Problems 2 and 3 are past HW questions.

Hyperlinks

- True/false questions
- Medical testing
- DNA sequencing
- MLE and MoM comparison
- Gaussian mixture model

Problem 1 (True/False questions) {#true}

> 1. There exists an estimator $\hat{\theta}$ for which $\text{MSE}(\hat{\theta}) = (\text{Bias}(\hat{\theta}))^2$
> 2. If $\hat{\theta}$ is unbiased, then all other estimators are biased.
> 3. The sample mean is unbiased under all models where a mean exists.
> 4. The squared error loss of an estimator for its estimand is a random variable.
> 5. When data comes from a discrete distribution, the likelihood function is also discrete.

1. True. You can choose any zero-variance constant estimator.
2. False. For example, for X_1, X_2 i.i.d, using the first observation, the second observation, or $\bar{X}$ are all unbiased.
3. Partially false. It is unbiased for the mean but not necessarily other estimands.
4. True. The expression for SEL involves the estimator, an r.v.
5. False. Consider the likelihood function of a binomial distribution, which can take inputs from the real between 0 and 1.

Problem 2 (Medical testing) {#medtest}

This problem will be a review of conditioning

We are trying to estimate what proportion of the Harvard body (staff, grad students, and undergraduates) have some unspecified pandemic illness. The administration conducts frequent tests to determine the prevalence and incidence of infection.

(a) Perfect test

> Assume that the test is perfect and that whether a person tests positive is distributed i.i.d. Bernoulli with proportion p.
>
> Determine the MLE and produce an estimate of p given the observations. Assume the data is in the vector $\vec{x}$ of length n.
>
> *Note: Why is i.i.d. Bernoulli a poor choice of model here?**

Likelihood function:

$$L(p \mid \vec{x}) = \prod_{j=1}^{n} p^{x_j}(1-p)^{1-x_j} = p^k(1-p)^{n-k},$$

where k is $\sum_{j=1}^{n} x_j$.

Log-likelihood:

$$\ell(p \mid \vec{x}) = k \ln p + (n-k)\ln(1-p)$$

MLE after setting the derivative of the above to 0:

$$\frac{\partial \ell(p \mid \vec{X})}{\partial p} = \frac{k}{p} + \frac{k-n}{1-p} = 0$$

We can then solve for p to get the MLE

$$-k + pk = pk - np \to \hat{p} = \frac{k}{n},\ k = \sum_{j=1}^{n} X_j.$$

Note that we can observe this is a maximum because of the second partial derivative (work left to reader).

(b) Imperfect test

> The **sensitivity** of a test is the probability that someone who has the condition tests positive.
>
> The specificity of a test is the probability that someone who does not have the condition tests negative.
>
> Let sensitivity be a and specificity by b. Correct the likelihood function to accommodate for error in the test.

Calculate the probability someone tests positive $P(T)$ with the probability that someone has the disease, $P(D)$.

$$P(T) = P(T \mid D)P(D) + P(T \mid D^c)P(D^c) = ap + (1-b)(1-p) = 1 + p(a+b-1) - b.$$

Replace p in the likelihood function with $P(T)$.

$$L(p \mid \vec{x}) = (1 + p(a+b-1) - b)^k (b - p(a+b-1))^{n-k}.$$

Problem 3 (DNA sequence) {#dna}

A DNA sequence can be composed of four different possible base pairs. For example, consider the following sequence:

CTACCTTCAATTGCTGGAACG

(a) Multinomial model

> For simplicity, assume that DNA base pairs are selected from a Multinomial distribution (ignore other properties of DNA base pairs).
>
> Let the probabilities corresponding to the base pair selection be represented as $\theta = (p_a, p_c, p_g, p_t)$.
>
> Write the log-likelihood of θ.

Let W_x be the count of base-pairs with outcome x. We have that

$$(W_a, W_c, W_g, W_t) \sim \mathrm{Multinomial}(n, \theta).$$

We have the following expressions given observations w_x (and $x \in \{a, c, g, t\}$):

$$L(\theta) = \prod_{i=1}^{n} \prod_{x} p_x^{I_{x_i = x}} = \prod_{x} p_x^{w_x},$$

and log-likelihood

$$\ell(\theta) = \sum_{x} w_x \log p_x.$$

(b) Markov chain

> We can use a Markov chain model to accommodate for possible violations of independence in the genome. Let us model the sequence as a lag-1 Markov model. In a lag-1 Markov model, we have $\Pr(X_s \mid X_{s-1}, \ldots, X_1) = \Pr(X_s \mid X_{s-1})$. This model can be represented with a 4×4 transition matrix:

$$T = \begin{pmatrix} \tau_{aa} & \tau_{ac} & \tau_{ag} & \tau_{at} \\ \tau_{ca} & \tau_{cc} & \tau_{cg} & \tau_{ct} \\ \tau_{ga} & \tau_{gc} & \tau_{gg} & \tau_{gt} \\ \tau_{ta} & \tau_{tc} & \tau_{tg} & \tau_{tt} \end{pmatrix}$$

> How many parameters does this matrix have?

Only 12 because the last entry in a row is determined by the previous three (or, in any situation, you only need three of the entries of a row to determine the fourth). By definition of a Markov chain, the entries in a row need to add to 1.

> Assume that the marginal distribution $\Pr(X_1 = x), x \in \{t, c, g, a\}$ is known. Write the log-likelihood of this model.

The joint likelihood of the data (stored in $\vec{x}$)

$$f_{\vec{X}}(\vec{x} \mid T) = f(x_1 \mid T) \prod_{i=2}^{10} f(x_t \mid x_{t=1}, T) = P(X_1 = x_1 \mid T) \prod_{i=2}^{10} P(X_i = x_i \mid X_{i=1} = x_{i=1}, T)$$

where we have

$$P(X_i = y \mid X_{i=1} = x, T) = \tau_{xy}.$$

Recall that we assume we know $f(x_1 \mid T) = P(X_1 = x_1 \mid T)$.

Let W_{xy} be the number of observations where the base pair x is followed by y.

We then have the log-likelihood

$$\ell(T \mid \vec{x}) = \sum_{x,y} w_{xy} \log \tau_{xy} : x, y = a, c, g, t.$$

This is constrained so that each row of T sums to 1.

The summarizing statistic is the number of each type of base pair change.

Problem 4 (MLE and MoM) {#mlemom}

Emily and her mom buy a chihuahua from the pet store. The storekeeper, a statistician, mentions that Chihuahua weights can be reasonably modeled as i.i.d. Normal with a standard deviation of one pound. Unfortunately, the storekeeper forgot the average weight of a chihuahua.

(a) MLE

> Based on a single observation, y, of her new dog's weight, Emily estimates the mean of the data-generating distribution via maximum likelihood. What estimator does she use? How would Emily estimate $\theta = \mu^2$, where μ is the mean?

The MLE is the single observation, $\hat{\mu}_{MLE} = Y$. By invariance, $\hat{\theta}_{MLE} = Y^2$.

(b) MoM

> Emily's mom prefers the method of moments. Using the first and second moments respectively, she derives estimators for the same two estimands that Emily estimated. What are Mom's estimators?

The MoM is the same as the MLE, $\hat{\mu}_{MoM} = Y$. We can use the variance to define the MoM for the second moment:

$$\begin{aligned}
\mathbb{E}[Y^2] &= \mathrm{Var}(Y) + (\mathbb{E}Y)^2 \\
&= 1 + \mu^2 \\
&= 1 + \theta \\
\hat{\theta}_{MoM} &= Y^2 - 1.
\end{aligned}$$

(c) MSE

> For any estimators that disagree, determine whether Emily or Mom gave the estimator with lower MSE? Can you determine this without actually computing the MSE?

Note that the MSE for the first moment estimators are identical. In the estimator of the second moment, we see that the MoM estimator is unbiased while the variance is the same, so the MSE is lower.

Problem 5 (Gaussian mixture model) {#mixture}

Suppose $Y_1, \ldots, Y_n$ are drawn i.i.d. from a Gaussian mixture model, with $n > 1$. In particular, the data come from a standard Normal distribution with probability $\frac{1}{2}$ and otherwise come from a Normal distribution with unknown mean and variance, $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ respectively.

(a) Likelihood calculations

> What is the likelihood? What is the log-likelihood?

Note that with LOTP, the probability of either case explaining an observation has the same weighted proportion. So, writing the likelihood, we can remove all the multiplicative scalars and write

$$L(\mu, \sigma^2) = \prod_{j=1}^{n} \left(e^{-Y_j^2/2} + \frac{1}{\sigma} e^{-(Y_j-\mu)^2/(2\sigma^2)} \right)$$

and log-likelihood

$$\ell(\mu, \sigma^2) = \sum_{j=1}^{n} \ln \left(e^{-Y_j^2/2} + \frac{1}{\sigma} e^{-(Y_j-\mu)^2/(2\sigma^2)} \right).$$

(b) Arbitrary likelihood

> Show that the likelihood can be made arbitrarily large. Hint: Set $\mu = Y_1$.

Observe that if we follow the hint, we have

$$\ell(\mu, \sigma^2) = \ln \left(e^{-Y_1^2/2} + \frac{1}{\sigma} \right) + \sum_{j=2}^{n} \ln \left(e^{-Y_j^2/2} + \frac{1}{\sigma} e^{-(Y_j-\mu)^2/(2\sigma^2)} \right).$$

By sending σ toward zero, notice the LH term becomes arbitrarily large. We also see

$$\lim_{\sigma \to 0} \ln \left(e^{-Y_j^2/2} + \frac{1}{\sigma} e^{-(Y_j-\mu)^2/(2\sigma^2)} \right) = -Y_j^2/2.$$

To see this, note that the RH term tends to zero. We have

$$\frac{1}{\sigma} e^{-(Y_j-\mu)^2/(2\sigma^2)} \to \frac{1}{\sigma} \frac{1}{e^{1/\sigma^2}} \to 0.$$

Thus, the likelihood contribution from the first data point can be driven arbitrarily high and outpaces the cost of explaining the other variables.

For an intuition of this, we can think of how the PDF of the Normal changes when σ shrinks. We can create a peak at the one value we observe that corresponds to μ and drive the "height" of the PDF at that point arbitrarily high while still having the PDF integrate to 1. Because of the known Normal(0,1) in the mixture, the other points have a fixed likelihood of being generated from the other part of the mixture and will not affect this runaway increase.

(c) Non-arbitrary likelihood

> Why would we not achieve arbitrarily large likelihood using the same technique if the model were $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$?

If we look at the log-likelihood of a non-mixture Gaussian, we have

$$\ell(\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{j=1}^{n} (Y_j - \mu)^2,$$

so we see that taking an arbitrarily low σ^2 will lower the likelihood.