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 Wednesday, 11am, 29th May, 2013 **at the latest.**

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

1. Suppose $X_1, X_2, \ldots, X_{10}$ 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 $X_1, X_2, \ldots, X_{10}$ has monotone likelihood ratio in $\sum_{i=1}^{10} X_i$.

   b) Derive the UMP unbiased size $\alpha = 0.1$ 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 the power function for *all* values of μ on the real axis.

   d) Calculate the density $f_{X_{(2)}}(x)$ of the second order statistic $X_{(2)}$ under H_0. Hence find numerically $P(X_{(2)} < 0)$. (You could use the `integrate` function in R or S-PLUS).

2. In a sequence of consecutive years $1, 2, \ldots, T$, 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 $n_1, n_2, \ldots, n_T$. 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 $a = 3, b = 2$. If the counts within the last seven years were $2, 4, 7, 3, 4, 4, 5$ find the estimate of λ for this data.

   c) The bank claims that the yearly intensity λ is less than 4. Test the bank's claim via Bayesian testing with a zero-one loss, using the data from b).

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) Use the above functions to estimate γ_1 and γ_2 for the variable `time` from the `lung` data set in S-PLUS (this variable records survival times of lungcancer patiemts form the Mayo clinic in the USA). Verify your results by comparing the estimates with the ones obtained when using the S-PLUS in-built functions *skewness* and *kurtosis*. Examining the help of these in-built functions, explain the differences (if any).

c) Bootstrap the $\hat{\gamma}_1$ and $\hat{\gamma}_2$ estimators by using $B = 2000$ 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 `time` data? Double check your claim graphically using `qqnorm`. 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 and the lmRob procedures in isolating outliers. Suppose your student number contains the numbers XXXXXXX, in order that you generate data that is unique to your student number, 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 $y = x + 2$ and 30% are clustered around (6,3).

```
>set.seed(round(log(XXXXXXX)))
>x70<-runif(70,0.5,4)
>e70<-rnorm(70,mean=0,sd=0.2)
>y70<-2+x70+e70
>x30<-rnorm(30,mean=6,sd=0.5)
>y30<-rnorm(30,mean=3,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, produce and include in your assignment the following graphs:

i) **Graph 1**. 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. Label the lines appropriately.

ii) **Graph 2.** Using the `ltsreg.formula` option you can override the default value of `quan` (the number of residuals included in the LTS calculations)). Modify the default LTS regression by asking that only 70 residuals be included in the calculation. Redraw the graph with the new LTS regression line replacing the old one.

iii) Suppose however that you did not know the amount of contamination and used 85 residuals instead of 70 in ii) (i.e., some outliers are still influencing the LTS fit). Try the LTS estimator again. Does it deliver a good fit?

Now, the lmRob procedure is claimed to adjust "automatically" to the amount of contamination after starting with a high-breakdown point estimator. Apply the lmRob procedure with its default settings to fit the data and comment on whether or not the claim could be believed to be true. Attach the graph (**Graph 3**).

5. a) 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 $JK(T_n) = nT_n - \frac{n-1}{n}\sum_{i=1}^{n} T_n^{(i)}$.

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

i) Show that $EX_{(n-1)} = \frac{(n-1)\theta}{n+1}$. Using this fact, or otherwise, show 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) The bootstrap-based bias adjustment of T_n happens via $2T_n - E_{\hat{F}}T_n^*$ where the expectation is calculated under the empirical distribution. If $n = 3$, this distribution puts equal weights of $1/3$ at $X_{(1)}, X_{(2)}$ and $X_{(3)}$. Show that the the bootstrap bias-adjusted estimator of θ is then approximately equal to

$$1.2963X_{(3)} - 0.25926X_{(2)} - 0.037X_{(1)}.$$

Hence show that the bootstrap-based bias adjustment has reduced the magnitude of the bias of the original estimator T_3.