Methods of Applied Stat Lab 3, Rudraharsh Tewary

28/01/24

Branches on git

Branches on git are useful when you have more than one person working on the same file, or when you are experimenting with different code etc that may not work. So far we've just been pushing to the 'main' branch, but you can also create other branches within your repo, do some work, save and push, and then if you're happy, merge that work back into the 'main' branch. The idea is that the 'main' branch is always kept clean and working, while other branches can be tested and deleted.

Before merging work into the main branch, it's good practice to do a 'pull request' – this flags that you want to make changes, and alerts someone to review your code to make sure it's all okay.

For this week, I would like you to save this .qmd file to your class repo, then create a new branch to make your edits to the file. Then, once you are happy with this week's lab submission, on GitHub, create a 'pull request' and assign me to be the reviewer.

Question 1

Consider the happiness example from the lecture, with 118 out of 129 women indicating they are happy. We are interested in estimating θ, which is the (true) proportion of women who are happy. Calculate the MLE estimate $\hat{\theta}$ and 95% confidence interval.

Answer 1

Through the given data, and knowing that the MLE of the binomial distribution is given as:

$$\hat{\theta} = \frac{x}{n}$$

We get the estimates and confidence intervals as

```
x <- 118
n <- 129
mle <- x/n
z <- 1.96
se <- sqrt((mle*(1-mle))/n)
lower_bound <- mle - z*se
upper_bound <- mle + z*se
confidence_interval <- c(lower_bound,upper_bound)
mle
```

```
[1] 0.9147287
```

```
confidence_interval
```

```
[1] 0.8665329 0.9629244
```

Question 2

Assume a Beta(1,1) prior on θ. Calculate the posterior mean for $\hat{\theta}$ and 95% credible interval.

Answer 2

We can do the above task using the following code block:

```
alpha_posterior <- 1 + x
beta_posterior <- 1 + n-x
posterior_mean <- alpha_posterior/(alpha_posterior+beta_posterior)
credibility_interval <- qbeta(c(0.025,0.975),alpha_posterior,beta_posterior)
posterior_mean
```

```
[1] 0.9083969
```

```
credibility_interval
```

```
[1] 0.8536434 0.9513891
```

Question 3

Now assume a Beta(10,10) prior on θ. What is the interpretation of this prior? Are we assuming we know more, less or the same amount of information as the prior used in Question 2?

Answer 3

When we use a Beta(10,10) Prior on θ means we estimate the amount of women who are aged 65+ and are happy is around 50% or $\theta = 0.5$. We are assuming we know more information than the prior we assumed in question 2. If we use a Beta(10,10) Prior we get the following results:

```
alpha_posterior_new <- 10 + x
beta_posterior_new <- 10 + n-x
posterior_mean_new <- alpha_posterior_new/(alpha_posterior_new + beta_posterior_new)
credibility_interval_new <- qbeta(c(0.025,0.975),alpha_posterior_new,beta_posterior_new)
posterior_mean_new
```

```
[1] 0.8590604
```

```
credibility_interval_new
```

```
[1] 0.7990363 0.9099708
```

Question 4

Create a graph in ggplot which illustrates

- The likelihood (easiest option is probably to use `geom_histogram` to plot the histogram of appropriate random variables)
- The priors and posteriors in question 2 and 3 (use `stat_function` to plot these distributions)

Comment on what you observe.

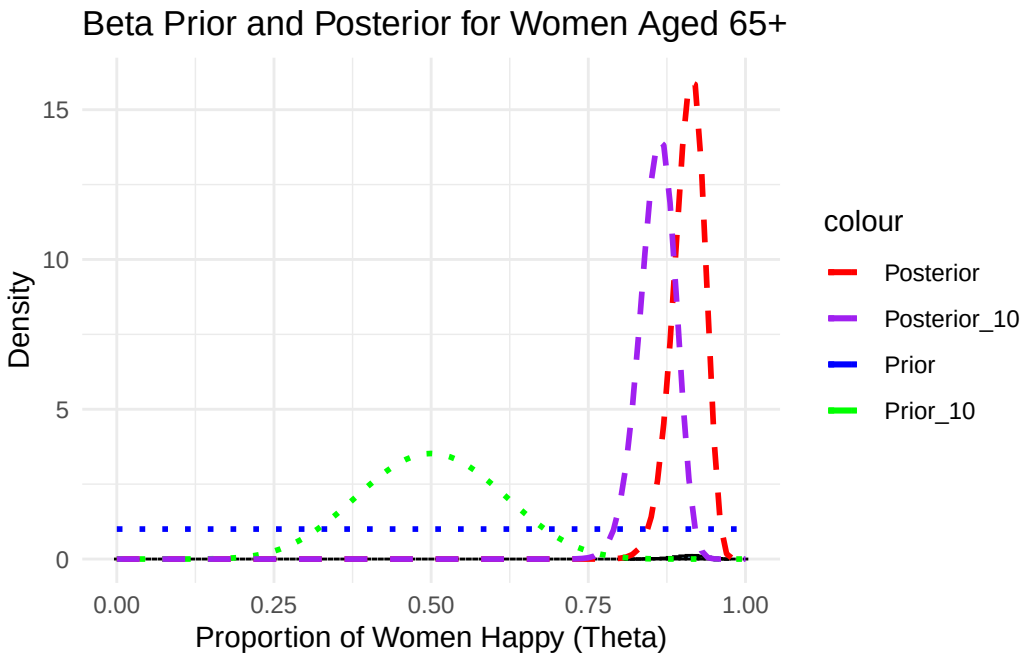
Answer 4

We can get the graph through the following code block :

```r
library(ggplot2)

theta_values <- seq(0, 1, by = 0.01)
likelihood <- dbinom(x, size = n, prob = theta_values)
data <- data.frame(theta = theta_values, likelihood = likelihood)

beta_pdf <- function(theta, alpha, beta) {
  dbeta(theta, shape1 = alpha, shape2 = beta)
}

ggplot(data, aes(x = theta, y = likelihood)) +
  geom_histogram(stat = "identity", fill = "skyblue", color = "black", bins = 30) +
  stat_function(fun = beta_pdf, args = list(alpha = 1, beta = 1),
                geom = "line", aes(color = "Prior"), size = 1, linetype = "dotted") +
  stat_function(fun = beta_pdf, args = list(alpha = alpha_posterior, beta = beta_posterior
                geom = "line", aes(color = "Posterior"), size = 1, linetype = "dashed") +

  stat_function(fun = beta_pdf, args = list(alpha = 10, beta = 10),
                geom = "line", aes(color = "Prior_10"), size = 1, linetype = "dotted") +
  stat_function(fun = beta_pdf, args = list(alpha = alpha_posterior_new, beta = beta_poste
                geom = "line", aes(color = "Posterior_10"), size = 1, linetype = "dashed")
  labs(title = "Beta Prior and Posterior for Women Aged 65+",
       x = "Proportion of Women Happy (Theta)",
       y = "Density") +
  scale_color_manual(values = c("Prior" = "blue", "Posterior" = "red","Prior_10"="green","
  theme_minimal()
```

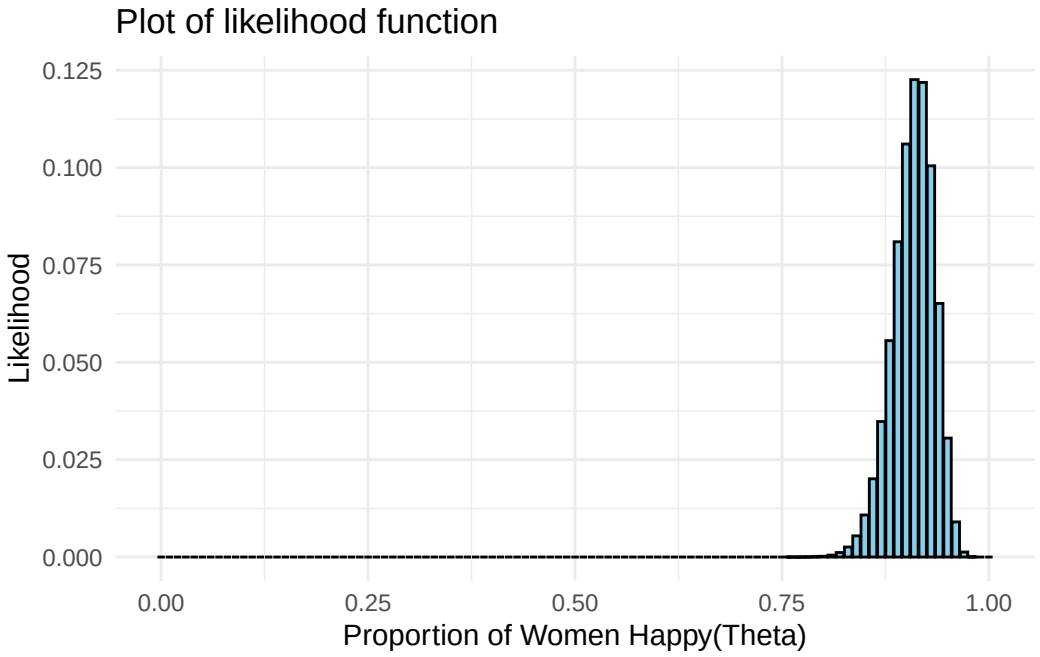
```
Warning in geom_histogram(stat = "identity", fill = "skyblue", color = "black",
: Ignoring unknown parameters: `binwidth`, `bins`, and `pad`
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```




```
# We can also get a seperate plot of likelihood as follows:
ggplot(data, aes(x = theta, y = likelihood)) +
  geom_histogram(stat = "identity", fill = "skyblue", color = "black", bins = 30) +
  labs(title = "Plot of likelihood function",
           x = "Proportion of Women Happy(Theta)",
           y = "Likelihood")+
  theme_minimal()
```

```
Warning in geom_histogram(stat = "identity", fill = "skyblue", color = "black",
: Ignoring unknown parameters: `binwidth`, `bins`, and `pad`
```



We get the following interpretations on both plots :- 1) Likelihood plot :- Upon a cursory examination of the binomial likelihood plot, we can infer that the likelihood suggests that based on the data, 85-90% of women aged 65+ would be happy, with the range of 85-90% being the most likely proportion.

2) Priors and Posteriors plot :- Upon looking at the prior and posterior plot for the Beta(1,1) and Beta(5,5) priors. We see that the Beta(1,1) prior agrees to the likelihood plot that we obtained above. While, a Beta(10,10) prior introduces the notion that we have prior information suggesting that the real proportion of 65+ women who should be happy is around 50%, causing the posterior distribution of θ based on the data to shift left, causing the new suggested proportion to be around 80%.

Question 5

Laplace was interested in calculating the probability that observing a male birth was less than 0.5, given data he observed in Paris. Calculate this probability, assuming a uniform prior on observing a male birth and using data given in the slides.

Answer 5

We use a uniform prior (Which is a beta(1,1)) on the probability of a male birth and get the following result from our code block:

```r
b = 251527
g = 241945

probability <- pbeta(0.5,b+1,g+1)
probability
```

```
[1] 1.146058e-42
```

Question 6

(No R code required) A study is performed to estimate the effect of a simple training program on basketball free-throw shooting. A random sample of 100 college students is recruited into the study. Each student first shoots 100 free-throws to establish a baseline success probability. Each student then takes 50 practice shots each day for a month. At the end of that time, each student takes 100 shots for a final measurement. Let θ be the average improvement in success probability. θ is measured as the final proportion of shots made minus the initial proportion of shots made.

Given two prior distributions for θ (explaining each in a sentence):

- A noninformative prior, and
- A subjective/informative prior based on your best knowledge

Answer 6

I would give the following distributions as priors in both scenarios:

1) Non-informative prior :- In the case of a non-informative prior, I would select a uniform (0,1), that is, a Beta(1,1) distribution as my prior. Because I dont know anything about the effect of practice on success probabilites and consider all values of θ likely.
2) Informative prior :- In the case of a informative prior, one possible distribution I would suggest is a Beta(5,5), centered around 0.5, assuming that prior research and expert opinion tells us that there is atleast a moderate chance for improvement in free-throw success.