Assignment-1

1)

	Zip Code	Age	Nationality	Disease
1	476**	2*	*	Heart Disease
2	476**	2*	*	Viral Infection
3	476**	2*	*	Cancer
4	476**	2*	*	Cancer
5	4790*	≥40	*	Viral Infection
6	4790*	≥40	*	Heart Disease
7	4790*	≥40	*	Viral Infection
8	4790*	≥40	*	Cancer
9	476**	3*	*	Cancer
10	476**	3*	*	Cancer
11	476**	3*	*	Viral Infection
12	476**	3*	*	Heart Disease

a) Value of K:-

Group1: Zip code 476** and Age 2* - 4 records
Group2: Zip code 4790* and Age ≥40 - 4 records
Group3: Zip code 476** and Age 3* - 4 records

Therefore K=4

b) l_1- diversity:-

Group1: {Heart Disease, Viral Infection, Cancer} - 3 Unique diseases
Group2: {Viral Infection, Heart Disease, Cancer} - 3 Unique diseases
Group3: {Cancer, Viral Infection, Heart Disease} - 3 Unique diseases

Here, there are 3 Unique diseases, hence the value of $l_1 = 3$

c) l_2- diversity:-

	Heart Disease	Viral Infection	Cancer
Group1 –	1/4	1/4	2/4
Group2 –	1/4	2/4	1/4
Group3 –	1/4	1/4	2/4

Entropy $= -\sum p_i \log p_i$

Group1 $= -\left[\frac{1}{4}\log\left(\frac{1}{4}\right) + \frac{1}{4}\log\left(\frac{1}{4}\right) + \frac{2}{4}\log\left(\frac{2}{4}\right)\right]$

$= -(-1.0397) = 1.0397$

$\log_e l = $

$l = e^{(1.039)} = 2.826$

Group2: $-\left[\frac{1}{4}\log\left(\frac{1}{4}\right)+\frac{2}{4}\log\left(\frac{2}{4}\right)+\frac{1}{4}\log\left(\frac{1}{4}\right)\right]$

$\log l = -(-1.0397) = 1.0397$

$l = e^{1.0397} = 2.826$

Group3: $-\left[\frac{1}{4}\log\left(\frac{1}{4}\right)+\frac{1}{4}\log\left(\frac{1}{4}\right)+\frac{2}{4}\log\left(\frac{2}{4}\right)\right]$

$\log l = -(-1.0397) = 1.0397$

$l = e^{1.0397} = 2.826$

Therefore, all groups have same value $l_2 = 2.826$

d) t_1-closeness:-

Global Probabilities of
HD = 3/12
VI - 4/12
Cancer = 5/12

	Heart Disease	Viral Infection	Cancer
Group1	1/4	1/4	2/4
Group2	1/4	2/4	1/4
Group3	1/4	1/4	2/4

$$D_{KL}(P\|Q) = \sum p(x)\log\left(\frac{p(x)}{q(x)}\right)$$

$$= \frac{1}{4}\log\left(\frac{1/4}{3/12}\right)+\frac{1}{4}\log\left(\frac{1/4}{4/12}\right)+\frac{2}{4}\log\left(\frac{2/4}{5/12}\right)$$

$$= 0 + \frac{1}{4}\log\left(\frac{3}{4}\right)+\frac{2}{4}\log\left(\frac{1}{\cancel{2}} * \frac{\cancel{12}^{6}}{5}\right)$$

$$= \frac{1}{4}\log\left(\frac{3}{4}\right)+\frac{1}{2}\log\left(\frac{6}{5}\right) = 0.01924$$

$$D_{KL}(P\|Q) = \sum p(x)\log\left(\frac{p(x)}{q(x)}\right)$$

$$= \frac{1}{4}\log\left(\frac{1/4}{3/12}\right)+\frac{2}{4}\log\left(\frac{2/4}{4/12}\right)+\frac{1}{4}\log\left(\frac{1/4}{5/12}\right)$$

$$= \frac{1}{2}\log\left(\frac{3}{2}\right)+\frac{1}{4}\log(3/5) = 0.0750$$

$D_{KL}(P||Q) = \sum p(x) \log\left(\frac{p(x)}{q(x)}\right)$

$= \frac{1}{4}\log\left(\frac{1/4}{3/12}\right) + \frac{1}{4}\log\left(\frac{1/4}{4/12}\right) + \frac{2}{4}\log\left(\frac{2/4}{5/12}\right)$

$= 0 + 0.25 * \log\left(\frac{3}{4}\right) + \frac{2}{4}\log\left(\frac{\cancel{12}}{5} \times \frac{2}{4}\right)$

$= 0.0192$

KL-Divergence $= \max(D_{KL_{G_1}}, D_{KL_{G_2}}, D_{KL_{G_3}})$

$= \underline{0.0750}$

2) $X \in R^{100 \times 2} \Rightarrow \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ \vdots & \\ 0 & 0 \end{bmatrix} \rightarrow$ 100 Rows

$q(x) = \frac{1}{100} X^T X$

$||x_i||_2 = 1$

<u>Laplacian</u> :-

a) $f(x) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$; $b = \frac{\Delta}{\varepsilon}$

l_1-sensitivity $= ||q(x) - q(x')||$

$q(x) = \frac{X^T X}{100} = \frac{1}{100}\sum_{i=1}^{100} x_i^T . x_i$

$q(x') = \frac{X'^T . X}{100} = \frac{1}{100}\sum_{i=1}^{100} x_i'^T x_i'$

Δf (l_1-Sensitivity) $= ||q(x) - q(x')||$

$= ||\frac{1}{100}\sum_{i=1}^{100} x_i^T . x_i - \frac{1}{100}\sum_{i=1}^{100} x_i'^T . x_i'||$

$= \frac{1}{100}(x_m^T . x_m - x_m'^T . x_m')$ [$\because$ m is the value that differs from 2 datasets]

<u>Eg</u>:- Let $x_m = [0 \ 1]$ $\quad x_m' = [1 \ 0]$

$x_m^T = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ $\quad x_m'^T = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$x_m^T . x_m = \begin{bmatrix} 0 \\ 1 \end{bmatrix}[0 \ 1] = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ $\quad x_m'^T . x_m' = \begin{bmatrix} 1 \\ 0 \end{bmatrix}[1 \ 0] = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$

$\Delta f = \frac{1}{100}(x_m^T . x_m - x_m'^T . x_m') = \frac{1}{100}\left[\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\right] = \frac{1}{100}(|-1| + |0| + |0| + |1|) = 2/100$

$\therefore$ l_1 - sensitivity $= 2/100$

pdf of calibrated distributed $= \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$

$$b = \frac{\Delta}{\varepsilon} = \frac{2}{100(\varepsilon)}$$

$$\Rightarrow \frac{1}{2\left(\frac{2}{100\varepsilon}\right)} * \exp\left(\frac{-|x|}{\frac{2}{100\varepsilon}}\right) = \frac{25}{\varepsilon} * \exp\left(\frac{-|x|50}{\varepsilon}\right)$$

here 1 - differential privay $\varepsilon = 1$

$$\Rightarrow \frac{25}{1} * \exp(-50|x|)$$

b) Let $x_1 = [1\ 0]$ $\quad$ $x_1 = (0\ 1]$

$x_2 = [0\ 1]$ $\quad$ $x_2 = (0, 1]$

$\underbrace{\qquad}_{\text{dataset } X}$ $\quad$ $\underbrace{\qquad}_{\text{dataset } X'}$

$$q(x) = \frac{1}{100} X^T X \qquad q(x') = \frac{1}{100} X'^T X$$

$$= \frac{1}{100}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad = \frac{1}{100}\begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$$

$$= \frac{1}{100}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad = \frac{1}{100}\begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}$$

$$\Delta_2 f = \| q(x) - q(x') \|_2 = \frac{1}{100}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = \frac{1}{100}\sqrt{1 + 1}$$

$$= \sqrt{2}/100$$

$$\text{Gaussian pdf} = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\sigma = \frac{c\Delta_2 f}{\varepsilon} = \frac{\sqrt{2\ln\left(\frac{1.25}{\delta}\right)} \cdot \frac{\sqrt{2}}{100}}{\varepsilon} = \frac{\sqrt{4\ln\left(\frac{1.25}{10^{-5}}\right)}}{(1)\,100}$$

$$= 0.2347$$

$$\text{pdf} = \frac{1}{(0.2347)\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x - \cancel{0}}{0.2347}\right)^2}$$

$$= \frac{1}{1.0427} * e^{\frac{-x^2}{0.1101}}$$

3). Given $s(X,p) = -\left|\sum_{i\in[n]} \text{sign}(x_i - p)\right|$ ③

a) Rationale of defined score function:-

For each individual, the salary is compared to median value:

i) if $x_i > p$ then $\text{Sign}(x_i - p) = 1$

ii) if $x_i = p$ then $\text{Sign}(x_i - p) = 0$

iii) if $x_i < p$ then $\text{Sign}(x_i - p) = -1$

This can be explained as

- If p is too small (many salaries greater than p) ⇒ the sign (x_i-p) is positive
- If p is too large (many salaries is less than p) ⇒ the sign $(x_i - p)$ is negative
- If p is the median, the sign $(x_i - p)$ is close to zero,

The theme is to minimize sign$(x_i - p)$, which corresponds to find a candidate p that balances the no. of salaries greater than p and no. of salaries less than p. This is exactly the property of Median

Optimal Value:-

- When at Median:- If p is the median, then approximately half the salaries $x_i > p$, half salaries $< p$. The sum$((x_i - p))$ will be close to '0'
- When not a Median:- If p is not a median, the sum$(x_i - p)$ is not zero, then sign $(x_i - p)$ will be either positive/negetive, where it would be too large or too small

The score function is minimized when p is median.

b) Sensitivity of Score function:-

Sensitivity of score function $\Delta S = \max_{x,x'} |S(x,p) - S(x',p)|$

Possible scenarios:-

a) if $x_i > p$ and $x_i' < p$ then $Sign(x_i - p) - Sign(x_i' - p) = 1 - (-1) = 2$

b) if $x_i < p$ and $x_i' > p$ then $Sign(x_i - p) - Sign(x_i' - p) = -1 - 1 = -2$

c) if $x_i < p$ and $x_i' < p$ then $Sign(x_i - p) - Sign(x_i' - p) = -1 - (-1) = 0$

d) if $x_i > p$ and $x_i' > p$ then $Sign(x_i - p) - Sign(x_i' - p) = 1 - 1 = 0$

e) if $x_i = p$ and $x_i' = p$ then $Sign(x_i - p) - Sign(x_i' - p) = 0 + 0 = 0$

Therefore, maximum value = 2

Therefore, Sensitivity value of Score function = 2

c) $Pr\left(\left| index(\tilde{p}, X^{\uparrow}) - \frac{n}{2} \right| \geq 4 \frac{\ln(m/\alpha)}{\varepsilon} \right) \leq \alpha$

The above inequality explains :-

$\tilde{p}$ - output of exponential Mechanism

X - database X

$index(\tilde{p}, X^{\uparrow})$ - position of $\tilde{p}$ in the sorted database $X^{\uparrow}$

$n/2$ - index of true median $X^{\uparrow}$

m - range of possible salary values $x_i \in [m] = \{1, 2, \dots m\}$

ε - privacy budget.

α - failure probability

Consider a tail bound exponential mechanism:-

$$Pr(S(x, \tilde{p}) \leq S(x, p^*) - t) \leq exp\left(\frac{-\varepsilon t}{2\Delta S}\right)$$

where $S(x, p)$ is score function

$\tilde{p}$ is the output of mechanism

p^* is true median.

$\Delta S = 2$, Sensitivity of Score function

t is the deviation in the score

Let $t = 4 \frac{\ln(m/\alpha)}{\varepsilon}$

$$\Rightarrow Pr\left(S(x, \tilde{p}) \leq S(x, p^*) - \frac{4\ln(m/\alpha)}{\varepsilon}\right) \leq exp\left(-\frac{\varepsilon \cdot \frac{4\ln(m/\alpha)}{\varepsilon}}{2 \cdot 2}\right)$$

$$\Rightarrow Pr\left(S(x, \tilde{p}) \leq S(x, p^*) - \frac{4\ln(m/\alpha)}{\varepsilon}\right) \leq exp\left(-\frac{4\ln(m/\alpha)}{4}\right)$$

$$\Rightarrow exp(-\ln(m/\alpha))$$

$$\Rightarrow \alpha/m$$

$\Rightarrow$ There are m possible values of $p \Rightarrow p \in [m]$. Applying the union bound over all m values the probability that any p deviates by more than t is:

$$Pr\left(\left| index(\tilde{p}, X^{\uparrow}) - \frac{n}{2} \right| \geq t \right) \leq m \cdot \frac{\alpha}{m} = \alpha$$

4) a) Given $y_i := \begin{cases} x_i & \text{with probability } 1/2 + \nu \\ 1 - x_i & \text{with probability } 1/2 - \nu \end{cases}$

$HIV+ \Rightarrow y_i = \frac{1}{2} + \nu \ (x_i)$

$HIV- \Rightarrow y_i = \frac{1}{2} - \nu \ (1 - x_i)$

As random response acheives differential privacy:-

$$\frac{Pr[Y = y \mid x]}{Pr[Y = y \mid x']} \leq e^{\varepsilon}$$ when $HIV+ \Rightarrow x_i = 1$
$HIV- \Rightarrow x_i = 0$

2 Cases:-

Case 1:-

$$\frac{Pr[Y_i = 1 \mid x_i = 1]}{Pr[Y_i = 1 \mid x_i = 0]} = \frac{\frac{1}{2} + \nu}{\frac{1}{2} - \nu}$$

Case 2:-

$$\frac{Pr[Y_i = 0 \mid x_i = 1]}{Pr[Y_i = 0 \mid x_i = 0]} = \frac{\frac{1}{2} - \nu}{\frac{1}{2} + \nu}$$

To determine ε, we consider

$$\frac{Pr[Y = y \mid x]}{Pr[Y = y \mid x']} = \max\left(\frac{\frac{1}{2} + \nu}{\frac{1}{2} - \nu}, \frac{\frac{1}{2} - \nu}{\frac{1}{2} + \nu} \right) = e^{\varepsilon}$$

$$\Rightarrow \varepsilon = \max\left(\ln\left(\frac{\frac{1}{2} + \nu}{\frac{1}{2} - \nu} \right), \ln\left(\frac{\frac{1}{2} - \nu}{\frac{1}{2} + \nu} \right) \right)$$

$$\left[\because \frac{\frac{1}{2} + \nu}{\frac{1}{2} - \nu} > 1 \text{ and } \frac{\frac{1}{2} - \nu}{\frac{1}{2} + \nu} < 1 \right]$$

$$\varepsilon = \ln\left(\frac{1 + 2\nu}{1 - 2\nu} \right)$$

$\therefore$ The value of ε depends on ν

As ν increases, less privacy $[\nu = 0, \varepsilon = 0 \Rightarrow$ perfect privacy$]$

ν decreases, more privacy $[\nu = 1/2, \varepsilon = \infty \Rightarrow$ No privacy$]$

b) Given $\tilde{p} = \frac{1}{N} \sum_{i=1}^{N} y_i$

$$E[\tilde{P}] = \frac{1}{N}\sum_{i=1}^{N} E[Y_i]$$

$$\left(\because \text{ here } E[Y_i] = \left(\tfrac{1}{2}+v\right)x_i + \left(\tfrac{1}{2}-v\right)(1-x_i) \right)$$

$$= \cancel{\tfrac{1}{2}x_i} + v x_i + \tfrac{1}{2} - \cancel{\tfrac{1}{2}x_i} - v + v x_i$$

$$= \tfrac{1}{2} + 2v x_i - v$$

$$= \tfrac{1}{2} + v[2x_i - 1] = \tfrac{1}{2} + \frac{2v}{n}\sum_{i=1}^{N} x_i - 1$$

$$= \tfrac{1}{2} + (2p-1)v$$

$$E(\tilde{P}) = \tfrac{1}{2} + (2p-1)v$$

$$\frac{E(\tilde{P}) - 1/2}{v} = (2p-1) \Rightarrow \frac{\left[\frac{E(\tilde{P})-1/2}{v} + 1\right]}{2} = p$$

$$\Rightarrow \left[\frac{E(\tilde{P}) - 1/2}{2v} + \frac{1}{2}\right] = p$$

$$\boxed{\therefore\ \tilde{P}_0 = \boxed{\frac{\tilde{P} - 1/2}{2v} + 1/2}}$$

$\tilde{P}_0$ – is the unbiased term

$$E(\tilde{P}_0) = E\left[\frac{\tilde{P} - 1/2}{2v} + 1/2\right]$$

$$= \frac{E[\tilde{P}] - 1/2}{2v} + 1/2 = \frac{\cancel{1/2} + (2p-1)\cancel{v} - \cancel{1/2}}{2\cancel{v}} + 1/2$$

$$= \frac{(2p-1)}{2} + 1/2 = \frac{2p-1+1}{2} = \frac{2p}{2} = p \,//$$

$\therefore\ \tilde{P}_0$ is an unbiased estimator of 'p'

c) $\operatorname{var}[\tilde{P}_0] \leq \frac{1}{16v^2 N}$

As in above solution $\tilde{P}_0 = \frac{\tilde{P} - 1/2}{2v} + 1/2$

$$\operatorname{Var}(\tilde{P}_0] = \operatorname{Var}\left[\frac{\tilde{P} - 1/2}{2v} + 1/2\right] = \operatorname{Var}\left[\frac{\tilde{P} - 1/2}{2v}\right]$$

We know that $Var(ax) = a^2 Var(x)$

$$\therefore Var\left(\frac{\tilde{p} - 1/2}{2v}\right) = \frac{1}{4v^2} var(\tilde{p} - 1/2)$$

$$= \frac{1}{4v^2} var(\tilde{p})$$

We know $\tilde{p} = \frac{1}{N}\sum_{i=1}^{N} y_i$

Variance of $\tilde{p}$ $Var(\tilde{p}) = \frac{1}{N^2}\sum_{i=1}^{N} Var(y_i)$

$$E[Y_i] = 1/2 + v(2x_i - 1)$$

$$Var(y_i) = E(y_i)(1 - E(y_i))$$

$$= \left(\frac{1}{2} + v(2x_i - 1)\right)\left(\frac{1}{2} - v(2x_i - 1)\right)$$

$$var(y_i) = \frac{1}{4} - v^2(2x_i - 1)^2$$

$$= \frac{1}{4} - v^2 \quad (\because (2x_i - 1)^2 = 1)$$

$$Var(\tilde{p}) = \frac{1}{N}\left(\frac{1}{4} - v^2\right)$$

$$Var(\tilde{p}_0) = \frac{1}{4v^2} var(\tilde{p})$$

$$= \frac{1}{4v^2}\left(\frac{1}{N}\right)\left(\frac{1}{4} - v^2\right)$$

$$= \frac{1}{4v^2 N}\left(\frac{1}{4} - v^2\right) = \frac{1}{4v^2 N \cdot 4} = \frac{1}{16v^2 N}$$

$$\therefore Var[\tilde{p}_0] \leq \frac{1}{16v^2 N}$$

d) $\tilde{p}$ is computed directly from the randamized response y_i which already satisfy DP. $\tilde{p}_0$ is just a mathematical adjustment of $\tilde{p}$. It doesn't use any additional info / randmness. As DP is immune to post processing means, Once data y_i is released with a DP guarantee, any further computation (like calculating $\tilde{p}_0$) doesn't weaken / strengthen the privacy guarantee. Hence $\tilde{p}_0$ and $\tilde{p}$ have exact same DP.

c) We have $\tilde{P}_U = \dfrac{\tilde{P} - 1/2}{2\nu} + 1/2$

$$E[\tilde{P}_U] = p$$

$$Var[\tilde{P}_U] \leq 1/16\nu^2 N$$

Apply Chebyshev's Inequality:-

$$X = \tilde{P}_U ; \mu = p ; \sigma^2 = Var[\tilde{P}_U]$$

$$Pr[\, |\tilde{P}_U - p| \geq k\sigma) \leq 1/k^2$$

$$\left[\because \sigma = \sqrt{Var[\tilde{P}_U]} \leq \frac{1}{4\nu\sqrt{N}} \right]$$

$$Pr\left[|\tilde{P}_U - p| \geq k \frac{1}{4\nu\sqrt{N}} \right] \leq 1/k^2$$

Let $k = 1/\delta$

$$Pr\left[|\tilde{P}_U - p| \geq 1/\delta \left(\frac{1}{4\nu\sqrt{N}}\right)\right] \leq \delta^2$$

$$|\tilde{P}_U - p| \leq \frac{1}{4\nu\delta\sqrt{N}} \Rightarrow |\tilde{P}_U - p| = O\left(\frac{1}{\nu\sqrt{N}}\right)$$

$$\boxed{\therefore |\tilde{P}_U - p| = O\left(\frac{1}{\nu\sqrt{N}}\right)}$$