Coding Homework 3

Jacqueline Heitmann
25562334

```r
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0
──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.1     ✔ tibble    3.2.1
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.0.2
── Conflicts ──────────────────────────────────────────── tidyverse_conflicts()
──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

For reference: θ | p(θ) 0.0 | 0.01 0.2 | 0.40 0.4 | 0.34 0.6 | 0.17 0.8 | 0.03 1.0 | 0.01

1a)

```r
# make theta, or the possible values that can be taken on by theta which happen
to be proportion of zener cards guessed correctly
theta <- c(0, 0.2, 0.4, 0.6, 0.8, 1)

# prior (ie ptheta) which are the presumed probabilities of getting a specific
theta value
ptheta <- c(0.01, 0.4, 0.34, 0.17, 0.03, 0.01)
```

1b-d) Use theta and your Zener card data ({yi}) to make the following vector: lik = p({yi}|θ) = the likelihood of the data for each value of θ.

Hint: Count the number of correct guesses and use Equation 5.11 from the textbook.

5.11 Equation for reference: $\theta^{#heads} (1-\theta)^{#tails}$ (idk why the latex code isn't working)

```r
# load in data
zener <- read.csv("../data/Zener_data.csv")
```

```r
k <- sum(zener$correct)
n <- nrow(zener)
wrong <- n - k

# k = number of correct zener cards; size = total number of trials; prob = theta
# lik is the equivalent of p(y|theta)
lik <- (theta^k) * ((1 - theta)^wrong)

# compute marginal likelihood - p(y) - which is the likelihood * the probability
of theta
marg_lik <- sum(ptheta * lik)
marg_lik
```

```
[1] 2.04374e-81
```

```r
# trying to find the probability of theta given y p(y|theta)
# p(theta|y) = ( p(y|theta) * p(theta) / p(y) ) where the denominator is the
marginal likelihood
# and
post <- (lik * ptheta) / marg_lik
```

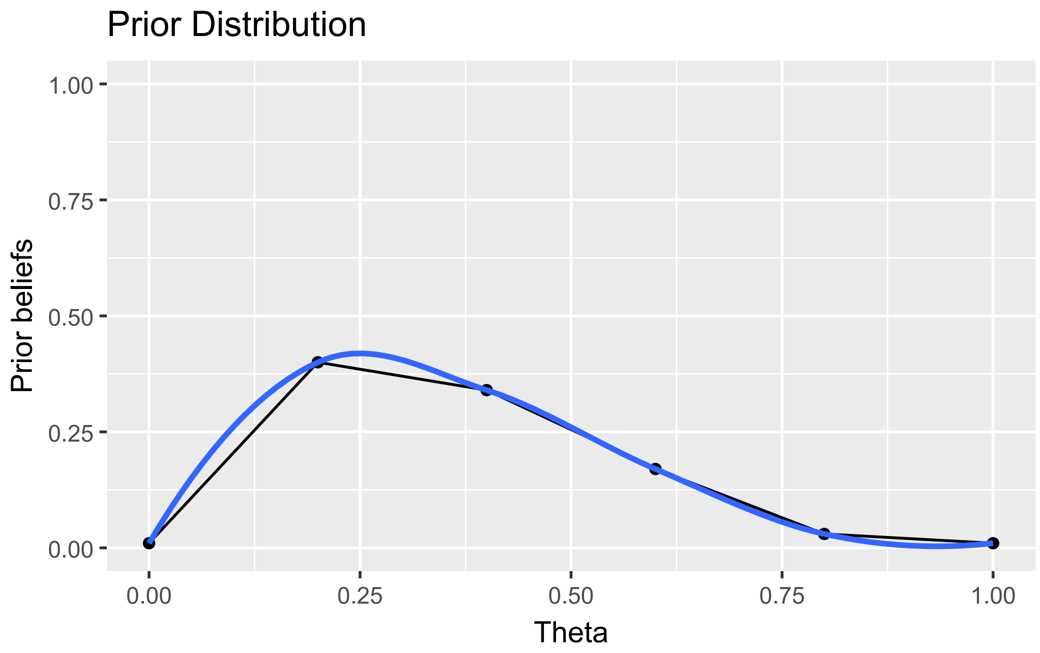
2a) (a) (10 points) Use ggplot2 with geom line and geom point to plot the prior distribution. Label the plot "Prior distribution" (use ggtitle) and make sure the y-axis is scaled from 0 to 1 (use ylim) for the sake of consistency with the second plot. Hint: The code will be similar to what you used in the first homework, except that you are plotting post instead of the binomial distribution obtained using dbinom.

```r
# prior dist df
prior_df <- data.frame(theta, ptheta)

ggplot(prior_df, aes(x = theta, y = ptheta)) +
  geom_line() +
  geom_point() +
  geom_smooth() +
  ylim(0, 1) +
  labs(title = "Prior Distribution", x = "Theta", y = "Prior beliefs")
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

```
Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
-Inf
```



2b) Plot the posterior distribution in the same way, following all the guidelines for the prior distribution plot (obviously give it the right title).

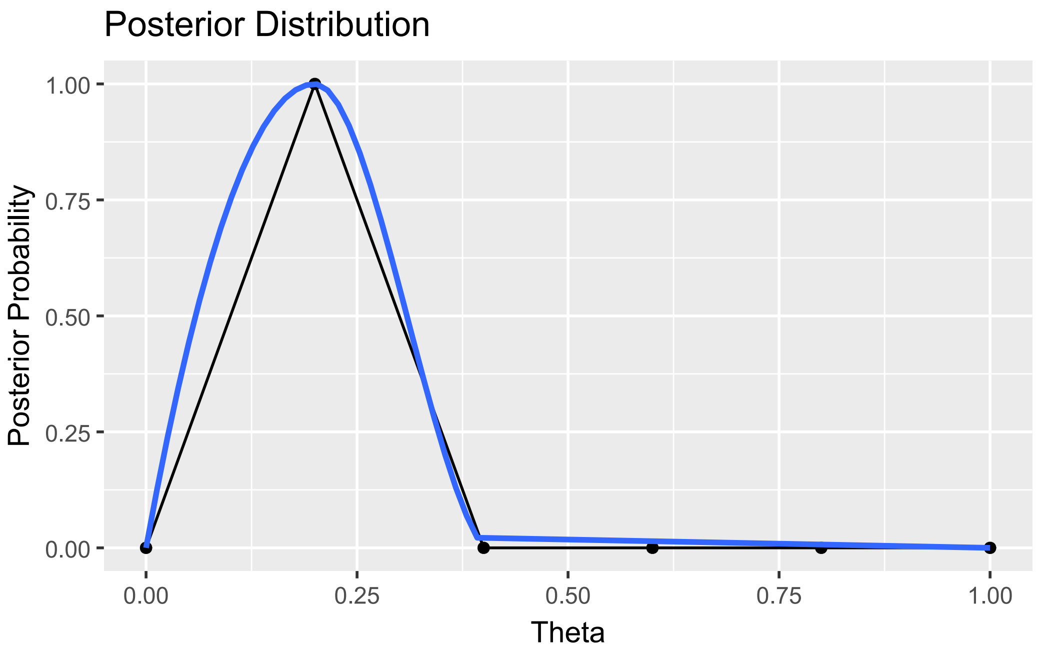
```
# posterior dist df
post_df <- data.frame(theta, post)

ggplot(prior_df, aes(x = theta, y = post)) +
  geom_line() +
  geom_point() +
  geom_smooth() +
  ylim(0, 1) +
 labs(title = "Posterior Distribution", x = "Theta", y = "Posterior Probability")
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

```
Warning: Removed 47 rows containing missing values or values outside the scale
range
(`geom_smooth()`).
```

```
Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
-Inf
```



2c) In the prior distribution, there is greater variability across values of theta, reflecting a broader range of possible beliefs about theta. The highest probability density is around theta = 0.20, which corresponds to a prior belief of 40% for that value.

In contrast, the posterior distribution is more defined, with the probability mass concentrated between theta = 0 and theta = 0.35, peaking around theta = 0.15. This shift demonstrates how the data has updated our beliefs, making the posterior distribution more defined and specific compared to the prior. There is also much less variation beyond values of 0.35 with little cases reaching greater values of theta.

3. Finally, we will practice summarizing the prior and prior distributions.

(a) (10 points) Compute the prior mean/expected value of θ (notation: "E[θ]") and posterior mean/expected value of θ (notation: "E[θ|y]"). Compare the two: did the mean of θ increase or decrease? Hint: This computation will be similar to what you did in homework2.

```
eprior_theta <- sum(theta * ptheta)
eprior_theta
```

```
[1] 0.352
```

```
epos_theta <- sum(theta*post)
epos_theta
```

```
[1] 0.2
```

The mean decreased.

(b) (10 points) Compute the prior variance of θ (notation: “$V[\theta]$”) and posterior variance of θ (notation: “$V[\theta|y]$”). Compare the two: did the variance of θ increase or decrease? Hint: This computation will be similar to what you did in homework 2.

```r
prior_var <- sum(ptheta * (theta - ptheta)^2)
prior_var
```

```
[1] 0.076246
```

```r
post_var <- sum(post * (theta - post)^2)
post_var
```

```
[1] 0.64
```

The variance increased!

(c) (10 points) The posterior mean of θ (i.e. $E[\theta|y]$) is one way of boiling down the posterior distribution into a single “best guess” about θ, i.e. a point estimate. Another point estimate is the posterior mode, i.e. the value of θ with the highest posterior probability. Use the graph of the posterior distribution to identify the posterior mode. Note: The posterior mode is often called the maximum a posteriori or MAP estimate. You don’t need to remember this for the class, but it’s worth mentioning

The maximum would be around theta values of 0.15.