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. Arsenic is a chemical element (As) found naturally in ground water. Excessive levels may result from contamination caused by hazardous waste or industries that make or use arsenic. Environmental engineers collected a random sample of $n = 102$ water wells in Texas and recorded arsenic concentrations (in parts per billion, ppb). I have put the data on the course web site (under the Data sets section, called "arsenic").

(a) Based on the information stated above, what do you think the population is? There are no "right" answers here, but there are certainly good and bad answers.
(b) Prepare a histogram of the data. Prepare a boxplot of the data. Amend the code we used in Example 6.1 in the notes; you can change the `xlim` option for these data or just take it out. To see where the five-number summary comes from in the boxplot, type `quantile(arsenic)` in R.
(c) Calculate an estimate of the population mean concentration μ. Calculate an estimate of the population standard deviation concentration σ.
(d) Based on the shape of the histogram and boxplot, what continuous probability distribution might be a reasonable model to describe the population? Explain your reasoning.
(e) How might you estimate the proportion of wells in the population whose arsenic levels exceed 20 ppb? I can think of an easy way to do this and a more interesting way based on answers to part (c) and (d). See if you can come up with two ways to do it.

2. The World Health Organization uses a normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 15$ to describe the systolic blood pressure (SBP) of all American males aged 18 and over (i.e., this is the population).

(a) Use R to graph this population distribution. Use the 68-95-99.7 rule to form intervals 1, 2, and 3 standard deviations from the mean. Interpret.
(b) Calculate the probability a randomly selected individual from this population has SBP larger than 140.
(c) Suppose one observes a random sample of $n = 25$ American males from this population. What is the probability the sample mean SBP $\overline{Y}$ is larger than 140? Why is this answer different from your answer in part (b)?

3. A shock absorber is a suspension component that controls the up-and-down motion of a vehicle's wheels. The data on the next page are $n = 38$ distances (in km) to failure for a specific brand of shock absorber observed under "extreme" driving conditions.

(a) Prepare a histogram of the data. Prepare a boxplot of the data. Based on the

shape of the histogram and boxplot, what continuous probability distribution might be a reasonable model to describe the population? Explain your reasoning. What might even constitute "the population" in this problem?

6700	6950	7820	9120	9660	9820	11310	11690	11850	11880
12140	12200	12870	13150	13330	13470	14040	14300	17520	17540
17890	18450	18960	18980	19410	20100	20100	20150	20320	20900
22700	23490	26510	27410	27490	27890	28100	30050		

(b) Use R to calculate the sample mean $\overline{y}$ and the sample standard deviation s for these data.
(c) The data above have been collected by the manufacturer of the shock absorber as part of a quality control program to assess the reliability of its parts. One member of the manufacturer's quality assessment team believes that their shock absorber population mean distance to failure is 18,000 km under "extreme" driving conditions. Under this assumption, calculate the t statistic

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

and plot your t statistic's value on the $t(37)$ density (like Figure 6.5 in the notes). Is your t statistic an unusual value from this distribution? If so, what does this suggest about the team member's claim that $\mu = 18000$ km? If not, what does this suggest?
(d) Did you suggest a normal population distribution in part (a)? Prepare a normal qq plot for the data to see if you should have. Note that a normal population distribution assumption is needed for the t statistic in part (c) to be distributed according to $t(37)$. What does the qq plot suggest? Does this affect your analysis in part (c)? Comment on robustness here.