An introduction to maximum likelihood estimation

Statistical Computing and Empirical Methods
Unit EMATM0061, Data Science MSc

Rihuan Ke
rihuan.ke@bristol.ac.uk

Teaching Block 1, 2024

University of BRISTOL

What we will cover today

We will introduce an important method for statistical estimation: maximum likelihood estimation -- this is a method based on the likelihood function that measures how well a model fits a data set

The likelihood function is a function of the unknown parameters. By maximising the likelihood function, we aim to find the parameters such that the model best fits the data set

We will work on several examples of maximum likelihood estimators, where the likelihood is maximised for specific models.

We will give an overview of the maximum likelihood method's favourable properties.

Some key concepts

Tasks. We need to estimate the parameters θ in our model based upon a sample $X_1, \cdots, X_n \sim \mathbb{P}_\theta$.

We estimate our parameters based on sample statistics,
which are functions of the samples $\hat{\theta} = g(X_1, \cdots, X_n)$ that don't depend on θ.

A sample statistic $\hat{\theta} = g(X_1, \cdots, X_n)$ of a population parameter θ is consistent if for all $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\{|g(X_1, \cdots, X_n) - \theta| > \epsilon\} = 0$.

The bias of an estimator $\hat{\theta} = g(X_1, \cdots, X_n)$ of a population parameter θ is
$\text{Bias}\left(\hat{\theta}\right) := \mathbb{E}\left(\hat{\theta}\right) - \theta$.

The variance of an estimator $\hat{\theta} = g(X_1, \cdots, X_n)$ of a population parameter θ is
$\text{Var}\left(\hat{\theta}\right) := \mathbb{E}\left\{\left(\hat{\theta} - \mathbb{E}\left(\hat{\theta}\right)\right)^2\right\}$.

A minimum variance unbiased estimator $\hat{\theta}$ has minimal variance over all possible unbiased estimators.

Today's focus

Problem formulation: Given a probabilistic model $\mathbb{P}_\theta$ (parametrized by θ) and a sample $X_1, \cdots, X_n$, can we find an algorithm/method to compute an estimate of θ from the sample?

Of course, we can use the sample mean to estimate the population mean, and use the sample variance to estimate the population variance, but ...

We would like a general strategy for finding (near) optimal estimators $\hat{\theta}$ for population parameters θ.

Find the parameter θ such that the model $\mathbb{P}_\theta$ best fit the dataset (sample)

1. The likelihood function

Problem formulation: Given a probabilistic model $\mathbb{P}_\theta$ (parametrized by θ) and a sample $X_1, \cdots, X_n$, can we find an algorithm to compute an estimate of θ for the sample?

Search the unknown parameter θ in the parameter space Θ.

The likelihood function $L : \Theta \to [0, \infty)$ is a function of the parameter θ. It maps each θ to a single number which measures the goodness of fit to the data $X_1, \cdots, X_n$.

Suppose that $\mathbf{X} = (X_1, \cdots, X_n)$ is a sequence of i.i.d. random variables. The likelihood function is defined as below:

Case 1: Discrete random variables. $L(\theta; \mathbf{X}) := \prod_{i=1}^n p_\theta(X_i)$ where p_θ is the probability mass function of X_i.

Case 2: Continuous random variables. $L(\theta; \mathbf{X}) := \prod_{i=1}^n f_\theta(X_i)$ where p_θ is the probability density function of X_i.

Example 1 (discrete random variables)

Example 1. Suppose $X_1, \cdots, X_n \sim \mathcal{B}(q_0)$ are i.i.d. Bernoulli random variables with an unknown parameter $\mathbb{E}(X_i) = q_0$.

Every observation X_i has the probability mass function $p_q : \mathbb{R} \to [0, 1]$ given by

$$p_q(x) = q^x \cdot (1-q)^{1-x} \cdot \mathbb{1}_{\{0,1\}}(x) = \begin{cases} 1-q, & \text{if } x = 0 \\ q, & \text{if } x = 1 \\ 0, & \text{Otherwise.} \end{cases}$$

So $q \in \Theta := [0, 1]$, and the likelihood function $L : [0, 1] \to [0, \infty)$ is given by

$$L(q; \mathbf{X}) = \prod_{i=1}^{n} p_q(X_i) = \prod_{i=1}^{n} \left\{ q^{X_i} \cdot (1-q)^{1-X_i} \right\} = q^{\sum_{i=1}^n X_i} \cdot (1-q)^{n-\sum_{i=1}^n X_i}.$$

Here we assume that the value of X_i is either 0 or 1, so $\mathbb{1}_{\{0,1\}}(X_i) = 1$.

Example 2 (continuous random variables)

Example 2. Suppose $X_1, \cdots, X_n \sim \mathcal{N}(\mu_0, \sigma_0^2)$ are i.i.d. Gaussian random variables with unknown parameters (μ_0, σ_0^2).

Every observation X_i has the probability density function $f_{\mu,\sigma} : \mathbb{R} \to [0, \infty]$ given by

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\} \text{ for all } x \in \mathbb{R}$$

So $(\mu, \sigma) \in \Theta := \mathbb{R} \times (0, \infty)$, and the likelihood function $L : \mathbb{R} \times (0, \infty) \to [0, \infty)$ is given by

$$L(\mu, \sigma; \mathbf{X}) = \prod_{i=1}^{n} f_{\mu,\sigma}(X_i) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \right)$$

2. Maximum likelihood estimation

Problem formulation: Given a probabilistic model $\mathbb{P}_\theta$ (parametrized by θ) and a sample $X_1, \cdots, X_n$, can we find an algorithm to compute an estimate of θ for the sample?

Answer: we can maximise the likelihood function $L(\theta; \mathbf{X})$ for an estimate of θ

Maximum likelihood estimate

The maximum likelihood estimate $\hat{\theta}(X_1, \cdots, X_n)$ for a parameter $\theta_0 \in \Theta$ is defined to be the parameter value which maximizes the likelihood:

$$\hat{\theta}(X_1, \cdots, X_n) = \text{argmax}_{\theta \in \Theta} L(\theta; \mathbf{X}).$$

This formalizes the idea of choosing a parameter such that the model best fits the data.

Example 1

Example 1. Suppose $X_1, \cdots, X_n \sim \mathcal{B}(q_0)$ are i.i.d. Bernoulli random variables with an unknown parameter $\mathbb{E}(X_i) = q_0$.

$$L(q; \mathbf{X}) = \prod_{i=1}^{n} p_q(X_i) = \prod_{i=1}^{n} \left\{ q^{X_i} \cdot (1-q)^{1-X_i} \right\} = q^{\sum_{i=1}^{n} X_i} \cdot (1-q)^{n-\sum_{i=1}^{n} X_i}.$$

Let $\overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$. We aim to find the maximizer of

$$\log L(q; \mathbf{X}) = \log\left(q^{n\overline{X}} \cdot (1-q)^{n-n\overline{X}} \right) = n\overline{X}\log(q) + (n - n\overline{X})\log(1-q).$$

To find the maximizer, setting $\frac{\partial \log(L(q;\mathbf{X}))}{\partial q} = n\left(\frac{\overline{X}}{q} - \frac{1-\overline{X}}{1-q} \right) = 0$

Solving this equation, we get the maximum likelihood estimate $\hat{q}_{\text{MLE}} = \overline{X}$.

So the maximum likelihood estimate for q_0 is also an MVUE!

Example 2

Example 2. Suppose $X_1, \cdots, X_n \sim \mathcal{N}(\mu_0, \sigma_0^2)$ are i.i.d. Gaussian random variables with unknown parameters (μ_0, σ_0^2).

$$L(\mu, \sigma; \mathbf{X}) = \prod_{i=1}^{n} f_{\mu,\sigma}(X_i) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2\right)$$

By taking the logarithm and differentiating we see that

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{is the MLE for } \mu_0 \qquad \text{(this is also a MVUE)}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad \text{is the MLE for } \sigma_0^2 \qquad \text{(this is not unbiased)}$$

Recall that the MVUE of σ^2 is given by $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, which is unbiased.

3. Simulation studies

In what follows, we will conduct simulation studies to see

- The maximum likelihood estimator with Gaussian distributions & Cauchy distributions, respectively
- How we can compute the maximum likelihood estimate using the idea of numerical optimisation when there is no closed-form solution
- Visualisation of the distribution of the maximum likelihood estimate (as a random variable)

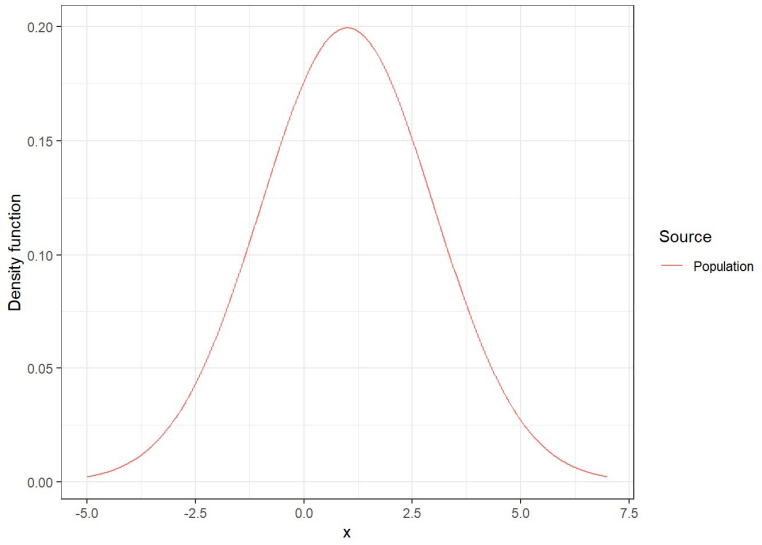
Simulation 1: maximum likelihood

Example 2. Suppose $X_1, \cdots, X_n \sim \mathcal{N}(\mu_0, \sigma_0^2)$ are i.i.d. Gaussian random variables with unknown parameters (μ_0, σ_0^2).

```
mu <- 1 # choose a mean
sigma <- 2 # choose a standard deviation

# 1. generate some x indices
x<-seq(mu-3*sigma, mu+3*sigma, sigma*0.01)
# 2. data frame with population density
df_gaussian <- data.frame(x, Density=dnorm(x, mean=mu, sd=sigma), Source="Population")
# 3. plot the density function
df_gaussian %>% ggplot(aes(x=x, y=Density, color=Source)) +
  geom_line() + ylab("Density function") + theme_bw()
```

The density function of the Gaussian random variable:



Simulation 1: maximum likelihood

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{is the MLE for } \mu_0 \qquad \text{(this is also a MVUE)}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad \text{is the MLE for } \sigma_0^2 \qquad \text{(this is not unbiased)}$$

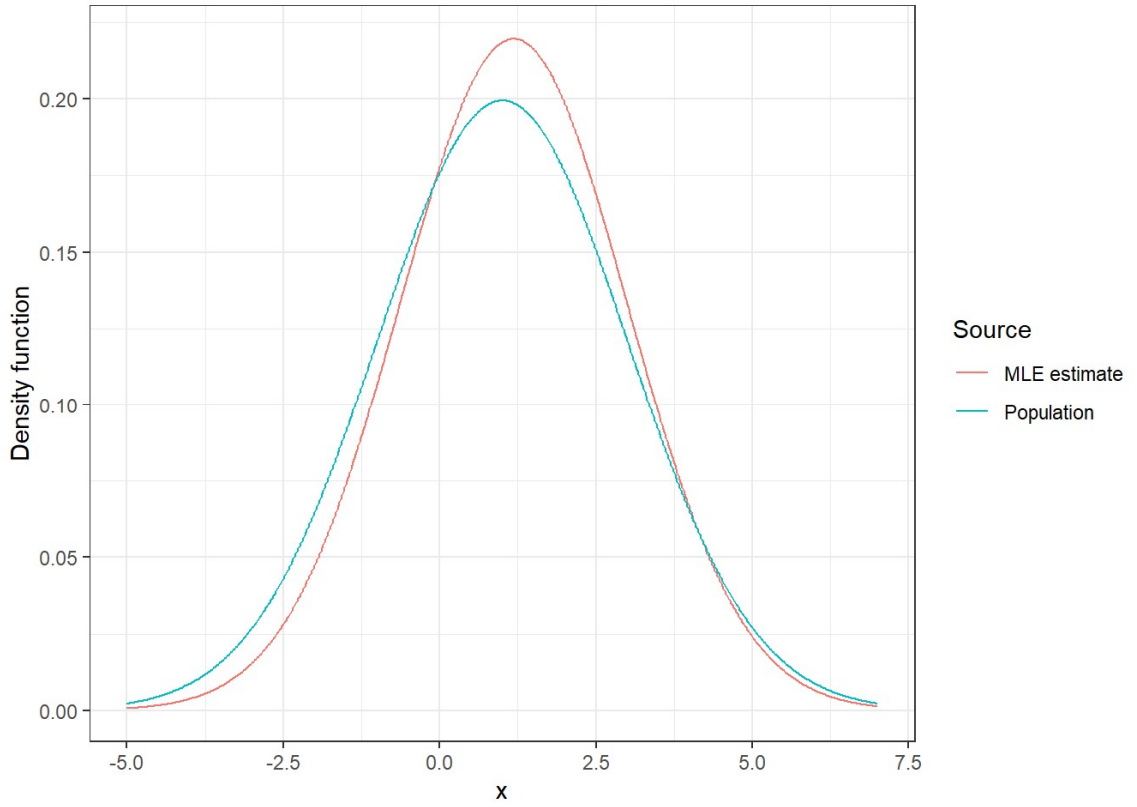
MLE for **Gaussian distribution**; Plot the parametric Gaussian model, fitted with MLE. Then compare it with the true density

```
set.seed(123)
sample_size <- 100

# 4. generate a sample of Gaussian random variables
sample_data <- rnorm(sample_size, mu, sigma)
# 5. ML estimate of mu and sigma
mu_mle <- mean(sample_data)
sigma_mle <- sd(sample_data)*sqrt((sample_size-1)/sample_size)
# 6. add estimated density function to df
df_gaussian <- df_gaussian %>%
  rbind(data.frame(x, Density=dnorm(x,mean=mu_mle,sd=sigma_mle),
                   Source="MLE estimate"))
# 7. plot the true and estimated density functions
df_gaussian %>% ggplot(aes(x=x, y=Density, color=Source)) +
  geom_line() + ylab("Density function")+theme_bw()
```

Simulation 1: maximum likelihood

MLE for **Gaussian distribution**; Plot the parametric Gaussian model, fitted with MLE. Then compare it with the true density



Simulation 2: maximum likelihood with penguins data

Let's fit a **Gaussian model** to the weights of Gentoo penguins

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{is the MLE for } \mu_0 \qquad \text{(this is also a MVUE)}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad \text{is the MLE for } \sigma_0^2 \qquad \text{(this is not unbiased)}$$

```r
library(palmerpenguins)
```

```r
# 1. sample of Gentoo weights
gentoo_weights <- penguins %>%
  filter(species=='Gentoo') %>%
  pull(body_mass_g) # extract the column of Gentoo weights

# 2. ML estimates of mu and sigma
n <- length(gentoo_weights) # sample size
mu_mle <- mean(gentoo_weights, na.rm=TRUE) # mle mean
sigma_mle <- sd(gentoo_weights, na.rm=TRUE) * sqrt((n-1)/n) # mle standard deviation
```

Simulation 2: maximum likelihood with penguins data

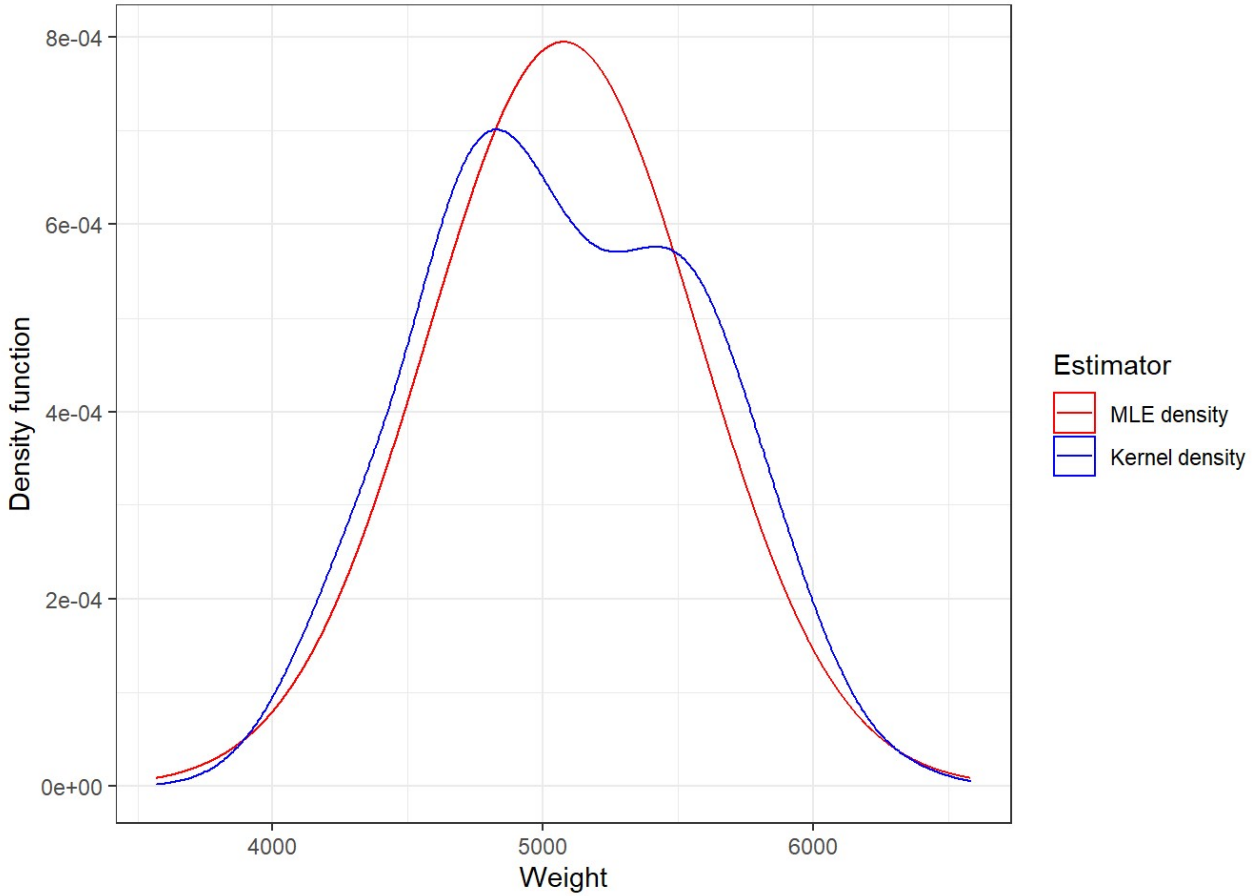
Let's plot our parametric Gaussian model, fitted with MLE, and our kernel density plot

```
# 4. generate indices
weights <- seq(mu_mle-3*sigma_mle, mu_mle+3*sigma_mle, sigma_mle*0.001)

# 5.1 plot estimated density functions (MLE density)
colors <- c("MLE density"="red", "Kernel density"="blue") # set color legend
estimated_density = data.frame(Weight=weights,
                               Density=dnorm(weights, mean=mu_mle, sd=sigma_mle))
plot_obj <- ggplot() + geom_line(data=estimated_density,
                     aes(x=Weight, y=Density, color="MLE density"))
# 5.2 kernel density plot of the sample
plot_obj + geom_density(data=tibble(gentoo_weights), aes(x=gentoo_weights, color="Kernel density")) +
  labs(y="Density function", color="Estimator") + theme_bw() + scale_color_manual(values=colors)
```

Simulation 2: maximum likelihood with penguins data

Let's plot our parametric Gaussian model, fitted with MLE, and our kernel density plot

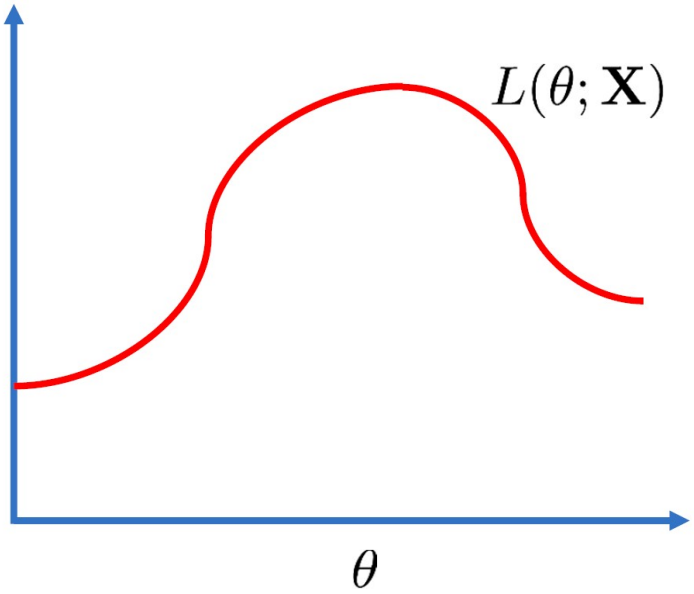


Finding the maximizer

In many other cases, there is no closed-form solution.

Recall that: The maximum likelihood estimate $\hat{\theta}(X_1, \cdots, X_n)$ for a parameter $\theta_0 \in \Theta$ is defined to be the parameter value which maximizes the likelihood: $\hat{\theta}(X_1, \cdots, X_n) = \text{argmax}_{\theta \in \Theta} L(\theta; \mathbf{X})$.

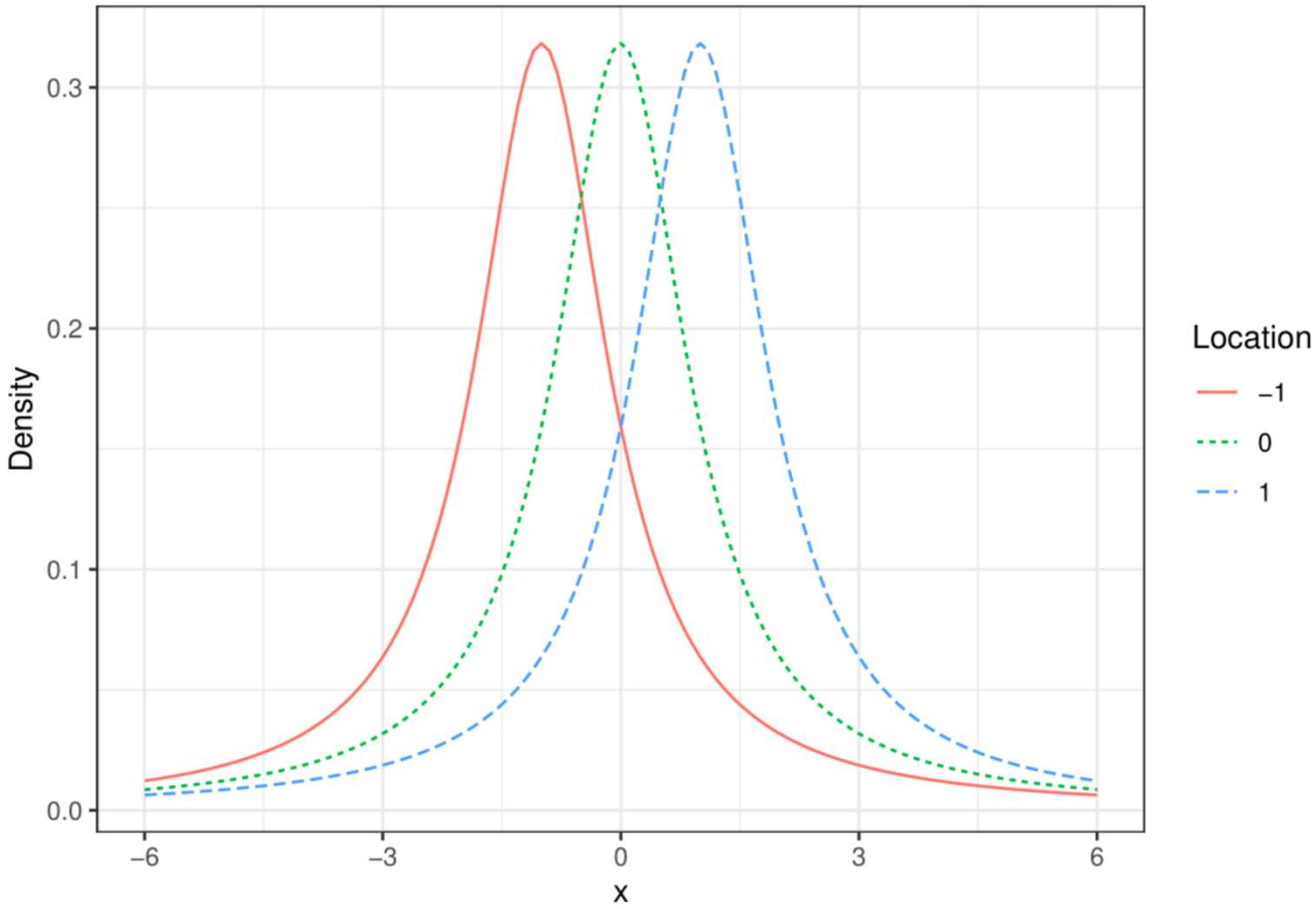
We use optimization techniques to maximize the likelihood function $\theta \to L(\theta; \mathbf{X})$ numerically.



Simulation 3: maximum likelihood with Cauchy distribution

A random variable X has a Cauchy distribution with location parameter θ if its density is

$$f_\theta(x) := \frac{1}{\pi \cdot (1 + (x - \theta)^2)}.$$



Simulation 3: maximum likelihood with Cauchy distribution

A random variable X has a Cauchy distribution with location parameter θ if its density is

$$f_\theta(x) := \frac{1}{\pi \cdot (1 + (x - \theta)^2)}.$$

Suppose that we have i.i.d. random variables $X_1, \cdots, X_n \sim f_{\theta_0}$. We aim to estimate the unknown parameter θ_0.

The likelihood function is given by

$$L(\theta; \mathbf{X}) = \prod_{i=1}^{n} f_\theta(X_i) = \prod_{i=1}^{n} \frac{1}{\pi(1 + (X_i - \theta)^2)}.$$

The log-likelihood function is given by

$$\log L(\theta; \mathbf{X}) = -n\log(\pi) - \sum_{i=1}^{n} \log\big(1 + (X_i - \theta)^2\big).$$

Unfortunately, we do not have an analytical solution to $\hat{\theta}_n \in \text{argmax} L(\theta; \mathbf{X})$

We can compute the optimisation problem numerically.

Simulation 3: maximum likelihood with Cauchy distribution

Our goal is to maximise the log-likelihood
$\log L(\theta; \mathbf{X}) = -n\log(\pi) - \sum_{i=1}^{n} \log\left(1 + (X_i - \theta)^2\right)$.

We can use the function `optimise()`.

An example of finding the MLE numerically:

```
set.seed(0)
sample_size <- 100
theta_0 <- 5 # True parameter theta

# 1. Generate sample of a sequence of Cauchy R. V.
cauchy_sample<-rcauchy(n=sample_size, location=theta_0)

# 2. The log likelihood function
log_lik_cauchy <- function(theta,sample_X){return(-sum(log(1+(sample_X-theta)**2)))}
log_lik_cauchy_X <- function(theta){return(log_lik_cauchy(theta,cauchy_sample) )}

# 3. optimise the log likelihood function
theta_ml_est <- optimise(f=log_lik_cauchy_X, interval=c(-1000,1000), maximum=TRUE)$maximum
theta_ml_est
```

```
## [1] 4.906282
```

Simulation 3: maximum likelihood with Cauchy distribution

Our goal is to maximise the log-likelihood

$$\log L(\theta; \mathbf{X}) = -n\log(\pi) - \sum_{i=1}^{n} \log\left(1 + (X_i - \theta)^2\right).$$

Next, let's conduct a simulation to study the distribution of the maximum likelihood estimate, by using 100000 trials

```
set.seed(0)

num_trials <- 100000
sample_size <- 100
theta_0 <- 5 # True parameter theta

# 1. log likelihood function
log_lik_cauchy <- function(theta,sample_X){return(-sum(log(1+(sample_X-theta)**2)))}

# 2. mapping a sample to ML estimate of theta
theta_ml<-function(sample_X){
  log_lik_cauchy_X <- function(theta){return(log_lik_cauchy(theta,sample_X))}
  theta_ml_est <- optimise(f=log_lik_cauchy_X,interval=c(-10,18),maximum=TRUE)$maximum
  return(theta_ml_est)
}
```

Simulation 3: maximum likelihood with Cauchy distribution

Our goal is to maximise the log-likelihood

$$\log L(\theta; \mathbf{X}) = -n\log(\pi) - \sum_{i=1}^{n} \log\left(1 + (X_i - \theta)^2\right).$$

Next, let's conduct a simulation to study the distribution of the maximum likelihood estimate, by using 100000 trials

```
# 3.1 create num_trials samples; the size of each sample is sample_size
df <- data.frame(trial=seq(num_trials)) %>%
  mutate(sample=map(.x=trial,~rcauchy(sample_size,location=theta_0)))
# 3.2 for each sample, compute ML est and Median est
cauchy_simulation_df <- mutate(df, ml_est=map_dbl(.x=sample, .f=theta_ml)) %>%
  mutate(med_est=map_dbl(.x=sample,.f=median))

# 4. pivot
plot_df<-cauchy_simulation_df %>%
  pivot_longer(cols=c(ml_est,med_est) ) %>%
  mutate(name=map_chr(.x=name,~case_when(.x=="med_est"~"Median",
                                         .x=="ml_est"~"Maximum likelihood")))

# 5. create plot
ggplot(plot_df, mapping=aes(x=value, color=name, linetype=name) ) +
  geom_density() + theme_bw() + xlim(c(4,6)) +
  labs(color="", linetype="") + xlab("Estimate" ) + ylab("Density")
```

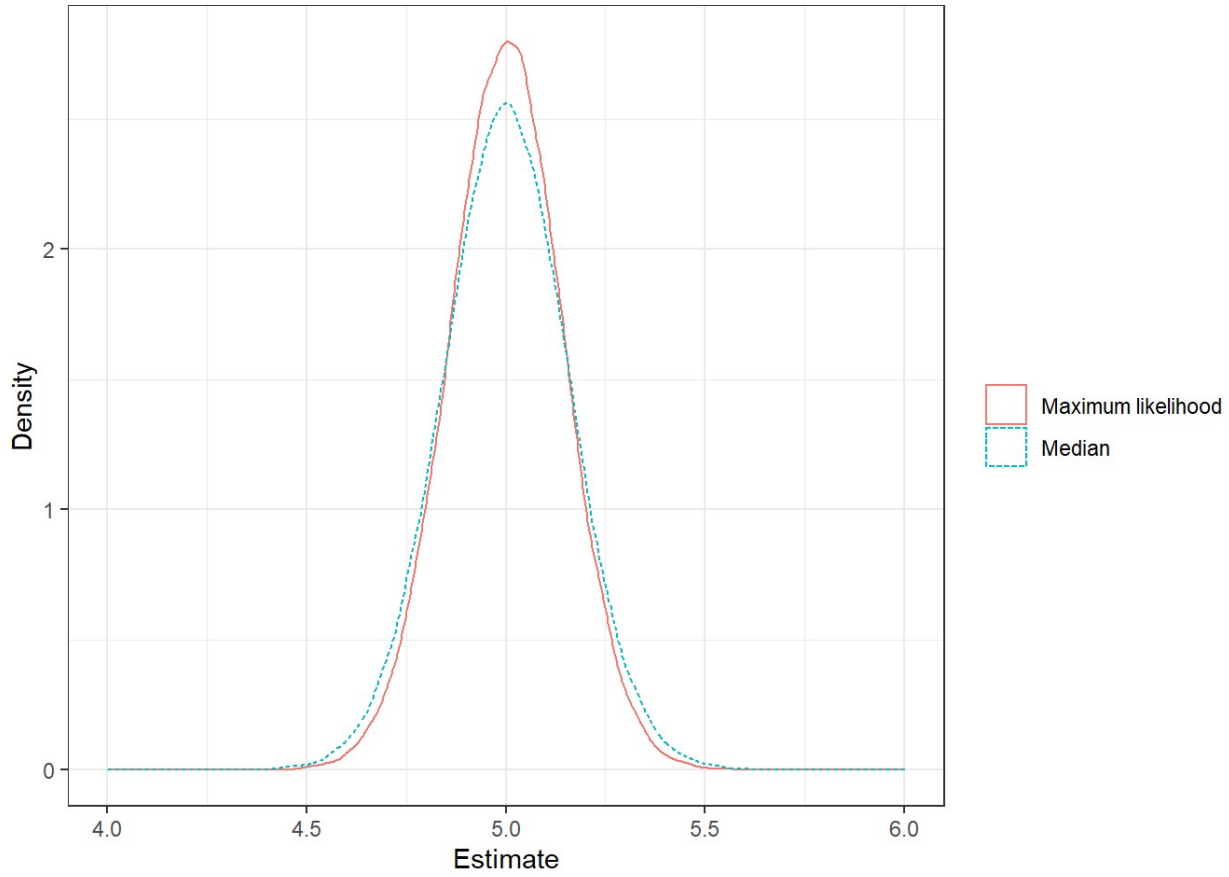
Simulation 3: maximum likelihood with Cauchy distribution

Our goal is to maximise the log-likelihood

$$\log L(\theta; \mathbf{X}) = -n\log(\pi) - \sum_{i=1}^{n} \log\left(1 + (X_i - \theta)^2\right).$$

Next, let's conduct a simulation to study the distribution of the maximum likelihood estimate, by using 100000 trials

The distribution of MLE compared with the distribution of the sample median.



Simulation 3: maximum likelihood with Cauchy distribution

Our goal is to maximise the log-likelihood

$$\log L(\theta; \mathbf{X}) = -n\log(\pi) - \sum_{i=1}^{n} \log\left(1 + (X_i - \theta)^2\right).$$

Next, let's conduct a simulation to study the distribution of the maximum likelihood estimate, by using 100000 trials

Compute the mean square error for MLE and for the sample median as estimates of the location parameter:

```
msefunc <- function(x){return (mean( (x-theta_0)^2 ) )}

med_estimate_mean_sqr_error <- cauchy_simulation_df %>%
  pull(med_est)  %>% msefunc

med_estimate_mean_sqr_error
```

```
## [1] 0.02533741
```

```
ml_estimate_mean_sqr_error <- cauchy_simulation_df %>%
  pull(ml_est)  %>% msefunc

ml_estimate_mean_sqr_error
```

```
## [1] 0.02062467
```

MLE has a slightly smaller MSE in the simulation study

4. Property: Is MLE consistent?

It is known that the maximum likelihood estimate (MLE) is consistent under some natural conditions.

That means, we have $\hat{\theta}(X_1, \cdots, X_n) \to \theta_0 \in \Theta$ as $n \to \infty$.

Example 1. Suppose $X_1, \cdots, X_n \sim \mathcal{B}(q_0)$ are i.i.d. Bernoulli random variables with an unknown parameter $\mathbb{E}(X_i) = q_0$.

The MLE $\hat{q}_{\text{MLE}} = \overline{X} \quad \to \quad q_0$ as $n \to \infty$.

Example 2. Suppose $X_1, \cdots, X_n \sim \mathcal{N}(\mu_0, \sigma_0^2)$ are i.i.d. Gaussian random variables with unknown parameters (μ_0, σ_0^2).

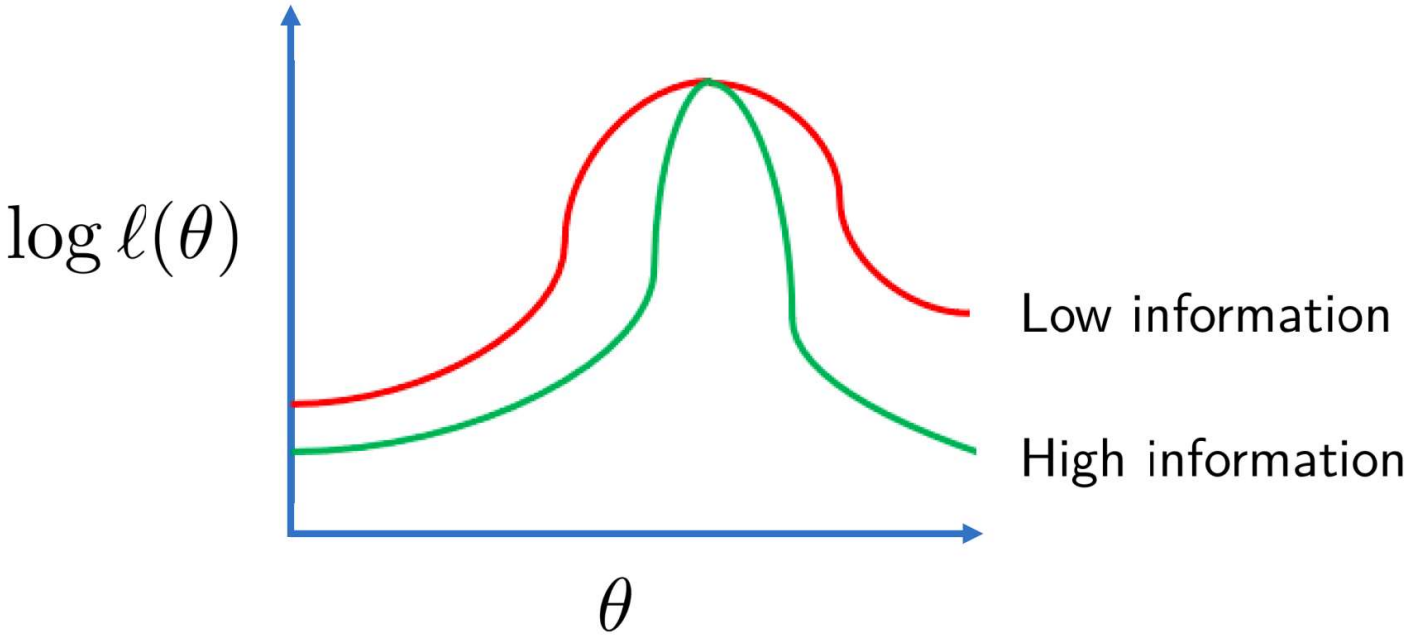
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \to \qquad \mu_0 \qquad \text{as} \qquad n \to \infty$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad \to \qquad \sigma_0^2 \qquad \text{as} \qquad n \to \infty$$

~~5. Maximum likelihood and Fisher information (*optional)~~

A useful quantity for understanding maximum likelihood estimation is the Fisher information given by

$$\mathcal{I}(\theta) := -\mathbb{E}\left\{ \frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right\} \quad \text{where} \quad X \sim f_\theta.$$



5. Maximum likelihood and Fisher information (*optional)

A useful quantity for understanding maximum likelihood estimation is the Fisher information given by

$$\mathcal{I}(\theta) := -\mathbb{E}\left\{ \frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right\} \quad \text{where} \quad X \sim f_\theta.$$

Theorem

Let $\hat{\theta}$ be the maximum likelihood estimator based on a sample $X_1, \cdots, X_n \sim f_\theta$.

Let $Z \sim \mathcal{N}(0,1)$ be a standard Gaussian random variable. For a sequence of suitably well-behaved i.i.d. random variables $X_1, \cdots, X_n \sim f_\theta$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n\mathcal{I}(\theta_0)}(\hat{\theta}_n - \theta_0) \le x \right) = \mathbb{P}(Z \le x).$$

Remarks: 1. The variance of $\hat{\theta}_n$ is approximately at the level of $\frac{1}{n\mathcal{I}(\theta_0)}$ for large n.

2. The result implies that $\hat{\theta}_n$ is a consistent estimator for θ_0.

Cramer and Rao showed that the variance level $\frac{1}{n\mathcal{I}(\theta_0)}$ is the best possible.

What have we covered?

We introduced the likelihood function for measuring how well a model fits a data set.

We introduced the method of maximum likelihood estimation.

We considered several examples where the likelihood can be analytically maximized.

We discussed the use of numerical methods when analytic methods are unavailable.

We also discussed some of the maximum likelihood methods favourable properties.

University of BRISTOL

Thanks for listening!

Dr. Rihuan Ke

rihuan.ke@bristol.ac.uk

Statistical Computing and Empirical Methods
Unit EMATM0061, MSc Data Science