CS6015: Linear Algebra and Random Processes
Assignment - 2
Course Instructor : Prashanth L.A.
Due : Nov-6, 2017

Let $X_1, \ldots, X_N$ denote a sequence of i.i.d. random variables (r.v.s), each with mean μ and variance σ^2. Let $\overline{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ denote the sample mean.

1. What is $\mathbb{E}(\overline{X}_N)$ and Var$(\overline{X}_N)$? (1 mark)

2. While a concentration inequality was derived for Bernoulli r.v.s in the class, a similar result holds for bounded r.v.s and we present the well-known Hoeffding inequality below.

   Theorem 1. *Let $X_1, \ldots, X_N$ denote a sequence of i.i.d. random variables (r.v.s) with $X_i \in [a, b]$, for all i, where $-\infty < a \leq b < \infty$. Letting $\overline{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ and μ denote $\mathbb{E}X_i$, for all i, we have*

   $$\mathbb{P}\left(\overline{X}_N - \mu \geq \epsilon\right) \leq \exp\left(-\frac{2N\epsilon^2}{(b-a)^2}\right) \text{ and } \mathbb{P}\left(\overline{X}_N - \mu \leq -\epsilon\right) \leq \exp\left(-\frac{2N\epsilon^2}{(b-a)^2}\right). \tag{1}$$

   Use (1) to arrive at the following equivalent form: For $\delta \in (0, 1)$ and $\epsilon' > 0$,

   $$\mathbb{P}\left(\mu \in [\overline{X}_N - \epsilon', \overline{X}_N + \epsilon']\right) \geq 1 - \delta. \tag{2}$$

   Give an explicit expression for ϵ' as a function of N and δ.
   Hint: Notice that the probability of the event, which is complementary to the one on LHS of (2), is at most δ. Compare this with the form in (1) and pick a suitable δ using the RHS of (1). (2 marks)

3. Write a program (in your favorite language) to obtain N samples from a Poisson distribution with parameter $\lambda = 10$.

   (a) Choose the number of samples N from the set $\{10, 100, 1000, 10000\}$.
   (b) For each value of N, repeat the experiment 10000 times.
   (c) Store the sample mean value $\overline{X}_N$ from each of the 10000 replications.
   (d) Plot the histogram of the sample mean $\overline{X}_N$, with 1000 bars. (5 marks)

   Interpret the numerical results and answer the following:

   (a) Is the sample mean close to the true mean? Why is this expected? Justify your answer. (2 marks)
   (b) How many times was the sample mean in the interval $[9.99, 10.01]$?
   How about $[9.9, 10.1]$? Answer this for various choices of N. (2 marks)
   (c) Calculate a 95% confidence interval for the sample mean using the numerical results. How many times did the true mean fall outside the confidence interval? (2 marks)
   (d) BONUS: Why isn't Theorem 1 applicable for Poisson r.v.s? Approximate Poisson with parameter λ by a Binomial distribution with parameters n and λ/n. Apply the equivalent Hoeffding bound (from the answer to question 2) to the latter distribution and calculate the 95% confidence interval. Compare the latter theoretical confidence interval with those obtained numerically. (3 marks)

(e) If one wants an accuracy of 0.1 (i.e., the absolute difference between sample mean and true mean), how many samples N would be necessary? If the accuracy is to be 0.01, by how much would the number of samples N increase? Generalize the answer, i.e., if the accuracy increases by a decimal place, what would be the corresponding jump in N? (3 marks)

4. Consider a random variable X that takes values $\pm 1, \pm 2, \ldots$. with p.m.f. f defined as

$$f(k) = \frac{A}{k^2} \text{ for } k = \pm 1, \pm 2, \ldots$$

(a) For what choice of A would f be a valid p.m.f., i.e., $\sum_{k \neq 0} f(k) = 1$? Justify your answer. (1 mark)

(b) Generate N samples using the p.m.f. f defined above[1] and plot the histogram of the sample mean $\overline{X}_N$ as in the previous question. Interpret the results you obtain for $N = 1000$ and $N = 10000$. In particular, answer if the sample mean concentrates around, i.e., stays close to, some value? Compare the 95% confidence intervals for the two choices of N. Is this behavior of sample mean expected? Justify your answer. (4 marks)

Here is what you have to submit:

1 Hand-written (or typed) answer.

2 Hand-written (or typed) answer.

3 Submit the source code, preferably one that is readable with some comments. Also, include all the histograms in a document or submit printouts of plots.

3a Hand-written (or typed) answer.

3b Tabulated results for various N.

3c Tabulated results for various N.

3d Hand-written (or typed) answer.

3e Hand-written (or typed) answer .

4 Submit the source code, preferably one that is readable with some comments. Also, include all the histograms in a document or submit printouts of plots.

4a Hand-written (or typed) answer.

4b Hand-written (or typed) answer.

* For each hand-written (or typed) answer, provide concrete justification.
** Barring the bonus question, the total marks in this assignment is 20 and during course grading, the score obtained would be halved, leading to a contribution of 10% in the final grade. However, half of the marks obtained for the bonus question would be added to the total score separately in the grade calculation.

[1] For simulating from the p.m.f. f, use the procedure described in example 100 of `http://math.iisc.ernet.in/~manju/UGstatprob16/statprob.pdf`