1.8:x

Binomial Distribution formula:

$$: \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Let k be number of red ball we get in the draw in sample N

Since N = 10, and v need to less or equal to 0.1 means v = 0.1 or v =0;

So k must be equal to 0 for 1;

So P|v<=0.1| = P|x = 1| + P|x = 0|

P|x=0| = (10, 10) * (0.1^10) *(1-0.1)^10 = 1 *10^-10

P|x=1| = (10, 9) * (0.1*9)*(1-0.1)^1 = 9 * 10^-9

P|x = 1| + P|x = 0| = 9.1 *10 ^ -9

1.9:

Answer: μ = 0.9, and v <= 0.1, so |μ - v| <=0.8. Therefore ϵ = 0.8.

P|μ - v < ϵ| <= 2e^(-2$\epsilon^2 N$), where size is 10, so N = 10,

2e^(-2$\epsilon^2 N$) = 2e^(-2 * 0.8^2 * 10) = 0.00000552

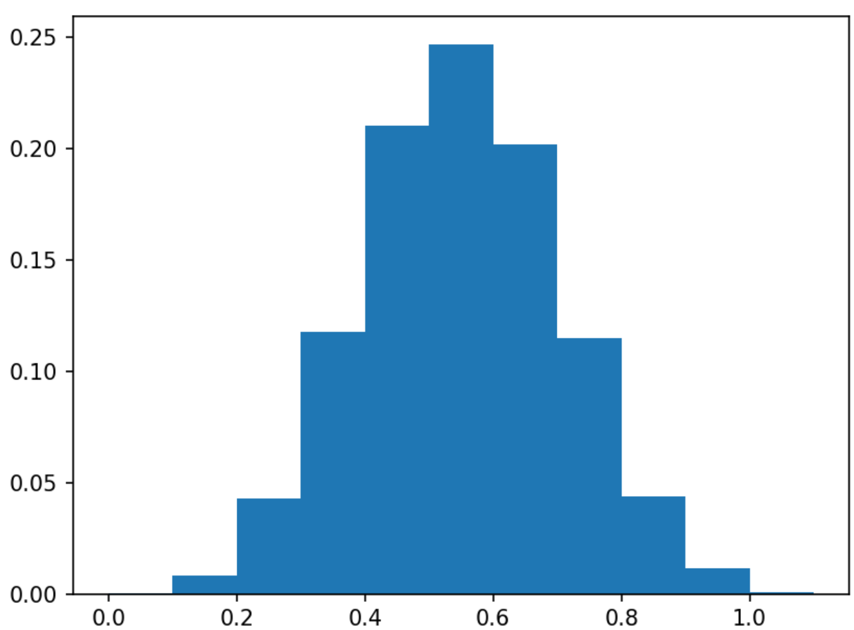
1.10:

Part a:

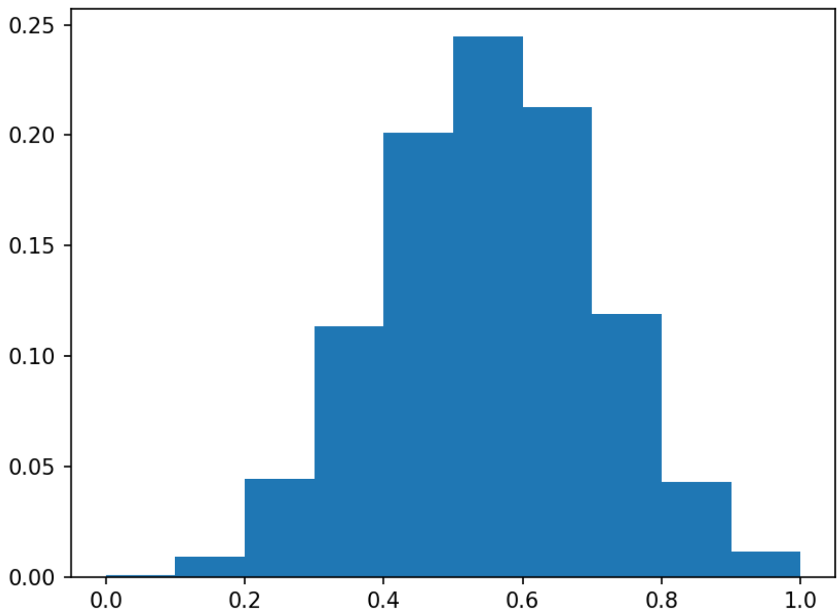
μ is 0.5

Part b:

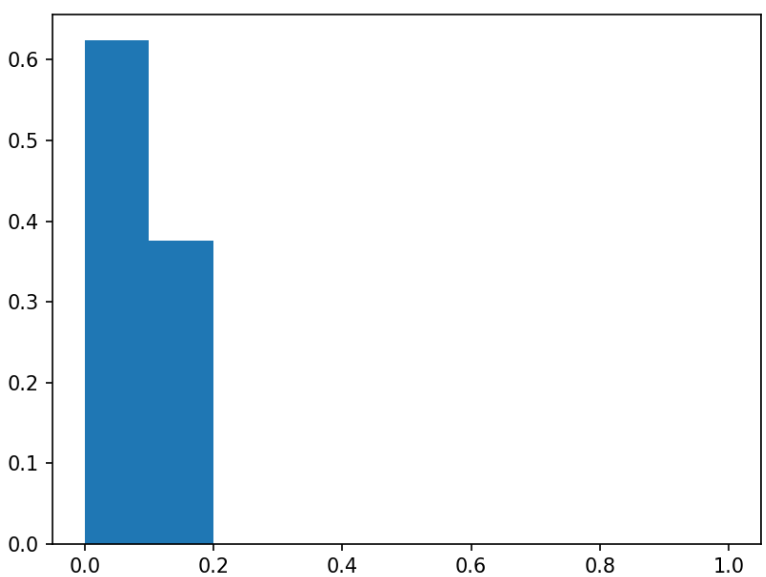
V1:



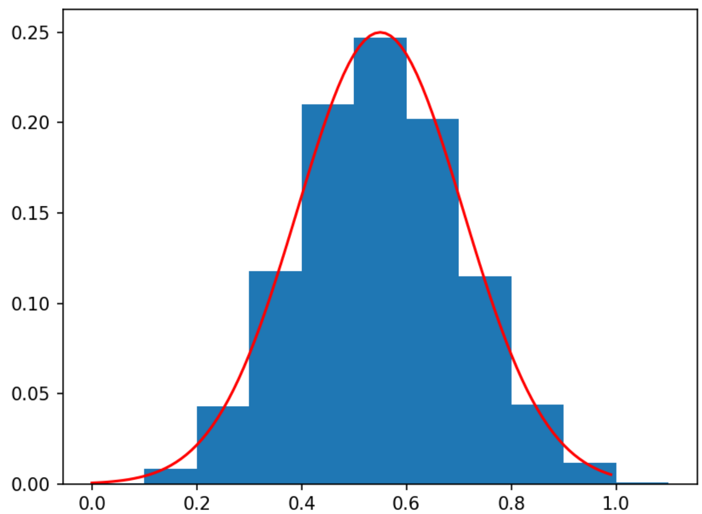
Vr:


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

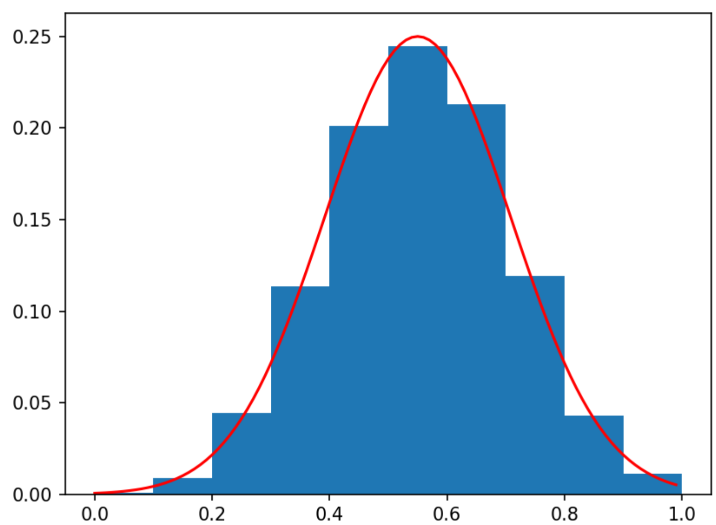
Vm:


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

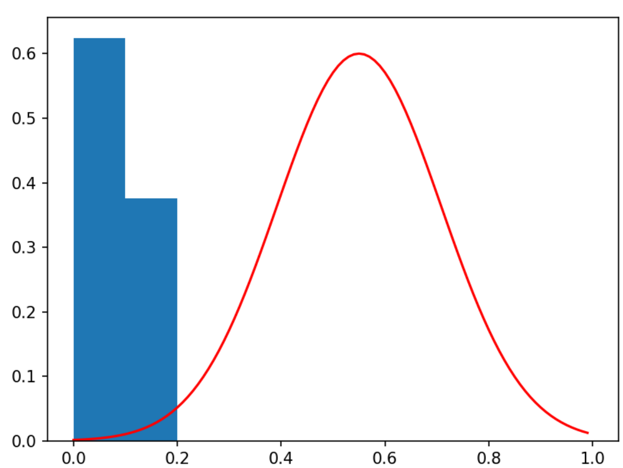
Part c:
V1:


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

Vr:


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

Vm:


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

Part d:
C1 and Cr obey the Hoeffding bound, Cm doesn't, because Cmin's sample didn't pick IID.

Part e:
There are 1000 bins used in the experiment, each bin has a huge number of red and the same number of green as red. For each bin we take 10 balls, let C1 be the number of red balls taken out from the first bin, Crand be the number of red balls taken out from a random bin, Cmin the the minimum number of balls taken out from a bin. In this experiment the new C1, Crand will still obey the Hoeffding bound, and Cmin doesn't.

1.11:
Part a:
Answer: No, it can't be guaranteed. Because we can't guarantee anything outside of D, the sample chosen might be bad and make S performance good in the D but performance not good outside of D.
Part b:
Answer: Yes. Even if p is less than 0.5 which means there is more -1 the +1, but there is still a very small chance that we pick all +1 in D. For example, if p = 0.1, then the probability of getting all +1 is 0.1^25, it's a very small number but it's possible.
Part c:
For S to make a better H than C, we need have at least 13 of 25 data to be +1, so x can be 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25
By Binomial Distribution:
P|x=13| = (25,13) * (0.9)^13 * (1 - 0.1)* (25-13)
P|x=14| = (25,14) * (0.9)^14 * (1 - 0.1)* (25-14)
P|x=15| = (25,15) * (0.9)^15 * (1 - 0.1)* (25-15)
P|x=16| = (25,16) * (0.9)^16 * (1 - 0.1)* (25-16)
P|x=17| = (25,17) * (0.9)^17 * (1 - 0.1)* (25-17)
P|x=18| = (25,18) * (0.9)^18 * (1 - 0.1)* (25-18)
P|x=19| = (25,19) * (0.9)^19 * (1 - 0.1)* (25-19)
P|x=20| = (25,20) * (0.9)^20 * (1 - 0.1)* (25-20)
P|x=21| = (25,21) * (0.9)^21 * (1 - 0.1)* (25-21)
P|x=22| = (25,22) * (0.9)^22 * (1 - 0.1)* (25-22)
P|x=23| = (25,23) * (0.9)^23 * (1 - 0.1)* (25-23)
P|x=24| = (25,24) * (0.9)^24 * (1 - 0.1)* (25-24)
P|x=25| = (25,25) * (0.9)^25 * (1 - 0.1)* (25-25)
P|x=13|+P|x=14|+P|x=15|.....P|x=25| = 0.9999998
Part d:
Answer: No, Because S is making hypotheses based on the simple, so S will always have a better one than C or the same one. Even if p = 0.5, S and C will have the same hypothesis.
1 .12
Choice a:
Wrong, because we can't guarantee anything outside of data samples, so we can't guarantee we can find g approximate to f

Choice b:
Wrong, in some extreme case we might fail to find g
Choice c:
Correct, because we can’t be sure we can learn a g. Even if we find g we can’t guarantee it is close to f, but with high probability it is approximate to f.

Problem 1.3
Part a:
Since w* is already the optimal set of weight, so for any n, y_n and w* x_n will always have the same sign. Because if y_n and w*x_n has different sign means x been misjudged by w, but since w is optimal this won’t happen. Since y_n and w* x_n always have the same sign, so their product has to be a positive number, no matter what n we take. Therefore, $\min_{1\leq n\leq N} y_n(\mathbf{w}^{*\mathrm{T}}\mathbf{x}_n)$ >0, so p > 0.
Part b:

From PLA algorithm: $\mathbf{w}(t+1) = \mathbf{w}(t) + y(t)\mathbf{x}(t).$
w(t) = w(t-1) + y(t-1)x(t-1)
(w(t-1) + y(t-1)x(t-1))^T * (w*) >= w(t-1)^T * (w*) + p
We can get rid off w(t-1)^T * (w*) on both side
y(t-1)x(t-1)^T * (w*) >= p
Substitute p in:
y(t-1)x(t-1)^T * (w*) >= $\min_{1\leq n\leq N} y_n(\mathbf{w}^{*\mathrm{T}}\mathbf{x}_n)$
Since p is chosen n to make $y_n(\mathbf{w}^{*\mathrm{T}}\mathbf{x}_n)$ the smallest,
So y(t-1)x(t-1)^T * (w*) must be greater or equal to p.

Prove $\mathbf{w}^{\mathrm{T}}(t)\mathbf{w}^* \geq t\rho.$ by induction:
Base case: t =0: w^T(0)*(w*) = 0, because w(0) = 0, tp = 0 because t = 0. Therefore, w^T(0)*(w*) >= tp when t = 0.
Induction step: when t = n, assume w^T(n)*(w*) >= np, need to approve w^T(n+1)*(w*) >= (n+1)p

From first half of part b, we know $\mathbf{w}^{\mathrm{T}}(t)\mathbf{w}^* \geq \mathbf{w}^{\mathrm{T}}(t-1)\mathbf{w}^*+\rho$ we plug n+1 for t, we get:
w^T(n+1)w* >= w^T(n)w* + p
Since we have
w^T(n)*(w*) >= np
w^T(n+1)w* >= w^T(n)w* + p >= np + p
Claim proved
Part c:
Since w(t+1) = w(t) + y(t)x(t)
||w(t)||^2 = (||w(t-1)|| + ||y(-1)x(t-1)||)^2
||w(t)||^2 = ||w(t-1)||^2 + ||y(-1)x(t-1)||^2 + 2w(t-1)^T*y(-1)x(-1)

Because x(t-1) misclassified by w(t-1), so 2w(t-1)^T*y(-1)x(-1) <=0
So ||w(t)||^2 = ||w(t-1)||^2 + ||y(-1)x(t-1)||^2 + 2w(t-1)^T*y(-1)x(-1) <= ||w(t-1)||^2 + ||y(-1)x(t-1)||^2
Since y(-1) can only be 1 or -1, in absolute value will always equal to 1
Therefore ||w(t)||^2 <= ||w(t-1)||^2 + ||x(t-1)^2
Claim proved
Part d:
Base case: t = 0: w(0) = 0 = ||w(0)||^2 <= 0*R^2
Induction step: Assume when t = n, ||w(n)||^2 <= nR^2, need to prove ||w(n)+1)||^2 <= (n+1)R^2
(n+1)R^2 = n*R^2 + R^2
From part we know $\|\mathbf{w}(t)\|^2 \leq \|\mathbf{w}(t-1)\|^2 + \|\mathbf{x}(t-1)\|^2.$:
So ||w(n)+1||^2 <= ||w(n)||^2 + ||x(n)||^2
Since we already know ||w(n)||^2 <= nR^2, and since R is the Max of x so R^2 >= ||x(n)||^2\
Therefore: ||w(n)+1||^2 <= ||w(n)||^2 + ||x(n)||^2 <= n*R^2 + R^2 = (n+1)R^2
Claim proved
Part e:

Form part d we have $\|\mathbf{w}(t)\|^2 \leq tR^2$
||w(t)||<=(t^(½))*R

From part b we have $\mathbf{w}^{\mathrm{T}}(t)\mathbf{w}^* \geq t\rho.$, divide rule we got from part b by rule we get from part d, we get: w^T(t)w*/(||w(t)||) >= tp/((t^(½))*R) = t^(1/2)*p/R

w^T(t)w*/(||w(t)||) = ||w*||
||w*|| >= t^(1/2)*p/R
||w*||*R/P >= t^(1/2)
t <= ||w*||^2*R^2/P
Claim proved

Problem 1.7:
Part a:

$$P[k \mid N, \mu] = \binom{N}{k} \mu^k (1-\mu)^{N-k}.$$

μ = 0.05 N = 10 to make v =0 , k =0,
For 1 coin:
1 - (1 - (10,0) * 0.05^0*(1-0.05)^(10-0))^1 = 0.95^10 = 0.5987
For 1000 coin:

1- (1- (10,0) * 0.05^0*(1-0.05)^(10-0))^1000 approximate equal to 1(can't get anything from calculator)
For 1000000 coin:
1- (1- (10,0) * 0.05^0*(1-0.05)^(10-0))^1000000 approximate equal to 1
For μ = 0.8
For 1 coin:
1 - (1 - (10,0) * 0.8^0 * (1-0.8)^(10-0))^1 = 0.2^10 = 1.024*10^-7
For 1000 coin:
1 - (1 - (10,0) * 0.8^0 * (1-0.8)^(10-0))^1000 = 1.02394763*10^(-4)
For 1000000 coin:
1 - (1 - (10,0) * 0.8^0 * (1-0.8)^(10-0))^1000000 = 0.09733159268

Part b:
V =N/k, N = 6, μ = 0.5
Since there is 2 coin
P|max vi - μi <= ε] = 1 - (P(|v - μ| <= ε))^2 = 1 - (P(|k/N - 0.5| <= ε))^2 = 1 - (P(|k/6 - 0.5| <= ε))^2
= 1 - (P(|k - 3| <= 6ε))^2 = 1 - (P(| 3 - 6ε <= k <=3 + 6ε))^2

3 - 6ε <= k <=3 + 6ε = $\sum_{3 - 6\varepsilon <= k <=3 + 6\varepsilon} (\frac{N}{k})\, \mu^{k}(1 - \mu)^{1-k}$

P|max vi - μi <= ε] = 1 - ($\sum_{3 - 6\varepsilon <= k <=3 + 6\varepsilon} (\frac{N}{k})\, \mu^{k}(1 - \mu)^{1-k}$)^2

Graph: