Note: This homework assignment covers Chapter 5.

Disclaimer: If you use R, include all R code and output as attachments. Do not just "write in" the R code you used. Also, don't just write the answer and say this is what R gave you. If my grader can't see how you got an answer, it is wrong. I want to see your code and your answers accompanying your code (like in the notes).

1. In a recent study, biomedical engineers modeled the survival time (T, measured in months) of dialysis patients with chronic kidney disease using a Weibull distribution with shape parameter $\beta = 1.39$ and scale parameter $\eta = 17.55$.

(a) Graph the probability density function (pdf) of T. Compute $E(T)$, the mean survival time, and place an "$\times$" on the pdf indicating where $E(T)$ is.
(b) Find the median of this distribution. Interpret what the median represents.
(c) Find $S_T(t)$, the survivor function of T and graph it. Calculate $S_T(24)$ and interpret what this means in words.
(d) Ninety-five percent of the patients will die before what time?
(e) Is the hazard function of T increasing, decreasing, or constant? Explain what this implies about the population of dialysis patients with chronic kidney disease.

2. The data below are taken from Xu et al. (2003, *Applied Soft Computing*), who describe a reliability study on turbochargers in diesel engines. These are failure time data for $n = 40$ turbochargers; the failure times T are measured in 1000s of hours.

1.6	2.0	2.6	3.0	3.5	3.9	4.5	4.6	4.8	5.0
5.1	5.3	5.4	5.6	5.8	6.0	6.0	6.1	6.3	6.5
6.5	6.7	7.0	7.1	7.3	7.3	7.3	7.7	7.7	7.8
7.9	8.0	8.1	8.3	8.4	8.4	8.5	8.7	8.8	9.0

(a) Under a Weibull assumption for T, calculate the maximum likelihood estimates $\widehat{\beta}$ and $\widehat{\eta}$ for the data above. Use these values (along with the Weibull assumption) to answer the following questions.
(b) Calculate $P(T < 5.0)$. Interpret what this probability means in words.
(c) Find the 90th percentile of the time to failure distribution. Interpret in words.
(d) Plot the estimated hazard function for T. Explain, in plain English, what information this graph reveals.
(d) Does the Weibull model seem reasonable for these data? Construct a Weibull qq plot. Interpret this plot. In the light of your answer here, how do you feel about the accuracy of your answers to parts (b), (c), and (d)?

3. A probability distribution well known as a "competitor" to the Weibull distribution for modeling lifetime data is the **lognormal distribution**, whose pdf is given by

$$f_T(t) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma t} \exp\left\{-\dfrac{1}{2}\left(\dfrac{\ln t - \mu}{\sigma}\right)^2\right\}, & t > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The parameters μ and σ^2 are not the mean and variance in this distribution (like they are in the normal distribution).

LOGNORMAL R CODE: Suppose that $T \sim \text{lognormal}(\mu, \sigma^2)$.

$F_T(t) = P(T \leq t)$	ϕ_p
`plnorm(t,`μ`,`σ`)`	`qlnorm(p,`μ`,`σ`)`

Suppose that the lifetime T (measured in hours) of a semiconductor laser has a lognormal distribution with $\mu = 10$ and $\sigma^2 = 2.25$.

(a) Use this code to graph the probability density function (pdf) of T:

```
# Plot PDF
t = seq(0,300000,1)
pdf = dlnorm(t,10,sqrt(2.25))
plot(t,pdf,type="l",xlab="t",ylab="f(t)")
abline(h=0)
```

(b) The mean of a lognormal random variable is not μ. Instead, it is

$$E(T) = e^{\mu + \sigma^2/2}.$$

Calculate the mean semiconductor laser lifetime. Use R to find the median lifetime. Which one is larger: the mean or the median?
(c) What is the probability that a semiconductor laser lifetime exceeds 100,000 hours?
(d) Ten percent of the semiconductor lasers will fail before what time?
(e) The hazard function for $T \sim \text{lognormal}(\mu, \sigma^2)$ is given by

$$h_T(t) = \frac{\dfrac{1}{\sqrt{2\pi}\sigma t} \exp\left\{-\dfrac{1}{2}\left(\dfrac{\ln t - \mu}{\sigma}\right)^2\right\}}{1 - F_Z\left(\dfrac{\ln t - \mu}{\sigma}\right)},$$

for $t > 0$, where $F_Z(\cdot)$ denotes the $\mathcal{N}(0,1)$ cumulative distribution function. Comparing this to the Weibull hazard function, why do you think engineers might prefer the Weibull

distribution over the lognormal distribution when modeling lifetimes?
(f) Of course, just because the Weibull distribution is used more doesn't mean that it is always a good model. Suppose that an engineer (incorrectly) assumed that the semiconductor laser lifetime distribution was Weibull when, in reality, it is lognormal. What could be the consequences of using an incorrect model?

4. A recent article in the *Journal of Engineering Manufacture* described a study to examine the performance of a microdrill when holes are drilled into a certain brass alloy (CuZn38). A sample of $n = 50$ drills was used. On each one, engineers recorded

$$T = \text{ number of holes a drill machines before it breaks.}$$

Here the data recorded in the study:

11	14	20	23	31	36	39	44	47	50
59	61	65	67	68	71	74	76	78	79
81	84	85	89	91	93	96	99	101	104
105	105	112	118	123	136	139	141	148	158
161	168	184	206	248	263	289	322	388	512

For example, the observation "11" means that the drill machined 11 holes successfully before it broke. Therefore, the number of holes drilled describes the lifetime of the drill.

(a) Fit a Weibull distribution to these data; i.e., calculate the maximum likelihood estimates $\widehat{\beta}$ and $\widehat{\eta}$ under a Weibull assumption for T.
(b) Use the `qweibull` function in R to calculate the median lifetime (in terms of the number of holes drilled into CuZn38) based on the Weibull model fit in part (a). Compare this to the sample median of the $n = 50$ data values above, which is

$$\frac{91 + 93}{2} = 92 \text{ holes.}$$

Why are these medians different?
(c) One of the claims made by the authors of this article was that the Weibull distribution wasn't the best model for the data and that the lognormal distribution provided a better fit.

1. Display the qq plot for the data based on your Weibull fit in part (a). You should be able to see why the authors were suspect of the Weibull model for T.

2. Fit the lognormal model to these data using the following code:

```
microdrill.data = c(11,14,...,512) # Enter data
fitdist(microdrill.data,"lnorm") # lnorm stands for lognormal
```

When you do this, you should get this output:

```
> fitdist(microdrill.data,"lnorm")
Fitting of the distribution ' lnorm ' by maximum likelihood
Parameters:
         estimate Std. Error
meanlog 4.4992858 0.11031837
sdlog   0.7800686 0.07800629
```

The estimates reported in this table (under `estimate`) are $\widehat{\mu}$ and $\widehat{\sigma}$. If you get different answers here, you probably entered the data incorrectly! Go back and check, because then your answers to parts (a) and (b) are wrong.

3. Use the `qlnorm` function in R to calculate the median lifetime (in terms of the number of holes drilled into CuZn38) based on the lognormal model fit. Compare this to your answer in part (b). Are these close?

4. Display the qq plot for the data based on the lognormal fit. Use this code:

```
meanlog = 4.499
sdlog = 0.780
qqPlot(microdrill.data,distribution="lnorm",meanlog=meanlog,sdlog=sdlog,
        xlab="Lognormal quantiles",ylab="Microdrill data",pch=16)
```

   Do you agree with the authors' claim that the lognormal model is a better fit to these data?

(d) Want to see what a really bad model fit looks like? Display the qq plot for the data under a normal (Gaussian) assumption:

```
qqPlot(microdrill.data,distribution="norm",xlab="N(0,1) quantiles",
        ylab="Microdrill data",pch=16)
```

This is what "a serious departure" looks like! Specifically, the normal model fits the data very poorly in the upper tail because the data are skewed to the high side. This is why the normal distribution is rarely used for lifetime data; i.e., a symmetric model cannot handle skewness, which is a common feature of lifetime data.