Homework 1

Due 2019 September 13

Report instructions

Please upload *one* PDF file to Gradescope (Entry code: ME62YG).

For this homework you will implement code (on PrairieLearn) to perform statistical analysis of data, which is representative of analyzing simulation output. You specifically should implement functions that compute the mean value, the standard deviation, the auto-correlation time, and the standard error (please see PrairieLearn for details and instructions). For your report you will be asked to discuss a couple of specific problems related to your analysis.

Data set 1

- Discuss the change in the estimated error of the mean, when you use only the first 500 data points to calculate the mean value.
- Discuss the change in the standard deviation, when you only use the first 500 data points.
- How would you expect these quantities to change as you gradually remove more and more data points from the data set (until only 500 are left)? Why?

Data set 2

- Discuss the change in the error of the estimated mean when you include/exclude correlation.

Data set 3

- Discuss where you would set the initial cutoff for this data set and why.
-
- Discuss whether the difference between the two mean values (with and without cutoff) is significant and why.

Data set 4

A data set (download here) was sampled from the distribution

$$P(x) = b/(|x|^a + c)$$

with $a = 2.2$ and $c = 1.0$. The constant b is determined by the normalization.

- Based on the analytic expression given above, what do you expect for the mean and why?
- What do you expect for the variance? *(Hint: For the variance, the behavior at large x matters, so one can use an approximation for the denominator of the distribution function.)*
- Look at the convergence of the mean and σ by computing these values for five "end cutoffs" from 1000 to 5000 (i.e., use data points 0–999, 0–1999, 0–2999, etc). Do the same for your data sets of the problems "Dataset 1: Mean value" and "Dataset 2: Autocorrelation time" on PrairieLearn and compare the convergence behavior!

Central Limit Theorem

Given a population with a mean μ and a finite, non-zero variance σ, the sampling distribution of the mean approaches a normal distribution with a mean of μ and a variance of $\sigma' = \sigma/(N-1)$ as N, the sample size, increases.

This is the Central Limit Theorem and implies that the estimated mean approaches a Gaussian distribution as more points are used.

Now suppose you have you have 2 versions A and B of a code that you're debugging. You run each code 6 times to try to determine if they give the same answers. This yields:

$$A = \{1.12, 1.52, 1.33, 1.09, 1.20, 1.26\}$$

$$B = \{1.44, 1.34, 1.19, 1.13, 1.56, 1.45\}$$

- Compute the mean, variance, and the estimate of the error of the mean for A and B separately, assuming each run is uncorrelated with the others and report these results.
- Show that the probability that the two runs are (NOT) drawn from the same distribution is about 29% (71%).

  To do this, first find how many standard deviations the difference is from zero; do this by dividing the "estimate of the difference" by the "estimate of the error of the difference." From this number determine the probability that the two are from the same distribution using a Normal Standard Probability Distribution Table (often referred to as $P(0, x)$ or the Error Function $\mathrm{erfc}(x/\sqrt{2})$). A detailed explanation of how to do this can be found here/PDF.