STU33009 Week 8 Assignment

Efeosa Louis Eguavoen - 17324649

March 26, 2020

1 Question 1

Question 1. I want to estimate what fraction of TCD CS students are "studying to pass". To do this I email a poll to the students in third year taking the ST3009 module, and N students reply. Let Xi be a random variable which takes value 1 when the i'th student who replies says they are studying to pass and 0 otherwise. I then estimate the fraction studying to pass as Y =1/N * (Sum of Xi where i =1) i.e. the fraction of respondees who reply that they arestudying to pass.
(a) Discuss two ways in which this approach may lead to Y being a poor estimate of the fraction of students studying to pass.
(b) Discuss the random experiment here i.e. the experiment that we can repeat many times and so use the frequency interpretation of probability. How does this relate to your answer in (a)? What might be a better way to design the experiment?

(a) It's a poor estimation on how many are studying to pass as there's no obligation for the students to reply truthfully to the poll which would skew our results, also we may not get enough people to answer the poll to get a good estimate of the overall population.
(b) The random experiment here of polling students to see if they're studying to pass must be done independantly so students don't know how other students are voting to make sure results aren't skewed, also there should be some sort of blind testing done so students can't be identified. This would prevent the problem of students not answering truthfully. As for getting a large enough sample size so we can use the frequency of interpretation of probability, we should poll randomly in person instead of emailing a poll to increase the response rate.

2 Question 2

Suppose I have N = 100 independent and identically distributed Bernoulli random variables X1, . . .,XN with mean $\mu = 0.1$.
(a) Using the definition of independence etc, state in mathematical terms what it means for two Bernoulli random variables to be "independent and identically distributed".
(b) Let Y = 1 N PN i=1 Xi . Is Y a random variable ? If so, what is its mean and variance ?

(c) Use Chebyshev's inequality to give a 95 percnet confidence interval for Y .
(d) Compare your answer in (c) with the confidence interval obtained using the Central Limit Theorem. Discuss the pros and cons of these two approaches (Chebyshev and CLT) to deriving a confidence interval.
(e) Write a short matlab simulation that generates N independent and identically distributed Bernoulli random variables and calculates their empirical mean. By running this simulation 10,000 times plot an estimate of the PMF of the emprical mean (include both the code and the plot in your submitted answer). Using this, estimate a 95 confidence interval and compare this with the confidence intervals calculated in (c) and (d).

(a) For 2 Bernoulli random variables to be both independant and identically distrubuted where X = Berx(p),Y=Bery(p): X = Y for them to be identically distributed
P(intersection(X,Y)) = X*Y
(b) Yes, Y is a Binomial random Variable, it's mean is 1/N*sum(N,E(Xk) where k =1) = μ = 0.1.
The variance is equal to

$$\sigma^2/N, andgivenp = \mu = 0.1, weget(1-p)(0-p)^2 + p(1-p)^2 = 0.09$$

$$\sigma^2/N = 0.09/100$$

which is our variance.
(c)

$$0.1 - 0.3/\sqrt{.05(100)} \leq samplemean(xhat) \leq 0.1 + 0.3/\sqrt{.05(100)}$$

(d)

$$-2\sigma \leq X - \mu \leq 2\sigma$$

$$-0.6 \leq X - 0.1 \leq 0.6$$

Chebyshev inequality gives an actual bound for the data instead of an approximation and works for all values of N unlinke CLT but it gives a very loose bound on the data. CLT requires only the mean and standard deviation to describe the distribution but we don't know how accurate it is as we may not know for sure how large N should be.
(e)

```
vals = zeros(100);
for i=1:10000
    p=0.1;
    A=(rand(10)<p);
    meanc = mean(A,'all');
    vals(i) = meanc;
    disp(vals(i));
end
disp(vals);
```

```
histogram(vals);
```