Homework #06 • MATH495/STAT490 • Variances of Estimated PDFs & CDFs

- submit your write-up before 12 noon on **Thursday** 23 October.
- page limits will be enforced.
- highlight major results.
- **please indicate partners in collaborative efforts. Thank you.**
- to aid the grader, please begin each lettered problem on a new page.

A) Variance of Histograms (2 pages + 2 plots, 10 pts) The goal of this problem is to understand the nature of the variations encountered when trying estimate the PDF by making a histogram. The context for this study will be the simple histogram technique as demonstrated by the script *w07unif.m* for $N = 500$ uniformly distributed random variables on $(0, 1)$ and $M = 10$ disjoint bins. The bin widths are $\Delta b = 1/M$ and the bin centres are located at $x_k = (k - 1/2)/M$ for $k = 1 \ldots M$. Recall the estimated PDF formula

$$f(x_k) \approx \tilde{f}(x_k) = \tilde{f}_k = \frac{\#\ \text{rv's in bin } k}{N\Delta b} \ .$$

- explain why the # of rv's in bin j is a binomial random variable, then use the scaling properties of the mean and variance to obtain $\mathrm{E}[f_k]$ and $\mathrm{Var}[f_k]$;
- modify the given script to generate several estimated PDFs and verify, by simulation, the theoretical dependence of $\mathrm{Var}[f_k]$ on N and M;
- make a loglog plot of $\mathrm{Var}[f_k]$ versus N with fixed M, and another versus $M-1$ with fixed N. (Do you see why you may choose any value of k?)

B) Variance of Empirical CDFs (3 pages + 2 plots, 10 pts) The goal of this problem is to understand the nature of the variations encountered when producing an empirical CDF. The context for this study will be the method (of HW #03) as demonstrated by the script *w07cdf.m* for $N = 500$ uniformly distributed random variables x_j on $(0, 1)$. Recall the empirical CDF formula

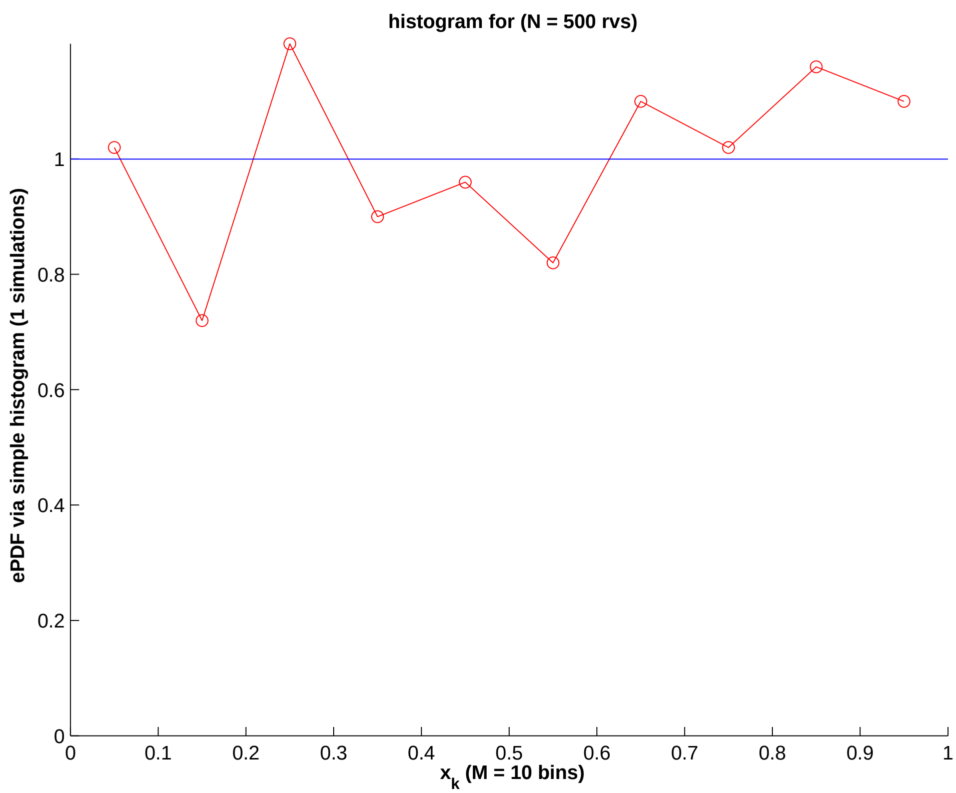
$$\tilde{F}_k = \tilde{F}(\tilde{x}_k) = k/N$$

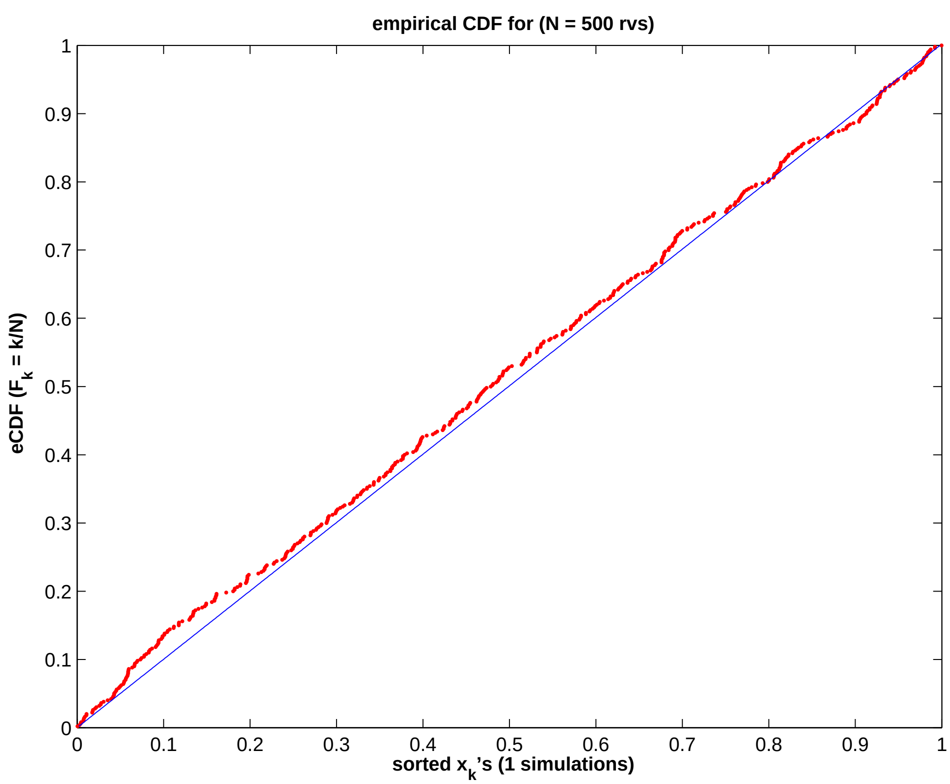
where $0 < \tilde{x}_1 < \tilde{x}_2 < \ldots < \tilde{x}_N < 1$ are the sorted random variables.

- explain why the sorted index j for $\tilde{x}_k = x_1 = y$ is a binomial random variable, then give the conditional probability $P\{\tilde{x}_k = x_1 | x_1 = y\}$;
- give the probability $P\{\tilde{x}_k = y\}$ and quote a result from the 06 October lecture to verify that the integral over all y is one;
- use the previous probability to calculate $\mathrm{E}[\tilde{x}_k]$ and $\mathrm{Var}[\tilde{x}_k]$ (you might also note the strange similarity to Problem #88 in Chapter 3 in Ross — although i think there may be a typo in part (b), *Explain how this proves the result of Section 3.6.3*?);
- modify the script *w07cdf.m* to generate several estimated CDFs and verify, by simulation, the theoretical dependence of $\mathrm{Var}[\tilde{x}_k]$ on N and k;

Bonus: Comment briefly on the reason why the CDF method seems to give more satisfying results than the histogram PDF.

C) Rolling (2 pages, 10 pts) Like Problem #91 in Ross, but easier. Read Case 1 of Section 3.6.4 and give the expected number of rolls of a single die until the pattern $1, 2, 3, 4, 5, 6$ in arises consecutive rolls.


histogram for (N = 500 rvs)
ePDF via simple histogram (1 simulations)
x_k (M = 10 bins)
0
0.2
0.4
0.6
0.8
1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


empirical CDF for (N = 500 rvs)
eCDF (F_k = k/N)
sorted x_k's (1 simulations)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1