Lecture 4: Interval Estimation & Goodness of Estimation

STA261 – Probability & Statistics II

Ofir Harari

Department of Statistical Sciences

University of Toronto

Outline

point estimation does not reveal uncertainty

Confidence Intervals

- The last couple of lectures dealt with *point estimation*: finding an estimator $\widehat{\theta}$ with good properties (e.g. consistency) that will hopefully land "in the ballpark" of θ.

- But we will inevitably err –

$$\mathbb{P}(\hat{\theta} = \theta) = 0 \text{ (for continuous data)}$$

  – and then what...?

because MLE estimator has normal distribution (continuous)

- We have learned about the notion of standard error (SE) of an estimator –
  - Could report the point estimate along with its SE – a good start
  - Is that what the "margin of error: ±4 percentage points" in the newspapers is all about ?
  - Somewhat misleading if the sampling distribution of the estimator is asymmetrical

Confidence Intervals (cont.)

- The idea of confidence intervals is to provide a range of plausible values for θ, rather then a single number.

Definition

Let $X_1, \dots, X_n \sim f_\theta$. A $100(1-\alpha)\%$ *confidence interval* for θ is a pair of statistics $L = L(X_1, \dots, X_n)$ and $U = U(X_1, \dots, X_n)$ such that

$$\mathbb{P}(L \leq \theta \leq U) = 1 - \alpha.$$

We call $100(1-\alpha)\%$ the *confidence level*.

note theta is fixed, L and U are random

Example: Normal mean with known variance

Example

1. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where σ^2 is assumed to be known. Find a $100(1-\alpha)\%$ confidence interval for μ.

2. Assuming $\sigma = 5$, find a 95% confidence interval for μ, if $n = 16$ and $\overline{X} = 175$.

Solution:

1. Recall that $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, or, equivalently: $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$.

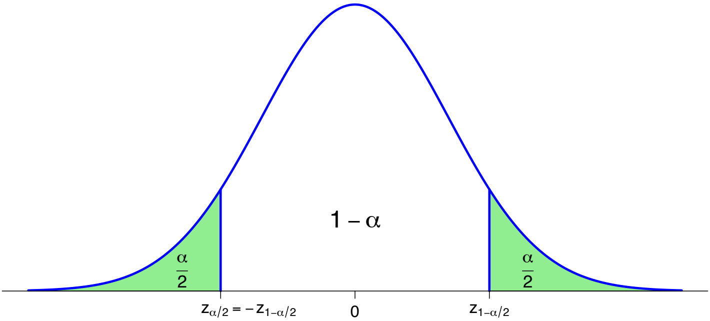
   Think of a pair of numbers, a and b, that satisfy –

$$\mathbb{P}\left(a \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq b\right) = 1 - \alpha$$

   – infinitely many options, but a natural choice would be $a = z_{\alpha/2}$ and $b = z_{1-\alpha/2}$ – the quantiles of the standard Normal distribution.

the symmetric range over normal curve

Normal mean with known variance (cont.)



$$1 - \alpha = \mathbb{P}\left(-z_{1-\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right)$$

$$= \mathbb{P}\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right)$$

UNIVERSITY OF TORONTO

Normal mean with known variance (cont.)

We have shown that $\mathbb{P}\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right) = 1 - \alpha$, hence

$$\left[\overline{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\,,\ \overline{X} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right]$$

is a $100(1-\alpha)\%$ confidence interval for μ.

2. Here $\alpha = 0.05 \Longrightarrow 1 - \dfrac{\alpha}{2} = 0.975$. Substitute

$$\overline{X} \pm \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} = 175 \pm \frac{5}{\sqrt{16}} z_{0.975} = 175 \pm 1.25 \times 1.96$$

$$\Longrightarrow \begin{cases} U = 177.45, \\ L = 172.55, \end{cases}$$

thus $[172.55, 177.45]$ is a 95% confidence interval for μ in this case.

idea is find a pivot that approximates parameter
in this case the pivot is the standardization of sample mean

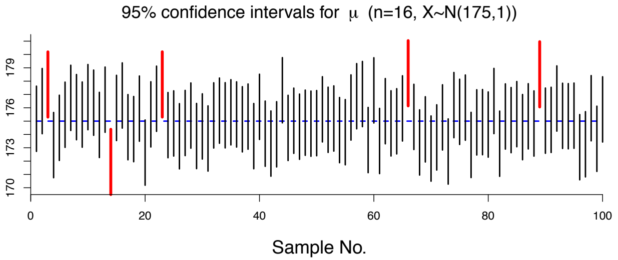
the true population mean is always the center of sampling distribution

Understanding confidence intervals

- So, $[172.55, 177.45]$ is a 95% confidence interval for μ
- Surely that means "μ has a 95% chance of lying between 172.55 and 177.45"...?
  - An outrageous statement! μ is a fixed scalar (albeit an unknown one)
  - What is the chance of 5 lying between 4 and 6? Between 3 and 4?
- In the construction of confidence intervals, it is the interval itself that is random
- A 95% Confidence level suggests that if we had infinitely many random samples and calculated the confidence limits for each, 95% of the resultant intervals would include the true parameter value
- Can only hope that the one we have is a good one...

R simulation

```
> N_Samples <- 100 #No. of random samples
>
> x <- matrix(rnorm(16*N_Samples, mean=175, sd=5), ncol=16) #100 samples of size 16
> xBar <- apply(x, 1, mean) #vector of sample means
> U <- xBar + qnorm(.975)*sigma/4 #upper interval limits
> L <- xBar - qnorm(.975)*sigma/4 #lower interval limits
> uncovered <- which((L>175)|(U<175)) #locating "bad" intervals
>
> plot(c(1:N_Samples), rep(175, N_Samples), type='l', lty=2, col=4, lwd=2)
> segments(1:N_Samples, L, 1:N_Samples, U, lwd=2)
> segments(uncovered, L[uncovered], uncovered, U[uncovered], lwd=4, col=2)
```



The pivotal method

Definition

A *pivotal quantity* (or simply "a pivot") is a function $g(X_1, \ldots, X_n; \theta)$ of the data and parameter of interest, whose distribution does not depend on any unknown parameter.

note sigma and n are all given in the question

- In the last example, $\overline{X} - \mu \sim \mathcal{N}\left(0, \dfrac{\sigma^2}{n}\right)$ served as a pivot
- The pivotal method for confidence interval goes as follows:
  1. Find a pivot $g(X_1, \ldots, X_n; \theta)$ and identify its distribution
  2. Find a and b such that $\mathbb{P}\left(a \leq g(X_1, \ldots, X_n; \theta) \leq b\right) = 1 - \alpha$
  3. Find L and U such that $\mathbb{P}\left(L \leq \theta \leq U\right) = 1 - \alpha$

Example: Normal mean with unknown variance

Example

Repeat the last example, this time with σ^2 unknown, and assuming $S^2 = 25$.

- This time $\overline{X} - \mu$ is no longer a pivot – because σ^2 is unknown.
- However, in the first lecture we verified that $\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$, and is therefore a pivot.

note no population param in pivot

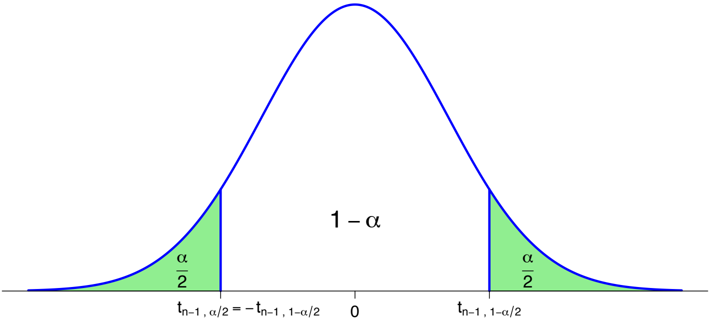
- Now if we look for a and b to satisfy

$$\mathbb{P}\left(a \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq b\right) = 1 - \alpha,$$

we can choose $a = t_{n-1,\alpha/2}$ and $b = t_{n-1,1-\alpha/2}$ – the quantiles of the t_{n-1} distribution!

note n-1 d.f.

Normal mean with unknown variance (cont.)



$$1-\alpha = \mathbb{P}\left(-t_{n-1,1-\alpha/2} \leq \frac{\overline{X}-\mu}{S/\sqrt{n}} \leq t_{n-1,-\alpha/2}\right)$$

$$= \mathbb{P}\left(\overline{X} - \frac{S}{\sqrt{n}} t_{n-1,1-\alpha/2} \leq \mu \leq \overline{X} + \frac{S}{\sqrt{n}} t_{n-1,-\alpha/2}\right)$$

Normal mean with unknown variance (cont.)

- We just showed that

$$\left[\overline{X} - \frac{S}{\sqrt{n}} t_{n-1,1-\alpha/2}\,,\, \overline{X} + \frac{S}{\sqrt{n}} t_{n-1,1-\alpha/2}\right]$$

  is a $100(1-\alpha)\%$ confidence interval for μ.

- For our data

$$\overline{X} \pm \frac{S}{\sqrt{n}} t_{n-1,1-\alpha/2} = 175 \pm \frac{5}{\sqrt{16}} t_{15,0.975} = 175 \pm 1.25 \times 2.131$$

$$\Longrightarrow \begin{cases} U = 177.66, \\ L = 172.34, \end{cases}$$

- Interval of length 5.32 compared to 4.9 when σ^2 was assumed to be known

CI gets larger compared to if sigma^2 is known.

Example: CI for Normal variance

Example

Find a $100(1-\alpha)\%$ confidence interval for σ^2, based on $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$.

Solution:

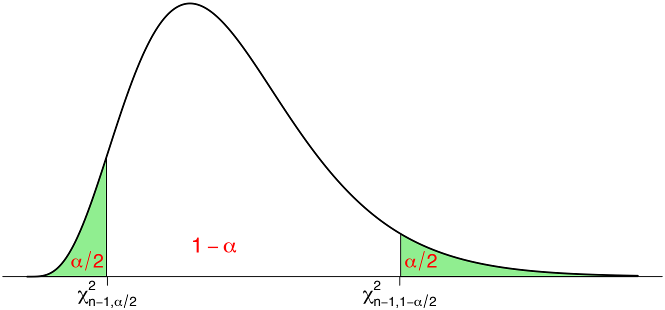
- Recall that $\dfrac{(n-1)S^2}{\sigma^2} = \displaystyle\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 \sim \chi^2_{n-1}$ (a pivot).
- We need to find a and b such that $\mathbb{P}\left(a \leq \dfrac{(n-1)S^2}{\sigma^2} \leq b \right) = 1 - \alpha$

problem chi squared not symmetric

- Ideally, choose them such that the length of the eventual CI is minimized
- A hard optimization problem – not always worth the trouble
- Simply choose $a = \chi^2_{n-1,\alpha/2}$ and $b = \chi^2_{n-1,1-\alpha/2}$,then a $(1-\alpha)100\%$ CI for σ^2 will be

$$\left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \right]$$

The χ^2 quantiles



$$1-\alpha = \mathbb{P}\left(\chi^2_{n-1,\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1,1-\alpha/2}\right)$$

$$= \mathbb{P}\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right).$$

Asymptotic confidence intervals

- When pivots are hard to find, one can invoke large sample theory, namely:

$$\widehat{\theta}_{\text{MLE}} \sim AN\left(\theta, \mathcal{I}^{-1}(\hat{\theta}_{\text{MLE}})\right)$$

plugging in

- Can be taken advantage of to construct $100(1-\alpha)\%$ *asymptotic confidence interval* of the form

this is CI for normal 's mean

$$\boxed{\left[\widehat{\theta}_{\text{MLE}} - \frac{z_{1-\alpha/2}}{\sqrt{\mathcal{I}(\hat{\theta}_{\text{MLE}})}}, \widehat{\theta}_{\text{MLE}} + \frac{z_{1-\alpha/2}}{\sqrt{\mathcal{I}(\hat{\theta}_{\text{MLE}})}}\right].}$$

- For example, for $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$ we calculated $\widehat{\lambda}_{\text{MLE}} = 1/\overline{X}$ and $\mathcal{I}(\lambda) = n/\lambda^2$. A $100(1-\alpha)\%$ confidence interval for θ would then be

substitute mle estimator for true param by plugin principle

$$\left[\frac{1}{\overline{X}} - \frac{z_{1-\alpha/2}}{\overline{X}\sqrt{n}}, \frac{1}{\overline{X}} + \frac{z_{1-\alpha/2}}{\overline{X}\sqrt{n}}\right].$$

Comparing different estimators

- So far we have covered two methods of parameter estimation: the Method of Moments and the Maximum Likelihood principle
- Various other methods exist: Bayesian estimation, Least-Squares estimation etc.
- How do we choose between the different types of estimators then?
- Consider the following *loss function*:

$$\mathscr{L}(\hat{\theta}, \theta) = (\theta - \widehat{\theta})^2 \quad \text{(the \textit{squared error loss})}$$

  - Inflicts harsh penalties on large deviations from the true parameter value
  - Forgiving when it comes to small deviations
  - Overall a good candidate for a measure of estimation accuracy – except that... it's a random variable!

The Mean Squared Error

Definition

The *Mean Squared Error* of an estimator $\widehat{\theta}$ of a parameter θ is

$$\mathrm{MSE}(\hat{\theta}, \theta) = \mathbb{E}\left\{(\hat{\theta} - \theta)^2\right\}.$$

- By and large, we use the MSE to assess goodness-of-estimation out of mathematical convenience
- It could be argued that a more appropriate measure would be the *Mean Absolute Error* $\mathbb{E}\left\{|\theta - \widehat{\theta}|\right\}$, but the latter is not differentiable at the origin
- It does not have the following lovely property either –

The Bias-Variance decomposition

Proposition

Let $\widehat{\theta}$ be an estimator of a parameter θ, and denote

$$b(\hat{\theta}, \theta) = \mathbb{E}[\hat{\theta}] - \theta \quad \text{(the } \textit{bias} \text{ of } \widehat{\theta}\text{)}.$$

Then

$$\text{MSE}(\hat{\theta}, \theta) = b^2(\hat{\theta}, \theta) + \text{Var}[\hat{\theta}].$$

Proof:

note \hat{\theta} is the RV here, \theta is just a constant

$$\text{MSE}(\hat{\theta}, \theta) = \mathbb{E}\left\{(\hat{\theta} - \theta)^2\right\} = \mathbb{E}\left\{\left(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta\right)^2\right\}$$

recognize that bias is a constant

recognize this is the variance

$$= \mathbb{E}\left\{\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2\right\} + \mathbb{E}\left\{\left(\mathbb{E}[\hat{\theta}] - \theta\right)^2\right\} + 2\mathbb{E}\left\{\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)\left(\mathbb{E}[\hat{\theta}] - \theta\right)\right\}$$

$$= b^2(\hat{\theta}, \theta) + \text{Var}[\hat{\theta}] + 2b(\hat{\theta}, \theta)\mathbb{E}\left\{\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)\right\} = b^2(\hat{\theta}, \theta) + \text{Var}[\hat{\theta}].$$

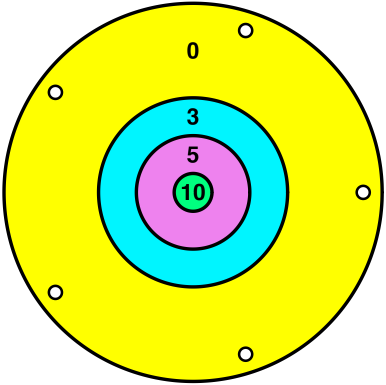
Making sense of the Bias-Variance decomp.

- Think of an Olympic shooter, trying to earn her bread at a competition

sportskeeda.com

The Bias-Variance decomposition (cont.)

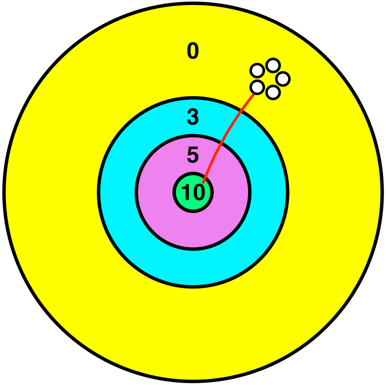
- A shaky hand will not win her any medals



- This is the variance!

The Bias-Variance decomposition (cont.)

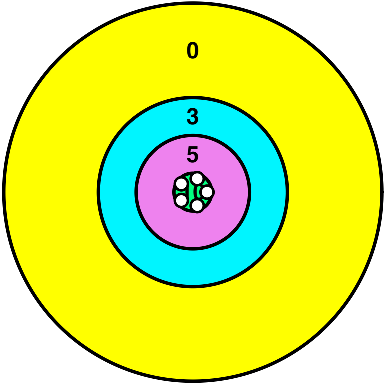
- But if her rifle is out of whack, not even the steadiest of hands will save her



- This is the bias!

The Bias-Variance decomposition (cont.)

- High accuracy requires both a steady hand and zeroed sights



- This is the MSE!

Example: Bernoulli trials

Example

Suppose that we observe a series of Bernoulli trials $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Binom}(1, p)$. Compare the following estimators of p (in terms of their MSE):

1. $\widehat{p}_1 = \overline{X}$ (MME and MLE)
2. $\widehat{p}_2 = \dfrac{\sum_{i=1}^n X_i + 1}{n+2}$ (Bayesian estimator)
3. $\widehat{p}_3 = X_1$

Solution:

1. As always with the sample mean, $\mathbb{E}[\hat{p}_1] = \mathbb{E}[\bar{X}] = \mathbb{E}[X] = p$. The MSE thus reduces to the variance (why?):

$$\text{MSE}(\hat{p}_1, p) = \underline{\text{Var}[\hat{p}_1] = \text{Var}[\bar{X}]} = \frac{\text{Var}[X]}{n} = \frac{p(1-p)}{n}.$$

Bernoulli trials (cont.)

Solution (cont.):

2. First, let us calculate

$$\mathbb{E}[\hat{p}_2] = \mathbb{E}\left[\frac{\sum_{i=1}^n X_i + 1}{n+2}\right] = \frac{\sum_{i=1}^n \mathbb{E}\left[X_i\right] + 1}{n+2} = \frac{np+1}{n+2},$$

and so the bias is $b(\hat{p}_2, p) = \dfrac{np+1}{n+2} - p = \dfrac{1-2p}{n+2}$. As for the variance,

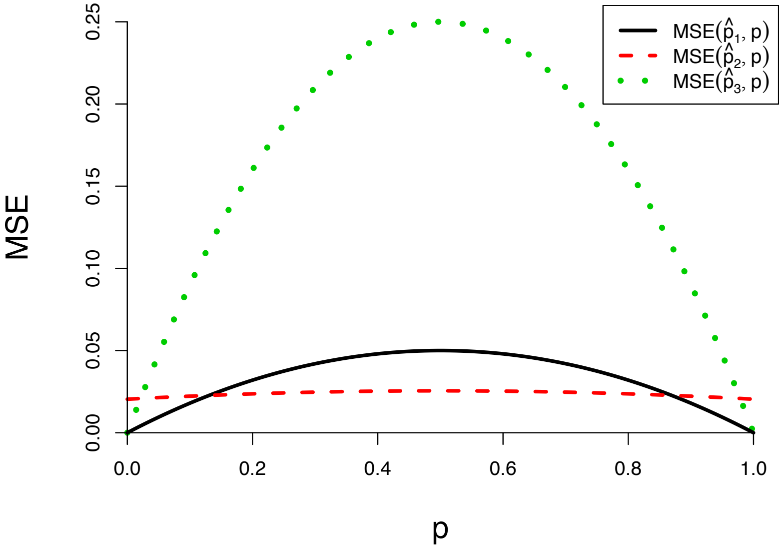
$$\mathrm{Var}[\hat{p}_2] = \mathrm{Var}\left[\frac{\sum_{i=1}^n X_i + 1}{n+2}\right] = \frac{\sum_{i=1}^n \mathrm{Var}\left[X_i\right]}{(n+2)^2} = \frac{np(1-p)}{(n+2)^2},$$

and finally

$$\mathrm{MSE}(\hat{p}_2, p) = b^2(\hat{p}_2, p) + \mathrm{Var}[\hat{p}_2] = \frac{(1-2p)^2 + np(1-p)}{(n+2)^2}.$$

3. Trivially, $\mathbb{E}[\hat{p}_3] = p$, therefore $\mathrm{MSE}(\hat{p}_3, p) = \mathrm{Var}[\hat{p}_3] = p(1-p)$.

Bernoulli trials (cont.)



UNIVERSITY OF TORONTO

Example: variance of a Normal population

Example

For $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, Compare the following estimators of σ^2:

1. $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ (the sample variance)
2. $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ (MME and MLE)

Solution: 1. easy to calculate because we find a pivot for S^2

1. Recall that $\dfrac{(n-1)S^2}{\sigma^2} = \displaystyle\sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \sim \chi^2_{n-1}$, therefore

Text

$$\mathbb{E}\left[S^2\right] = \frac{\sigma^2}{n-1}\underline{\mathbb{E}\left[\chi^2_{n-1}\right]} = \frac{\sigma^2}{n-1}\cdot(n-1) = \sigma^2,$$ note S^2 is unbiased

hence

since chi squared with n d.f. is gamma(n/2, 1/2) with mean n and variance 2n

$$\mathrm{MSE}(S^2, \sigma^2) = \mathrm{Var}\left[S^2\right] = \frac{\sigma^4}{(n-1)^2}\mathrm{Var}\left[\chi^2_{n-1}\right] = \frac{\sigma^4 \cdot 2(n-1)}{(n-1)^2} = \frac{2\sigma^4}{n-1}.$$

Variance of a Normal population (cont.)

Solution (cont.):

use the fact that this estimator is a transformation of S^2

2. Clearly $\boxed{\widehat{\sigma}^2 = \frac{(n-1)S^2}{n}}$, thus

so mme and mle are biased

$$\mathbb{E}\left[\widehat{\sigma}^2\right] = \frac{n-1}{n}\mathbb{E}\left[S^2\right] = \frac{(n-1)\sigma^2}{n},$$

therefore

the asymptotic normality still holds n-> infinity

$$b(\widehat{\sigma}^2, \sigma^2) = \frac{(n-1)\sigma^2}{n} - \sigma^2 = -\frac{\sigma^2}{n}.$$

In addition,

$$\mathrm{Var}\left[\widehat{\sigma}^2\right] = \frac{(n-1)^2}{n^2}\mathrm{Var}\left[S^2\right] = \frac{(n-1)^2}{n^2} \cdot \frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2},$$

and finally

$$\mathrm{MSE}(\widehat{\sigma}^2, \sigma^2) = b^2(\widehat{\sigma}^2, \sigma^2) + \mathrm{Var}\left[\widehat{\sigma}^2\right]$$

$$= \frac{(2n-1)\sigma^4}{n^2} < \frac{2\sigma^4}{n-1} = \mathrm{MSE}(S^2, \sigma^2) \quad \text{for any } n \geq 2.$$

o mme and mle estimator is more accurate: bias not necessarily bad

Unbiased estimators

Definition

We say that $\widehat{\theta}$ is an *unbiased* estimator of θ if $\mathbb{E}[\hat{\theta}] = \theta$ (i.e. $b(\hat{\theta}, \theta) = 0$).

- $\overline{X}$ is always an unbiased estimator of $\mu = \mathbb{E}[X]$ by LLN
- S^2 is always an unbiased estimator of $\sigma^2 = \mathrm{Var}[X]$ (Practice Problem Set 1)
- Can always correct bias by scaling or shifting – not always beneficial in terms of the MSE
- Unbiased estimators are not necessarily superior to biased ones – yet we love them. Mostly because
  1. For an unbiased $\widehat{\theta}$,
  $$\mathrm{MSE}(\hat{\theta}, \theta) = \mathrm{Var}[\hat{\theta}]$$
  – compact!
  2. We have some seriously nice theory for unbiased estimators

The Cramér–Rao lower bound

Harald Cramér, 1893-1985
Source: insurancehalloffame.org

Calyampudi R. Rao, 1920–
Source: isical.ac.in

The Cramér–Rao lower bound (cont.)

Theorem

Let $X_1, \ldots, X_n \sim f_\theta$, and let $\widehat{\theta}$ be an unbiased estimator of θ. Under some regularity conditions

$$\mathrm{Var}[\hat{\theta}] \geq \mathcal{I}^{-1}(\theta),$$

where $\mathcal{I}(\theta)$ is the Fisher Information.

variane for mle are as good as it gets
for unbiased estimator asymptotically

Proof:

Denoting $\underline{x} = (x_1, \ldots, x_n)$, we have

score

$$\ell'(\theta) = \frac{\partial \log f(\underline{x}|\theta)}{\partial \theta} = \frac{\frac{\partial f(\underline{x}|\theta)}{\partial \theta}}{f(\underline{x}|\theta)} \implies \frac{\partial f(\underline{x}|\theta)}{\partial \theta} = \ell'(\theta) f(\underline{x}|\theta) = u(\theta) f(\underline{x}|\theta),$$

where $u(\theta)$ is the Score statistic. Now, since $\widehat{\theta}$ is unbiased, we know that

$$\theta = \mathbb{E}[\hat{\theta}] = \int \hat{\theta}(\underline{x}) f(\underline{x}|\theta) \mathrm{d}\underline{x},$$

uses the fact that theta is unbiased here

UNIVERSITY OF TORONTO

The Cramér–Rao lower bound (cont.)

theta hat is not a function of theta, so skip..

Proof (cont.):

Having established that

$$\theta = \mathbb{E}[\hat{\theta}] = \int \hat{\theta}(\underline{x}) f(\underline{x}|\theta)\mathrm{d}\underline{x},$$

we can differentiate to obtain

$$1 = \frac{\partial\theta}{\partial\theta} = \frac{\partial}{\partial\theta}\int \hat{\theta}(\underline{x}) f(\underline{x}|\theta)\mathrm{d}\underline{x} = \int \hat{\theta}(\underline{x})\frac{\partial f(\underline{x}|\theta)}{\partial\theta}\mathrm{d}\underline{x}$$

Text

Cov(X,Y) = E[XY] - E[X]E[Y]

by previously

$$= \int \hat{\theta}(\underline{x}) u(\theta) f(\underline{x}|\theta)\mathrm{d}\underline{x} = \underline{\mathbb{E}[\hat{\theta}\cdot u(\theta)] = \mathrm{Cov}\left(\hat{\theta}, u(\theta)\right) \quad \text{(why?)}}$$

+ E[theta]E[u(theta)], which is 0 because E[u(theta)] = 0

$$\leq \sqrt{\mathrm{Var}[\hat{\theta}]}\cdot\sqrt{\mathrm{Var}[u(\theta)]} = \sqrt{\mathrm{Var}[\hat{\theta}]}\cdot\sqrt{\mathcal{I}(\theta)},$$

since we proved last week that $\cancel{\mathrm{Var}[\hat{\theta}]} = \mathcal{I}(\theta)$, which completes the proof.

this is true by the fact that Corr(X,Y) = Cov(X,Y)/sqrt{Var{X}Var{Y}} <= 1 i.e. correlation is between -1 and 1

Note Var[u(theta)] = I(theta)

Example: the Poisson distribution

- For $X_1 \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Pois}(\lambda)$ we have already calculated the log-likelihood

$$\ell(\lambda) = n\overline{X} \log \lambda - n\lambda + \text{const}$$

and concluded that the MLE of λ was $\widehat{\lambda}_{\text{MLE}} = \overline{X}$. In particular, it is unbiased.

- Further calculations yield

$$\ell'(\lambda) = \frac{n\overline{X}}{\lambda} - n \quad \text{and} \quad \ell''(\lambda) = -\frac{n\overline{X}}{\lambda^2}$$

- Note that $n\overline{X} = \sum_{i=1}^{n} X_i \sim \text{Pois}(n\lambda)$, thus $\mathbb{E}[n\overline{X}] = \text{Var}[n\overline{X}] = n\lambda$.

poisson processes

- The Fisher Information is therefore

$$\mathcal{I}(\lambda) = -\mathbb{E}[\ell''(\lambda)] = \mathbb{E}\left[\frac{n\overline{X}}{\lambda^2}\right] = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}.$$

Example: Poisson distribution (cont.)

- We have calculated $\mathcal{I}(\lambda) = \frac{n}{\lambda}$
- The CR bound guarantees that for any unbiased estimator $\widehat{\lambda}$ of λ

$$\operatorname{Var}[\hat{\lambda}] \geq \mathcal{I}^{-1}(\lambda) = \frac{\lambda}{n}.$$

unbiased

- However, for $\widehat{\lambda}_{\text{MLE}} = \overline{X}$ we have

$$\operatorname{Var}[\hat{\lambda}_{\text{MLE}}] = \operatorname{Var}[\bar{X}] = \frac{\operatorname{Var}[X]}{n} = \frac{\lambda}{n}.$$

- The MLE achieves the CR bound in this case!
- We know for sure then that no unbiased estimator of λ outperforms $\overline{X}$.

achieves CR bound: allows to prove
optimality of unbiase estimators

Example: Normal distribution

- For $X_1 \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ we have already calculated the log-likelihood

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 + \text{const.}$$

- $\dfrac{\partial \ell}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2(\sigma^2)^2} \displaystyle\sum_{i=1}^{n} (X_i - \mu)^2.$

- $\dfrac{\partial^2 \ell}{\partial (\sigma^2)^2} = \dfrac{n}{2\sigma^4} - \dfrac{\cancel{n}}{\sigma^6} \displaystyle\sum_{i=1}^{n} (X_i - \mu)^2 = \dfrac{n}{2\sigma^4} - \dfrac{1}{\sigma^4} \displaystyle\sum_{i=1}^{n} \left(\dfrac{X_i - \mu}{\sigma} \right)^2$

- Recall that $\displaystyle\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2_n$, then easier to find expected value…

$$\mathcal{I}(\sigma^2) = -\mathbb{E}\left\{ \frac{\partial^2 \ell}{\partial (\sigma^2)^2} \right\} = -\frac{n}{2\sigma^4} + \frac{1}{\sigma^4} \mathbb{E}\left\{ \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \right\}$$

$$= -\frac{n}{2\sigma^4} + \frac{n}{\sigma^4} = \frac{n}{2\sigma^4}.$$

sigma^2 is the unit of differentiation

Example: Normal distribution (cont.)

- We just calculated: $\mathcal{I}(\sigma^2) = \dfrac{n}{2\sigma^4}$
- The CR bound for any unbiased estimator $\widehat{\sigma}^2$ of σ^2 is thus

$$\mathrm{Var}[\hat{\sigma}^2] \geq \mathcal{I}^{-1}(\sigma^2) = \frac{2\sigma^4}{n}$$

- The sample variance S^2 is unbiased, and we calculated

$$\mathrm{Var}[S^2] = \frac{2\sigma^4}{n-1} \Longrightarrow \text{does not achieve the CR bound.}$$

- However, $\lim\limits_{n\to\infty} \dfrac{\mathrm{Var}[S^2]}{\mathcal{I}^{-1}(\sigma^2)} = 1.$

note, S^2 does not achieve CR bound
but its negligible. We say S^2 is asymptotically efficient

Efficiency

Definition

1. We say that an unbiased estimator $\widehat{\theta}$ of a parameter θ is *finite sample efficient* (or simply "efficient") if

$$\mathrm{Var}[\hat{\theta}] = \mathcal{I}^{-1}(\theta).$$

(i.e. it achieves the CR lower bound).

2. We say that $\widehat{\theta}$ is *asymptotically efficient* if

$$\lim_{n\to\infty} \frac{\mathrm{Var}[\hat{\theta}]}{\mathcal{I}^{-1}(\theta)} = 1.$$

3. The *Relative Efficiency* of an unbiased estimator $\widehat{\theta}_1$ of θ with respect to another unbiased estimator $\widehat{\theta}_2$ is

$$\mathrm{eff}(\widehat{\theta}_1, \widehat{\theta}_2) = \frac{\mathrm{Var}[\hat{\theta}_2]}{\mathrm{Var}[\hat{\theta}_1]}.$$

Efficiency (cont.)

- In the Poisson example, $\widehat{\lambda}_{\mathrm{MLE}} = \overline{X}$ achieved the CR lower bound, hence it is efficient.
- In the Normal example
$$\lim_{n\to\infty} \frac{\mathrm{Var}[S^2]}{\mathcal{I}^{-1}(\sigma^2)} = 1,$$
thus S^2 is asymptotically efficient.

note S^2 is not an MLE, but still asymptotically efficient

- When we learned about large sample properties of Maximum Likelihood Estimators, we proved that (under some conditions)
$$\widehat{\theta}_{\mathrm{MLE}} \sim AN(\theta, \mathcal{I}^{-1}(\theta)),$$
therefore MLEs are *asymptotically unbiased* and asymptotically efficient.

doesnt imply that finite sample of MLE is efficient still have to check

Muon decay example

- X was the cosine of the angle at which electrons are released, with pdf

$$f(x|\alpha) = \frac{1+\alpha x}{2}, \quad -1 \le x \le 1, \quad -1 \le \alpha \le 1.$$

- We calculated $\mathbb{E}[X] = \frac{\alpha}{3}$. Similarly,

$$\mathbb{E}[X^2] = \int_{-1}^{1} x^2 \frac{1+\alpha x}{2} \mathrm{d}x = \frac{1}{3}$$

question from HW

$$\Longrightarrow \boxed{\mathrm{Var}[X]} = \mathbb{E}[X^2] - \{\mathbb{E}[X]\}^2 = \frac{1}{3} - \frac{\alpha^2}{9} = \frac{3-\alpha^2}{9}.$$

- The Method of Moments estimator was found to be $\widehat{\alpha}_{\mathrm{MME}} = 3\overline{X}$, with

$$\mathbb{E}[\hat{\alpha}_{\mathrm{MME}}] = 3\mathbb{E}[\bar{X}] = 3\mathbb{E}[X] = \alpha \Longrightarrow \text{unbiased,}$$

and

$$\mathrm{Var}[\hat{\alpha}_{\mathrm{MME}}] = 9\mathrm{Var}[\bar{X}] = \frac{9\mathrm{Var}[X]}{n} = \frac{3-\alpha^2}{n}.$$

method of moments estimator

UNIVERSITY OF TORONTO

Muon decay example (cont.)

remember we used newton raphson previously

- The Maximum Likelihood estimator, $\widehat{\alpha}_{\text{MLE}}$, is not given in a closed form: cannot calculate its exact sampling distribution.

- We do know that for large samples, $\widehat{\alpha}_{\text{MLE}} \sim \mathcal{N}\left(\alpha, \mathcal{I}^{-1}(\alpha)\right)$ (approximately).

by asymptotic normality

- Calculate

$$\mathcal{I}(\alpha) = n\mathcal{I}^*(\alpha) = -n\mathbb{E}\left[\frac{\partial^2 \log f(x|\alpha)}{\partial \alpha^2}\right] = -n\int \frac{\partial^2 \log f(x|\alpha)}{\partial \alpha^2} f(x|\alpha)\mathrm{d}x$$

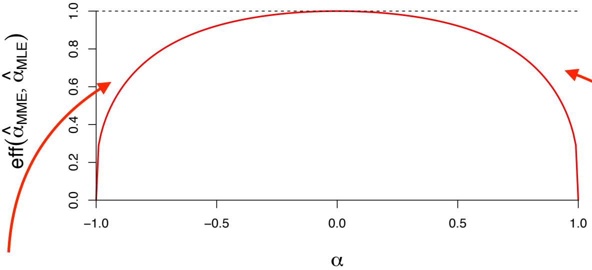
$$= n\int_{-1}^{1} \frac{x^2}{(1+\alpha x)^2}\frac{1+\alpha x}{2}\mathrm{d}x = \begin{cases} \dfrac{n\left(\log\dfrac{1+\alpha}{1-\alpha} - 2\alpha\right)}{2\alpha^3} & , \quad \alpha \neq 0, \\ \dfrac{n}{3} & , \quad \alpha = 0. \end{cases}$$

can also calculate fisher info with I*

Muon decay example (cont.)

- The *asymptotic* relative efficiency is thus

$$\mathrm{eff}(\hat{\alpha}_{\mathrm{MME}}, \hat{\alpha}_{\mathrm{MLE}}) = \frac{\mathrm{Var}[\hat{\alpha}_{\mathrm{MLE}}]}{\mathrm{Var}[\hat{\alpha}_{\mathrm{MME}}]} = \frac{2\alpha^3}{3-\alpha^2}\left(\log\frac{1+\alpha}{1-\alpha} - 2\alpha\right)^{-1} \quad (\alpha \neq 0).$$



Var[\alpha_{MME}] increasing toward boundary

- Note how much efficiency the MME loses (relative to the MLE) close to the boundary of the parameter space!