

STAT 687 Homework #2 - Solution

From the text problem # 3.5

The supplier of an electromechanical control for a household appliance ran an accelerated life test on sample controls. In the test, 25 controls were put on test and run until 30 thousand cycles had been accumulated. Failures occurred at 5, 21, and 28 thousand cycles. The other 22 controls did not fail by the end of the test.

- (a) Compute and plot a nonparametric estimate for $F(t)$.

A nonparametric estimate for $F(t)$ can be easily produced using a ‘life data’ object from the SMRD package. Life data objects are created with the function `frame.to.ld` which requires the data to be formatted into a data frame.

If the data set is pre-loaded as part of the SMRD package (as is the case for the data set in this exercise), it has already been formatted into a data frame. In this case the life data object can be created by simply entering the name of the data set into `frame.to.ld`, for example

```
prob3_5.ld <- frame.to.ld(frame = prob3_5,
                          response.column = 1,
                          censor.column = 2,
                          time.units = "Kilocycles")
```

NOTE: The complete list of pre-loaded data sets in the SMRD package, can be found by running `data(package = 'SMRD')` in the R console.

A plot of the nonparametric estimate for $F(t)$ is produced by plotting the life.data object, as shown in Figure 1. This exercise does not require us to plot the confidence bands along with the estimate, so we set the `band.type` argument to “none”. The `y.range` argument is added because of the way R dynamically sets axis values for a plot. Try running following code without the `y.range` argument. This problem could easily be fixed such that the `y.range` argument would not need to be set... but the fix would cause bigger issues for other exercises.

```
plot(prob3_5.ld, band.type = "none", ylim = c(0,0.1))
```

To compute the values of $F(\hat{t}_i)$ and present them in a \LaTeX table the following code is used.

NOTE: to render the \LaTeX table you must add the chunk option `results='asis'`.

```
xtab <- xtable::xtable(print(cdfest(prob3_5.ld))$table,
                        align = c('cccccc'),
                        digits = rep(c(0,4), c(3,4)),
                        caption = "Nonparametric CDF Estimate for Problem 3.5")
print(xtab,
      include.rownames = F,
      caption.placement = 'top',
      type = switch(output, 'html' = 'html', 'pdf' = 'latex'),
      table.placement = 'h',
      comment = F)
```

- (b) Compute an approximate 95% confidence interval for the probability that an electromechanical device from the same production process, tested in the same way, would fail before 30 thousand cycles. Use the conservative binomial approach.

In Example 3.1 we see that the nonparametric estimate for $\hat{F}(30 \text{ kilocycles})$ with singly censored

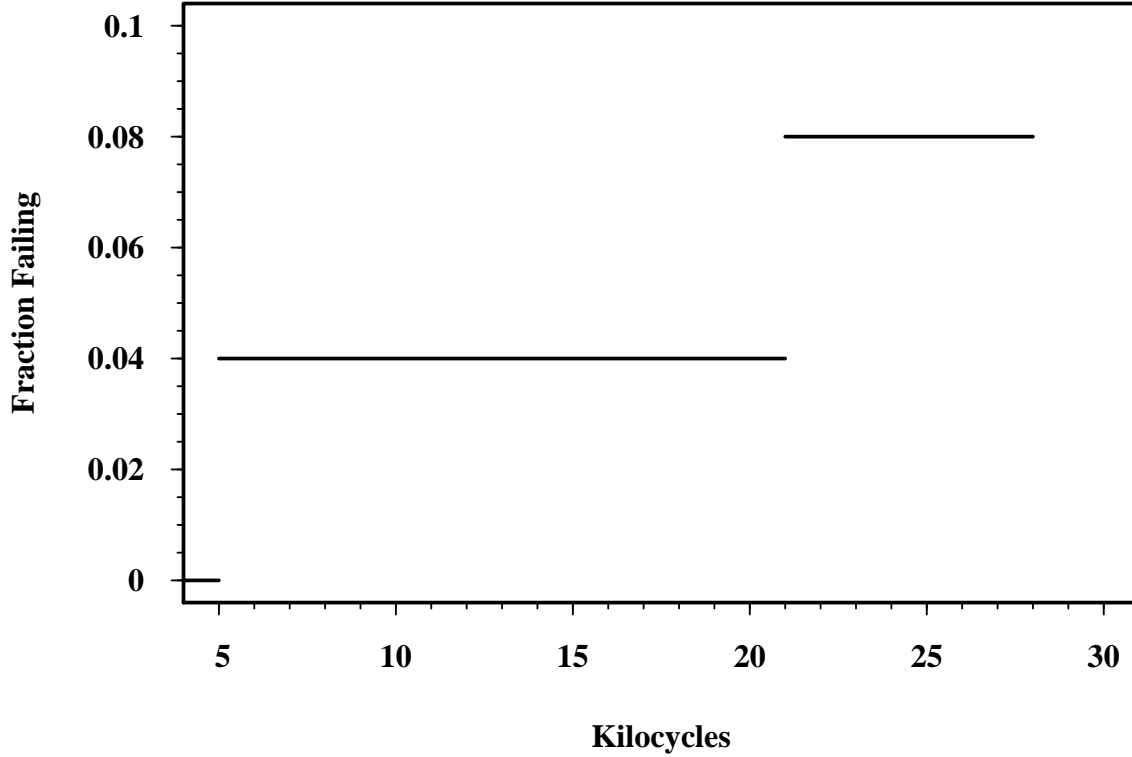


Figure 1: Plot of the nonparametric estimate of the cdf for Problem 3.5

data is

$$\frac{\# \text{ failed up to 30 kilocycles}}{\# \text{ at risk}} = \frac{3}{25}.$$

The Clopper-Pearson ‘exact’ confidence interval based on the cumulative probabilities of the binomial distribution (i.e., the exact distribution rather than an approximation) gives a lower 95% confidence bound of

$$\left(1 + \frac{(25 - 3 + 1) \times \mathcal{F}(.975, 50 - 6 + 2, 6)}{3}\right)^{-1} = 0.02546$$

and an upper 95% confidence bound of

$$\left(1 + \frac{25 - 3}{(3 + 1) \times \mathcal{F}(.925, 6 + 2, 50 - 6)}\right)^{-1} = 0.31219$$

where $\mathcal{F}(p, df_1, df_2)$ denotes the p^{th} quantile of an F distribution with df_1 and df_2 degrees of freedom.

-
- (c) Compute an approximate 95% confidence interval for the probability that an electromechanical device from the same production process, tested in the same way, would fail before 30 thousand cycles. Use the normal-approximation method based on $Z_{\hat{F}(30)} \sim NOR(0, 1)$

Table 1: Nonparametric CDF Estimate for Problem 3.5

Kilocycles-lower	Kilocycles-upper	Fhat	SE_Fhat	95% Lower	95% Upper
0	5	0.0000	0.0000	0.0000	0.0000
5	21	0.0400	0.0392	0.0056	0.2355
21	28	0.0800	0.0543	0.0201	0.2694
28	30	0.1200	0.0650	0.0392	0.3130

Using the $Z_{\hat{F}(30)} \sim NOR(0, 1)$ approximate method results in a lower 95% confidence bound of

$$\frac{3}{25} - 1.96 \times \sqrt{\frac{3}{25} \times \left(1 - \frac{3}{25}\right) \times \frac{1}{25}} = -0.007$$

and an upper 95% confidence bound of

$$\frac{3}{25} + 1.96 \times \sqrt{\frac{3}{25} \times \left(1 - \frac{3}{25}\right) \times \frac{1}{25}} = 0.247$$

-
- (d) Explain why, in this situation, the approach in part (b) would be preferred to the approach in part (c)

The CI in part (b) is conservative as it based on the exact distribution of the value of $F(t)$ (i.e. the binomial) while the method used in part (c) is an approximation of the binomial distribution. Further, the approximation in part (c) has a negative lower bound which is not possible.

- (e) The appliance manufacturer is really interested in the probability of the number of days to failure for its product. Use-rate differs from household to household, but the average rate is 2.3 cycles per day. What can the manufacturer say about the proportion of devices that would fail in 10 years of operation (the expected technological life of the product)?

We can convert usage in days to usage in cycles by making the following conversion

$$10 \text{ years} \times \left(365 \frac{\text{days}}{\text{year}}\right) \times \left(2.3 \frac{\text{cycles}}{\text{day}}\right) \approx 8.4 \times 10^3 \text{ cycles.}$$

Therefore, we want to compute a 95% CI for $F(8.4)$ where usage is measured in kilocycles. From the plot in part a) the manufacturer can say that with 95% confidence the proportion of electromechanical control failures after 10 years of use at a use rate of 2.3 cycles per day is in the interval (0,0.24)

Alternatively, if the manufacturer uses the binomial (Clopper-Pearson) approximation, the reported 95% confidence interval would be (0.01, 0.204).

If the manufacturer uses the normal approximation, the reported 95% confidence interval would be (0.01, 0.204).

Finally, if the manufacturer used R to compute the 95% confidence interval by plotting the Kaplan-Meier estimate from a “Surv” object, the reported interval would be (0, 0.114).

If the electromechanical controls experienced only one of two usage rates r_1 or r_2 , where $r_1 > r_2$ such that the lifetime of any control device follows either $F_1(t|\theta_1)$ or $F_2(t|\theta_2)$. The lifetime of the overall population would then depend on the proportion of units experiencing each usage rate denoted $\zeta_i, i = 1, 2$, where $\zeta_2 = 1 - \zeta_1$. The probability of failure for the population would then be expressed as

$$F_T(t) = \sum_{i=1}^2 \zeta_i F_i(t|\theta_i)$$

-
- (f) Refer to part (e). Describe an appropriate model to use when use-rate varies in the population of units. To simplify, start by assuming that there are only two different use-rates. Discuss, using appropriate expressions

From the text problem # 3.7

Consider the Plant 1 heat exchanger data in Figure 3.1

- (a) Write the likelihood for these data in terms of π_1, π_2, π_3 , and π_4

In terms of $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)$, the likelihood can be written as

$$L(\boldsymbol{\pi}) = C[\pi_1]^1[\pi_2]^2[\pi_3]^2[\pi_4]^{95}$$

where C is a constant that does not depend on π .

-
- (b) Write the likelihood for these data in terms of p_1, p_2, p_3 , and p_4

In terms of $\mathbf{p} = (p_1, p_2, p_3, p_4)$, the likelihood can be written as

$$L(\mathbf{p}) = C[p_1]^1[p_2(1 - p_1)]^2[p_3(1 - p_1)(1 - p_2)]^2[(1 - p_1)(1 - p_2)(1 - p_3)]^{95}$$

because $p_4 = 1$ in this case.

From the text problem # 3.18

Explain why the nonparametric estimate of $F(t)$ is a set of points for the heat exchanger data in Example 3.6 but a step-function for the shock absorber data in Example 3.9

The difference in these examples is what is known when a failure occurs. In the heat exchanger problem, failures are only known to occur during the inspections at the end of the year. If a heat exchanger pipe failed prior to the inspection, the exact failure time would be unknown. It would be incorrect to plot the CDF as a step function between the inspections, since zero, one or many failures may have occurred during the year, we just don't know when.

The battery data set lists exact failure times, so if no failures were observed between times t_1 and t_2 , we know that no units failed between these two times and it is reasonable to assume that the value of nonparametric estimate of the cdf is constant between these times.

From the text problem # 3.23

Consider the relationship $S(t_i) = \exp[-H(t_i)]$, where $H(t)$ is the cumulative hazard function. Note that a nonparametric ML estimator (based on the product-limit estimator) of $H(t)$ without assuming a distributional form is

$$\hat{H}(t_i) = -\sum_{j=1}^i \log(1 - \hat{p}_j) \approx \sum_{j=1}^i \hat{p}_j = \sum_{j=1}^i \frac{d_j}{n_j} = \hat{H}(t_i)$$

$\hat{\hat{H}}(t_i)$ is known as the Nelson-Aalen estimator of $H(t_i)$. Thus $\hat{\hat{F}}(t_i) = 1 - \exp[-\hat{\hat{H}}(t_i)]$ is another nonparametric estimator for $F(t_i)$.

-
- (a) Give conditions to assure a good agreement between $\hat{H}(t_i)$ and $\hat{\hat{H}}(t_i)$ and thus between $F(\hat{t}_i)$ and $\hat{\hat{F}}(t_i)$.

To have good agreement we must have

$$-\sum_{j=1}^i \log(1 - \hat{p}_j) \approx \sum_{j=1}^i \hat{p}_j.$$

Looking at the Taylor series expansion

$$\log(1 - \hat{p}_j) = -\sum_{n=1}^{\infty} \frac{\hat{p}_j^n}{n}, \text{ if } \hat{p}_j < 1$$

From this we can see if \hat{p}_j is small the first term in the Taylor series will accurately approximate it. Therefore we see that we'll have good agreement if these probabilities are relatively small.

-
- (b) Use the delta method to compute approximate expressions for $Var[\hat{H}(t_i)]$ and $Var[\hat{\hat{H}}(t_i)]$. Comment on the expressions you get.

From equation 3.9 (Greenwood's Formula) we know

$$\hat{Var}(\hat{F}(t_i)) = (\hat{S}(t_i))^2 \sum_{j=1}^i \frac{\hat{p}_j}{n_j(1 - \hat{p}_j)}.$$

From the delta method we also know $\hat{F}(t_i)$ and $\hat{H}(t_i)$ are related by

$$\hat{H}(t_i) = g(\hat{F}(t_i)) = -\log(1 - \hat{F}(t_i)),$$

thus

$$\begin{aligned} Var(\hat{H}(t_i)) &\approx (g'(\hat{F}(t_i)))^2 \hat{Var}(\hat{F}(t_i)) = \frac{1}{(\hat{S}(t_i))^2} (\hat{S}(t_i))^2 \sum_{j=1}^i \frac{\hat{p}_j}{n_j(1 - \hat{p}_j)} \\ &= \sum_{j=1}^i \frac{\hat{p}_j}{n_j(1 - \hat{p}_j)} \end{aligned}$$

Using similar arguments as on top of pg. 55 we can find the approximate variance for $\hat{H}(t_i)$

$$Var(\hat{\hat{H}}(t_i)) = Var\left(\sum_{j=1}^i \hat{p}_j\right) \approx \sum_{j=1}^i Var(\hat{p}_j) = \sum_{j=1}^i \frac{\hat{p}_j(1 - \hat{p}_j)}{n_j}$$

These look pretty similar.

- (c) Compare the Nelson-Aalen estimate of $F(t)$ and compare with the estimate computed in Exercise 3.20. Describe similarities and differences.

Run the following R-Code to obtain the plots and tables. (the fleming-harrington method is the Nelson-Aalen estimate). The similarities occur at the beginning, and after a few failures they estimates start to diverge. As expected, the Nelson-Aalen estimate for $\hat{S}(t_i)$ are always greater than (or equal to) the Kaplan-Meier estimate, see part (d)

```
fan_dat <- matrix(c(
450.000,1,460.000,0,1150.000,1,1150.000,1,1560.000,0,1600.000,1,
1660.000,0,1850.000,0,1850.000,0,1850.000,0,1850.000,0,1850.000,0,
2030.000,0,2030.000,0,2030.000,0,2070.000,1,2070.000,1,2080.000,1,
2200.000,0,3000.000,0,3000.000,0,3000.000,0,3000.000,0,3100.000,1,
3200.000,0,3450.000,1,3750.000,0,3750.000,0,4150.000,0,4150.000,0,
4150.000,0,4150.000,0,4300.000,0,4300.000,0,4300.000,0,4300.000,0,
4600.000,1,4850.000,0,4850.000,0,4850.000,0,4850.000,0,5000.000,0,
5000.000,0,5000.000,0,6100.000,0,6100.000,0,6100.000,0,6100.000,1,
6300.000,0,6450.000,0,6450.000,0,6700.000,0,7450.000,0,7800.000,0,
7800.000,0,8100.000,0,8100.000,0,8200.000,0,8500.000,0,8500.000,0,
8500.000,0,8750.000,0,8750.000,0,8750.000,1,9400.000,0,9900.000,0,
10100.000,0,10100.000,0,10100.000,0,11500.000,0),
byrow = T, ncol = 2)

s1 <- Surv(fan_dat[,1],fan_dat[,2],type='right')

f1 <- survfit(s1~1,type="kaplan-meier")
f2 <- survfit(s1~1,type="fleming-harrington")

plot(f1,
      ylim = c(.5,1),
      xlab = 'Time (t)',
      ylab = 'Fraction Surviving',
      las = 1)
par(new = T)
plot(f2,
      col = 2,
      ylim = c(.5,1),
      axes = FALSE)

legend("bottomleft",
      c(expression("Kaplan-Meier"),
        expression("Nelson-Aalen")),
      col = 1:2, bty = "n",
      lwd = 2, xjust = 0.5)
```

```
summary(f1)
```

```
Call: survfit(formula = s1 ~ 1, type = "kaplan-meier")
```

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
450	70	1	0.986	0.0142	0.958	1.000

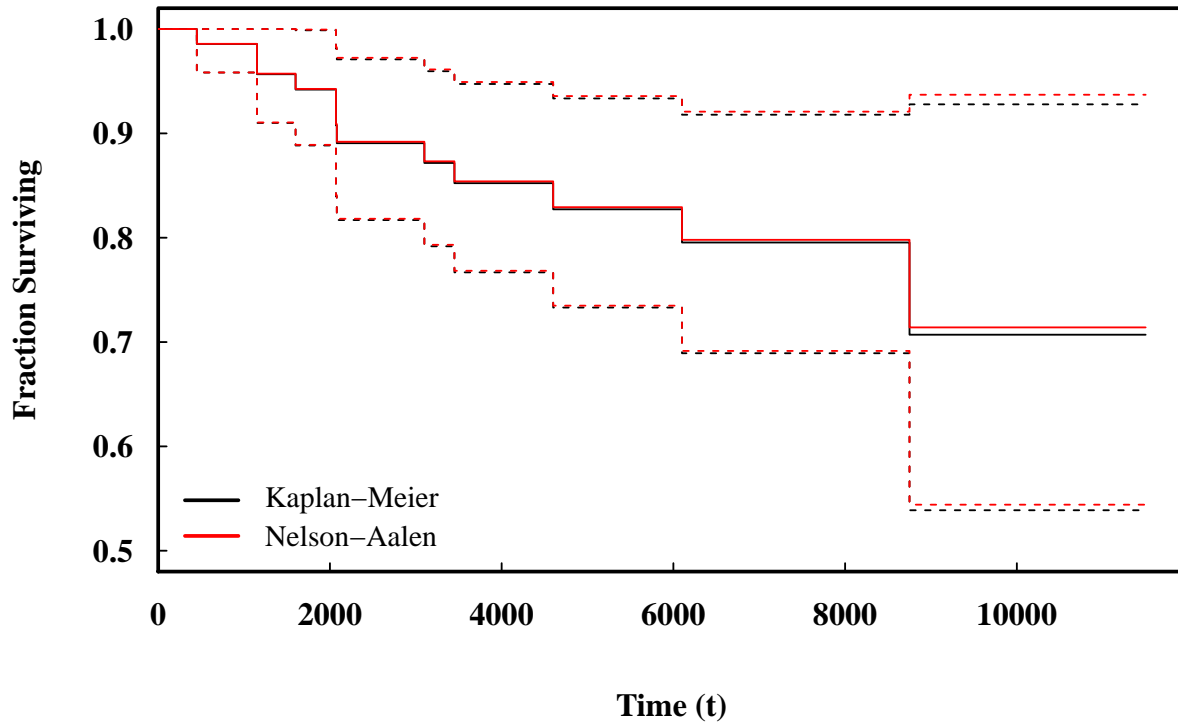


Figure 2: Comparison of Nelson-Aalen and Kaplan-Meier Estimates for the Diesel Engine Fan Data Set (Table C.1)

1150	68	2	0.957	0.0244	0.910	1.000
1600	65	1	0.942	0.0282	0.888	0.999
2070	55	2	0.908	0.0361	0.840	0.981
2080	53	1	0.891	0.0392	0.817	0.971
3100	47	1	0.872	0.0427	0.792	0.960
3450	45	1	0.852	0.0460	0.767	0.947
4600	34	1	0.827	0.0510	0.733	0.933
6100	26	1	0.795	0.0581	0.689	0.918
8750	9	1	0.707	0.0980	0.539	0.928

```
summary(f2)
```

```
Call: survfit(formula = s1 ~ 1, type = "fleming-harrington")
```

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
450	70	1	0.986	0.0142	0.958	1.000
1150	68	2	0.957	0.0245	0.910	1.000
1600	65	1	0.943	0.0282	0.889	0.999
2070	55	2	0.909	0.0361	0.841	0.983
2080	53	1	0.892	0.0393	0.818	0.972
3100	47	1	0.873	0.0428	0.793	0.961
3450	45	1	0.854	0.0461	0.768	0.949
4600	34	1	0.829	0.0511	0.735	0.936
6100	26	1	0.798	0.0583	0.691	0.921
8750	9	1	0.714	0.0990	0.544	0.937

(d) Show that $\hat{\hat{H}}(t_i) < \hat{H}(t_i)$ and that $\hat{\hat{F}}(t_i) < \hat{F}(t_i)$.

We know that $\hat{H}(t_i) = -\sum_{j=1}^i \log(1 - \hat{p}_j)$. Using the the Taylor Series expansion for $\log(1 - \hat{p}_j)$ shown in part (a) results in

$$\hat{H}(t_i) = \sum_{j=1}^i \left(\sum_{n=1}^{\infty} \frac{\hat{p}_j^n}{n} \right).$$

Because $\hat{p}_j > 0, \forall j$ we know that $\hat{H}(t_i) > \sum_{j=1}^i \hat{p}_j = \hat{\hat{H}}(t_i)$.

Therefore,

$$\begin{aligned} \hat{H}(t_i) &> \hat{\hat{H}}(t_i) \\ -\exp[\hat{H}(t_i)] &> -\exp[\hat{\hat{H}}(t_i)] \\ 1 - \exp[\hat{H}(t_i)] &> 1 - \exp[\hat{\hat{H}}(t_i)] \\ \hat{F}(t_i) &> \hat{\hat{F}}(t_i) \end{aligned}$$

From the text problem # 4.5

Let $T \sim WEIB(\mu, \sigma)$, $\eta = \exp(\mu)$, and $\beta = 1/\sigma$.

- (a) For $m > 0$, show that $E(T^m) = \eta^m \Gamma(1 + m/\beta)$, where $\Gamma(x)$ is the gamma function.

Need to show that after implementing the parameter substitutions suggested in the text, the m^{th} moment for the random variable $T \sim Weibull(\eta, \beta)$

$$E[T^m] = T^m \int_0^\infty \frac{\beta}{\eta} \left(\frac{t}{\eta}\right)^{\beta-1} e^{-(\frac{t}{\eta})^\beta} dt = \eta^m \int_0^\infty x^{\frac{m}{\beta}} e^{-x} dx$$

First, define $a = \left(\frac{1}{\eta}\right)^\beta$ and distribute T^m giving

$$E[T^m] = \int_0^\infty a\beta (t)^{m+\beta-1} e^{-at^\beta} dt$$

Next, introduce the substitution $x = at^\beta \rightarrow t = \left(\frac{x}{a}\right)^{\frac{1}{\beta}} \rightarrow dt = \frac{1}{a\beta} \left(\frac{x}{a}\right)^{-1+\frac{1}{\beta}} dx$ such that

$$\begin{aligned} E[T^m] &= \int_0^\infty a\beta \left(\frac{x}{a}\right)^{\frac{m+\beta-1}{\beta}} \exp(-x) \frac{1}{a\beta} \left(\frac{x}{a}\right)^{-1+\frac{1}{\beta}} dx \\ &= \int_0^\infty \left(\frac{x}{a}\right)^{\frac{m}{\beta}+1-\frac{1}{\beta}-1+\frac{1}{\beta}} \exp(-x) dx \\ &= a^{-\frac{m}{\beta}} \int_0^\infty x^{\frac{m}{\beta}} \exp(-x) dx \\ &= a^{-\frac{m}{\beta}} \Gamma\left(\frac{m}{\beta} + 1\right) \\ E[T^m] &= \eta^m \Gamma\left(\frac{m}{\beta} + 1\right). \end{aligned}$$

- (b) Use the result in (a) to show that $E(T) = \eta\Gamma(1 + 1/\beta)$ and $Var(T) = \eta^2[\Gamma(1 + 2/\beta) - \Gamma^2(1 + 1/\beta)]$.

Using the result from (a) for $m = 1$, $E[T] = \eta\Gamma\left(\frac{1}{\beta} + 1\right)$. Further, it can be shown that

$$\begin{aligned} Var[T] &= E[T^2] - (E[T])^2 \\ &= \eta^2 \Gamma\left(\frac{2}{\beta} + 1\right) - \left(\eta\Gamma\left(\frac{1}{\beta} + 1\right)\right)^2 \\ &= \eta^2 \left[\Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right) \right] \end{aligned}$$

From the text problem # 4.7

Consider the Weibull $h(t)$. Note that when $\beta = 1$, $h(t)$ is constant and that when $\beta = 2$, $h(t)$ increases linearly. Show that:

- (a) If $0 < \beta < 1$, then $h(t)$ is decreasing in t

First, note that for the Weibull distribution

$$\begin{aligned}h(t) &= \frac{\beta}{\eta^\beta} t^{\beta-1}, & t > 0, \beta > 0, \eta > 0 \\h'(t) &= \frac{\beta(\beta-1)}{\eta^\beta} t^{\beta-2}, & t > 0, \beta > 0, \eta > 0 \\h''(t) &= \frac{\beta(\beta-1)(\beta-2)}{\eta^\beta} t^{\beta-3}, & t > 0, \beta > 0, \eta > 0\end{aligned}$$

Thus, for $0 < \beta < 1$, $h'(t) < 0$, and $h(t)$ is decreasing for $\beta \in (0, 1)$

- (b) If $1 < \beta < 2$, then $h(t)$ is concave increasing.

For $\beta \in (1, 2)$, $h'(t) > 0$ and $h''(t) < 0$, thus $h(t)$ is concave increasing.

- (c) If $\beta > 2$, then $h(t)$ is convex increasing

For $\beta \in (2, \infty)$, $h'(t) > 0$ and $h''(t) > 0$, thus $h(t)$ is convex increasing

From the text problem # 4.15

The coefficient of variation, γ_2 , is a useful scale-free measure of relative variability for a random variable.

- (a) Derive an expression for the coefficient of variation for the Weibull distribution

Using the results from Exercise 4.5, we can express the Weibull coefficient of variation as

$$\gamma_2 = \sqrt{\frac{\sigma^2}{\mu^2}} = \left[\Gamma\left(1 + \frac{2}{\beta}\right) \Gamma^{-2}\left(1 + \frac{1}{\beta}\right) - 1 \right]^{0.5}$$

- (b) Compute γ_2 for all combinations of $\beta = 0.5, 1, 3, 5$ and $\eta = 50, 100$. Also, draw (or use the computer to draw) a graph of the Weibull pdfs for the same combinations of parameters.

As seen in (a), the value of γ_2 for the Weibull distribution depends only on the value of the shape parameter β . For a given value of β the value of γ_2 can be found using the following R code

```
beta <- c(0.5,1,3,5)

CV <- matrix(NA, ncol = 1, nrow = 4)

colnames(CV) <- c("CV");
rownames(CV) <- c("eta(0.5) =", "eta(1.0) =", "eta(3.0) =", "eta(5.0) =")

CV[1:4] <- (gamma(1+2/beta[1:4]))*(gamma(1+1/beta[1:4]))^-2-1)^0.5
CV
```

	CV
eta(0.5) =	2.2360680
eta(1.0) =	1.0000000
eta(3.0) =	0.3634465
eta(5.0) =	0.2290529

The plots of the respective pdfs may then be found using

```
beta <- rep(c(0.5,1,3,5),2)
eta <- rep(c(50,100),each=4)

for (i in 1:8) {
  curve(dweibull(x,beta[i],eta[i]),
        from = 1,
        to = 100,
        ylim = c(0,.04),
        col = i,
        lty = i,
        xlab = if(i==1) "Time (t)",
        ylab = if(i==1) "f(t)",
        add = if(i>1) TRUE,
        las = 1)
}
legend("topright",
      c(expression(eta~beta=="0.5, 50"),expression(eta~beta=="1.0, 50"),
        expression(eta~beta=="3.0, 50"),expression(eta~beta=="5.0, 50"),
        expression(eta~beta=="0.5,100"),expression(eta~beta=="1.0,100"),
        expression(eta~beta=="3.0,100"),expression(eta~beta=="5.0,100")),
      lty = 1:8,lwd = 2,col = 1:8,
      bty = "n",y.intersp = .9)
```

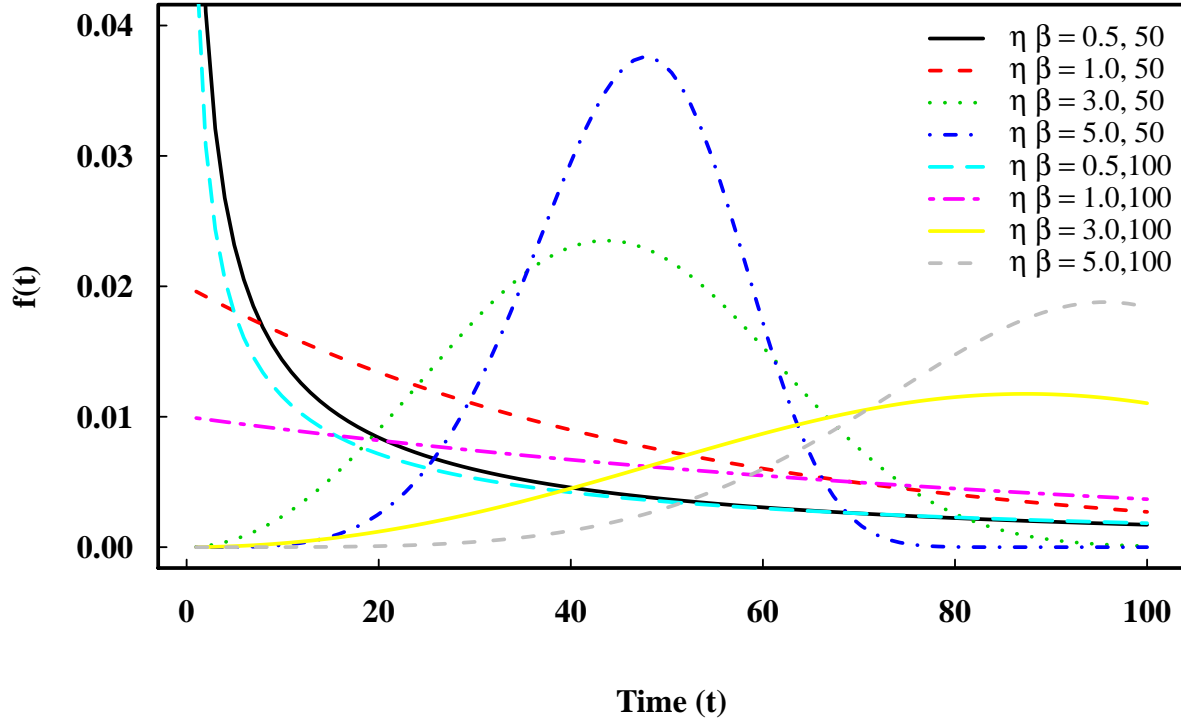


Figure 3: Plots of various Weibull density functions

- (c) *Explain the effect that changes in η and β have on the shape of the Weibull density and the effect that they have on γ_2 .*

In problem 4.7 we discussed the effect changing the shape parameter β has on the shape of the Weibull distribution, thus we need not discuss it again here. Increasing the scale parameter η or “characteristic life” increases the skewness of the distribution by shifting the 0.632 quantile to the right.

The coefficient of variation for the Weibull distribution is a function of the shape parameter alone. The figure below illustrates the γ_2 β relationship

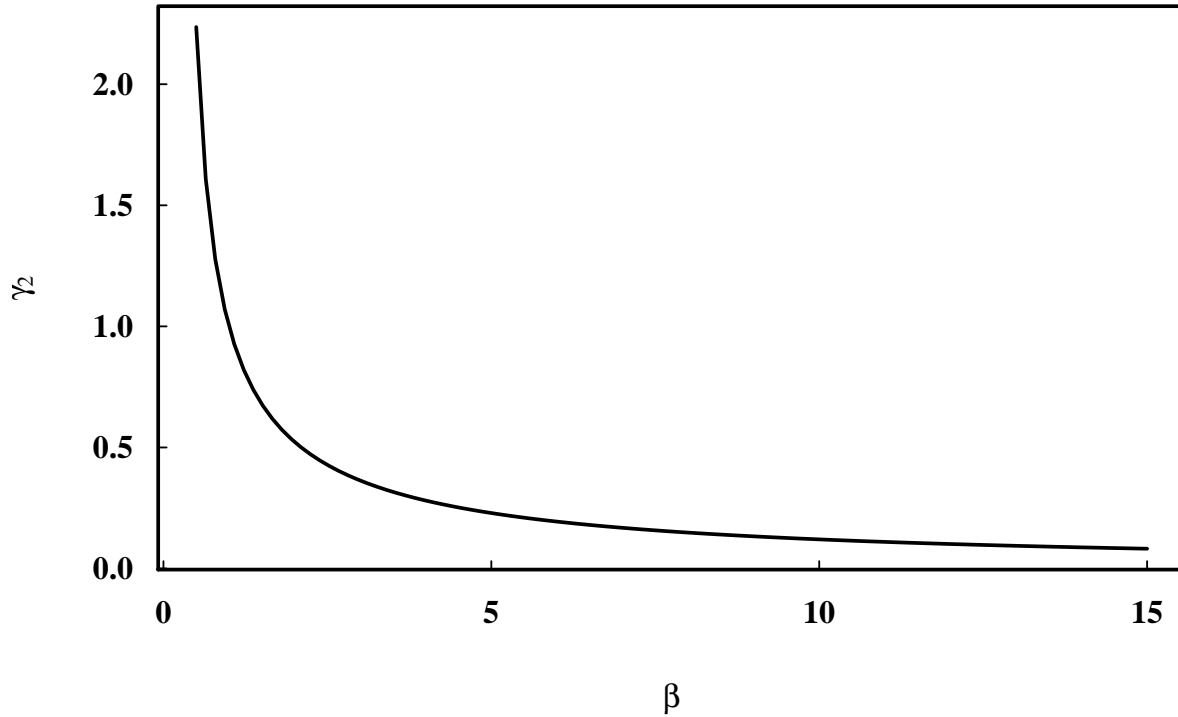


Figure 4: Weibull coefficient of variation vs shape parameter

From the text problem # 5.2

In some applications a sample of failure times comes from a mixture of subpopulations.

- (a) Write down the expression for the cdf_ $F(t)$ for a mixture of two exponential distributions with means $\theta_1 = 1$ and $\theta_2 = 10$ (subpopulations 1 and 2, respectively) with ζ being the proportion from subpopulation 1.

In reliability, mixture distributions can arise when multiple failure mechanisms generate failures for which the respective distributions are sufficiently distinct.

For the mixture model discussed in this exercise, the pdf is expressed as

$$f_x(x) = \eta \frac{1}{\theta_1} e^{-\left(\frac{x}{\theta_1}\right)} + (1 - \eta) \frac{1}{\theta_2} e^{-\left(\frac{x}{\theta_2}\right)}$$

Integrating wrt x gives

$$F_x(x) = \int_0^x f_x(t) dt = \eta(1 - e^{-\left(\frac{x}{\theta_1}\right)}) + (1 - \eta)(1 - e^{-\left(\frac{x}{\theta_2}\right)})$$

and finally,

$$F_x(x) = 1 - \eta e^{-x} - (1 - \eta) e^{-\left(\frac{x}{10}\right)}.$$

- (b) For $\zeta = 0, 0.1, 0.5, 0.9$, and 1, compute the mixture $F(t)$ for a number of values of t ranging between 0 and 30. Plot these distributions on one graph.

The figure below shows the CDF plots for the requested mixture distributions. This plot may be generated with the following R code.

```
MIX <- function(zeta,t) 1-zeta*exp(-t)-(1-zeta)*exp(-(t/10))

curve(MIX(0.0,x),from = 0,to = 30,xlab = 't',ylab = 'F(t)', las = 1)
curve(MIX(0.1,x),lty = 2,col = 2, add = T)
curve(MIX(0.5,x),lty = 3,col = 3,add = T)
curve(MIX(0.9,x),lty = 4,col = 4,add = T)
curve(MIX(1.0,x),lty = 5,col = 5,add = T)

legend("bottomright",
      c(expression(zeta==0.0),
        expression(zeta==0.1),
        expression(zeta==0.5),
        expression(zeta==0.9),
        expression(zeta==1.0)),
      lty = 1:5,lwd = 2, col = 1:5,
      bty = "n",xjust = 0.5)
```

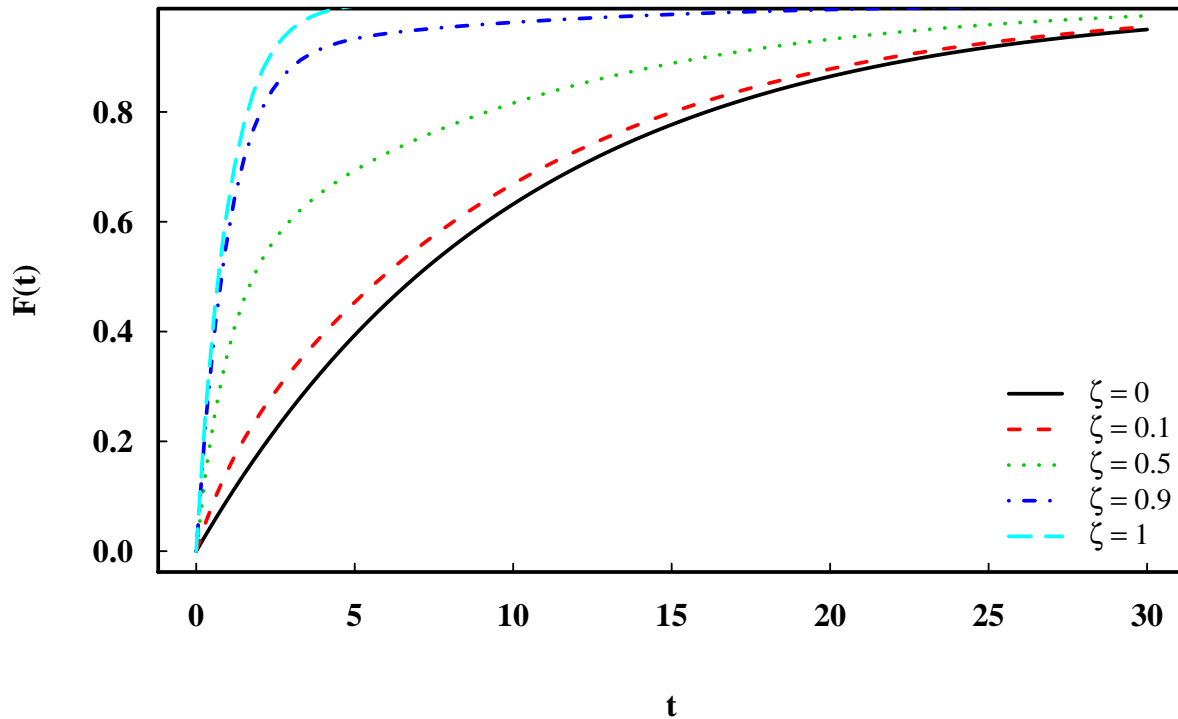


Figure 5: CDF plots for exponential mixture distributions

- (c) Plot $\log(t)$ versus $\log[-\log[1 - F(t)]]$ for each $F(t)$ computed in part (b). Comment on the shapes of the mixtures of exponential distributions, relative to a pure exponential distribution or a Weibull distribution.

The same code shown for (b) can be used for (c) applying the transformations.

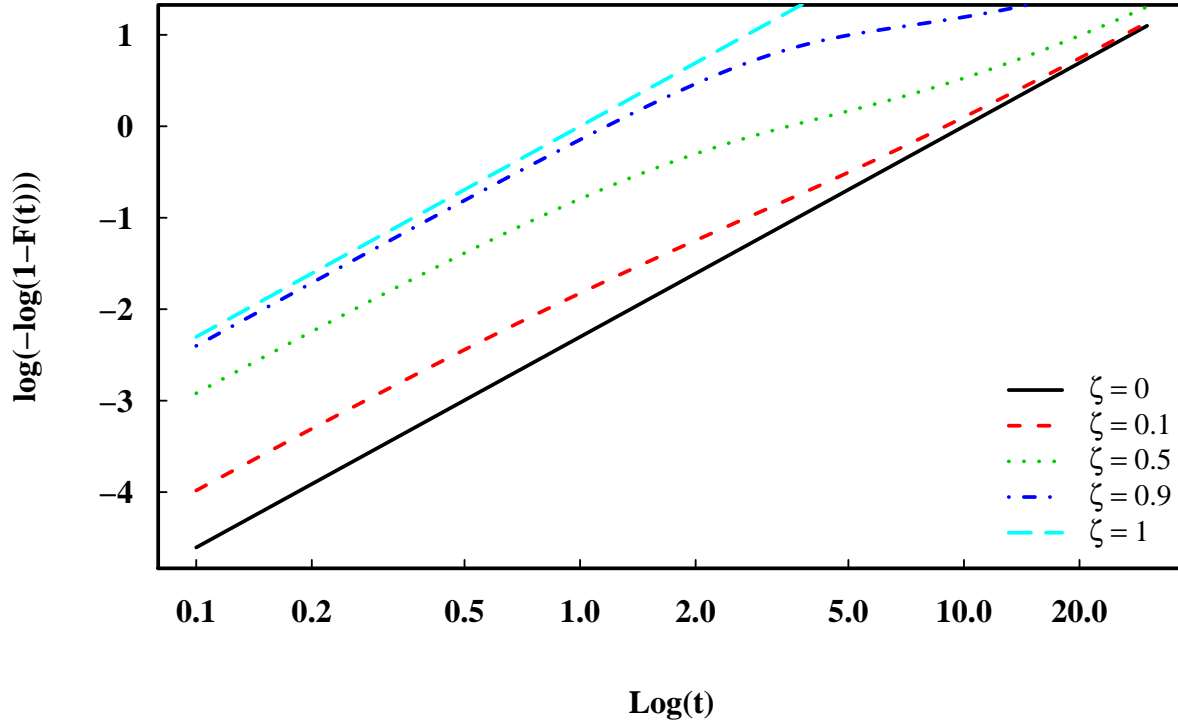


Figure 6: Weibull-transformed quantile plots for the mixture distributions

- (d) Plot the hazard function $h(t)$ on the mixture distributions in part (b).

As discussed, the hazard function can be expressed as $h(t) = \frac{f_X(x)}{1-F_X(x)}$. Though this exercise began with the pdf for the mixture model, the following code can be used to find $f_X(x)$ symbolically when only the cdf is known.

```
DMIX<-function(zeta,t) {
  d.1<-body(MIX)[[2]]      #Extract equation from function object as a call object
  d.2<-as.expression(d.1)  #Convert extracted equation to expression object
  d.3<-D(d.2,"t")          #Take derivative of expression
  d.4<-as.expression(d.3)  #Convert derivative to expression object
  d.5<-eval(d.4)           #Convert derivative expression back to a function object
}
```

Adjusting the plotting code used in parts (b) and (c) results in the following plot of $h(t)$.

- (e) Qualitatively, what do the Weibull plots in part (c) suggest about the hazard function of a mixture of two exponential distributions?

The plots shown in (b), (c) and (d) suggest that the mixture of two exponential distributions with $\eta \in (0, 1)$ is not in the exponential family.

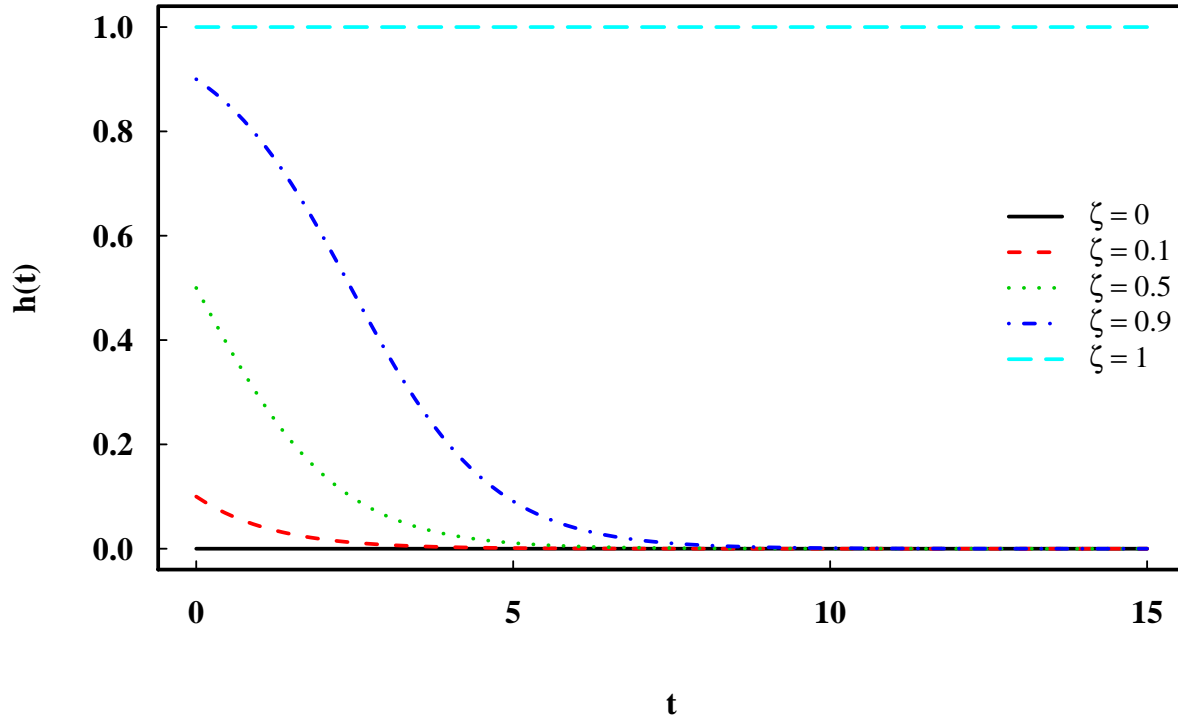


Figure 7: Hazard plots for exponential mixture distributions

From the text problem # 5.4

Refer to Exercise 5.1. Show that a mixture of two exponential distributions with different θ values with always have a decreasing hazard function

For a mixture of exponentials, where $\theta_1 \neq \theta_2, \theta_1 > \theta_2$ this problem asks us to show that

$$\frac{dh(x)}{dt} = \left(\frac{f(x)}{S(x)} \right)^2 + \frac{f'(x)}{S(x)} < 0, x > 0.$$

Rearranging this expression gives $f(x)^2 < -f'(x)S(x)$. Substituting the density and survival functions for the mixture distribution in this expression results in

$$\left(\frac{\eta}{\theta_1} e^{-\frac{x}{\theta_1}} + \frac{(1-\eta)}{\theta_2} e^{-\frac{x}{\theta_2}} \right) \left(\frac{\eta}{\theta_1} e^{-\frac{x}{\theta_1}} + \frac{(1-\eta)}{\theta_2} e^{-\frac{x}{\theta_2}} \right) < \left(\frac{\eta}{\theta_1} e^{-\frac{x}{\theta_1}} + \frac{(1-\eta)}{\theta_2} e^{-\frac{x}{\theta_2}} \right) \left(\eta e^{-\frac{x}{\theta_1}} + (1-\eta) e^{-\frac{x}{\theta_2}} \right)$$

Expanding this expression gives

$$\frac{\eta^2}{\theta_1^2} e^{-\frac{2x}{\theta_1}} + \frac{2\eta(1-\eta)}{\theta_1\theta_2} e^{-\left(\frac{x}{\theta_1} + \frac{x}{\theta_2}\right)} + \frac{(1-\eta)^2}{\theta_2^2} e^{-\frac{2x}{\theta_2}} < \frac{\eta^2}{\theta_1^2} e^{-\frac{2x}{\theta_1}} + \frac{\eta(1-\eta)}{1} e^{-\left(\frac{x}{\theta_1} + \frac{x}{\theta_2}\right)} \left(\frac{1}{\theta_1^2} + \frac{1}{\theta_2^2} \right) + \frac{(1-\eta)^2}{\theta_2^2} e^{-\frac{2x}{\theta_2}}$$

Simplifying gives

$$\frac{2}{\theta_1\theta_2} < \frac{1}{\theta_1^2} + \frac{1}{\theta_2^2}$$

$$0 < \frac{1}{\theta_1^2} - \frac{2}{\theta_1\theta_2} + \frac{1}{\theta_2^2}$$

$$\left(\frac{1}{\theta_1} - \frac{1}{\theta_2}\right)^2$$

As we have defined $\theta_1 > \theta_2$ the quantity $h'(t) < 0$.

From the text problem # 5.12

Let $T_{(1)}$ denote the minimum of m independent Weibull random variables with parameters $\mu_i, i = 1, \dots, m$, and constant σ . Show that $T_{(1)}$ has a Weibull distribution.

Let $T_{(1)} = \min(T_1, \dots, T_m)$ denote the minimum failure time observed from m independent Weibull random variables having distinct scale parameters $\mu_i, i = 1, \dots, m$ and an equivalent shape parameter σ . Then

$$\begin{aligned}
 F_{T_{(1)}}(t) &= Pr(T_{(1)} < t) \\
 &= 1 - Pr(T_{(1)} > t) \\
 &= 1 - \prod_{j=1}^m Pr(T_j > t) \\
 &= 1 - \prod_{j=1}^m \exp \left[- \left(\frac{t}{\exp(\mu_j)} \right)^{\frac{1}{\sigma}} \right] \\
 &= 1 - \exp \left[- \sum_{j=1}^m \left(\frac{t}{\exp(\mu_j)} \right)^{\frac{1}{\sigma}} \right] \\
 &= 1 - \exp \left[- t^{\frac{1}{\sigma}} \times \sum_{j=1}^m \left(\frac{1}{\exp(\mu_j)} \right)^{\frac{1}{\sigma}} \right] \\
 &= 1 - \exp \left[- \left(\frac{t}{\exp[\mu_{(1)}]} \right)^{\frac{1}{\sigma}} \right]
 \end{aligned}$$

which is in the form of a Weibull cdf for which the scale parameter is

$$\exp[\mu_{(1)}] = \left[\sum_{j=1}^m \left(\frac{1}{\exp(\mu_j)} \right)^{\frac{1}{\sigma}} \right]^{-\sigma}$$