

Advanced Data Analysis HW6

Ao Liu, al3472

1.

A random variable T is said to have a Weibull distribution if its survival function is given by $S(t) = e^{-(\alpha t)^\beta}$ where $\alpha > 0$ and $\beta > 0$.

(a)

Find the density, $f_T(t)$ of T

Answer:

$$f_T(t) = -\frac{dS(t)}{dt} = \beta\alpha^\beta t^{\beta-1} e^{-(\alpha t)^\beta}$$

(b)

Find the hazard function $\lambda(t)$ of T

Answer:

$$\lambda(t) = \frac{f(t)}{S(t)} = \beta\alpha^\beta t^{\beta-1}$$

(c)

Show that

$$\log(-\log(S(t))) = \beta \log(\alpha) + \beta \log(t)$$

Based on this, describe a graphical method for checking whether or not the data is from a Weibull distribution.

Answer:

$$\begin{aligned} \log(-\log(S(t))) &= \log((\alpha t)^\beta) \\ &= \beta \log(\alpha t) \\ &= \beta \log(\alpha) + \beta \log(t) \end{aligned}$$

Here we get a graphical method for checking whether or not the data is from a Weibull distribution: By plotting all the $(\log(t), \log(-\log(S(t))))$, the points from the same Weibull distribution should lie in a straight line.

(d)

Consider the following data

143, 164, 188, 188, 190, 192, 206, 209, 213, 216, 220, 227, 230, 234, 246, 265, 304

and use as an estimate of $S(t(i))$

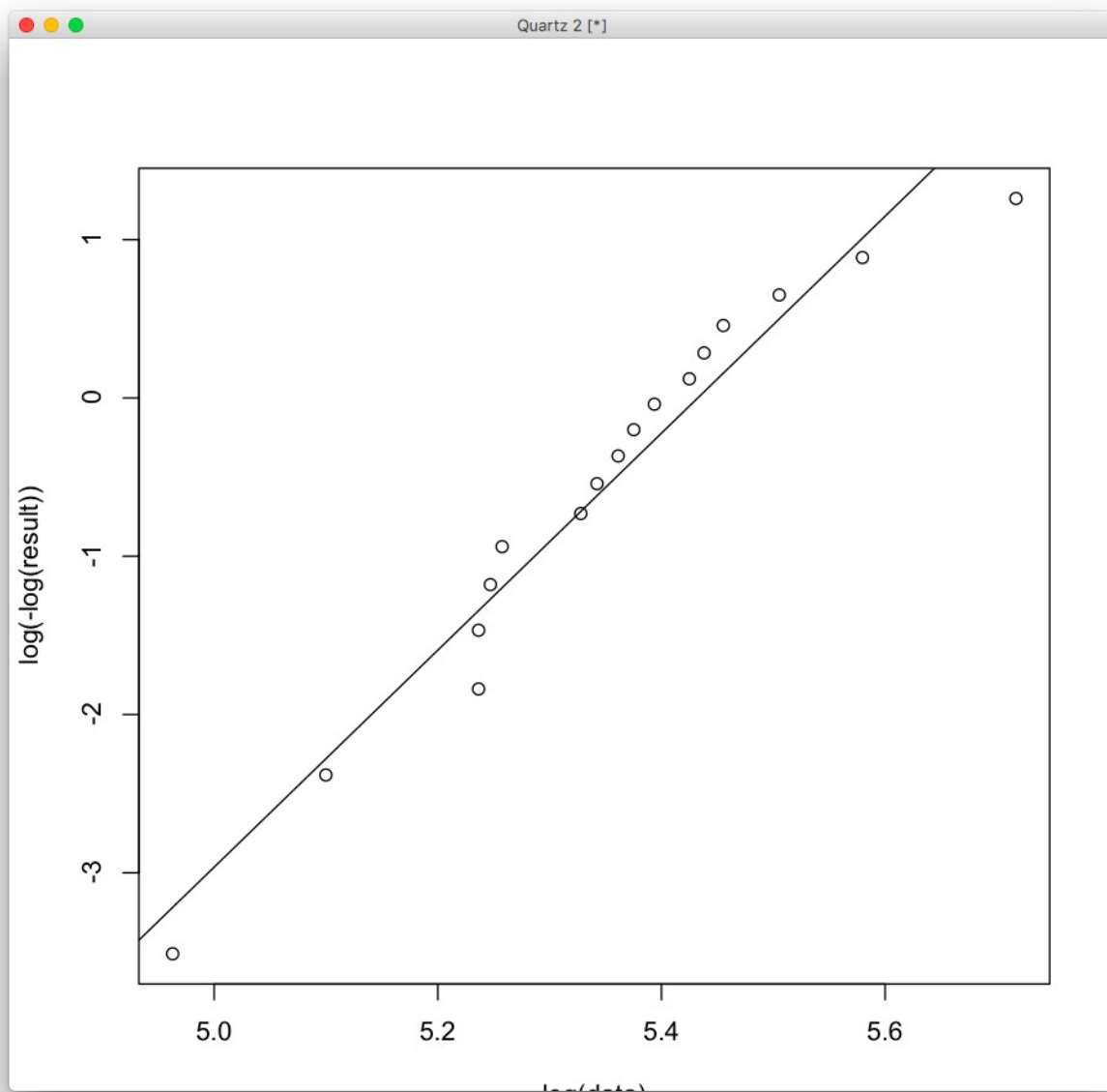
$$S(t(i)) = 1(i - 0.5)/n$$

where $t(i)$ is the i th ordered value and n is the sample size. Use the graphical technique in the previous question to check if a Weibull distribution is appropriate for these data

Answer:

Follow the technique in (c), we plot all the $(\log(t_i), \log(-\log(\hat{S}(t_i))))$ in an axis:

```
1 > data = c(143, 164, 188, 188, 190, 192, 206, 209, 213, 216, 220, 227, 230, 234, 246,  
2         265, 304)  
2 > result = c()  
3 > n = length(data)  
4 > for (i in 1:17){  
5     result[i] = 1-(i-0.5)/n  
6 }  
7 > plot(log(-log(result))~log(data))  
8 > abline(lm(log(-log(result))~log(data)))
```



We can tell from the plot that a Weibull distribution is appropriate for these data.

(e)

Assume that the Weibull distribution is a good fit, use least squares approach to estimate its parameters.

Answer:

```
1 > fit = lm(log(-log(result))~log(data))
2 > abline(fit)
3 > summary(fit)
```

```
1 Call:
2 lm(formula = log(-log(result)) ~ log(data))
3
4 Residuals:
5      Min       1Q   Median       3Q      Max
6 -0.68997 -0.12226  0.09174  0.19153  0.30116
7
8 Coefficients:
9             Estimate Std. Error t value Pr(>|t|)
10 (Intercept)  -37.2330     2.1806  -17.07 3.08e-11 ***
11 log(data)     6.8538     0.4073   16.83 3.80e-11 ***
12 ---
13 Signif. codes:  0      ***      0.001      **      0.01      *      0.05      .      0.1      1
14
15 Residual standard error: 0.2871 on 15 degrees of freedom
16 Multiple R-squared:  0.9497, Adjusted R-squared:  0.9463
17 F-statistic: 283.1 on 1 and 15 DF, p-value: 3.796e-11
```

Assume that the Weibull distribution is a good fit, by using least squares approach, we have the following estimation for its parameters:

$$\beta = 6.8538$$

$$\alpha = e^{\frac{-37.2330}{\beta}} = 0.0044$$

2.

The data below show survival times in months of patients with Hodgkins disease who were treated with nitrogen mustard. Group A patients received little or no prior therapy whereas Group B patients received heavy prior therapy. Starred are observations are censoring times.

GroupA :1.25, 1.41, 4.98, 5.25, 5.38, 6.92, 8.89, 10.98, 11.18, 13.11, 13.21, 16.33, 19.77, 21.08, 21.84*, 22.07, 31.38, 32.61*, 37.18*, 42.92

GroupB :1.05, 2.92, 3.61, 4.20, 4.49, 6.72, 7.31, 9.08, 9.11, 14.49*, 16.85, 18.82*, 26.59*, 30.26*, 41.34*

(a)

Obtain and plot the Kaplan Meier estimates of S_A and S_B , the corresponding survival functions.

Answer:

For group A, we have:

$y_A(i)$	1.25	1.41	4.98	5.25	5.38	6.92	8.89	10.98	11.18	13.11	13.21	16.33	19.77	21.08	22.07	42.92
$d_A(i)$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
$N_A(i)$	20	19	18	17	16	15	14	13	12	11	10	9	8	7	5	1

So the Kaplan-Meier estimator of $S(t)$ is:

$$\hat{S}_A(t) = \prod_{y_A(j) \leq t} (1 - \frac{d_A(j)}{N_A(j)})$$

For group B, we have:

$y_B(i)$	1.05	2.92	3.61	4.20	4.49	6.72	7.31	9.08	9.11	16.85
$d_B(i)$	1	1	1	1	1	1	1	1	1	1
$N_B(i)$	15	14	13	12	11	10	9	8	7	5

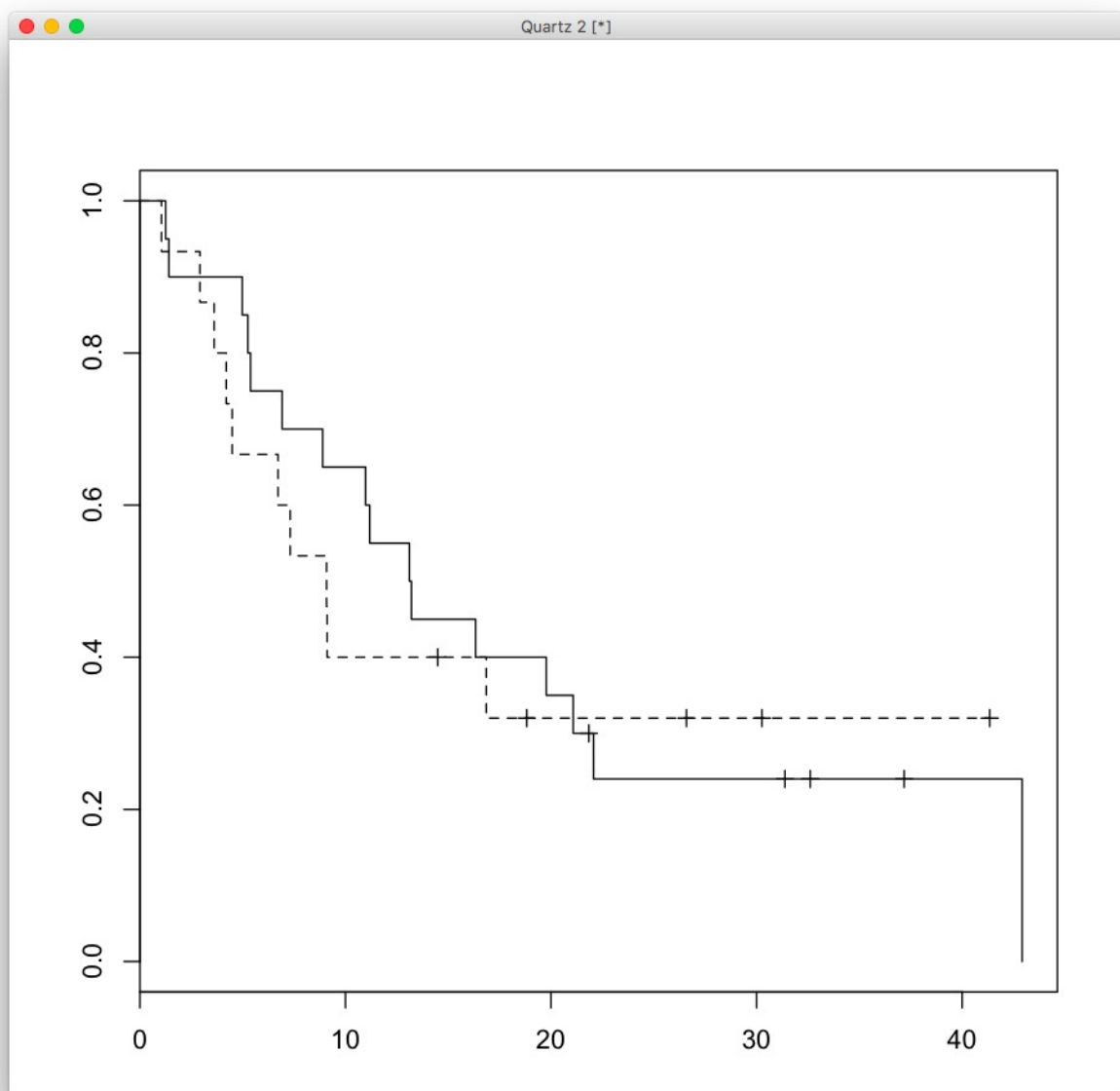
So the Kaplan-Meier estimator of $S(t)$ is:

$$\hat{S}_B(t) = \prod_{y_B(j) \leq t} \left(1 - \frac{d_B(j)}{N_B(j)}\right)$$

```

1 > library(survival)
2 > time = c(1.25, 1.41, 4.98, 5.25, 5.38, 6.92, 8.89, 10.98, 11.18, 13.11, 13.21, 16.33,
            19.77, 21.08, 21.84, 22.07, 31.38, 32.61, 37.18, 42.92, 1.05, 2.92, 3.61, 4.20,
            4.49, 6.72, 7.31, 9.08, 9.11, 14.49, 16.85, 18.82, 26.59, 30.26, 41.34)
3 > status = c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1,
              1, 1, 1, 1, 0, 1, 0, 0, 0, 0)
4 > group = c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2)
5 > fit <- survfit(Surv(time,status)~group,type = "kaplan-meier")
6 > plot(fit, lty=1:2)

```



(b)

Estimate $S_A(10)$ and $S_B(10)$ using a 95% confidence interval.

Answer:

```
1 > summary( fit )
1 Call: survfit(formula = Surv(time, status) ~ group, type = "kaplan-meier")
2
3      group=1
4  time n.risk n.event survival std.err lower 95% CI upper 95% CI
5  1.25    20      1     0.95  0.0487    0.859    1.000
6  1.41    19      1     0.90  0.0671    0.778    1.000
7  4.98    18      1     0.85  0.0798    0.707    1.000
8  5.25    17      1     0.80  0.0894    0.643    0.996
9  5.38    16      1     0.75  0.0968    0.582    0.966
10 6.92    15      1     0.70  0.1025    0.525    0.933
11 8.89    14      1     0.65  0.1067    0.471    0.897
12 10.98   13      1     0.60  0.1095    0.420    0.858
13 11.18   12      1     0.55  0.1112    0.370    0.818
14 13.11   11      1     0.50  0.1118    0.323    0.775
15 13.21   10      1     0.45  0.1112    0.277    0.731
16 16.33    9      1     0.40  0.1095    0.234    0.684
17 19.77    8      1     0.35  0.1067    0.193    0.636
18 21.08    7      1     0.30  0.1025    0.154    0.586
19 22.07    5      1     0.24  0.0980    0.108    0.534
20 42.92    1      1     0.00    NaN      NA      NA
21
22      group=2
23  time n.risk n.event survival std.err lower 95% CI upper 95% CI
24  1.05    15      1     0.933  0.0644    0.815    1.000
25  2.92    14      1     0.867  0.0878    0.711    1.000
26  3.61    13      1     0.800  0.1033    0.621    1.000
27  4.20    12      1     0.733  0.1142    0.540    0.995
28  4.49    11      1     0.667  0.1217    0.466    0.953
29  6.72    10      1     0.600  0.1265    0.397    0.907
30  7.31    9      1     0.533  0.1288    0.332    0.856
31  9.08    8      1     0.467  0.1288    0.272    0.802
32  9.11    7      1     0.400  0.1265    0.215    0.743
33 16.85    5      1     0.320  0.1239    0.150    0.684
```

According to the output from above,

$$\hat{S}_A(10) = \hat{S}_A(8.89) = 0.65$$

the 95% CI for $\hat{S}_A(10)$ is:

$$(0.471, 0.897)$$

$$\hat{S}_B(10) = \hat{S}_B(9.11) = 0.40$$

the 95% CI for $\hat{S}_B(10)$ is:

$$(0.215, 0.743)$$

(c)

Test $H_0 : S_A = S_B$ against $H_a : S_A \neq S_B$. Use $\alpha = 0.05$.

Answer:

To test the hypothesis, we do the following test in R:

```
1 > survdiff(Surv(time, status) ~ group, rho=0)
```

```

1  Call:
2  survdiff(formula = Surv(time, status) ~ group, rho = 0)
3
4           N Observed Expected (O-E)^2/E (O-E)^2/V
5 group=1  20         16   16.66    0.0261    0.0749
6 group=2  15         10    9.34    0.0466    0.0749
7
8  Chisq= 0.1  on 1 degrees of freedom, p= 0.784

```

Since the p-value is $0.784 > 0.05$, we cannot reject the Null Hypothesis that $S_A = S_B$.