



# *Statistical Estimation with Censored Data*

March 7 2008, ETH Zürich

Marloes Maathuis

# Outline

- Introduction to statistical estimation
  - Why do we study properties of estimators?
- Censored data
  - What are censored data, and where do they arise?
- Statistical estimation with censored data
  - Properties of estimators are non-standard
  - Some examples of what we know and don't know yet

## A basic problem in statistics

- Want to know something (a parameter) about a certain population

## A basic problem in statistics

- Want to know something (a parameter) about a certain population
- Example: What is the % of democratic voters in the next US presidential election?
  - Population: Voters in the next US presidential election
  - Parameter: % of democratic voters

## A basic problem in statistics

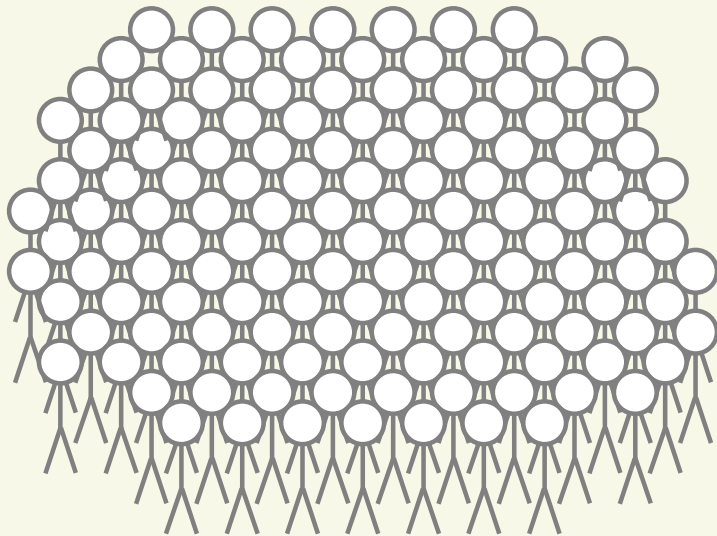
- Want to know something (a parameter) about a certain population
- Example: What is the % of democratic voters in the next US presidential election?
  - Population: Voters in the next US presidential election
  - Parameter: % of democratic voters
- Problem: It is infeasible to consider the entire population

# A basic problem in statistics

- Want to know something (a parameter) about a certain population
- Example: What is the % of democratic voters in the next US presidential election?
  - Population: Voters in the next US presidential election
  - Parameter: % of democratic voters
- Problem: It is infeasible to consider the entire population
- Solution:
  - Take a random sample
  - Use the sample to estimate what is going on in the population

# Population versus sample in election example

Population:

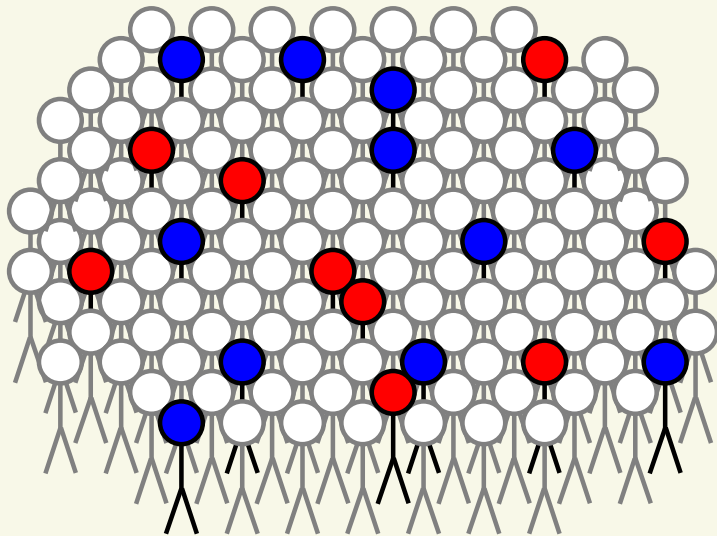


Parameter: population %

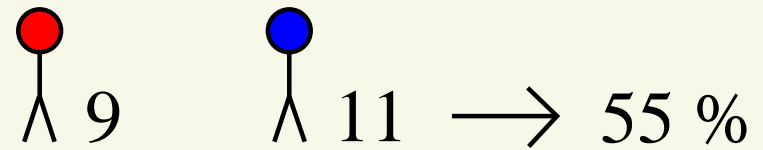
- fixed number
- unknown

# Population versus sample in election example

Population:



Sample:



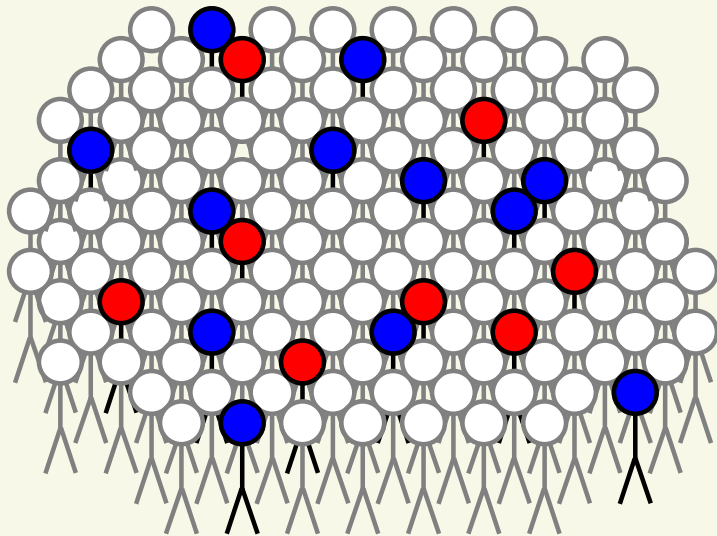
Parameter: population %

- fixed number
- unknown

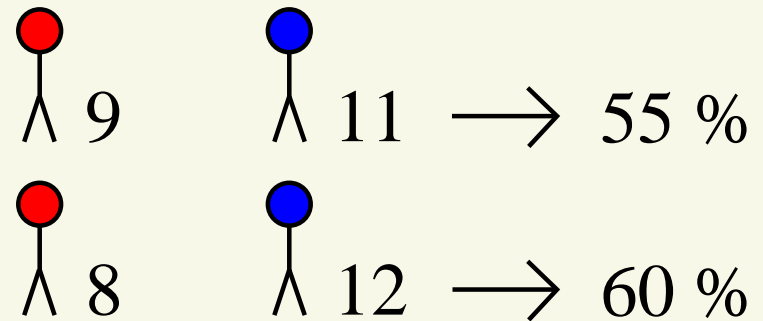
Estimate: sample %

# Population versus sample in election example

Population:



Sample:



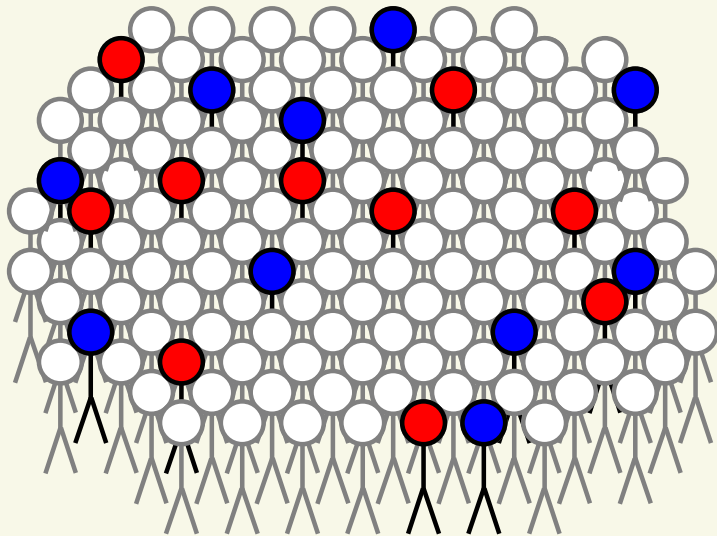
Parameter: population %

- fixed number
- unknown

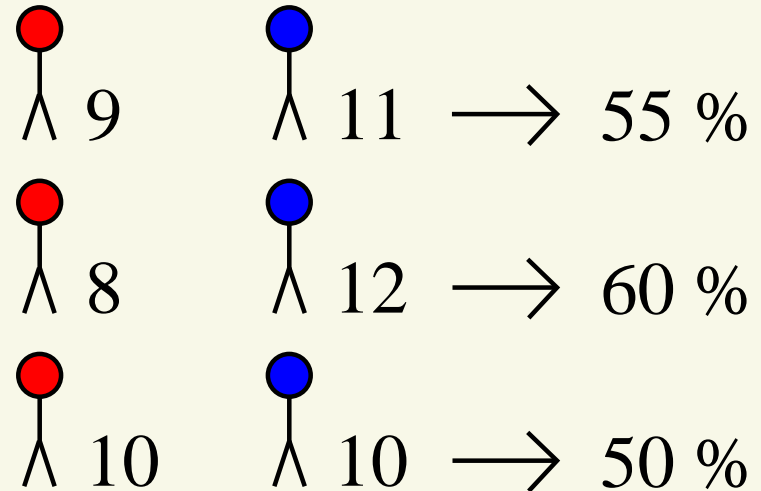
Estimate: sample %

# Population versus sample in election example

Population:



Sample:



Parameter: population %

- fixed number
- unknown

Estimate: sample %

- random
- known for any given sample

## Random error

Sample % = population % + random error

# Random error

Sample % = population % + random error



© Mike Baldwin / Cornered

# Random error

Sample % = population % + random error

- Example 1:
  - Sample percentage is 52%
  - Size of random error is of the order of 0.1%

# Random error

Sample % = population % + random error

- Example 1:
  - Sample percentage is 52%
  - Size of random error is of the order of 0.1%
  - We predict that the democrats will win

# Random error

Sample % = population % + random error

- Example 1:
  - Sample percentage is 52%
  - Size of random error is of the order of 0.1%
  - We predict that the democrats will win
- Example 2:
  - Sample percentage is 52%
  - Size of random error is of the order of 5%
  - We cannot predict a winner with confidence

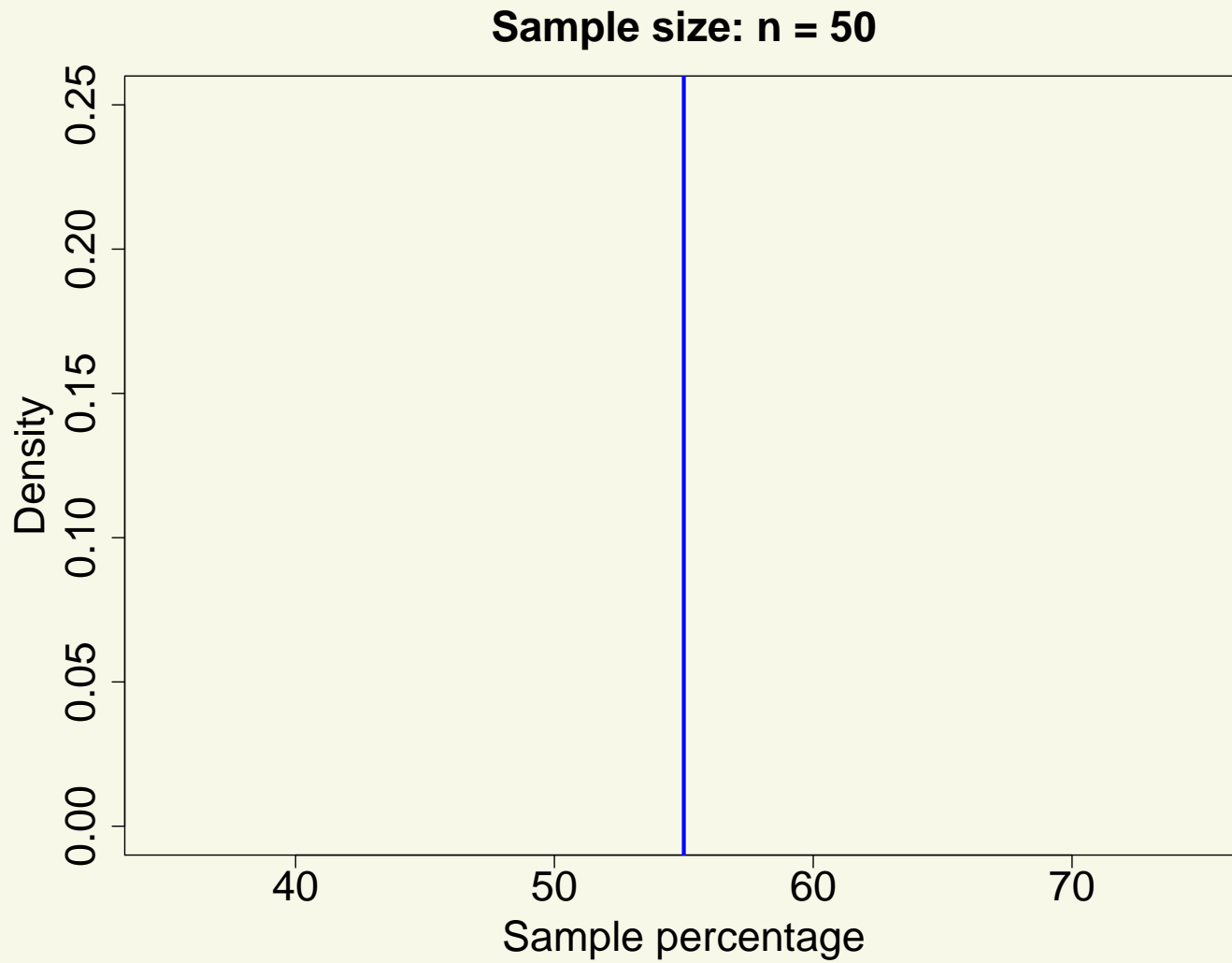
# Random error

Sample % = population % + random error

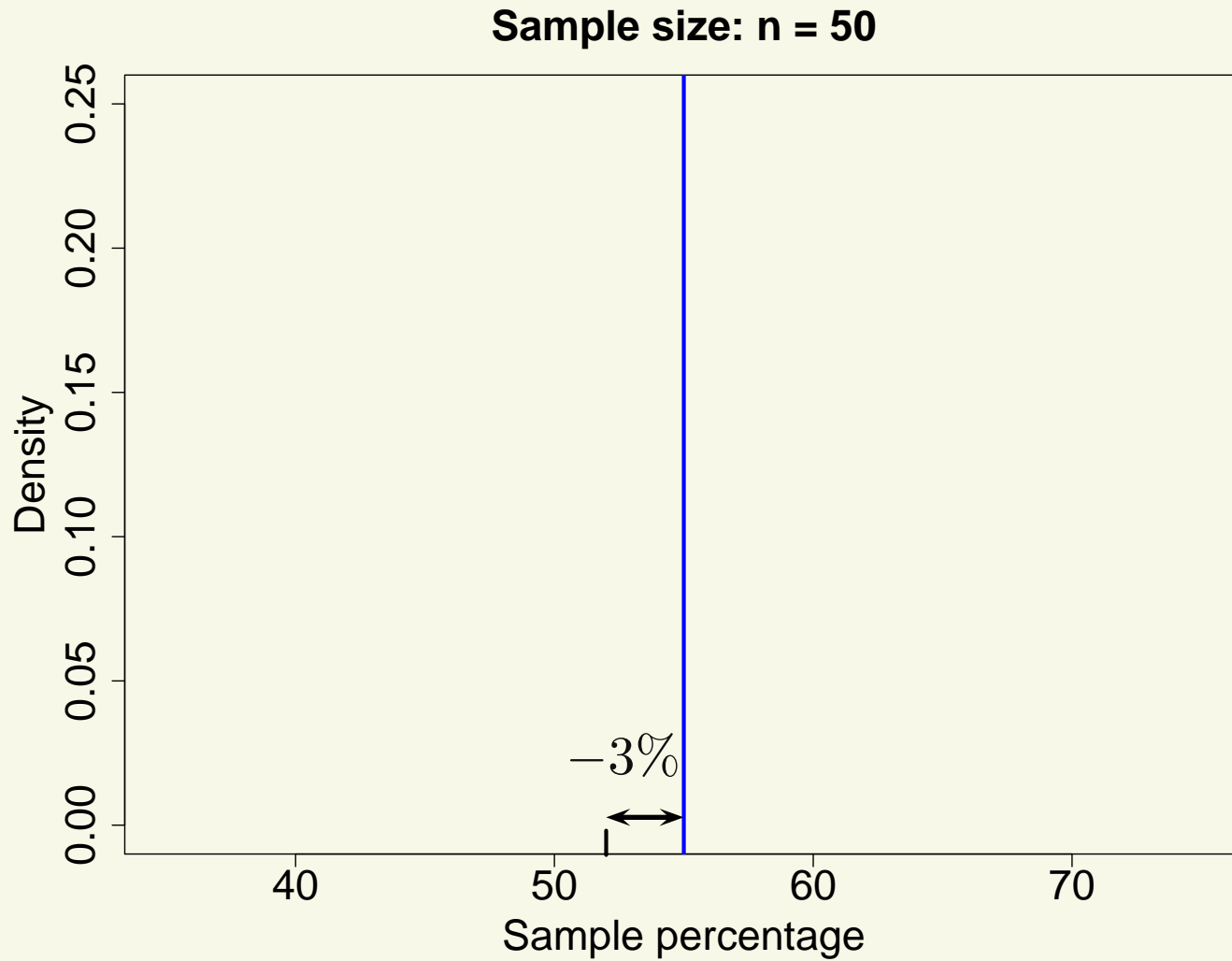
- Example 1:
  - Sample percentage is 52%
  - Size of random error is of the order of 0.1%
  - We predict that the democrats will win
- Example 2:
  - Sample percentage is 52%
  - Size of random error is of the order of 5%
  - We cannot predict a winner with confidence

Conclusion: We need to study the precision of the estimate

# Computer simulation

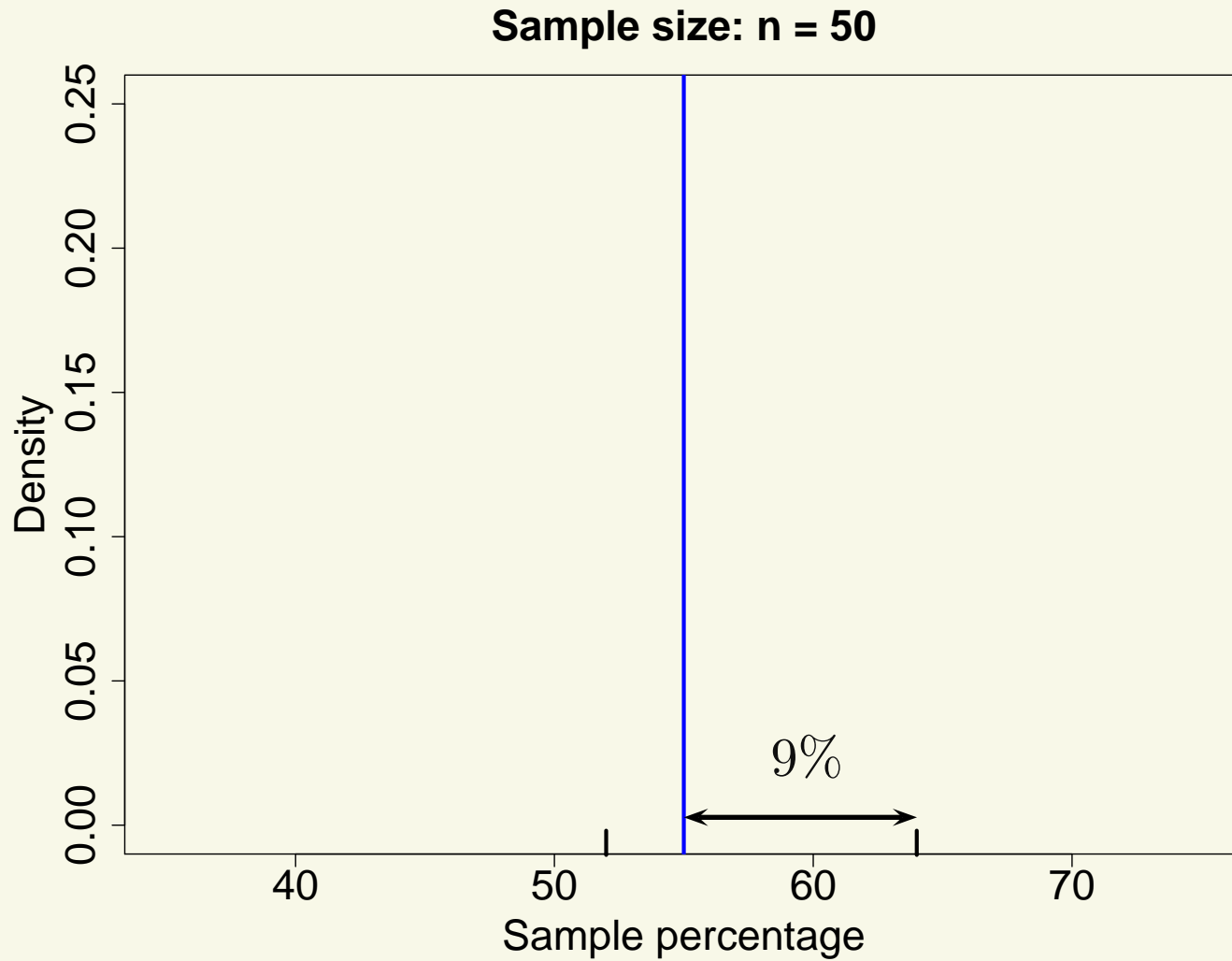


# Computer simulation



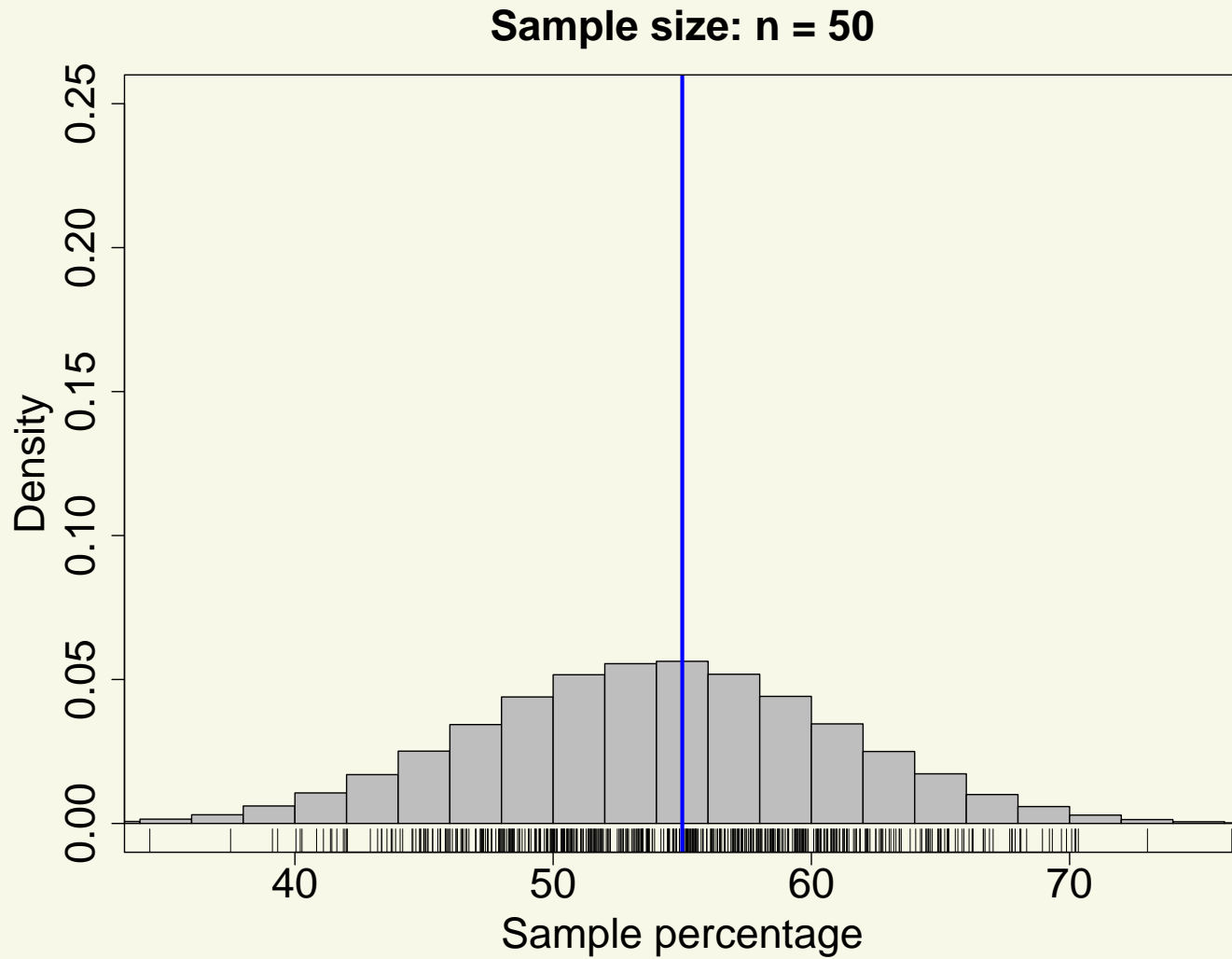
Sample % = population % + random error

# Computer simulation



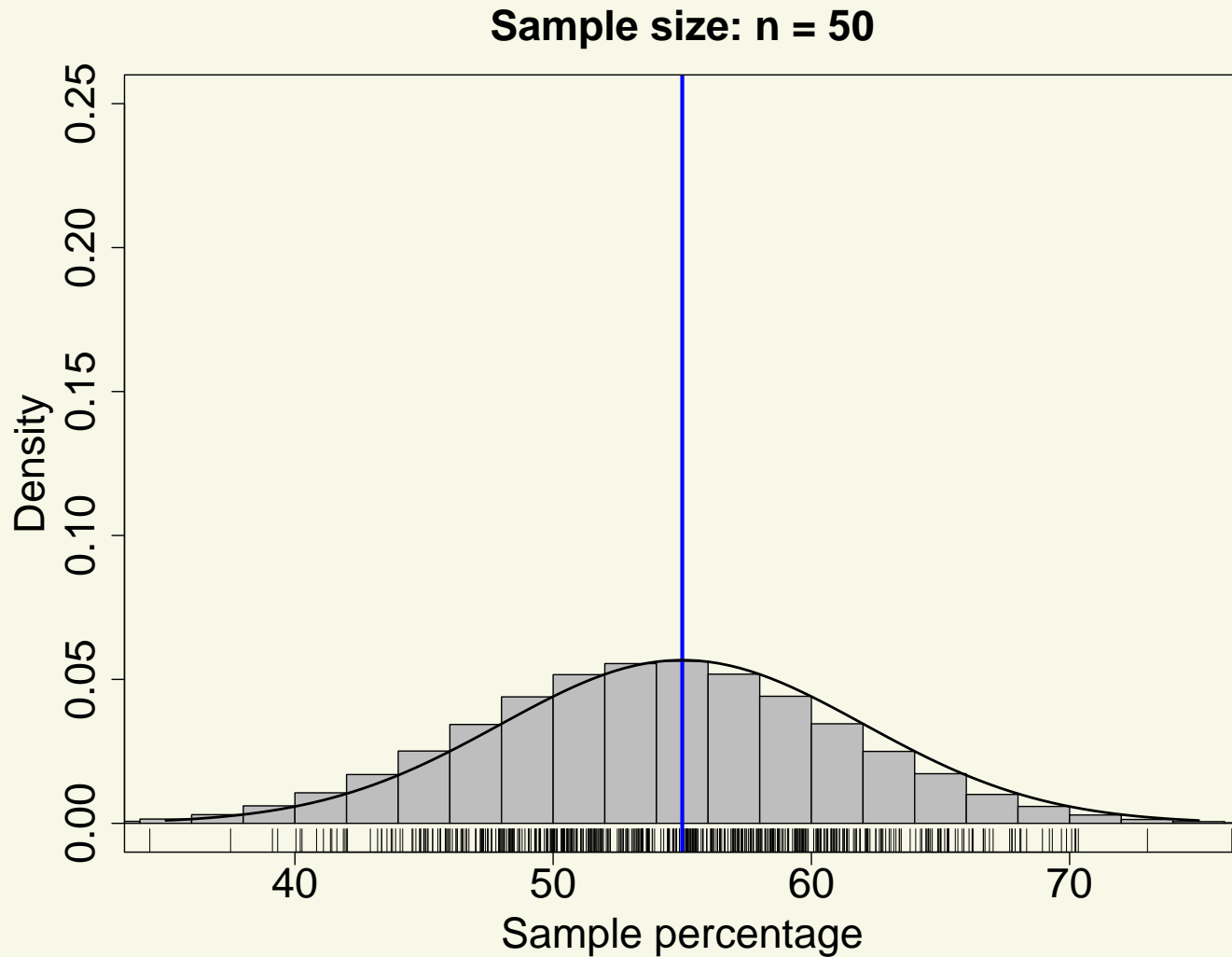
Sample % = population % + random error

# Computer simulation



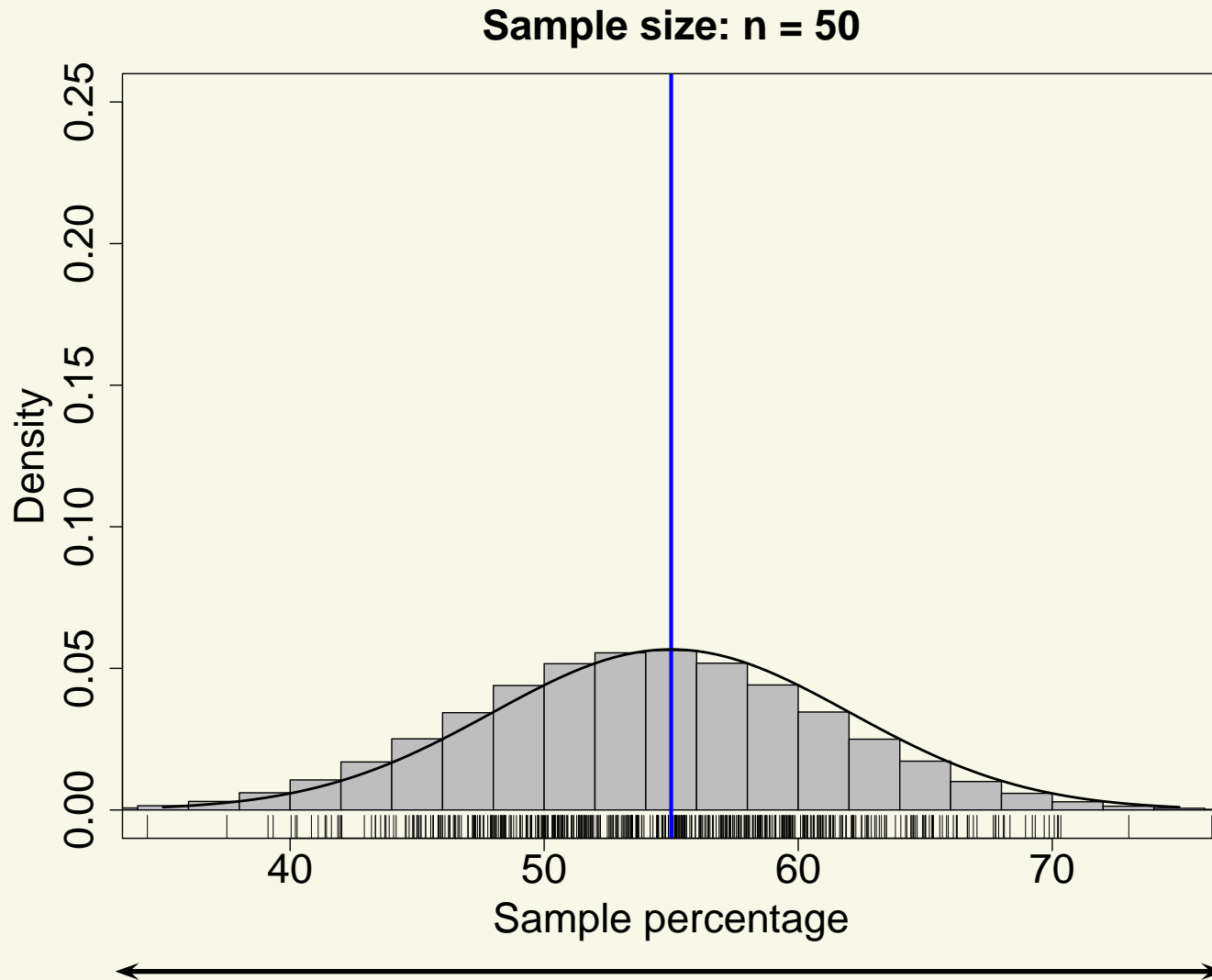
Sample % = population % + random error

# Computer simulation



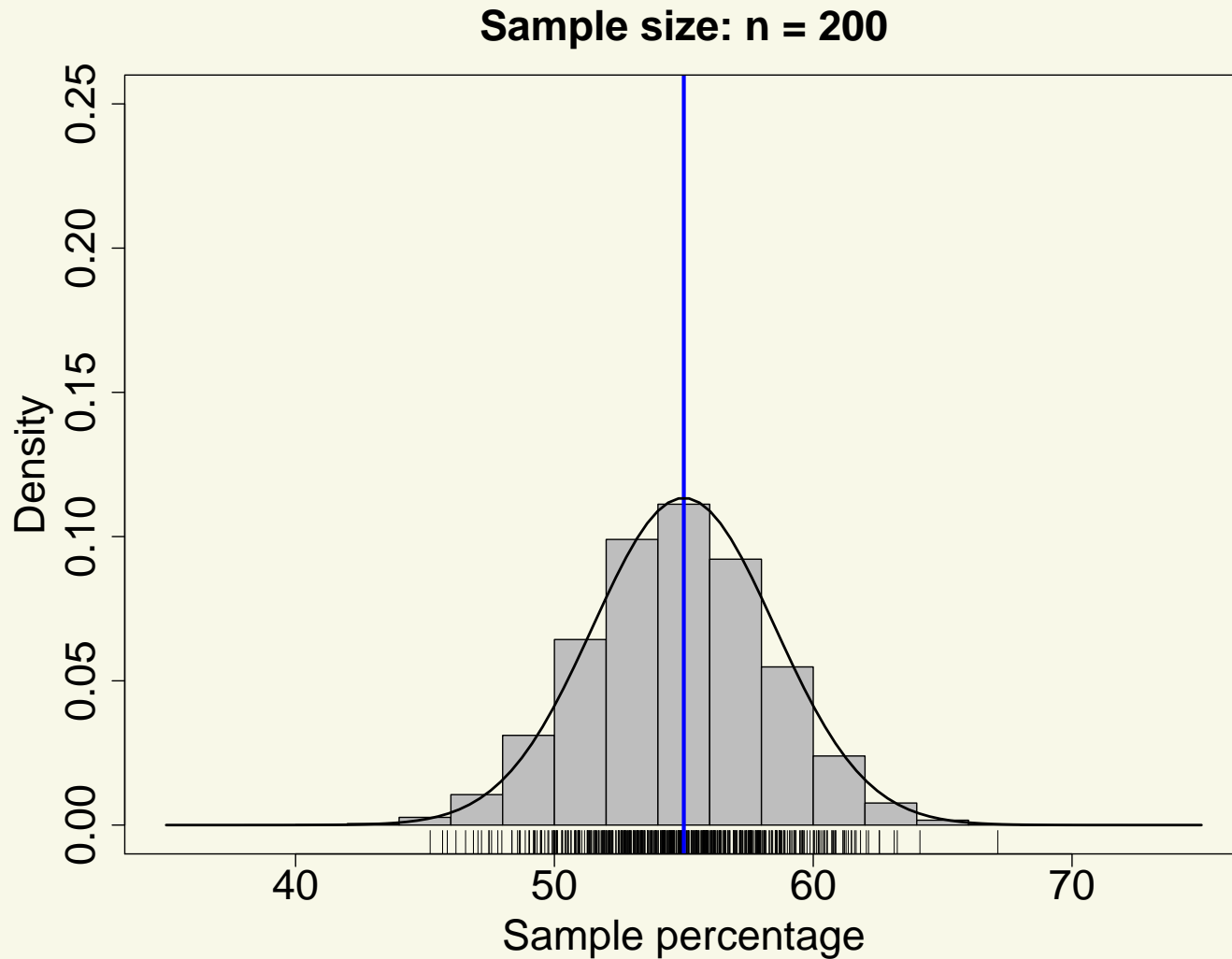
$$\text{Sample \%} = \text{population \%} + (1/\sqrt{n})N(0, \sigma^2)$$

# Computer simulation



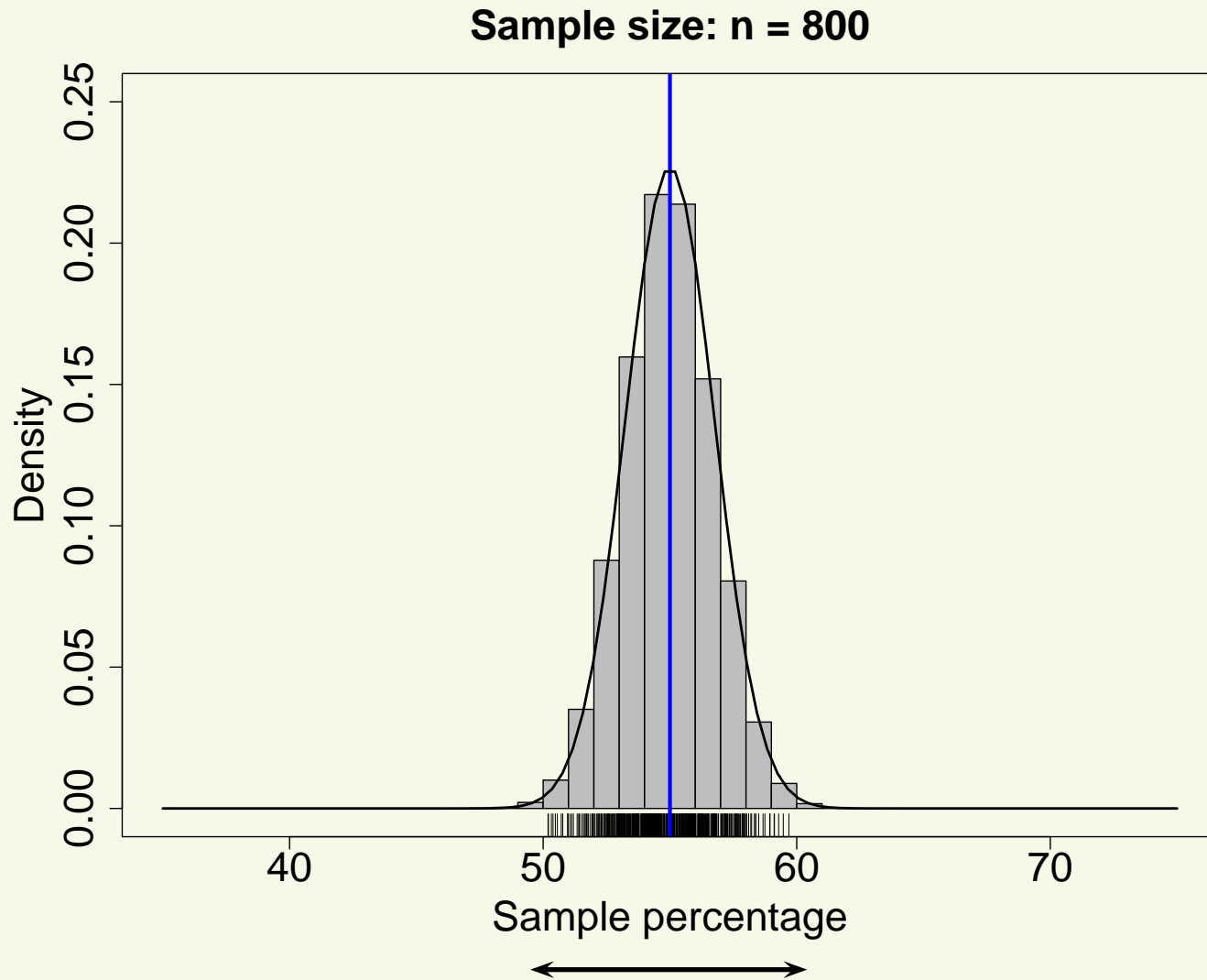
$$\text{Sample \%} = \text{population \%} + (1/\sqrt{n})N(0, \sigma^2)$$

# Computer simulation



$$\text{Sample \%} = \text{population \%} + (1/\sqrt{n})N(0, \sigma^2)$$

# Computer simulation



$$\text{Sample \%} = \text{population \%} + (1/\sqrt{n})N(0, \sigma^2)$$

# Central Limit Theorem

Estimate = population parameter +  $(1/\sqrt{n})N(0, \sigma^2)$

- Contains three parts:
  - Estimate converges to population parameter as  $n \rightarrow \infty$
  - Rate of convergence:  $1/\sqrt{n}$
  - Shape of error distribution:  $N(0, \sigma^2)$

# Central Limit Theorem

Estimate = population parameter +  $(1/\sqrt{n})N(0, \sigma^2)$

- Contains three parts:
  - Estimate converges to population parameter as  $n \rightarrow \infty$
  - Rate of convergence:  $1/\sqrt{n}$
  - Shape of error distribution:  $N(0, \sigma^2)$
- Some practical consequences:
  - Given a sample size  $n$ , we know the precision of the estimate. Hence, we can draw conclusions.
  - Given a certain error tolerance, we can compute the sample size we need.

## Time to event data

Time period between a certain beginning and a certain event:

- Lifetime of people
  - Useful for pension plans, demographic projections

## Time to event data

Time period between a certain beginning and a certain event:

- Lifetime of people
  - Useful for pension plans, demographic projections
- Age at HIV infection
  - Useful for monitoring of epidemic, prevention campaigns

# Time to event data

Time period between a certain beginning and a certain event:

- Lifetime of people
  - Useful for pension plans, demographic projections
- Age at HIV infection
  - Useful for monitoring of epidemic, prevention campaigns
- Remission time of cancer
  - Useful for patient information

## Time to event data

Time period between a certain beginning and a certain event:

- Lifetime of people
  - Useful for pension plans, demographic projections
- Age at HIV infection
  - Useful for monitoring of epidemic, prevention campaigns
- Remission time of cancer
  - Useful for patient information
- Time to failure of a machine
  - Useful for maintenance schedules

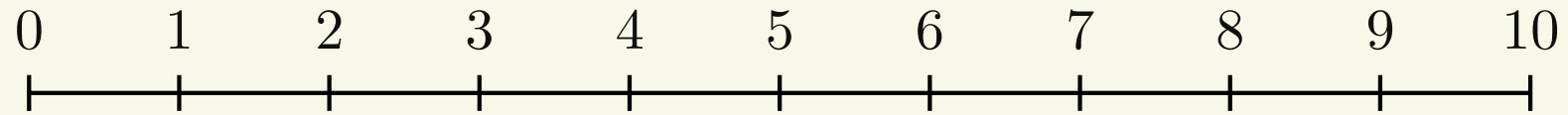
## Time to event data

Time period between a certain beginning and a certain event:

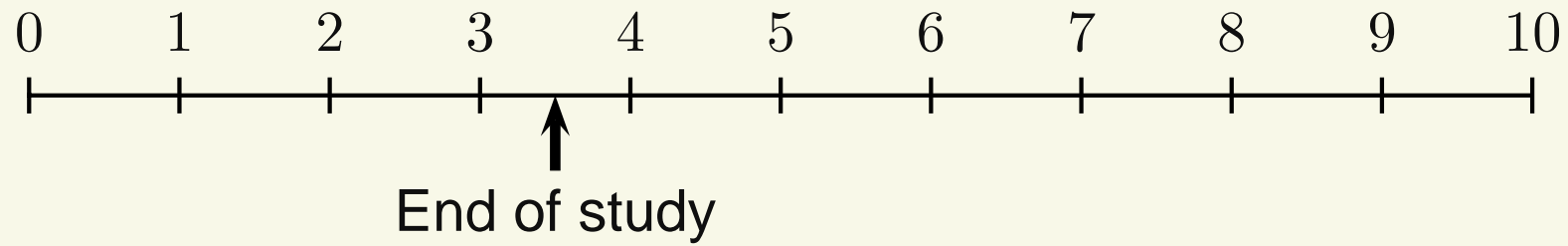
- Lifetime of people
  - Useful for pension plans, demographic projections
- Age at HIV infection
  - Useful for monitoring of epidemic, prevention campaigns
- Remission time of cancer
  - Useful for patient information
- Time to failure of a machine
  - Useful for maintenance schedules
- Unemployment time
  - Useful for monitoring the job market

The Worst  
Part of  
Censorship  
is ~~what~~  
~~what~~

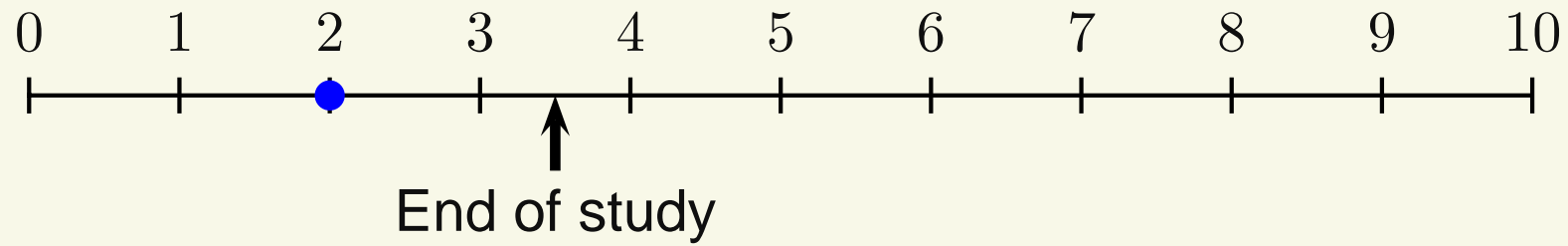
## Right censored data



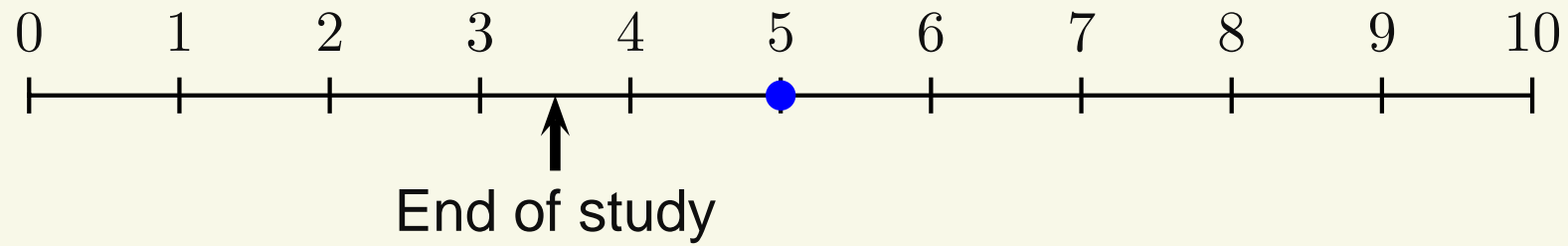
## Right censored data



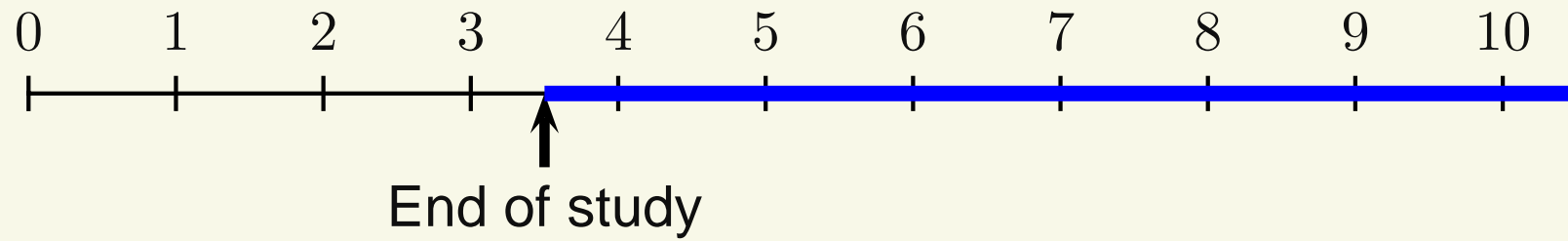
# Right censored data



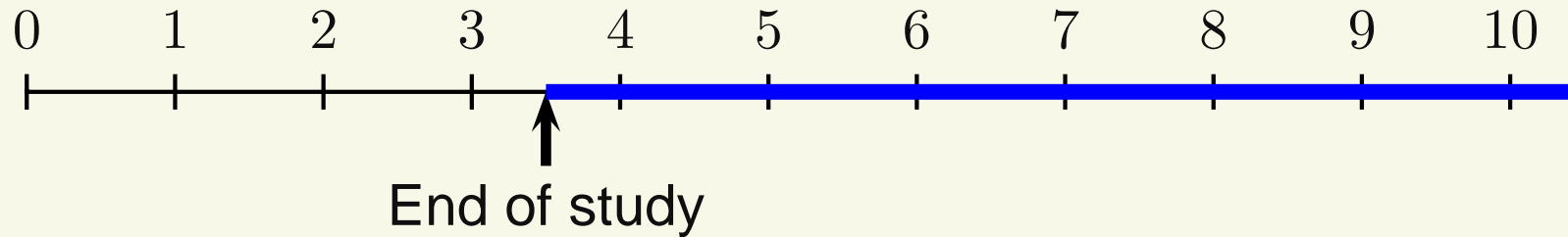
## Right censored data



## Right censored data



## Right censored data

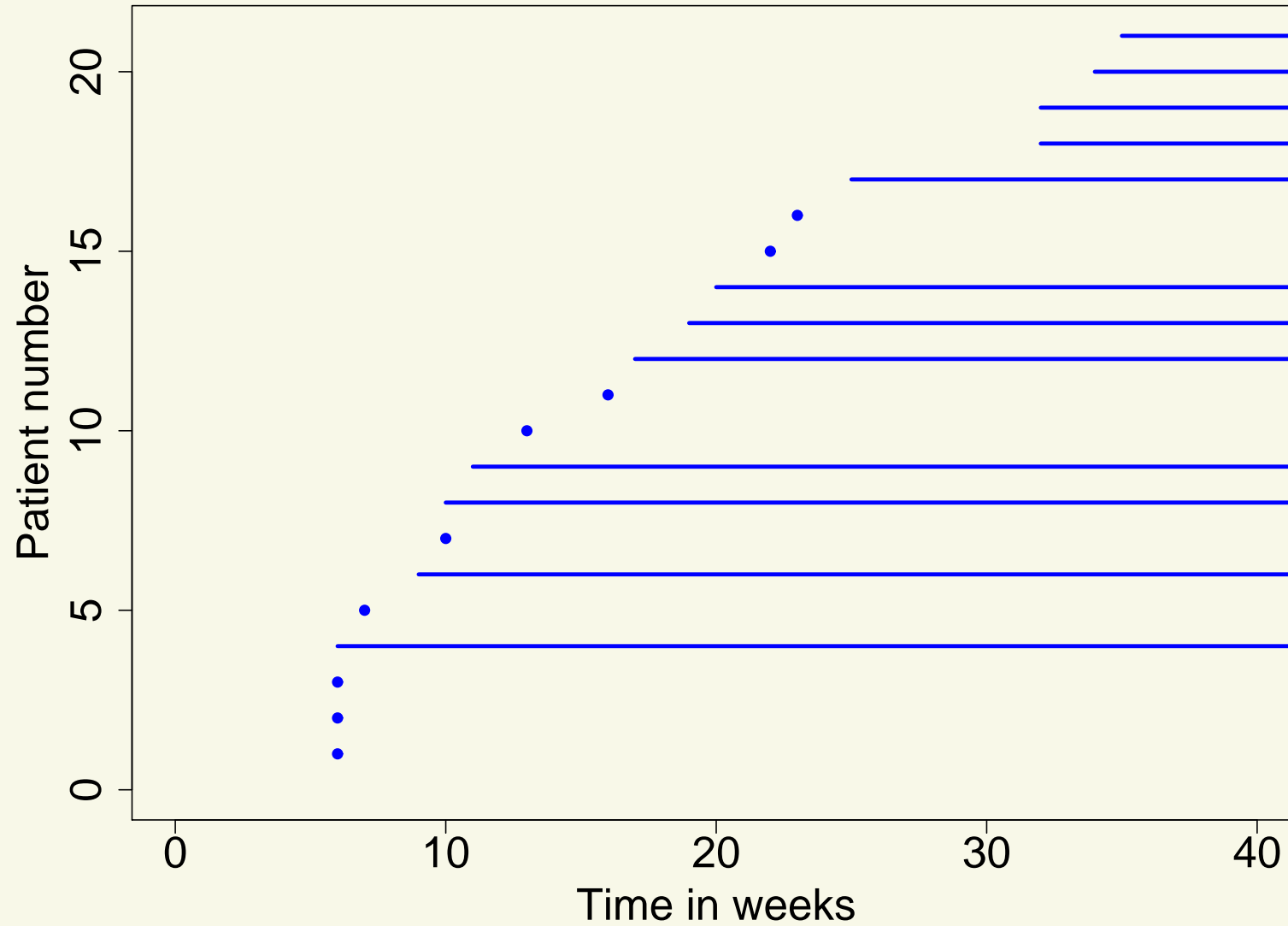


General situation:

- Interested in the time to an event
- The study has a finite time length and/or subjects drop out

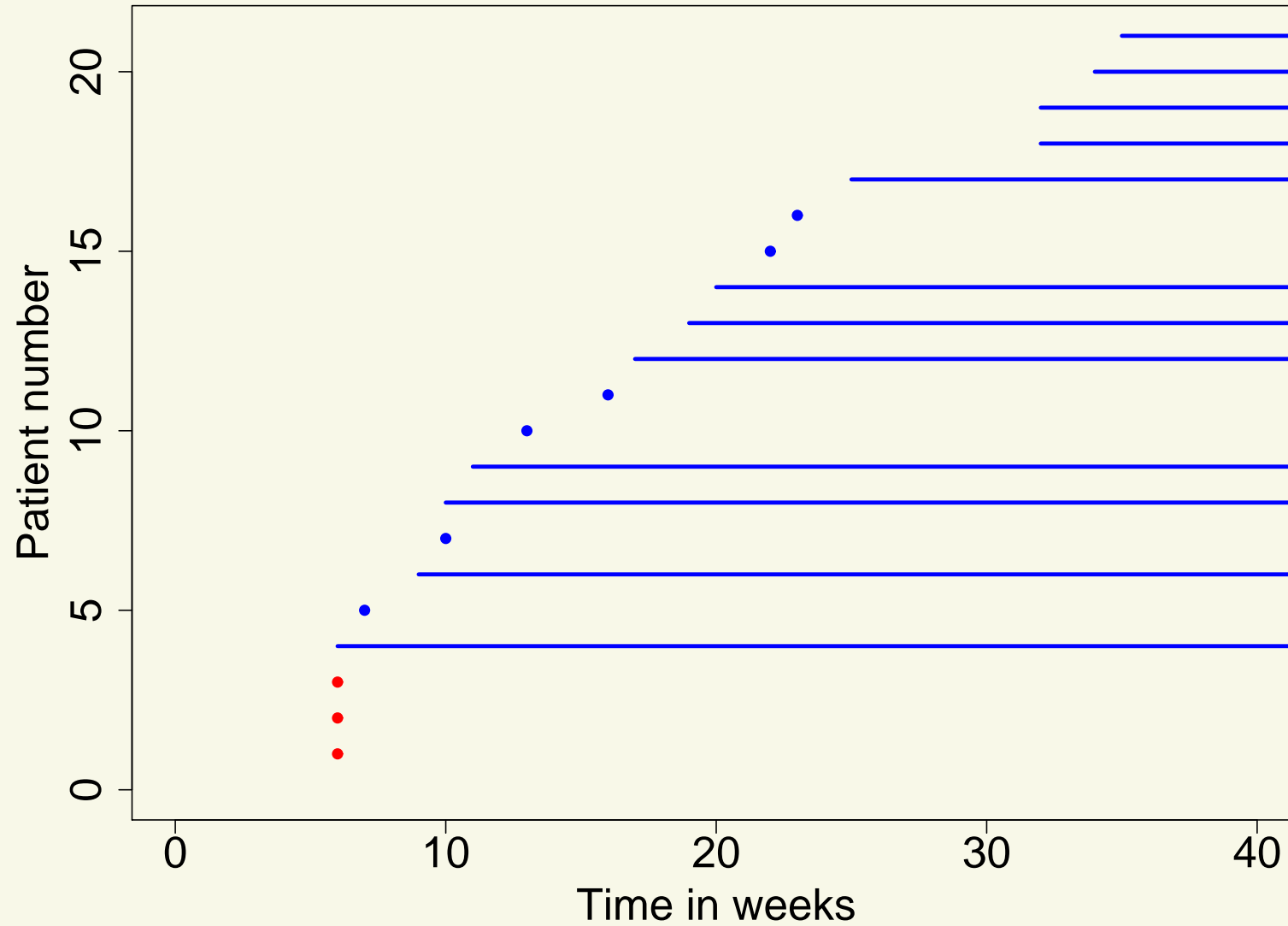
# Right censored data

Remission times of leukemia (Freireich et al., 1963)



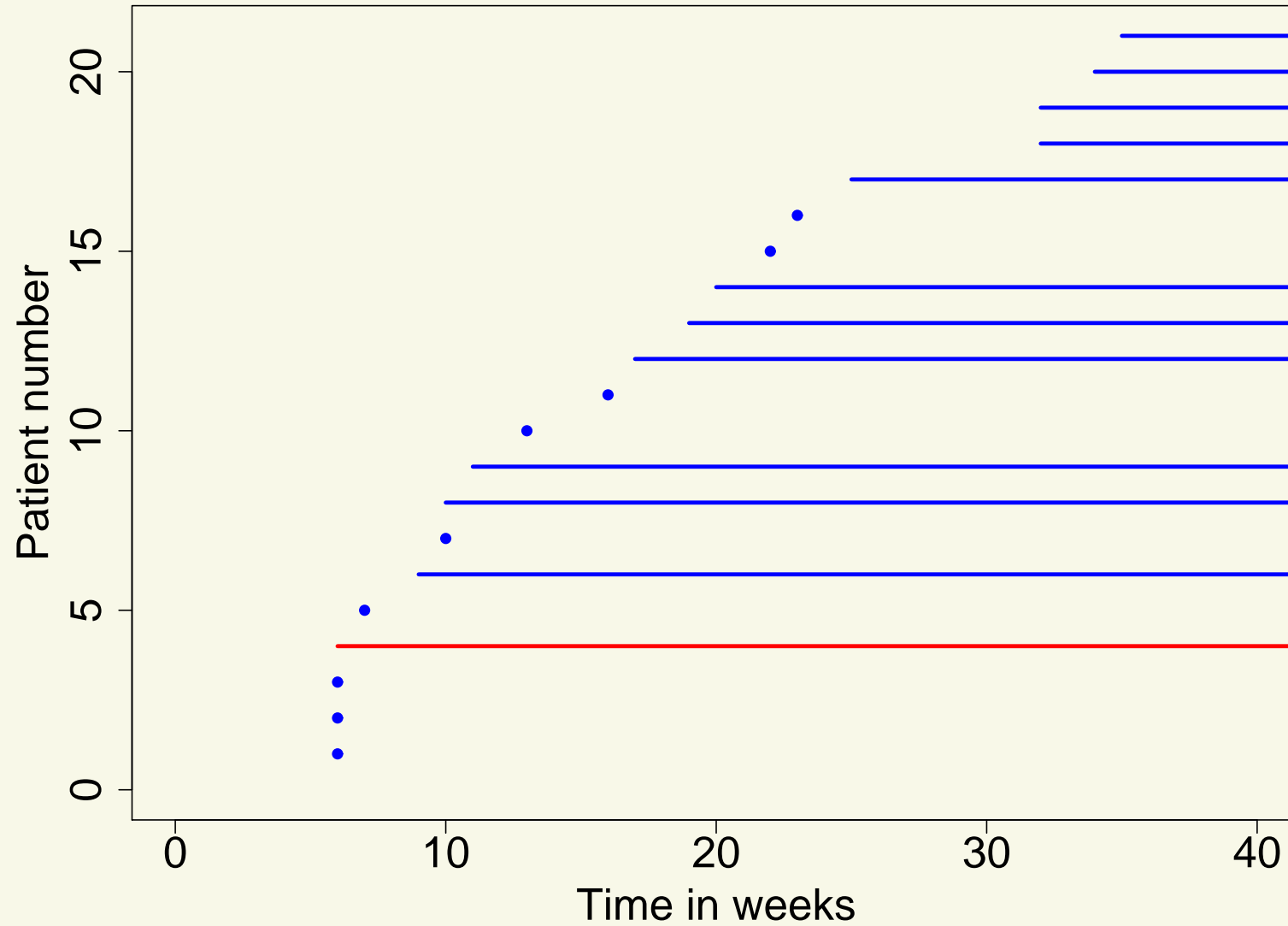
# Right censored data

Remission times of leukemia (Freireich et al., 1963)

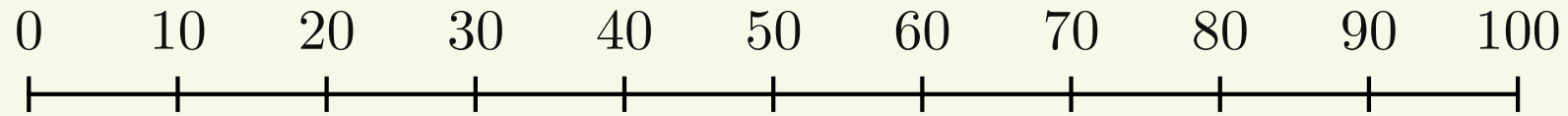


# Right censored data

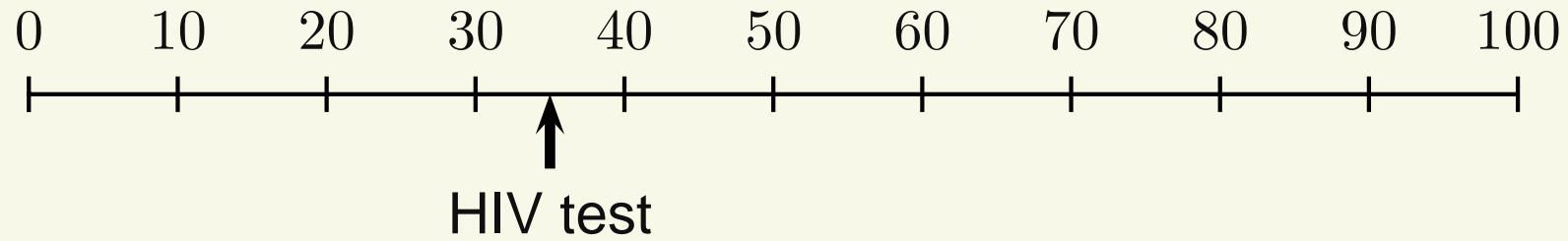
Remission times of leukemia (Freireich et al., 1963)



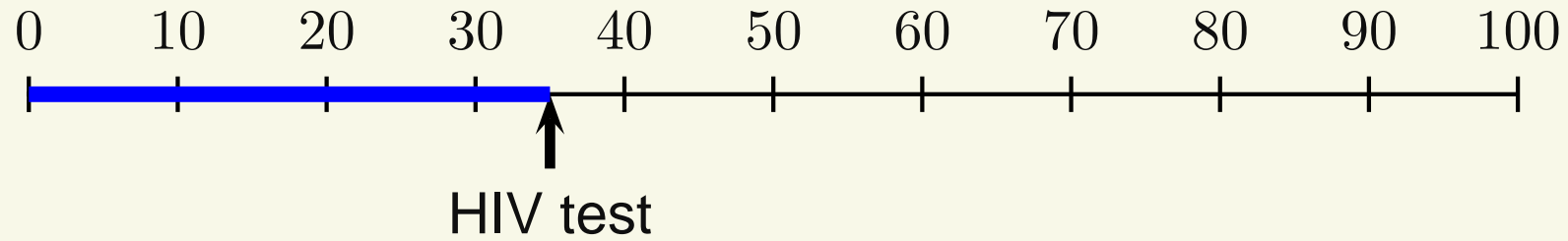
## Interval censored data with 1 observation time



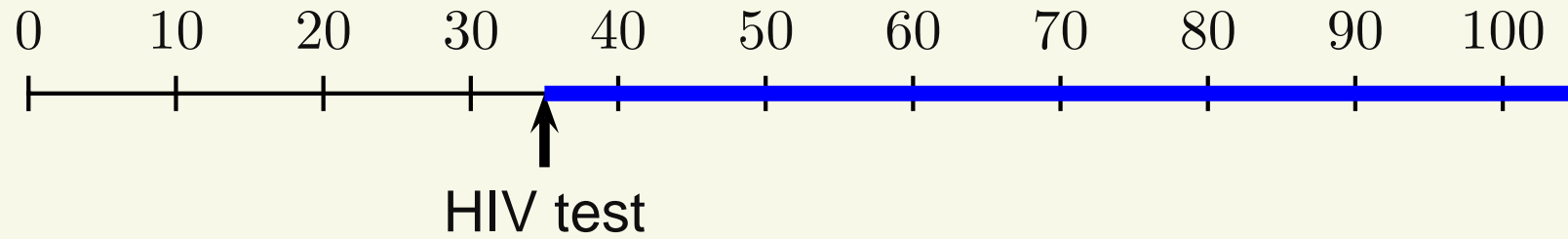
## Interval censored data with 1 observation time



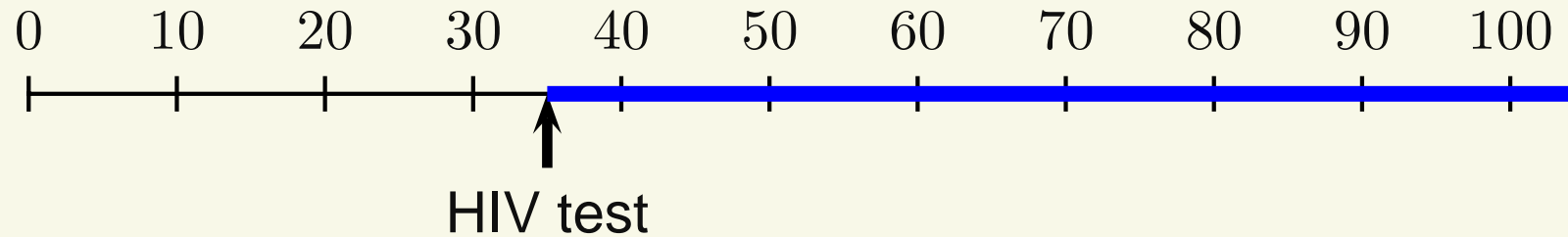
# Interval censored data with 1 observation time



## Interval censored data with 1 observation time



## Interval censored data with 1 observation time

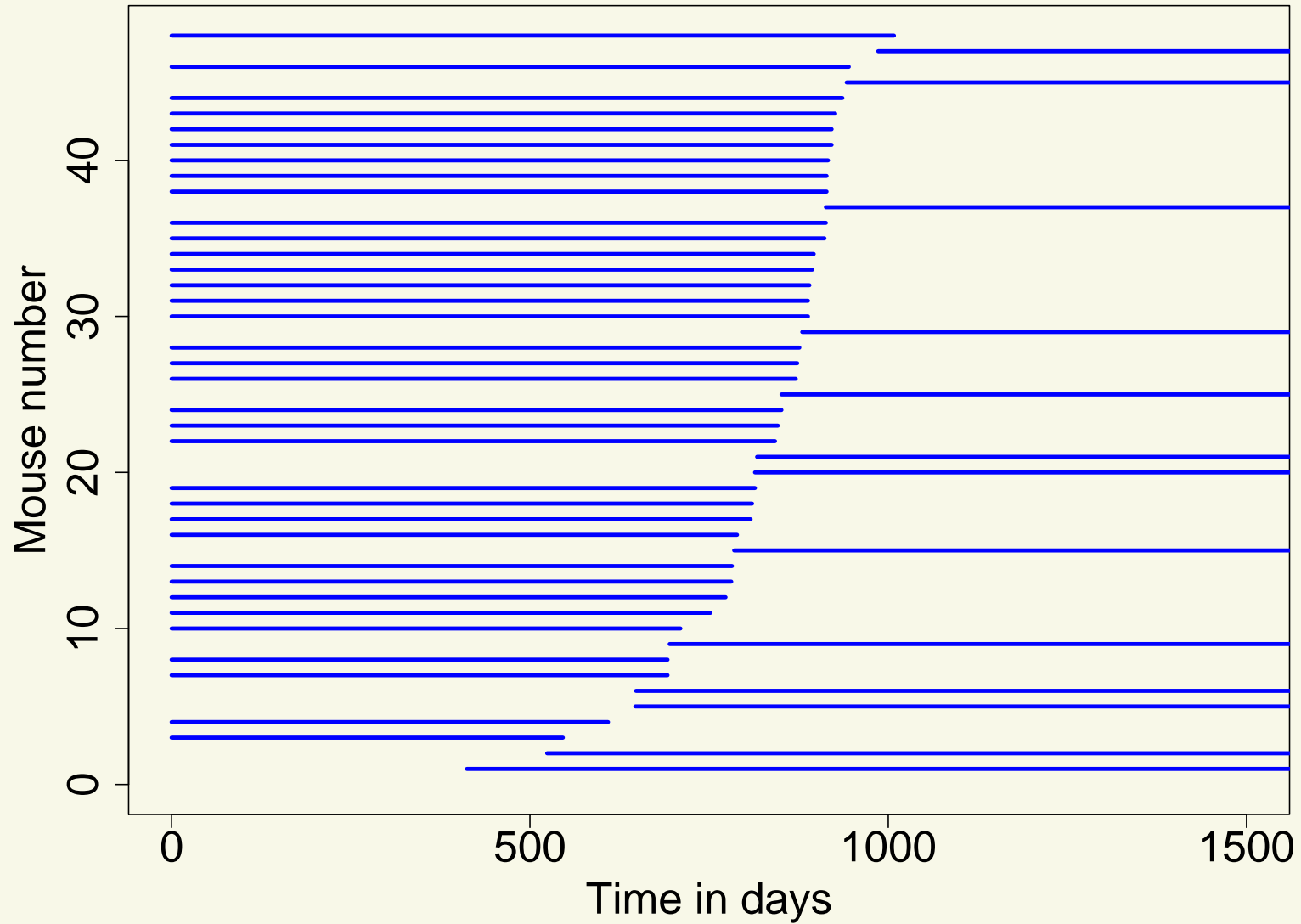


General situation:

- Interested in the time to an event
- Only know the status at one point

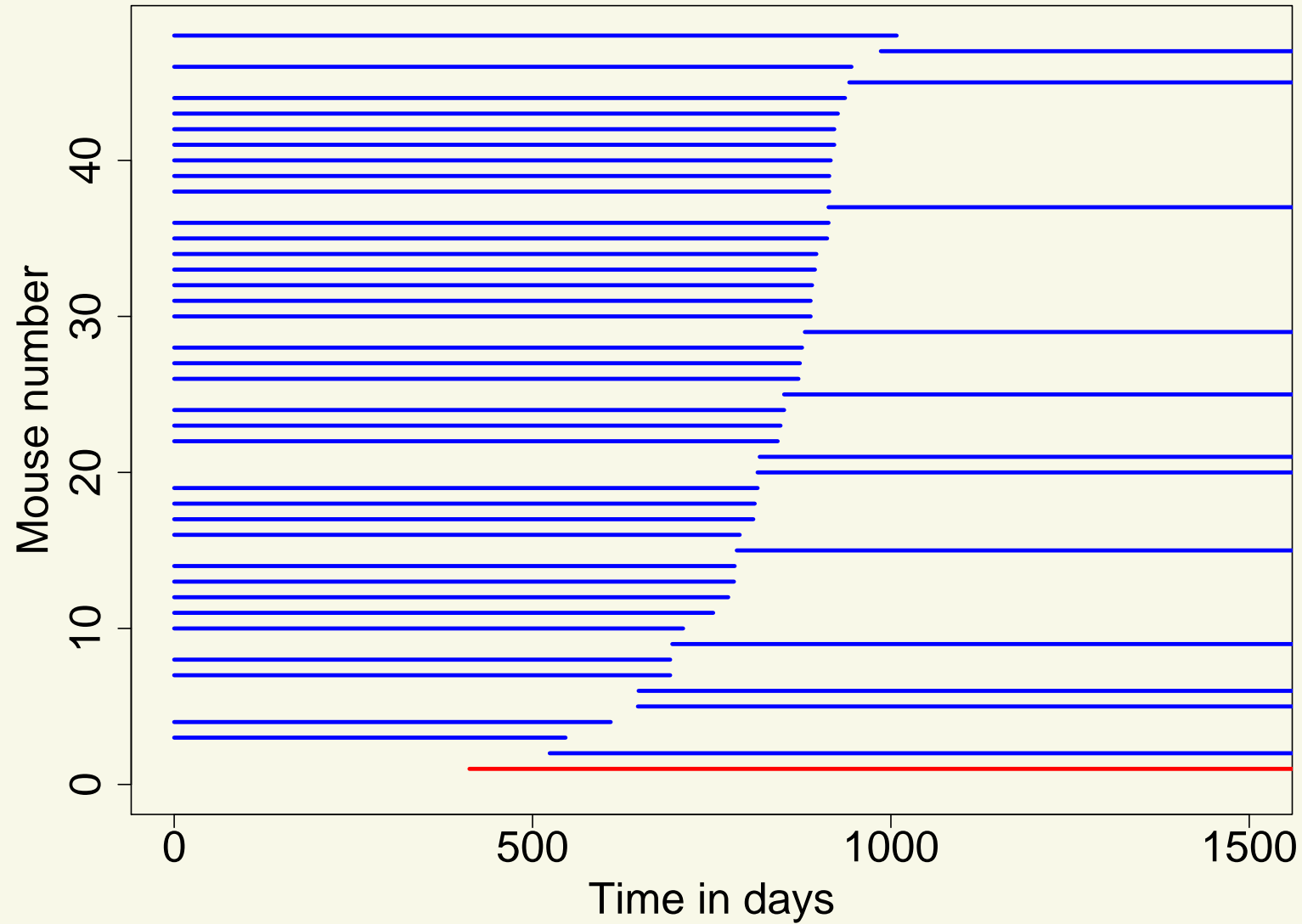
# Interval censored data with 1 observation time

Time to lung tumor (Hoel and Walberg, 1972)



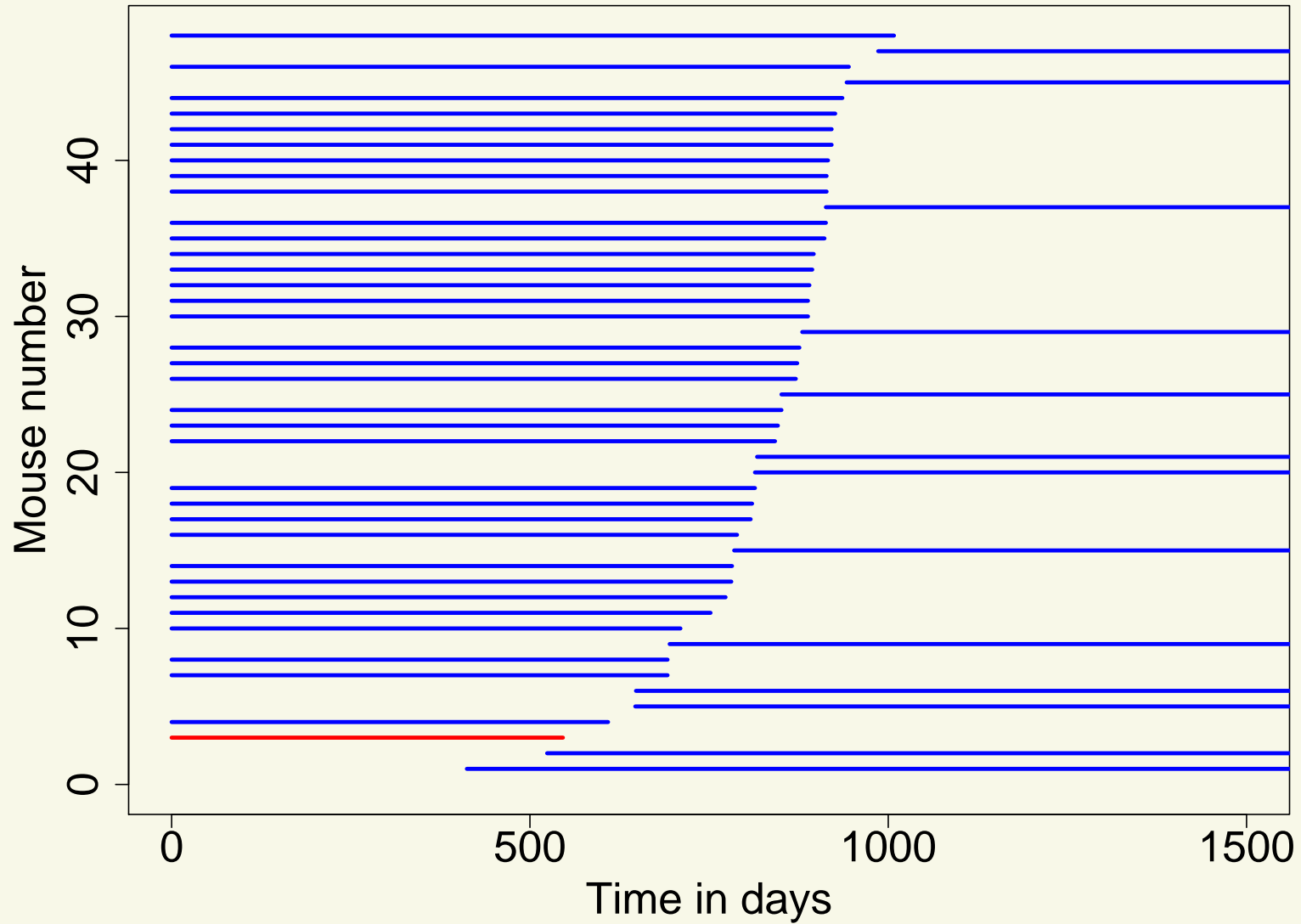
# Interval censored data with 1 observation time

Time to lung tumor (Hoel and Walberg, 1972)

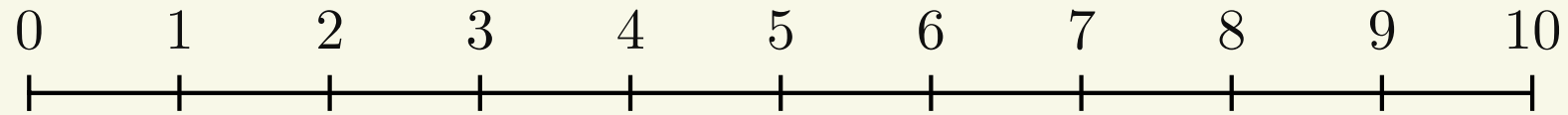


# Interval censored data with 1 observation time

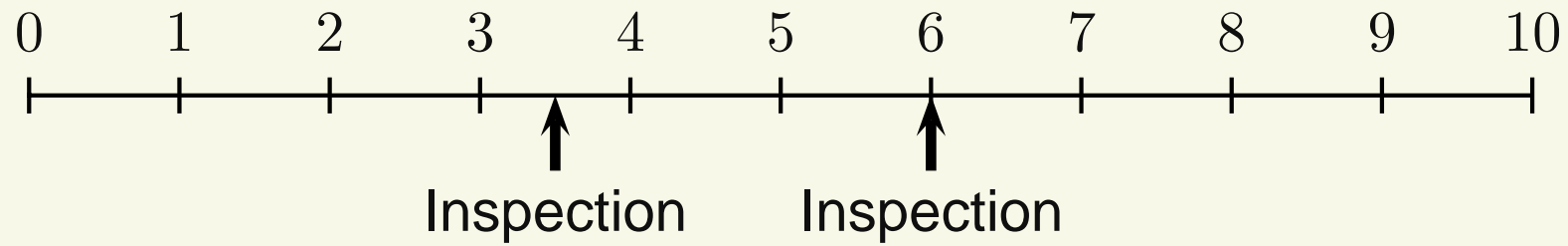
Time to lung tumor (Hoel and Walberg, 1972)



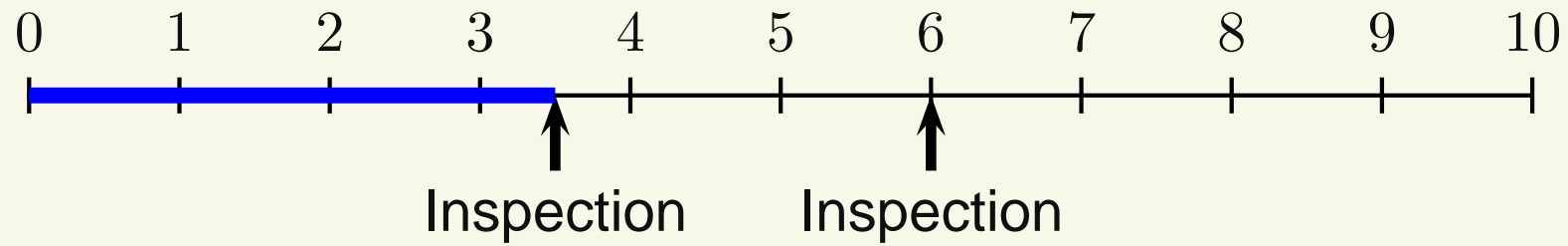
# Interval censored data with several observation times



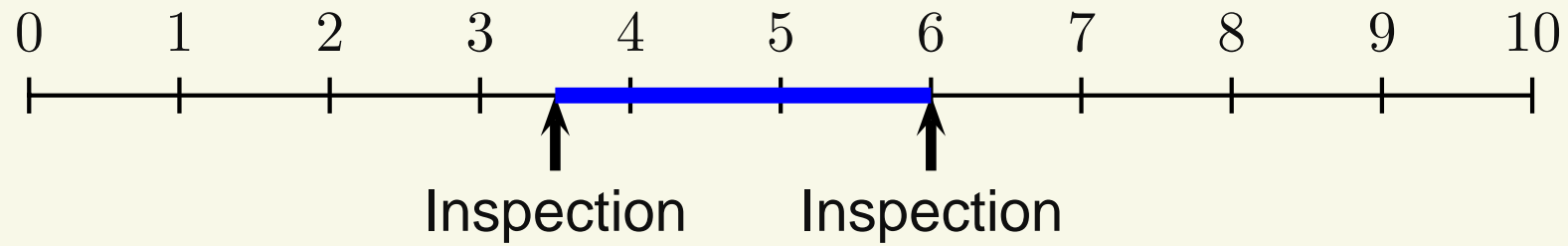
# Interval censored data with several observation times



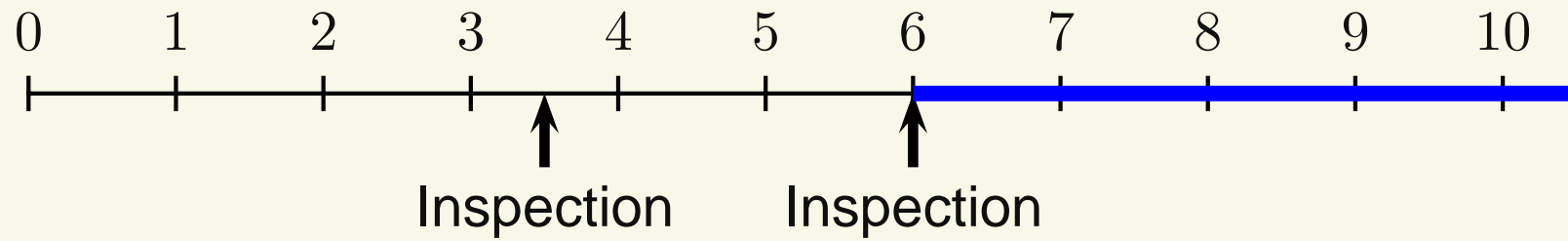
# Interval censored data with several observation times



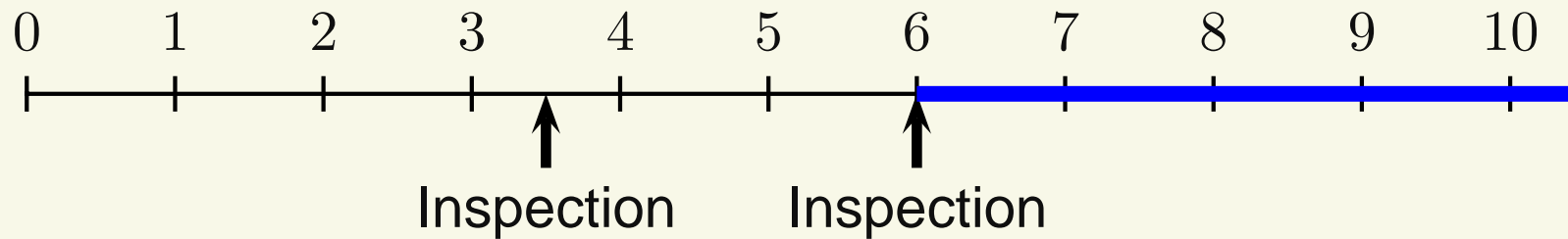
# Interval censored data with several observation times



# Interval censored data with several observation times



## Interval censored data with several observation times

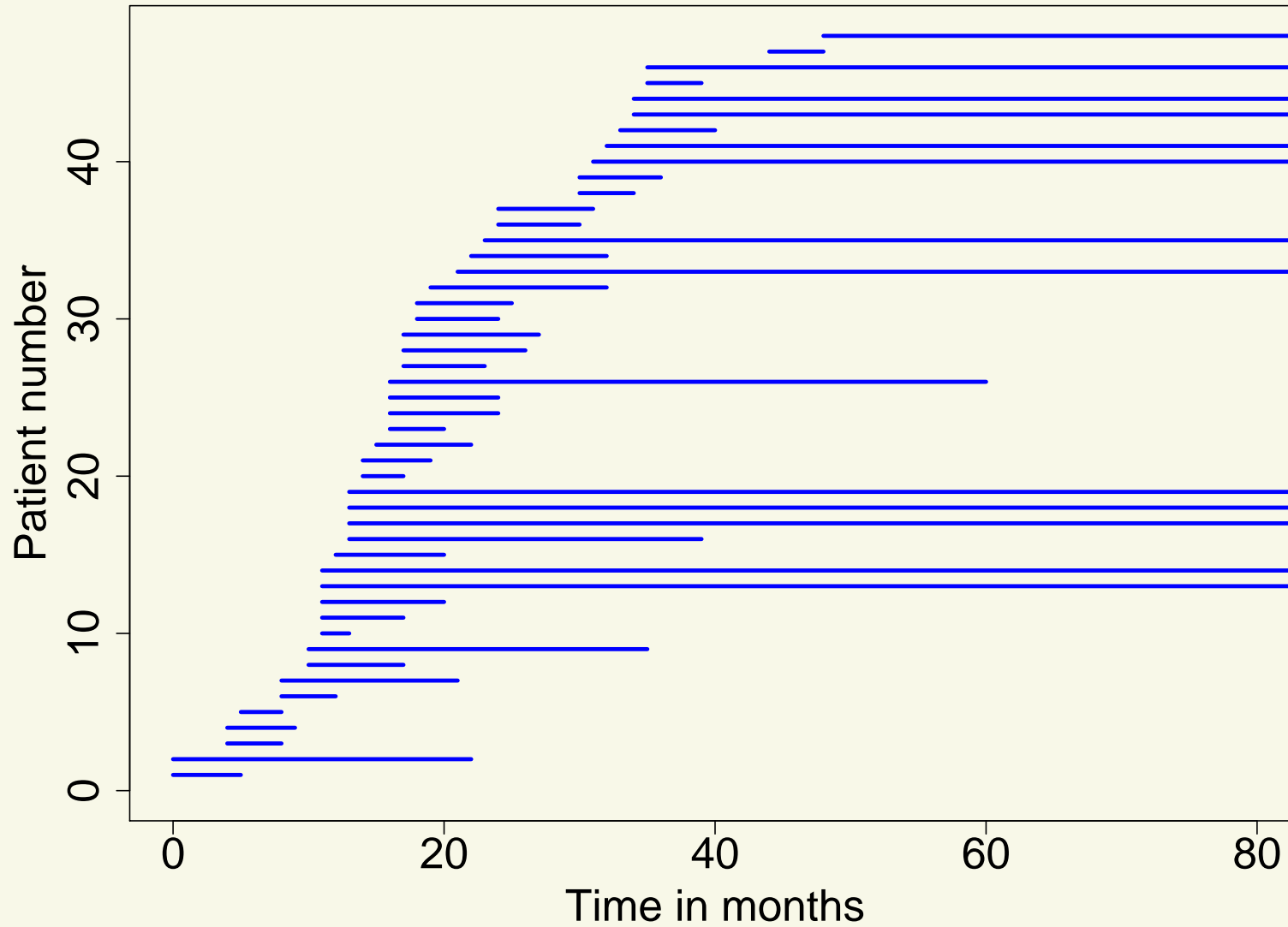


General situation:

- Interested in the time to an event
- Only have data from periodic inspections

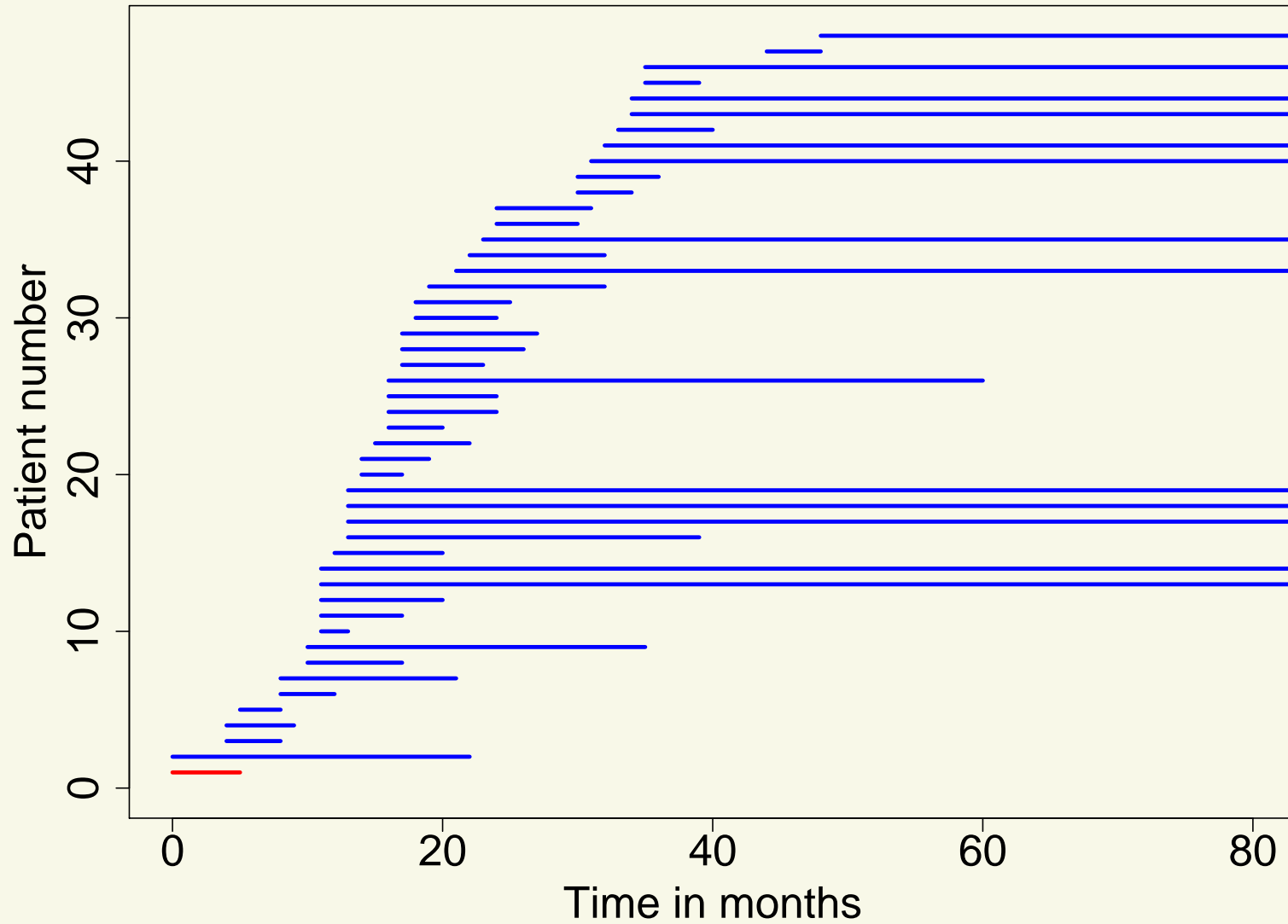
# Interval censored data with several observation times

Time to cosmetic deterioration (Finkelstein and Wolfe, 1985)



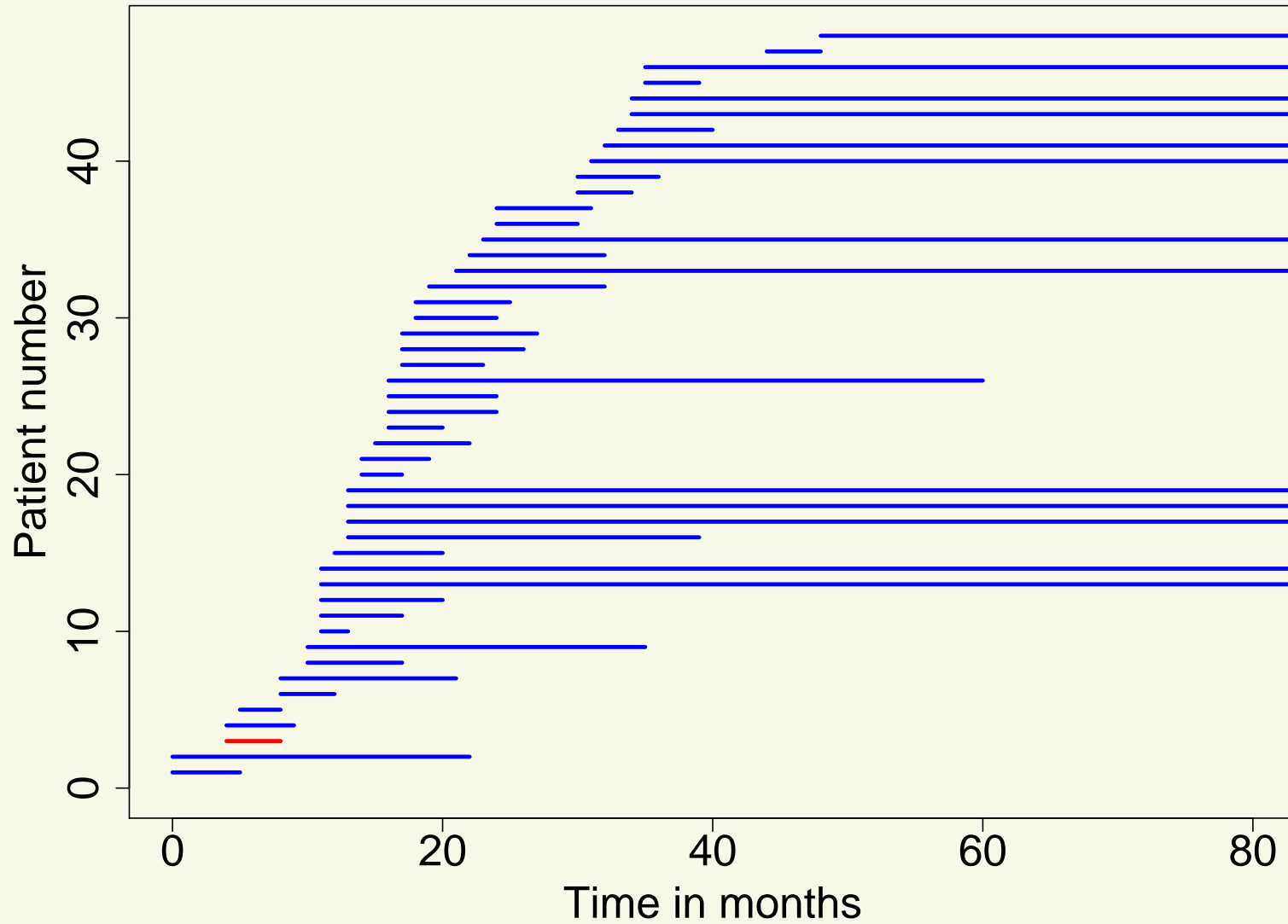
# Interval censored data with several observation times

Time to cosmetic deterioration (Finkelstein and Wolfe, 1985)



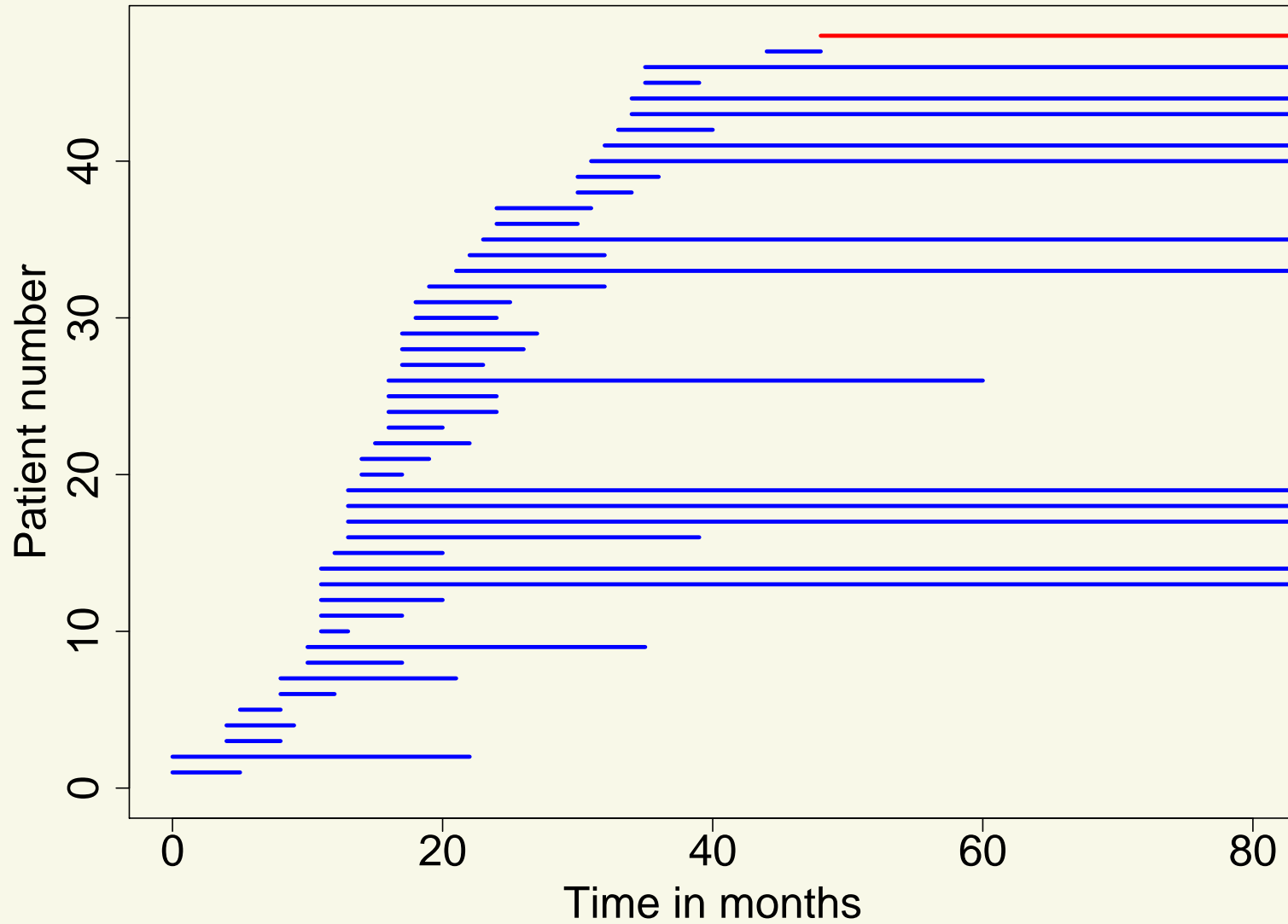
# Interval censored data with several observation times

Time to cosmetic deterioration (Finkelstein and Wolfe, 1985)

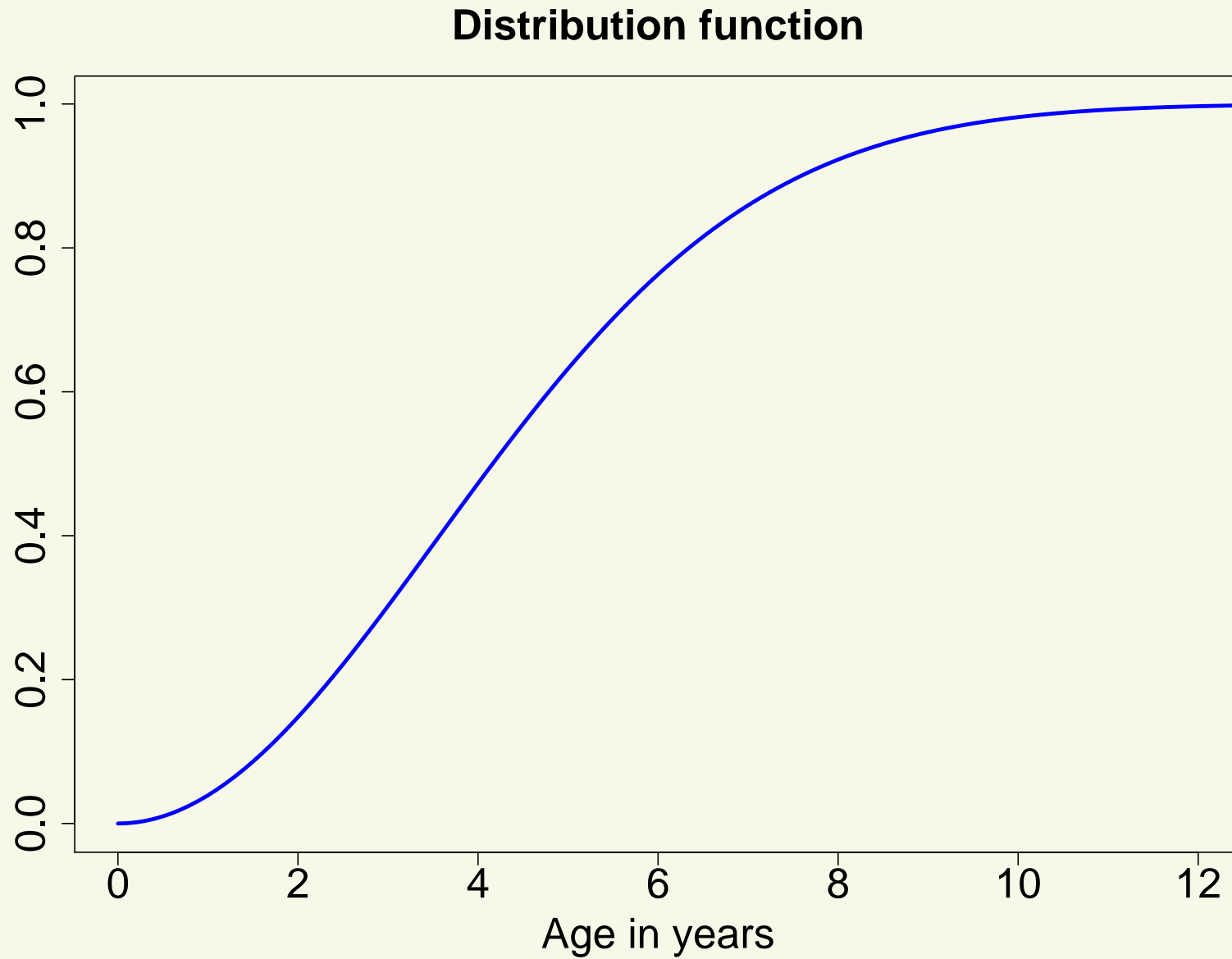


# Interval censored data with several observation times

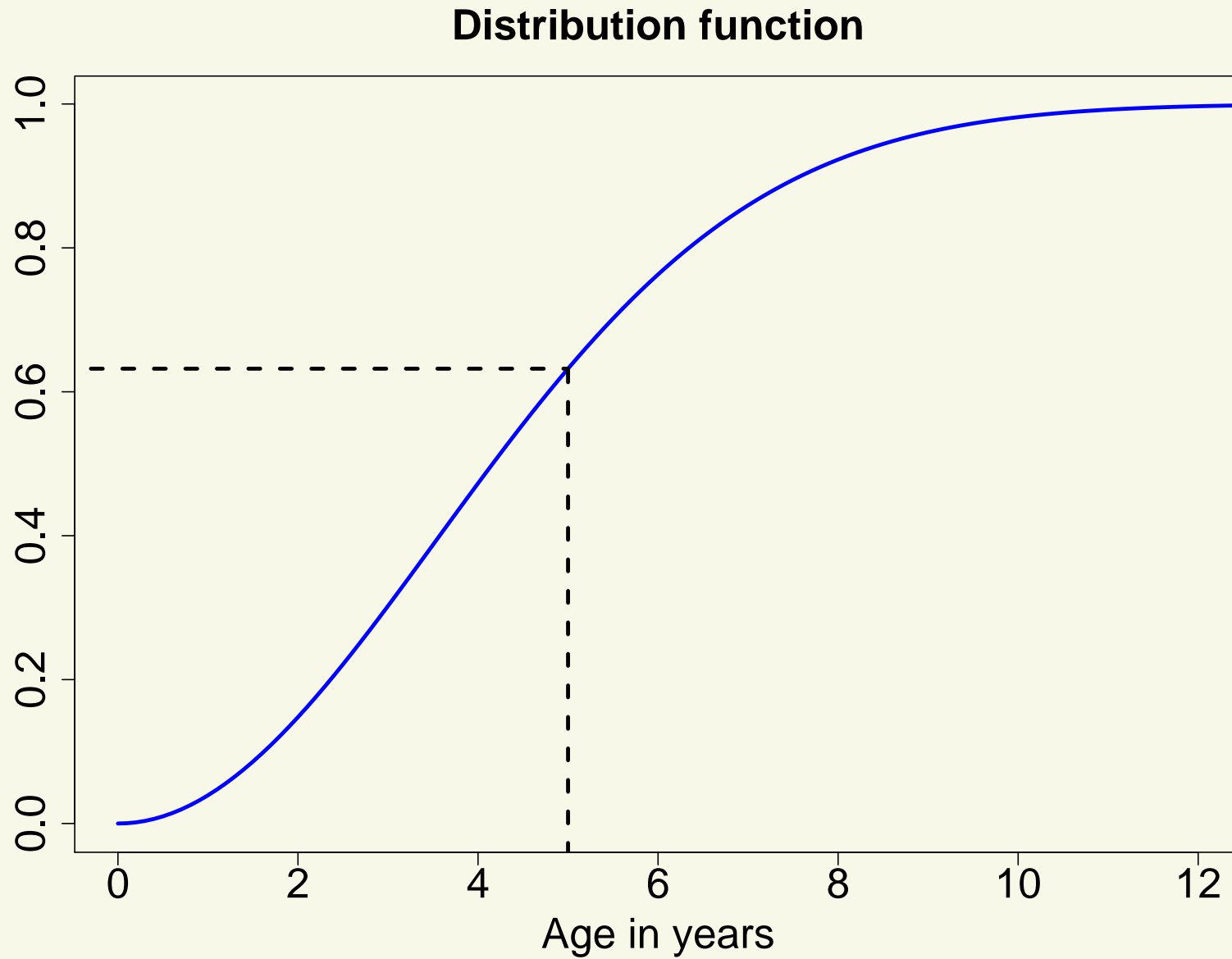
Time to cosmetic deterioration (Finkelstein and Wolfe, 1985)



# Distribution function



# Distribution function



## Problem formulation

- Want to know the distribution function of event times in a population

# Problem formulation

- Want to know the distribution function of event times in a population
- Example:
  - Population: Swiss children born after 2005
  - Want to know: Distribution function of the age at first tooth cavity

# Problem formulation

- Want to know the distribution function of event times in a population
- Example:
  - Population: Swiss children born after 2005
  - Want to know: Distribution function of the age at first tooth cavity
- As before:
  - We cannot observe the entire population, so we cannot know the distribution function exactly

# Problem formulation

- Want to know the distribution function of event times in a population
- Example:
  - Population: Swiss children born after 2005
  - Want to know: Distribution function of the age at first tooth cavity
- As before:
  - We cannot observe the entire population, so we cannot know the distribution function exactly
  - We take a random sample, and use it to estimate the distribution function

# Problem formulation

- Want to know the distribution function of event times in a population
- Example:
  - Population: Swiss children born after 2005
  - Want to know: Distribution function of the age at first tooth cavity
- As before:
  - We cannot observe the entire population, so we cannot know the distribution function exactly
  - We take a random sample, and use it to estimate the distribution function
  - We want to understand the precision of this estimate

# Problem formulation

- Want to know the distribution function of event times in a population
- Example:
  - Population: Swiss children born after 2005
  - Want to know: Distribution function of the age at first tooth cavity
- As before:
  - We cannot observe the entire population, so we cannot know the distribution function exactly
  - We take a random sample, and use it to estimate the distribution function
  - We want to understand the precision of this estimate
- Extra complications:
  - Event times in the sample may be subject to censoring
  - Want to estimate function
  - Want to do this under minimal assumptions

# Let the data speak for themselves

- Approach:
  - Do not assume a parametric family
  - Do not make assumptions about smoothness or shape

# Let the data speak for themselves

- Approach:
  - Do not assume a parametric family
  - Do not make assumptions about smoothness or shape
- Some situations in which this nonparametric approach is useful:
  - We do not know what assumptions or parametric form are suitable
  - We want to evaluate a parametric fit

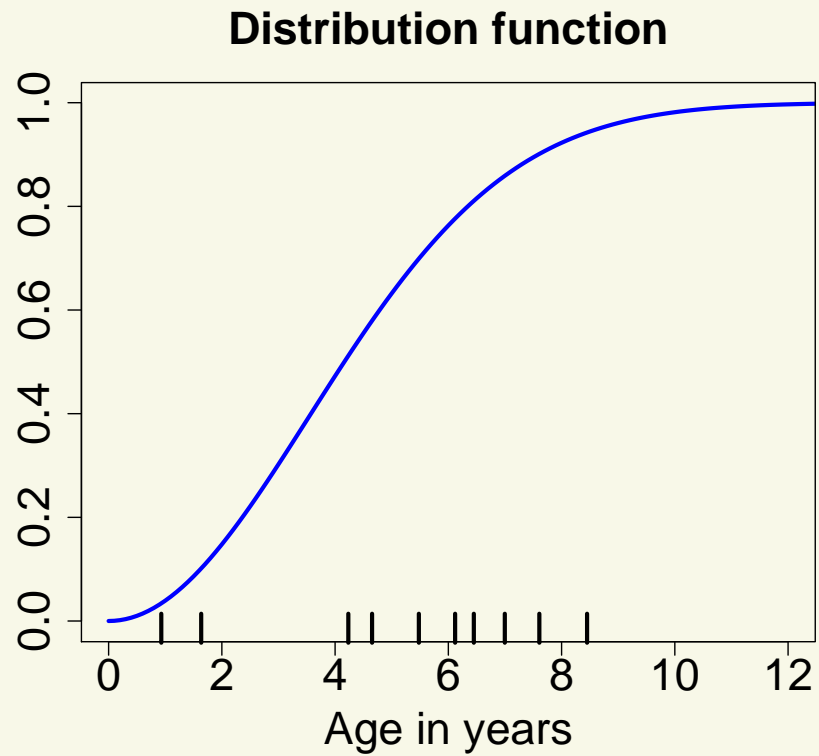
# Let the data speak for themselves

- Approach:
  - Do not assume a parametric family
  - Do not make assumptions about smoothness or shape
- Some situations in which this nonparametric approach is useful:
  - We do not know what assumptions or parametric form are suitable
  - We want to evaluate a parametric fit
- Nonparametric maximum likelihood estimator (MLE):

$$\hat{F}_n = \operatorname{argmax}_{F \in \mathcal{F}} l_n(F)$$

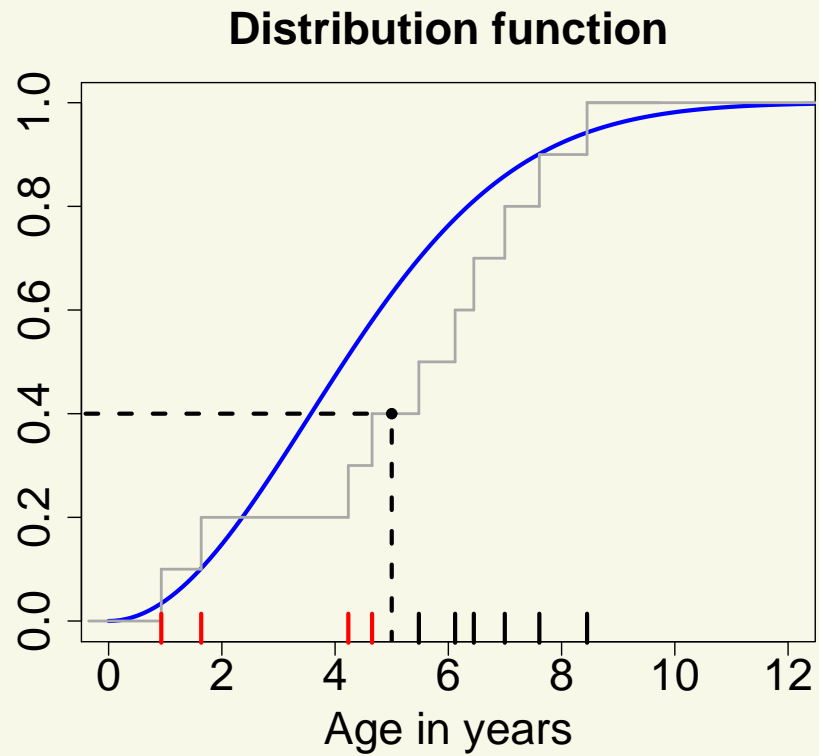
# Uncensored data

Sample size:  $n = 10$



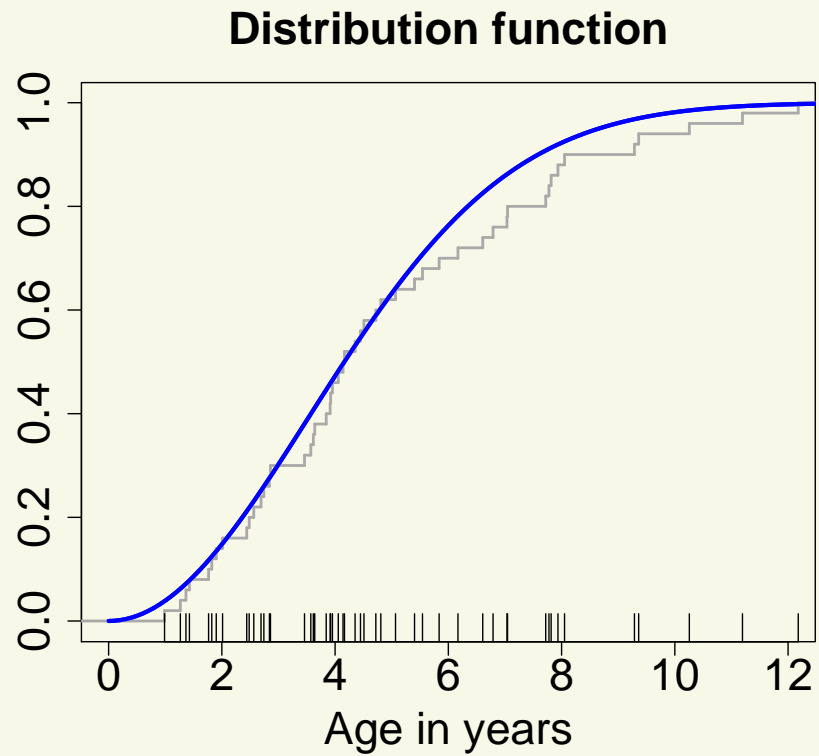
# Uncensored data

Sample size:  $n = 10$



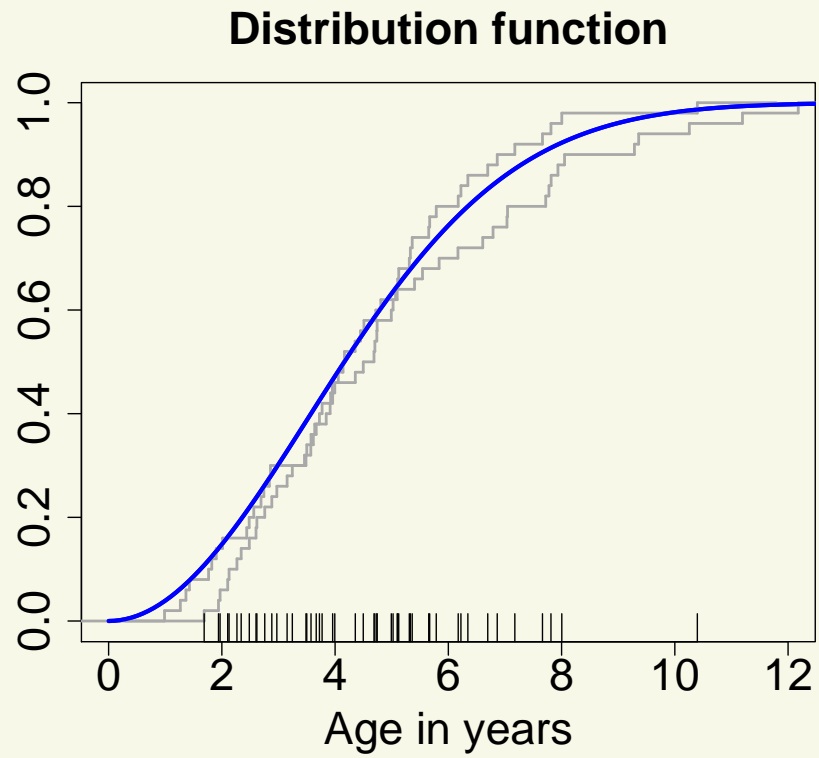
# Uncensored data

Sample size:  $n = 50$



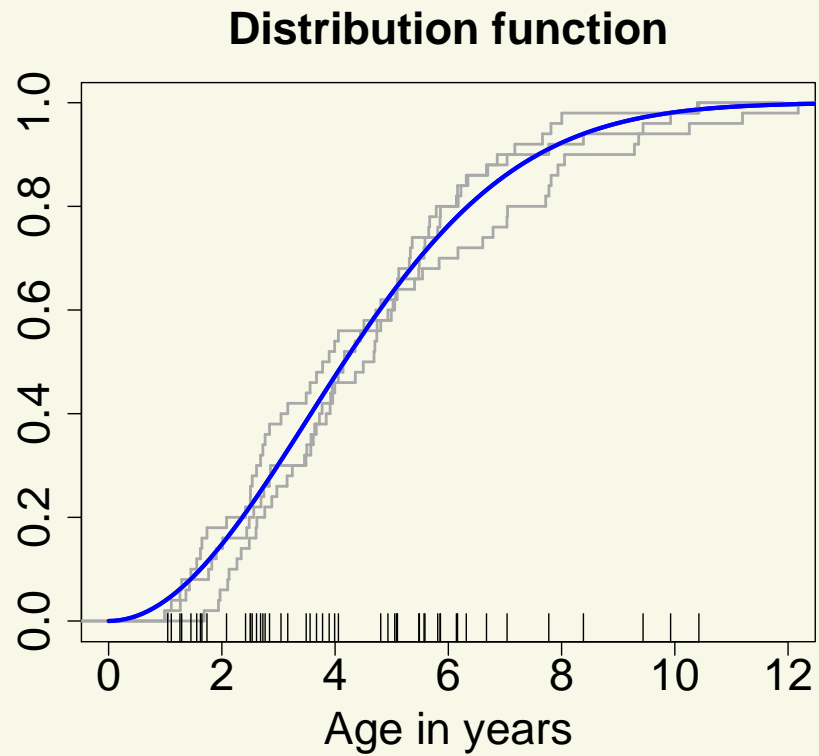
# Uncensored data

Sample size:  $n = 50$



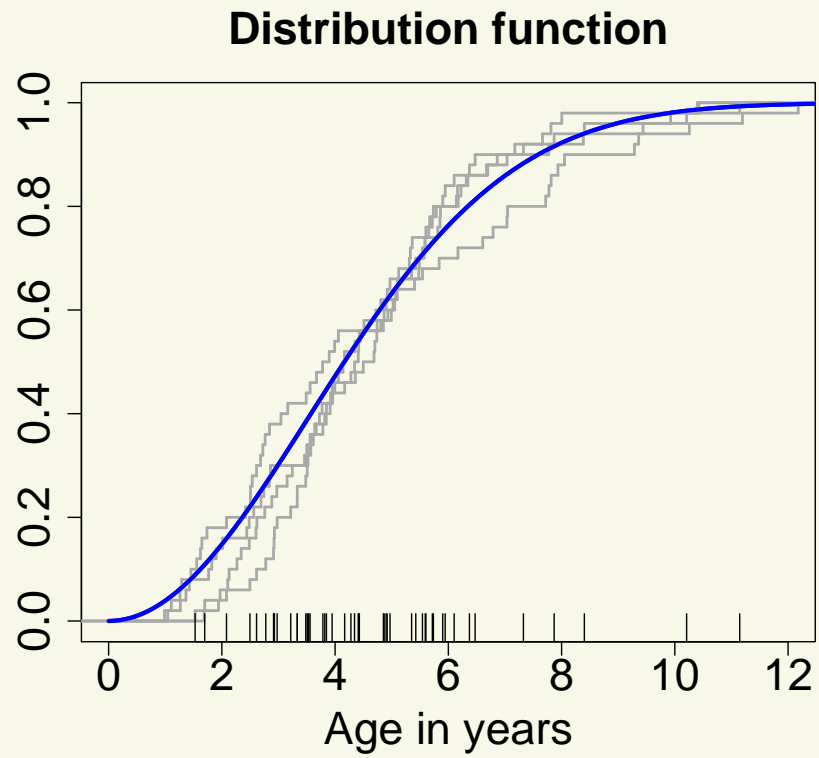
# Uncensored data

Sample size:  $n = 50$



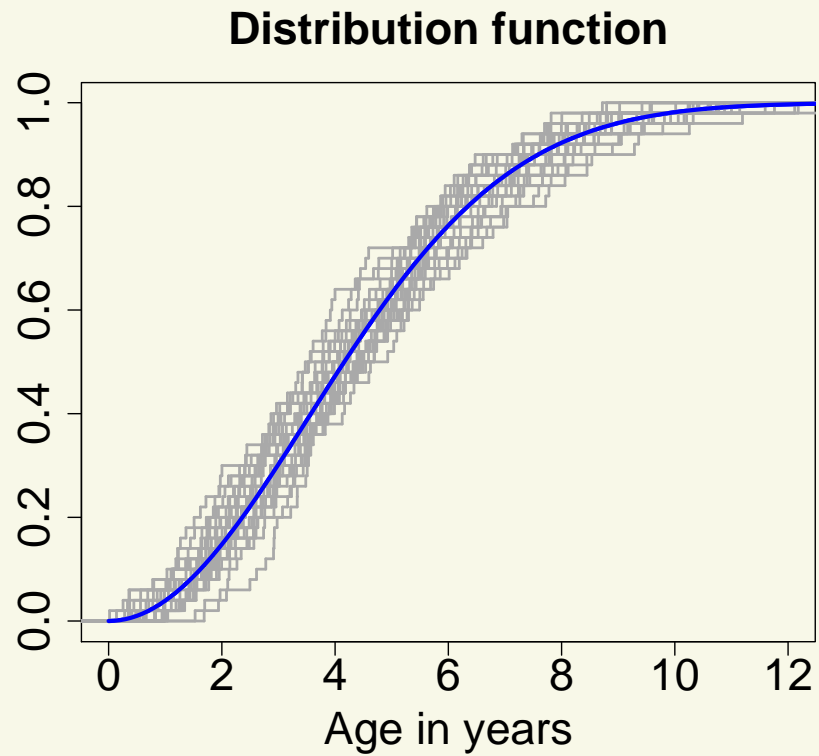
# Uncensored data

Sample size:  $n = 50$



# Uncensored data

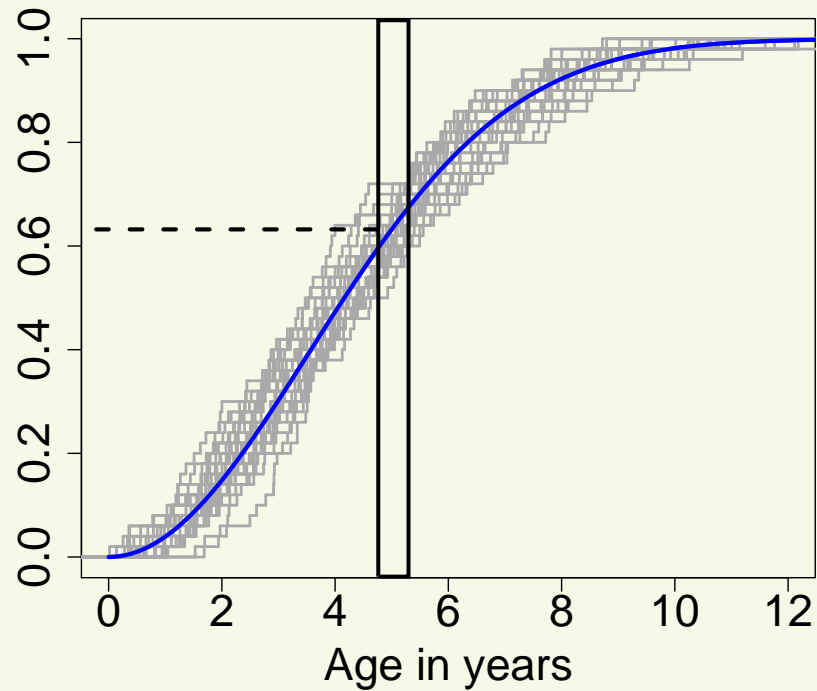
Sample size:  $n = 50$



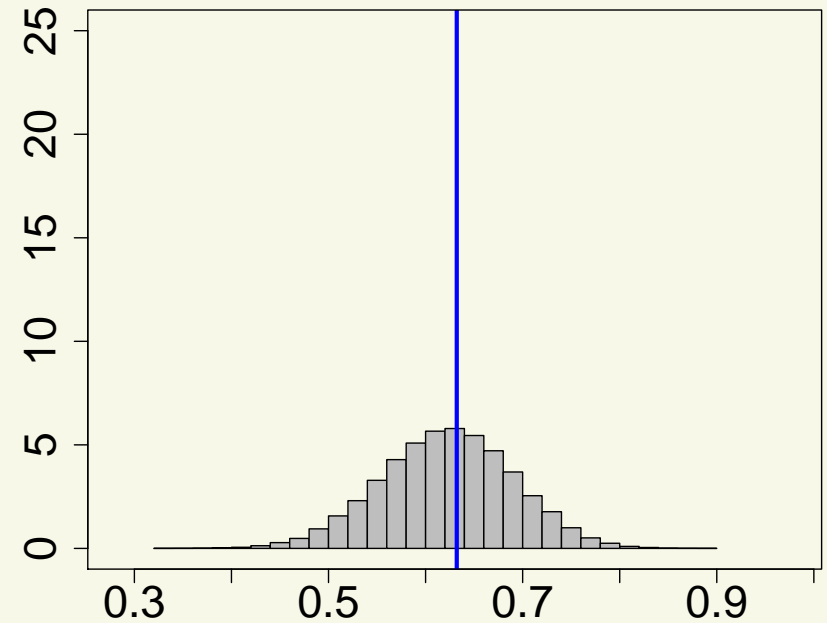
# Uncensored data

Sample size:  $n = 50$

Distribution function



Estimate at 5 years

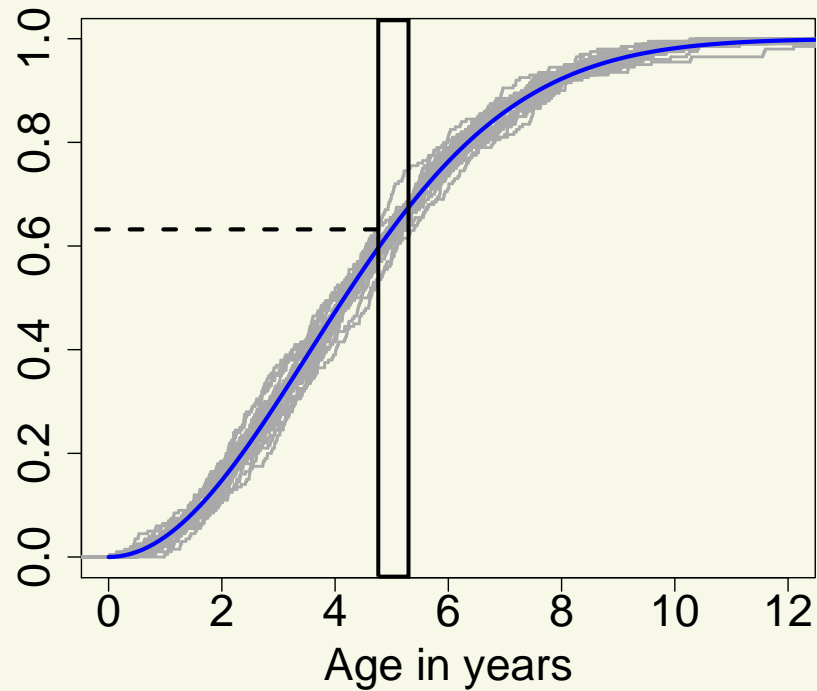


Estimate at point = population value at point +  $(1/\sqrt{n})N(0, \sigma^2)$

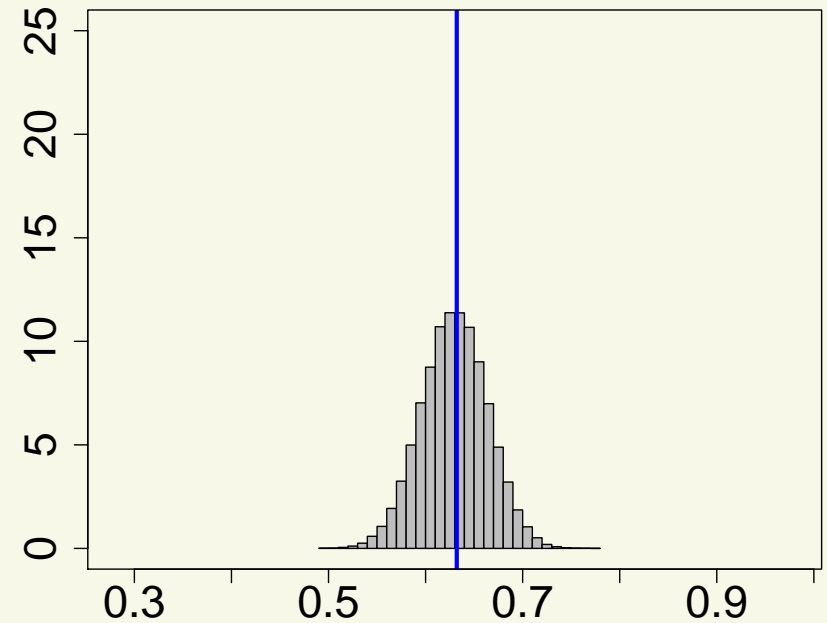
# Uncensored data

Sample size:  $n = 200$

Distribution function



Estimate at 5 years

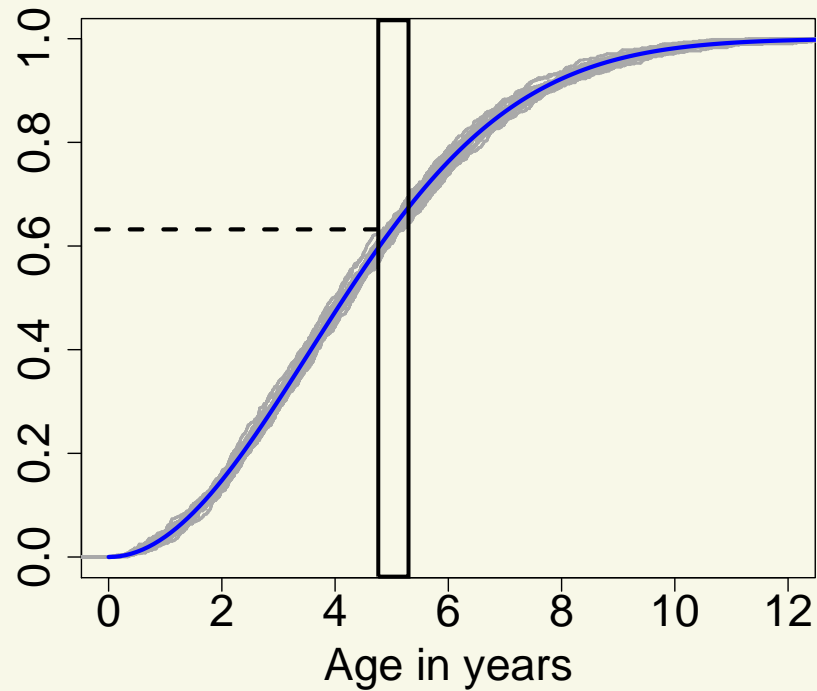


Estimate at point = population value at point +  $(1/\sqrt{n})N(0, \sigma^2)$

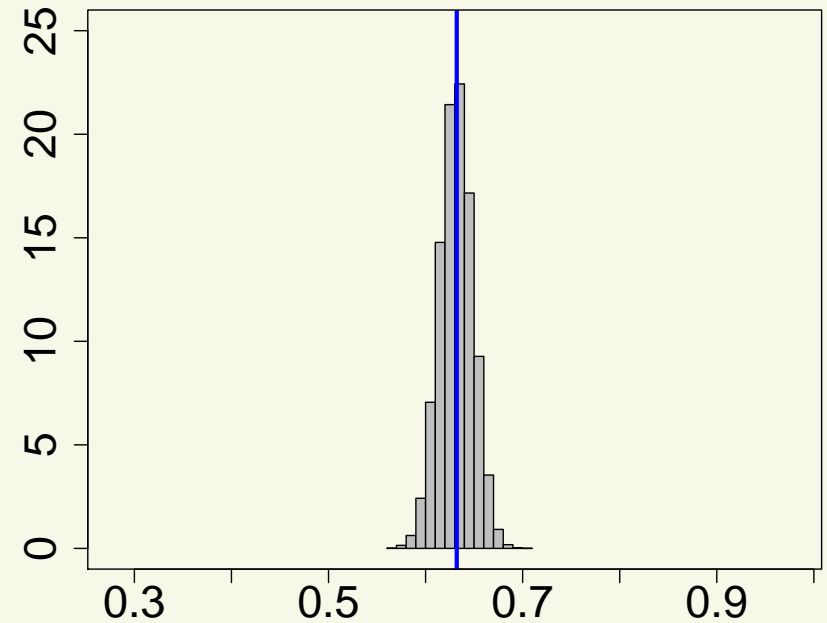
# Uncensored data

Sample size:  $n = 800$

**Distribution function**



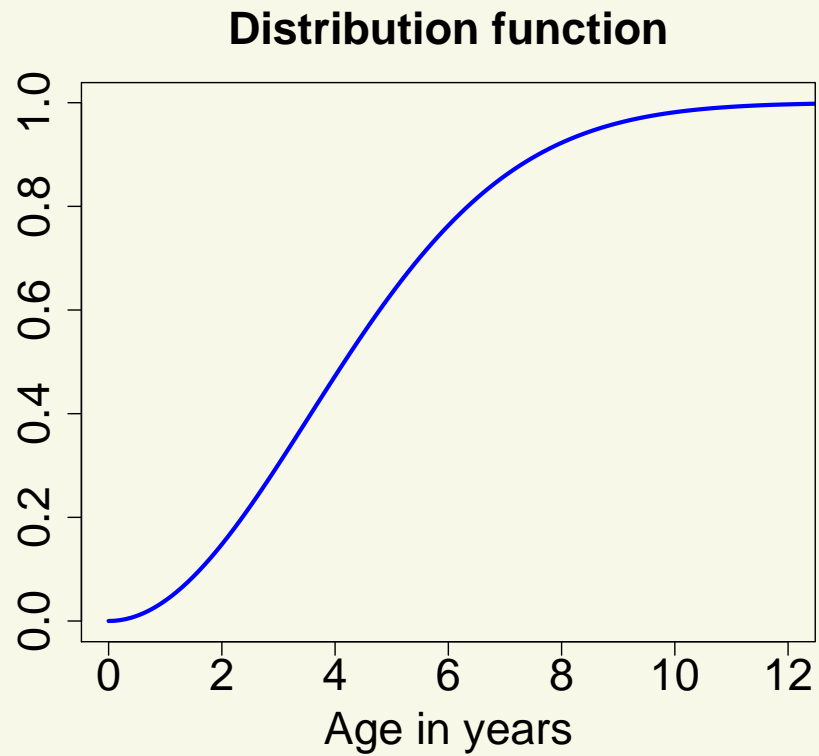
**Estimate at 5 years**



Estimate at point = population value at point +  $(1/\sqrt{n})N(0, \sigma^2)$

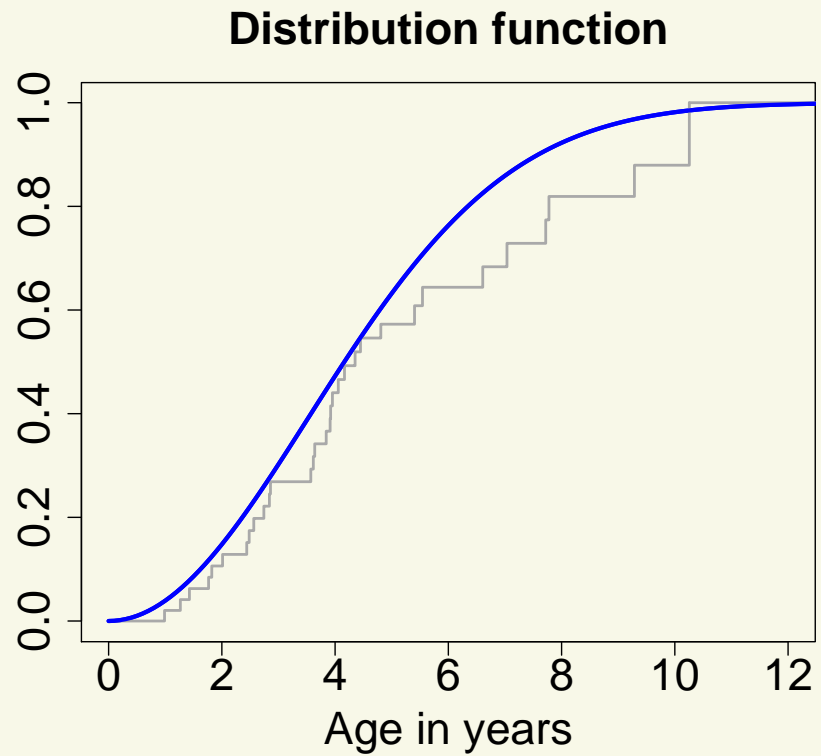
# Right censored data

Sample size:  $n = 50$



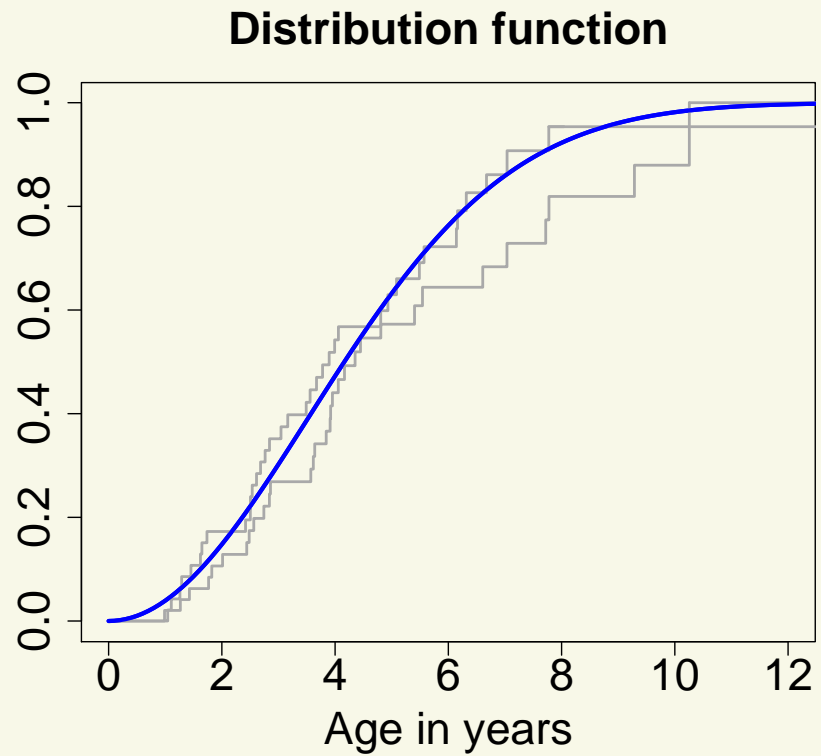
# Right censored data

Sample size:  $n = 50$



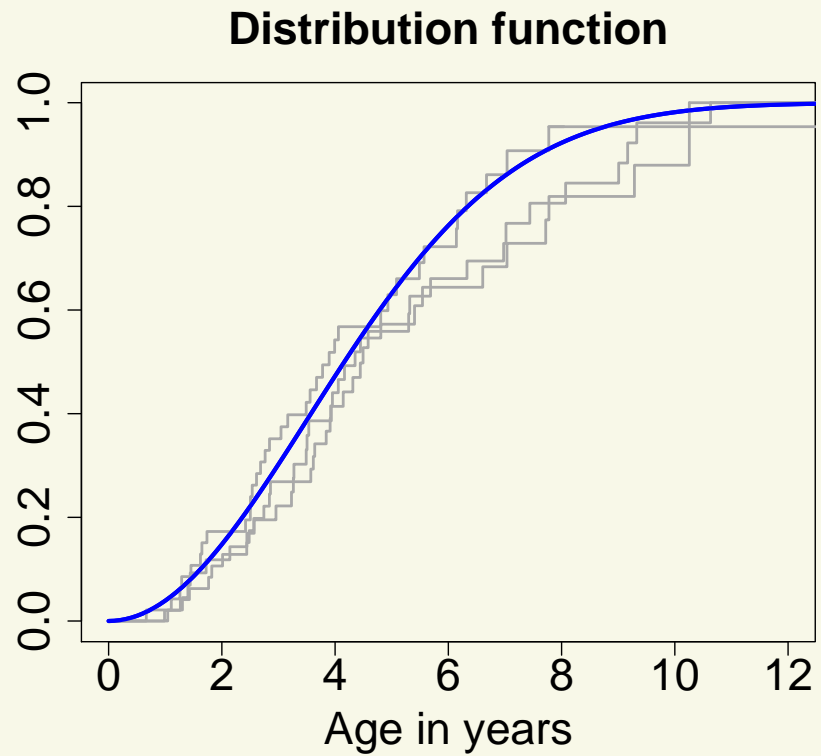
# Right censored data

Sample size:  $n = 50$



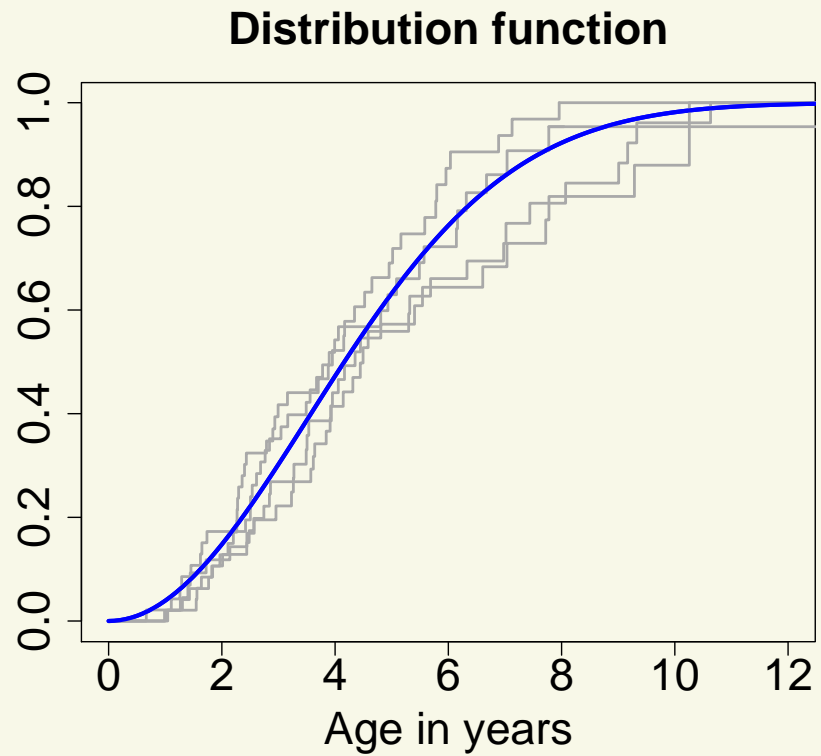
# Right censored data

Sample size:  $n = 50$



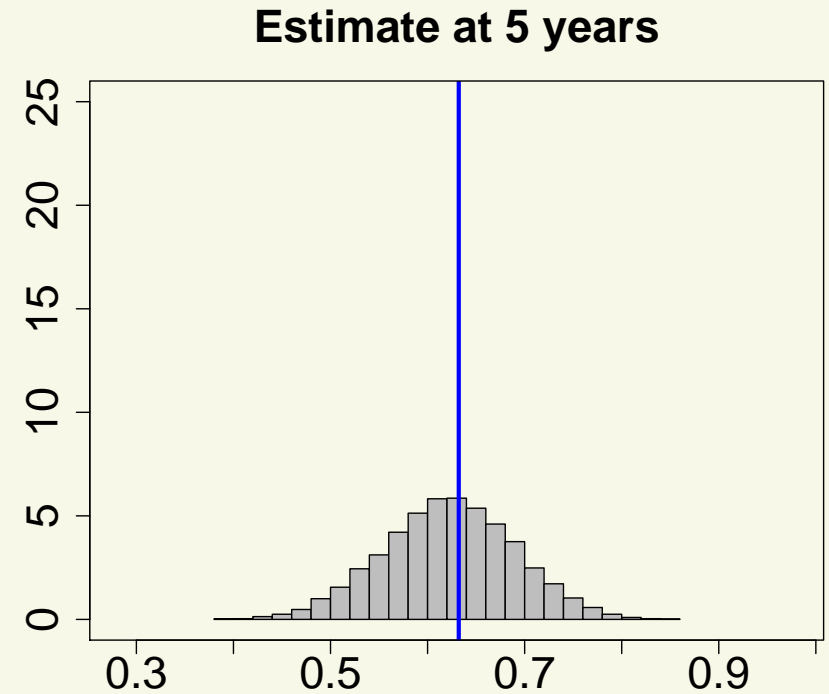
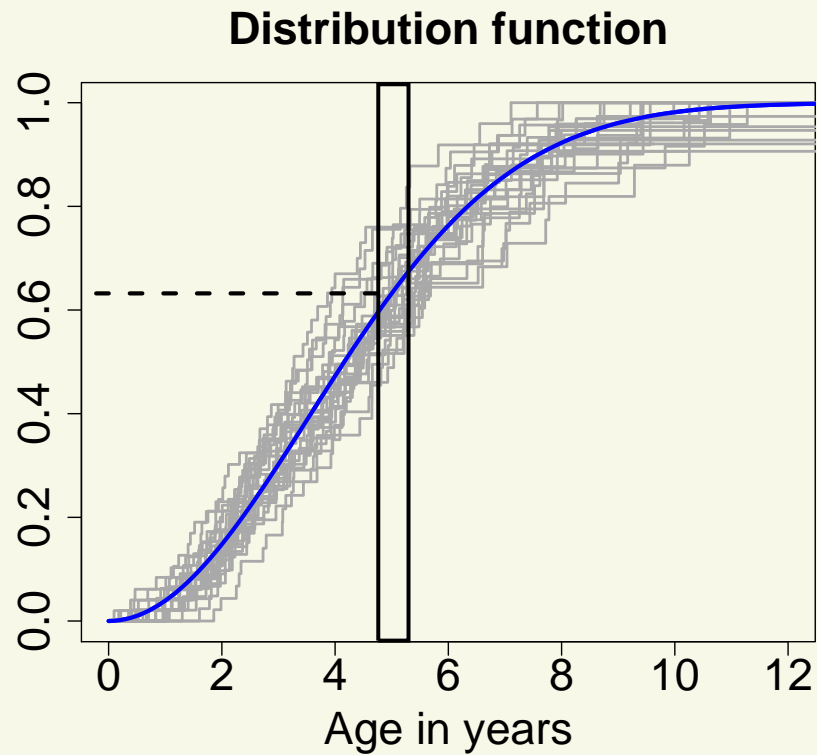
# Right censored data

Sample size:  $n = 50$



# Right censored data

Sample size:  $n = 50$

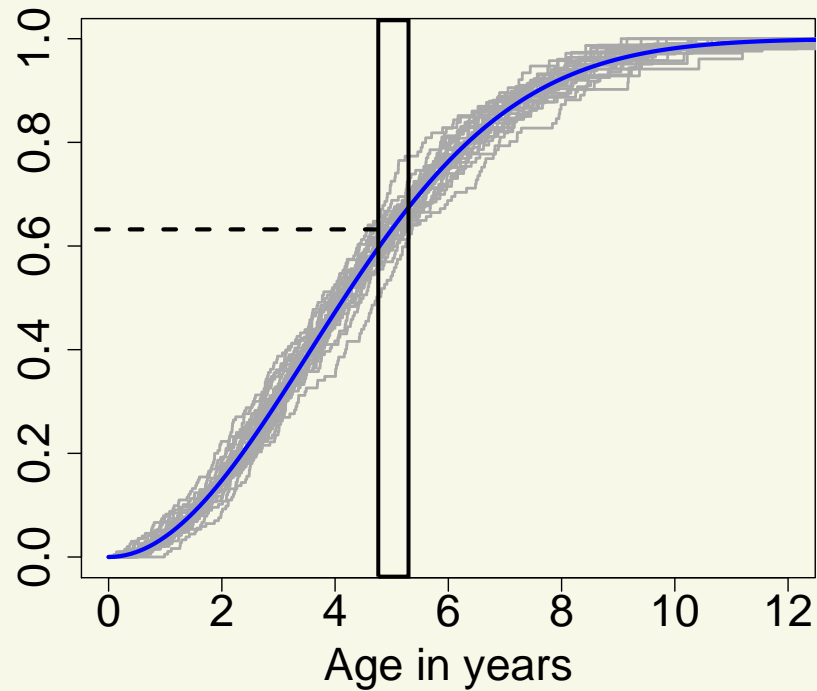


Estimate at point = population value at point +  $(1/\sqrt{n})N(0, \sigma^2)$

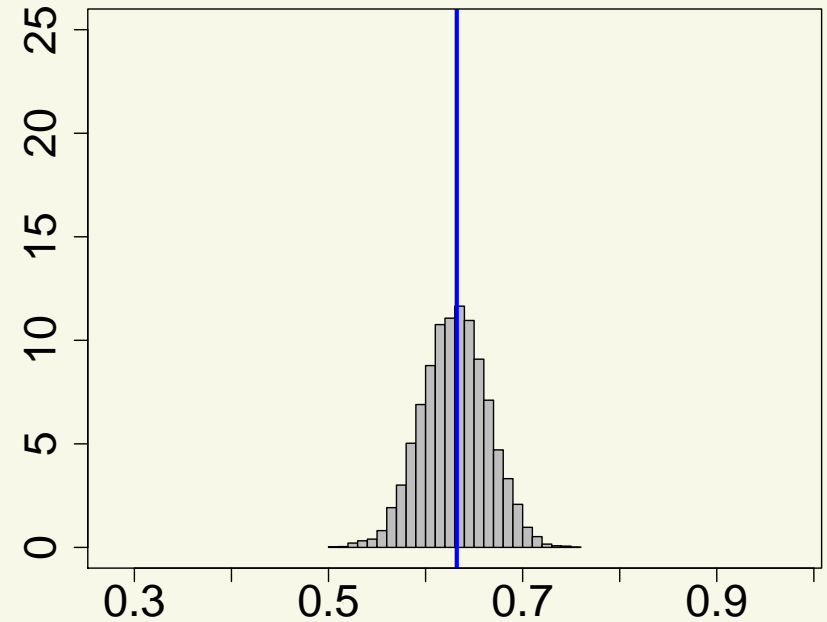
# Right censored data

Sample size:  $n = 200$

**Distribution function**



**Estimate at 5 years**

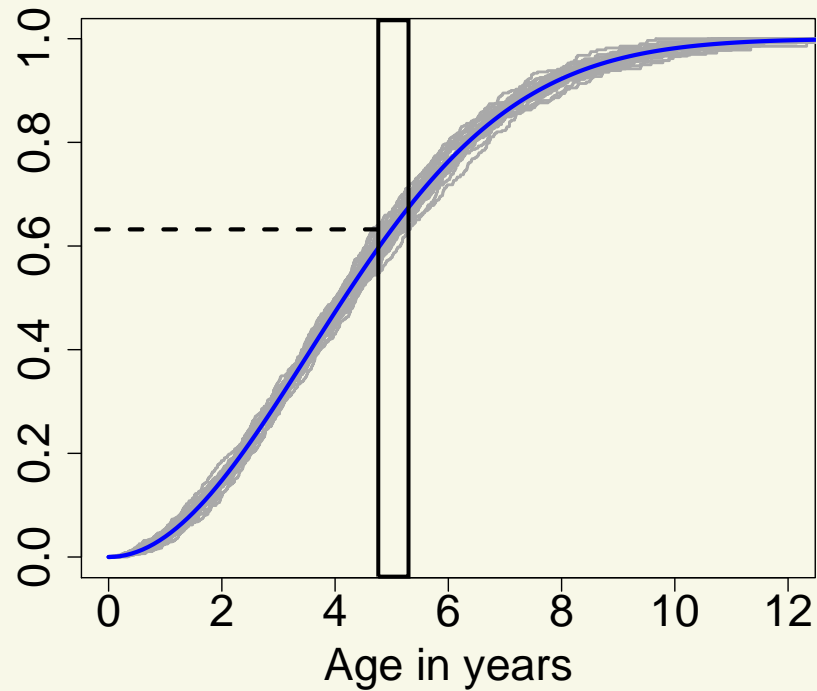


Estimate at point = population value at point +  $(1/\sqrt{n})N(0, \sigma^2)$

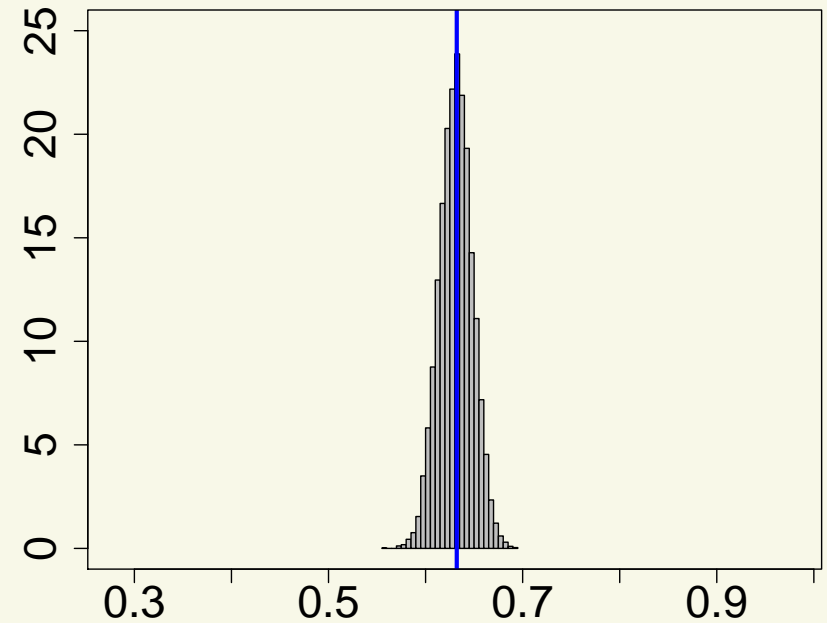
# Right censored data

Sample size:  $n = 800$

Distribution function



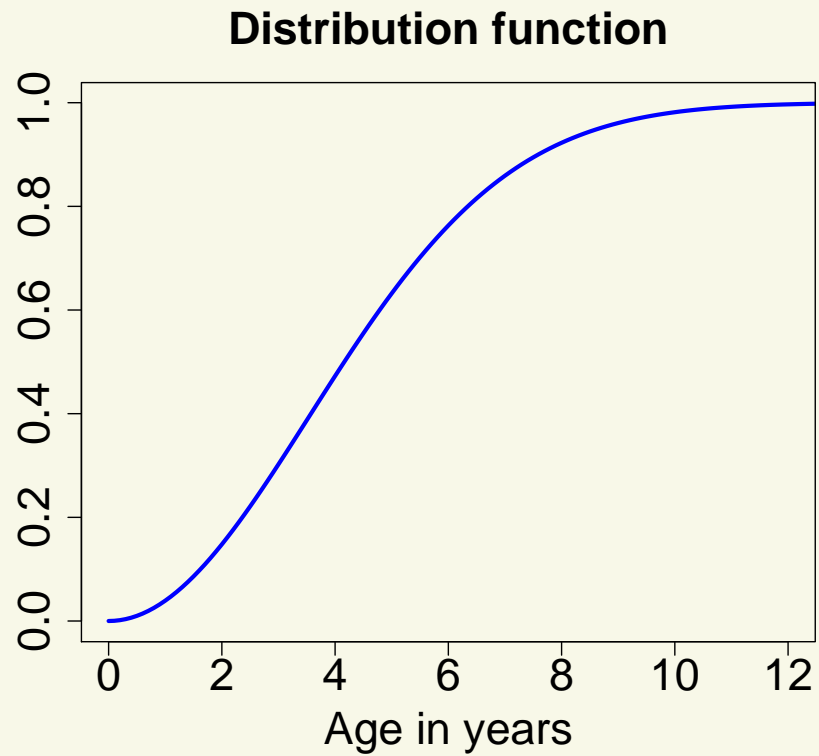
Estimate at 5 years



Estimate at point = population value at point +  $(1/\sqrt{n})N(0, \sigma^2)$

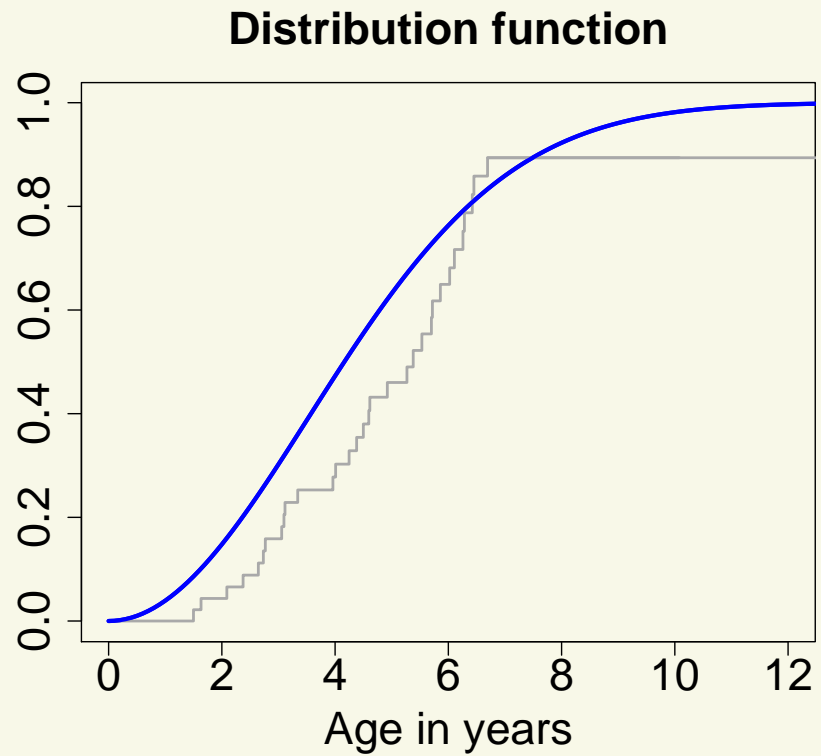
# Interval censored data, imputed

Sample size:  $n = 50$



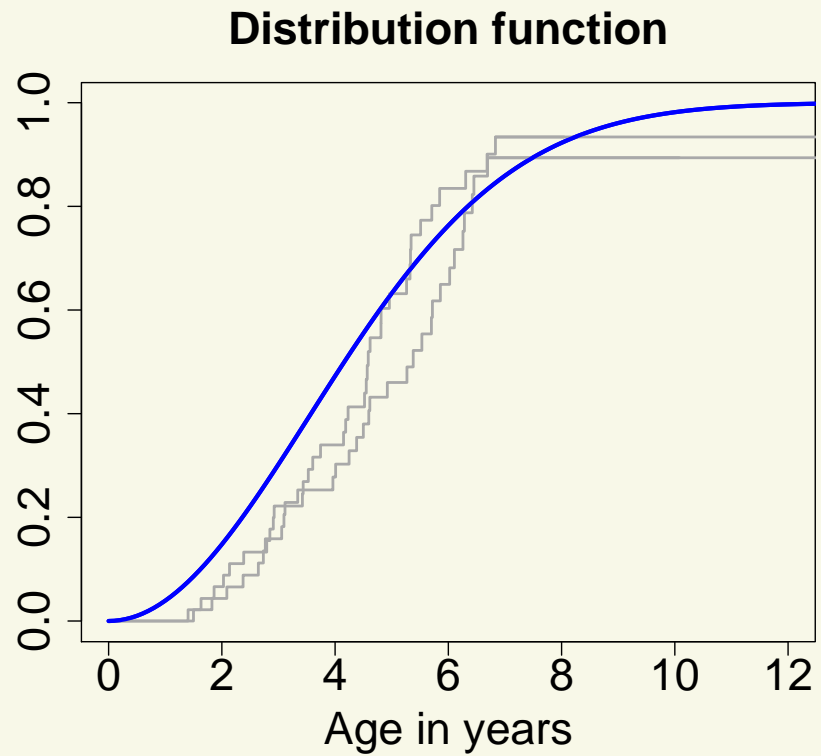
# Interval censored data, imputed

Sample size:  $n = 50$



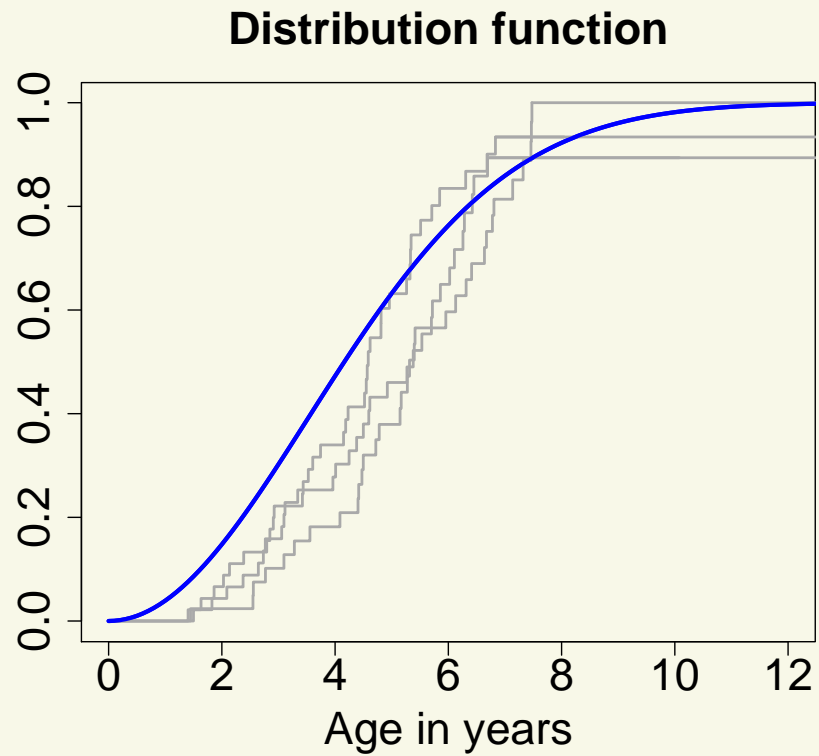
# Interval censored data, imputed

Sample size:  $n = 50$



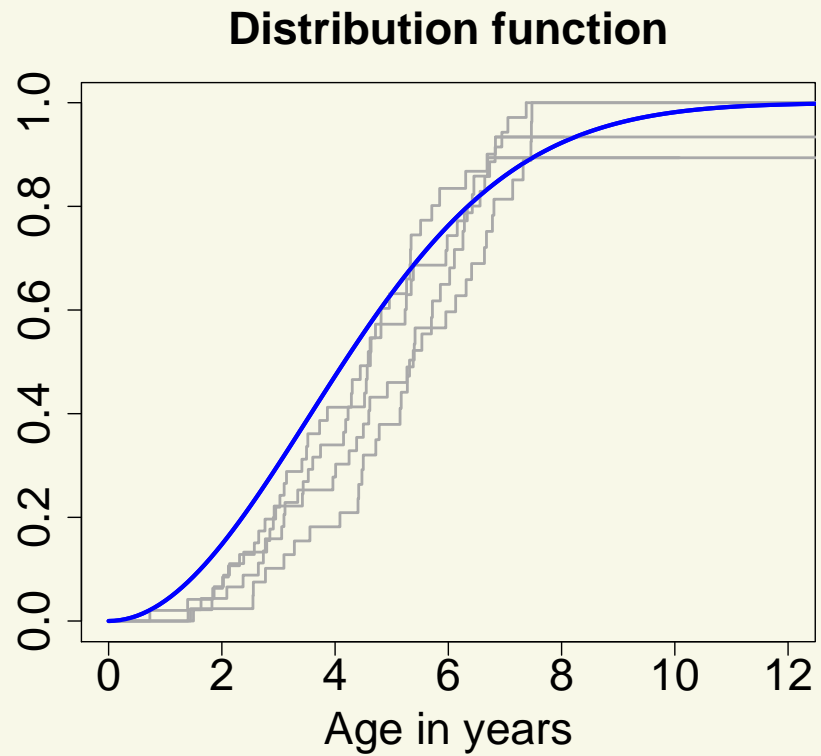
# Interval censored data, imputed

Sample size:  $n = 50$



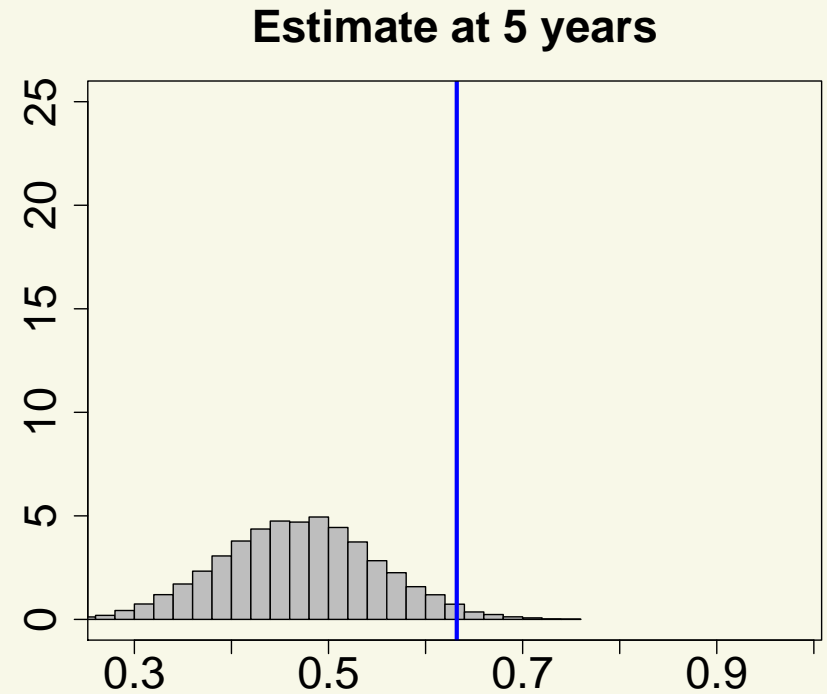
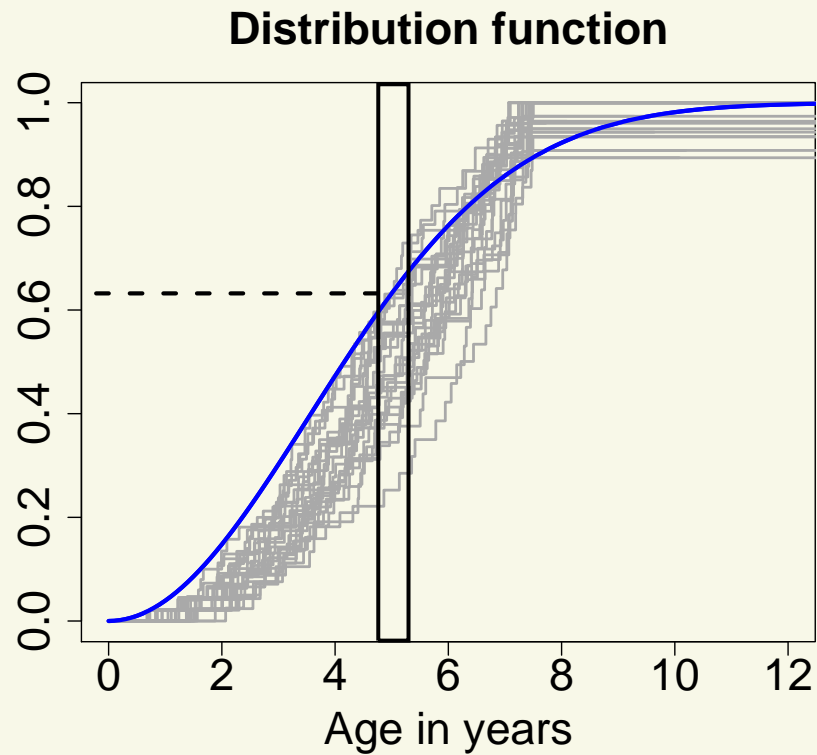
# Interval censored data, imputed

Sample size:  $n = 50$



# Interval censored data, imputed

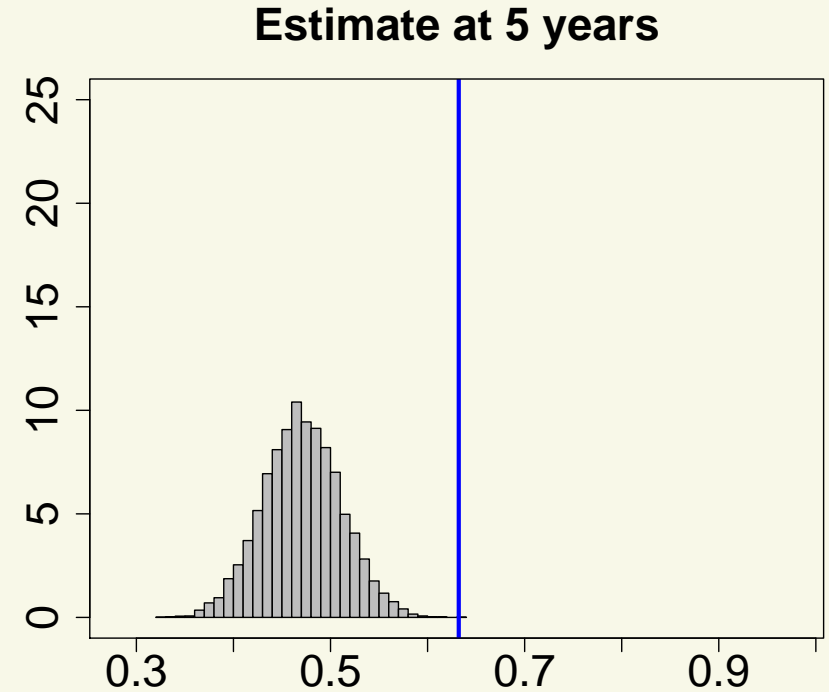
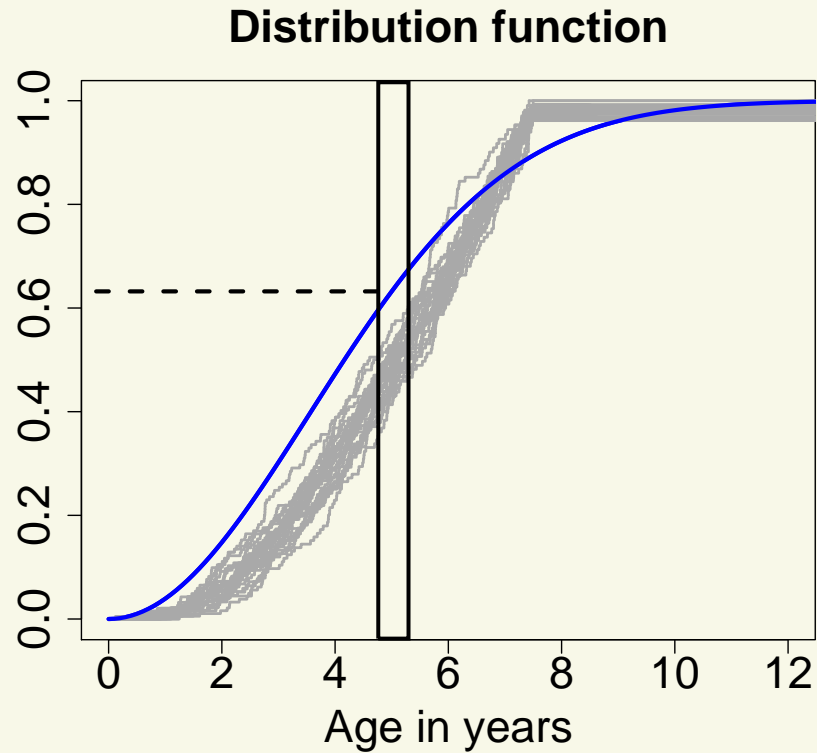
Sample size:  $n = 50$



Estimate at point = population value at point + **bias** +  $(1/\sqrt{n})N(0, \sigma^2)$

# Interval censored data, imputed

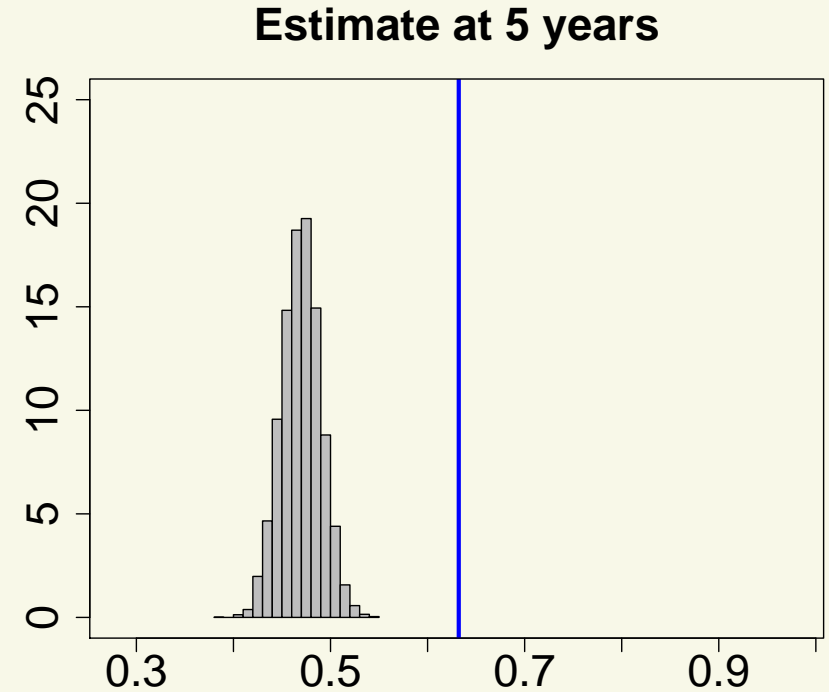
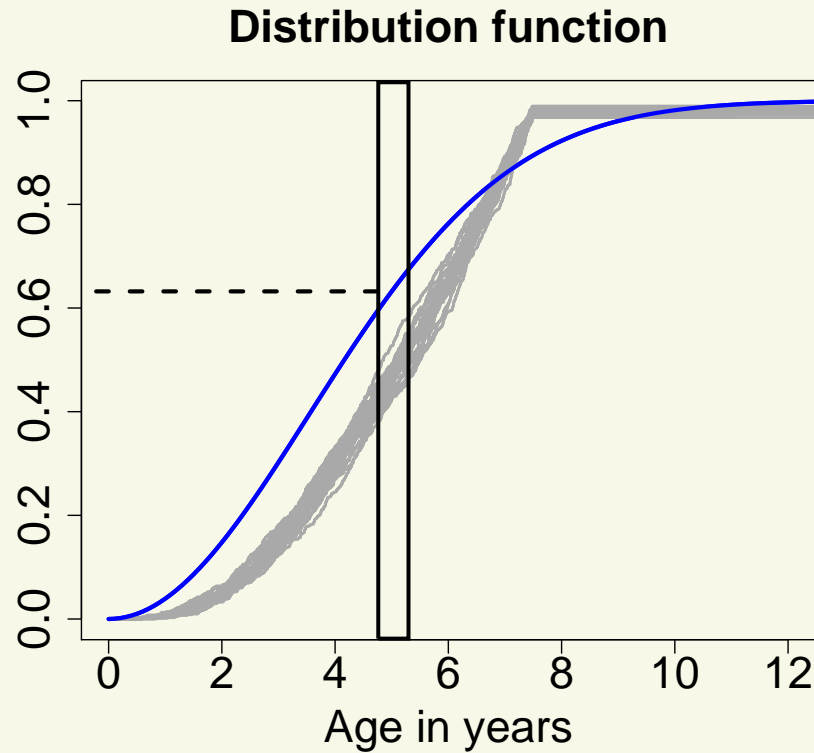
Sample size:  $n = 200$



Estimate at point = population value at point + **bias** +  $(1/\sqrt{n})N(0, \sigma^2)$

# Interval censored data, imputed

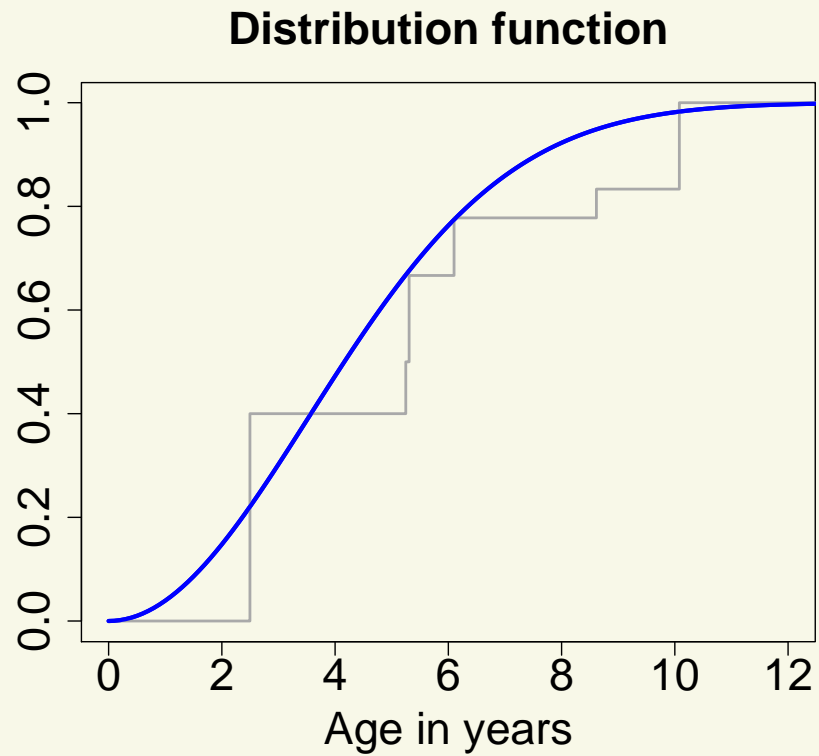
Sample size:  $n = 800$



Estimate at point = population value at point + **bias** +  $(1/\sqrt{n})N(0, \sigma^2)$

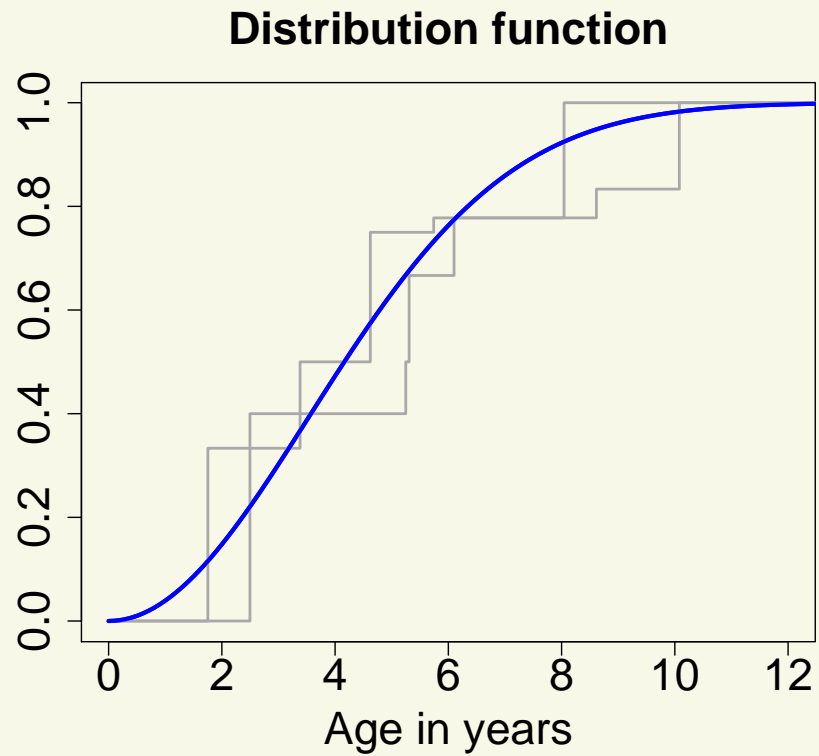
# Interval censored data, MLE

Sample size:  $n = 50$



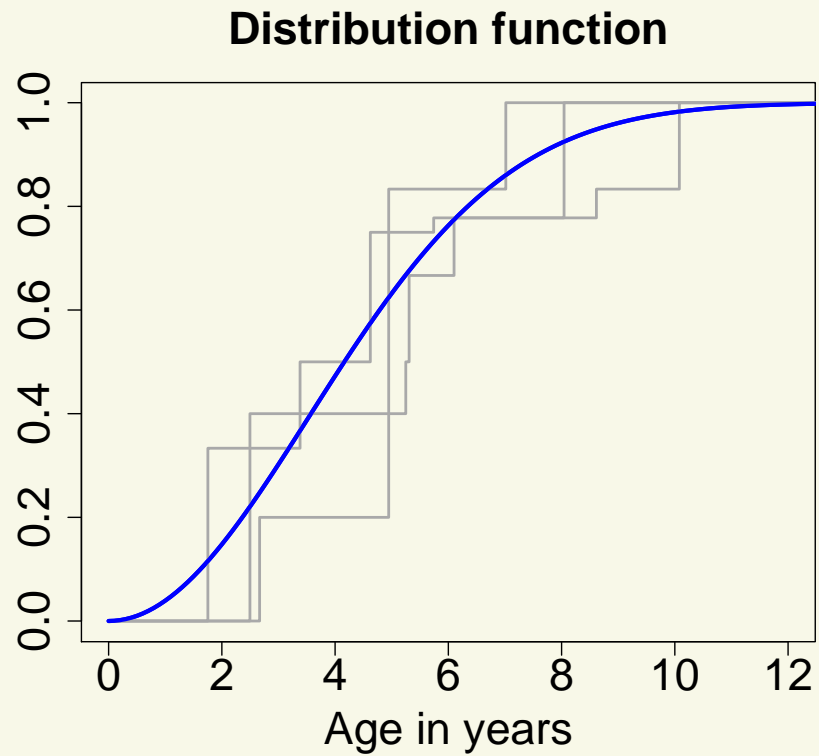
# Interval censored data, MLE

Sample size:  $n = 50$



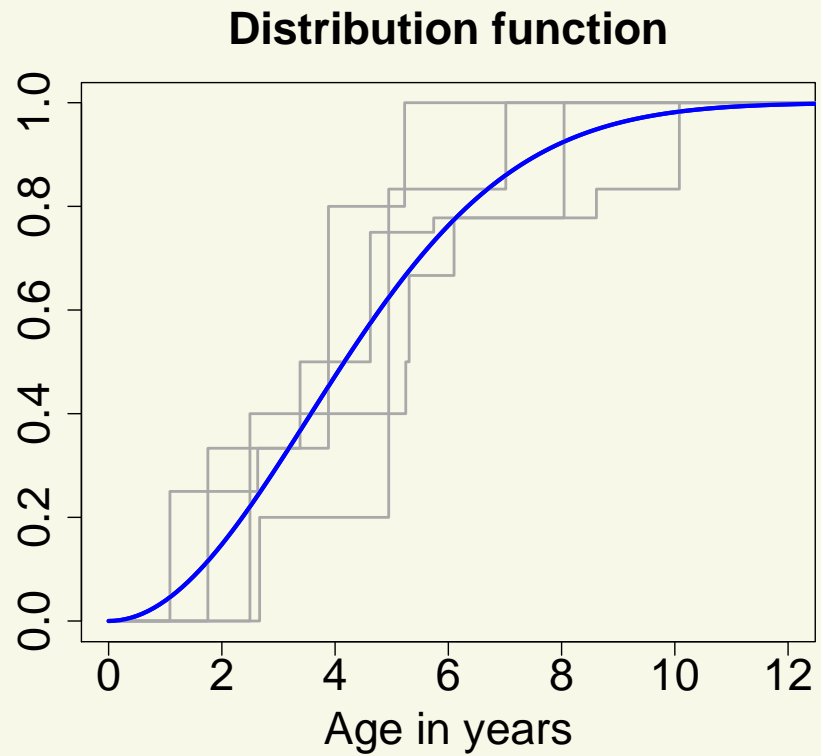
# Interval censored data, MLE

Sample size:  $n = 50$



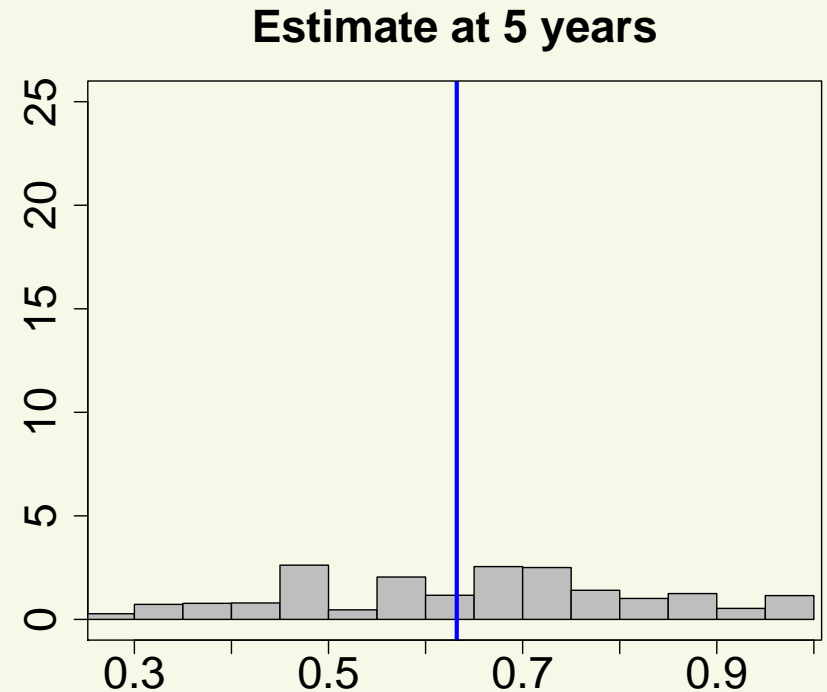
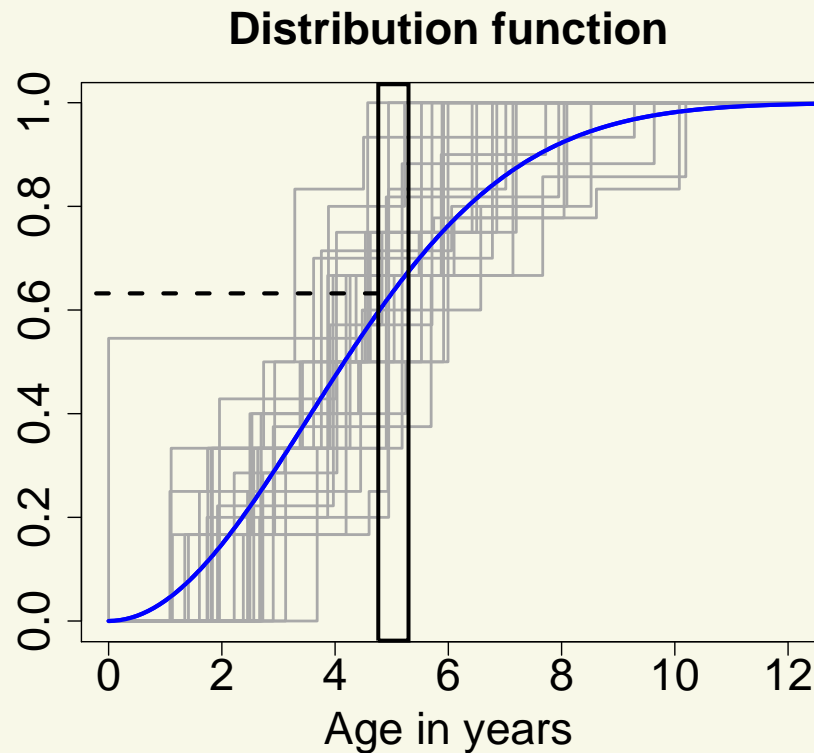
# Interval censored data, MLE

Sample size:  $n = 50$



# Interval censored data, MLE

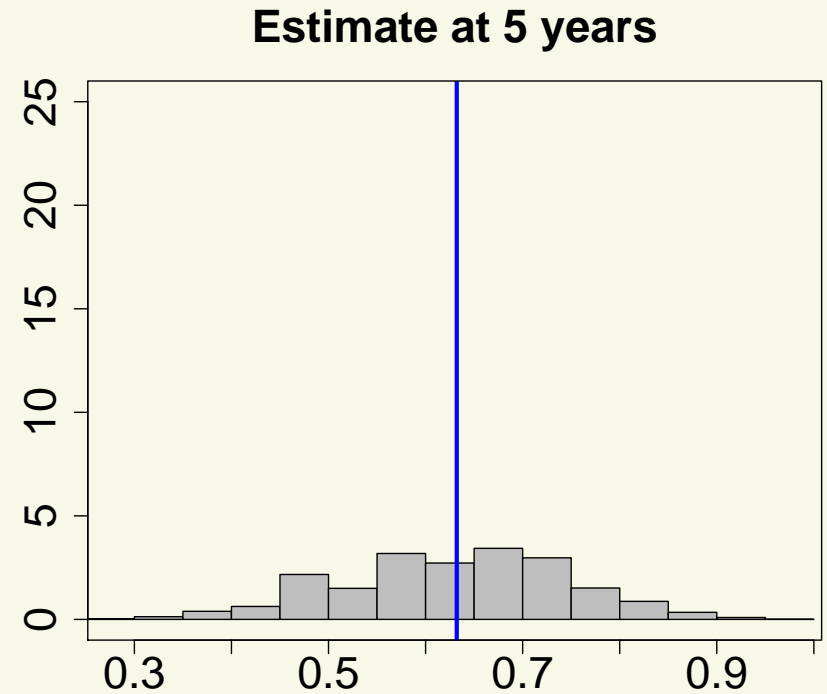
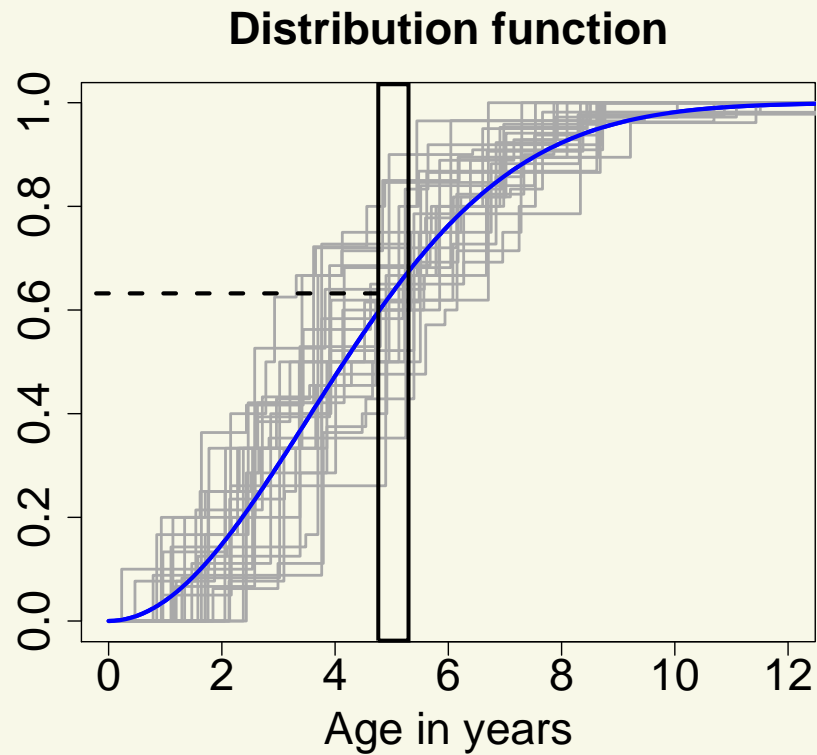
Sample size:  $n = 50$



Estimate at point = population value at point +  $(1/\sqrt[3]{n})$  Chernoff's Distr.

# Interval censored data, MLE

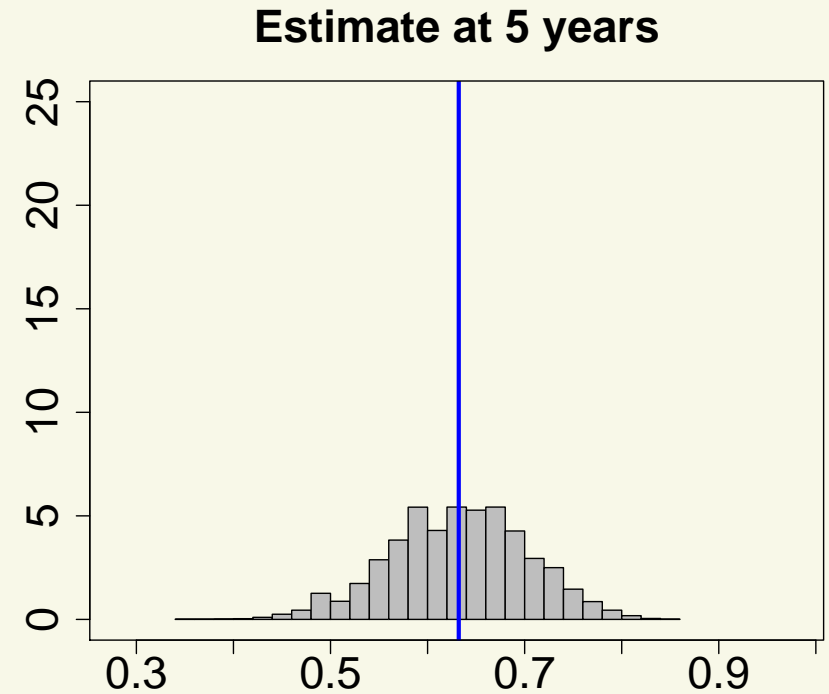
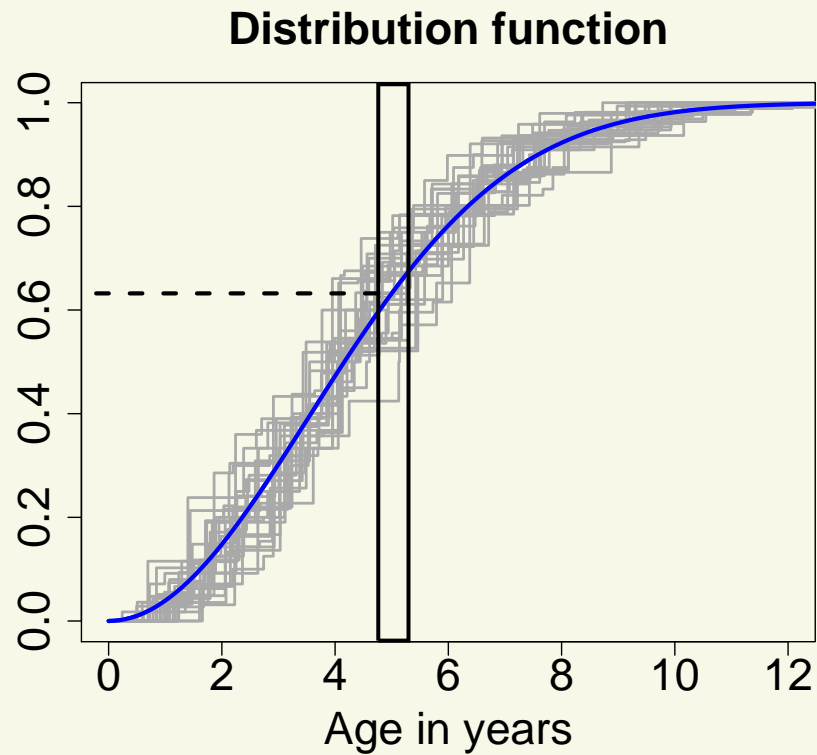
Sample size:  $n = 200$



Estimate at point = population value at point +  $(1/\sqrt[3]{n})$  Chernoff's Distr.

# Interval censored data, MLE

Sample size:  $n = 800$

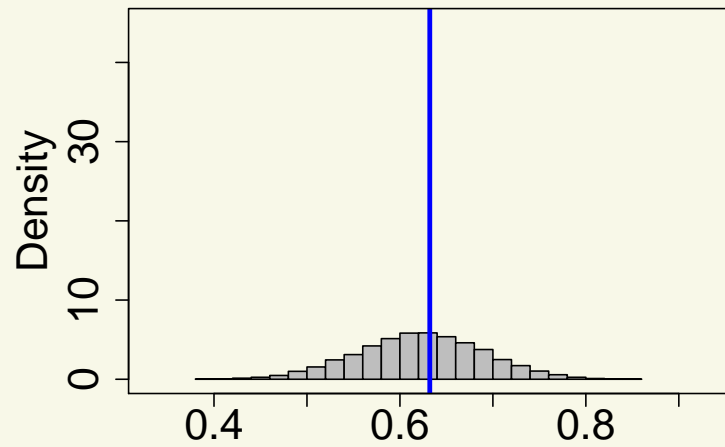


Estimate at point = population value at point +  $(1/\sqrt[3]{n})$  Chernoff's Distr.

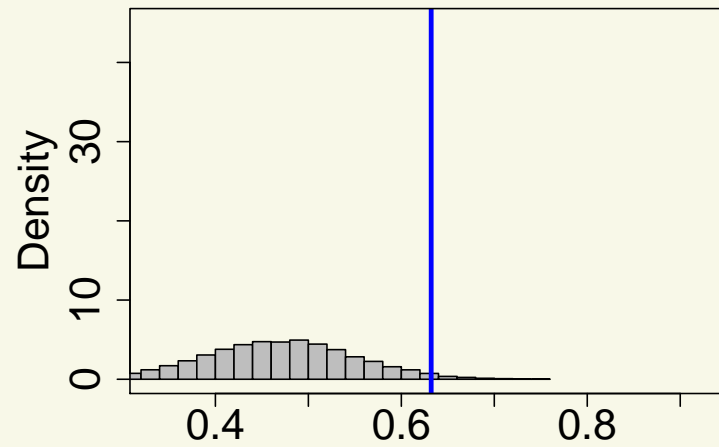
# Simulation: estimation of distribution function at a point

Sample size:  $n = 50$

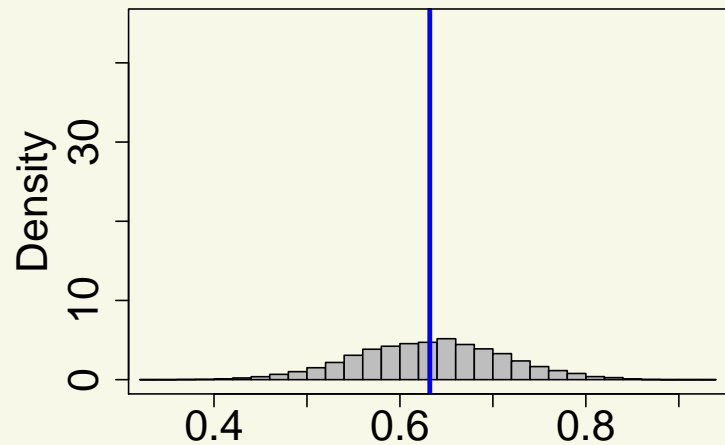
**Uncensored**



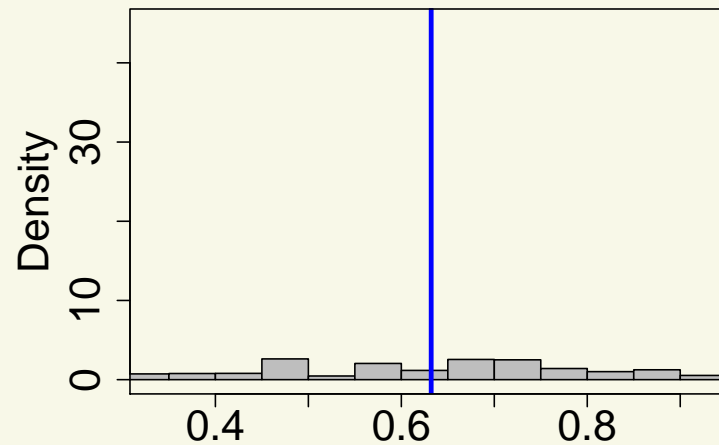
**Interval censored, imputation**



**Right censored**



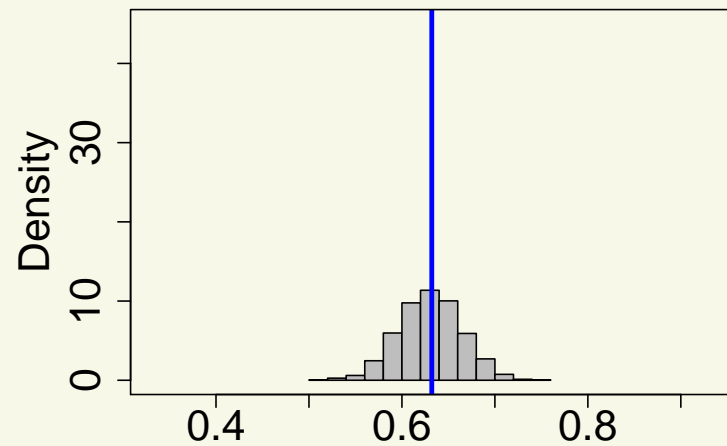
**Interval censored, MLE**



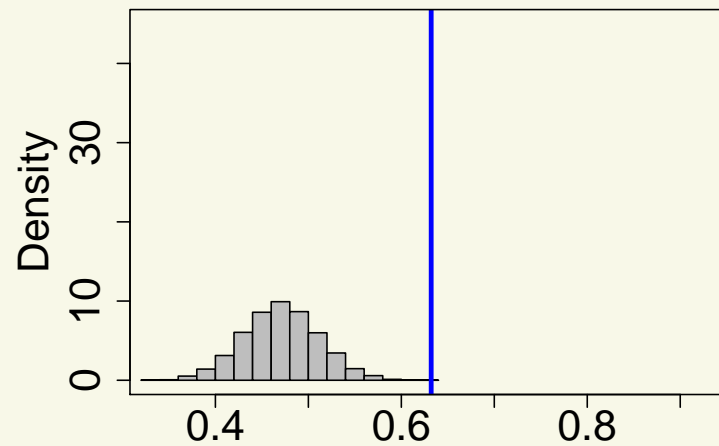
# Simulation: estimation of distribution function at a point

Sample size:  $n = 200$

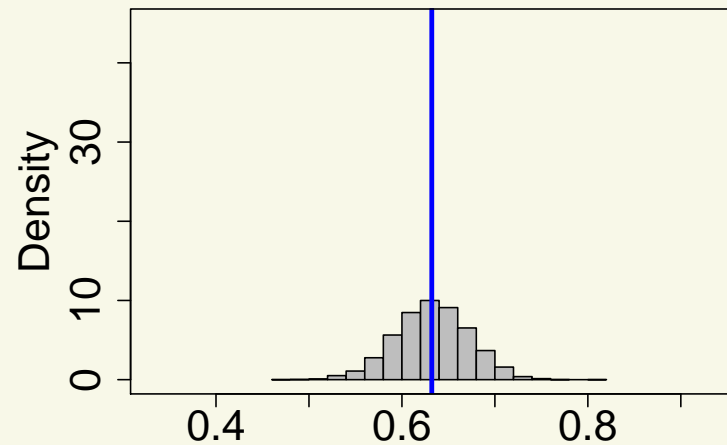
**Uncensored**



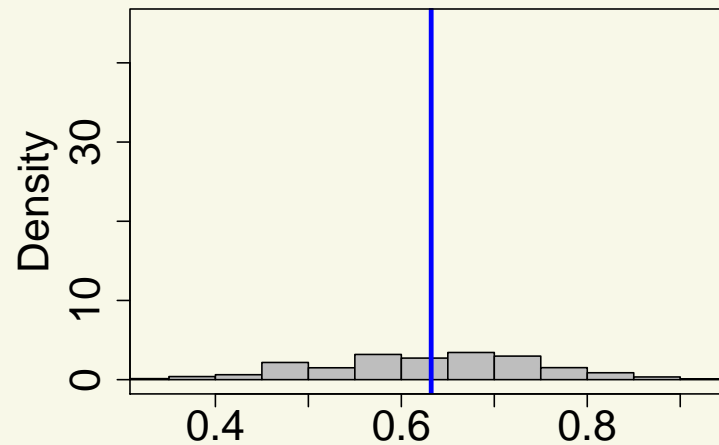
**Interval censored, imputation**



**Right censored**



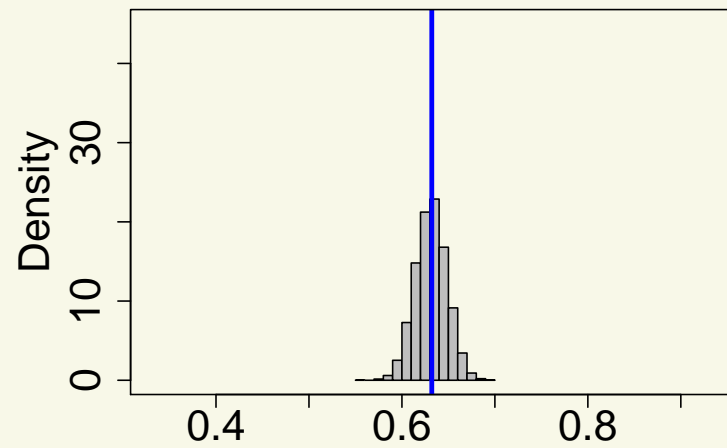
**Interval censored, MLE**



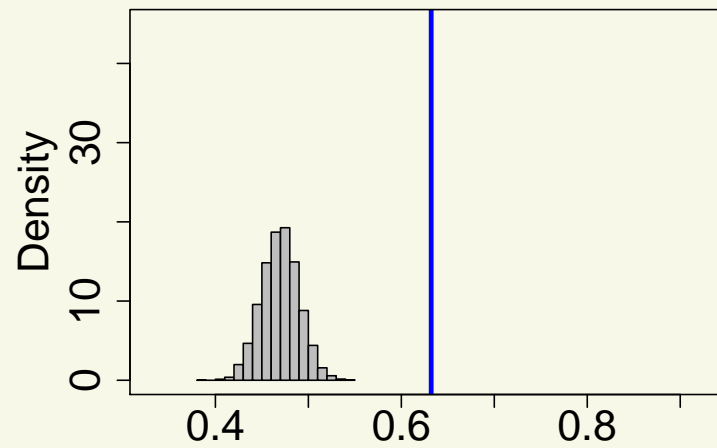
# Simulation: estimation of distribution function at a point

Sample size:  $n = 800$

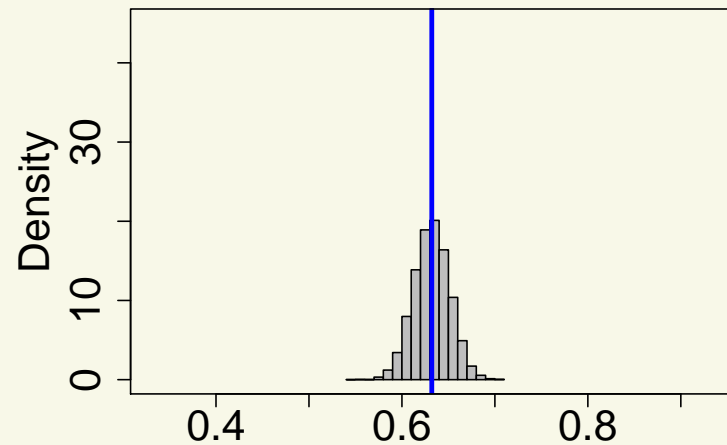
**Uncensored**



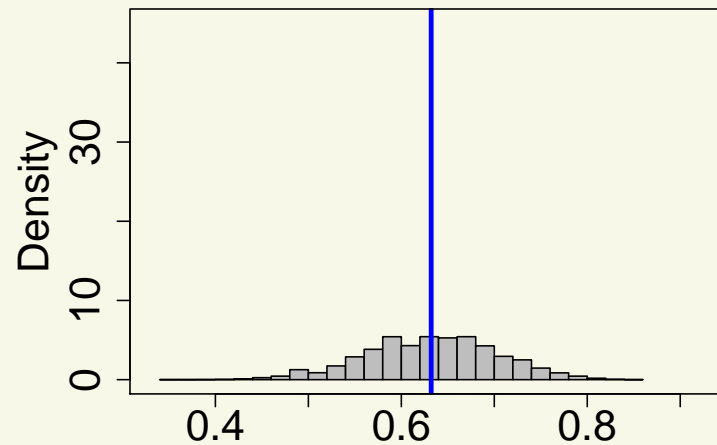
**Interval censored, imputation**



**Right censored**



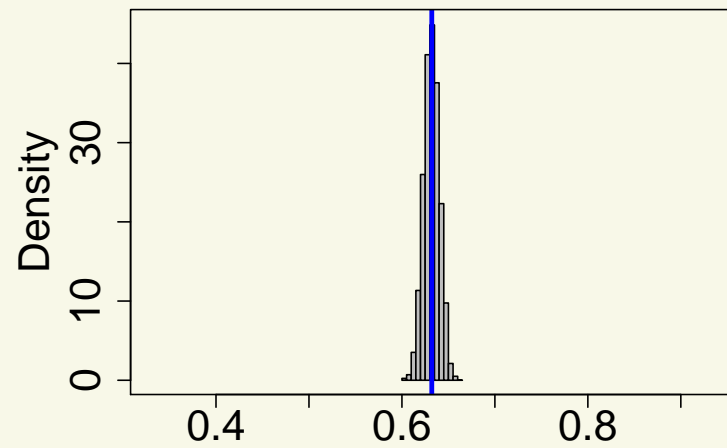
**Interval censored, MLE**



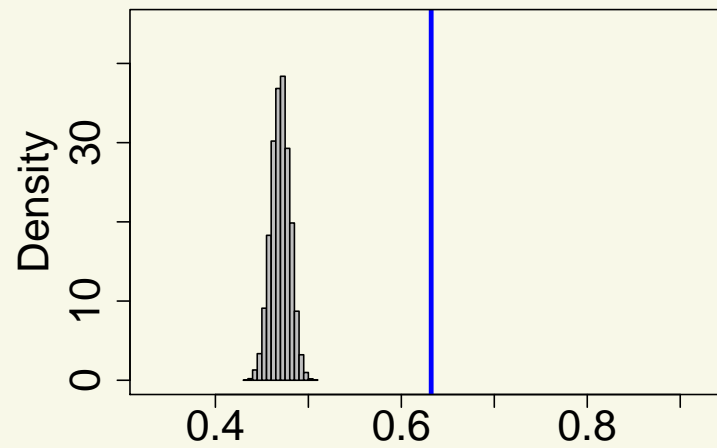
# Simulation: estimation of distribution function at a point

Sample size:  $n = 3200$

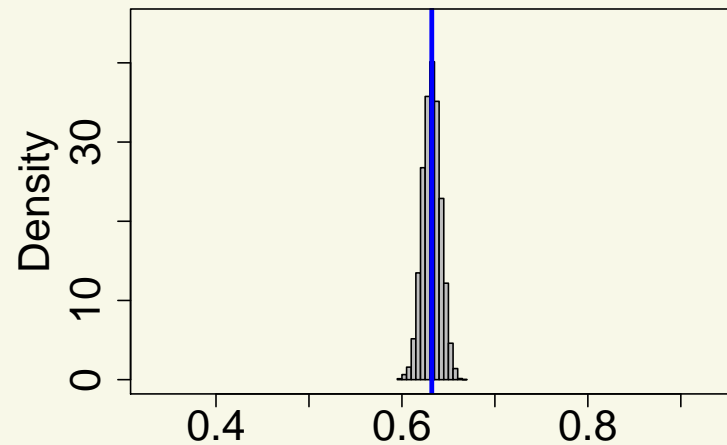
**Uncensored**



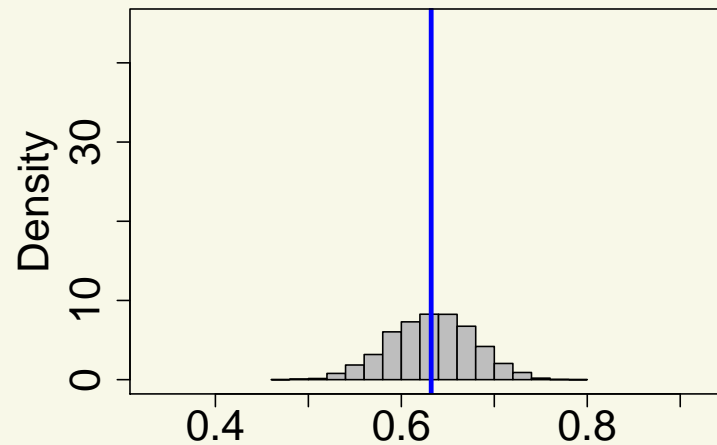
**Interval censored, imputation**



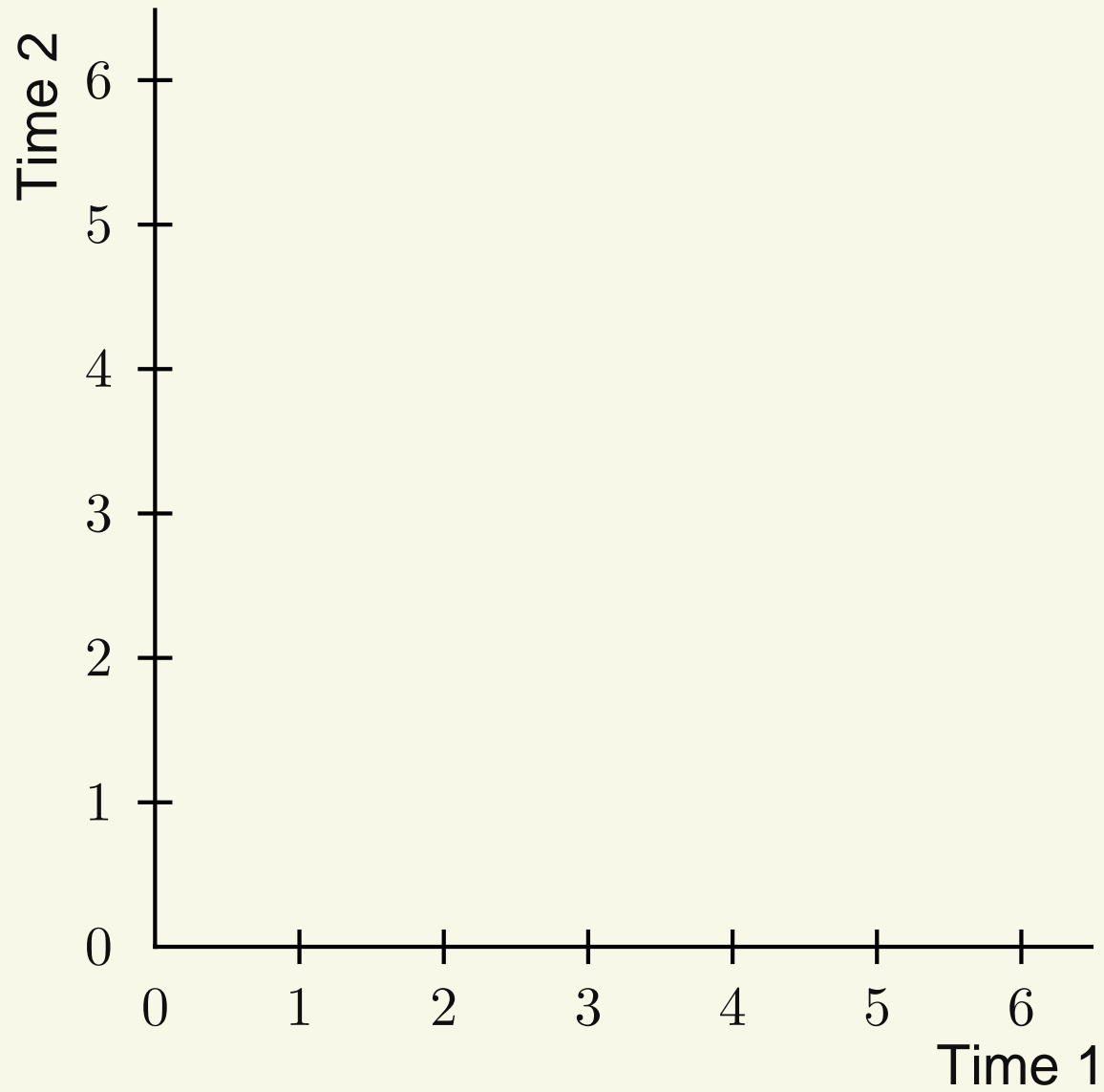
**Right censored**



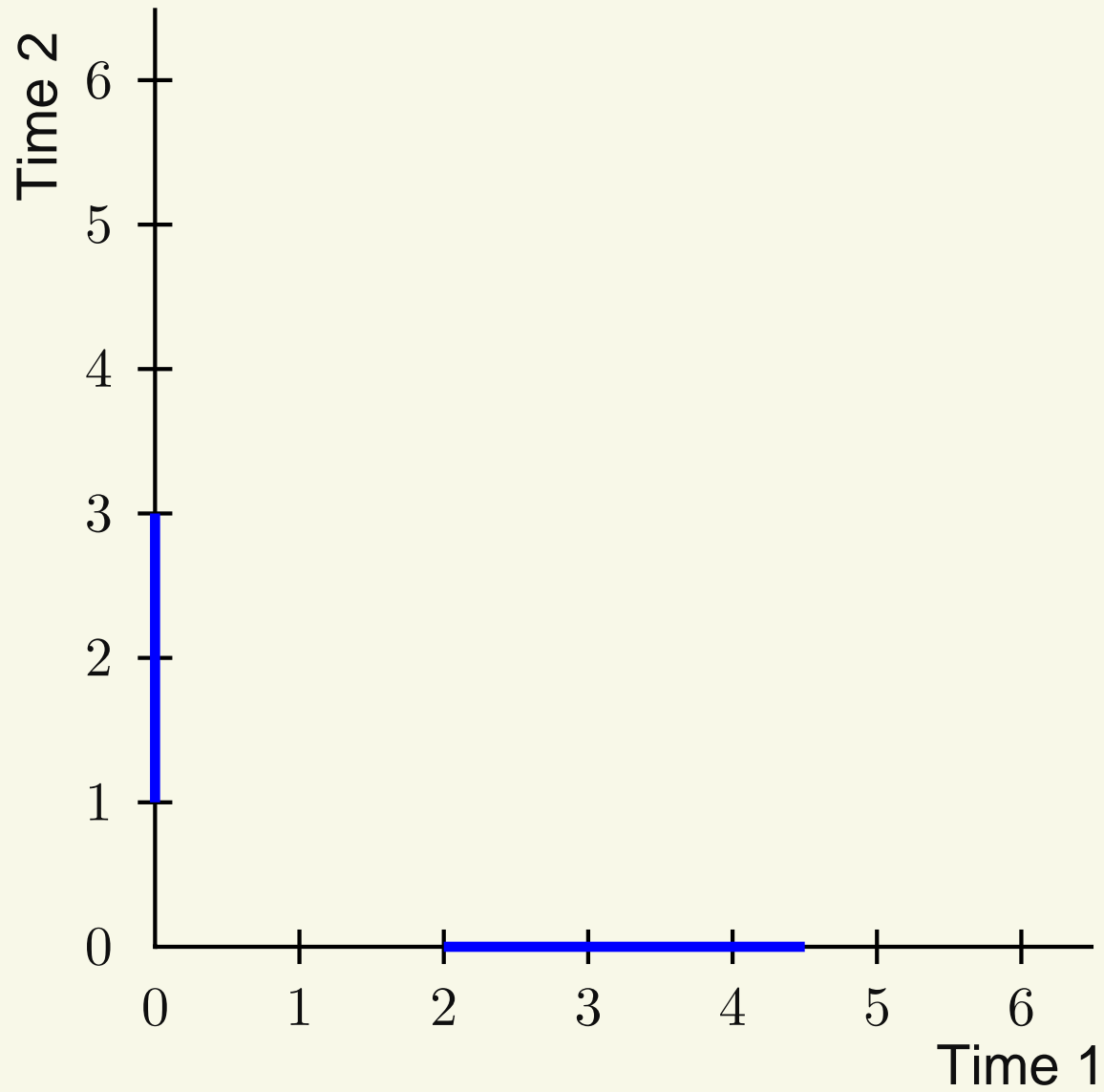
**Interval censored, MLE**



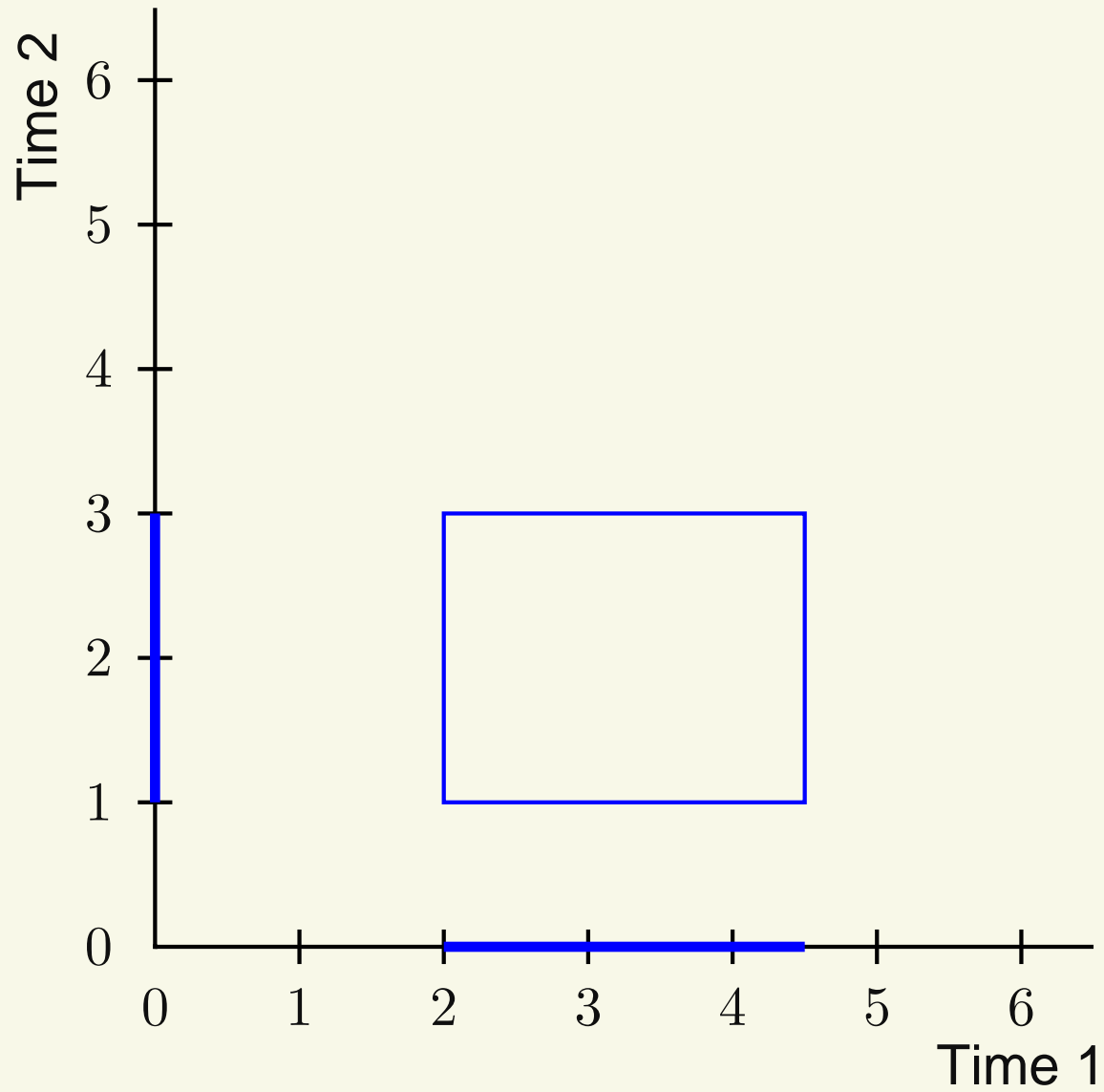
# Generalizations to higher dimensions



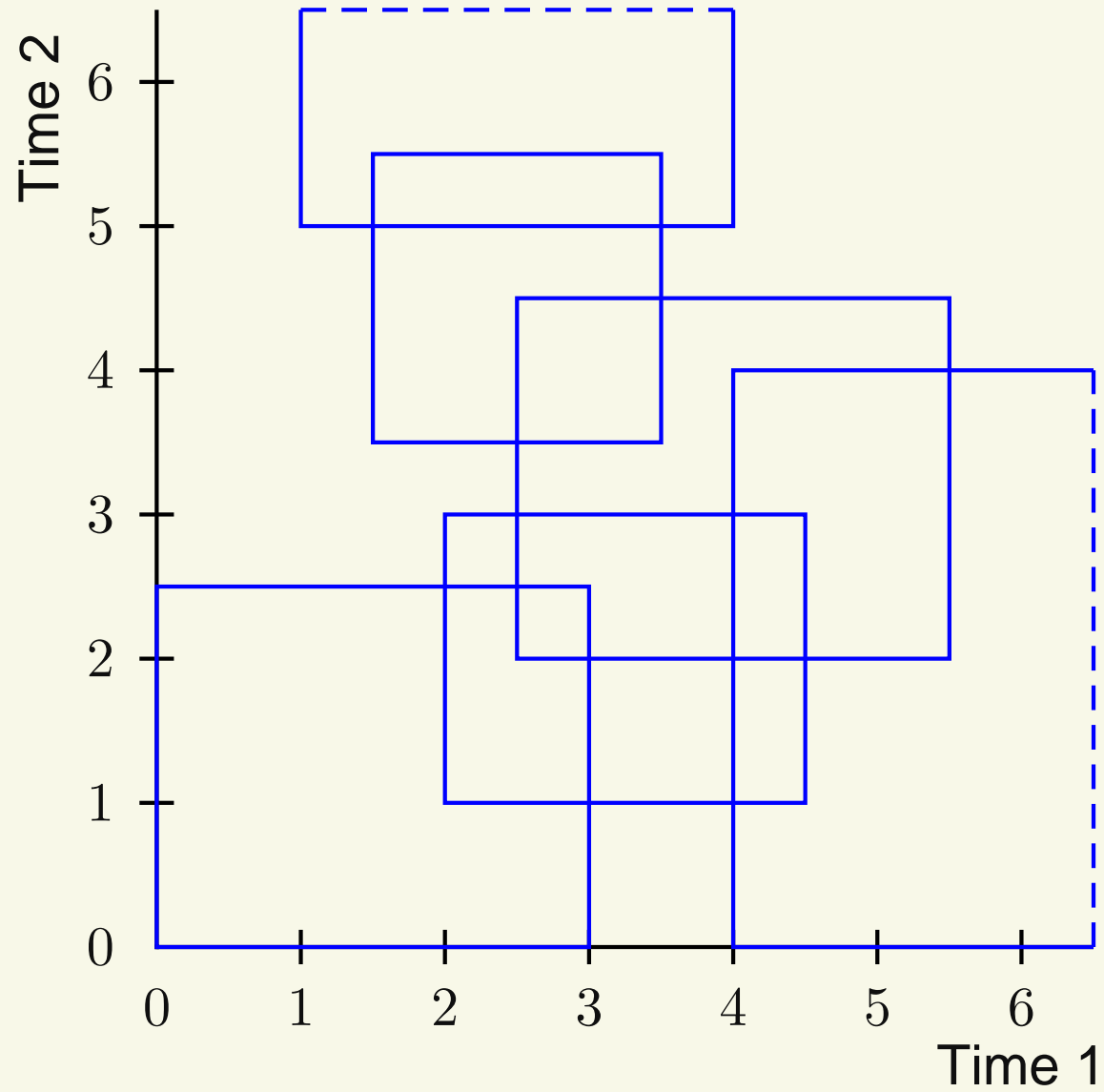
# Generalizations to higher dimensions



# Generalizations to higher dimensions



# Generalizations to higher dimensions



# Seminar for Statistics



Thank you for your attention



Apéro in the Dozentenfoyer