M3S12 BIOSTATISTICS - EXERCISES 3

1. *(Challenging)* Suppose that $X_1, ..., X_n$ are i.i.d $N(\mu, \sigma^2)$ random variables.

   (a) Find the m.l.e. of $\theta = (\mu, \sigma^2)$, and the Fisher Information matrix for θ.

   (b) Compute the Likelihood Ratio, Rao(Score) and Wald test statistics for testing

   $$\begin{aligned} H_0 &: \mu = 0 \\ H_1 &: \mu \neq 0 \end{aligned}$$

   with σ^2 not specified. For the Score and Wald tests, use the componentwise tests (p. 34 in the Course Notes).

   (c) Compute the above test statistics for testing

   $$\begin{aligned} H_0 &: \left(\mu, \sigma^2\right) = \left(0, \sigma_0^2\right) = \theta_0 \\ H_1 &: \left(\mu, \sigma^2\right) \neq \theta_0 \end{aligned}$$

   using the full test statistics (pp 32-33).

2. The leukaemia count, Y, in a particular area is often assumed to be distributed as a Poisson random variable with mean $N\theta$. Here N denotes the expected number of counts based on the leukaemia rates in the health region in which the area is located and on the population of the area (i.e. the number of people at risk).

   Assuming a gamma prior $Gamma(\alpha, \beta)$ for θ, where $\alpha, \beta > 0$, show that the posterior distribution for θ is a gamma distribution also. Hence determine the posterior mean, mode and variance.

   Consider the following real data; there were $y = 4$ observed leukaemia cases in Seascale, a small village near to a nuclear power plant in Cumbria, NorthWest England. Here the expected number of counts, based on population studies and other demographics, was $N = 0.25$.

   Obtain the posterior distribution under gamma prior distributions with parameters $\alpha = \beta = 0.1$, $\alpha = \beta = 1.0$ and $\alpha = \beta = 10$. Determine the posterior mean, mode and variances in each of these cases and comment.

3. Consider a Bayesian analysis of the data in the 2×2 table given below and studied in Exercises 1, Q4. Recall that the data arise from a case-control study carried out to investigate the relationship between oesophagel cancer (diseased or not, F, F') and alcohol consumption ($\geq 80g/< 79g$ on average per day for exposed/unexposed groups, E, E').

	E	E'
F	96	104
F'	109	666

   Let $\gamma_1 = P(E|F)$ and $\gamma_0 = P(E|F')$. Assume that the prior distribution for each of the unknown γs is the same Beta distribution $Be(\alpha, \beta)$ and determine the posterior distribution for each. For the case $\alpha = \beta = 1$ determine the posterior mean, mode and variance and compare with the pure likelihood-based approach. Does this seem a sensible prior distribution ?

   How could other functionals of the parameters (γ_0, γ_1) can be studied using **simulation-based** methods, that is, if SPLUS can be used to produce a large sample from the posterior distribution for each parameter ?

4. In the Normal Linear Regression model, prove that the Bayesian joint posterior distribution of (β, σ^2) given data vectors (x, y) is multivariate normal-Inverse Gamma, if the prior distribution is also multivariate normal-Inverse Gamma. That is, for data (x_i, y_i) for $i = 1, ..., n$

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, ..., n$$

where $\varepsilon_i \sim N(0, \sigma^2)$, $Y|X, \beta, \sigma^2 \sim N_n\left(X\beta, \sigma^2 1_n\right)$, and where

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad 1_n \text{ is the } n \times n \text{ identity matrix}$$

and

$$\begin{aligned} p_{\beta,\sigma^2}\left(\beta, \sigma^2\right) &= p_{\beta|\sigma^2}\left(\beta|\sigma^2\right) p_{\sigma^2}\left(\sigma^2\right) \equiv N_{K+1}\left(\theta, \sigma^2\Sigma\right) \times IGamma\left(\alpha, \beta\right) \\ \Longrightarrow p_{\beta,\sigma^2|X,Y}\left(\beta, \sigma^2|X, y\right) &= p_{\beta|\sigma^2,X,Y}\left(\beta|\sigma^2 X, y\right) p_{\sigma^2|X,Y}\left(\sigma^2|X, y\right) \\ &\equiv N_{K+1}\left(\theta^*, \sigma^2\Sigma^*\right) \times IGamma\left(\alpha^*, \beta^*\right) \end{aligned}$$

for known prior parameters $\theta, \Sigma, \alpha, \beta$ and posterior parameters $\theta^*, \Sigma^*, \alpha^*, \beta^*$ to be identified.

5. The following data are believed to follow a linear regression model:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

x	0.54	2.03	3.15	3.96	6.25	8.17	11.08	12.44	14.04	14.34	18.71	19.90
y	11.37	11.21	11.61	8.26	14.08	16.25	11.00	14.94	16.91	15.78	21.26	20.25

Explore the relationship between predictor and response using SPLUS; find estimates, standard errors, and confidence intervals, and construct the ANOVA table for a fit. Use the **SPLUS code for Q5** attached.

6. The data set in the file **Oxygen.xls** that you will receive by email, contains the oygen uptake (in ml) and expired ventilation (in litres) for patients in a lung capacity experiment. Explore the relationship between the two variables, assuming a regression model, with the expired ventilation as the response variable.

SPLUS code for Q5

```
#Data
x<-c(0.54,2.03,3.15,3.96,6.25,8.17,11.08,12.44,14.04,14.34,18.71,19.90)
y<-c(11.37,11.21,11.61,8.26,14.08,16.25,11.00,14.94,16.91,15.78,21.26,20.25)

#Plot
plot(x,y)

#Fit the model using the "lm" function
xy.lm<–lm(y∼x)
summary(xylm)

#Plot the line of best fit
abline(xylm$coeff[1],xylm$coeff[2])

#Analysis of Variance using the "aov" function
xy.aov<-aov(y∼x)
xy.aovsummary<–summary(xy.aov)
print(xy.aovsummary)

#xy.aovsummary is a matrix containing the ANOVA table
p.value<-xy.aovsummary[1,5]
print(paste("P-value is :",round(p.value,6)))
```