TRINITY COLLEGE DUBLIN

School of Computer Science and Statistics

Week 8 Questions ST3009: Statistical Methods for Computer Science

For each problem, explain/justify how you obtained your answer in order to obtain full credit. In fact, most of the credit for each problem will be given for the derivation/model used as opposed to the final answer.

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 X_i 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 $Y = \frac{1}{N}\sum_{i=1}^{n} X_i$ i.e. the fraction of respondees who reply that they are studying 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?

Question 2. Suppose I have $N = 100$ independent and identically distributed Bernoulli random variables $X_1, \ldots, X_N$ 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 $Y = \frac{1}{N}\sum_{i=1}^{N} X_i$. Is Y a random variable ? If so, what is its mean and variance ?

(c) Use Chebyshev's inequality to give a 95% 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).