Stat 641 Fall 2012

Solutions for Assignment 4

I. (45 points) Using the following R code and various other R functions, we obtain the following:

```
library(MASS)
yS = c(0.42,    0.86,    0.88,    1.11,    1.34,    1.38 ,   1.42,    1.47,    1.63,
   1.73,    2.17,    2.42,    2.48,    2.74,    2.74,    2.79,    2.90,    3.12,
   3.18,    3.27,    3.30,    3.61 ,   3.63,    4.13 ,   4.40,    5.00,    5.20,
   5.59,    7.04,    7.15,    7.25,    7.75,    8.00,    8.84,    9.30 ,   9.68,
   10.32,   10.41,   10.48,   11.29,   12.30,   12.53,   12.69,   14.14,   14.15,
   14.27 ,  14.56,   15.84,   18.55,   19.73,   20.00)
yL =
 c( 0.94 ,  1.26 ,  1.44  , 1.49 ,  1.63 ,  1.80 ,  2.00 ,  2.00 ,  2.56,
    2.58 ,  3.24 ,  3.39  , 3.53 ,  3.77 ,  4.36 ,  4.41 ,  4.60 ,  4.67,
    5.39 ,  6.25 ,  7.02 ,  7.89 ,  7.97 ,  8.00 ,  8.28 ,  8.83 ,  8.91,
    8.96 ,  9.92 , 11.36 , 12.15,  14.40 , 16.00 , 18.61 , 18.75 , 19.05,
   21.00 , 21.41 , 23.27 , 24.71,  25.00 , 28.75 , 30.23 , 35.45 , 36.35)
yL  = sort(yL)
yLt = yL[c(-1,-2,-3,-4,-length(yL),-(length(yL)-1),-(length(yL)-2),-(length(yL)-3))]
meanL = mean(yL)
trim.meanL = mean(yLt)
stdL = sd(yL)
medL = median(yL)
iqrL = quantile(yL,.75) - quantile(yL,.25)
madL = mad(yL)
outL = c(meanL, stdL, medL, madL)
meanS = mean(yS)
stdS = sd(yS)
medS = median(yS)
madS = mad(yS)
outS = c(meanS, stdS, medS, madS)
weib  = log(yL)
n  = length(weib)
i = 1:n
ui = (i-.5)/n
QW = log(-log(1-ui))
postscript("u:/meth1/homework/solutions/Assign4ProbI.ps",height=6,horizontal=F)
plot(QW,weib,abline(lm(weib~QW)),
       main="Weibull Reference Plot",cex=.75,lab=c(7,11,7),
       xlab="Q(u) = log(-log(1-ui))",
       ylab="log(yL(i))")
legend(-4,3.0,"y = 2.3938 + 0.7785 Q(u)")
mle_weib <- fitdistr(yL,"weibull")
shape = mle_weib$estimate[1]
scale = mle_weib$estimate[2]
mean_weib = scale*gamma(1+1/shape)
std_weib = sqrt(scale^2*(gamma(1+2/shape)-(gamma(1+1/shape))^2))
median_weib = scale*((-log(.5))^(1/shape))
iqr_weib = scale*((-log(.25))^(1/shape)) - scale*((-log(.75))^(1/shape))
out_weib = c(mean_weib,std_weib,median_weib,iqr_weib)
out_weib
outL
outS
```

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

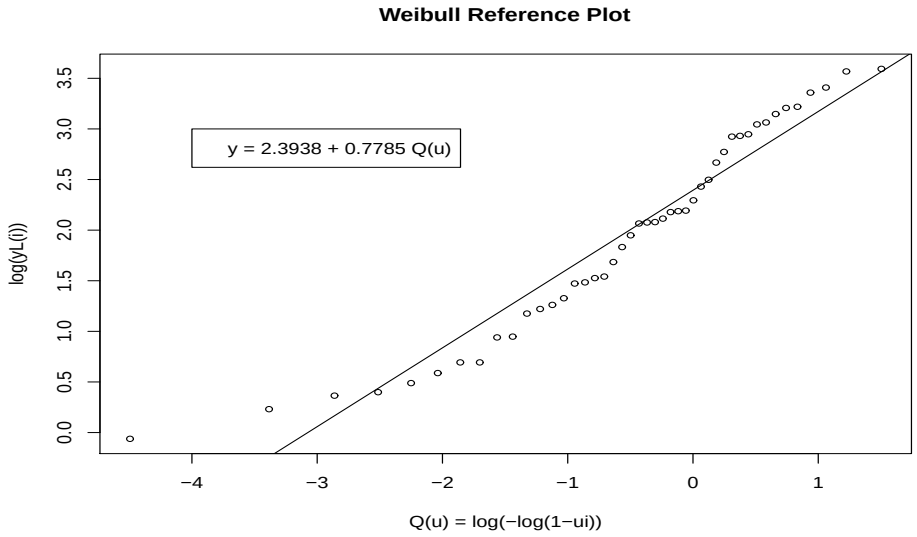
   $K(.10) = 45 - [(45)(.1)] - [(45)(.1) + 1] + 1 = 45 - [4.5] - [5.5] + 1 = 37$ values in the data set yielding:

   $\hat{\mu}_{(.1)} = \frac{1}{37}\sum_{i=5}^{41} Y_{(i)} = \frac{1}{37}(357.67) = 9.667$ whereas the untrimmed mean is

   $\hat{\mu} = \frac{1}{45}\sum_{i=1}^{45} Y_{(i)} = \frac{1}{45}(493.58) = 10.968$

   The untrimmed mean is somewhat larger than the 10% trimmed mean which would indicate that there a few large outliers in the data. In fact, examining the sorted data, we have three relatively large data values in the data set: 30.23, 35.45, 36.35.

2. A Weibull reference distribution plot is displayed:



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

$$\hat{\gamma} = 1/.7785 = 1.2845 \qquad \hat{\alpha} = e^{2.3938} = 10.9550$$

The MLE of the Weibull parameters from R are

$$\hat{\gamma} = \text{Weibull Shape} = 1.1326673 \qquad \hat{\alpha} = \text{Weibull Scale} = 11.4982087$$

3. Using the MLE estimates: $P[Y_L > 30] = 1 - F(30) \approx e^{-(30/11.4982087)^{1.1326673}} = .0517$

   Using the Graphical estimates: $P[Y_L > 30] = 1 - F(30) \approx e^{-(30/10.9550)^{1.2854}} = .0260$

   The distribution-free estimate would be $P[Y > 3000] = 1 - \hat{F}(30) = 1 - 42/45 = .0667$

4. The distribution-free estimates are

   $\hat{\mu} = \text{sample mean} = \bar{Y} = 10.9684 \qquad \hat{\sigma} = \text{sample stand. dev.} = S_Y = 9.8369$

   Using the formulas on page 3 in HO 6, we have for the Weibull distribution, with MLE's from R:

$$\widehat{\mu} = \widehat{\alpha}\Gamma\left(1 + \tfrac{1}{\widehat{\gamma}}\right) = (11.4982087)\Gamma\left(1 + \tfrac{1}{1.1326673}\right) = 10.9924$$

$$\widehat{\sigma} = \widehat{\alpha}\sqrt{\Gamma\left(1 + \tfrac{2}{\widehat{\gamma}}\right) - \Gamma^2\left(1 + \tfrac{1}{\widehat{\gamma}}\right)} = 11.4982087\sqrt{\Gamma\left(1 + \tfrac{2}{1.1326673}\right) - \Gamma^2\left(1 + \tfrac{1}{1.1326673}\right)} = 9.7250$$

   The MLE estimates of μ and σ are very close to the distribution-free estimates thus lending evidence that the Weibull model is the correct model for this data. However, the closeness of the fit of two moments thus not justify the fit of the distribution. We will investigate further in a later assignment.

5. The distribution-free estimates are

   $\hat{\tilde{\mu}} = \hat{Q}(.5) = \text{sample median} = Y_{(23)} = 7.97 \qquad$ Using R-function, quantile(yL,.5) = 7.97

   $\widehat{IQR} = \text{sample IQR} = \hat{Q}(.75) - \hat{Q}(.25) = Y_{(.75n+.5)} - Y_{(.25n+.5)} = Y_{(34.25)} - Y_{(11.75)} = 18.825 - 3.353 = 15.47$

   Using R-function, $quantile(yL, .75) - quantile(yL, .25) = 18.61 - 3.39 = 15.22$

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

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

$\widehat{\tilde{\mu}} = \hat{Q}(.5) = \hat{\alpha}(-log(1-.5))^{1/\hat{\gamma}} = 11.4982087(-log(1-.5))^{1/1.1326673} = 8.32$

$\widehat{IQR} = \hat{Q}(.75) - \hat{Q}(.25) = \hat{\alpha}(-log(1-.75))^{1/\hat{\gamma}} - \hat{\alpha}(-log(1-.25))^{1/\hat{\gamma}} = 11.51$

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

$\widehat{\tilde{\mu}} = qweibull(.5, 1.1326673, 11.4982087) = 8.32$

$\widehat{IQR} = qweibull(.75, 1.1326673, 11.4982087) - qweibull(.25, 1.1326673, 11.4982087) = 11.51$

The MLE estimate of the median based on the Weibull model is close to the distribution-free estimate (8.32 vs 7.97) but there is substantial difference between the two estimates of the IQR (11.51 vs 15.47). This may be due to the IQR reflecting only the fit of the data in the middle of the distribution.

6. For Large Litter Data: $\hat{\mu}_L = 10.97 \;\; \hat{\sigma}_L = 9.84$

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

7. For Large Litter Data: $\hat{\tilde{\mu}}_L = 7.27 \;\; \hat{MAD}_L = 8.02$

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

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

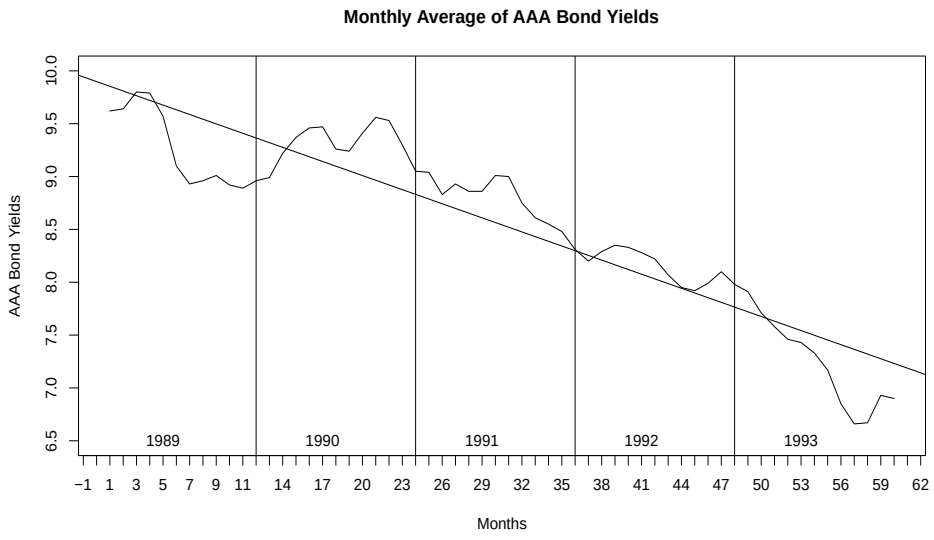
   For the Large Litter Data, the pdf appeared to be just more right skewed so the mean should be larger than the median (10.96 vs 7.97) and the standard deviation somewhat larger than MAD (9.84 vs 8.02). I was somewhat surprised that there was not a larger difference between S and MAD considering the 4 or 5 rather large values in the Large Litter data set.

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

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

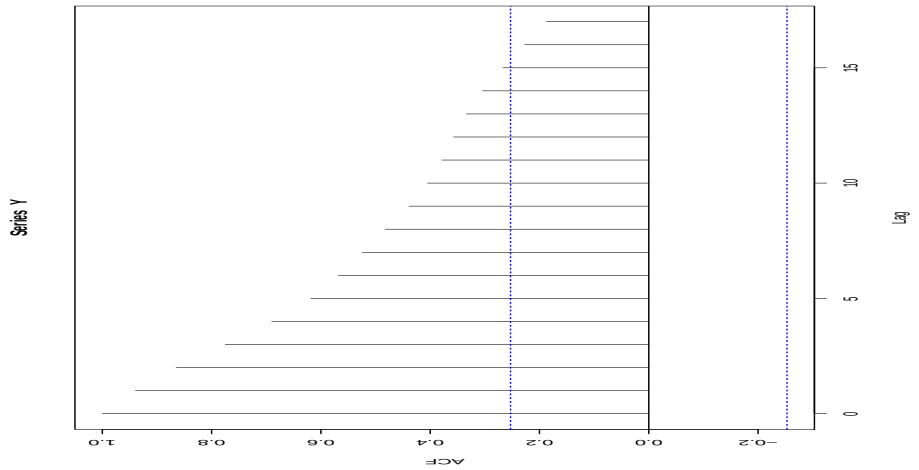
II. (20 points) Using the monthly average daily yields of the bonds the following results are obtained:

1. The following time series plot is obtained using the provided R code:



2. The lag k autocorrelations are given below. Based on these correlations and the plot it would appear that the adjacent monthly sales have a strong positive correlation. There appears to be a very slow decline in the autocorrelations with a pattern such as $\rho_k = (\rho_1)^k = (.939)^k$ for $k = 1, 2, \cdots, 17$, as would be seen in an AR(1) model. This would indicate that the monthly average yields of the AAA bonds are strongly correlated.

i	0	1	2	3	4	5	6	7	8	9	10	11	12
$\hat{\rho}_i$	1.000	0.939	0.864	0.775	0.690	0.618	0.568	0.524	0.482	0.438	0.404	0.378	0.357
i	13	14	15	16	17								
$\hat{\rho}_i$	0.333	0.304	0.266	0.226	.187								



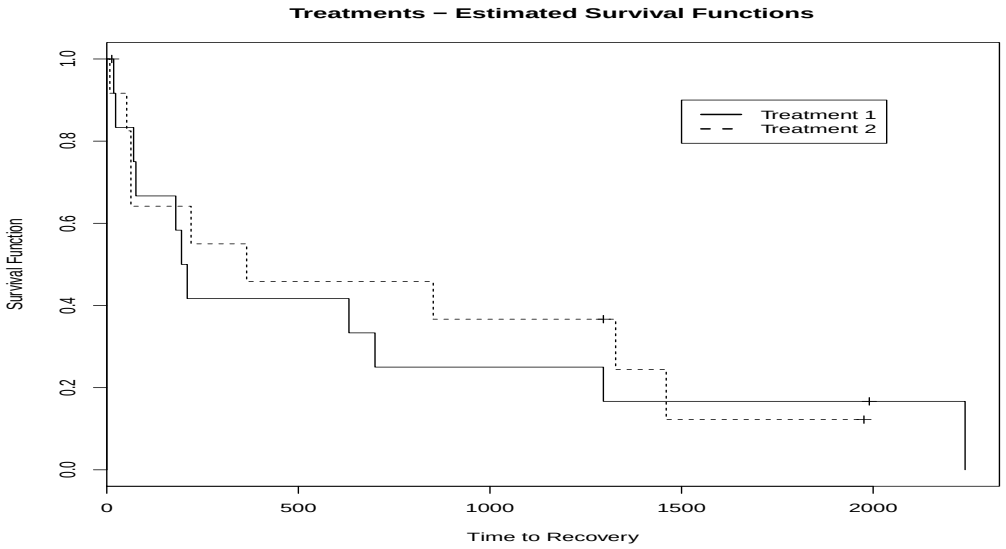
3. The mean and standard deviations of the monthly sales over the five years are given in the following table:

Month	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
Mean	8.752	8.738	8.806	8.780	8.722	8.584	8.482	8.384	8.352	8.332	8.340	8.240
St.Dev.	0.690	0.760	0.871	0.927	0.889	0.808	0.857	1.007	1.119	1.085	0.907	0.872

The sales appear non-stationary with an overall decline in yields along with a somewhat cyclic behavior over the five years. However, the monthly means and standard deviations over the five years are somewhat stable with a pattern of higher values for January through May and then lower values through the remaining months.

III. (20 points) Using the times to recovery (or censoring) for the 25 patients we obtain:

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



2. From the R output we have

```
              G=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   18     12       1    0.917  0.0798       0.7729        1.000
   23     11       1    0.833  0.1076       0.6470        1.000
   70     10       1    0.750  0.1250       0.5410        1.000
   76      9       1    0.667  0.1361       0.4468        0.995
  180      8       1    0.583  0.1423       0.3616        0.941
  195      7       1    0.500  0.1443       0.2840        0.880
  210      6       1    0.417  0.1423       0.2133        0.814
  632      5       1    0.333  0.1361       0.1498        0.742
  700      4       1    0.250  0.1250       0.0938        0.666
 1296      3       1    0.167  0.1076       0.0470        0.591
 2240      1       1    0.000     NaN           NA           NA

              G=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    8     12       1    0.917  0.0798       0.7729        1.000
   52     10       1    0.825  0.1128       0.6311        1.000
   63      9       2    0.642  0.1441       0.4132        0.996
  220      7       1    0.550  0.1499       0.3224        0.938
  365      6       1    0.458  0.1503       0.2410        0.872
  852      5       1    0.367  0.1456       0.1684        0.798
 1328      3       1    0.244  0.1392       0.0801        0.746
 1460      2       1    0.122  0.1110       0.0206        0.724

> print(results, print.rmean=TRUE)
Call: survfit(formula = Surv(T, ST) ~ G)

    records n.max n.start events *rmean *se(rmean) median 0.95LCL 0.95UCL
G=1      13    13      13     11    635        216    202      76      NA
G=2      12    12      12      9    747        226    365      63      NA
```

Note that for G=1, the table reports the median as 202 but $\hat{S}(195) = .5$ from the output of the K-M estimator of the survival function. According to our definition of the quantile function, $\hat{Q}(u) = inf\{t : \hat{S}(t) \leq 1-u\}$, the median would be 195. The output from SAS is identical except the output has the estimated median listed as 202.5 which equals (195+210)/2. In the SAS output, estimates of the lower and upper quartiles are also displayed. Their values are given as $\hat{Q}(.25) = 73$, the average of 70 and 76, and $\hat{Q}(.75) = 998$ which is the average of 700 and 1296. For G=2, SAS reports $\hat{Q}(.25) = 63$, $\hat{Q}(.75) = 1328$ which are what we obtain using $\hat{Q}(u) = inf\{t : \hat{S}(t) \leq 1-u\}$. The logic of taking the average of the endpoints of the flat region at level u in $\hat{S}$, as is done in R and SAS, is equivalent to using a smoothed version of the estimated survival function. That is, taking the average of the smallest and largest values of t that satisfy S(t) = u, which is the midpoint of the flat region at level u.

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

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

IV. (15 points)

1. B because the true stress for the censored specimens are greater than or equal to $t_C = 500$ psi
2. D because the true time to learn for the puppies in group I is less than the age of the puppy at the start of the study
3. E because the true time to learn for the puppies in group II is recorded for each of the puppies in this group
4. A because the true time to learn for the puppies in group III is greater than t_C age of puppy at the end of the study.

   The correct answer would be "right censoring" if the puppies were of differing ages at the start of the study and Type I only if all of the puppies were of the same age at the start of the study.
5. E because brake failure mileage for the censored automobiles are greater than the miles traveled at the end of the study. The censoring is not Type I censoring because the researcher is recording miles driven not time whereas the stopping rule was time.

V. (5 Bonus points) Let Y have a distribution which is symmetric about its median $\tilde{\mu}_Y$.

Let $W = |Y - \tilde{\mu}|$. Then MAD equals the median of W.

1. By the symmetry of the distribution of Y, $Q(.75) - \tilde{\mu}_Y = \tilde{\mu}_Y - Q(.25) = -(Q(.25) - \tilde{\mu}_Y)$
2. By the definition of $Q(.75)$, the median$(Y - \tilde{\mu}_Y$ for $Y \geq \tilde{\mu}_Y) = Q(.75) - \tilde{\mu}_Y$
3. By the symmetry of the distribution of Y, median$(Y - \tilde{\mu}_Y$ for $Y \geq \tilde{\mu}_Y)$ = median$(-(Y - \tilde{\mu}_Y)$ for $Y \leq \tilde{\mu}_Y)$

   MAD = median$(|Y - \tilde{\mu}_Y|)$ = median$(Y - \tilde{\mu}_Y$ for $Y \geq \tilde{\mu}_Y) = Q(.75) - \tilde{\mu}_Y$
4. Therefore, MAD $= Q(.75) - \tilde{\mu}_Y = \frac{1}{2}(Q(.75) - \tilde{\mu}_Y - (Q(.25) - \tilde{\mu}_Y)) = \frac{1}{2}(Q(.75) - Q(.25)) =$ SIQR