Bayesian inference 2017
Exercise session 3 (20.–23.11.2017)

1. Assume a random variable X that has a continuous distribution. The change of variables formula for random variables says that the density of the transformed random variable $Y = g(X)$ for a transformation[1] $g : A \to B$ is obtained by the formula

$$f_Y(y) = \begin{cases} f_X(h(y))|h'(y)|, & \text{when}\, y \in B \\ 0, & \text{otherwise,} \end{cases}$$

where $h : B \to A$ is an inverse function of transformation g:

$$h(y) := g^{-1}(y).$$

(a) We used a log-normal distribution as a prior for the Poisson likelihood in the lectures. A density function of the log-normal distributed random variable $Y \sim \text{Log-normal}(\mu, \sigma^2)$ is

$$f_Y(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} e^{-\frac{(\log y - \mu)^2}{2\sigma^2}}, \quad \text{when } y > 0.$$

If a random variable X follows a normal distribution $N(\mu, \sigma^2)$, then a its transformation e^X has a log-normal distribution[2] $\text{Log-normal}(\mu, \sigma^2)$.

Use the change of variables formula to derive the density f_Y of the log-normal distribution from the density of the normal distribution:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

(b) Uniform priors are not in general uniform any more for the transformed parameters. Consider for example the uniformly distributed parameter $\theta \sim \text{U}(0, 1)$ which means that its density is

$$f_\Theta(\theta) = \begin{cases} 1 & \text{when } 0 < \theta < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Logit-transfomation maps the interval $(0, 1)$ onto the whole real axis. Compute the distribution $f_\Phi(\phi)$ of the logit-transformation

$$\phi = \text{logit}(\theta) = \log \frac{\theta}{1 - \theta}$$

of the parameter θ using the change of variables formula. Check by simulating that your analytical result is correct: generate (for all the simulations in these exercises you can just use `R` functions, such as `runif` here) a sample $\theta_1, \ldots, \theta_S$ from $U(0, 1)$, and draw a histogram[3] of the logit transform $\text{logit}(\theta)$ of the sample. Then draw a curve of the density function f_Φ on top of this histogram to check that they match each other. You can also observe that the prior is not uniform any more in the transformed space.

[1] A transformation g has also to be a diffeomorphism, which means that it is: 1) bijection from A to B, and 2) both g and g^{-1} must be continuously differentiative. However, these conditions hold in this exercise for $A = \mathbb{R}$, and $B = (0, \infty)$ in (a), and $A = (0, 1)$, $B = \mathbb{R}$ in (b).

[2] And correspondingly, if $Y \sim \text{Log-normal}(\mu, \sigma^2)$, then $\log X \sim N(\mu, \sigma^2)$; hence the name of the distribution.

[3] Use an argument `probability = TRUE` when drawing a histogram so that it plots normalized counts. Also because you have a large sample, it may be useful to set the number of bins of the histogram higher, for example: `breaks=50`, so that the histogram has a sharper resolution.

2. There is a tram stop near your home; you don't know the timetable, but you know that the time interval between the trams is constant. Every day you go to school, you mark down how long you have to wait for the tram. This week you observed following waiting times (as minutes; numbers after the dot are decimals, not seconds):

```
y <- c(1.36, 7.47, 7.31, 7.48, 10.33)
```

You want to find out what is an interval between the trams, and decide to model your waiting times as uniformly distributed (you assume that your arrival times to the stop are quite random) random variables $Y_1, \ldots Y_n \sim \mathrm{U}(0, \theta) \perp\!\!\!\perp \mid \theta$. The parameter θ is the interval between the trams.

A conjugate prior for this sampling distribution is a Pareto distribution[4], so you decide to use a Pareto prior $\theta \sim \mathrm{Pareto}(b, K)$. A density function of the Pareto distribution is

$$p(\theta) = \begin{cases} \frac{Kb^K}{\theta^{K+1}} & \text{if } \theta \geq b \\ 0 & \text{otherwise.} \end{cases}$$

(a) You decide to use prior parameters $b = 1$ and $K = 1$. Generate a sample of 10^4 points from your prior distribution. Draw a histogram of your sample. Why does it look a bit strange? What is the median, and what is the maximum value of your sample? What do you think is meant when the Pareto distribution is said to have ”heavy tails”?

(b) Show that the posterior distribution for this model is Pareto $(c, n + K)$, where $c = \max\{b, y_1, \ldots, y_n\}$.

(c) Compute 80% equal-tailed credible interval for the parameter θ with prior parameters $b = 1$ and $K = 1$, and the observations $\mathbf{y}$. Plot the posterior density, and color the area under the curve on the credible interval (or otherwise mark the credible interval into the plot). Draw a second plot, but this time with a HPD interval. Which one of the credible intervals do you think makes more sense with this posterior and why?

(d) Compute a posterior mean $E[\theta \mid \mathbf{Y} = \mathbf{y}]$. What is a posterior mode $\hat{\theta}_{\mathrm{MAP}} = \underset{\theta}{\operatorname{argmax}}\, p(\theta|\mathbf{y})$ for these observations? Which one do you think makes more sense as a point estimate for the interval between the trains? An expected value of the Pareto-distributed random variable X is

$$EX = \frac{Kb}{K-1}, \quad \text{when } K > 1.$$

(e) Compute the probabilities $P(\theta > 11 \mid \mathbf{Y} = \mathbf{y})$ and $P(11 < \theta < 13 \mid \mathbf{Y} = \mathbf{y})$.

(f) You mark down your waiting times for 3 more weeks. Now the observed waiting times for 4 weeks are:

```
y <- c(1.36, 7.47, 7.31, 7.48, 10.33, 7.68, 0.11, 2.79, 7.99, 6.17, 8.32,
       6.54, 3.39, 11.08, 3.51, 10.05, 3.43, 3.20, 2.24, 2.79, 3.80, 3.63,
       1.91, 0.48, 2.63)
```

Plot the posterior density with an 80% credible interval (choose the one you considered better in the part c) using these new observations. Compute also the posterior mean $E[\theta \mid \mathbf{Y} = \mathbf{y}]$ and the probabilities $P(\theta > 11 \mid \mathbf{Y} = \mathbf{y})$ and $P(11 < \theta < 13 \mid \mathbf{Y} = \mathbf{y})$ from this new posterior. How did the new observations influence your opinion about the probable values of the interval between the trams?

[4] `R` package `VGAM` contains functions `dpareto`, `ppareto`, `qpareto` and `rpareto`, which may be useful in this exercise.

3. Finally we have a real data set! (Exercise 2.21 from Bayesian Data Analysis, 3:rd edition).

A file `pew_research_center_june_elect_wknd_data.dta` contains the results of the poll made before year 2008 Presidential elections of USA, and a file `2008ElectionResult.csv` contains the results of these elections for each state.

Load the data sets and compute the proportions of 'very liberal' (a value `'very liberal'` of the variable `'ideo'`) participants of the poll for each of the states. Draw a scatterplot with the proportion of very liberals of the poll participants on the x-axis, and the proportion of Obama voters of the state on the y-axis. Draw points of the scatterplot as abbreviations of the states; these can be found from `state.abb`-vector.

Draw also a similar scatterplot with the number of participants of the poll in the state on the x-axis, and the proportion of very liberals participants of the state on the y-axis.

Some tips:

- A `dta`-file can be read into `R` with a `read.dta`-function of the package `foreign`.
- States are in alphabetical order both in the poll and the election results. You can use this fact to simplify the merging of the poll and the election results.
- If you remove Alaska (`'AK'`) and Hawaii (`'HI'`) from `state.abb`-vector, also it is in the same order as the results, though it is missing Washington D.C. / District of Columbia, which is not a real state. You can either remove it from the results, or add it to the correct place int the `state.abb`-vector (for example with an abbreviation `'DC'`).
- **Bonus**: You can also color the state abbreviations according to their region; regions are found on the `state.region`-vector.

4. Continued from the previous exercise. This is a preview of the hierarchical models; we will examine them in a more detail later in the course. Basically hierarchical models are just models with a multi-level hierarchy, so that the priors of the parameters may also have priors. However, this time let's use a little shortcut known called "empirical Bayes", where we do not set priors for the prior parameters, but estimate the prior parameters from the data.

Let's model the proportions of 'very liberals' (random variables $Y_1, \dots Y_{49}$) as samples from binomial distributions which have their own parameter θ_j for each state, and these parameters are a sample from the common beta distribution:

$$Y_j | \theta_j \sim \text{Binom}(n_j, \theta_j), \quad \theta_j \sim \text{Beta}(\alpha, \beta)$$

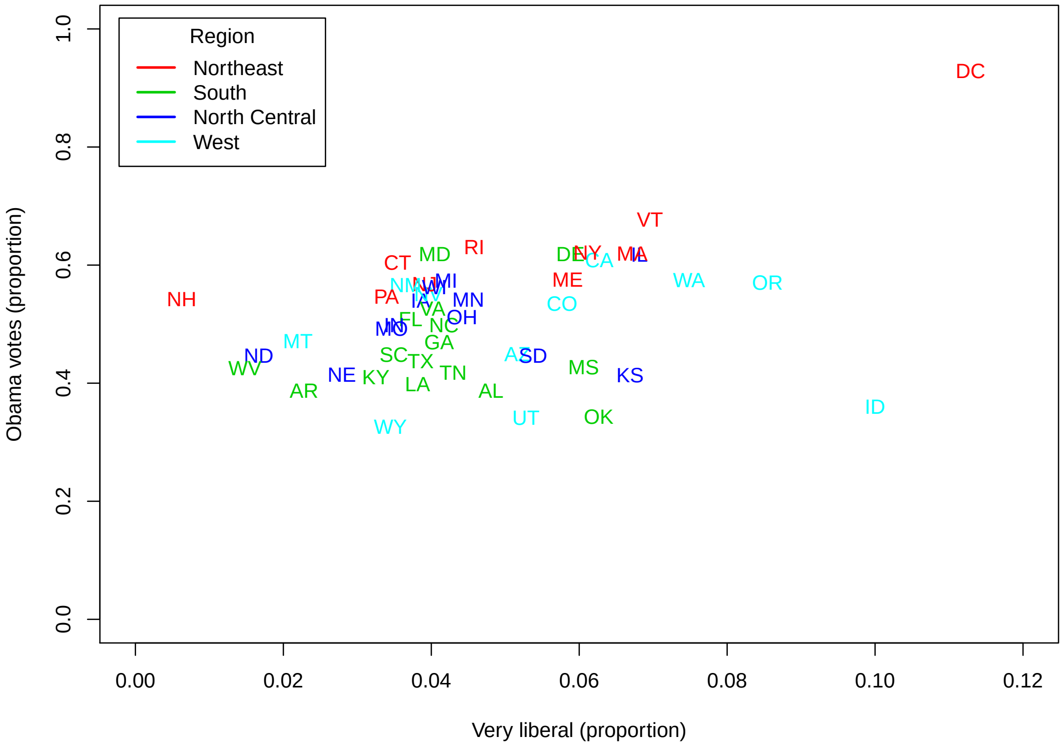
for all $j = 1, \dots, 49$.

(a) Estimate the parameters of the prior beta distribution from the whole data set, that is, from the proportions of very liberal of all the participants. You can use the method of moments, or if you are feeling lazy, just compute the maximum likelihood estimates for the parameters α and β as follows:

```
install.packages('VGAM')
library(VGAM)

# negative log likelihood of data given alpha and beta
ll <- function(alpha, beta) {
  -sum(dbetabinom.ab(y, n, alpha, beta, log = TRUE))
}
```

Figure 1: An example of what pictures could look like: this is the first picture from Exercise 3. The colors and regions corresponding to them are extra and not necessary.



```
mm <- mle(ll, start = list(alpha = 1, beta = 10), method = "L-BFGS-B")
alpha <- coef(mm)[1]
beta <- coef(mm)[2]
```

In this code snippet we assumed that `y` is a vector containing number of very liberal participants for each of the states, and `n` is a vector containing the number of participants for each of the states.

(b) Derive the posterior distribution for the proportions of very liberal θ_j for each of the states, and compute the posterior means for these proportions. Draw a scatterplot with a number of participants of the state on the x-axis, and the posterior mean of proportion of the very liberals of the state on the y-axis.

Compare this figure to the corresponding figure (2nd figure, not the Obama one) from the previous exercise. Both plots have a different quantity on their y-axis: what is the conceptual difference between the quantities plotted on the y-axes? What explains the differences between their values (you can take a look at the Figures 2.8 and 2.9 in page 50 of BDA3 for inspiration)?