ST5201: Basic Statistical Theory
Chapter 8: Estimation of Parameters and Fitting of Probability Distributions

CHOI Yunjin
stachoiy@nus.edu.sg

Department of Statistics and Applied Probability
National University of Singapore (NUS)

17th October, 2017

Outline

- Review
- Desired properties of estimators
- Cramer-Rao Lower bound

Review: Method of moments

- The method of moments
  1. Suppose $\theta = (\theta_1, \theta_2, \cdots, \theta_K) \in \mathbb{R}^K$
  2. Calculate K lower order moments in terms of θ:

$$E(X) = h_1(\theta), E(X^2) = h_2(\theta), \cdots, E(X^K) = h_K(\theta)$$

  3. Find the inverse function of h's to express the parameter θ's.

$$\begin{aligned} \theta_1 &= f_1(E(X), E(X^2), \cdots, E(X^K)) \\ &\cdots \\ \theta_K &= f_K(E(X), E(X^2), \cdots, E(X^K)) \end{aligned}$$

  4. Insert the sample moments into the expressions, thus obtaining the estimators $\hat{\theta}$:

$$\begin{aligned} \hat{\theta}_1 &= f_1\left(\frac{1}{n}\sum_{i=1}^n X_i, \frac{1}{n}\sum_{i=1}^n X_i^2, \cdots, \frac{1}{n}\sum_{i=1}^n X_i^K\right) \\ &\cdots \\ \hat{\theta}_K &= f_K\left(\frac{1}{n}\sum_{i=1}^n X_i, \frac{1}{n}\sum_{i=1}^n X_i^2, \cdots, \frac{1}{n}\sum_{i=1}^n X_i^K\right) \end{aligned}$$

- Remark: The method of moments
  - $\mathbb{P}_\theta$ can be any distribution indexed by θ. It is not required to belong to the known density functions.
  - Advantages:
    - Generally, the estimator is easy to calculate
    - The estimator is consistent
  - Disadvantages:
    - Existence of moments required
    - Sometimes, it is hard to find the limiting distribution of $\hat{\theta}$
    - It does not consider the parameter space Θ

Review: Maximum likelihood estimator

- Maximum Likelihood Estimator (MLE)
  - $\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} L_n(\theta)$
    or equivalently, $\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} l_n(\theta)$
    where $L_n(\theta) := \prod_{i=1}^n f(x_i|\theta)$ is a likelihood function and
    $l_n(\theta) := \sum_{i=1}^n \log f(x_i|\theta)$ is a log likelihood function
  - Advantages:
    - MLE is a consistent estimator. i.e., $\hat{\theta}_n \xrightarrow{p} \theta$
    - Limiting distribution is clear
    - Considers the parameter space Θ
    - Allow relationship between samples as long as $L_n(\theta)$ is known.
  - Disadvantages:
    - Calculating the maximizer of a function can be complicated, or even impossible in exact way.

- Maximum Likelihood Estimator (MLE)
  - Score Function: $\frac{\partial}{\partial\theta} \ln f(X|\theta)$.
  - Fisher Information:
    - $I(\theta) = E\left[\left(\frac{\partial}{\partial\theta} l(\theta)\right)^2\right]$.
    - Under appropriate smoothness condition on f, $I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2} \ln f(X|\theta)\right]$
    - Fisher information of the random i.i.d. sample $\mathbf{X} = (X_1, \cdots, X_n)$: $I_n(\theta) = E\left[\left(\frac{\partial}{\partial\theta} l_n(\theta)\right)^2\right] = -E\left[\frac{\partial^2}{\partial\theta^2}\left(\sum_{i=1}^n \ln f(X_i|\theta)\right)\right] = nI(\theta)$
  - Asymptotic Normality of MLE
    For i.i.d. samples $X_1, \cdots, X_n$ from $f(x|\theta_0)$,
    the limiting distribution of the MLE $\hat{\theta}_n$ is normal:
    $\sqrt{I_n(\theta_0)}(\hat{\theta}_n - \theta_0) = \sqrt{nI(\theta_0)}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, 1)$

Recall a previous example: Suppose that X is a discrete r.v. with

$$P(X=0)=\frac{2}{3}\theta,\ P(X=1)=\frac{1}{3}\theta,\ P(X=2)=\frac{2}{3}(1-\theta),\ P(X=3)=\frac{1}{3}(1-\theta),$$

and $P(X=x)=0$ for $x\notin\{0,1,2,3\}$, where $0\le\theta\le 1$ is a parameter. Here are 10 indept. observations taken from such a distribution: (3, 0, 2, 1, 3, 2, 1, 0, 2, 1). Find the MLE of θ.

Solution.

- MLE: $\hat{\theta}_{MLE}=\frac{N_0+N_1}{N_0+N_1+N_2+N_3}=\frac{1}{2}$, where $N_k=\#$ observations with $X_i=k$, $1\le i\le n$, $k=0,1,2,3$.
- MM: $\hat{\theta}_{MM}=(\bar{X}_n-7/3)/(-2)=\frac{5}{12}$
- More estimators, e.g., $\hat{\theta}=\frac{N_1}{N_0+N_1+N_2+N_3}$, $1/N_1$, etc
- Which is better?

- Estimator from Method of Moments
- Maximum Likelihood Estimator
- Sometimes, they are the same (e.g., Bernoulli, Poisson, ...)
- Sometimes, they may be different
  - Example in the previous slide. MM estimator has value $5/12$, and MLE gives estimate as $1/2$
- We could define even more estimators, say, $\frac{N_1}{N_0+N_1+N_2+N_3}$, $1/N_1$, etc.
- Question: **Which is the best estimator?**

1) Consistency

- An estimator is consistent if the estimate $\hat{\theta}$ is guaranteed to converge to the true parameter value θ_0 as the quantity of data increases.
- Nearly always a desirable property for a statistical estimator
  - MM is consistent
  - MLE is consistent
  - For the above problem, $\frac{N_1}{N_0+N_1+N_2+N_3} \to \frac{\theta}{3}$ is not consistent

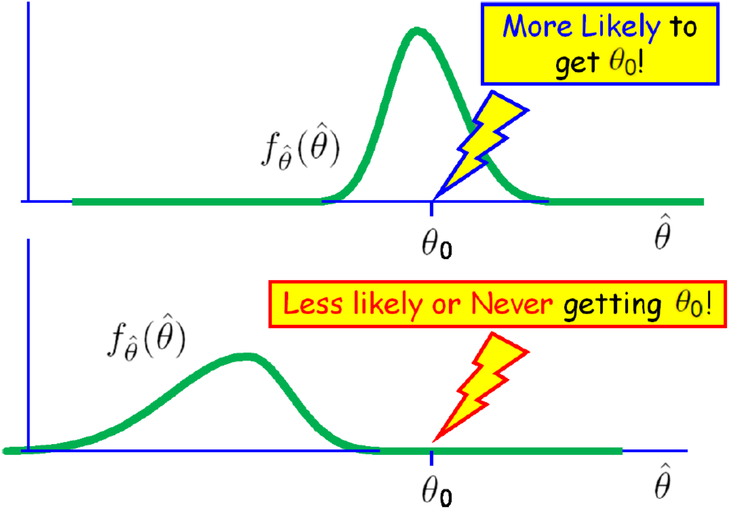
- All estimates $\hat{\theta}$ of parameter θ are statistics, i.e., r.v.'s, such that

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

Sampling distribution of $\hat{\theta}$

The probability distribution of any estimate $\hat{\theta}$ of a parameter θ of an underlying probability model of interest is called the *sampling distribution of $\hat{\theta}$*. We denote its density by $f_{\hat{\theta}}$.

- The asymptotic distribution is the approximation of the sampling distribution of $\hat{\theta}$ when $n \to \infty$
- For an estimator, we hope $\hat{\theta}$ has "*center*" near to θ and small "*spread*"



[1] θ_0 is true value of θ

2) Bias

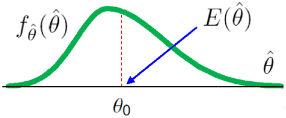
- To make sure that the "center" of $\hat{\theta}_n$ is near to θ_0, we define $Bias{=}E(\hat{\theta}_n) - \theta_0$.
- It is desirable that the bias of $\hat{\theta}_n$ is 0 ($E(\hat{\theta}_n) = \theta_0$) or at least small ($E(\hat{\theta}_n) \approx \theta_0$).
- $\hat{\theta}_n$ is called unbiased (or, an unbiased estimator) if it has zero bias ($E(\hat{\theta}_n) = \theta$)
- For the example on page 7:
  - MLE is unbiased: $E(\hat{\theta}_{MLE}) = E(\frac{N_0+N_1}{N_0+N_1+N_2+N_3}) = \theta$
  - MM is unbiased: $E(\hat{\theta}_{MLE}) = E(\frac{\bar{X}_n-7/3}{-2}) = \theta$
  - It is possible that an estimator is consistent but biased, e.g., the estimator $\hat{\theta} = \frac{N_0+N_1+1}{N_0+N_1+N_2+N_3}$ has expectation $\frac{n\theta+1}{n} \neq \theta$, but converges to θ when $n \to \infty$

- "Center" of the sampling distribution of an estimate $\hat{\theta}$ is represented by its mean value

$$\mu_{\hat{\theta}} = E(\hat{\theta}) = \int_{\Theta} x f_{\hat{\theta}}(x) dx$$

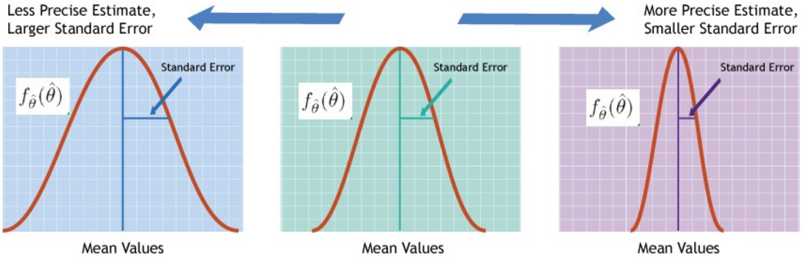
Unbiasedness means this center *equals to* the truth

- Unbiasedness is a very important criterion to address an estimate due to the long-run interpretation of $E(\hat{\theta}) = \theta_0$.



- Unbiased estimate may not be unique. For example, in the above problem, both MLE and MM are unbiased estimators.

Sampling Distribution [2]



[2] θ_0 is true value of θ

3) Variance

- To make sure the estimator we got is "close" to θ_0, we hope the "spread" of $f_{\hat{\theta}}$ is small
- The spread can be expressed by the variance of $\hat{\theta}_n$, $\mathrm{Var}(\hat{\theta}_n)$
- Statistical Metric–$\mathrm{Var}(\hat{\theta})$
- For the example on page 7:
  - MLE: $\mathrm{Var}(\hat{\theta}_{MLE}) = \mathrm{Var}(\frac{N_0+N_1}{N_0+N_1+N_2+N_3}) = \mathrm{Var}(\frac{N_0+N_1}{n}) = \frac{\theta(1-\theta)}{n}$ ($N_0 + N_1 \sim \mathrm{Bin}(n, \theta)$, n is fixed)
  - MM: $\mathrm{Var}(\hat{\theta}_{MM}) = \mathrm{Var}(\frac{\bar{X}_n - 7/3}{-2}) = \frac{1}{4}\mathrm{Var}(\bar{X}_n) = \frac{1}{4n}\mathrm{Var}(X_1) = \frac{\theta(1-\theta)+5/36}{n}$
  - MLE has slightly smaller variance

- The variance of $\hat{\theta}$ is usually denoted by $\sigma^2_{\hat{\theta}}$, and the corresponding standard deviation of $\hat{\theta}$ (also called as *standard error*) is denoted by $\sigma_{\hat{\theta}}$
- "spread" of the sampling distribution of an estimate $\hat{\theta}$

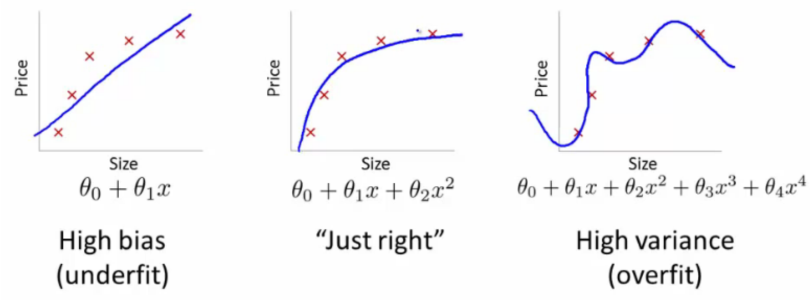
$$\sigma^2_{\hat{\theta}} = \mathrm{Var}(\hat{\theta}) = E(\hat{\theta}^2) - (E(\hat{\theta}))^2 = \int_{\Theta} u^2 f_{\hat{\theta}}(u)du - \mu^2_{\hat{\theta}}$$

- In practice, θ is unknown. Usually, after we obtain the formula for $\sigma_{\hat{\theta}}$ as $\sqrt{h(\theta)}$, we introduce $\hat{\theta}$ into the formula $\sqrt{h(\theta)}$ as an estimate for the standard error

Definition

When $\sigma^2_{\hat{\theta}} = h(\theta)$, $s_{\hat{\theta}} = \sqrt{h(\hat{\theta})}$ is often used to replace $\sigma_{\hat{\theta}}$, and is called *estimated standard error of* $\hat{\theta}$

Bias-variance Example



- Each of bias and variance comes at the cost of the other
- The choice depends on the relative importance of expected accuracy versus reliability in the task at hand

- We hope to have only ONE value for assessing an estimator, instead of both bias and variance. Then we choose the estimator with SMALLEST number
- That number must assure both small bias and variance
- One generally used criteria is mean squared error

Definition: Mean Square Error

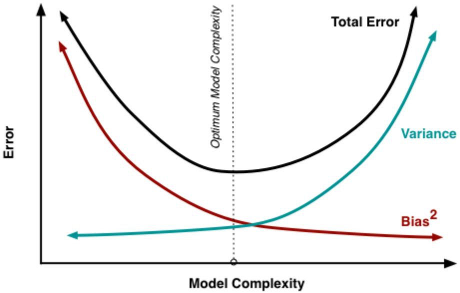
The *mean square error (MSE) of an estimator $\hat{\theta}$* of a parameter θ is defined by

$$\begin{aligned} MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 &= \mathrm{Var}(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2 \\ &= \text{Variance of } \hat{\theta} + \text{Squared Bias of } \hat{\theta} \end{aligned}$$

Remark: averaging over the data X, not the θ.

Decomposition of MSE

$$
\begin{aligned}
MSE(\hat\theta) = E(\hat\theta-\theta)^2 &= E(\hat\theta - E(\hat\theta) + E(\hat\theta) - \theta)^2 \\
&= E\left[(\hat\theta - E(\hat\theta))^2 + 2(\hat\theta - E(\hat\theta))(E(\hat\theta)-\theta) + (E(\hat\theta)-\theta)^2\right] \\
&= E(\hat\theta - E(\hat\theta))^2 + 2E\left[(\hat\theta - E(\hat\theta))(E(\hat\theta)-\theta)\right] + E(E(\hat\theta)-\theta)^2 \\
&= \mathrm{Var}(\hat\theta) + 2(E(\hat\theta)-\theta)E(\hat\theta - E(\hat\theta)) + (E(\hat\theta)-\theta)^2 \\
&= \mathrm{Var}(\hat\theta) + 2(E(\hat\theta)-\theta)(E(\hat\theta) - E(\hat\theta)) + \mathrm{Bias}^2(\hat\theta) \\
&= \mathrm{Var}(\hat\theta) + \mathrm{Bias}^2(\hat\theta)
\end{aligned}
$$



- It combines both bias and variance. Here, it uses Bias^2 so that it has the same unit with the variance
- If an estimator $\hat{\theta}_n$ has MSE $\to 0$, then this estimator is consistent
- MSE is easy to calculate, compare to Bias+Standard Deviation

Proof for Consistency:

If $\hat{\theta}_n$ has MSE converges to 0, then $E(\hat{\theta}_n - \theta_0)^2 \to 0$ when $n \to \infty$.

$$
\begin{aligned}
P(|\hat{\theta}_n - \theta_0| > \epsilon) &= P(|\hat{\theta}_n - E(\hat{\theta}_n) + E(\hat{\theta}_n) - \theta_0| > \epsilon) \\
&\leq P(|\hat{\theta}_n - E(\hat{\theta}_n)| > \epsilon - |E(\hat{\theta}_n) - \theta_0|) \\
&\leq \frac{\mathrm{Var}(\hat{\theta}_n)}{[\epsilon - (E(\hat{\theta}_n) - \theta_0))]^2} = \frac{\mathrm{Var}(\hat{\theta}_n)}{(\epsilon - \mathrm{Bias})^2}
\end{aligned}
$$

- As MSE converges to 0, the bias converges to 0, so the denominator converges to ϵ^2.
- As MSE converges to 0, $\mathrm{Var}(\hat{\theta})_n)$ converges to 0, so the numerator converges to 0.
- So, $P(|\hat{\theta}_n - \theta_0| > \epsilon) \to 0$

We know that the MLE's for $Ber(p)$ is $\hat{p} = \bar{X}_n$

- Sampling distribution

$$\hat{p}_n = \frac{1}{n}Y, \qquad Y \sim Bin(n, p)$$

- Unbiasedness: $E(\hat{p}_n) = p$ (unbiased)
- Standard error: $\sigma_{\hat{p}_n} = \sqrt{p(1-p)/n} \to 0$ as $n \to \infty$
- MSE: $MSE(\hat{p}_n) = \text{Var}(\hat{p}_n) + 0 = \frac{p(1-p)}{n}$.
- Consistency: $\hat{p}_n$ is consistent estimate of p as MSE converges to 0

Remark: It is possible that the sampling distribution is unknown, and we have to use asymptotic distribution

Example: Assessing of MLEs of $N(\mu, \sigma^2)$

We know that the MLE's for $N(\mu, \sigma^2)$ are

$$\hat{\mu} = \bar{X}, \ \ \hat{\sigma^2} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- Sampling distribution

$$\hat{\mu} \sim N(\mu, \frac{\sigma^2}{n}), \ \ \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1} \text{ (introduced later)}$$

- Unbiasedness
  - $E(\hat{\mu}) = \mu$(unbiased)
  - $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$(biased), converges to σ^2 when $n \to \infty$
- Standard error (and as $n \to \infty$)
  - $\sigma_{\hat{\mu}} = \sigma/\sqrt{n} \to 0$
  - $\mathrm{Var}(\hat{\sigma}^2) = 2(n-1)\sigma^4/n^2 \to 0$

Example: Assessing of MLEs of $N(\mu, \sigma^2)$ - cont'd

- MSE: both going to 0 as $n \to \infty$

$$
\begin{aligned}
MSE(\hat{\mu}) &= \mathrm{Var}(\hat{\mu}) + 0 = \frac{\sigma^2}{n} \\
MSE(\hat{\sigma}^2) &= \mathrm{Var}(\hat{\sigma}^2) + \mathrm{Bias}^2(\hat{\sigma}^2) = \mathrm{Var}(\hat{\sigma}^2) + (E(\hat{\sigma}^2) - \sigma^2)^2 \\
&= 2(n-1)\sigma^4/n^2 + (\sigma^2/n)^2 = (2n-1)\sigma^4/n^2
\end{aligned}
$$

- Consistency (usually implied by MSE $\to 0$)
    - Both are consistent estimate of μ and σ^2.
    - Remark: it suffices to check if the MSE of an estimate goes to 0 as $n \to \infty$ to prove consistency

- Sampling Distribution
- Desirable properties
  - Unbiasedness: $E(\hat{p}_n) = p$
  - Standard error: $\sigma_{\hat{p}_n} \to 0$ as $n \to \infty$
  - MSE: $MSE(\hat{p}_n) = \text{Var}(\hat{p}_n) + Bias(\hat{p})^2 \to 0$.
  - Consistency: $\hat{p}_n$ is consistent estimate of p if MSE converges to 0

What is the best we can do?

Cramer-Rao Inequality

Let $X_1, X_2, \cdots, X_n$ be an i.i.d. sample with PDF/PMF $f(x|\theta_0)$. For any unbiased estimator $\hat{\theta}_n$ of the parameter θ, under smoothness assumptions on $f(x|\theta)$, there is

$$\mathrm{Var}(\hat{\theta}_n) \geq \frac{1}{nI(\theta_0)}$$

where $I(\theta_0)$ is the Fisher Information for $f(x|\theta_0)$. This lower bound is called *Cramer-Rao Lower Bound (CRLB)*.

- Provide a benchmark of how good an unbiased estimate is
- However, Cramer-Rao lower bound is not necessarily achieved
- For biased estimator, similar results show that the lower bound for variance is $\frac{(1+b'(\theta_0))^2}{nI(\theta_0)}$, where $b'(\theta)$ is the bias of $\hat{\theta}$. So the lower bound for MSE with biased estimator can also be constructed.
- CRLB decreases when the sample size increases

Definition: Efficienct Estimator

Let $X_1, \cdots, X_n$ be an i.i.d. sample with density $f(x|\theta_0)$. An unbiased estimator $\hat{\theta}_n$ of θ is said to be *efficient* when $\mathrm{Var}(\hat{\theta}) = [nI(\theta_0)]^{-1}$

Definition: Efficiency

Let $X_1, \cdots, X_n$ be an i.i.d. sample with density $f(x|\theta_0)$. For any unbiased estimator $\hat{\theta}_n$ of θ, the *efficiency of $\hat{\theta}_n$* is defined as

$$\mathrm{e}(\hat{\theta}_n) = \frac{[nI(\theta_0)]^{-1}}{\mathrm{Var}(\hat{\theta}_n)}.$$

- Obvisouly, for any unbiased estimator $\hat{\theta}_n$, $\mathrm{e}(\hat{\theta}_n) \leq 1$, which means the variance of $\hat{\theta}_n$ is always larger than CRLB.
- Recall that, the asymptotic variance of the MLE of θ is given by $[nI(\theta_0)]^{-1}$, which means the MLE is asymptotically efficient

Suppose $X_1, \cdots, X_n$ form a sample from $N(\mu, \theta)$ with parameter μ is given but θ is unknown. Calculate the CRLB.

Solution: For normal distribution,

$$f(x|\theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{(x-\mu)^2}{2\theta}}$$

$$l(x|\theta) = \ln f(x|\theta) = -\frac{(x-\mu)^2}{2\theta} - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\theta$$

$$l'(x|\theta) = \frac{(x-\mu)^2}{2\theta^2} - \frac{1}{2\theta} \text{ and } l''(x|\theta) = -\frac{(x-\mu)^2}{\theta^3} + \frac{1}{2\theta^2}$$

It follows that the Fisher information is

$$I(\theta) = -E[l^{''}(X|\theta)] = -E(-\frac{(X-\mu)^2}{\theta^3} + \frac{1}{2\theta^2}) = \frac{E(X-\mu)^2}{\theta^3} - \frac{1}{2\theta^2}$$
$$= \frac{\theta}{\theta^3} - \frac{1}{2\theta^2} = \frac{1}{2\theta^2},\ (\theta > 0)$$

So, we have the CRLB $\frac{2\theta^2}{n}$.

Suppose $X_1, \cdots, X_n$ form a iid sample from Poisson distribution,

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

Find the CRLB for $\hat{\lambda}$.

Solution: For the Poisson distribution,

$$l(\lambda) = X \ln \lambda - \lambda - \ln X!$$

$$l^{'}(\lambda) = \frac{X}{\lambda} - 1 \text{ and } l^{''}(\lambda) = -\frac{X}{\lambda^2}$$

$$I(\lambda) = \frac{E(X)}{\lambda^2} = \frac{1}{\lambda}$$

Finally, we have the CRLB $\frac{\lambda}{n}$.

Example Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. Find the CRLB and, in cases 1. and 2. check whether it is equalled, for the variance of an unbiased estimator of

1. μ when σ^2 is known,
2. σ^2 when μ is known
3. μ when σ^2 is unknown
4. σ^2 when μ is unknown

Solution: The sample joint p.d.f. is

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}\left(x_i - \mu\right)^2/\sigma^2)$$

and

$$\log f_{\mathbf{X}}(\mathbf{x}|\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2/\sigma^2$$

1. *When* σ^2 *is known* $\theta = \mu$ and

$$\log f_{\mathbf{X}}(\mathbf{x}|\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2}\sum_{i=1}^{n}\left(x_i - \theta\right)^2/\sigma^2$$

$$S(\mathbf{x}) = \frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{x}|\theta) = \sum_{i=1}^{n}\left(x_i - \theta\right)/\sigma^2 = \frac{n}{\sigma^2}\left[\bar{x} - \theta\right]$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased estimator of $\theta = \mu$ whose variance equals the CRLB and that $\frac{n}{\sigma^2} = I(\theta)$ i.e. CRLB $= \frac{\sigma^2}{n}$. Thus $\bar{X}$ is a most efficient estimator.

2. *When μ is known* but σ^2 is unknown, $\theta = \sigma^2$ and

$$\log f_{\mathbf{X}}(\mathbf{x}|\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\theta) - \frac{1}{2\theta}\sum_{i=1}^{n}(x_i-\mu)^2$$

Hence

$$\begin{aligned} S(\mathbf{x}) &= \frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{x}|\theta) = -\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n}(x_i-\mu)^2 \\ &= \frac{n}{2\theta^2}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - \theta\right] \end{aligned}$$

$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ is an unbiased estimator of $\theta = \sigma^2$ and $\frac{n}{2\theta^2} = I(\theta)$ i.e. the $CRLB = \frac{2\theta^2}{n} = \frac{2\sigma^4}{n}$

3. and 4. *Case both μ and σ^2 unknown* Here $\boldsymbol{\theta} = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$ i.e. $\theta_1 = \mu$ and $\theta_2 = \sigma^2$

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}\left(x_i - \mu\right)^2/\sigma^2) \\ &\propto \theta_2^{-n/2} \exp\left(-\frac{1}{2\theta_2}\sum_{i=1}^{n}\left(x_i - \theta_1\right)^2\right) \end{aligned}$$

and

$$\log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) = -\frac{n}{2}\log\theta_2 - \frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2$$

Thus

$$\frac{\partial}{\partial\theta_1}\log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)$$

$$\frac{\partial^2}{\partial\theta_1^2}\log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) = -\frac{n}{\theta_2}$$

$$\frac{\partial^2}{\partial\theta_2\partial\theta_1}\log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) = -\frac{1}{\theta_2^2}\sum_{i=1}^{n}(x_i - \theta_1)$$

$$\frac{\partial}{\partial\theta_2}\log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) = -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^{n}(x_i - \theta_1)^2$$

$$\frac{\partial^2}{\partial \theta_2^2} \log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) = \frac{n}{2\theta_2^2} - \frac{1}{\theta_2^3} \sum_{i=1}^{n} (x_i - \theta_1)^2$$

Consequently

$$I_{11}(\boldsymbol{\theta}) = -\mathbf{E}\left(-\frac{n}{\theta_2}\right) = \frac{n}{\theta_2}$$

$$I_{12}(\boldsymbol{\theta}) = -\mathbf{E}\left(-\frac{1}{\theta_2^2} \sum_{i=1}^{n} (X_i - \theta_1)\right) = 0$$

$$I_{22}(\boldsymbol{\theta}) = -\mathbf{E}\left(\frac{n}{2\theta_2^2} - \frac{1}{\theta_2^3} \sum_{i=1}^{n} (X_i - \theta_1)^2\right) = \frac{n}{2\theta_2^2}$$

i.e.

$$I(\boldsymbol{\theta}) = \begin{bmatrix} \frac{n}{\theta_2} & 0 \\ 0 & \frac{n}{2\theta_2^2} \end{bmatrix}$$

and

$$[I(\boldsymbol{\theta})]^{-1} = J(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\theta_2}{n} & 0 \\ 0 & \frac{2\theta_2^2}{n} \end{bmatrix} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

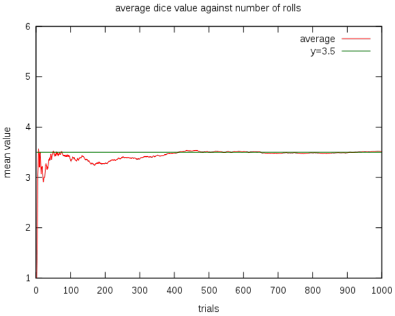
Consequently, for unbiased estimators $\hat{\mu}$, $\hat{\sigma}^2$ of μ and σ^2 respectively

$$Var(\hat{\mu}) \geq \frac{\sigma^2}{n}$$

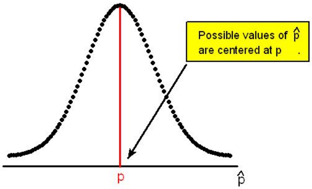
and

$$Var\left(\hat{\sigma}^2\right) \geq \frac{2\sigma^4}{n}$$

- The bias, variance, and MSE help us to select a *good* estimator
- However, even the *best* estimator cannot tell us the truth with 100 percent gaurantee (CRLB)
- Small MSE only helps us to control the estimate to be close to the truth.
  - It is impossible for n to be infinity due to cost, restrictions, etc
  - Even for large n, the estimate is still a r.v. with some fluctuation
- With the estimate, what result can we claim about the truth?



Coin toss problem with $\text{Ber}(p)$ and estimator $\hat{p} = \bar{X}_n$. Say that with $n = 100$ tosses, we got $\bar{X}_n = 0.6$. What can we infer from this result?



- $\hat{p} \sim \frac{1}{100} Bin(100, p)$, asymptotically, $\hat{p} \sim N(p, \frac{p(1-p)}{100})$
- $\hat{p} = 0.6$ is a realization from the distribution $\frac{1}{100} Bin(100, p)$
- For one realization, the probability that "the distance between it and the truth is $\leq c$" (i.e., $P(|\hat{p} - p| \leq c)$) can be calculated, e.g., for $c = 0.1$,

$$\begin{aligned} P(|\hat{p} - p| \leq 0.1) &= P\left(\frac{|\hat{p}-p|}{\sqrt{p(1-p)/100}} \leq \frac{0.1}{\sqrt{p(1-p)/100}}\right) \\ &\approx \Phi\left(\frac{0.1}{\sqrt{p(1-p)/100}}\right) - \Phi\left(-\frac{0.1}{\sqrt{p(1-p)/100}}\right) \end{aligned}$$

- Say $p = 0.5$, then $P(|\hat{p} - p| \leq 0.1) \approx .95$ for one realization $\hat{p}$. With probability 0.95, $|\hat{p} - p| \leq 0.1 \Leftrightarrow \hat{p} - 0.1 \leq p \leq \hat{p} + 0.1$ for 95% of such realizations

Definition: Confidence Interval

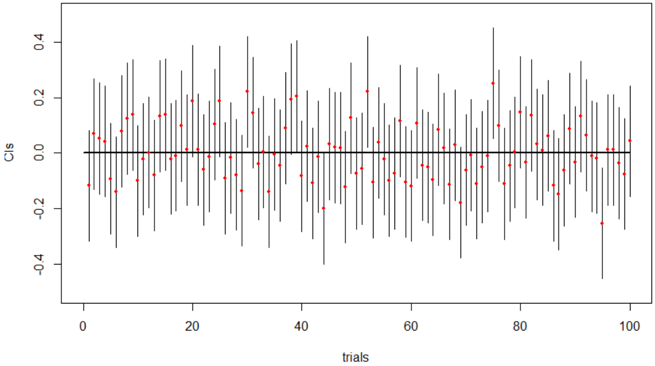
Let $X_1, X_2, \cdots, X_n$ be an i.i.d. random sample with density function $f(x|\theta_0)$, where θ_0 is unknown. For a constant $0 < \alpha < 1$, the $100(1-\alpha)\%$ *confidence interval (CI) of θ* is a random interval (L, U), s.t.

$$P(\theta_0 \in (L, U)) \geq 1 - \alpha.$$

Here, $1 - \alpha$ is called the *confidence level* for this interval.

- Usually, the construction of the confidence interval (L, U) depends on $X_1, X_2, \cdots, X_n$, which means that both L and U are functions of these r.v.'s. So the interval is also random.
- Obviously, to construct CI, we need an estimator and its distribution
- In the coin toss example, the interval $(\hat{p} - 0.1, \hat{p} + 0.1) = (0.5, 0.7)$ is the 95% confidence interval for p, and the confidence level is 0.95.
- Confidence intervals with the same confidence level can be different, depending on different ways of construction

- $L = g(X_1, X_2, \cdots, X_n)$, $U = h(X_1, X_2, \cdots, X_n)$. For one i.i.d sample $X_1, X_2, \cdots, X_n$, we have one interval (L, U)
- If we repeat the procedure M times, and we have M confidence intervals. Then approximately $(1-\alpha)M$ of these CIs contain θ_0
- For one realization of the CI (L, U), θ_0 is either in this interval or not. No probability for this.



Construct a Confidence Interval

- To construct a CI, we need an estimator and its distribution
- As MLE is asymp. normal dist. with known mean/variance, so MLE is generally used
- Note: other estimators also work, as long as the dist can be found

A popular procedure for CI:

1. Find the MLE $\hat{\theta}$
2. Find the variance $\sigma_n^2 = \text{Var}(\hat{\theta})$. If impossible, then find the fisher information $I(\theta_0)$ and let $\sigma_n^2 = [nI(\theta_0)]^{-1}$.
3. Construct the $100(1-\alpha)\%$ CI for θ as

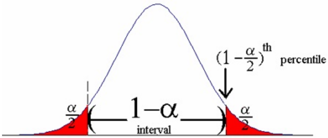
$$(\hat{\theta} - z_{\alpha/2}\sigma_n, \hat{\theta} + z_{\alpha/2}\sigma_n),$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile for standard normal distribution s.t. $\Phi(z_{\alpha/2}) = 1 - \alpha/2$.

Construct a Confidence Interval

- Note that $\hat{\theta} \sim N(\theta_0, \sigma_n^2)$, then

$$\begin{aligned} P(\hat{\theta} - z_{\alpha/2}\sigma_n \le \theta_0 \le \hat{\theta} + z_{\alpha/2}\sigma_n) &= P(-z_{\alpha/2}\sigma_n \le \theta_0 - \hat{\theta} \le z_{\alpha/2}\sigma_n) \\ &= P(-z_{\alpha/2}\sigma_n \le \hat{\theta} - \theta_0 \le z_{\alpha/2}\sigma_n) \\ &= P(-z_{\alpha/2} \le \tfrac{\hat{\theta}-\theta_0}{\sigma_n} \le z_{\alpha/2}) \\ &\approx \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = 1 - \alpha \end{aligned}$$



- When $I(\theta_0)$ or σ_n depend on the unknown θ_0, we use $I(\hat{\theta})$ instead, or the estimated standard error $s_{\hat{\theta}}$
- Popular choices for α are .05, .01
- When n increases, σ_n^2 decreases, and the confidence interval becomes narrower $\Leftrightarrow$ Increase sample size makes the result more precise
- When α increases, z_α decreases, and the confidence interval becomes narrower $\Leftrightarrow$ Narrower interval has smaller confidence level

Example: Confidence Interval – Poisson

Suppose we have an iid sample $X_1, \cdots, X_n$ from Poisson distribution with parameter λ. Build the confidence interval for λ.

Solution: Recall that the MLE for Poisson distribution is $\bar{X}_n$. As we know that $n\bar{X}_n \sim Pois(n\lambda)$, so we can use this derivation directly or use Fisher information.
Here we use fisher information as an example. By recalling $f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$, the log likelihood and the second derivative are:

$$\ln f(x|\lambda) = x \ln \lambda - \lambda - \ln x!$$
$$\frac{\partial^2}{\partial \lambda^2} \ln f(X|\lambda) = -\frac{X}{\lambda^2}$$

Thus,

$$I(\lambda) = \frac{X}{\lambda} = -E(-\frac{X}{\lambda}) = E(\frac{X}{\lambda}) = \frac{E(X)}{\lambda} = \frac{1}{\lambda}$$

The approximated Fisher information is

$$nI(\hat{\lambda}) = \frac{n}{\hat{\lambda}} = \frac{n}{\bar{X}}, \quad \text{where } \hat{\lambda}_{MLE} = \bar{X}$$

and the asymptotic variance $\frac{1}{nI(\hat{\lambda})} = \frac{\bar{X}}{n}$. The $100(1-\alpha)\%$ confidence interval for λ is

$$\bar{X} \pm z_{\alpha/2}\sqrt{\frac{\bar{X}}{n}}$$

Note that the asymptotic variance in this case is actually EXACT variance,

$$\text{Var}(\hat{\lambda}) = \text{Var}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{1}{n^2}(n\lambda) = \frac{\lambda}{n}$$

But the confidence interval is approximate as the sampling distribution of $\bar{X}$ is approximately normal.

A machine fills cups with a liquid, and is supposed to be adjusted so that the content of the cups is 250g of liquid. The content the machine fill every cup is denoted as a r.v. $X \sim N(\mu, 2.5^2)$. To determine if the machine is adequately calibrated, a sample of $n = 25$ cups of liquid are chosen at random and the cups are weighed. The resulting measured masses of liquid are $X_1, \cdots, X_{25}$, a random sample from X with mean 250.2g. What is the 95% confidence interval for μ?

Solution: Obviously, the MLE is $\bar{X}_n = 250.2$, and the sample size $n = 25$. Note that $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n}) = N(\mu, \frac{2.5^2}{25})$, so we have

$$\sigma_n = \sqrt{\mathrm{Var}(\bar{X}_n)} = \sqrt{\frac{2.5^2}{25}} = 0.5.$$

The 95% confidence interval for it is

$$(\bar{X}_n - z_{\alpha/2}\sigma_n, \bar{X}_n + z_{\alpha/2}\sigma_n) = (250.2 - 0.5 * z_{\alpha/2}, 250.2 - 0.5 * z_{\alpha/2}),$$

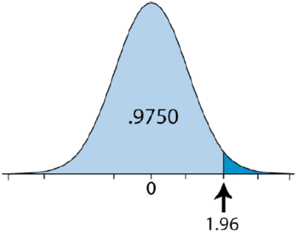
where $\alpha = 1 - 0.95 = 0.05$.

Confidence Interval - a Practical Example

With $\alpha = 0.05$, $z_{\alpha/2} = z_{0.025}$, which is the 0.975 quantile of standard normal distribution.
Check Z-table, and we find $z_{0.025} = 1.96$. So the 95% confidence interval is

$$\begin{aligned}(250.2 - 0.5 * 1.96, 250.2 + 0.5 * 1.96) &= (250.2 - 0.98, 250.2 + 0.98) \\ &= (249.22, 251.18).\end{aligned}$$

	.00	.01	.02	.03	.04	.05	.06
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750



Confidence Interval - A sample mean example

A sample of size $n = 100$ produced the sample mean of $\bar{X} = 16$. Assuming the population standard deviation $\sigma = 3$ (which means the standard deviation for one single observation is 3), compute a 95% confidence interval for the population mean μ.

Solution:
From the central limit theorem, we have $\bar{X} \sim N\left(\mu, \frac{9}{100}\right)$ approximately. Therefore, the confidence interval becomes

$$\left(16 - z_{0.05/2}\frac{3}{10}, 16 + z_{0.05/2}\frac{3}{10}\right) = (15.412, 16.588)$$

where $z_{0.025} = 1.96$.

Assuming the population standard deviation $\sigma = 3$, how large should a sample be to estimate the population mean μ so that the 95% confidence interval has width not exceeding 1?

Solution:
We have $\bar{X} \sim N\left(\mu, \frac{9}{n}\right)$ approximately from CLT, and the 95% confidence interval is $\left(\bar{X} - 1.96\frac{3}{\sqrt{n}}, \bar{X} + 1.96\frac{3}{\sqrt{n}}\right)$. Therefore, the width of the 95% confidence interval is $2 \cdot 1.96 \cdot \frac{3}{\sqrt{n}}$. As we want this width to be smaller that 1, we have

$$\begin{aligned} 2 \cdot 1.96 \cdot \frac{3}{\sqrt{n}} \leq 1 \quad &\Leftrightarrow \quad \sqrt{n} \geq 11.76 \\ &\Leftrightarrow \quad n \geq 138.2976. \end{aligned}$$

Thus the sample should be larger than or equal to 139.