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

Midterm Examination SOLUTIONS

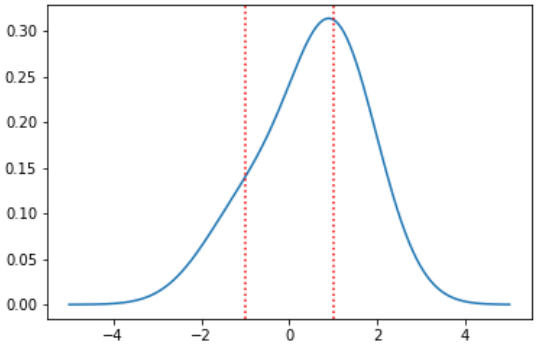
Problem 1. [50 pts] We have access to a data set $X_1, X_2, \cdots, X_n$ where X_i's are generated i.i.d. according to a distribution with the following pdf:

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

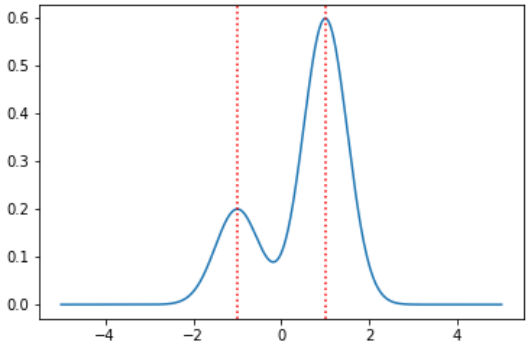
where the parameters p is known to be between 0 and 1.

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 $p = 1/4$.



Accept answers that look like



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

We first observe that the pdf $f(x)$ is a combination of two gaussian pdfs corresponding

to $\mathcal{N}(-a,1)$ and $\mathcal{N}(a,1)$. We recall the formula $\mathrm{var}(X) = \mathbb{E}[X^2] + \mathbb{E}[X]^2$. We evaluate

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x \left(p \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(x+a)^2}{2}} + (1-p) \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2}} \right) dx \\
&= p \int x \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x+a)^2}{2}} \right) dx + (1-p) \int x \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2}} \right) dx \\
&= p\mathbb{E}_{\mathcal{N}(-a,1)}[X] + (1-p)\mathbb{E}_{\mathcal{N}(a,1)}[X] \\
&= p(-a) + (1-p)a \\
&= a(1-2p).
\end{aligned}$$

When $a = 1$, we have $\mathbb{E}[X] = 1 - 2p$. Similarly, we write

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &= \int_{-\infty}^{\infty} x^2 \left(p \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(x+a)^2}{2}} + (1-p) \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2}} \right) dx \\
&= p \int x^2 \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x+a)^2}{2}} \right) dx + (1-p) \int x^2 \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2}} \right) dx \\
&= p\mathbb{E}_{\mathcal{N}(-a,1)}[X^2] + (1-p)\mathbb{E}_{\mathcal{N}(a,1)}[X^2] \\
&= p(1+a^2) + (1-p)(1+a^2) \\
&= 1 + a^2,
\end{aligned}$$

where we used the provided hint that $\mathbb{E}[X^2] = \sigma^2 + \mu^2 = 1 + a^2$ for normal distributions $\mathcal{N}(\pm a, 1)$. Therefore, when $a = 1$, the variance is

$$\mathrm{var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = 2 - (1-2p)^2$$

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

Recall that when $a = 1$, we derived the mean of X as

$$\mathbb{E}[X] = 1 - 2p \iff p = \frac{1 - \mathbb{E}[X]}{2}.$$

Therefore, we can write the method of moments estimator of p as

$$\hat{p}_{\mathrm{MoM}} = \frac{1 - \bar{X}}{2}.$$

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

We could immediately observe that since $\hat{p}_{\mathrm{MoM}}$ is a *linear* function of $\bar{X}$, it will be an unbiased estimator of p. In other words

$$\begin{aligned}
\mathbb{E}[\hat{p}_{\mathrm{MoM}}] &= \mathbb{E}\left[\frac{1 - \bar{X}}{2}\right] \\
&= \frac{1 - \mathbb{E}[X]}{2} = p.
\end{aligned}$$

(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:

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

We recall $\mu = 1 - 2p$, $p = \frac{1-\mu}{2}$. Writing

$$\begin{aligned}
\mu \in [\hat{\mu} - \beta, \hat{\mu} + \beta] &\iff \hat{\mu} - \beta \le \mu \le \hat{\mu} + \beta \\
&\iff \frac{1 - (\hat{\mu} + \beta)}{2} \le \frac{1 - \mu}{2} \le \frac{1 - (\hat{\mu} - \beta)}{2} \\
&\iff \frac{1 - \hat{\mu}}{2} - \frac{\beta}{2} \le p \le \frac{1 - \hat{\mu}}{2} + \frac{\beta}{2}.
\end{aligned}$$

We recall from the previous two parts that $\hat{p} = \frac{1-\hat{\mu}}{2}$, and therefore

$$\Pr\big(\mu \in [\hat{\mu} - \beta, \hat{\mu} + \beta]\big) = \Pr\big(p \in [\hat{p} - \beta/2, \hat{p} + \beta/2]\big).$$

This implies $\beta' = \beta/2$.

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

If we wanted to find a $1-\alpha$ confidence interval for $\hat{\mu}$, by CLT we would set $\beta = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, where

$$\sigma^2 = \operatorname{var}(X) = 2 - (1 - 2p)^2,$$

such that

$$\beta = z_{\alpha/2}\sqrt{\frac{2 - (1 - 2p)^2}{n}}.$$

From the previous part, we know $\beta' = \beta/2$ and thus the $1 - \alpha$ confidence interval for p using $\hat{p}$ is

$$\left[\hat{p} - z_{\alpha/2}\frac{\sqrt{2 - (1 - 2p)^2}}{2\sqrt{n}}, \hat{p} + z_{\alpha/2}\frac{\sqrt{2 - (1 - 2p)^2}}{2\sqrt{n}}\right].$$

In other words, there is a $1 - \alpha$ probability that p is contained in the aforementioned confidence interval.

Accept answers that use $\hat{p}$ in place of p in the variance, complete with Bessel's correction (justification unnecessary). Make note if Bessel's correction not present (no deduction).

(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.)

Observe from the formulas for the first and second moments derived in part b:

$$\begin{aligned} \mathbb{E}[X] &= a(1-2p) \\ \mathbb{E}\left[X^2\right] &= 1 + a^2. \end{aligned}$$

From the second equation, we can write $a^2 = \mathbb{E}[X^2] - 1$, $a = \sqrt{\mathbb{E}[X^2] - 1}$ (since a assumed non-negative). Plugging this into the first equation, we have

$$\begin{aligned} &\mathbb{E}[X] = \sqrt{\mathbb{E}[X^2] - 1}(1-2p) \\ \iff\ &p = \frac{1}{2} - \frac{\mathbb{E}[X]}{2\sqrt{\mathbb{E}[X^2] - 1}}. \end{aligned}$$

Thus, denoting $\hat{\mu}$ as earlier and $\hat{M} = \frac{1}{2}(X_1^2 + \cdots + X_n^2)$, our method of moments estimates for a, p are

$$\begin{aligned} \hat{a} &= \sqrt{\hat{M} - 1} \\ \hat{p} &= \frac{1}{2} - \frac{\hat{\mu}}{2\sqrt{\hat{M} - 1}}. \end{aligned}$$

Accept answers that only derive formulas for a, p.

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).

For this problem, assume $\sigma_0 < \sigma_1$.

(a) Explain why

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

$X_1, X_2, \ldots, X_2$ are independent, and thus their joint density is the product of their individual densities.

(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.

Using part a, we can write

$$\begin{aligned}
\frac{1}{n} \log f(X_1, \ldots, X_n|\sigma_0) &= \frac{1}{n} \sum_{i=1}^{n} \log f(X_i|\sigma_0) \\
&= \frac{1}{n} \sum_{i=1}^{n} (-\log(2\sigma_0) - |X_i|/\sigma_0) \\
&= -\log(2\sigma_0) - \sigma_0^{-1} \frac{1}{n} \sum_{i=1}^{n} |X_i|
\end{aligned}$$

(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.

Adapting the formula from part b, we write

$$\begin{aligned}
T(X_1, X_2, \cdots, X_n) &= \frac{1}{n} \log f(X_1, X_2 \cdots, X_n|\sigma_0) - \frac{1}{n} \log f(X_1, X_2 \cdots, X_n|\sigma_1) \\
&= (\log(2\sigma_1) - \log(2\sigma_0)) + (\sigma_1^{-1} - \sigma_0^{-1}) \frac{1}{n} \sum_{i=1}^{n} |X_i|.
\end{aligned}$$

The only part that depends on data is $\sum |X_i|$. Under the null hypothesis, we know that $\mathbb{E}[|X|] = \sigma_0$. To compute the variance $\text{var}(|X|)$, we observe that since $\mathbb{E}[X] = 0$,

$$\begin{aligned}\text{var}(X) &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \\ &= \mathbb{E}\left[|X|^2\right] - 0 = 2\sigma_0^2.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\text{var}(|X|) &= \mathbb{E}\left[|X|^2\right] - \mathbb{E}[|X|]^2 \\ &= 2\sigma_0^2 - \sigma_0^2 = \sigma_0^2.\end{aligned}$$

We may now apply the CLT to claim that $\frac{1}{n}\sum_{i=1}^n |X_i|$ approximately follows the normal distribution $\mathcal{N}\left(\sigma_0, \frac{\sigma_0^2}{n}\right)$. Using the properties of gaussian distributions, this tells us that

$$T(X_1, X_2, \cdots, X_n) = (\log(2\sigma_1) - \log(2\sigma_0)) + (\sigma_1^{-1} - \sigma_0^{-1})\frac{1}{n}\sum_{i=1}^n |X_i|$$

is approximately normally distributed with mean

$$(\log(2\sigma_1) - \log(2\sigma_0)) + (\sigma_1^{-1} - \sigma_0^{-1})\sigma_0 = (\log(2\sigma_1) - \log(2\sigma_0)) + \left(\frac{\sigma_0}{\sigma_1} - 1\right)$$

and variance

$$(\sigma_1^{-1} - \sigma_0^{-1})^2\frac{\sigma_0^2}{n}$$

(d) Given a significance level α, design the acceptance/rejection regions for the above hypothesis testing problem and test statistic T.

From the previous part, we determined that $T(X_1, \ldots, X_n)$ is approximately normally distributed with a certain mean μ and variance s^2, under the assumption of H_0.

We observe that in the case $\sigma_0 < \sigma_1$, $\sigma_1^{-1} - \sigma_0^{-1} < 0$, and thus $T(X_1, \cdots, X_n)$ is smaller when $\sum_{i=1}^n |X_i|$ is larger, which is more likely to happen when X_i follow the alternate hypothesis. Therefore, the rejection region is one-sided of the form $\{T(X_1, \ldots, X_n) < C_\alpha\}$. From the previous part, since T is asymptotically normal, C_α takes form $C_\alpha = \mu - z_\alpha s$, where μ and s^2 are the corresponding mean and variance of the asymptotic normal distribution. This ensures that

$$\mathbb{P}[\mu - z_\alpha s \leq T(X_1, \ldots, X_n)] = 1 - \alpha.$$

Plugging in the values we derived in the previous part, we get the acceptance region

$$\left[(\log(2\sigma_1) - \log(2\sigma_0)) + \left(\frac{\sigma_0}{\sigma_1} - 1\right) - z_\alpha\left|\sigma_1^{-1} - \sigma_0^{-1}\right|\frac{\sigma_0}{\sqrt{n}}\,, \quad \infty\right].$$