STATISTICS 641 - ASSIGNMENT 6

DUE DATE: 11:59pm (CST), MONDAY, NOVEMBER 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.

ASSIGNMENT 6 - DUE 11:59pm (CST), MONDAY, NOVEMBER 8, 2021

- Read Handout 11
- Supplemental Reading from Devore book: Chapters 7 & 9 & Sections 15.3
- The data for Problems I., IV. and V. are on our Google Drive in Data folder

  `Assignment06_Fall2021_data.txt`

P1. (16 Points) An experiment was designed to evaluate whether or not rainfall can be increased by treating clouds with silver iodide. Rainfall was measured from 60 clouds, of which 30 were chosen randomly to be seeded with silver iodide. The objective is to describe the effect that seeding has on rainfall. The measurements are the amounts of rainfall in acre-feet from the 60 clouds.

Seeded: 151, 450, 124, 235, 357, 110, 302, 671, 118, 115, 275, 275, 2550, 243, 201, 199, 130, 119, 92, 91, 92, 98, 1650, 1200, 1180, 900, 700, 460, 340, 330

Unseeded: 246, 268, 275, 348, 305, 311, 206, 279, 426, 269, 257, 299, 337, 329, 319, 312, 327, 342, 351, 205, 151, 426, 154, 353, 396, 441, 254, 263, 278, 281

NOTE: In what follows, you should first check whether the data are Normally distributed. If not, apply a Box-Cox transformation.

1. Place 95/95 lower tolerance intervals on the amount of rainfall amounts from both seeded and unseeded clouds. If you had to transform one or both of the datasets, create a bound for the transformed data, then back-transform to get a bound on the original scale.
2. Place 95% confidence intervals on the average rainfall from both seeded and unseeded clouds. If you had to transform one or both of the datasets, use the studentized bootstrap, because our confidence interval procedures for a mean are not appropriate for transformed data.
3. Place 95% confidence intervals on the median rainfall from both seeded and unseeded clouds. Note that, since with the Normal distribution, the median equals the mean, you can just apply confidence interval approaches for a mean. If you had to transform one or both of the datasets, go ahead and use the confidence interval approach for a mean and back-transform.
4. What can you conclude about the effect of the seeding on the amount of rainfall?

P2. (10 Points) Twenty-eight bundles of impregnated carbon fibers of length 20 mm are exposed to gradually increasing stress until they finally fail. The stress at failure are recorded as follows. The maximum stress that can be applied to the fibers is 3 and four of the fibers had **not** failed at that stress so a value of 3 was assigned to the four fibers:

2.526, 2.546, 2.628, 2.669, 2.869, 2.710, 2.731, 2.751, 2.771, 2.772, 2.782, 2.789, 2.793, 2.834,
2.844, 2.854, 2.875, 2.876, 2.895, 2.916, 2.919, 2.957, 2.977, 2.988, 3, 3, 3, 3

1. Estimate with a 95% confidence interval the average stress to failure for the carbon fibers without specifying the distribution of the stress to failure values. Do this two ways: (i) using an asymptotic CI based on the results of the R function `survfit`, and (ii) using the studentized bootstrap, treating the censored observations as true stress values (i.e., ignoring the censoring).
2. Estimate with a 95% confidence interval the average stress to failure for the carbon fibers assuming the distribution of the stress to failure values has a Weibull distribution. Do this using the parametric bootstrap. To estimate the Weibull parameters, use the `survreg` function:

```
fit <- survreg(Surv(x, delta) ~ 1, dist = "weibull")
shape_est <- 1 / fit$scale
scale_est <- exp(fit$coef)
```

P3. (9 Points) The National Institute for Standards and Technology conducted a study to develop standards for asbestos concentration. Asbestos dissolved in water was spread on a filter, and punches of 3-mm diameter were taken from the filter and mounted on a transmission electron microscope. An operator counted the number of asbestos fibers on each of 200 grid squares yielding the following counts: (the researcher no longer had the original counts just the following grouped data and the mean of the 200 counts $\bar{Y} = \frac{1}{200}\sum_{i=1}^{200} Y_i = 4940/200 = 27.7$)

	Grouped Counts							
	0-10	11-15	16-20	21-24	25-27	28-30	31 or more	Total
O_i	2	1	36	52	50	39	20	200
E_i	.12	2.43	34.62	57.51	45.36	32.62	27.34	200
$\frac{(O_i-E_i)^2}{E_i}$	30.25	.85	.06	.53	.48	1.25	1.97	35.39

1. The consulting statistician computed the Chi-square GOF and obtained $p - value < .001$ and then stated that the Poisson model provided an inadequate fit to the data. Do you agree with his results? If not correct his computations and reassess the fit of the Poisson model.
2. Assuming that the statistician was in fact correct and the Poisson model provided a reasonable fit to the counts, construct a 95% confidence interval on the average number of asbestor fibers per 3-mm diameter area. (Hint: If $Y_1, \ldots, Y_n$ are iid Poisson(λ) r.v.'s, then by the central limit theorem, the distribution of $\frac{\sqrt{n}(\bar{Y}-\lambda)}{\sqrt{\lambda}}$ is approximately $N(0,1)$ for large n.

P4. (15 Points) The space shuttle uses epoxy spherical vessels in an environment of sustained pressure. A study of the lifetimes of epoxy strands subjected to sustained stress was conducted. The data giving the lifetimes (in hours) of 100 strands tested at a prescribed level of stress is given in the following table.

.18	3.1	4.2	6.0	7.5	8.2	8.5	10.3	10.6	24.2
29.6	31.7	41.9	44.1	49.5	50.1	59.7	61.7	64.4	69.7
70.0	77.8	80.5	82.3	83.5	84.2	87.1	87.3	93.2	103.4
104.6	105.5	108.8	112.6	116.8	118.0	122.3	123.5	124.4	125.4
129.5	130.4	131.6	132.8	133.8	137.0	140.2	140.9	148.5	149.2
152.2	152.9	157.7	160.0	163.6	166.9	170.5	174.9	177.7	179.2
183.6	183.8	194.3	195.1	195.3	202.6	220.0	221.3	227.2	251.0
266.5	267.9	269.2	270.4	272.5	285.9	292.6	295.1	301.1	304.3
316.8	329.8	334.1	346.2	351.2	353.3	369.3	372.3	381.3	393.5
451.3	461.5	574.2	656.3	663.0	669.8	739.7	759.6	894.7	974.9

1. Estimate with a 99% confidence interval the probability that an epoxy strand subjected to the prescribed stress will survive for 300 hours. Use the Agresti-Coull approach.
2. Estimate with 99% certainty the time, $L_{.95}$, such that at least 95% of epoxy strands under the prescribed stress would have lifetimes greater than $L_{.95}$. You can assume that the lifetimes follow an exponential distribution.
3. Using the above data, predict with 95% certainty the lifetime of a strand subjected to the prescribed stress. Again, assume an exponential distribution.

P5. (20 Points) An experiment was conducted to determine the strength of a certain type of braided cord after weathering. The strengths of 56 pieces of cord that had been weathered for 30 days were investigated. The 56 pieces of cord were placed simultaneously in a tensile strength device. The device applies an increasing amount of force until the cord fails. The following strength readings (psi) are given below.

19.7	21.6	21.9	23.5	24.2	24.4	24.9	25.1
26.4	26.9	27.6	27.7	27.9	28.4	29.8	30.7
31.1	31.1	31.7	31.8	32.6	34.0	34.8	34.9
35.1	36.6	37.0	37.7	38.7	38.7	39.0	39.6
40.0	41.4	41.4	41.8	42.2	43.5	44.5	45.0
45.5	45.9	46.3	46.7	46.7	47.0	47.0	47.4
47.6	48.6	48.8	57.9	58.3	67.9	84.2	97.3

1. The manufacturer of the cords would like to estimate the proportion of weathered cords having tensile strength less than 50 psi. Provide a 95% confidence interval based on the information from the failure data from the 56 cords.
2. The manufacturer is planning a sales campaign to promote its cord and would like to state a tensile strength value for its cord. Provide the manufacturer with a 95% confidence interval that would provide an estimate of the median tensile strength value for the weathered braided cord.
3. In order to determine if the braided cord has tensile strength that falls within federal specifications, the manufacturer wants to determine an interval of strength values, (T_L, T_U), such that the manufacturer would be 95% confident that the interval would contain at least 90% of all strength values for its braided cords.

P6. (30 points) **Multiple Choice Questions** Select the letter of the **BEST** answer.

(1) A pipeline engineer is investigating the strength of pipe used to transport gasoline. The company has a warehouse of pipe and wants to determine an interval of values which will have 95% confidence of containing 90% of the strength readings. The engineer will construct the interval using a random sample of 275 pipes and recording their strengths: $Y_1, \ldots, Y_{275}$. A plot of the data reveals that the data is highly right skewed. The Shapiro-Wilk statistic for the transformation $X_i = \sqrt{\frac{1}{Y_i}}$ yields p-value=0.3467. Which of the following methods would be best for constructing the interval of strength values?

A. Use a studentized Bootstrap prediction interval.

B. Estimate the population pdf using a kernel density estimator and then use the MLEs of the parameters in the kernel density estimator.

C. Use a distribution-free prediction interval because $n = 275$ is very large.

D. Use the inverse of the end points of the interval constructed using the X_is: $\left(1/U_X^2, 1/L_X^2\right)$.

E. None of the above procedures would be acceptable.

(2) The bootstrap procedure for constructing a C.I. for a parameter θ is often used instead of a distribution-based procedure in constructing the C.I. when

A. the researcher wants just a rough idea of the values of θ.

B. the parametric procedure produces a very wide C.I.

C. a nonparametric and parametric procedure yield very different C.I.s.

D. when the conditions for applying the parametric procedure are not satisfied.

E. all of the above are true

(3) A study is to be conducted to estimate the mean tensile strength in newtons per square meter (N/m^2)of a new alloy. What sample size is needed to ensure that the sample mean will estimate the average tensile strength to within 10 N/m^2 with a reliability of 99%. Tensile strength generally has a normal distribution with a standard deviation of approximately 30 N/m^2.

A. 35

B. 49

C. 60

D. 538

E. cannot be determined since σ is unknown

(4) An industrial process produces piston rings having a nominal diameter of 9 cm. A 95% confidence interval for the mean diameter of the piston rings produced during July was calculated and a 95%/95% tolerance interval was calculated for the diameters of the piston rings produced during July. Which one of the following statements is true?

A. The probability is .95 that the mean diameter will fall within the confidence interval.

B. If the engineer wanted to set limits such that 95% of the output was within these limits, then the tolerance interval would be more informative than the confidence interval.

C. The width of the 95% confidence interval is generally narrower than the width of the 95%/95% tolerance interval and hence is a more precise estimator.

D. The tolerance interval for the piston diameters could be used to determine if the mean diameter for July's output was equal to 9 cm or not.

E. All of the statements A-D are false.

(5) In using the studentized bootstrap procedure to construct confidence intervals, there are two types of approximations involved in the procedure.

A. the sample size n and the level of confidence α

B. the estimation of the population cdf, F using the edf $\widehat{F}$ and the estimation of the edf using the bootstrap cdf $\widehat{F}_B$

C. the sample from the population and the sample from the resample

D. the bias and variance of the estimator

E. the estimation of the population cdf F and pdf f

(6) Suppose that $X_1, \cdots, X_n$ are highly positively correlated with a $N(\mu, \sigma^2)$ distribution. A 95% confidence interval for μ was constructed using the formula $\bar{X} \pm (t_{\alpha/2,n-1})(s/\sqrt{n})$. The true coverage probability of this confidence interval

A. is 0.95.

B. is very close to 0.95.

C. is much less than 0.95.

D. is much greater than 0.95.

E. may be greater or less than 0.95.

(7) A 95/95 tolerance interval is to be constructed for a population having pdf $f(\cdot)$. Which one of the following statements is true?

A. The tolerance interval is always wider than a 95% confidence interval for the population parameter.

B. It is necessary to specify the family for $f(\cdot)$ in order to construct the tolerance interval.

C. The normal based tolerance interval will have approximately the correct probabilities provided the sample size is large enough for the central limit theorem to be valid.

D. The distribution-free tolerance interval will generally be wider than the tolerance interval based on a specified family for $f(\cdot)$.

E. All of the above statements are true..

(8) Let $X_1,\ X_2, \ldots, X_{10}$ be iid observations from a population having pdf f. The researcher wants a 95% confidence interval on σ the population standard deviation. The researcher does not know the exact form of f but states that it is highly right skewed. What method of constructing the C.I. would you recommend?

A. use a distribution-free approach.

B. use the Box-Cox transformation.

C. apply a normal based procedures because normal based procedures are usually robust to departures from normality.

D. construct a bootstrap C.I. for σ.

E. recommend to the experimenter that more data would be required to construct the C.I.

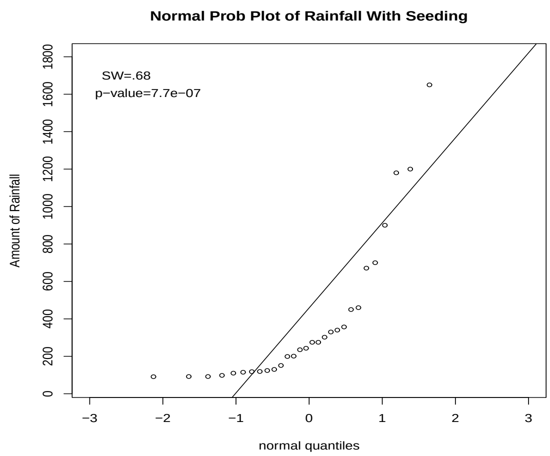
Stat 641 Fall 2021
Solutions for Homework 6

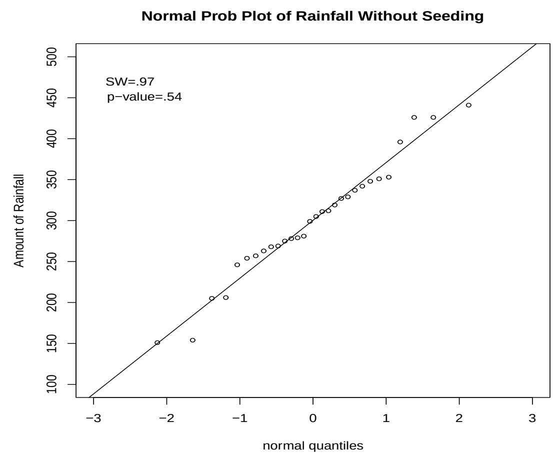
Problem P1. (16 points)

Need to determine if the two data sets come from a normal distribution and if not, can we transform data to normality:

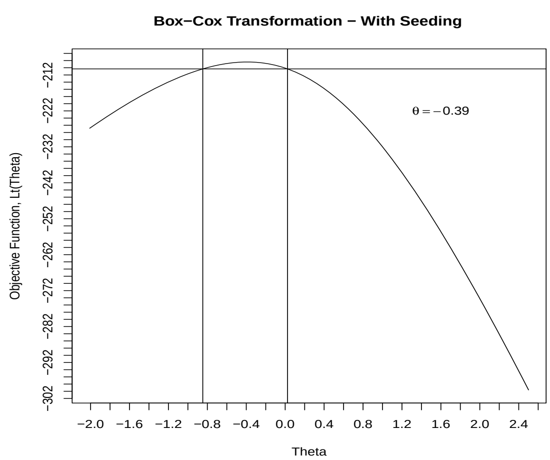
i. Clouds With Seeding: non-normal from plot, SW=.679 and p-value=0.0000008 from Shapiro-Wilk.

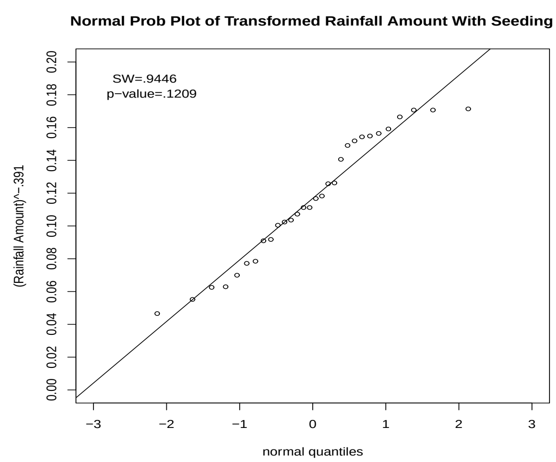
ii. Clouds Without Seeding: normal from plot, SW=.970 and p-value=0.539 from Shapiro-Wilk.





iii. Box-Cox transformation for With Seeding Data and Normal Probability Plot - Transformed With Seeding Data:





1. For a population with a normal distribution, 95/95 lower tolerance bound and 95/95 upper tolerance bound are respectively,

$$L_{0.95,0.95} = \bar{X} - K_{0.95,0.95}S, \qquad L_{0.95,0.95} = \bar{X} + K_{0.95,0.95}S,$$

where for n=30, $K_{0.95,0.95} = 2.220$ and $S = \sqrt{\frac{1}{n-1}\sum_i (X_i - \bar{X})^2}$.

i. For the With Seeding data (Y_1), using the Box-Cox transformation we have that the Shapiro-Wilk's test of $X = g(Y_1) = Y_1^{-.391}$ has a p-value = .121 along with the normal reference plot on the previous page indicates that the transformed rainfall has approximately a normal distribution.

This is a decreasing function, therefore, a 95/95 lower tolerance bound for the distribution of Y_1 is the inverse transformation of the 95/95 upper tolerance bound for the distribution of X:

$$\bar{X} + K_{0.95,0.95}S = 0.1168 + (2.220)(0.0382) = 0.202.$$

Therefore, 95/95 lower tolerance interval on the rainfall from clouds With Seeding, Y_1, is

$$(g^{-1}(0.2016) = (0.2016)^{-1/0.39}, \infty) = (60.72,\ \infty).$$

- Note the sample size for our data is too small, n=30, to use a distribution free technique when $\gamma = .95$, P=.95 and n=30 from Table.

ii. Without Seeding data (Y_2) is normally distributed. Thus, 95/95 lower tolerance bound is

$$\bar{Y_2} - K_{0.95,0.95}S = 300.2667 - (2.220)(71.3031) = 141.97.$$

Therefore, 95/95 lower tolerance interval on the rainfall from clouds Without Seeding is $(141.97,\ \infty)$.

2. i. With Seeding data is not normally distributed and transformations are not in general appropriate for generating C.I. for a population mean, the studentized bootstrap C.I. will be implemented.

the 95% CI for the mean (μ_1) rainfall from clouds With Seeding is (thL, thU) = (293.21, 817.37).

If you assumed that $n = 30$ was large enough to apply asymptotic results you will obtain the following results:

$$\bar{Y_1} \pm t_{29,0.025}\frac{S_{y_1}}{\sqrt{n}} = 458.6 \pm (2.045)\frac{552.2462}{\sqrt{30}} = (253.41,\ 665.79).$$

I would not be very confident in the validity of this CI due to the data being highly skewed with a small sample size n = 30. Note that the asymptotic CI and the studentized bootstrap CI are very different: (293.21, 817.37) versus (253.41, 665.79).

ii. Under the normality of Without Seeding data, the 95% CI for the mean (μ_2) rainfall Without Seeding is

$$\bar{Y_2} \pm t_{29,0.025}\frac{S_{y_2}}{\sqrt{n}} = 300.2667 \pm (2.045)\frac{71.3031}{\sqrt{30}} = (273.64,\ 326.89).$$

If you apply the studentized bootstrap method to the Without Seeding data, the 95% CI for the mean (μ_2) rainfall Without Seeding is (274.19, 326.90) versus the parametric CI, (273.64, 326.89), a close agreement between the two C.I.s.

3. For the median we construct a CI for the transformed data and then apply the inverse transformation to this CI to obtain the CI for the median in the original scale. This method works for any quantile because if $X = g(Y)$ then $Q_X(u) = g(Q_Y(u))$ thus if $(L_X,\ U_X)$ is a $100(1-\alpha)\%$ C.I. for $Q_X(u)$ then $(g^{-1}(L_X),\ g^{-1}(U_X)$ is a C.I. for $Q_X(u)$ provide that g is an increasing function. If g is a decreasing function then reverse the endpoints to obtain the CI on $Q_Y(u)$.

i.a From P1- part 1. of this assignment, we have shown that the With Seeding data (Y_1), $X = g(Y_1) = Y_1^{-.391}$ has approximately a normal distribution. In the normal distribution, the mean and median are the same parameter so we can use the formula for a CI on the mean of a normally distributed population for the transformed data:

$$\bar{X_1} \pm t_{29,0.025}\frac{S_{x_1}}{\sqrt{n}} = .1168 \pm (2.045)\frac{.0382}{\sqrt{30}} = (.1025375,\ 0.1310625).$$

Thus, we have that $((0.1310625)^{-1/.391},\ (.1025375)^{-1/.391}) = (180.75,\ 338.62)$ is a 95% C.I. on the median rainfall With Seeding.

i.b Alternatively, a distribution-free C.I. on the median is given by: Using our R-code or Table VII.3, we have $r = 10$. Thus a distribution-free 95% CI on the median rainfall With Seeding is

$$(Y_{(r)}, Y_{(n-r+1)}) = (Y_{(10)}, Y_{(21)}) = (130,\ 357).$$

- Note the two C.I.'s are somewhat different: (180.75, 338.62) vs (130, 357) but fairly close considering we have only n=30 data values.

ii. For the Without Seeding rainfall, the data is from normal distribution and so the mean is equal to the median. Thus, the answer should be the same as in part 2, (273.64, 326.89).

The distribution-free 95% CI on the median rainfall With Seeding is $(Y_{(r)}, Y_{(n-r+1)}) = (Y_{(10)}, Y_{(21)}) = (269,\ 329)$

Note the C.I. for the median, (269, 329), is in close agreement with the C.I. for the mean, (273.64, 326.89) for the Without Seeding data. We would expect this to be true because the Without Seeding data has a normal distribution and hence the mean and median are the same parameter.

However, this is not true for the With Seeding data, (293.21, 817.37) for the mean and (130, 357) for the median. This is due to the heavy skewness in the With Seeding data.

4. It would appear that Seeding produces a few very large rainfalls (1180, 1200, 1650, 2550) compared to Unseeded (all values less than 500) however the median rainfalls are about the same for Seeding and Unseeded. Furthermore, with the exception of a few very large rainfalls, the distribution of the Seeded clouds tended to produce smaller rainfalls than the Unseeded clouds.

Problem P2. (10 points) Failure Stress of impregnated carbon fibers:

1. First analysis without specifying the distribution of the stress to failure values using the Kaplan-Meier Estimator.

   i. Using the provided R code, the estimate of the average stress to failure for the carbon fibers from the R code is 2.8311 with a standard error of 0.0249. We could obtain a 95% CI for μ using the expression

   $$\hat{\mu} \pm z_{\alpha/2}\widehat{SE}(\hat{\mu}) = 2.8311 \pm 1.96(.0249) = (2.78, 2.88)$$

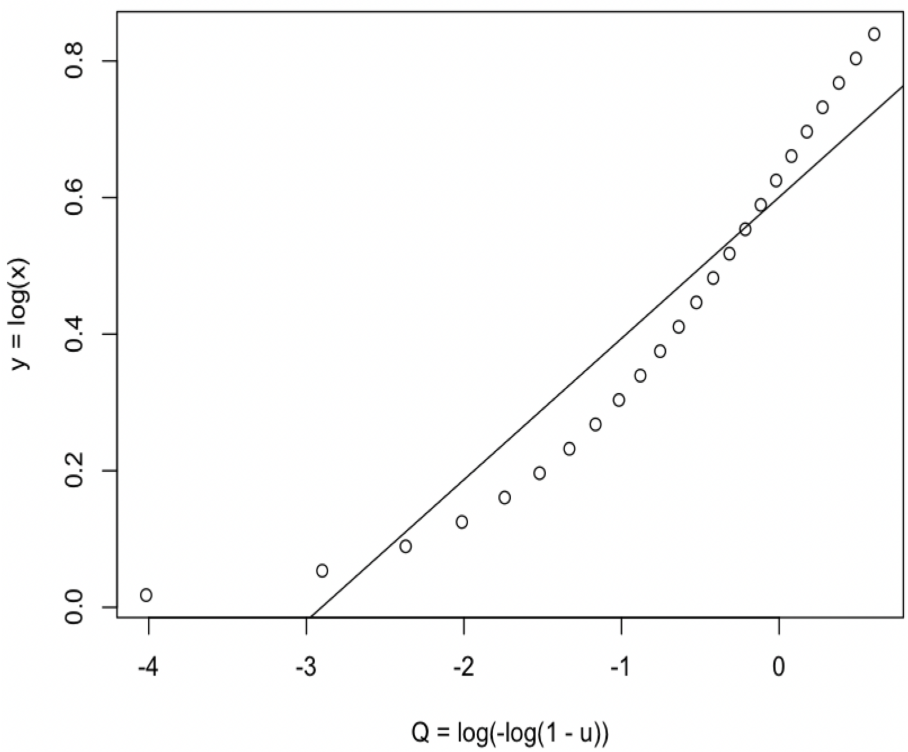
   The problem with the CI is that it is an asymptotic results and with $n = 28$ an asymptotic result may be questionable.

   ii. The provided R code yields (thL, thU) = (2.784, 2.880) as a 95% for the mean which is nearly identical to the asymptotic result. This is very surprising in that the bootstrap just averaged the censored values as if their actual value was 3.

2. The Weibull Reference plot will be modified to take into account the fact that 4 of the 28 data values are Type I censored.

   For the probability plot, make the transformation $Y_i = ln(W_i)$ and then plot $\big(Q(u_i), Y_{(i)}\big)$ for the 24 uncensored values using $u_i = \frac{i-.5}{28}$ $i = 1, \ldots, 24$ and the quantile function for the standard extreme value distribution: $Q(u_i) = log(-log(1-u_i))$. Note that we only plotted the 24 uncensored values but used n=28 in computing u_i.

   From the plot, it would appear that the Weibull distribution provides a fair fit to the data.



Also, using the modified Anderson-Darling statistic on just the uncensored values yields a value of AD = 0.1508 with a p-value between 0.474 and 0.637. This, together with the Reference Distribution plot suggests that there is a reasonable fit of the Weibull model to the data set.

i. **Exact C.I.** To obtain a confidence interval explicitly for the mean from a Weibull distribution is very involved due to the complex relationship between the mean μ and the parameters α and γ:

   $\mu = \alpha\Gamma\left(1 + \frac{1}{\gamma}\right)$ and variance $\sigma^2 = \alpha^2\Gamma\left(1 + \frac{2}{\gamma}\right) - \left(\alpha\Gamma\left(1 + \frac{1}{\gamma}\right)\right)^2$.

   The parametric bootstrap 95% C.I. for the mean would be (2.729, 2.884). This C.I. has endpoints nearly identical to the Kaplan-Meier C.I.: (2.782, 2.880).

Problem P3. (9 points)

1. The expected count for the Group $0 - 10$ is less than 1 which invalidates the Chi-square approximation. Need to combine groups $0 - 10$ and $11 - 15$ to obtain $E_1 = .12 + 2.43 = 2.55$; $O_1 = 2 + 1 = 3$; $\frac{(O_1 - E_1)^2}{E_1} = \frac{(3-2.55)^2}{2.55} = .08 \quad \Rightarrow \quad \chi^2 = 35.39 - 30.25 - .85 + .08 = 4.37$ with $df = 6 - 1 = 5$. Thus, $p - value = Pr[\chi^2_5 \geq 4.37] \approx pchisq(4.37, 5) = 0.497 \quad \Rightarrow$ Poisson model provides an excellent fit to the data.

2. $\hat{\lambda} = \bar{Y} = 27.7$

   $Pr\left[\frac{\sqrt{n}(\bar{Y} - \lambda)}{\sqrt{\lambda}} \leq Z_{\alpha/2}\right] \approx 1 - \alpha$. Thus, we have

   $\sqrt{n}|\bar{Y} - \lambda| \leq \sqrt{\lambda} Z_{\alpha/2} \quad \Rightarrow \lambda = \frac{2\bar{Y} + \frac{1}{n}Z^2_{\alpha/2} \pm \sqrt{(2\bar{Y} + \frac{1}{n}Z^2_{\alpha/2})^2 - 4\bar{Y}^2}}{2} \quad \Rightarrow$

   An approximate $100(1 - \alpha)\%$ C.I. for λ is

   $\bar{Y} + \frac{1}{2n}Z^2_{\alpha/2} \pm Z_{\alpha/2}\sqrt{\frac{1}{n}\bar{Y} + \frac{1}{4n^2}Z^2_{\alpha/2}} = 27.7 + \frac{1}{2*200}1.96^2 \pm 1.96\sqrt{\frac{1}{200}27.7 + \frac{1}{4*200^2}1.96^2} = (26.98, 28.44)$

A less accurate approximate $100(1-\alpha)\%$ C.I. for λ is the basic Wald C.I. with $\hat{\lambda} = \bar{Y}$:

$\bar{Y} \pm Z_{\alpha/2}\sqrt{\bar{Y}/n} = 27.7 \pm 1.96\sqrt{27.7/200} = (26.97,\ 28.43)$

Problem P4. (15 points)

1. Let Y be the lifetime of epoxy strands and p be the probability that an epoxy strand will survive for 300 hours. Therefore, the parameter to be estimated is $p = P(Y \geq 300)$.

   Because $\min\{n\widehat{p}, n(1-\widehat{p})\} \geq 5$ and $n > 40$, we can use the Agresti-Coull CI for p:

   $\widehat{p} = 22/100 = 0.22$ and so $\tilde{X} = 22 + (2.58^2)/2 = 25.3282$, and $\tilde{n} = 100 + 2.58^2 = 106.6564$.

   Thus, $\tilde{p} = \tilde{X}/\tilde{n} = 0.2375$ and the 99% Agresti-Coull CI on p is

$$0.2375 \pm (2.58)\sqrt{\frac{(0.2375)(0.7625)}{106.6564}} = (0.1312, 0.3438).$$

- Alternatively, from the Kaplan-Meier Estimator:

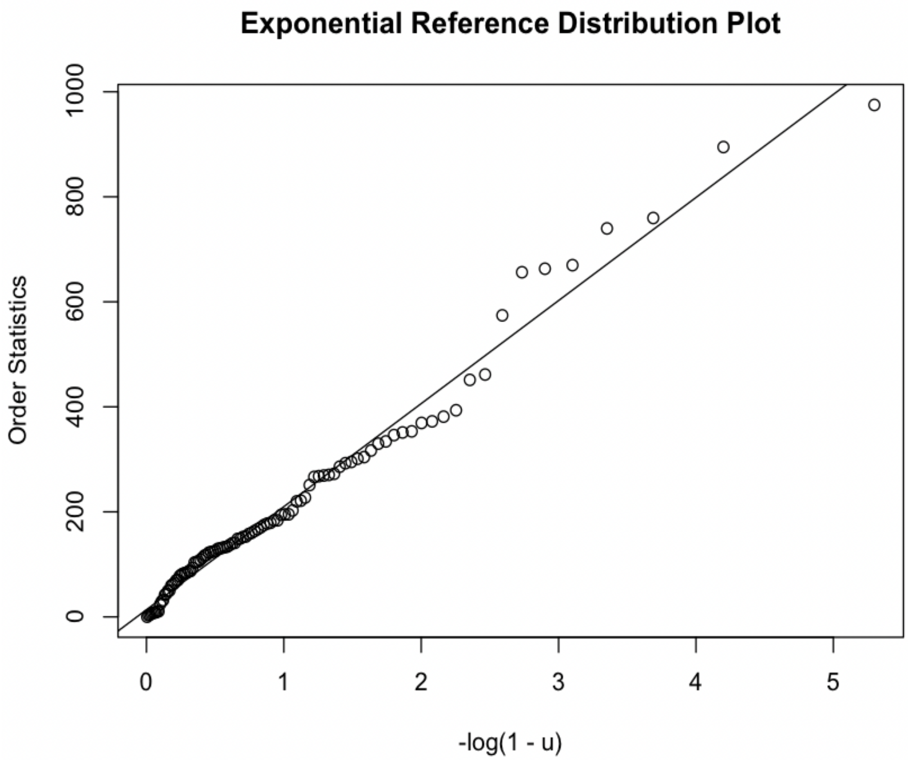
```
L = c(data)
n =  length(L)
wc = c(rep(1,100))
cordsurv = survfit(Surv(L,wc)~1,conf.type="log-log")
summary(cordsurv)
print(cordsurv,print.rmean=TRUE)
Call: survfit(formula = Surv(L, wc) ~ 1, conf.type = "log-log")

   time n.risk n.event survival std.err lower 95% CI upper 95% CI
  295.10     23       1     0.22 0.04142     0.144852       0.3053
  301.10     22       1     0.21 0.04073     0.136546       0.2943
```

   Using linear interpolation, $\hat{p} = .212$ with 95% C.I. (.138, .296)

2. From the following quantile plot, the exponential model appears to fit

   AD=2.115 which yields $.005 < p - value < .010$. This indicates a poor fit of an exponential model with $\hat{\beta} = \bar{T} = 209.1838$.



i. Using the results for an exponential distribution, a 95/99 lower tolerance bound is

$$L_{0.95,0.99} == -\hat{\beta}\left[\frac{2n}{\chi^2_{1-.99}}\right] log(.95) = -(209.1838)\left[\frac{200}{156.432}\right](\log 0.95) = 13.718.$$

Therefore, a 95/99 lower tolerance interval on the average lifetime of the epoxy strands is $(13.718,\ \infty)$.

ii. A distribution-free 95/99 lower tolerance interval for n=100 has m=1 from Table on page 51 in Handout 11.

Thus, the distribution-free lower tolerance interval would be $(Y_{(1)},\ \infty)\ =\ (.18,\ \infty)$.

With n=100, the distribution-free interval is not very informative.

3. Using the methodology for an exponential distribution, a 95% PI for Y_{101} is given by

$$(\bar{Y}F_{0.025,2,200},\ \bar{Y}F_{0.975,2,200})\ =\ ((209.1838)(0.0253),\ (209.1838)(3.758))\ =\ (5.3,\ 786.1).$$

Problem P5. (20 points) Strength of Braided Cord:

1. The parameter of interest is $p = P[S < 50]$. From the data, $\hat{p} = \frac{51}{56} = .91$, $n \cdot min(\hat{p}, 1-\hat{p}) = 5.1 > 5$, and $n = 56 > 40 \quad \Rightarrow$ Use Agresti-Coull C.I.

$\tilde{X} = 51 + (1.96^2)/2 = 25.3282$, and $\tilde{n} = 56 + 1.96^2 = 59.8416$.

Thus, $\tilde{p} = \tilde{X}/\tilde{n} = 0.8843$ and the 95% Agresti-Coull CI on p is

$$0.8843 \pm (1.96)\sqrt{\frac{(0.8843)(1-.8843)}{59.8416}}\ =\ (0.803,\ 0.965).$$

Alternatively, the Clopper-Pearson C.I. would be obtained from y=51, n=56,

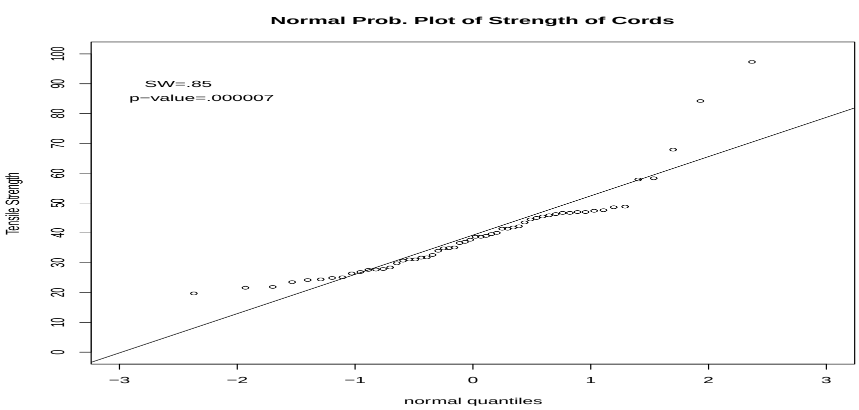
$$P_L = \frac{1}{1 + \left(\frac{6}{51}\right) F_{12,102,.025}} = \frac{1}{1 + \left(\frac{6}{51}\right)(2.074684)} = .804$$

$$P_U = \frac{\left(\frac{52}{5}\right) F_{104,10,.025}}{1 + \left(\frac{52}{5}\right) F_{104,10,.025}} = \frac{\left(\frac{52}{5}\right)(3.149015)}{1 + \left(\frac{52}{5}\right)(3.149015)} = .970$$

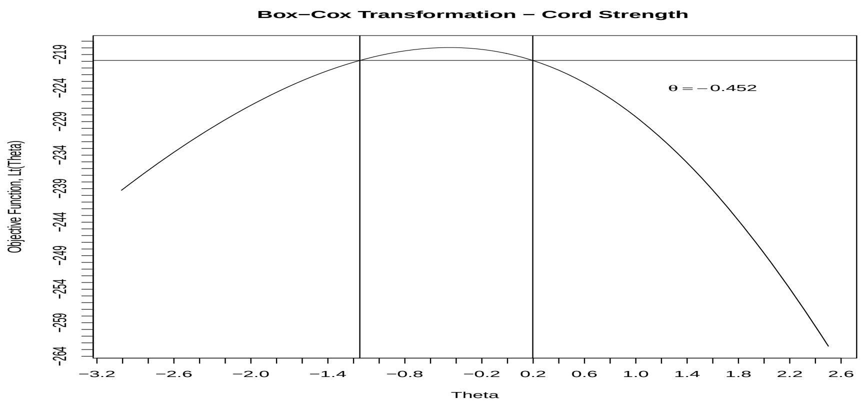
Clopper-Pearson C.I is (.804, .970)

- The two C.I.'s are in close agreement: (0.803, 0.965) vs (.804, .970) which agree to the second decimal place.
- However, the Wald C.I. is $0.911 \pm 1.96\sqrt{(.911)(1-.991)/56} = (.836,\ .985)$ which is considerable different from the other two C.I.s.

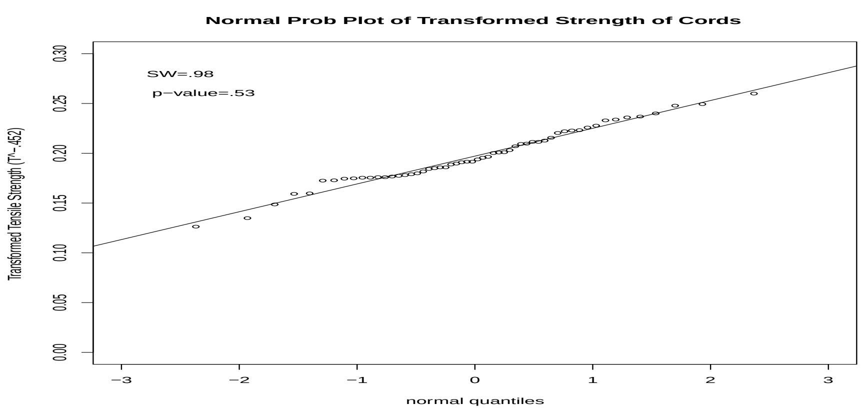
2. The normal probability plot and a value for the Shapiro-Wilk's test of W = .853 which yields a p-value of 6.915e-06 indicate that a normal distribution would not be a good model for data.



As was done in P1-3, a Box-Cox transformation will be obtained and a CI will be placed on the median of the transformed which will be inverted to obtain a CI for the median for the cord strength.



From the Box-Cox methodology, the transformation $X = Y^{-.452}$ yields a p-value of .52 from the Shapiro-Wilk test which indicates an excellent fit of a normal distribution to the data. The following normal reference distribution plot confirms the fit:



In the normal distribution, the mean and median are the same parameter so we can use the formula for a CI on the mean of a normally distributed population for the transformed data:

$$\bar{X} \pm t_{55,0.025}\frac{S_x}{\sqrt{n}} = .1971 \pm (1.673)\frac{.02813}{\sqrt{56}} = (0.1908111,\ 0.2033889).$$

Thus, we have that $((0.2033889)^{-1/.452},\ (0.1908111)^{-1/.452}) = (33.90,\ 39.05)$ is a 95% C.I. on the median of the cord strengths.

- Alternatively, a nonparametric C.I. will be constructed.

Using Table VII.3 in Handout 11 or the R code on page 31 of Handout 11, we obtain k=21, therefore a 95% C.I. on Median is

$[Y_{(21)},\ Y_{(56-21+1)}] = [Y_{(21)},\ Y_{(36)}] = (32.6,\ 41.8)$ with a coverage of 95.6%.

The two CI's are in somewhat agreement: (33.90, 39.05) vs (32.6, 41.8)

3. Using the transformation: $X = Y^{-.452}$, the Tolerance Interval will be obtained using the normal based procedures:

   A (.90, .95) T.I. for the distribution of X is given by

   $$\bar{X} \pm K_{0.90,0.95} S_X = 0.1971307 - (1.98)(0.02813393) = (0.1414255,\ 0.2528359)$$

   Inverting the endpoints of the T.I. for the X distribution yields the following T.I. for the strength distribution:

   $$((0.2528359)^{-1/.452},\ (0.1414255)^{-1/.452}) = (20.95,\ 75.75)$$

- Alternatively, a Nonparametric T.I. is obtained as follows. From the Table in Handout 11 for Nonparametric Tolerance Interval, we obtain for n=56 that m=2 which implies that

  the $(P, \gamma) = (.9, .95)$ tolerance interval is $[Y_{(1)}, Y_{(56)}] = [19.7,\ 97.3]$

- Note that the distribution-free T.I. is considerable wider than the parametric T.I.:

  [19.7, 97.3] versus (20.95, 75.75). With n=56, the distribution-free T.I. is not very informative in that it uses the minimum and maximum values in the data set to provide the specified coverage.

Problem P6. (3 points each) Multiple Choice:

1. **- D** The transformation to normality yields an excellent fit therefore just invert the endpoints to obtain the requested Tolerance Interval.

2. **- D** The bootstrap is often used when there is not an appropriate parametric procedure.

3. **- C** $n = \hat{\sigma}^2\ Z^2_{\alpha/2}/\Delta^2 = ((30)(2.576)/10)^2 = 59.7$. Therefore, take n=60.

4. **- B**

5. **- B** See page 60 in Handout 11

6. **- C** If the data are highly correlated, then S^2/n underestimates $Var(\bar{X})$. This results in a C.I. for μ which is too narrow and hence has a smaller coverage probability than the nominal level of confidence.

7. **- C or D** See page 53 in Handout 11.

8. **- D or E** If the population distribution is highly right skewed, the sampling distribution of $(n-1)S^2/\sigma^2$ does not have a chi-squared distribution. In fact, the distribution will be more right skewed than the chi-squared distribution. This will result in a C.I. having lower coverage probability than the stated level of confidence. The asymptotic results would not hold for n=10 observations. There are no distribution-free methods for construction a C.I. for σ. Transformations are generally not advisable for constructing C.I.'s for μ or σ. The bootstrap C.I. for σ would not be very useful based on only n=10 observations from a population with a highly skewed distribution. Probably the best option is to tell the researcher that the sample size needs to be increased.

```
####
#### (1)
####

x_y <- c(151, 450, 124, 235, 357, 110, 302, 671, 118, 115, 275, 275, 2550, 243, 201, 199,
  130, 119, 92, 91, 92, 98, 1650, 1200, 1180, 900, 700, 460, 340, 330)
x_n <- c(246, 268, 275, 348, 305, 311, 206, 279, 426, 269, 257, 299, 337, 329, 319, 312,
  327, 342, 351, 205, 151, 426, 154, 353, 396, 441, 254, 263, 278, 281)
n_y <- length(x_y)
n_n <- length(x_n)

##
## Are data normally distributed?
##

## Normal quantile plot and Shapiro-Wilks test for x_y
x_y <- sort(x_y)
u <- (1:n_y - 0.5) / n_y
z <- qnorm(u)

plot(z, x_y, main = "Normal Prob Plot of Rainfall With Seeding",
  xlab = "Normal Quantiles", ylab = "Amount of Rainfall", ylim = c(50, 1800),
  xlim = c(-3, 3))
abline(lm(x_y ~ z))

shapiro.test(x_y)
text(-2.5, 1700, "SW = 0.68")
text(-2.2, 1600, "p-value = 7.7e-07")

## Normal quantile plot and Shapiro-Wilks test for x_n
x_n <- sort(x_n)
u <- (1:n_n - 0.5) / n_n
z <- qnorm(u)

plot(z, x_n, main = "Normal Prob Plot of Rainfall Without Seeding",
  xlab = "Normal Quantiles", ylab = "Amount of Rainfall", ylim = c(100, 500),
  xlim = c(-3, 3))
abline(lm(x_n ~ z))

shapiro.test(x_n)
text(-2.5, 470, "SW = 0.97")
text(-2.3, 450, "p-value = 0.54")

##
## Box-Cox transformation for With Seeding data
##

l <- 0
theta_seq <- seq(-3, 3, by = 0.001)

y_y <- log(x_y)
s_0 <- sum(y_y)
v_0 <- var(y_y)
for(i in 1:length(theta_seq)) {
```

```
  if(abs(theta_seq[i]) < 1e-10) {
    l[i] <- -s_0 - (n_y / 2) * (log(2 * pi * v_0) + 1)
    theta_seq[i] <- 0
  } else {
    x_theta <- (x_y ^ theta_seq[i] - 1) / theta_seq[i]
    v_1 <- var(x_theta)
    l[i] <- (theta_seq[i] - 1) * s_0 - (n_y / 2) * (log(2 * pi * v_1) + 1)
  }
}

i_max <- which.max(l)
theta_seq[i_max]

plot(theta_seq, l, xlab = "theta", ylab = "L(theta)", type = "l")
abline(v = theta_seq[i_max])

## 95% CI for theta
which_ci <- (1:length(theta_seq))[l[i_max] - l <= 0.5 * qchisq(0.95, 1)]
theta_ci <- theta_seq[c(min(which_ci), max(which_ci))]
abline(v = theta_ci[1])
abline(v = theta_ci[2])

## Normal quantile plot of transformed With Additive data
tx_y <- x_y ^ -0.391
tx_y <- sort(tx_y)
u <- (1:n_y - 0.5) / n_y
z <- qnorm(u)

plot(z, tx_y, main = "Normal Prob Plot of Transformed Rainfall With Seeding",
  xlab = "Normal Quantiles", ylab = "(Rainfall Amount)^-0.391", ylim = c(0, 0.2),
  xlim = c(-3, 3))
abline(lm(tx_y ~ z))

shapiro.test(tx_y)
text(-2.3, 0.19, "SW = 0.9446")
text(-2.2, 0.18, "p-value = 0.1209")

##
## (1)
##

K <- 2.220

## For seeded data, create upper bound for transformed data, then back-transform to
## original scale to get lower bound.
U_y <- mean(tx_y) + K * sd(tx_y)
L_y <- U_y ^ (-1 / 0.391)

## Unseeded data (already Normally distributed).
L_n <- mean(x_n) - K * sd(x_n)

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

```
## Studentized bootstrap for With Additive data
th_est_D <- mean(x_y)
sd_est_D <- sd(x_y)

a <- (1 - 0.95) / 2
B <- 9999

th_est_B <- sd_est_B <- Z_B <- numeric(B)
for(b in 1:B) {
  x_B <- sample(x_y, replace = TRUE)
  th_est_B[b] <- mean(x_B)
  sd_est_B[b] <- sd(x_B)
  Z_B[b] <- sqrt(n_y) * (th_est_B[b] - th_est_D) / sd_est_B[b]
}

Z_B <- sort(Z_B)
L_Z <- Z_B[(B + 1) * a]
U_Z <- Z_B[(B + 1) * (1 - a)]
th_L <- th_est_D - U_Z * sd_est_D / sqrt(n_y)
th_U <- th_est_D - L_Z * sd_est_D / sqrt(n_y)

th_L
th_U

## Without Additive data already Normally distributed, so just use t-based interval
mean(x_n) + c(-1, 1) * qt(0.975, n_n - 1) * sd(x_n) / sqrt(n_n)

## For illustration purposes, here is the studentized bootstrap CI
th_est_D <- mean(x_n)
sd_est_D <- sd(x_n)

a <- (1 - 0.95) / 2
B <- 9999

th_est_B <- sd_est_B <- Z_B <- numeric(B)
for(b in 1:B) {
  x_B <- sample(x_n, replace = TRUE)
  th_est_B[b] <- mean(x_B)
  sd_est_B[b] <- sd(x_B)
  Z_B[b] <- sqrt(n_n) * (th_est_B[b] - th_est_D) / sd_est_B[b]
}

Z_B <- sort(Z_B)
L_Z <- Z_B[(B + 1) * a]
U_Z <- Z_B[(B + 1) * (1 - a)]
th_L <- th_est_D - U_Z * sd_est_D / sqrt(n_n)
th_U <- th_est_D - L_Z * sd_est_D / sqrt(n_n)

th_L
th_U

##
## (3)
```

```
##

## Apply t-based interval to transformed With Additive data to get an interval on its
## median, then back-transform.
med_interval_ty <- mean(tx_y) + c(-1, 1) * qt(0.975, n_y - 1) * sd(tx_y) / sqrt(n_y)
med_interval_y <- c(med_interval_ty[2] ^ (-1 / 0.391), med_interval_ty[1] ^ (-1 / 0.391))

## Again, t-based interval for Without Additive data
mean(x_n) + c(-1, 1) * qt(0.975, n_n - 1) * sd(x_n) / sqrt(n_n)

## Distribution-free CIs
L <- 0.95
P <- 0.5

f_df_CI <- function(n, L, P) {
  s <- ceiling(n * P) - 1
  r <- floor(n * P) + 1
  cov <- 0

  while(s < n - 1 && r > 1 && cov < L) {
    s <- s + 1
    cov <- pbinom(s - 1, n, P) - pbinom(r - 1, n, P)
    if(cov >= L)
      break;
    r <- r - 1
    cov <- pbinom(s - 1, n, P) - pbinom(r - 1, n, P)
  }

  return(list("r" = r, "s" = s, "cov" = cov))
}

f_df_CI(n_y, L, P)
f_df_CI(n_n, L, P)

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

library(MASS)
library(survival)

x <- c(2.526, 2.546, 2.628, 2.669, 2.869, 2.710, 2.731, 2.751, 2.771, 2.772, 2.782,
  2.789, 2.793, 2.834, 2.844, 2.854, 2.875, 2.876, 2.895, 2.916, 2.919, 2.957, 2.977,
  2.988, 3, 3, 3, 3)
n <- length(x)
delta <- c(rep(1, n - 4), rep(0, 4))

##
## (1)
##

## Asymptotic CI based on survfit
surv_fit <- survfit(Surv(x, delta) ~ 1, conf.type = "log-log")
summary(surv_fit)
```

```
print(surv_fit, print.rmean = TRUE)

2.8311 + c(-1, 1) * 1.96 * 0.0249

## Studentized bootstrap
theta_D <- mean(x)
SD_D <- sd(x)
alpha <- 0.025
B <- 9999

Z_B <- theta_B <- SD_B <- numeric(B)
for(b in 1:B) {
  x_B <- sample(x, replace = TRUE)
  theta_B[b] <- mean(x_B)
  SD_B[b] <- sd(x_B)
  Z_B[b] <- sqrt(n) * (theta_B[b] - theta_D) / SD_B[b]
}
Z_B <- sort(Z_B)
Z_L <- Z_B[(B + 1) * alpha]
Z_U <- Z_B[(B + 1) * (1 - alpha)]
CI_L <- theta_D - SD_D * Z_U / sqrt(n)
CI_U <- theta_D - SD_D * Z_L / sqrt(n)

##
## (2)
##

## Weibull reference distribution plot, only using the uncensored values
y <- log(x)[1:24]
y <- sort(y)
u <- (1:24 - 0.5) / n
Q_u <- log(-log(1 - u))
plot(Q_u, u, xlab = "Q = log(-log(1 - u))", ylab = "y = log(x)",
  main = "Weibull Reference Plot - Stress to Failure")
abline(lm(u ~ Q_u))

## Anderson-Darling GOF test, only using the uncensored values
weib_fit <- fitdistr(x[1:24], "weibull")
U <- pweibull(sort(x[1:24]), shape = weib_fit$est[1], scale = weib_fit$est[2])
AD <- -24 - (1 / 24) * sum((2 * (1:24) - 1) * log(U)) -
  (1 / 24) * sum((2 * 24 + 1 - 2 * (1:24)) * log(1 - U))
AD_adj <- AD * (1 + 0.2 / sqrt(24))

## Parametric bootstrap. We will use the censored values here and survreg to estimate the
## Weibull parameters.
surv_reg <- survreg(Surv(x, delta) ~ 1, dist = "weibull")
summary(surv_reg)
names(surv_reg)
lambda_D <- exp(surv_reg$coef)
k_D <- 1 / surv_reg$scale

mean_D <- lambda_D * gamma(1 + 1 / k_D)
var_D <- lambda_D ^ 2 * (gamma(1 + 2 / k_D) - (gamma(1 + 1 / k_D)) ^ 2)
```

```
B <- 9999
k_B <- lambda_B <- mean_B <- var_B <- Z_B <- numeric(B)
for(b in 1:B) {
  x_B <- rweibull(n, shape = k_D, scale = lambda_D)

  ## Create censoring
  x_B[x_B >= 3] <- 3
  delta_B <- rep(1, n)
  delta_B[x_B == 3] <- 0

  surv_reg_B <- survreg(Surv(x_B, delta_B) ~ 1, dist = "weibull")
  lambda_B[b] <- exp(surv_reg_B$coef)
  k_B[b] <- 1 / surv_reg_B$scale

  mean_B[b] <- lambda_B[b] * gamma(1 + 1 / k_B[b])
  var_B[b] <- lambda_B[b] ^ 2 * (gamma(1 + 2 / k_B[b]) - (gamma(1 + 1 / k_B[b])) ^ 2)
  Z_B[b] <- sqrt(n) * (mean_B[b] - mean_D) / sqrt(var_B[b])
}
Z_B <- sort(Z_B)
Z_L <- Z_B[250]
Z_U <- Z_B[9750]
mean_L <- mean_D - sqrt(var_D) * Z_U / sqrt(n)
mean_U <- mean_D - sqrt(var_D) * Z_L / sqrt(n)

####
#### (3)
####

n <- 200
lambda_hat <- 27.7

CI <- lambda_hat + (1 / (2 * n)) * 1.96 ^ 2 + c(-1, 1) * 1.96 *
  sqrt((1 / n) * lambda_hat + (1 / (4 * n ^ 2)) * 1.96 ^ 2)

####
#### (4)
####

x <- c(.18, 3.1, 4.2, 6.0, 7.5, 8.2, 8.5, 10.3, 10.6, 24.2, 29.6, 31.7, 41.9, 44.1, 49.5,
  50.1, 59.7, 61.7, 64.4, 69.7, 70.0, 77.8, 80.5, 82.3, 83.5, 84.2, 87.1, 87.3, 93.2,
  103.4, 104.6, 105.5, 108.8, 112.6, 116.8, 118.0, 122.3, 123.5, 124.4, 125.4, 129.5,
  130.4, 131.6, 132.8, 133.8, 137.0, 140.2, 140.9, 148.5, 149.2, 152.2, 152.9, 157.7,
  160.0, 163.6, 166.9, 170.5, 174.9, 177.7, 179.2, 183.6, 183.8, 194.3, 195.1, 195.3,
  202.6, 220.0, 221.3, 227.2, 251.0, 266.5, 267.9, 269.2, 270.4, 272.5, 285.9, 292.6,
  295.1, 301.1, 304.3, 316.8, 329.8, 334.1, 346.2, 351.2, 353.3, 369.3, 372.3, 381.3,
  393.5, 451.3, 461.5, 574.2, 656.3, 663.0, 669.8, 739.7, 759.6, 894.7, 974.9)
n <- length(x)

X <- sum(x >= 300)
p_hat <- X / n

##
## (1)
##
```

```
## Agresti-Coull CI
X_tilde <- X + 0.5 * qnorm(0.995) ^ 2
n_tilde <- n + qnorm(0.995) ^ 2
p_tilde <- X_tilde / n_tilde

CI <- p_tilde + c(-1, 1) * qnorm(0.995) * sqrt(p_tilde * (1 - p_tilde) / n_tilde)

## Kaplan-Meier CI. Linear interpolation of CIs for survival times 295.1 and 301.1.
delta <- rep(1, n)
surv_fit <- survfit(Surv(x, delta) ~ 1, conf.type = "log-log")
summary(surv_fit)

plot(c(0.1449, 0.1365), c(0.3053, 0.2943))
lines(c(0.1449, 0.1365), c(0.3053, 0.2943))

w <- (300 - 295.1) / (301.1 - 295.1)
LO <- 0.1449 * (1 - w) + 0.1365 * w
HI <- 0.3053 * (1 - w) + 0.2943 * w
points(LO, HI, pch = 20)

##
## (2)
##

## Exponential distribution reference plot
u <- (1:n - 0.5) / n
Q <- -log(1 - u)
x_sort <- sort(x)

plot(Q, x_sort, xlab = "-log(1 - u)", ylab = "Order Statistics",
  main = "Exponential Reference Distribution Plot")
abline(lm(x_sort ~ Q))

## Anderson-Darling GOF statistic
beta_hat <- mean(x)
U <- pexp(x_sort, 1 / beta_hat)

AD <- -n - (1 / n) * sum((2 * (1:n)) * log(U)) -
  (1 / n) * sum((2 * n + 1 - 2 * (1:n)) * log(1 - U))
AD_adj <- AD * (1 + 0.6 / n)

## Lower tolerance bound assuming exponential distribution
K <- 2 * n / qchisq(0.01, 2 * n)
W <- -beta_hat * K * log(0.95)

##
## (3)
##

LO <- beta_hat * qf(0.025, 2, 2 * n)
HI <- beta_hat * qf(0.975, 2, 2 * n)

####
```

```
#### (5)
####

x <- c(19.7, 21.6, 21.9, 23.5, 24.2, 24.4, 24.9, 25.1, 26.4, 26.9, 27.6, 27.7, 27.9,
  28.4, 29.8, 30.7, 31.1, 31.1, 31.7, 31.8, 32.6, 34.0, 34.8, 34.9, 35.1, 36.6, 37.0,
  37.7, 38.7, 38.7, 39.0, 39.6, 40.0, 41.4, 41.4, 41.8, 42.2, 43.5, 44.5, 45.0, 45.5,
  45.9, 46.3, 46.7, 46.7, 47.0, 47.0, 47.4, 47.6, 48.6, 48.8, 57.9, 58.3, 67.9, 84.2,
  97.3)
n <- length(x)

##
## (1)
##

X <- sum(x < 50)
p_hat <- X / n

## Agresti-Coull CI
X_tilde <- X + 0.5 * qnorm(0.975) ^ 2
n_tilde <- n + qnorm(0.975) ^ 2
p_tilde <- X_tilde / n_tilde

CI <- p_tilde + c(-1, 1) * qnorm(0.975) * sqrt(p_tilde * (1 - p_tilde) / n_tilde)

##
## (2)
##

## If the data are Normally distributed, can just compute a CI for the mean, since it is
## the same parameter as the median in that case.
x_sort <- sort(x)
u <- (1:n - 0.5) / n
Q <- qnorm(u)

plot(Q, x_sort, xlab = "Normal Quantiles", ylab = "Sample Quantiles",
  main = "Normal Reference Distribution Plot\nRaw Data")
abline(lm(x_sort ~ Q))

shapiro.test(x)

## Do Box-Cox transformation
l <- 0
theta_seq <- seq(-3, 3, by = 0.001)

y <- log(x)
s_0 <- sum(y)
v_0 <- var(y)
for(i in 1:length(theta_seq)) {
  if(abs(theta_seq[i]) < 1e-10) {
    l[i] <- -s_0 - (n / 2) * (log(2 * pi * v_0) + 1)
    theta_seq[i] <- 0
  } else {
    x_theta <- (x ^ theta_seq[i] - 1) / theta_seq[i]
    v_1 <- var(x_theta)
```

```
    l[i] <- (theta_seq[i] - 1) * s_0 - (n / 2) * (log(2 * pi * v_1) + 1)
  }
}

i_max <- which.max(l)
theta_seq[i_max]

plot(theta_seq, l, xlab = "theta", ylab = "L(theta)", type = "l")
abline(v = theta_seq[i_max])

## 95% CI for theta
which_ci <- (1:length(theta_seq))[l[i_max] - l <= 0.5 * qchisq(0.95, 1)]
theta_ci <- theta_seq[c(min(which_ci), max(which_ci))]
abline(v = theta_ci[1])
abline(v = theta_ci[2])

## Transformed data
y <- x ^ (-0.452)
shapiro.test(y)

plot(Q, sort(y), xlab = "Normal Quantiles", ylab = "Sample Quantiles",
  main = "Normal Reference Distribution Plot\nTransformed Data")
abline(lm(sort(y) ~ Q))

## Normal-based CI on mean / median
CI_y <- mean(y) + c(-1, 1) * qt(0.975, n - 1) * sd(y) / sqrt(n)
CI_x <- c(CI_y[2] ^ (-1 / 0.452), CI_y[1] ^ (-1 / 0.452))

## Nonparametric CI
L <- 0.95
P <- 0.50
s <- ceiling(n * P) - 1
r <- floor(n * P) + 1
cov <- 0

while(s < n - 1 && r > 1 && cov < L) {
  s <- s + 1
  cov <- pbinom(s - 1, n, P) - pbinom(r - 1, n, P)
  if(cov >= L)
    break;
  r <- r - 1
  cov <- pbinom(s - 1, n, P) - pbinom(r - 1, n, P)
}

r
s
x_sort[c(r, s)]
cov

##
## (3)
##

## Tolerance interval on Normal transformed data
```

```
mu_hat <- mean(y)
sd_hat <- sd(y)

## Linear interpolation of constant K
w <- (56 - 50) / (60 - 50)
K <- 1.999 * (1 - w) + 1.960 * w

CI_y <- mu_hat + c(-1, 1) * K * sd_hat

## Back-transform to original units
CI_x <- c(CI_y[2] ^ (-1 / 0.452), CI_y[1] ^ (-1 / 0.452))
```