Machine Learning from Data CSCI 4100
Assignment 2
Jae Park (661994900)

1. Exercise 1.8

If $v \leq 0.1$, it means that either one or zero red marbles can exist. Using binomial distribution, the probability is

$${}_{10}C_1 \times 0.9^1 \times 0.1^9 + {}_{10}C_0 \times 0.9^0 \times 0.1^{10} = 9 \times 10^{-9} + 1 \times 1 \times 10^{-10} = 9.1 \times 10^{-10}$$

2. Exercise 1.9

Given: $\mu = 0.9$, $v \leq 0.1$. That means we are looking for ε that is $0.8 < |\mu - v|$.

$$P[|\mu - v| > 0.8] \leq 2e^{-2\times(0.8)^2 \times 10} = 5.522\ldots \times 10^{-6.}$$

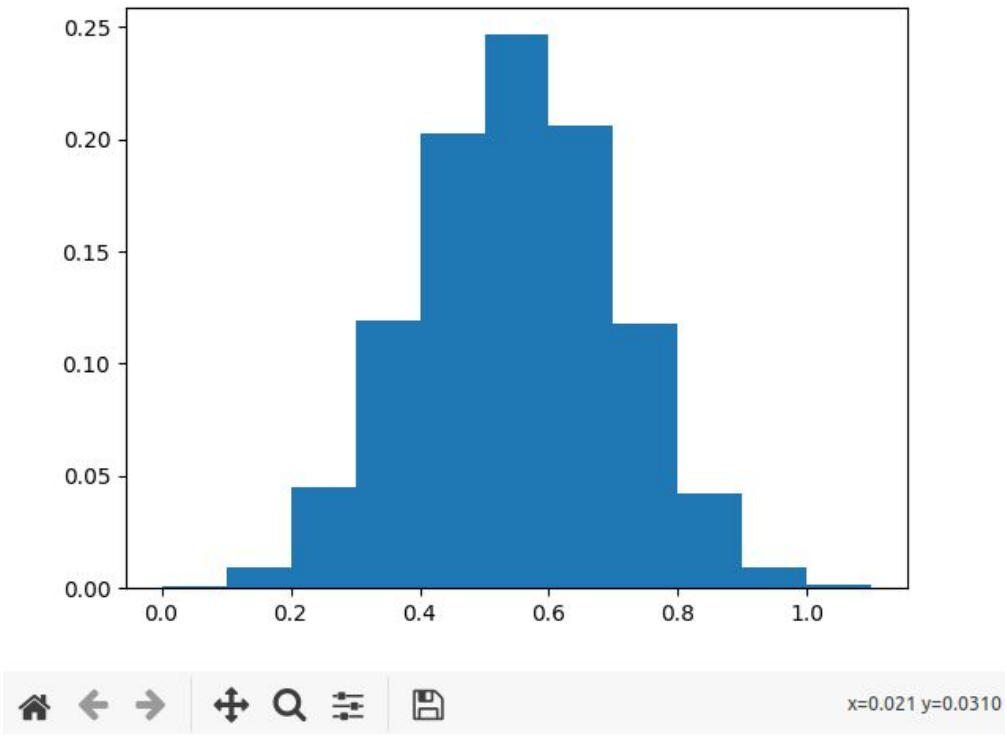
To quantify the relationship between v and μ, we use a simple **upper** bound, the *Hoeffding Inequality*, so it make sense that the result is greater than the actual probability given by 1.8

3. Exercise 1.10

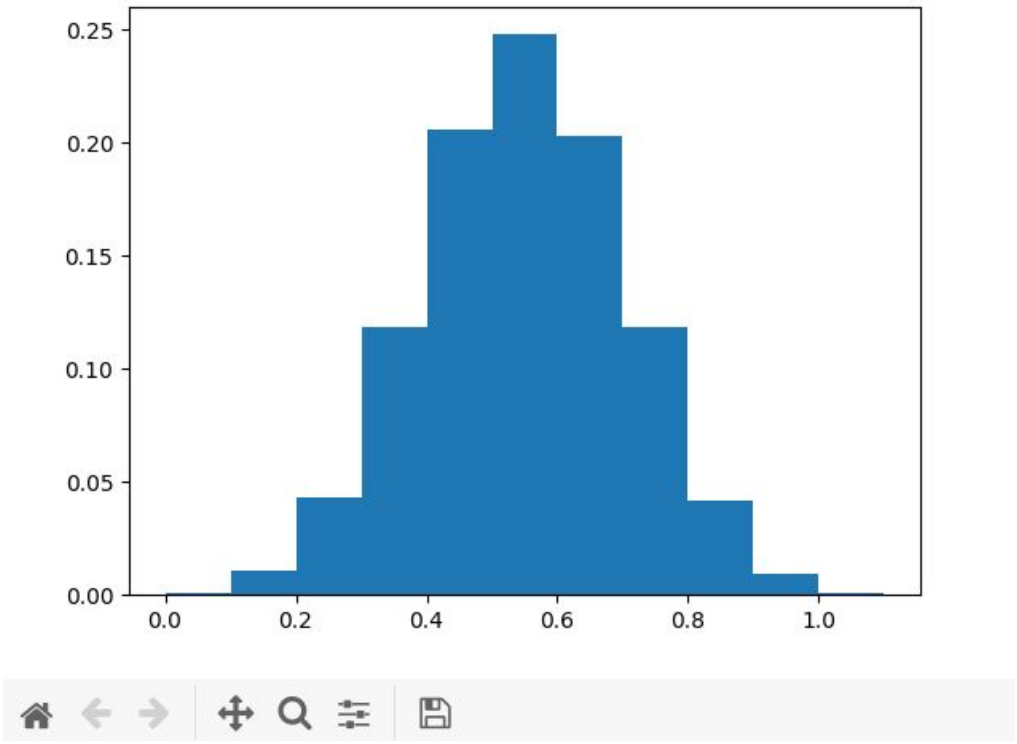
(a) μ for the three coins selected is 0.5 because the probability of getting heads for each coin is 0.5.

(b)

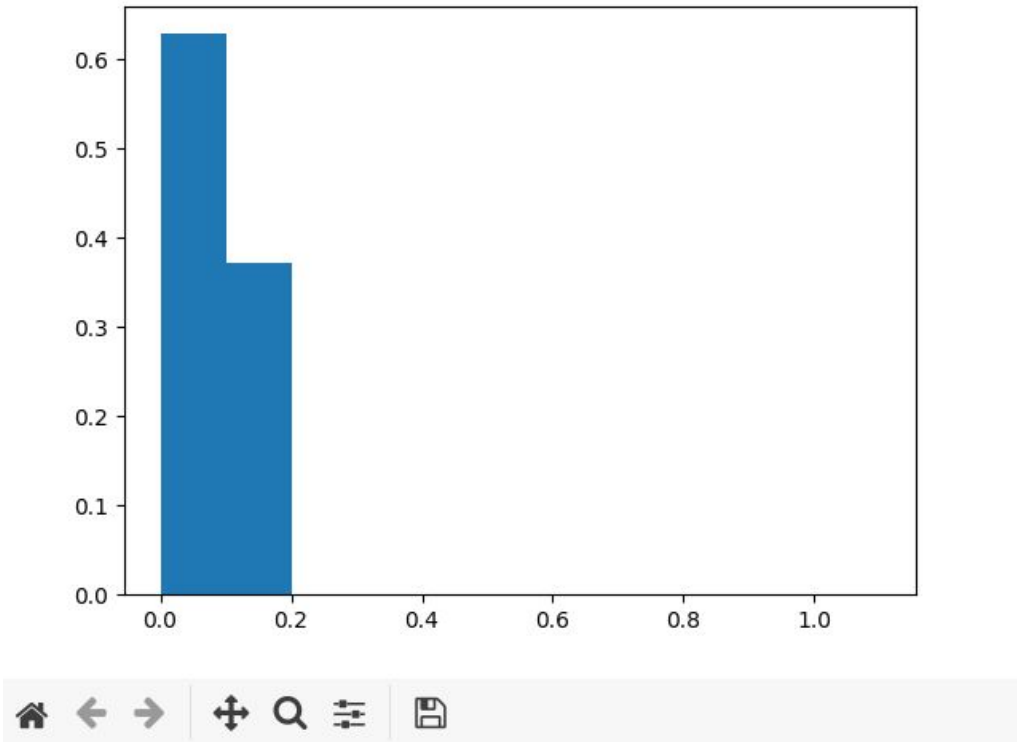
v_1



v_{rand}

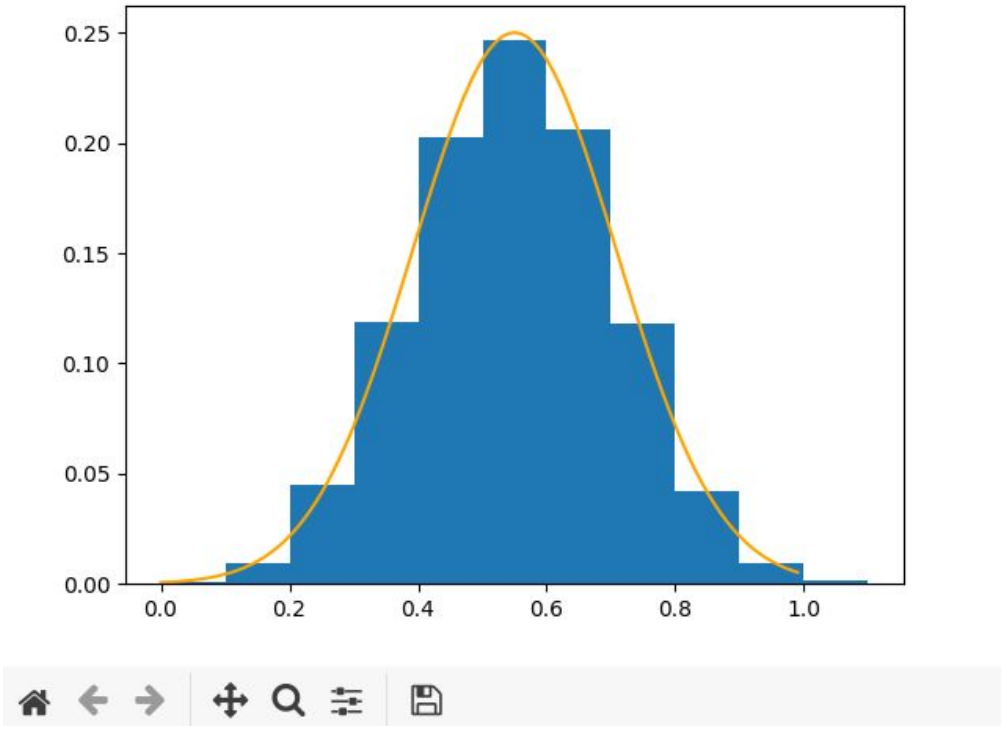

0.25
0.20
0.15
0.10
0.05
0.00
0.0
0.2
0.4
0.6
0.8
1.0

v_{min}

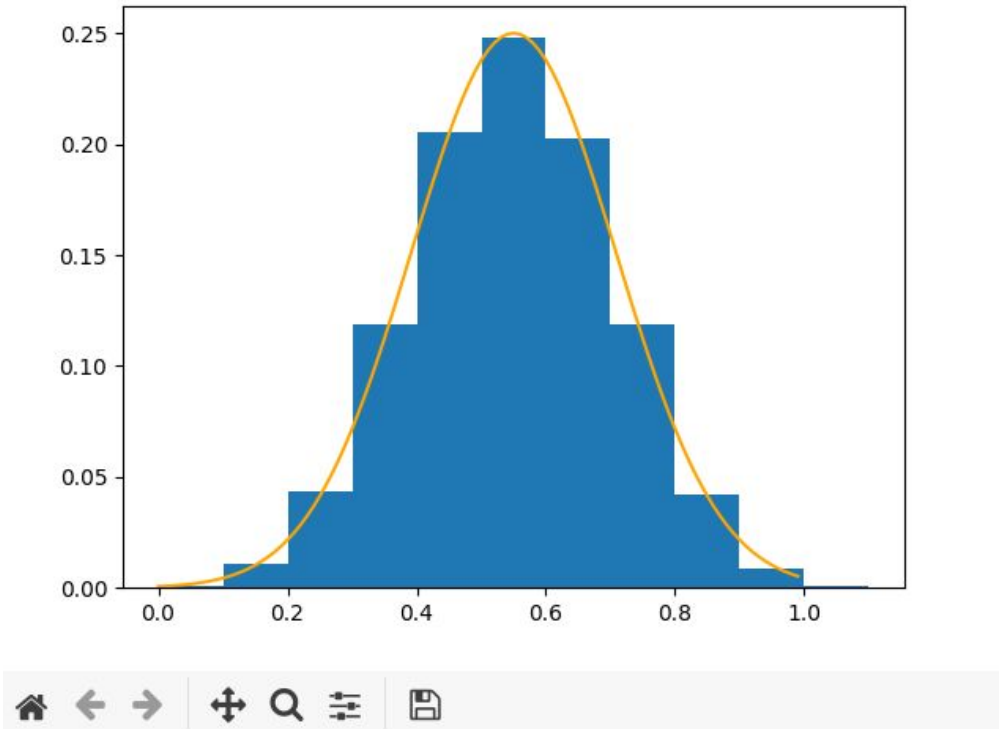

0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.2
0.4
0.6
0.8
1.0

(c)

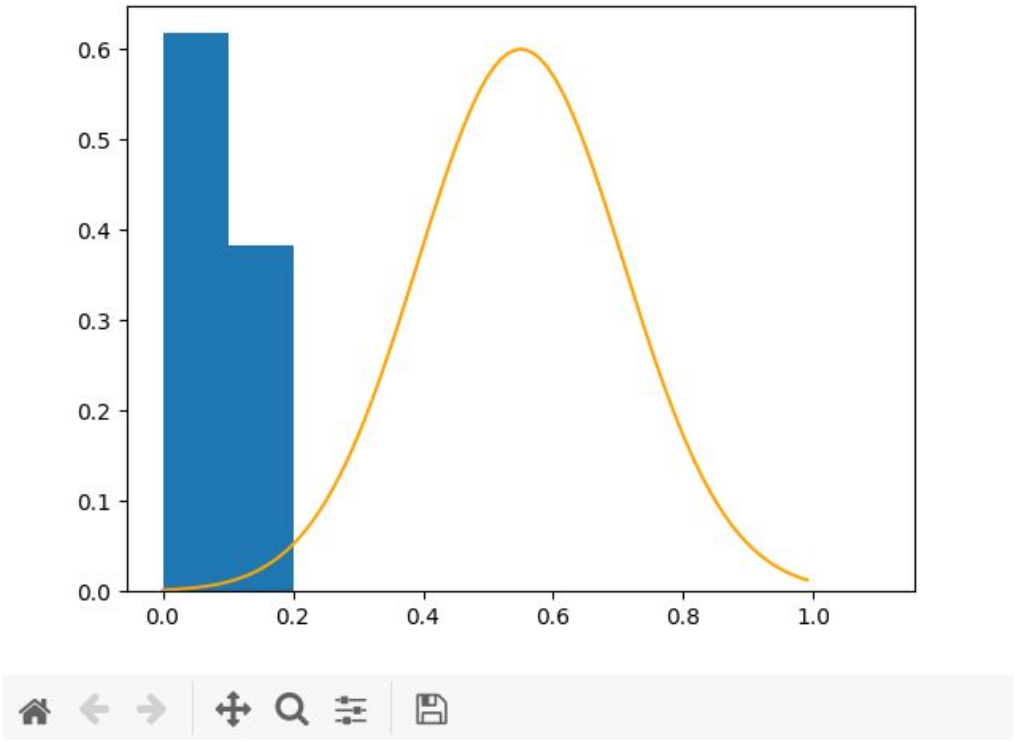
scaled, adjusted distribution of Hoeffding for v_1



scaled, adjusted distribution of Hoeffding for v_{rand}



scaled, adjusted distribution of Hoeffding for v_{min}



(d)

C_1 and c_{rand} obey the Hoeffding bound, but c_{min} didn't. The difference between c_1/c_{rand} and c_{min} is that, for c_{min}, it chooses the minimum frequency of heads, i.e., the coins are not fixed during the experiment. For that matter, c_1 and c_{rand} had a similar performance.

(e)

For the c_1 and c_{rand} case, we may say that each coin represents a bin (model) where the hypothesis h is set before the data is generated. However, as the coins are not fixed, c_{min} cannot be regarded as the bin model.

4. Exercise 1.11

(a)

No. As discussed in the lecture, there is no *guarantee* that S will choose a hypothesis that always performs better than the counterpart outside of the training data. It could have been a sheer luck that a hypothesis was chosen such that it performs well within the

in-sample data set D, but that was done through the process of verification, not real learning.

(b)

Yes. Suppose an extreme case where p is a very small probability, meaning there is very little +1's compared to -1's in Y. And somehow, through random selection, D only contained +1, so our hypothesis S is chosen which says y is always +1. However, we have more -1's in Y, so from the probabilistic point of view, C will perform much better outside of the training data, which tells us y always -1 (opposite of S).

(c)

In order for S to be considered a better hypothesis than C, it has to be the case that more +1's are present out of the 25 training examples than the -1's. So the probability that 13 to 25 +1's will be present out of 25 sample with p 0.9 is

$${}_{25}C_{13} \times 0.9^{13} \times 0.1^{12} + {}_{25}C_{14} \times 0.9^{14} \times 0.1^{11} + \ \ldots \ + {}_{25}C_{25} \times 0.9^{25} \times 0.1^{0} \approx 0.9999998379$$

(d)

No. Assuming that the data set D is "well-chosen" (i.e., IID), it is fair to say that the smart algorithm S will choose a hypothesis that agrees the most with D and will approximate v to μ. Therefore, regardless of what the p value is, S will always choose a better hypothesis between h_1 and h_2, and C will be left to choose whichever the worse hypothesis is.

5. Exercise 1.12

(c) is what we can promise. We cannot guarantee that we will find a hypothesis that achieves $E_{out}(g) \approx E_{in}(g)$, but at least we will know if we find it (pg.25). If we did find it, we can say that with some probabilistic certainty, g will approximate f outside of the sample.

6. Problem 1.3

(a)

Showing $\min_{1 \le n \le N} y_n(w^{*T}x_n) > 0$ is basically same as saying that all of $y_n(w^{*T}x_n)$ is positive for every n, as the minimum of them has to be greater than 0. So the question is asking is $y_n(w^{*T}x_n)$ always positive for every iteration n? And the answer is yes because we said w^* is an optimal set of weight that separates the data. This means that y_n and $w^{*T}x_n$ will always have the same sign. This makes $y_n(w^{*T}x_n)$ to be always positive.

(b)

$$w^T(t)w^* \ge w^T(t-1)w^* + p \qquad \ldots (1)$$

Using the known update rule $w(t+1) = w(t) + y(t)x(t)$,

$$w^T(t-1)w^* + y(t-1)x(t-1)w^* \ge w^T(t-1)w^* + p$$

and replacing p with the proven above in part (a),

$$w^T(t-1)w^* + y(t-1)x(t-1)w^* \geq w^T(t-1)w^* + \min_{1\leq n\leq N}[y_n(t-1)(w^{*T}x_n(t-1)]$$

Cancel out the equivalent left terms and we get a base case that is,

$$y(t-1)x(t-1)w^* \geq \min_{1\leq n\leq N}[y_n(t-1)(w^{*T}x_n(t-1)] \quad \ldots (2)$$

This is certainly true because at iteration *t-1*, or at any iteration as a matter of fact, the min out of $1 \leq n \leq N$ will be smaller than what is on the left hand side of the inequality, in this case $y * x * w^*$ at iteration t-1. Given this holds true, we go back to inequality (1) where the form is simpler.

Base case:

$$w^T(t)w^* \geq w^T(t-1)w^* + p$$

Next step: we look at an iteration before the update, at iteration t-2

$$w^T(t)w^* \geq w^T(t-1)w^* + p_{t-1} \geq w^T(t-2)w^* + p_{t-1} + p_{t-2}$$

where the subscript of p denotes the iteration of update in which p takes place. As demonstrated by inequality (2), we can see that this relationship also hold true because (2) shows that p is of something smaller or equal to than what is actually being updated.

Iterative step:

$$w^T(t)w^* \geq w^T(t-1)w^* + p_{t-1} \geq \ldots \geq w^T(0)w^* + t * p$$

We know that $w(0) = 0$, so $w^T(t)w^* \geq tp$

(c)

Using a simple property $(a+b)^2 = a^2 + 2ab + b^2$,

$||w(t)||^2 = ||w(t-1) + y(t-1)x(t-1)||^2 = ||w(t-1)||^2 + 2 * (y(t-1)w^T(t-1)x(t-1)) + ||y(t-1)x(t-1)||^2$

We know $y(t-1)w^T(t-1)x(t-1) \leq 0$ because x(t-1) was misclassified by w(t-1)

Therefore, $||w(t)||^2 = ||w(t-1)||^2 + 2 *$ (negative vector) $+ ||y(t-1)x(t-1)|| \leq ||w(t-1)||^2 + ||x(t-1)||^2$

(d)

Base case: $||w(0)||^2 \leq 0 * R^2$ holds true because $w(0) = 0$.

Inductive step:

Assume step t-1 holds true, that is

$$||w(t-1)||^2 \leq (t-1)R^2 \quad \ldots.(1)$$

Now, so that it is still true for step t:

From the proof of (c) above, we know that

$$||w(t)||^2 \leq ||w(t-1)||^2 + ||x(t-1)||^2 \quad \ldots(2)$$

We can combine (1) and (2) by substituting $||w(t-1)||^2$ appropriately, and we get

$$||w(t)||^2 \leq (t-1)R^2 + ||x(t-1)||^2$$

$R = \max_{1\leq n\leq N}||x_n||$, so $||x(t-1)||^2$ can be written as R^2.

$$||w(t)||^2 \leq (t-1)R^2 + ||x(t-1)||^2 = (t-1)R^2 + R^2 = (t)R^2$$

$$\therefore ||w(t)||^2 \leq tR^2$$

This shows that induction holds true for the t^{th} step.

(e)

$$w^T(t)w^* \geq w^T(t-1)w^* + p$$
$$\therefore w^T(t)w^* - w^T(t-1)w^* \geq p \quad ...(1)$$

$||w(t)||^2 \leq tR^2$ (square root on both sides)

$\therefore \sqrt{t}R \leq 0 \leq ||w(t)|| \leq \sqrt{t}R$, and the zero bound is given because $||w(t)||$ is always positive, but we will write it as $||w(t)|| \leq \sqrt{t}R$ for the sake of the proof, and this partial inequality is also certainly true.

$$||w(t)|| \leq \sqrt{t}R$$
$$\frac{1}{||w(t)||} \geq \frac{1}{\sqrt{t}R}$$
$$\frac{1}{||w(t)||} \geq \frac{\sqrt{t}}{tR}$$
$$\frac{t}{||w(t)||} \geq \frac{\sqrt{t}}{R} \quad ...(2)$$

multiply (1) and (2) and we get

$$(w^T(t)w^* - w^T(t-1)w^*) \times \frac{t}{||w(t)||} \geq p * \frac{\sqrt{t}}{R}$$

rewritten,

$$\frac{w^T(t)w^* - w^T(t-1)w^*}{||w(t)||} \geq \sqrt{t} * \frac{p}{R}$$

We can see that $\frac{w^T(t)w^*}{||w(t)||} \geq \frac{w^T(t)w^* - w^T(t-1)w^*}{||w(t)||}$ because $w^T(t-1)w^*$ is not negative.

$$\frac{w^T(t)w^*}{||w(t)||} \geq \frac{w^T(t)w^* - w^T(t-1)w^*}{||w(t)||} \geq \frac{p}{\sqrt{t}R}$$
$$\therefore \frac{w^T(t)w^*}{||w(t)||} \geq \frac{\sqrt{t}p}{R}$$

$$\frac{w^T(t)w^*}{||w(t)||} * \frac{R}{p} \geq \sqrt{t} \quad (\frac{R}{p} \text{ is not negative since both positive})$$
$$\frac{R\, w^T(t)w^*}{p\, ||w(t)||} \geq \sqrt{t}$$

use transpose matrix property $a^T b \leq$ ll a ll x ll b ll

$$\frac{R \quad ||w(t)|| * ||w^*||}{p \quad ||w(t)||} \geq \sqrt{t}$$

$$\frac{R \quad ||w^*||}{p} \geq \sqrt{t}$$
$$\frac{R^2 \, ||w^*||^2}{p^2} \geq t$$

7. Problem 1.7

(a)

P[one coin of the sample (10) has v = 0] = p = $(1 - \mu)^{10}$

When $\mu = 0.05$, p = 0.59873...

P[at least one coin of 1 coin(s) has v = 0] = 1 – (1 – p) = 0.598...

P[at least one coin of 1,000 coin(s) has v = 0] = $1 - (1 - p)^{1000} = 1.000...$

P[at least one coin of 1,000,000 coin(s) has v = 0] = $1 - (1 - p)^{1000000} = 1.000...$

When $\mu = 0.8$, p = 1.024×10^{-7}

P[at least one coin of 1 coin(s) has v = 0] = $1 - (1 - p) = 1.024 \times 10^{-7}$

P[at least one coin of 1,000 coin(s) has v = 0] = $1 - (1 - p)^{1000} = 1.024 \times 10^{-4}$

P[at least one coin of 1,000,000 coin(s) has v = 0] = $1 - (1 - p)^{1000000} = 0.09733...$

(b)

i	$\lvert v - \mu \rvert$	$P[\lvert v - \mu \rvert]$
0	1/2	0.0156
1	1/3	0.0938
2	1/6	0.2344
3	0	0.3125
4	1/6	0.2344
5	1/3	0.0938
6	1/2	0.0156

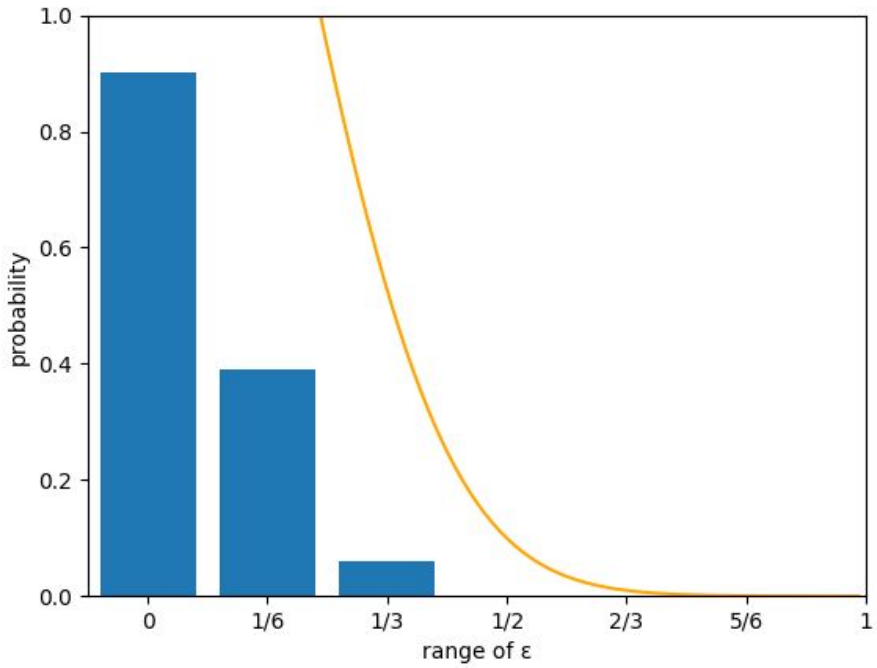
For $0 \leq \varepsilon < 1/6$, i = 3, P = $1 - (0.3125)^2 = 0.9023....$

For $1/6 \leq \varepsilon < 1/3$, i = 2 or 3 or 4 , P = $1 - (0.2344+0.3125+0.2344)^2 = 0.3896...$

For $1/3 \leq \varepsilon < 1/2$, i = 1 or 2 or 3 or 4 or 5,

P = $1 - (0.0938+0.2344+0.3125+0.2344+0.0938)^2 = 0.0612...$

For $1/2 \leq \varepsilon \leq 1$, i = 0 or1 or 2 or 3 or 4 or 5 or 6, P = $1 - (1)^2 = 0$



8. Problem 1.8

(a)

E(t) = $\sum_{x=0}^{\infty} t * P(T = t)$

for some number a,

E(t) = $\sum_{t=0}^{a-1} t * P(T = t) + \sum_{t=a}^{\infty} t * P(T = t)$

$\geq \sum_{t=a}^{\infty} t * P(T = t)$

$\geq \sum_{t=a}^{\infty} a * P(T = t)$ because a = $min_t\ t$

$E(t) \geq$ a * $\sum_{t=a}^{\infty} P(T = t)$ = a * P(T $\geq$ a)

$\frac{E(t)}{a} \geq$ P(T $\geq$ a)

(b)

Let t = $(u-\mu)^2$. Note that E(t) = E($(u-\mu)^2$) = Var(u).

$|u-\mu| \geq \alpha$ is exactly same as t = $(u-\mu)^2 \geq \alpha^2$

Therefore, P[$|u-\mu| \geq \alpha$] = P[t $\geq \alpha^2$].

Y is always non-negative, so applying (a), we get

$$P[(u-\mu)^2 \geq \alpha^2] = P[t \geq \alpha^2] \leq \frac{E(t)}{a^2} = \frac{Var(u)}{a^2} = \frac{\sigma^2}{a^2}$$

Reducing α^2 to α, we get

$$P[(u-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{a}$$

(c)

$$P[(u-\mu)^2 \geq \alpha] = P[t \geq \alpha] \leq \frac{E(t)}{a} = \frac{Var(u)}{a}$$

$$P[(u-\mu)^2 \geq \alpha] \leq \frac{Var(u)}{a} \quad ...(1)$$

$$Var(u) = \frac{1}{N^2} \sum_{i=1}^{N} var(u_i) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N} \quad ... (2)$$

plug in $Var(u)$ ...(2) into ... (1)

$$P[(u-\mu)^2 \geq \alpha] \leq \frac{Var(u)}{a} = \frac{\sigma^2}{N} * \frac{1}{a}$$

$$P[(u-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{Na}$$