STATISTICS 641 - ASSIGNMENT 6

DUE DATE: Noon (CDT), FRIDAY, MARCH 27, 2015

Name _____

Email Address _____

Please TYPE your name and email address. Often we have difficulty in reading the handwritten names and email addresses. Make this cover sheet the first page of your Solutions.

STATISTICS 641 - ASSIGNMENT 6 - Due Noon (CDT), FRIDAY - 03/27/2015

• Read Handouts 9 & 10 and Chapter 5, Sections 6.1, 14.6.2, and 15.1 in the Textbook

I. (28 points) A company has designed a new battery system for electric powered automobiles. To estimate the lifetime of the system, the design engineers place the batteries in 25 electric powered cars and test them under simulated city driving. Let Y_i be the time to failure of the batteries of the ith car, $i = 1, \ldots, 25$. The failure times are recorded in units of 10,000 miles. The company wants to know the probability that the sample mean based on 25 observations will estimate the true mean within a margin of error of $\pm.2$ (2000 miles), provided that the true mean has a value of 5 (50,000 miles), that is, approximate, $P[-0.2 \leq \bar{Y} - 5 \leq 0.2]$.

1. From past studies, the distribution of the time to failure of the braking system is exponential with a mean value of 5 (50,000 miles). Let $\bar{Y}$ be the sample mean time to failure of the 25 cars.
   a. What is the **exact** distribution of $\bar{Y}$ if the exponential model is still valid?
   b. What is the mean and standard deviation of $\bar{Y}$ if the exponential model is still valid?
2. Simulate 1000 random samples of size 25 from the an exponential distribution with $\beta = 5$. Compute the sample mean from each of the 1000 samples. Display a normal distribution reference plot for the 1000 sample means. Does the plot suggest that the sampling distribution of $\bar{Y}$ is approximately normal? Perform a GOF test of normality using your 1000 sample means and report your results.
3. Compute or estimate $P[-0.2 \leq \bar{Y} - 5 \leq 0.2]$ in each of the following manners:
   a. Using the exact distribution of $\bar{Y}$, compute $P[-0.2 \leq \bar{Y} - 5 \leq 0.2]$.
   b. Using the central limit theorem, compute $P[-0.2 \leq \bar{Y} - 5 \leq 0.2]$.
   c. Using your simulated 1000 $\bar{Y}$s, approximate, $P[-0.2 \leq \bar{Y} - 5 \leq 0.2]$.
4. Compare the three computations/estimators of $P[-0.2 \leq \bar{Y} - 5 \leq 0.2]$.

II. (24 points) A major problem in the Gulf of Mexico is the excessive capture of game fish by shrimpers. A random sample of the catch of 50 shrimpers yield the following data concerning the catch per unit effort (CPUE) of Red Snappers, a highly sought game fish. Let C_i be the CPUE for the ith shrimper. The data, $C_1, C_2, \ldots, C_{50}$ is given next.

0.6	0.7	1.1	1.3	1.8	2.0	2.3	2.7	2.9	3.1
3.9	4.3	4.4	4.9	5.2	5.4	6.1	6.8	7.1	8.0
9.4	10.3	12.9	15.9	16.0	22.0	22.2	22.5	23.0	23.1
23.9	26.5	26.7	28.4	28.5	32.2	40.2	42.5	47.2	48.3
55.8	57.0	57.2	64.9	67.6	71.3	79.5	114.5	128.6	293.5

1. Use the R program from Handout 10 (or any other program of your choice) to draw 5000 bootstrap samples from the CPUE data. From the 5000 samples, estimate the standard error of the sample mean for the $Y_i = log(C_i)$, data. Compare this estimate to the usual estimate $\frac{S_Y}{\sqrt{n}}$, where S_Y is the sample standard deviation computed from the $n = 50$ values of $Y_i = log(C_i)$.
2. Use your bootstrap samples to estimate the mean and standard deviation of the following sample statistics for $Y = log(C)$
   a. The sample median, $\widehat{Q}_Y(.5)$
   b. The sample standard deviation, S_Y
   c. The sample MAD, $\widehat{MAD}_Y$
3. Historically, the log(CPUE) data was modeled as a random sample from a $N(3, (1.5)^2)$ distribution. Compare your bootstrap estimates of the mean and standard deviation of $\widehat{Q}_Y(.5)$ and S_Y from part 2. of this problem to the theoretical mean and standard deviation of $\widehat{Q}_Y(.5)$ and S_Y based on $Y = log(C)$ having a $N(3, (1.5)^2)$ distribution.

III. (24 points) Answer the following questions:

1. In an extrasensory perception (ESP) experiment, five choices are offered for each question. Assume that a person without ESP guesses randomly and thus correctly answers with probability 1/5. Further assume that the responses are independent. Suppose that 100 questions are asked.

   a. What are the mean and standard deviation of the number of correct answers?

   b. What are the mean and standard deviation of the proportion of correct answers?

   c. What is the probability that a person without ESP will correctly answer at least 30 or the 100 questions? Compute the probability both exactly and using a normal approximation.

2. If a random sample of size n is drawn from a normal distribution, the standardized sample variance $U = (n-1)S^2/\sigma^2$ has a chi-squared distribution with df=n-1. Use a simulation to investigate the impact of deviations from normality on the distribution of U.

   a. Generate 1000 random samples of size 10 from an $N(20, 5^2)$ distribution. Calculate $U_i = (10-1)S_i^2/25$ from each sample, $i = 1, 2, \ldots, 1000$. Compare the sample quantiles for the distribution of U, $\hat{Q}(u)$ for $u = .1, .25, .5, .9, .95, .99$ to the corresponding percentiles for a chi-square distribution with df=9.

   b. Repeat a., but now draw your samples from a Y = 18.4+15.8*G where G has gamma distribution with parameters: $(\alpha = .1, \beta = 1)$ (E(Y) = 20, Var(Y)=25)

   c. Repeat a., but now draw your samples from a X = 20+2.9*T where T has a t-distribution with df=3 (E(X) = 20, Var(X)=25)

   d. What is the impact of skewness and heavy tails on the sampling distribution of U?

3. In sample surveys, if people are asked a sensitive or a potentially stigmatizing question (e.g., "Have you ever shoplifted?" or "Do you use illegal drugs?"), the respondent often give an evasive answer or refuse to answer at all, which introduces a bias in the estimate of the proportion, p, of people in the population belonging to the sensitive group. To avoid this bias and still protect the privacy of the survey respondents, a technique called **randomized response** is implemented. The respondent uses a random device to select one of the following two questions:

   Q1. I belong to the sensitive group: Answer "Yes" or "No."

   Q2. I do not belong to the sensitive group: Answer "Yes" or "No."

   The interviewer does not see the which question the respondent receives from the random device and hence does not know which question was answered thus protecting the respondent's privacy. However, the interview does know the probability $\theta \neq 1/2$ that the random device selects Question 1.

   Let π denote the probability of a "Yes" answer by the respondent and let p be the population proportion of people belonging to the sensitive group.

   a. Assuming the respondent answers truthfully, show that

   $$\pi = \text{P(Yes)} = p\theta + (1-p)(1-\theta) = p(2\theta - 1) + (1-\theta)$$

   b. Suppose $\hat{\pi}$ is the proportion of "Yes" responses in a random sample of n people. Show that

   $$\hat{p} = \frac{\hat{\pi} - (1-\theta)}{2\theta - 1}$$

   is an unbiased estimator of p.

   c. Show that

   $$Var(\hat{p}) = \frac{[p(2\theta-1) + (1-\theta)][\theta - p(2\theta-1)]}{n(2\theta-1)^2} = \frac{p(1-p)}{n} + \frac{\theta(1-\theta)}{n(2\theta-1)^2}$$

   d. What is the impact on $Var(\hat{p})$ as θ moves away from 1/2 towards 0 or 1?

IV. (24 points) **Multiple Choice Questions** Select the letter of the **BEST** answer. Justify your answer using 20 words or less.

1. A researcher is attempting to estimate the average number of miles between recharging of the batteries for a battery operated car. There are several potential estimators of the average. The *best* approach for selecting an estimator would be to

   A. always select the unbiased estimator because its average value equals the true value of the parameter.

   B. select the estimator with the smallest variance because then the estimator would have the least change from sample to sample.

   C. select the estimator with the smallest average squared distance from the parameter.

   D. select the estimator with the smallest bias.

   E. all the above are appropriate answers

2. The bootstrap procedure for obtaining the sampling distribution for an estimator of the parameter θ is often used instead of an exact derivation or an asymptotic result when

   A. the researcher wants just a rough idea of the sampling distribution

   B. the exact procedure is too involved mathematically

   C. when the conditions for applying the asymptotic procedure are not satisfied.

   D. a nonparametric and parametric procedure yield very different distributions

   E. all of the above are true

3. The two factors which affect the accuracy of using the bootstrap procedure to estimate the percentiles of the sampling distribution of the Maximum Likelihood Estimator (MLE) are

   A. how well the edf $\widehat{F}$ matches the population cdf F and the number of bootstrap samples B

   B. the sample size n and the number of bootstrap samples B

   C. the sample size n and how well the edf $\widehat{F}$ matches the population cdf F

   D. the number of bootstrap samples B and the shape of the population cdf F

   E. All of the above are equally important

4. A random sample of n units is selected from a process having pdf $f(\cdot)$. If $f(\cdot)$ is symmetric about the parameter θ then we can conclude that

   A. the sample mean is a better estimator of θ than is the sample median.

   B. the sample mean and sample median are equivalent estimators of θ.

   C. the sample mean has a smaller variance than the sample median as an estimator of θ.

   D. the sample median has a smaller MSE than the sample mean as an estimator of θ.

   E. the estimator having smallest MSE depends on the form of $f(\cdot)$.

5. A researcher takes a random sample of n units from a population and wants to estimate the population parameter θ using the statistics $\hat{\theta}$. Which one of the following statements are true concerning the sampling distribution of $\hat{\theta}$?

   A. If n is large enough, we can use the normal distribution to accurately approximate the sampling distribution of $\hat{\theta}$.

   B. The bootstrap procedure will yield an accurate approximation of the the sampling distribution of $\hat{\theta}$ provided we generate a very large number of bootstrap samples.

   C. We can always remove the dependency of the sampling distribution of $\hat{\theta}$ on θ by using the studentized version $(\hat{\theta} - \theta)/\hat{\sigma}_{\hat{\theta}}$.

   D. All of the above

   E. None of the above

6. A cancer researcher is studying a parameter θ associated with the effectiveness of a new treatment, namely, the residual time a drug spends in the kidney of test rats. She wants to determine an estimator of θ based on injecting 100 rats with a standard dosage of the drug. There are a number of possible estimators of θ in the literature. In attempting to select the most **effective** estimator for the researcher, you recommend that she

   A. use the estimator having the smallest variance because then this estimator would not be very different from study to study.

   B. use the unbiased estimator because then on the average she will have a correct estimator.

   C. use the estimator with the smallest bias because then the mean squared error would be small.

   D. use the estimator with the smallest mean squared error because then the estimator would have the highest concentration of values about θ.

   E. use the sample mean when the sample size is very large.

7. A sociologist is interested in the range of scores on a test used to determine the level of aggression in juveniles who have been arrested at least once before the age of 12. She has a random sample of 247 such juveniles and estimates the range using the statistic $R = A_{(247)} - A_{(1)}$, where $A_{(i)}$s are ordered 247 test scores. The sampling distribution of the R could be obtained by

   A. using the Central Limit Theorem and a normal distribution approximation.

   B. using an Exact Derivation of the distribution of R from the theory of order statistics.

   C. using a large Simulation study with $B > 10000$ samples of size $n = 247$ and then approximating the sampling distribution of R using the empirical distribution function of the B values of R.

   D. using a Bootstrap procedure with $B > 10000$ samples of size $n = 247$ and then approximating the sampling distribution of R using the empirical distribution function of the B values of R.

   E. None of the above would be appropriate

8. The sample estimator $\hat{\theta}$ of a population parameter θ is unbiased. Unbiased means that the estimator $\hat{\theta}$

   A. is 95% certain of being close to θ.

   B. has a sampling distribution which is symmetric about θ.

   C. has a smaller mean squared error than a biased estimator.

   D. has a sampling distribution with mean value θ.

   E. all the above