Samuel Petit
17333946

Week 8 - St3009

Question 1 -

a) The first way in which this approach may lead to Y being a poor estimate is that students may not be honest when filling the poll. This would lead to invalid data and any conclusions that come out of that study could be questioned.
Secondly, not all students will fill the poll and the poll is only sent to 3rd year students. It means that the data would be the fraction of a fraction so the obtained data may be quite small and ~~so~~ could not be an accurate representation of the students trying to pass.

b) Repeating the experiment many times would still have to deal with the above problems, human error would still be a problem and the people surveyed still represent a very small representation of the population.
Instead we could repeat the poll multiple

times during the year, honesty in the answers would still be a problem but we could instead observe the amount of people answering the poll.
We could also make the poll mandatory and open it to all years/classes to have more data.

Question 2.

a) If we have iid (Bernoulli) random variables, then we have:

$$\sum_{k=1}^{N} X_k \sim Bin(n, p).$$

b) Y is not a random variable.

c) Chebyshev's inequality gives us:

$$\mu - \frac{\sigma}{\sqrt{0.05N}} \leq Y \leq \mu + \frac{\sigma}{\sqrt{0.05N}}$$

$$\Leftrightarrow 0.1 - \frac{\sqrt{0.1(1-0.1)}}{\sqrt{100 \times 0.05}} \leq Y \leq 0.1 + \frac{\sqrt{0.1(1-0.1)}}{\sqrt{100 \times 0.05}}$$

$$\Leftrightarrow -0.034164 \leq Y \leq 0.234164.$$

d)

Q2 d) We have:

$$\frac{\frac{X_1+\dots+X_n}{n}-\mu}{\frac{\sigma}{\sqrt{N}}} = \frac{Y-\mu}{\frac{\sigma}{\sqrt{N}}}$$

With a 95 % confidence interval we use -1.96 and 1.96.

$$-1.96 \le \frac{Y-\mu}{\frac{\sigma}{\sqrt{N}}} \le 1.96$$

$$\Leftrightarrow \quad -1.96 \le \frac{Y-0.1}{\frac{\sqrt{0.1(1-0.1)}}{\sqrt{100}}} \le 1.96$$

$$-1.96 \cdot \frac{\sqrt{0.1(1-0.1)}}{\sqrt{100}} + 0.1 \le Y \le 1.96 \cdot \frac{\sqrt{0.1(1-0.1)}}{\sqrt{100}} + 0.1$$

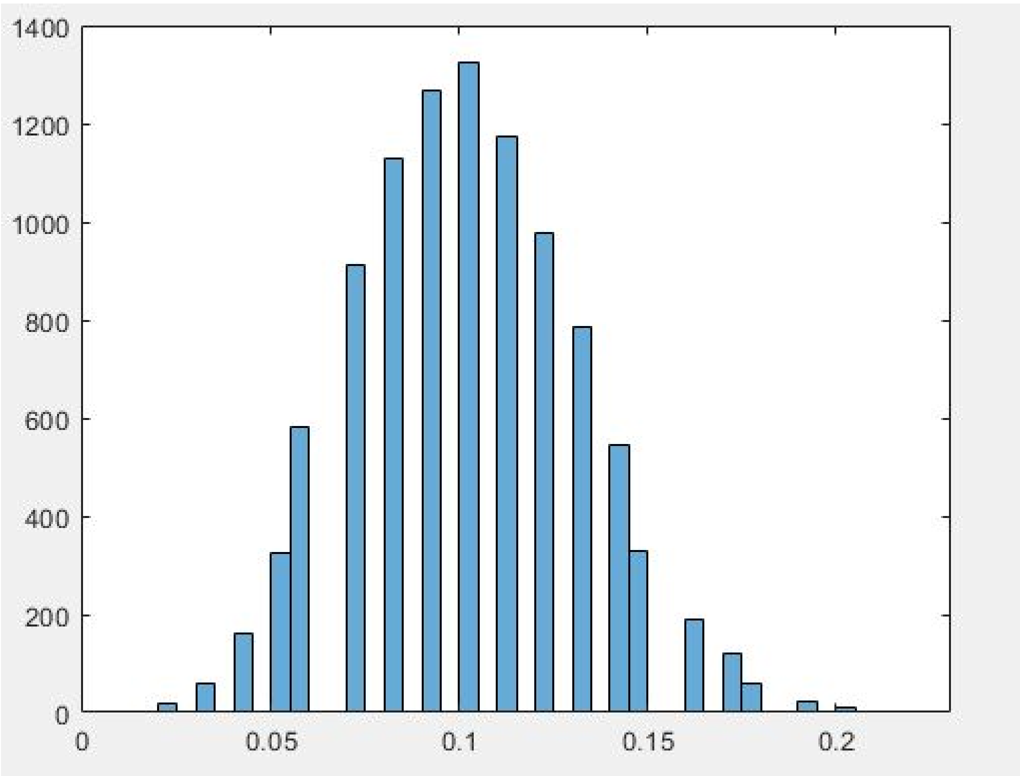
$$= 0.0412 \le Y \le 0.1588.$$

CLT gives a full distribution of $\bar{X}$ (or in this case Y), it only works for finite Ns however, it only requires a mean and a variance.
Chebyshev provides a bound instead of an approximation, works for all Ns but is loose in general.

Question e)

```
Size = 10000;
N = 100;
finalMean = zeros(1, Size);
for c = 1:Size
    currentSample = binornd(N,0.1);
    finalMean(c) = mean(currentSample) / N;
end
histogram(finalMean);

sampleMean = mean(finalMean);
sampleStandardDeviation = std(finalMean);
sampleStandardErrorOfMean = sampleStandardDeviation/sqrt(Size);
confidenceIntervalOfMean = sampleMean + 1.96*sampleStandardErrorOfMean*[1; -1];
disp(confidenceIntervalOfMean);
disp(sampleMean);
```




```
>> testmatlab
    0.1004
    0.0992
```

We notice that the confidence interval generated by matlab is indeed included in our own calculated one in parts c and d which are slightly wider.