STATISTICS 641 - EXAM I

Student's Name _____

Student's Email Address _____

INSTRUCTIONS FOR STUDENTS:

(1) The exam consists of 10 pages including this cover page and 13 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 - **2-pages,** $8.5"$**x**$11"$**, write/type/paste on both sides of the two sheets**
- Tables for Exam 1 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_____

Part I. (40 points) CIRCLE One of (A, B, C, D, or E) corresponding to the **BEST** answer to each question.

1. A metallurgist has developed a new alloy for use in bridge support beams. She prepares 30 independent batches of the alloy and then pours 25 beams from each of the 30 batches. The beams are then placed in a device which measures their tensile strength, $T_1, T_2, \ldots, T_{25}$. From each batch the metallurgist measures the minimum tensile strength for the 25 beams: $M_i = min\{T_1, T_2, \ldots, T_{25}\}$ obtaining the data: $M_1, M_2, \ldots, M_{30}$. In order for an alloy to be certified for bridge support beams, its tensile strength must exceed 2500 units. The metallurgist determines N, the number of batches out of the 30 batches which had M_i exceeding 2500. A reasonable distribution for N is

   A. binomial distribution
   B. Weibull distribution
   C. negative binomial distribution
   D. hypergeometric distribution
   E. None of the above would be appropriate

2. The EPA is required to estimate the yearly amount of sulfur in the discharges from coal powered electricity generating plants. The level of relative humidity in the air around the plants has an effect on the sulfur level in the discharges. In order to adjust for the impact of relative humidity, the EPA determined the average annual relative humidity for each utility plant and then classified all such plants into ten groups based on their average relative humidity. A random sample of 20 plants was then selected from each of the 10 groups. The average yearly sulfur discharge was recorded at each of the 200 power plants. 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 stratified multistage cluster random sample.
   E. a multistage cluster random sample.

3. A geologist has determined that minor earthquakes caused by hydraulic fracturing in oil exploration occur according to a Poisson process with an average rate of 1 earthquake per day. The geologist wants to generate a sequence of the numbers of earthquakes that may occur in randomly selected 5 day periods. Let E be the number of earthquakes in a random 5 day period. A single realization of E is generated using $U = .43$ as a random observation from a Uniform on (0,1) distribution. The corresponding value of E is

   A. 1
   B. 2
   C. 3
   D. 4
   E. 5

4. Which of the statements A-D about population parameters in which both μ and σ exist is **FALSE**?

   A. For highly left skewed distributions, the median would generally be used in place of μ as a representation of a distribution's "center".

   B. For highly right skewed distributions, MAD would be preferred over σ as a measure of spread.

   C. The population median $\tilde{\mu}$ is related to the population mean by $|\tilde{\mu} - \mu| \leq \sigma$.

   D. A 5% trimmed mean averages the middle 90% of the population values.

   E. None of the statements are False.

5. The following 30 data values are observations on the random variable Y having pdf $f(y)$:

1.1	1.4	1.7	1.9	2.2	2.6	2.8	3.2	3.4	3.6	4.2	4.3	4.8	5.0	5.4
5.7	7.1	10.7	14.9	18.8	20.4	21.3	23.9	25.1	26.2	34.4	37.2	38.1	45.1	50.7

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

$$\widehat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y - Y_i}{h}\right) \;\; ; \;\; \text{with kernel} \;\; K(u) = \begin{cases} (1 - |u|) & \text{for } |u| < 1 \\ 0 & \text{for } |u| \geq 1 \end{cases} \;\; ; \;\; \text{bandwidth} \;\; h = 2$$

   What is the contribution of $Y_{15} = 5.4$ to the kernel density estimate of $f(4)$?

   A. .01125

   B. .0050

   C. .0283

   D. 1/30

   E. cannot be computed with the given information

$$\frac{1}{30(2)}\left(1 - \left|\frac{4 - 5.4}{2}\right|\right) = \frac{1}{60}(0.3)$$

$$= 0.005$$

6. A researcher decides to use a Guassian kernel density estimator, $\hat{f}(y)$, as an estimator of a continuous population pdf, $f(y)$. She wants to know how the selection of the bandwidth will affect the performance of the estimator. Which ONE of the following statements is TRUE?

   A. The smoothness of $\hat{f}(y)$ increases with increasing values of the bandwidth.

   B. The smoothness of $\hat{f}(y)$ decreases with increasing values of the bandwidth.

   C. The area under the curve $\hat{f}(y)$ increases with increasing values of the bandwidth.

   D. The area under the curve $\hat{f}(y)$ decreases with increasing values of the bandwidth.

   E. All of the above statements are False.

(7.) The quantile function $Q(\cdot)$ is to be estimated using a random sample of $n = 20$ data values:

7.7	8.4	9.0	10.8	12.3	13.2	13.4	16.6	17.3	18.9
20.2	21.2	21.3	21.6	22.2	26.6	27.1	30.7	35.9	43.9

The following estimator of $Q(u)$ was selected

$$\widehat{Q}(u) = Y_{((n+1)u)}$$

as the estimator $Q(u)$. The value of $\widehat{Q}(.80)$ that would be obtained using this estimator is

A. 27.00
B. 26.60
C. 27.10
D. 26.70
E. 26.85

$\widehat{Q}(.80) = Y_{((21)(0.80))} = Y_{(16.8)}$
$= Y_{(16)}(0.2) + Y_{(17)}(0.8)$
$= 27$

8. In estimating the variability in a population, MAD is preferred to the standard deviation, σ as a measure of population dispersion when the population distribution

A. has absolutely no outliers.
B. has normal-like tails.
C. has a symmetric distribution.
D. has a heavy-tailed distribution.
E. has a bi-modal distribution.

9. Which of the following statements about population parameters in which both μ and σ exists is **TRUE**?

A. If a population has its kurtosis equal to 3, then the population distribution has a pdf very similar to a normal pdf.
B. For a symmetric distribution, MAD is preferred to the SIQR as a measure of dispersion.
C. If a population has its skewness equal to 0, then the population is symmetric about its median.
D. Two random variables with correlation equal to zero are independent random variables.
E. None of the above are true.

10. An entomologist is interested in the ability of ticks to conserve water in a very dry environment, relative humidity less than 10%. She randomly selects 100 Lone Star ticks from a large collection of Lone Star ticks and places them in a water-free container in which the temperature is maintained at $30°C$ 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

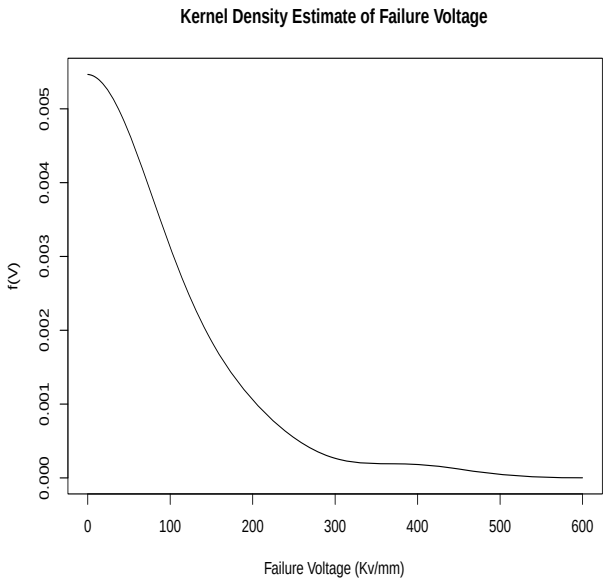
A. Right censored
B. Left censored
C. Type I censored
D. Type II censored
E. Uncensored

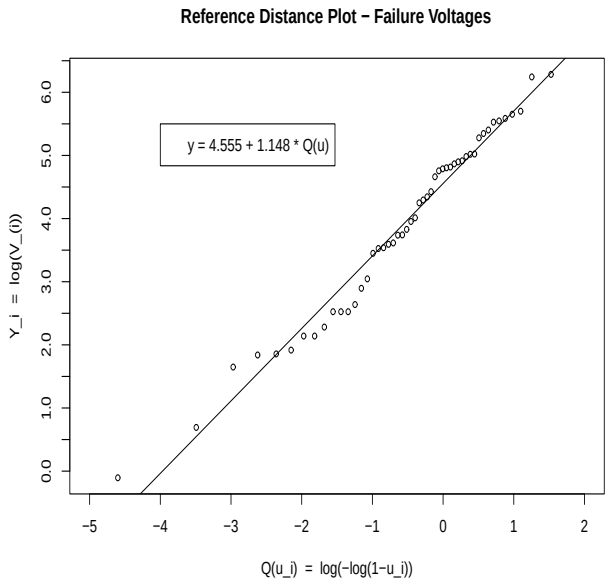
Part II. (60 points) A manufacturer of tablet computers has developed a new model which weighs much less that the previous version of the tablet due to a change in the case material. The process engineer is concerned that the new tablet will be less resilient to power surges. An experiment was conducted to measure the voltage level which would result in a failed tablet. The study consisted of 50 tablets and the data given in the following table is failure voltages in Kv (kilovolts). The process engineer informs you that for any voltage reading greater than 200 Kv/mm in the data set, the true voltage is greater than the recorded voltage due to a calibration problem with the instruments.

0.9	2.0	5.2	6.3	6.4	6.8	8.5	8.5	9.8	12.5	12.5	12.5
14.0	18.1	21.0	31.5	34.0	34.4	36.3	37.1	42.0	42.2	46.0	52.2
55.3	70.0	73.3	77.0	83.6	105.9	116.4	120.1	122.1	123.4	129.8	134.1
136.4	145.9	151.1	151.8	196.1	210.0	222.2	251.3	256.4	266.9	285.0	298.9
514.5	535.7										

Summary statistics, a plot of the estimated pdf for the failure voltages, V_i, and a plot of $Y_i = log(V_{(i)})$ versus $Q_o(u_i) = log(-log(1-u_i))$ are given below, where $u_i = (i-.5)/50$ for $i = 1, \cdots, 50$.

Minimum	$\widehat{Q}(.25)$	$\widehat{Q}(.5)$	Mean	$\widehat{Q}(.75)$	Maximum	SIQR	S	MAD
0.9	15.0	62.65	106.70	143.5	535.70	64.25	121.03	81.54





1. The electronics firm wants to estimate the median failure voltage, $\mathbf{Q}(.\mathbf{5})$, the probability of a failure voltage greater than 80 Kv/mm, $\mathbf{S}(\mathbf{80})$ and the warranty failure voltage, $\mathbf{V_W}$, a value such that at least 90% of the manufactured tablets will have failure voltage greater than $\mathbf{V_W}$, using the Weibull model. Compute your estimates of the $\mathbf{Q(.5)}$, $\mathbf{S}(\mathbf{80})$, and $\mathbf{V}_W$ using the Weibull model with the information about failure voltages greater than 200 Kv/mm incorporated in your estimates.

   Use the SAS output given on Pages 8-10 in obtaining your answers.

   The Weibull cdf has the form

$$F(v) = 1 - e^{-(v/\alpha)^\gamma}$$

   a. Estimate Median Failure Voltage:

*from table below: $\alpha = 118.3669$, $\gamma = .6940$

$Q(v) = F^{-1}(v)$

$Q(v) = \alpha[-\ln(1-v)]^{1/\gamma}$

$Q(0.5) = 118.3669[-\ln(0.5)]^{\frac{1}{.6940}} = 69.80$

   b. Estimate $S(80)$:

$S(80) = 1 - F(80)$

$= e^{-\left(\frac{80}{118.3669}\right)^{0.6940}} = 0.46676$

   c. Estimate V_W:

$V_W = \min\{v : P(V \le v) \le 0.10\}$

$Q(0.10) = 118.3669[-\ln(0.9)]^{\frac{1}{0.6940}}$

$V_W = 4.62365$

2. After considering that some of the 50 measurements were censored, the process engineer questions the validity of the Weibull model. Use the Kaplan-Meirer Product Limit estimator of the survival function, as displayed in the output from SAS on pages 8 - 10, to estimate the three parameters: $\mathbf{Q}(.5)$, $\mathbf{S}(80)$, and $\mathbf{V}_W$.

   a. Estimate Median Failure Voltage:

$$\hat{Q}(0.5) = \inf\{v : \hat{P}[V \le v] \ge 0.5\}$$

$$\hat{Q}(0.5) = 55.30$$

   b. Estimate $S(80)$:

$$\hat{S}(80) = 0.42\left(\frac{(80-63.6)(0.44-0.42)}{97-93.6}\right)$$

$$= 0.437$$

   c. Estimate V_W:

$$\hat{Q}(0.10) = \inf\{v : \hat{S}(v) \le 1 - 0.13$$

$$= 64.$$

3. Based on the plots and your answers in parts 1. and 2., does the Weibull model appear to be an appropriate model for the failure voltages?

See sols.

SAS PROGRAM:

```
data failurevoltage;
input V C  @@;
label V  = 'Failure Voltage' C ='Censoring (0=Yes)';
cards;
       0.9 1   2.0 1   5.2 1   6.3 1   6.4 1   6.8 1   8.5 1   8.5 1   9.8 1  12.5 1  12.5 1  12.5 1
      14.0 1  18.1 1  21.0 1  31.5 1  34.0 1  34.4 1  36.3 1  37.1 1  42.0 1  42.2 1  46.0 1  52.2 1
      55.3 1  70.0 1  73.3 1  77.0 1  83.6 1 105.9 1 116.4 1 120.1 1 122.1 1 123.4 1 129.8 1 134.1 1
     136.4 1 145.9 1 151.1 1 151.8 1 196.1 1 210.0 0 222.2 0 251.3 0 256.4 0 266.9 0 285.0 0 298.9 0
     514.5 0 535.7 0
;
run;

proc lifereg data=failurevoltage;
 model  V = /dist=weibull covb;
run;

proc lifereg data=failurevoltage;
 model  V*C(0) = /dist=weibull covb;
run;

proc lifetest data=failurevoltage outsurv=a plots=(s);
 time V*C(0) ;
 run;
```

SAS OUTPUT:

MAXIMUM LIKELIHOOD ESTIMATES - IGNORING CENSORING

Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	4.5820	0.1766	4.2359	4.9281	673.19	<.0001
Scale	1	1.1842	0.1322	0.9515	1.4739		
Weibull Scale	1	97.7093	17.2552	69.1219	138.1198		
Weibull Shape	1	0.8444	0.0943	0.6785	1.0509		

MAXIMUM LIKELIHOOD ESTIMATES - USING CENSORING INFORMATION

The LIFEREG Procedure

Number of Observations	50
Noncensored Values	41
Right Censored Values	9
Left Censored Values	0
Interval Censored Values	0
Name of Distribution	Weibull

Number of Observations Read	50
Number of Observations Used	50

Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	4.7738	0.2261	4.3307	5.2169	445.81	<.0001
Scale	1	1.4408	0.1849	1.1204	1.8529		
Weibull Scale	1	118.3669	26.7620	75.9939	184.3665		
Weibull Shape	1	0.6940	0.0891	0.5397	0.8926		

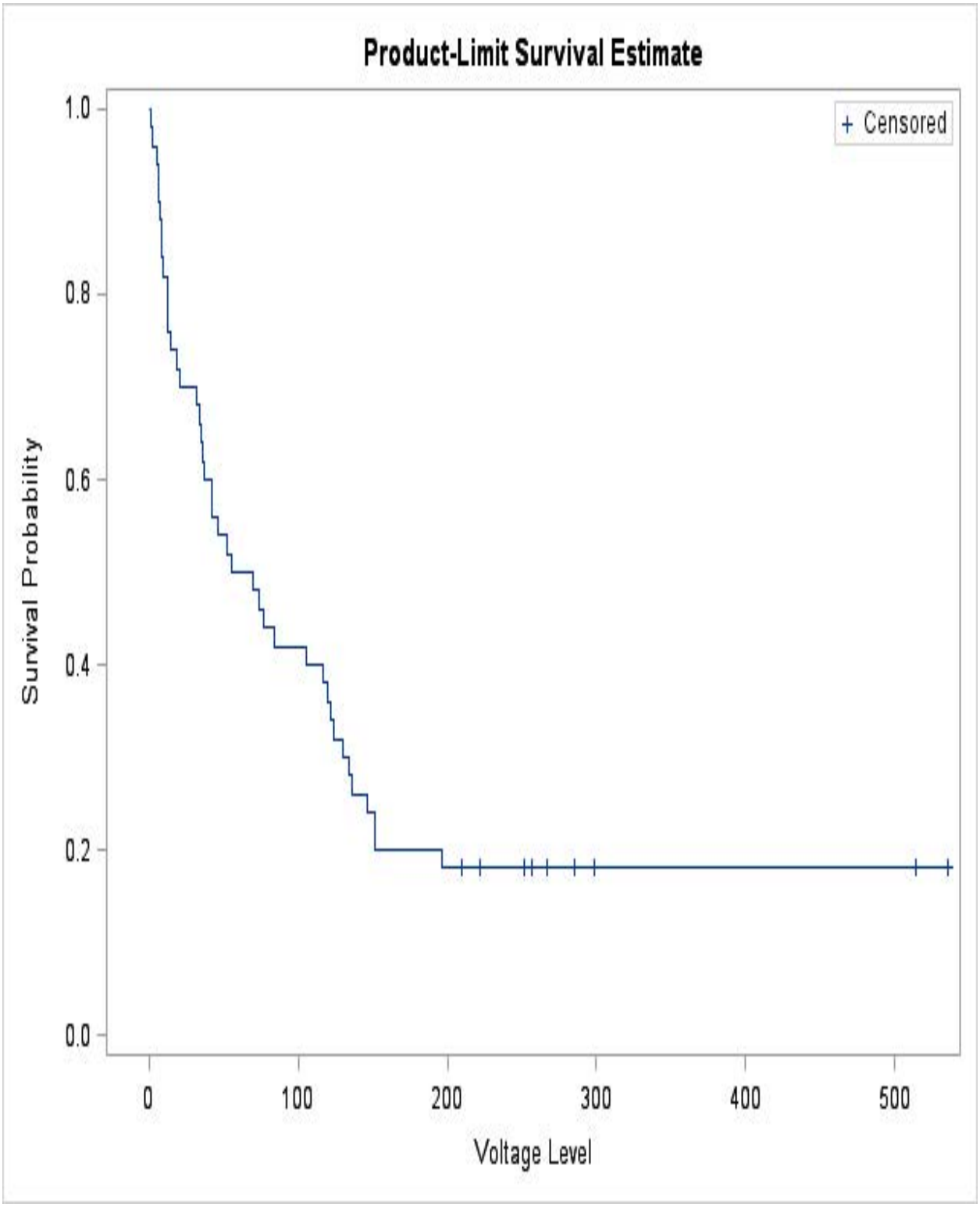
Voltage Levels at Failure

The LIFETEST Procedure

Product-Limit Survival (Kaplan-Meirer) Estimates

V	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.000	1.0000	0	0	0	50
0.900	0.9800	0.0200	0.0198	1	49
2.000	0.9600	0.0400	0.0277	2	48
5.200	0.9400	0.0600	0.0336	3	47
6.300	0.9200	0.0800	0.0384	4	46
6.400	0.9000	0.1000	0.0424	5	45
6.800	0.8800	0.1200	0.0460	6	44
8.500	.	.	.	7	43
8.500	0.8400	0.1600	0.0518	8	42
9.800	0.8200	0.1800	0.0543	9	41
12.500	.	.	.	10	40
12.500	.	.	.	11	39
12.500	0.7600	0.2400	0.0604	12	38
14.000	0.7400	0.2600	0.0620	13	37
18.100	0.7200	0.2800	0.0635	14	36
21.000	0.7000	0.3000	0.0648	15	35
31.500	0.6800	0.3200	0.0660	16	34
34.000	0.6600	0.3400	0.0670	17	33
34.400	0.6400	0.3600	0.0679	18	32
36.300	0.6200	0.3800	0.0686	19	31
37.100	0.6000	0.4000	0.0693	20	30
42.000	0.5800	0.4200	0.0698	21	29
42.200	0.5600	0.4400	0.0702	22	28
46.000	0.5400	0.4600	0.0705	23	27
52.200	0.5200	0.4800	0.0707	24	26
55.300	0.5000	0.5000	0.0707	25	25
70.000	0.4800	0.5200	0.0707	26	24
73.300	0.4600	0.5400	0.0705	27	23
77.000	0.4400	0.5600	0.0702	28	22
83.600	0.4200	0.5800	0.0698	29	21
105.900	0.4000	0.6000	0.0693	30	20
116.400	0.3800	0.6200	0.0686	31	19
120.100	0.3600	0.6400	0.0679	32	18
122.100	0.3400	0.6600	0.0670	33	17
123.400	0.3200	0.6800	0.0660	34	16
129.800	0.3000	0.7000	0.0648	35	15
134.100	0.2800	0.7200	0.0635	36	14
136.400	0.2600	0.7400	0.0620	37	13
145.900	0.2400	0.7600	0.0604	38	12
151.100	0.2200	0.7800	0.0586	39	11
151.800	0.2000	0.8000	0.0566	40	10
196.100	0.1800	0.8200	0.0543	41	9
210.000*	.	.	.	41	8
222.200*	.	.	.	41	7
251.300*	.	.	.	41	6
256.400*	.	.	.	41	5
266.900*	.	.	.	41	4
285.000*	.	.	.	41	3
298.900*	.	.	.	41	2
514.500*	.	.	.	41	1
535.700*	.	.	.	41	0

NOTE: The marked survival times are censored observations.


Product-Limit Survival Estimate
+ Censored
Survival Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
Voltage Level

STAT 641 - Partial Solutions - EXAM I

Part I. (40 points)

1. **A:** N is the number of batches out of the 30 batches having $M > 2500$. This is a sequence of iid Bernoulli trials. Thus, the binomial distribution with n=30 and $p = P[M > 2500]$ would be a reasonable model for N.

2. **C:** This study is a stratified simple random sample with Strata being the ten Relative Humidity Levels and Sampling Unit is a Power Plant.

3. **D:** The value of E corresponding to $U = 0.43$ is obtained from the Poisson distribution with $\lambda = (5)(1) = 5$. Using the Poisson tables with $\lambda = 5$ the following values are obtained from the cdf, F(y), for the Poisson distribution or are obtained by summing the terms in the Poisson pmf $f(y) = \lambda^y e^{-\lambda}/y!$ for $y = 0, 1, 2, \ldots$ with $\lambda = 5$.

y	0	1	2	3	4
f(y)	.007	.033	.085	.140	.175
F(y)	.007	.040	.125	.265	.440

   Therefore, the randomly generated value of E is 4, the smallest value of y such that $F(y) = P[Y \leq y] \geq 0.43$.

4. **E:** All of the statements are true.

5. **B:** The contribution of $Y_{15} = 5.4$ to the kernel density estimate of $f(4)$ is obtained from

$$\frac{1}{(30)(2)} K\left(\frac{4-5.4}{2}\right) = \frac{1}{(30)(2)}\left(1 - \left|\frac{4-5.4}{2}\right|\right) = .005$$

6. **A:** As the bandwidth, h, increases $\hat{f}(y)$ incorporates more of the data values into the estimate and hence smoothes the estimate.

7. **A:** $\widehat{Q}(u) = Y_{(n+1)u}$ along with n=20, yields

   $\widehat{Q}(.8) = Y_{(20+1).8} = Y_{16.8} = Y_{16} + .8(Y_{17} - Y_{16}) = 26.6 + .8(27.1 - 26.6) = 27.0$

8. **D:** MAD is preferred to standard deviation as a measure of population dispersion when the population distribution has a heavy-tailed distribution.

9. **E:** All of the statements are false.

10. **B:** The recorded value of the amount of retained water for the censored ticks will be greater than their value at 90 days thus the type of censoring is Left censored

Part II. (60 points)

1. (9 points each) The survival function is $S(v) = 1 - F(v) = e^{-(v/\alpha)^{\gamma}}$ and the quantile function is

   $Q(u) = \alpha[-log(1-u)]^{1/\gamma}$.

   From the SAS output, the estimators of the Weibull parameters which incorporate censoring are given by $\hat{\alpha} = 118.3669$ and $\hat{\gamma} = 0.694$. Therefore,

   a. $\hat{\tilde{\mu}} = \hat{Q}(.5) = \hat{\alpha}[-log(1-u)]^{1/\hat{\gamma}} = (118.3669)(-log(1-.5))^{1/.694} = 69.80$

   b. $\hat{S}(80) = e^{-(80/118.3669)^{.694}} = .467$:

   c. V_W is that value of V such that $P[V \geq V_W] = .9$, that is, $P[V < V_W] = .1$, which yields $V_W = Q_V(.1)$. Therefore, the MLE estimate is

   $\hat{V}_W = \hat{Q}_V(.1) = \hat{\alpha}[-log(1-.1)]^{1/\hat{\gamma}} = (118.3669)(-log(1-.1))^{1/.694} = 4.623$

2. (9 points each) Using the Kaplan-Meirer Product Limit estimator produced by SAS, the three estimates are

   a. $\hat{Q}(u) = inf\left\{v : \hat{S}(v) \leq 1-u\right\} \Rightarrow \hat{Q}(.5) = inf\left\{v : \hat{S}(v) \leq .5\right\} = 55.3$

   b. $\hat{S}(80) = .42 + \frac{(80-83.6)(.44-.42)}{.77-83.6} = 0.431$

   c. $V_W = Q(.1) \quad \Rightarrow \quad \hat{V}_W = \hat{Q}(.1) = inf\left\{v : \hat{S}(v) \leq 1-.1\right\} = 6.4$

3. (6 points) Ignoring that some of the data values are right censored, the Weibull model appears to provide a reasonable model for the failure voltages for the following reasons:

   i. The estimated pdf for V is a highly right skewed pdf

   ii. The plotted points in the Weibull reference plot are reasonable close to a straight line.

   - However, ignoring the censoring could result in a very misleading conclusion. Thus the reference distribution plot should be modified to take into account the censoring. Using the estimates which take into account the censoring we have the following results:

     The Weibull based estimates of the median and V_W are not very close to the distribution-free estimates obtained from the Product Limit Estimator of S(t). Likewise, the Kaplan-Meirer PL estimate of V_W is about 50% larger than the Weibull based estimator. Therefore, I would tentatively conclude that the Weibull model is not a good fit to the voltage data.

     Median: 69.8 vs 55.3 S(80): .467 vs .43 V_W: 4.623 vs 6.4

Exam 1 Scores for STAT 641 - Fall 2019

Min = 37, Q(.25) = 61.3, Q(.5) = 83.5, Mean = 75.7, Q(.75) = 91, Max = 100