Statistical Methods for Computer Science

Assignment 8 — STU33009

1. (a)
   - It is not necessary that all of the students in the class would respond to the poll. The students who do respond are likely to be the ones that are more involved in the module and, therefore, are less likely to be the ones simply “studying to pass". This could significantly skew the estimate.
   - Since the aim is to estimate the fraction of TCD CS students that are “studying to pass", taking a sample from the ST3009 module mightn’t account for all of the TCD CS students.

   (b)
   - Survey TCD CS students from all modules.
   - Make it mandatory to respond to the survey so that all students respond.

2. (a) Two events are independent if the occurrence of one event does not effect the occurrence of the other. The independence of two Bernoulli random variables, X and Y, can be represented as follows,

$$P(X = 1 \cap Y = 1) = P(X = 1) \cdot P(Y = 1)$$

Two Bernoulli random variables are identically distributed if their probability distributions are the same.

(b) Y is not a random variable. It is the mean of the Bernoulli random variables.

(c) By Chebyshev’s inequality, we can say,

$$P(|Y - \mu| \geq \epsilon) \leq \frac{\sigma^2}{N\epsilon^2}$$

Since we want 95% confidence, we can say,

$$P(|Y - \mu| \geq \epsilon) \leq 0.05$$

Therefore,

$$\frac{\sigma^2}{N\epsilon^2} = 0.05$$

We want ϵ, so we can make it the subject of the equation,

$$\epsilon = \sqrt{\frac{\sigma^2}{0.05N}}$$

Since this is a Bernoulli distribution, we can say,

$$Var(X) = \sigma^2 = \mu \cdot (1 - \mu)$$

$$\sigma^2 = 0.1 \cdot (1 - 0.1) = 0.1 \cdot 0.9 = 0.09$$

Subbing in the values, we get,

$$\epsilon = \sqrt{\frac{0.09}{0.05 \cdot 100}} = \sqrt{0.018} = \frac{3\sqrt{5}}{50} = 0.1341640786$$

Subbing this into the Chebyshev's inequality, we get,

$$P\left(|Y - \mu| \geq \frac{3\sqrt{5}}{50}\right) \leq 0.05$$

We can re-write this as an inequality with 95% confidence, like so,

$$|Y - \mu| \leq \frac{3\sqrt{5}}{50}$$

$$-\frac{3\sqrt{5}}{50} \leq Y - \mu \leq \frac{3\sqrt{5}}{50}$$

$$\mu - \frac{3\sqrt{5}}{50} \leq Y \leq \mu + \frac{3\sqrt{5}}{50}$$

Subbing in for μ, we get,

$$0.1 - \frac{3\sqrt{5}}{50} \leq Y \leq 0.1 + \frac{3\sqrt{5}}{50}$$

$$-0.03416407865 \leq Y \leq 0.2341640786$$

(d) A 95% confidence interval with CLT is given by two standard deviations, like so,

$$P(-2\sigma \leq Y - \mu \leq 2\sigma) \approx 0.95$$

where,

$$2\sigma = 2\sqrt{\frac{\sigma^2}{N}}$$

From the previous part, we already know that $\sigma^2 = 0.09$. Filling in the values, we get,

$$2\sigma = 2\sqrt{\frac{0.09}{100}} = 2\sqrt{0.0009} = 2 \cdot 0.03 = 0.06$$

Therefore,

$$-0.06 \leq Y - \mu \leq 0.06$$

$$\mu - 0.06 \leq Y \leq \mu + 0.06$$

Subbing in for μ, we get,

$$0.1 - 0.06 \leq Y \leq 0.1 + 0.06$$

$$0.04 \leq Y \leq 0.16$$

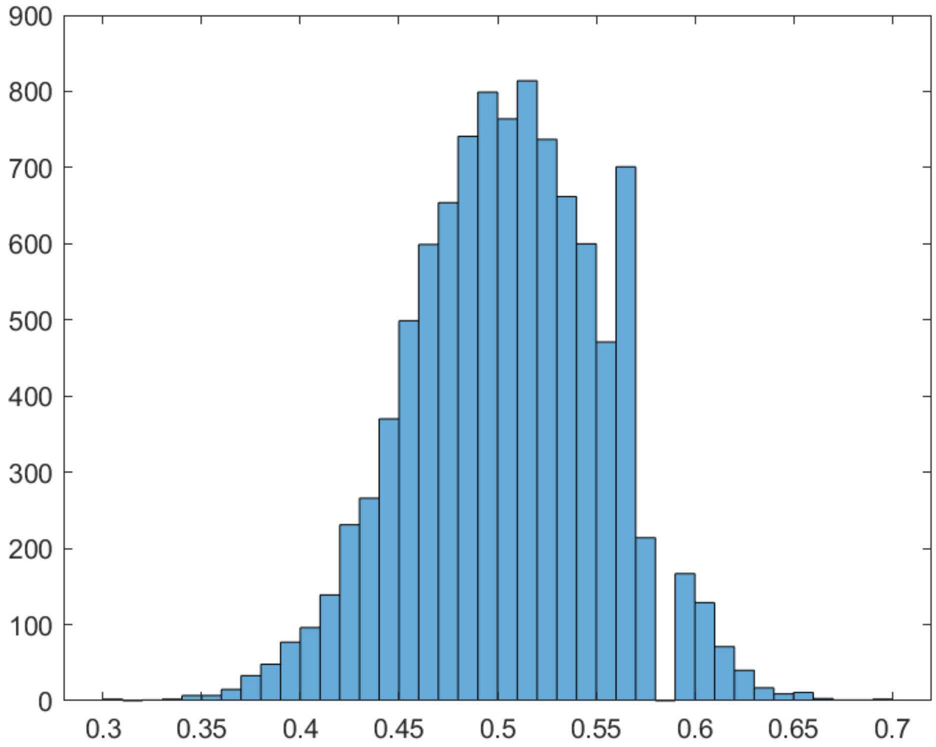
From the results, it is evident that the Central Limit Theorem tends to be more optimistic in its confidence intervals when compared to Chebyshev's Inequality.

(e) For the CLT, the 95% error bound is given by,

$$2\sigma = 2\sqrt{\frac{\sigma^2}{N}}$$

This can be translated into our matlab code as a function.

```
N = 100;
iters = 10000;

sample_means = zeros(1, iters);

for i = 1:iters
    sample = randi([0,1], 1, N);
    sample_means(i) = mean(sample);
end

histogram(sample_means)

m = mean(sample_means);     % mean of sample means
v = var(m);                 % variance
e = clt_error(v, iters)     % error bound

fprintf('%f <= Y <= %f\n', m-e, m+e)

% Bernoulli variance formula
function v = var(mean)
    v = mean*(1-mean);
end

% 95% clt error formula
function e = clt_error(var, N)
    e = 2*sqrt(var/N);
end
```



```
>>
e =

    0.0100

0.490288 <= Y <= 0.510288
```

It's evident that our error is much smaller when compared to our previous confidence intervals. For this reason, our confidence interval is much more optimistic in its estimation.