ESE 402/542 :Statistics for Data Science
Instructor: Hamed Hassani
Fall 2020

Midterm Examination

NAME	

Note: Each Multiple Choice (MC) question has 10 points, Problem 1 has 30 points, and Problem 2 has 40 points. For the multiple choice questions, you are only required to write the item that you've chosen (e.g. Question 1: (d) or Question 5: (e)).

Additional Information: If X is distributed according to the Gaussian distribution with mean μ and variance σ^2, then $\mathbb{E}[X^2] = \mu^2 + \sigma^2$.

Also, for $\beta > 0$ and $n > 1$ we have $\int_0^\infty \frac{1}{(x+\beta)^n} dx = \frac{1}{(n-1)\beta^{n-1}}$.

	Grade (y/n)	Score	Max. Score
Multiple Choice			30
Problem 1			30
Problem 2			40
TOTAL			100

10pts **MC-Question 1:** Assume we use the sample mean estimator to estimate the true mean of data that is generated i.i.d. from some distribution. Consider the following statements about the sample mean estimator:

(1) It is always an unbiased estimator of the mean.

(2) It is always a minimum-variance-unbiased-estimator of the mean.

(3) It always achieves the Cramer-Rao bound.

Which of the above is wrong?

(a) 1, 3

(b) 2

(c) 3

[✓ (d)] 2,3

(e) 1,2

10pts **MC-Question 2:** The formula that we derived in class for the confidence interval of the mean (using the sample mean estimator) has the variance of the data in it. However, we oftentimes do not know the true variance of the data. Hence, to fix the formula we need to:

(a) Estimate he variance from data and further incorporate the error due to the estimation of the variance into the formula.

[✓ (b)] Just estimate the variance from data and and use it instead of the true variance.

(c) Use half of the data to estimate the variance and the other half for the confidence interval.

(d) We need to know the true variance; Otherwise, the formula is highly inaccurate.

10pts **MC-Question 3**: Which of the following is correct?

1. The maximum-likelihood estimator is unbiased.
2. The method-of-moments estimator is unbiased.
3. [✓ The maximum-likelihood estimate will become close to the true parameter as the number of samples grows.]
4. The maximum-likelihood estimator always achieves the Cramer-Rao bound.

Problem 1. [30 pts] We have access to a data set $X_1, X_2, \cdots, X_n$ where X_i's are generated i.i.d. according to a distribution with the following pdf:

$$f(x|a) = \frac{2a^2}{(x+a)^3}\mathbf{1}\{x \geq 0\},$$

where the parameter a is known to be positive.

the steps of deriving the integral are not important. only the final outcome is.

15 pts > (a) Use the method of moments to estimate the parameter a from data.

$$\mu_1 = E[X] = \int_0^\infty \frac{2a^2 x}{(x+a)^3}dx = 2a^2\int_0^\infty \frac{x+a-a}{(x+a)^3}dx :$$

$$\Rightarrow \boxed{\mu_1 = a}$$

$$= 2a^2\int_0^\infty \frac{1}{(x+a)^2}dx - 2a^3\int \frac{1}{(x+a)^3}dx = 2a - a = a$$

$$\Rightarrow \hat{a} = \hat{\mu}_1 = \bar{X} \qquad \left(\hat{a} = \frac{X_1 + \cdots + X_n}{n}\right)$$

→ the estimator obtained from the method of moments

15 pts (b) Use the method of maximum likelihood to estimate a from data.

(Note that you can express the outcome of the estimator as the solution of some specific equation.)

$$lik(a) = \prod_{i=1}^n f(X_i|a) = (2a^2)^n \prod_{i=1}^n \frac{1}{(X_i+a)^3}$$

$$\Rightarrow \ell(a) = \log(lik(a)) = n \ln 2a^2 - 3\sum_{i=1}^n \ln(X_i+a)$$

$$\ell'(a) = 0 \Rightarrow \frac{2n}{a} = 3\sum_{i=1}^n \frac{1}{(X_i+a)}$$

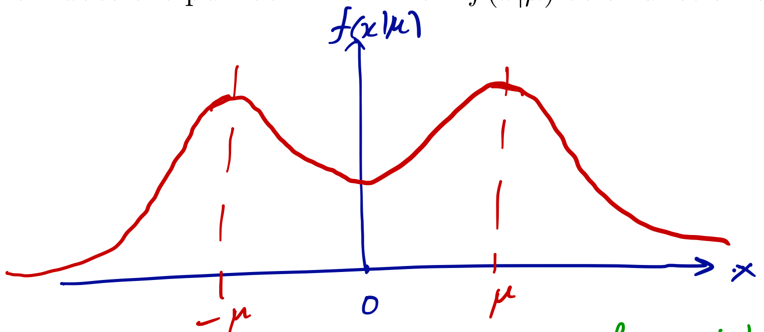
→ $\hat{a}_{ML}$ would be the solution of this equation

Problem 2. [40 pts] We know that μ is a positive number but we do not know its value. We have access to a data set $X_1, X_2, \cdots, X_n$ where X_i's are generated i.i.d. according to a distribution with the following pdf:

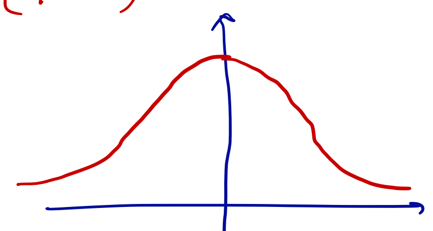
$$f(x|\mu) = \frac{1}{2}\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2}} + \frac{1}{\sqrt{2\pi}}e^{-\frac{(x+\mu)^2}{2}}\right).$$

8 pts (a) How does the pdf look like? Draw $f(x|\mu)$ as a function of x.

$(\mu > 1)$



$(\mu \leq 1)$



(to get the full credit, either the case for $\mu > 1$ should be given or both cases for $\mu \leq 1$ and $\mu > 1$ should be plotted.

8pts (b) Find the mean and variance of the distribution in terms of μ. Note that you don't really need to do any integration here. To compute the variance, pay attention to the fact that the pdf is the average of two Gaussian pdfs: one with mean μ and the other with mean $-\mu$.

if $X \sim f(x|\mu)$, the $E[X] = 0$ (because the pdf is symmetric).

$$\mathrm{var}(X) = E[X^2] = \frac{1}{2}\int_{-\infty}^{\infty} x^2 \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}} + \frac{1}{\sqrt{2\pi}} e^{\frac{(x+\mu)^2}{2}}\right)dx = \frac{1}{2}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x+\mu)^2}{2}} dx + \frac{1}{2}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2}}$$

$$= \frac{1}{2}\left(\mu^2+1 + \mu^2+1\right) = \mu^2+1$$

8pts (c) Design an <u>unbiased</u> estimator for the parameter $\theta = \mu^2$ using the sample data $X_1, \cdots, X_n$.

given part (b), we simply have: $\theta = E[X^2] - 1$, hence:

$$\hat{\theta} = \frac{X_1^2 + \dots + X_n^2}{n} - 1$$

Since X_i's are iid

8pts (d) Compute the variance of your estimator.

$$Var(\hat{\theta}) = Var\left(\frac{X_1^2+\cdots+X_n^2}{n} - 1\right) = Var\left(\frac{\sum X_i^2 - 1}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i^2-1) = \frac{Var(X^2-1)}{n}$$

$$Var(X^2-1) = E[(X^2-1)^2] - (E[X^2-1])^2 = E[X^4-2X^2+1] - \mu^4$$

$$= E[X^4] - 2E[X^2] + 1 - \mu^4 = E[X^4] - 2(1+\mu^2) + 1 - \mu^4$$

$$= E[X^4] - \mu^4 - 2\mu^2 - 1 = \mu^4 + 6\mu^2 + 3 - 2\mu^2 - 1 - \mu^4 = 4\mu^2+2 \Rightarrow \boxed{Var(\hat{\theta}) = \frac{4\mu^2+2}{n}}$$

$$\left(E[X^4] = \frac{1}{2}\int_{-\infty}^{\infty} \frac{x^4 e^{-\frac{(x-\mu)^2}{2}}}{\sqrt{2\pi}}dx + \frac{1}{2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}} x^4 e^{-\frac{(x+\mu)^2}{2}}dx = \mu^4 + 6\mu^2 + 3 \right)$$

→ I gave a hint about this in Piazza

8pts (e) Provide a 95% confidence interval for the parameter θ using your estimator and the data sample (assume that the number of data points, n, is large).

$$\left[\hat{\theta} - z_{\alpha/2}\sqrt{\frac{4\mu^2+2}{n}}, \hat{\theta} + z_{\alpha/2}\sqrt{\frac{4\mu^2+2}{n}}\right] \quad \text{where } \alpha = 0.05$$

we can also estimate μ^2 from data (e.g. μ^2 in the above formula can be replaced with $\hat{\theta}$).

(the confidence interval (CI) will get the full credit).

(f) Can you comment on how good your estimator is? i.e. do you think you might be able to find another estimator with smaller variance?

Omitted, but if the connection to the Cramer-Rao bound and computation of the Fisher Information is mentioned, the +2 marks should be given as extra credit.

(so one needs to computed $\frac{1}{nI(\theta)}$ and compare it with the variance computed in part (d). But computing $I(\theta)$ is very difficult).