University of BRISTOL

An introduction to maximum likelihood estimation

Henry W J Reeve

henry.reeve@bristol.ac.uk

Statistical Computing & Empirical Methods (EMATM0061)

MSc in Data Science, Teaching block 1, 2021.

What will we cover today?

- We will introduce the likelihood function for measuring how well a model fits a data set.
- We will introduce a very flexible method known as maximum likelihood estimation.
- We considered a variety of examples where the likelihood can be maximized for specific models.
- We will also give an overview of the maximum likelihood method’s favourable properties.

Statistical estimation

Random sample

Reality

Data

Estimation

Population

Sample

$$\theta$$

$$\hat{\theta} = g(X_1, \cdots, X_n)$$

Consistency, bias, variance and error

Given data $X_1, \ldots, X_n \sim \mathbb{P}_\theta$ we are interested in statistical estimators $\hat{\theta}_n$ of parameters θ.

A sample statistic $\hat{\theta}_n$ is a **consistent** estimator of θ if for all $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left(|\hat{\theta}_n - \theta| \geq \epsilon\right) = 0$;

The **bias** of $\hat{\theta}_n$ is given by $\text{Bisa}(\hat{\theta}_n) := \mathbb{E}(\hat{\theta}_n) - \theta$;

The **variance** of $\hat{\theta}_n$ is given by $\text{Var}(\hat{\theta}_n) := \mathbb{E}\big[\{\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)\}^2\big]$.

The **mean squared error** of $\hat{\theta}_n$ is $\text{MSE}(\hat{\theta}_n) := \mathbb{E}\{(\hat{\theta}_n - \theta)^2\} = \text{Bias}(\hat{\theta}_n)^2 + \text{Var}(\hat{\theta}_n)$.

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

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

The likelihood function

Let $X_1, \cdots, X_n$ be a sample from some parametric model $\mathbb{P}_{\theta_0}$ with unknown parameter $\theta_0 \in \Theta$

The **likelihood function** $\ell : \Theta \to [0, \infty)$ associates to each parameter $\theta \in \Theta$,

a single number which measures the goodness of fit to the data $X_1, \cdots, X_n$.

The likelihood function for discrete random variables

Let $X_1, \cdots, X_n$ be a sample from some parametric model $\mathbb{P}_{\theta_0}$ with unknown parameter $\theta_0 \in \Theta$

The **likelihood function** $\ell : \Theta \to [0, \infty)$ associates to each parameter $\theta \in \Theta$,

a single number which measures the goodness of fit to the data $X_1, \cdots, X_n$.

<u>Case 1: Discrete random variables</u>

Suppose that $X_1, \cdots, X_n$ are i.i.d. discrete random variables with probability mass function p_{θ_0}.

Then $$\ell(\theta) := \prod_{i=1}^{n} p_\theta(X_i).$$

The likelihood function for discrete random variables

Example 1

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

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

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

The likelihood function $\ell : [0, 1] \to [0, \infty)$ is given by

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

The likelihood function for continuous random variables

Let $X_1, \cdots, X_n$ be a sample from some parametric model $\mathbb{P}_{\theta_0}$ with unknown parameter $\theta_0 \in \Theta$

The **likelihood function** $\ell : \Theta \to [0, \infty)$ associates to each parameter $\theta \in \Theta$,

a single number which measures the goodness of fit to the data $X_1, \cdots, X_n$.

<u>Case 2: Continuous random variables</u>

Suppose that $X_1, \cdots, X_n$ are i.i.d. continuous random variables with density f_{θ_0}.

Then $$\ell(\theta) := \prod_{i=1}^{n} f_\theta(X_i).$$

The likelihood function for continuous random variables

Example 2

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

Every observation X_i has the probability density function $f_{\mu,\sigma} : \mathbb{R} \to (0, 1)$ given by

$$f_{\mu,\sigma}(x) := \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad \text{for all} \quad x \in \mathbb{R} \quad \text{with} \quad (\mu, \sigma^2) = (\mu_0, \sigma_0^2)$$

The likelihood function $\ell : \mathbb{R} \times (0, \infty) \to [0, \infty)$ is given by

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

Maximum likelihood estimation

Let $X_1, \cdots, X_n$ be a sample from some parametric model $\mathbb{P}_{\theta_0}$ with unknown parameter $\theta_0 \in \Theta$

The **likelihood function** $\ell : \Theta \to [0, \infty)$ associates to each parameter $\theta \in \Theta$,

a single number which measures the goodness of fit to the data $X_1, \cdots, X_n$.

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:

$$\hat{\theta}(X_1, \cdots, X_n) = \mathrm{argmax}_{\theta \in \Theta} \{\ell(\theta)\} .$$

This formalizes the idea of choosing a parameter which best fits the data.

Maximum likelihood estimation

The **maximum likelihood estimate** (MLE) $\hat{\theta}(X_1, \cdots, X_n)$ for a parameter $\theta_0 \in \Theta$ is defined to be the parameter value which maximizes the likelihood: $\hat{\theta}(X_1, \cdots, X_n) = \mathrm{argmax}_{\theta \in \Theta} \{\ell(\theta)\}$.

<u>Example 1</u>

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

The likelihood function $\ell : [0,1] \to [0,\infty)$ is given by $\ell(q) = q^{\sum_{i=1}^n X_i} \cdot (1-q)^{n - \sum_{i=1}^n X_i}$.

The maximum likelihood estimate for q is $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ which is also an MVUE!

Maximum likelihood estimation

Example 1

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

The likelihood function $\ell : [0,1] \to [0, \infty)$ is given by $\ell(q) = q^{\sum_{i=1}^n X_i} \cdot (1-q)^{n - \sum_{i=1}^n X_i}$.

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

$$\log(l(q)) = \sum_{i=1}^n X_i \cdot \log(q) + \left(n - \sum_{i=1}^n X_i\right) \log(1-q)$$

$$= n\left\{\overline{X}\log(q) + (1-\overline{X})\log(1-q)\right\}.$$

Maximum likelihood estimation

Example 1

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

The likelihood function $\ell : [0,1] \to [0,\infty)$ is given by $\ell(q) = q^{\sum_{i=1}^n X_i} \cdot (1-q)^{n-\sum_{i=1}^n X_i}$.

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

$$\log(l(q)) = \sum_{i=1}^n X_i \cdot \log(q) + \left(n - \sum_{i=1}^n X_i\right) \log(1-q)$$

$$= n\left\{\overline{X}\log(q) + (1-\overline{X})\log(1-q)\right\}.$$

$$\Rightarrow \quad \frac{\partial \log(l(q))}{\partial q} = n\left\{\frac{\overline{X}}{q} - \frac{(1-\overline{X})}{1-q}\right\}$$

Maximum likelihood estimation

Example 1

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

$$\log(l(q)) = \sum_{i=1}^{n} X_i \cdot \log(q) + \left(n - \sum_{i=1}^{n} X_i\right)\log(1-q) = n\left\{\overline{X}\log(q) + (1-\overline{X})\log(1-q)\right\}.$$

$$\Rightarrow \quad \frac{\partial \log(l(q))}{\partial q} = n\left\{\frac{\overline{X}}{q} - \frac{(1-\overline{X})}{1-q}\right\}$$

We find the MLE by setting $\frac{\partial \log(l(q))}{\partial q} = n\left\{\frac{\overline{X}}{q} - \frac{(1-\overline{X})}{1-q}\right\} = 0$

By rearranging we obtain $\hat{q}_{\mathrm{MLE}} = \overline{X}$

Maximum likelihood estimation

<u>Example 2</u>

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

The likelihood function $\ell : \mathbb{R} \times (0, \infty) \to [0, \infty)$ is given by

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

By taking the logarithm and differentiating we see that

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the MLE for μ_0 (this is also a MVUE)

$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ is the MLE for σ_0^2 (this has non-zero bias).

Maximum likelihood estimation

Maximum likelihood estimation has the property of **functional invariance.**

Suppose that $\hat{\theta}_n$ is a maximum likelihood estimator for a parameter $\theta_0 \in \Theta$.

Let $g : \Theta \to \tilde{\Theta}$ be a bijective (i.e. one-to-one) function.

Then $g\left(\hat{\theta}_n\right)$ is a maximum likelihood estimator for $g\left(\theta_0\right)$.

Maximum likelihood estimation

Maximum likelihood estimation has the property of **functional invariance.**

Suppose that $\hat{\theta}_n$ is a maximum likelihood estimator for a parameter $\theta_0 \in \Theta$.

Let $g : \Theta \to \tilde{\Theta}$ be a bijective (i.e. one-to-one) function.

Then $g\left(\hat{\theta}_n\right)$ is a maximum likelihood estimator for $g\left(\theta_0\right)$.

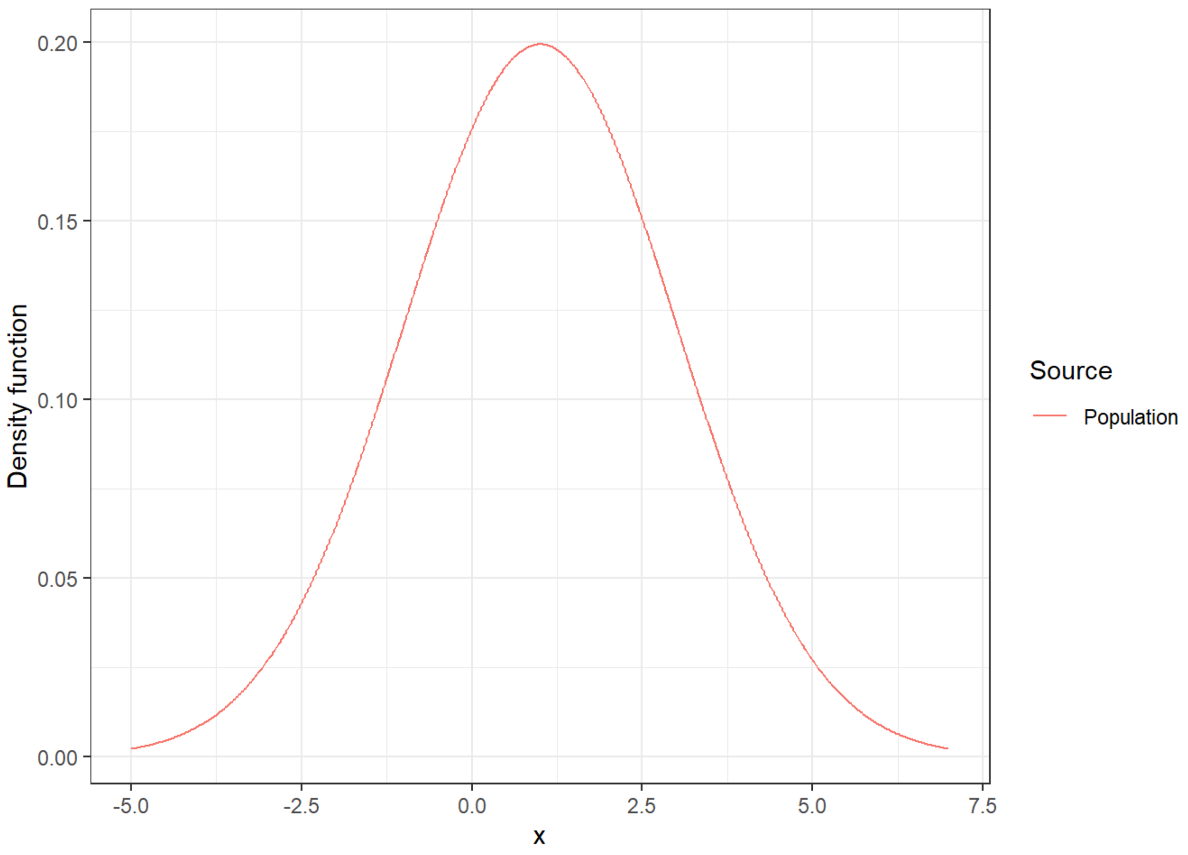
<u>Example</u> Suppose $X_1, \cdots, X_n \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$ are i.i.d. Gaussian

$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ is the maximum likelihood estimate for σ_0^2 .

$\Rightarrow$ $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$ is the maximum likelihood estimate for σ_0 .

Maximum likelihood simulation

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



Maximum likelihood simulation

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

We can generate simulated data from a Gaussian distribution to test the MLE method

```r
set.seed(123) # choose a random seed for reproducibility
```

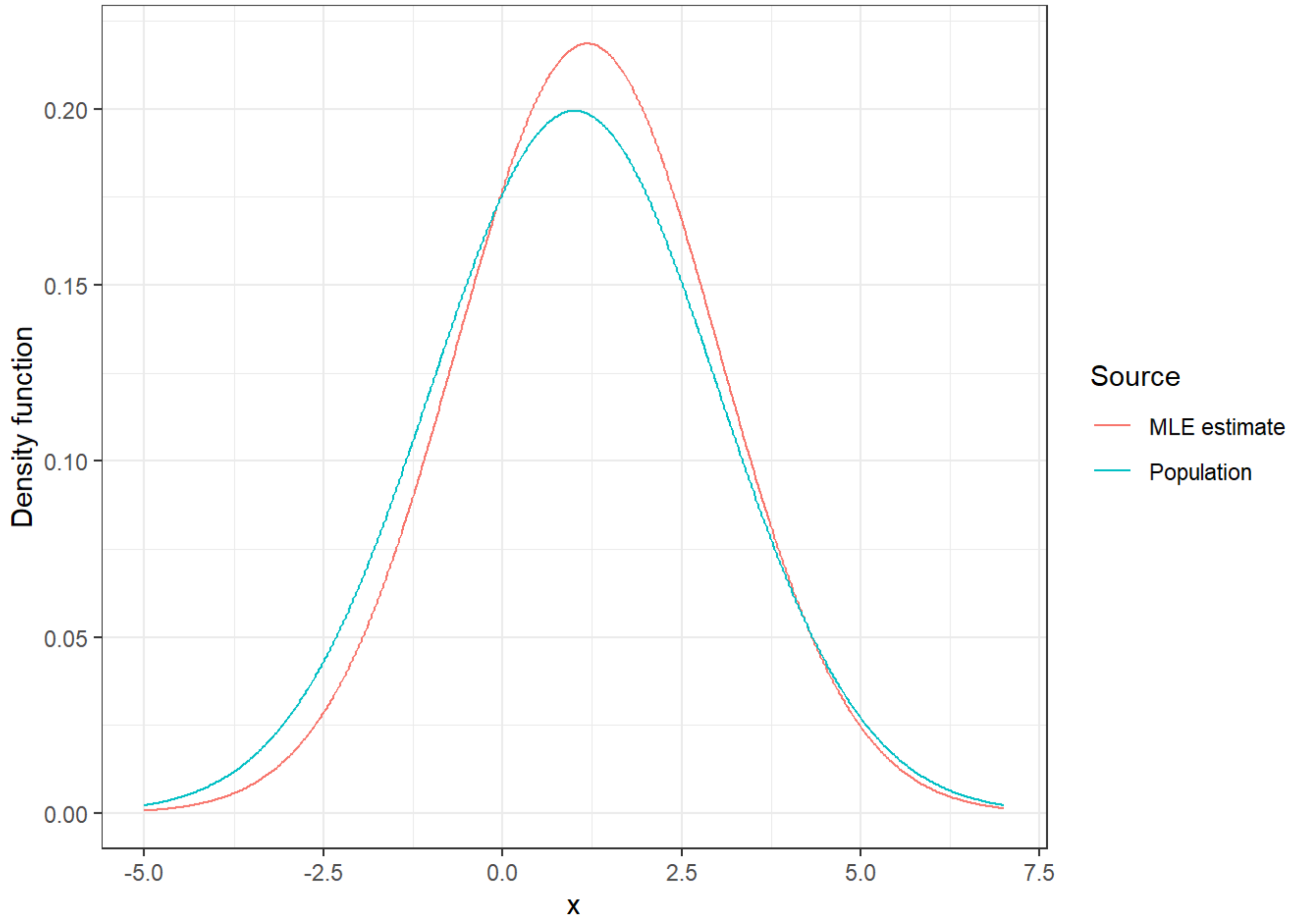
```r
sample_size<-100 # choose a sample size
sample_data <- rnorm(sample_size,mu,sigma) # generate some random data
```

```r
mu_mle<-mean(sample_data)
sigma_mle<-sd(sample_data)*sqrt((sample_size-1)/sample_size)
```

```r
df_gaussian<-df_gaussian%>%
  rbind(data.frame(x,Density=dnorm(x,mean=mu_mle,sd=sigma_mle),Source="MLE estimate")) # add in mle density
```

Maximum likelihood simulation

```
df_gaussian%>%ggplot(aes(x=x,y=Density,color=Source))+geom_line()+ylab("Density function")+theme_bw() # plot
```



Maximum likelihood with penguins data

Let’s fit a Gaussian model to the wights of Gentoo penguins

```
gentoo_weights<-penguins%>%
  filter(species=="Gentoo")%>%
   pull(body_mass_g) # extract the column of gentoo weights
```

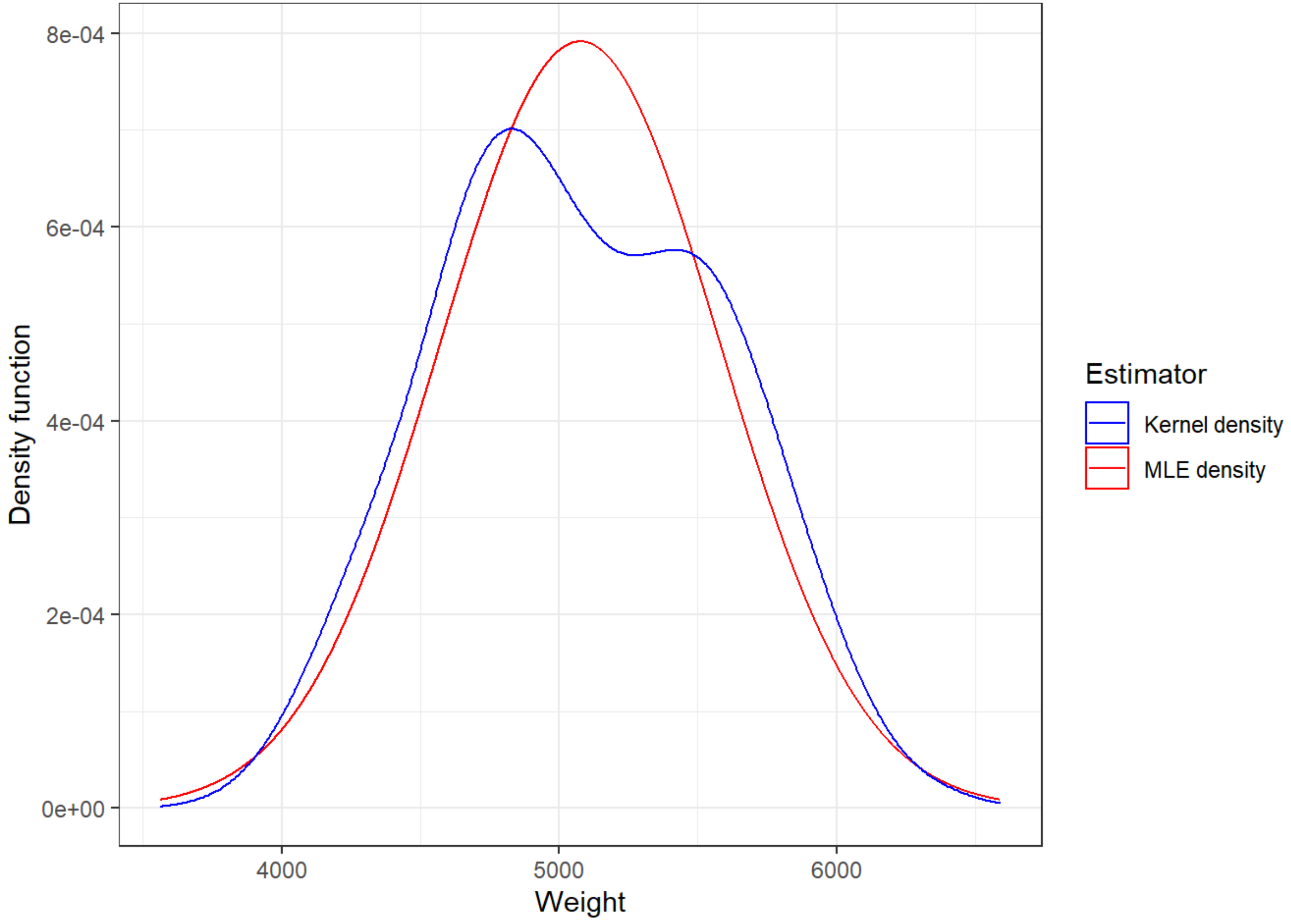
```
n<-length(gentoo_weights) # sample size
mu_mle_peng<-mean(gentoo_weights,na.rm=1) # compute mle mean
sigma_mle_peng<-sd(gentoo_weights,na.rm=1)*sqrt((n-1)/n) # compute mle standard deviation
```

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

```
weights<-seq(mu_mle_peng-3*sigma_mle_peng,mu_mle_peng+3*sigma_mle_peng,sigma_mle_peng*0.001) # generate indicies
colors<-c("MLE density"="red","Kernel density"="blue")  # set color legend
ggplot()+ geom_line(data=data.frame(Weight=weights,Density=dnorm(weights,mean=mu_mle_peng,sd=sigma_mle_peng)),
  aes(x=Weight,y=Density,color="MLE density"))+ # plot MLE
  geom_density(data=tibble(gentoo_weights),aes(x=gentoo_weights,color="Kernel density"))+ # plot kernel density
  labs(y="Density function",color="Estimator")+theme_bw()+scale_color_manual(values=colors)
```

Maximum likelihood with penguins data

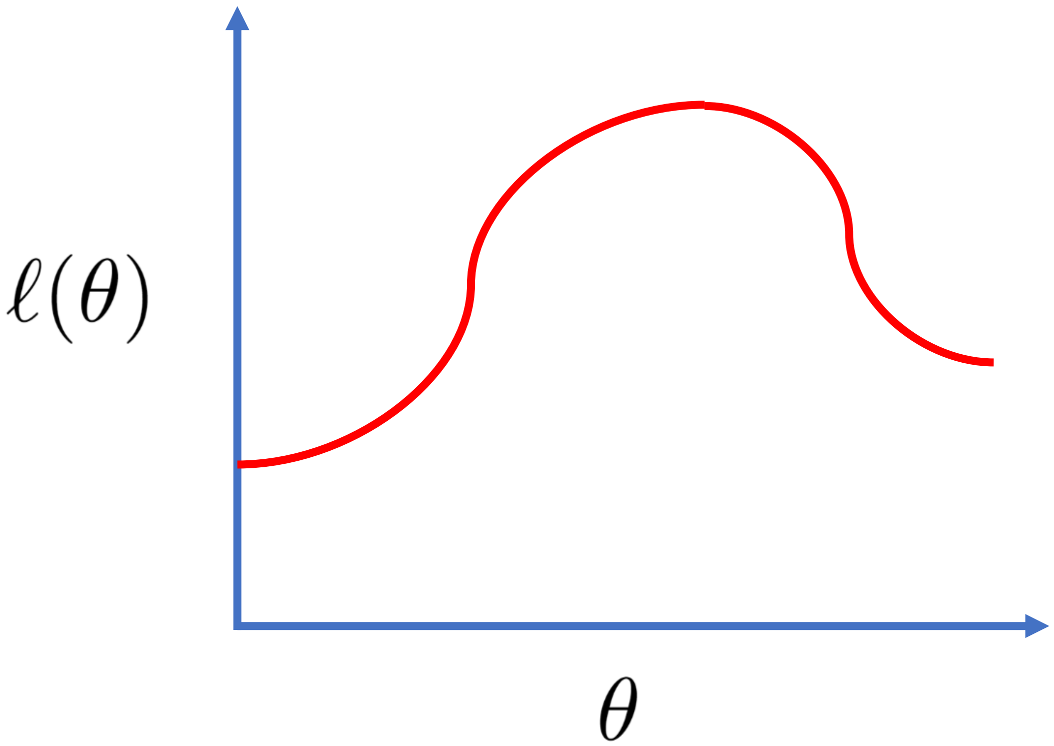
The parametric Gaussian model fitted with maximum likelihood estimation and a kernel density plot



Maximum likelihood estimation

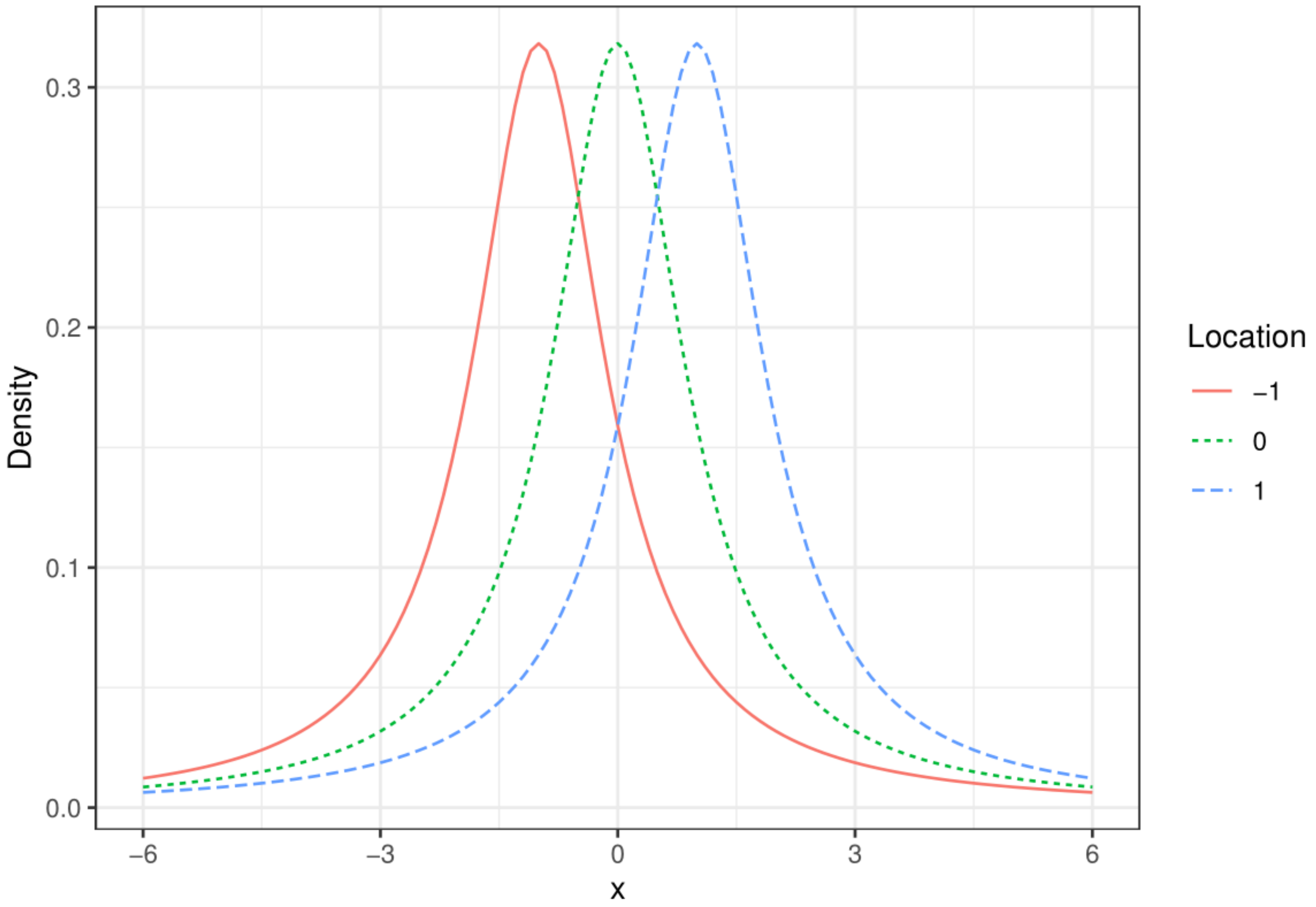
In many other cases there is no closed form solution.

We use techniques from optimization to maximize the likelihood function $\theta \mapsto \ell(\theta)$ numerically.



Maximum likelihood with a Cauchy distribution

A Cauchy random variable with location parameter θ has density $f_\theta(x) := \frac{1}{\pi\{1+(x-\theta)^2\}}$.



Maximum likelihood with a Cauchy distribution

A Cauchy random variable with location parameter θ has density $f_\theta(x) := \frac{1}{\pi\left\{1+(x-\theta)^2\right\}}$.

Let's suppose we have i.i.d. data $X_1, \ldots, X_n \sim f_{\theta_0}$ and we want to estimate θ_0.

Maximum likelihood with a Cauchy distribution

A Cauchy random variable with location parameter θ has density $f_\theta(x) := \frac{1}{\pi\left\{1+(x-\theta)^2\right\}}$.

Let's suppose we have i.i.d. data $X_1, \ldots, X_n \sim f_{\theta_0}$ and we want to estimate θ_0.

The likelihood function is given by

$$\ell(\theta) = \prod_{i=1}^{n} f_\theta(X_i) = \prod_{i=1}^{n} \frac{1}{\pi\left\{1 + (X_i - \theta)^2\right\}}.$$

Maximum likelihood with a Cauchy distribution

A Cauchy random variable with location parameter θ has density $f_\theta(x) := \frac{1}{\pi\left\{1+(x-\theta)^2\right\}}$.

Let's suppose we have i.i.d. data $X_1, \ldots, X_n \sim f_{\theta_0}$ and we want to estimate θ_0.

The likelihood function is given by

$$\ell(\theta) = \prod_{i=1}^{n} f_\theta(X_i) = \prod_{i=1}^{n} \frac{1}{\pi\left\{1+(X_i-\theta)^2\right\}}.$$

Equivalently, the log-likelihood function is given by

$$\log \ell(\theta) = \sum_{i=1}^{n} \log\left\{f_\theta(X_i)\right\} = -\sum_{i=1}^{n} \log\left\{1+(X_i-\theta)^2\right\} - n\log(\pi).$$

Maximum likelihood with a Cauchy distribution

A Cauchy random variable with location parameter θ has density $f_\theta(x) := \frac{1}{\pi\{1+(x-\theta)^2\}}$.

Let's suppose we have i.i.d. data $X_1, \ldots, X_n \sim f_{\theta_0}$ and we want to estimate θ_0.

The likelihood function is given by

$$\ell(\theta) = \prod_{i=1}^{n} f_\theta(X_i) = \prod_{i=1}^{n} \frac{1}{\pi\left\{1 + (X_i - \theta)^2\right\}}.$$

Equivalently, the log-likelihood function is given by

$$\log \ell(\theta) = \sum_{i=1}^{n} \log\{f_\theta(X_i)\} = -\sum_{i=1}^{n} \log\left\{1 + (X_i - \theta)^2\right\} - n\log(\pi).$$

Unfortunately, we do not have an analytic solution to $\hat{\theta}_n \in \operatorname{argmax}\{\ell(\theta)\}$.

Maximum likelihood with a Cauchy distribution

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

```
set.seed(0)
sample_size<-100
theta_0<-5
```

Maximum likelihood with a Cauchy distribution

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

```
set.seed(0)
sample_size<-100
theta_0<-5
```

```
cauchy_sample<-rcauchy(n=sample_size,location=theta_0)
# generate cauchy data
```

Maximum likelihood with a Cauchy distribution

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

```
set.seed(0)
sample_size<-100
theta_0<-5
```

```
cauchy_sample<-rcauchy(n=sample_size,location=theta_0)
# generate cauchy data
```

```
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))}
# the log likelihood function
```

Maximum likelihood with a Cauchy distribution

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

```
set.seed(0)
sample_size<-100
theta_0<-5
```

```
cauchy_sample<-rcauchy(n=sample_size,location=theta_0)
# generate cauchy data
```

```
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))}
# the log likelihood function
```

```
theta_ml_est<-optimise(f=log_lik_cauchy_X,interval=c(-1000,1000),maximum = TRUE)$maximum
# numerical optimisation to compute the maximum likelihood estimate
theta_ml_est
```

Maximum likelihood with a Cauchy distribution

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

```
set.seed(0)
sample_size<-100
theta_0<-5
```

```
cauchy_sample<-rcauchy(n=sample_size,location=theta_0)
# generate cauchy data
```

```
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))}
# the log likelihood function
```

```
theta_ml_est<-optimise(f=log_lik_cauchy_X,interval=c(-1000,1000),maximum = TRUE)$maximum
# numerical optimisation to compute the maximum likelihood estimate
theta_ml_est
```

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

Maximum likelihood with a Cauchy distribution

```r
set.seed(0)
num_trials<-100000
sample_size<-100
theta_0<-5
```

```r
log_lik_cauchy<-function(theta,sample_X){return(-sum(log(1+(sample_X-theta)^2)))} # log likelihood
```

Maximum likelihood with a Cauchy distribution

```
set.seed(0)
num_trials<-100000
sample_size<-100
theta_0<-5
```

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

```
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(-10,10),maximum = TRUE)$maximum

  return(theta_ml_est)

} # compute the maximum likelihood estimate
```

Maximum likelihood with a Cauchy distribution

```r
set.seed(0)
num_trials<-100000
sample_size<-100
theta_0<-5
```

```r
log_lik_cauchy<-function(theta,sample_X){return(-sum(log(1+(sample_X-theta)^2)))} # log likelihood
```

```r
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(-10,10),maximum = TRUE)$maximum

  return(theta_ml_est)

} # compute the maximum likelihood estimate
```

```r
cauchy_simulation_df<-data.frame(trial=seq(num_trials))%>%
  mutate(sample=map(.x=trial,~rcauchy(sample_size,location=theta_0)))%>%
  mutate(ml_est=map_dbl(.x=sample,.f=theta_ml))%>%
  mutate(med_est=map_dbl(.x=sample,.f=median))
```

Maximum likelihood with a Cauchy distribution

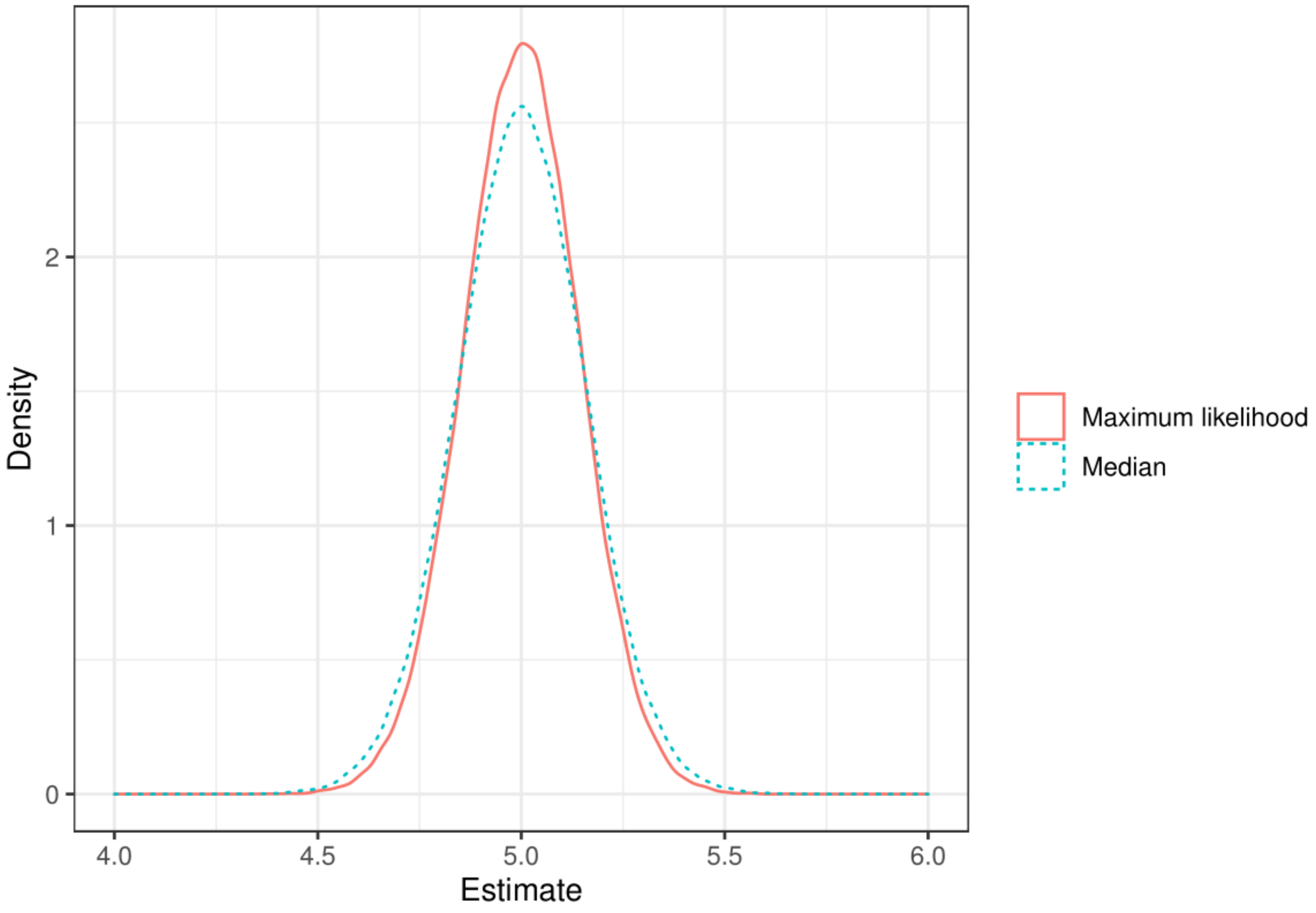
```r
theta_ml<-function(sample_X){

  log_lik_cauchy_X<-function(theta){return(log_lik_cauchy(theta,sample_X))}

  theta_ml_est<-optimise(f=log_lik_cauchy_X,interval=c(-10,10),maximum = TRUE)$maximum

  return(theta_ml_est)

} # compute the maximum likelihood estimate
```

```r
cauchy_simulation_df<-data.frame(trial=seq(num_trials))%>%
  mutate(sample=map(.x=trial,~rcauchy(sample_size,location=theta_0)))%>%
  mutate(ml_est=map_dbl(.x=sample,.f=theta_ml))%>%
  mutate(med_est=map_dbl(.x=sample,.f=median))
```

```r
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")))%>%
  ggplot(mapping=aes(x=value,color=name,linetype=name))+
  geom_density()+theme_bw()+xlim(c(4,6))+
  labs(color="",linetype="")+xlab("Estimate")+ylab("Density")
```

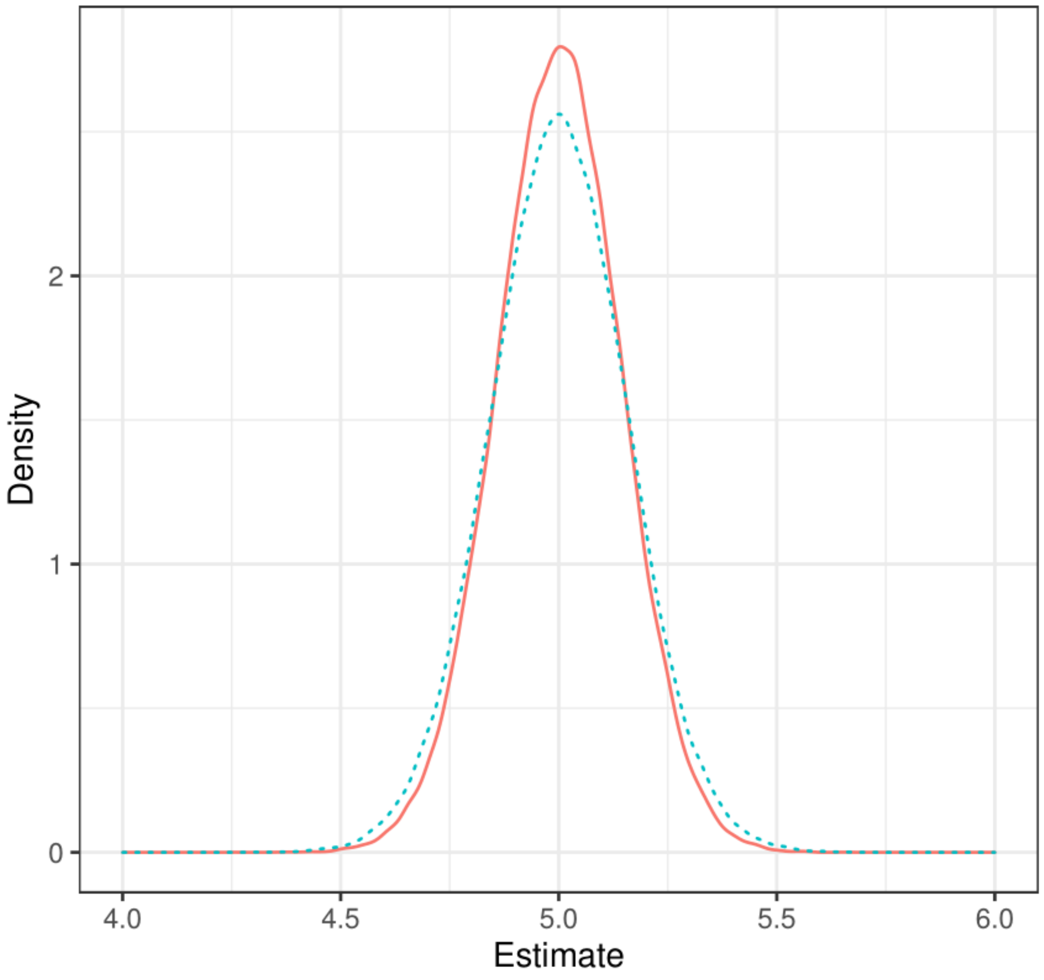
Maximum likelihood with a Cauchy distribution

We use numerical methods to maximise $\log \ell(\theta) = -\sum_{i=1}^{n} \log\left\{1 + (X_i - \theta)^2\right\} - n\pi.$



Maximum likelihood with a Cauchy distribution

We use numerical methods to maximise $\log \ell(\theta) = -\sum_{i=1}^{n} \log\left\{1 + (X_i - \theta)^2\right\} - n\pi.$




```r
med_estimate_mean_sqr_error<-cauchy_simulation_df%>%
  pull(med_est)%>%
  (function(x){return(mean((x-theta_0)^2))})
```

```r
med_estimate_mean_sqr_error
```

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

```r
ml_estimate_mean_sqr_error<-cauchy_simulation_df%>%
  pull(ml_est)%>%
  (function(x){return(mean((x-theta_0)^2))})
```

```r
ml_estimate_mean_sqr_error
```

```
## [1] 0.02062478
```

Maximum likelihood estimation

The maximum likelihood estimate (MLE) is **consistent** under natural conditions.

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

Maximum likelihood estimation

The maximum likelihood estimate (MLE) is **consistent** under natural conditions.

We have $\hat{\theta}(X_1, \cdots, X_n) \rightarrow \theta_0 \in \Theta$ as $n \rightarrow \infty$.

Example 1 Suppose $X_1, \cdots, X_n \sim \mathcal{B}(q_0)$. Then

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow q_0 \quad \text{as} \quad n \rightarrow \infty$$

Example 2 Suppose $X_1, \cdots, X_n \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$ are i.i.d. Gaussian

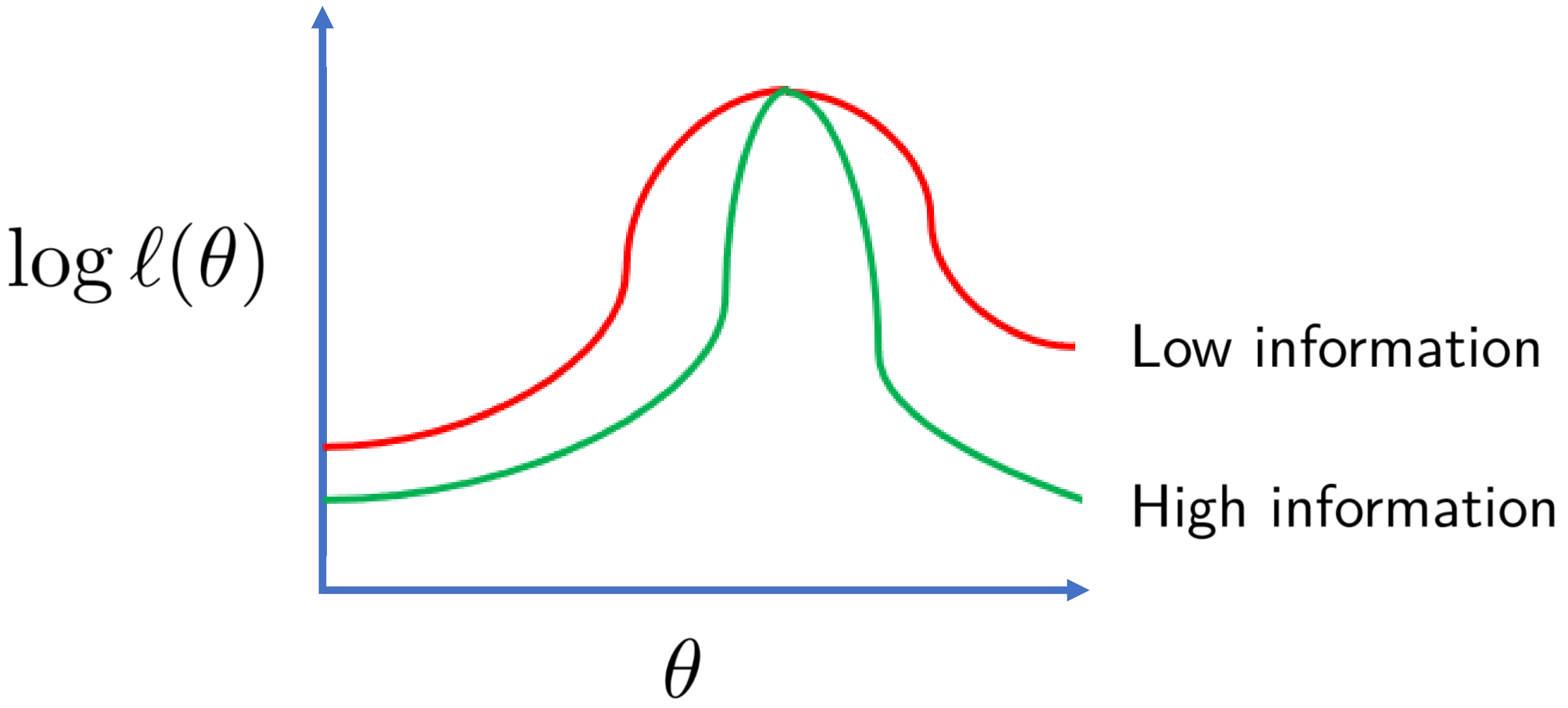
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow \mu_0 \quad \text{as} \quad n \rightarrow \infty$$

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

Maximum likelihood and Fisher information (*)

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

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



Maximum likelihood and Fisher information (*)

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

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

Let $\hat{\theta}_n$ be the maximum likelihood estimator based on a sample $X_1, \ldots, X_n \sim f_{\theta_0}$.
Let $Z \sim \mathcal{N}(0,1)$ be a standard Gaussian random variable and take $x \in \mathbb{R}$.
For a suitably well-behaved random variables $X_1, \ldots, X_n \sim f_{\theta_0}$ we have

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

Maximum likelihood and Fisher information (*)

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

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

Let $\hat{\theta}_n$ be the maximum likelihood estimator based on a sample $X_1, \ldots, X_n \sim f_{\theta_0}$.
Let $Z \sim \mathcal{N}(0,1)$ be a standard Gaussian random variable and take $x \in \mathbb{R}$.
For a suitably well-behaved random variables $X_1, \ldots, X_n \sim f_{\theta_0}$ we have

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

This implies that the maximum likelihood estimator $\hat{\theta}_n$ is a consistent estimator for θ_0.

Maximum likelihood and Fisher information (*)

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

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

Let $\hat{\theta}_n$ be the maximum likelihood estimator based on a sample $X_1, \ldots, X_n \sim f_{\theta_0}$.
Let $Z \sim \mathcal{N}(0,1)$ be a standard Gaussian random variable and take $x \in \mathbb{R}$.
For a suitably well-behaved random variables $X_1, \ldots, X_n \sim f_{\theta_0}$ we have

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

This implies that the maximum likelihood estimator $\hat{\theta}_n$ is a consistent estimator for θ_0.

Cramer and Rao showed that the variance level $\frac{1}{n \cdot \mathcal{I}(\theta_0)}$ 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 alternatives when analytic methods are unavailable.
- We also discussed some of the maximum likelihood method's favourable properties.

University of BRISTOL

Thanks for listening!

Henry W J Reeve

henry.reeve@bristol.ac.uk

Include EMATM0061 in the subject of your email.

Statistical Computing & Empirical Methods (EMATM0061)

MSc in Data Science, Teaching block 1, 2021.