MIKE LEDERLE
STATISTICS 641 - ASSIGNMENT #4 - NOON (CDT) Friday - 10/5/2012

• Read Handouts 6 & 7 and Chapters 2 & 4 in the Textbook.
• Submit for grading the following problems:
I. (45 points) A researcher is studying the relative brain weights (brain weight divided by body weight) for 51 species of mammal whose average litter size is less than 2 and for 45 species of mammal whose average litter size is greater than or equal to 2. The researcher was interested in determining what evidence that brain sizes tend to be different for the two groups. (Data from *The Statistical Sleuth* by Fred Ramsey and Daniel Schafer).

RELATIVE BRAIN WEIGHTS - SMALL LITTER SIZE

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			

RELATIVE BRAIN WEIGHTS - LARGE LITTER SIZE

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

1. For the Large Litter Size mammals, Compute a 10% trimmed mean, and compare it to the untrimmed sample mean. Does this comparison suggest any extreme values in the data?

```
large <- c(0.94, 1.26, 1.44, 1.49, 1.63, 1.8, 2, 2, 2.56, 2.58, 3.24, 3.39, 3.53, 3.77, 4.36,
    4.41, 4.6, 4.67, 5.39, 6.25, 7.02, 7.89, 7.97, 8, 8.28, 8.83, 8.91, 8.96, 9.92, 11.36,
    12.15, 14.4, 16, 18.61, 18.75, 19.05, 21, 21.41, 23.27, 24.71, 25, 28.75, 30.23, 35.45,
    36.35)
mean(large, 0.1)

## [1] 9.667

mean(large)

## [1] 10.97
```

Yes, it suggests there are extreme values, since when we remove the tails, we get a value different (12% different) than the untrimmed mean.

2. The researcher suggested a Weibull distribution to model the data for the Large Litter Size mammals. Assuming that the Weibull distribution is an appropriate model for the Large Litter Size data, obtain the MLE estimates of the Weibull parameters for the Large Litter Size data.

Can calculate via `MASS` library (as given in notes):

```
library(MASS)
(f <- fitdistr(large, "weibull"))

## Warning:  NaNs produced

##     shape     scale
##    1.1327   11.4982
##  ( 0.1317) ( 1.6007)

str(f)

## List of 5
##  $ estimate: Named num [1:2] 1.13 11.5
##   ..- attr(*, "names")= chr [1:2] "shape" "scale"
##  $ sd      : Named num [1:2] 0.132 1.601
##   ..- attr(*, "names")= chr [1:2] "shape" "scale"
##  $ vcov    : num [1:2, 1:2] 0.0174 0.0687 0.0687 2.5622
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "shape" "scale"
##   .. ..$ : chr [1:2] "shape" "scale"
##  $ loglik  : num -152
##  $ n       : int 45
##  - attr(*, "class")= chr "fitdistr"

shape <- f$estimate[1]
scale <- f$estimate[2]
```

3. Estimate the probability that a randomly selected mammal with a litter size of 5 will have a relative brain weight greater than 30.

```
pweibull(30, shape, scale, lower.tail = FALSE)

## [1] 0.05166
```

4. Compare the MLE estimates of μ and σ based on the Weibull model to the distribution-free estimates of μ and σ for the Large Litter Size data.

The weibull mean and standard deviation are

```r
mu.weibull <- scale * gamma(1 + 1/shape)
names(mu.weibull) <- NULL
mu.weibull

## [1] 10.99

sigma.weibull <- sqrt(scale^2 * (gamma(1 + 2/shape) - gamma(1 + 1/shape)^2))
names(sigma.weibull) <- NULL
sigma.weibull

## [1] 9.725
```

```r
mean(large)

## [1] 10.97

sd(large)

## [1] 9.837
```

5. Compare the MLE estimates of median and IQR based on the Weibull model to the distribution-free estimates of median and IQR for the Large Litter Size data.

For the weibull model:

```r
first.weibull <- qweibull(0.25, shape, scale)
median.weibull <- qweibull(0.5, shape, scale)
third.weibull <- qweibull(0.75, shape, scale)
(IQR.weibull <- third.weibull - first.weibull)

## [1] 11.51

median.weibull

## [1] 8.32
```

Distribution-free estimates:

```
IQR(large)

## [1] 15.22

median(large)

## [1] 7.97
```

6. Without any assumed model, estimate the mean and standard deviation of the relative brain weights for both Large and Small litter sizes.

NOTE: I am making the assumption that we are going distribution-free for the rest of the problem.

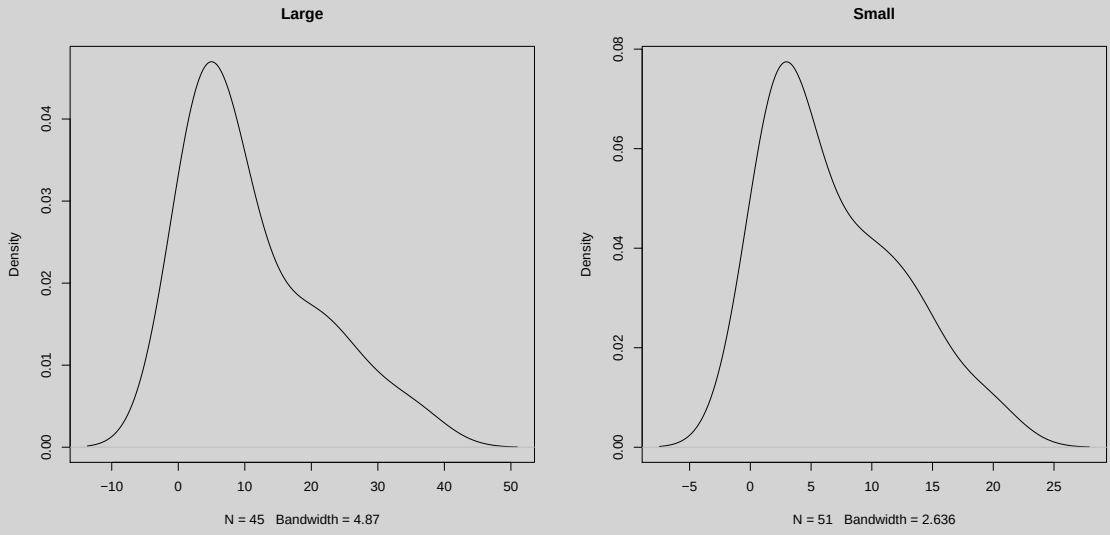
```
small <- c(0.42, 0.86, 0.88, 1.11, 1.34, 1.38, 1.42, 1.47, 1.63, 1.73, 2.17, 2.42, 2.48, 2.74,
    2.74, 2.79, 2.9, 3.12, 3.18, 3.27, 3.3, 3.61, 3.63, 4.13, 4.4, 5, 5.2, 5.59, 7.04, 7.15,
    7.25, 7.75, 8, 8.84, 9.3, 9.68, 10.32, 10.41, 10.48, 11.29, 12.3, 12.53, 12.69, 14.14,
    14.15, 14.27, 14.56, 15.84, 18.55, 19.73, 20)

mean(large)

## [1] 10.97

mean(small)

## [1] 6.886

sd(large)

## [1] 9.837

sd(small)

## [1] 5.46
```

7. Estimate the median and MAD of the relative brain weights for both Large and Small litter sizes.

```r
MAD <- function(samp) {
    median(abs(samp - median(samp)))/0.6745
}

median(large)

## [1] 7.97

MAD(large)

## [1] 8.021

median(small)

## [1] 5

MAD(small)

## [1] 5.234
```

8. Based on your plots from Assignment #3, which pair of estimates of the center and spread in the two data sets best represents the center and spread in the two populations of relative brain weights?

```r
plot(density(large, window = "g", bw = "nrd"), type = "l", main = "Large")
plot(density(small, window = "g", bw = "nrd"), type = "l", main = "Small")
```



Because the data are skewed, use the median & MAD.

9. Using your answers from the previous three questions, suggest a relationship (if any) between litter size and relative brain weights.

Larger litters tend to have larger brain size.

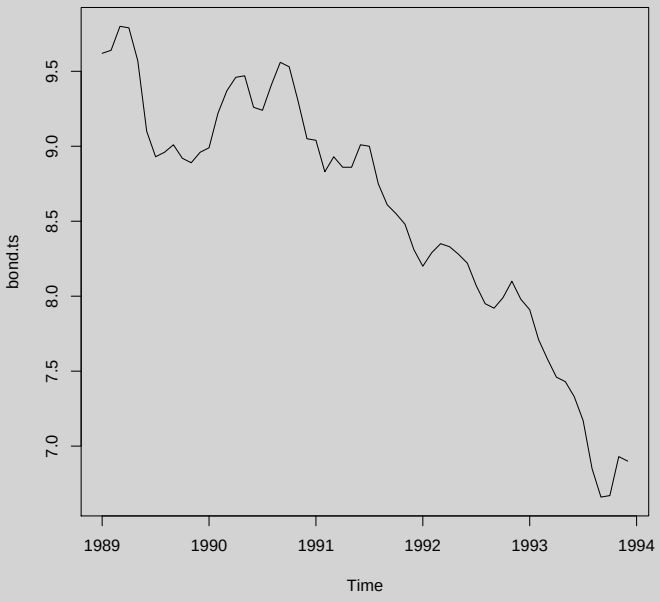
II. (20 points) The following data is the monthly average of daily yields of Moody's AAA bonds for the years 1989 to 1993.

Year	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.
1989	9.62	9.64	9.80	9.79	9.57	9.10	8.93	8.96	9.01	8.92	8.89	8.96
1990	8.99	9.22	9.37	9.46	9.47	9.26	9.24	9.41	9.56	9.53	9.30	9.05
1991	9.04	8.83	8.93	8.86	8.86	9.01	9.00	8.75	8.61	8.55	8.48	8.31
1992	8.20	8.29	8.35	8.33	8.28	8.22	8.07	7.95	7.92	7.99	8.10	7.98
1993	7.91	7.71	7.58	7.46	7.43	7.33	7.17	6.85	6.66	6.67	6.93	6.90

The R code **Assignment04ProbII_2012.R** provided in Files/Homework Assignments will be very useful in this problem.

1. Create a time series plot of the data.

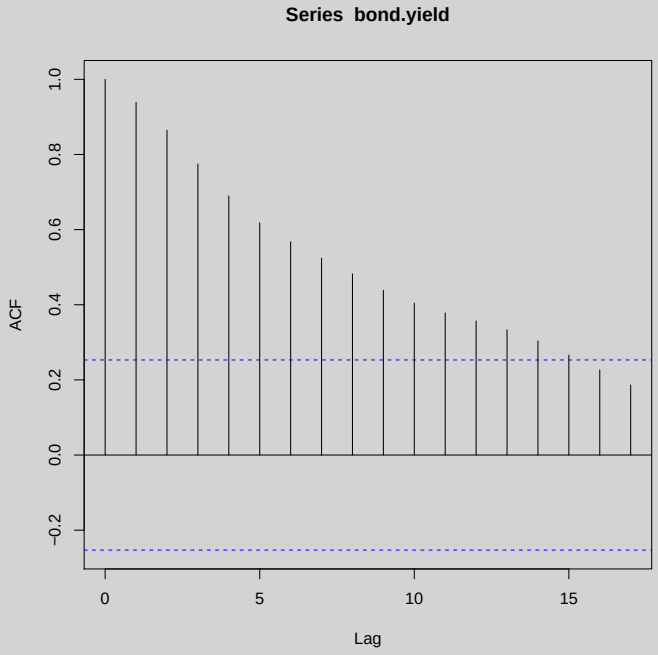
```
bond.yield <- c(9.62, 9.64, 9.8, 9.79, 9.57, 9.1, 8.93, 8.96, 9.01, 8.92, 8.89, 8.96, 8.99,
    9.22, 9.37, 9.46, 9.47, 9.26, 9.24, 9.41, 9.56, 9.53, 9.3, 9.05, 9.04, 8.83, 8.93, 8.86,
    8.86, 9.01, 9, 8.75, 8.61, 8.55, 8.48, 8.31, 8.2, 8.29, 8.35, 8.33, 8.28, 8.22, 8.07, 7.95,
    7.92, 7.99, 8.1, 7.98, 7.91, 7.71, 7.58, 7.46, 7.43, 7.33, 7.17, 6.85, 6.66, 6.67, 6.93,
    6.9)
bond.ts <- ts(bond.yield, frequency = 12, start = c(1989, 1))
plot(bond.ts)
```



2. Calculate the values of ρ_k, the autocorrelation coefficients. What conclusions can you draw?

```
bond.acf <- acf(bond.yield, plot = TRUE)
# str(bond.acf)
bond.acf$acf
```

```
## , , 1
##
##          [,1]
##  [1,] 1.0000
##  [2,] 0.9390
##  [3,] 0.8645
##  [4,] 0.7747
##  [5,] 0.6896
##  [6,] 0.6181
##  [7,] 0.5677
##  [8,] 0.5241
##  [9,] 0.4821
## [10,] 0.4385
## [11,] 0.4044
## [12,] 0.3782
## [13,] 0.3569
## [14,] 0.3334
## [15,] 0.3038
## [16,] 0.2662
## [17,] 0.2264
## [18,] 0.1866
```



Conclusion is strong autocorrelation, meaning the data are not random when comparing near-adjacent values.

3. Does the time series appear to be stationary? That is, do the mean and variance appear to remain constant over time.

The above `acf` plot indicates non-stationary behavior.

III. (20 points) Twenty-five patients diagnosed with rare skin disease are randomly assigned to two drug treatments. The following times are either the time in days from the point of randomization to either a complete recovery or censoring (as indicated by the status variable: 0 means censored, i.e., time at which patient left study prior to a complete recovery, 1 means patient's time to recovery).

	Treatment 1												
Time	180	632	2240	195	76	70	13	1990	18	700	210	1296	23
Status	1	1	1	1	1	1	0	0	1	1	1	1	1
	Treatment 2												
Time	8	852	52	220	63	8	1976	1296	1460	63	1328	365	
Status	0	1	1	1	1	1	0	0	1	1	1	1	

The R code **Assignment04ProbIII_2012.R** provided in Files/Homework Assignments will be very useful in this problem.

1. Estimate the survival function for the two treatments.

The survival function is estimated in the `survival` column of the call to `summary`.

```
library(survival)
T <- c(180, 632, 2240, 195, 76, 70, 13, 1990, 18, 700, 210, 1296, 23, 8, 852, 52, 220, 63,
    8, 1976, 1296, 1460, 63, 1328, 365)

ST <- c(1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1)
G <- c(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)
# out <- cbind(T,ST,G) S <- Surv(T, ST) str(S)
results <- survfit(Surv(T, ST) ~ G)
```

```
summary(results)

## Call: survfit(formula = Surv(T, ST) ~ G)
##
##                 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
```

2. Compare the mean and median time to death for the two treatments

Compare the mean and median columns below:

```
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
##     * restricted mean with upper limit =  2108
```

3. Which treatment appears to be most effective in the treatment of the skin disease?

Treatment 1; both the mean and median time to recovery are less.

IV. (15 points) **Select** the letter of the **BEST** answer. Justify your answer with at most 20 words.

1. An experiment involves putting specimens of steel under stress until the specimen fractures. The machine increases the stress until the specimen fractures. The maximum stress that the machine can place on a specimen is 500 psi. Out of the 35 specimens used in the experiment, 5 did not fracture at 500 psi. This type of censoring is called

   A. Random censoring

   B. Type I censoring

      Stress force replaces time; any greater than 500 are censored.

   C. Type II censoring

   D. Left censoring

   E. Right censoring

2. A veterinarian designed a study to determine the age at which Labrador retrievers learned how to swim. There was three groups of puppies:

   Group I: Puppies who knew how to swim prior to the beginning of the study;

   Group II: Puppies who learned how to swim during the study;

   Group III: Puppies who had not yet learned how to swim at the conclusion of the study.

   The age at which each puppy learned how to swim was recorded. The values recorded for the Group I puppies are

   A. Type I censored

   B. Type II censored

   C. Random censored

   D. Left censored

      Only have upper bound on the age at which swimming was learned.

   E. Uncensored

3. Refer to Problem 2. The values recorded for the Group II puppies are

   A. Type I censored

   B. Type II censored

   C. Random censored

   D. Left censored

   E. Uncensored

      Complete data set.

4. Refer to Problem 2. The values recorded for the Group III puppies are

   A. Type I censored

      The study ended at t; only know that group III has $T > t$.

   B. Type II censored

   C. Random censored

   D. Left censored

   E. Uncensored

5. An engineer for an automotive manufacturer is studying the occurrence of a defective in the braking system for a newly designed braking system. She randomly selects 100 automobiles for study and plans to record the distance traveled prior to a failure in the braking system. However, she needs to conclude the study 12 months after its inception. For each of the 100 automobiles she recorded the mileage at which a failure occurred in the braking system or the mileage driven during the 12 month study for those automobiles that did not have a failure. We would describe the data from this type of study as having

   A. Type I censoring

      We only know lower bound on those systems that didn't fail at the study completion.

   B. Type II censoring
   C. Random censoring
   D. Left censoring
   E. Right censoring

V. Bonus Problem for 5 points (attempt this problem only if you have extra time).

   Prove the following statement:

   If Y has a symmetric distribution with $\mu < \infty$ and median $\tilde{\mu}$,

   then, the median of $W = |Y - \tilde{\mu}|$, equals the SIQR of Y.