Prof. Ahlstrom
MATH 2310 Sec. 01

Probability and Statistics for the Sciences and Engineering (MATH 2310)
Lab 5: Continuous distributions and the method of moments
20 points

This lab will focus on continuous probability distributions, using the built-in functionality in R for calculating probabilities. In this assignment, we will be using the functions for several continuous distributions. Each distribution has four associated functions (examples given for Weibull distribution):

- Probability density (for example: `dweibull`)
- Cumulative distribution (for example: `pweibull`)
- Quantile function (for example: `qweibull`)
- Random sampling function (for example: `rweibull`)

It may be helpful to refer to R documentation for functions like these as you approach this lab.

1. **Wave height**
   The file `cape_elizabeth_2023_data.csv` contains historical data from a buoy near Cape Elizabeth in the U.S. state of Washington (source: https://www.ndbc.noaa.gov/). One of the columns included in the file provides data about **significant wave height** in meters. This entire assignment will be focused on the significant wave height data from this column.

   In this assignment, you will be generating **probability plots** for *five different types of continuous distributions*. You can construct a probability plot, which is also called a "quantile–quantile" or Q–Q plot, to assess how closely a data set follows a distribution. If the data set follows the distribution closely, then the probability plot should roughly look like a straight line. If the plot follows a more curved pattern, that is an indication that the distribution is a worse fit for the data set. The code below shows an example of generating a probability plot for a data set `x` against a Weibull distribution with parameters of scale $\lambda = 1$ and shape $k = 1.5$.

```
qqplot(
  qweibull(
    ppoints(length(x)),
    shape = 1.5,
```

```
    scale = 1
  ),
  x
)
```

Here, the quantile function of the Weibull distribution is plotted against the sorted observations of the data point x, comparing where each data observation falls percentile-wise relative to where it would be expected according to the distribution. If the general slope trend between nearby points doesn't change much on the probability plot—that is, if the plot forms approximately a straight line—then that is evidence that the sample data set appears to follow that probability distribution fairly well.

(a) Calculate the **mean** $\overline{x}$ and the **variance** s^2 of the wave height sample data.

(b) The **method of moments** is an approach that can be used to find a probability distribution with particular attributes. Using this method, if you know for instance what mean and variance you want the distribution to have, you can use that to determine the parameters for the distribution.

The **normal distribution**, which is a two-parameter distribution, is perhaps the easiest one to do this with. You can determine a normal distribution with mean μ and variance σ^2 simply by having its mean parameter be that same μ and its scale (standard deviation) parameter σ as the square root of the desired variance.

Generate a probability plot using the `qnorm` function, with parameters `mean` as the $\overline{x}$ you found in part (a), and `sd` as the square root of the variance s^2 from part (a).

(c) The **uniform distribution** is also a two-parameter distribution, generally represented with a and b as the lower and upper limits at which the probability density function (pdf) is nonzero. You can also describe the distribution by its center c and radius r. If we want a uniform distribution with a specific mean μ and variance σ^2, we can obtain it by using the following settings for the distribution's parameters:

- $c = \mu$
- $r = \sqrt{3\sigma^2}$
- $a = c - r$
- $b = c + r$

Generate a probability plot using the `qunif` function, with parameters as described above, matching the desired mean and variance to the sample values you calculated in part (a).

(d) The **exponential distribution** is a one-parameter distribution, so it can be uniquely determined just by specifying a desired mean (so you don't need to use the variance for this one). The rate parameter can be set as $\lambda = 1/\mu$ in order to obtain an exponential distribution with mean μ.

Generate a probability plot using the `qexp` function, with parameters as described above, matching the desired mean to the sample mean you calculated in part (a).

(e) The **gamma distribution** is a two-parameter distribution, with `shape` parameter α and `scale` parameter β. (The R function for this distribution also allows you to specify `rate` instead of `scale`, where the `rate` is just the multiplicative inverse of the `scale`, but that's not how our textbook defines the distribution, so **I would recommend not using the `rate` argument**.) A gamma distribution with a desired mean μ and variance σ^2 can be obtained using the following parameter settings:

- Shape: $\alpha = (\mu)^2/(\sigma^2)$
- Scale: $\beta = (\sigma^2)/\mu$

Generate a probability plot using the `qgamma` function, with parameters as described above, matching the desired mean and variance to the sample values you calculated in part (a).

(f) The **lognormal distribution** is another two-parameter distribution. This one is a bit more complicated, since it is usually parametrized using the parameters of the normal distribution you would have to take the logarithm of to obtain the given lognormal distribution. This can be characterized in R using the `meanlog` and `sdlog` arguments. In order to obtain a lognormal distribution with a specific desired mean μ and variance σ^2, its parameters can be set as follows. (**Note** that when "log" is used in the following expressions, it refers to the natural logarithm ln; that is, it is the logarithm whose base is Euler's number $e \approx 2.71828$.)

- `meanlog` $= \log\left(\mu/\sqrt{\frac{\sigma^2}{(\mu)^2}+1}\right)$
- `sdlog` $= \sqrt{\log\left(\frac{\sigma^2}{(\mu)^2}+1\right)}$

Generate a probability plot using the `qlnorm` function, with parameters as described above, matching the desired mean and variance to the sample values you calculated in part (a).

(g) You have now generated five probability plots. One of the five distributions used should look closer to a straight line than any of the other ones. We can conclude that out of these five options, our sample data follows this distribution the most closely.

Based on your probability plots, answer the question: Which of these five distributions is the best fit for our sample data?

(h) You can plot the probability density function (pdf) of a distribution to examine its shape. For example, to plot the pdf of the Weibull distribution described earlier, we could use the following code:

```
x_points = seq(0, max(x), along.with = x)
plot(x_points,
     dweibull(x_points, shape = 1.5, scale = 1))
```

Plot **the pdf** of the distribution you found as the best fit for the wave height sample data, and plot a **histogram** of the sample data next to it (either horizontally or vertically).

Compare the shapes of the two plots. How do they compare to each other? You may adjust the width of the histogram intervals to aid in this comparison.

(i) **Use the distribution** you decided on in (g) to answer the following questions. Assume that significant wave height follows this distribution in order to answer them. Note that in R, there are functions you can use to find probabilities for a given distribution. For example, considering the Weibull distribution example given above, we could use the following code to calculate the probability that a sampled value X from this distribution attains a value greater than or equal to 2:

```
pweibull(2, shape = 1.5, scale = 1, lower.tail = FALSE)
```

Output of this command:

```
[1] 0.05910575
```

Which means:
$P(X \geq 2) \approx 0.059$

Or we could use the following code to calculate the probability that a sampled value X from this distribution attains a value less than or equal to 2:

```
pweibull(2, shape = 1.5, scale = 1, lower.tail = TRUE)
```

Output of this command:

```
[1] 0.9408943
```

Which means:
$P(X \leq 2) \approx 0.941$

Note that we can adjust our input for the `lower.tail` argument to change whether we are performing a greater than or less than operation.

Returning to the specific scenario for this lab, **using the distribution you determined in part (g)**, what is the probability that at a randomly selected point in time, the significant wave height h will be greater than or equal to 5 meters?

(j) We can use the R quartile probability functions, the functions whose names start with `q`, to answer the same type of question as in part (i), but in the opposite direction. For example, the following command:

```
qweibull(0.95, shape = 1.5, scale = 1, lower.tail = TRUE)
```

which has the following output:

```
[1] 2.078111
```

tells us that the value of x for which $P(X \leq x) = 0.95$ is approximately given by $x \approx 2.078$.

Using this same kind of idea, and **using the distribution you determined in part (g)**, what is the height h for which 95% of significant wave height readings would be expected to fall below h? In other words, what is h such that $P(X \leq h) = 0.95$?

(k) *[**Extra credit - optional**, 5 points total] The method of moments is much more difficult to apply to a Weibull distribution, so we haven't dealt with that in this assignment. But a Weibull distribution is probably actually a better fit for this data set than any of the five other continuous distributions we considered in this lab. Specifically, significant wave height is thought to follow a Rayleigh distribution. A Rayleigh distribution is the same as a Weibull distribution with shape parameter $\alpha = 2$, given which the scale parameter ς of the Rayleigh distribution is related to the scale parameter of the Weibull distribution β according to the relation $\beta = \varsigma\sqrt{2}$.*

The Rayleigh distribution, then, is a one-parameter distribution, whose mean in terms of its scale parameter ς is given by $\mu = \varsigma\sqrt{\pi/2}$. Given this information, use the method of moments to generate a probability plot for the significant wave height data set aganist a Rayleigh / Weibull distribution. Repeat parts (h), (i), and (j) using this distribution instead, and compare how similar or different your answers are.