Assignment 4

Bayesian models & data analysis

1 Using Bayes' theorem for statistical inference I

1.1 Assumptions

Suppose you do an experiment and collect data y. A researcher assumes the following:

(a) The data in your experiment is generated by a probability mass function $f(x;\theta)$.

$x \sim f(x;\theta)$

where,

$f(x;\theta) = \frac{10!}{x!(10-x)!}\theta^x(1-\theta)^{10-x}$

(b) The parameter θ of the probability mass function (associated with the generative process) has a real number value somewhere between 0.5 and 1, and each value between 0.5 and 1 is equally likely. This assumption can be expressed by the following probability density function:

$$f(\theta) = \begin{cases} 2 & \text{when } 0.5 \leq \theta \leq 1 \\ 0 & \text{when } \theta < 0.5 \text{ or } \theta > 1 \end{cases}$$

The assumption (a) is called the **likelihood assumption**. When the data has been collected, you can express the probability mass function as a likelihood function. Suppose the collected data is y, you can write the likelihood function:

$$\mathcal{L}(\theta|y) = \frac{10!}{y!(10-y)!}\theta^y(1-\theta)^{10-y} \tag{1}$$

The above likelihood function is a function of θ when y is known, fixed, i.e., you can determine a likelihood for each value of θ.

The assumption (b) is called the **prior assumption** about the parameters. In this case, it is a prior assumption about the parameter θ. This prior assumption about θ has been expressed by a probability density function $f(\theta;a,b)$; you can call this function, the **the prior distribution**. You can write the prior distribution of θ as

$$p(\theta) = \begin{cases} 2 & \text{when } 0.5 \leq \theta \leq 1 \\ 0 & \text{when } \theta < 0.5 \text{ or } \theta > 1 \end{cases} \tag{2}$$

1.2 Statistical inference using Bayes' theorem

Our goal is to infer what value(s) of the parameter θ has generated the data y given the above two assumptions.

We do not only want to know which values of θ could have generated y, we also want to find out which values of θ are more likely and which are less likely to have generated the data y.

More precisely, we want to know what is the probability density of a particular value of θ, say θ_i, that it would generate the data y under the assumptions (a) and (b). This probability density assigned to a particular value of θ given the data y and the likelihood and prior assumptions is called the **posterior density**.

We want to know the approximate function that assigns the posterior densities to each value of θ. This posterior density function is called the **posterior distribution** of θ and is represented by $p(\theta|y)$.

As the likelihood function $\mathcal{L}(\theta|y)$ assigns likelihood to each value of θ given the data y, the posterior distribution $p(\theta|y)$ assigns a probability density to each value of θ given the data y, the likelihood assumption, and the prior assumption $p(\theta)$.

A key step in Bayesian modeling is to somehow estimate the posterior distribution of the parameters of the model. This step is called **parameter estimation**.

Suppose y is the data collected from the experiment,

$\mathcal{L}(\theta|y)$ is the likelihood function assumed by the researcher, and $p(\theta)$ is the prior distribution of the parameter θ assumed by the researcher. The posterior distribution of θ, i.e., $p(\theta|y)$ can be given by the Bayes' rule:

$$p(\theta|y) = \frac{\mathcal{L}(\theta|y)p(\theta)}{\int \mathcal{L}(\theta|y)p(\theta)\, d\theta} \tag{3}$$

You already know the likelihood function $L(\theta|y)$ (see Equation 1) and the prior density function $p(\theta)$ (see Equation 2).

1.3 Data

Suppose you are given that

- The data: $y = 7$
- The marginal likelihood: $\int \mathcal{L}(\theta|y)p(\theta)\, d\theta = \frac{227}{1408}$

You can use the above information and the Bayes' rule (Equation 3) to calculate the posterior density $p(\theta|y)$ for each value of θ.

1.4 Exercises

1. Estimate the posterior density for the following values of θ.
    - $\theta = 0.75$
    - $\theta = 0.25$

- $\theta = 1$

2. Graph the posterior distribution of θ.

   (Hint: Create a vector containing a lot of equidistant values of θ between say 0 and 1; calculate posterior density for each value in the vector; plot a graph with values of θ on the x-axis and associated posterior densities $p(\theta|y)$ on the y-axis.)

3. What value of θ has the maximum posterior density?

4. Compare the graphs of the likelihood function, the prior distribution, and the posterior distribution.

2 Model building in the Bayesian framework

2.1 The research problem

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.

You ask the following question:

Does it take longer to recognize the non-words compared to the words?

Technically,

Is the mean recognition time for the non-words larger than the mean recognition time for the words?

2.2 Hypotheses

Null hypothesis: The mean recognition time for the words is equal to the mean recognition time for the non-words.

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

2.3 Models

The hypotheses are about the underlying cognitive processes that generate the observed recognition times for the words and the non-words. The mean recognition time should be a parameter in the underlying generative process.

A model is a set of statistical assumptions about the underlying generative process. A model should be able to generate data corresponding to a particular value(s) of the parameter(s) and it should be able to generate predictions based on the prior assumptions about the parameters.

We will express our models using our assumptions about the likelihood and the priors.

Null hypothesis model

T_w is the vector of word recognition times; T_{nw} is the vector of non-word recognition times. T_{w_i} is the recognition time for the i^{th} word; T_{nw_j} is the recognition time for the j^{th} non-word. Likelihood:

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

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

Priors:

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

$$\delta = 0$$

$$\sigma = 60$$

Lexical-access model

Likelihood:

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

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

Priors:

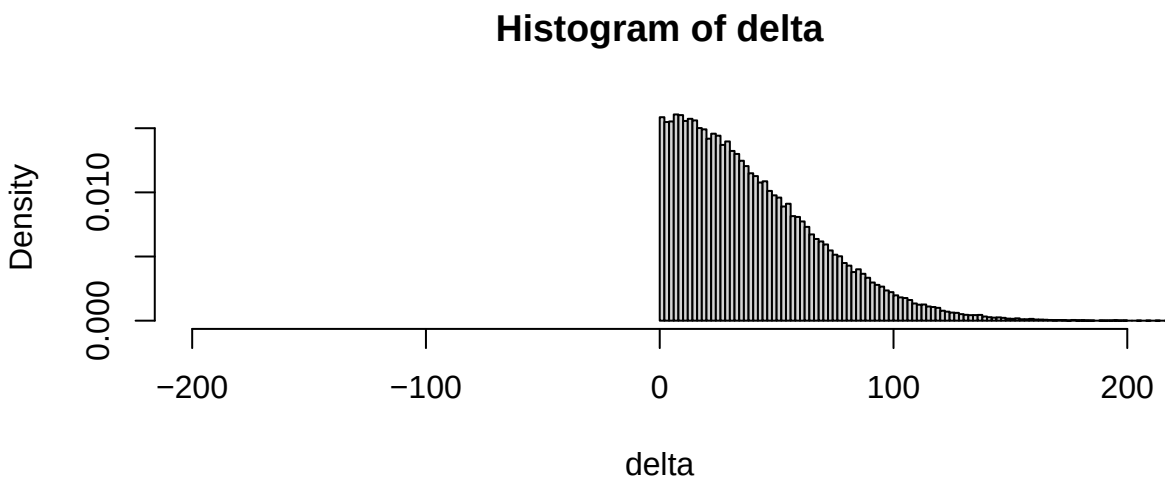
$$\mu \sim Normal(300, 50)$$

$$\delta \sim Normal_+(0, 50)$$

where $Normal_+()$ represent a truncated normal distribution such that δ would always be larger than 0.

$$\sigma = 60$$

```r
library(truncnorm)
# You can generate from a truncated normal distribution using rtruncnorm
delta <- rtruncnorm(100000,a=0,b=Inf,mean=0,sd=50)
hist(delta,xlim = c(-200,200),probability = T,breaks = 100)
```




```r
# You can calculate the probability density of obtaining a value x
# from the truncated normal distribution.
x <- 20
density_x <- dtruncnorm(x,a=0,b=Inf,mean=0,sd=50)
```

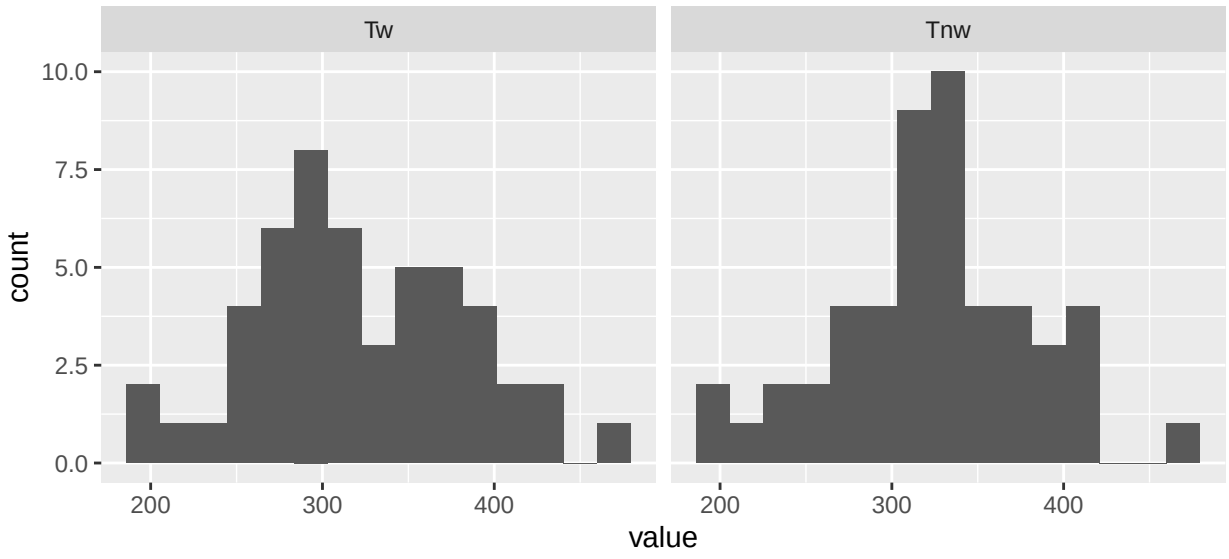
2.4 Data

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

```r
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
```



2.5 Exercises

The unnormalized posterior density of μ for the Null hypothesis model is given by:

$$p'(\mu|T_w, T_{nw}) = \mathcal{L}(\mu, \sigma|T_w)\mathcal{L}(\mu, \sigma|T_{nw})p(\mu)$$

You can calculate the term $\mathcal{L}(\mu, \sigma|T_w)\mathcal{L}(\mu, \sigma|T_{nw}$ using the code

$prod(dnorm(T_w, mean = \mu, sd = \sigma)) * prod(dnorm(T_{nw}, mean = \mu, sd = \sigma))$

and, $p(\mu)$ can be calculated using $dnorm(\mu, 300, 50)$.

1. Graph the unnormalized posterior distribution of μ for the Null hypothesis model.
2. Generate the prior predictions from the lexical-access model.

   (Hint: Draw a vector of values for μ from its prior distribution $\mathcal{N}(300, 50)$; Draw a vector of values for δ from its prior $\mathcal{N}_+(0, 50)$. Plug each set of values of μ and δ in the generative process $\mathcal{N}(\mu + \delta, \sigma = 60)$ to generate non-word recognition times and plug them in $\mathcal{N}(\mu, \sigma = 60)$ to generate the word recognition times. Plot the recognition times as a histogram.)
3. Compare the prior predictions of the null hypothesis model and the lexical access model.
4. Compare the prior predictions of each model against the observed data T_w and T_{nw}. Which model seems more consistent with the data?
5. Graph the unnormalized posterior distribution of δ for the lexical-access model.