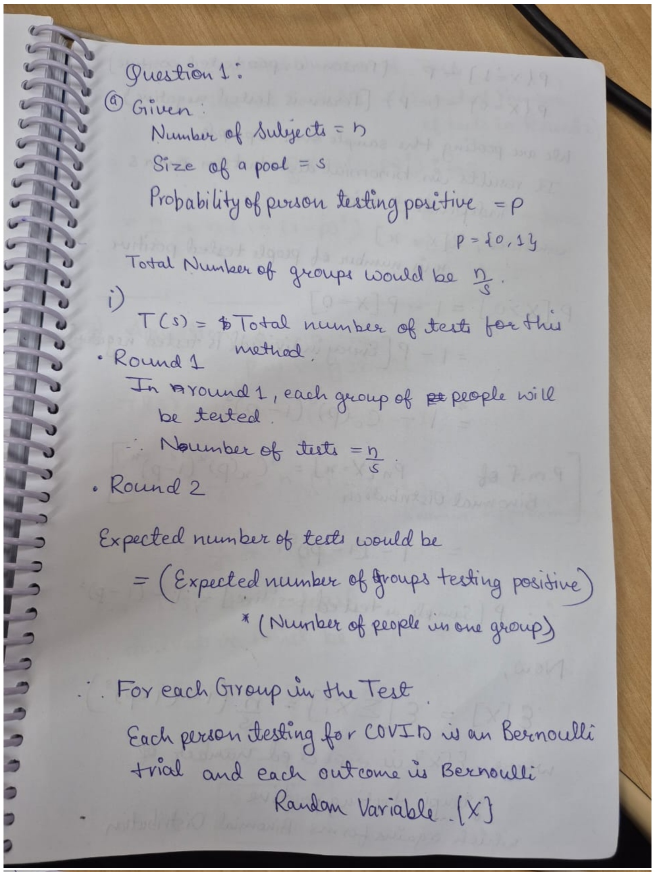
Assignment 2: CS 215

Mahant Sakhare - 24B0956, Hitansh Baria - 24B1075, Harshal Walke - 24B0954

1. **(a) Solution Images**



Question 1:

ⓐ Given:

Number of Subjects $= n$

Size of a pool $= s$

Probability of person testing positive $= p$

$p = \{0, 1\}$

Total Number of groups would be $\frac{n}{s}$.

i)

$T(s) =$ Total number of tests for this method.

- Round 1

In round 1, each group of people will be tested.

$\therefore$ Number of tests $= \frac{n}{s}$.

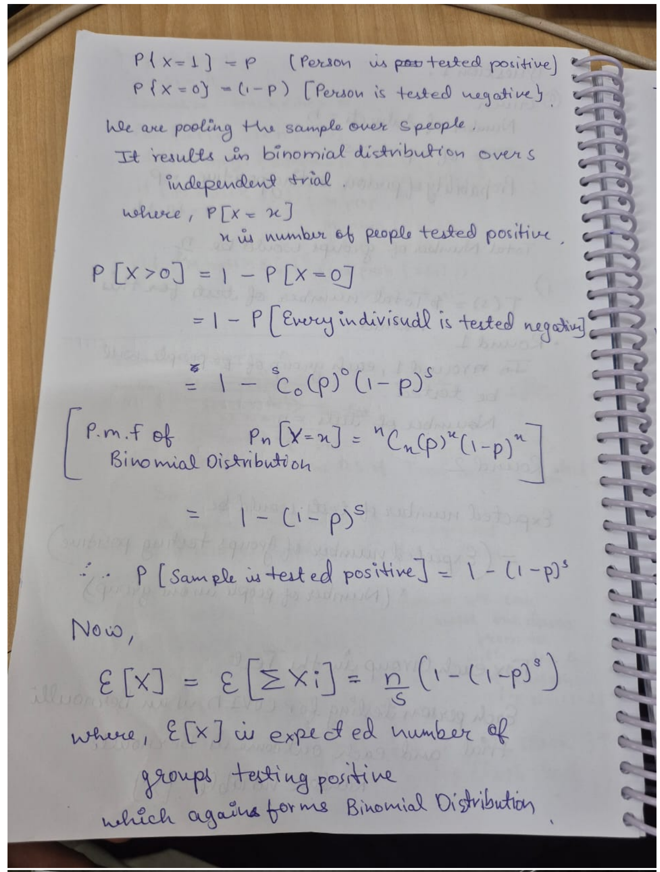
- Round 2

Expected number of tests would be

$= $ (Expected number of groups testing positive) * (Number of people in one group)

$\therefore$ For each Group in the Test.

Each person testing for COVID is an Bernoulli trial and each outcome is Bernoulli Random Variable $\{X\}$



$P\{X=1\} = p$ [Person is tested positive]

$P\{X=0\} = (1-p)$ [Person is tested negative].

We are pooling the sample over s people.

It results in binomial distribution over s independent trial.

where, $P[X=x]$

x is number of people tested positive.

$P[X>0] = 1 - P[X=0]$

$= 1 - P$[Every indivisual is tested negative]

$= 1 - {}^{s}C_0 (p)^0 (1-p)^s$

[P.m.f of Binomial Distribution $\quad P_n[X=x] = {}^{n}C_x (p)^x (1-p)^x$]

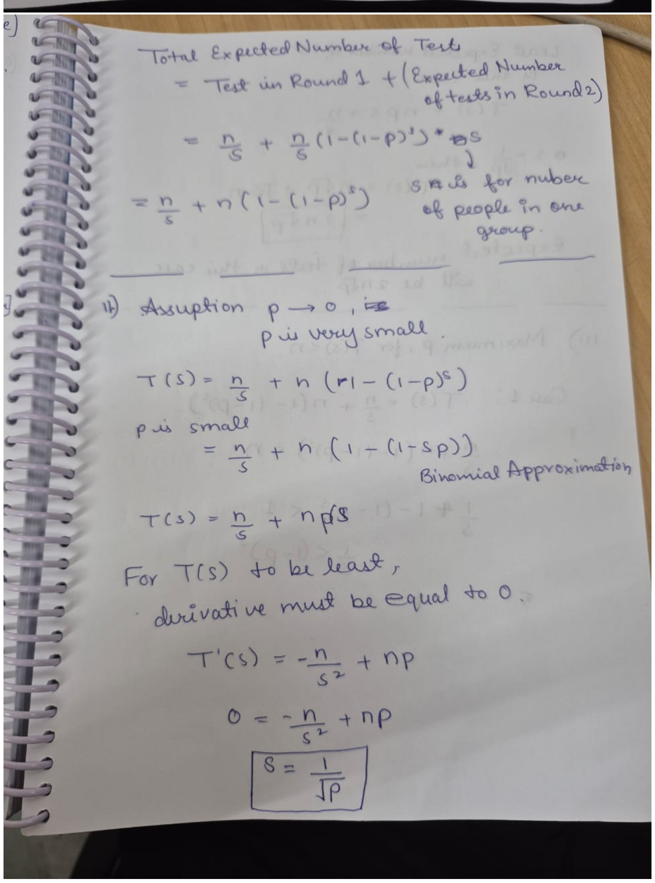
$= 1 - (1-p)^s$

$\therefore P$[Sample is tested positive] $= 1 - (1-p)^s$

Now,

$\mathcal{E}[X] = \mathcal{E}[\sum X_i] = \frac{n}{s}(1-(1-p)^s)$

where, $\mathcal{E}[X]$ is expected number of groups testing positive which again forms Binomial Distribution.



Total Expected Number of Test

$=$ Test in Round 1 $+$ (Expected Number of tests in Round 2)

$= \frac{n}{s} + \frac{n}{s}(1-(1-p)^s) * s$

s is for nuber of people in one group.

$= \frac{n}{s} + n(1-(1-p)^s)$

ii) Assuption $p \to 0$, p is very small.

$T(s) = \frac{n}{s} + n(1-(1-p)^s)$

p is small

$= \frac{n}{s} + n(1-(1-sp))$

Binomial Approximation

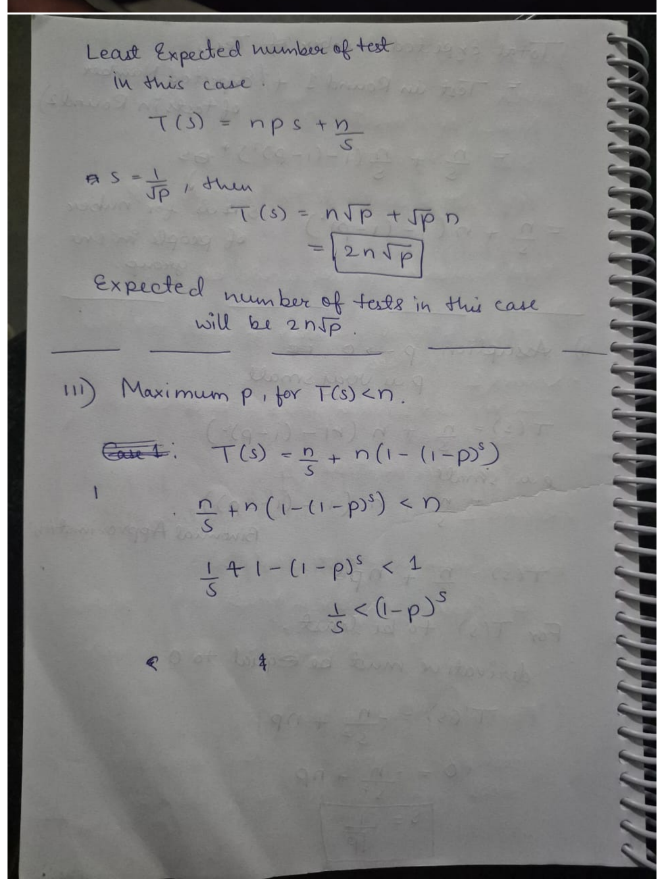
$T(s) = \frac{n}{s} + nps$

For $T(s)$ to be least, derivative must be equal to 0.

$T'(s) = -\frac{n}{s^2} + np$

$0 = -\frac{n}{s^2} + np$

$\boxed{s = \frac{1}{\sqrt{p}}}$



Least Expected number of test in this case.

$T(s) = nps + \frac{n}{s}$

as $s = \frac{1}{\sqrt{p}}$, then

$T(s) = n\sqrt{p} + \sqrt{p}\,n$

$= \boxed{2n\sqrt{p}}$

Expected number of tests in this case will be $2n\sqrt{p}$.

iii) Maximum p, for $T(s) < n$.

$T(s) = \frac{n}{s} + n(1-(1-p)^s)$

$\frac{n}{s} + n(1-(1-p)^s) < n$

$\frac{1}{s} + 1 - (1-p)^s < 1$

$\frac{1}{s} < (1-p)^s$

The code for the 3rd part is in the provided folder named as Q1a.m .

(b)

b-i) Probability a healthy subject participates in a negative pool

Consider a healthy subject and a pool they join. The pool tests negative if no diseased subject joins it.

Each other subject joins the pool with probability π and is diseased with probability p, so the probability a diseased subject joins is $p\pi$. Across $n-1$ other subjects, the probability that no diseased subject joins is approximately:

$$P_{\text{negative}} \approx (1-p\pi)^{n-1} \approx e^{-np\pi} \quad \text{(for large } n \text{ and small } p\pi\text{)}$$

.

b-ii) Optimal π

Let $K \sim \text{Binomial}(T_1, \pi)$ be the number of pools a healthy subject joins. Let $q = P_{\text{negative}} \approx e^{-np\pi}$.

The probability that the healthy subject participates in at least one negative pool is:

$$P_{\text{at least one negative}} \approx 1 - \exp(-\lambda q) = 1 - \exp(-T_1 \pi e^{-np\pi})$$

To maximize this probability with respect to π, define

$$f(\pi) = \pi e^{-np\pi}.$$

Derivative set to zero gives the optimal participation probability:

$$\pi^* = \frac{1}{np}$$

b-iii) Probability all pools a healthy subject participates in are positive

The complement of having at least one negative pool:

$$P_{\text{all positive}} \approx \exp(-T_1 \pi e^{-np\pi})$$

At $\pi = \pi^*$:

$$P_{\text{all positive}}(\pi^*) \approx \exp\left(-\frac{T_1}{enp}\right)$$

b-iv) Expected total number of tests

- **Round 1:** T_1 tests
- **Round 2:**
  - Positive subjects: np tests (all positives are tested)
  - "Unlucky" healthy subjects: $n(1-p) \cdot \exp(-T_1/(enp))$

Hence, the expected total number of tests:

$$E[T_{\text{total}}] = T_1 + np + n(1-p)\exp\left(-\frac{T_1}{enp}\right)$$

b-v) Optimal T_1 and minimal expected tests

Goal: minimize

$$g(T_1) = T_1 + np + n(1-p)\exp\left(-\frac{T_1}{enp}\right)$$

Derivative:

$$g'(T_1) = 1 - \frac{n(1-p)}{enp}\exp\left(-\frac{T_1}{enp}\right) = 0$$

Solve for T_1^*:

$$T_1^* = enp \cdot \ln\frac{n(1-p)}{ep}$$

Expected total tests at optimum:

$$E[T_{\text{total,opt}}] = T_1^* + np + n(1-p)\cdot\frac{ep}{n(1-p)} = T_1^* + np + ep$$

(c) We now compare the expected number of tests for:

- **Method (a):** Simple pooling (non-overlapping pools)

$$T_a(p) = \min_s \left[\frac{n}{s} + n\left(1-(1-p)^s\right)\right]$$

  Approximation for small p: $T_a \approx 2n\sqrt{p}$, $s^* \approx 1/\sqrt{p}$.
- **Method (b):** Overlapping pools

$$T_b(p) \approx enp\ln\frac{n(1-p)}{ep} + np + ep$$

  with $\pi^* = 1/(np)$ and $T_1^* = enp\ln\frac{n(1-p)}{ep}$.

The code for the plot is provided in the folder named as Q1c.

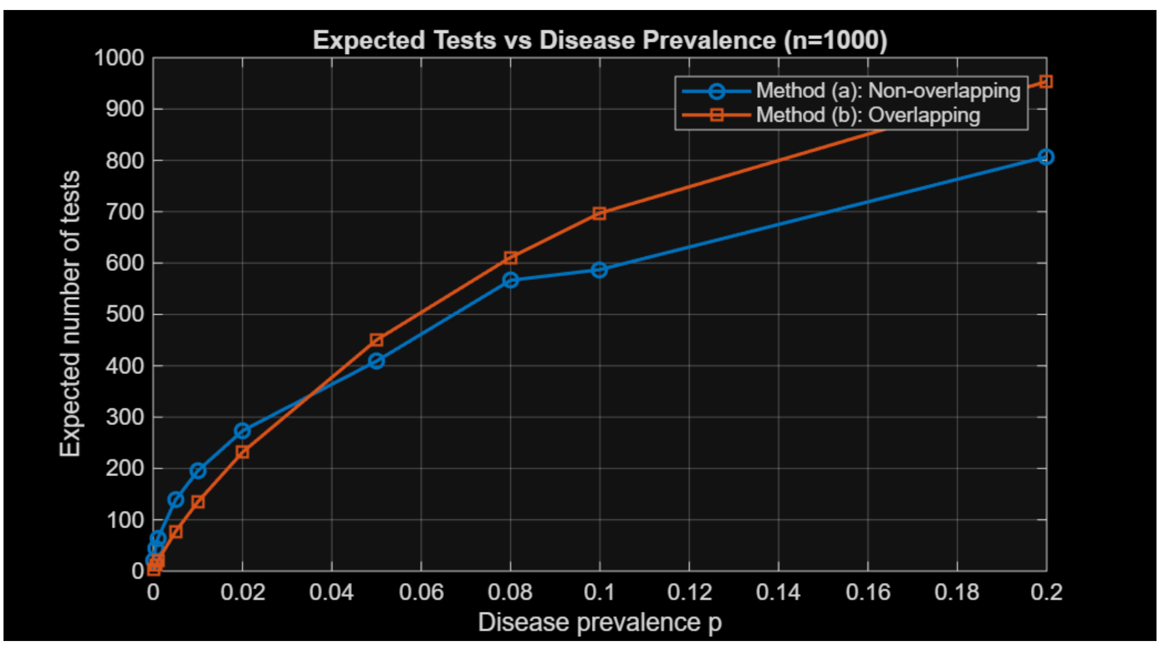


Figure 1: Expected number of tests vs disease prevalence p for methods (a) and (b). Method (b) generally reduces tests at moderate p compared to naive pooling.

MATLAB

We generate expected tests for $n = 1000$ and $p \in \{1\text{e-}4, 5\text{e-}4, 0.001, 0.005, 0.01, 0.02, 0.05, 0.08, 0.1, 0.2\}$. Include the generated plot as an image:

Interpretation

- For very small p, both methods significantly reduce tests relative to n.
- Method (b) may be slightly better for moderate p due to overlapping pool optimization.
- For large p (e.g., $p > 0.1$), pooling becomes less beneficial as almost all pools are positive.

2. Given two independent random variables X and Y with PDFs $f_X(.)$ and $f_Y(.)$ respectively, and define $Z = XY$. We derive the PDF of Z using its CDF.

$$F_Z(z) = P(Z \le z) = P(XY \le z).$$

Depending on the sign of X, we have two cases.

Case1: $X > 0$.

If $X = x > 0$, then the event $XY \le z$ is equivalent to $Y \le z/x$. Hence

$$F_Z(z) = \int_0^\infty \int_{-\infty}^{z/x} f_{X,Y}(x,y)\, dy\, dx.$$

Since X and Y are independent, $f_{X,Y}(x,y) = f_X(x) f_Y(y)$, so

$$F_Z(z) = \int_0^\infty \int_{-\infty}^{z/x} f_X(x)\, f_Y(y)\, dy\, dx.$$

Differentiating with respect to z (By Leibniz's rule), we obtain

$$\begin{aligned}
f_Z(z) &= \frac{\partial}{\partial z} F_Z(z) \\
&= \int_0^\infty \frac{\partial}{\partial z} \left[\int_{-\infty}^{\frac{z}{x}} f_X(x) \cdot f_Y(y) dy \right] dx \\
&= \int_0^\infty f_X(x) \cdot f_Y\left(\frac{z}{x}\right) \cdot \frac{1}{x} dx \\
&= \int_0^\infty f_X(x) \cdot f_Y\left(\frac{z}{x}\right) \cdot \frac{1}{|x|} dx
\end{aligned}$$

Case2: $X < 0$.

If $X = x < 0$, then the event $XY \le z$ is equivalent to $Y \ge z/x$. Hence

$$F_Z(z) = \int_{-\infty}^0 \int_{z/x}^\infty f_{X,Y}(x,y)\, dy\, dx = \int_{-\infty}^0 \int_{z/x}^\infty f_X(x)\, f_Y(y)\, dy\, dx.$$

Differentiating with respect to z gives

$$\begin{aligned}
f_Z(z) &= \frac{\partial}{\partial z} F_Z(z) \\
&= \int_{-\infty}^{0} \frac{\partial}{\partial z}\left[\int_{\frac{z}{x}}^{\infty} f_X(x) f_Y(y) dy\right] dx \\
&= \int_{-\infty}^{0} f_X(x) \cdot \left[0 - f_Y\left(\frac{z}{x}\right)\right] \cdot \frac{1}{x} dx \\
&= \int_{-\infty}^{0} f_X(x) \cdot f_Y\left(\frac{z}{x}\right) \cdot \frac{-1}{x} dx \\
&= \int_{-\infty}^{0} f_X(x) \cdot f_Y\left(\frac{z}{x}\right) \cdot \frac{1}{|x|} dx
\end{aligned}$$

Combining both cases, we get

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) \cdot f_Y\left(\frac{z}{x}\right) \cdot \frac{1}{|x|} dx$$

which is the desired PDF of $Z = XY$.

3. The correct estimate for $E(x)$ is $\hat{x} = \sum_{i=1}^{n} \frac{x_i}{n}$.

   REASON: We know that each x_i is a random variable in itself, since x_i is a sample from the PDF $f_X(.)$. Hence, each x_i will have same PDF as X. Also it is given to us that the x_i's are independent, which means that they are independent and identically distributed random variables each having mean μ.

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

Since, the expectation of each x_i is same that of X, we have

$$E(X) = E(x_i) = \mu$$

Taking n to be a large number and applying Weak Law of Large Numbers, we get

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

$\big(\because\ P\{|\sum_{i=1}^{n} \frac{x_i}{n} - \mu| > \epsilon\} \to 0 \text{ as } n \to \infty\ \forall\ \epsilon > 0\big)$

$$\therefore E(\hat{x}) = \mu = E(X)$$

Reason Other Option is incorrect:

Let $\{y_i\}_{i=1}^{n}$ be the other set of random variable, where $y_i = x_i f_X(x_i)$.

The y_is are iids with expectation

$$E(y_i) = E(x_i f_X(x_i))$$

As we know that x_i and X are identically distributed, we have

$$\begin{aligned}
E(y_i) &= E(X f_X(x)) \\
&= \int (x f_X(x)) f_X(x) \\
&= \int x f_X(x)^2 dx
\end{aligned}$$

By WLLN, we know

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

$$\therefore E(y_i) = \int x f_X(x)^2 dx$$

This is estimating some other random variable with PDF $f_X(.)^2$. Hence, it is incorrect estimate for $E(X)$ as X has PDF $f_X(.)$.

4. Dependence Measures

(a) **Correlation Coefficient (ρ)**

$$\rho = \frac{\text{cov}(I_1, I_2)}{\sigma_{I_1}\sigma_{I_2}}, \quad -1 \leq \rho \leq 1$$

(b) **Quadratic Mutual Information (QMI)**

$$\text{QMI} = \sum_{i_1}\sum_{i_2}\left(p_{I_1 I_2}(i_1, i_2) - p_{I_1}(i_1)p_{I_2}(i_2)\right)^2$$

(c) **Mutual Information (MI)**

$$\text{MI} = \sum_{i_1}\sum_{i_2} p_{I_1 I_2}(i_1, i_2)\log\left(\frac{p_{I_1 I_2}(i_1, i_2)}{p_{I_1}(i_1)p_{I_2}(i_2)}\right)$$

The joint histogram $p_{I_1 I_2}$ is computed with a bin-width of 10. Marginals p_{I_1}, p_{I_2} are derived by summing over rows/columns of the joint histogram.

Experimental Scenarios

(a) **Original images:** $I_1 = \text{T1}$, $I_2 = \text{T2}$.

(b) **Negative image:** $I_2 = 255 - I_1$.

(c) **Squared image:** $I_2 = 255 \times \frac{(I_1)^2}{\max((I_1)^2)} + 1$.

The codes of the following three plots are provided in the folder named as Q4a , Q4b, Q4c for first ,second and third plot respectively.

Case 1: Original MR Images

Observations:

- **QMI and MI:** Both peak sharply at $t_x = 0$, confirming maximum dependence at perfect alignment.
- **Correlation (ρ):** Shows a minimum at $t_x = 0$, reflecting the fact that T1 and T2 intensities are not linearly related. Still, the extremum indicates correct alignment.

Case 2: Negative Image

Observations:

- **Correlation (ρ):** At $t_x = 0$, $\rho \approx -1$, as expected from a perfect negative linear relation.
- **QMI and MI:** Still peak at $t_x = 0$, highlighting robustness to the sign of dependence.

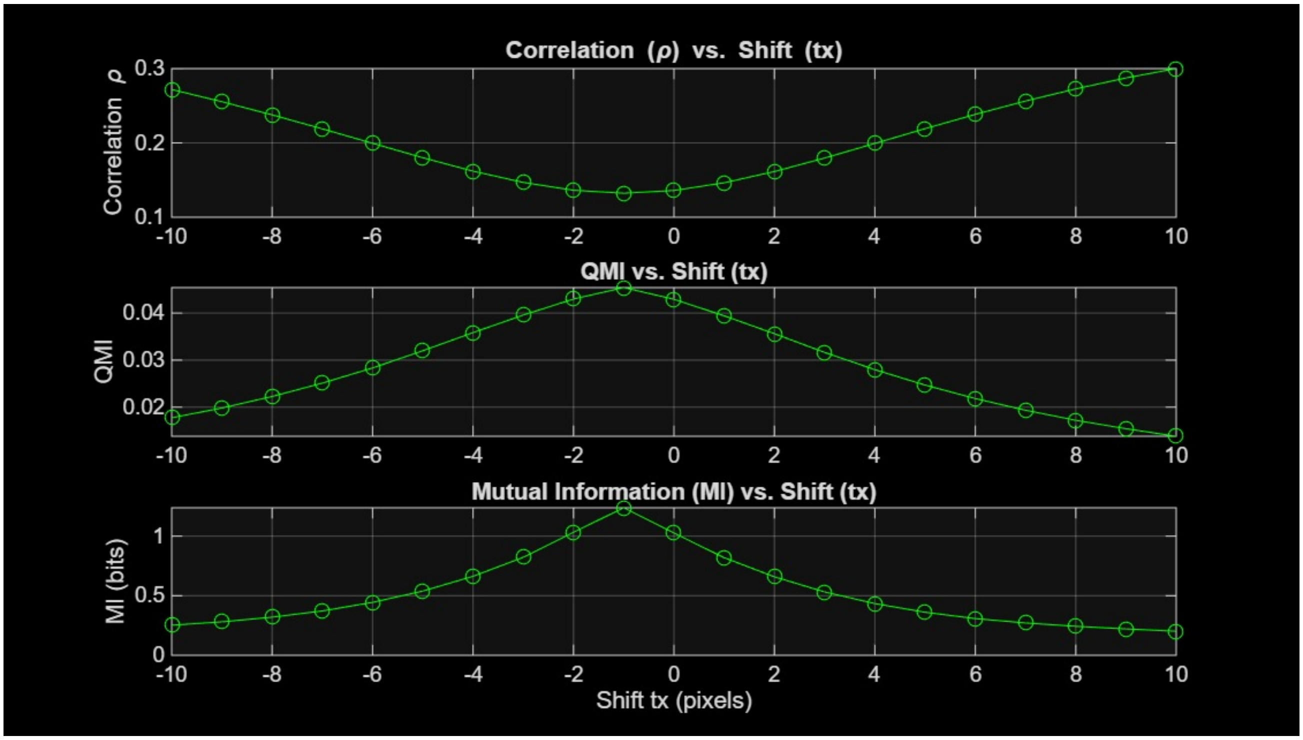


Figure 2: Dependence measures vs. shift (t_x) for original MR images.

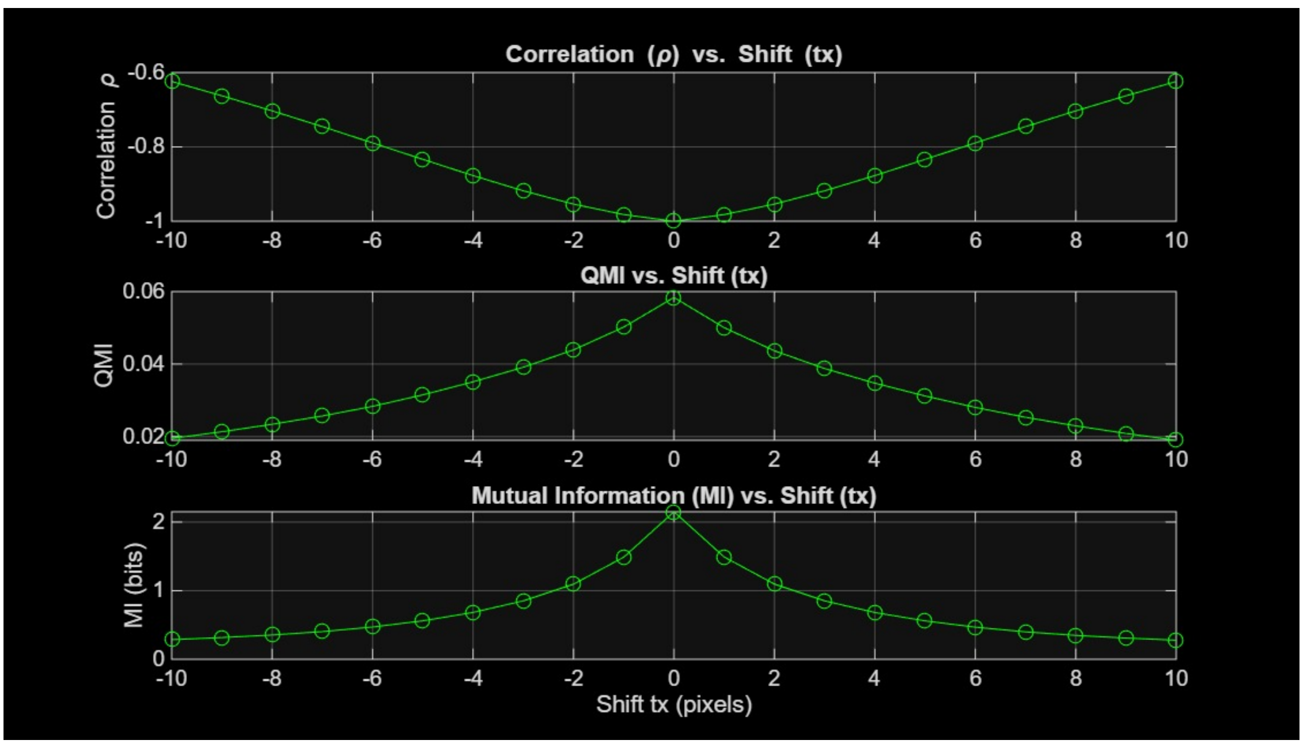


Figure 3: Dependence measures vs. shift (t_x) for I_1 and its negative $I_2 = 255 - I_1$.

Case 3: Squared Image

Observations:

- **Correlation (ρ):** Does not show a clear peak at $t_x = 0$, confirming unreliability under nonlinear intensity transformations.

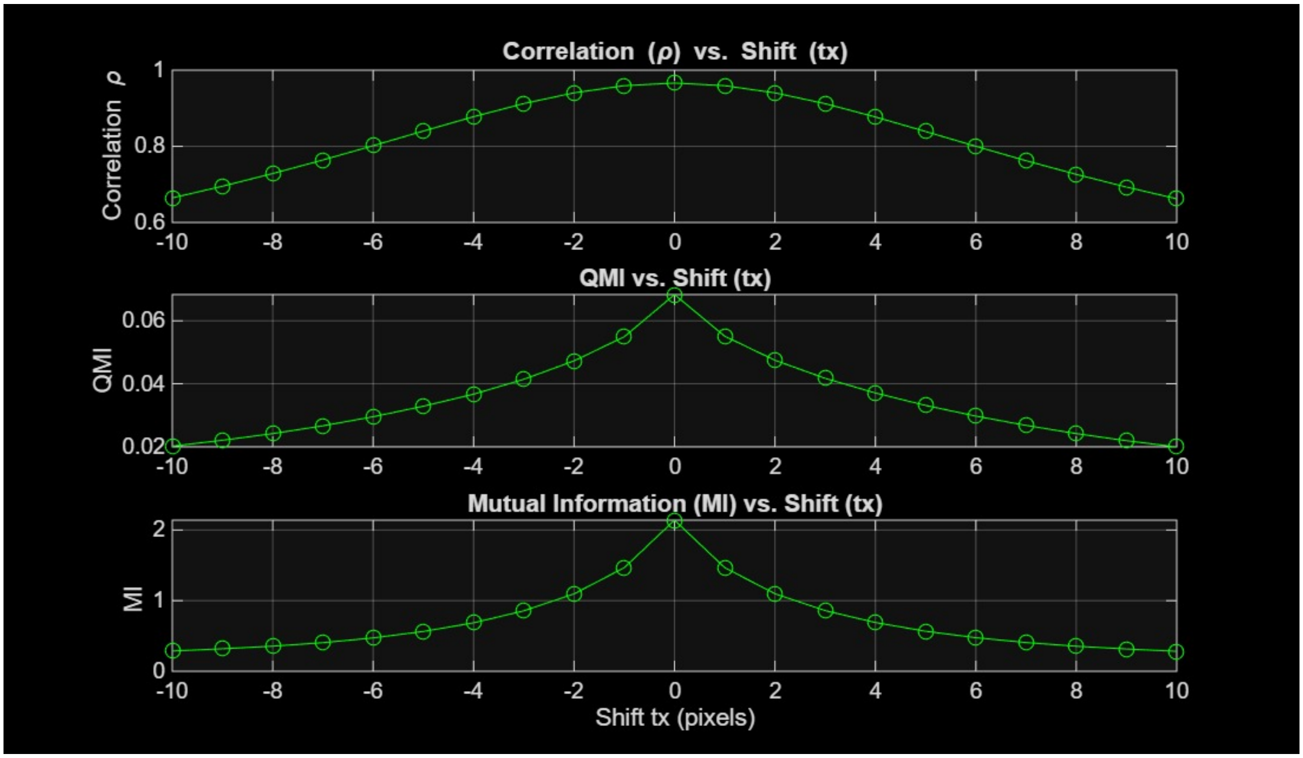


Figure 4: Dependence measures vs. shift (t_x) for squared image $I_2 = 255 \times \frac{(I_1)^2}{\max((I_1)^2)} + 1$.

- **QMI and MI:** Both exhibit distinct maxima at $t_x = 0$, successfully capturing nonlinear dependence and providing correct alignment.

5. When $t > 0$:

$$P(X \geq x) \leq P\big(e^{tX} \geq e^{tx}\big) \leq \frac{\mathbb{E}[e^{tX}]}{e^{tx}} \quad \text{(By Markov's inequality)}$$

$$P(X \geq x) \leq e^{-tx}\,\varphi_X(t), \quad \text{when } t > 0$$

where $\varphi_X(t) = \mathbb{E}[e^{tX}]$ is the moment generating function (MGF).

When $t < 0$:

$$P(X \leq x) \leq P\big(e^{tX} \geq e^{tx}\big) \leq \frac{\mathbb{E}[e^{tX}]}{e^{tx}} \quad \text{(By Markov's inequality)}$$

$$P(X \leq x) \leq e^{-tx}\,\varphi_X(t), \quad \text{when } t < 0$$

Let

$$X = \sum_{i=1}^{n} X_i, \quad \mathbb{E}[X_i] = p_i, \quad \mathbb{E}[X] = \mu.$$

We know that

$$P(X \geq x) \leq \frac{\mathbb{E}[e^{tX}]}{e^{tx}} \quad \text{for } t > 0.$$

Now, let $x = (1+\delta)\mu$, then

$$P(X \geq (1+\delta)\mu) \leq \frac{\mathbb{E}[e^{tX}]}{e^{(1+\delta)\mu t}}.$$

Since the X_i are independent:

$$\mathbb{E}[e^{tX}] = \mathbb{E}\left[e^{t\sum_{i=1}^{n} X_i}\right] = \prod_{i=1}^{n} \mathbb{E}[e^{tX_i}].$$

For a Bernoulli random variable X_i:

$$\mathbb{E}[e^{tX_i}] = (1-p_i)e^{t\cdot 0} + p_i e^{t\cdot 1} = 1 - p_i + p_i e^t = 1 + p_i(e^t - 1).$$

Using the inequality $1 + x \leq e^x$:

$$1 + p_i(e^t - 1) \leq e^{p_i(e^t-1)}.$$

Therefore:

$$\prod_{i=1}^{n} \mathbb{E}[e^{tX_i}] \leq \prod_{i=1}^{n} e^{p_i(e^t-1)} = e^{(e^t-1)\sum_{i=1}^{n} p_i} = e^{(e^t-1)\mu}.$$

Hence, the tail bound becomes:

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

$$\prod_{i=1}^{n} \mathbb{E}[e^{tX_i}] \leq e^{(e^t-1)\mu}$$

Thus,

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

6. Let $P(X = i)$ be the probability of occurrence of first head at i^{th} trial.

   Probability of occurrence of heads is p.

   Hence probability of occurrence of tails is $(1-p)$.

$$\therefore\ P(X = i) = (1-p)^{i-1}p$$

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

Since p is constant:

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

Expanding:

$$= p\Big[1(1-p)^0 + 2(1-p)^1 + 3(1-p)^2 + \cdots + n(1-p)^{n-1}\Big]$$

Now, let us simplify the series terms.

$$\sum_{i=1}^{n}(1-p)^{i-1} = \frac{1-(1-p)^n}{p}$$

$$\sum_{i=2}^{n}(1-p)^{i-1} = \frac{(1-p)-(1-p)^n}{p} = \frac{(1-p)-(1-p)^n}{p}$$

$$\sum_{i=3}^{n}(1-p)^{i-1} = \frac{(1-p)^2-(1-p)^n}{p}$$

$$\sum_{i=n}^{n}(1-p)^{i-1} = \frac{(1-p)^{n-1}-(1-p)^n}{p}$$

Let $P(X=i)$ be the probability of occurrence of the first head at the i^{th} trial.

The probability of occurrence of a head is p,
hence the probability of occurrence of a tail is $(1-p)$.

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

The expectation is:

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

Since p is constant:

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

Expanding:

$$= p\Big[1(1-p)^0 + 2(1-p)^1 + 3(1-p)^2 + \cdots + n(1-p)^{n-1}\Big]$$

Now, recall that

$$\sum_{i=1}^{n}(1-p)^{i-1} = \frac{1-(1-p)^n}{p}$$

$$\sum_{i=2}^{n}(1-p)^{i-1} = \frac{(1-p)-(1-p)^n}{p}$$

$$\sum_{i=3}^{n}(1-p)^{i-1} = \frac{(1-p)^2-(1-p)^n}{p}$$

$$\sum_{i=n}^{n}(1-p)^{i-1} = \frac{(1-p)^{n-1}-(1-p)^n}{p}$$

Hence,

$$E(T) = p\left(\frac{(1-p)-(1-p)^n}{p} + \frac{(1-p)^2-(1-p)^n}{p} + \cdots + \frac{(1-p)^{n-1}-(1-p)^n}{p}\right)$$

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

$$E(T) = \frac{(1-p)\left(1-(1-p)^{n-1}\right)}{p} - (n-1)(1-p)^n$$