Note: This homework assignment covers Chapter 6.

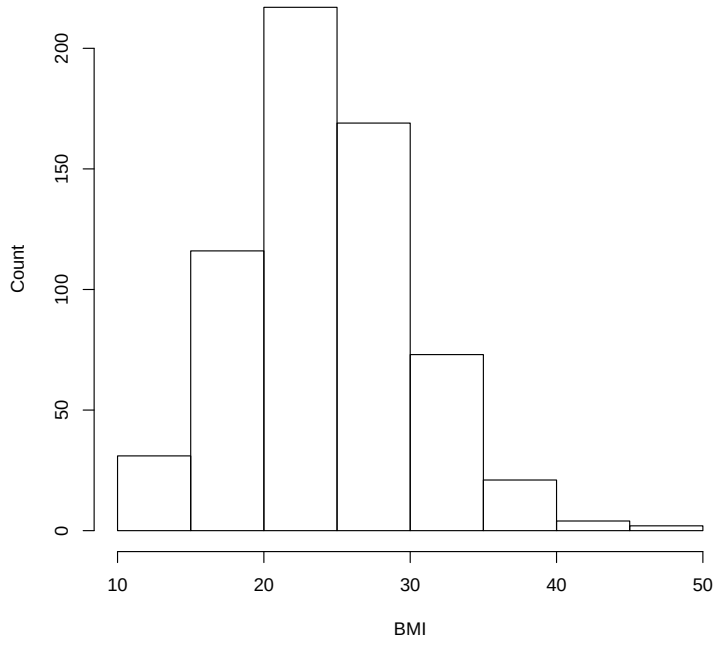
Disclaimer: If you use R, include all R code and output as attachments. Do not just "write in" the R code you used. Also, don't just write the answer and say this is what R gave you. If my grader can't see how you got an answer, it is wrong. I want to see your code and your answers accompanying your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

1. The time to failure (in hours) of a bearing used in a mechanical shaft is under investigation. A sample of $n = 84$ bearings produced the observations in Table 1 shown in Homework 4. For this problem, assume these observations form a random sample from a large population of bearings produced (e.g., a manufacturing process used to produce the bearings).
(a) Provide a point estimate of the population mean time to failure. Provide a point estimate of the population standard deviation time to failure. State the units attached to your estimates.
(b) Provide an estimate of the standard error of the sample mean $\overline{Y}$.

2. The School Breakfast Program (SBP) and National School Lunch Program (NSLP) are federally assisted meal programs that operate in participating public and private schools. In recent years, there has been an emerging concern that participation in these school-meal programs may be related to childhood obesity. In one study, there were $n = 633$ fourth-grade students selected from schools in Columbia, SC. Each student's body mass index (BMI, denoted by Y) was measured. For those of you who are not familiar with BMI, it is calculated as follows:

$$\text{BMI} = \frac{\text{weight (in kg)}}{[\text{height (in m)}]^2}.$$

I created a histogram of the $n = 633$ fourth-grade student BMI measurements:



(a) What is the population in this problem? What is the sample? **Note:** I don't think the answer to first question is all that clear-cut. There are certainly bad answers.
(b) Based on the histogram, which continuous probability distribution (from Chapter 4) seems like a reasonable probability model for BMI? Give a sound explanation.

(c) I cannot give you access to the data themselves because of confidentiality reasons. However, I calculated the sample mean to be $\overline{y} \approx 24.26$ and the sample standard deviation (sample variance) to be $s \approx 5.91$ ($s^2 \approx 34.94$). Using these values, estimate

- μ and σ^2 under a normal assumption for BMI
- α and λ under a gamma assumption for BMI

Hint: For the gamma distribution, what are the theoretical (model) values of $E(Y)$ and $\text{var}(Y)$? Note that $\overline{y}$ and s^2 are estimates of these theoretical values. Therefore, set the theoretical values equal to the sample values and solve for the model parameters in each case.
(d) Under a gamma assumption, estimate $P(Y > 30)$. How close is your answer to 100/633? (There were 100 students out of 633 with BMIs greater than 30 in the data set). If your answer is/is not "close," what does this suggest? **Note:** The value "30" is the cutoff for being classified as obese (according to the Centers for Disease Control and Prevention).

3. A cheese manufacturer is concerned that a supplier is adding water to their milk to increase profits. Adding water to milk raises the mean freezing temperature, which is $\mu = -0.545$ deg C (without water being added). A random sample of $n = 10$ batches of the supplier's milk yielded the following freezing temperatures:

−0.541	−0.538	−0.532	−0.533	−0.526
−0.543	−0.537	−0.528	−0.538	−0.549

(a) Use R to calculate the sample mean $\overline{y}$ and the sample standard deviation s for these data.
(b) If the supplier is not adding water to the milk, then the mean freezing temperature of $\mu = -0.545$ should be correct. Under this assumption, calculate the t statistic

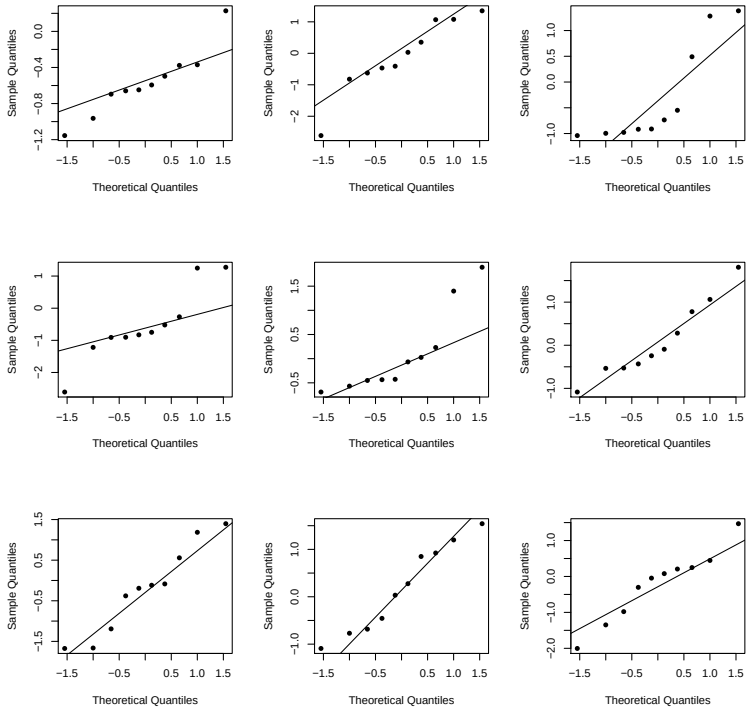
$$t = \frac{\overline{y} - \mu}{s/\sqrt{n}},$$

and plot your t statistic's value on the $t(9)$ density (like Figure 6.5 in the notes). Is your t statistic an "unusual" value from this distribution? If so, what does this suggest? If not, what does this suggest?
(c) Prepare a normal qq plot for the data. Note that a normal distribution assumption is needed for the t statistic in part (b) to be distributed according to $t(9)$. What does this plot suggest?

Note: With a small sample size like $n = 10$, the information we get from qq plots should always be analyzed with a grain of salt. *It's hard to make comfortable assessments with limited amounts of information.* This has motivated me to write Problem 4.

4. *Misinterpreting qq plots.* Analysts often erroneously place too much faith in qq plots when assessing whether a distribution adequately represents a data set (especially when the sample size is small). My goal in writing this problem is to illustrate to you the dangers that can arise.

In this problem, you will be generating your own data using R. The nice thing about R is that there are internal functions that generate random samples from most distributions. For specificity, you will use R to generate multiple random samples of $\mathcal{N}(0, 1)$ observations, each sample being of size $n = 10$. For example, I did this in R and got the following:




```
> B = 9 # number of simulated data sets
> n = 10 # sample size
> data = matrix(round(rnorm(n*B,0,1),3), nrow = B, ncol = n)
> data
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9]  [,10]
 [1,] -0.964  0.228 -0.659 -0.594 -0.696 -0.497 -0.648 -0.370 -1.155 -0.377
 [2,]  1.068  1.350 -0.822  1.079  0.031 -0.410 -0.467 -0.626 -2.614  0.353
 [3,] -0.549 -0.996 -0.912 -0.978  0.491  1.383 -0.736 -1.040 -0.917  1.281
 [4,] -0.909  1.274 -1.219 -0.829 -0.522 -0.904 -0.268  1.248 -2.608 -0.750
 [5,] -0.690  0.231  1.398  0.026 -0.435 -0.067 -0.427  1.891 -0.568 -0.451
 [6,] -1.085  0.779 -0.093  1.063 -0.433  1.809 -0.536 -0.529 -0.244  0.281
 [7,] -0.192 -1.663 -0.379  0.558 -1.674 -1.192 -0.117 -0.085  1.398  1.186
 [8,] -0.456  1.542 -1.091  0.276  1.199 -0.770  0.924 -0.684  0.033  0.850
 [9,] -1.348  0.207  0.446  0.248 -0.979 -0.046 -2.004  1.469 -0.303  0.079
```

Note that the rows contain 9 separate samples. Each row has $n = 10$ observations (sample size). I used the `round(,3)` function in R to round all observations to three decimal places. **Under a normal assumption,** I display each sample's qq plot in the figure above.

You will now use R to simulate the process of drawing repeated random samples from a $\mathcal{N}(0, 1)$ distribution and then creating qq plots like I have just done. Unfortunately, this will require some R programming on your part. However, before you storm into my office with tar and feathers, you can relax because I have written everything for you.

(a) Generate your own data and create a qq plot for each sample using this R code:

```
# create 3 by 3 figure
par(mfrow=c(3,3))
```

```
B = 9 # number of simulated data sets
n = 10 # sample size

# create matrix to hold all the data
# I removed the round function
# you don't have to print off the data
data = matrix(rnorm(n*B,0,1), nrow = B, ncol = n)
# for loop
# this creates a qq plot for each sample of data
for (i in 1:B){
    qqnorm(data[i,],pch=16,main="")
    qqline(data[i,])
    }
```

Print off your figure (for the data/plots you generated) and mark the qq plot that appears to violate the normal assumption the most (you don't have to print off your data values). **Remember:** In theory, all of these plots should display perfect linearity. Why? **Answer:** Because we are generating the data from a normal distribution.

> *Therefore, even when we create normal qq plots with normally distributed data, we can get plots that don't look perfectly linear.*

This is a consequence of natural sampling variability. This illustrates why you don't want to rush to discount a distribution as being plausible based on a single plot, especially when the sample size n is small (like $n = 10$).

(b) Now I want you to do some experimenting:

(i) Increase your sample size to $n = 100$ and repeat. What happens? What if $n = 1000$? Just change n in the R code above and re-run.

(ii) Take $n = 100$, replace

```
data = matrix(rnorm(n*B,0,1), nrow = B, ncol = n)
```

with

```
data = matrix(rexp(n*B,1), nrow = B, ncol = n)
```

and re-run. By doing this, you are changing the underlying population distribution from $\mathcal{N}(0, 1)$ to exponential(1). What do these normal qq plots look like now? What is the difference?

In each of (i) and (ii) above, write a short summary of what you observe and what you think is going on.