UNIVERSITEIT VAN DIE VRYSTAAT
UNIVERSITY OF THE FREE STATE

STSB 6810 and 7910

WISKUNDIGE STATISTIEK & AKTUARIËLE WETENSKAP/
MATHEMATICAL STATISTICS & ACTUARIAL SCIENCE

Test 2 — 5 May 2017

MEMORANDUM

TYD/TIME: 220 Minutes PUNTE/MARKS: 40

INSTRUCTIONS:

- Answer all questions in a single Word document but convert the final version to .pdf before submitting (backup the Word document and all working files separately).
- Label questions clearly, as it is done on this question paper.
- All results accurate to 2 decimal places.
- Show all derivations, formulas, code, sources and reasoning.
- Intervals should cover 95% probability unless stated otherwise.
- No communication software, devices or websites may be accessed prior to submission.

Question 1

Counts of employee absenteeism on weekdays are recorded for a month at a factory:

Day	Number Absent
Monday	7
Tuesday	5
Wednesday	4
Thursday	4
Friday	5

Assume a $Dirichlet(0.5, 0.5, 0.5, 0.5, 0.5)$ prior on the vector of unknown proportions $\mathbf{p} = (p_1, p_2, p_3, p_4, p_5)$. Assume a Multinomial likelihood with fixed n.

(a) Simulate 80,000 vectors from the posterior distribution and calculate the probability that absenteeism is higher on Mondays than on Wednesdays. [5]

$$\mathbf{p}|\mathbf{x} \sim Dirichlet(7.5, 5.5, 4.5, 4.5, 5.5) \checkmark$$

```r
rDirichlet <- function(n,alphas) {
# Simulations from the Dirichlet Distribution, according to the method of Wikipedia
k <- length(alphas)
gams <- matrix(rgamma(n*k,c(alphas)),n,k,byrow=TRUE)
gamtotal <- matrix(rowSums(gams),n,k)
sim <- gams/gamtotal
return(sim)
}
sims <- rDirichlet(80000,c(7.5,5.5,4.5,4.5,5.5))
mean(sims[,1]>sims[,3])
```

Sims code correct ✔ ✔ ✔Probability≈ 0.817 ✔

Total for Question 1: 5

Question 2

You are given a set of insurance claims from two locations in thousands of Rands. Assume no excess so that the minimum claim is just above zero. Further note that the last claim is censored since it is not yet fully developed, we only know it is at least 440.

Your task is to fit four specific models and determine which model is most parsimonious. In all cases use $Normal(0, precision = 0.0001)$ priors for location parameters (truncated if necessary) and $Gamma(0.0001, 0.0001)$ priors for scale parameters.

LocationA	190	140	610	380	214	404	285	294	331
LocationB	472	193	242	458	430	274	39	76	440*

(a) Enter the data in R and show that the average without the censored value is 296. [1]

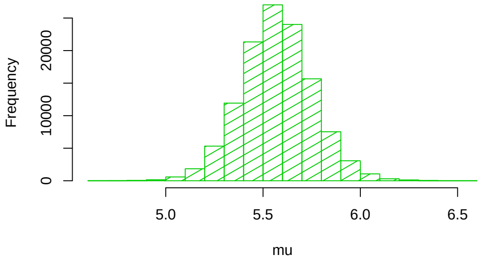
```r
(x <- c(190,140,610,380,214,404,285,294,331,472,193,242,458,430,274,39,76,NA))
mean(x,na.rm=TRUE)
```

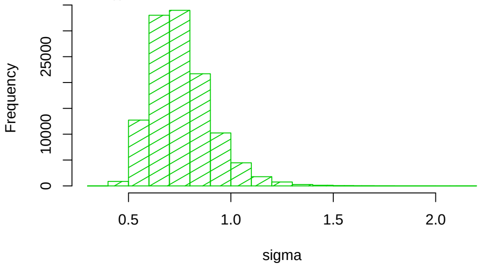
✔

(b) Model 1: assign a LogNormal distribution to all the values, assuming a single mean and variance. Use the ‘bugs’ function to fit the given model. Give histograms approximating the posterior densities of μ and σ. [5]

```
library(R2OpenBUGS)
model1 <- function() {
for (i in 1:17) {
  y[i] ~ dnorm(mu, tau)
}
y[18] ~ dnorm(mu,tau)%_%C(6.086775,)
mu ~ dnorm(0, 0.0001)
tau ~ dgamma(0.0001, 0.0001)
sigma <- 1/sqrt(tau)
}
write.model(model1,’model1.txt’)
BUGSdata <- list(y=log(x))
inits <- function(){list(mu=mean(log(x[1:17])),tau=1/(sd(log(x[1:17]))^2))}
out1 <- bugs(BUGSdata,inits,c(’mu’,’tau’,’sigma’),80000,’model1.txt’)
windows(6,4)
hist(out1$sims.list$mu,main=’’,xlab=’mu’,col=3,angle=30,density=10)
windows(6,4)
hist(out1$sims.list$sigma,main=’’,xlab=’sigma’,col=3,angle=30,density=10)
```

Model code ✔ ✔. A lot of simulations ✔. Output looks neat and correct ✔ ✔.

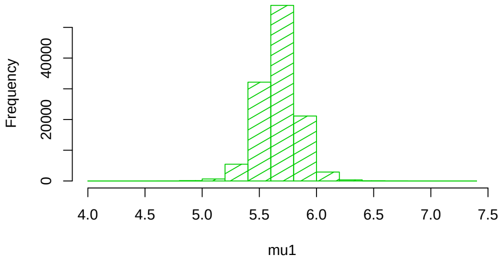


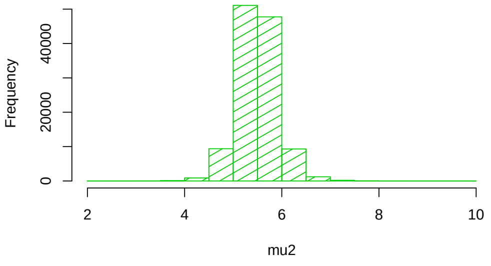


(c) Model 2: assign a LogNormal distribution to all the values, assuming a different mean and different variance for each location (4 parameters). Use the ‘bugs’ function to fit the given model. Give histograms of the two μ parameters. [4]

```
model2 <- function() {
for (i in 1:17) {
  y[i] ~ dnorm(mu[group[i]], tau[group[i]])
}
y[18] ~ dnorm(mu[group[18]], tau[group[18]])%_%C(6.086775,)
for (i in 1:Ngroups) {
mu[i] ~ dnorm(0, 0.0001)
tau[i] ~ dgamma(0.0001, 0.0001)
sigma[i] <- 1/sqrt(tau[i])
}
}
write.model(model2,’model2.txt’)
Ngroups <- 2
BUGSdata <- list(y=log(x),Ngroups=Ngroups,group=rep(1:Ngroups,each=9))
inits <- function(){list(mu=rep(mean(log(x[1:17])),Ngroups),tau=rep(1/(sd(log(x
    [1:17]))^2),Ngroups))}
out2 <- bugs(BUGSdata,inits,c(’mu’,’tau’,’sigma’),80000,’model2.txt’)
windows(6,4)
hist(out2$sims.list$mu[,1],main=’’,xlab=’mu1’,col=3,angle=30,density=10)
windows(6,4)
hist(out2$sims.list$mu[,2],main=’’,xlab=’mu2’,col=3,angle=30,density=10)
```

Model code and initial values changed correctly ✓ ✓. Output looks neat and correct ✓ ✓.





(d) Model 3: assign a Gamma distribution to all the values, assuming a single mean and variance. Use the ‘bugs’ function to fit the given model. Give summary statistics of the parameters. [4]

```
model3 <- function() {
for (i in 1:17) {
  y[i] ~ dgamma(alpha, lambda)
}
y[18] ~ dgamma(alpha, lambda)%_%C(440,)
alpha ~ dnorm(0, 0.0001)%_%T(0,)
lambda ~ dgamma(0.0001, 0.0001)
}
write.model(model3,'model3.txt')
BUGSdata <- list(y=x)
inits <- function(){list(alpha=(mean(x[1:17])/sd(x[1:17]))^2,lambda=mean(x[1:17])/(sd
    (x[1:17])^2))}
out3 <- bugs(BUGSdata,inits,c('alpha','lambda'),80000,'model3.txt')
summary(out3$sims.list$alpha)
summary(out3$sims.list$lambda)
```

```
Min.   1st Qu. Median Mean   3rd Qu. Max.
0.6244 2.3650  2.9480 3.0560 3.6390  9.0770

Min.     1st Qu.  Median   Mean     3rd Qu.  Max.
0.001363 0.007319 0.009406 0.009786 0.011850 0.028740
```

Model code ✔ ✔Initial values ✔. Output looks correct ✔.

(e) Model 4: assign a Gamma distribution to all the values, assuming a different mean and different variance for each location (4 parameters). Use the 'bugs' function to fit the given model. Give summary statistics of the means and variances. [2]

```
model4 <- function() {
for (i in 1:17) {
  y[i] ~ dgamma(alpha[group[i]], lambda[group[i]])
}
y[18] ~ dgamma(alpha[group[18]], lambda[group[18]])%_%C(440,)
for (i in 1:Ngroups) {
alpha[i] ~ dnorm(0, 0.0001)%_%T(0,)
lambda[i] ~ dgamma(0.0001, 0.0001)
mu[i] <- alpha[i]/lambda[i]
sigma2[i] <- alpha[i]/lambda[i]/lambda[i]
}
}
write.model(model4,'model4.txt')
Ngroups <- 2
BUGSdata <- list(y=x,Ngroups=Ngroups,group=rep(1:Ngroups,each=9))
inits <- function(){list(alpha=rep((mean(x[1:17])/sd(x[1:17]))^2,Ngroups),lambda=rep(
    mean(x[1:17])/(sd(x[1:17])^2),Ngroups))}
out4 <- bugs(BUGSdata,inits,c('alpha','lambda','mu','sigma2'),80000,'model4.txt')
summary(out4$sims.list$mu)
summary(out4$sims.list$sigma2)
```

.

```
Min. : 143.4 Min. : 92.89
1st Qu.: 291.5 1st Qu.: 272.20
Median : 318.3 Median : 320.90
Mean : 323.0 Mean : 341.91
3rd Qu.: 348.6 3rd Qu.: 384.40
Max. :1234.0 Max. :4860.00

Min. : 3204 Min. : 5634
1st Qu.: 11350 1st Qu.: 33110
Median : 15780 Median : 51610
Mean : 19863 Mean : 86621
3rd Qu.: 23060 3rd Qu.: 87400
Max. :1483000 Max. :56690000
```

Model code changed correctly ✔. Output looks correct ✔.

(f) Compare the 4 models using the Deviance Information Criterion and draw a conclusion. [2]

```
DICs <- c(out1$DIC,out2$DIC,out3$DIC,out4$DIC)
```

42.22 41.70 226.10 226.90 ✔. Since the second DIC is the smallest we conclude that the LogNormal model with 2 groups is the most parsimonious ✔(best balance of accuracy and complexity).

Total for Question 2: 18

Question 3

At an archery tournament you are asked to model the distances between 9 arrows on a target and the bullseye of the target. Assuming Normal variation in all dimensions you know you can approximate the distances using a Rayleigh distribution, with density $f(x|\tau) = \tau x e^{-0.5\tau x^2}$. τ is the precision parameter and represents the archer's skill. Potentially this approach could be used to more accurately differentiate archers, but for now you only wish to model one sample with $\sum x_i^2 = 160$.

(a) Show that the Jeffreys prior is $\pi(\tau) \propto \tau^{-1}$. [3]

$$g = -\ln f(x|\tau) = -\ln\tau - \ln x + 0.5\tau x^2 \text{ ✔}$$

$$\text{So } \frac{dg}{d\tau} = -\tau^{-1} + 0.5x^2 \text{ and } \frac{d^2g}{d\tau^2} = \tau^{-2} \text{ ✔}.$$

$$\therefore \pi(\tau) \propto \sqrt{E_X[\tau^{-2}]} \text{ ✔} = \tau^{-1}.$$

(b) Show that the log posterior is $\ln\pi(\tau|\mathbf{x}) = (n-1)\ln\tau - 0.5\tau\sum_{i=1}^{n} x_i^2 + c$. [2]

$$\ell = \sum_{i=1}^{n} \ln f(x|\tau) = \sum_{i=1}^{n} \left(\ln\tau + \ln x_i - 0.5\tau x_i^2\right) = n\ln\tau + \sum_{i=1}^{n}\ln x_i - 0.5\tau\sum_{i=1}^{n} x_i^2 \checkmark$$

$$\therefore \ln\pi(\tau|\mathbf{x}) = n\ln\tau - 0.5\tau\sum_{i=1}^{n} x_i^2 - \ln\tau + c \checkmark$$

(c) Derive the posterior mode for τ analytically. [1]

$$\frac{d\ln\pi(\tau|\mathbf{x})}{d\tau} = \frac{n-1}{\tau} - 0.5\sum_{i=1}^{n} x_i^2$$

$$\therefore \hat{\tau} = \frac{2(9-1)}{160} = 0.1 \checkmark$$

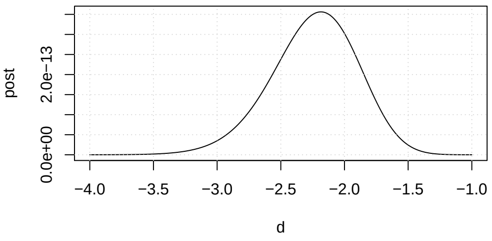
(d) Define $\delta = \ln\tau$ and show that the log posterior is $\ln\pi(\delta|\mathbf{x}) = 9\delta - 80e^{\delta} + c$. [1]

Since $\delta = \ln\tau$ we have that $\tau = e^{\delta}$ and $\left|\frac{d\tau}{d\delta}\right| = e^{\delta}$.

$$\therefore \ln\pi(\delta|\mathbf{x}) = 8\delta - 80e^{\delta} + \ln e^{\delta} + c \checkmark$$

(e) Draw a graph of the posterior in δ and determine crude estimates of the mean and standard deviation. [3]

```
d <- seq(-5,0,0.001)
post <- exp(d)^9*exp(-80*exp(d))
plot(d,post,type='l')
grid()
post <- post/sum(post)
(m <- sum(post*d))
(s <- sqrt(sum((d-m)^2*post)))
```

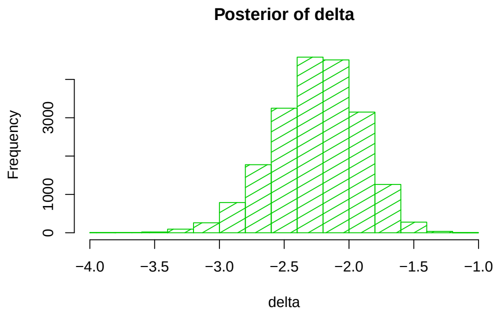


✓ ✓Mean about -2.2 and sd about 0.34. ✓

(f) Draw δ values from the posterior using the Metropolis method. Work on the log scale and use jumps that are Normal with the standard deviation you calculated. Use the above mode as a starting point. Discard the first 10000 values as burn-in, and drop every second value, to obtain 20000 final simulations. Draw a histogram of the simulations to visually compare it to the graph you drew earlier. [7]

```
lp <- function(d) { 9*d-80*exp(d) }
sims <- rep(0,60000)
d0 <- -2.2
lp0 <- lp(d0)
counter <- 0
while (counter < 60000) {
d1 <- rnorm(1,d0,0.34)
lp1 <- lp(d1)
if (lp1 - lp0 > log(runif(1))) {
  counter <- counter + 1
  sims[counter] <- d1
  d0 <- d1
  lp0 <- lp1
}
}
sims <- sims[(1:30000)*2]
sims <- sims[-(1:10000)]
windows(6,4)
hist(sims,main='Posterior of delta',xlab='delta',col=3,angle=30,density=10)
```

Sims code correct ✔ ✔ ✔ ✔ ✔Burn-in and thinning ✔Graph ✔



(g) **STSB7910 (Masters) only [5 marks]:**
Simulate 200 such samples from a Rayleigh distribution with $\tau = 1$ and apply the above procedure to each one. For each sample calculate a 95% interval for τ and finally calculate the coverage over the 200 samples. []

Simulating from distribution correct ✔ ✔Check new mean and sd. ✔Apply to each sample correctly. ✔Reasonable coverage. ✔

Total for Question 3: 17

Total half marks on memo = 90 vs. 80 = Double total margin points (=40).