Statistics Assignment Week 8 - John Sinclair - 16325734

Q1.

(a) Y is a poor estimate of the fraction of students studying to pass because:
   (i) N is not representative of the entire class, many students would likely not respond or complete the poll leading to an incorrect representation of the class.
   (ii) Of those who respond, a portion of those would likely respond untruthfully and so Y would be a poor estimate of the fraction of students who are just studying to pass.

(b) Even with many repetitions the experiment fails to account for those who don't fill out the poll and for those who lie, a better design for the experiment would be to make it mandatory, ensuring N is an accurate representation of the class and make it anonymous, reducing the risk of untruthful responses.

Q2.

(a) Definition of independence: P(AnB) = P(A)P(B), so two events are said to be independent if the occurrence of one event does not affect the probability of occurrence of the other. If the events are identically distributed, the probability distribution of both the events is equal.

(b) Y is not a random variable, it is the sum of the random variables divided by the number of random variables, i.e. the mean of the random variables.

(c) Chebyshev's inequality:
We know μ = 0.1, σ = sqrt(μ(1 - μ)) and N = 100
95% confidence interval:

μ - (σ / sqrt(0.05 * N)) <= Y <= μ + (σ / sqrt(0.05 * N))

σ / sqrt(0.05 * N) = (sqrt(0.1 * (1 - 0.1)) / sqrt(0.05 * 100)) = 0.13416

0.1 - 0.13416 <= Y <= 0.1 + 0.13416
-0.03416 <= Y <= 0.23416

(d) CLT formula: (Y - μ) / (σ / sqrt(N))
μ = 0.1
σ = sqrt(μ(1 - μ)) = sqrt(0.1 * (1 - 0.1)) = 3 / 10
sqrt(N) = sqrt(100) = 10
σ / sqrt(N) = (3 / 10) / 10 = 0.03

-1.96 * 0.03 <= Y - 0.1 <= 1.96 * 0.03
-1.96 * 0.03 + 0.1 <= Y <= 1.96 * 0.03 + 0.1
0.0412 <= Y <= 0.1588

While the Central Limit Theorem only requires the mean and variance to fully describe, it only provides an approximation whereas Chebychev provides an actual bound and work for all sample sizes.

(e) Code:

```
N = 100;
iterations = 10000;
averages = zeros(1,iterations);
for i = 1:iterations
 randoms = randi([0 1], 1, N);
 tempAvg = 0;
for j = 1:N
 tempAvg = tempAvg + randoms(j);
end
 averages(i) = tempAvg / N;
end
histogram(averages)
```

PMF Plot: