Report submissions are accepted in the following formats: One report file in pdf format denoted **name_lab2.pdf**. *Also submit an email with your R-files, with a file named* **name_lab2.R** *that can be used to run your analysis, remember to submit* **all** *of the files you have created, and also that the codes are possible to run. Email the files to* **jonas.wallin@stat.lu.se**.

Discussion between groups is permitted (and encouraged), as long as your answers and code reflects your own work.

Deadline: Wednesday 22/11 at 23.59

Bullets, indicates mandatory exercises, whereas star indicates voluntarily.

Basic probabilist part 2

- In this exercise we are studying a test for Celiac disease. Generate 100'000 individuals, give the person value 1 if it the person has Celiac disease which occurs with probability 0.01. Let each, imaginary, person take a Celiac test which is positive with probability 0.97 if the person has Celiac disease (Sensitivity) and the test is positive with probability $p = 1 - 0.935$ if the person does not has Celiac disease (1-Specificity).

  (probabilities taken from DN article. For definition of Sensitivity and Specificity see wiki.)

  - Use the sample to estimate the probability of having Celiac disease given that the test is positive.
  - Use the sample to estimate the probability of not having Celiac disease given that the test is negative.
  - Is it a good test?

  [2.5p]

*) Let $(X_1, X_2, ..., X_n)$ be at independent random sample from $N(\mu, \sigma)$. Assume σ is known, and put a Normal prior on μ, i.e. $\mu \sim N(\mu_0, \sigma_\mu)$. Recall that the posterior distribution $p(\mu|x_1, ..., x_n) \propto p(x_1, \ldots, x_n|\mu, \sigma) \cdot p(\mu)$. Show that the posterior distribution of μ is a Normal distribution, and derive mean and standard deviation for the distribution. [2.5p]
Hint: It enough to show that the density is proportional to that of a Normal distribution.

Linear normal distribution

- The R-package NHANES (National Health and Nutrition Examination Survey, 1999-2004) contains the body shape and related measurements from the US. Here we will examine linear models for a persons weight.

  - Load the data using:

```
library(NHANES)
data("NHANES")
```

```
NHANES = NHANES[duplicated(NHANES$ID)==F,]
NHANES = NHANES[NHANES$Age >20, ]
NHANES = NHANES[,c("Height","Weight","DirectChol","Tot
NHANES = NHANES[complete.cases(NHANES),]
NHANES = data.frame(NHANES)
```

- Build a linear model for weight using height, DirectChol and TotChol as covariate. Estimate the parameters using map. Choose and motivative your own priors. Build a **map** object so that one can sample from the posterior distribution. Present histograms of the posterior distribution for the parameters.
- Repeat the previous exercise but with log($weight$) instead of $weight$.
- Create a Counterfactual plots for either DirectChol, TotChol. Use either the log($weight$) or the $weight$ model.
- Compare the two models (log vs regular) using WAIC. Don't use WAIC command from rethinking but build your own WAIC code. Note that for the model using log($weight$) one should use **dlnorm**. This since if one uses log($weight$) as observations this can not be compared with $weights$, since they are on different scales. Also give the model weights for the models (see page 199 in the book). Which model fits the data best?

[5p]

- The data set NAEP.txt contains national assessment of educational progress math scores (1992) for students from Nebraska and New jersey. You are to study the educational difference between the states. The state variable
  - Setup a Bayesian model using the category variable state. Is there a 89% significant difference between the states educational level? State: 0- Nebraska, 1- New Jersey.
  - Expand the model using also the second category variable race. Does your model report a 91% significant difference for educational ? Is there a change in the difference? if so motivate why this difference occurred? Gender: 1- white, 2- Black, 3-other.

[5p]

* The R-package HistData contains the dataset GaltonFamilies. The dataset contains the original data that Galton used to analyze the relation between parent and children height. In this exercise you are supposed to build a Bayesian model for the data.

- The data is in inches transform it to cm.
- Implement Bayesian version of Galton's original model, and fit it using map. In Galton models he adjusted the observations by scaling female parameters by 1.08. Don't adjust the data but adjust the parameter (with the fixed

factor $\frac{1}{1.08}$) if the height is from a female. The prior for the parameters is up to you. Galton model is:

$$\begin{aligned} y_i &\sim N(\mu_i, \sigma), \text{if person i is male} \\ \frac{y_i}{1.08} &\sim N(\mu_i, \sigma), \text{if person i is female} \\ \mu_i &= \alpha + \beta midparent_i \end{aligned}$$

- Build and estimate a model which includes a parameters that correct for the gender difference.
- For your model analyze posterior distribution see if the coefficient Galton used for gender is reasonable.
- Compare the two models you built using WAIC.

[5p]

Marginal distribution

* Again we are studying a bag of blue and white marbles. Suppose that in a bag you don't know how many marbles it is. However, you can see that that the number of marbles, N, is somewhere from $[1, 5]$ (given the size of the bag). In the file marble2.txt

  it contains draws from the bag, where one indicates blue marble, and zero white marble. Solve the following problems:

  - Define a prior which is uniform over all possible blue and white marbles. Hint: a convenient way to handle the distribution is as a Matrix.
  - Define one prior such that the marginal distribution of the number of marbles are all equal. For your prior what is the marginal distribution of blue marbles?
  - Compute the posterior distribution of the number of marbles given the observations, Y, for both the prior. Display the posterior marginal distribution both of number of blue marbles and number of marbles in the bag.

[5p]