You can use Mathematica as an aid for many of the computations, however make sure to do hand calculations where suitable as well.

1. Consider the exponential distribution with probability density function $f(x) = \lambda e^{-\lambda x}$ defined on $x \geq 0$ and with parameter $\lambda > 0$.

   (a) Show that $f(x)$ is a valid probability density function by showing that the integral over $[0, \infty)$ is unity.

   (b) Use integration to show that the mean of the distribution is $\frac{1}{\lambda}$.

   (c) Use integration to show that the variance of the distribution is $\frac{1}{\lambda^2}$.

   (d) Determine the median of the distribution. The median is the number M such that,
   $$\int_0^M f(x)\,dx = \frac{1}{2}.$$

   (e) The quantile function of the distribution, $q(u)$ for $u \in [0, 1)$, is defined as follows: For each u, we should have,
   $$\int_0^{q(u)} f(x)\,dx = u.$$
   Determine an expression for $q(u)$.

   (f) Say that U is a uniformly distributed random variable on $[0, 1]$. If you set a new random variable X, via $X = q(U)$, then the distribution of X is exponential (for $q(\cdot)$ evaluated for an exponential distribution as in the item above). Show this empirically for $\lambda = 3$ by generating 10^6 uniform random variables, and comparing the empirical quantile of this data with $q(\cdot)$.

2. Consider the normal probability distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$. The probability density is,
   $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

   (a) Showing that $f(x)$ is a valid probability density function is not immediate. Do this first numerically for $\mu = 2$ and $\sigma = 3$ by approximating the integral via a discretization sum over 100, 1,000, 10,000, and 10^5 terms. You should observe that as the number of terms grows, the value of the sum approaches 1.

   (b) Use integration to show that the mean of the distribution is μ.

   (c) Use integration to show the variance of the distribution is σ^2.

   (d) The k'th moment of the distribution, denoted m_k for $k = 1, 2, 3, \ldots$, is
   $$m_k = \int_{-\infty}^{\infty} x^k f(x)\,dx.$$
   Based on the previous items, $m_0 = 1$, $m_1 = \mu$, and $m_2 = \mu^2 + \sigma^2$. Show that for higher valued $k > 2$, we have,
   $$m_k = \mu\, m_{k-1} + (k-1)\sigma^2 m_{k-2}.$$

   (e) Use this recurrence relation to compute m_4 for $\mu = 2$ and $\sigma = 3$. Compare this value to a numerical computation of the integral both using a discretization, and Mathematica's in-built, `NIntegrate[]` function.

(f) Show analytically that,

$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

You may use material from the lecture or online material, but must justify and explain your calculations.

3. Assume you are presented with univariate data of a random sample, $x_1, \ldots, x_n$ and wish to find a single number, x^* that summarizes $x_1, \ldots, x_n$ as best as possible. One way to specify this in terms of a loss function is to seek a value x^* that minimizes,

$$L(u) = \sum_{i=1}^{n} (x_i - u)^2.$$

Analytically, it is very easy to show that $x^* = \sum_{i=1}^{n} x_i/n$, the sample mean. Nevertheless, it is good to see how this number can be reached via a gradient descent algorithm. Set $\eta > 0$, start with some arbitrary initial $x(0)$. Then you get a sequence of points $x(t)$, for $t = 1, 2, 3, \ldots$ via,

$$x(t+1) = x(t) - \eta \nabla L\big(x(t)\big).$$

(a) Show that,

$$x(t) = \alpha^t x(0) + \beta \frac{1-\alpha^t}{1-\alpha}.$$

for some α and β (specify these values in terms of of the problem parameters and data).

(b) Determine the range of η values for which $x(t)$ will converge to x^*.

4. Consider the simple linear regression problem where you are presented with data points $(x_1, y_1), \ldots, (x_n, y_n)$. You seek β_0 and β_1 to fit the line,

$$y = \beta_0 + \beta_1 x,$$

by minimizing the loss function

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

This minimization is often carried out by methods others than gradient descent, but for the purposes of this exercise you will use gradient descent.

(a) Compute an expression for the gradient $\nabla L(\beta_0, \beta_1)$.

(b) Return to question 5 from Assignment 1. In that question you dealt with data points,

$$(2.4, 3.1), (4.7, 2.7), (4.9, 4.8), (2.9, 7.6), (8.1, 5.4),$$

and fit a line parameterized by β_0 and β_1. Do this now numerically using a gradient descent algorithm using the expression for the gradient you developed above. Make sure that the learning rate (step size rate), η is small enough for convergence. Illustrate your numerical experiments.