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 by Thursday, 10am, 28th May, 2015 **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 $\alpha = 0.05$ test of $H_0 : \mu = 3$ versus $H_1 : \mu \neq 3$.

   c) Find the power function of this test (using the cdf of the standard normal or otherwise) and evaluate the power numerically for $\mu = 1, 1.5, 2.5, 3.5$ and 5. Sketch a graph.

   d) Check directly analytically that $\frac{\partial}{\partial \mu} power(\mu)_{|\mu=3} = 0$ holds for your test.

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 $a = 2, b = 2$. 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) 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 lung-cancer patients 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 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 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.

ii) **Graph 2.** 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) **Graph 3.** Suppose that you did *not* know the amount of contamination and used 90 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 lmRob procedure with its default settings to the data and comment on whether or not the claim could be trusted. Attach the resulting graph and comment.

5. a) Two independent random samples are given: $X_1, X_2, \ldots, X_n$ with a density $f_X(x;\theta) = \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right),\ x > 0$ and $Y_1, Y_2, \ldots, Y_m$ with a density $f_Y(y;\mu) = \frac{1}{\mu}\exp\left(-\frac{y}{\mu}\right), y > 0.$

i) Formulate precisely the Generalized Likelihood Ratio Test (GLRT) for $H_0 : \theta = \mu$ vs $H_1 : \theta \neq \mu$;

ii) Intuitively, closeness of the statistic

$$T = \left(\sum_{i=1}^{n} X_i \right) / \left(\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j \right)$$

to $n/(n+m)$ should indicate that the null hypothesis could be trusted hence a reasonable test that only expolits the data via the value of T should be a reasonable one. Show that the GLRT can indeed be expressed by only using the sample sizes n, m and the value of T.

b) Suppose that two independent samples from two Bernoulli populations are given: a sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from the first population, with a parameter $\theta = \theta_1$, and another independent sample sample $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ form the second population with a parameter θ_2. Show that for the MLE $\hat{h}$ of the ratio $h(\theta_1, \theta_2) = \sqrt{\frac{\theta_1}{\theta_2}}$ we have

$$\sqrt{n}(\hat{h} - h(\theta_1, \theta_2)) \to^d N(0, \frac{\theta_1(1-\theta_2) + \theta_2(1-\theta_1)}{4\theta_2^2}).$$