Biostatistics I: Basic Concepts

Dimitris Rizopoulos
Department of Biostatistics, Erasmus University Medical Center

d.rizopoulos@erasmusmc.nl

@drizopoulos

Contents

Chapter 1

Probability Concepts

1.1 Random Variables

Most scientific questions revolve around the *understanding* of phenomena

- Why do we have seasons?
- What will be the weather tomorrow?
- How does the age of patients affect their blood pressure?

1.1 Random Variables (cont'd)

To *understand* phenomena we need to *measure* them

1.1 Random Variables (cont'd)

- Definition: A *variable* is a quantification of a phenomenon
  - ▷ i.e., we assign a *numeric* value to an observable event

- We can have two types of phenomena *Deterministic* and *Stochastic/Random*
  - ▷ *Deterministic*: future values of the phenomenon **can** be exactly predicted from current conditions
  - ▷ *Stochastic*: future values of the phenomenon **cannot** be exactly predicted from current conditions

1.1 Random Variables (cont'd)

- Examples:
  - ▷ Why do we have seasons?
    - * the earth circles around the sun on a tilted axis
    - * this movement remains the same every year
      ⇒ *Deterministic* phenomenon
  - ▷ How the age of patients affects their blood pressure?
    - * as you age, the vascular system changes
    - * but we cannot predict the exact blood pressure of a person based on his/her age
      ⇒ *Stochastic* phenomenon

1.1 Random Variables (cont'd)

- **Definition:** A *random variable* is a numeric quantification of a *stochastic* phenomenon

- **Examples:**
  - ▷ Blood pressure
    - * Phenomenon: circulating blood pressures the walls of blood vessels
    - * Random variable: the numeric value in mm Hg of this pressure
  - ▷ Asthma attack
    - * Phenomenon: tightening of muscles around the airways
    - * Random variable: '1' occurrence of this phenomenon, and '0' otherwise

1.1 Random Variables (cont'd)

- We have two types of random variables: *Continuous* and *Discrete*

- *Continuous* random variables take an *uncountable* number of possible values
  - ▷ cholesterol levels
  - ▷ BMI
  - ▷ VAS pain score
  - ▷ ...

1.1 Random Variables (cont'd)

- *Discrete* random variables take a *countable* number of possible values
  - ▷ death from cancer, 'yes' or 'no' (*dichotomous*)
  - ▷ none, mild, moderate and severe symptoms (*ordinal*)
  - ▷ patients' race (*nominal*)
  - ▷ the number of asthma attacks in a period (*count*)

1.1 Random Variables (cont'd)

- Notation: We typically denote random variables with uppercase Latin letters, e.g.,
  - ▷ Y, X, T, etc.
- With the same corresponding lowercase letter we denote the *realizations (i.e., observed values)* of these random variables, e.g.,
  - ▷ y, x, t, etc.
- Example: Let X denote the random variable describing the blood pressure phenomenon,
  - ▷ the *specific* value we observe when we measure the blood pressure is denoted by x

1.1 Random Variables (cont'd)

- The different types of random variables contain different amount of information

$$\text{dichotomous} < \text{ordinal/nominal} < \text{count} < \text{continuous}$$

- I know more about blood pressure if I know its exact value than only knowing that it was below, e.g., 140 mm Hg
  - ▷ a blood pressure of 138 mm Hg is different than 110 mm Hg, even though both are below the limit

Note: This is why, in general, it is not a good idea to categorize continuous variables
$\Rightarrow$ **by categorizing we lose information**

1.2 Distribution Functions

Our aim is to understand *stochastic* phenomena
$\Rightarrow$ **Challenging** because we cannot exactly predict them

- Despite the random nature of *stochastic* phenomena, often there are patterns in randomness,
  - ▷ i.e., some values of random variables are more **probable** than others

1.2 Distribution Functions (cont'd)

- Definition: *Probability* is a numerical description of how likely an event is to occur

Properties

- ▷ it is constrained between 0 and 1
- ▷ 0 indicates impossibility of the event
- ▷ 1 indicates certainty

1.2 Distribution Functions (cont'd)

- **Definition:** We give probabilities of occurrence to all different possible values of a random variable (that correspond to different possible outcomes of the phenomenon under study)
  - ▷ this collection of probabilities define the *probability distribution* of the random variable

- **Example:** We toss a fair coin
  - ▷ we denote by X the random variable of the possible outcomes

$$X = \begin{cases} 0, & \text{if tails} \\ 1, & \text{if head} \end{cases}$$

1.2 Distribution Functions (cont'd)

- Example: We toss a fair coin
  - ▷ the distribution of X

$$\Pr(X = x) = \begin{cases} \Pr(X = 0) = 0.5, & \text{if tails} \\ \Pr(X = 1) = 0.5, & \text{if head} \end{cases}$$

1.2 Distribution Functions (cont'd)

- Example: We are interested in the severity of complications after a surgery
  - ▷ we denote by X the random variable of the possible outcomes

$$X = \begin{cases} 0, & \text{if no complications} \\ 1, & \text{if mild complications} \\ 2, & \text{if moderate complications} \\ 3, & \text{if severe complications} \end{cases}$$

- Example: We are interested in the severity of complications after a surgery
  - ▷ the distribution of X

$$\Pr(X = x) = \begin{cases} \Pr(X = 0) = 0.4, & \text{if no complications} \\ \Pr(X = 1) = 0.3, & \text{if mild complications} \\ \Pr(X = 2) = 0.1, & \text{if moderate complications} \\ \Pr(X = 3) = 0.2, & \text{if severe complications} \end{cases}$$

1.2 Distribution Functions (cont'd)

- Example: We are interested in the cholesterol levels of a group of patients
  - ▷ we denote by X the random variable of the possible outcomes

$$X = \begin{cases} 180 \text{ mg/dL}, \\ 180.1 \text{ mg/dL}, \\ 180.12 \text{ mg/dL}, \\ 180.123 \text{ mg/dL}, \\ 180.1234 \text{ mg/dL}, \\ \dots \end{cases}$$

We have an **infinite** and **uncountable** set of possible values

1.2 Distribution Functions (cont'd)

- For a continuous random variable the probability that it will take a particular value is zero, i.e.,

$$\Pr(X = x) = \begin{cases} \Pr(X = 180 \text{ mg/dL}) = 0, \\ \Pr(X = 180.1 \text{ mg/dL}) = 0, \\ \Pr(X = 180.12 \text{ mg/dL}) = 0, \\ \dots \end{cases}$$

1.2 Distribution Functions (cont'd)

- For such random variables it is more relevant to speak about an interval, i.e.,

$$\begin{cases} \Pr(180 \text{ mg/dL} < X < 190 \text{ mg/dL}) = 0.3, \\ \Pr(190 \text{ mg/dL} < X < 210 \text{ mg/dL}) = 0.2, \\ \Pr(X > 210 \text{ mg/dL}) = 0.25, \\ \ldots \end{cases}$$

1.2 Distribution Functions (cont'd)

- We have some key functions to describe distributions

- Definition: The *cumulative distribution function* (CDF) denotes the probability that a random variable X takes a value less or equal to x

$$F_X(x) = \Pr(X \leq x)$$

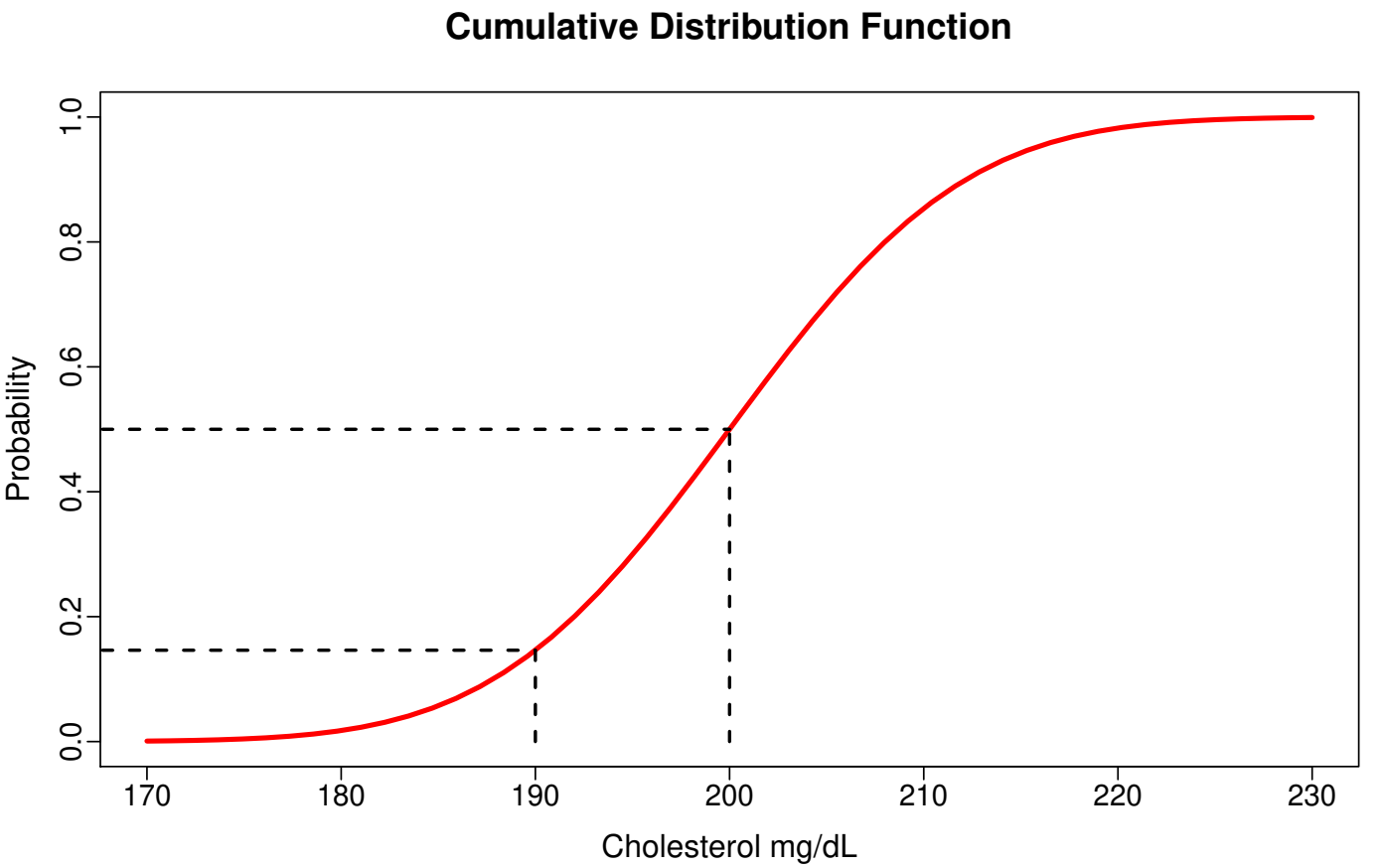
Properties

▷ it is constrained between 0 and 1

▷ it is an increasing function of x

▷ it is defined for both continuous and discrete random variables

1.2 Distribution Functions (cont'd)



1.2 Distribution Functions (cont'd)

- **Definition:** The *probability mass function* denotes the probability that a *discrete* random variable X takes a particular value x

$$p_X(x) = \Pr(X = x)$$

Properties

▷ it is constrained between 0 and 1

▷ the sum of the probabilities for all possible values of X is one

$$\begin{aligned} \sum_x p_X(x) &= \Pr(X = 0) + \Pr(X = 1) + \Pr(X = 2) + \ldots \\ &= 1 \end{aligned}$$

1.2 Distribution Functions (cont'd)

- Definition: The *probability density function* (PDF) is used to specify the probability that a *continuous* random variable X takes values in particular interval (a, b)

$$\Pr(a < X < b) = \int_a^b f_X(x)\, dx$$

Properties

▷ it is non-negative

▷ the integral over all possible values of X is one

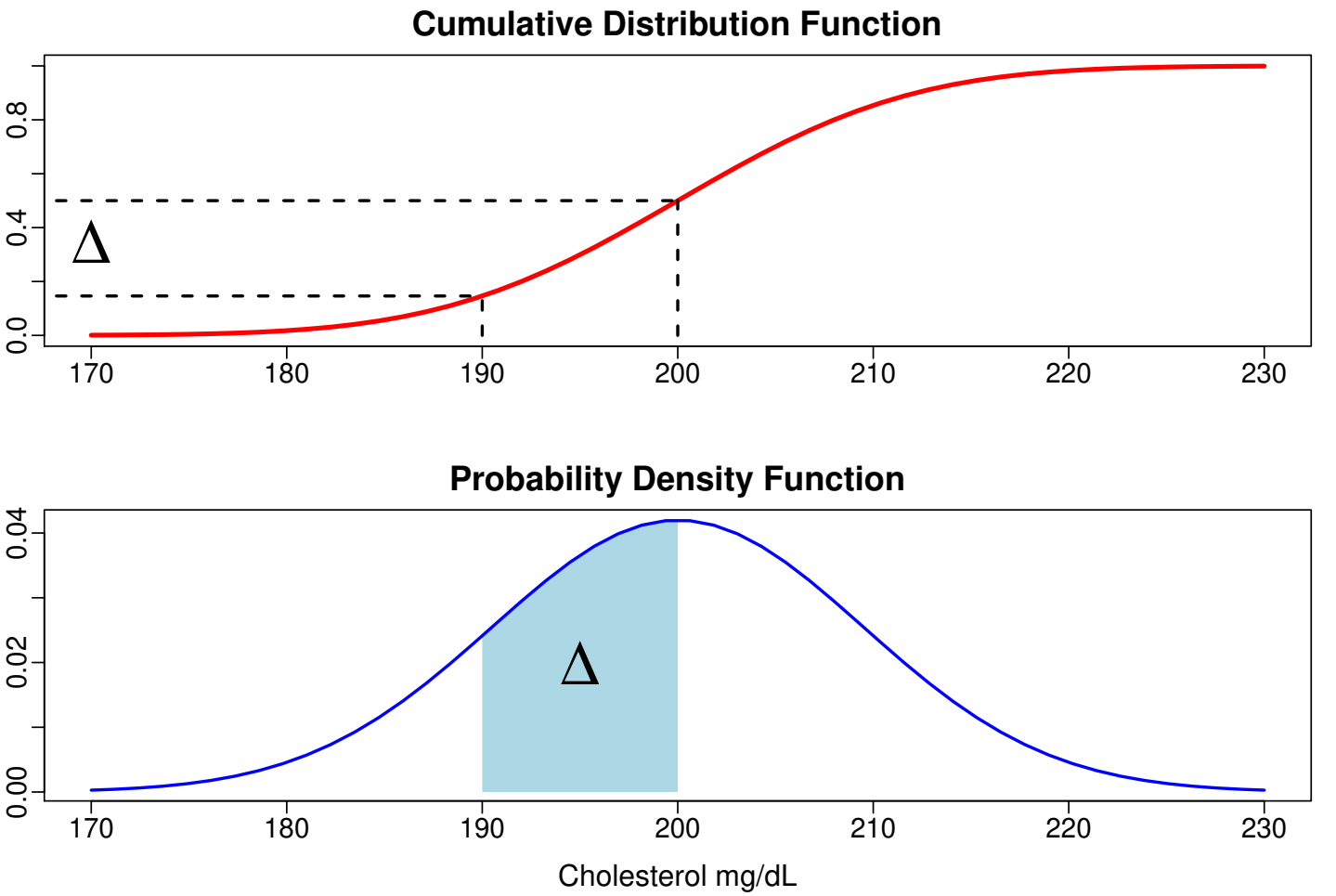
$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1$$

1.2 Distribution Functions (cont'd)

- For continuous random variables the *cumulative distribution function* and the *probability density function* are linked via

$$f_X(x) = \frac{dF_X(x)}{dx} \quad \text{and} \quad F_X(x) = \int_{-\infty}^{x} f_X(s)\, ds$$

1.2 Distribution Functions (cont'd)



1.3 Expectation & Quantiles

- We have seen how we can describe the *whole* distribution of random variables using the
  - ▷ probability mass function
  - ▷ probability density function
  - ▷ cumulative distribution function

Most often we would like to summarize the distribution by some *representative* quantities

1.3 Expectation & Quantiles (cont'd)

- **Definition:** The *expected value* of a random variable X is the mean of its distribution

$$E(X) = \begin{cases} \sum_x x \Pr(X = x), & \text{if } X \text{ is discrete} \\ \\ \int x f_X(x)\, dx, & \text{if } X \text{ is continuous} \end{cases}$$

1.3 Expectation & Quantiles (cont'd)

- Notes:
  - ▷ the expected value is a *weighted* average of the random variable's values
  - ▷ weighted because different values have different probabilities of occurrence
  - ▷ (for continuous random variables different *intervals of values* have different probabilities)

1.3 Expectation & Quantiles (cont'd)

- Example: We are interested in the severity of complications after a surgery
  - ▷ we denote by X the random variable of the possible outcomes
  - ▷ its probability mass function is

$$\Pr(X = x) = \begin{cases} \Pr(X = 0) = 0.4, & \text{if no complications} \\ \Pr(X = 1) = 0.3, & \text{if mild complications} \\ \Pr(X = 2) = 0.1, & \text{if moderate complications} \\ \Pr(X = 3) = 0.2, & \text{if severe complications} \end{cases}$$

- **What is the expected values of X?**

1.3 Expectation & Quantiles (cont'd)

$$
\begin{aligned}
E(X) &= \sum_x x \Pr(X = x) \\
&= 0 \times \Pr(X = 0) + 1 \times \Pr(X = 1) + 2 \times \Pr(X = 2) + 3 \times \Pr(X = 3) \\
&= 0 \times 0.4 + 1 \times 0.3 + 2 \times 0.1 + 3 \times 0.2 \\
&= 1.1
\end{aligned}
$$

1.3 Expectation & Quantiles (cont'd)

- The expected value is a *location* measure of the distribution of a random variable X
  - ▷ it gives us some information where the mean of the distribution is located

- Another set of useful location measures is the *quantiles* of the distribution

1.3 Expectation & Quantiles (cont'd)

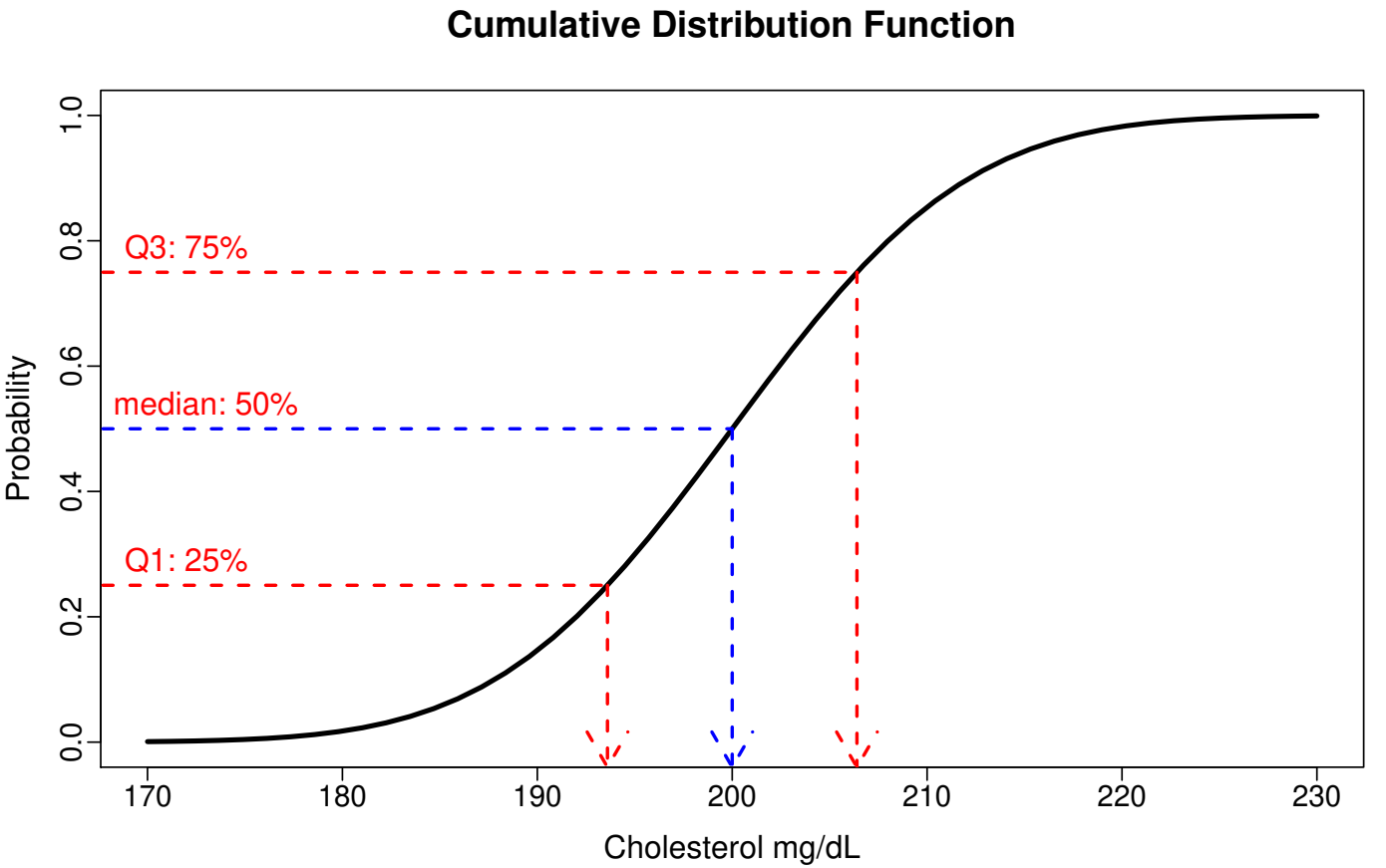
- **Definition:** The *k-th quantile* of a random variable X is the value below which a given $k\%$ of values in its distribution fall

$$Q_X(k) = \{x : F_X(x) = k\}, \quad 0 \leq k \leq 1$$

i.e., the value x for which the CDF equals k

- The most-used quantiles are
  - ▷ *median*: the value x below which 50% of the observations fall
  - ▷ *1st quartile*: the value x below which 25% of the observations fall
  - ▷ *3rd quartile*: the value x below which 75% of the observations fall

1.3 Expectation & Quantiles (cont'd)



1.4 Variance

- The expected value and the quantiles give us information about the location of the distribution of a random variable

- Another important quantity is the spread of the distribution
  - ▷ i.e., how far away are the values of a random variable located from each other

1.4 Variance (cont'd)

- **Definition:** The *variance* of a random variable X measures how far its values are spread out from their mean (i.e., expected value)

$$\text{var}(X) = E\Big[\big\{X - E(X)\big\}^2\Big] = \begin{cases} \displaystyle\sum_x \big\{x - E(X)\big\}^2 \Pr(X = x), & \text{if } X \text{ is discrete} \\ \displaystyle\int \big\{x - E(X)\big\}^2 f_X(x)\, dx, & \text{if } X \text{ is continuous} \end{cases}$$

1.4 Variance (cont'd)

- Notes:
  - ▷ the variance is always positive
  - ▷ the reason why we compute *squared differences* is because otherwise the positive and negative differences would cancel out
  - ▷ the fact that we calculate square differences means that the variance is on the squared scale of the random variable
    * e.g., the variance of the blood pressure random variable is in mm Hg^2

    this is why most often we also calculate the *standard deviation*

$$\text{sd}(X) = \sqrt{\text{var}(X)}$$

1.5 Covariance & Correlation

- The variance measures how far a set of numbers is spread out for a single random variable

- However, often we are interested in measuring the spread of pairs of random variables, and how these spreads are associated with each other, e.g.,
  - ▷ how are changes in blood pressure associated with changes in age?
  - ▷ how are changes in cholesterol associated with changes in BMI?

1.5 Covariance & Correlation (cont'd)

▷ **Definition:** The *covariance* is a measure of how much two random variables change together

$$\text{cov}(X, Y) = E\Big[\{X - E(X)\}\ \{Y - E(Y)\}\Big]$$

1.5 Covariance & Correlation (cont'd)

- Notes:
  - ▷ it can be positive or negative
    - * if the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values, the covariance is positive
    - * in the opposite case, the covariance is negative
  - ▷ the magnitude of the covariance is not easy to interpret because it depends on the magnitudes of the variables
  - ▷ the variance is a special case of the covariance

$$\text{var}(X) = \text{cov}(X, X)$$

1.5 Covariance & Correlation (cont'd)

- **Definition:** The *correlation* is a standardized version of the covariance and is a measure of the *linear* correlation (dependence) between two variables

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\text{sd}(X)\ \text{sd}(Y)}$$

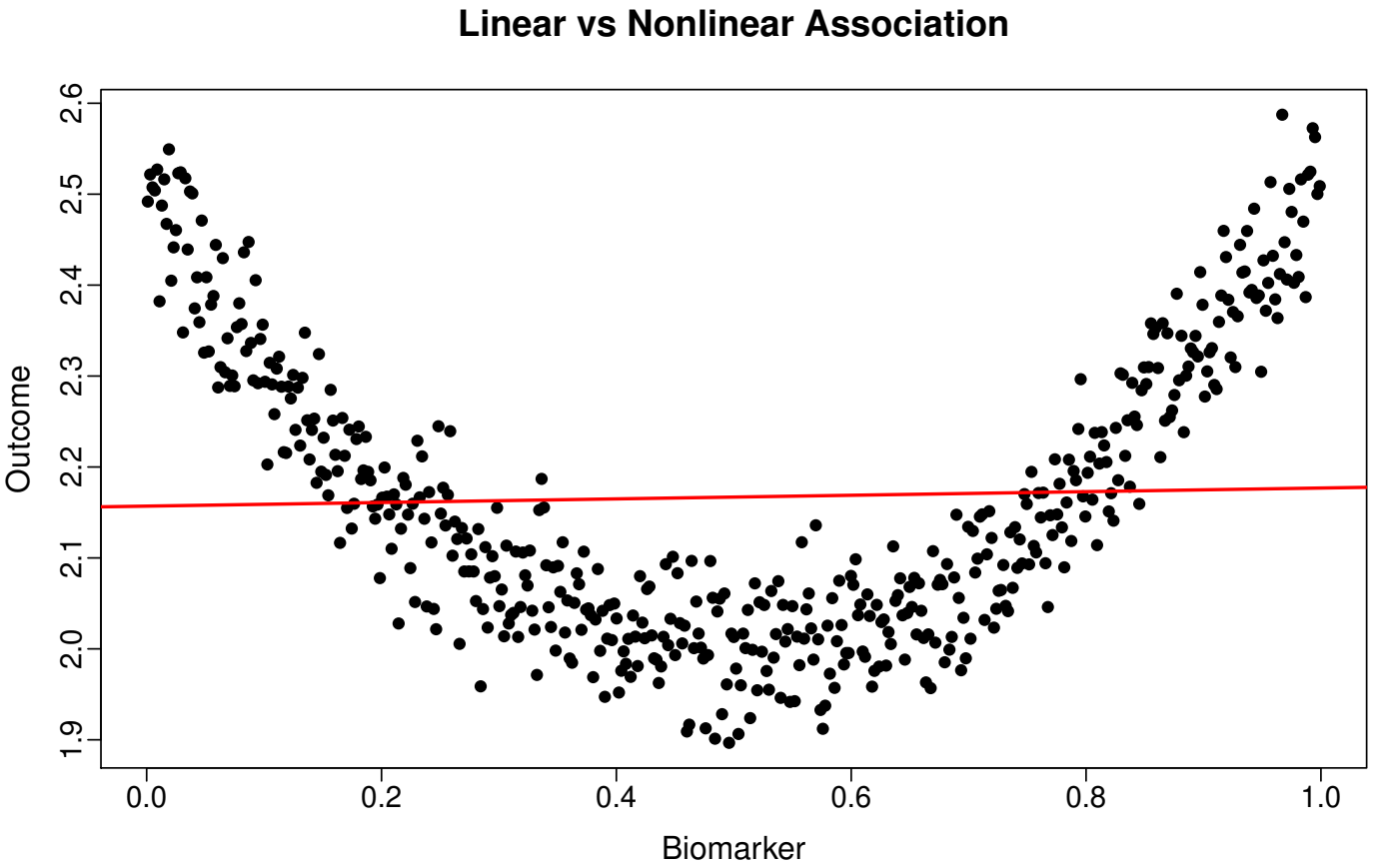
Properties

- ▷ it is constrained between -1 and 1
- ▷ 0 means no correlation between the two variables
- ▷ -1 means perfect negative correlation
- ▷ 1 mean perfect positive correlation

1.5 Covariance & Correlation (cont'd)

- Notes:
  - ▷ the correlation is a measure of *linear* association
  - ▷ two variables may have zero linear correlation but still be (strongly) associated

1.5 Covariance & Correlation (cont'd)



1.6 Standard Distributions

- We have seen that the distribution of a random variable describes in a generic manner the probability of certain events

- Often we use distributions that place some restrictions on their shape
  - ▷ these distributions have *parameters* that control key quantities of the distribution,
  - ▷ typically, the mean and variance of the distribution
  - ▷ parameters are typically denoted with Greek letters

1.6 Standard Distributions (cont'd)

- Definition: The *normal (Gaussian)* distribution has a probability density function given by the equation

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Notes:

▷ we write $X \sim \mathcal{N}(\mu, \sigma^2)$ to denote that X follows the normal distribution

▷ the parameter μ denotes the mean of the distribution (i.e., $E(X) = \mu$)

▷ the parameter σ^2 denotes the variance of the distribution (i.e., var$(X) = \sigma^2$)

▷ it is a symmetric distribution around $\mu \Rightarrow$ the median is also μ

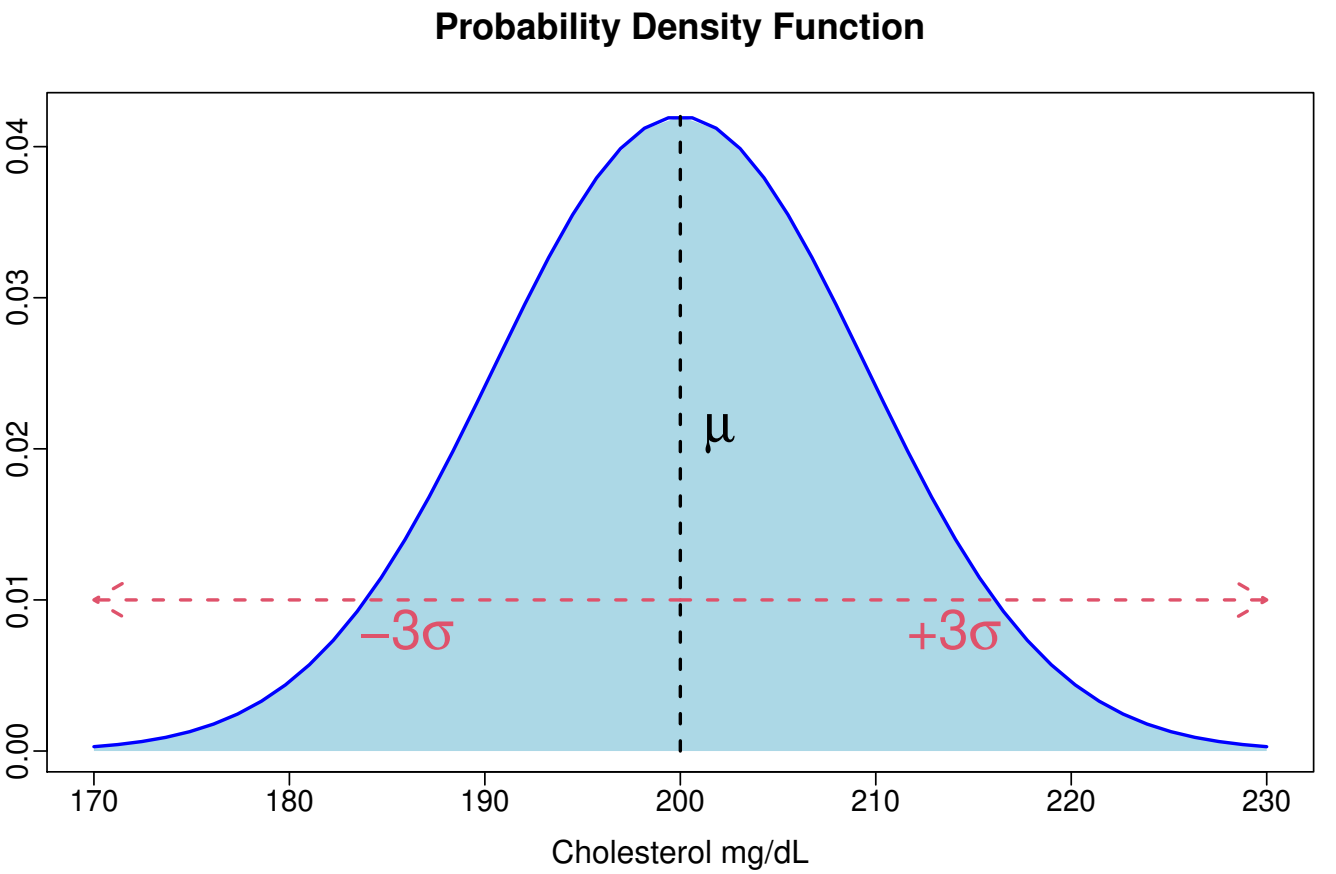
▷ (Note: π is not a parameter here but the number $\pi = 3.14159$)

1.6 Standard Distributions (cont'd)

Notes:

- ▷ it describes phenomena for which the majority of the observations is located around the mean, and as we get further away from the mean, fewer and fewer observations are to be found
- ▷ **it plays a very central role in the analysis of continuous random variables**

1.6 Standard Distributions (cont'd)



1.6 Standard Distributions (cont'd)

- **Definition:** The *binomial* distribution has a probability mass function given by the equation

$$p_X(x) = \binom{N}{x} \pi^x (1-\pi)^{N-x}$$

Notes:

- ▷ we write $X \sim Bin(N, \pi)$ to denote that X follows the binomial distribution
- ▷ it describes the number of 'successes' out of N independent trials, where the probability of success of each trial is π
- ▷ the mean is $E(X) = N\pi$ and the variance var$(X) = N\pi(1-\pi)$
- ▷ (Note: the first term is the binomial coefficient giving the number of ways x successes can be distributed in N trials)

1.6 Standard Distributions (cont'd)

Notes:

▷ when we have one trial, i.e., $N = 1$ we get the *Bernoulli* distribution

$$p_X(x) = \pi^x(1-\pi)^{1-x}, \quad x \text{ takes the values 0 or 1}$$

▷ **it plays a very central role in the analysis of dichotomous and ordinal random variables**

1.6 Standard Distributions (cont'd)

- **Definition:** The *Poisson* distribution has a probability mass function given by the equation

$$p_X(x) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

Notes:

▷ we write $X \sim Pois(\lambda)$ to denote that X follows the Poisson distribution

▷ it describes the number of 'events' that occur in a given period of time

▷ the mean and variance are equal to λ, i.e., $E(X) = \text{var}(X) = \lambda$

▷ **it plays a very central role in the analysis of count variables**

▷ (Note: $x!$ is the factorial, e.g., $5! = 1 \times 2 \times 3 \times 4 \times 5$)

Chapter 2

Statistics Concepts

2.1 Population & Sample

Aim: Understand Phenomena

- Research questions
  - ▷ will the new treatment for hypertension work better than the standard one?
  - ▷ which are risk factors for IC admission due to COVID-19?
  - ▷ are genetic factors related to the onset of breast cancer?
  - ▷ ...

2.1 Population & Sample (cont'd)

These questions are generically formulated
For which specific patients are we talking about?

- Research questions
  - ▷ will the new treatment for hypertension work better than the standard one?
    * patients older than 50 years old
    * who had blood pressure higher than 160 mm Hg for two consecutive days
    * no family history of hypertension

2.1 Population & Sample (cont'd)

- Definition: The *target population* is the precise definition of the total group of individuals for whom we want to draw conclusions
  - ▷ this is achieved by formulating the *inclusion criteria* for the study

- **Ideally**, we collect *data* from the whole population (i.e., from all subjects), and proceed to analyze them
  - ▷ *data* are actually the realizations from the random variables of interest,
  - ▷ e.g., blood pressure measurements

2.1 Population & Sample (cont'd)

- **However**, the problem is that it is infeasible to collect data from all subjects in the population
  - ▷ simply because the population contains too many subjects

- To proceed, we work with a *sample* (i.e., a small subset) from the population

> A **well-chosen** sample will contain most of the information about a particular population characteristic

2.1 Population & Sample (cont'd)

- Definition: When all subjects from the population have the same chance to be included in the sample we obtain a *random sample*
  - ▷ such a sample is guaranteed to provide us with valid statements about the target population

- **However**, most often, we cannot take a random sample but rather a *"convenience"* sample, e.g.,
  - ▷ subjects from a hospital's registry
  - ▷ subjects from a specific area (Rotterdam study)
  - ▷ ...

2.1 Population & Sample (cont'd)

- Problems with non-random samples
  - ▷ (academic) hospital patients are not the same as the ones seen in the community
  - ▷ patients who return questionnaires are different than those who do not
  - ▷ ...

These problems can lead to **sampling bias**,
i.e., the results of the analysis can be wrongly attributed

2.1 Population & Sample (cont'd)

- To be able to make generalizations from our "convenience" sample, we want it to be sufficiently *representative* of the target population

- Definition: A *representative sample* is a group of subjects from the target population that adequately replicates the population according to whatever characteristic or quality is under study

- A representative sample parallels key variables and characteristics of the larger population, e.g., sex, age, education level, socioeconomic status, etc.

2.2 Estimation & Sampling Variability

- The fact that we can only work with a (representative) sample from our target population causes an important complication

> **Sampling Error**: There will be a difference between the characteristic we measure in the sample and the same characteristic in the population

- Note: a larger sample size reduces the likelihood of sampling errors and increases the likelihood that the sample accurately reflects the target population

2.2 Estimation & Sampling Variability (cont'd)

- Definition: *Sampling variability* is the variability in the analysis results caused by the fact that we work with the sample and not the whole population

  ▷ we often work with a particular study/sample

  ▷ however, this is just one sample from our target population

  ▷ in principle, we could take many different samples from the population

  ▷ *each sample would yield different results*

2.2 Estimation & Sampling Variability (cont'd)

- Let's return to our research questions
  - ▷ will the new treatment for hypertension work better than the standard one?

- But at whom is this question targeted?
  - ▷ our specific sample
  - ▷ or our target population?

If we want to draw some conclusions from the sample at hand regarding the population, we need to quantify and account for the sampling variability

2.2 Estimation & Sampling Variability (cont'd)

- But how can we say anything about the variability from different samples, given that we have only a single sample at hand?

Under some assumptions and the statistical theory,
we can determine the magnitude of sampling variability of samples such as our own,
but based only on the data in our single sample

2.2 Estimation & Sampling Variability (cont'd)

- Example:
  - ▷ what is the average blood pressure for a specific group of patients?
  - ▷ let's formulate this question more precisely
    - * the characteristics of the group define our target population
    - * we denote by X the random variable describing the blood pressure values
    - * this random variable will follow a distribution, with mean denoted by a parameter μ
    - * our aim is to estimate μ

- Definition: The *estimand* is the parameter of the target population we wish to estimate from a sample

2.2 Estimation & Sampling Variability (cont'd)

- Example:
  - ▷ let's assume that we will obtain a representative sample from this population of size n (i.e., the number of patients in our sample)
  - ▷ **Important:** We think generically, we do not have the sample yet!
    * if I will get a sample, how can I use it to estimate μ?

2.2 Estimation & Sampling Variability (cont'd)

- Example:
  - ▷ each subject in the sample has a random variable X_i describing his/her blood pressure levels
  - ▷ we could then estimate μ using the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Definition: The *estimator* is a rule for estimating the parameter in the population using the data we will collect in a sample

2.2 Estimation & Sampling Variability (cont'd)

- Example:
  - ▷ when we have available *specific* values x_i from a *realized* sample, we calculate the realized value of the sample mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where x_i denotes the blood pressure measurements for patient i

- Definition: The estimate of a particular population characteristic we obtain from a specific sample using an estimator is called the *point estimate*

2.2 Estimation & Sampling Variability (cont'd)

- But what would be the variability of this point estimate in *all* different samples of size n from this target population?

$$\text{se}(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

where

▷ σ is the standard deviation of blood pressure values in our population

2.2 Estimation & Sampling Variability (cont'd)

- Definition: The *standard error* is the standard deviation of the results from all possible samples from the target population
  - ▷ the collection of the possible results from all different samples is called the *sampling distribution*, and
  - ▷ the *standard error* is the standard deviation of this distribution

2.3 Maximum Likelihood Estimation

- **Aim:** We want to estimate key characteristics for the population from a representative sample we have at hand

- There are several methods to obtain such estimates, but one of the most popular ones is

The Maximum Likelihood Estimation Method

2.3 Maximum Likelihood Estimation (cont'd)

- To find the most optimal values of a distribution's parameters, we need a measure of how likely specific values of these parameters are in light of the data

- As measure of likelihood we use the *probability density function* (or the *probability mass function* if we have discrete data) but we treat it as a function of the parameters given the sample at hand

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

where

  - ▷ x_i denotes the data (e.g., blood pressure values)
  - ▷ θ are the parameters (e.g., mean and variance)

2.3 Maximum Likelihood Estimation (cont'd)

- The most 'likely' parameter values in the light of the data are the values that maximize the likelihood function

- For numerical reasons, it is more convenient to work with the log-likelihood function

$$\ell(\theta) = \log\{L(\theta)\} = \log\Big\{\prod_{i=1}^{n} f(x_i;\theta)\Big\} = \sum_{i=1}^{n} \log f(x_i;\theta)$$

- The value of θ that maximizes $L(\theta)$ also maximizes $\ell(\theta)$
  $\Rightarrow$ sufficient to maximize $\ell(\theta)$

2.3 Maximum Likelihood Estimation (cont'd)

- Example: We have a sample of blood pressure values for a group of patients – we consider the following simple setting

$$\text{BP}_i \sim \mathcal{N}(\mu, 1)$$

where

  - ▷ BP_i is the blood pressure value for patient i $(i = 1, \ldots, n)$
  - ▷ μ denotes the mean of the blood pressure values
  - ▷ the variance σ^2 is set to 1

2.3 Maximum Likelihood Estimation (cont'd)

- The log-likelihood function is

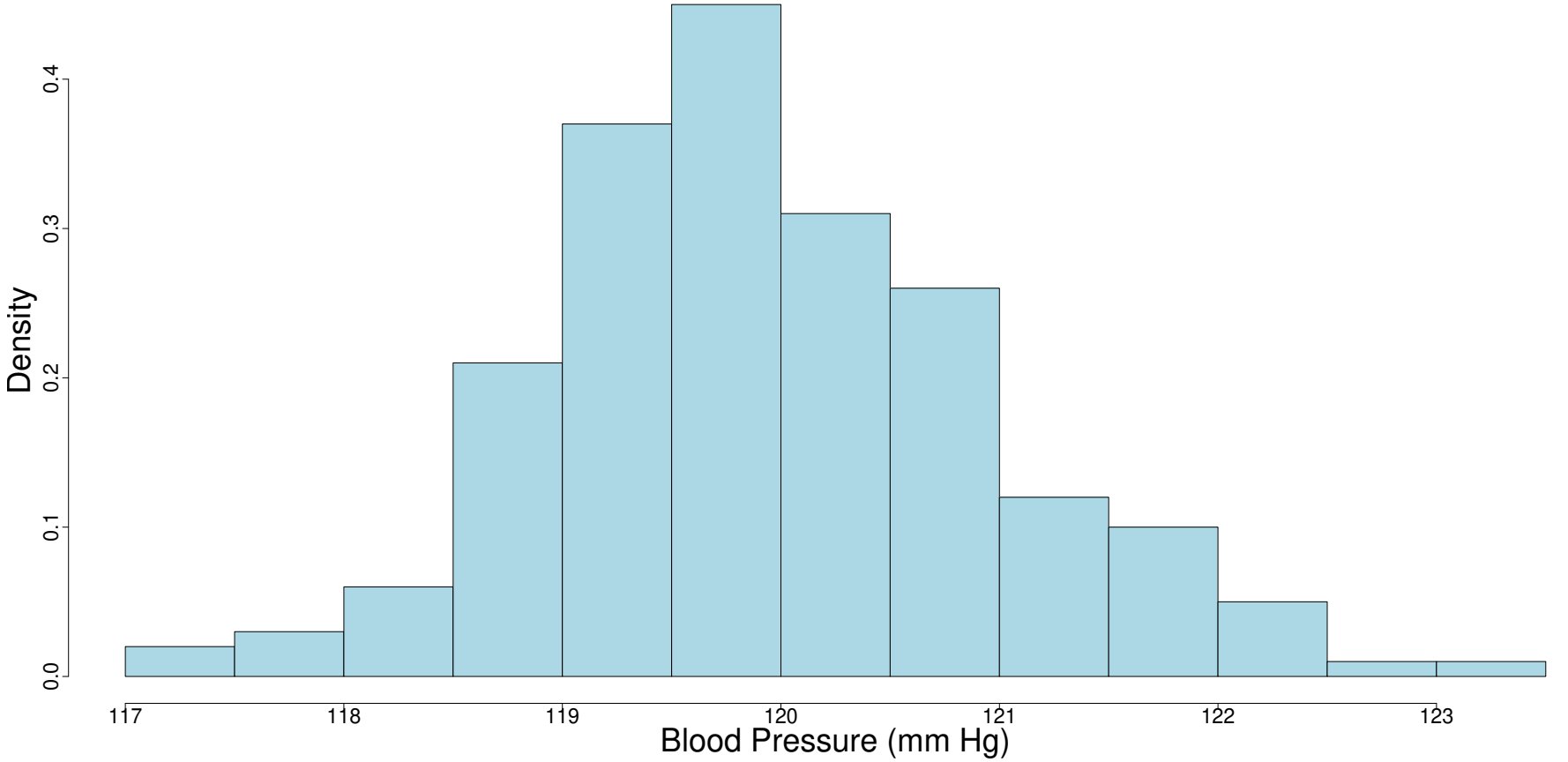
$$\ell(\mu) = \sum_{i=1}^{n} \log f(\mathsf{BP}_i; \mu)$$

where

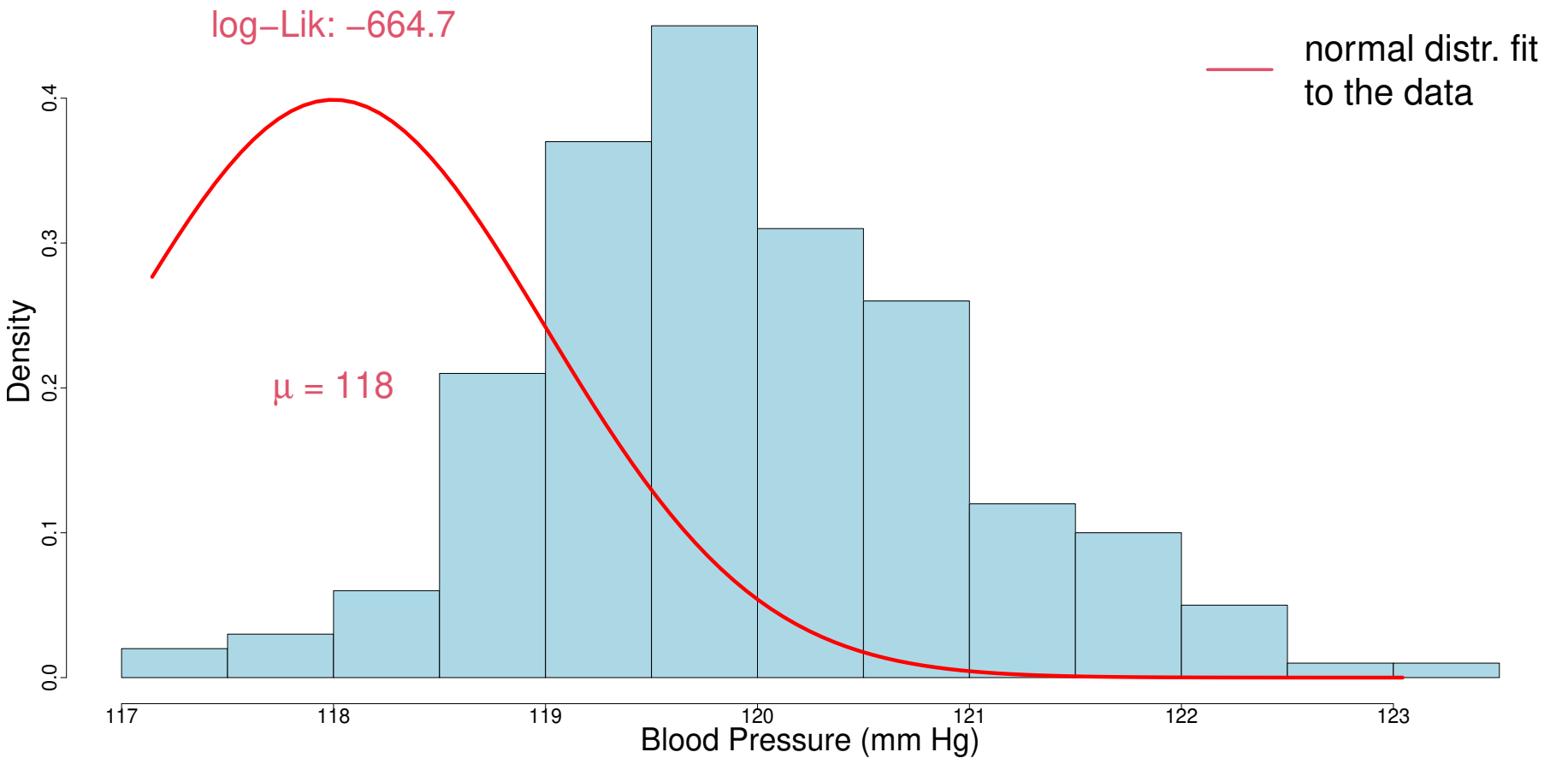
▷ $f(\mathsf{BP}_i; \mu)$ denotes here the *probability density function* of the normal distribution with mean μ and variance 1

$$f(x_i; \mu) = \frac{1}{\sqrt{2\pi}} \exp\Big\{-\frac{1}{2}(\mathsf{BP}_i - \mu)^2\Big\}$$

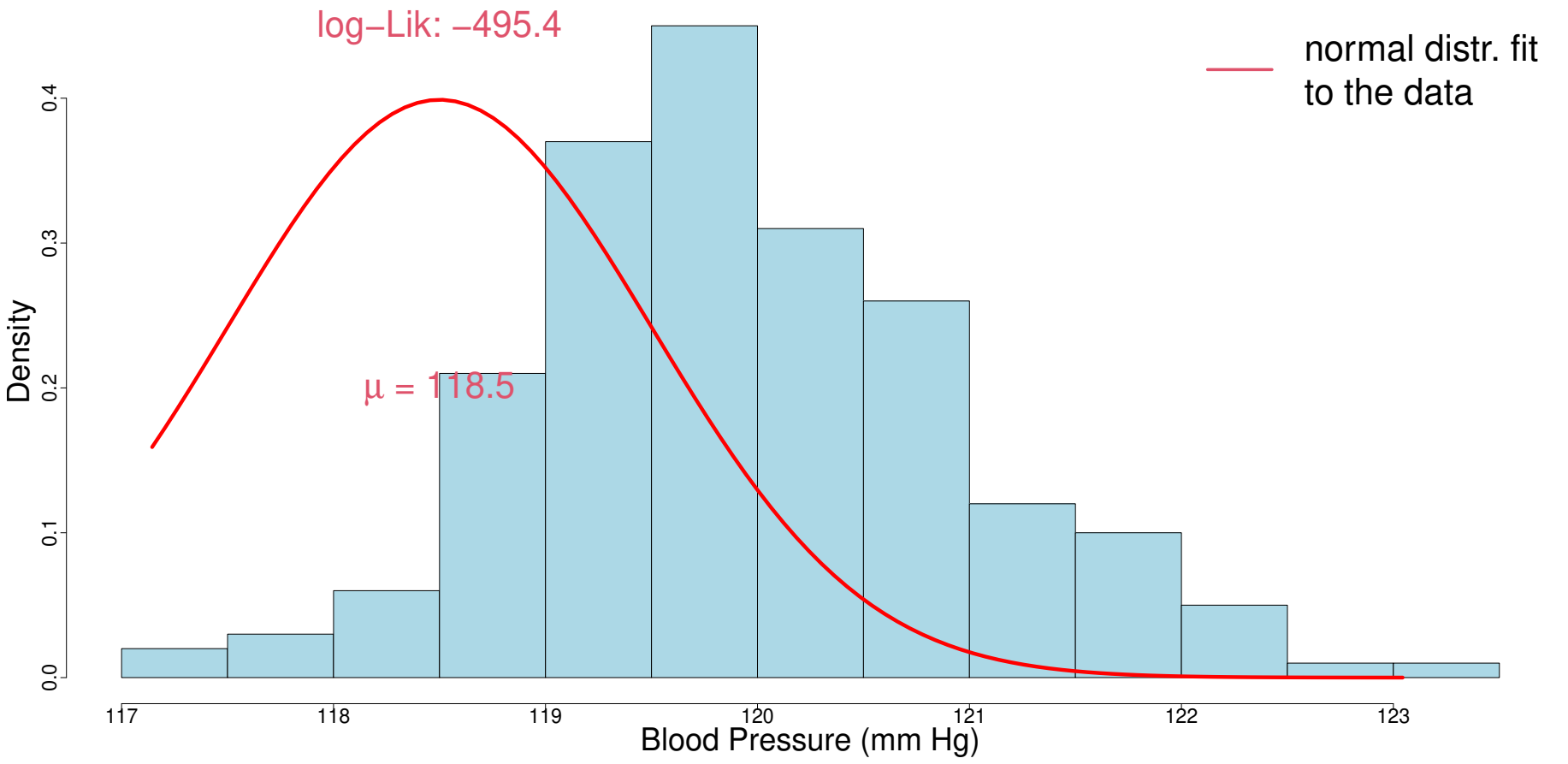
2.3 Maximum Likelihood Estimation (cont'd)



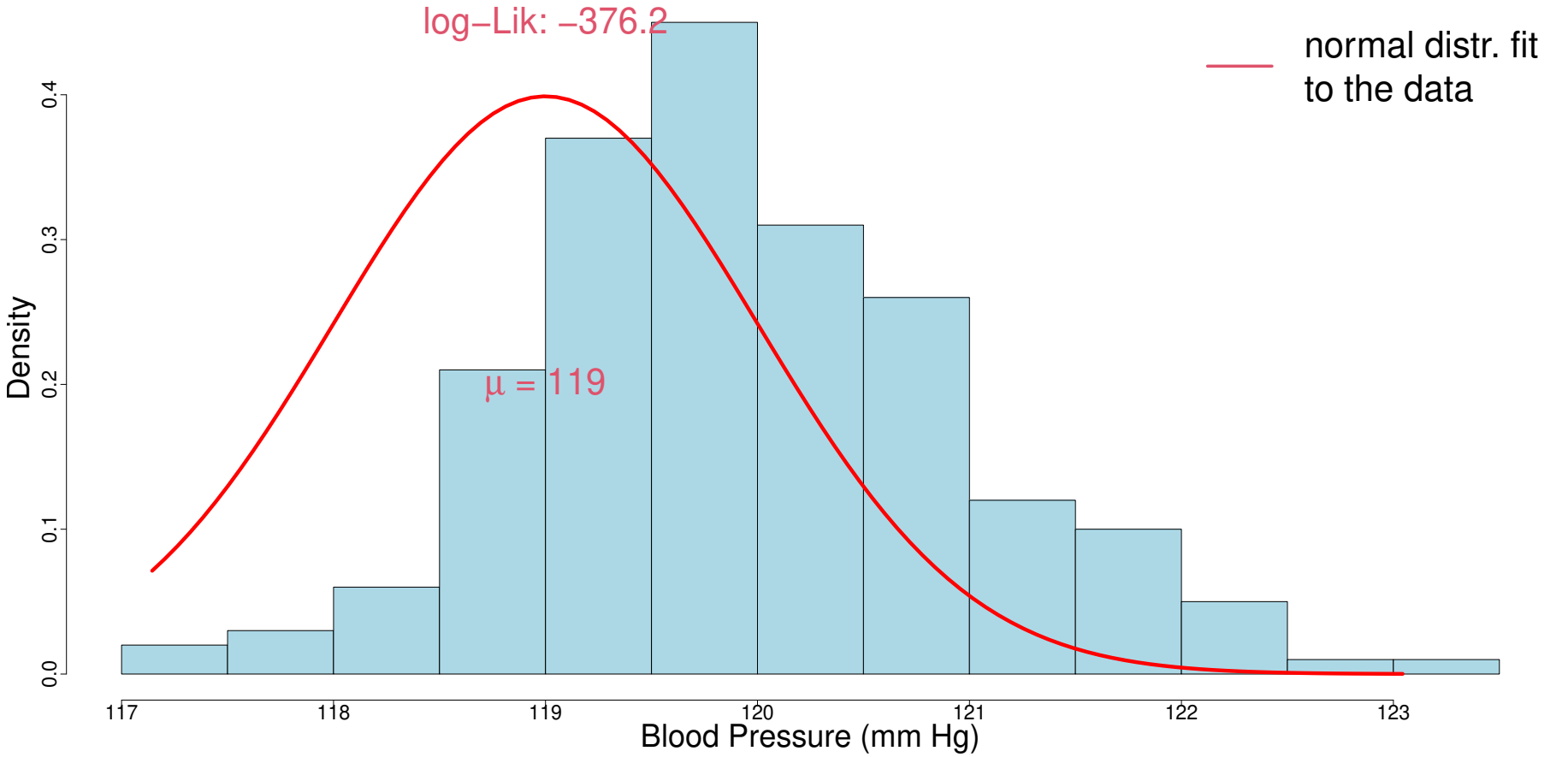
2.3 Maximum Likelihood Estimation (cont'd)



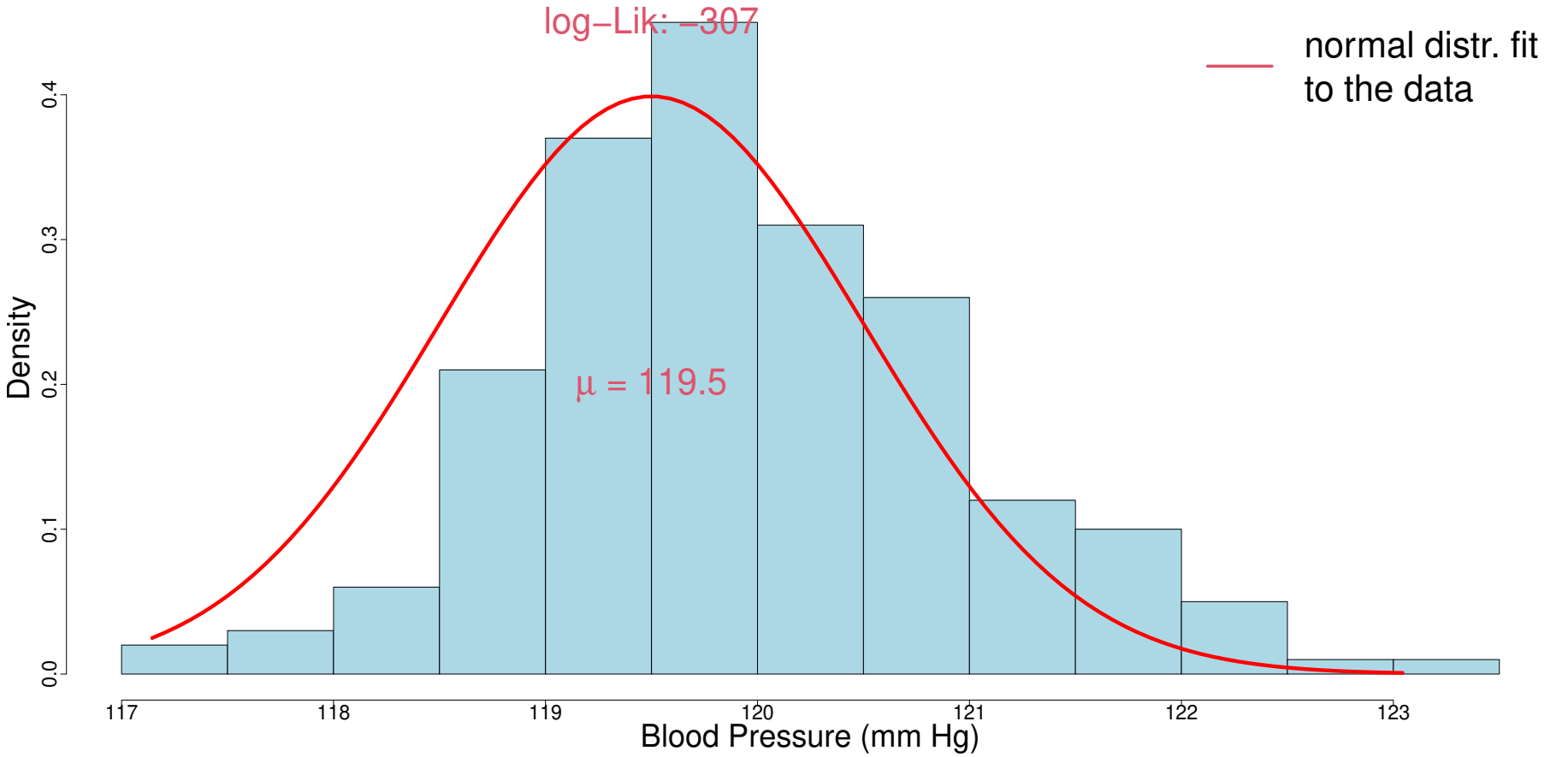
2.3 Maximum Likelihood Estimation (cont'd)



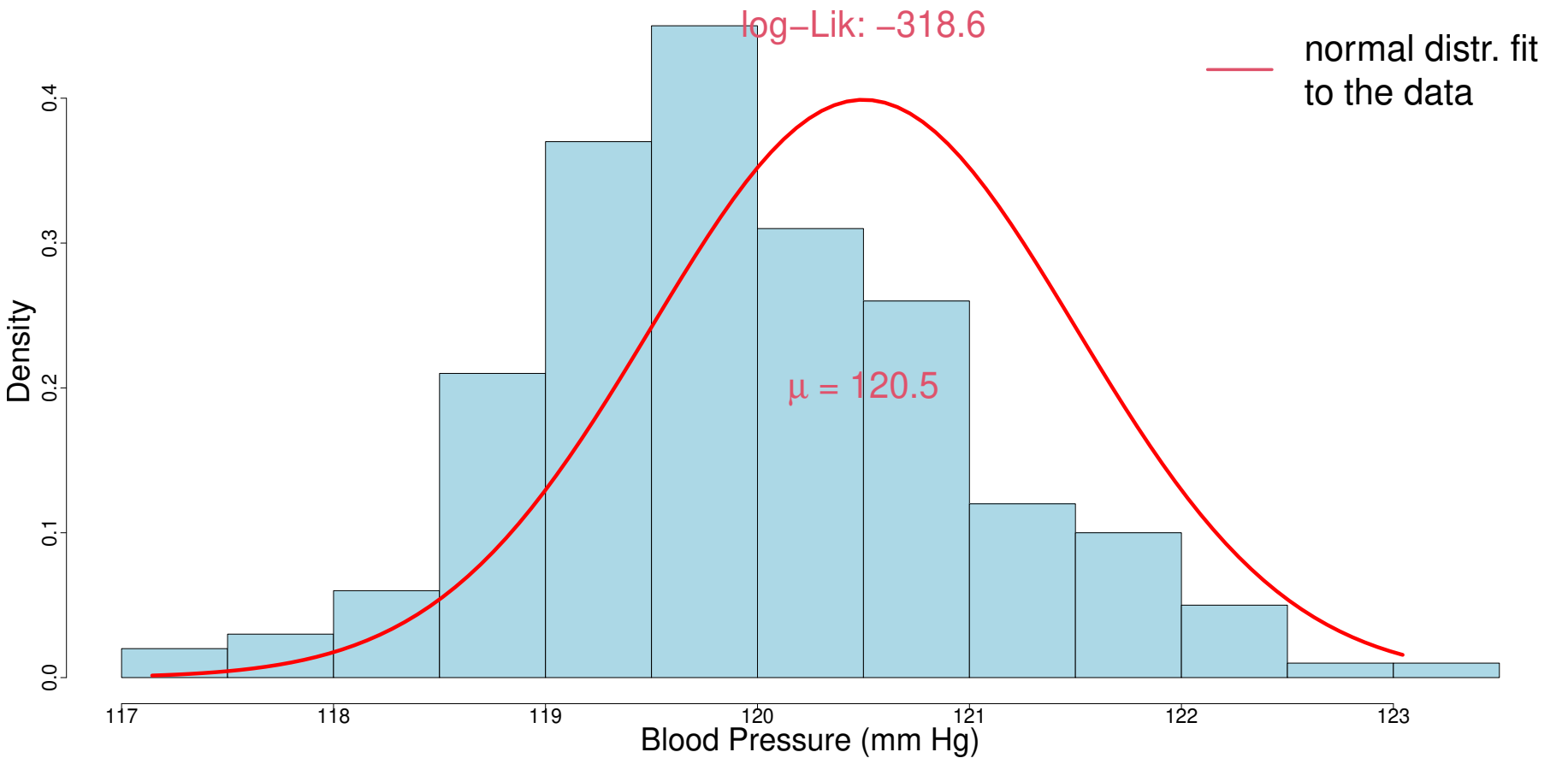
2.3 Maximum Likelihood Estimation (cont'd)



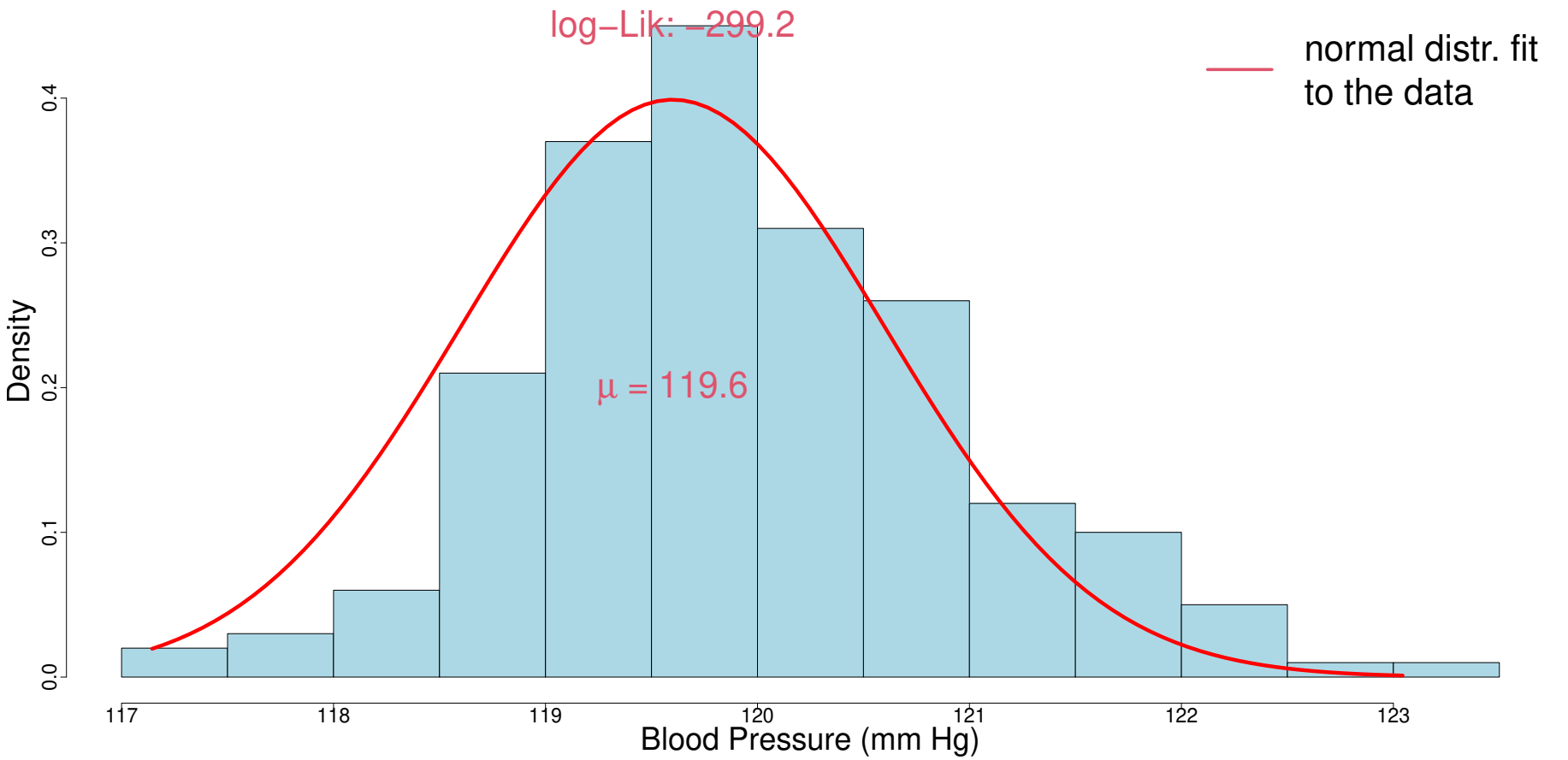
2.3 Maximum Likelihood Estimation (cont'd)



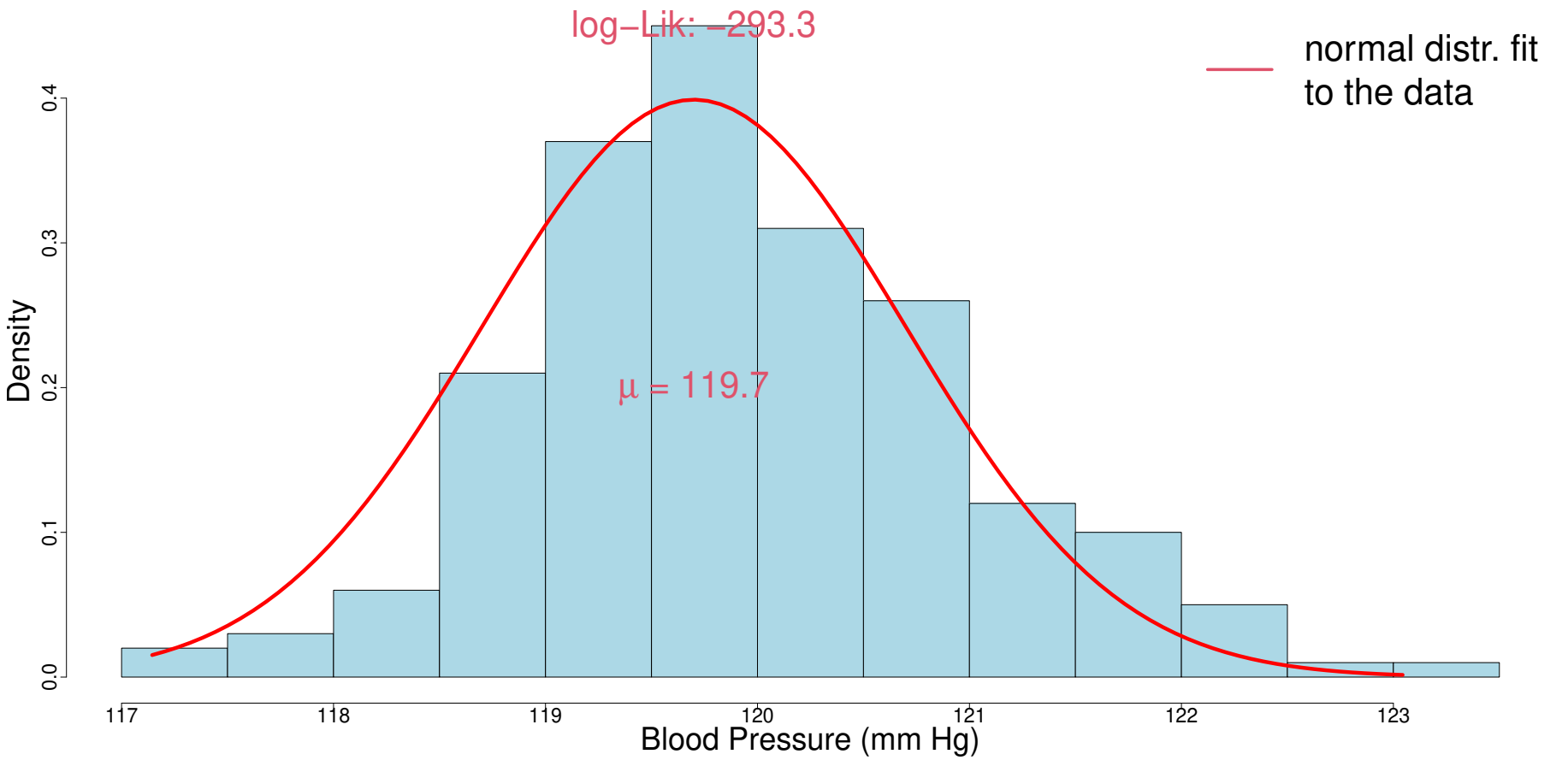
2.3 Maximum Likelihood Estimation (cont'd)



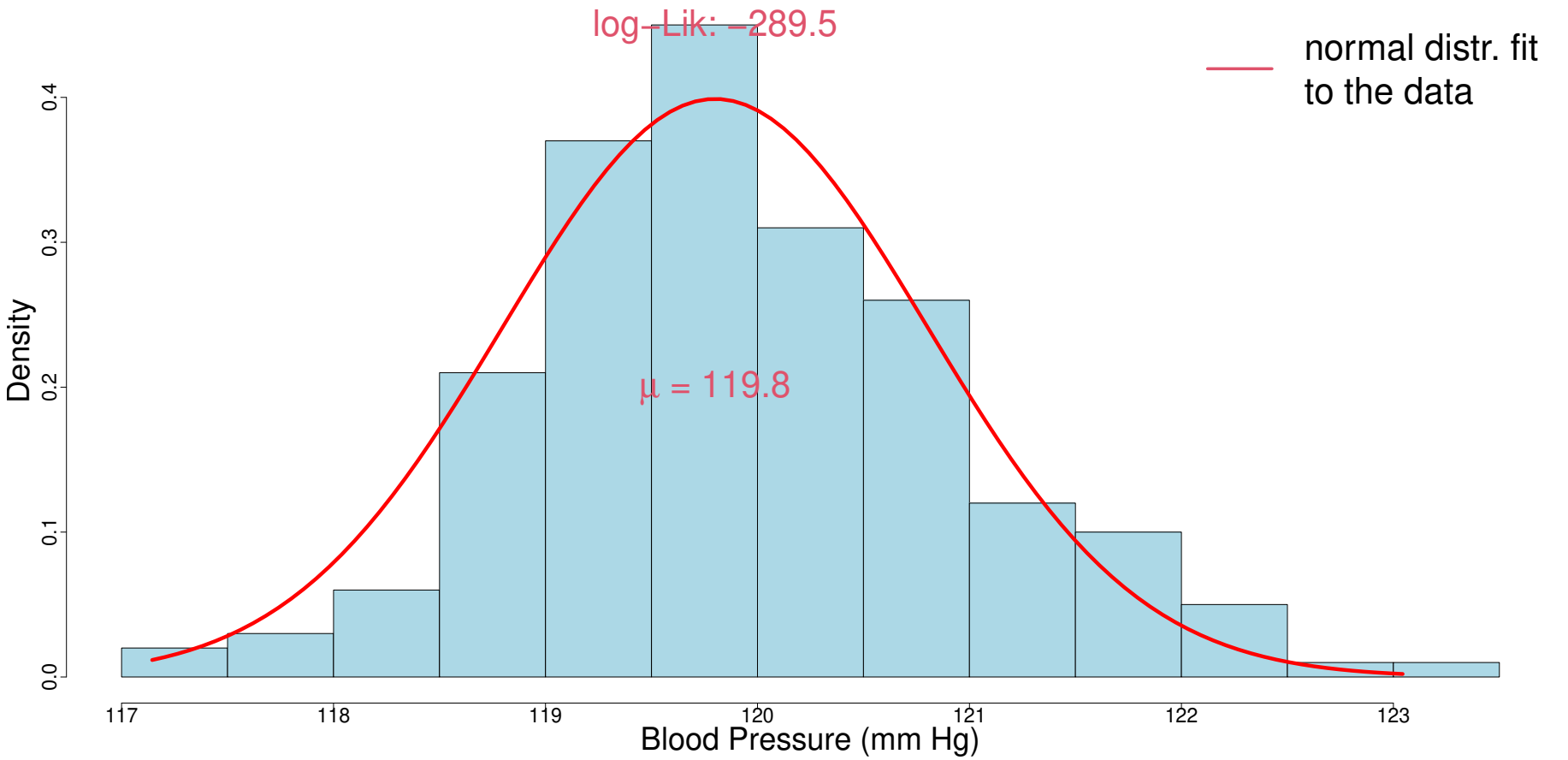
2.3 Maximum Likelihood Estimation (cont'd)



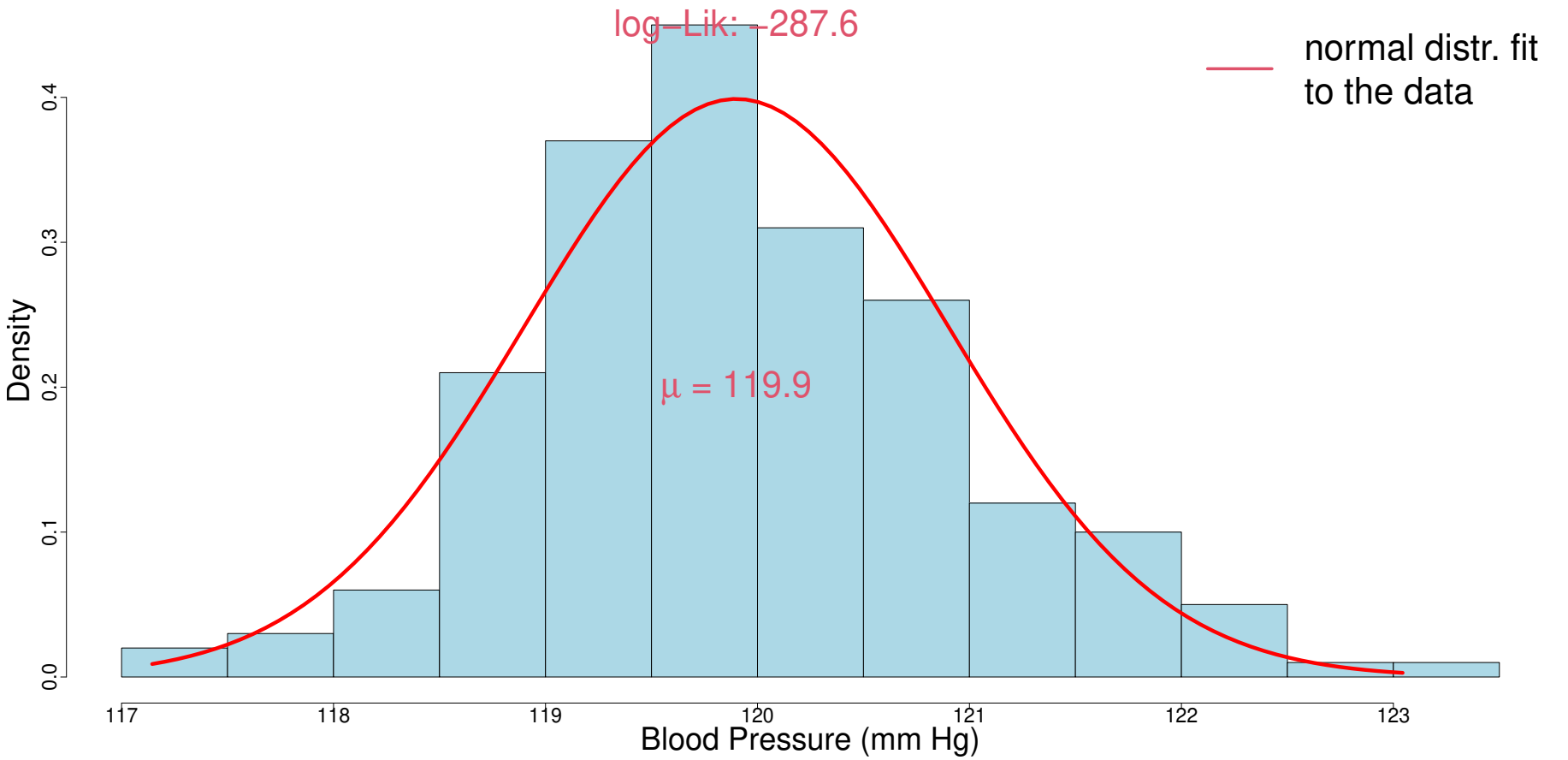
2.3 Maximum Likelihood Estimation (cont'd)



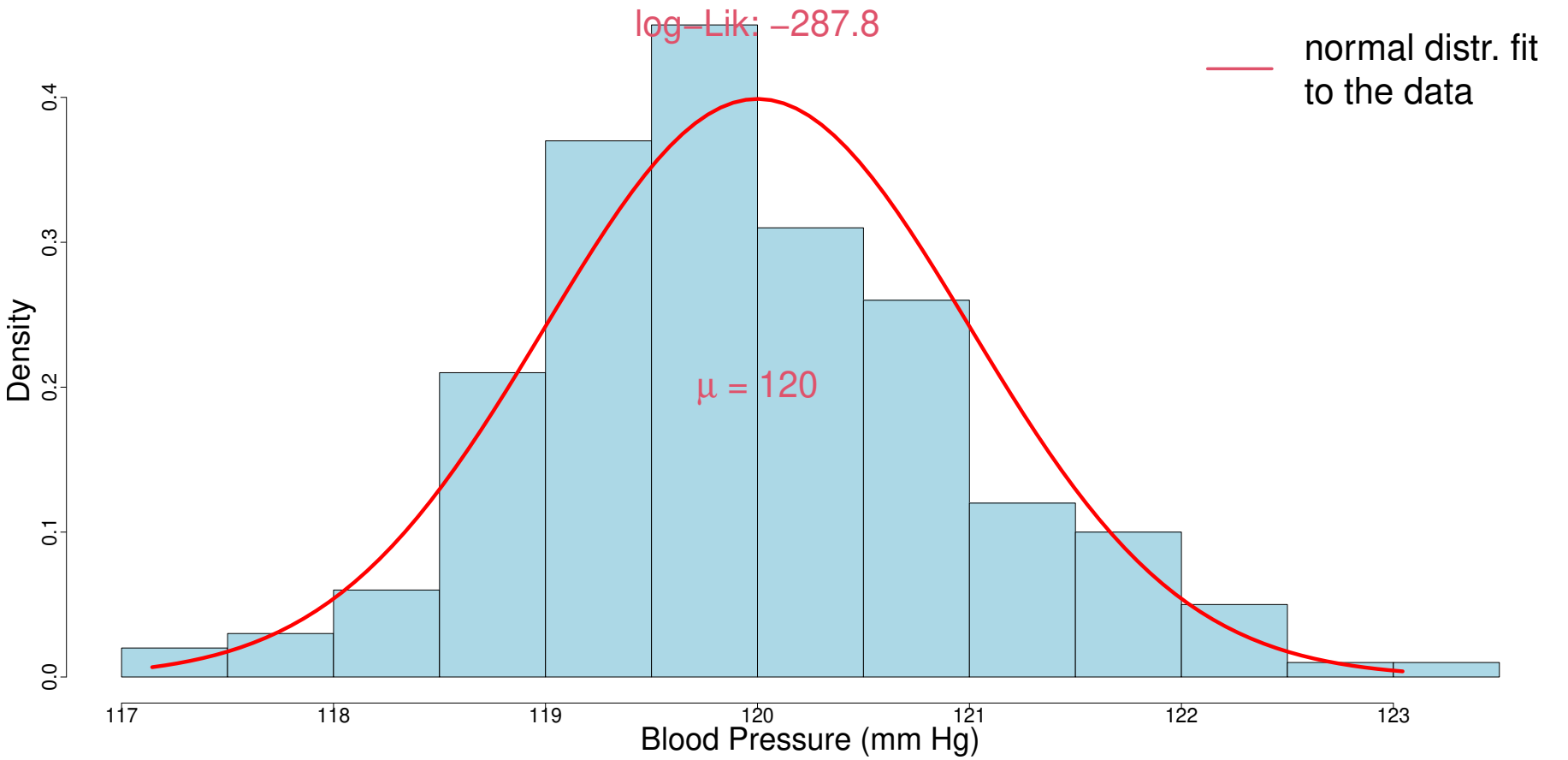
2.3 Maximum Likelihood Estimation (cont'd)



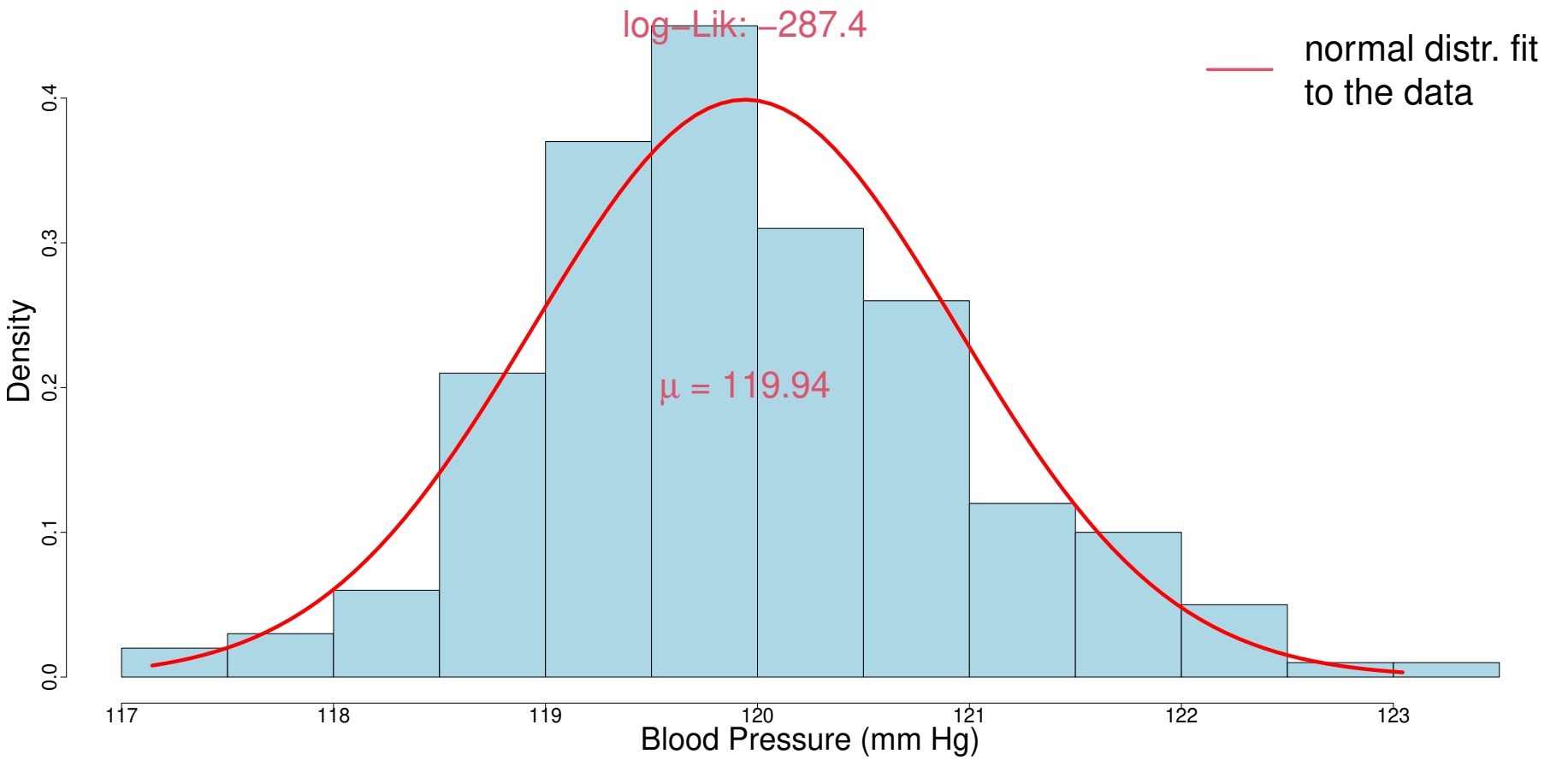
2.3 Maximum Likelihood Estimation (cont'd)



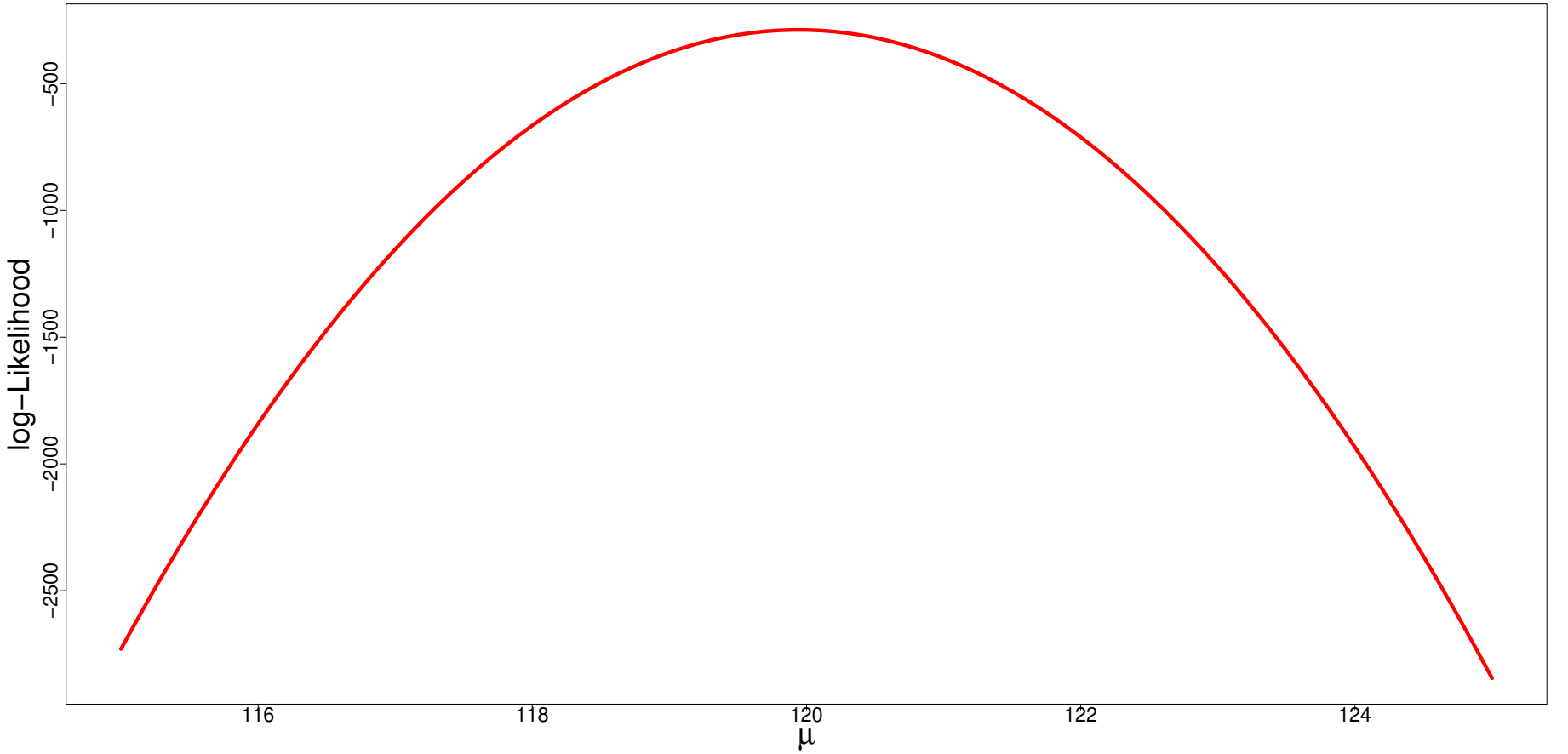
2.3 Maximum Likelihood Estimation (cont'd)



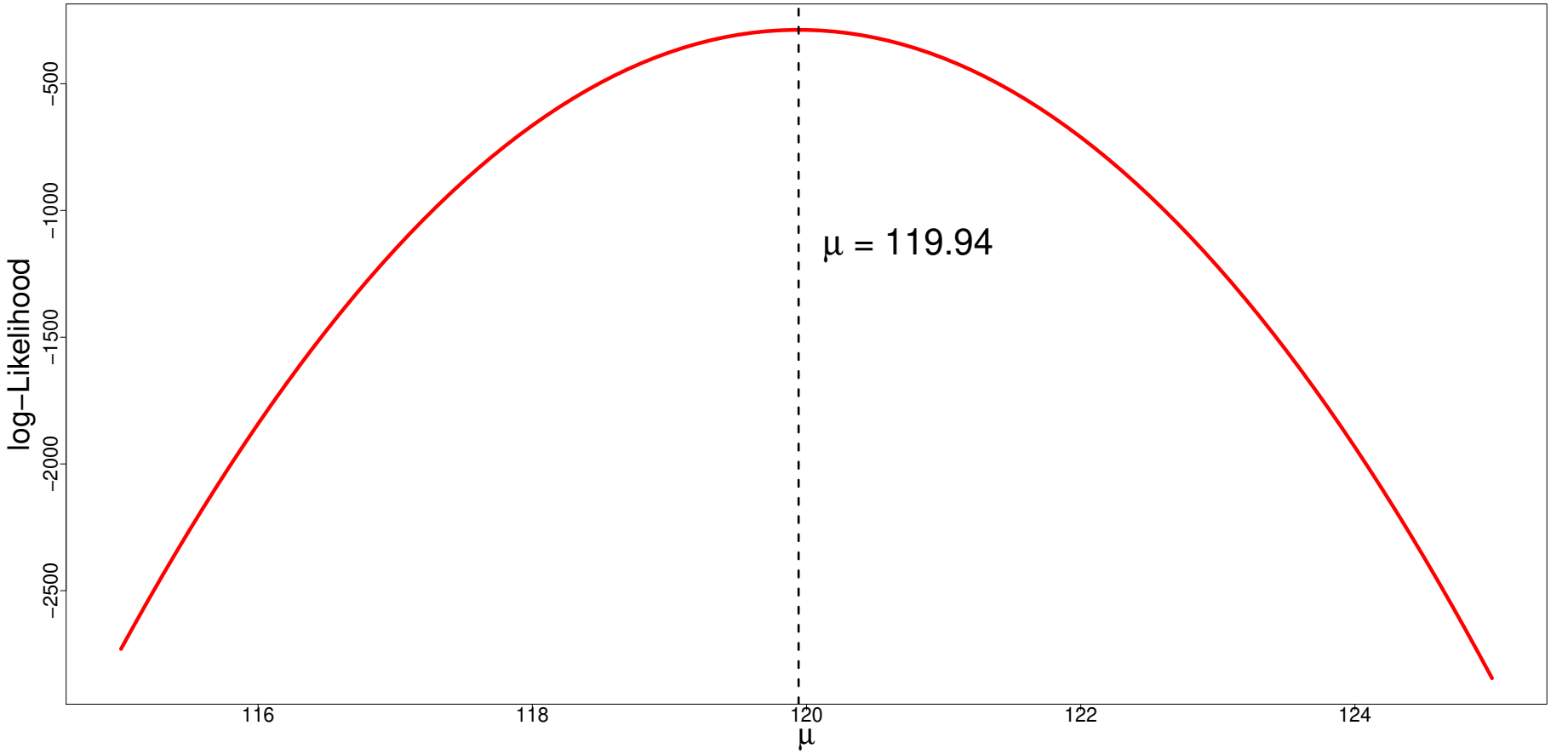
2.3 Maximum Likelihood Estimation (cont'd)



2.3 Maximum Likelihood Estimation (cont'd)



2.3 Maximum Likelihood Estimation (cont'd)



2.3 Maximum Likelihood Estimation (cont'd)

- The same idea can be extended in the case where we have many parameters to estimate

- What statistical software does for us
  - ▷ when we had one parameter we could search 'by hand' to find the value that maximized the log-likelihood;
  - ▷ however, when there are many parameters it is obvious that this **cannot** be done that easily
  - ▷ statistical packages implement efficient algorithms to find these values

2.4 Properties of MLEs

- The values of the parameters which maximize the log-likelihood value are known as the *Maximum Likelihood Estimates (MLEs)*

- Relevant question: We have found the MLEs for the sample of patients we have at hand, but how are these values related to the values from the population?

- Remember: We use statistics to say something about the target population, *which we (almost) never have*, using the sample, *which we do have*

2.4 Properties of MLEs (cont'd)

- Provided that we have a representative sample from the target population and under some general regularity conditions, the MLEs have some nice asymptotic properties (i.e., for large enough sample sizes)

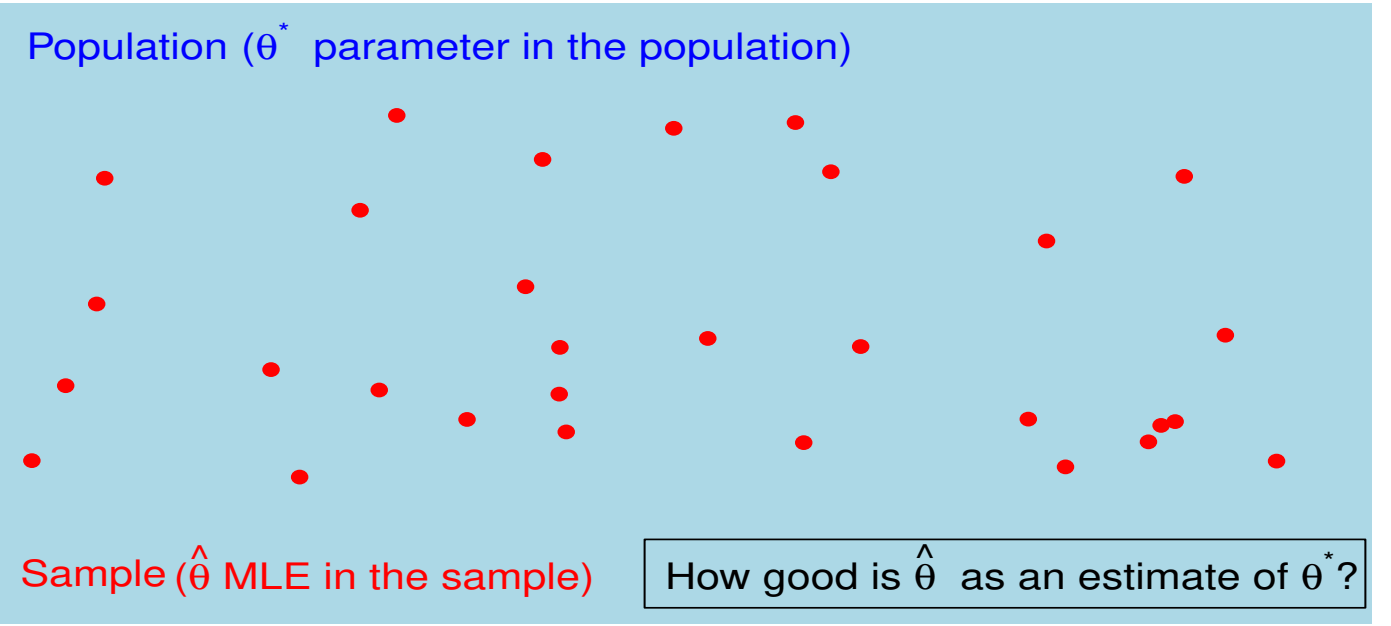
- Say that the phenomenon we wish to study is described by a distribution with parameters θ^*

  ▷ θ^* denotes the true value of the parameters, i.e., the value of the parameters in the population, which we do not have

2.4 Properties of MLEs (cont'd)

- We have a representative sample from this population $x_1, \ldots, x_n$ of size n, based on which we find the MLEs denoted by $\hat{\theta}$

2.4 Properties of MLEs (cont'd)



2.4 Properties of MLEs (cont'd)

- We have a representative sample from this population $x_1, \ldots, x_n$ of size n, based on which we find the MLEs denoted by $\hat{\theta}$

1. *Consistency*: Having a sufficiently large number of observations n, it is possible to find the value of θ^* with arbitrary precision

$$\hat{\theta} \xrightarrow{P} \theta^*$$

(in mathematical terms we say that as n goes to infinity $\hat{\theta}$ converges in probability to its true value θ^*)

▷ This means that the MLEs are asymptotically unbiased

2.4 Properties of MLEs (cont'd)

2. *Efficiency*: The maximum likelihood method estimates the quantity of interest (i.e., θ^*) in the "best possible" manner (with respect to a quadratic loss function)

   (in mathematical terms and as n goes to infinity the MLEs achieve the Cramér–Rao lower bound)

▷ This means that no other asymptotically unbiased estimator has lower asymptotic mean squared error than the MLE, i.e., we estimate θ^* as accurately as possible

3. *Asymptotic Normality*: As the sample size increases, the distribution of the MLE tends to a (multivariate) normal distribution with mean θ^* and covariance matrix equal to the inverse of the Fisher information matrix $\mathcal{I}_{\theta^*}$

$$\hat{\theta} \sim \mathcal{N}\left(\theta^*, \mathcal{I}_{\theta^*}^{-1}\right)$$

where

$$\mathcal{I}_{\theta^*} = E\left\{ -\sum_{i=1}^{n} \frac{\partial^2 \log f(x_i;\theta)}{\partial\theta^\top \partial\theta}\bigg|_{\theta=\theta^*} \right\}$$

▷ The standard errors that are reported by the statistical software are based on an estimate of $\mathcal{I}_{\theta^*}^{-1}$

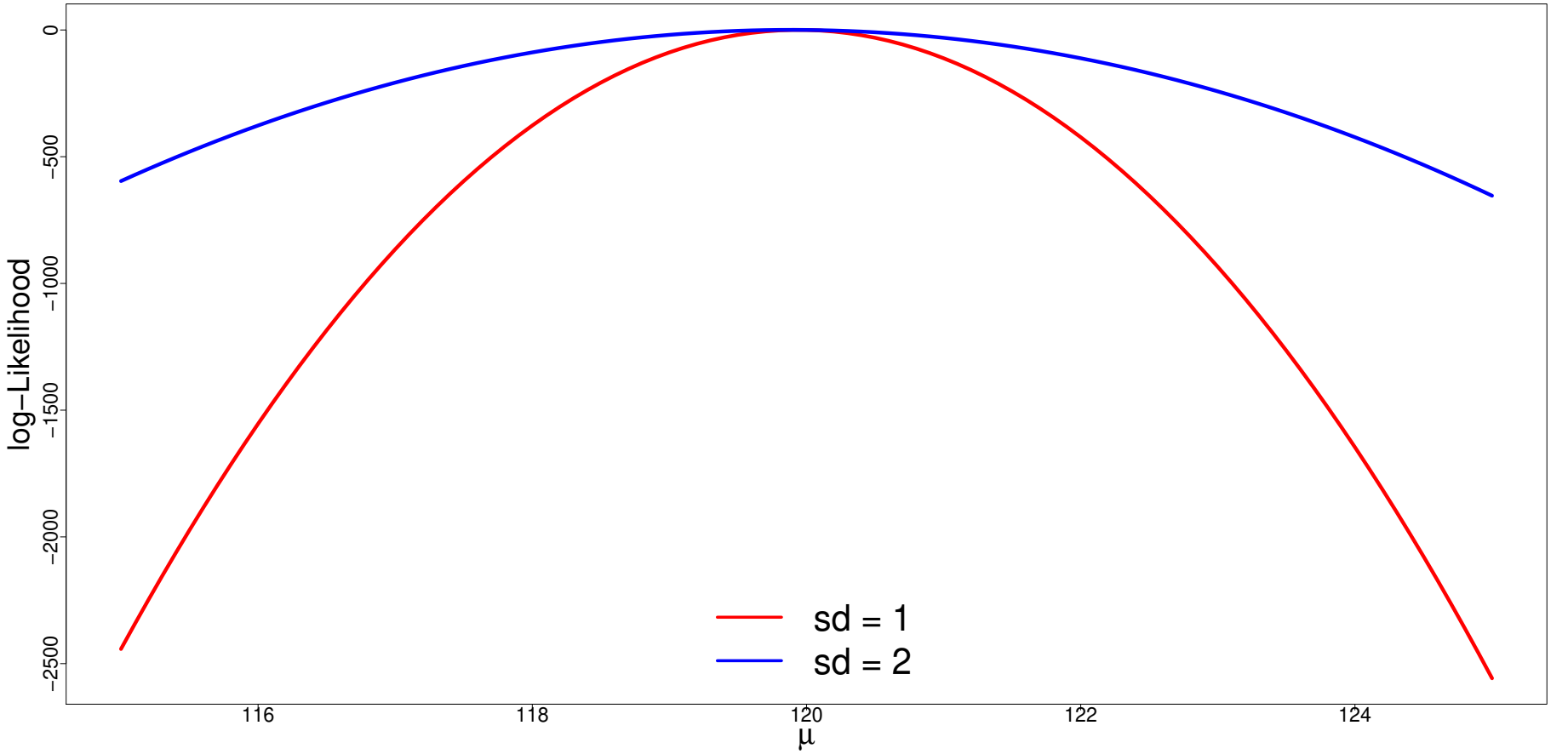
2.4 Properties of MLEs (cont'd)

- The inverse of the Fisher information matrix $\mathcal{I}_{\theta^*}$ describes the local curvature of the log-likelihood function in the neighborhood of θ^*, i.e.,
  - ▷ how peaked or flat the log-likelihood surface is

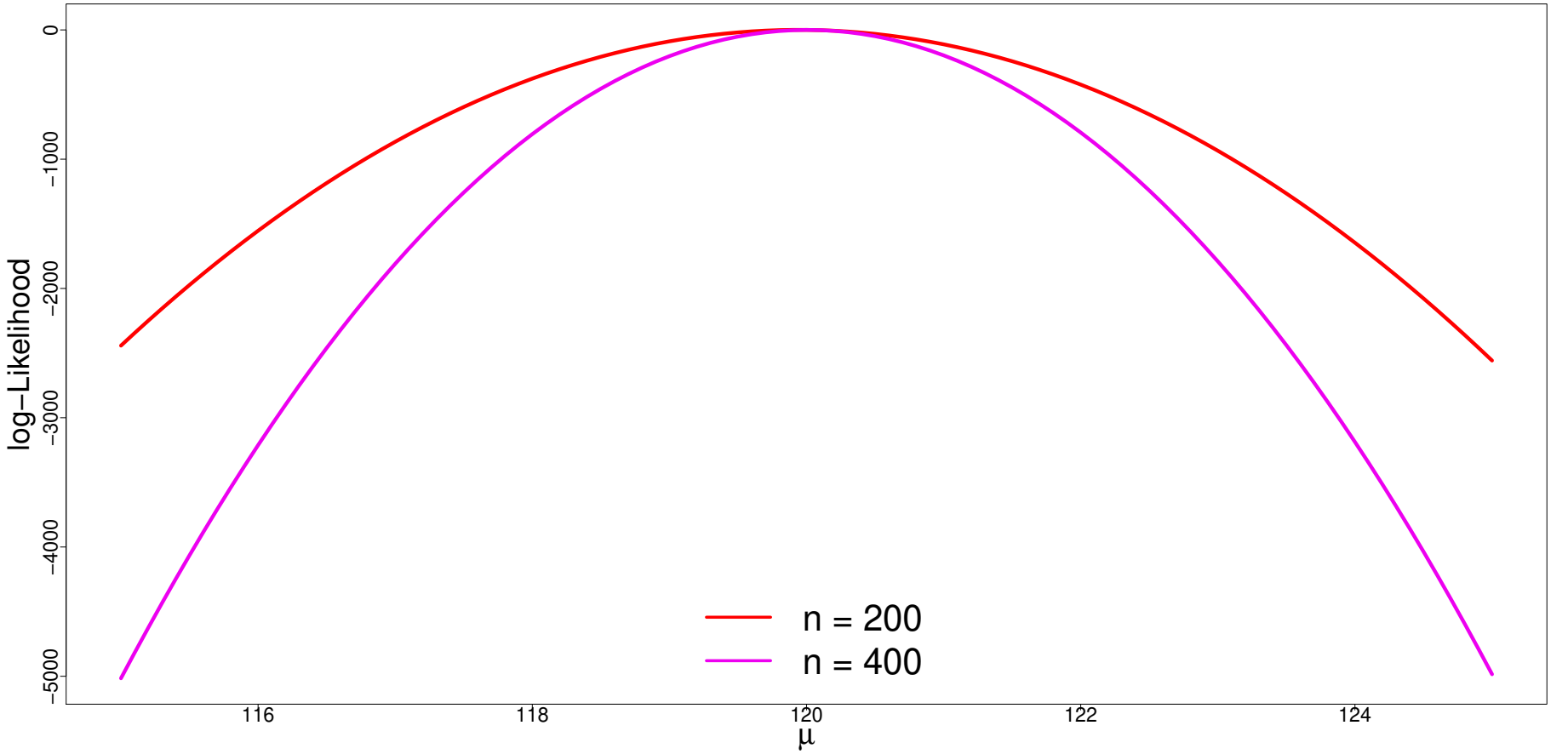
- The more peaked the log-likelihood is, the more accurate the maximum likelihood estimate $\hat{\theta}$ is
  - ▷ smaller standard errors

- The log-likelihood is more peaked as
  - ▷ the variance of the data decreases (something we cannot control)
  - ▷ the sample size increases (something we can control)

2.4 Properties of MLEs (cont'd)



2.4 Properties of MLEs (cont'd)



2.4 Properties of MLEs (cont'd)

- Why are these properties of the MLEs important?

$$\Downarrow$$

Based on these properties we can **make inference**

- In particular, we can
  1. construct confidence intervals
  2. do hypothesis testing and calculate p-values

2.5 Confidence Intervals

- We often want to summarize the information available in the sample for the parameters of interest

- Having established the properties of the MLEs, we can construct intervals for the true parameter θ^* of the population

- Based on the asymptotic normality of $\hat{\theta}$, i.e.,

$$\hat{\theta} \sim \mathcal{N}\left(\theta^*, \mathcal{I}_{\theta^*}^{-1}\right)$$

2.5 Confidence Intervals (cont'd)

A $(1-\alpha)\%$ *Confidence Interval (CI)* for θ^* is constructed as

$$\left\{\hat{\theta} - Z_{1-\alpha/2} \times \text{se}(\hat{\theta}),\quad \hat{\theta} + Z_{1-\alpha/2} \times \text{se}(\hat{\theta})\right\}$$

where

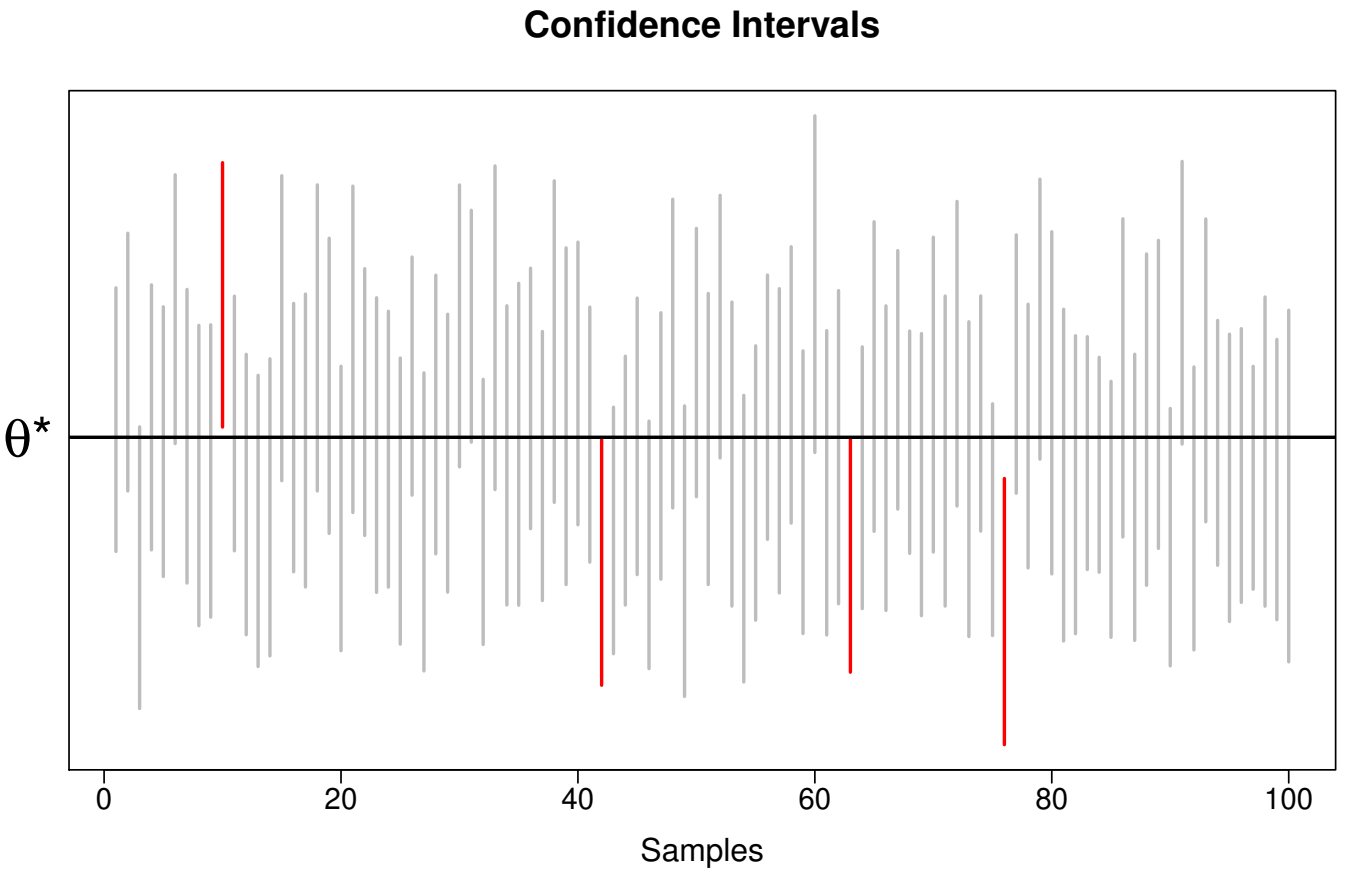
- $Z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal distribution (for a 95% CI, $\alpha = 0.05$, and $Z_{1-\alpha/2} = 1.96$)

- $\text{se}(\hat{\theta})$ denotes the standard error of $\hat{\theta}$, which is the square root of the corresponding diagonal element of the inverse Fisher information matrix $\mathcal{I}^{-1}_{\theta^*}$

2.5 Confidence Intervals (cont'd)

- Notes:
  - ▷ **Interpretation:** The confidence interval width quantifies the degree of precision characterizing the point estimate of interest
    - * e.g., a very wide estimated CI suggests that the results have high degree of variability $\Rightarrow$ low precision
  - ▷ The interval we constructed is random, i.e., if we had another random sample from the same population, we would have obtained another interval
    - * e.g., if we had 100 samples and for each one we construct a 95% CI, we would expect about 95 of them to contain θ^*

2.5 Confidence Intervals (cont'd)



2.5 Confidence Intervals (cont'd)

- **Misinterpretations** of confidence intervals
  $\Rightarrow$ *the following are some incorrect statements about confidence intervals*

  ▷ A 95% confidence level means that 95% of the sample data lie within the confidence interval

  ▷ A 95% confidence level means that for a given realized interval there is a 95% probability that the population parameter lies within the interval

(Note: https://dx.doi.org/10.3758/s13423-013-0572-3)

2.6 Hypothesis Testing

- Very often the motivation behind an analysis is to test a specific hypothesis
  - ▷ Example: Is gender an important risk factor for high blood pressure, having corrected for BMI and Age?

- In general, we are interested in

$$H_0 : \theta = \theta_0$$
$$H_a : \theta \neq \theta_0$$

- Because we have estimated the parameters using maximum likelihood, we have the following three options
  - ▷ Wald test
  - ▷ Likelihood Ratio test
  - ▷ Score test

2.6 Hypothesis Testing (cont'd)

- The Wald test is defined as

$$(\hat{\theta}_a - \theta_0)^\top \{\hat{\text{var}}(\hat{\theta}_a)\}^{-1} (\hat{\theta}_a - \theta_0) \sim \chi^2_p$$

where

  - $\hat{\theta}_a$ the maximum likelihood estimate under the alternative hypothesis
  - $\hat{\text{var}}(\hat{\theta}_a) = \left[E\left\{ -\partial^2 \ell(\theta) / \partial\theta^\top \partial\theta \big|_{\theta=\hat{\theta}_a} \right\} \right]^{-1}$ denotes the covariance matrix of the MLEs
  - p denotes the number of parameters being tested

- The Wald test requires estimating the distribution's parameter under the alternative hypothesis

2.6 Hypothesis Testing (cont'd)

- Note:
  - ▷ the χ^2_p distribution is the distribution of the sum of squares of p standard normal distributions (i.e., std. normal = normal distribution with mean 0 and variance 1)
  - ▷ it is one of the most widely used probability distributions in inferential statistics

2.6 Hypothesis Testing (cont'd)

- The score test is defined as

$$\mathcal{S}(\hat{\theta}_0)^\top \hat{\text{var}}(\hat{\theta}_0)\mathcal{S}(\hat{\theta}_0) \sim \chi^2_p$$

  where
  - ▷ $\hat{\theta}_0$ the maximum likelihood estimate under the null hypothesis
  - ▷ $\mathcal{S}(\theta) = \partial\ell(\theta)/\partial\theta^\top$ denotes the score vector
  - ▷ $\hat{\text{var}}(\hat{\theta}_0)$ denotes the covariance matrix of the MLEs
  - ▷ p denotes the number of parameters being tested

- The score test requires estimating the distribution's parameter under the null hypothesis

2.6 Hypothesis Testing (cont'd)

- The likelihood ratio test (LRT) is defined as

$$-2 \times \{\ell(\hat{\theta}_0) - \ell(\hat{\theta}_a)\} \sim \chi^2_p$$

  where

  ▷ $\ell(\cdot)$ the value of the log-likelihood function

  ▷ $\hat{\theta}_0$ the maximum likelihood estimate under the null hypothesis

  ▷ $\hat{\theta}_a$ the maximum likelihood estimate under the alternative hypothesis

  ▷ p denotes the number of parameters being tested

- The LRT requires estimating the distribution's parameter under both the null & alternative hypotheses

2.6 Hypothesis Testing (cont'd)

- Asymptotically (i.e., for large samples) these three tests are equivalent

- **Advice:** Prefer to use the likelihood ratio test over the other two

- Why:
  - ▷ it has better theoretical properties
  - ▷ it makes you carefully think about the hypotheses being tested

2.6 Hypothesis Testing (cont'd)

- Based on either of the three tests and the formulated null and alternative hypotheses, we can calculate a p-value

- How does this work and what is a p-value?

  1. We start by assuming that the null hypothesis is correct, i.e., $\theta = \theta_0$

  2. Under this assumption, we determine the sampling distribution of our test statistic
     - ▷ what are the values for the test statistic from all possible samples from the target population in which the null hypothesis holds
     - ▷ this is actually the χ^2_p distribution

- How does this work and what is a p-value?

3. We then calculate the probability of obtaining test results at least as extreme as the results observed in the original sample, under the assumption that the null hypothesis is correct

$$p\text{-value} = \Pr(\text{extreme test results} \mid \text{null is correct})$$

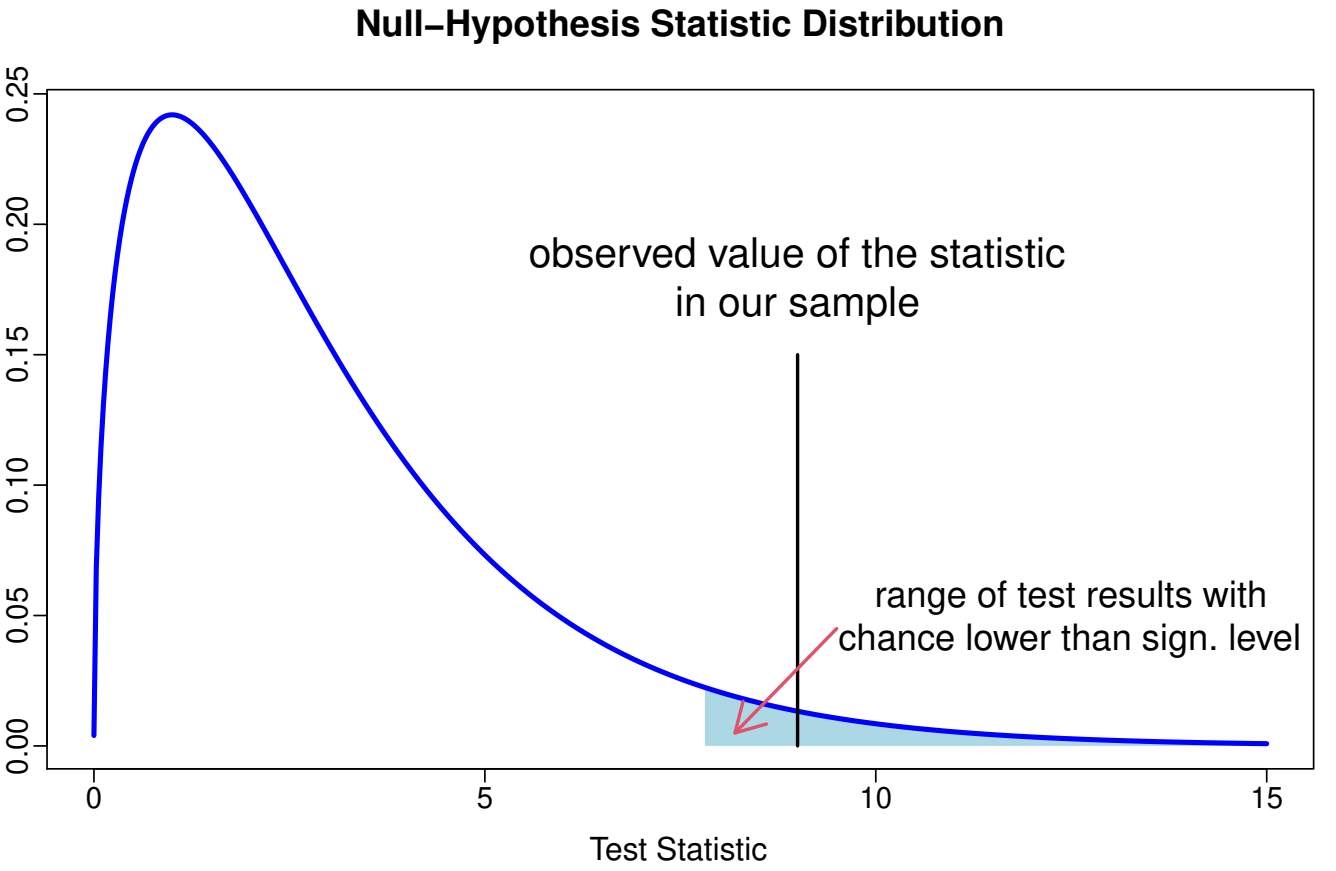
2.6 Hypothesis Testing (cont'd)

- To translate this probability estimate into a decision we need a cutoff point
  - ▷ if the p-value is sufficiently small, then this would be considered as evidence against the null hypothesis

$$p\text{-value} = \Pr(\text{extreme test results} \mid \text{null is correct}) < \alpha$$

α is called the *significance level*

2.6 Hypothesis Testing (cont'd)



2.6 Hypothesis Testing (cont'd)

The manner we take this decision may lead to two possible errors

2.6 Hypothesis Testing (cont'd)

- Definition: *Type I error* occurs when we falsely reject the null hypothesis (i.e., the null hypothesis is true and we reject it)
  - ▷ the chance of this error equals the significance level α
  - ▷ it is typically pre-specified by the researchers

- Definition: *Type II error* occurs when we falsely do not reject the null hypothesis (i.e., we do not reject the null hypothesis despite that it is false)
  - ▷ the chance of this error decreases with increasing sample size
  - ▷ it is typically denoted by β
  - ▷ the power of a study is $1 - \beta$

2.6 Hypothesis Testing (cont'd)

- **Misinterpretations** of p-values
  $\Rightarrow$ *the following are some incorrect statements about p-values*

  ▷ The p-value is the probability that the null hypothesis is true, or the probability that the alternative hypothesis is false

  ▷ The p-value is the probability that the observed effects were produced by random chance alone

  ▷ Smaller p-values imply the presence of a larger or more important effects

(Note: http://dx.doi.org/10.1080/00031305.2016.1154108)