Week 10 – Maximum Likelihood Estimation

Review - Unbiased Estimators

$X_1, \dots X_n$ a random sample from distribution F, and θ a parameter of interest about F (e.g. mean)

- An estimator T of a parameter θ is unbiased if $E[T] = \theta$
- An unbiased estimator has no systematic tendency to produce estimates that are larger than or smaller than the target parameter θ
- $\overline{X_n}$ is an unbiased estimator of the population mean μ.
- $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X_n} \right)^2$ is an unbiased estimator of the population variance σ^2
- Unbiasedness is not preserved under transformations.

Last Class- Choosing among unbiased estimators. Example

$X_1, \ldots, X_n \overset{i.i.d}{\sim} U[0, \theta]$

$f_X(x) = \frac{1}{\theta} I_{[0,\theta]}(x)$

$E[X_i] = \frac{\theta}{2}$, so $E[\overline{X_n}] = \frac{\theta}{2}$

Then $\widetilde{\theta} = 2\overline{X_n}$ is an unbiased estimator for θ

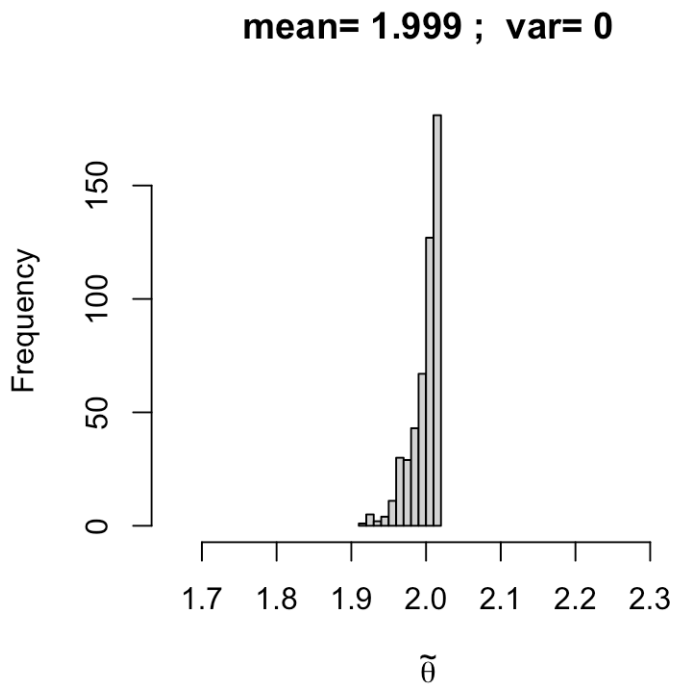
Intuitively a natural estimator for θ would also be based on $T = \max\{X_1, \ldots, X_n\}$

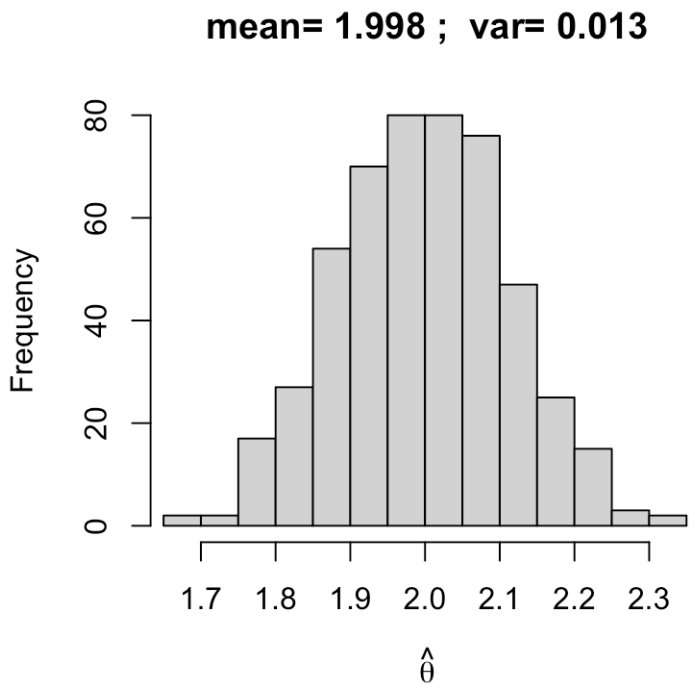
$f_T(t) = n\frac{t^{n-1}}{\theta^n} I_{[0,\theta]}(t)$

$E[T] = \frac{n}{n+1}\theta$ so $\widehat{\theta} = \frac{n+1}{n} T$ is an unbiased estimator of θ

So, both $\widetilde{\theta}$ and $\widehat{\theta}$ are unbiased estimators of θ, which one is better?

Last Class - Choosing among unbiased estimators-Example





Last Class - Choosing among unbiased estimators example.

- We see empirically that $\widetilde{\theta}$ is much less variable than $\widehat{\theta}$
- But also theoretically, $Var[\widetilde{\theta}] = \frac{\theta^2}{n(n+2)}$ and $Var[\widehat{\theta}] = \frac{\theta^2}{3n}$
- So, $Var[\widetilde{\theta}] < Var[\widehat{\theta}]$, for $n \geq 2$ and goes much faster to zero!
- Additionally, $\widehat{\theta}$ can take values way over the true θ
- So, $\widetilde{\theta}$ is a much better estimate of θ than $\widehat{\theta}$

Last Class- Mean square error

- Although unbiasedness is a desirable property, unbiased estimators do not always exist
- Even when they exist, requiring unbiadness maybe too stringent (i.e. there can be other good estimators that are biased)
- A general performance of an estimator can be judged by the way it spreads around the parameter to be estimated:

  If T is an estimator for a parameter θ, the mean squared error of T is the number:

  $MSE(T) = E[(T - \theta)^2]$.

- It's easy to show that $MSE(T) = Var(T) + Bias(T)^2$, where $Bias(T) = E[T] - \theta$
- A biased estimator with a small bias could be more useful than an unbiased estimator with a large variance.
- Better to use $MSE(T)$ to choose between estimators
- When $Bias(T) = 0, MSE(T) = Var(T)$

Last Class- Mean square error example

$$X_1, \ldots, X_n \overset{i.i.d}{\sim} Poisson(\mu)$$

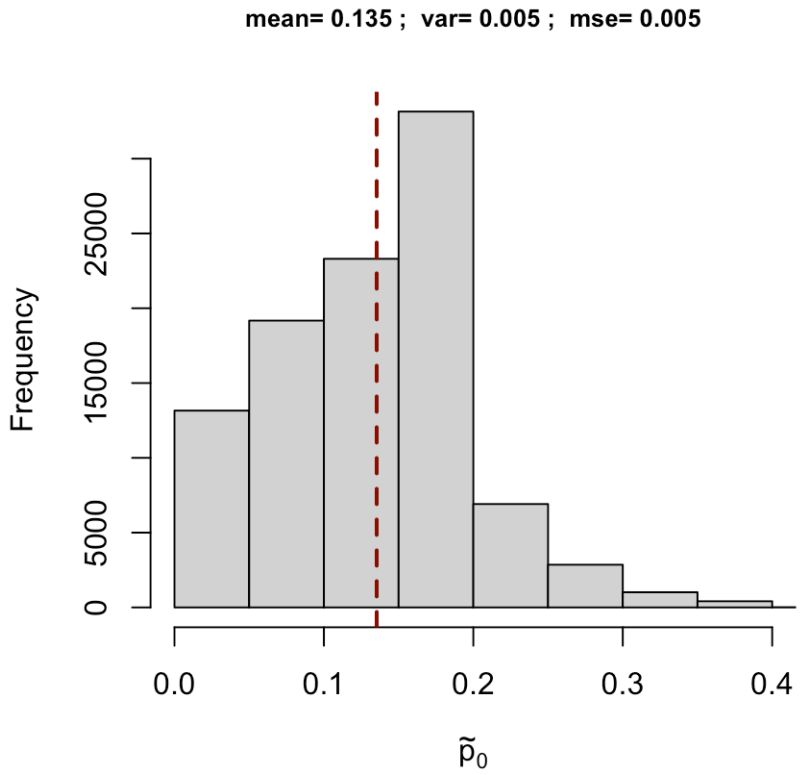
Two candidate estimators for $p_0 = P(X_i = 0) = e^{-\mu}$

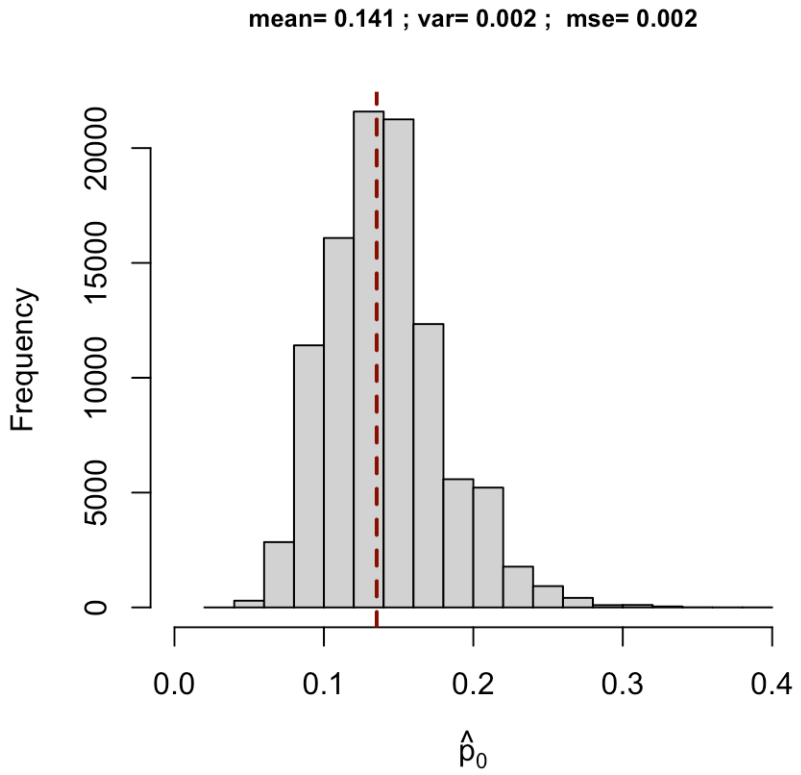
$\widetilde{p}_0 = \frac{\text{number of } X_i = 0}{n}$ and $\widehat{p}_0 = e^{-\overline{X_n}}$

$\widetilde{p}_0$ is unbiased but $\widehat{p}_0$ is not? Which one is better?

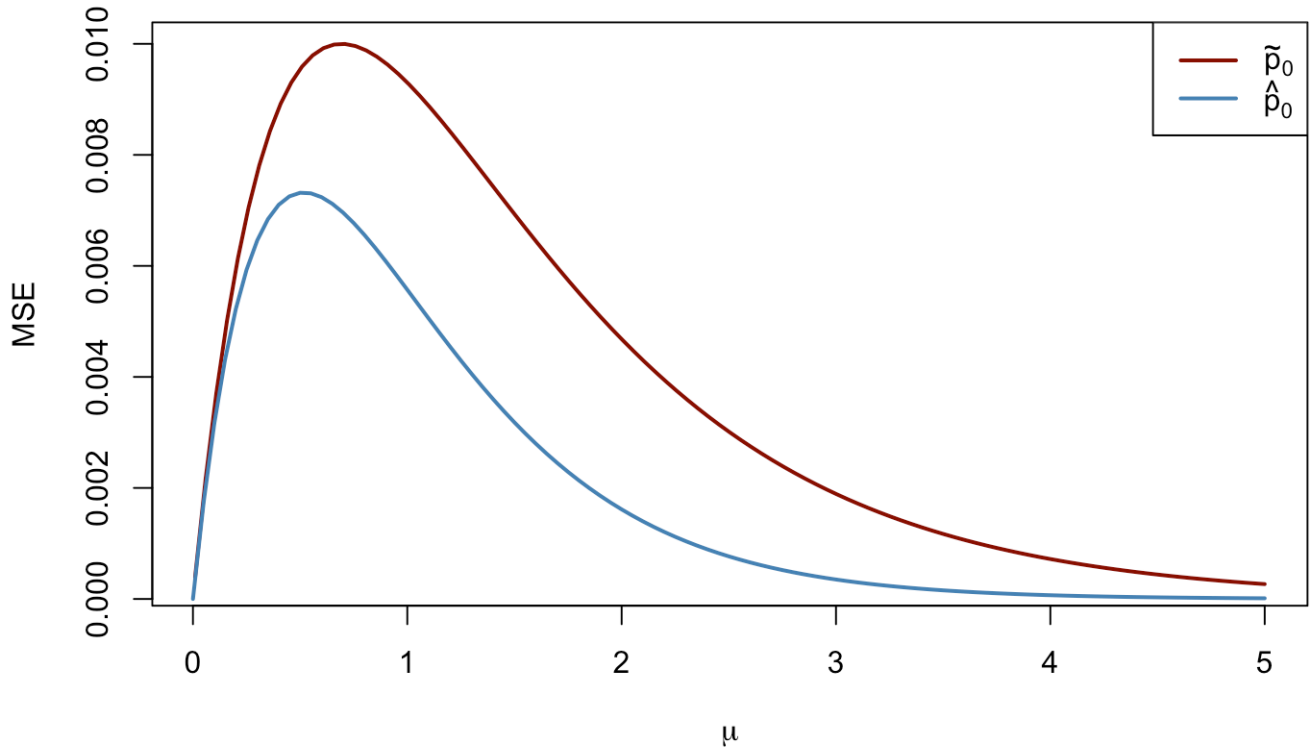
Last Class - Mean square error example

- Simulation to see how the two estimators behave
- True $p_0 = e^{-2} = 0.135$





Last Class - Mean square error example



Maximum likelihood estimation

- Previous examples of estimators were sample analogs of population parameters (e.g.sample mean and sample variance) or intuitive (e.g. estimator based on $\max(X_1, \dots, X_n)$ for $U[0, \theta]$)
- **Maximum likelihood provides a very general way to construct estimators with good properties. The idea is intuitive:**
- **Choose as the estimator the value of θ that maximizes the chances to produce the observed data.**
- Example: An infusion pump is a medical device that delivers fluids — typically medications, nutrients, or saline — into a patient's body in controlled amounts over time. We track **time to failure** (in hours) for five identical infusion pumps used in a hospital.
- Observed data in hours: $120, 85, 200, 60, 150$
- Assuming independent and identically distributed lifetimes following an $Exponential(\lambda)$ model. Compute the maximum likelihood estimate of λ.

MLE – exponential example

Data: $t_1 = 120, t_2 = 85, t_3 = 200, t_4 = 60, t_5 = 150$

Model: i.i.d data with density $f(t|\lambda) = \lambda e^{-\lambda t}$

Likelihood:

$$L(\lambda) = \prod_{i=1}^{5} \lambda e^{-\lambda T_i}$$

$$= (\lambda e^{-\lambda \cdot 120})(\lambda e^{-\lambda \cdot 85})(\lambda e^{-\lambda \cdot 200})(\lambda e^{-\lambda \cdot 60})(\lambda e^{-\lambda \cdot 150})$$

$$= \lambda^5 \exp[-\lambda(120 + 85 + 200 + 60 + 150)] = \lambda^5 \exp(-615\lambda)$$

log-likelihood:

$$\ell(\lambda) = \log L(\lambda) = 5 \log \lambda - 615\, \lambda$$

MLE – exponential example

Setting the score to zero to find the candidate maxima of the log-likelihood :

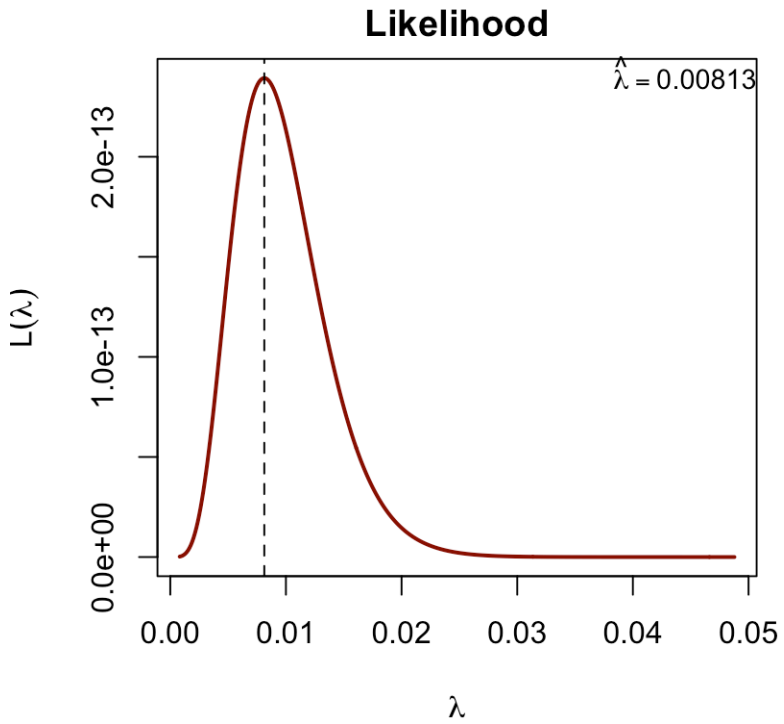
$$\frac{d\ell}{d\lambda} = \frac{5}{\lambda} - 615$$

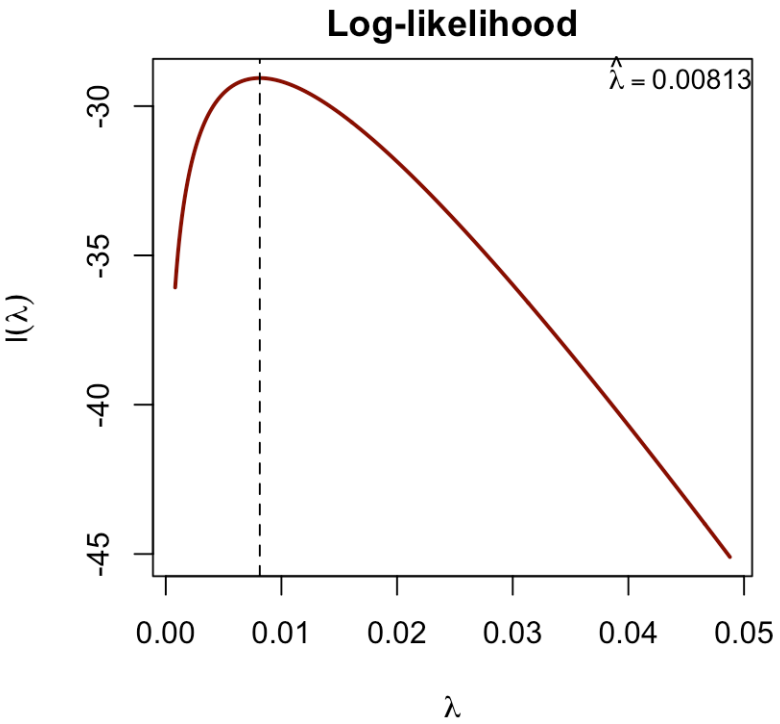
$$\frac{5}{\lambda} - 615 = 0 \quad \Rightarrow \quad \hat{\lambda} = \frac{5}{615} = 0.00813$$

Checking 2^{nd} derivative:

$$\frac{d^2\ell}{d\lambda^2} = -\frac{5}{\lambda^2} < 0$$

Maximum likelihood estimation – exponential example





Maximum likelihood estimation – exponential example

Let $T_1, \ldots, T_n \overset{iid}{\sim} \mathrm{Exp}(\lambda)$ with density $f(t;\lambda) = \lambda e^{-\lambda t}$ for $t > 0$. The likelihood and log-likelihood are

$$L(\lambda) \;=\; \prod_{i=1}^{n} \lambda e^{-\lambda T_i} \;=\; \lambda^n \exp\Big(-\lambda \sum_{i=1}^{n} T_i \Big),$$

$$\ell(\lambda) \;=\; \log L(\lambda) \;=\; n \log \lambda \;-\; \lambda \sum_{i=1}^{n} T_i .$$

Differentiate and set to zero:

$$\frac{d\ell}{d\lambda} \;=\; \frac{n}{\lambda} \;-\; \sum_{i=1}^{n} T_i \;=\; 0 \quad\Longrightarrow\quad \widehat{\lambda} \;=\; \frac{n}{\sum_{i=1}^{n} T_i}.$$

MLE - a more complex example

Number of cycles until pregnancy:

Table 21.1. Observed numbers of cycles up to pregnancy.

Number of cycles	1	2	3	4	5	6	7	8	9	10	11	12	>12
Smokers	29	16	17	4	3	9	4	5	1	1	1	3	7
Nonsmokers	198	107	55	38	18	22	7	9	5	3	6	6	12

Source: C.R. Weinberg and B.C. Gladen. The beta-geometric distribution applied to comparative fecundability studies. *Biometrics*, 42(3):547–560, 1986.

MLE - a more complex example

- Want to estimate $p = P(X = 1)$, where X_i is the number of cycles up to pregnancy among the i^{th} smoker
- We know that $S = \frac{\text{number of } X_i=1}{n}$ is an unbiased estimator for p
- This yields the estimate $\widetilde{p} = \frac{29}{100} = 0.29$
- But it uses only a fraction of the data discarding the rest. Maybe there is a better estimator using all the data?

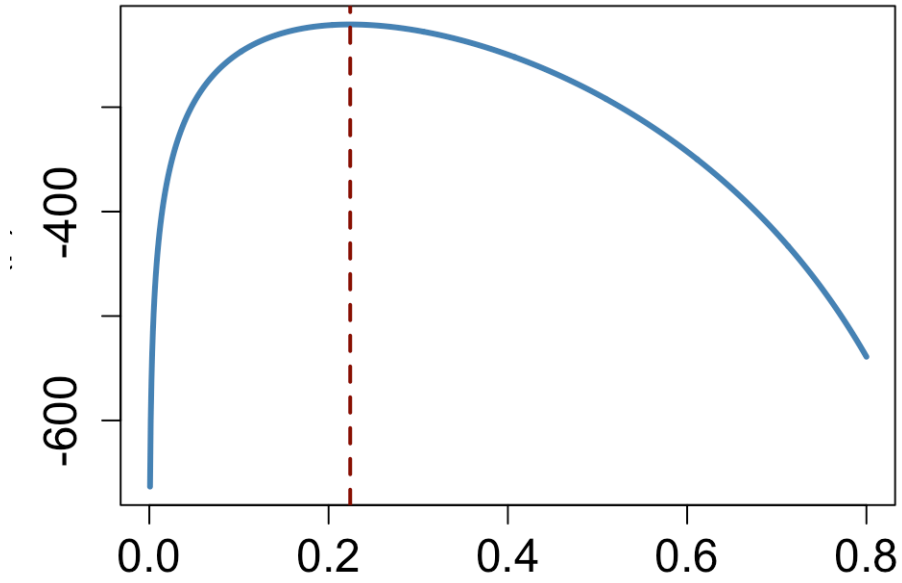
MLE - a more complex example

- Let X_i denote the number of cycles up to pregnancy for the i^{th} smoker. We can assume the $X_i \sim Geometric(p)$, and the X_i are independent. There are 415 smokers
- The probability of observing the data in the first row of the table is then given by:

$$\begin{aligned} L(p) &= P(X_1 = x_1 \mid p) \times P(X_2 = x_2 \mid p) \times \ldots P(X_{100} = x_{100} \mid p) \\ &= p^{29} . (p(1-p))^{16} \ldots \left((1-p)^{11} p\right)^3 . \left((1-p)^{12}\right)^7 = \\ &= p^{93}(1-p)^{322} \end{aligned}$$

- We want to find $0 \leq p \leq 1$ that makes $L(p)$ as large as possible, i.e. find the value of p that maximizes $L(p)$,
- Equivalent to finding the maximizer of $l(p) = log(L(p))$

MLE - a more complex example



$$l(p) = 93 \log(p) + 322 \log(1 - p)$$

$$0 = l'(p) = \frac{93}{p} - \frac{322}{1-p} \iff p = 93/415$$

$\hat{p} = 93/415 = 0.224$ is called the maximum likelihood estimate of p.

Maximum likelihood estimation

- The maximum likelihood estimate of θ is the value $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$ that maximizes the log-likelihood function $l(\theta)$.
- The corresponding random variable $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ is called the maximum likelihood estimator of θ.
- (We often use the same notation, $\hat{\theta}$, for both the estimate and the estimator. We understand which one we are referring to based on context)

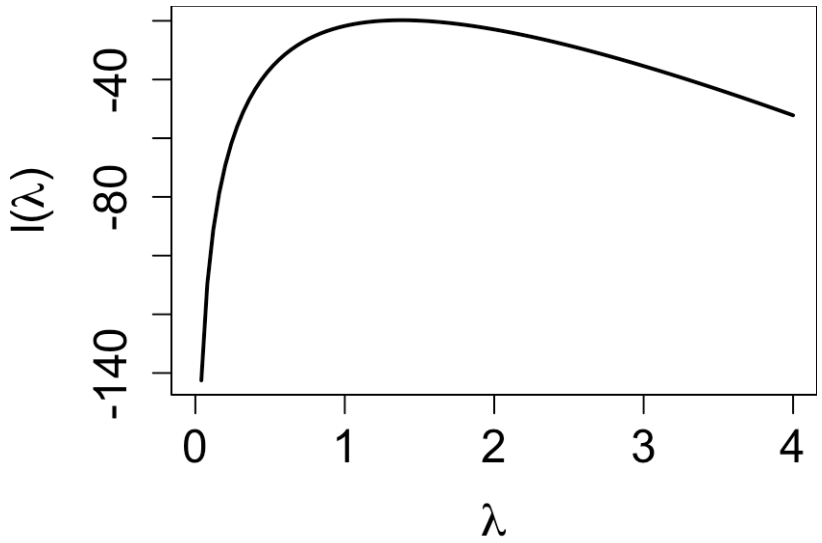
MLE with no analytical solution

Dataset: $x_1, \dots, x_n$ modeled as a realization of random variables $X_1, \dots, X_n$.

Where the pdf of X_i is a mixture of exponentials:

$X_i \sim Exp(\lambda)$ with probability $\frac{1}{2}$ and $Exp(2\lambda)$ with probability $\frac{1}{2}$

$f(x) = \frac{1}{2}\lambda e^{-\lambda x} + \frac{1}{2} 2\lambda e^{-2\lambda x}$



MLE with no analytical solution

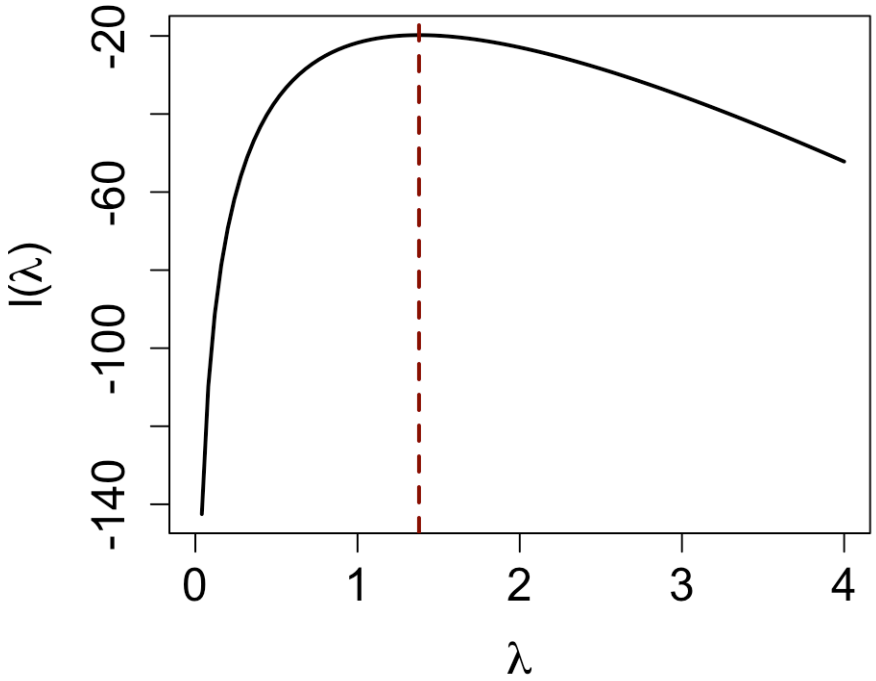
```
set.seed(101)
n = 50
mix = rbinom(n=n, size=1, prob=1/2)
rate=1
dataset = mix*rexp(n, rate=rate) + (1-mix)*rexp(n, rate=2*rate)

negloglik = function(x, data){-sum(log(0.5*dexp(data, rate=x) + 0.5*dexp(data, rate=2*x)))}

optimize(f=negloglik, interval=c(0.01, 10), data=dataset)
```

```
## $minimum
## [1] 1.37906
##
## $objective
## [1] 19.78651
```

MLE with no analytical solution



Properties of Maximum likelihood estimators

The maximum likelihood estimator (MLE) is consistent:

- $\widehat{\theta}_n \xrightarrow{P} \theta$ as $n \to +\infty$.
- i.e. as the sample size gets larger the MLE gets closer and closer to the true parameter θ.
- Consistency is much stronger that asymptotic unbiasedness.

Invariance:

- If $\widehat{\theta}$ is the MLE of θ then $g(\widehat{\theta})$ is the MLE of $g(\theta)$.

Asymptotic minimum variance No other consistent estimator has lower asymptotic mean squared error

Asymptotic normality - In most situations of interest (but not always, e.g. $U[0, \theta]$) the sampling distribution of the MLE gets closer and closer to a normal distribution as $n \to \infty$

With few exceptions MLE's are generally biased estimators

Asymptotic variance of the MLE

- Let $X_1, \ldots, X_n \sim f(x; \theta)$, and the log-likelihood function: $l(\theta) = \sum_{i=1}^{n} \log f(X_i; \theta)$
- The Fisher's information of θ (for one sample) is defined as:

$$I(\theta) = -E\left[\frac{\partial^2 l(\theta, X)}{\partial \theta^2}\right]$$

- If $f(x; \theta)$ is 'well behaved' (e.g. its support does not depend on θ)

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{1}{I(\theta))}\right)$$

- $\frac{1}{nI(\theta))}$ is the asymptotic variance of $\hat{\theta}$
- The standard error of the MLE is $se(\hat{\theta}) = \frac{1}{\sqrt{nI(\theta))}}$
- We can estimate the standard error of the MLE by $\widehat{se}(\hat{\theta}) = \frac{1}{\sqrt{nI(\hat{\theta})}}$

Next class

- Inference for the mean – confidence intervals and hypothesis tests.
- Read PSD 8.1-8.3