Normal (Gaussian) Distribution

Question 1) Normal Distribution

We say *x* is a normal or Gaussian random variable with parameter μ and σ^2 if its density function is given by:

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

and its distribution function is given by:

$$F(x;\mu,\sigma^2) = \int_{-\infty}^{x} f(y;\mu,\sigma^2)dy$$

We can express $F(x;\mu,\sigma^2)$ in term of the error function (erf) as follows:

$$F(x;\mu,\sigma^2) = \frac{1}{2}\mathrm{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right) + \frac{1}{2}$$

The probability density function (pdf) and cumulative distribution function (cdf) of normal distribution can also be calculated using two built-in functions *norm.pdf* and *norm.cdf* from the *scipy.stats* package in Python.

a) Write two Python function on your own based on above equations, one for calculating normal pdf and one for calculating normal cdf. (treating x, μ, σ^2 as inputs of the functions)
b) With x=[-6, 6], calculate the pdf and cdf using the functions you wrote above, and plot them for the following pairs of (μ, σ^2): (0, 1), (0,10^{-1}), (0, 10^{-2}), (-3, 1), (-3, 10^{-1}), (-3, 10^{-2}). (Please plot them in two figures: one contains all the pdf curves, and one contains all the cdf curves)
c) What can you observe about the affect of $\mu\ and\ \sigma^2$ on normal pdf and cdf curves?

Question 2) Central Limit Theorem

Assuming $X_1, X_2, \dots, X_n$ are independent random variables having the same probability distribution with mean μ and standard deviation σ, consider the sum $S_n = X_1 + X_2 + \cdots + X_n$.

This sum S_n is a random variable with mean $\mu_{S_n} = n\mu$ and standard deviation $\sigma_{S_n} = \sigma\sqrt{n}$.

The Central Limit Theorem states that as the probability distribution of the random variable S_n will approach a normal distribution with mean μ_{S_n} and standard deviation σ_{S_n}, regardless of the original distribution of the random variables $X_1, X_2, \dots, X_n$.

It is noted that the PDF of the normally distributed random variable S_n is given by:

$$f(S_n) = \frac{1}{\sigma_{S_n}\sqrt{2\pi}} e^{-\frac{(x-\mu_{S_n})^2}{2\sigma_{S_n}{}^2}}$$

This problem will help you get more understanding about the Central Limit Theorem. After plotting the required plots, you can see that even if the individual distributions of a RV do not look anything like Gaussian, when you add enough of the identical RVs together, the result is a Gaussian with a mean equal to the sum of the individual means of the RVs, and a standard deviation equal to the square root of the sum times the individual RV's standard deviation.

Below is the question:

Consider a collection of books, each of which has thickness W. The thickness W is a random variable, uniformly distributed between a minimum of a=1 and a maximum of b=3 cm. use the values of a and b that were provided to you, and calculate the mean and standard deviation of the thickness. Use the following table to report the results:

Mean thickness of a single book (cm)	Standard deviation of thickness (cm)
$\mu_W =$	$\sigma_W =$

The books are piled in stacks of n=1, 5, 10, or 15 books. The width S_n of a stack of n books is a random variable (the sum of the widths of the n books). This random variable has a mean $\mu_{S_n} = n\mu$ and a standard deviation of $\sigma_{S_n} = \sigma\sqrt{n}$.

Calculate the mean and standard deviation of the stacked books, for the different values of n=1, 5, 10, or 15. Use the following table to report the results:

Number of books n	Mean thickness of a stack of n books (cm)	Standard deviation of the thickness for n books
n=1	μ_{S_n}=	σ_{S_n}=
n=5	μ_{S_n}=	σ_{S_n}=
n=15	μ_{S_n}=	σ_{S_n}=

Perform the following simulation experiments, and plot the results.

a) Make n=1 and run N=10,000 experiments, simulating the random variable $S = W_1$.
b) After the N experiments are completed, create and plot a probability histogram of the random variable S.
c) On the same figure, plot the normal distribution probability function f(x), and compare the probability histogram with the plot of f(x)

$$f(S_n) = \frac{1}{\sigma_{S_n}\sqrt{2\pi}} e^{-\frac{(x-\mu_{S_n})^2}{2\sigma_{S_n}{}^2}}$$

d) Make n=5 and repeat steps (a)-(c)
e) Make n=15 and repeat steps (a)-(c)

Notice: For question 2, you need to submit:

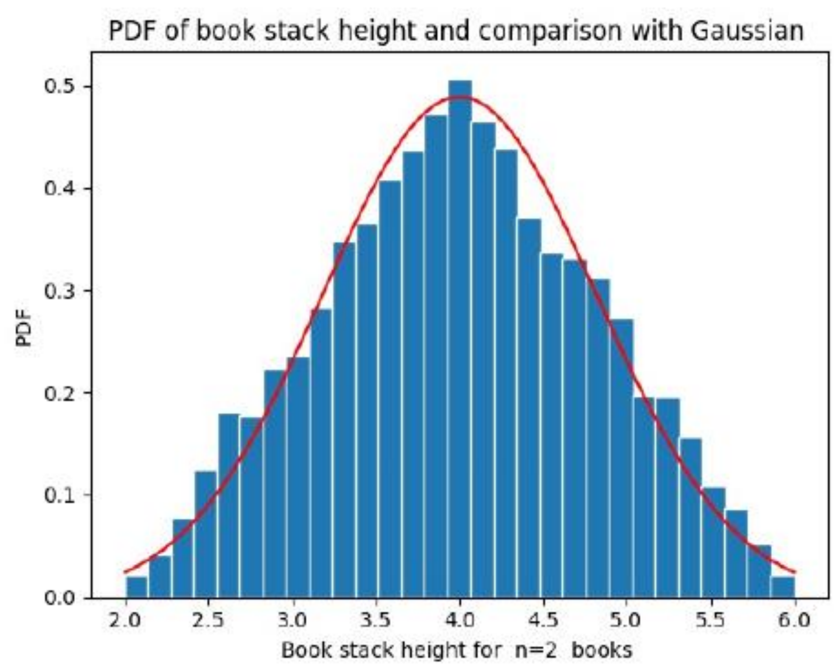
The above tables

The histogram for n={1,5,15} and the overlapping normal probability distribution plots.

Make sure that the graphs are properly labeled.

An example of creating the PDF graph for n=2 is shown below. The code below provides a suggestion on how to generate a bar graph for a continuous random variable X, which represents the total bookwidth for n=2, a=1, b=3.

Note that the value of "barwidth" is adjusted as the number of bins changes, to provide a clear and understandable bar graph.
Also note that the "density=True" parameter is needed to ensure that the total area of the bargraph is equal to 1.0.




```
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
# Generate the values of the RV X
```

```
N=100000; nbooks=2; a=1; b=3;
mu_x=(a+b)/2 ; sig_x=np.sqrt((b-a)**2/12)
X=np.zeros((N,1))
for k in range(0,N):
  x=np.random.uniform(a,b,nbooks)
  w=np.sum(x)
  X[k]=w
# Create bins and histogram
nbins=30; # Number of bins
edgecolor='w'; # Color separating bars in the bargraph
#
bins=[float(x) for x in np.linspace(nbooks*a, nbooks*b,nbins+1)]
h1, bin_edges = np.histogram(X,bins,density=True)
# Define points on the horizontal axis
be1=bin_edges[0:np.size(bin_edges)-1]
be2=bin_edges[1:np.size(bin_edges)]
b1=(be1+be2)/2
barwidth=b1[1]-b1[0] # Width of bars in the bargraph
plt.close('all')
# PLOT THE BAR GRAPH
fig1=plt.figure(1)
plt.bar(b1,h1, width=barwidth, edgecolor=edgecolor)
#PLOT THE GAUSSIAN FUNCTION
def gaussian(mu,sig,z):
   f=np.exp(-(z-mu)**2/(2*sig**2))/(sig*np.sqrt(2*np.pi))
   return f
f=gaussian(mu_x*nbooks,sig_x*np.sqrt(nbooks),b1)
plt.plot(b1,f,'r')
plt.show()
```

Question 3) Distribution of the sum of exponential random variables

This problem involves a battery-operated critical medical monitor. The lifetime (T) of the battery is a random variable with an exponentially distributed lifetime. A battery lasts an average of $\beta = 45\ days$. Under these conditions, the PDF of the battery lifetime is given by:

$$f_T(t\,;\beta) = \begin{cases} \dfrac{1}{\beta}\exp(-\dfrac{1}{\beta}t), & t \geq 0 \\ 0, & t < 0 \end{cases}$$

The mean and variance of the random variable T are:

$$\mu_T = \beta \qquad \sigma_T = \beta$$

When a battery fails it is replaced immediately by a new one. Batteries are purchased in a carton of 24. The objective is to simulate the RV representing the lifetime of a carton of 24 batteries, and create a histogram. To do this, follow the steps below.

a) Create a vector of 24 elements that represents a carton. Each one of the 24 elements in the vector is an exponentially distributed random variable (T) as shown above, with mean lifetime equal to β. Use the same procedure as in the previous problem to generate the exponentially distributed random variable T.
Use the Python function "numpy.random.exponential(beta,n)" to generate n values of the random variable T with exponential probability distribution. Its mean and variance are given by:

$$\mu_T = \beta \quad ; \quad \sigma_T = \beta$$

b) The sum of the elements of this vector is a random variable (C), representing the life of the carton, i.e.

$$C = T_1 + T_2 + \cdots + T_{24}$$

where $T_j, j=1,2,\ldots,24$ each is an exponentially distributed random variable. Create the random variable C, i.e simulate one carton of batteries. This is considered one experiment.

c) Repeat this experiment for a total of N=10,000 times, i.e. for N cartons. Use the values from the N=10,000 experiments to create the experimental PDF of the lifetime of a carton, *f(c)*.

d) According to the Central Limit Theorem the PDF for one carton of 24 batteries can be approximated by a normal distribution with mean and standard deviation given by:

$$\mu_C = 24\mu_T = 24\,\beta \quad ; \quad \sigma_C = \sigma_T\sqrt{24} = \beta\sqrt{24}$$

Plot the graph of normal distribution with mean μ_C and standard deviation σ_C over plot of the experimental PDF on the same figure, and compare the results.

e) Create and plot the CDF of the lifetime of a carton, F(c) . To do this use the Python "`numpy.cumsum`" function on the values you calculated for the experimental PDF. Since the CDF is the integral of the PDF, you must multiply the PDF values by the `barwidth` to calculate the areas, i.e. the integral of the PDF.
If your code is correct the CDF should be a nondecreasing graph, starting at 0.0 and ending at 1.0.

Answer the following questions:

1. Find the probability that the carton will last longer than three years, i.e. $P(S > 3*365) = 1 - P(S \leq 3*365) = 1 - F(1095)$. Use the graph of the CDF F(t) to estimate this probability.

2. Find the probability that the carton will last between 2.0 and 2.5 years (i.e between 730 and 912 days): $P(730 < S < 912) = F(912) - F(730)$ .Use the graph of the CDF F(t) to estimate this probability.