Question 1 (40). Random number generation for modeling data. The first questions on the exam deal with the sinc function. In these problems the sinc function will be the spatial amplitude distribution due to the diffraction of light from a rectangular aperture. For diffraction the function takes the form,

$$sinc\left(\frac{x}{x_0}\right) = \frac{\sin\left(\frac{x}{x_0}\right)}{\left(\frac{x}{x_0}\right)}$$

where $x_0$ is the distance from the center of the distribution to the first node. A spectroscopist would be interested in the distribution of photons, not amplitude. The distribution of photons is given by the square of the distribution of amplitude, thus the photons will follow a probability density function proportional to sinc².

a. (3) Let's work our way up to the sinc function first. Using the example MathCAD worksheet from the Chem 621 website (under tutorials) create a random number generator for the function $f(x) = x^{1/2}/x$. Demonstrate the validity of your random number generator by generating $10^5$ random numbers, create a histogram of the generated random numbers and overlay with the PDF.

b. (17) The function in part A is easily invertible, not all functions are easily invertible. Now to tackle the next step, we need to write a program (or combinations of programs) to numerically calculate the inverse cdf of an arbitrary pdf, including functions that are not easily invertible, and perform the random number generation. You will test this first on function used in part in part A to confirm its proper function. Because we will have to perform the random number generation so many times later on, the calculations will be made much more efficient by first constructing a table of the numerical cdf with user-defined resolution. If you do this right, the table for the cdf will range from 0 to 1. Now, in a separate program, generate a uniform random number between 0 and 1. Write a program that searches for where this random number falls in the table for the cdf and return the value of $\theta$. This is analogous to inputting the uniform random number into an analytical cdf. Repeat this $10^4$ times to confirm that your numerical result recovers the analytical result from part A. Also, for generating a large number of random numbers later on, it is desired, but not required, that this searching program be as efficient as possible.

c. (2) Generate random values obeying the non-trivially invertible functions sinc($x$)² in the range -5.00 ≤ $x$ ≤ +5.00 using the random number generator you developed in part. Demonstrate that your random number generator works by generating 10,000 values. Bin them with a resolution of $\Delta x = 0.1$. Plot the histogram, overlay the pdf. Depending on how you created your random number generator for part a you might need to modify your program slightly, or modify the sinc($x$)² function.
d. (18) In a high resolution optical spectrometer, the width of the optical grating determines the
maximum possible resolution. The grating diffracts light in the same way as a plane wave passing
through a rectangular aperture. Thus, the pattern of light at the exit slit of the spectrometer follows a
\( \text{sinc}^2 \) distribution. When a spectrometer output follows a \( \text{sinc}^2 \) distribution, the spectrometer resolution
is ordinarily given as the difference in wavelength between the \( \text{sinc}^2 \) center and the maximum of the
first node. This is called the Rayleigh criterion. For the equation we will be using, the resolution would
be \( \Delta \lambda = 1.5 \cdot x_0 \). This definition is most often used to denote when two spectral lines are just visually
separated by the spectrometer. Unfortunately, this definition is carelessly used when quantitation is
performed by measuring the area under the peak. This problem will use the random number generator
created in part (b) to investigate how far apart two peaks need to be for peak area to be a valid estimate
of concentration when performing quantitative atomic emission. Make two random number generators
that simulate atomic emission that has passed through the spectrometer. The first will be centered at a
wavelength of 510 nm and consist of 100 photons randomly distributed with \( x_0 = 1 \) nm. To estimate the
area of this peak add together all photons falling between the center of the peak and the first nodes,
e.g. \(~509\) to \(511\) nm. Obviously these limits will not recover 100% of the emitted photons, but the
percentage should be proportional to the true area. The second random number generator will start \(3\Delta\lambda\)
lower in wavelength than the sample peak and follow a \( \text{sinc}^2 \) distribution with \( x_0 = 1 \) nm. This output will
represent emission from an interference. Keep the analyte wavelength and number of emitted photons
fixed throughout the simulation. Vary the strength of the interference from 1, 10, 100, 1000, and 10,000
photons. For each interference strength compute the percent recovery of the analyte signal. Change the
interference wavelength so that it is \(2\Delta\lambda\) lower than the sample peak and repeat the experiment varying
the number of interference photons. Finally, change the interference so that it is \(1\Delta\lambda\) lower than the
sample peak and repeat the experiment varying the number of interference photons. Put your recovery
data into a table and include one graph for each interference wavelength that shows how the two
probability density functions are overlapping. Make some cogent statements about how far apart the
two peaks need to be for accurate quantitation with varying levels of interference emission. Does your
data support the use of the Rayleigh criterion for predicting the performance of quantitation in the
presence of an interference?

Question 2 (45).

(Will U B mine, within a 95% confidence interval?) Lognormal probability
densities appear commonly in data analysis, and are therefore worthy of special consideration.

a. (5) The parameters for a lognormal pdf are most conveniently expressed in terms of the mean and
standard deviation (SD) not of \( x \), but of the mean and SD of \( \ln(x) \), given by \( \mu_L \) and \( \sigma_L \), respectively. Using
parameters \( \mu_L = 5 \) and \( \sigma_L = 1.5 \), generate at least \( 10^4 \) lognormally distributed random numbers using the
\text{rlnorm} \ function in MathCad. Generate a histogram with 1000 bins ranging from \( x = 0 \) to \( 1000 \), and
overlay the theoretical pdf with the histogram (they should agree beautifully). The MathCad tutorials on
generating histograms from the 621 webpage may be useful to review.

b. (7) Let’s start by assuming we know the experimental standard deviation \( \sigma_L \) \textit{a priori} through a large
number of previously performed measurements. In this case, we can calculate the CDF two equivalent
ways by starting with a normal distribution in \( \phi \), where \( \phi(x) = \ln(x) \). In the first, we can substitute
\( d\phi = (d\phi/dx)dx \) to change the variable of integration to \( x \) as outlined in the lecture notes. Confirm that the mean of \( x \), \( \mu_x = \exp(\mu_L + \sigma_L^2/2) \) and that \( \sigma_x^2 = \left( e^{\sigma_L^2} - 1 \right) e^{2\mu_L + \sigma_L^2} \), given by eq. 4.33D and 4.34D of Fred Lytle’s textbook-in-progress. Alternatively, we can simply leave \( \phi \) as the variable of integration and change the limits of integration (hint: the limit of \( x = -\infty \) corresponds to \( \ln(x) = 0 \)). Using the known 95% double-sided confidence limits for a normal distribution in \( \phi \), calculate the corresponding confidence limits for the lognormal distribution in \( x \). From the \( 10^4 \) random lognormal values generated in part a, calculate the fraction that fall below the lower CL and the fraction that fall above the upper CL. How well do they compare to the theoretical values of 2.5% for each?

c. (8) Next, let’s consider an average of 20 lognormally distributed numbers where neither the average nor the variance are known \textit{a priori} and are instead estimated from the data. The Central Limit Theorem (CLT) suggests that the sum should approach a normal distribution; such that the confidence limits can be evaluated based on a normal pdf. Let’s see how well this works. Write a program that takes the sum of 20 random numbers from a lognormal distribution described by \( \mu_L \) and \( \sigma_L \) and repeats this process \( 10^4 \) times to return a vector of \( 10^4 \) sets of \( \bar{x} \) and \( s_x^2 \) values. Using the \( t \)-parameter in the lecture notes for \( N = 20 \) to determine the expected upper and lower confidence limits assuming a normal distribution in \( \bar{x} \), and determine the fraction of the “measured” averages that fall above and below the upper and lower confidence limits, respectively. How do these compare to the expected results, in which 2.5% of the observed means should fall below the lower limit and 2.5% should fall above the upper limit? Interpret.

d. (10) Let’s see if we can get around the limitations of the CLT by evaluating the confidence intervals in terms of normal distributions of the natural logs of variables. If \( \bar{\phi} \equiv \ln(\bar{x}) \), it can be shown (see part e.) that the mean values expected for \( \bar{\phi} \) and for \( \sigma_{\bar{\phi}}^2 \) are reasonably approximated by:

\[
\mu_{\bar{\phi}} = \mu_L + \frac{\sigma_L^2}{2} \quad \text{and} \quad \mu_{\sigma_{\bar{\phi}}^2} = \frac{\sigma_L^2}{N} + \frac{\sigma_L^4}{2(N-1)}
\]

Consequently, the values of \( \bar{\phi} \) will be described by a distribution with an average mean of \( \mu_{\bar{\phi}} \) and an average variance of \( \mu_{\sigma_{\bar{\phi}}^2} \). If we assume that the distribution in \( \bar{\phi} \) can be approximated by a normal pdf with a known mean and standard deviation, we can calculate the confidence intervals using the \( \mu_{\bar{\phi}} \pm 1.96 \sqrt{\mu_{\sigma_{\bar{\phi}}^2}} \) expression derived from the z-statistic. In principle, we could in turn use these values to generate \( t \)-parameters, but in practice the values we obtain depend on \( \mu_L \) and \( \sigma_L \) in a way that is not trivially related back to the z-statistic. Rather than generating a unique \( t \)-value for many possible combinations of \( \mu_L \) and \( \sigma_L \), it makes more sense just to numerically evaluate the confidence limits on a case-by-case basis. Confirm that the above approximate analytical expressions provide a general solution for recovering confidence intervals within \( \pm 1\% \) of the targeted 2.5% on the high and low ends (it will still not agree perfectly) by determining the fraction of random \( \bar{\phi} \) values falling below the lower cut-off and the fraction falling above the upper cut-off. Now, increase and decrease \( N \) to see how well the above approximation compares to the results expected by the CLT in part c).

e. (15) Next, derive the expressions for \( \mu_{\bar{\phi}} \) used in the preceding question following the approach described in Section 3.3 of the lecture notes by finding the expectation value for \( \ln(\bar{x}) \), given that
\[\bar{x} = e^{\mu_{\phi} + \frac{1}{2} \sigma_{\phi}^2}, \text{ in which } s_{\phi}^2 = \frac{1}{N-1} \sum (\phi_i - \overline{\phi})^2.\] Please note the equalities relating different gamma densities in the proof. To help in this process, it is useful to note that sums of normally distributed variables are also normally distributed, sums of lognormally distributed values are often well-approximated as lognormally distributed, and the natural log of a lognormally distributed value is normally distributed. State all approximations as appropriate. Building on your success in recovering the expressions for \(\mu_{\phi}\), use the toolkit you have developed to derive the expression for the expectation value of \(s_{\phi}^2\).

**Question 3 (40).** (U + math ≈ ☀, within the range of validity of the approximation). Taylor Series propagation of variance in nonlinear functions. Many experimental measurements are based on sensitive detection of subtle modulations in large signals. Let’s compare two approaches used to evaluate the difference. In the first, the amplitude of modulation (AoM) can be conceptually written as a difference between a set of measured voltages in a “high” state \(x_H\) and a “low” state \(x_L\), where the variance in the difference is given by the summed variance of the high and low measurements \(\sigma_H^2 + \sigma_L^2\). Let’s contrast this with a second technique based on the depth of attenuation (DoA), in which the AoM is normalized by dividing by \(x_H\), such that \(\text{DoA} = 1 - \frac{\mu_L}{\mu_H}\).

a.(7) Use propagation of errors from a first-order Taylor series expansion to generate an expression for the variance in the DoA in terms of \(\sigma_H\) and \(\sigma_L\). Calculate an analytical expression for the RSD.

b.(13) Assess the sensitivities of the two measurements to 1/f noise. Using the Noise.xmcd file in the MathCad Tutorials from the 621 webpage, generate a \(2 \times 10^3\) element vector of 1/f noise with a standard deviation of 0.1. Now, generate a signal vector of comparable length, in which every even element corresponds to a random variable for \(x_H\) from a Poisson distribution with a mean of 1000, and every odd element containing a random Poisson value of \(x_L\) with a mean of 900 (alternatively, you could generate two separate arrays for \(x_L\) and \(x_H\), and add sequential elements of the 1/f noise vector). Cycle through the array for each pair to calculate \(10^3\) values of AoM and DoA. Plot histograms of the AoM and DoA values and determine the mean and experimental standard deviation. Interpret any similarities or differences. Now, repeat this evaluation and plotting for the addition of normally distributed noise with a standard deviation of 0.1 instead of 1/f noise and interpret any similarities or differences.

c.(20) Considering the simplest case of no significant 1/f or normal electronic noise with only the intrinsic Poisson noise in \(x_i\) and \(x_{H_i}\), where would you expect the Taylor series expansion for the DoA to fail and why? Test your assumptions by exploring different regimes by varying the value of \(\mu_i\) leaving all other parameters fixed and identifying regimes in which the linear approximation fails based on departures between analytical theory and simulations of Poisson-distributed numbers \(10^4\) replicate “measurements” of \(V_H\) and \(V_L\). In order to extend the range of validity, perform a second-order Taylor series expansion for the mean of the DoA (you can safely neglect the need for second-order corrections of the variance). Derive a concise analytical expression for the RSD based on the second-order Taylor
series expansion for the mean. Assume for this question that \( \sigma_H \cong \sigma_L \cong \sigma \). Generate a 3D contour plot of the bias introduced by neglect of the second-order correction. The plot should be the bias (defined in Section 4.2 of the lectures) as a function of \((\mu_L / \mu_H)\) and of \((\mu_H - \mu_L)/\sigma\), such that the one plot can universally describe virtually all DoA measurements in which \( \sigma_H \cong \sigma_L \cong \sigma \). Provide an explanation for the regions in which the bias in the first-order Taylor series is significant. Based on this explanation, predict locations where second-order expansion may also be expected to fail.

This exam is due in class on Thursday February 21st.
Exams will be collected from 10:00 to 10:15am.
This is an “open world” exam on one condition. You may not discuss any aspect of the exam with anyone except Prof. Simpson and Shane Sullivan. This includes not viewing the computer files or written work of others taking the exam. You may use all or parts of the MathCad worksheets made by Prof. Lytle on the course website. Reference this material if you use it. Because of the size of the class, you will only be able to seek help from Prof. Simpson or Shane Sullivan once per day. Both of us will post office hours online. Make your questions as concise as possible to ensure that everyone will have time to receive clarification. Please be neat and organize your work. This will be essential for assigning partial credit. Please box all numerical answers and equations so it is obvious which response to grade. Type a summary of your answers on a separate Word document, including example figured, etc. All MathCad worksheets have to be generated by version 14 or greater. It is always a good idea to keep a backup of your worksheets, but make sure that the backup is not on a communal hard drive. All MathCad files will be turned in on a flashdrive, CD or DVD, etc. Be sure that your personal name is in the MathCAD file names. Be sure that your last name and a page number are in the top right corner of each page of your exam answers you turn in.