KU LEUVEN

Concepts of Bayesian Data Analysis Assignment

Frida Trepça
Lennert Vanhaeren
Jonas Kasimir van Nifterik
Joshua White
Axl Wynants

Academic year: 2022-2023

1.3: Obstetrics

This dataset contains birth weight of infants along with the smoking status of their mother during pregnancy. The nonsmoker group contains 7 observations and the smoker group contains 8 observations.

Table 1: Obstetrics dataset

W	Smoker
7.5	0
6.2	0
6.9	0
7.4	0
9.2	0
8.3	0
7.6	0
6.2	1
6.8	1
5.7	1
4.9	1
6.2	1
7.1	1
5.8	1
5.4	1

Birth weights are assumed to be normally distributed within each group.

1.3.1

The posterior distributions of μ_1 (mean birth weight of infants of nonsmokers) and μ_2 (mean birth weight of infants of smokers) are derived analytically using a non-informative prior. Birth weights are assumed normal within each group. However, since both the mean and variance of the groups are unknown, this is actually a multi-parameter problem. Because this project concerns the means, we are only interested in the marginal distribution of μ. Under these conditions it can be derived that the marginal posterior for μ with a non-informative prior is a scaled and shifted t-distribution.

$$p(\mu|\boldsymbol{y}) = t_{n-1}\left(\bar{y}, \frac{s^2}{n}\right) \tag{1}$$

This distribution is implemented in the `metRology` package in R. Plotting these distributions leads to the analytical posteriors for the two groups as depicted in figure 1.

1.3.2

Summary measures of these distributions can also be derived analytically

Table 2: Summary measures for the analytical posterior densities of μ_1 and μ_2

	Mean	**SD**	**Equal tail 2.5%**	**Equal tail 97.5%**
$\boldsymbol{\mu_1}$	7.5857	0.4452	6.4965	8.6750
$\boldsymbol{\mu_2}$	6.0250	0.2998	5.3161	6.7339

Note that since the posterior is symmetric and unimodal, the posterior median and posterior mode are equal to the posterior mean. For the same reason, it doesn't matter whether HPD or equal tail intervals are used, as they lead to the same result. However these intervals are calculated using the quantiles of the posterior so we refer to them as equal tail intervals.

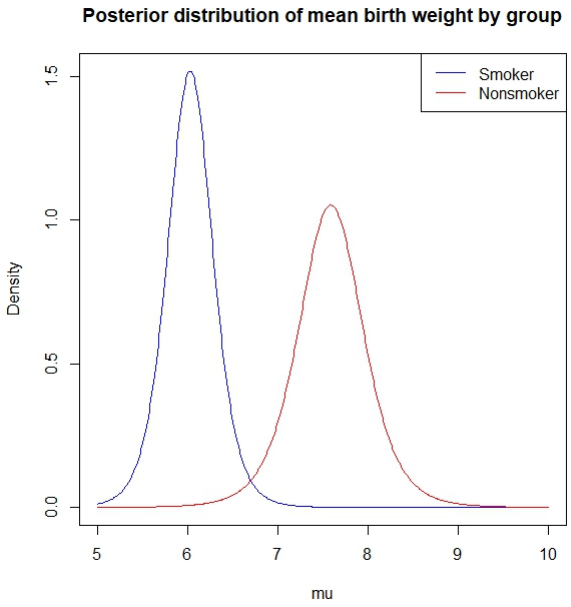

Figure 1: Analytical posteriors of μ_1 and μ_2

1.3.3

In general, linear combinations of t-distributed random variables are not themselves t-distributed. Since the posterior of the difference doesn't seem to have a straightforward analytical form, we turn to sampling to solve this problem. Random variates were generated from the previous analytical posteriors. Their difference was computed to obtain a sample from the posterior distribution of the difference in means.

$$\mu_1 - \mu_2$$

This results in the following empirical distribution for $n = 100,000$.

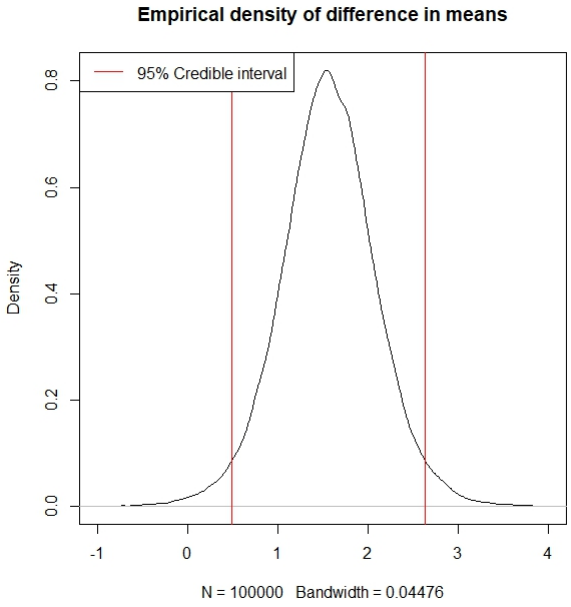

Figure 2: Sampled posterior of $\mu_1 - \mu_2$ with 95% credible interval

With summary measures derived from the sample

Table 3: Summary measures for the posterior density of $\mu_1 - \mu_2$

	Mean	SD	Equal tail 2.5%	Equal tail 97.5%
$\boldsymbol{\mu_1 - \mu_2}$	1.5610	0.5377	0.4932	2.6255

Because 0 is outside of the 95% credible interval, we can conclude that there is an association between smoking status and birthweight.

1.3.4

After analytically deriving the posterior distribution and its summary measures in the questions above, it would be interesting to obtain and check an MCMC sample for the problem. We do this by running a jags code.

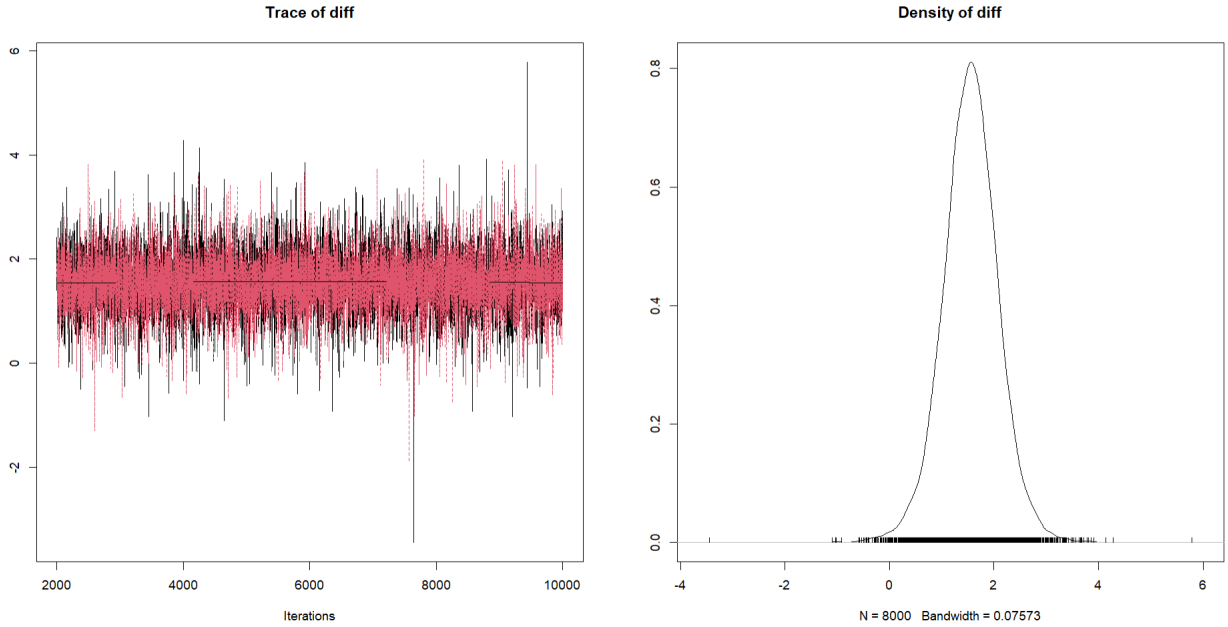

Figure 3: The Trace and Density of $\mu_1 - \mu_2$

In Figure 3 above, we see that the specific path for the Markov chain of the difference in means is shown. We can see from the trace plot of the difference between the two means $(\mu_1 - \mu_2)$, that the chain is relatively stable and does not appear to contain a pattern, so the chain doesn't stay in one place for too long. The trace plot is relatively stable. Therefore, we can conclude that both of the chains have converged to their stationary distribution. This means that a burn-in of 2,000 MCMC iterations is enough.

1.3.5

As seen on the above trace plot, the auto-correlation seems to be low. However, to see the convergence of MCMC chains in more detail, we need to properly test for it. In the Figure 4 below, we see, via the auto-correlation plot, that the latter is quite low, which normally suggests good convergence of the MCMC chains.

Additionally, we check the Gelman-Rubin diagnostic plot. It measures whether there is significant difference between the variance within several chains, called the scale reduction factor. A factor of 1 means that between variance and within chain variance are equal. As a rule of thumb we say that values below 1.1 are allowed. For our model the scale reduction factors both are 1. Therefore we can say the model has converged. The Gelman-Rubin plot gives the scale reduction factors for each parameter and shows the development of the factor over time (in chain-steps). This allows us to see whether the chains are stable or not, we expect this to happen after the burn-in part of the chain (2000 for our model).

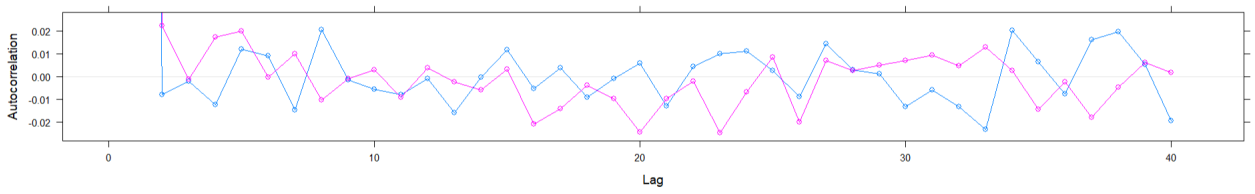

Figure 4: ACF Plot

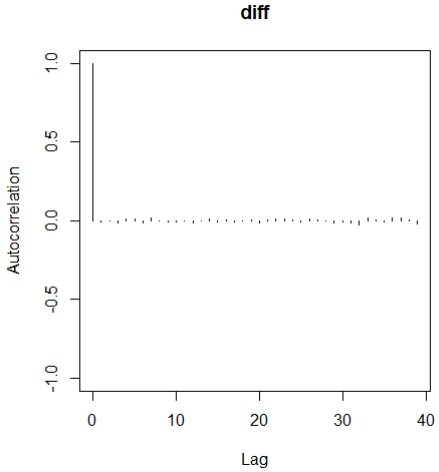

Figure 5: Autocorrelation of the MCMC chain

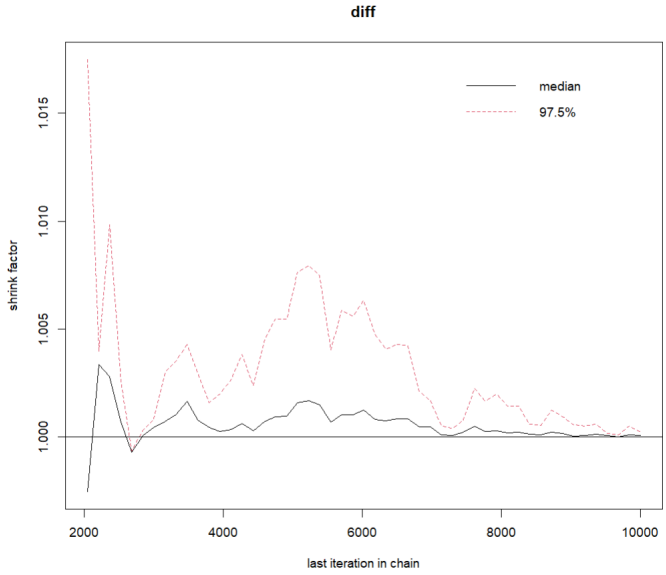

Figure 6: The Gelman Rubin Diagnostic Plot

Additionally, we check for the effective size of the MCMC of the mean difference, which measures the information content, or effectiveness of a sample chain. It results to be equal to 16412.92, meaning that

the chain has a high information content. This indicates again towards good convergence of the chain.

1.3.6

Next, in a comparison of the summary measures obtained from the MCMC chain to those derived analytically, in questions 1-3, we see that they do not differ by much from each other (in fact we see they are very close to each other).

Table 4: Summary measures for the MCMC Chain of $\mu_1 - \mu_2$

	Mean	**SD**	**Equal tail 2.5%**	**Equal tail 97.5%**
$\boldsymbol{\mu_1 - \mu_2}$	1.5594	0.5347	0.4853	2.6012

The density appears symmetrical from the plot but without an analytical form this cannot be confirmed for sure, hence we cannot claim that the results of HPD and equal tail procedures would be the same. Here we give the equal tail interval since it was used in the previous questions.

0 remains outside the 95% credible interval, which means we can keep the (previous) conclusion that there is an association between smoking status and birthweight.

1.3.7

It is also interesting to find out what the relative difference in weight is between smokers and non-smokers, which will be defined as:

$$\Delta = \frac{\mu_1 - \mu_2}{\mu_1}$$

Due to this formula, it is to note that when setting up the code, we can not initialize μ_1 to 0, as dividing by the latter would not make sense. Thus, we list it as 0.1. After initializing the new MCMC model (again via jags code), the below figure shows the trace and density of the relative difference.

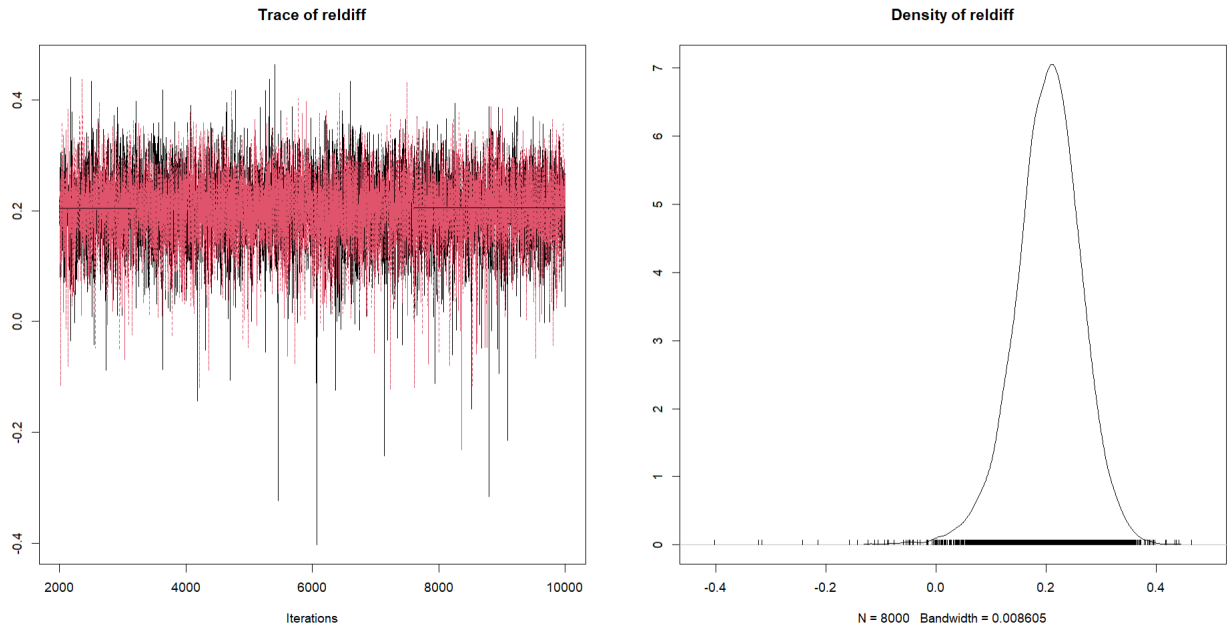

Figure 7: The Trace and Density of Δ

As seen from the figure, the posterior density seems to be symmetrically distributed. Additionally, from the trace of the relative difference MCMC chains, we note that the auto-correlation is also low, as the plot does not seem to stay in one place for too long, so it is relatively stable.(The chains converge). Additionally, in table 5 we will see the summary statistics of the relative difference of the means.

As seen from the table 5, the mean relative difference according to MCMC is around 0.2035. Additionally the HPD interval for the relative difference of the means would be [0.0804; 0.3218].

Table 5: Summary measures for the posterior density of $\Delta = \frac{\mu_1 - \mu_2}{\mu_1}$

	Mean	SD	Equal tail 2.5%	Equal tail 97.5%
Δ	0.2035	0.0619	0.0733	0.3182

2.1: Dose-response model

2.1.1

Parameters α and β from the following logistic regression model

$$\text{logit}(\pi) = \alpha + \beta d, \tag{2}$$

where d is the DYME dosage, were estimated using Gibbs sampling. Specifically, two Markov chains were run, with 10000 iterations each (including a burn-in of 1000 iterations). The starting values for (α, β) were $(-2, 0.1)$ for one chain and $(4, -0.1)$ for the other. Both parameters were given a uniform (vague) prior distribution on the interval $[-100, 100]$.

The convergence of the chains was determined using the method of Gelman, which uses a scale reduction factor to test whether inferences from each individual chain are different from inferences using all available information [1]. This method uses Bayesian credible intervals for the convergence statistic, with wide intervals indicating that more simulations are needed to reach convergence. The intervals for α and β were $[1.00, 1.00]$ and $[1.00, 1.01]$, respectively, both of which indicate that the chains converged.

2.1.2

Figure 8 shows graphical results of the two MCMC chains. Specifically, the paths followed for both α and β are shown for each chain. The posterior densities are also shown in the figure, both symmetrically distributed. Note that the posterior distributions have been obtained by combining both chains. Summary measures for each of the densities is given in Table 6.

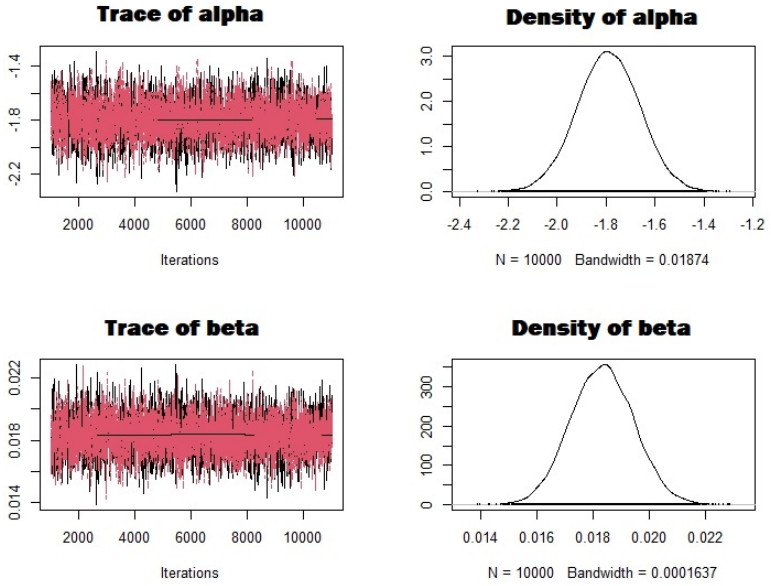

Figure 8: Trace plots of the two chains for α and β with red and black indicating the two chains, and plots of the posterior densities of α and β.

From the densities in Figure 8 and the summary measures in Table 6, the visual observation that the posterior densities are symmetric is supported by the fact that the mean and median posterior values agree, for both α and β. This is more evidence that the number of samples is sufficient to make valid

Table 6: Summary measures for the posterior densities of α and β obtained by combining both chains. The headings HPD 2.5% and HPD 97.5% refer to the upper and lower estimates for 95% highest posterior density intervals of each parameter.

	Mean	**SD**	**Median**	**HPD 2.5%**	**HPD 97.5%**
$\boldsymbol{\alpha}$	-1.7918	0.1303	-1.7889	-2.0433	-1.5324
$\boldsymbol{\beta}$	0.0183	0.0011	0.0183	0.0162	0.0207

conclusions about the parameters. From the trace plots, it is seen that each iteration is relatively independent from previous iterations, since the chain does not stay in one location for many consecutive iterations.

The posterior summary measures, such as the mean and median, represent parameter estimates for model (2), on the logit scale. Both parameters are significantly different from zero, according to their HPD intervals. Using the posterior mean, $\hat{\alpha} = -1.7918$ and $\hat{\beta} = 0.0183$. The intercept, α, can be interpreted as the log-odds of malformation at a dose of zero. The slope parameter β represents the increase in the log-odds of malformation for a unit increase in dose of DYME. This parameter can also be interpreted as an odds ratio: a 100 unit increase in dose is estimated to increase the odds of malformation by a factor of $\exp(100 \cdot 0.0183) = 6.23$.

2.1.3

Figure 9 shows the probability of malformation predicted by the fitted model using the mean of each parameter's posterior distribution as the estimate (see Table 6), along with the observed proportions of malformation at specific doses. The predicted probability of malformation was obtained using the following equation:

$$\pi(d) = \frac{\exp(\alpha + \beta d)}{1 + \exp(\alpha + \beta d)} = \frac{\exp(-1.7918 + 0.0183 \cdot d)}{1 + \exp(-1.7918 + 0.0183 \cdot d)}, \tag{3}$$

which models the probability of malformation, π, as a function of the dose of DYME, d.

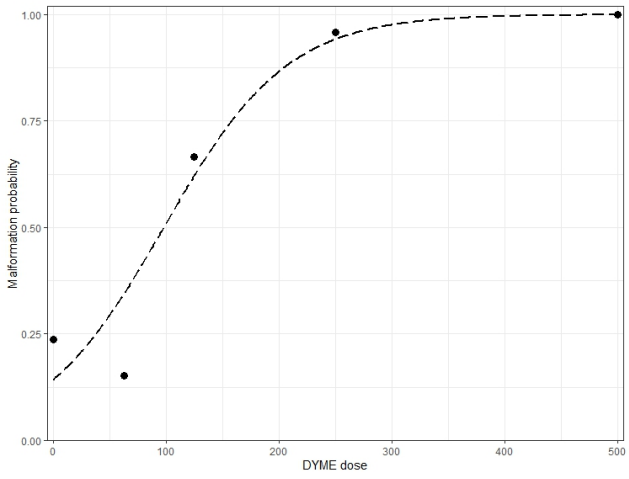

Figure 9: Fitted malformation probabilities for the range of DYME dosages (dotted line) and observed malformation probabilities for previously tested dosages (bold points) for model 2.

2.1.4

A safe level of exposure can be defined as a dose corresponding to a very small increase in excess risk of q, e.g. $q = 0.05$. This is called the Benchmark dose (BMD). The posterior median BMD_M of the safe level of exposure for DYME corresponding with an excess risk of $q = 0.05$ is equal to:

$$\text{BMD}_M \approx 17.1405$$

We obtain a sample of BMD-values by computing the following quantity, using each sampled pair of α and β values from section 2.1.1:

$$\text{BMD} = \frac{\text{logit}(q^*) - \alpha}{\beta}$$

with $q* = q(1 - P(0)) + P(0)$ and for $P(0)$, we fill in $\pi(0)$. The histogram of the sampled BMD-values is shown in figure 10. The median was calculated by taking the mean of these sampled values.

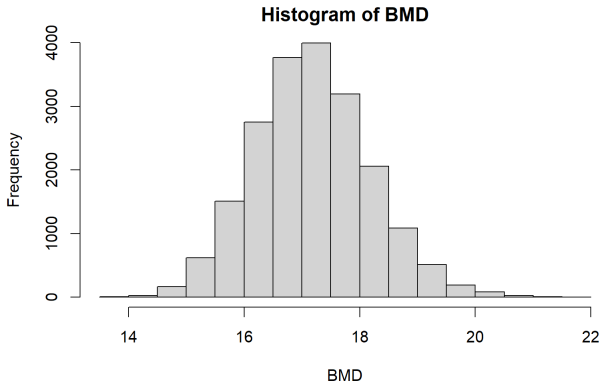

Figure 10: Histogram of sampled BMD-values, based on sampled values for α and β.

2.1.5

A safe level of exposure can, alternatively, be obtained from a threshold model:

$$y \sim \text{binomial}(N, \pi) \tag{4}$$

$$\text{logit}(\pi) = \alpha + \beta(d - \tau)I(d > \tau) \tag{5}$$

with τ the threshold dose below which there is no excess risk.

Once again, two Markov chains were run with 10000 iterations each and a burn-in of 1000 iterations. The starting values for (α, β, τ) were (-2, 0.1, 100) and (4, -0.1, 200) respectively. So we keep the same starting values for α and β as when we fitted this model without τ (to make the comparison between boths models fair). In the same manner, we use the same (vague) uniform priors to α and β on the interval [-100, 100]. For τ, we use a vague uniform prior as well, but on the interval [0, 500]. This is because the dosage can not be negative and the maximal dosage in our dataset is 500.

Figure 11 shows graphical results of the two MCMC chains. On the left, the paths followed for the parameters (α, β and τ) are shown for each chain. On the right, the posterior densities are shown. All of them seem symmetric, which is confirmed by how close the posterior medians are to the posterior means. Summary measures for each of the densities are given in Table 7.

Table 7: Summary measures for the posterior densities of α, β and τ obtained by combining both chains. The headings HPD 2.5% and HPD 97.5% refer to the upper and lower estimates for 95% highest posterior density intervals of each parameter.

	Mean	**SD**	**Median**	**HPD 2.5%**	**HPD 97.5%**
$\boldsymbol{\alpha}$	-1.2930	0.1122	-1.2911	-1.5202	-1.0796
$\boldsymbol{\beta}$	0.0272	0.0021	0.0271	0.0233	0.0314
$\boldsymbol{\tau}$	61.0357	4.6504	61.5055	51.2350	70.2527

The fact that the posterior distributions are symmetric is evidence that the number of samples is sufficient to make valid conclusions about the parameters. The trace plots in Figure 11 show the iterations are independent from each other, since the chain moves quickly through the posterior distribution. The convergence of the chains was once again determined using the method of Gelman. The point estimates

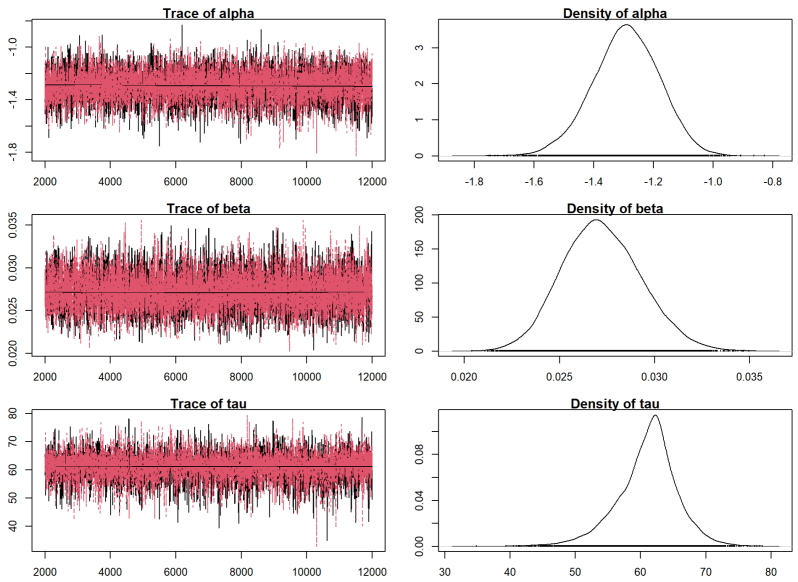

Figure 11: Trace plots of the two chains for α, β and τ with red and black indicating the two chains, and plots of the posterior densities of α, β and τ.

for the scale reduction factor are 1 for all three of the parameters. The intervals for α and β are [1.00, 1.00], indicating convergence. The interval for τ is [1.00, 1.01], also indicating convergence.

The posterior summary measures, such as the mean and median, represent parameter estimates for model (5), on the logit scale. Using the posterior mean, $\hat{\alpha} = -1.2930$, $\hat{\beta} = 0.0272$ and $\hat{\tau} = 61.0357$. The intercept, α, can be interpreted as the expected log-odds of malformation at a dose smaller than or equal to the threshold dose, i.e. $0 \leq d \leq \tau$. The slope parameter β, on the other hand, represents, ceteris paribus, how the log-odds of malformation change for a unit increase in dose of DYME when the dose is above the threshold value τ. It can also be interpreted as an odds ratio: if $d > \tau$, a 100 unit increase in dose is estimated to increase the odds of malformation by a factor of $\exp(100 \cdot 0.0272) = 5.16$.

The values of α and β only change by a small amount compared to the model without a threshold, i.e. model 2. Therefore, they both remain significantly different from zero (0 is not in the HPD CI). The main difference is the change in interpretation of α, which is now the expected log-odds of malformation at a dose **smaller than or equal to the threshold dose** (which is approximately 61), whereas before it was the expected log-odds at a dose of zero. So up until a dose of larger than ≈ 61, the probability of malformation will be constant. This results in a different plot for the predicted probability of malformation for a certain DYME dose, see Figure 12.

The predicted probabilities of malformation for the plot were obtained using the following equation:

$$\pi(d) = \frac{\exp\left(\alpha + \beta(d-\tau)I(d>\tau)\right)}{1+\exp(\alpha+\beta(d-\tau)I(d>\tau))} = \frac{\exp(-1.2930 + 0.0272 \cdot (d-61.0357)I(d>61.0357))}{1+\exp(-1.2930+0.0272\cdot(d-61.0357)I(d>61.0357))},$$

meaning that we have:

$$\begin{cases} \pi(d) = \dfrac{\exp(-1.2930)}{1+\exp(-1.2930)} & \text{for } d < 61.0357 \\ \pi(d) = \dfrac{\exp(-1.2930+0.0272\cdot(d-61.0357))}{1+\exp(-1.2930+0.0272\cdot(d-61.0357))} & \text{for } d > 61.0357 \end{cases} \tag{6}$$

which can be simplified to:

$$\begin{cases} \pi(d) \approx 0.2147 & \text{for } d < 61.0357 \\ \pi(d) \approx \dfrac{\exp\left(-2.9522 + 0.0272 \cdot d\right)}{1+\exp\left(-2.9522+0.0272\cdot d\right)} & \text{for } d > 61.0357 \end{cases} \tag{7}$$

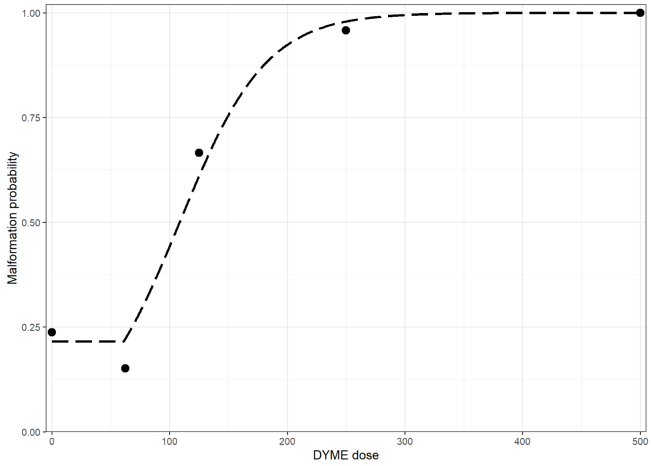

Figure 12: Fitted malformation probabilities for the range of DYME dosages (dotted line) and observed malformation probabilities for previously tested dosages (bold points) for the threshold model, i.e. model 5.

References

[1] Stephen P. Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

Appendix

Code

QA

```
# Let's define the two groups and their sample birth weights
nonsmoker <- c(7.5, 6.2, 6.9, 7.4, 9.2, 8.3, 7.6)
smoker <- c(6.2, 6.8, 5.7, 4.9, 6.2, 7.1, 5.9, 5.4)

# Question 1
# Marginal posterior of the mean is a scaled, shifted t-distribution

library(metRology)

mean_nonsmoker <- mean(nonsmoker)
mean_smoker <- mean(smoker)

n_nonsmoker <- length(nonsmoker)
n_smoker <- length(smoker)

sd_nonsmoker <- sqrt(var(nonsmoker))
sd_smoker <- sqrt(var(smoker))


mu <- seq(5, 10, 0.01)
likelihood_nonsmoker <- dt.scaled(mu, df = n_nonsmoker - 1, mean = mean_nonsmoker,
sd = sd_nonsmoker/sqrt(n_nonsmoker))
likelihood_smoker <- dt.scaled(mu, df = n_smoker - 1, mean = mean_smoker,
sd = sd_smoker/sqrt(n_smoker))


posterior_nonsmoker <- dt.scaled(mu, df = n_nonsmoker - 1, mean = mean_nonsmoker,
sd = sd_nonsmoker/sqrt(n_nonsmoker))
posterior_smoker <- dt.scaled(mu, df = n_smoker - 1, mean = mean_smoker,
sd = sd_smoker/sqrt(n_smoker))


plot(mu, likelihood_nonsmoker, type = 'l', col = 'blue')
lines(mu, posterior_nonsmoker, type = 'l', col = 'red')
# As the posterior and likelihood are the same, the two lines overlap


# The two posteriors in the same plot
plot(mu, posterior_smoker, type = 'l', col = 'blue', ylab = 'Density',
     main = 'Posterior distribution of mean birth weight by group')
lines(mu, posterior_nonsmoker, type = 'l', col = 'red')
legend("topright", c('Smoker', 'Nonsmoker'), col = c("Blue", "Red"), lty = c(1,1))

v1 = n_nonsmoker - 1
v2 = n_smoker - 1

# Question 2


# The posterior mean is equal to the mean of the observed data (= posterior median and mode
#bc it's symmetrical and unimodal)
mean_nonsmoker
mean_smoker
```

```r
# The posterior variance is v/(v-2) the variance of the observed data divided by
#the sample size due to the properties of the t-dist
postsd1 <- sqrt((v1 / (v1 - 2)) * var(nonsmoker)/n_nonsmoker)
postsd2 <- sqrt((v2 / (v2 - 2)) * var(smoker)/n_smoker)

#Credible intervals
nonsmoker_lower <- mean_nonsmoker - qt(df = v1,0.975)*postsd1
nonsmoker_upper <- mean_nonsmoker + qt(df = v1,0.975)*postsd1

smoker_lower <- mean_smoker - qt(df = v2,0.975)*postsd2
smoker_upper <- mean_smoker + qt(df = v2,0.975)*postsd2

nonsmoker_int <- cbind(nonsmoker_lower, nonsmoker_upper)
nonsmoker_int

smoker_int <- cbind(smoker_lower, smoker_upper)
smoker_int


# Question 3

#Sampling
samplenonsmoker <- rt.scaled(100000, df = v1,mean_nonsmoker, sd_nonsmoker/sqrt(n_nonsmoker))
samplesmoker <- rt.scaled(100000, df = v2, mean_smoker, sd_smoker/sqrt(n_smoker))
samplediff <- samplenonsmoker - samplesmoker

ddiff = density(samplediff)
credint <- quantile(samplediff, c(0.025, 0.975))
plot(ddiff, main = 'Empirical density of difference in means', xlim = c(-1, 4))
abline(v = credint[1], col = "red")
abline(v = credint[2], col = "red")
legend("topleft", c('95% Credible interval'), col = c("Red"), lty = c(1))

mean(samplediff)
sqrt(var(samplediff))

#Credible interval for difference

diff_int <- cbind(credint[1], credint[2])
diff_int

#0 is not within the credible interval so we can conclude there is an association


# Question 4

library(readr)
library(coda)
library(runjags)
library(MCMCvis)
library(ggmcmc)
library(basicMCMCplots)
library(rjags)


Nchains <- 2
```

```
model.data <- list('nonsmoker' = nonsmoker, 'smoker' = smoker, 'N_nonsmoker' = n_nonsmoker,
                   'N_smoker' = n_smoker)

model.inits <- model.inits <- list(mu1 = 0, mu2 = 0, tau1 = 1, tau2 = 1)

cat("model
  {
    for (i in 1:N_nonsmoker){
      nonsmoker[i] ~ dnorm(mu1, tau1)

    }
    for (i in 1:N_smoker){
    smoker[i] ~ dnorm(mu2, tau2)
    }
    mu1 ~ dnorm(0, 1.0E-6)
    mu2 ~ dnorm(0, 1.0E-6)
    tau1 ~ dgamma(1.0E-3, 1.0E-3)
    tau2 ~ dgamma(1.0E-3, 1.0E-3)
    diff <- mu1 - mu2
  }", file="meandiff.txt")

jags <- jags.model('meandiff.txt',
                   data = model.data,
                   inits = model.inits,
                   n.chains = Nchains)

update(jags,2000)
diff.sim <- coda.samples(jags,
                         c('diff'),
                         n.iter=8000,
                         thin=1)

print(diff.sim,digits=3)
plot(diff.sim) #Trace doesn't appear to contain a pattern, low autocorrelation which is good
diff.mcmc <- as.mcmc.list(diff.sim)


# Question 5
#Convergence tests

acfplot(diff.mcmc)
autocorr.plot(diff.mcmc) #Low autocorrelation suggests good convergence

gelman.diag(diff.mcmc)
gelman.plot(diff.mcmc,ask=FALSE) #Gelman-Rubin statistic is 1 -> good convergence

effectiveSize(diff.mcmc) #Effective size very high, = real size, good convergence again


# Question 6

plot(diff.mcmc)
summary(diff.mcmc)
HPDinterval(diff.mcmc) #0 still isn't in the interval, interval corresponds to the analytical one


# Question 7
#Can't initialize mu1 to 0 because that would mean dividing by 0 when calculating reldiff
```

```
model.inits2 <- list(mu1 = 0.1, mu2 = 0.1, tau1 = 1, tau2 = 1)

cat("model
  {
    for (i in 1:N_nonsmoker){
      nonsmoker[i] ~ dnorm(mu1, tau1)

    }
    for (i in 1:N_smoker){
    smoker[i] ~ dnorm(mu2, tau2)
    }
    mu1 ~ dnorm(0, 1.0E-6)
    mu2 ~ dnorm(0, 1.0E-6)
    tau1 ~ dgamma(1.0E-3, 1.0E-3)
    tau2 ~ dgamma(1.0E-3, 1.0E-3)
    reldiff <- (mu1 - mu2)/mu1
  }", file="meanreldiff.txt")

jags2 <- jags.model('meanreldiff.txt',
                    data = model.data,
                    inits = model.inits2,
                    n.chains = Nchains)

update(jags2,2000)
reldiff.sim <- coda.samples(jags2,
                            c('reldiff'),
                            n.iter=8000,
                            thin=1)

print(reldiff.sim,digits=3)
plot(reldiff.sim)
reldiff.mcmc <- as.mcmc.list(reldiff.sim)

summary(reldiff.mcmc)
HPDinterval(reldiff.mcmc)
#Mean relative difference according to MCMC is ~0.19-0.20
```

QB

```
library(rjags)
library(coda)
library(tidyverse)
library(runjags)
library(gtools)

y=c(67,34,193,250,141)
n=c(282,225,290,261,141)
x=c(0,62.5,125,250,500)

iterations <- 10000
burnin <- 1000
chains <- 2

model_code <- "model
  {
    #likelihood
    for (i in 1:5){
```

```
    y[i] ~ dbinom(p[i],n[i])
    logit(p[i]) = alpha + beta*x[i]
    }
    #priors
    alpha ~ dunif(-100,100)
    beta ~ dunif(-100,100)
  }"
cat(model_code, file = "model.txt")

model.fit <- jags.model(file="model.txt",
          data=list(n=n,y=y,x=x), n.chains = chains,
          inits=list(list(alpha=-2,beta=0.1),
                     list(alpha=4,beta=-0.1)))
model.samples <- coda.samples(model.fit, c("alpha", "beta"), n.iter=iterations)
summary(window(model.samples, start = burnin))
plot(model.samples, trace=TRUE, density = TRUE)
gelman.diag(model.samples,confidence=0.95)

combined.samples = combine.mcmc(model.samples)
HPDinterval(combined.samples,prob=0.95)

## Observed vs. fitted malformation probability plot
observed=data.frame(dose_obs=c(0,62.5,125,250,500),
    p_obs=c(67/282,34/225,193/290,250/261,141/141))
fitted=data.frame(dose_fitted=0:500,p_fitted=exp(-1.79177+0.01832*dose_fitted)/
    (1+exp(-1.79177+0.01832*dose_fitted)))
plot = ggplot()+geom_point(data=observed,aes(x=dose_obs,y=p_obs),size=3)+
  geom_line(data=fitted,aes(x=dose_fitted,y=p_fitted),lwd=1,linetype="longdash")+
  theme_bw()+xlab("DYME dose")+ylab("Malformation probability")+
  scale_x_continuous(expand = c(0, 0), limits = c(-5,505)) +
  scale_y_continuous(expand = c(0, 0),limits=c(0,1.02))+
  guides(fill = guide_legend(keywidth = 2, keyheight = 2),
         linetype=guide_legend(keywidth = 3, keyheight = 2))
plot

q = 0.05
alphas = c(model.samples[[1]][,1], model.samples[[2]][,1])
betas = c(model.samples[[1]][,2], model.samples[[2]][,2])
BMD = c()

# There's 20 000 samples after combining both chains
for(i in 1:20000) {
  P0 = exp(alphas[i])/(1+exp(alphas[i]))
  BMD = c(BMD, (logit(q*(1-P0)+P0)-alphas[i])/betas[i])
}
hist(BMD)
median(BMD) # posterior median = 17.11037

max(x) # tau can range from 0 to max(x) = 500
model_code <- "model
  {
    for (i in 1:5){
    y[i] ~ dbinom(p[i],n[i])
    logit(p[i]) = alpha + beta*(x[i]-tau)*(x[i]>tau)
    }
    alpha ~ dunif(-100,100)
    beta ~ dunif(-100,100)
    tau ~ dunif(0, 500)
  }"
```

```
cat(model_code, file = "threshold_model.txt")

threshold.fit <- jags.model(file="threshold_model.txt",
                        data=list(n=n,y=y,x=x),
                        n.chains = chains,
                        inits=list(list(alpha=-2,beta=0.1,tau=100),
                                   list(alpha=4,beta=-0.1,tau=200)))
update(threshold.fit, burnin)
model.samples <- coda.samples(threshold.fit, c("alpha", "beta", "tau"), n.iter=iterations)

summary(window(model.samples, start = burnin))
par(mar = c(5, 5, 2, 2))
plot(model.samples, trace=TRUE, density = TRUE)
gelman.diag(model.samples,confidence=0.95)

combined.samples = combine.mcmc(model.samples)
HPDinterval(combined.samples,prob=0.95)

## Observed vs. fitted malformation probability plot
observed=data.frame(dose_obs=c(0,62.5,125,250,500),
                    p_obs=c(67/282,34/225,193/290,250/261,141/141))
dose_fitted=0:500
fitted=data.frame(dose_fitted=dose_fitted,p_fitted=exp(-1.2930+0.0272*((dose_fitted)-61.0357)*
((dose_fitted)>61.0357))/
                    (1+exp(-1.2930+0.0272*((dose_fitted)-61.0357)*((dose_fitted)>61.0357))))
plot = ggplot()+geom_point(data=observed,aes(x=dose_obs,y=p_obs),size=3)+
  geom_line(data=fitted,aes(x=dose_fitted,y=p_fitted),lwd=1,linetype="longdash")+
  theme_bw()+xlab("DYME dose")+ylab("Malformation probability")+
  scale_x_continuous(expand = c(0, 0), limits = c(-5,505)) +
  scale_y_continuous(expand = c(0, 0),limits=c(0,1.02))+
  guides(fill = guide_legend(keywidth = 2, keyheight = 2),
         linetype=guide_legend(keywidth = 3, keyheight = 2))
plot
```