STATISTICS 641 - ASSIGNMENT #3 - Due NOON (CDT) Wednesday - 9/26/2012

• Read Handouts 4 & 5, Chapter 2 in the Textbook, and the following sections from Chapter 4 in the Textbook: 4.1, 4.2, and 4.3.4.

• Submit for grading the following problems:

I. (15 Points) Let Y have a 3-parameter Weibull distribution, that is, Y has pdf and cdf in the following form with $\alpha > 0,\ \gamma > 0,\ \theta > 0$:

$$f(y) = \begin{cases} \frac{\gamma}{\alpha^\gamma}(y-\theta)^{\gamma-1} e^{-\left(\frac{y-\theta}{\alpha}\right)^\gamma} & \text{for} \quad y \geq \theta \\ 0 & \text{for} \quad y < \theta \end{cases} \qquad F(y) = \begin{cases} 1 - e^{-\left(\frac{y-\theta}{\alpha}\right)^\gamma} & \text{for} \quad y \geq \theta \\ 0 & \text{for} \quad y < \theta \end{cases}$$

(a.) Derive the survival function for Y.

The survival function is given by $S(t) = \Pr[T > t] = 1 - F(t)$. So

$$S(y) = 1 - F(y) = \begin{cases} e^{-\left(\frac{y-\theta}{\alpha}\right)^\gamma} & \text{for} \quad y \geq \theta \\ 0 & \text{for} \quad y < \theta \end{cases}$$

(b.) Derive the hazard function for Y

The hazard function is given by $h(t) = f(t)/S(t)$. The exponential terms cancel, leaving

$$h(y) = \frac{f(y)}{S(y)} = \begin{cases} \frac{\gamma}{\alpha^\gamma}(y-\theta)^{\gamma-1} & \text{for} \quad y \geq \theta \\ 0 & \text{for} \quad y < \theta \end{cases}$$

II. (15 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 mamma 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). The quantile function $Q(u)$ is to be estimated using the 20 data values :

BRAINSIZE - 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			

BRAINSIZE - 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

A software package uses the estimator $\widehat{Q}(u) = Y_{((n-1)u+1)}$ as the estimator of Q(u).

- Calculate the estimates of the Quartiles: of $Q(.25)$, $Q(.5)$, $Q(.75)$ for just the **Small Litter Size** using the given formula.

Read in data; fix it up:

```
small <- read.table("~/Courses/STAT 641b/STAT-641/hw/03/small.txt", quote = "\"")
colnames(small)[1] <- "BRAIN.SIZE"
small <- small[ordered(small$BRAIN.SIZE), ]
head(small)

## [1] 0.42 0.86 0.88 1.11 1.34 1.38
```

Define $\hat{Q}$:

```
Q.hat <- function(vec, u) {
    index <- (length(vec) - 1) * u + 1
    k <- floor(index)
    r <- index - k
    vec[k] + r * (vec[k + 1] - vec[k])
}
```

Do calculations; compare to `quantile` function:

```
Q.hat(small, 0.25)

## [1] 2.61

Q.hat(small, 0.5)

## [1] 4.4

Q.hat(small, 0.75)

## [1] 10.37

quantile(small, 0.25)

##  25%
## 2.61

quantile(small, 0.5)

## 50%
## 4.4

quantile(small, 0.75)

##   75%
## 10.37
```

III. (20 points) Using the data from Problem II for just the **Small Litter Size**, we want to estimate the pdf $f(y)$ for the relative brain weights of the 51 species of mammal.

The kernel density estimate of $f(y)$ is given by

$$\widehat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y - Y_i}{h}\right),$$

Suppose we use the Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$ and a bandwidth of $h = 3$.

(a.) (11 points) Estimate $f(3)$ and $f(16)$ using the kernel density estimator.

(b.) (3 points) Using a relative frequency histogram with bin width of 5, estimate the values of $f(3)$ and $f(16)$.

(c.) (3 points) Which data value provides the smallest contribution to the kernel density estimator at y=16, $\widehat{f}(16)$?

(d.) (2 points) Which data value provides the largest contribution to the kernel density estimator at y=16, $\widehat{f}(16)$?

IV. (20 points) Using the relative Brain Weight data, answer the following questions:

(a.) Produce the following plots of the data: estimates of the pdf, cdf, and quantile function for both Small and Large litter sizes.

(b.) Describe the underlying distribution of the relative brain weights for both Small and Large litter sizes.

(c.) Based on the graphs, what are your conclusions about the relationship between litter size and relative brain weights?

V. (30 Points) **Select** the letter of the **best** answer for each question and provide a short explanation for your selection (20 words or less).

1. The function which provides the most detailed description for the realizations of a random variable is
   A. the cumulative distribution function, cdf $F(\cdot)$
   B. the probability density (mass) function, pdf $f(\cdot)$
   C. the quantile function, $Q(\cdot)$
   D. the survival function, $S(\cdot)$
   E. all the above functions are equivalent

2. A relative frequency histogram having classes of greatly different class widths was used as an estimator of a continuous population pdf. The relative frequency was plotted versus the class intervals. This plot will not be an appropriate estimator of the population pdf because
   A. all the intervals are not the same width.
   B. the area under the curve for each class is not an estimator of the probability of that class.
   C. the area under the curve is not proportional to one.
   D. the relative frequency varies greatly by class width.
   E. In fact it is an unbiased estimator of the pdf.

3. A relative frequency histogram having classes of greatly different class widths was used as an estimator of a continuous population pdf. The relative frequency was plotted versus the class intervals. The plot will result in a graphical distortion. The plot can be corrected by
   A. making all the intervals have the same width.
   B. increasing the sample size.
   C. making sure that the area under the curve adds to one
   D. plotting the relative frequency divided by class width.
   E. In fact there will not be a distortion since it is an unbiased estimator of the pdf.

4. A kernel density estimator was used as an estimator of a continuous population pdf, $f(y)$. The kernel density estimator is generally a vastly improved estimator over a relative frequency histogram (plot of
$\frac{N_i/n}{h_i}$ vs Class i) because
   A. in using the histogram, it is necessary to select the number of bins, bin widths, and their location.
   B. the kernel density estimator makes use of all the data in estimating $f(y)$ whereas the histogram only uses those data values in the same bin as y.
   C. the area under the curve adds to 1 for the kernel density estimator.
   D. there are too many spurious modes using the histogram
   E. all of the above
5. A random sample of n data values is obtained from a process having an absolutely continuous cdf of unknown shape. The metallurgist wants to select the best fitting distribution amongst several candidate cdfs. She decides to select the distribution which has mean and variance most closely matching the corresponding sample mean and variance. The major weakness in this approach is
   A. the mean and variance may be highly inflated by outliers
   B. there are many distribution having the same mean and variance but very different shapes
   C. she should have used robust estimators of the location and scale parameters
   D. the empirical distribution function contains more information about the tails of the distribution than does the mean and variance
   E. the moments of a distribution determine the distribution, hence there is no weakness in the approach
6. The skewness and kurtosis parameters are generally thought to represent the following characteristics of the population cdf, respectively,
   A. the center and spread in the distribution
   B. the heaviness of the tails and deviation from normality of the distribution
7. The median is a trimmed mean with level of trimming equal to
   A. 0%
   B. 50%
   C. 25%
   D. 75%
   E. none of the above
8. The standard deviation is preferred to MAD as a measure of population dispersion when the population distribution
   A. has absolutely no outliers.
   B. has a normal distribution.
   C. has a lognormal distribution.
   D. has a skewed but short-tailed distribution.
   E. cannot be determined with the given information.

9. Alternatives to σ for measuring the dispersion in a distribution are $SIQR$ and MAD. Which of the following statements about these measures are **TRUE**?
   A. All three measures are equal if the pdf for the distribution is symmetric.
   B. $SIQR$ is preferred to MAD if the distribution has very heavy tails
   C. For the normal distribution, $SIQR$ is preferred to MAD
   D. all of the above
   E. none of the above
10. A government study of the average monthly nitrate levels in the Mississippi river, N_t, just prior to its entry into the Gulf of Mexico is modeled as

    $N_t = 22.3 + .6N_{t-1} + e_t$ where $e_t's$ are iid $E[e_t] = 0,\ Var[e_t] = 2.8,\ e_t's$ are independent of $X_t's$

    The mean and variance of N_t is given by
    A. $\mu = 22.3,\ \sigma^2 = 2.8$
    B. $\mu = 22.3,\ \sigma^2 = 4.375$
    C. $\mu = 34.84,\ \sigma^2 = 2.8$
    D. $\mu = 55.75,\ \sigma^2 = 4.375$
    E. The values of μ and σ^2 would change from month to month.