Assgn1: Student's goals

Salvador Garcia, s1655274

27 January 2016

1 Description of the problem

229 students (109 boys and 120 girls) aged 7-13 from 9 schools were asked whether popularity or sporting ability was most important to them. The outcome is summarized in the table below. The question is to determine whether there is a difference between the importance of popularity and of sporting ability for girls and boys.

	Sporting ability	Popularity
Boys	59	50
Girls	30	90

2 Likelihood

Introduce the following random variables. Consider the ith boy, $i = 1, 2, ..., n = 109$, and set $X_i = 1$ if popularity is more important to him than sports, and $X_i = 0$ otherwise. Similarly, for the jth girl, $j = 1, 2, ..., m = 120$, set $Y_j = 1$ if popularity is more important to her than sports, and $Y_j = 0$ otherwise. Possible likelihood (independently θ_1 and θ_2 respectively):

$$\begin{aligned} X_i|\theta_1 &\sim Bern(\theta_1), \quad \theta_1 \in (0,1), i = 1, 2, ..., n \\ Y_i|\theta_2 &\sim Bern(\theta_2), \quad \theta_2 \in (0,1), i = 1, 2, ..., m \end{aligned} \tag{1}$$

Also, X_i and Y_j are independent for all i, j. This data, can be summarized as a binomial variable, with two parameters n and p with the next distributions:

$$\begin{aligned} S_X &= \sum_i^n X_i \sim Bin(n, \theta_1) \\ S_Y &= \sum_i^m Y_i \sim Bin(m, \theta_2) \end{aligned} \tag{2}$$

As S_X and S_Y are sufficient statistics for the binomial distribution, then there is no loss of information of θ_1 and θ_2 respectively. Now, θ_1 and θ_2 can be estimated as:

$$\begin{aligned} \widehat{\theta}_1 &= \frac{\sum_i^n X_i}{n} \\ \widehat{\theta}_2 &= \frac{\sum_i^m Y_i}{m} \end{aligned} \tag{3}$$

As can be seen from the above formulas, the parameters θ_1 and θ_2 are the average of the outcomes X_i and Y_i respectively. Then, used in the binomial distribution, this is the probability that boys or girls prefer popularity over sporting ability. In order to discover if there is a difference between the preferences of the preferences of boys and girls, one option is to test the hypothesis $\theta_1 = \theta_2$ or equivalently, $\theta_1 - \theta_2 = 0$.

In this example, the values of the MLE of θ_1 and θ_2 can be calculated with the next expressions: $\sum_i^n X_i = 50$ and $\sum_i^n Y_i = 90$, $m = 120$, $n = 109$.

3 Prior distribution

3.1 Available prior information

The binomial distributions is made with two parameters: n and p. The first is the number of trials and the second the probability of success. The number of trials for θ_1 and θ_2 is determined by n and m and are known. Now, the problem is to estimate p. For this section three prior distributions for θ_1 will be considered. The first two are not informative priors: the first is an uniform distribution over $[0, 1]$, the second the Jeffrey's prior for the binomial likelihood. The last one, is a informative prior that was obtained from a previous study. For the θ_2, only two priors will be considered and are the same non-informative priors used for θ_1.

The informative prior that is used in this example has important implications. It is stated that this distribution is the posterior distribution made from a *boys-only school*. It is not stated when this study was made, so there could be a bias because of the year of study. Generations change between each other in the time. Another bias that can be introduced from this posterior distribution is that the school is a $boys - only$ school, so each boy have only contact with more boys, so the perception of popularity and sport ability could be different. The third factor of bias is that it is not stated if the where was this school or if this was a private or public school. Such differences can potentially bias the study. So make the assumption that is the same population could be dangerous.

3.2 Prior distribution(s)

The following two uninformative prior distributions can be used for the distribution:

1. $p(\theta_i) = 1$, $\theta_i \in [0, 1]$
2. $\theta \sim beta(1/2, 1/2)$

Like $\widehat{\theta_i}$ is a probability, its range is between 0 and 1. Then, a uniform distribution can be used for this example. The second one is the Jeffrey prior for θ_i.

This Jeffrey prior can be derived as follows:

$$\begin{aligned} p(x|\theta_i) =& \binom{n}{x} \theta_i^x (1-\theta_i)^{n-x} \\ \log p(x|\theta_i) =& \log \binom{n}{x} + x \log \theta_i + (n-x) \log(1-\theta_i) \end{aligned} \tag{4}$$

Then deriving with respect θ_i:

$$\begin{aligned} \frac{d \log p(x|\theta_i)}{d\theta_i} =& + \frac{x}{\theta_i} - \frac{(n-x)}{(1-\theta_i)} \\ \frac{d^2 \log p(x|\theta_i)}{d\theta_i^2} =& - \frac{x}{\theta_i^2} - \frac{(n-x)}{(1-\theta_i)^2} \end{aligned} \tag{5}$$

Finally, $E_{\theta_i}(x) = n\theta_i$ so:

$$\begin{aligned} -E_\theta \left(\frac{d^2 \log p(x|\theta_i)}{d\theta_i^2} \right) =& \frac{n\theta_i}{\theta_i^2} + \frac{(n - n\theta_i)}{(1-\theta_i)^2} \\ =& \frac{n}{\theta_i} + \frac{(n)}{(1-\theta_i)} \\ =& \frac{n}{\theta_i(1-\theta_i)} \end{aligned} \tag{6}$$

Taking square root, the last expression is proportional to $\theta_i^{-\frac{1}{2}}(1-\theta_i)^{-\frac{1}{2}}$ i.e. equivalent to the kernel of a beta distribution with parameters $(\frac{1}{2}, \frac{1}{2})$.

The third prior to be used will be an informative one, this is the posterior distribution obtained from a previous analysis. The school was a boys-only school and this distribution is a *Beta(21,10)*.

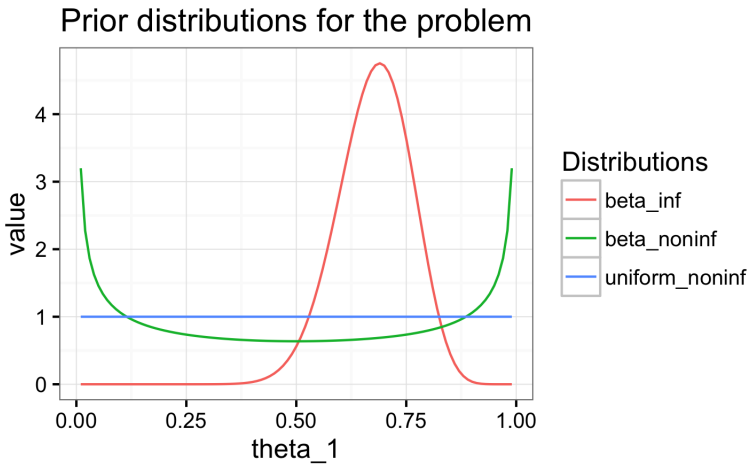

Figure 1: Priors distributions used for the study

4 Posterior inference

The prior distribution will be represented as $p(\theta_i)$ with $i = 1, 2$. Then, the likelihood is a binomial distribution that can be written as (with $\bar{X}$ the sample):

$$p(\bar{X}|\theta_i) \propto \theta_i^{\sum x_i}(1-\theta_i)^{n-\sum x_i} \tag{7}$$

Then, for the first prior (uniform), the posterior takes this form:

$$\begin{aligned} p(\theta_i|\bar{X}) \propto & \theta_i^{\sum_i X_i}(1-\theta_i)^{n-\sum_i x_i} \\ \propto & \theta_i^{50}(1-\theta_i)^{59} \end{aligned} \tag{8}$$

That is the kernel of a $Beta(51, 60)$. For the second prior distribution (Jeffrey's prior), the posterior takes the form:

$$\begin{aligned} p(\theta_i|\bar{X}) \propto & \theta_i^{\sum_i X_i - \frac{1}{2}}(1-\theta_i)^{n-\sum x_i - \frac{1}{2}} \\ \propto & \theta_i^{49.5}(1-\theta_i)^{58.8} \end{aligned} \tag{9}$$

That is the kernel of a $Beta(50.5, 59.5)$. For the last prior distribution (Informative), the posterior distribution takes the form:

$$\begin{aligned} p(\theta_i|\bar{X}) \propto & \theta_i^{\sum_i X_i + 21}(1-\theta_i)^{n-\sum x_i + 10} \\ \propto & \theta_i^{71}(1-\theta_i)^{69} \end{aligned} \tag{10}$$

That is the kernel of a $Beta(72, 70)$.

For the Y_i the computation is similar, so just the results are stated. For the uniform prior, the posterior takes the form of $Beta(91, 31)$, for the Jeffrey's distribution is a $Beta(90.5, 30.5)$.

4.1 Posterior distribution

Then, these three priors were used to find the posterior given the data of the study. The posterior distribution are plotted below:

4.2 Posterior summaries and plots

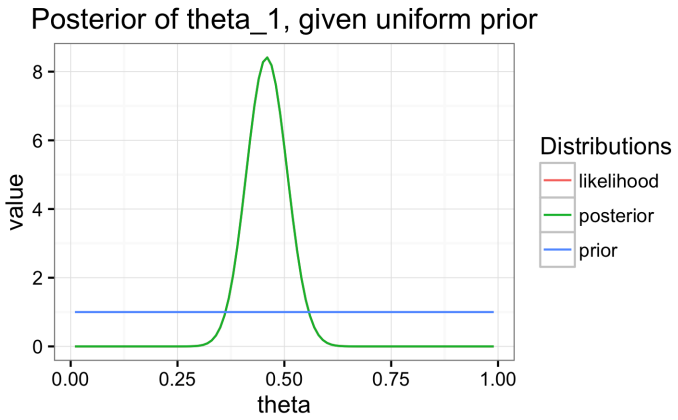

Figure 2: Likelihood, prior and posterior distribution of θ_1 given an uniform prior.

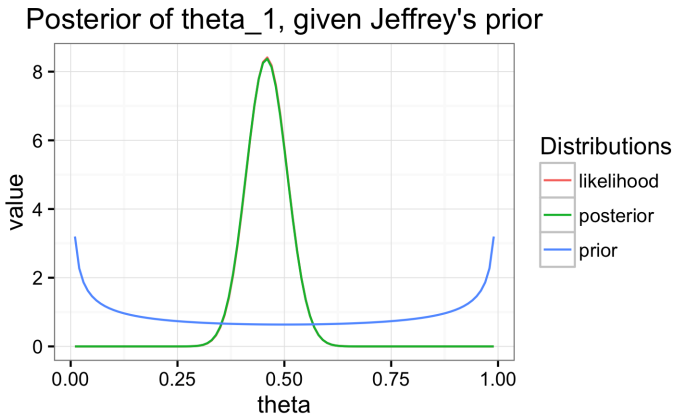

Figure 3: Likelihood, prior and posterior distribution of θ_1 given a Jeffrey's prior.

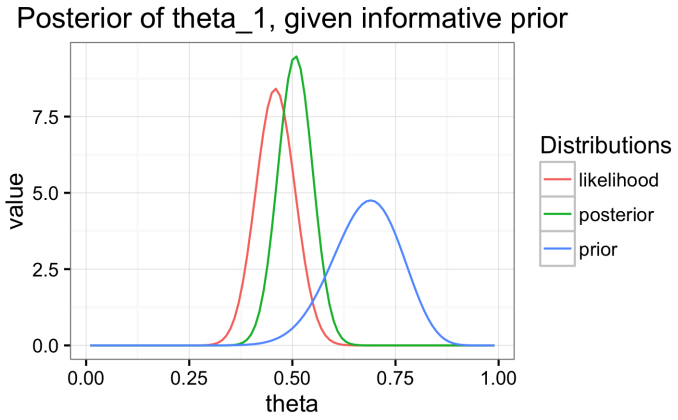

Figure 4: Likelihood, prior and posterior distribution of θ_1 given an informative prior.

With these graphs is easy to see that the first two are very similar. This is because both are created with a non-informative prior. The third one seems that the prior distribution and the likelihood are different (but is closer to the likelihood), so it is possible that the prior is not very correct. In this text, the Jeffrey's prior is the used for the analysis. Below are the equivalent graphs for θ_2 (just 2 non-informative priors were taken).

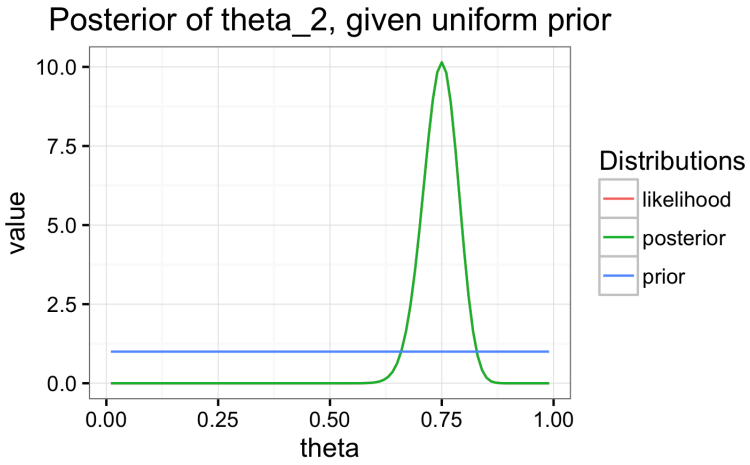

Figure 5: Likelihood, prior and posterior distribution of θ_2

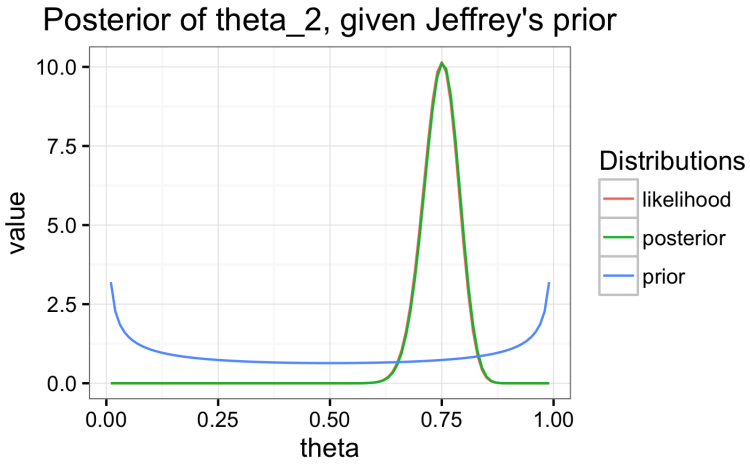

Figure 6: Likelihood, prior and posterior distribution of θ_2

In the example of θ_2, two non-informative priors were taken, and as can be in the plots above, the distributions are very similar.

4.3 Posterior predictive summaries and plots

Now, a comparison of the posterior summaries was made:

Table 1: Summary of distributions

Distribution	Mean	Std	Median	CI_low	CI_high
posterior of θ_1	0.4590909	4.081275e-05	0.4588421	0.3984535	0.5200510
posterior of θ_2	0.7479339	2.941329e-05	0.7493035	0.6966089	0.7974833

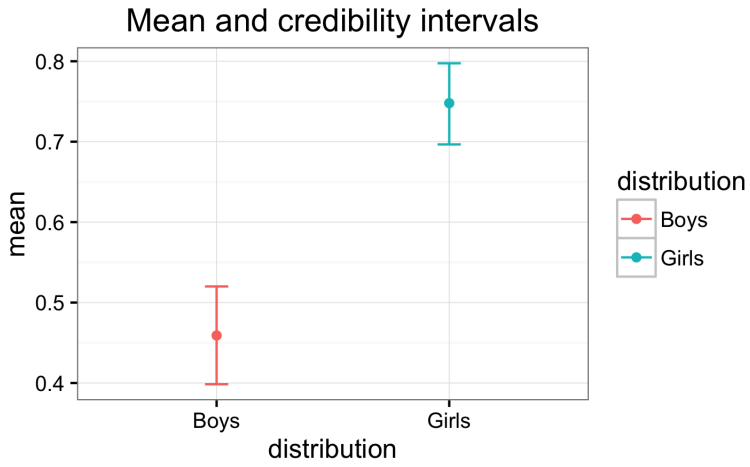

Figure 7: Credibility intervals of 90%

As can be seen in the two credibility intervals, the difference is big, so it can be concluded that the girls prefer popularity more than boys. Also, the standard deviation of θ_2 is less than the boys. One part of this lower standard deviation could be explained because of the factor of the root square of n that is in its formula.

4.4 Decisions

Now, with this graphic is easy to see how different are both posterior distributions:

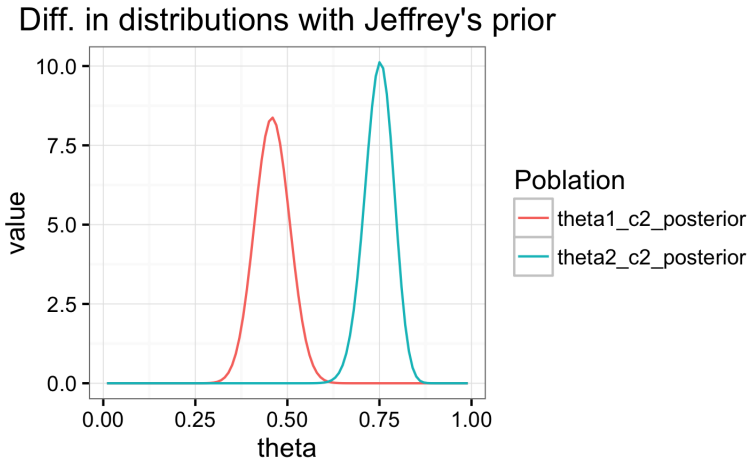

Figure 8: Comparison of posterior distribution of θ_1 and θ_2

The assumption that the data is binomial is questionable. Is hard to measure the preferences in a scale of 0-1, because maybe all the people that select 1 is different and some of them, could prefer much more to be popular than to have sporting ability. For example, a person that is almost indifferent for both decisions should be considered less strong than one that deeply prefers to be popular. For example if the study is made with a 1-5 preference scale, then this differences can be taken into account. This will change completely the distribution of the data. (and maybe this big difference between θ_1 and θ_2 would be smaller!).

The independence of each X_i and Y_i can also be questioned. For example, if inside the study there were people that belongs to small groups of friends, then is possible that they think very similar. But this assumption is hard to test if just a $0-1$ variable is given and also without more information. The third thing that could be not a good idea is the use of another prior distribution (at least as the one given). The reason have been stated above.

5 Conclusions

The difference between θ_1 and θ_2 is very big. For example, the 90% interval of each distribution doesn't overlap. So it is very probable that the $|\theta_1 - \theta_2|$ also doesn't overlap. Given these distributions (and given that both are β distributions, is easy to make an hypothesis test (It is not done because is too evident), but can be done. So, with this data, it is easy to conclude that girls prefer more popularity than boys.

A R graphs code

```r
library(tidyverse)

# 0) functions for plot ---------------------------------
posterior_plot <- function(prior, likelihood, posterior, title, filename){
  post_df <- data.frame(theta = n,
                        likelihood = likelihood,
                        prior = prior,
                        posterior = posterior)
  plot <- post_df %>%
      gather(Distributions, value, likelihood:posterior) %>%
      ggplot(aes(x = theta, y = value, group = Distributions, color = Distributions))+
      geom_line() +
      labs(title = title) +
      theme_bw()
  ggsave(filename = filename,
         plot = plot, width = 5, height = 3)
}
# 1) priors comparison ----------------------------------
dir.create("~/Repositories/UoE_2_BDA/11_Assignment1/")
n = seq(.01, .99, .01)
beta_noninf = dbeta(n, 1/2, 1/2)
uniform_noninf = dunif(n, 0, 1)
beta_inf = dbeta(n, 21, 10)
priors <- data.frame(theta_1 = n,
                     beta_noninf,
                     uniform_noninf,
                     beta_inf)

(gg_priors <- priors %>%
  gather(Distributions, value, beta_noninf:beta_inf) %>%
  ggplot(aes(x = theta_1, y = value, group = Distributions, color = Distributions)) +
  geom_line() +
  labs(title = "Prior distributions for the problem") +
  theme_bw())

ggsave(filename = "~/Repositories/UoE_2_BDA/11_Assignment1/priors_dist.png",
       plot = gg_priors, width = 5, height = 3)

# 2) Posterior for parameter \theta_1 -------------------
theta1_likelihood = dbeta(n, 51, 60)

# uniform prior
theta1_c1_posterior = dbeta(n, 51, 60)

posterior_plot(prior = uniform_noninf,
               likelihood = theta1_likelihood,
               posterior = theta1_c1_posterior,
               title = "Posterior of theta_1, given uniform prior",
               filename = "~/Repositories/UoE_2_BDA/11_Assignment1/theta1_c1_posterior.png")
```

```
# Jeffrey's prior
theta1_c2_posterior = dbeta(n, 50.5, 59.5)

posterior_plot(prior = beta_noninf,
               likelihood = theta1_likelihood,
               posterior = theta1_c2_posterior,
               title = "Posterior of theta_1, given Jeffrey's prior",
               filename = "~/Repositories/UoE_2_BDA/11_Assignment1/theta1_c2_posterior.png")

# Informative prior
theta1_c3_posterior = dbeta(n, 72, 70)

posterior_plot(prior = beta_inf,
               likelihood = theta1_likelihood,
               posterior = theta1_c3_posterior,
               title = "Posterior of theta_1, given informative prior",
               filename = "~/Repositories/UoE_2_BDA/11_Assignment1/theta1_c3_posterior.png")
# 3) Posterior for parameter \theta_2 ------------------
theta2_likelihood = dbeta(n, 91, 31)

# uniform prior
theta2_c1_posterior = dbeta(n, 91, 31)

posterior_plot(prior = uniform_noninf,
               likelihood = theta2_likelihood,
               posterior = theta2_c1_posterior,
               title = "Posterior of theta_2, given uniform prior",
               filename = "~/Repositories/UoE_2_BDA/11_Assignment1/theta2_c1_posterior.png")

# Jeffrey's prior
theta2_c2_posterior = dbeta(n, 90.5, 30.5)

posterior_plot(prior = beta_noninf,
               likelihood = theta2_likelihood,
               posterior = theta2_c2_posterior,
               title = "Posterior of theta_2, given Jeffrey's prior",
               filename = "~/Repositories/UoE_2_BDA/11_Assignment1/theta2_c2_posterior.png")

# 4) summaries ------------------------------------------
# Selected priors: Jeffrey priors

theta1_c2_posterior = dbeta(n, 50.5, 59.5)
theta1_c2_mean = 50.5/(50.5+59.5)
theta1_c2_std = sqrt(50.5*59.5)/(((50.5+59.5)^2)*(50.5+59.5+1))
theta1_c2_median = qbeta(.5, 50.5, 59.5)
theta1_c2_CI = qbeta(c(.1, .9), 50.5, 59.5)

theta2_c2_posterior = dbeta(n, 90.5, 30.5)
theta2_c2_mean = 90.5/(90.5+30.5)
theta2_c2_std = sqrt(90.5*30.5)/(((90.5+30.5)^2)*(90.5+30.5+1))
theta2_c2_median = qbeta(.5, 90.5, 30.5)
theta2_c2_CI = qbeta(c(.1, .9), 90.5, 30.5)
```

```
CI_gg <- data.frame(distribution = c("Boys", "Girls"),
            mean = c(theta1_c2_mean, theta2_c2_mean),
            std = c(theta1_c2_std, theta2_c2_std),
            median = c(theta1_c2_median, theta2_c2_median),
            CI_low = c(theta1_c2_CI[1], theta2_c2_CI[1]),
            CI_high = c(theta1_c2_CI[2], theta2_c2_CI[2])) %>%
  ggplot(aes(x = distribution)) +
  geom_point(aes(y = mean, color = distribution)) +
  geom_errorbar(aes(ymin = CI_low, ymax = CI_high,
                    width = .1, color = distribution)) +
  labs(title = "Mean and credibility intervals") +
  theme_bw()

ggsave(filename = "~/Repositories/UoE_2_BDA/11_Assignment1/CI_intervals.png",
       plot = CI_gg, width = 5, height = 3)

# 5) Posterior distribution comparisons ----------------
diff_gg <- data.frame(theta = n,
            theta1_c2_posterior,
            theta2_c2_posterior) %>%
  gather(Poblation, value, theta1_c2_posterior:theta2_c2_posterior) %>%
  ggplot(aes(x = theta, y = value, group = Poblation, color = Poblation)) +
  geom_line() +
  labs(title = "Diff. in distributions with Jeffrey's prior") +
  theme_bw()
ggsave(filename = "~/Repositories/UoE_2_BDA/11_Assignment1/diff.png",
       plot = diff_gg, width = 5, height = 3)
```