Assignment 5

Bayesian models & data analysis

1 Estimating the posterior distribution using different computational methods

In a visual word recognition experiment, a participant has to recognize whether a string shown on the screen is a meaningful word (e.g., "book") or a non-word (e.g., "bktr"). The participant is asked to answer "yes" if the shown string is a meaningful word, and "no" if it is a meaningless non-word. Suppose a participant is shown n words and n non-words on the screen one by one and you record the recognition time for each word/non-word.

Say, T_w is the vector of word recognition times, and T_{nw} is the vector of non-word recognition times.

We are interested in testing the following Lexical-access hypothesis.

Lexical-access hypothesis: The mean recognition time for the words is larger than the mean recognition time for the non-words.

We will test whether the data support the above hypothesis or not using a statistical model(s). We can express the lexical-access assumption using the following model.

The lexical-access model

Likelihood:

$$T_{w_i} \sim Normal(\mu, \sigma)$$

$$T_{nw_j} \sim Normal(\mu + \delta, \sigma)$$

Priors:

$$\mu \sim Normal(300, 50)$$

$$\delta \sim Normal(0, 50)$$

$$\sigma = 60$$

If the data are consistent with our hypothesis's predictions, the estimated posterior for the δ parameter should be positive.[1]

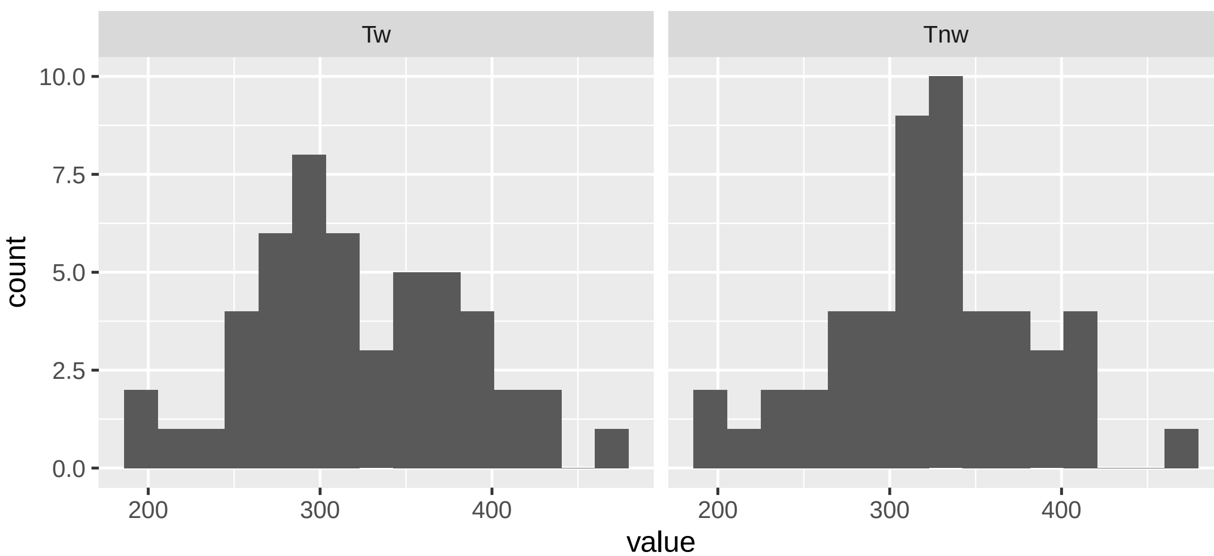
Data

The file *recognition.csv* contains the recognition times data for the words and non-words represented by the columns Tw and Tnw.

```
dat <- read.table("recognition.csv",sep=",",header = T)[,-1]
head(dat)
```

```
##         Tw      Tnw
## 1 285.0780 296.8060
## 2 267.5184 280.1157
## 3 289.9203 310.4417
## 4 399.0674 324.8276
## 5 359.9884 373.8152
## 6 403.3993 269.8220
```

```
## No id variables; using all as measure variables
```



Use the above information to solve the following exercises.

Exercises

1. Use grid approximation to estimate the posterior distribution of δ and graph the estimated posterior distribution.
2. Use monte carlo integration to estimate the marginal likelihood of the model.

   (Hint: Draw a lot of samples from the priors, compute likelihood for each sample, and calculate the average of all the likelihoods.)

[1] However, the parameter estimate does not give any evidence for the hypothesis, it is just tells us about true values of the parameter given the data and assumptions. To evaluate the evidence or to test the hypothesis, we must compare this model against a competitor model. In Bayesian framework, hypothesis testing always requires model comparison.

3. Use importance sampling to draw samples from the posterior distribution of δ.

   (Hint: Draw N samples for each parameter from some reasonable proposal densities, compute the likelihood and prior for each sample, and store the samples and their corresponding weights as {likelihood*prior/proposal density} in a dataframe/vectors. Now, select $N/2$ samples from the vector of N samples based on their weights. These new $N/2$ samples are the samples from the posterior.

   Suppose **proposals** is the vector containing N samples from the proposal density and **weights** is the vector contaning their respective likelihood*prior/proposal_density values. You can select $N/2$ samples using

```r
post_samples <- sample(proposals,size=N/2,prob=weights)
```

4. Use Markov chain Monte Carlo method to estimate the posterior distribution of δ.

   (Hint: You can refer to the Metropolis-Hastings algorithm code given on page 25 in the lecture notes.)

5. Graphically compare the posterior distributions of δ obtained using:

   - Importance sampling
   - Grid approximation
   - Markov chain Monte Carlo

6. Does the data support the lexical-access theory's prediction?

   (Hint: Check whether the estimated values of δ parameter are positive.)