STAT-S 470/670 Homework 2

Tentative Due Date: Wednesday 9/16, in class

1. Assume that we draw n data values independently from an exponential distribution with rate parameter λ, i.e. $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$. The exponential distribution is a positive and right skewed distribution function (see the Wikipedia entry here and the R help output here) and the probability distribution function is given by

$$f(x; \lambda) = \lambda \exp(-\lambda x).$$

It is well known that the mean of the exponential distribution is $\text{E}[X] = \mu = \frac{1}{\lambda}$ and the variance is $\text{Var}[X] = \sigma^2 = \frac{1}{\lambda^2}$. The median of the exponential distribution is given by $x_{0.5} = \frac{\log(2)}{\lambda}$. Now a statistician is considering using two statistics to estimate the mean of the distribution $\mu = \frac{1}{\lambda}$. The two statistics are:

- The sample mean $T_1 = \bar{X} = \frac{\sum_{i=1}^n X_i}{n}$
- The sample median divided by $log(2)$, i.e. $T_2 = \tilde{X}/\log(2)$

The statistician knows that because $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.), $\text{Var}(T_1) = \frac{1}{n^2}\sum_{i=1}^n \text{Var}(X_i) = \frac{n\text{Var}(X_1)}{n^2} = \frac{\sigma^2}{n} = \frac{1}{n\lambda^2}$.

(a) Using the central limit theorem, what is the asymptotic distribution of $\sqrt{n}(\bar{X} - \frac{1}{\lambda})$?

(b) Using the lecture notes, what is the asymptotic distribution of $\sqrt{n}(\tilde{X} - \frac{\log(2)}{\lambda})$?

(c) Using the property that $\text{Var}(cT) = c^2\text{Var}(T)$ for any statistic T and scalar c, along with the answer in part a, derive the formula for the asymptotic variance $\text{Var}(T_2)$.

(d) If we were to compare the two statistics, what is the asymptotic relative efficiency $\text{ARE}(T_1, T_2) = \frac{\text{Var}(T_1)}{\text{Var}(T_2)}$?

(e) Based upon part c, which statistic is the most efficient statistic to estimate $\frac{1}{\lambda}$? Broadly speaking, discuss the advantages and disadvantages of both statistics?

(f) Generate 1000 Exp(1) random variables using $x < -\text{rexp}(1000, 1)$ in R. Report what the out the mid-summary values are using R. What can you infer from the sequence of mid-summary values?

2. The following data set is from the Colorado Department of Transportation (CDOT) Monthly Injury Rates, January 1997 – December 2005. Use the R function scan() or the function c() to type in the data. If you have a vector named "A" which contains the data in R, use the function $B = \text{ts}(A, \text{frequency} = 12, \text{start} = c(1997, 1))$ to create a time series object in R.

19.50	16.72	20.92	16.42	21.22	15.40	20.68	14.55	20.23
15.11	20.95	16.68	14.67	16.50	22.15	20.14	18.33	14.20
11.61	22.24	18.75	14.22	15.03	22.07	13.34	12.73	19.23
19.74	19.74	20.60	19.29	18.22	23.65	17.44	13.07	19.00
18.44	17.25	19.19	12.77	14.10	16.69	16.92	21.92	20.84
18.43	19.54	23.61	21.40	28.34	20.43	20.43	15.58	16.58
22.44	14.59	18.70	16.79	14.12	13.67	15.94	24.04	15.42
16.26	17.74	12.37	16.87	16.28	17.97	19.56	13.56	16.13
18.20	17.29	19.38	20.47	16.75	16.69	15.93	14.73	17.83
19.78	15.78	16.17	17.18	13.90	15.33	16.10	12.03	17.92
23.56	11.35	19.10	12.91	18.32	19.24	11.57	14.33	13.60
13.12	11.19	14.33	16.91	13.03	17.32	10.70	12.56	16.04

   (a) Plot this time series data

   (b) Present a stem-and-leaf display for the data. Briefly describe the distribution.

   (c) Present a letter value display for the data. What can you tell about the symmetry of the data from this display?

   (d) Present a QQ plot for the data. What about the normality?

   (e) Did you find any outside values/outliers?

3. About how many outside values should we expect (on Gaussian theory)

   (a) In a single batch of 120 observations?

   (b) In total for two batches of 60?

   (c) In total for batches of 40, 30, 20, 10, 5, 5, 5, and 5?

4. Load package "cluster" and pre-loaded dataset "votes.repub"

```
> library(cluster)
> data(votes.repub)
> help(votes.repub)
votes.repub   package:cluster   R Documentation
Votes for Republican Candidate in Presidential Elections
 A data frame with the percents of votes given to the
 republican candidate in presidential elections from
 1856 to 1976.  Rows: 50 states  Cols: 31 elections
> votes.repub[c(1:3,11),26:31]   # 1952-1976 only
          X1952 X1956 X1960 X1964 X1968 X1972 X1976
Alabama   35.02 39.39 41.75  69.5  14.0  72.4 43.48
Alaska       NA    NA 50.94  34.1  45.3  58.1 62.91
Arizona   58.35 60.99 55.52  50.4  54.8  64.7 58.62
Hawaii       NA    NA 49.97  21.2  38.7  62.5 48.72
```

1. Plot the data by year for (1st, 2nd, 3rd, 4th, 5th) 10 states all on one graph (connect the points by lines). Plot a solid horizontal line at 50% for visual comparison (*anchoring*). Do any lines appear different from the others? Which ones? Compare plots.

2. Would a more sensible grouping of states make sense?

   1–Northeast: CT DE ME MA NH NJ NY PA RI VT

   2–Mid-Atlantic/East-Central: KY MD NC SC TN VA WV

   3–South: AL AR FL GA LA MS OK TX

   4–Midwest: IL IN IA KS MI MN MO NE OH WI

   5–Rockies: CO ID MT ND SD UT WY

   6–West: AK AZ CA HI NV NM OR WA

   Plot 6 groups on 1 page (`par(mfrow=c(2,3))`), also with `abline(h=50)`. What do you notice?

3. Boxplots, QQ-plots of the data (all & by groups)