STATISTICS 641 - ASSIGNMENT 1

DUE DATE: Noon (CDT), WEDNESDAY, September 8, 2021

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 #1 - Due Noon (CDT) WEDNESDAY, September 8, 2021

- Read Handouts 1 and 2
- Problems to Submit for Grading:

1. (8 points) What are two major problems in the initial analysis of the O-ring failures data described in Handout 1?
2. (8 points) What is one of the most frequent misinterpretations of statistical findings?
3. (12 points) In each of the following studies, (i) State whether the study is experimental or observational; (ii) State whether the study is comparative or description; (iii) If the study is comparative, identify the response variable and the explanatory variable.
   - **Study 1:** A study monitors the occurrence of heart disease over a 5-year period in men randomized to each high fiber or low fiber diets.
   - **Study 2:** An industrial pump manufacturer monitors warranty claims and surveys customers to assess the failure distribution of its pumps.
   - **Study 3:** A biologist randomly selects fish in a river to determine the proportion of fish which show signs of health problems due to pollutants that were poured in the river upstream by a chemical plant.
   - **Study 4:** A study from hospital records found that women who had low weight gain during their pregnancy were more likely to have low birth weight babies than women who had high weight gains during their pregnancy. The researchers also recorded the age and ethnicity of the women.
4. (12 points) Brazos county plans to survey 1000 out of the 60,000 registered voters in the county regarding their preference on the county paying a portion of the building of a new football stadium for Texas A&M University. A complete alphabetical list of the registered voters is available for selecting the 1000 participants. In each of the following scenarios, identify by name the type of sampling method being used.

   a. Out of the first 60 names on the list, one name is randomly selected. That person and every 60th person on the list after that person are then included in the survey.

   b. Each voter is randomly assigned a number between 1 and 60,000 with no repeats. The voters' names are then ordered from smallest to largest based on their assigned number. The first 1000 voters on the list are selected for the survey.

   c. The list of 60,000 voters is divided by into 10 separate lists by voting districts within the county. The 60,000 voters are randomly assigned a number between 1 and 60,000 with no repeats. The names on each of the 10 lists are then ordered from smallest to largest by the number assigned the voter. The first 100 names on each of the ten ordered lists are selected to be in the survey.

   d. The list of 60,000 voters is first divided into four lists consisting where the voter lived, the East, West, North, or South regions of the county. There are 120 voting precincts in the county with 30 precincts in each region. A random sample of 10 precincts is taken within each region and then a simple random sample of 25 votes is taken within each of the randomly selected precincts. The resulting 1000 voters are then interviewed in a personal survey.

5. (12 points) For each of the following studies,

   i. state whether the study is a survey, a prospective study, or a retrospective study;

   ii. state whether the study is comparative or descriptive;

   iii. if the study is comparative, identify the response and explanatory variable(s).

   a. A sample of members listed in the directory from a professional organization is used to estimate the proportion of female members.

   b. To assess the effect of smoking during pregnancy on premature delivery, mothers of preterm infants are matched by age and number of previous pregnancies to mothers of full term infants and then both are asked about their smoking habits during pregnancy.

   c. A sociologist interviews juvenile offenders to find out what proportion live in foster care.

   d. A marketing study uses the registration card mailed back to the company following the purchase of a video game to gauge the percentage of purchasers who learned about the purchased game from advertising on facebook, television, or by word of mouth.

6. (16 points) Consider the experiment discussed in class concerning the determination of DMZ in a product. Suppose the experiment was repeated and the following data was obtained:

			Chemical Analysis				
Operator	Specimen	Run	1	2	Run Mean	Specimen Mean	Operator Mean
1	1	1	136.75	138.75	137.75	138.00	142.75
		2	137.75	139.75	138.75		
		3	136.25	138.75	137.50		
	2	4	146.75	148.75	147.75	147.50	
		5	149.25	145.25	147.25		
		6	146.75	148.25	147.50		
2	3	7	147.25	149.25	148.25	148.50	143.50
		8	150.75	148.75	149.75		
		9	146.25	148.75	147.50		
	4	10	136.65	138.85	137.75	138.50	
		11	137.55	139.55	138.55		
		12	140.45	137.95	139.20		
3	5	13	166.75	168.25	167.50	163.50	143.25
		14	157.25	161.25	159.25		
		15	162.25	165.25	163.75		
	6	16	122.75	124.75	123.75	123.00	
		17	124.25	127.25	125.75		
		18	118.75	120.25	119.50		

   a. Modify the R code - DMZplot.R in Canvas: R Files, to produce a diagram similar to the diagram given in Handout 1.

   b. Which of the following sources of variation in the 36 observations do you think has the largest source of variation and which source has the smallest?

      - Operator - O
      - Specimen within Operator - S(O)
      - Run within Specimen and Operator - R(S,O)
      - Analysis within Run, Specimen, and Operator - A(R,S,O)

- **In Questions 7-10 on the next page, select the BEST answer.**

7. (8 points) The NRC commissioned a study to evaluate the impact of nuclear power plants on the temperature of the water down stream from the discharge from the plant. The researcher was concerned about the possible effects of climate on the study so she divide the USA into 5 regions and randomly selected 10 power plants from each region. During a five month period of time, the water temperature was measured daily at each of the 50 power plants at a location above and below the point of discharge into the stream. A total of 150 measurements are taken at each of these two locations. This type of study is an example of

   A. a simple random sample.
   B. a simple random cluster sample.
   C. a stratified cluster random sample.
   D. a stratified simple random sample.
   E. a multistage cluster random sample.

8. (8 points) An advocacy group for improved health care in the US wants to estimate the cost of treating elderly patients. To obtain information about doctors across the US, the group randomly selected 100 counties from a list of the 1000 largest counties in the US. In each county, the county public health administrator is contacted and asked to randomly select 20 doctors in their county. Each selected doctor then randomly selected 25 of their patients over the age of 60 and determined the annual medical cost for each patient. This study is an example of

   A. a simple random sample.
   B. a simple random cluster sample.
   C. a stratified simple random sample.
   D. a multistage cluster random sample.
   E. a stratified cluster random sample.

9. (8 points) A study was designed to evaluate the effects of pine bark beetles on native pine trees in East Texas. The researcher divided East Texas into 35 regions. Within each of these regions, she randomly selected 15 native pine trees, each of the 15 trees was examined, and an identifer was placed on those limbs where a pine bark beetle was located. Six months later she returned and determined the amount of damage to each of the limbs where an identifer had been placed. This type of study/sampling method is an example of

   A. a stratified simple random sample.
   B. a stratified cluster random sample.
   C. a multistage cluster random sample.
   D. a factorial random experiment.
   E. an observational prospective study.

10. (8 points) FEMA wanted an assessment of the amount of damage caused by hurricane Ike to individual homes on the coast of Texas. A random sample of 50 homes was taken from each of the twenty-five coastal counties of Texas. An evaluation of the amount of damage to each of the 1250 homes was made. These 1250 measurements were then summarized into an overall average amount of damage per home. This type of study is an example of

    A. a simple random sample.
    B. a simple cluster sample.
    C. a stratified simple random sample.
    D. a stratified cluster random sample.
    E. a multistage cluster random sample.

STAT 641 Fall 2021

Solutions for Assignment # 1

- Problem 1 - (8 Points) Two major problems are
  - Ignoring the launches in which there were no O-ring failures and
  - Extrapolating the data from previous launches in which the temperature was above 50 degrees to a launch in which the temperature would be in the low 30's.

- Problem 2 - (8 Points) One of the most frequent misinterpretations of statistical findings is attributing a "causal" relationship between two events when only a strong correlation exists between the events.

- Problem 3 - (12 Points)

  Study 1: (i) Experimental. (ii) Comparative (iii) Response: Occurence of heart disease.
  Explanatory: Amount of fiber in diet.

  Study 2: (i) Observational. (ii) Descriptive.
  The manufacturer is recording why pumps fail.

  Study 3: (i) Observational. (ii) Descriptive.
  The biologist is acquiring health data on the collected fish

  Study 4: (i) Observational. (ii)Comparative.(iii) Response: Baby's birthweight.
  Explanatory: Mother's weight gain during pregnancy.

- Problem 4 - (12 Points)

  (a.) Systematic sampling.

  (b.) Simple random sampling.

  (c.) Stratified random sampling with the strata being the voting districts.

  (d.) Stratified Multi-Stage Cluster sampling. The voting precincts are stratified into one of four regions, and then 10 voting precincts are randomly selected from each of the four regions. The selected precincts consist of clusters of voters and a random sample of 25 voters is selected from each of the selected precincts.

- Problem 5 - (12 Points)

  (a) (i) Survey. (ii) Descriptive

  (b) (i) Retrospective Study. (ii) Comparative. (iii) Response: Term of pregnancy. Explanatory: Mother's smoking habits.

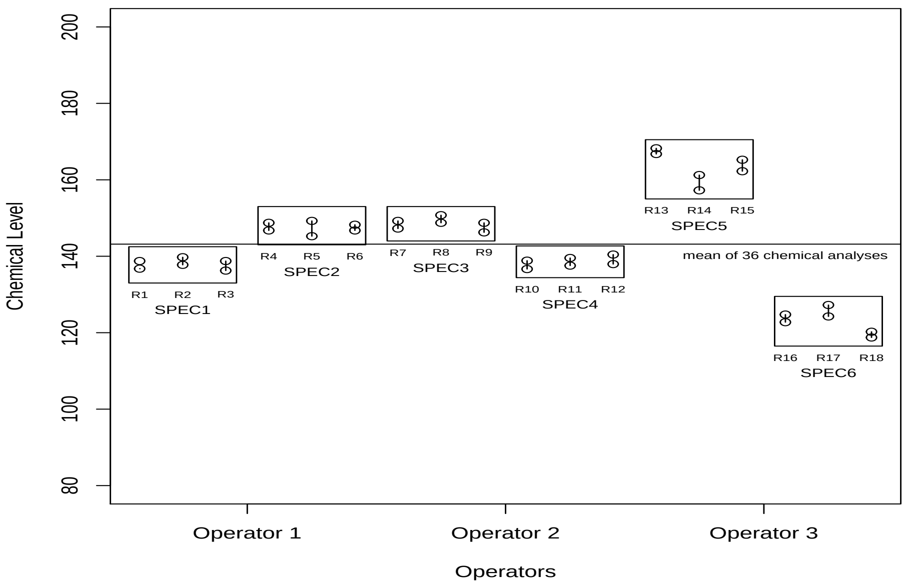
  (c) (i) Survey. (ii) Descriptive

  (d) (i) Survey. (ii) Descriptive

- Problem 6 - (16 Points)

  (a.) See the following plot. The code is at the end of this document.

Figure 3: Chemical Variation Plot



  (b.) S(O) means show the greatest amount of variability in comparison to the other three sources and the O means have the least amount of variability. In fact, 97% of the overall variability in the 36 response is attributable to S(O); 2.2% of the variability is attributable to R(S,O); 0.5% of the variability is attributable to A(R,S,O); 0.0% of the variability is attributable to O, and .3% of the variability is attributable to all other sources. We will demonstrate how to obtain these percentages in STAT 642.

7. - (8 points) **C**. The strata are the 5 Regions with a SRS of 10 Power Plants selected in each Region. The Power Plants are clusters with the daily measurements being the units within the clusters.

8. - (8 Points) **D**. The First Stage Clusters are Counties which are clusters of doctors. The Second Stage Clusters are Doctors with patients being the units within the clusters.

9. - (8 Points) **B**. The strata are the 35 Regions with a SRS of 15 Pine Trees selected in each Region. The Pine Trees are clusters with the limbs having pine bark beetles being the units within the clusters.

10. - (8 Points) **C**. The strata are the 25 Costal Counties with a SRS of 50 homes selected in each County.

```
run = c(1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,11,11,12,12,13,13,14,14,
15,15,16,16,17,17,18,18)
Res = c(
 136.75, 138.75,
 137.75, 139.75,
 136.25, 138.75,
 146.75, 148.75,
 149.25, 145.25,
 146.75, 148.25,
 147.25, 149.25,
 150.75, 148.75,
 146.25, 148.75,
 136.65, 138.85,
 137.55, 139.55,
 140.45, 137.95,
 166.75, 168.25,
 157.25, 161.25,
 162.25, 165.25,
 122.75, 124.75,
 124.25, 127.25,
 118.75, 120.25)
spec = seq(1,6)


plot(run,Res,type="p",xlab="Operators",ylab="Chemical Level",
        main="Figure 3: Chemical Variation Plot ",cex=.99,
        ylim=c(80,200),xaxt="n")
rect(0.75,133,3.25,142.5)
segments(1,136.75,1,136.75)
segments(2,137.75,2,139.75)
segments(3,136.25,3,138.75)
text(1,130,"R1",cex=.55)
text(2,130,"R2",cex=.55)
text(3,130,"R3",cex=.55)
text(2,126,"SPEC1",cex=.75)

rect(3.75,143,6.25,153)
segments(4,146.75,4,148.75)
segments(5,149.25,5,145.25)
segments(6,146.75,6,148.25)
text(4,140,"R4",cex=.55)
text(5,140,"R5",cex=.55)
text(6,140,"R6",cex=.55)
text(5,136,"SPEC2",cex=.75)

rect(6.75,144,9.25,153)
segments(7,147.24,7,149.25)
segments(8,150.75,8,148.75)
segments(9,146.25,9,148.75)
text(7,141,"R7",cex=.55)
text(8,141,"R8",cex=.55)
text(9,141,"R9",cex=.55)
text(8,137,"SPEC3",cex=.75)

rect(9.75,134.4,12.25,142.7)
segments(10,136.65,10,138.85)
segments(11,137.55,11,139.55)
segments(12,140.45,12,137.95)
text(10,131.4,"R10",cex=.55)
text(11,131.4,"R11",cex=.55)
text(12,131.4,"R12",cex=.55)
text(11,127.4,"SPEC4",cex=.75)

rect(12.75,155,15.25,170.5)
segments(13,166.75,13,168.25)
segments(14,157.25,14,161.25)
segments(15,162.25,15,165.25)
text(13,152,"R13",cex=.55)
text(14,152,"R14",cex=.55)
text(15,152,"R15",cex=.55)
text(14,148,"SPEC5",cex=.75)

rect(15.75,116.5,18.25,129.5)
segments(16,122.75,16,124.75)
segments(17,124.25,17,127.25)
segments(18,118.75,18,120.25)
text(16,113.5,"R16",cex=.55)
text(17,113.5,"R17",cex=.55)
text(18,113.5,"R18",cex=.55)
text(17,109.5,"SPEC6",cex=.75)

axis(side=1,at=c(3.5,9.5,15.5),
labels=c("Operator 1","Operator 2","Operator 3"))
abline(143.1667, 0)
text(16,140,"mean of 36 chemical analyses", cex = 0.7)
```

STATISTICS 641 - ASSIGNMENT 2

DUE DATE: Noon (CDT), MONDAY, SEPTEMBER 20, 2021

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 #2 - Due Noon, Monday - 9/20/21

- Read Handout 3
- Supplemental reading: Chapter 2, 3, 4 in the Devore's book
- Hand in the following Problems:

(1.) (10 points) Assume that the random variable Y has pmf with parameter $p,\ 0 < p < 1$:

$$f(y) = \begin{cases} p(1-p)^y & \text{for } \ y = 0, 1, 2, 3, \ldots \\ 0 & \text{otherwise} \end{cases}$$

(a.) Find the cdf, $F(y)$ for Y
Hint: $\sum_{k=0}^{m} ab^k = a\frac{1-b^{m+1}}{1-b}$

(b.) Find the 80th percentile of $F(y)$ if $p = .4$. That is, evaluate $Q(.8)$ for $p = .4$

(2.) (20 points) Let Y have a 3-parameter Weibull distribution, that is, Y has pdf and cdf in the following form with $\alpha > 0,\ \gamma > 0,\ \theta > 0$:

$$f(y) = \begin{cases} \frac{\gamma}{\alpha^\gamma}(y-\theta)^{\gamma-1}e^{-\left(\frac{y-\theta}{\alpha}\right)^\gamma} & \text{for} \quad y \geq \theta \\ 0 & \text{for} \quad y < \theta \end{cases} \qquad F(y) = \begin{cases} 1 - e^{-\left(\frac{y-\theta}{\alpha}\right)^\gamma} & \text{for} \quad y \geq \theta \\ 0 & \text{for} \quad y < \theta \end{cases}$$

(a.) Verify that the pair $(\theta,\ \alpha)$ are location-scale parameters for this family of distributions.

(b.) Derive the quantile function for the three parameter Weibull family of distributions.

(c.) What is the probability that a random selected value from a Weibull distribution with $\theta = 5,\ \gamma = 4$ and $\alpha = 20$ has value greater than 23?

(d.) Compute the 25th percentile from a Weibull distribution with with $\theta = 5,\ \gamma = 4$ and $\alpha = 20$.

(3.) (10 points) An alternative form of the 2-parameter Weibull distribution is given as follows with parameters $\beta > 0,\ \gamma > 0$

$$f(y) = \begin{cases} \frac{\gamma}{\beta}y^{\gamma-1}e^{-y^\gamma/\beta} & \text{for} \quad y \geq 0 \\ 0 & \text{for} \quad y < 0 \end{cases} \qquad F(y) = \begin{cases} 1 - e^{-y^\gamma/\beta} & \text{for} \quad y \geq 0 \\ 0 & \text{for} \quad y < 0 \end{cases}$$

(a.) Show that β is not a scale parameter for this family of distributions.

(b.) Show that $\alpha = \beta^{1/\gamma}$ is a scale parameter for this family of distributions.

(4.) (10 points) An experiment measures the number of particle emissions from a radioactive substance. The number of emissions has a Poisson distribution with rate $\lambda = .15$ particles per week.

(a.) What is the probability of at least 2 emission occurring in a randomly selected week?

(b.) What is the probability of at least 2 emission occurring in a randomly selected year?

(5.) (10 points) Let $Z_1,\ Z_2,\ Z_3,\ Z_4,\ Z_5,\ Z_6,\ Z_7,\ Z_8$ be independent N(0,1) r.v.'s. Identify the distributions of the random variables, A, B, C, D by providing the name of the distribution and the appropriate degrees of freedom, if needed.

(a.) $A = Z_7/\sqrt{[Z_1^2 + Z_2^2 + Z_3^2]/3}$.

(b.) $B = Z_5/Z_6$.

(c.) $C = Z_1^2 + Z_2^2 + Z_3^2$

(d.) $D = 3(Z_4^2 + Z_5^2 + Z_6^2 + Z_7^2)/4[Z_1^2 + Z_2^2 + Z_3^2]$

(e.) $E = 3Z_1^2/[Z_2^2 + Z_3^2 + Z_4^2)]$.

(6.) (10 points) Let $U = .26$ be a realization from a Uniform on (0,1) distribution.

Express a single realization from each of the following distributions using just the fact $U = .26$.

(a.) W = Weibull(γ=4, α=1.5)

(b.) N = NegBin(r=8, p=.7)

(c.) B = Bin(20, .4)

(d.) P = Poisson(λ=3)

(e.) U = Uniform on (0.3, 2.5)

(7.) (30 points) For each of the following situations described below, select the distribution which best models the given situation (you may need to use some distributions multiple times). Provide a very short justification for your answer.

Hypergeometric
Uniform
Chi-square
Equally Likely
Gamma
Lognormal
Poisson
Exponential
Cauchy
Binomial
Chi-square
Weibull
Geometric
Lognormal
F
Negative Binomial
Gamma
t
Normal
Exponential
Beta

(a.) In an epidemiological study of the incidence of skin cancer for those individuals who use artificial tanning procedures, a researcher randomly selected 100 of the 10,000 customers of a large tanning facility. The 100 customers were examined and the number N who had developed skin cancer was recorded. The distribution of N is is __?

(b.) A civil engineer has determined that micro-cracks in bridge support columns occur according to a Poisson process with an average rate of 30 micro-cracks in a 10 feet length of column. She wants to assess the possible distance, D, between micro-cracks in a column. The distribution of D is is __?

(c.) In an epidemiological study of the incidence of chronic fatigue within those individuals who are fully employed and pursing a college degree, a researcher decided to interview 500 of the 100,000 students enrolled in a distance learning program at a well known university. The number S of students who had been diagnosed with chronic fatigue was recorded and used in the analysis. The distribution of S is ____?

(d.) A metallurgist designed a study to estimate the distribution of cracks in the cooling pipes at nuclear power plants. She randomly selected 200 sections of pipes within the cooling systems at various nuclear power plants. An x-ray is taken of each pipe and the number of cracks, C, of length greater than 20mm is recorded. A possible probability model for C, the number of cracks of length greater than 20mm in a randomly selected pipe, is ____?

(e.) The automobile industry has thousands of small suppliers of parts for its assembly process. In a randomly selected delivery of parts to the assembly plant, let p be the proportion of parts in the shipment that are not within specification. From previous studies, it is known that around 80% of the suppliers produce parts with values of p in the 0 to 0.03 range but the remaining suppliers have values of p between 0.03 and 0.25. A possible probability model for p is _____?

(f.) For each day during a six month period in Stamford, Connecticut, the maximum daily ozone reading R was recorded. The distribution of R is _____?

(g.) A mechanical engineer for a natural gas distributor is investigating the occurrence of leaks in gas pipelines. From 30 years of data she finds that the number of major cracks in any 100 feet of pipe appears to be independent of the number of major cracks in any adjacent 100 feet of pipe. From the 30 year data set, she determines that the average number of major cracks per 100 feet of pipe is relatively constant. Let **C** be the number of major cracks in a randomly selected 100 feet section of pipe. The distribution of the C is _____?

(h.) A biostatistician wishes to investigate the distribution of defective genes occurring in the kidney of mice exposed to a toxin. She uses 100 mice in a lab study and determines for each mouse the number of defective genes in their kidney. A possible probability model for G, the number of defective genes in the kidney of a mouse exposed to the toxin, is _____?

(i.) A geologist is studying the frequency of occurrence of minor earthquakes in Texas. From hundreds of years of data she finds that the number of earthquakes in any 24 hour period is independent of the number of earthquakes in any other 24 period of time. Over the past 100 years, the daily earthquake rate has been relatively constant. Let **Q** be the number of days in which there were no earthquakes in a randomly selected year. The distribution of the Q is _____?

(j.) In an epidemiological study of the incidence of skin cancer for those individuals who use artificial tanning procedures, a researcher needs at least 100 subjects in the study for it to have validity. The researcher interviews subjects to determine if they fit the criteria to be included in the study. Let S be the number of subjects interviewed until 100 subjects are determined to be acceptable for the study. The distribution of S is _____?

(k.) A quality control engineer wants to document major problems in a process which produces ball bearings. He measures the difference D between the nominal diameter of a 5 cm ball bearing and the true bearing diameter. He finds that the bearings are equally likely to have a diameter larger than or smaller than 5 cm. Furthermore, approximately 20% of the bearings have diameters which deviate more than 3 standard deviations from 5 cm. The distribution of D is _____?

(l.) The wings on an airplane are subject to stresses which cause cracks in the surface of the wing. After 1000 hours of flight the wing is inspected with an x-ray machine and the number of cracks N are recorded. The distribution of N is _____?

(m.) Suppose small aircraft arrive at a certain airport according to a Poisson process with rate 8 aircraft per hour. For the next 100 days, the length of time, T, until the 15th aircraft arrives each day is recorded. The distribution of T is _____?

(n.) An entomologist is studying the number of ticks on cattle in a feed lot. From 20 years of tick inspection data, she finds that the number of ticks on a randomly selected cow appears to be independent of the number of ticks on any other randomly selected cow in the feed lot. The density of ticks per cow appears to be fairly constant over the past 20 years. Let **T** be the number of cows in a random selection of 100 cows in the feed lot having fewer than 5 ticks on their torso. The distribution of T is _____?

(o.) An administrator is studying the quality of high-school curriculums in Michigan. She randomly selects 50 high schools out of the 357 high schools in Michigan for the study. A careful examination of their curriculum is performed. Let X be the number of high schools in which the curriculum was found to be unsatisfactory. The distribution of X is ____?

Stat 641 Fall 2021

Solutions for Assignment 2

(1.) (10 points) In the following expressions let m be a non-negative integer. Using the expression $\sum_{k=0}^{m} ab^k = a\frac{1-b^{m+1}}{1-b}$, we have with a = p, b = 1- p, [y] = greatest integer $\leq$ y

(a.) For $y < 0$, F(y) = 0; for $y \geq 0$,

$$F(y) = P[Y \leq y] = \sum_{k=0}^{[y]} p(1-p)^k = 1-(1-p)^{[y]+1} = \begin{cases} 0 & \text{if } y < 0 \\ p & \text{for } 0 \leq y < 1 \\ 1-(1-p)^2 & \text{for } 1 \leq y < 2 \\ 1-(1-p)^3 & \text{for } 2 \leq y < 3 \\ 1-(1-p)^4 & \text{for } 3 \leq y < 4 \\ \vdots & \end{cases}$$

(b.)

$$\text{Using the definition of } Q(u),\ Q(u) = inf(y : F(y) \geq u) \ \Rightarrow$$

$$Q(u) = \text{ smallest nonnegative integer } y_u \text{ such that } 1-(1-p)^{y_u+1} \geq u \text{ with } Q(0) = 0$$

$$Q(u) = \text{ smallest nonnegative integer } y_u \text{ such that } y_u \geq \frac{log(1-u)}{log(1-p)} - 1 \text{ with } Q(0) = 0$$

That is,

$$Q(u) = \begin{cases} 0 & \text{if } u \leq 0 \\ 0 & \text{for } 0 < u \leq p \\ 1 & \text{for } p < u \leq 1-(1-p)^2 \\ 2 & \text{for } 1-(1-p)^2 < u \leq 1-(1-p)^3 \\ 3 & \text{for } 1-(1-p)^3 < u \leq 1-(1-p)^4 \\ \vdots & \end{cases}$$

Using the above expression with p = .4, Q(.8) = 3

(2.) (20 Points) (a.) Let $W = \frac{Y-\theta}{\alpha}$ then the pdf of W is

$$f_W(w) = \alpha f(\theta+\alpha w) = \alpha\frac{\gamma}{\alpha^\gamma}((\theta+\alpha w)-\theta)^{\gamma-1}e^{-\left(\frac{(\theta+\alpha w)-\theta}{\alpha}\right)^\gamma} \quad \text{for } \theta+\alpha w \geq \theta \ \Rightarrow$$

$$f_W(w) = \begin{cases} \gamma w^{\gamma-1}e^{-w^\gamma} & \text{for } w \geq 0 \\ 0 & \text{for } w < 0 \end{cases}$$

Because the expression for $f_W(w)$ does not contain $(\theta,\ \alpha)$, we can conclude that $(\theta,\ \alpha)$ are location-scale parameters for the family of distributions.

(b.) To find the quantile function, set

$$u = F(y_u) = 1 - e^{-\left(\frac{y_u-\theta}{\alpha}\right)^\gamma}$$

and solve for y_u. In this case,

$$y_u = \theta + \alpha(-log(1-u))^{1/\gamma} \quad \Rightarrow \quad Q(u) = \theta + \alpha(-log(1-u))^{1/\gamma}$$

(c.) With $\theta = 5,\ \gamma = 4,\ \alpha = 20,\quad P(Y > 23) = 1 - P(Y \leq 23) = 1 - F(23) = e^{-\left(\frac{23-5}{20}\right)^4} = .5189$

(d.) With $\theta = 5,\ \gamma = 4,\ \alpha = 20,\quad Q(.25) = 5 + 20(-log(1-.25))^{1/4} = 19.647$

(3.) (10 points) (a.) Let $W = Y/\beta$. The pdf of W is

$$f_W(w) = \beta f(\beta w) = \beta\frac{\gamma}{\beta}(\beta w)^{\gamma-1}e^{-(\beta w)^\gamma/\beta} = \gamma\beta^{\gamma-1}w^{\gamma-1}e^{-\beta^{\gamma-1}w^\gamma} \quad \text{for } \ w > 0$$

Because the expression for the pdf of W contains β, β cannot be a scale parameter for the given family of distributions.

(b.) With $W = Y/\alpha$, the pdf of W is given by

$$f_W(w) = \alpha f(\alpha w) = \alpha\frac{\gamma}{\alpha}(\alpha w/\alpha)^{\gamma-1}e^{-(\alpha w/\alpha)^\gamma} = \gamma w^{\gamma-1}e^{-w^\gamma} \quad \text{for } \ w > 0$$

The expression for f_W is free of α, therefore α is a scale parameter.

(4.) (10 points)

(a.) Let X be the number of emissions in a week. X has a Poisson distribution with $\lambda = 0.15$.

$$P(X \geq 2) = 1 - P(X=0) - P(X=1) = 1 - \frac{e^{-0.15}(0.15)^0}{0!} - \frac{e^{-0.15}(0.15)^1}{1!} = 1 - .8607 - .1291 = 0.0102$$

Using the R-function dpois, $P(X \geq 2) = 1 - dpois(0, .15) - dpois(1, .15) = 1 - .8607 - .1291 = 0.0102$

(b) Let Y be the number of emissions in a year. Y has a Poisson distribution with $\lambda = 0.15 \times 52 = 7.8$.

$$P(Y \geq 2) = 1 - P(Y=0) - P(Y=1) = 1 - \frac{e^{-7.8}(7.8)^0}{0!} - \frac{e^{-7.8}(7.8)^1}{1!} = 1 - .0004097 - .003200 = .9964$$

Using the R-function dpois, $P(X \geq 2) = 1 - dpois(0, 7.8) - dpois(1, 7.8) = 1 - .0004097 - .003200 = .9964$ or

Using the R-function ppois, $P(X \geq 2) = 1 - P(X \leq 1) = 1 - ppois(1, 7.8) = 1 - .0036 = .9964$

(5.) (10 points)

(a.) A has a t-distribution with df = 3 (A is the ratio of a N(0,1) r.v. and the square root of a Chi-square r.v. divided by its df. with the numerator and denominator r.v's having independent distributions)

(b.) B has a Cauchy distribution with location = 0 and scale =1 (B is the ratio of two independent N(0,1) r.v.'s)

(c.) C has a chi-squared distribution with df = 3 (C is the sum of independent squared N(0,1) r.v.s)

(d.) D has an F-distribution with $df_1 = 4$, $df_2 = 3$ (An F-distribution is the ratio of two independent Chi-square r.v.'s divided by their df's.)

(e.) E has an F-distribution with $df_1 = 1$, $df_2 = 3$ (An F-distribution is the ratio of two independent Chi-square r.v.'s divided by their df's.)

(6.) (10 points) Let U = .26 be a realization from a Uniform on (0,1) distribution.

(a.) W = Weibull(γ=4,α=1.5): $Q(u) = 1.5[-log(1-u)]^{1/4} \Rightarrow$

$W = Q(.26) = 1.5[-log(1-.26)]^{1/4} = 1.111$

(b.) $N = NegBin(r = 8, p = 0.7)$. Recall that the R functions for Negative Binomial are modeling the number of failures. Using the R function **pnbinom(x,8,.7)** with

x=c(0,1,2,3), we obtain the cdf, F(x), for X equal to the number failures before the 8th success:

$$F(x) = \begin{cases} 0.05764801 & x = 0 \\ 0.19600323 & x = 1 \\ 0.38278279 & x = 2 \\ 0.56956234 & x = 3 \end{cases}$$

Thus, with U=.26, we obtain $X = 2$ because $F(1) = .196 < .26 < .383 = F(2)$.

Therefore, N, the number of trials before the 8th success, $N = X + 8 = 2 + 8 = 10$.

(c.) B = Bin(20,.4): Using the R function **pbinom(x,20,.4)** with

x=c(5,6,7), we obtain

$$F(x) = \begin{cases} .1256 & x = 5 \\ .2500 & x = 6 \\ .4159 & x = 7 \end{cases}$$

Thus, with U=.26, we obtain $B = 7$ because $F(6) = .25 < .26 < .4159 = F(7)$

(d.) P = Poisson(λ=3): Using the R function **ppois(x,3)** with x=c(0,1,2), we obtain

$$F(x) = \begin{cases} .04978707 & x = 0 \\ .19914827 & x = 1 \\ .42319008 & x = 2 \end{cases}$$

Thus, with U=.26, we obtain $P = 2$ because $F(1) = .1991 < .26 < .4232 = F(2)$.

(e.) Y = Uniform on (0.3,2.5). Then the pdf is $f(y) = 1/(2.5-.3)$ for $.3 < y < 2.5$; 0 otherwise. Therefore, the cdf is given by

$$F(y) = 0 \text{ for } y \leq .3;\ F(y) = 1 \text{ for } y \geq 2.5;\ \text{For } .3 < y < 2.5,\ F(y) = \int_{.3}^{y} \frac{1}{2.5-.3} dy = \frac{1}{2.5-.3}(y-.3)$$

Let $u = F(y_u) = \frac{1}{2.5-.3}(y_u - .3)$, then solve for y_u yields $Q(u) = y_u = .3 + (2.5-.3)u$.

Therefore, with $U = .26$, $Y = .3 + (2.5-.3)(.26) = 0.872$

(7.) (30 points)

(a.) Hypergeometric - Sampling from finite population in which there are two types of units

(b.) Exponential - Distance between events in a Poisson process

(c.) Binomial (assuming 100,000 is very large) or Hypergeometric - Sampling from finite population in which there are two types of units

(d.) Poisson - counting the number of cracks larger than 20mm in a randomly selected pipe

(e.) Beta- Values of p are within (0,1) and have a skewed distribution

(f.) Weibull - Modeling extremes, maximum daily ozone level

(g.) Poisson - number of events occurring in a fixed length of pipe

(h.) Poisson - number of events occurring in a large number of genes

(i.) Binomial - number of occurrences of independent events in a fixed number of trials

(j.) Negative Binomial - number of trials until success, 100 subjects are accepted into the study

(k.) Cauchy - symmetric distribution with a large number of extreme values

(l.) Poisson - recording number of events in space - cracks on wing

(m.) Gamma - Time until 15th event in a Poisson process

(n.) Binomial - Counting number of trials (cows) in which a success occurs (5 or fewer ticks)

(o.) Hypergeometric - Sampling from finite population in which there are two types of units

STATISTICS 641 - ASSIGNMENT 3

DUE DATE: NOON (CDT), MONDAY, SEPTEMBER 27, 2021

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 #3 - Due NOON (CDT) Monday - 9/27/2021

- Read: Handouts 4 & 5
- Supplemental Reading in Devore book: Chapters 1 & 4

- Submit for grading the following problems:

P1. (10 Points) Let Y have a double exponential distribution, that is, Y has pdf and cdf in the following form with parameters $\theta,\ \beta > 0$:

$$f(y) = \frac{1}{2\beta} e^{-\left(\left|\frac{y-\theta}{\beta}\right|\right)} \text{ for } \quad -\infty \le y \le \infty \qquad F(y) = \begin{cases} \frac{1}{2} e^{-\left(\frac{\theta-y}{\beta}\right)} & \text{for} \quad y < \theta \\ & \\ 1 - \frac{1}{2} e^{-\left(\frac{y-\theta}{\beta}\right)} & \text{for} \quad y \ge \theta \end{cases}$$

(a.) Derive the quantile function for Y

(b.) Derive the survival function for Y

(c.) Derive the hazard function for Y

P2. (10 Points) A researcher is studying the relative brain weights 1000 times the ratio of brain weight to body weight for 51 species of mammal whose average litter size is less than 2 and for 44 species of mammal whose average litter size is greater than or equal to 2. The researcher was interested in determining what evidence that brain sizes tend to be different for the two groups. (Data from *The Statistical Sleuth* by Fred Ramsey and Daniel Schafer). The data is in the Homework Assignment Folder: Assign3-BrainSize

BRAINSIZE - SMALL LITTER SIZE

0.42	0.86	0.88	1.11	1.34	1.38	1.42	1.47	1.63
1.73	2.17	2.42	2.48	2.74	2.74	2.79	2.90	3.12
3.18	3.27	3.30	3.61	3.63	4.13	4.40	5.00	5.20
5.59	7.04	7.15	7.25	7.75	8.00	8.84	9.30	9.68
10.32	10.41	10.48	11.29	12.30	12.53	12.69	14.14	14.15
14.27	14.56	15.84	18.55	19.73	20.00			

BRAINSIZE - LARGE LITTER SIZE

0.94	1.26	1.44	1.49	1.63	1.80	2.00	2.00	2.56
2.58	3.24	3.39	3.53	3.77	4.36	4.41	4.60	4.67
5.39	6.25	7.02	7.89	7.97	8.00	8.28	8.83	8.91
8.96	9.92	11.36	12.15	14.40	16.00	18.61	18.75	19.05
21.00	21.41	23.27	24.71	25.00	28.75	30.23	35.45	

A software package uses the estimator $\widehat{Q}(u) = Y_{((n-1)u+1)}$ as the estimator of Q(u).

- Calculate the estimates of the Quartiles: of $Q(.25),\ Q(.5),\ Q(.75)$ for just the **Large Litter Size** using the software package's formula.

P3. (24 points) Using the data from Problem 2 for just the **Large Litter Size**, we want to estimate the pdf $f(y)$ for the relative brain weights of the 44 species of mammal.

The kernel density estimate of $f(y)$ is given by

$$\widehat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y - Y_i}{h}\right),$$

Suppose we use the Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$ and a bandwidth of $h = 3$.

(a.) Estimate $f(3)$ and $f(16)$ using the kernel density estimator.

(b.) Using a relative frequency histogram with bin width of 5, estimate the values of $f(3)$ and $f(16)$.

(c.) Which data value provides the smallest contribution to the kernel density estimator at y=16, $\widehat{f}(16)$?

(d.) Which data value provides the largest contribution to the kernel density estimator at y=16, $\widehat{f}(16)$?

P4. (28 points) Using the relative Brain Weight data, answer the following questions:

(a.) Produce the following plots of the data: estimates of the pdf, cdf, and quantile function for both Small and Large litter sizes.

(b.) Describe the underlying distribution of the relative brain weights for both Small and Large litter sizes.

(c.) Based on the graphs, what are your conclusions about the relationship between litter size and relative brain weights?

P5. (28 Points) **Select** the letter of the **best** answer for each question. No explanation is needed for your selection.

1. The function which provides the most detailed description for the realizations of a random variable is

   A. the probability density (mass) function, pdf $f(\cdot)$

   B. the quantile function, $Q(\cdot)$

   C. the survival function, $S(\cdot)$

   D. the cumulative distribution function, cdf $F(\cdot)$

   E. all the above functions are equivalent

2. A relative frequency histogram having classes of greatly different class widths was used as an estimator of a continuous population pdf. The relative frequency was plotted versus the class intervals. This plot will not be an appropriate estimator of the population pdf because

   A. all the intervals are not the same width.

   B. the relative frequency varies greatly by class width.

   C. the area under the curve is not proportional to one.

   D. the area under the curve for each class is not an estimator of the probability of that class.

   E. In fact it is an unbiased estimator of the pdf.

3. A relative frequency histogram having classes of greatly different class widths was used as an estimator of a continuous population pdf. The relative frequency was plotted versus the class intervals. The plot will result in a graphical distortion. The plot can be corrected by

   A. making all the intervals have the same width.

   B. plotting the relative frequency divided by class width.

   C. making sure that the area under the curve adds to one

   D. increasing the sample size.

   E. In fact there will not be a distortion since it is an unbiased estimator of the pdf.

4. A kernel density estimator was used as an estimator of a continuous population pdf, $f(y)$. The kernel density estimator is generally a vastly improved estimator over a density histogram (plot of
   $\frac{N_i/n}{h_i}$ vs Class i) because

   A. in using the histogram, it is necessary to select the number of bins, bin widths, and their location.

   B. there are too many spurious modes using the histogram

   C. the area under the curve adds to 1 for the kernel density estimator.

   D. the kernel density estimator makes use of all the data in estimating $f(y)$ whereas the histogram only uses those data values in the same bin as y.

   E. all of the above

5. In using a kernel density estimator to estimate a population pdf based on a random sample $Y_1, \cdots, Y_n$, the design factor which is **least** crucial in determining the effectiveness of the estimator is

   A. the sample size, n

   B. the number of plotting points, m provided $m > 50$

   C. the bandwidth, h

   D. the kernel $k(\cdot)$

   E. all four factors are equally crucial

6. A kernel density estimator is an estimator of a population pdf, $f(y)$. The bandwidth of the kernel density estimator is selected by

   A. using the uniform distribution to randomly select a value between 0 and 1.

   B. taking the value which minimizes the asymptotic integrated mean square error.

   C. taking the value which produces a curve having area closest to 1.

   D. taking the value of the bandwidth which yields maximum entropy.

   E. asking Dr. Sheather.

7. A random sample of n data values is obtained from a process having an absolutely continuous cdf of unknown shape. The metallurgist wants to select the best fitting distribution amongst several candidate cdfs. She decides to select the distribution which has mean and variance most closely matching the corresponding sample mean and variance. The major weakness in this approach is

   A. the mean and variance may be highly inflated by outliers

   B. the empirical distribution function contains more information about the tails of the distribution than does the mean and variance

   C. she should have used robust estimators of the location and scale parameters

   D. there are many distribution having the same mean and variance but very different shapes

   E. the moments of a distribution determine the distribution, hence there is no weakness in the approach

8. The skewness and kurtosis parameters are generally thought to represent the following characteristics of the population cdf, respectively,

   A. the center and spread in the distribution
   B. the heaviness of the tails and deviation from normality of the distribution
   C. the deviation from symmetry and concentration in the tails of the distribution
   D. the deviation from symmetry and concentration in the tails and/or the peakedness of the distribution
   E. none of the above

9. The median is a trimmed mean with level of trimming equal to

   A. 0%
   B. 25%
   C. 50%
   D. 75%
   E. none of the above

10. The standard deviation is preferred to MAD as a measure of population dispersion when the population distribution

    A. has absolutely no outliers.
    B. has a skewed but short-tailed distribution.
    C. has a lognormal distribution.
    D. has a normal distribution.
    E. cannot be determined with the given information.

11. The coefficient of kurtosis

    A. measures the dispersion of the distribution about two values $\mu \pm \sigma$.
    B. measures the peakedness of a distribution.
    C. measures the concentration of the mass of a distribution in the tails of the distribution.
    D. measures the difference between a distribution and the normal distribution.
    E. all of the above

12. Alternatives to σ for measuring the dispersion in a distribution are $SIQR$ and MAD. Which of the following statements about these measures are **TRUE**?

    A. All three measures are equal if the pdf for the distribution is symmetric.
    B. $SIQR$ is preferred to MAD if the distribution has very heavy tails
    C. For the normal distribution, $SIQR$ is preferred to MAD
    D. all of the above
    E. none of the above

13. A government study of the average monthly nitrate levels in the Mississippi river, N_t, just prior to its entry into the Gulf of Mexico is modeled as

    $N_t = 22.3 + .6N_{t-1} + e_t$ where $e_t's$ are iid r.v.s, $E[e_t] = 0,\ Var[e_t] = 2.8,\ e_t's$ are independent of $N_t's$

    The mean and variance of N_t are given by

    A. $\mu = 22.3,\ \sigma^2 = 2.8$
    B. $\mu = 55.75,\ \sigma^2 = 4.375$
    C. $\mu = 34.84,\ \sigma^2 = 2.8$
    D. $\mu = 22.3,\ \sigma^2 = 4.375$
    E. The values of μ and σ^2 would change from month to month.

Stat 641 Fall 2021

Solutions for Assignment 3

P1. (10 points) Let Y have a double exponentiall distribution.

(a.) The quantile function

$$Q(u) = \begin{cases} \theta + \beta \; log(2u) & \text{for} \quad u \le .5 \\ \\ \theta - \beta \; log(2(1-u)) & \text{for} \quad u \ge .5 \end{cases}$$

(b.) The survival function is given by

$$S(y) = P(Y > y) = 1 - F(y) \;\Rightarrow \qquad S(y) = \begin{cases} 1 - \frac{1}{2}e^{-\left(\frac{\theta - y}{\beta}\right)} & \text{for} \quad y < \theta \\ \\ \frac{1}{2}e^{-\left(\frac{y-\theta}{\beta}\right)} & \text{for} \quad y \ge \theta \end{cases}$$

(c.) The hazard function is given by

$$h(y) = \frac{f(y)}{S(y)} \;\Rightarrow \qquad h(y) = \begin{cases} \frac{\frac{1}{\beta}e^{\left(\frac{y-\theta}{\beta}\right)}}{2 - e^{\left(\frac{y-\theta}{\beta}\right)}} & \text{for} \quad y \le \theta \\ \\ \frac{1}{\beta} & \text{for} \quad y > \theta \end{cases}$$

P2. (10 points) $n = 44 \;\Rightarrow\; \widehat{Q}(u) = Y_{(43u+1)} \;\Rightarrow$

- $\widehat{Q}(.25) = Y_{(11.75)} = .25Y_{(11)} + .75Y_{(12)} = .25(3.24) + .75(3.39) = 3.35$
- $\widehat{Q}(.5) = Y_{(22.5)} = .5Y_{(22)} + .5Y_{(23)} = .5(7.89) + .5(7.97) = 7.93$
- $\widehat{Q}(.75) = Y_{(33.25)} = .75Y_{(33)} + .25Y_{(34)} = .75(16.00) + .25(18.61) = 16.65$

P3. (24 points) Using the R code:

```
y = c(0.94,    1.26,    1.44,    1.49,    1.63,    1.80,    2.00,    2.00,    2.56,
    2.58,    3.24,    3.39,    3.53,    3.77,    4.36,    4.41,    4.60,    4.67,
    5.39,    6.25,    7.02,    7.89,    7.97,    8.00,    8.28,    8.83,    8.91,
    8.96,    9.92,   11.36,   12.15 ,  14.40,   16.00,   18.61,   18.75,   19.05,
   21.00,   21.41,   23.27,   24.71,   25.00,   28.75,   30.23,   35.45   )
h=3
n=length(y)
deni <- function(x){
  (1/sqrt(2*pi))*exp(-((x-y)/h)^2/2)/(n*h)
}
f3 = sum(sapply(3,deni))
f16 = sum(sapply(16,deni))
f16i = sapply(16,deni)
min = min(f16i)
imin = which(f16i==min)
ymin=y[imin]
max = max(f16i)
imax=which(f16i==max)
ymax=y[imax]
```

(a.) The value for $\widehat{f}(3)$ is f3 = 0.059703 and for $\widehat{f}(16)$ is f16 = 0.01669353

(b.) Using a relative frequency histogram with a bin width of 5, with

$n_j = \#Y_i$'s in $[0.94 + 5(j-1), 0.94 + 5j)$, we have $n_1 = 19,\; n_2 = 10,\; n_3 = 3,\; n_4 = 4,\; n_5 = 5,\; n_6 = 2,\; n_7 = 1$.

Therefore, the estimates are $\widehat{f}(3) = 19/(44 \times 5) = 0.08636364$ and for $\widehat{f}(16) = 4/(44 \times 5) = 0.01818182$. A fairly close agreement between the estimates obtained by the two methods.

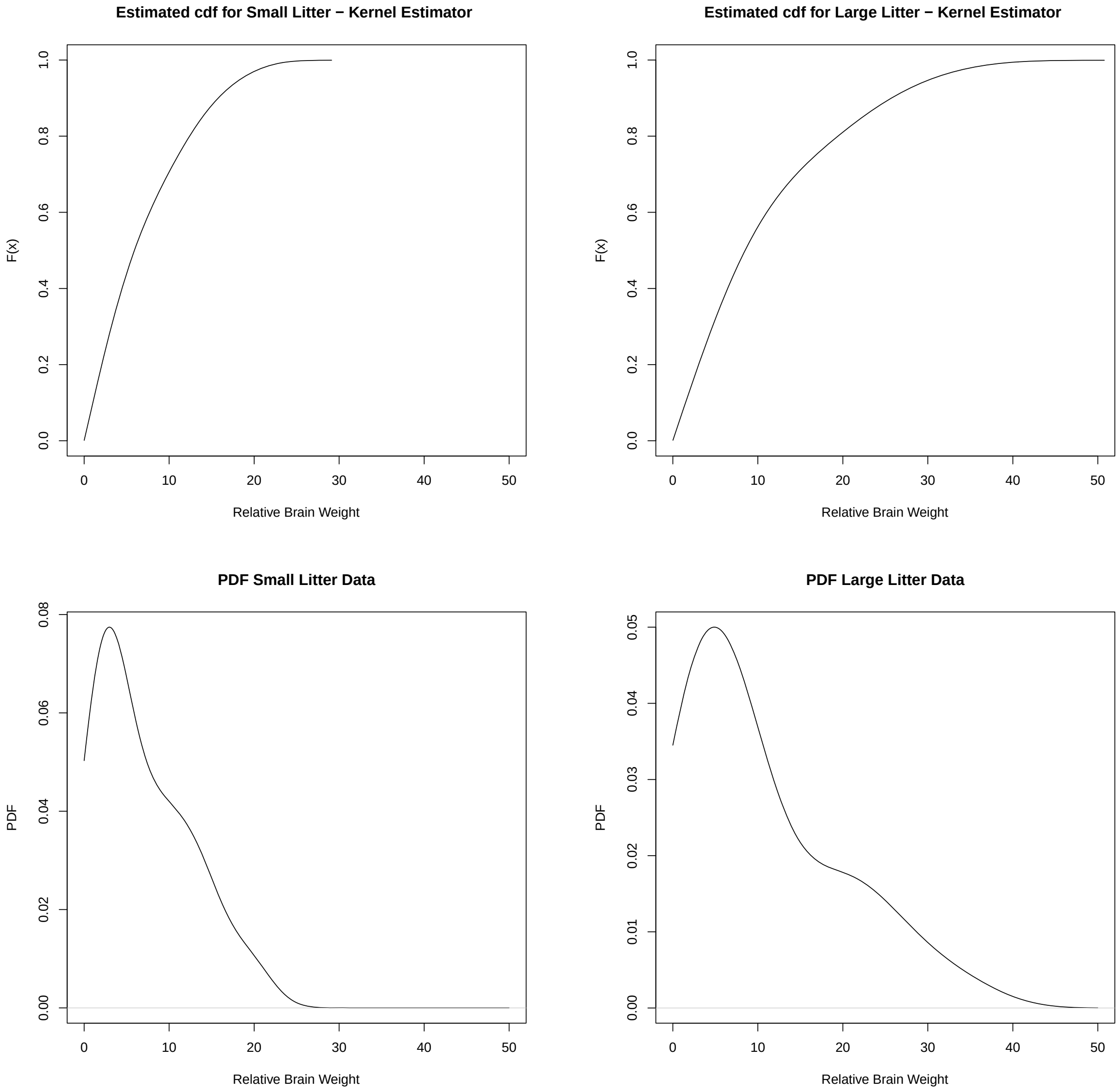
(c.) The data value provides the smallest contribution to the estimator at y=16, $\widehat{f}(16)$ is the data value furthest from 16, which is y = 35.45 with a contribution of 2.253479e-12 to $\widehat{f}(16)$=0.01669353. This is obtained by computing by hand:

$$\frac{1}{nh}K\left(\frac{y-Y_i}{h}\right) = \frac{1}{44*3}K\left(\frac{16-35.45}{3}\right) = \left(\frac{1}{132}\right)\left(\frac{1}{2*\pi}\right)e^{-\left(\frac{16-35.45}{3}\right)^2/2} = 2.253479e^{-12}$$

(d.) The data value provides the largest contribution to the estimator at y=16, $\widehat{f}(16)$ is the data value closest to 16, which is y = 16 with a contribution of 0.00302229 to $\widehat{f}(16)$=0.01669353. This is obtained by computing by hand:

$$\frac{1}{nh}K\left(\frac{y-Y_i}{h}\right) = \frac{1}{44*3}K\left(\frac{16-16}{3}\right) = \left(\frac{1}{132}\right)\left(\frac{1}{2*\pi}\right)e^{-\left(\frac{16-16}{3}\right)^2/2} = .00302229$$

P4. (28 points) (a.) Plots of pdfs (kernel density estimator), edf (smoothed), and quantile (smoothed):

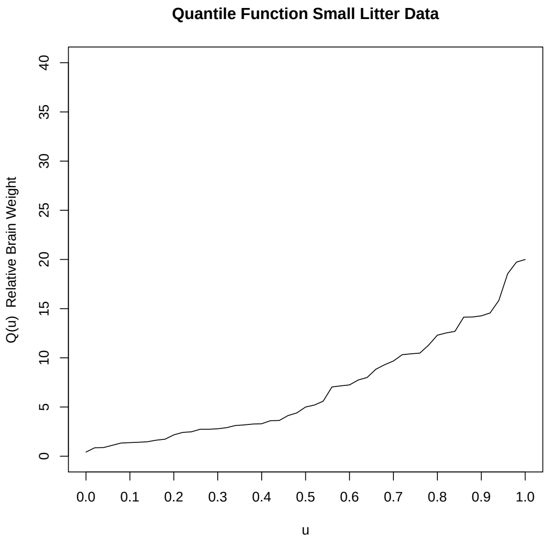


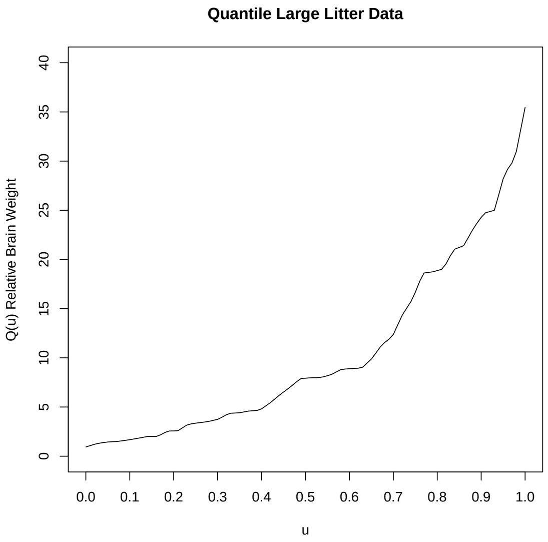
Estimated cdf for Small Litter − Kernel Estimator

Estimated cdf for Large Litter − Kernel Estimator

PDF Small Litter Data

PDF Large Litter Data





(a)

See code at end of document.

(b.) Small Litter: Relative brain weights are somewhat right skewed which indicates that a few species of mammals with small average litters have large brains relative to their body weights.

Large Litter: Relative brain weights are highly right skewed which indicates that sizeable proportion of the species of mammals with large average litters have large brains relative to their body weights.

(c.) Based on the graphs, I would conclude that there is a positive relation ship between average litter size and relative brain weights. However, it would be more informative to have the actual litter sizes associated with each species to draw a more concrete conclusion.

P5. (28 points) Multiple Choice Questions:

1. **E** Given any one of the four functions then you can derive the other three from the given function
2. **D** See page 24 in Handout 4
3. **B** See page 24 in Handout 4
4. **D** See pages 30 & 32 in Handout 4
5. **B** See page 50 in Handout 4
6. **B** See page 37 in Handout 4
7. **D** See pages 16 & 17 in Handout 5
8. **C or D** See page 13 & 14 in Handout 5
9. **C** See page 23 in Handout 5
10. **A or D** See page 27 in Handout 5
11. **E** See page 14 in Handout 5
12. **E** See pages 25-27 in Handout 5

A. is false because σ does not exist for Cauchy which is symmetric whereas both SIQR and MAD exist and are equal

B. is false because MAD is nearly always preferred to SIQR

C. is false because for a normal distribution MAD=SIQR

13. **B** See page 33 in Handout 5:

$\theta = 22.3,\ \rho = .6,\ \sigma_e^2 = 2.8\ \Rightarrow \mu_X = \frac{\theta}{1-\rho} = \frac{22.3}{1-.6} = 55.75$

$\sigma_X^2 = \frac{\sigma_e^2}{1-\rho^2} = \frac{2.8}{1-.36} = 4.375$

```
##
## (2)
##

dta <- read.csv("Assign3_BrainSize.csv")

y <- dta[, 2]
y <- y[!is.na(y)]
n <- length(y)

y_s <- sort(y)

## 0.25: (n - 1) * 0.25 + 1 = 11.75
(n - 1) * 0.25 + 1
y_s[11] + 0.75 * (y_s[12] - y_s[11])

## 0.5: (n - 1) * 0.5 + 1 = 22.5
(n - 1) * 0.5 + 1
y_s[22] + 0.5 * (y_s[23] - y_s[22])

## 0.75: (n - 1) * 0.5 + 1 = 33.25
(n - 1) * 0.75 + 1
y_s[33] + 0.25 * (y_s[34] - y_s[33])

##
## (3)
##

dd <- density(y_s)
plot(dd, type = "l")

h <- 3

## (a)
K_u <- function(u) {
  return(dnorm(u))
}

## f(3)
sum(K_u((3 - y_s) / h)) / (n * h)

## f(16)
sum(K_u((16 - y_s) / h)) / (n * h)

## (b)
brks <- y_s[1] + 5 * (0:7)
n_j <- hist(y_s, prob = FALSE, breaks = brks)$counts
R_j <- n_j / n
f_hat_j <- R_j / 5

hist(y_s, prob = TRUE, breaks = brks)$density

## (c)
kk <- K_u((16 - y) / h) / (n * h)
y[which.min(kk)]

## (d)
```

```
y[which.max(kk)]

##
## (4)
##

y_small <- dta[, 1]
y_large <- dta[, 2]
y_large <- y_large[!is.na(y_large)]

## (a)

## PDFs
par(mfrow = c(1, 2))
plot(density(y_small), xlab = "y", ylab = "f", main = "Small Litters", xlim = c(-10, 50),
  cex.axis = 0.75)
plot(density(y_large), xlab = "y", ylab = "f", main = "Large Litters", xlim = c(-10, 50),
  cex.axis = 0.75)

## EDFs
qq_small <- quantile(y_small, probs <- seq(0, 1, by = 0.01))
qq_large <- quantile(y_large, probs)

plot(qq_small, probs, type = "s", xlab = "y", ylab = "F", main = "Small Litters",
  xlim = c(0, 35), cex.axis = 0.75)
plot(qq_large, probs, type = "s", xlab = "y", ylab = "F", main = "Large Litters",
  xlim = c(0, 35), cex.axis = 0.75)

## Quantile functions
plot(probs, qq_small, type = "s", xlab = "Q(u)", ylab = "u", main = "Small Litters",
  ylim = c(0, 35), cex.axis = 0.75)
plot(probs, qq_large, type = "s", xlab = "Q(u)", ylab = "u", main = "Large Litters",
  ylim = c(0, 35), cex.axis = 0.75)
```

STATISTICS 641 - ASSIGNMENT 4

DUE DATE: NOON (CDT), WEDNESDAY, OCTOBER 6, 2021

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 #4 - NOON (CDT) Wednesday - 10/6/2021

- Read: Handouts 6 and 7
- Supplemental Reading: Chapter 1 & Sections 4.6, 6.1 in Devore book and *Applied Survival Analysis Using R*
- Submit for grading the following problems:

P1. (50 points) A researcher is studying the relative brain weights (brain weight divided by body weight) for 51 species of mammals whose litter size is 1 and for 44 species of mammals whose average litter size is greater than or equal to 2. The researcher was interested in determining what evidence that brain sizes tend to be different for the two groups. (Data from *The Statistical Sleuth* by Fred Ramsey and Daniel Schafer).

RELATIVE BRAIN WEIGHTS - SMALL LITTER SIZE

0.42	0.86	0.88	1.11	1.34	1.38	1.42	1.47	1.63
1.73	2.17	2.42	2.48	2.74	2.74	2.79	2.90	3.12
3.18	3.27	3.30	3.61	3.63	4.13	4.40	5.00	5.20
5.59	7.04	7.15	7.25	7.75	8.00	8.84	9.30	9.68
10.32	10.41	10.48	11.29	12.30	12.53	12.69	14.14	14.15
14.27	14.56	15.84	18.55	19.73	20.00			

RELATIVE BRAIN WEIGHTS - LARGE LITTER SIZE

0.94	1.26	1.44	1.49	1.63	1.80	2.00	2.00	2.56
2.58	3.24	3.39	3.53	3.77	4.36	4.41	4.60	4.67
5.39	6.25	7.02	7.89	7.97	8.00	8.28	8.83	8.91
8.96	9.92	11.36	12.15	14.40	16.00	18.61	18.75	19.05
21.00	21.41	23.27	24.71	25.00	28.75	30.23	35.45	

1. For the Small Litter Size mammals, answer the following questions: The data is given in the file: Brain Weight Data.txt in Canvas
   a. Compute a 10% trimmed mean, and compare it to the untrimmed sample mean. Does this comparison suggest any extreme values in the data?
   b. The researcher suggested a Weibull distribution to model the data for the Small Litter Size mammals. Assuming that the Weibull distribution is an appropriate model for the Small Litter Size data, obtain the MLE estimates of the Weibull parameters for the Small Litter Size data.
   c. Estimate the probability that a randomly selected mammal with a litter size of 1 will have a relative brain weight greater than 15, first using the Weibull model and secondly using a distribution-free estimate.
   d. Compare the MLE estimates of μ and σ based on the Weibull model to the distribution-free estimates of μ and σ for the Small Litter Size data.
   e. Compare the MLE estimates of median and IQR based on the Weibull model to the distribution-free estimates of median and IQR for the Small Litter Size data.
2. Without any assumed model, estimate the mean and standard deviation of the relative brain weights for both Large and Small litter sizes.
3. Estimate the median and MAD of the relative brain weights for both Large and Small litter sizes.
4. Based on your plots from Assignment #3, which pair of estimates of the center and spread in the two data sets best represents the center and spread in the two populations of relative brain weights?
5. Using your answers from the previous three questions, suggest a relationship (if any) between litter size and relative brain weights.

P2. (30 points) Twenty-five patients diagnosed with rare skin disease are randomly assigned to two drug treatments. The following times are either the time in days from the point of randomization to either a complete recovery or censoring (as indicated by the status variable: 0 means censored, i.e., time at which patient left study prior to a complete recovery, 1 means patient's time to recovery).

	Treatment 1												
Time	180	632	2240	195	76	70	13	1990	18	700	210	1296	23
Status	1	1	1	1	1	1	0	0	1	1	1	1	1
	Treatment 2												
Time	8	852	52	220	63	8	1976	1296	1460	63	1328	365	
Status	0	1	1	1	1	1	0	0	1	1	1	1	

1. Estimate the survival function for the two treatments.

2. Compute the mean and median time to recovery for the two treatments using the estimated survival function.

3. Which treatment appears to be most effective in the treatment of the skin disease?

4. Estimate the mean and median time to recovery ignoring the censoring and compare these values to values obtained in part 2.

P3. (20 points) **Select** the letter of the **BEST** answer.

1. An experiment involves putting specimens of steel under stress until the specimen fractures. The machine increases the stress until the specimen fractures. The maximum stress that the machine can place on a specimen is 500 psi. Out of the 35 specimens used in the experiment, 5 did not fracture at 500 psi. This type of censoring is called

   A. Right censoring

   B. Type I censoring

   C. Type II censoring

   D. Random censoring

   E. Left censoring

2. An entomologist is interested in the ability of ticks to conserve water in very dry condition, relative humidity less than 10%. She randomly selects 100 Lone Star ticks for a large collection of Lone Star ticks and places them in a water-free container in which the temperature is maintained at $30^{\circ}C$ with a relative humidity of 10%. The amount of water in the ticks will gradually decline over time. The amount of water retained by the ticks is measured after 90 days. Twelve of the 100 ticks did not survive until the end of the study but their water contents were recorded at the time of their death.

   We would describe the data from this study as being

   A. Right censored

   B. Type I censored

   C. Type II censored

   D. Left censored

   E. Uncensored

3. A chemist employed at a large cosmetic firm designs a study to assess the toxicity of a new skin conditioner. He simultaneously feeds 100 mice a large volume of the conditioner. The time to death is recorded for the mice. The study was terminated after 30 days at which time twelve of the mice were still alive. The data from this type of study is best described as having

   A. Right censoring

   B. Type I censoring

   C. Type II censoring

   D. Random censoring

   E. Left censoring

4. The product engineer for an automobile safety testing agency is evaluating the likelihood of a fire in the batteries of electric automobiles. She randomly selects 100 electric vehicles for testing and the cars were driven by the employees of the agency. The study was terminated when the 20th vehicle had a fire in its battery. The engineer recorded the number of miles each vehicle was driven until either the vehicle's battery caught on fire or until the study was terminated. What type of censoring took place, if any?

   A. Right censoring

   B. Type I censoring

   C. Type II censoring

   D. Random censoring

   E. There is no censoring because a mileage was recorded for each vehicle.

5. An engineer for an automotive manufacturer is studying the occurrence of a defective in the braking system for a newly designed braking system. She randomly selects 100 automobiles for study and plans to record the distance traveled prior to a failure in the braking system. However, she needs to conclude the study 12 months after its inception. For each of the 100 automobiles she recorded the mileage at which a failure occurred in the braking system or the mileage driven during the 12 month study for those automobiles that did not have a failure. We would describe the data from this type of study as having

   A. Right censoring

   B. Type I censoring

   C. Type II censoring

   D. Random censoring

   E. Left censoring

Bonus Problems for 10 points (attempt problems only if you have extra time).

A. Bonus Problem 1 (5 points) Let a random variable Y have a continuous strictly increasing cdf F with pdf f.

Let $\mu_{(\alpha)}$ be the $\alpha-$trimmed mean of Y, that is,

$$\mu_{(\alpha)} = \frac{1}{1-2\alpha}\int_{Q(\alpha)}^{Q(1-\alpha)} yf(y)dy.$$

Prove that

$$\lim_{\alpha\to .5} \mu_{(\alpha)} = \tilde{\mu}, \qquad \text{where} \quad \tilde{\mu} = Q(.5), \quad \text{the median of the distribution of } \ Y$$

Hint 1: Use l'Hôpital's rule in your proof and

the fact that $\frac{d}{dx}\int_{g(x)}^{h(x)} tf(t)dt = h(x)f(h(x))h^{'}(x) - g(x)f(g(x))g^{'}(x)$

Hint 2: Also use the fact that $F(Q(u)) = u \Rightarrow \frac{d}{du}F(Q(u)) = \frac{d}{du}u = f(Q(u))Q'(u) = 1$

B. Bonus Problem 2 (5 points)

Let $T_1, T_2, \ldots, T_{20}$ be the miles to failure of 20 wheel bearings in 20 newly manufacturered trucks. The values were obtained in an accelerated life testing study and are as follows in units of 1000 miles of use:

37.52	30.33	42.82	31.14	31.49	38.04	38.31	36.15	34.08	30.76
45.14	44.81	33.30	45.73	30.00	33.97	33.65	30.60	43.07	31.26

The 20 values are modeled as a shifted exponential distribution with p.d.f. given as follows:

$$g(t;\beta,\theta) = \beta e^{-\beta(t-\theta)} \quad \text{for} \quad t \geq \theta$$

i. Write the likelihood function of β and θ

ii. Determine the MLE of β and θ assuming both parameters are unknown

iii. Estimate the probability that the miles to failure for a randomly selected truck is greater than 40,000 miles. Note that $G(t;\beta,\theta) = 1 - e^{\beta(t-\theta)}$.

Stat 641 Fall 2021

Solutions for Assignment 4

P1. (50 points)

1. See R code at the end of this document. For the Small Litter Size we obtain:

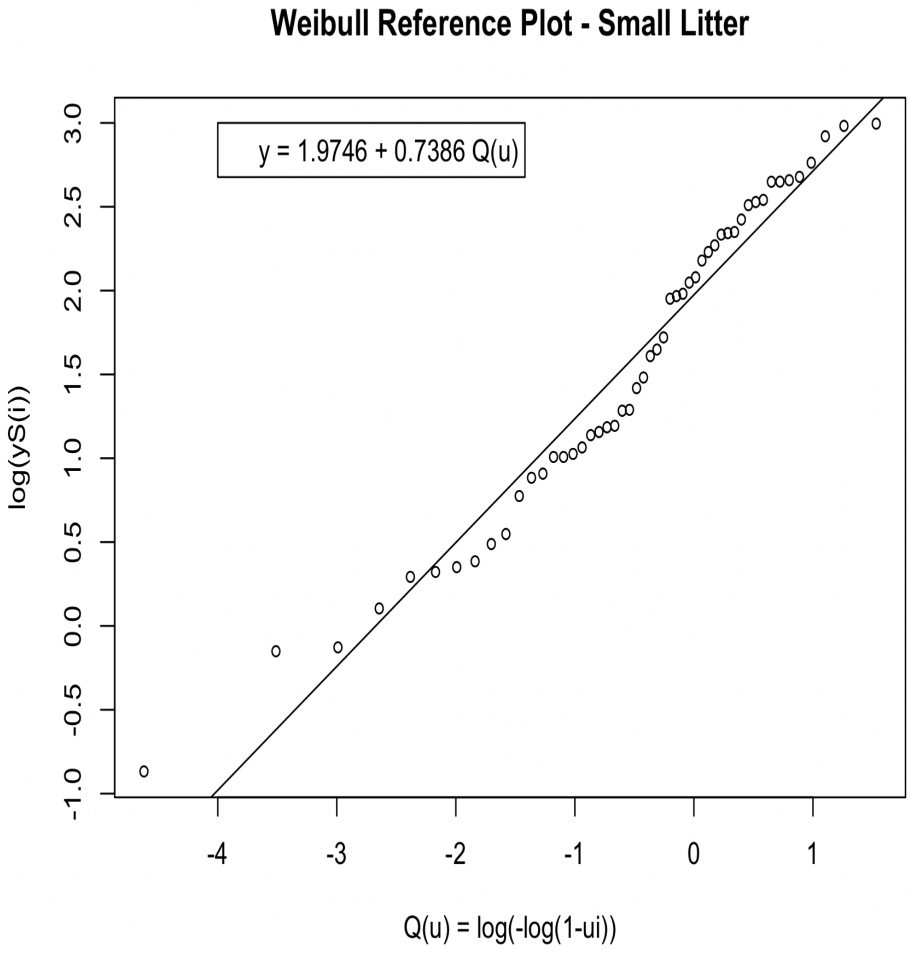
   a. A 10% trimmed mean would involve averaging the middle

   $K(.10) = 51 - [(51)(.1)] - [(51)(.1) + 1] + 1 = 51 - [5.1] - [6.1] + 1 = 41$ values in the data set yielding:

   $\hat{\mu}_{(.1)} = \frac{1}{41}\sum_{i=6}^{46} Y_{(i)} = \frac{1}{41}(257.89) = 6.29 = mean(yS, trim = .1)$ whereas the untrimmed mean is $\hat{\mu} = \frac{1}{51}\sum_{i=1}^{51} Y_{(i)} = \frac{1}{51}(351.18) = 6.89$

   The untrimmed mean is somewhat larger than the 10% trimmed mean which would indicate that there a few large outliers in the data. In fact, examining the sorted data, we have three relatively large data values in the data set: 18.55, 19.73, 20.00

   b. A Weibull reference distribution plot is displayed:



The plot indicates a reasonably good fit of the Small Litter data to a Weibull distribution. The graphical estimates are

$$\hat{\gamma} = 1/.7386 = 1.3539 \qquad \hat{\alpha} = e^{1.9746} = 7.2037$$

The MLE of the Weibull parameters from R are

$$\hat{\gamma} = \text{Weibull Shape} = 1.265583 \qquad \hat{\alpha} = \text{Weibull Scale} = 7.4268$$

A fairly good match to the graphical estimates.

c. Using the MLE estimates: $P[Y_L > 15] = 1 - F(15) \approx e^{-(15/7.4268)^{1.265583}} = .0877$

Using the Graphical estimates: $P[Y_L > 15] = 1 - F(15) \approx e^{-(30/7.2037)^{1.3539}} = .0672$ about 23% less than the MLE

The distribution-free estimate would be $P[Y > 15] = 1 - \hat{F}(15) = 1 - 47/51 = .0784$ about 11% less than the MLE

d. The distribution-free estimates are

$\hat{\mu}$ = sample mean = $\bar{Y}_S = 6.8859$ $\quad$ $\hat{\sigma}$ = sample stand. dev. = $S_{Y_S} = 5.46$

Using the formulas on page 6 in HO 6, we have for the Weibull distribution, with MLE's from R:

$$\widehat{\mu} = \widehat{\alpha}\Gamma\left(1 + \frac{1}{\widehat{\gamma}}\right) = (7.4268)\Gamma\left(1 + \frac{1}{1.265583}\right) = 6.8981$$

$$\widehat{\sigma} = \widehat{\alpha}\sqrt{\Gamma\left(1 + \frac{2}{\widehat{\gamma}}\right) - \Gamma^2\left(1 + \frac{1}{\widehat{\gamma}}\right)} = 7.4268\sqrt{\Gamma\left(1 + \frac{2}{1.265583}\right) - \Gamma^2\left(1 + \frac{1}{1.265583}\right)} = 5.4882$$

The MLE estimates of μ and σ are very close to the distribution-free estimates thus lending evidence that the Weibull model is the correct model for this data.

e. The distribution-free estimates are

$\hat{\tilde{\mu}} = \hat{Q}(.5)$ = sample median = $Y_{(26)} = 5.00$ $\quad$ Using R-function, quantile(yS,.5,type=5) = 5.00

$\widehat{IQR}$ = sample IQR = $\hat{Q}(.75) - \hat{Q}(.25) = Y_{(.75n+.5)} - Y_{(.25n+.5)} = Y_{(38.75)} - Y_{(13.25)} \Rightarrow$

$\hat{Q}(.75) - \hat{Q}(.25) = (.25 * Y_{(38)} + .75 * Y_{(39)}) - (.75 * Y_{(13)} + .25 * Y_{(14)}) = 10.4625 - 2.545 = 7.9175$

Using R-function, $quantile(yL, .75, type = 5) - quantile(yL, .25, type = 5) = 10.4625 - 2.545 = 7.9175$

Using the formula for the quantile function from a Weibull distribution:

$Q(u) = \alpha(-log(1 - u))^{1/\gamma}$ along with MLE from R for α and γ we have

$\widehat{\tilde{\mu}} = \hat{Q}(.5) = \hat{\alpha}(-log(1 - .5))^{1/\hat{\gamma}} = 7.42681(-log(1 - .5))^{1/1.265583} = 5.56$

$\widehat{IQR} = \hat{Q}(.75) - \hat{Q}(.25) = \hat{\alpha}(-log(1 - .75))^{1/\hat{\gamma}} - \hat{\alpha}(-log(1 - .25))^{1/\hat{\gamma}} = 6.8387$

Equivalently, using the R quantile function for the Weibull distribution, we have

$\widehat{\tilde{\mu}} = qweibull(.5, 1.265583, 7.42681) = 5.56$

$\widehat{IQR} = qweibull(.75, 1.265583, 7.42681) - qweibull(.25, 1.265583, 7.42681) = 6.8387$

The MLE estimate of the median based on the Weibull model is close to the distribution-free estimate (5.56 to 5.00) but there is substantial difference between the two estimates of the IQR (6.8387 to 7.9175). This may be due to the IQR reflecting only the fit of the data in the middle of the distribution.

2. For Large Litter Data: $\hat{\mu}_L = 10.39$ $\;\hat{\sigma}_L = 9.15$

   For Small Litter Data: $\hat{\mu}_S = 6.89$ $\;\hat{\sigma}_S = 5.46$

3. For Large Litter Data: $\hat{\tilde{\mu}}_L = 7.93$ $\;M\hat{A}D_L = 7.95$

   For Small Litter Data: $\hat{\tilde{\mu}}_S = 5.00$ $\;M\hat{A}D_S = 5.23$

4. For the Small Litter Data, the pdf appeared to be just slightly right skewed so the mean should be only slightly larger than the median (6.89 vs 5.00) and the standard deviation somewhat larger than MAD (5.46 vs 5.23). The larger than expected difference in the Mean and Median was very surprising considering that S and MAD were so close in value.

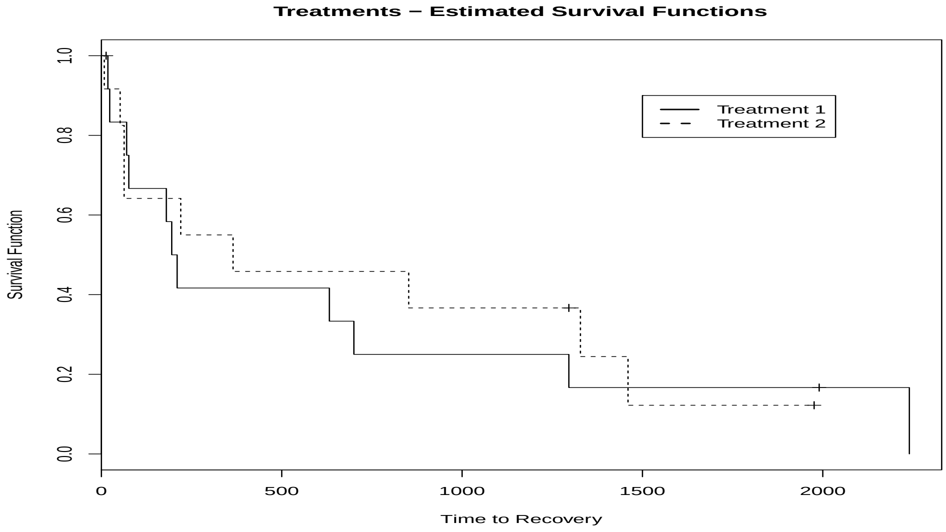
   For the Large Litter Data, the pdf appeared to be just more right skewed so the mean should be larger than the median (10.39 vs 7.93) and the standard deviation somewhat larger than MAD (9.15 vs 7.95). I was somewhat surprised that there was not a larger difference between S and MAD considering the 4 or 5 rather large values in the Large Litter data set.

Based on the right skewness of the estimated pdf for the Large Litter data and the goal of the study was to compare the Small to the Large Litter relative brain weights, I would select (Median, MAD) to represent the location and scale in the two data sets.

5. From the given data, it would appear that larger relative brain weights are associated with Larger Litter sizes. It would be much more informative to have the actual litter sizes associated with each relative brain weight as opposed to having the groupings into just small and large litters.

P2. (30 points) Using the times to recovery (or censoring) for the 25 patients we obtain:

1. The estimate survival functions for the two Treatments are given in the following plot:



2. From the R output we have

```
                 G=1
time n.risk n.event survival std.err lower 95% CI upper 95% CI
  18     12       1    0.917  0.0798       0.7729        1.000
  23     11       1    0.833  0.1076       0.6470        1.000
  70     10       1    0.750  0.1250       0.5410        1.000
  76      9       1    0.667  0.1361       0.4468        0.995
 180      8       1    0.583  0.1423       0.3616        0.941
 195      7       1    0.500  0.1443       0.2840        0.880
 210      6       1    0.417  0.1423       0.2133        0.814
 632      5       1    0.333  0.1361       0.1498        0.742
 700      4       1    0.250  0.1250       0.0938        0.666
1296      3       1    0.167  0.1076       0.0470        0.591
2240      1       1    0.000     NaN           NA           NA
                 G=2
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   8     12       1    0.917  0.0798       0.7729        1.000
  52     10       1    0.825  0.1128       0.6311        1.000
  63      9       2    0.642  0.1441       0.4132        0.996
 220      7       1    0.550  0.1499       0.3224        0.938
 365      6       1    0.458  0.1503       0.2410        0.872
```

```
  852      5       1    0.367  0.1456       0.1684        0.798
 1328      3       1    0.244  0.1392       0.0801        0.746
 1460      2       1    0.122  0.1110       0.0206        0.724
> print(results, print.rmean=TRUE)
Call: survfit(formula = Surv(T, ST) ~ G)
    records n.max n.start events *rmean *se(rmean) median 0.95LCL 0.95UCL
G=1      13    13      13     11    657        229    202      76      NA
G=2      12    12      12      9    731        216    365      63      NA
```

Note that for G=1, the table reports the median as 202 but $\hat{S}(195) = .5$ from the output of the K-M estimator of the survival function. According to our definition of the quantile function, $\hat{Q}(u) = inf\{t : \hat{S}(t) \leq 1 - u\}$, the median would be 195.

The estimated mean and median are smaller for Treatment 1 (G=1) than for Treatment 2 (G=2).

3. Based on the median time to recovery, Treatment 1 would be the more effective treatment. The mean times to recovery are much larger than the median times due to a few very large values in both treatment groups. But, Treatment 1 still has a smaller mean the Treatment 2. However, as we will discuss in future handouts, when the standard errors of the estimators are taken into account, there may not be significant evidence of a difference in the two treatments.

P4. (20 points)

1. **A or B -** Because the true stress for the censored specimens are greater than or equal to $t_C = 500$ psi

2. **D -** Because the amount of water retained at 90 days would be less than the amount of water retained at death.

3. **A or B -** Because the study is terminated at a fixed time, 30 days

4. **A or C -** Because the study was terminated after a pre-selected number of fires

5. **A -** Because brake failure mileage for the censored automobiles are greater than the miles traveled at the end of the study.

V. (10 Bonus points)

- Bonus 1. (5 points)

$$\lim_{\alpha\to .5}\mu_{(\alpha)} = \frac{\lim_{\alpha\to .5}\int_{Q(\alpha)}^{Q(1-\alpha)} yf(y)dy}{\lim_{\alpha\to .5}(1-2\alpha)} = \frac{0}{0}$$

Apply $l^{'}$Hopital's Rule:

$$\begin{aligned}\lim_{\alpha\to .5}\mu_{(\alpha)} &= \frac{\lim_{\alpha\to .5}\frac{d}{d\alpha}\int_{Q(\alpha)}^{Q(1-\alpha)} yf(y)dy}{\lim_{\alpha\to .5}\frac{d}{d\alpha}(1-2\alpha)} \\ &= \frac{\lim_{\alpha\to .5}[Q(1-\alpha)f(Q(1-\alpha))(-1)Q^{'}(1-\alpha) - Q(\alpha)f(Q(\alpha))Q^{'}(\alpha)]}{-2} \\ &= \frac{-2Q(.5)f(Q(.5))Q^{'}(.5)}{-2} = Q(.5)\end{aligned}$$

Therefore, $Q^{'}(u) = \frac{1}{f(Q(u))}$

- Bonus 2. (5 points)

i. The likelihood function is given by

$$L(\beta,\theta;t_1,t_2,\ldots,t_n) = \prod_{i=1}^{n} g(t_i;\ \beta,\theta) = \prod_{i=1}^{n}\beta e^{-\beta(t_i-\theta)}I(t_i\geq\theta)$$

ii. As a function of θ, the likelihood increases as θ increase, until $\theta \geq min(t_1,t_2,\ldots,t_n)$ after which the likelihood becomes 0.

- Therefore, the MLE for θ is $\hat{\theta} = min(t_1,t_2,\ldots,t_n)$

The log-likelihood function is given by

$$l(\beta,\theta;t_1,t_2,\ldots,t_n) = log(L(\beta,\theta;t_1,t_2,\ldots,t_n)) = nlog(\beta) - \beta\left(\sum_{i=1}^{n} t_i - n\theta\right) \quad \text{for all} \quad t_i \geq \theta$$

The log-likelihood evaluated at $\hat{\theta}$ is

$$l = l(\beta,\hat{\theta};t_1,t_2,\ldots,t_n) = nlog(\beta) - \beta\left(\sum_{i=1}^{n} t_i - n\hat{\theta}\right) \quad \text{for all} \quad t_i \geq \hat{\theta}$$

$$\frac{dl}{d\beta} = \frac{n}{\beta} - \sum_{i=1}^{n}(t_i - \hat{\theta}) \quad \text{and} \quad \frac{d^2l}{d\beta^2} = \frac{-n}{\beta^2} < 0$$

Setting $\frac{dl}{d\beta}$ equal to 0 and solving for β yields

$$\hat{\beta} = \frac{n}{\sum_{i=1}^{n}(t_i-\hat{\theta})} \quad \text{which is a maximum because 2nd derivative was negative}$$

iii. $P[T > 40] = e^{-\hat{\beta}(40-\hat{\theta})} = e^{-.1637(40-30)} = 0.1946$

```
####
#### P1
####

##
## (1)
##

library(MASS)
yS = c(0.42,    0.86,    0.88,    1.11,    1.34,    1.38 ,   1.42,    1.47,    1.63,
   1.73,    2.17,    2.42,    2.48,    2.74,    2.74,    2.79,    2.90,    3.12,
   3.18,    3.27,    3.30,    3.61 ,   3.63,    4.13 ,   4.40,    5.00,    5.20,
   5.59,    7.04,    7.15,    7.25,    7.75,    8.00,    8.84,    9.30 ,   9.68,
   10.32,   10.41,   10.48,   11.29,   12.30,   12.53,   12.69,   14.14,   14.15,
   14.27 ,  14.56,   15.84,   18.55,   19.73,   20.00)
yL =
 c( 0.94 ,  1.26 ,  1.44  , 1.49 ,  1.63 ,  1.80 ,  2.00 ,  2.00 ,  2.56,
    2.58 ,  3.24 ,  3.39  , 3.53 ,  3.77 ,  4.36 ,  4.41 ,  4.60 ,  4.67,
    5.39 ,  6.25 ,  7.02 ,  7.89 ,  7.97 ,  8.00 ,  8.28 ,  8.83 ,  8.91,
    8.96 ,  9.92 , 11.36 , 12.15,  14.40 , 16.00 , 18.61 , 18.75 , 19.05,
   21.00 , 21.41 , 23.27 , 24.71,  25.00 , 28.75 , 30.23 , 35.45 )

nS <- length(yS)
nL <- length(yL)

## (a)
yS  = sort(yS)
ySt = yS[c(-(1:5),-(nS - 0:4))]
meanS = mean(yS)
trim.meanS = mean(ySt)
trimmedmean = mean(yS,trim=.10)

## (b)
i <- 1:nS
ui <- (i - 0.5) / nS
QW <- log(-log(1 - ui))
plot(QW, log(yS), main="Weibull Reference Plot - Small Litter",cex=.75,lab=c(7,11,7),
       xlab="Q(u) = log(-log(1-ui))",
       ylab="log(yS(i))")
abline(lm(log(yS) ~ QW))
legend(-4,3.0,"y = 1.9746 + 0.7386 Q(u)")

mle_weib <- fitdistr(yS,"weibull")
gamma_hat <- 1.2655827
alpha_hat <- 7.4268097

## (c)
exp(-(15 / alpha_hat) ^ gamma_hat)

## (d)
mu_hat <- alpha_hat * gamma(1 + 1 / gamma_hat)
sigma_hat <- sqrt(alpha_hat ^ 2 * (gamma(1 + 2 / gamma_hat) -
  (gamma(1 + 1 / gamma_hat)) ^ 2))

mean(yS)
sd(yS)

## (e)
med_hat <- alpha_hat * (-log(1 - 0.5)) ^ (1 / gamma_hat)
Q1_hat <- alpha_hat * (-log(1 - 0.25)) ^ (1 / gamma_hat)
Q3_hat <- alpha_hat * (-log(1 - 0.75)) ^ (1 / gamma_hat)
IQR_hat <- Q3_hat - Q1_hat

0.5 * nS + 0.5
0.25 * nS + 0.5
0.75 * nS + 0.5
med_df <- yS[26]
Q1_df <- 0.75 * yS[13] + 0.25 * yS[14]
Q3_df <- 0.25 * yS[38] + 0.75 * yS[39]
median(yS)
quantile(yS, c(0.25, 0.75), type = 5)
```

```
IQR_df <- Q3_df - Q1_df

##
## (2)
##

xbar_S <- mean(yS)
s_S <- sd(yS)
xbar_L <- mean(yL)
s_L <- sd(yL)

##
## (3)
##

med_L <- mean(yL)
iqr_L <- quantile(yL, 0.75) - quantile(yL, 0.25)
mad_L <- mad(yL)
mad_S <- mad(yS)

####
#### P2
####

library(survival)

T = c( 180, 632, 2240, 195, 76, 70, 13, 1990, 18, 700, 210, 1296, 23, 8, 852,  52, 220, 63,   8, 1976,1296,1460,63,1328,365)
ST = c(  1,   1,    1,   1,  1,  1,  0,    0,  1,   1,   1,    1,  1, 0,   1,   1,   1,  1,   1,    0,   0,   1, 1,   1,  1)
G =  c( rep(1,13),rep(2,12))

out = cbind(T,ST,G)
Surv(T, ST)

results <- survfit(Surv(T, ST) ~ G)
summary(results)
print(results, print.rmean=TRUE,rmean="individual",mark.time=True)

par(lab=c(15,20,4))
plot(results,ylab="Survival Function",xlab="Time to Recovery",mark.time=TRUE,
main="Treatments - Estimated Survival Functions",lty=1:2 )
legend(1500,.9,c("Treatment 1","Treatment 2"),lty=1:2,lwd=2)
```