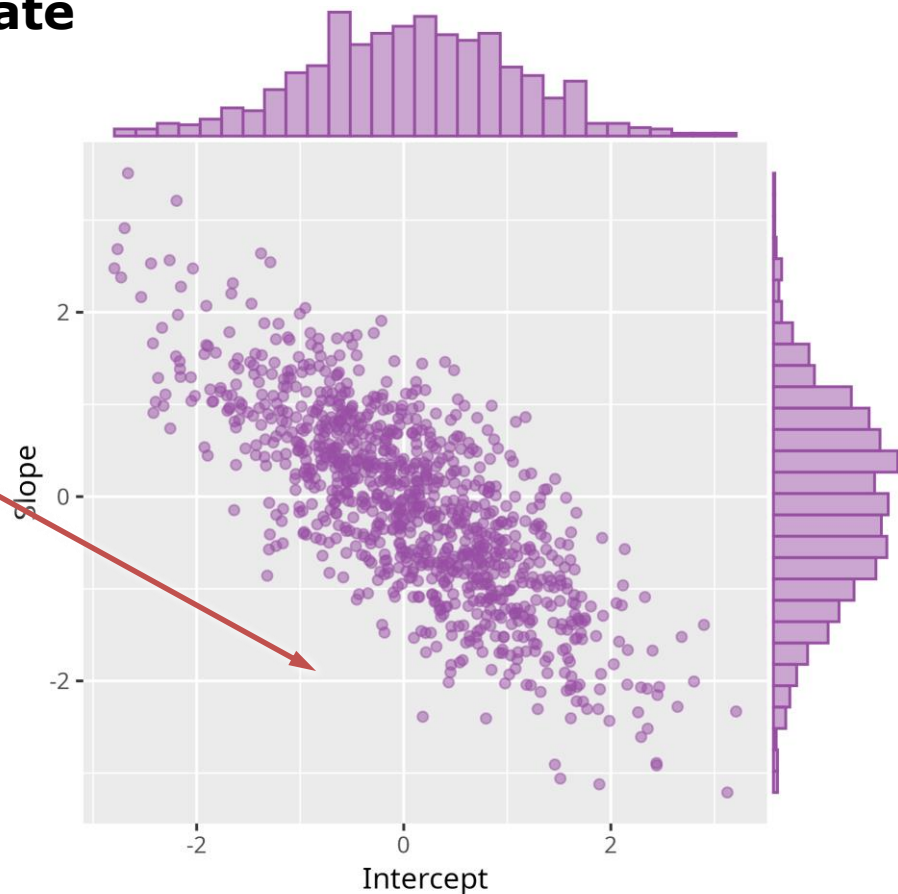


A random row is part of the posterior sample

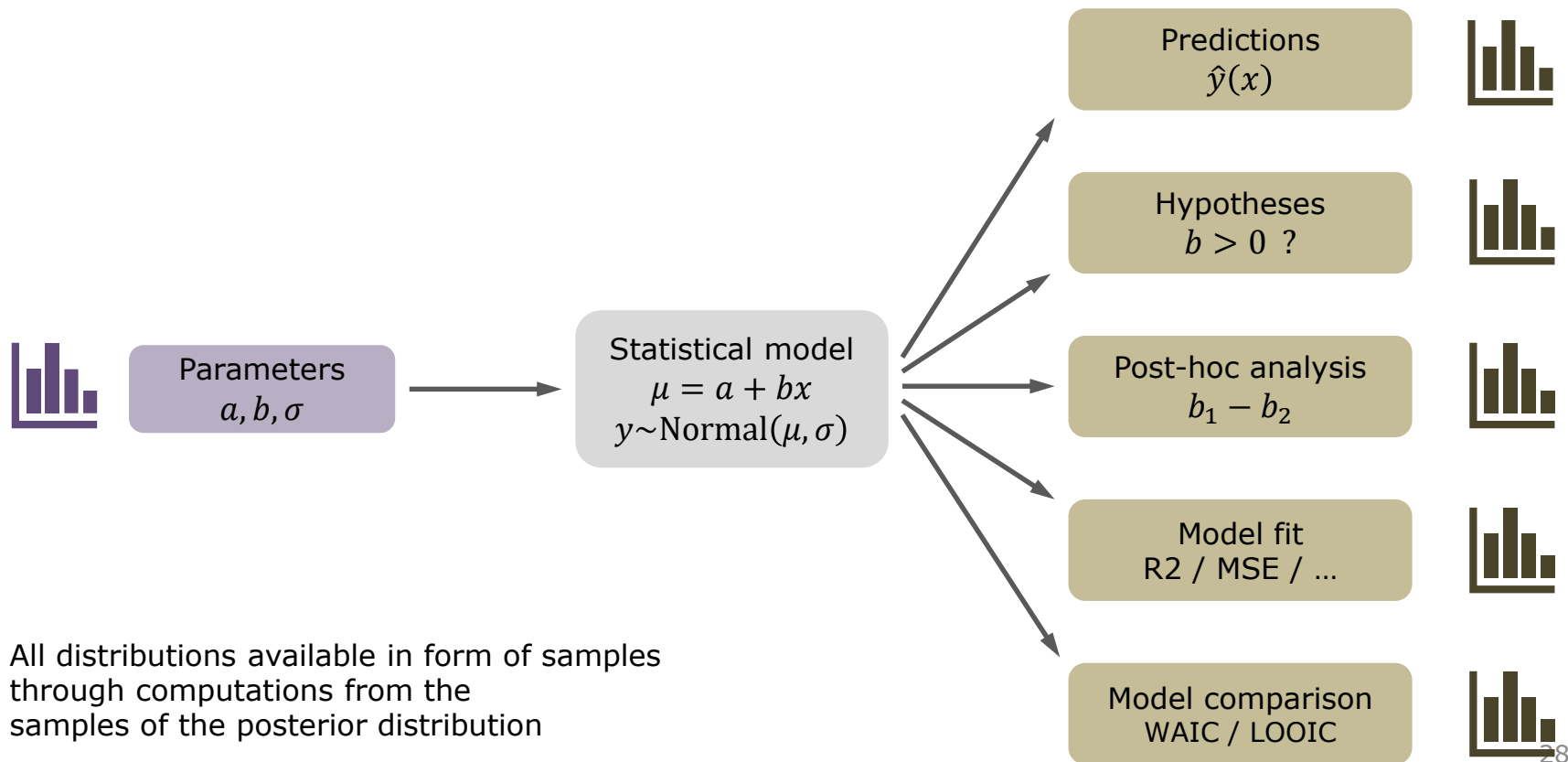
# The posterior sample is multivariate

|       | Intercept | Slope  |
|-------|-----------|--------|
| [1,]  | -0.524    | 0.638  |
| [2,]  | -0.703    | 0.473  |
| [3,]  | -0.200    | 0.942  |
| [4,]  | 0.610     | -0.357 |
| [5,]  | 0.928     | -0.581 |
| [6,]  | 0.807     | -0.941 |
| [7,]  | 1.905     | -1.924 |
| [8,]  | -0.136    | -0.079 |
| [9,]  | -0.799    | 0.403  |
| [10,] | 0.098     | -1.532 |

Random entries of each column (mixed) are **not** part of the posterior sample



# Everything is a distribution !



*Posterior predictions*

## Example:

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

Deterministic part:

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

Stochastic part:

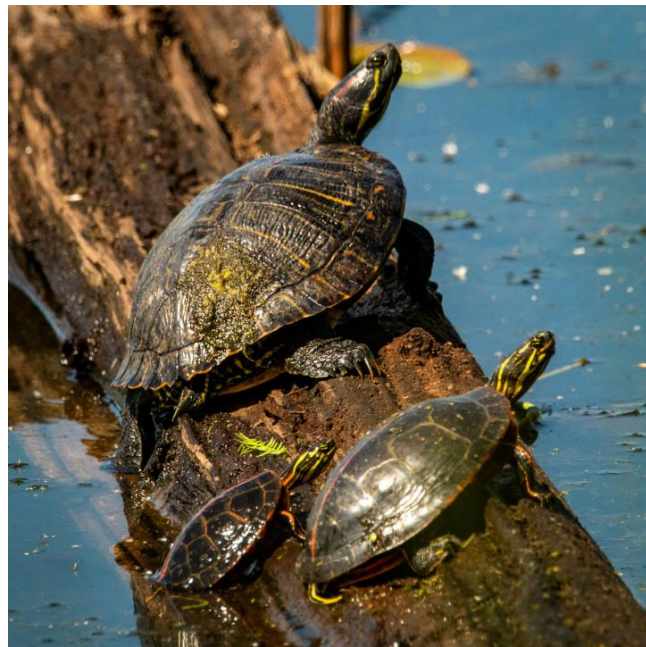
$$y \sim \text{Normal}(\mu, \sigma)$$

Parameters:

$a$  intercept

$b$  slope

$\sigma$  standard deviation



# Model output

```
> fit1 = brm(weight ~ age, data=data)
> summary(fit1)
```

Family: gaussian  
Links: mu = identity; sigma = identity  
Formula: weight ~ age  
Data: data (Number of observations: 8)  
Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;  
total post-warmup draws = 4000

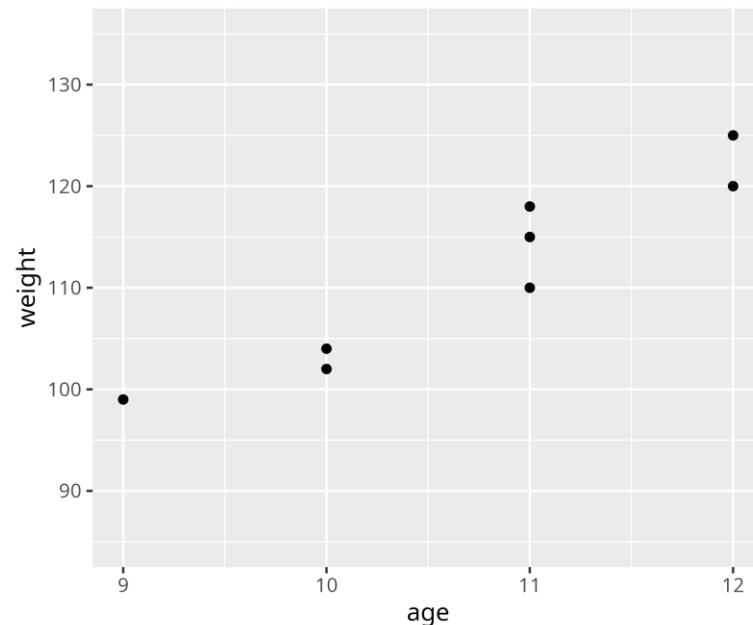
## Regression Coefficients:

|           | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|-----------|----------|-----------|----------|----------|------|----------|----------|
| Intercept | 19.99    | 16.24     | -13.79   | 53.23    | 1.00 | 2848     | 1916     |
| age       | 8.53     | 1.50      | 5.46     | 11.59    | 1.00 | 2817     | 1838     |

## Further Distributional Parameters:

|       | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|-------|----------|-----------|----------|----------|------|----------|----------|
| sigma | 4.01     | 1.43      | 2.19     | 7.73     | 1.00 | 1606     | 1766     |

Draws were sampled using sampling(NUTS). For each parameter, Bulk\_ESS and Tail\_ESS are effective sample size measures, and Rhat is the potential scale reduction factor on split chains (at convergence, Rhat = 1).



mean and sdev of parameters'  
posterior distribution

# How to predict ?

What is the predicted weight at age  $x = 10.5$  ?

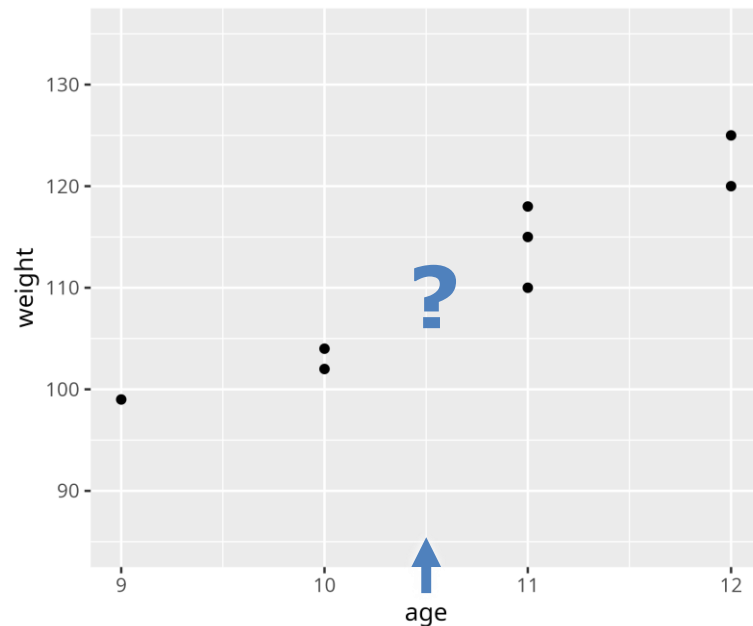
Deterministic part:  $\mu(x) = a + b \cdot x$

We have mean values for

intercept  $a = 19.99$  and slope  $b = 8.53$

However, in Bayesian statistics, we **don't** use mean parameter values to make prediction.

We use the **whole posterior distribution** to quantify prediction uncertainty correctly !

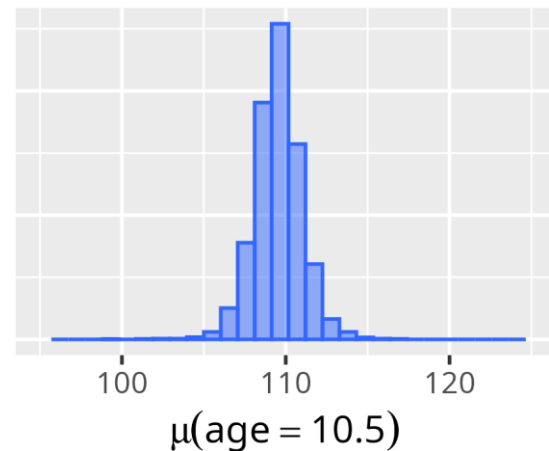




## The fitted distribution

| draw | b_Intercept | b_age  | fitted(age=10.5) |
|------|-------------|--------|------------------|
| 1    | 7.473       | 9.738  | 109.720          |
| 2    | 30.477      | 7.614  | 110.425          |
| 3    | 22.153      | 8.273  | 109.022          |
| 4    | 34.008      | 7.310  | 110.763          |
| 5    | 39.144      | 6.667  | 109.147          |
| 6    | 12.472      | 9.328  | 110.415          |
| 7    | 7.005       | 9.674  | 108.579          |
| 8    | -3.196      | 10.633 | 108.446          |
| 9    | 23.362      | 8.309  | 110.602          |
| 10   | 23.745      | 8.247  | 110.342          |

Deterministic part  
 $\mu = a + b \cdot x$

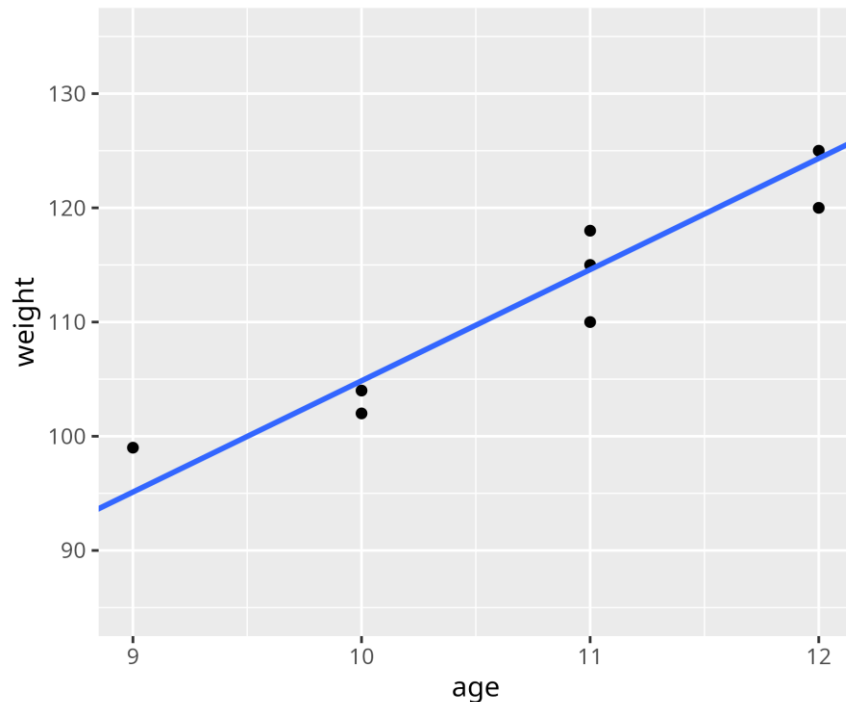


Each sample from the posterior  $(a_i, b_i)$  generates 1 sample for,  $\mu_i(\text{age} = 10.5) = a_i + b_i \cdot 10.5$

```
> posterior_epred(fit1, newdata=data.frame(age=10.5))
```

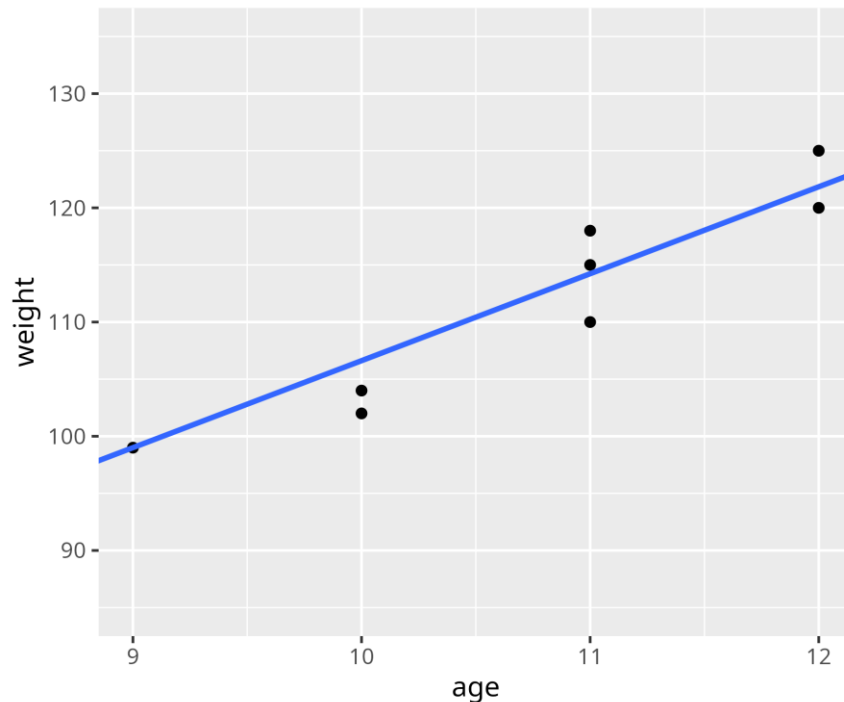
## Each posterior sample generates a regression line

| draw | b_Intercept | b_age  | sigma |
|------|-------------|--------|-------|
| 1    | 7.473       | 9.738  | 4.128 |
| 2    | 30.477      | 7.614  | 3.456 |
| 3    | 22.153      | 8.273  | 3.474 |
| 4    | 34.008      | 7.310  | 4.200 |
| 5    | 39.144      | 6.667  | 4.456 |
| 6    | 12.472      | 9.328  | 3.671 |
| 7    | 7.005       | 9.674  | 4.045 |
| 8    | -3.196      | 10.633 | 4.117 |
| 9    | 23.362      | 8.309  | 3.519 |
| 10   | 23.745      | 8.247  | 3.336 |
| 11   | 27.194      | 7.948  | 4.249 |
| 12   | 18.373      | 8.566  | 3.351 |
| 13   | 20.358      | 8.522  | 3.447 |
| 14   | 10.307      | 9.322  | 3.007 |
| 15   | 24.486      | 8.224  | 3.408 |



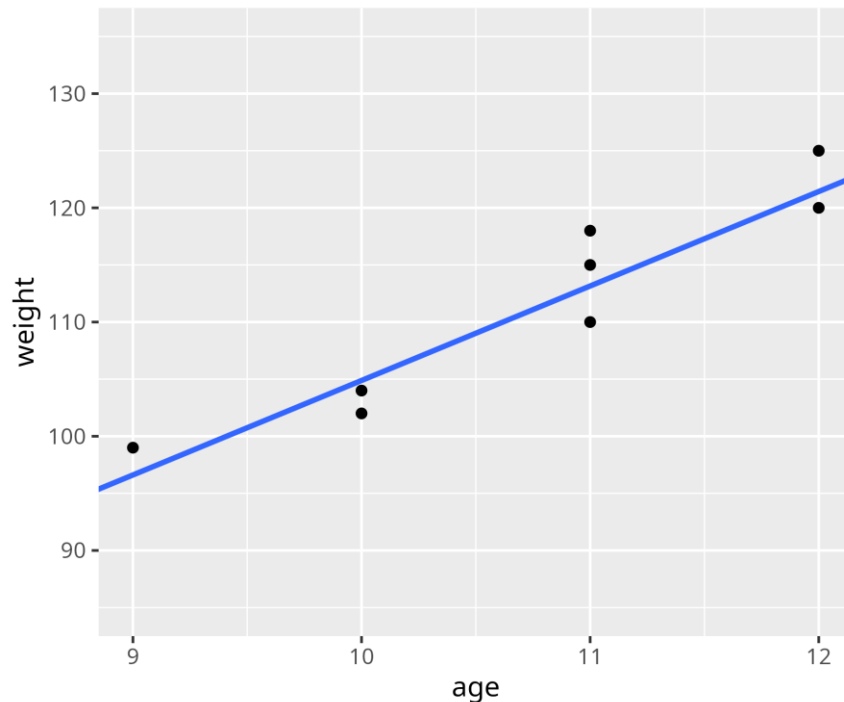
## Each posterior sample generates a regression line

| draw | b_Intercept | b_age  | sigma |
|------|-------------|--------|-------|
| 1    | 7.473       | 9.738  | 4.128 |
| 2    | 30.477      | 7.614  | 3.456 |
| 3    | 22.153      | 8.273  | 3.474 |
| 4    | 34.008      | 7.310  | 4.200 |
| 5    | 39.144      | 6.667  | 4.456 |
| 6    | 12.472      | 9.328  | 3.671 |
| 7    | 7.005       | 9.674  | 4.045 |
| 8    | -3.196      | 10.633 | 4.117 |
| 9    | 23.362      | 8.309  | 3.519 |
| 10   | 23.745      | 8.247  | 3.336 |
| 11   | 27.194      | 7.948  | 4.249 |
| 12   | 18.373      | 8.566  | 3.351 |
| 13   | 20.358      | 8.522  | 3.447 |
| 14   | 10.307      | 9.322  | 3.007 |
| 15   | 24.486      | 8.224  | 3.408 |



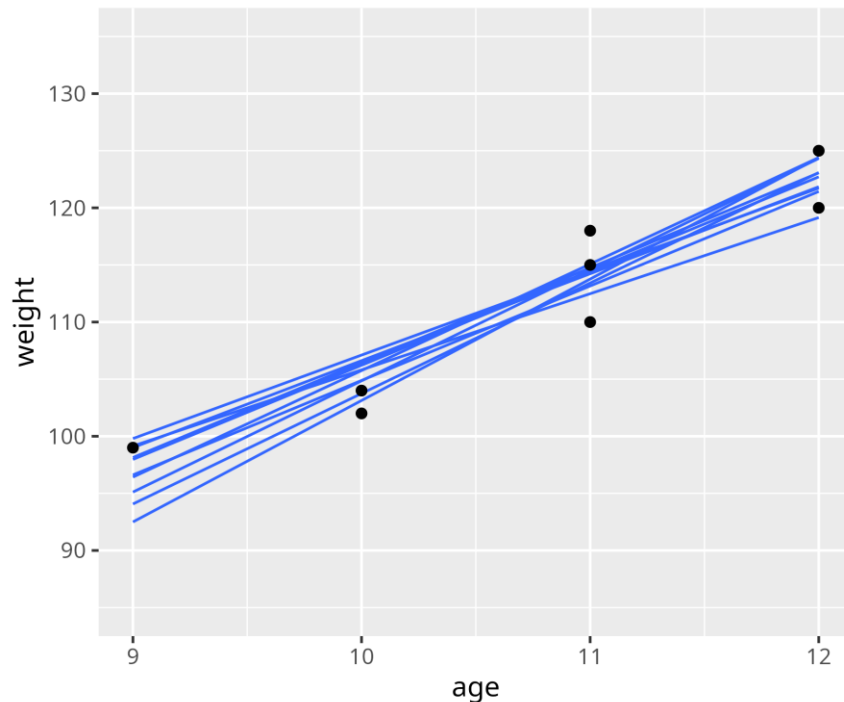
## Each posterior sample generates a regression line

| draw | b_Intercept | b_age  | sigma |
|------|-------------|--------|-------|
| 1    | 7.473       | 9.738  | 4.128 |
| 2    | 30.477      | 7.614  | 3.456 |
| 3    | 22.153      | 8.273  | 3.474 |
| 4    | 34.008      | 7.310  | 4.200 |
| 5    | 39.144      | 6.667  | 4.456 |
| 6    | 12.472      | 9.328  | 3.671 |
| 7    | 7.005       | 9.674  | 4.045 |
| 8    | -3.196      | 10.633 | 4.117 |
| 9    | 23.362      | 8.309  | 3.519 |
| 10   | 23.745      | 8.247  | 3.336 |
| 11   | 27.194      | 7.948  | 4.249 |
| 12   | 18.373      | 8.566  | 3.351 |
| 13   | 20.358      | 8.522  | 3.447 |
| 14   | 10.307      | 9.322  | 3.007 |
| 15   | 24.486      | 8.224  | 3.408 |



## Each posterior sample generates a regression line

| draw | b_Intercept | b_age  | sigma |
|------|-------------|--------|-------|
| 1    | 7.473       | 9.738  | 4.128 |
| 2    | 30.477      | 7.614  | 3.456 |
| 3    | 22.153      | 8.273  | 3.474 |
| 4    | 34.008      | 7.310  | 4.200 |
| 5    | 39.144      | 6.667  | 4.456 |
| 6    | 12.472      | 9.328  | 3.671 |
| 7    | 7.005       | 9.674  | 4.045 |
| 8    | -3.196      | 10.633 | 4.117 |
| 9    | 23.362      | 8.309  | 3.519 |
| 10   | 23.745      | 8.247  | 3.336 |
| 11   | 27.194      | 7.948  | 4.249 |
| 12   | 18.373      | 8.566  | 3.351 |
| 13   | 20.358      | 8.522  | 3.447 |
| 14   | 10.307      | 9.322  | 3.007 |
| 15   | 24.486      | 8.224  | 3.408 |



# Credible intervals

Distribution of fitted values / regression lines  
with **deterministic model part only**

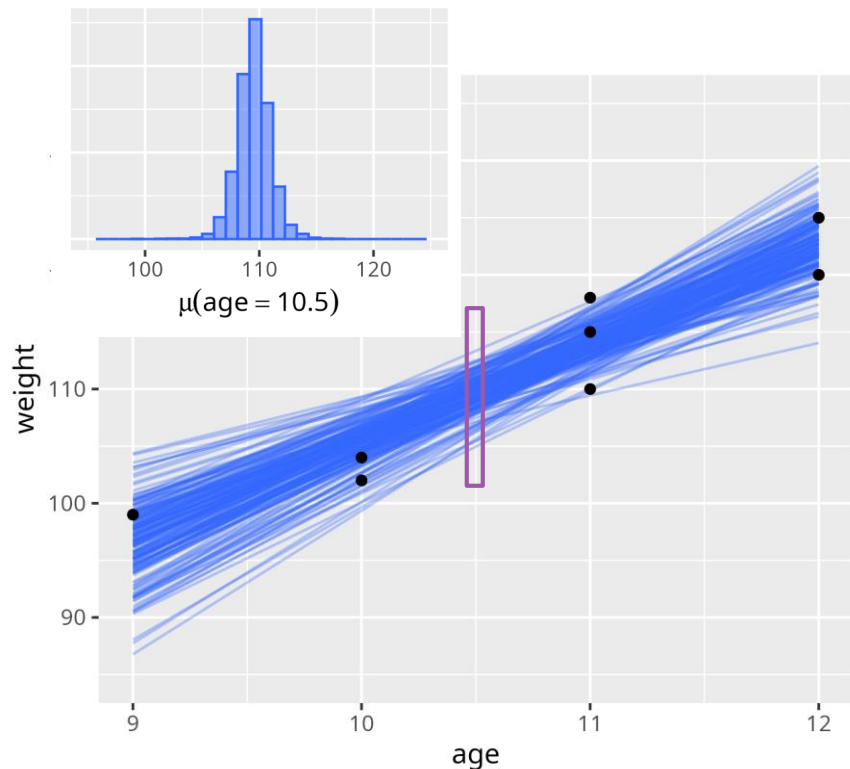
$$\mu_i(x) = a_i + b_i \cdot x \quad (i = 1, \dots, 1000)$$

Mean fitted value

$$\bar{\mu}(x) = \text{mean}(\mu_1(x), \dots, \mu_{1000}(x))$$

95% intervals are called **credible intervals**.  
They quantify uncertainty of the regression line.

There is nothing magical about 95%,  
can also choose other intervals, e.g. 90%



```
> plot(conditional_effects(fit1, spaghetti=TRUE, ndraws=200))
```

# Credible intervals

Distribution of fitted values / regression lines  
with **deterministic model part only**

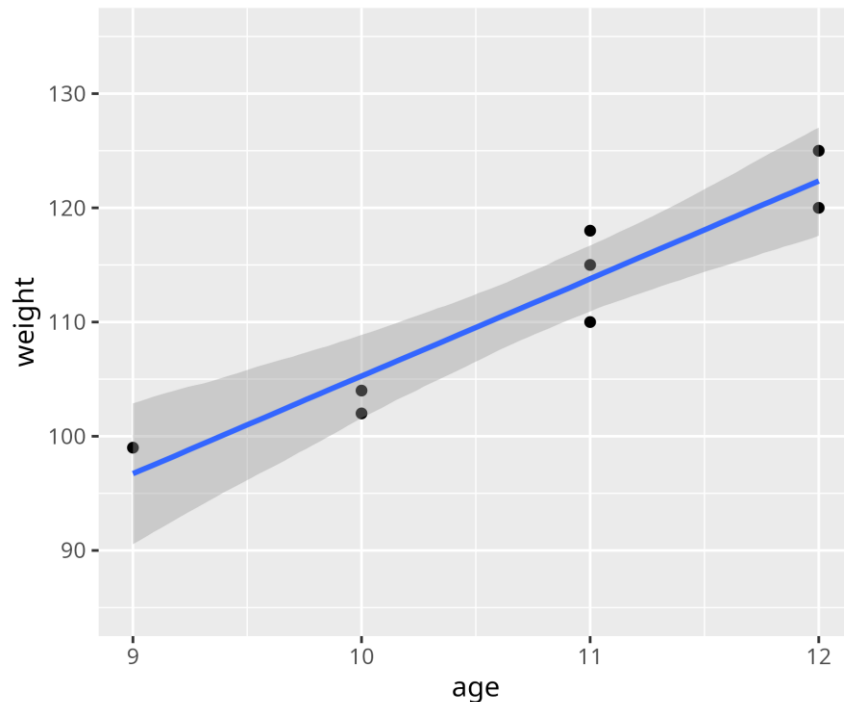
$$\mu_i(x) = a_i + b_i \cdot x \quad (i = 1, \dots, 1000)$$

Mean fitted value

$$\bar{\mu}(x) = \text{mean}(\mu_1(x), \dots, \mu_{1000}(x))$$

95% intervals are called **credible intervals**.  
They quantify uncertainty of the regression line.

There is nothing magical about 95%,  
can also choose other intervals, e.g. 90%

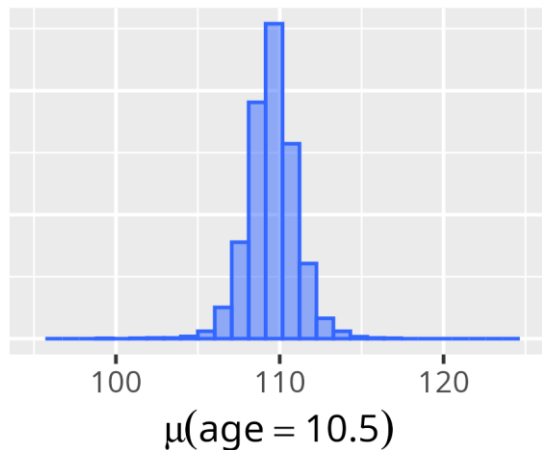


```
> plot(conditional_effects(fit1), points=TRUE)
```

# The predictive distribution

| b_Intercept | b_age  | fitted(age=10.5) |
|-------------|--------|------------------|
| 7.473       | 9.738  | 109.720          |
| 30.477      | 7.614  | 110.425          |
| 22.153      | 8.273  | 109.022          |
| 34.008      | 7.310  | 110.763          |
| 39.144      | 6.667  | 109.147          |
| 12.472      | 9.328  | 110.415          |
| 7.005       | 9.674  | 108.579          |
| -3.196      | 10.633 |                  |
| 23.362      | 8.309  |                  |
| 23.745      | 8.247  |                  |

Deterministic part  
 $\mu = a + b \cdot x$



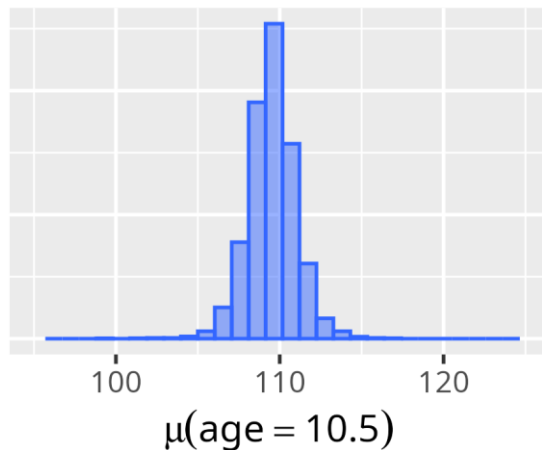
```
> posterior_epred(fit1, newdata=data.frame(age=10.5))
```



# The predictive distribution

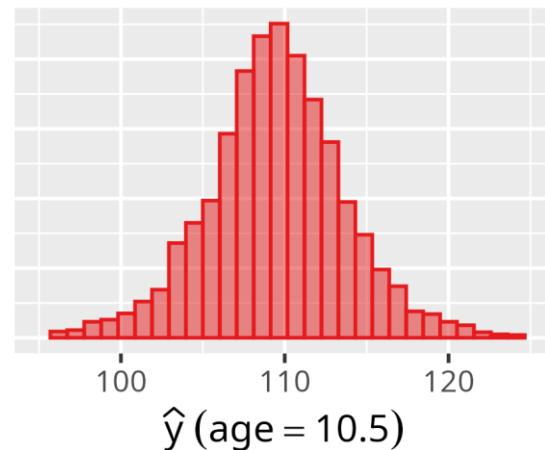
| b_Intercept | b_age  | fitted(age=10.5) | sigma | predict(age=10.5) |
|-------------|--------|------------------|-------|-------------------|
| 7.473       | 9.738  | 109.720          | 4.128 | 119.538           |
| 30.477      | 7.614  | 110.425          | 3.456 | 105.864           |
| 22.153      | 8.273  | 109.022          | 3.474 | 110.969           |
| 34.008      | 7.310  | 110.763          | 4.200 | 111.090           |
| 39.144      | 6.667  | 109.147          | 4.456 | 110.522           |
| 12.472      | 9.328  | 110.415          | 3.671 | 109.970           |
| 7.005       | 9.674  | 108.579          | 4.045 | 115.610           |
| -3.196      | 10.633 |                  | 4.117 |                   |
| 23.362      | 8.309  |                  | 3.519 |                   |
| 23.745      | 8.247  |                  | 3.336 |                   |

Deterministic part  
 $\mu = a + b \cdot x$



> posterior\_epred(fit1, newdata=data.frame(age=10.5))

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



> posterior\_predict(fit1, newdata=data.frame(age=10.5))

# Prediction intervals

Predictions add random residual error to fitted values

Distribution of predicted values

with **deterministic and stochastic model part**

$$\hat{y}_i(x) = \mu_i(x) + \varepsilon_i \quad (i = 1, \dots, 1000)$$

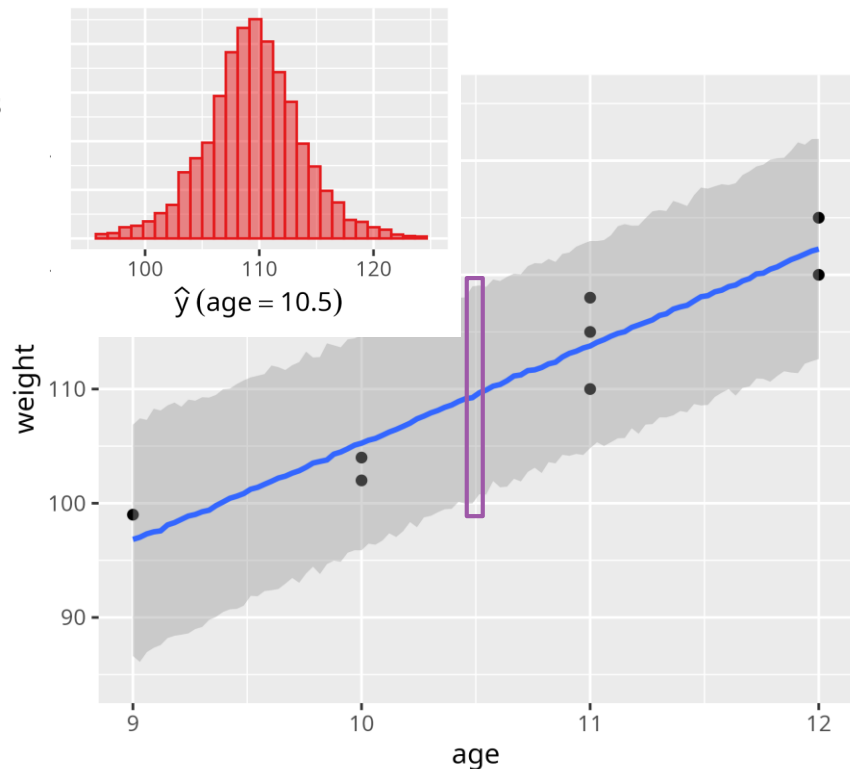
$$\varepsilon_i \sim \text{Normal}(0, \sigma_i)$$

Same as:  $\hat{y}_i(x) \sim \text{Normal}(\mu_i(x), \sigma_i)$

95% intervals are called **prediction intervals**.

They quantify uncertainty of newly predicted data.

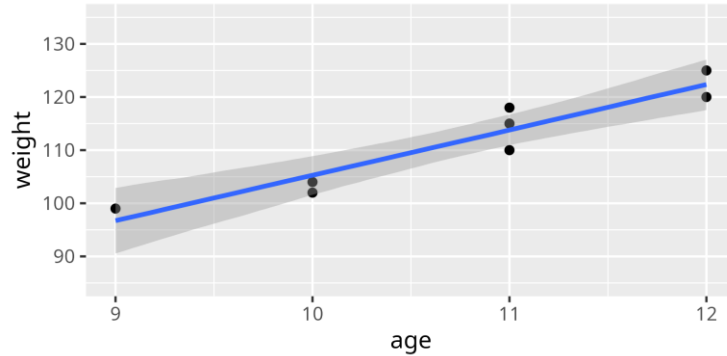
(Should contain around 95% of observed data.)



```
> plot(conditional_effects(fit1, method="posterior_predict"),  
       points=TRUE)
```

# Fitted vs. predictive

**Fitted**

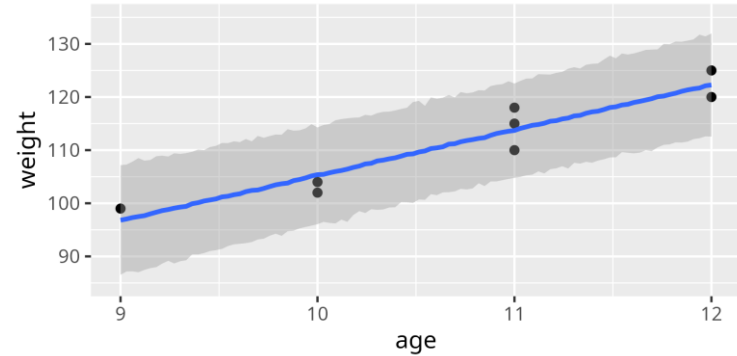


Mean regression line / curve under parameter uncertainty

„Credible intervals“

Uses deterministic model part only

**Predicted**



Predictive data distribution under parameter uncertainty and model residuals

„Prediction intervals“

Uses deterministic and stochastic model parts

# Fitted vs. predictive

## Fitted

```
> fitted(fit1)
```

|      | Estimate  | Est.Error | Q2.5      | Q97.5    |
|------|-----------|-----------|-----------|----------|
| [1,] | 105.25324 | 1.813620  | 101.59634 | 108.8769 |
| [2,] | 122.30603 | 2.343272  | 117.55827 | 127.0220 |
| [3,] | 113.77964 | 1.458628  | 110.92357 | 116.6928 |
| [4,] | 113.77964 | 1.458628  | 110.92357 | 116.6928 |
| [5,] | 96.72685  | 2.995963  | 90.55468  | 102.8774 |
| [6,] | 113.77964 | 1.458628  | 110.92357 | 116.6928 |
| [7,] | 105.25324 | 1.813620  | 101.59634 | 108.8769 |
| [8,] | 122.30603 | 2.343272  | 117.55827 | 127.0220 |

Mean regression line / curve under parameter uncertainty

„Credible intervals“

Uses deterministic model part only

## Predicted

```
> predict(fit1)
```

|      | Estimate  | Est.Error | Q2.5      | Q97.5    |
|------|-----------|-----------|-----------|----------|
| [1,] | 105.22283 | 4.549012  | 95.94294  | 114.3628 |
| [2,] | 122.18553 | 4.887509  | 112.25245 | 131.7162 |
| [3,] | 113.76152 | 4.490723  | 104.56416 | 122.8968 |
| [4,] | 113.70727 | 4.512039  | 104.60886 | 122.9418 |
| [5,] | 96.82354  | 5.324598  | 86.61397  | 107.5571 |
| [6,] | 113.75514 | 4.443002  | 104.59945 | 122.6224 |
| [7,] | 105.24951 | 4.562843  | 95.97297  | 114.4820 |
| [8,] | 122.42399 | 4.799248  | 112.77401 | 132.2151 |

Predictive data distribution under parameter uncertainty and model residuals

„Prediction intervals“

Uses deterministic and stochastic model parts

## *Posterior predictive checks*

# Linear regression assumptions

1. Independent observations.

Systematic differences in  $y$  are because of  $x$  !

2. Trend of  $y$  follows (linear) prediction model

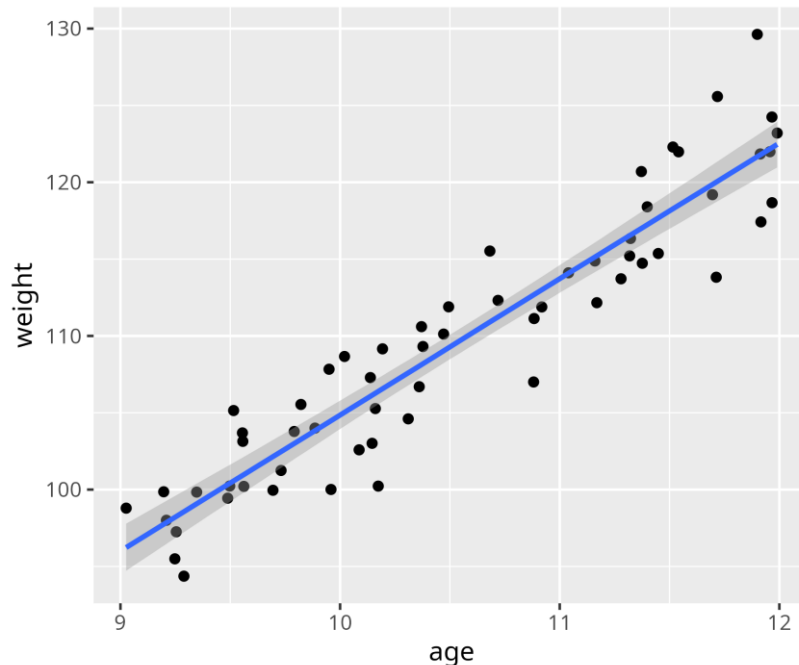
$$\mu(x) = a + b \cdot x$$

3. Residuals follow normal distribution

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

4. Constant variance (standard deviation)

across whole range of  $x$



# Model checking

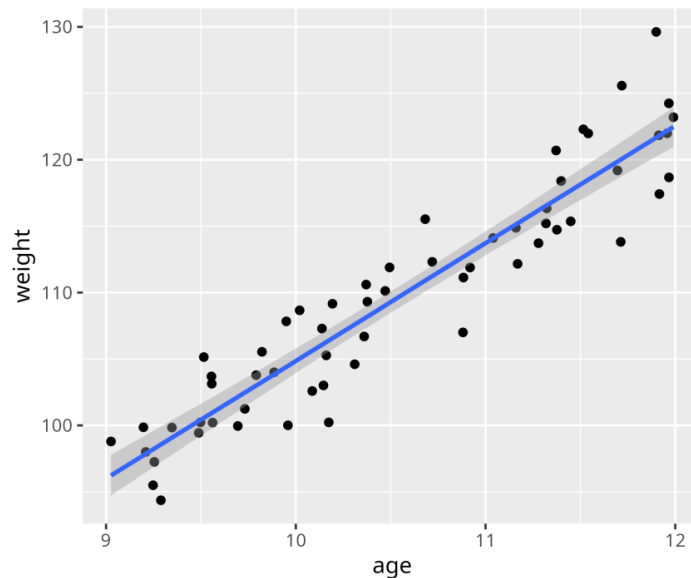
Visualization is easy when you have just one predictor!

Need alternative visual tools when dealing with multiple predictors.

Response / prediction is just 1 variable

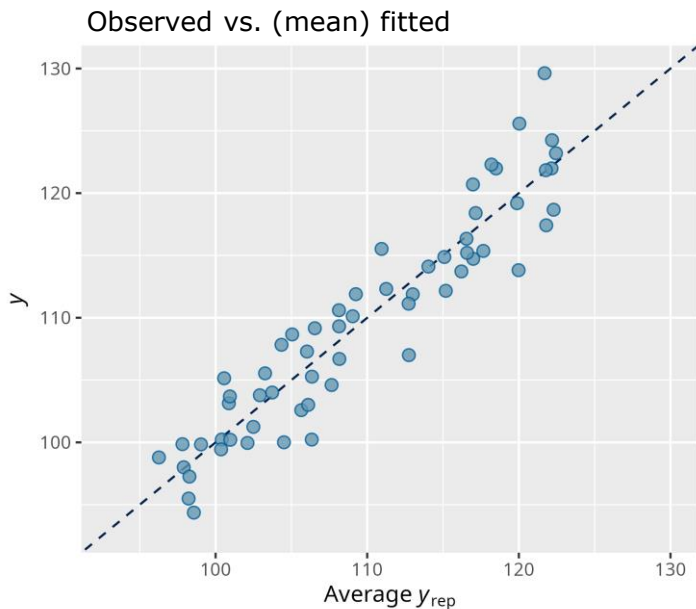
→ Compare and plot against each other:

- observations
- (mean) predictions
- residuals (observed - predicted)

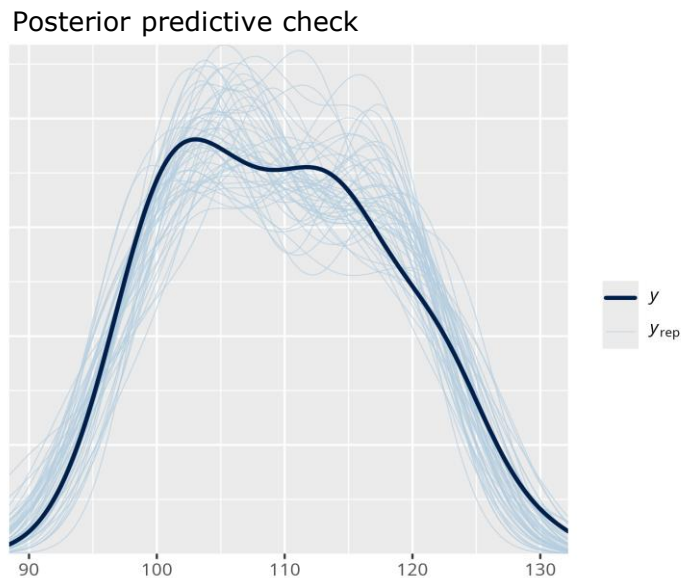


```
> plot(conditional_effects(fit1), points=TRUE)
```

# Model checking (from brms package)



```
> pp_check(fit1, type="scatter_avg")
```



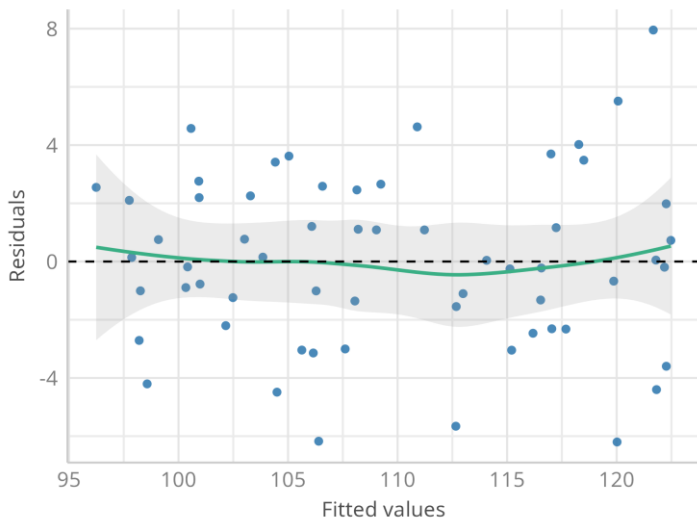
```
> pp_check(fit1, ndraws=50)
```



# Model checking (from performance package)

## Linearity

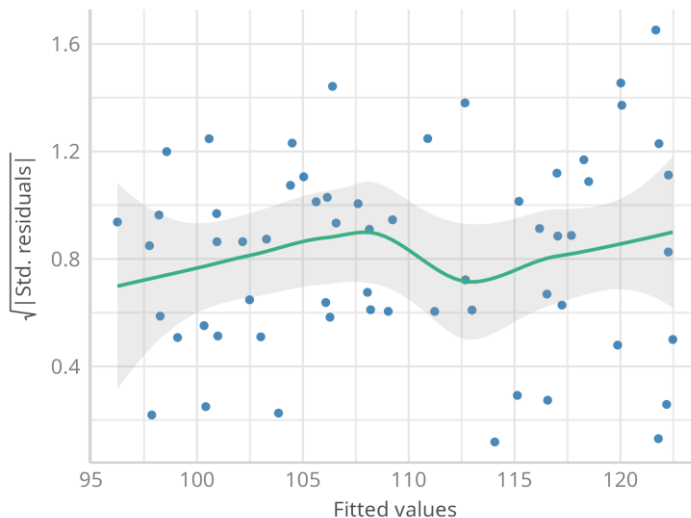
Reference line should be flat and horizontal



```
> check_model(fit1, check=„linearity“)
```

## Homogeneity of Variance

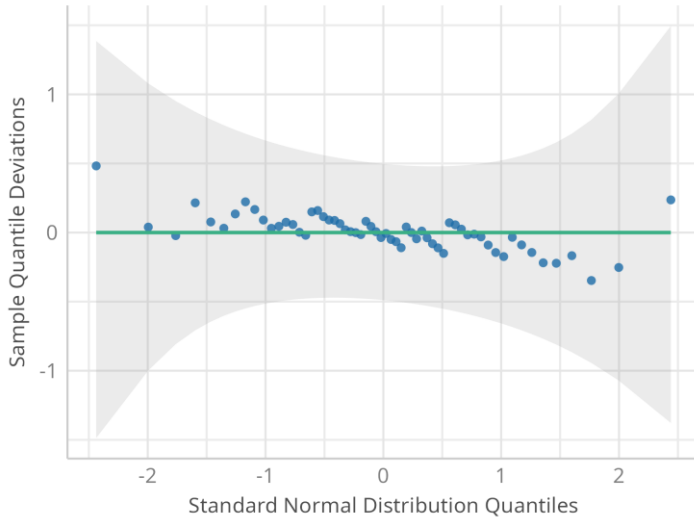
Reference line should be flat and horizontal



```
> check_model(fit1, check=„homogeneity“)
```

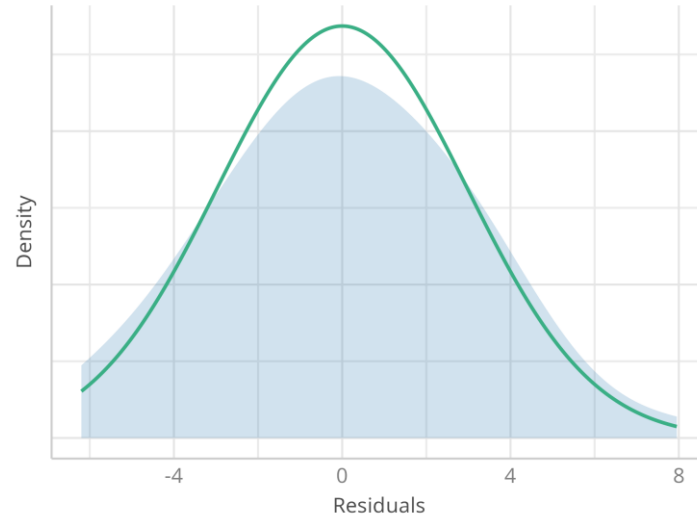
# Model checking (from performance package)

Normality of Residuals  
Dots should fall along the line



```
> check_model(fit1, check=„qq“)
```

Normality of Residuals  
Distribution should be close to the normal curve



```
> check_model(fit1, check=„normality“)
```

# Pitfalls of prediction: Multivariate posterior

If your posterior (parameters  $a, b$ ) was shaped like Croatia, (nonlinear correlation), then the mean  $(\bar{a}, \bar{b})$  in 2d-space would not be part of the posterior sample

Parameter combination  $(\bar{a}, \bar{b})$  is **highly unlikely**

Prediction  $\mu(\bar{a}, \bar{b})$  is not the mean prediction, but rather meaningless!

→ **Always use full posterior for making predictions!**



Due to its shape, the centre of Croatia is actually located in Bosnia and Herzegovina

[Post übersetzen](#)



# Pitfalls of prediction: Jensen's inequality

For linear models  $\mu$  (as in linear regression  $\mu = a + bx$ )

$$\text{mean}(\mu(\theta_1), \dots, \mu(\theta_{1000})) = \mu(\text{mean}(\theta_1, \dots, \theta_{1000}))$$

$\underbrace{\hspace{10em}}$

Distribution of  
predictions

$\underbrace{\hspace{10em}}$

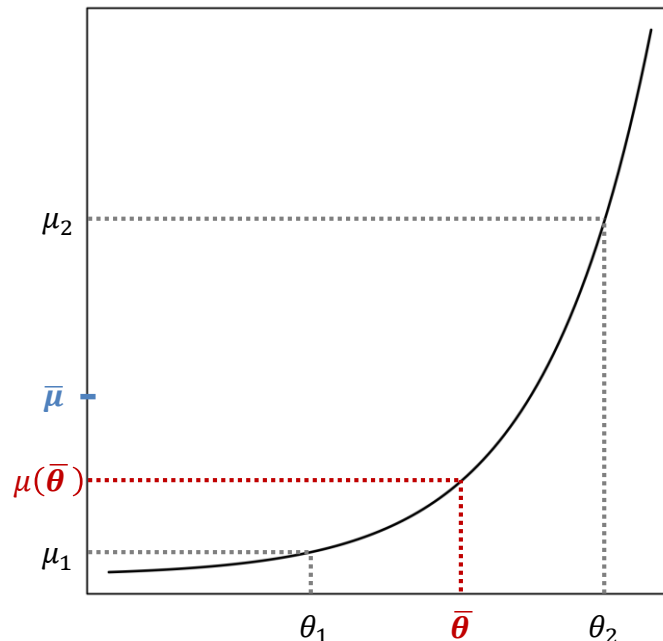
Prediction with  
point estimate  $\bar{\theta}$

For **nonlinear functions**  $\mu$

$$\text{mean}(\mu(\theta_1), \dots, \mu(\theta_{1000})) \neq \mu(\text{mean}(\theta_1, \dots, \theta_{1000}))$$

Relevant for nonlinear statistical models, GLMs,  
link functions, parameter or data transformations

→ **Always use full posterior for making predictions!**



# *Summary*

# Summary

- Priors → you choose !
  - Likelihood → given by data & statistical model
  - MCMC samples from posterior → **check convergence** !
  - Informative priors can decrease uncertainty in posterior
  - More datapoints can decrease uncertainty in posterior
  - Use posterior predictions to **check model assumptions and model fit**
  - In Bayesian statistics, everything is a distribution
- Use full posterior (samples) for everything

## Further reading

Banner, K. M., Irvine, K. M., & Rodhouse, T. J. (2020). The use of Bayesian priors in Ecology: The good, the bad and the not great. *Methods in Ecology and Evolution*, 11(8), 882–889. <https://doi.org/10.1111/2041-210X.13407>

Bürkner, P. (2024). The brms Book [in progress]. <https://paulbuerkner.com/software/brms-book/>

Conn, P. B., Johnson, D. S., Williams, P. J., Melin, S. R., & Hooten, M. B. (2018). A guide to Bayesian model checking for ecologists. *Ecological Monographs*, 88(4), 526–542. <https://doi.org/10.1002/ecm.1314>

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, 182(2), 389–402. <https://doi.org/10.1111/rssa.12378>

Lemoine, N. P. (2019). Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, 128(7), 912–928. <https://doi.org/10.1111/oik.05985>

McElreath, R. (2020). Statistical Rethinking: A Bayesian Course with Examples in R and STAN (2nd ed.). *Chapman and Hall/CRC*. <https://doi.org/10.1201/9780429029608>

van de Schoot, R., Depaoli, S., King, R., et al. (2021). Bayesian statistics and modelling. *Nature Reviews. Methods Primers*, 1(1), 1–26. <https://doi.org/10.1038/s43586-020-00001-2>

Wesner, J. S., & Pomeranz, J. P. F. (2021). Choosing priors in Bayesian ecological models by simulating from the prior predictive distribution. *Ecosphere*, 12(9), e03739. <https://doi.org/10.1002/ecs2.3739>

<https://github.com/stan-dev/stan/wiki/prior-choice-recommendations>