STATISTICS 641 - EXAM II

Student's Name ____

Student's Email Address ____

INSTRUCTIONS FOR STUDENTS:

(1) The exam consists of 7 pages including the cover page, and 20 pages of Tables.

(2) You have exactly **70 minutes** to complete the exam.

(3) Show *ALL* your work on the exam pages.

(4) Do not discuss or provide information to anyone concerning the questions on this exam or your solutions until I post the solutions to the exam.

(5) You may use the following:

- Calculator - Your device cannot facilitate a connection to the internet or to send text messages
- Summary Sheets - (**4-pages,** 8.5"**x**11"**, write/type/paste on both sides of the four sheets**)
- Tables for Exam 2 which are attached.

(6) Do not use any other written material except for your summary sheets and the attachments to the exam.

(7) Do not use a computer, cell phone, or any other electronic device (other than a calculator).

I attest that I spent no more than 70 minutes to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.

Student's Signature________________________________

I. (40 points) CIRCLE (A, B, C, D, or E) corresponding to the **BEST** answer. Only one letter is allowed per question. Partial credit will be given on problems if you show your calculations.

(1.) A metallurgist is studying the tensile strength of a new alloy. She is concerned that this alloy may have highly variable strengths which can be detrimental in its use in airplane wings. She measured the tensile strengths of a large number of batches of the new alloy and wants to establish a lower bound on the tensile strength for this alloy such that the lower bound would have a very high degree of certainty that 99% of all batches of the alloy would have a tensile strength that would exceed this bound.

A. The computed boundl would be a **lower confidence bound**.

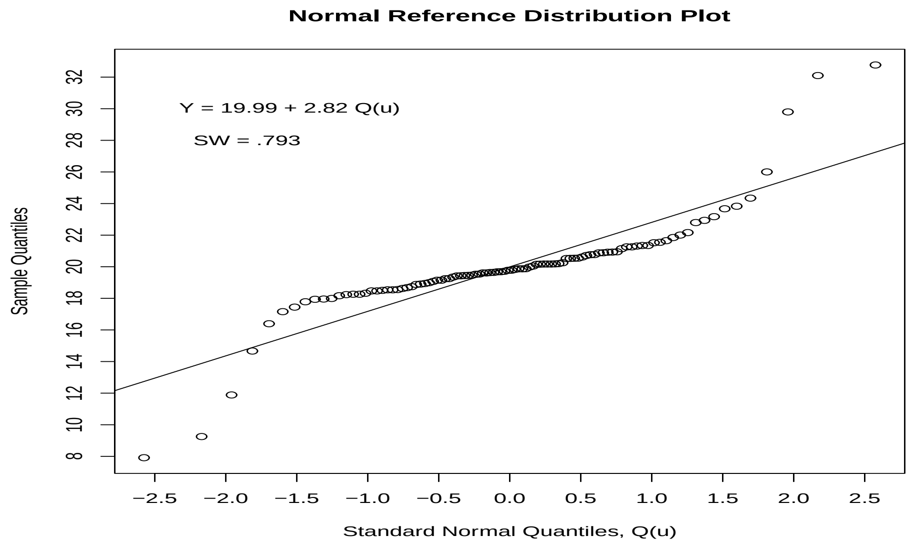
B. The computed interval would be a **lower prediction bound**.

[✓ C.] The computed interval would be a **lower tolerance bound** .

D. The computed interval would be a **lower natural bound**.

E. None of the above would be appropriate.

(2.) An entomologist conducts an experiment to study the side effects of a new pesticide. She simultaneously exposes 100 mice to the pesticide. The amount of pesticide in their bladder (ppm) was recorded 30 days after the exposure. The 100 values are displayed in the following normal reference distribution plot. The Shapiro-Wilk test yielded SW = .793 and plotted line is $\widehat{Q}(u) = 19.99 + 2.82 Q_Z(u)$. Select the distribution which best describes the cdf, F, for the amount of pesticide in the mice's bladder.



A. $F(\cdot)$ has a N(20,3) distribution

[✓ B.] $F(\cdot)$ has a shifted t-distribution with df=3 and centered at 20 ppm

C. $F(\cdot)$ has a shifted t-distribution with df=40 and centered at 20 ppm

D. $F(\cdot)$ has an Exponential($\beta = 20$) distribution

E. None of the above are an appropriate model based on the 100 data values

(3.) A 95% confidence interval for the population mean is reported as (2.4, 3.6). The correct interpretation of this interval is

[✓ A.] In repeated sampling, 95% of the confidence intervals produced will contain the population mean.

B. There is a 95% probability that the population mean is between 2.4 and 3.6.

C. Approximately 95% of the time the population mean is between 2.4 and 3.6.

D. There is a 95% probability that the sample mean will be between 2.4 and 3.6.

E. All of the above are correct interpretations.

(4.) A biased estimator with a small variance may not be desirable because

A. its mean square error would be larger than the mean square error of an unbiased estimator

B. its average value would not be close to the parameter being estimated

[✓ C.] its sampling distribution would be highly concentrated about a value other than the parameter being estimated

D. its sampling distribution would be unknown whereas the sampling distribution of unbiased estimators are approximately normal

E. none of the above

(5.) An environmentalist samples the outflow from a chemical plant's discharge pipe daily for 85 days and obtains the following lead levels in the output: $L_1, \cdots, L_{85}$. She determines that the daily lead levels, L_t, are related by $L_t = \theta + \rho L_{t-1} + e_t$, where the e_ts have independent $N(0, \sigma_e^2)$ distributions and $\rho = .91$. The engineer constructs a nominal 90% confidence interval for the average daily lead level, μ, using the formula $\bar{L} \pm t_{.05,84}(S/\sqrt{85})$, where $\bar{L}$ and S are the sample mean and standard deviation for the 85 lead levels. The true coverage probability of this confidence interval

A. is 0.90.

[✓ B.] is much less than 0.90.

C. is very close to 0.90.

D. is much greater than 0.90.

E. may be greater than 0.90 or less than 0.90 depending on the distribution of the L_t's.

(6.) A researcher takes a random sample of n units from a population and wants to estimate the population parameter θ using the statistic $\hat{\theta}$. Which ONE of the following statements is TRUE concerning the sampling distribution of $\hat{\theta}$?

A. If $\hat{\theta}$ is the MLE of θ and 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 use the studentized version of the estimator.

C. The sampling distribution of the sample quantile $\hat{Q}(u)$ can be adequately approximated by a normal distribution provided $n \geq 30$.

D. The Box-Cox procedure will nearly always yield a transformation of the data such that the sampling distribution of $\hat{\theta}$ is very nearly a normally distribution.

[✓ E.] None of the above statements are true.

(7.) A large organization wants to compare the job satisfaction of people with college degrees to people with a terminal high school education. A random sample of size $n_1 = 123$ is taken from the 13,000 people with a college degree and a random sample of size $n_2 = 350$ is taken from the 1,950,000 persons with just a high school education. The sample means $\bar{Y}_1$ and $\bar{Y}_2$ are computed from the two samples. Which **ONE** of the following statements is **TRUE** about the bias in using $\bar{Y}_1$ as an estimator of μ_1, the population mean of college degree holders and $\bar{Y}_2$ as an estimator μ_2, the population mean of High School grads?

A. $\bar{Y}_1$ has a smaller bias than $\bar{Y}_2$.

B. $\bar{Y}_1$ has a larger bias than $\bar{Y}_2$.

C. $\bar{Y}_1$ and $\bar{Y}_2$ have the same positive bias.

D. The estimator having greatest bias depends on the shape of the population pdf's.

[✓ E.] None of the above statements are true

(8.) A plant physiologist is studying the infestation rate of potato bud insects on genetically altered potato plants. The researcher measures the infestation rate, $Y_1, \cdots, Y_{300}$, of potato bud insects on 300 randomly selected genetically altered plants. The researcher wants to determine whether the cdf of Y, F has a logistic distribution. However, the technician recorded the exact infestation rate if the number of insects per plant was less than 25 insects but just recorded $Y_i = 25$ for those plants having more than 25 insects. The appropriate method to modify the reference distribution plot to take into account the under-reporting of Y_i is to

A. plot just the uncensored data values

B. plot both the censored and uncensored data values because the sample size is so large ($n = 300$)

C. use the Chi-square Goodness-of-Fit statistic because the data is discrete

[✓ D.] plot just the values of $Y_i < 25$ but use $n = 300$ in computing the logistic quantiles, $Q(u_i)$, $u_i = (i - .5)/n$

E. None of the above would be appropriate modifications

(9.) A medical researcher wants to construct a $(P, \gamma) = (.99, 95)$ tolerance interval for the systolic blood pressures of elderly males based on a random sample of 250 subjects $Y_1, \ldots, Y_{250}$. A plot of the data reveals that the data is highly right skewed. You apply a Box-Cox transformation and compute a p-value = .456 from the Shapiro-Wilk statistic for the transformation $X = Y^{.25}$. Which of the following methods would you recommend for constructing the tolerance interval?

A. A studentized Bootstrap procedure because tolerance intervals cannot be inverted.

B. Use the normal-based tolerance interval because $n = 250$ is very large: $\left(\bar{Y} - K_{(.99,.95)} S_Y, \bar{Y} + K_{(.99,.95)} S_Y\right)$

C. Use a nonparametric tolerance interval because $n = 250$ is very large

[✓ D.] Use the inverse of the normal-based tolerance interval: $\left(\left[\bar{X} - K_{(.99,.95)} S_X\right]^4, \left[\bar{X} + K_{(.99,.95)} S_X\right]^4\right)$

E. None of the above procedures would work very well

(10.) A political scientist wants to estimate the proportion, p, of registered voters that are members of the Socialist party. From previous elections, it is known that p is less than 0.2. How large a random sample of registered voters must the scientist select in order to be 95% confident that the sample proportion is within .05 of the true value of p?

A. 97

[✓ B.] 246

C. 385

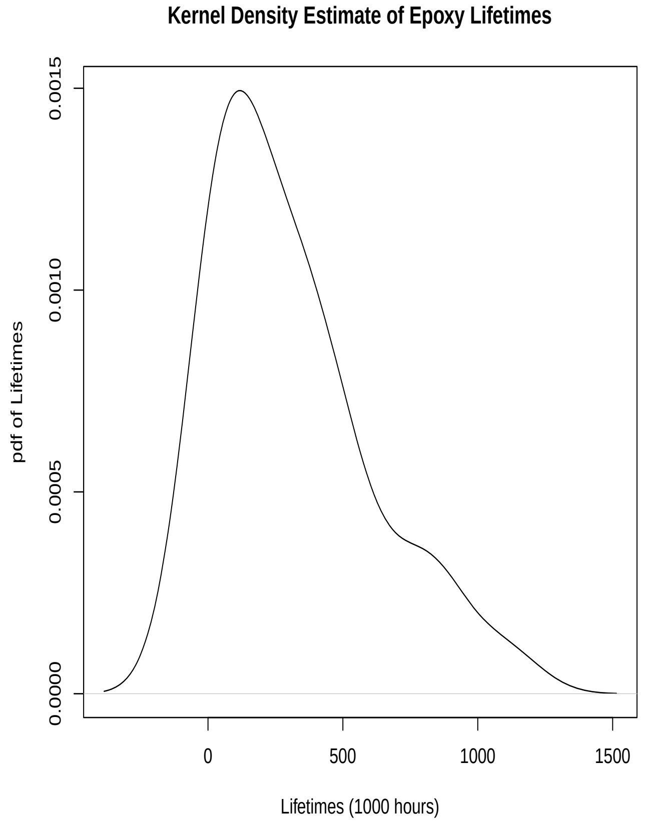
D. 425

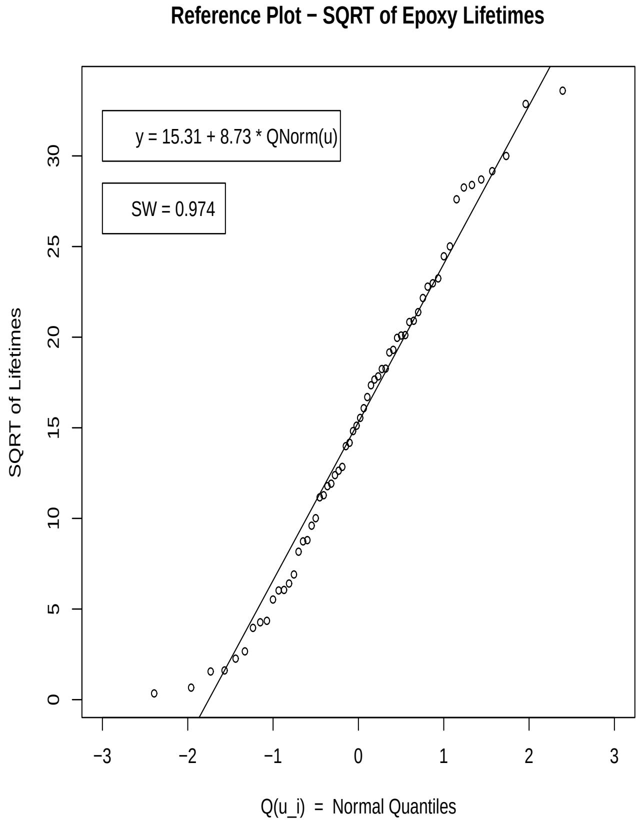
E. cannot be determined without further information

Part II. (60 points) The International Space Station uses epoxy vessels as storage devices. The epoxy needs to be able to function in an environment of sustained pressure. A study of a newly developed epoxy subjected to sustained stress was commissioned. Sixty strands of the epoxy are placed on test at a prescribed stress. The lifetimes in **thousands of hours** of the 60 strands are given in the following table. The space agency wants to estimate various quantities associated with the lifetimes of the stressed epoxy strands.

0.1	0.4	2.4	2.6	5.1	7.1	15.7	18.3	19.0	30.6	36.3	36.6
41.0	47.8	66.6	76.3	77.5	92.1	100.3	124.6	127.2	138.7	142.0	153.5
159.6	164.9	195.7	200.8	219.6	228.4	241.7	258.5	278.8	301.0	311.9	318.1
332.9	333.5	366.9	372.5	398.4	403.6	404.6	434.2	437.0	457.0	490.9	519.5
527.7	540.3	598.4	625.6	762.2	798.5	806.5	823.8	850.4	899.7	1080.4	1129.2

Summary statistics, a plot of the estimated pdf for the lifetimes, Y_i, and a plot of $X_i = \sqrt{Y_{(i)}}$ versus $Q_o(u_i) =$ Normal quantiles are given below, where $u_i = (i - .5)/60$ for $i = 1, \cdots, 60$.





Summary Statistics for Lifetimes, Y

Minimum	$\widehat{Q}(.25)$	$\widehat{Q}(.5)$	Mean	$\widehat{Q}(.75)$	Maximum	SIQR	S	MAD
0.1	73.9	235.1	310.6	442.0	1129.2	184.1	291.0	282.7

Summary Statistics for SQRT(Lifetimes), X

Minimum	$\widehat{Q}(.25)$	$\widehat{Q}(.5)$	Mean	$\widehat{Q}(.75)$	Maximum	SIQR	S	MAD
0.34	8.59	15.33	15.31	21.02	33.60	6.22	8.80	9.73

(1.) (15 points) The Shapiro-Wilk test for using a normal distribution to model the square root of epoxy lifetimes produced W = 0.974. Based on the Shapiro-Wilk test, the data summaries, and the plots on page 5, does a normal distribution appear to be an appropriate distribution for the square root transformation of the lifetime data? Justify your answer.

(2.) (15 points) Provide the researchers with an interval of values $(L_{Life},\ U_{Life})$ such that the researchers would be 95% confident that the interval $(L_{Life},\ U_{Life})$ would contain the lifetimes for least 90% of all epoxy cords exposed to the prescribed stress.

(3.) (15 points) Estimate using a 95% C.I. the proportion of cords receiving the prescribed stress that would have a lifetime greater than 750,000 hours.

(4.) (15 points) A future study has been proposed to obtain more precise estimates of the mean time to failure of the stressed epoxy cords. How large a sample size is needed to obtain an estimate of the mean lifetime for which the researchers would be 99% confident that the estimated mean lifetime is within 5000 hours of the true mean lifetime?

SOLUTIONS STAT 641 - EXAM II

I. (40 points) CIRCLE (A, B, C, D, or E) corresponding to the **BEST** answer. Only one letter is allowed per question. Partial credit will be given on problems if you show your calculations.

(1.) **C.** the metallurgist wanted a lower bound with 99% of population values greater than the bound

(2.) **B.** $F(\cdot)$ has a shifted t-distribution with df=3 because the plot reveals a symmetric distribution with tails heavier than a normal distribution.

(3.) **A.** This is the definition of a C.I.

(4.) **C.** If the estimator is biased with a small variance then most of its values will be near its expected value

(5.) **B.** Because of the positive correlation, $S/\sqrt{n}$ will under estimate the true standard error of $\hat{L}$ and hence the confidence interval will be too narrow to be a 90% C.I.

(6.) **E.**

A. If $\hat{\theta} = Y_{(n)}$ then the asymptotic distribution is an extreme value distribution

B. The accuracy of the bootstrap procedure also depends on how well the edf, $\hat{F}$, matches the population cdf F

C. Recall the examples in HO 10.

D. The Box-Cox procedure works well only with data from skewed or heavy-tailed distributions but not from mutli-modal distributions

E. Thus, none of A-D are true

(7.) **E.** None of the above statements are true because $\bar{Y}$ is an unbiased estimator of μ for all sample sizes.

(8.) **D.**

(9.) **D.**

(10.) **B.** $n = \frac{Z^2_{.025}\hat{p}(1-\hat{p})}{(.05)^2} \leq \frac{(1.96)^2.\hat{2}(1-.2)}{(.05)^2} = 245.9 \Rightarrow$ use n = 246.

Part II. (60 points)

1. The square root transformation provides an excellent fit to the data because:

   a. The plotted points $(X_{(i)},\ Q_Z(u_i)), i = 1, \ldots, 60$ are very close to a straight line

   b. The SW test yields W=.974 which has an associated .10 <p-value < 0.50 for the X = $\sqrt{Y}$ data values

(2.) The $P = .90,\ \gamma = .95$ tolerance interval would be obtained by first noting that the $X = \sqrt{Y}$ has a normal distribution and computing a (P=.9, γ= .95) tolerance interval for the distribution of X:

$\bar{X} \pm K_{.90,.95}S_X = 15.31 \pm (1.96)(8.80) = 15.31 \pm 17.248 = (-1.938,\ 32.558) = (0,\ 32.558)$, X is a nonnegative random variable.

Next, invert the endpoints of the tolerance interval on the distribution of X to obtain the tolerance interval for the distribution of lifetimes, $Y = X^2$:

$(0,\ 32.558^2) = (0,\ 1060.023)$ which implies $(0,\ 1060023)$ is a (.9, .95) tolerance interval for the distribution of the lifetimes.

A less efficient distribution-free Tolerance Interval considering n=60 would be obtained by using $m = 2$ from the table yielding $(Y_{(1)},\ Y_{(60)}) = (.1,\ 1129.2)$. Thus, $(100,\ 1129200)$ would be a less efficient tolerance interval for the distribution of lifetimes.

(3.) A 95% C.I. on the proportion of cords receiving the prescribed stress that would have a lifetime greater than 750,000 hours is given by

Let Y be the number of cords out of the 60 having lifelength greater than 750,000. From the data B=8. Because, $n = 60 > 40$ and $\min(n\hat{p},\ n(1-\hat{p}) = 8 > 5$, the Agresti-Coull C.I. would be appropriate.

$\tilde{Y} = B + .5(1.96)^2 = 9.9208 \quad \tilde{n} = n + (1.96)^2 = 63.8416 \ \Rightarrow\ \tilde{p} = 9.9208/63.8416 = .1554$

The A-C 95% C.I. on p is $.1554 \pm 1.96\sqrt{.1554(1-.1554)/63.8416} = .1554 \pm .0889 = (.067,\ .244)$

(4.) Find n such that we are 99% confident that $\bar{Y}$ is within 5000 hours of μ_Y.

From the data, an estimate of σ_Y would be $S = 291.0$. Also, tentatively using the Central Limit Theorem to approximate the sampling distribution of $\frac{|\bar{Y}-\mu|}{\sigma_Y/\sqrt{n}}$ we obtain the following:

$$.99 = P\left[|\bar{Y}-\mu| < 5000/1000\right] = P\left[\frac{|\bar{Y}-\mu|}{\sigma_Y/\sqrt{n}} < \frac{5}{\sigma_Y/\sqrt{n}}\right] \approx P\left[|Z| < \frac{5}{\sigma_Y/\sqrt{n}}\right] \text{ also, } P\left[|Z| < 2.576\right] = .99 \ \Rightarrow$$

$$\frac{5}{\sigma_Y/\sqrt{n}} = 2.576 \ \Rightarrow\ n \approx \frac{(2.576)^2(\hat{\sigma}_Y)^2}{(5)^2} = \frac{(2.576)^2(291.0)^2}{(5)^2} = 22476.97 \ \Rightarrow\ \text{ would need a sample size of } n = 22477$$

Exam 2 Scores for STAT 641

Min = 52, Q(.25) = 77, Q(.5) = 83, Mean = 81.9 , Q(.75) = 90, Max = 100