Maximum Likelihood Estimation

Lecture 7

Please, sign in on iClicker

Outline

1. Independence for continuous random variable
2. Random samples
3. Estimating true parameters
4. Maximum likelihood estimation (MLE)
    - MLE using a range of potential values
    - Analytical method for MLE
    - Numerical methods for MLE

1. Independence for Continuous Random Variable

- In the discrete case, we have that:

$$P(X = x \cap Y = y) = P(X = x) \cdot P(Y = y).$$

- Equivalently,

$$P(Y = y \mid X = x) = P(Y = y).$$

Definition in the Continuous Case

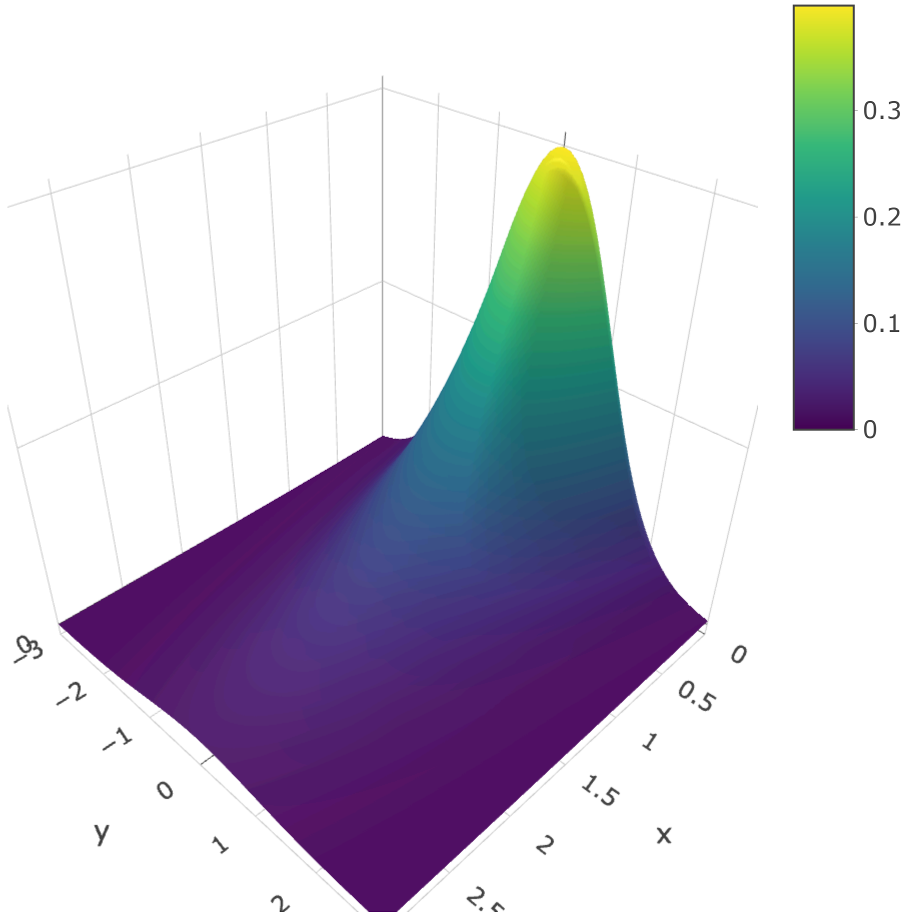
- Probabilities become densities!
- The definition becomes

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y).$$

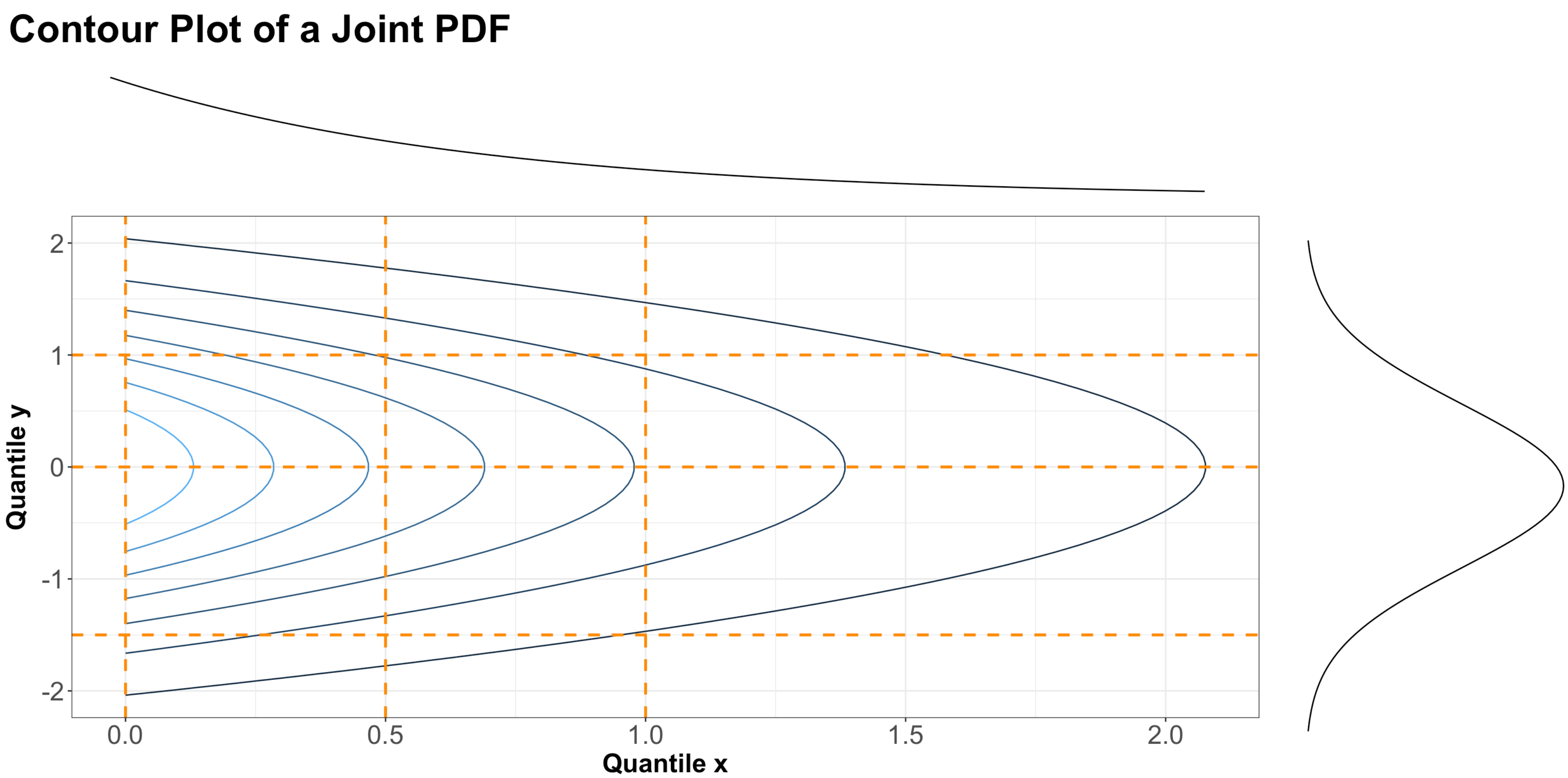
- Equivalently,

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{f_X(x) \cdot f_Y(y)}{f_X(x)} = f_Y(y).$$

Recall that we can represent a bivariate density function in 3D



The previous 3D plot also has a contour version

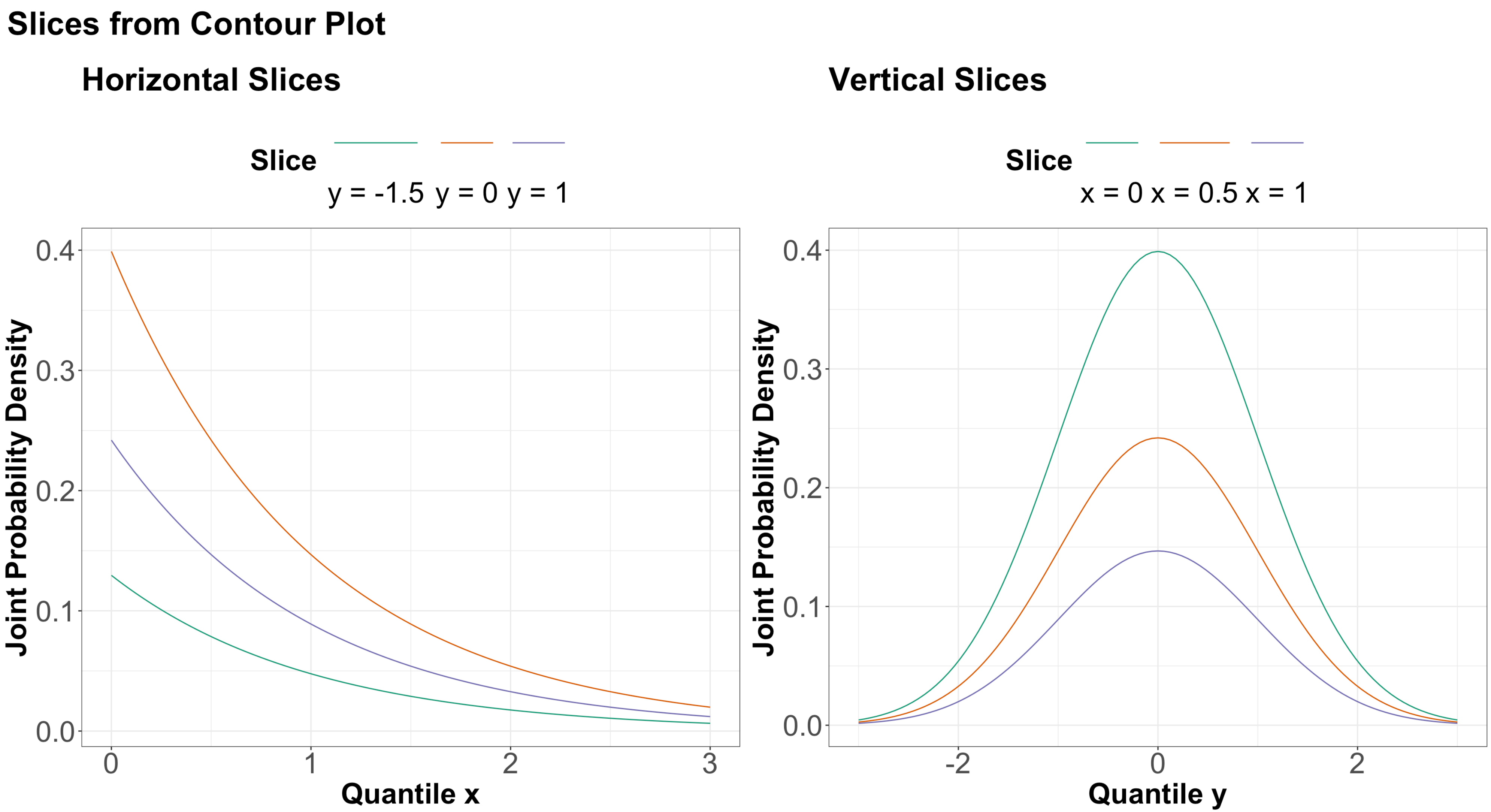


- Are X and Y independent?

Independence Visualized

- We can tell whether two random variables are independent, by looking at "slices" in a contour plot of a bivariate density function.

Viewing the Slices



What is going on with the slices?

- Every slice has the same shape.
- They are all the same as the marginal density after normalization.
- Recall conditional density

$$f_{Y|X}(y|x) = f_Y(y).$$

So they are independent!!

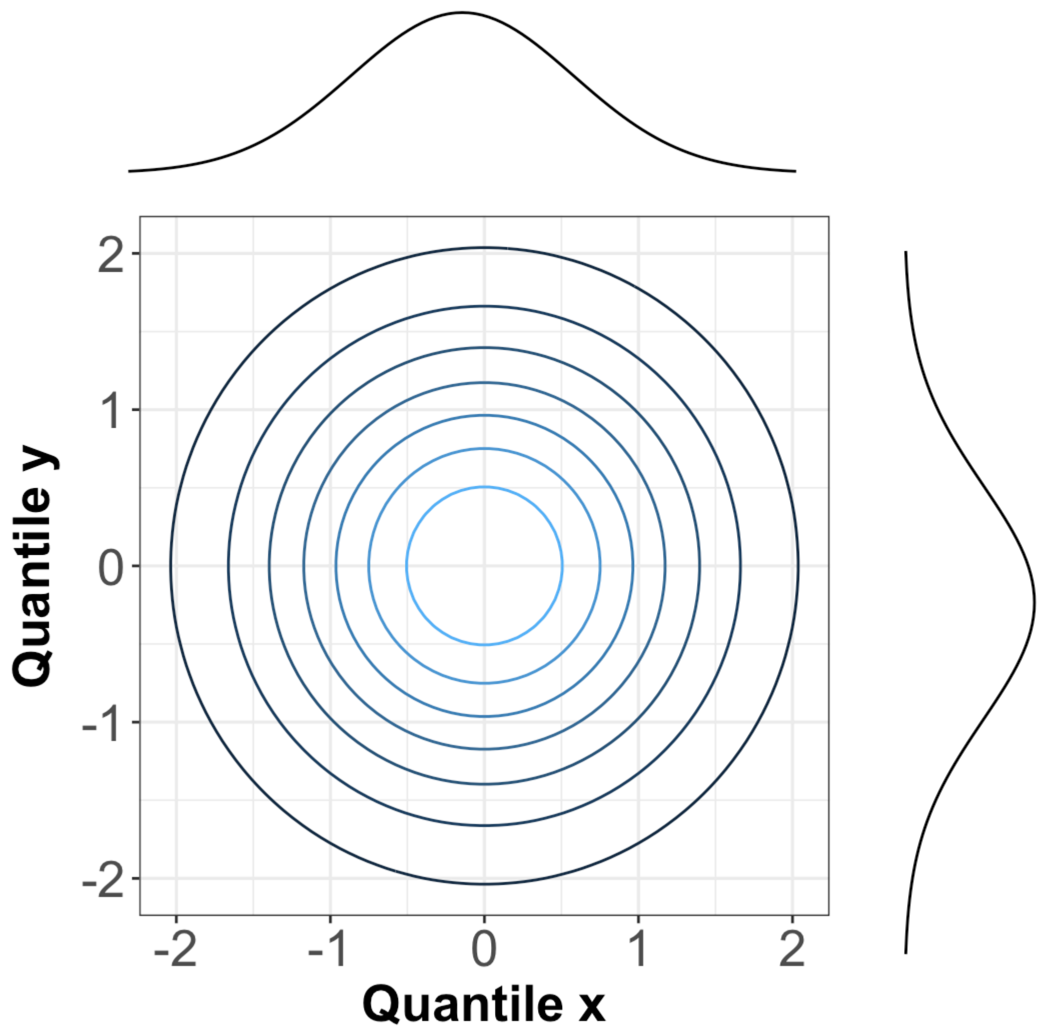
- In fact, X and Y are two independent random variables:

$$X \sim \text{Exponential}(\lambda = 1)$$
$$Y \sim \mathcal{N}\left(\mu = 0, \sigma^2 = 1\right).$$

with the joint PDF:

$$\begin{aligned} f_{X,Y}(x, y) &= [\lambda \exp(-\lambda x)] \times \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2} \right] \right\} \\ &= \exp(-x) \times \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2} \right) \right]. \end{aligned}$$

Another Example



They are also independent!!

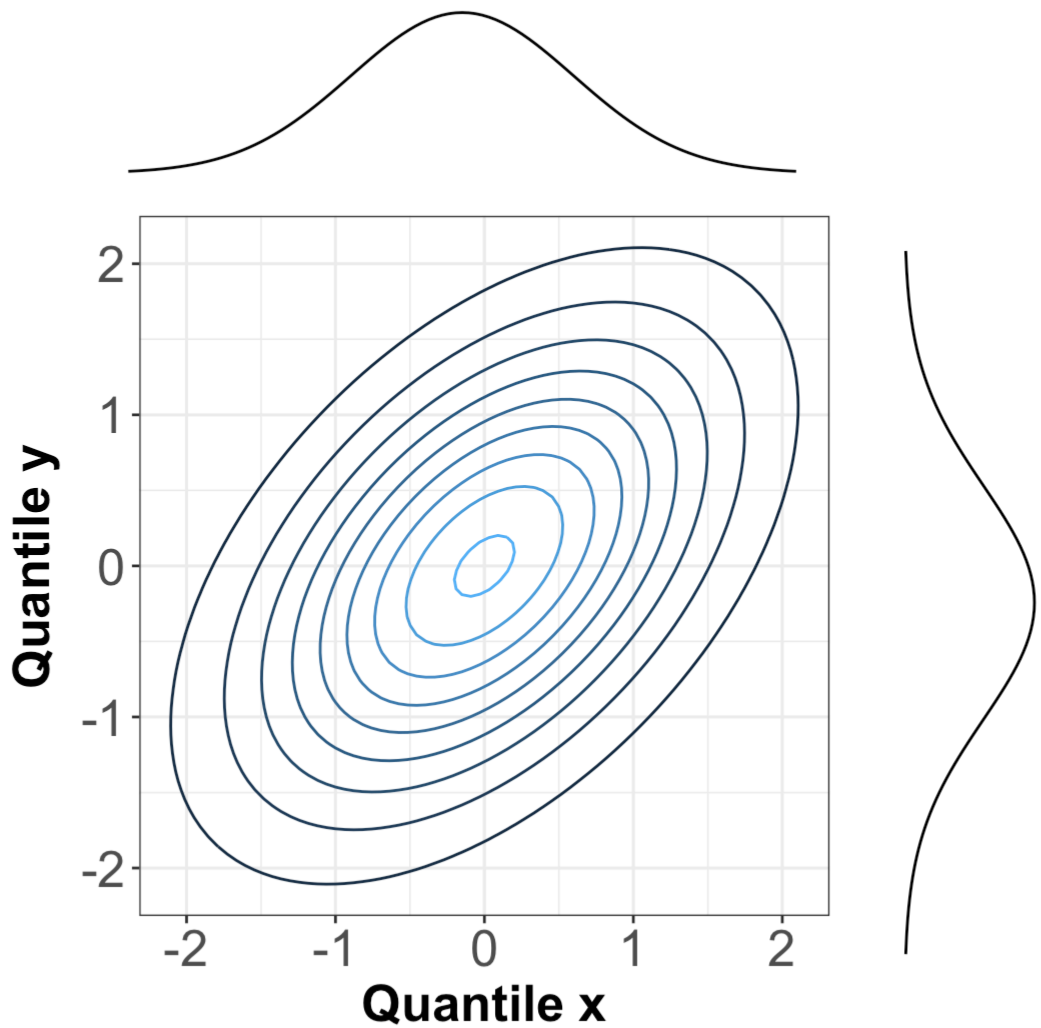
- X and Y are two independent Standard Normal distributions:

$$\begin{aligned} X &\sim \mathcal{N}(0,1) \\ Y &\sim \mathcal{N}(0,1). \end{aligned}$$

with the joint PDF:

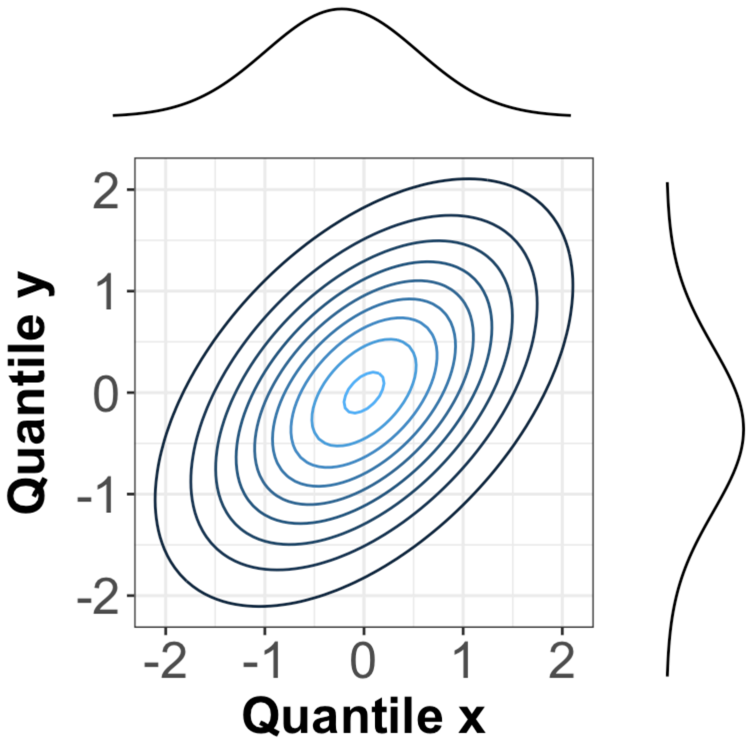
$$f_{X,Y}(x,y) = \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)\right] \times \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)\right].$$

Another example



They are NOT independent

- If X and Y are not independent, then these contours would appear in a diagonal pattern.



2. Random Samples

- A **random sample** is a collection of random variables.
  - For example, $X_1, X_2, \ldots, X_n$.
- We think of data as being a random sample.

Independent and identically distributed

- We often assume a random sample is **independent and identically distributed** (or **iid**):
    - Each pair of random variables are **independent**.
    - Each random variable follows **the same distribution.**

3. Estimating true parameters

- We can model real-world data with specific distributions.
  - Wait time can be modeled by a Exponential distribution.
- We can calculate probabilistic quantities from a distribution.
  - The mean of a Exponential RV is β.
- However, in practice, we will never know these true parameters. We need to estimate them.

How can we estimate the true parameter?

- Using data: **a random sample of n observations**
- Given the random sample $Y_1, \ldots, Y_n$, calculate the sample mean:

$$\bar{Y} = \sum_{i=1}^{n} \frac{Y_i}{n}$$

- We hope $\bar{Y}$ is a good estimator for the mean β.

Finding good estimators can be a difficult task

- For example, β_0 and β_1 from a linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon.$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

- How can we estimate these parameters?

Overview of Estimation Methods

- Point estimation
  - **Maximum likelihood estimation**
  - Method of moments
  - Bayesian estimation (Inference II)
  - Least squares estimation (Regression I)
  - EM algorithm (Unsupervised Learning)
- Interval estimation (Inference I)

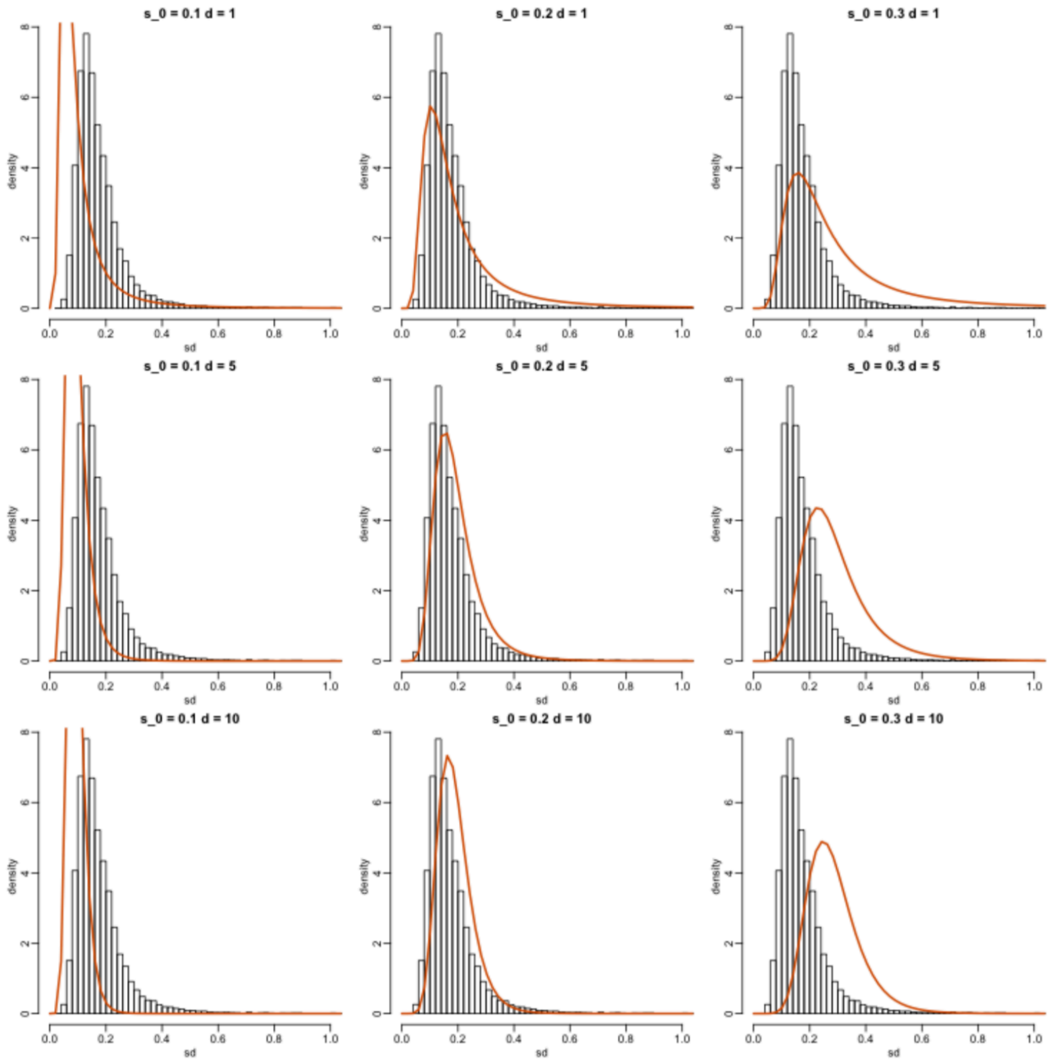
A note on the scope of the course

- In this course (including `lab4` and `quiz2`), we will only focus on estimation in univariate cases.

4. Maximum likelihood estimation

- Given **observed data and distributional assumptions**, we want to find the values of the parameters that would make the observed data **most likely to have occurred**.

Which orange line (PDF) fits the histogram (observed data) better?



Histograms of sample standard deviations and densities of estimated distributions.

Likelihood Function

- The likelihood function represents the probability of the observed data as a function of the parameter(s).
- A likelihood function is constructed from the joint PDF/PMF:

$$\mathcal{L}(\beta) = f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \beta)$$

Overview of MLE

- Collect data
- Make a distributional assumption for the data
- Use the **likelihood function** to find the parameters that best fits the data

Example

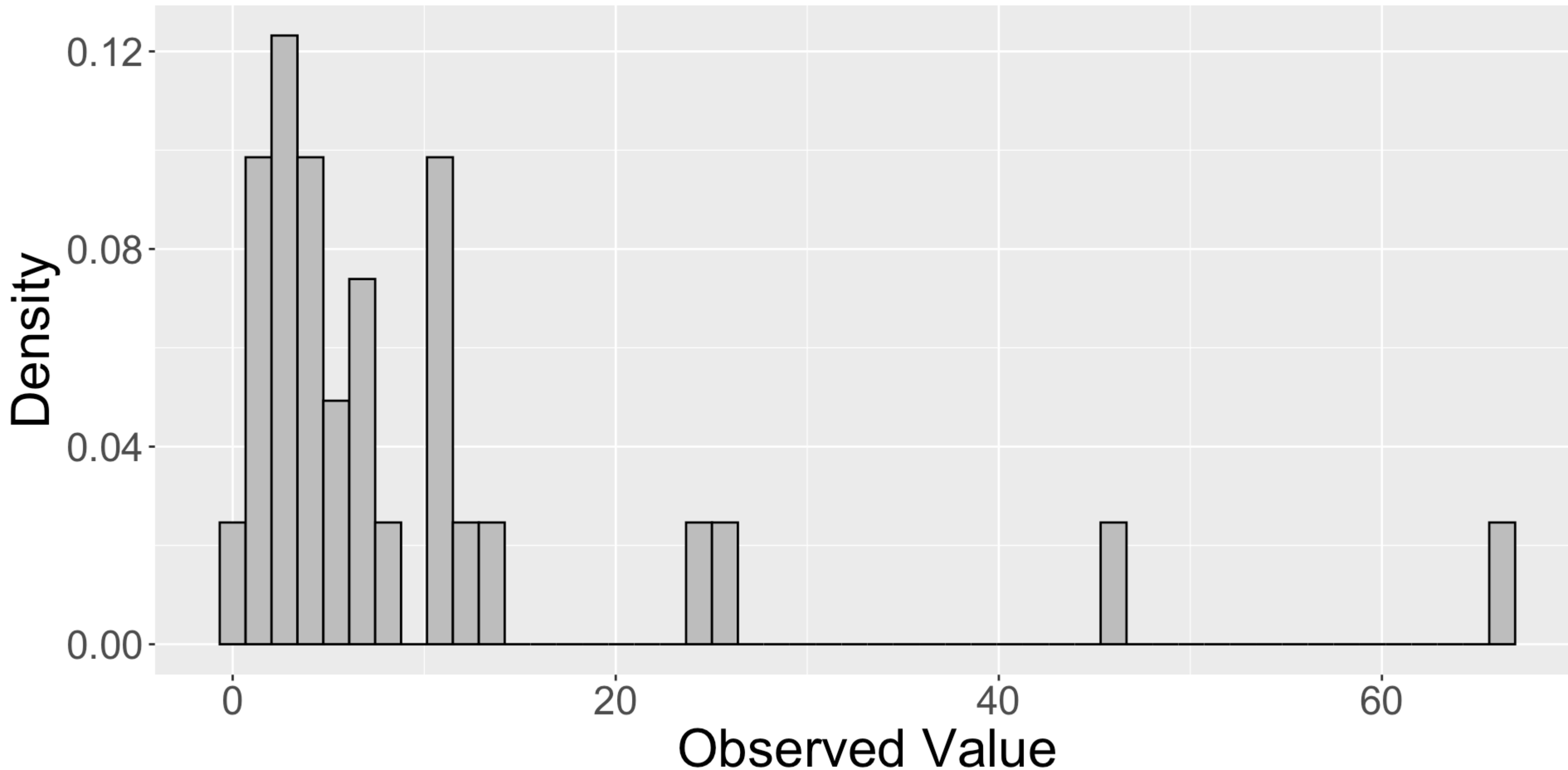
- Suppose you own many ice cream carts.
- You want to estimate **the average wait time from one customer to the next one in a given cart.**

How can we estimate the mean wait time β?

- Step 1: Collect data!
- You implemented a simple random sampling and got a sample of size $n = 30$ wait times (in minutes)

```
sample_n30 <- tibble(values = c(
  24.9458614574341, 7.23174970992907, 4.16136401519179, 5.60304128237143,
  5.37929488345981, 1.40547217217847, 7.0701988485075, 2.84055356831115,
  0.894746121019125, 2.9016381111011, 3.19011222943664, 11.093013768209 9,
  3.49700326472521, 46.2914818498428, 2.00653892990149, 2.87363994969391,
  11.4050390862658, 11.6616687767937, 12.8855835341646, 3.88483320176601,
  0.406148910522461, 25.7642258988289, 8.4743227359272, 4.17410666868091,
  1.84968510270119, 2.15972620035141, 10.5289600339151, 6.44162824716339,
  10.603532313964 5, 66.6861112673485
))
```

The histogram shows the empirical distribution:



Step 2: Choose the right distribution

- What distribution would you choose to model the distribution?

Discussion

- **Besides the Exponential distribution**, what other suitable distribution can we used?
  A. Poisson
  B. Log-Normal
  C. Binomial
  D. Weibull
- We aim to model **continuous and non-negative** data.
- Log-Normal and Weibull have been used to **model wait times for something to happen**.

Mathematical formulation

- We have a random sample of $n = 30$ iid random variables:

$$Y_1, Y_2, Y_3, \ldots, Y_{28}, Y_{29}, Y_{30}$$

- We make the distributional assumption:

$$Y_i \sim \text{Exponential}(\beta) \quad \text{for } i = 1, 2, \ldots 30$$

- We want to estimate β.

Step 3: Compute likelihood function from the data

Compute Likelihood using R

- Let us try to choose some values for β, and then calculate the likelihood.

```r
likelihood_20 <- prod(dexp(sample_n30$values, rate = 1 / 20))
likelihood_20
```

```
[1] 1.880207e-46
```

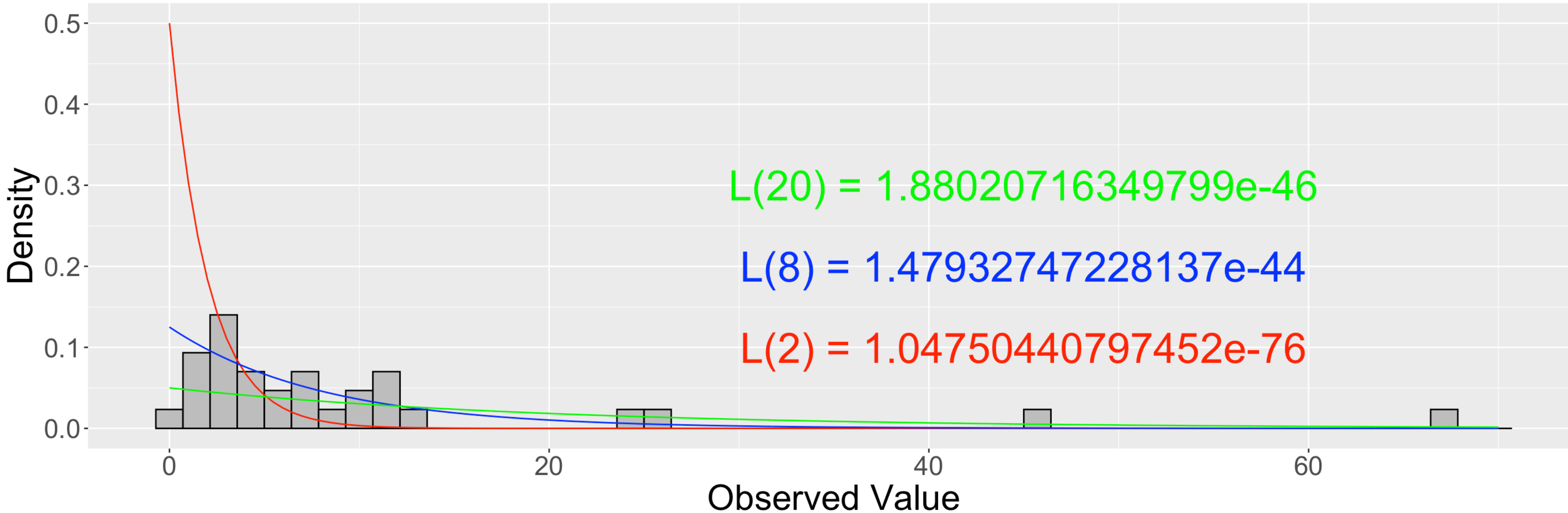
```r
likelihood_8 <- prod(dexp(sample_n30$values, rate = 1 / 8))
likelihood_8
```

```
[1] 1.479327e-44
```

```r
likelihood_2 <- prod(dexp(sample_n30$values, rate = 1 / 2))
likelihood_2
```

```
[1] 1.047504e-76
```

Compare Exponential(β) to the empirical distribution



- $\beta = 8$ minutes has the LARGEST likelihood and fits the observed data.

Issues with likelihood


```r
likelihood_20 # beta = 20
```

```
[1] 1.880207e-46
```

```r
likelihood_8 # beta = 8
```

```
[1] 1.479327e-44
```

```r
likelihood_2 # beta = 2
```

```
[1] 1.047504e-76
```

- The values are very small!

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \underbrace{\frac{1}{\beta}\exp(-y_i/\beta)}_{<1} \to 0.$$

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \underbrace{\frac{1}{\beta}\exp(-y_i/\beta)}_{>1} \to \infty.$$

We often use the log-likelihood (with base e): $\log \mathcal{L}(\beta)$

```r
round(log(likelihood_20), 4)
```

```
[1] -105.2875
```

```r
round(log(likelihood_8), 4)
```

```
[1] -100.9222
```

```r
round(log(likelihood_2), 4)
```

```
[1] -174.9501
```

- The use of the log-likelihood function is common **for numerical stability**.

Now, how can we find the parameters with the maximum log-likelihood?

- Finding the maximum log-likelihood from a set of values
- Analytical solution (closed-form solution)
- Numerical methods

4.1 MLE from a set of potential values

- We can calculate the log-likelihood for a range of β from 5 to 50 by 0.5.

```r
exp_values <- tibble(
  possible_betas = seq(5, 50, 0.5),
  likelihood = map_dbl(1 / possible_betas, ~ prod(dexp(sample_n30$values, .
  log_likelihood = map_dbl(1 / possible_betas, ~ log(prod(dexp(sample_n30$v
)
head(exp_values)
```

```
# A tibble: 6 × 3
  possible_betas likelihood log_likelihood
           <dbl>      <dbl>          <dbl>
1            5     1.78e-48          -110.
2            5.5   2.78e-47          -107.
3            6     2.18e-46          -105.
4            6.5   1.03e-45          -104.
5            7     3.30e-45          -102.
6            7.5   7.85e-45          -102.
```

The MLE estimator of the mean wait time is 10.5 minutes

```r
empirical_MLE <- exp_values %>%
  arrange(desc(likelihood)) %>%
  slice(1)
empirical_MLE
```

```
# A tibble: 1 × 3
  possible_betas likelihood log_likelihood
           <dbl>      <dbl>          <dbl>
1           10.5   4.09e-44          -99.9
```

The log-likelihood plot and MLE

- We plot all possible β's with the corresponding log-likelihood, and blue line indicates the MLE estimator.

4.2 Analytical solution for MLE

- Even though we try a wide range of β values, this does not guarantee an optimal solution.

- In some cases, we can find the closed-form solution.

MLE as an optimization problem

- MLE is an optimization problem:

$$\max_{\beta} \log \mathcal{L}(\beta).$$

- We take the first derivative, set the derivative equal to zero, and solve for β (critical points):

$$\frac{\partial}{\partial \beta} \log \mathcal{L}(\beta) = 0.$$

Analytical solution for Exponential distribution

- The likelihood for β in the Exponential distribution:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \frac{1}{\beta} \exp(-y_i/\beta) = \frac{1}{\beta^n} \exp\left(-\frac{1}{\beta}\sum_{i=1}^{n} y_i\right).$$

- Log-likelihood function:

$$\log \mathcal{L}(\beta) = -n \log(\beta) - \frac{1}{\beta}\sum_{i=1}^{n} y_i.$$

From

$$\log \mathcal{L}(\beta) = -n \log(\beta) - \frac{1}{\beta} \sum_{i=1}^{n} y_i.$$

We take the first partial derivative with respect to β:

$$\begin{aligned}
\frac{\partial}{\partial \beta} \log \mathcal{L}(\beta) &= -\frac{n}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^{n} y_i \\
&= \frac{1}{\beta} \left(-n + \frac{1}{\beta} \sum_{i=1}^{n} y_i \right).
\end{aligned}$$

Set this derivative equal to zero and solve for β

$$\frac{1}{\beta}\left(-n + \frac{1}{\beta}\sum_{i=1}^{n} y_i\right) = 0$$

$$\rightarrow -n + \frac{1}{\beta}\sum_{i=1}^{n} y_i = 0$$

$$\rightarrow \frac{1}{\beta}\sum_{i=1}^{n} y_i = n \rightarrow \beta = \sum_{i=1}^{n} y_i/n = \bar{y}$$

Second derivative test (optional)

- If the second derivative is less than zero evaluated at the MLE estimate, then **the estimate is a local maximum.**

$$
\begin{aligned}
\frac{\partial^2}{\partial \beta} \log \mathcal{L}(\beta) &= \frac{n}{\beta^2} - \frac{2}{\beta^3} \sum_{i=1}^{n} y_i \\
&= \frac{n}{\left(\frac{\sum_{i=1}^{n} y_i}{n} \right)^2} - \frac{2}{\left(\frac{\sum_{i=1}^{n} y_i}{n} \right)^3} \sum_{i=1}^{n} y_i \\
&= \frac{n^3}{\left(\sum_{i=1}^{n} y_i\right)^2} - \frac{2n^3}{\left(\sum_{i=1}^{n} y_i\right)^2} = -\frac{n^3}{\left(\sum_{i=1}^{n} y_i\right)^2} < 0.
\end{aligned}
$$

Sample mean is the MLE for β!

- For the Exponential distribution, we have shown that

$$\hat{\beta} = \bar{Y} = \sum_{i=1}^{n} \frac{Y_i}{n},$$

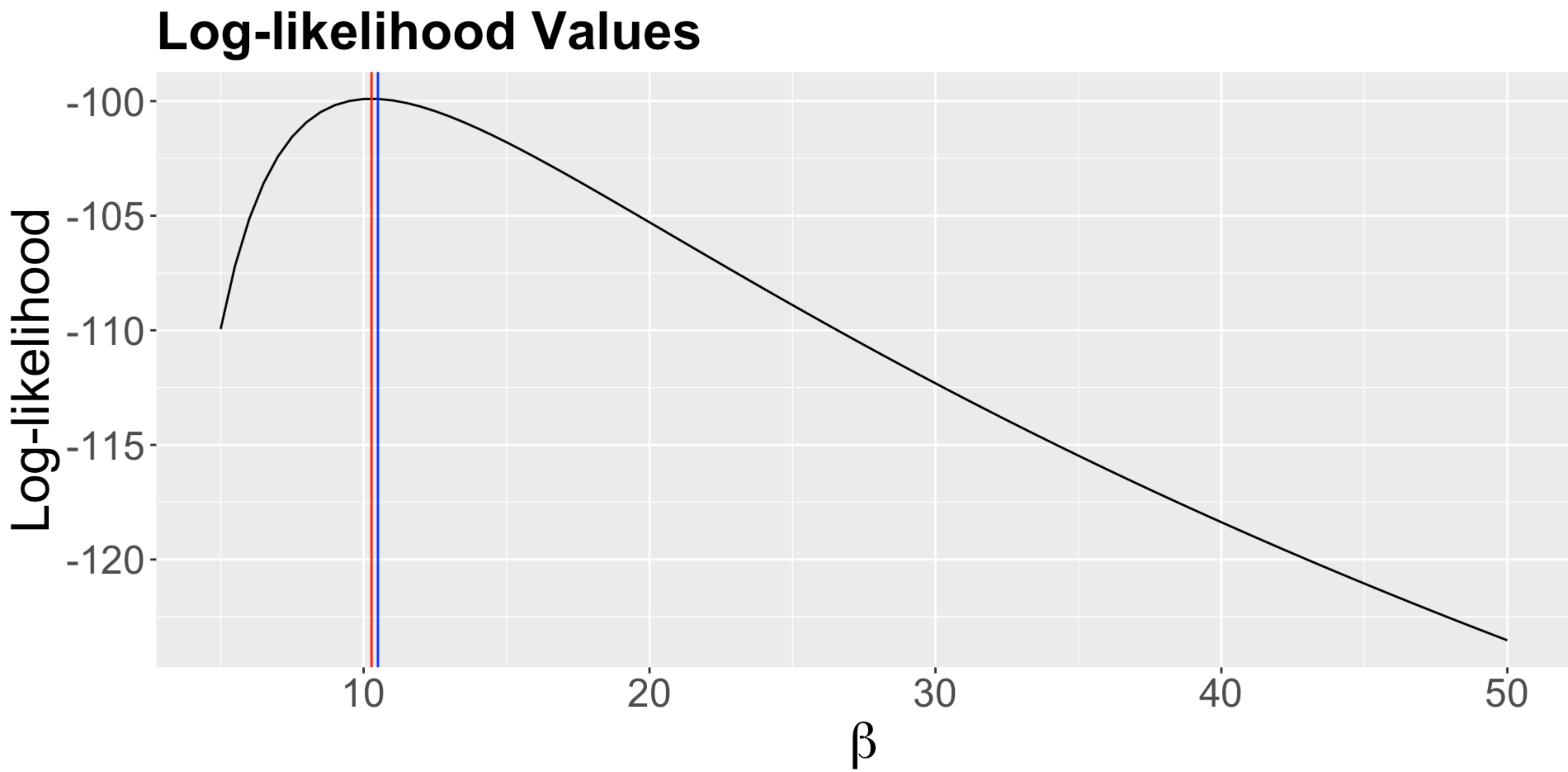
  is the MLE estimator for β.

- We often use **the hat symbol** to denotes estimators, and estimators are random variables.

Compute and plot the analytical solution

```r
analytical_MLE <- mean(sample_n30$values) # We use the sample mean() functi
round(analytical_MLE, 4)
```

```
[1] 10.277
```



Conclusion for the ice cream example

- Our pilot study with $n = 30$ sampled wait times indicates that the **estimated wait time** between each ice cream customer is **10.277 minutes**.

4.3 Numerical methods for MLE (optional)

- Some likelihood functions do not have a **closed-form solution**.
- We need to use **numerical optimization methods**:
  - Gradient descent (Supervised Learning I)
  - Newton's method
  - more

`optimize()` in R

> optimize(f, interval, maximum = TRUE, …)

- `f` is the function to be optimized
- `interval` is a range of values
- `maximum` indicates minimum or the maximum

```r
LL <- function(beta) log(prod(dexp(sample_n30$values, rate = 1 / beta)))
optimize(LL, c(5, 50), maximum = TRUE)
```

```
$maximum
[1] 10.27704

$objective
[1] -99.89738
```

Summary of MLE

- MLE is a very useful method **to estimate population parameters** from a **random sample**.
- These estimators are
  - Consistent: $\hat{\beta} \to \beta$ as $n \to \infty$
  - Asymptotically optimal under certain conditions
- However, MLE relies on strong assumptions:
  - Distributional assumptions
  - iid assumption

Steps for MLE from a set of values

- Collect data
- Choose the right distribution
- Obtain the likelihood function (joint PDF/PMF)
- Obtain the log-likelihood function
- Compute log-likelihood for a set of parameter values
- Choose the parameter value that achieve the maximum log-likelihood

Steps for Analytical MLE

- From the log-likelihood function, calculate the partial derivative with respect to the parameter
- Set the partial derivative equal to zero and solve for the optimal value
- Check the second partial derivative

Today's Learning Goals

By the end of this lecture, we will be able to…

- Identify the graphical and mathematical relationship between two independent continuous random variables.
- Explain the concept of maximum likelihood estimation.
- Apply maximum likelihood estimation for univariate cases.