ESE 402/542: Statistics for Data Science
Instructor: Hamed Hassani
Fall 2021

Midterm Examination

NAME	

Additional Information:

- The pdf of a Gaussian, $\mathcal{N}(\mu, \sigma^2)$, is $f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- $\int_{-\infty}^{\infty} \frac{x^2}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)}{2\sigma^2}} dx = \mu^2 + \sigma^2$, $\int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}\sigma}, e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu$,
  $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}, e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1$
- Linearity of expectation is your friend.
- $\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2$, i.e. $\sigma^2 = \mathbb{E}\left[X^2\right] - \mu^2$.
- If X continuous ($p(x)$ is its pdf), $\mathbb{E}[g(X)] = \int g(x)p(x)dx$
- If X discrete ($p(x)$ is its pmf), $\mathbb{E}[g(X)] = \sum_x g(x)p(x)$

	Grade (y/n)	Score	Max. Score
Problem 1			50
Problem 2			50
TOTAL			100

Problem 1. [50 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:

$$f(x|a,p) = p \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(x+a)^2}{2}} + (1-p) \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2}}.$$

where the parameters p is known to be between 0 and 1.

In the following, for parts (a)-(f) we assume that $a = 1$ is given, but the value of p is to be estimated from data.

(a) Draw the pdf $f(x|a,p)$ as a function of x for the case $p = 1/4$.

(b) Find the variance of X_i in terms of p.

(c) Use the method of moments to estimate the parameter p from data. Let's denote this estimator by $\hat{p}$.

(d) Is $\hat{p}$ an unbiased estimator?

(e) Let $\mu = \mathbb{E}[X_1]$ and also let $\hat{\mu}$ be the empirical mean of the data (i.e. $\hat{\mu} = (X_1 + \cdots + X_n)/n$). For any $\beta > 0$, find β' such that the following holds:

$$\Pr\big(\mu \in [\hat{\mu} - \beta, \hat{\mu} + \beta]\big) = \Pr\big(p \in [\hat{p} - \beta', \hat{p} + \beta']\big).$$

(f) Use part (e) to find the $1 - \alpha$ confidence interval for p using the estimate $\hat{p}$.

(g) Let us now assume that the value of a is also unknown (in addition to p). Use the method of moments to estimate both a and p from data. Assume $a \geq 0$. (Hint: treat this as solving a system of equations with two variables.)

Problem 2. [50 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:

$$f(x|\sigma) = \frac{1}{2\sigma} e^{-\frac{|x|}{\sigma}}. \tag{1}$$

We are given that $\mathbb{E}[X_i] = 0$, $\mathbb{E}[|X_i|] = \sigma$, and $\mathrm{Var}(X_i) = 2\sigma^2$.

We consider a hypothesis testing problem with $H_0 : \sigma = \sigma_0$ and $H_a : \sigma = \sigma_1$. For this setting, we consider the following test statistic:

$$T(X_1, X_2, \cdots, X_n) = \frac{1}{n} \log f(X_1, X_2 \cdots, X_n|\sigma_0) - \frac{1}{n} \log f(X_1, X_2 \cdots, X_n|\sigma_1),$$

where $f(X_1, X_2 \cdots, X_n|\sigma_0)$ is the joint density of $X_1, \cdots, X_n$ given $\sigma = \sigma_0$ (and the other term is defined similarly). You may assume that $\sigma_0 < \sigma_1$.

(a) Explain why

$$f(X_1, X_2 \cdots, X_n|\sigma_0) = f(X_1|\sigma_0) \times f(X_2|\sigma_0) \times \cdots \times f(X_n|\sigma_0).$$

(b) Using part (a) and (1) expand and simplify the term $\frac{1}{n} \log f(X_1, X_2 \cdots, X_n|\sigma_0)$ as much as you can.

(c) Derive an approximate formula for the distribution of $T(X_1, \cdots, X_n)$ in the case that H_0 is the true hypothesis.

(d) Given a significance level α, design the acceptance/rejection regions for the above hypothesis testing problem and test statistic T. (Remember: $\sigma_0 < \sigma_1$).