Point Estimation

Ricardo Dahis

PUC-Rio, Department of Economics

Summer 2023

estimation?

- in statistics, we often want to recover some population parameter, but we only have a sample drawn from that population.

- in other words, we want to estimate the population parameters
  - a point estimator is any statistic $T(X_1, \ldots, X_N)$ of the sample
  - an estimate is the realization of T for a given sample $x = (x_1, \ldots, x_n)$

- and derive some important properties about estimators
  - is it biased? consistent? asymptotic distribution?
  - is it the most efficient estimator available?
  - how to conduct hypothesis testing? (coming soon...)

- in this lecture, we will discuss
  1. methods for finding estimators
  2. methods for assessing estimators

Contents

1. estimation methods

1.1 method of moments

1.2 maximum likelihood

1.3 Bayesian approach

2. evaluating estimators

2.1 Cramér-Rao inequality

2.2 attainment

2.3 Rao-Blackwell

3. exercises

Contents

1. estimation methods

1.1 method of moments

1.2 maximum likelihood

1.3 Bayesian approach

2. evaluating estimators

2.1 Cramér-Rao inequality

2.2 attainment

2.3 Rao-Blackwell

3. exercises

how to find an estimator?

- method of moments
  - estimate population moments with their sample analogs
  - simple and intuitive
- maximum likelihood
  - describe data fully, not only moments
  - efficient, invariant, full of desirable properties
- Bayesian approach
  - fundamentally different approach: θ is a quantity with possibly non-zero variation, rather than fixed
  - sample updates the belief about the parameter

Contents

1. estimation methods

1.1 method of moments

1.2 maximum likelihood

1.3 Bayesian approach

2. evaluating estimators

2.1 Cramér-Rao inequality

2.2 attainment

2.3 Rao-Blackwell

3. exercises

sample analog of moments: intuition

- let $\{X_1, \ldots, X_n\}$ denote a random sample from a population with pmf/pdf $f(x|\theta)$ where $\theta = [\theta_1, \cdots, \theta_k]'$.

- the pdf has population moments $\mathbb{E}(X^j) = \mu_j$ and we assume that at least k moments exist.

- population moments are typically known functions of the parameters

$$\mu_j = h_j(\theta)$$

  e.g. if $X \sim N(\mu, \sigma^2)$, then $\mu_1 = \mu$ and $\mu_2 = \sigma^2 + \mu^2$.

- so we equate the sample moments to population moments, and then solve for the parameters

$$\begin{aligned} \hat{\mu}_1 &\equiv n^{-1}\sum_{i=1}^{n} X_i &= \hat{\mu} \\ \hat{\mu}_2 &\equiv n^{-1}\sum_{i=1}^{n} X_i^2 &= \hat{\sigma}^2 + \hat{\mu}^2 \end{aligned}$$

  so the method of moments estimates are $\hat{\mu} = \hat{\mu}_1$ and $\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2$.

sample analog of moments: intuition

- rewriting,

$$\begin{aligned}
\hat{\mu} &= \hat{\mu}_1 = n^{-1}\sum_{i=1}^{n} X_i = \bar{X}_n \\
\hat{\sigma}^2 &= \hat{\mu}_2 - \hat{\mu}_1^2 = n^{-1}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2 \\
&= n^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2
\end{aligned}$$

as we could expect.

sample analog of moments

- generally, the method of moments estimator $\hat{\theta} = [\hat{\theta}_1, \ldots, \hat{\theta}_k]'$ is the solution to the simultaneous equations that match the first k sample moments to the corresponding k population moments:

$$\begin{cases} \hat{\mu}_1 & \equiv & \frac{1}{n}\sum_{i=1}^{n} X_i & = & h_1(\hat{\theta}_1, \ldots, \hat{\theta}_k) \\ & \vdots & & & \\ \hat{\mu}_k & \equiv & \frac{1}{n}\sum_{i=1}^{n} X_i^k & = & h_k(\hat{\theta}_1, \ldots, \hat{\theta}_k) \end{cases}$$

- we need at least k moments to solve the system of equations for k parameters.

- even considering k moments, the system of equation may not have an unique solution for θ, but we will assume this case away.

- an interesting case arises when we have more equations than unknowns.
  - for example, if $X \sim \text{Poisson}(\lambda)$, we have that $\mathbb{E}(X) = \text{var}(X) = \lambda$
  - so we have two equations and one unknown!
  - this system might not have a solution at all.

method of moments for gamma

- example: suppose we have a sample $\{x_1, \ldots, x_n\}$ drawn from the gamma(α,β) distribution, given by

$$f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}$$

with $\mathbb{E}(X) = \alpha\beta$ and var$(X) = \alpha\beta^2$. So $\mathbb{E}(X^2) = \alpha\beta^2 + \alpha^2\beta^2$. Equating to sample moments,

$$\begin{aligned} \hat{\mu}_1 &\equiv n^{-1}\sum_{i=1}^{n} X_i = \hat{\alpha}\hat{\beta} \\ \hat{\mu}_2 &\equiv n^{-1}\sum_{i=1}^{n} X_i^2 = \hat{\alpha}\hat{\beta}^2 + \hat{\alpha}^2\hat{\beta}^2 \end{aligned}$$

and solving for the parameters in terms of quantities calculated from data,

$$\begin{aligned} \hat{\mu}_2 &= \hat{\alpha}\hat{\beta}^2 + \hat{\alpha}^2\hat{\beta}^2 \Rightarrow \hat{\mu}_2 - \hat{\mu}_1^2 = \hat{\alpha}\hat{\beta}^2 \Rightarrow \hat{\beta} = \frac{\hat{\mu}_2 - \hat{\mu}_1^2}{\hat{\mu}_1} \\ \hat{\alpha} &= \hat{\mu}_1 \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2} \end{aligned}$$

method of moments for exponential

- example: suppose we have a sample $\{x_1, \ldots, x_n\}$ drawn from exponential(λ) distribution, given by

$$f(x|\lambda) \quad = \quad \lambda e^{-\lambda x}$$

with $\mathbb{E}(X) = \lambda^{-1}$ and var$(X) = \lambda^{-2}$, so $\mathbb{E}(X^2) = 2\lambda^{-2}$. We can propose two method of moments estimators:

$$\hat{\mu}_1 \quad \equiv \quad n^{-1} \sum_{i=1}^{n} X_i \quad = \quad \hat{\lambda}^{-1} \quad \Rightarrow \quad \hat{\lambda} \quad = \quad \frac{1}{\hat{\mu}_1}$$

and

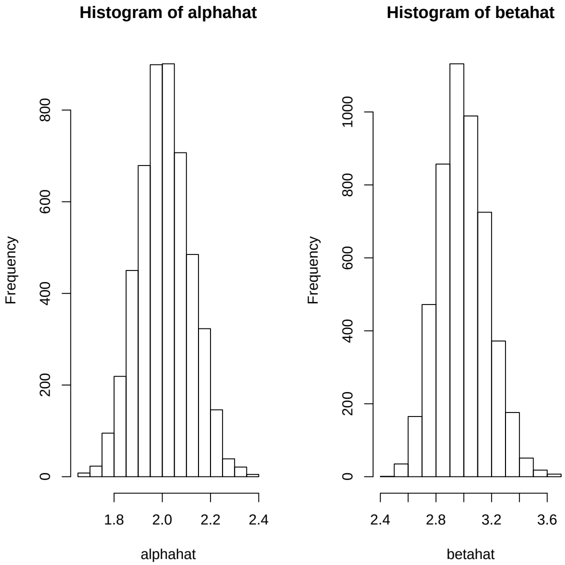
$$\hat{\mu}_2 \quad \equiv \quad n^{-1} \sum_{i=1}^{n} X_i^2 \quad = \quad 2\hat{\lambda}^{-2} \quad \Rightarrow \quad \hat{\lambda} \quad = \quad \sqrt{\frac{2}{\hat{\mu}_2}}$$

method of moments in practice - R codes

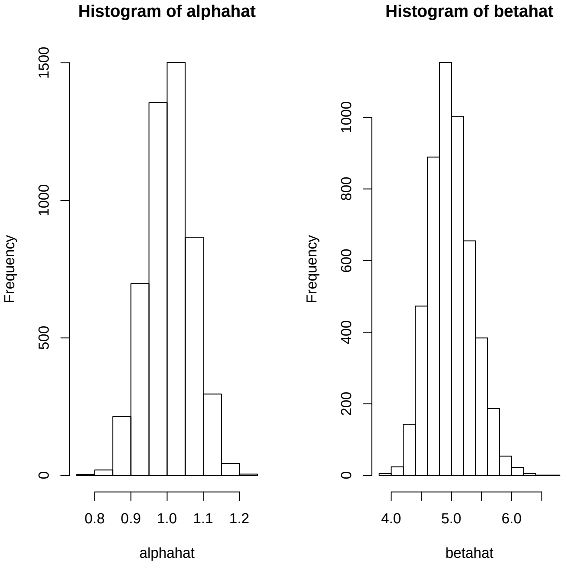
$$\begin{aligned}\hat{\alpha} &= \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}\\ \hat{\beta} &= \frac{\hat{\mu}_2 - \hat{\mu}_1^2}{\hat{\mu}_1}\end{aligned}$$

```
 MMsamplerGamma <- function(alpha,beta,n){
  hats <- matrix(0,5000,2)
  for (i in 1:5000){
    randomVec <- rgamma(n,scale=beta,shape=alpha)
    mu1 <- mean(randomVec)
    mu2 <- mean(randomVec^2)
    hats[i,1] <- mu1^2/(mu2-mu1^2)
    hats[i,2] <- (mu2-mu1^2)/mu1
  }
  hats
}
```

Gamma(2,3)


Histogram of alphahat
Frequency
0
200
400
600
800
1.8
2.0
2.2
2.4
alphahat
Histogram of betahat
Frequency
0
200
400
600
800
1000
2.4
2.8
3.2
3.6
betahat

Gamma(1,5)


Histogram of alphahat
Frequency
0
500
1000
1500
0.8
0.9
1.0
1.1
1.2
alphahat
Histogram of betahat
Frequency
0
200
400
600
800
1000
4.0
5.0
6.0
betahat

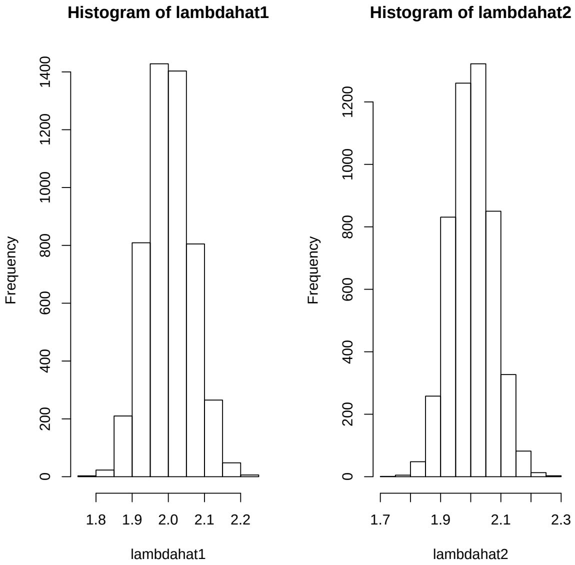
method of moments in practice - R codes

$$\hat{\lambda}_1 = \frac{1}{\hat{\mu}_1}$$

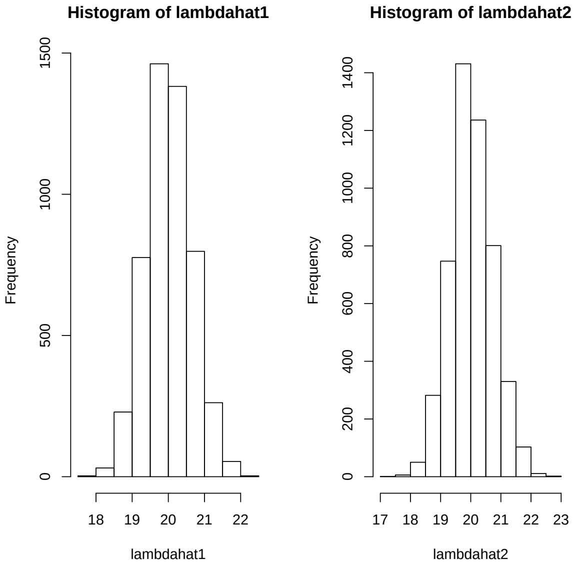
$$\hat{\lambda}_2 = \sqrt{\frac{2}{\hat{\mu}_2}}$$

```
 MMsamplerExponential <- function(lambda,n){
  hats <- matrix(0,5000,2)
  for (i in 1:5000){
    randomVec <- rexp(n,lambda)
    mu1 <- mean(randomVec)
    mu2 <- mean(randomVec^2)
    hats[i,1] <- 1/mu1
    hats[i,2] <- (2/mu2)^(0.5)
  }
  hats
}
```

exp(2)


Histogram of lambdahat1
Frequency
0
200
400
600
800
1000
1200
1400
1.8
1.9
2.0
2.1
2.2
lambdahat1
Histogram of lambdahat2
Frequency
0
200
400
600
800
1000
1200
1.7
1.9
2.1
2.3
lambdahat2

exp(20)


Histogram of lambdahat1
Frequency
0
500
1000
1500
18
19
20
21
22
lambdahat1
Histogram of lambdahat2
Frequency
0
200
400
600
800
1000
1200
1400
17
18
19
20
21
22
23
lambdahat2

Contents

1. estimation methods

1.1 method of moments

1.2 maximum likelihood

1.3 Bayesian approach

2. evaluating estimators

2.1 Cramér-Rao inequality

2.2 attainment

2.3 Rao-Blackwell

3. exercises

likelihood function

- definition: if $x = (x_1, \ldots, x_n)$ is a random sample from a population with pdf/pmf $f_X(x|\boldsymbol{\theta})$, the likelihood function is

$$\begin{aligned} \ell(\theta|x) &= \ell(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n) &=& f_X(x|\theta) \\ & &=& \prod_{i=1}^{n} f_X(x_i|\theta) \end{aligned}$$

  where $\theta \in \Theta$.

- discrete case: if we compare the likelihood functions at θ_1 and θ_2 and find that

$$\Pr_{\theta_1}(X = x) \;=\; \ell(\theta_1|x) \;>\; \ell(\theta_2|x) \;=\; \Pr_{\theta_2}(X = x),$$

  then the sample we observe is more likely to stem from $\theta = \theta_1$ than from $\theta = \theta_2$. in other words, θ_1 is more plausible that θ_2 given $X = x$.

- definition: the maximum likelihood estimator (MLE) of the parameter vector θ based on a sample $X = (X_1, \ldots, X_n)$. That is,

$$\hat{\theta}_n(X) \;=\; \arg\max_{\theta \in \Theta} \; \ell(\theta_1, \ldots, \theta_k|X_1, \ldots, X_n)$$

pdf vs likelihood function

- from $\ell(\theta|x) = f_X(x|\theta)$, it is key to note that
  - the pdf is a function of data, given parameters
  - the likelihood is a function of the parameters, given data

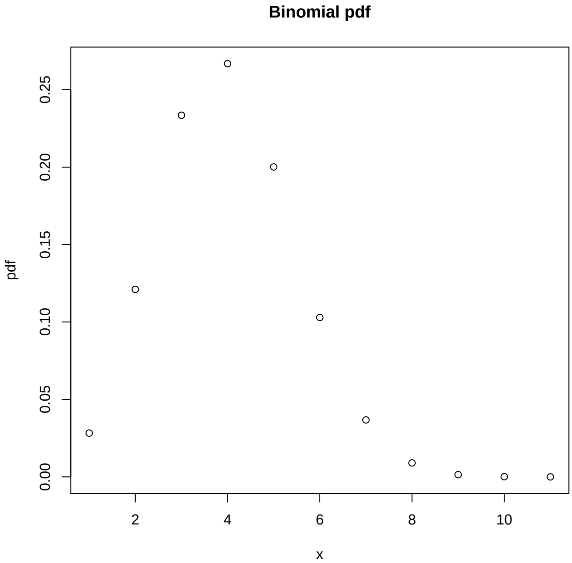
- example: let X have a binomial distribution. The p.d.f. is a function of x, given p,

$$f_X(x|p = 0.3) \quad = \quad \binom{10}{x}(0.3)^x(0.7)^{10-x}$$

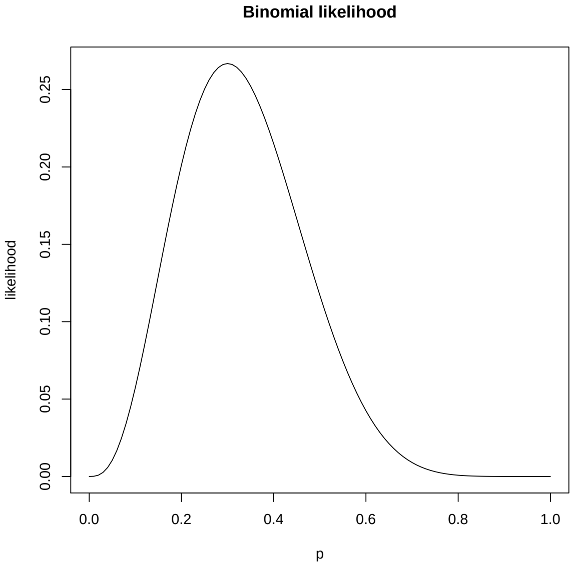
and the likelihood is a function of p given x

$$\ell(p|x = 3) \quad = \quad \binom{10}{3}p^3(1-p)^{10-3}$$

pdf vs likelihood function



pdf vs likelihood function



how to find and verify the global maximum?

- if the likelihood is differentiable in $\boldsymbol{\theta}$, possible candidates for the MLE are the values of $\boldsymbol{\theta}$ that solve $\frac{\partial}{\partial \theta_i} \ell(\boldsymbol{\theta}|\boldsymbol{x}) = 0$ for $i = 1, \ldots, k$

- only possible candidates because…
  - necessary, but not sufficient, condition for an interior maximum
  - first derivative may be nonzero if the extrema occur at the boundaries
  - check boundary separately and make sure the points at which first derivative is zero are not only local maxima, local/global minima or inflection points

Gaussian likelihood

- example: if $X_1, \dots, X_n$ are iid $N(\mu, 1)$, then

$$\begin{aligned} \ell(\mu|\boldsymbol{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x_i-\mu)^2\right] \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\ \frac{\mathrm{d}}{\mathrm{d}\mu}\ell(\mu|\boldsymbol{x})\bigg|_{\mu=\hat{\mu}} &= 0 \ \Rightarrow \ \sum_{i=1}^{n}(x_i-\hat{\mu}) = 0 \ \Rightarrow \ \hat{\mu} = \bar{x}_n \end{aligned}$$

  now, $\bar{x}_n$ is not only the unique solution of $\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0$, but also such that $\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\ell(\mu|\boldsymbol{x})\Big|_{\mu=\bar{x}_n} < 0$, so it is a global maximum.

- checking the likelihood at the boundaries: $\lim_{\mu\to\pm\infty}\ell(\mu|\boldsymbol{x}) = 0$

MLE for the Bernoulli distribution

- taking logs: easier to work with the log-likelihood function. Since ln is a monotone function,

$$\hat{\theta}_n(\boldsymbol{X}) \quad = \quad \arg\max_{\theta\in\Theta} \; \ell(\boldsymbol{\theta}|\boldsymbol{x}) \quad = \quad \arg\max_{\theta\in\Theta} \; \ln \ell(\boldsymbol{\theta}|\boldsymbol{x})$$

- example: let $(X_1, \ldots, X_n)$ be a random sample from a Bernoulli population with parameter $0 \leq p \leq 1$, then the likelihood function reads

$$\begin{aligned} \ell(p|\boldsymbol{x}) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{n\bar{x}_n}(1-p)^{n(1-\bar{x}_n)} = \ell(p|\bar{x}_n) \\ \ln \ell(p|\bar{x}_n) &= n\bar{x}_n \ln p + n(1-\bar{x}_n)\ln(1-p) \\ \frac{\mathrm{d}}{\mathrm{d}p} \ln \ell(p|\bar{x}_n)\bigg|_{p=\hat{p}} &= \frac{n\bar{x}_n}{\hat{p}} - \frac{n(1-\bar{x}_n)}{1-\hat{p}} = 0 \;\Rightarrow\; \hat{p} = \bar{x}_n \end{aligned}$$

- checking the boundaries: likelihood function is monotone in p given that $\ln \ell(p|\bar{x}_n = 0) = n\ln(1-p)$ and that $\ln \ell(p|\bar{x}_n = 1) = n \ln p$, and it is easy to verify that $\hat{p} = \bar{x}_n$, confirming the sample mean as the MLE of p.

MLE for the Gaussian distribution

- example: let $X_1, \ldots, X_n$ be iid $n(\theta, \sigma^2)$ with both θ and σ^2 unknown. Then

$$\begin{aligned}\ell(\theta, \sigma^2|\boldsymbol{x}) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{\sigma^2}\right\} \\ \ln \ell(\theta, \sigma^2|\boldsymbol{x}) &= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{\sigma^2}\end{aligned}$$

taking the first derivatives with respect to θ and σ^2 and equating to zero,

$$\begin{aligned}\frac{\partial}{\partial\theta}\ln \ell(\theta, \sigma^2|\boldsymbol{x}) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta) \;\Rightarrow\; \hat{\theta} = \bar{x}_n \\ \frac{\partial}{\partial\sigma^2}\ln \ell(\theta, \sigma^2|\boldsymbol{x}) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\theta)^2 \\ &\Rightarrow\; \hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2\end{aligned}$$

- you may want to show that this is indeed a local maximum

maximum likelihood invariance property

- say that we have a mapping $\theta \to \tau(\theta)$, $\eta = \tau(\theta)$ and that we are actually interested in $\hat{\eta}$, but $\hat{\theta}$ is easier to find or compute.

- example: in the betting odds example, we may find that the MLE of p is $\hat{p}$, but we are actually interested in $\frac{\hat{p}}{1-\hat{p}}$.

- if we are to apply maximum likelihood, we could
  - define a likelihood $\ell(\theta|x)$, but then we obtain $\hat{\theta}$ and we are interested in $\hat{\eta}$.
  - define an alternative likelihood $\ell^*(\eta|x)$. We obtain $\hat{\eta}$ directly, but this option is harder to compute.

- the maximum likelihood invariance property states that

$$\ell^*(\hat{\eta}) \quad = \quad \ell^*(\tau(\hat{\theta}))$$

  so, for example, MLE of $\frac{p}{1-p}$ is $\frac{\hat{p}}{1-\hat{p}}$.

- the invariance property is useful if we want to find out functions of estimators: e.g., if the MLE of θ is $\bar{X}_n$, then the MLE of θ^2 is $\bar{X}_n^2$.

maximum likelihood invariance property

if the mapping is bijective, then

- we can maximize the likelihood as a function of θ or $\eta = \tau(\theta)$:

$$\ell^*(\eta|x) \;=\; \prod_{i=1}^{n} f(x_i|\tau^{-1}(\eta)) \;=\; \ell(\tau^{-1}(\eta)|\boldsymbol{x})$$

- and the supremum unchanged

$$\sup_{\eta} \ell^*(\eta|\boldsymbol{x}) \;=\; \sup_{\eta} \ell(\tau^{-1}(\eta)|\boldsymbol{x}) \;=\; \sup_{\theta} \ell(\theta|\boldsymbol{x})$$

- the maximum of $\ell^*(\eta)$ is attained at $\hat{\eta} = \tau(\hat{\theta})$.

maximum likelihood invariance property

- the mapping τ not always is bijective
  - there may be more that one θ such that $\tau(\theta) = \eta$
  - define the induced likelihood ℓ^* given by

$$\ell^*(\eta|x) \quad = \quad \sup_{\theta:\tau(\theta)=\eta} \ell(\theta|x)$$

  and $\hat{\eta}$ that maximizes $\ell^*(\eta|x)$ is the MLE of η

- theorem (CB 7.2.10): if $\hat{\theta}$ is the MLE of θ, for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

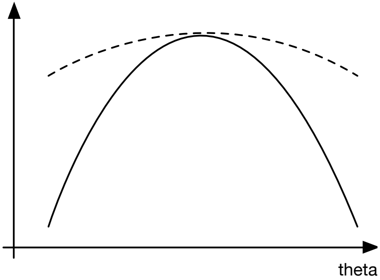
- proof: given the definition of the induced likelihood function,

$$\ell^*(\hat{\eta}|x) \quad = \quad \sup_{\eta} \sup_{\{\theta:\tau(\theta)=\eta\}} \ell(\theta|x) \quad = \quad \sup_{\theta} \ell(\theta|x) \quad = \quad \ell(\hat{\theta}|x)$$

whereas $\ell(\hat{\theta}|x) = \sup_{\{\theta:\tau(\theta)=\tau(\hat{\theta})\}} \ell(\theta|x) = \ell^*(\tau(\hat{\theta})|x)$. ■

computational issues

- MLE stems from a maximization process, so it is prone to numerical difficulties, especially if the log-likelihood is flat in the neighborhood of its maximum



- we also need to find the global maximum, which is not a trivial task.

- even finding local maxima may be numerically challenging, especially if there are non linearities in the parameters.

alternatives: direct maximization

- usually simpler algebraically, but sometimes harder to implement because there are no set rules to follow

- one general technique is to bound the likelihood from above and then establish that there is a unique point that attains the global upper bound.

- example: recall that $\sum_{i=1}^{n}(x_i - a)^2 \geq \sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ for any constant a, with equality if and only if $a = \bar{x}_n$, and hence

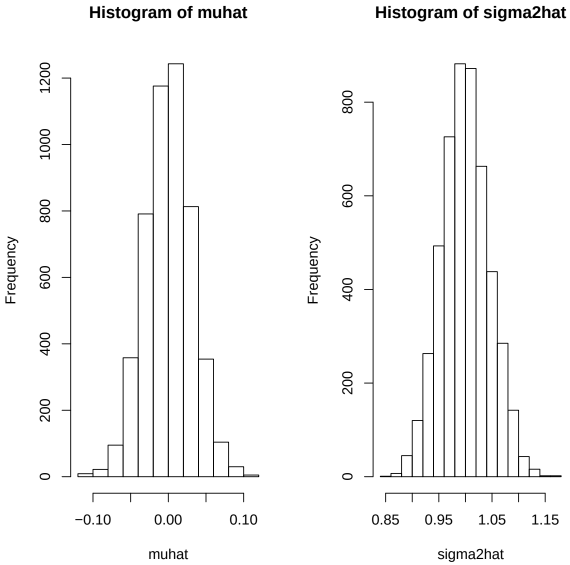
$$\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \leq \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\right]$$

  with equality if and only if $\mu = \bar{x}_n$ so $\bar{X}_n$ is the MLE.

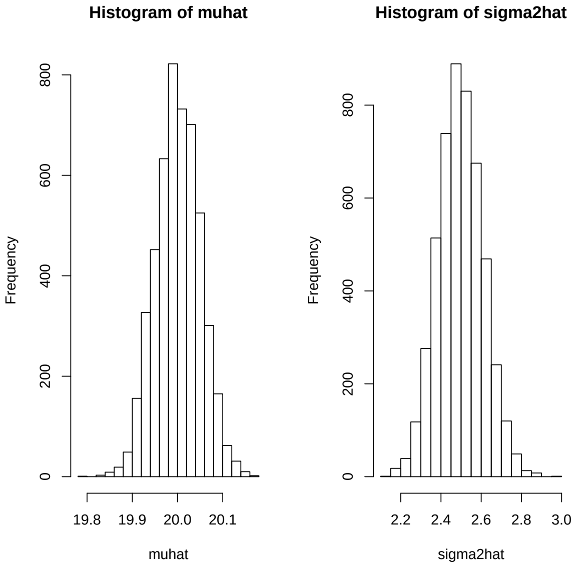
MLE in practice - R codes

```
 MLEsampler <- function(mu,sigma2,n){
  logLik <- function(theta){
    n <- length(randomVec)
    mu <- theta[1]
    sigma2 <- theta[2]
    logLik <- -0.5*n*log(sigma2)-0.5*sum((randomVec-mu)^2)/sigma2
    logLik <- -1*logLik
  }
  hats <- matrix(0,5000,2)
  for (i in 1:5000){
    randomVec <- rnorm(n,mean=mu,sd=sqrt(sigma2))
    thetahat <- optim(c(10,2),logLik)
    hats[i,1] <- thetahat$par[1]
    hats[i,2] <- thetahat$par[2]
  }
  hats
}
```

normal(0,1)


Histogram of muhat
Frequency
0
200
400
600
800
1000
1200
−0.10
0.00
0.10
muhat
Histogram of sigma2hat
Frequency
0
200
400
600
800
0.85
0.95
1.05
1.15
sigma2hat

normal(20,2.5)


Histogram of muhat
Frequency
0
200
400
600
800
19.8
19.9
20.0
20.1
muhat
Histogram of sigma2hat
Frequency
0
200
400
600
800
2.2
2.4
2.6
2.8
3.0
sigma2hat

Contents

1. estimation methods

1.1 method of moments

1.2 maximum likelihood

1.3 Bayesian approach

2. evaluating estimators

2.1 Cramér-Rao inequality

2.2 attainment

2.3 Rao-Blackwell

3. exercises

Bayesian statistics

- in classical statistics, the parameter θ is thought to be an unknown and fixed quantity: parameter θ exists in nature and we develop methods to estimate θ and conduct inference.

- frequentist probabilities: restrict the assignment of probabilities to statements that describe an outcome of an experiment that can be repeated. Then the probability is the limit when the experiment is repeated an infinite number of times.

- Bayesian approach: θ can be described by a probability distribution and reflects some subjective prior knowledge of the researcher
  - data updated the prior distribution, generating the posterior distribution
  - updating is done with the use of the Bayes' Rule

- subjective probabilities: probabilities applicable to any situation in which there is uncertainty, subject to respecting Kolmogorov's axioms

Bayesian statistics

- remember Bayes' rule: for any two events A and B,

$$\mathbb{P}(A \cap B) \;=\; \mathbb{P}(A|B)\mathbb{P}(B) \;=\; \mathbb{P}(B|A)\mathbb{P}(A) \;\Rightarrow\; \mathbb{P}(A|B) \;=\; \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

  which is also applicable to pdf's.

- Setting $A = \theta$ and $B = x$, we get

$$\pi(\theta|x) \;=\; \frac{p(x|\theta)\cdot\pi(\theta)}{p(x)}$$

  where
  - $\pi(\theta|x)$ is the posterior: updated distribution of the parameters given data
  - $p(x|\theta)$ is the likelihood: the probability that a given stretch of data is observed, conditional on the parameters
  - $\pi(\theta)$ is the prior: the distribution of the parameter which reflects the statistician's subjective views on the parameters
  - $p(x)$ is the marginal distribution of x, $p(x) = \int p(y,\theta)d\theta = \int p(y|\theta)\pi(\theta)d\theta$

Bayesian statistics

- intuition:
  1. statistician has certain beliefs about the parameters, reflected on the prior
  2. data x becomes available
  3. statistician updates his/her belief about the distribution of parameters conditional on data, generating the posterior

- if no data is available, statistician will report the beliefs, $\pi(\theta|x) = \pi(\theta)$.

- as the amount of data goes to infinity, it can be shown that the posterior becomes the likelihood $\pi(\theta|x) = p(x|\theta)$.

Bayesian statistics

- the equation

$$\pi(\theta|x) \quad = \quad \frac{p(x|\theta)\cdot\pi(\theta)}{p(x)}$$

  is the basis of Bayesian statistics and econometrics.

- note that $p(x)$ is invariant with respect to θ, so

$$\pi(\theta|x) \quad \propto \quad p(x|\theta)\cdot\pi(\theta)$$

  which is simpler to compute since $p(y)$ might be cumbersome to compute.

- $p(x|\theta)\cdot\pi(\theta)$ does not integrate to one, but maintains the shape of $\pi(\theta|x)$ and is much quicker to compute.

coin-tossing example, once more

- example: we are interested in estimating the probability of a head in a coin toss, which is coded as $\mathbb{P}(y_i = 1) = \theta$, so $\mathbb{P}(y_i = 0) = 1 - \theta$.

- a classical statistician would maximize the likelihood

$$\begin{aligned} p(y_1, \dots, y_n|\theta) &= \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{n-\sum_{i=1}^{n} y_i} \\ &= \theta^{n\bar{y}_n}(1-\theta)^{n(1-\bar{y}_n)} \end{aligned}$$

  which, as we have seen, gives $\hat{\theta} = n^{-1}\sum_{i=1}^{n} y_i = \bar{y}_n$.

- a Bayesian statistician would require a prior on the parameters $\pi(\theta)$ to compute

$$\pi(\theta|y_1, \dots, y_n) \quad \propto \quad p(y_1, \dots, y_n|\theta) \cdot \pi(\theta)$$

coin-tossing example, once more

- Statistician A believes that $\pi(\theta) = \mathcal{I}\left\{0 \leq \theta \leq 1\right\}$. This is also known as the uninformative prior. In this case,

$$\pi(\theta|y_1, \ldots, y_n) \quad \propto \quad \theta^{n\bar{y}_n}(1-\theta)^{n(1-\bar{y}_n)} \cdot \mathcal{I}\left\{0 \leq \theta \leq 1\right\}$$

  and the posterior is proportional to the likelihood function, given that $0 \leq \theta \leq 1$ is a natural range for θ.

- Statistician B believes that $\pi(\theta) = \frac{1}{h}\mathcal{I}\left\{\frac{1}{2} - \frac{1}{2}h \leq \theta \leq \frac{1}{2} + \frac{1}{2}h\right\}$ for a very small h. In this case,

$$\pi(\theta|y_1, \ldots, y_n) \quad \propto \quad \theta^{n\bar{y}_n}(1-\theta)^{n(1-\bar{y}_n)} \cdot \frac{1}{h}\mathcal{I}\left\{\frac{1}{2} - \frac{1}{2}h \leq \theta \leq \frac{1}{2} + \frac{1}{2}h\right\}$$

  which will also be very close to $\frac{1}{2}$. In particular, the posterior is zero for the regions of the prior such that $\pi(\theta) = 0$.

coin-tossing example, once more

- Statistician C proposes a more flexible specification: assumes that $\theta \sim \text{Beta}(\alpha, \beta)$, so

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

  with $0 \leq \theta \leq 1$ and $\alpha, \beta > 0$.

- α and β are referred to as hyper-parameters.

- Why a Beta distribution?
  - depending on choices of α and β, captures vast array of beliefs
  - as we will see, facilitates the computation of the posterior

coin-tossing example, once more

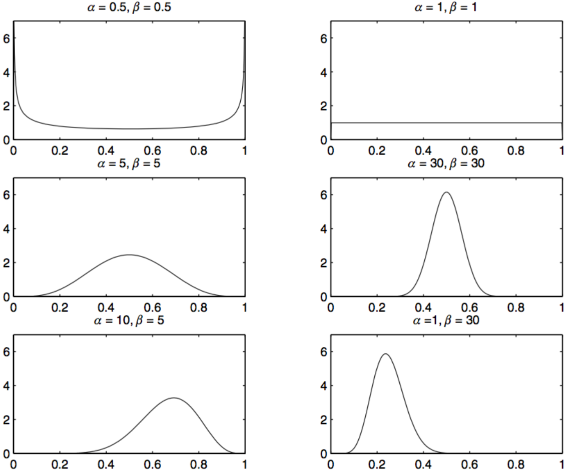


Figure 2.1. Beta distributions for various values of α and β.

coin-tossing example, once more

- the posterior distribution is

$$
\begin{aligned}
\pi(\theta|y) &\propto \theta^{n\bar{y}_n}(1-\theta)^{n(1-\bar{y}_n)} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&\propto \theta^{n\bar{y}_n}(1-\theta)^{n(1-\bar{y}_n)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&\propto \theta^{n\bar{y}_n+\alpha-1}(1-\theta)^{n(1-\bar{y}_n)+\beta-1}
\end{aligned}
$$

  so $\theta|y$ has the shape of a Beta(α_1, β_1), with $\alpha_1 = n\bar{y}_n + \alpha$ and $\beta_1 = n(1-\bar{y}_n) + \beta$, except for a normalization constant.

- using the fact that if $z \sim$ Beta$(\dot{\alpha}, \dot{\beta})$,

$$
\mathbb{E}(z) = \frac{\dot{\alpha}}{\dot{\alpha}+\dot{\beta}} \text{ and } \operatorname{var}(z) = \frac{\dot{\alpha}\dot{\beta}}{(\dot{\alpha}+\dot{\beta})^2(\dot{\alpha}+\dot{\beta}+1)}
$$

  then

$$
\mathbb{E}(\theta|y) = \frac{\alpha_1}{\alpha_1+\beta_1} \text{ and } \operatorname{var}(\theta|y) = \frac{\alpha_1\beta_1}{(\alpha_1+\beta_1)^2(\alpha_1+\beta_1+1)}
$$

conjugate priors

- expanding the expression for $\mathbb{E}(\theta|y)$,

$$\begin{aligned}\mathbb{E}(\theta|y) &= \frac{\alpha_1}{\alpha_1+\beta_1} = \frac{n\bar{y}_n+\alpha}{n\bar{y}_n+\alpha+n(1-\bar{y}_n)+\beta} \\ &= \frac{n\bar{y}_n+\alpha}{\alpha+\beta+n} = \frac{\alpha}{\alpha+\beta+n}+\frac{n}{\alpha+\beta+n}\bar{y}_n\end{aligned}$$

- as the sample size grows, $n \to \infty$,

$$\mathbb{E}(\theta|y) = \underbrace{\frac{\alpha}{\alpha+\beta+n}}_{\longrightarrow 0}+\underbrace{\frac{n}{\alpha+\beta+n}}_{\longrightarrow 1}\bar{y}_n$$

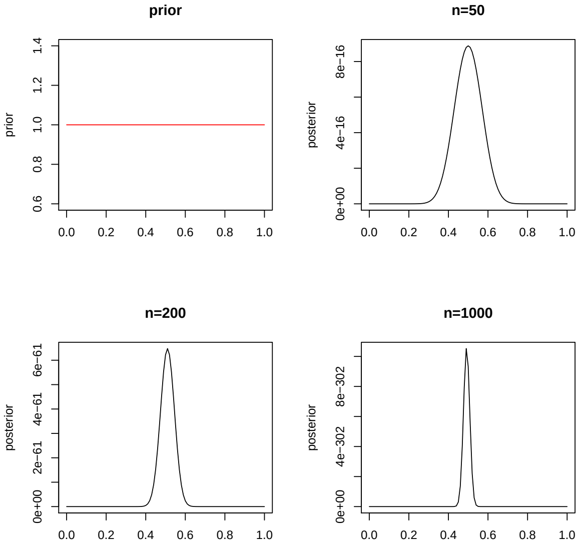
- in other words, as the sample size increases, the prior distribution has smaller relevance in determining the posterior distribution.

- this is an example of conjugate priors: smart choices of priors for a given shape of the likelihood such that it is easy to compute the posterior.

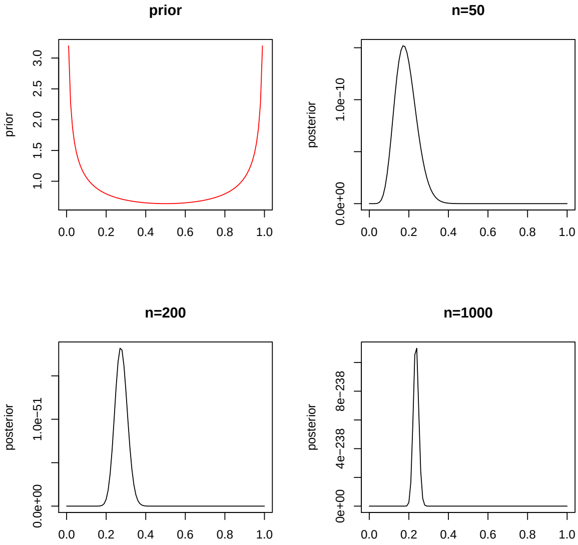
Bayes in practice - R codes

```
 plotPosterior <- function(alpha,beta,n){
  ybar <- mean(y[1:n])
  posteriorDist <-
    function(theta){theta^{n*ybar+alpha-1}*(1-theta)^{n*(1-ybar)+beta-1}}
  plot(seq(0,1,by=0.01),sapply(seq(0,1,by=0.01),posteriorDist),type=’l’,
    xlab=’’,ylab=’posterior’)
}

theta0 <- 0.5
alpha <- 1
beta <- 1
y <- rbinom(1000,1,theta0)

plotPosterior(alpha,beta,50)
plotPosterior(alpha,beta,200)
plotPosterior(alpha,beta,1000)
```

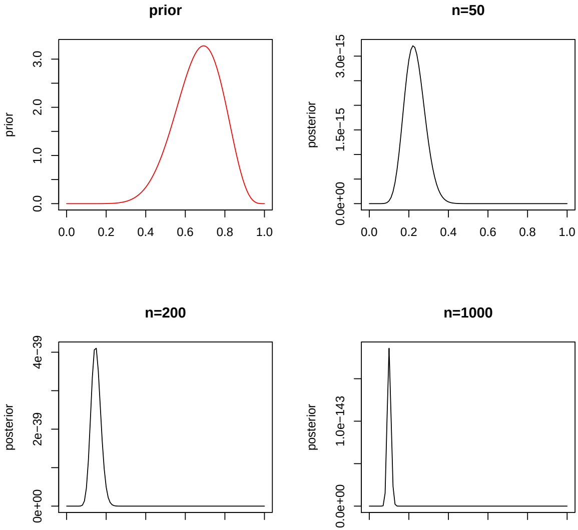
prior Beta(1,1) and $\theta_0 = 0.5$



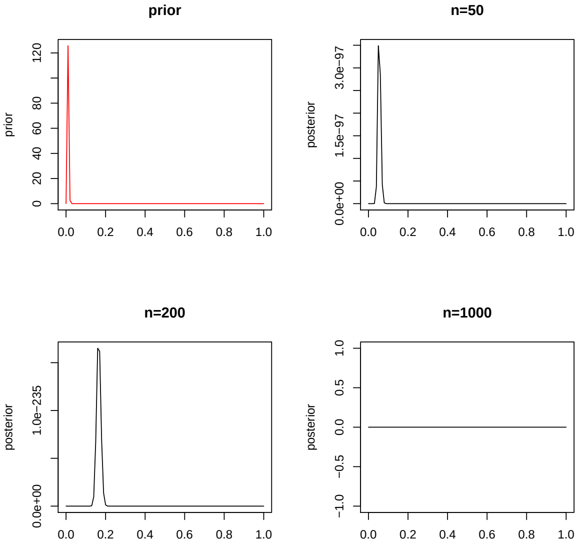
prior Beta(0.5,0.5) and $\theta_0 = 0.25$



prior Beta(10,5) and $\theta_0 = 0.1$



prior Beta(10,1000) and $\theta_0 = 0.95$



Contents

1. estimation methods

1.1 method of moments

1.2 maximum likelihood

1.3 Bayesian approach

2. evaluating estimators

2.1 Cramér-Rao inequality

2.2 attainment

2.3 Rao-Blackwell

3. exercises

basic criteria

- why to gauge performance? there are many ways to find reasonable point estimators for a given parameter and each method may lead to a different estimator

- definition: the mean squared error (MSE) of an estimator T of a parameter θ is the average squared difference between the estimator and the parameter: $\mathbb{E}_\theta(T-\theta)^2$
  - mostly any function of $|T-\theta|$ measures the goodness-of-fit of the estimator T
  - we will focus on mean squared error and unbiasedness

- advantages: analytically tractable, differentiable, easily interpretable

$$\begin{aligned}\mathbb{E}_\theta(T-\theta)^2 &= \mathbb{E}_\theta(T-\mathbb{E}_\theta(T)+\mathbb{E}_\theta(T)-\theta)^2 \\ &= (\mathbb{E}_\theta(T)-\theta)^2+\mathrm{var}_\theta(T) \;=\; \underbrace{\mathrm{bias}_\theta^2(T)}_{\text{accuracy}}+\underbrace{\mathrm{var}_\theta(T)}_{\text{precision}}\end{aligned}$$

MSE in a Gaussian model

- example: let $X_1, \ldots, X_n \sim \text{iid}N(\mu, \sigma^2)$. The sample mean and variance are both unbiased estimators of their population counterparts given that $\mathbb{E}_{\mu,\sigma^2}(\bar{X}_n) = \mu$ and $\mathbb{E}_{\mu,\sigma^2}(S_n^2) = \sigma^2$ for all μ and σ^2.

- Since the estimators are unbiased, MSE is equal to the variance, and

$$
\begin{aligned}
\mathbb{E}_{\mu,\sigma^2}(\bar{X}_n - \mu)^2 &= \text{var}(\bar{X}_n) \\
&\quad \bar{X}_n \sim N(\mu, \sigma^2/n) \Rightarrow \text{var}(\bar{X}_n) = \frac{\sigma^2}{n} \\
\mathbb{E}_{\mu,\sigma^2}(S_n^2 - \sigma^2)^2 &= \text{var}(S_n^2) \\
&\quad (n-1)S_n^2/\sigma^2 \sim \chi^2_{n-1} \Rightarrow \text{var}(S_n^2) = \frac{\sigma^4}{(n-1)^2} 2(n-1) = 2\frac{\sigma^4}{n-1}
\end{aligned}
$$

  given that the variance of χ^2_k is $2k$.

- Can we find any improvement?

MSE in a Gaussian model

- example (cont'd): what about the MLE estimator for the variance?

$$\begin{aligned} \mathbb{E}_{\mu,\sigma^2}(\hat{\sigma}^2) &= \mathbb{E}_{\mu,\sigma^2}\left(\frac{n-1}{n}\,S_n^2\right) &= \frac{n-1}{n}\sigma^2 \\ \mathrm{var}_{\mu,\sigma^2}(\hat{\sigma}^2) &= \mathrm{var}_{\mu,\sigma^2}\left(\frac{n-1}{n}\,S_n^2\right) &= \frac{2(n-1)}{n^2}\,\sigma^4 \end{aligned}$$

hence

$$\begin{aligned} \mathbb{E}(\hat{\sigma}^2-\sigma^2)^2 &= \left[\frac{n-1}{n}\sigma^2-\sigma^2\right]^2+\frac{2(n-1)}{n^2}\,\sigma^4 \\ &= \sigma^4\left[\frac{1}{n^2}+\frac{2(n-1)}{n^2}\right] \quad = \quad \sigma^4\left[\frac{2n-1}{n^2}\right] \end{aligned}$$

and note that $\frac{2n-1}{n^2} < \frac{2n}{n^2} = \frac{2}{n} < \frac{2}{n-1}$.

- since $\frac{n-1}{n} < 1$, the MLE mean squared error is smaller than that of pure unbiased estimators.

MSE drawbacks

- tradeoff: imposing unbiasedness does not suffice to control the MSE because sometimes it pays off to include some bias to achieve a larger decrease in the variance.

- drawbacks:

  (1) one should not necessarily abandon the sample variance given that the MLE will on average underestimate the true variance

  (2) in addition, MSE is a somewhat weird criterion for a scale parameter given that it penalizes equally underestimation and overestimation

Contents

1. estimation methods

1.1 method of moments

1.2 maximum likelihood

1.3 Bayesian approach

2. evaluating estimators

2.1 Cramér-Rao inequality

2.2 attainment

2.3 Rao-Blackwell

3. exercises

best unbiased estimators

- there is no best estimator in the MSE sense because the class of all estimators is way too vast, e.g., $\hat{\theta}(\boldsymbol{X}) = 1$ is definitely the best ever estimator for $\theta = 1$, but terrible otherwise.

- **how about restricting the class of estimators?**
  - suppose there is an estimator T^* with $\mathbb{E}(T^*) = \tau(\theta) \neq \theta$. Define the class of all estimators with equal bias at every θ,

$$C_\tau = \{T : \mathbb{E}_\theta(T) = \tau(\theta)\}$$

  - then for any $T_1, T_2 \in C_\tau$, $\text{bias}_\theta T_1 = \text{bias}_\theta T_2$, so comparison in MSE sense between T_1 and T_2 can be based on variance alone.
  - if T is unbiased, $\text{MSE}(T, \theta) = \text{var}_\theta(T)$

- definition: an estimator T^* is a best unbiased estimator of $\tau(\theta)$ if $\mathbb{E}_\theta(T^*) = \tau(\theta)$ for every θ and, for any other unbiased estimator T of $\tau(\theta)$,

$$\text{var}_\theta(T^*) \quad \leq \quad \text{var}_\theta(T)$$

for all θ. T^* is also called the uniformly minimum variance unbiased estimator of θ (UMVUE)

Cramér-Rao inequality

- finding the UMVUE could often be a daunting and endless task.
- it is very useful to have a lower bound on the variance of any estimator:
  - if an estimator reaches that bound, should be UMVUE. Search ends! ☺
  - but not reaching does not mean it is not UMVUE
  - we will also have results on when the bound is attained
- this lower bound exists and is known as Cramér-Rao lower bound.
- first, an example...

Poisson unbiased estimation

- example: let $X_1, \dots, X_n \sim$ iid Poisson(λ). For all λ, $\mathbb{E}_\lambda(X) = \text{var}_\lambda(X) = \lambda$

  $$\begin{aligned} \mathbb{E}_\lambda(\bar{X}_n) &= \lambda \\ \mathbb{E}_\lambda(S_n^2) &= \lambda \end{aligned}$$

  to compare them, we should look at their variances

  $$\text{var}_\lambda(\bar{X}_n) = \lambda/n \leq \text{var}_\lambda(S_n^2) \qquad \text{for all } \lambda$$

  after some lengthy calculations.

- consider now the unbiased estimator $T_a = a\bar{X}_n + (1-a)S_n^2$, is there some a for which $\text{var}_\lambda(T_a) \leq \text{var}_\lambda(\bar{X}_n)$ for every λ?

- is there any other better unbiased estimator lurking about?

Cramér-Rao inequality: univariate case

- theorem (**Cramér-Rao lower bound**): let $(X_1, \ldots, X_n)$ denote a sample with joint pdf $f(\boldsymbol{x}|\theta)$ and let $T(\boldsymbol{X})$ be any estimator satisfying

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta\big[T(\boldsymbol{X})\big] \;=\; \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}\big[T(\boldsymbol{x})f(\boldsymbol{x}|\theta)\big]\,\mathrm{d}\boldsymbol{x} \;\text{ and }\; \mathrm{var}_\theta\big[T(\boldsymbol{X})\big] < \infty,$$

  it then follows that

$$\mathrm{var}_\theta\big[T(\boldsymbol{X})\big] \;\geq\; \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta\big[T(\boldsymbol{X})\big]\right)^2}{\mathbb{E}_\theta\left[\frac{\partial \ln f(\boldsymbol{X}|\theta)}{\partial\theta}\right]^2}$$

Cramér-Rao inequality

$$\mathrm{var}_\theta\left[T(\boldsymbol{X})\right] \;\geq\; \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\,\mathbb{E}_\theta\left[T(\boldsymbol{X})\right]\right)^2}{\mathbb{E}_\theta\left[\frac{\partial \ln f(\boldsymbol{X}|\theta)}{\partial\theta}\right]^2}$$

- some comments:
  - if $\mathbb{E}_\theta(T(\boldsymbol{X})) = \theta$, $\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\,\mathbb{E}_\theta\left[T(\boldsymbol{X})\right]\right)^2 = 1$
  - $\mathbb{E}_\theta\left[\frac{\partial \ln f(\boldsymbol{X}|\theta)}{\partial\theta}\right]^2$ is known as the Fisher information of the sample for it entails a bound on the variance of the best unbiased estimator of θ
  - if we assume that $(X_1, \ldots, X_n)$ is iid with $f(x|\theta)$, then the lower-bound denominator reduces to (this is proposition 1 below)

    $$n \cdot \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\ \ln f(X_i|\theta)\right]^2$$

    as the information number grows, the more info we have on θ and so the lower the bound on the variance of the UMVU estimator.

Cramér-Rao inequality

$$\mathrm{var}_\theta\left[T(\boldsymbol{X})\right] \quad \geq \quad \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\,\mathbb{E}_\theta\left[T(\boldsymbol{X})\right]\right)^2}{\mathbb{E}_\theta\left[\frac{\partial \ln f(\boldsymbol{X}|\theta)}{\partial\theta}\right]^2}$$

- about Fischer information number:
  - we will also see that, if we can exchange integral and derivatives,

    $$\mathbb{E}_\theta\left[\frac{\partial \ln f(\boldsymbol{X}|\theta)}{\partial\theta}\right]^2 \quad = \quad -\mathbb{E}_\theta\left[\frac{\partial^2 \ln f(\boldsymbol{X}|\theta)}{\partial\theta^2}\right]$$

    (proposition 2 below)
  - so, if $T(\boldsymbol{X})$ is unbiased, and $(X_1, \ldots, X_n)$ is i.i.d. sample with $f(x|\theta)$, the Cramér-Rao inequality reads

    $$\mathrm{var}_\theta\left[T(\boldsymbol{X})\right] \quad \geq \quad -\frac{1}{n\cdot\mathbb{E}_\theta\left[\frac{\partial^2 \ln f(x_i|\theta)}{\partial\theta^2}\right]}$$

Cramér-Rao inequality

- lemma: the expected value of the derivative of the likelihood evaluated at the true parameter is zero, $\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\ln\ell(\theta|X)\right]=0$.

- proof:

$$\begin{aligned}
\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\ln\ell(X,\theta)\right] &= \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}\ln\ell(X,\theta)f(x|\theta)dx \\
&= \int_{-\infty}^{\infty}\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}f(x|\theta)dx \;=\; \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(x|\theta)dx \\
&= \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}f(x|\theta)dx \;=\; 0
\end{aligned}$$

- important remark: this is true <u>only</u> if the derivative of the likelihood is evaluated at the true value, or else we would have

$$\int_{-\infty}^{\infty}\frac{\frac{\partial}{\partial\theta}f(x|\theta^*)}{f(x|\theta^*)}f(x|\theta)dx \;\neq\; \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(x|\theta)dx$$

Cramér-Rao inequality

- proof (i): by the Cauchy-Schwartz inequality,

$$[\mathrm{cov}(X, Y)]^2 \leq (\mathrm{var}X)(\mathrm{var}Y) \implies \mathrm{var}X \geq \frac{[\mathrm{cov}(X, Y)]^2}{\mathrm{var}Y}$$

  we choose X as $T(\boldsymbol{X})$ and Y as $\frac{\partial}{\partial\theta} \ln f(\boldsymbol{X}|\theta)$

- proof (ii): given that $\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta} \ln f(\boldsymbol{X}|\theta)\right] = 0$, then

$$\mathrm{var}_\theta\left[\frac{\partial \ln f(\boldsymbol{X}|\theta)}{\partial\theta}\right] = \mathbb{E}_\theta\left[\frac{\partial \ln f(\boldsymbol{X}|\theta)}{\partial\theta}\right]^2$$

  so we only need to solve $\mathrm{cov}(X, Y)$.

Cramér-Rao inequality

- proof (iii): note that

$$
\begin{aligned}
\frac{d}{d\theta}\mathbb{E}_\theta T(\boldsymbol{X}) &= \frac{d}{d\theta}\int T(\boldsymbol{x})f(\boldsymbol{x}|\theta)dx \;=\; \int T(\boldsymbol{x})\left[\frac{\partial}{\partial\theta}f(\boldsymbol{x}|\theta)\right]dx \\
&= \int T(\boldsymbol{x})\left[\frac{\partial}{\partial\theta}f(\boldsymbol{x}|\theta)\right]\frac{f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x}|\theta)}dx \\
&= \mathbb{E}_\theta\left[T(\boldsymbol{X})\frac{\frac{\partial}{\partial\theta}f(\boldsymbol{X}|\theta)}{f(\boldsymbol{X}|\theta)}\right] \\
&= \mathbb{E}_\theta\left[T(\boldsymbol{X})\frac{\partial}{\partial\theta}\ln\ell(\boldsymbol{X}|\theta)\right]
\end{aligned}
$$

which is the cov between $T(\boldsymbol{X})$ and $\frac{\partial}{\partial\theta}\ln f(\boldsymbol{X}|\theta)$, since $\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\ln f(\boldsymbol{X}|\theta)\right] = 0$ ■

Cramér-Rao inequality

- proposition 1: if $X_1, \dots, X_n$ are independent and identically distributed,

$$\mathbb{E}_\theta \left[\frac{\partial \ln f(\boldsymbol{X}|\theta)}{\partial \theta} \right]^2 = n\, \mathbb{E}_\theta \left[\frac{\partial \ln f(X_i|\theta)}{\partial \theta} \right]^2$$

- proof: by independence, $\ln f(\boldsymbol{X}|\theta) = \ln \prod_{i=1}^n f(X_i|\theta) = \sum_{i=1}^n \ln f(X_i|\theta)$ so

$$\begin{aligned}
\mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} \ln f(\boldsymbol{X}|\theta) \right]^2 &= \mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} \sum_{i=1}^n \ln f(X_i|\theta) \right]^2 = \mathbb{E}_\theta \left[\sum_{i=1}^n \frac{\partial}{\partial \theta} \ln f(X_i|\theta) \right]^2 \\
&= \sum_{i=1}^n \mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} \ln f(X_i|\theta) \right]^2 \\
&\quad \underbrace{+ \sum_{i=1}^j \sum_{i \neq j} \mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} \ln f(X_i|\theta) \frac{\partial}{\partial \theta} \ln f(X_j|\theta) \right]}_{=0 \text{ by independence}}
\end{aligned}$$

■

Cramér-Rao inequality

- proposition 2:

$$\mathbb{E}_\theta\left[\frac{\partial \ln f(\boldsymbol{X}|\theta)}{\partial\theta}\right]^2 = -\mathbb{E}_\theta\left[\frac{\partial^2 \ln f(\boldsymbol{X}|\theta)}{\partial\theta^2}\right]$$

- proof: note that $\frac{\partial}{\partial\theta}\ln \ell(x,\theta) = \frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}$, so

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\ln \ell(x,\theta) &= \frac{\left[\frac{\partial^2}{\partial\theta^2}f(x|\theta)\right]f(x|\theta) - \left[\frac{\partial}{\partial\theta}f(x|\theta)\right]^2}{[f(x|\theta)]^2} \\
&= \frac{\frac{\partial^2}{\partial\theta^2}f(x|\theta)}{f(x|\theta)} - \left[\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right]^2 = \frac{\frac{\partial^2}{\partial\theta^2}f(x|\theta)}{f(x|\theta)} - \left[\frac{\partial}{\partial\theta}\ln \ell(x|\theta)\right]^2 \\
\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln \ell(x,\theta)\right] &= \mathbb{E}_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2}f(x|\theta)}{f(x|\theta)}\right] - \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\ln \ell(x|\theta)\right]^2
\end{aligned}$$

and $\mathbb{E}_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2}f(x|\theta)}{f(x|\theta)}\right] = \int \frac{\partial^2}{\partial\theta^2}f(x|\theta)dx = \frac{\partial^2}{\partial\theta^2}\int f(x|\theta)dx = 0.$ ■

some notation

- let's put some notation down
  - the $k \times 1$ score vector for the full sample $s(\boldsymbol{x}, \theta) = \frac{\partial}{\partial \theta} \ln \ell(\boldsymbol{x}, \theta)$
  - the $k \times 1$ score vector for one observation $s(x_i, \theta) = \frac{\partial}{\partial \theta} \ln \ell(x_i, \theta)$
  - the $k \times k$ Fisher information matrix $\mathcal{I}(\theta) = \mathbb{E}_\theta \left(s(\boldsymbol{x}, \theta) s(\boldsymbol{x}, \theta)' \right)$
  - the $k \times k$ Hessian matrix $\mathcal{H}(\boldsymbol{x}, \theta) = \frac{\partial^2}{\partial \theta \partial \theta'} \ln \ell(\boldsymbol{x}, \theta)$
  - the $k \times k$ expected Hessian $\mathcal{H}(\theta) = \mathbb{E}_\theta \left[\mathcal{H}(\boldsymbol{x}, \theta) \right]$
- in the results that follow, we are going to assume that one can swap integrals and derivatives.

Cramér-Rao inequality: multivariate case

- theorem (**Cramér-Rao lower bound**): let $T(\boldsymbol{X})$ be an unbiased estimator of θ. Then

$$\text{var}_\theta\left[T(\boldsymbol{X})\right] \quad \geq \quad \left[\mathcal{I}(\theta)\right]^{-1}$$

  and, given the results obtained before, one can obtain

  - $\mathcal{I}(\theta)$ by definition as the expected outer product of the scores, or
  - calculating the expected Hessian and using the identity $\mathcal{I}(\theta) = -\mathcal{H}(\theta)$

Cramér-Rao inequality

- proof (i): we first show that

$$\mathbb{E}_\theta\left(T(\boldsymbol{X})s(\boldsymbol{X},\theta)'\right) \quad = \quad I$$

where I is the identity matrix since

$$\begin{aligned}\mathbb{E}_\theta\left(T(\boldsymbol{X})s(\boldsymbol{X},\theta)'\right) &= \int_{-\infty}^{\infty} T(\boldsymbol{X})\frac{\frac{\partial}{\partial\theta'}f(x|\theta)}{f(x|\theta)}f(x|\theta)dx \\ &= \frac{\partial}{\partial\theta'}\int_{-\infty}^{\infty} T(\boldsymbol{X})f(x|\theta)dx \\ &= \frac{\partial}{\partial\theta'}\mathbb{E}_\theta(T(\boldsymbol{X})) \quad = \quad I\end{aligned}$$

- proof (ii): to get the Cramér-Rao lower bound, note that

$$\text{var}_\theta\begin{pmatrix} T(\boldsymbol{X}) \\ s(\boldsymbol{X},\theta)\end{pmatrix} \quad = \quad \begin{pmatrix}\text{var}_\theta T(\boldsymbol{X}) & I \\ I & \mathcal{I}(\theta)\end{pmatrix}$$

Cramér-Rao inequality

- proof (cont'd): to get the result, note any variance-covariance matrix is positive semi-definite. Therefore,

$$\mathrm{var}_\theta\, T(\boldsymbol{X}) \cdot \mathcal{I}(\theta) - I \quad \geq \quad 0$$

  which implies

$$\mathrm{var}_\theta\, T(\boldsymbol{X}) \quad \geq \quad [\mathcal{I}(\theta)]^{-1}$$

  establishing the final result. ■

Cramér-Rao inequality

- proposition 2, multivariate case: $\mathcal{I}(\theta) = -\mathcal{H}(\theta)$

- proof:

$$\frac{\partial^2}{\partial\theta\partial\theta'} \ln \ell(\boldsymbol{x}, \theta) = \frac{\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta'} \ln f(\boldsymbol{x}|\theta)\right) = \frac{\frac{\partial^2}{\partial\theta\partial\theta'} f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x}|\theta)} - \left[\frac{\partial}{\partial\theta} \ln f(\boldsymbol{x}|\theta)\right]\left[\frac{\partial}{\partial\theta'} \ln f(\boldsymbol{x}|\theta)\right]$$

applying to product rule. So the expected Hessian

$$\begin{aligned} \mathcal{H}(\theta) &= \mathbb{E}_\theta\left[\mathcal{H}(\boldsymbol{X}, \theta)\right] = \int_{-\infty}^{\infty} \frac{\partial^2}{\partial\theta\partial\theta'} \ln \ell(\boldsymbol{x}, \theta) f(\boldsymbol{x}|\theta) d\boldsymbol{x} \\ &= \int_{-\infty}^{\infty} \frac{\partial^2}{\partial\theta\partial\theta'} f(\boldsymbol{x}|\theta) d\boldsymbol{x} - \mathcal{I}(\theta) \\ &= \frac{\partial^2}{\partial\theta\partial\theta'} \int_{-\infty}^{\infty} f(\boldsymbol{x}|\theta) d\boldsymbol{x} - \mathcal{I}(\theta) \\ &= -\mathcal{I}(\theta) \end{aligned}$$

completing the proof. ■

example 1

- let $X_1, \dots, X_n$ be i.i.d. Poisson (λ). Then

$$
\begin{aligned}
f(x_i|\lambda) &= \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \\
\ln \ell(x_i|\lambda) &= -\lambda + x_i \ln \lambda - \ln x_i! \\
s(x_i, \lambda) &= -1 + \frac{x_i}{\lambda} \\
\mathcal{I}(\theta) &= n\,\mathbb{E}_\theta \left[\left(-1 + \frac{x_i}{\lambda} \right)^2 \right] \;=\; n\,\mathbb{E}_\theta \left(1 - 2\frac{x_i}{\lambda} + \frac{x_i^2}{\lambda^2} \right) \\
&= n \left(1 - 2\frac{\lambda}{\lambda} + \frac{\lambda + \lambda^2}{\lambda^2} \right) \;=\; n \left(1 - 2\frac{\lambda}{\lambda} + \frac{\lambda + \lambda^2}{\lambda^2} \right) \;=\; \frac{n}{\lambda}
\end{aligned}
$$

- so the unbiased estimator $T(\boldsymbol{X}) = \bar{X}$ achieves the Cramér-Rao lower bound $\frac{\lambda}{n}$ and therefore it is the UMVU estimator. ■

example 2

- let $X_1, \dots, X_n$ be i.i.d. from exponential(λ), so

$$\begin{aligned} f(x_i|\lambda) &= \lambda e^{-\lambda x_i} \\ \ln \ell(x_i|\lambda) &= \ln \lambda - \lambda x_i \\ s(x_i, \lambda) &= \frac{1}{\lambda} - x_i \\ \mathcal{H}(\lambda) &= \mathbb{E}_\theta \mathcal{H}_i(x_i|\lambda) = -\frac{n}{\lambda^2} \end{aligned}$$

  so the Cramér-Rao inequality states that, among the unbiased estimators, the lower bound on the variance is $\frac{\lambda^2}{n}$.

- Given that $\mathbb{E}(X^k) = \frac{k!}{\lambda^k}$

$$\hat{\lambda}_{MM,k} = \left[\frac{k!}{\hat{\mu}_k}\right]^{\frac{1}{k}}$$

  is as method-of-moment estimator, where $\hat{\mu}_k$ is the k-th sample moment. We also consider the maximum likelihood estimator.

example 2 - R codes

```
 CRsamplerExponential <- function(lambda,n){
   logLikExp <- function(theta){
      n <- length(randomVec)
      logLik <- n*log(theta)-theta*sum(randomVec)
      logLik <- -1*logLik
   }
   hats <- matrix(0,5000,6)
   for (i in 1:5000){
      randomVec <- rexp(n,lambda)
      mu1 <- mean(randomVec)
      mu2 <- mean(randomVec^2)
      mu3 <- mean(randomVec^3)
      mu4 <- mean(randomVec^4)
      mu5 <- mean(randomVec^5)
      hats[i,1] <- 1/mu1
      hats[i,2] <- (factorial(2)/mu2)^(1/2)
      hats[i,3] <- (factorial(3)/mu3)^(1/3)
      hats[i,4] <- (factorial(4)/mu4)^(1/4)
      hats[i,5] <- (factorial(5)/mu5)^(1/5)
      thetahat <- optim(lambda,logLikExp)
      hats[i,6] <- thetahat$par[1]
   }
   hats
}
```

example 2

Table: Cramér-Rao lower bound simulations

	$\lambda = 2,\ n = 1000$			$\lambda = 20,\ n = 1000$		
	λ^2/n	$\text{var}(\hat{\lambda})$	$\frac{\text{var}(\hat{\lambda})}{\lambda^2/n}$	λ^2/n	$\text{var}(\hat{\lambda})$	$\frac{\text{var}(\hat{\lambda})}{\lambda^2/n}$
$\hat{\lambda}_{MM,1}$	0.004	0.004	1.026	0.400	0.415	1.037
$\hat{\lambda}_{MM,2}$	0.004	0.005	1.261	0.400	0.504	1.259
$\hat{\lambda}_{MM,3}$	0.004	0.008	2.026	0.400	0.814	2.036
$\hat{\lambda}_{MM,4}$	0.004	0.014	3.532	0.400	1.432	3.580
$\hat{\lambda}_{MM,5}$	0.004	0.023	5.771	0.400	2.348	5.872
$\hat{\lambda}_{MLE}$	0.004	0.004	1.027	0.400	0.415	1.037

example 2

- there is one catch here: $\hat{\lambda}_{MM,1} = \frac{1}{\hat{\mu}_1}$ so

$$\mathbb{E}(\hat{\lambda}_{MM,1}) \;=\; \mathbb{E}\left(\frac{1}{\hat{\mu}_1}\right) \overset{\text{Jensen}}{\neq} \frac{1}{\mathbb{E}(\hat{\mu}_1)} \;=\; \lambda$$

  i.e., the estimator is biased.

- we could instead estimate $\varphi = \frac{1}{\lambda}$.
  - the method-of-moments estimators are $\hat{\varphi}_{MM,k} = \hat{\lambda}_{MM,k}$.
  - by the invariance principle, the maximum likelihood estimator is $\hat{\varphi}_{MLE} = \hat{\lambda}_{MLE}$.
  - the Cramér-Rao lower bound is $\frac{1}{\varphi^2 n}$.

example 2

Table: Cramér-Rao lower bound simulations

	$\varphi = 1/2,\ n = 1000$			$\varphi = 1/20,\ n = 1000$		
	λ^2/n	$\text{var}(\hat{\lambda})$	$\frac{\text{var}(\hat{\lambda})}{\lambda^2/n}$	λ^2/n	$\text{var}(\hat{\lambda})$	$\frac{\text{var}(\hat{\lambda})}{\lambda^2/n}$
$\hat{\lambda}_{MM,1}$	$2.5 \cdot e^{-5}$	$2.55 \cdot e^{-5}$	1.019	$2.5 \cdot e^{-6}$	$2.55 \cdot e^{-6}$	1.020
$\hat{\lambda}_{MM,2}$	$2.5 \cdot e^{-5}$	$3.15 \cdot e^{-5}$	1.253	$2.5 \cdot e^{-6}$	$3.09 \cdot e^{-6}$	1.239
$\hat{\lambda}_{MM,3}$	$2.5 \cdot e^{-5}$	$5.06 \cdot e^{-5}$	2.022	$2.5 \cdot e^{-6}$	$4.95 \cdot e^{-6}$	1.981
$\hat{\lambda}_{MM,4}$	$2.5 \cdot e^{-5}$	$9.01 \cdot e^{-5}$	3.604	$2.5 \cdot e^{-6}$	$8.64 \cdot e^{-6}$	3.456
$\hat{\lambda}_{MM,5}$	$2.5 \cdot e^{-5}$	$14.92 \cdot e^{-5}$	5.971	$2.5 \cdot e^{-6}$	$14.10 \cdot e^{-6}$	5.640
$\hat{\lambda}_{MLE}$	$2.5 \cdot e^{-5}$	$2.54 \cdot e^{-5}$	1.019	$2.5 \cdot e^{-6}$	$2.54 \cdot e^{-6}$	1.019

example 3

- let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. Then

$$\begin{aligned}
\ln \ell(x_i, \mu, \sigma^2) &= -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \sigma^2 - \frac{(x_i-\mu)^2}{2\sigma^2} \\
s(x_i, \mu, \sigma^2) &= \begin{pmatrix} \frac{x_i-\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} + \frac{(x_i-\mu)^2}{2\sigma^4} \end{pmatrix} \\
\mathcal{H}(x_i, \mu, \sigma^2) &= n \begin{pmatrix} -\frac{1}{\sigma^2} & -\frac{x_i-\mu}{\sigma^4} \\ -\frac{x_i-\mu}{\sigma^4} & \frac{1}{2\sigma^4} - \frac{(x_i-\mu)^2}{\sigma^6} \end{pmatrix} \\
\mathcal{H}(\mu, \sigma^2) &= \mathbb{E}_\theta\left[\mathcal{H}(x_i, \mu, \sigma^2)\right] = n \begin{pmatrix} -\frac{1}{\sigma^2} & 0 \\ 0 & -\frac{1}{2\sigma^4} \end{pmatrix}
\end{aligned}$$

therefore

$$\mathcal{I}(\theta) = -\mathcal{H}(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

so $\bar{X}$ achieves the Cramér-Rao lower bound, but the variance of the estimator S^2 for σ^2 is $2\sigma^4/(n-1)$ is strictly greater than the lower bound. Two important questions: is the lower bound with respect to σ^2 unattainable? How to find the best unbiased estimator of σ^2?

Contents

1. estimation methods

1.1 method of moments

1.2 maximum likelihood

1.3 Bayesian approach

2. evaluating estimators

2.1 Cramér-Rao inequality

2.2 attainment

2.3 Rao-Blackwell

3. exercises

attainment of Cramér-Rao lower bound

- bound follows from an application of the Cauchy-Schwarz inequality
  - conditions for attainment of the bound are the same for equality in the Cauchy-Schwarz inequality!

- theorem (attainment) (CB 7.3.15): let $X_1, \ldots, X_n \sim$ iid $f(x|\theta)$, with $f(x|\theta)$ satisfying the conditions for the CR inequality, then an unbiased estimator $T(\boldsymbol{X})$ of $\tau(\theta)$ attains the CR lower bound if and only if

$$a(\theta)\big[T(\boldsymbol{X}) - \tau(\theta)\big] \;=\; \frac{\partial \ln \prod_{i=1}^{n} f(X_i|\theta)}{\partial \theta}$$

  for some function $a(\theta)$ that depends exclusively on θ.

attainment of Cramér-Rao lower bound

- proof: the Cramér-Rao bound is equivalently written

$$
\begin{aligned}
\left[\operatorname{cov}_\theta\left(T(\boldsymbol{X}), \frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n} f(X_i|\theta)\right)\right]^2 &\leq [\operatorname{var}_\theta T(\boldsymbol{X})]\cdot\left[\operatorname{var}_\theta \frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n} f(X_i|\theta)\right] \\
&\Downarrow \\
\left[\operatorname{corr}_\theta\left(T(\boldsymbol{X}), \frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n} f(X_i|\theta)\right)\right]^2 &\leq 1 \\
&\Downarrow \\
\left|\operatorname{corr}_\theta\left(T(\boldsymbol{X}), \frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n} f(X_i|\theta)\right)\right| &\leq 1
\end{aligned}
$$

so holds with equality if, and only if, $T(\boldsymbol{X})$ is an affine function of $\frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n} f(X_i|\theta)$, that is, there are a and b such that

$$
T(\boldsymbol{X}) = a\cdot\frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n} f(X_i|\theta) + b
$$

attainment of Cramér-Rao lower bound

- proof (cont'd): taking expectations on both sides,

$$\tau(\theta) \quad = \quad b$$

since $\mathbb{E}_\theta\left[T(\boldsymbol{X})\right] = \tau(\theta)$ and $\mathbb{E}_\theta\left[\ln\prod_{i=1}^n f(X_i|\theta)\right] = 0$. It follows that

$$T(\boldsymbol{X}) - \tau(\theta) \quad = \quad a \cdot \frac{\partial}{\partial\theta}\ln\prod_{i=1}^n f(X_i|\theta)$$

that is, $T(\boldsymbol{X}) - \tau(\theta)$ is proportional to $\frac{\partial}{\partial\theta}\ln\prod_{i=1}^n f(X_i|\theta)$. ■

attainment of Cramér-Rao lower bound

- example: in the Gaussian case, we have

$$\begin{aligned}
\ell(\mu,\sigma^2|\boldsymbol{x}) &= \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \\
\ln\ell(\mu,\sigma^2|\boldsymbol{x}) &= -\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \\
\frac{\partial}{\partial\sigma^2}\ln\ell(\mu,\sigma^2|\boldsymbol{x}) &= \frac{n}{2\sigma^4}\left(\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n}-\sigma^2\right)
\end{aligned}$$

choosing $a(\theta)=\frac{n}{2\sigma^4}$, the best unbiased estimator is $\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$. It is calculable only if μ is known! If μ is not known, the Cramér-Rao lower bound cannot be achieved. ■

Contents

1. estimation methods

1.1 method of moments

1.2 maximum likelihood

1.3 Bayesian approach

2. evaluating estimators

2.1 Cramér-Rao inequality

2.2 attainment

2.3 Rao-Blackwell

3. exercises

Rao-Blackwell theorem

- what if allowable estimators do not attain the lower bound? sufficient statistics allow us to refine the search for efficient estimators

- theorem (**Rao-Blackwell**) let T denote any unbiased estimator of $\tau(\theta)$ and let S be a sufficient statistic for θ, then $\phi(S) = \mathbb{E}(T|S)$ is such that $\mathbb{E}_\theta\left[\phi(S)\right] = \tau(\theta)$ and $\text{var}_\theta \phi(S) \leq \text{var}_\theta(T)$ for all θ

- proof: $\phi(S)$ indeed is an estimator because it does not depend on θ due to the fact that S is sufficient (in other words, sufficiency ensures it remains a statistic depending exclusively on the sample and not on the parameter), whereas

$$\begin{aligned} \tau(\theta) &= \mathbb{E}_\theta(T) = \mathbb{E}_\theta\left[\mathbb{E}(T|S)\right] = \mathbb{E}_\theta\left[\phi(S)\right] \\ \text{var}_\theta(T) &= \text{var}_\theta\left[\mathbb{E}(T|S)\right] + \mathbb{E}_\theta\left[\text{var}(T|S)\right] \\ &= \text{var}_\theta\left[\phi(S)\right] + \mathbb{E}_\theta\left[\text{var}(T|S)\right] \\ &\geq \text{var}_\theta\left[\phi(S)\right] \end{aligned}$$

that is, conditioning improves the estimator. ■

conditioning on insufficient statistics

- it is generally true that conditioning improves the variance
  - if not conditioned on a sufficient statistic, the result might depend on the true parameter
  - so it is not an estimator!
- example: if $X_1, X_2 \sim \text{iid} N(\mu, 1)$, then $\bar{X}_2 \sim N(\mu, 1/2)$. Conditioning on the insufficient statistic X_1 then yields

$$\phi(X_1) \quad = \quad \mathbb{E}_\mu(\bar{X}_2|X_1) \quad = \quad \frac{\mathbb{E}_\mu(X_1|X_1) + \mathbb{E}_\mu(X_2|X_1)}{2} \quad = \quad \frac{X_1 + \mu}{2}$$

and so $\phi(X_1)$ is not an estimator as it depends on μ. ■

further results

- we answered the question of when the lower bound is attainable
- there is a large literature on finding the UMVU estimator, which we will not cover
- in particular, the Lehmann-Scheffe lemma shows that conditioning on a complete statistic, the Rao-Blackwell device yields the UMVU estimator.

Contents

1. estimation methods

1.1 method of moments

1.2 maximum likelihood

1.3 Bayesian approach

2. evaluating estimators

2.1 Cramér-Rao inequality

2.2 attainment

2.3 Rao-Blackwell

3. exercises

Reference:

- Casella and Berger, Ch. 7
- Greenberg, *Introduction to Bayesian Econometrics*, Ch. 2

Exercises:

- 7.1, 7.2(a), 7.4, 7.7–7.12, 7.14, 7.17–7.21, 7.37, 7.38(a), 7.40–7.42, 7.44, 7.46 (a,b,d), 7.49(a,b)