Question 1

You would like to survey students at Brown to determine the mean time $t$ (in hours) that they sleep each night. Assume that you will sample uniformly at random $n$ students from the population, and let $X_i$ be the number of hours reported by student $i$. The mean of their answers is then

$$M_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

(a) As an initial guess, you assume the standard deviation of $X_i$ is 2.0 hours ($\text{Var}(X_i) = 4.0$). How large should $n$ be so that Chebyshev’s inequality guarantees that the estimate $M_n$ is within 30 minutes of the true $t$, with probability of at least 0.99?

(b) You learn that all students have sleep times between 4 and 10 hours, and thus decide to set $\text{Var}(X_i) = (10 - 4)^2/12 = 3.0$, corresponding to a uniform distribution over that range. How should the value of $n$ obtained in part (a) be revised?

After learning about the central limit theorem, you assume that the number of surveyed students $n$ is sufficiently large that the distribution of $M_n$ is normal. The Matlab $\text{norminv}$ command may be helpful when answering this question.

(c) Suppose you assume that $\text{Var}(X_i)$ is 4.0, as in part (a). Assuming the distribution of $M_n$ is normal, how large should $n$ be so that the estimate $M_n$ is within 30 minutes of the true $t$, with probability of at least 0.99?

(d) Suppose you assume that the $\text{Var}(X_i)$ is 3.0, as in part (b). Assuming the distribution of $M_n$ is normal, how large should $n$ be so that the estimate $M_n$ is within 30 minutes of the true $t$, with probability of at least 0.99?

(e) Briefly discuss how and why the answers obtained using the central limit theorem differ from those obtained using Chebyshev’s inequality.

Answer 1
Question 2

(a) Your friend’s new puppy gets loose on Thayer street, and begins wandering aimlessly. Every minute, he travels south 5 meters with probability $\frac{1}{2}$, or north 5 meters with probability $\frac{1}{2}$. The directions of travel at successive minutes are independent. Using the central limit theorem, what is the mean and variance of the puppy’s location after 1 hour? Which location has the highest probability?

(b) Suppose that immediately after getting loose, the puppy smells food trucks to the south. Every minute, he travels south 5 meters with probability $\frac{2}{3}$, or north 5 meters with probability $\frac{1}{3}$. The directions of travel at successive minutes are independent. Using the central limit theorem, what is the mean and variance of the puppy’s location after 1 hour? Which location has the highest probability?

(c) In order to find the lost puppy, you and your friend decide to identify an interval of locations that contain the puppy with approximately 95% probability. Compute this interval using a centered 2-tailed Gaussian confidence interval using your Gaussian approximations from parts (a) and (b). What is the length of this interval (in meters) for the motion model in part (a)? What is its length for the motion model in part (b)? Explain any differences.

Answer 2
Question 3

In this problem we apply Monte Carlo simulation to estimate the value of the integral

\[ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx. \]

(a) Since the integral is computed over the infinite domain \((-\infty, \infty)\) we first bound analytically the contribution of the lower and upper tails. Prove all the steps in the following derivation for all \( L > 0 \). (Hint: Use power series to prove the last part.):

\[ \int_{L}^{\infty} e^{-\frac{x^2}{2}} dx \leq \sum_{j=L}^{\infty} \int_{j}^{j+1} e^{-\frac{t^2}{2}} dt \leq \sum_{j=L}^{\infty} e^{-\frac{j^2}{2}} \leq e^{-\frac{(L+1)^2}{2}} \leq e^{-\frac{(L)^2}{2}} \sum_{i=0}^{\infty} e^{-i} \leq 2e^{-\frac{(L)^2}{2}}. \]

Now choose an integer value \( L \) such that

\[ \int_{-\infty}^{-L} e^{-\frac{x^2}{2}} dx + \int_{L}^{\infty} e^{-\frac{x^2}{2}} dx \leq 0.0001. \]

(b) Next we approximate the value of the integral

\[ \frac{1}{\sqrt{2\pi}} \int_{-L}^{L} e^{-\frac{x^2}{2}} dx. \]

Using Matlab (or another language of your choice), write a simple Monte Carlo simulation to approximate the integral by using the approach outlined in Algorithm 1.

**Algorithm 1 Monte Carlo integral approximation**

\[
\begin{align*}
E_0 &\leftarrow 0 & \text{Estimator Initialization} \\
\text{for } i = 1, 2, \ldots, 1000 \text{ do} & \\
& \quad \text{Generate a random variable } x_i \text{ uniformly distributed on } [-L, L] \\
& \quad y_i \leftarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} \\
& \quad E_i \leftarrow E_{i-1} + y_i \\
& \quad \hat{\mu}_i \leftarrow 2LE_i \\
\text{end for}
\end{align*}
\]

Plot the values of \( \hat{\mu}_i \) (on the y axis) as the values of \( i \) increases (on the x axis). Give an interpretation of the plotted result.

(c) Modify your previous evaluation to compute also the empirical variance of your Monte Carlo approximation, which we denote denoted \( \hat{\sigma}_i^2 \) for \( n = 1, 2, \ldots, 1000 \). Your code should follow the pseudocode of Algorithm 2.

Provide an interpretation on why this code provides an estimate of the variance of \( 2L y_i \), for any \( y_i \) defined as above. Then explain why \( \hat{\sigma}_i^2 \) is an estimate of the variance of \( \hat{\mu}_i \). Finally, plot the value of the empirical standard deviation \( \hat{\sigma}_i \) of \( \mu_i \) for \( i = 1, 2, \ldots, 1000 \). Give an interpretation of the plotted result.
Algorithm 2: Empirical variance Monte Carlo integral approximation

\begin{align*}
E_0 & \leftarrow 0 \quad \triangleright \text{Estimator initialization} \\
S_0 & \leftarrow 0 \quad \triangleright \text{Second moment initialization} \\
\text{for } i = 1, 2, \ldots, 1000 \text{ do} & \\
\quad & \text{Generate a random variable } x_i \text{ uniformly distributed on } [-L, L] \\
\quad & \text{Calculate } y_i = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} \\
\quad & E_i \leftarrow E_{i-1} + 2Ly_i \\
\quad & S_i \leftarrow S_{i-1} + (2Ly_i)^2 \\
\quad & \hat{\mu}_i \leftarrow \frac{E_i}{i} \\
\quad & \hat{\sigma}_i^2 \leftarrow \frac{S_i}{i} - \hat{\mu}_i^2 \\
\text{end for}
\end{align*}

(d) Repeat the above experiment for \( L = 1, L = 4, \) and the \( L \) you chose above. For each choice of \( L \), plot \( M_i \) as a function of \( i \), and provide 1-standard deviation error bars of \( \pm \hat{\sigma}_i \). You may use the matlab function \texttt{errorbar}(\mu, \sigma). If your plotting environment does not support error bars (or you are unable to use them), then separately plot \( \mu_i, \mu_i + \sigma_i, \) and \( \mu_i - \sigma_i \), labeled appropriately, for each \( L \). How does the choice of \( L \) effect the performance of your estimator?

(e) Express \( \lim_{i \to \infty} \hat{\mu}_i \) and \( \lim_{i \to \infty} \frac{\hat{\sigma}_i^2}{i} \) algebraically as functions of \( L \). Your solution may contain integrals, but try to simplify them as much as possible. Is this consistent with the relationship you observed in the previous part?

Answer 3