DAI Assignment 2

Ansh Garg - 24B1066
Kashyap Khandelwal - 24B0932
Pranav Patil - 24B1074

August 2025

Question 1

a) i) In round 1, all pools will have to be tested. Thus n/s tests required. In round 2, a pool is further tested if it is . positive

$$P(\text{pool is positive}) = 1 - P(\text{Pool is negative})$$

For a pool to be negative, each member in the pool should be healthy. Probability for this is $(1-p)^s$. Thus, probability that the pool is positive is $1-(1-p)^s$. Number of tests for such a pool is s (Each person in the pool will have to be tested). The above holds for all n/s pool.s Thus, in round 2, expected number of tests is:

$$\sum P(\text{pool is positive}) \cdot s = \frac{n}{s} \cdot (1-(1-p)^s) \cdot s$$

Thus the expected total number of tests is $T(s) = n/s + n(1-(1-p)^s)$.
ii) If p is very small, $T(s) = n/s + nps$.
To find minimum of $T(s)$, differentiate with respect to s and put $T'(s) = 0$. Thus,

$$T'(s) = -n/s^2 + np = 0$$

So, $s = 1/\sqrt{p}$.
The expected number of tests, $T(1/\sqrt{p}) = 2n\sqrt{p}$.
iii) Putting $T(s) < n$,

$$1/s + 1 - (1-p)^s < 1$$

$$1/s < (1-p)^s$$

$$(1/s)^{1/s} < 1 - p$$

$$p < 1 - (1/s)^{1/s}$$

Thus $p_{max} = 1 - (1/s)^{1/s}$.

b) i) Given, a healthy person, the probability that it participates in a pool T_{i_k} is π. For a pool to be negative, it must not have any unhealthy person. For each such np people, the probability of they not being in the pool is $1-\pi$. Thus, the probability that a pool test negative is $(1-\pi)^{np}$. Thus, the required probability is just the product of these two, i.e $\pi \cdot (1-\pi)^{np}$.
ii) To find maximum of this, put $f'(\pi) = 0$.

$$(1-\pi)^{np} - \pi \cdot np \cdot (1-\pi)^{np-1} = 0$$

So, $\pi = 1/(np+1)$.
iii) Probability that all pools that a genuinely healthy person participates, in tests positive is complement of the event that it participates in some pool that tests negative. Thus, the probability is $1-\pi\cdot(1-\pi)^{np}$. This must be true for all T_1 pools. Thus, probability becomes $\{1-\pi\cdot(1-\pi)^{np})\}^{T_1}$.

Substituting our optimal π, we get, $\{1 - \frac{1}{np+1} \cdot (\frac{np}{np+1})^{np}\}^{T_1}$. Using approximation, $(1-x)^a = e^{-xa}$ for small $x > 0$, we get the probability is, $(1 - \frac{e^{-np/(np+1)}}{np+1})^{T_1}$.

iv) In round 1, number of tests is T_1.

In round 2, let us check for genuinely positive and genuinely negative persons separately.

For a healthy person to be tested again, all pools that it participated in must have tested positive, which has probability $\{1 - \pi \cdot (1-\pi)^{np})\}^{T_1}$.

For an infected person, the probability that the pool it participates in tests positive is 1. The probability that the person is in some pool is $1 - (1-\pi)^{T_1}$

Thus, the expected total number of tests is

$$T_1 + np[1 - (1-\pi)^{T_1}] + n(1-p) \cdot [1 - \pi(1-\pi)^{np}]^{T_1}$$

v) Using approximation $(1-x)^a = e^{-xa}$ for small $x > 0$, and using optimal value of $\pi = 1/np+1$, we can approximate $(1-\pi)^{np} = (np/np+1)^{np} \to 1$. Thus, expression simplifies to

$$E[N_T] = T_1 + np(1 - e^{-\pi T_1}) + n(1-p)(1-\pi)^{T_1}$$

Differentiating and setting it to 0, we get,

$$1 + np\pi e^{-\pi T_1} - n(1-p)\pi e^{-\pi T_1}$$

Thus, optimal T_1 is:

$$T_1 = \frac{\ln(n(1-2p)\pi)}{\pi}$$

Substituting in our formula, we get,

$$E[N_T] = T_1 + np + \frac{1}{\pi}$$

Substitute optimal value of π,

$$(np+1)\ln(\frac{n(1-2p)}{np+1}) + 2np + 1$$

c) Following are the plots obtained:

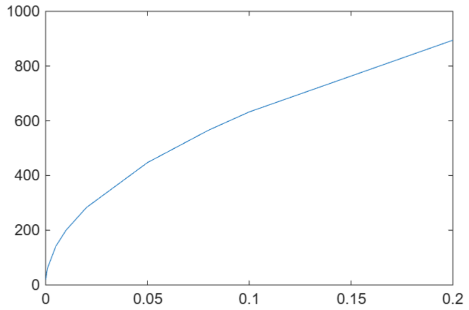


Figure 1: Method 1

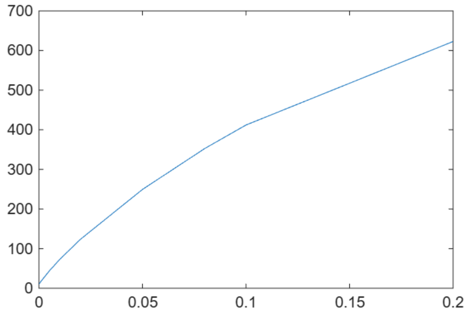


Figure 2: Method 2

Clearly, method 2 is far better. For all values of p used in the plit expected number of tests in the second case comes out to be less than that for the first one.
However, also note that as the prevalance of the disease increases, both methods turn out to be ineffective.

Question 2

We want to calculate the PDF of Z. We first calculate its CDF and then differentiate it to get the PDF.
We know that the probability that $a \leq X \leq a + da$ is $f_X(a)\ da$ where $f_X(\cdot)$ is the pdf of X.
$P(XY \leq z)$ is the CDF of Z. Here, we consider two cases.

- $a > 0$:
  Here $Y \leq \frac{z}{a}$, the probability of which is $F_Y\left(\frac{z}{a}\right)$. We multiply this with the probability of X being in the range and integrate over a to get F_Z:

$$F_{Z_1}(z) = \int_0^\infty F_Y\left(\tfrac{z}{a}\right) f_X(a)\, da.$$

- $a < 0$:
  Here $Y \geq \frac{z}{a}$, the probability of which is $1 - F_Y\left(\frac{z}{a}\right)$. We multiply this with the probability of X being in the range and integrate over a to get F_Z:

$$F_{Z_2}(z) = \int_{-\infty}^0 \left(1 - F_Y\left(\tfrac{z}{a}\right)\right) f_X(a)\, da.$$

Now, $F_Z(z) = F_{Z_1}(z) + F_{Z_2}(z)$. Since $f_Z(z) = \frac{dF_Z(z)}{dz}$, we get,

$$F_Z(z) = \int_{-\infty}^0 \left(1 - F_Y\left(\tfrac{z}{a}\right)\right) f_X(a)\, da + \int_0^\infty F_Y\left(\tfrac{z}{a}\right) f_X(a)\, da$$

Differentiating the above with respect to z, we get,

$$f_Z(z) = \int_{-\infty}^{0} \left(-\tfrac{1}{a}\right) f_Y\left(\tfrac{z}{a}\right) f_X(a)\, da + \int_{0}^{\infty} \left(\tfrac{1}{a}\right) f_Y\left(\tfrac{z}{a}\right) f_X(a)\, da$$

Simplifying,

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{|a|} f_Y\left(\tfrac{z}{a}\right) f_X(a)\, da$$

Question 3

We should consider the sample mean as the estimate for $E[X]$.
The estimator

$$\hat{x} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i$$

and therefore by linearity of expectation

$$E[\hat{x}] \;=\; E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \;=\; \frac{1}{n}\sum_{i=1}^{n} E[x_i]$$

Since each sample x_i is drawn from the same distribution as X, we have $E[x_i] = E[X]$ for every i. Hence

$$E[\hat{x}] \;=\; \frac{n\,E[X]}{n} \;=\; E[X]$$

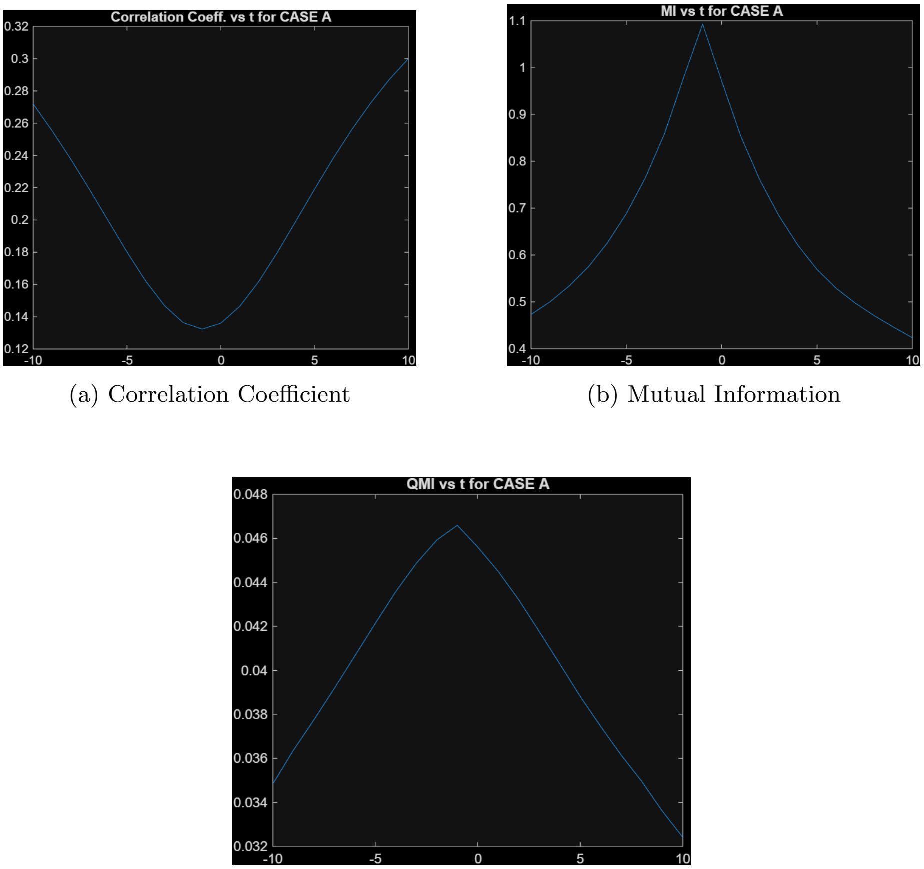
Thus the sample mean is an unbiased estimator of $E[X]$ (Note that according to the Law of Large Numbers, this converges to $E[X]$ as $n \to \infty$).
Why the other option is incorrect:

$$\hat{x} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \cdot f_X(x_i)$$

Thus the second option does not estimate $E[X]$ but rather the the expectation of some other random variable, $Y = Xf_X(X)$.

Question 4

CASE A
Measures of dependence between first image and shifted version of second image.
All three plots, ρ, QMI, MI have their corresponding Maxima/Minima at t = -1, which suggests that the images, T1.jpg and T2.jpg aren't perfectly aligned, T2 is originally off by 1 pixel to the right relative to T1.



(a) Correlation Coefficient

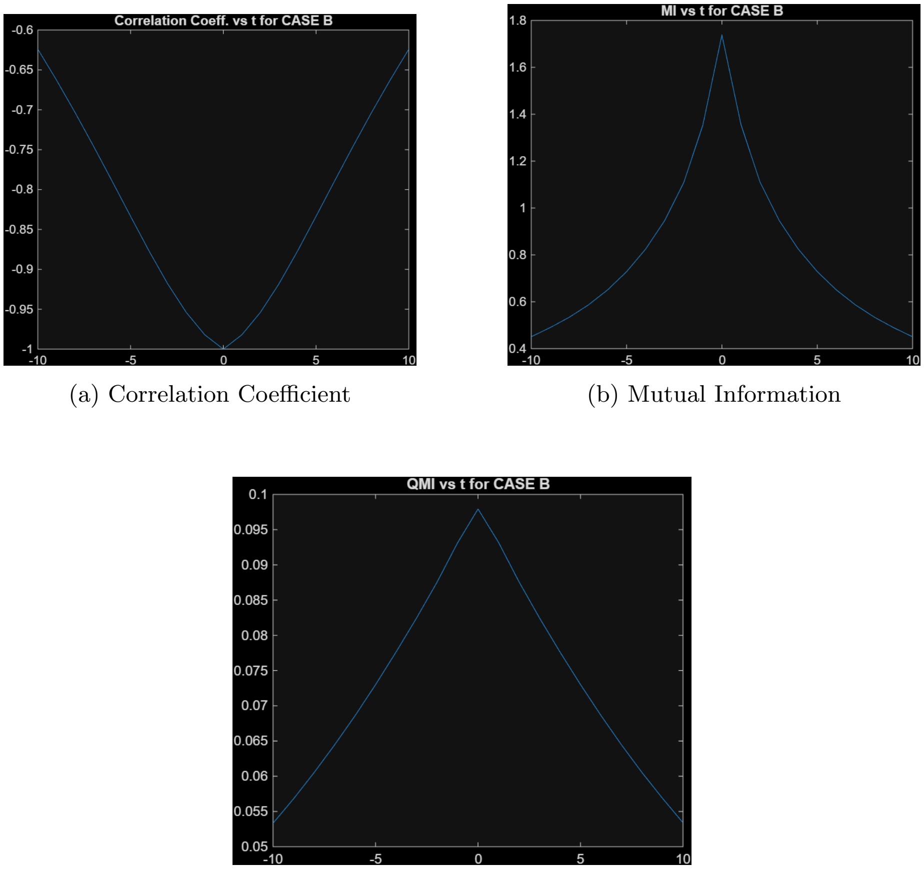
(b) Mutual Information

(c) Quadratic Mutual Information

QMI and MI are maximised at t=-1, where the images depend on each other the most. The correlation coeff minimizes at t = -1, and is positive overall.

CASE B

Measures of dependence between first image and shifted version of $(255 - I_1)$. The second image is obtained as a direct function of the first without any 'shifting' involved, all plots have their maxima/minima at t=0. Thus they are aligned.

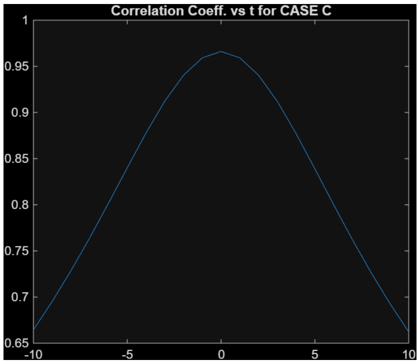


(a) Correlation Coefficient

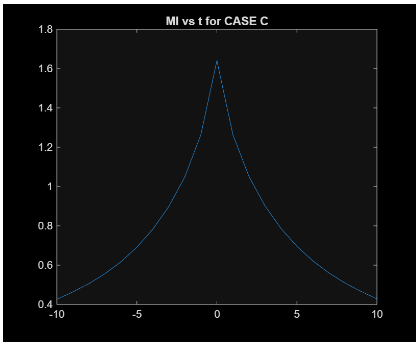
(b) Mutual Information

(c) Quadratic Mutual Information

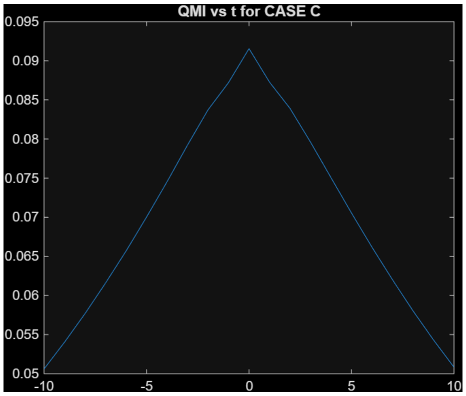
The correlation coefficient is negative because of the nature of the function. It is equal to -1 at t=0, which correctly represents the linear relation between I_1 and I_2 (also they are inversely related).
For MI and QMI, the peaks at t = 0, as at that position the images depend on each other the most (Mutual Information).
CASE C Measures of dependence between first image and shifted version of $I_2 = 255 \times (I_1)^2/\max((I_1)^2) + 1$.
The second image is obtained as a direct function of the first without any 'shifting' involved, all plots have their maxima/minima at t=0. Thus they are aligned.



(a) Correlation Coefficient



(b) Mutual Information



(c) Quadratic Mutual Information

The correlation coefficient is positive because of the nature of the function which positively correlates them.
For MI and QMI, the peaks at t = 0, as at that position the images depend on each other the most (Mutual Information).

Question 5

$\phi_X(t) = E[e^{tX}]$. Using Markov's inequality we can prove the required results:

- Right tail ($t > 0$):

$$P(X \geq x) = P(e^{tX} \geq e^{tx}) \leq E[e^{tX}] \cdot e^{-tx}$$

$$P(X \geq x) \leq e^{-tx} \cdot \phi_X(t)$$

- Left tail ($t < 0$):

$$P(X \leq x) = P(e^{tX} \geq e^{tx}) \leq E[e^{tX}] \cdot e^{-tx}$$

$$P(X \leq x) \leq e^{-tx} \cdot \phi_X(t)$$

The random variable $X = \sum_{i=1}^{n} X_i$ has MGF:

$$\phi_X(t) = \prod_{i=1}^{n} \phi_{X_i}(t)$$

Each X_i is a Bernoulli random variable with $E[X_i] = p_i$.
Thus, $\phi_{X_i}(t) = 1 - p_i + p_i e^t$
Now, using approximation $1 + x \leq e^x$, $1 + p_i(e^t - 1) \leq e^{p_i(e^t-1)}$. Thus,

$$\phi_X(t) = \prod_{i=1}^{n} \phi_{X_i}(t) \leq \prod_{i=1}^{n} e^{p_i(e^t-1)}$$

$$\phi_X(t) \leq e^{(e^t-1)\cdot\sum p_i}$$

Now using the above proven bound for $t > 0$ and putting $\sum p_i = \mu$,

$$P(X > (1+\delta)\mu) \leq P(X \geq (1+\delta)\mu) \leq e^{-(1+\delta)t\mu} \phi_X(t)$$

$$P(X > (1+\delta)\mu) \leq \frac{e^{(e^t-1)\cdot\mu}}{e^{(1+\delta)t\mu}}$$

To find the tightest bound, minimize the expression on the right side with respect to t. To minimize $e^t - 1 - t(1+\delta)$, differentiate and set to 0.

$$e^t - (1+\delta) = 0$$

$$t = \ln(1+\delta)$$

Thus, we get the bound as:

$$e^{\mu(\delta-(1+\delta)\ln(1+\delta))}$$

Question 6

Let the random variable $T = i$ denote the trial number at which we get the first head. Hence, all the previous $i - 1$ tosses must be tails.

$P(T = i) =$ (Probability of a head on the i-th toss)×(Probability of tails on all previous $i - 1$ tosses)

$$P(T = i) = p(1-p)^{i-1}.$$

We know that for a discrete random variable X,

$$E[X] = \sum_{i=1}^{n} x_i \, P(X = x_i)$$

Thus,

$$E[T] = \sum_{i=1}^{n} i\, p(1-p)^{\,i-1}$$

Let $S_n = \sum_{i=1}^{n} i\,(1-p)^{\,i-1}$
Put $q = 1{-}p$. Then $S_n = \sum_{i=1}^{n} i\, q^{\,i-1}$. This is a standard arithmetico–geometric progression (AGP). Consider

$$S_n - qS_n = (1 + q + q^2 + \cdots + q^{\,n-1}) - nq^n.$$

$$1 + q + q^2 + \cdots + q^{\,n-1} = \frac{1-q^n}{1-q}$$

Hence,

$$S_n(1-q) = \frac{1-q^n}{1-q} - nq^n$$

Substituting $1 - q = p$,

$$S_n = \frac{1 - q^n - nq^n(1-q)}{(1-q)^2} = \frac{1 - q^n\big(1 + n(1-q)\big)}{(1-q)^2}$$

Finally, substituting into the expectation formula:

$$E[T] = pS_n = \frac{1-(1-p)^n}{p} - n(1-p)^n$$