STATISTICS 641 - EXAM II

Student's Name _____

Student's Email Address _____

INSTRUCTIONS FOR STUDENTS:

(1) The exam consists of 4 pages of questions, cover page, and 20 pages of Tables.

(2) You have exactly **60 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 on both sides of the four sheets**)
- The attached tables.

(6) Do not use any other written material except for your summary sheets and the attachmented tables.

(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 60 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 ONE of (A, B, C, D, or E) corresponding to the **BEST** answer

(1.) An entomologist is investigating Y,the amount of water loss due to evaporation from Lone Star ticks. In order to conduct a simulation study of these ticks, the entomologist wants to determine if the distribution of Y, F, is equal to the cdf, F_o . The entomologist measures the water loss from 237 random selected ticks. Based on these values, the most appropriate procedure for accessing whether F is F_o is

A. is the Anderson-Darling statistic.
B. is the Kolmogorov-Smirnov statistic.
C. is the Chi-square Goodness-of-Fit statistic.
D. is the Shapiro-Wilks statistic.
E. depends on the shape of F_o.

(2.) A cardiologist is studying the distribution of pulse rates in junior tennis players. He is concerned that this group of teenagers may have enlarged hearts which can be reflected in inflated pulse rates. Using a large random sample of such athletes, he wants to establish a set of values of the resting pulse rates for these athletes such that the set of values would have a very high degree of certainty of containing 90% of the resting pulse rates for all such athletes.

A. The computed interval would be a **confidence interval**.
B. The computed interval would be a **tolerance interval**.
C. The computed interval would be a **prediction interval**.
D. The computed interval would be a **natural interval**.
E. None of the above would be appropriate.

(3.) An unbiased estimator of θ with a large variance may be preferred to a biased estimator of θ with a small variance because

A. its mean square error would be smaller than the mean square error of the biased estimator
B. its values would be closer to θ than would the values of the biased estimator
C. its sampling distribution would be scattered about θ whereas the biased estimator would be scattered about a value other than θ
D. its sampling distribution would be known whereas the sampling distribution of the biased estimator is unknown
E. unbiased estimators have a normal distribution whereas biased estimators have skewed distributions.

(4.) Let σ be the standard deviation of a highly skewed to the right population distribution. Suppose a researcher selects a random sample of 100 units from the population and constructs a 95% confidence interval on σ using the upper and lower .025 percentiles from the chi-squared distribution. The researcher's statistician points out that a C.I. on σ has the requirement that the population distribution has to be normally distributed. The coverage probability of the constructed 95% confidence interval on σ will be

A. close to 95% because the normal based confidence interval is robust to deviations from normality
B. close to 95% because the sample size is greater than 30
C. much greater than 95%
D. much less than 95%
E. None of the above statements are true.

(5.) A metallurgist has 3 potential cdf's F_1, F_2, F_3 which she is considering using in a simulation model for evaluating the tensile strength of an new alloy. She measured the tensile strength of 45 specimens of the alloy producing $Y_1, \ldots, Y_{45}$ which can be considered as iid observations on a continuous cdf F. The most appropriate procedure to use in selecting which of the 3 cdfs would provide the best representation of the process cdf F is

A. Maximum Likelihood Estimator (MLE)
B. Minimum Mean Square Error (MSE)
C. Shapiro-Wilk (SW) statistic
D. Anderson-Darling (AD) statistic
E. None of the above would be valid

(6.) A medical sociologist takes a random sample of lower income residents in Texas to estimate the percentage of income (PI) spent on medical insurance. The distribution of PI in Texas is is highly right skewed. She wants a 95% confidence interval on the median value of PI. You apply a Box-Cox transformation and obtain X = g(PI), where $g(t) = 1/\sqrt{t}$. A p-value = .456 is obtained from the Shapiro-Wilk statistic for the transformation data. Which of the following methods would you recommend for constructing the confidence interval?

A. A studentized Bootstrap procedure

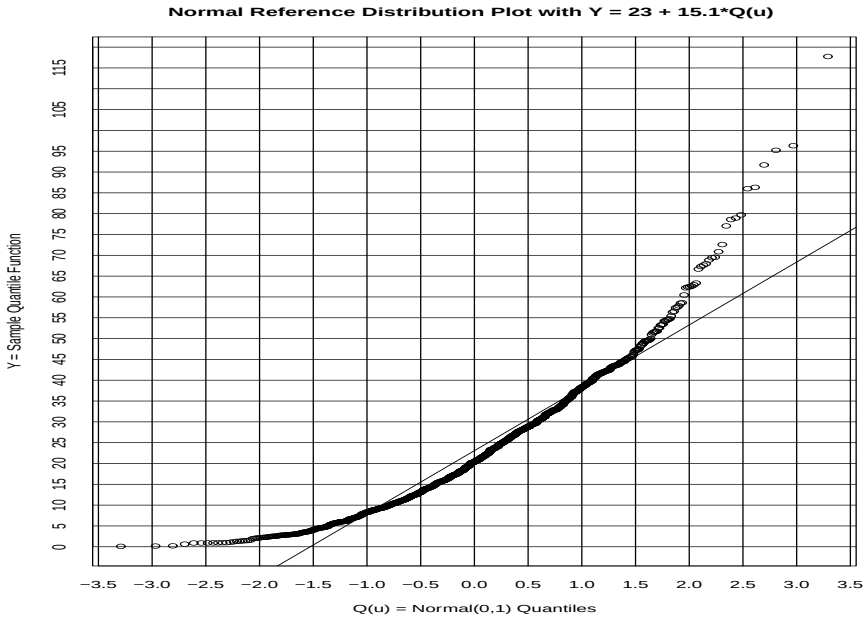
B. Use the inverse of the normal-based tolerance interval: $\left(\left[\bar{X} + t_{(.025)} S_x/\sqrt{n} \right]^{-2}, \left[\bar{X} - t_{(.025)} S_x/\sqrt{n} \right]^{-2} \right)$

C. Use a nonparametric confidence interval

D. Estimate the population pdf using a kernel density estimator and then obtain MLEs for the parameters in the kernel density estimator

E. None of the above procedures would work very well

(7.) A random sample of 1000 data values are selected from a process having cdf, $F(\cdot)$. The plot of the sample quantile versus a standard normal quantile is given below.



Use the above plot, to estimate the 50th percentile and 98th percentiles for the process from which the 1000 data values were selected.

A. (23, 54)

B. (20, 64)

C. (20, 37)

D. (0, 2.05)

E. Cannot be determined from this graph because data is too far from line

(8.) Select the distribution which best describes the 1000 data values depicted in Problem 7:

A. $F(\cdot)$ has a N(mean = 20, variance = $(15)^2$) distribution

B. $F(\cdot)$ has a Gamma(shape =.8, scale = 25) distribution

C. $F(\cdot)$ has a Weibull(shape = 1.5, scale = 25) distribution

D. $F(\cdot)$ has an Exponential(scale=29) distribution

E. Cannot be determined from this graph because data is too far from line

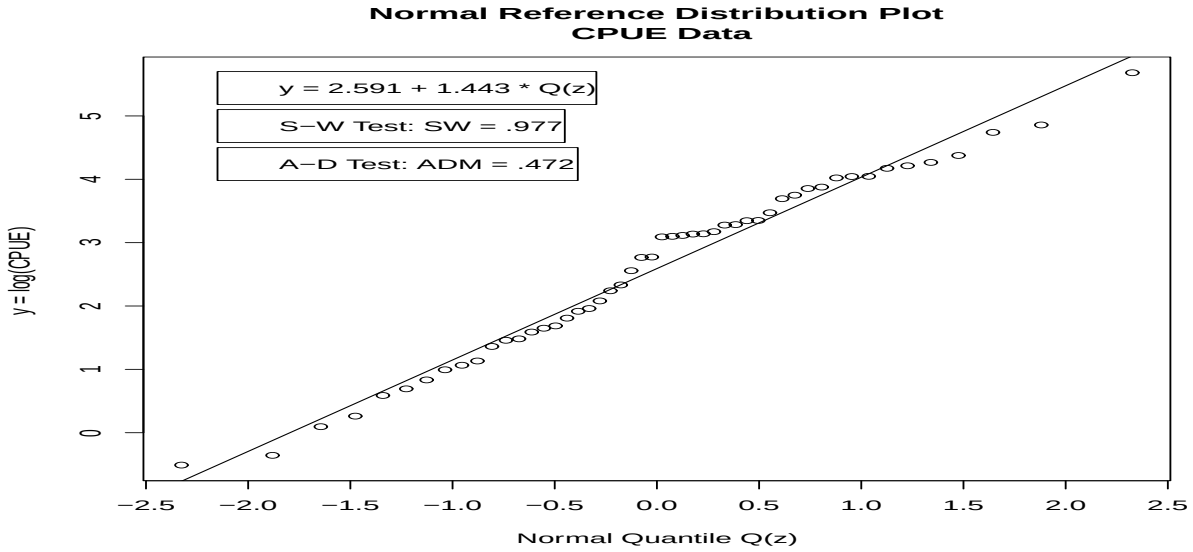
Question II. (60 points) Hurricane Harvey had a major impact on the Gulf of Mexico resulting in a reduction of the amount of fish caught by commercial fisherman. The Gulf of Mexico Fishery Commission designed a study to evaluate the size of the reduction in catch. A random sample of the catch from 50 commercial fisherman yield the following data concerning the catch per unit effort (CPUE) of Finfish. The data for $X_i = CPUE_i$, $i = 1, 2, \ldots, 50$ is given here. From previous data on CPUE in the Gulf of Mexico, the log-normal distribution provided an excellent fit.

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	

The following summary statistics were computed from $X_i = CPUE_i$ and $Y_i = log(X_i)$, $i = 1, \ldots, 50$:

	Mean	St.Dev.	Min.	Q(.25)	Q(.5)	Q(.75)	Max.
CPUE	31.720	47.701	0.600	4.525	19.050	41.920	293.100
log(CPUE)	2.5910	1.453	-0.5108	1.5090	2.9340	3.7360	5.6810

A normal reference plot for $Y = log(X)$ with values from the Shapiro-Wilk statistic and the adjusted Anderson-Darling statistic are given below:



(1.) (15 points) Does a LogNormal Distribution appear to provide an adequate fit to the current CPUE data? Justify your answer.

(2.) (15 points) Place a 95% confidence interval on the proportion of catches which would have a CPUE less than 55.

(3.) (15 points) Provide a 99% confidence interval on the median CPUE of Finfish based on the 50 catches.

(4.) (10 points) Provide the researchers with an interval of values (D_L, D_U) such that the researchers would be 90% confident that the interval would contain the CPUE for least 95% of the catches for all commercial fishermen in the Gulf of Mexico.

(5.) (5 points) The researchers are interested in conducting a much larger study. Determine the sample size n for the new study such that the researchers can be 95% confident that the estimated median CPUE, $\hat{\tilde{\mu}}$, from the new study is at most a 10% overestimation of the median CPUE for all Gulf of Mexico commercial fishermen, $\tilde{\mu}$.

STATISTICS 641 - EXAM II - SOLUTIONS

I. (40 points) CIRCLE ONE of the following letters **(A, B, C, D, or E)** corresponding to the best answer

(1.) **E.** Because the optimal statistic depends on the form of F_o, for example, if F_o is discrete then Chi-square GOF, if F_o is normal then S-W GOF, etc.

(2.) **B.** The cardiologist wants to have a high level of certainty that the interval contains 90% of the population hence he would require a Tolerance interval.

(3.) **C.** If $\hat{\theta}$ is a biased estimator θ with a small variance then the sampling distribution of $\hat{\theta}$ would be highly concentrated about $E[\hat{\theta}]$ which is not equal to θ, the parameter being estimated.

(4.) **D.** much less than 95% because the sampling distribution of $(n-1)S^2/\sigma^2$ would be more highly right skewed than when the population distribution has a normal distribution. This results in the true lower .025 percentile being less than the lower .025 percentile from a chi-squared distribution and the true upper .025 percentile being greater than the upper .025 percentile from a chi-squared distribution. Thus, the C.I. constructed using the percentiles from the chi-squared distribution would be too narrow to have 95% coverage.

(5.) **D.** F is a continuours cdf, use the A-D GOF statistic to select the best fitting cdf

(6.) **B :** A normal based C.I. on the mean of the transformed data and then invert back to original scale would be the best approach. Recall, the mean and median are equal for a normal distribution.

(7.) **B.** Because $Z_{.5} = 0$ on the horizontal axis corresponds to approximately 20 on the vertical axis

$Z_{.98} = 2.05$ on the horizontal axis corresponds to approximately 64 on the vertical axis.

(8.) **C.** The distribution is highly right skewed with $\hat{Q}(.5) \approx 20$ and $\hat{Q}(.98) \approx 64$.

Weibull, Exponential, and Gamma are all possible for $F(\cdot)$

Weibull(shape =1.5,scale = 25): $Q(.5) = 25 * (-log(.5))^{1/1.5} = 19.6,\ \ Q(.98) = 25 * (-log(1-.98))^{1/1.5} = 62.1$

Exponential(scale=29): $Q(.5) = 29 * (-log(.5)) = 20.1$ and $Q(.98) = 29 * (-log(1-.98)) = 113.5$

Gamma(shape =.8,scale = 25): mean is $\mu = \alpha\beta = 20 >> Q(.5)$ because distribution is right skewed.

Question II. (60 points)

(1.) (15 points) Does a LogNormal Distribution appear to provide an adequate fit to the data? Justify your answer.

- Because we are testing the normality of the data therefore use S-W test: p-value = .50 using Table A29. The closeness of the points to the line in the normal reference plot and the very large value for the p-value indicate that the normal distribution would provide an excellent fit to the distribution of log(CPUE).

(2.) (15 points) Place a 95% confidence interval on the proportion of catches which would have a CPUE less than 55, that is, on $p = P[X < 55]$

- Let B=number of $X_i < 55$. From the data, n=50, $\hat{p} = B/n = 40/50 = 0.8$.
- $n = 50 > 40$ and $n \cdot min(\hat{p}, 1-\hat{p}) = 10 > 5$, therefore we can use the 95% Agresti-Coull C.I.:

$\tilde{B} = B + .5(1.96)^2 = 41.9208;\ \tilde{n} = n + (1.96)^2 = 53.8416\ \Rightarrow$

$\tilde{p} = \tilde{B}/\tilde{n} = 41.9208/53.8416 = .7786\ \Rightarrow\ C.I. = .7786 \pm 1.96\sqrt{(.7786)(1-.7786)/53.8416}$

$\Rightarrow\ C.I. = .7786 \pm .1109 = (.6677,\ .88953)$

(3.) (15 points) Provide a 99% confidence interval on the median CPUE of Finfish based on the information from the 50 catches.

- $Y = log(X) \Rightarrow\ Q_Y(.5) = log(Q_X(.5))\ \Rightarrow$

  $.99 = P[L_Y \leq Q_Y(.5) \leq U_Y] = P[L_Y \leq log(Q_X(.5)) \leq U_Y] = P[e^{L_Y} \leq Q_X(.5) \leq e^{U_Y}]$

  $Y = log(CPUE)$ has approximately a normal distribution, therefore, $\mu_Y = Q_Y(.5)$,

  A 99% C.I. on $Q_Y(.5)$ is equivalent to a 99% C.I. on μ_Y

  which is given by $\bar{Y} \pm t_{.005,49} S_Y/\sqrt{n}$

  $L_Y = 2.591 - (2.68)(1.453)/\sqrt{50} = 2.0403; \quad U_Y = 2.591 + (2.68)(1.453)/\sqrt{50} = 3.1417$

  A 99% C.I. on $Q_X(.5)$ is $\left(e^{2.0403},\ e^{3.1417}\right) = (7.69,\ 23.14)$

- A less optimal answer would be to use the interval: $(X_{(k)},\ X_{(n-k+1)})$ where k = 16 from Table VII.3

  $(X_{(16)},\ X_{(35)}) = (5.4,\ 28.5)$

(4.) (10 points) Provide the researchers with an interval of values (D_L, D_U) such that the researchers would be 90% confident that the interval would contain the CPUE for least 95% of the catches for all commercial fishermen in the Gulf of Mexico.

- $Y = log(CPUE)$ has approximately a $N(\mu_Y, \sigma_Y)$ distribution, therefore a $(P = .95, \gamma = .90)$ Tolerance interval for the distribution of Y is

  $\bar{Y} \pm K_{P,\gamma} S_Y$, where $K_{.95,\ .90} = 2.285$ for n=50 from the Tolerance Interval Table. Thus, we have

  $2.591 \pm (2.285)(1.453) = (-0.7291,\ 5.9111)$

  Therefore, a $P = .95,\ \gamma = .90$ Tolerance interval for the distribution of X is

  $(e^{-0.7291},\ e^{5.9111}) = (0.48,\ 369.11)$

  - A less optimal answer would be to use the interval:

    $(X_{(r)},\ X_{(n-s+1)})$ with $r + s = m = 1$ from Tolerance Interval Table with $P = .95$, $\gamma = .90$, and $n = 50$

    $(X_{(1)},\ X_{(51)}) = (0.6,\ \infty)$ or $(X_{(0)},\ X_{(50)}) = (0,\ 293.5)$

(5.) (5 points) The value of n such that $P[\hat{\tilde{\mu}} \leq 1.1\tilde{\mu}] = .95$ is determined as follows:

If $Y = log(X)$ then $\tilde{\mu}_Y = log(\tilde{\mu}_X)$, where $\tilde{\mu}$ is the median.

- $.95 = P[\hat{\tilde{\mu}} \leq 1.1\tilde{\mu}] = P[log(\hat{\tilde{\mu}}) \leq log(1.1) + log(\tilde{\mu})] = P[\hat{\tilde{\mu}}_Y \leq log(1.1) + \tilde{\mu}_Y]$

  Because Y is log-normal, $\tilde{\mu}_Y = \mu_Y\ \Rightarrow\ \hat{\tilde{\mu}}_Y = \hat{\mu}_Y = \bar{Y} \Rightarrow$

  $.95 = P[\bar{Y} \leq log(1.1) + \mu_Y] = P[\frac{\bar{Y} - \mu_Y}{\sigma_Y/\sqrt{n}} \leq \frac{log(1.1)}{\sigma_Y/\sqrt{n}}] = P[Z \leq \frac{log(1.1)}{\sigma_Y/\sqrt{n}}]\ \Rightarrow$

  $\frac{log(1.1)}{\sigma_Y/\sqrt{n}} = Z_{.95} = 1.645\ \Rightarrow\ n = \frac{\sigma_Y^2 (1.645)^2}{(log(1.1))^2} \approx \frac{(1.453)^2 (1.645)^2}{(log(1.1))^2} = 628.9\ \Rightarrow\ n = 629$

EXAM II SCORES:

Min = 12, Q(.25) = 65.25, Q(.5) = 71.5, Mean = 72.64, Q(.75) = 82, Max = 100