THE UNIVERSITY OF NEW SOUTH WALES

DEPARTMENT OF STATISTICS

MATH3811/MATH3911- Statistical Inference/Higher Statistical Inference

ASSIGNMENT 2

A reminder that assignments count as part of your assessment. Please, declare on the first page that the assignment is your own work, except where acknowledged. State also that you have read and understood the University Rules regarding Academic Misconduct.

To be submitted to your tutor by Thursday, 3pm, 24th May, 2012 **at the latest.**

Math3811: Attempt the first four questions. Math3911: Attempt all questions.

1. Suppose $X_1, X_2, \dots, X_{12}$ are independent and identically distributed random variables from $N(\mu, 4)$ (each with a density $f(x;\mu) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{8}}$).

   a) Show that the joint density of $X_1, X_2, \dots, X_{12}$ has monotone likelihood ratio in $\sum_{i=1}^{12} X_i$.

   b) Derive the UMP unbiased size $\alpha = 0.05$ test of $H_0 : \mu = 1$ versus $H_1 : \mu \neq 1$.

   c) Find the power function of this test (using the cdf of the standard normal or otherwise) and evaluate the power numerically for $\mu = 0, 0.5, 1.5, 2.5,$ and 4. Sketch a graph.

   d) Calculate the density $f_{X_{(3)}}(x)$ of the third order statistic $X_{(3)}$ under H_0.

2. In a sequence of consecutive years $1, 2, \dots, T$, an annual number of high-risk events is recorded by a bank. The random number N_t of high-risk events in a given year is modelled via Poisson(λ) distribution. This gives a sequence of independent counts $n_1, n_2, \dots, n_T$. The prior on λ is Gamma(a, b) with known $a > 0, b > 0$:

$$\tau(\lambda) = \frac{\lambda^{a-1} e^{-\lambda/b}}{\Gamma(a) b^a}, \lambda > 0.$$

   a) Determine the Bayesian estimator of the intensity λ with respect to quadratic loss.

   b) Assume $a = 2, b = 2$. The bank claims that the yearly intensity λ is less than 2. Within the last six years counts were $0, 2, 3, 3, 2, 2$. Test the bank's claim via Bayesian testing with a zero-one loss.

3. Important measures in exploratory data analysis are the *skewness* $\gamma_1 = \frac{E(X-E(X))^3}{Var(X)^{3/2}}$ and the *kurtosis* $\gamma_2 = \frac{E(X-E(X))^4}{Var(X)^2} - 3$. One way of estimating them is by using their empirical counterparts $\hat{\gamma_1} = \frac{\sqrt{n}\sum_{i=1}^n (X_i - \bar{X})^3}{(\sum_{i=1}^n (X_i - \bar{X})^2)^{3/2}}$ and $\hat{\gamma_2} = \frac{n \sum_{i=1}^n (X_i - \bar{X})^4}{(\sum_{i=1}^n (X_i - \bar{X})^2)^2} - 3$, respectively.

   a) Similarly to the `mycorr` function from the bootstrap tutorial script, write down two S-PLUS (or R) functions *myskewness* and *mykurtosis* to get the estimates $\hat{\gamma_1}$ and $\hat{\gamma_2}$.

b) Load Library MASS in S-PLUS and find the data `galaxies` (this variable describes the velocities of 82 galaxies taken in the *Corona Borealis* region (a small constellation in the northern sky). Use your functions to estimate γ_1 and γ_2 for the variable `galaxies`. Verify your results by comparing the estimates with the ones obtained when using the S-PLUS in-built functions *skewness* and *kurtosis* and explain the differences (if any).

c) Bootstrap the $\hat{\gamma_1}$ and $\hat{\gamma_2}$ estimators by using $B = 5000$ replicates and report the resulting 95% confidence intervals using the BCa method.

d) For a normal population, theoretical skewness and kurtosis are both equal to zero. If the 95% confidence interval for either skewness or kurtosis excludes zero, the normality is in doubt. What is your conclusion about the normality of the `galaxies` data? Double check your claim graphically using `qqnorm`. Comment. Include your coding, and the output containing the BCa confidence intervals, in your assignment.

4. We simulate an example to demonstrate the strength of the LTS regression in isolating outliers when there is some idea about the amount of contamination in the data. Suppose your student number contains the seven numbers XXXXXXX , in order that you generate data that is unique to your student number, you should include the student number in the starting seed for random number generation as shown below. After setting the initial seed, generate pairs of observations $(x_i, y_i, i = 1, 2, \ldots, 100)$ of which 70% are scattered around the line $y = 0.8x + 2$ and 30% are clustered around the point (5,2).

```
>set.seed(round(log(XXXXXXX)))
>x70<-runif(70,0.5,4)
>e70<-rnorm(70,mean=0,sd=0.2)
>y70<-2+0.8*x70+e70
>x30<-rnorm(30,mean=5,sd=0.5)
>y30<-rnorm(30,mean=2,sd=0.5)
>x<-c(x70,x30)
>y<-c(y70,y30)
>simuldata<-data.frame(x,y)
...
```

Using the above commands as a starter and the help of SPLUS, do the following:

i) **Graph one**. Plot the `x,y` data to produce a scatterplot. Study the help of the `abline` command and apply it to superimpose three regression lines: the ordinary least squares line, the default M-estimate line and the default LTS line.

ii) **Graph two.** Using the `ltsreg.formula` option you are allowed to override the default value of `quan` (the number of residuals included in the LTS calculations)). Modify the default LTS regression by instructing that only 70 residuals be included in the calculation. Redraw the graph with the new LTS regression line replacing the old one, label properly the resulting regression lines on the new graph.

iii) Compare and comment on the two graphs and and include them in your assignment.

5. The *jackknife* is a general technique for reducing bias in an estimator T_n. In order to "jackknife" T_n, we calculate the n statistics $T_n^{(i)}, i = 1, 2, \ldots, n$ where $T_n^{(i)}$ is calculated just like T_n except that X_i is removed from the sample when calculating. Then the jackknife estimator $JK(T_n)$ of θ is $JK(T_n) = nT_n - \frac{n-1}{n}\sum_{i=1}^{n} T_n^{(i)}$.

a) It is known that $T_n = X_{(n)}$ is biased for the parameter θ of the uniform distribution in $[0, \theta)$, with a bias equal to $-\frac{1}{n+1}\theta$.

i) Check that the bias of $JK(T_n)$ is of smaller magnitude in comparison to $-\frac{1}{n+1}\theta$.
ii) (*) Compare the Mean squared errors of T_n and of $JK(T_n)$. Show that

$$MSE(JK(T_n)) < MSE(T_n)$$

holds for *any sample size* $n > 1$. Hence, in this example, jackknifing has also reduced the MSE of the estimator T_n.

b) Typically, the bias of a "regular" estimator can be expressed, with A, B not depending on n, by $bias(T_n) = \frac{A}{n} + \frac{B}{n^2} + o(\frac{1}{n^2})$. Here $c = o(\frac{1}{n^2})$ means that $\lim_{n\to\infty} cn^2 = 0$. Show that then $bias(JK(T_n))$ does not involve the $\frac{A}{n}$ term hence in general, jackknifing reduces the magnitude of the bias.