36-220 Lab #7
Point Estimation and Bootstrapping

Please write your name below, tear off this front page and give it to a teaching assistant as you leave the lab. It will be a record of your participation in the lab. Please remember to include whether you are in Section A or B. Keep the rest of your lab write-up as a reference for doing homework and studying for exams.

Name:

Section:

- The symbol ♣ at the beginning of a question means that, after you answer that question, you should raise your hand and have either the TA or lab assistant review your answer. Once they have reviewed your work they will place a check in the appropriate space in the table below. The purpose of this check is to be sure you have answered the question correctly.
- **You should try to complete as much of the lab exercise as possible. We understand that students work at different paces and have tried to structure the exercise so that it can be completed in the allotted time. If you work systematically through the handout and still don't complete every question don't worry. The important thing is that you understand what you are doing. Nonetheless, you are encouraged to complete the lab on your own.**

Check-Problem ♣	Instructor's Initials
Question 3	
Question 7	

36-220 Lab #7
Point Estimation and Bootstrapping

1 Bias of Point Estimators

A point estimator is **biased** if, on average, it over- or under- estimates the true parameter. We will illustrate this by trying to estimate the population variance.

1. Create a new MINITAB worksheet by accessing **File** → **New** from the pull-down menus and selecting "MINITAB Worksheet."

2. Create 10 columns of 500 random Normal(0,1) observations. To do this, select **Calc** → **Random Data** → **Normal** from the pull-down menus. Enter "500" in the "Generate ... rows of data" field. Enter "C1-C10" in the "Store in column(s)" field. Click **OK**.

3. Put the mean of each row in column 11. To do this, select **Calc** → **Row Statistics** from the pull-down menu. Under "Statistic" select "Mean". Under "Input Variables" type "C1-C10". Under "Store results in" type "C11". Click **OK**.

4. Recall that the sample variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \quad (1)$$

   We'd like to calculate the sample variance for each row. In order to do this in MINITAB, we must first calculate the sample standard deviation. This is done by selecting **Calc** → **Row Statistics** from the pull-down menu. Under "Statistic" select "Standard Deviation". Under "Input Variables" type "C1-C10". Under "Store results in" type "C12". **OK**.

5. The sample standard deviation is the square root of the sample variance. In order to get the sample variance, we must square the sample standard deviation. To do this, select **Calc** → **Calculator**. Under "Store result in variable", type "C13". Under "Expression", type "C12 * C12". Click **OK**. We now have the sample variances of our 500 sets of 10 Normal(0,1) observations stored in C13.

6. The sample variance, given in Equation 1 is an estimate of the true, population variance. Note that the fraction in front of the summation of sample variance is $\frac{1}{n-1}$. An alternative estimate of the sample variance,s_0^2, is given in Equation 2:

$$s_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (2)$$

You'll notice by looking at Equations 1 and 2 that $s_0^2 = \frac{n-1}{n}s^2$. Let's calculate s_0^2 in MINITAB by doing the following: Select **Calc → Calculator** from the pull-down menus. Under "Store result in variable", type "C14". Under "Expression", type ".9*C13" (**Note: Since our n is 10, $\frac{n-1}{n} = \frac{9}{10} = .9$, which accounts for the .9 in this formula**). Click **OK**.

Question #1: Take the first 100 sample means from column 11 and copy them into a blank column. Now consider two averages, first that of the 100 sample means and then that of the 500 sample means (you can do this by looking at the mean given by executing **Stat → Basic Statistics → Display Descriptive Statistics** and selecting the appropriate column as your variable). What values did you expect to get? Do the results surprise you?

Question #2: What is the average of your 500 estimates of the sample variance, s^2? What is the average of your 500 estimates of s_0^2?

♣**Question #3**: Which estimator, s^2 or s_0^2, is closer to the population variance? Does s_0^2 over-estimate or under-estimate the population variance? If so, why?

Question #4: Consider the sample variance of each estimator,s^2 or s_0^2, from the results generated during question #8. Which of the two estimators has a smaller sample variance?

2 Bootstrapping

Bootstrapping is a general method for approximating the error properties of estimators by means of computer simulation. If we knew the sampling distribution of an estimator, $\hat{\theta}$, we could then work out its bias, $\mathbf{E}\left[\hat{\theta}\right] - \theta$, its variance $\mathrm{Var}\left(\hat{\theta}\right)$, etc. In general, the sampling distribution is very complicated, and doesn't have a closed, analytical form. Bootstrapping gets around this by *simulating* many random samples, and applying the estimator to each one. This then gives us an approximation of the sampling distribution of the estimator, from which we can calculate properties like bias and standard error. If we want to approximate the standard error, for instance, in a parameter estimate $\hat{\theta}$, which we got from a sample of size n, we'd proceed as follows.

1. Generate n random numbers, following the probability distribution with parameter $\hat{\theta}$; call these $z^*_{1,1}, z^*_{1,2}, \ldots z^*_{1,n}$. Use these to calculate a new bootstrap estimate, $\hat{\theta}^*_1$
2. Generate another n random numbers, $z^*_{2,1}, z^*_{2,2}, \ldots z^*_{2,n}$, and calculate another bootstrap estimate, $\hat{\theta}^*_2$
3. Repeat B times to get bootstrap estimates $\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots \hat{\theta}^*_B$
4. The bootstrap standard error is

$$s_{\hat{\theta}} = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}^*_i - \overline{\theta^*}\right)^2} \tag{3}$$

By the law of large numbers, if B is large, then the distribution of the bootstrap estimates $\hat{\theta}^*$ will be very close to the true sampling distribution, so the bootstrap standard error (or any other reasonable function of the distribution) will be close to its true, population value.
(This is called *parametric* bootstrapping because we re-use the parameter value we estimated from our original data. There is a variant, *non-parametric* bootstrapping, where we treat our original sample as a complete population, and draw new samples from it. There are advantages and dis-advantages to both procedures. Parametric bootstrapping is, however, much easier to do in MINITAB!)

1. Open a new worksheet in MINITAB.
2. Label the first column "X". Fill it with 10 simulated random variables which have the exponential($\lambda = 1$) distribution. Use **Calc** → **Random Data** → **Exponential**.
3. Compute the mean of the values in the first column. If X has the exponential(λ) distribution, then $\mathbf{E}[X] = 1/\lambda$ and $\lambda = 1/\mathbf{E}[X]$. Hence a reasonable estimate of λ is $1/\overline{X}$.

Question #5: What is your estimate $\hat{\lambda}$ for λ? What is the squared error of your estimate?

4. Label the next ten columns "Z1", "Z2" and so on through "Z10".

5. Fill each of these ten columns with 1000 simulated random variables which have the exponential($\hat{\lambda}$) distribution, where $\hat{\lambda}$, again, is your estimate for λ. **N.B.**, you must set the "scale" parameter here, and in MINITAB, that is $1/\lambda$.

6. Label the 12th column "Zbar", and fill it with the sample mean for each row. In other words, the preceeding ten columns are the values of $Z_1, Z_2, \ldots Z_{10}$; now put $\overline{Z}$ in the 12th column. Use **Calc** $\rightarrow$ **Row Statistics**.

7. Label the next column "Lstar". This is where you will calculate the bootstrap estimate for each simulated sample. Since $\hat{\lambda} = 1/\overline{X}$, we want $\lambda^* = 1/\overline{Z}$. Use **Calc** $\rightarrow$ **Calculator**.

8. Calculate the mean and sample standard deviation of the values in the "Lstar" column.

Question #6: Why do we have ten columns "Z1" ... "Z10"?

♣ **Question #7**: What is the sample standard deviation of "Lstar"? What is your bootstrap estimate of the standard error?

Question #8: What is the mean of "Lstar"? Does this lead you to believe that $1/\overline{X}$ is an unbiased estimate of λ or not? (Explain!)