CS7DS3 Assignment 1 2024

Neimhin Robinson Gunning, 16321701, `nrobinso@tcd.ie`

March 11, 2024

Question 1

We consider three potential prior distributions, $p(\theta)$, of the probability that the Pistons win an NBA league basketball game. Games are either won or lost, no draws, and we model the win/lose variable as a binomial distribution, $y_i \sim \mathcal{B}(1, \theta)$.

$$\theta \sim \mathcal{B}e(1, 1) \tag{1}$$

$$\theta \sim \mathcal{B}e(8, 8) \tag{2}$$

$$\theta \sim \mathcal{B}e(7, 13) \tag{3}$$

The mean, mode, variance, and 95% confidence intervals of these three distributions are presented in the listing below.[1]

fig/beta-summaries.txt   Tue Mar 05 15:37:06 2024   1

```
a = 1 , b =  1
Mean: 0.5
Mode: NaN
Variance: 0.08333333
95% Confidence Interval: | 0.025 - 0.975 | =  0.95

a = 8 , b =  8
Mean: 0.5
Mode: 0.5
Variance: 0.01470588
95% Confidence Interval: | 0.2658613 - 0.7341387 | =  0.4682773

a = 7 , b =  13
Mean: 0.35
Mode: 0.3333333
Variance: 0.01083333
95% Confidence Interval: | 0.1628859 - 0.5655016 | =  0.4026157

a = 3 , b =  29
Mean: 0.09375
Mode: 0.06666667
Variance: 0.002574574
95% Confidence Interval: | 0.02041986 - 0.2142162 | =  0.1937963

```

[1] Source code for calculating descriptive statistics in appendix: `src/beta-summary.r`

The $\mathcal{B}e(1,1)$ distribution is the uniform distribution over $[0,1]$, and is least informative possible Beta prior.

The $\mathcal{B}e(8,8)$ has mean 0.5 which mathes our external knowledge that win percentages are centred around 50%. However, we also know that win percentages ranged from 20% to 70%, which constitutes evidence that distribution of win percentages is not balanced around 50%. The $\mathcal{B}e(8,8)$ does not reflect this imbalance, but it is possible that the observed imbalance (0.2 vs $1-0.3$) arose due to chance, so choosing a balanced prior $\mathcal{B}e(8,8)$ is reasonable. The 95% confidence interval of the $\mathcal{B}e(8,8)$ distributions aligns somewhat accurately with the stated knowledge that the range of actual observed win percentages (which equates roughly to θ, although the win percentages range over all teams, not just the Pistons) was between 20% and 70%, but hyperparameters could have been chosen to match this prior knowledge more accurately.

The $\mathcal{B}e(7,13)$ somewhat accurately expresses our belief that that the Pistons will win 34% of games, $7/(13+7) = 0.35$. However, this belief can be expressed with any of $\mathcal{B}e(7k,13k)$, where an increase in k represents an increase in our confidence in the Pistons winning 35% of the games. The choice of $k=1$ reflects the fact the pundit seemed 'very confident'. However, there seems to be no reason why we didn't model the belief precisely as $\mathcal{B}e(\frac{28}{82}k, \frac{54}{82}k)$, again with some choice of k, e.g. $k=20$, to reflect the pundit being 'very confident'.

So which of the $\mathcal{B}e(8,8)$ and $\mathcal{B}e(7,13)$ distributions is a more informative prior? There are several metrics we can use to think about this question, and in all cases $\mathcal{B}e(7,13)$ is the more 'informative' distribution, wholistically.

1) The 95% confidence interval is narrower for $\mathcal{B}e(7,13)$, i.e. range of values we have high confidence of observing is smaller, so we are expressing more certainty with $\mathcal{B}e(7,13)$.

2) The variance is less for $\mathcal{B}e(7,13)$. It is valid to compare the variances of the distributions because their PDFs both have the same domain. The lower variance again expresses greater certainty, i.e. more informativeness.

3) We know that the probability of team x winning a game when x is unknown and we have no other information is 0.5. The $\mathcal{B}e(7,13)$ is opinionated about θ being less than 0.5, whereas $\mathcal{B}e(8,8)$ assigns equal probability to $\theta < 0.5$ and $\theta > 0.5$.

3) The update rule for beta-binomial conjugate sets $a_\nu := a + \sum y_i$ and $b_\nu := b + n - \sum y_i$. The larger our value for a, the more dominating a is in the calculation of a_n, and similarly for b and b_n. In this sense the $\mathcal{B}e(8,8)$ has more influence on a_n than does $\mathcal{B}e(7,13)$, but $\mathcal{B}e(7,13)$ has more influence on b_n than does $\mathcal{B}e(8,8)$. From this more nuanced perspective we can take either distribution to be more informative, depending on which hyperparameter we focus on. To tip the scales, however, the difference in influence on b_n, $13-8=5$, is greater than the difference in influence on a_n, $8-7=1$, and so wholistically $\mathcal{B}e(7,13)$ is the more informative prior.

Question 2

My prefered prior is the uniform distribution $\mathcal{B}e(1,1)$ because of the three, this prior has the least influence on the posterior distribution, i.e. the observed data is given greater weight in the calculation of the posterior. The other priors, while they reflect some aspects of the stated beliefs, still contain numbers chosen arbitrarily, i.e. in an unprincipled fashion. Under the principle of choosing the least informative prior $\mathcal{B}e(1,1)$ is my preference since it essentially says "we know nothing".

Question 3

The posterior of the beta-binomial conjugate is

$$p(\theta|y,a,b) \sim \mathcal{B}e(a_\nu, b_\nu)$$

where $a_\nu := a + \sum_i^n y_i$ and $b_\nu := b + n - \sum_i^n y_i$, and $n = 30$ is the number of trials. We observe $\sum_i^n y_i = 2$. Therefore the posterior is $\mathcal{B}e(3,29)$. The expected value of θ under this distribution is close to 0.1. An important change compared to the prior is that the 95% confidence interval is now quite narrow (about 0.2), whereas the 95% confidence interval of the prior has a width of 0.95.

Question 4

To calculate the probability that the Pistons win at least 8 games out of 52 given our posterior distribution of $\theta \sim \mathcal{B}e(3,29)$ we consider the probability $p(W \geq 8)$ where $W \sim B(52,\theta)$. which can be simulated with Monte Carlo methods using the following process:

1: generate n samples of θ from $\mathcal{B}e(3,29)$
2: for each θ_i generate W_i from $\mathcal{B}(52,\theta_i)$
3: count the occurences of $W \geq 8$ and divide by n

The results of these calculations are presented in Figure 1.[2]

Question

Game 1 requires a stake of €100 and returns €$10 \cdot X$, where X is the number of games won by the Pistons out of the remaining 52 games. The expected return is $E[10 \cdot X - 100] = E[10 \cdot X] - 100 = 10 \cdot E[X] - 100$. Game 2 requires a stake of €1000 and returns €$10 \cdot X^2$, so the expected return is $10 \cdot E[X]^2 - 1000$.

We assume that $\theta \sim \mathcal{B}e(3,29)$ and need the mean of $X \sim \mathcal{B}(52,\theta)$. The Monte Carlo simulation of X is a two step procedure:

1): Randomly generate θ_i from $\mathcal{B}e(3,29)$.

[2]Source code for estimating $p(W \geq 8)$ in appendix: `src/p8.r`

fig/q4.txt Mon Mar 11 12:13:00 2024 1

```
Estimated probability: 0.1973767
Num samples: 1e+07
Estimated probability: 0.8062438
Num samples: 1e+07
Estimated probability: 0.656407
Num samples: 1e+07
Estimated probability: 0.1970951
Num samples: 1e+07
```

Figure 1: $P(W \geq 8)$ is estimated as 0.1969874 with 10,000,000 Monte Carlo samples.

fig/q5.txt Fri Mar 08 12:28:55 2024 1

```
 1: a = 10 , b = 36
 2: Expected return for Game 1: 13.044061
 3: Expected return for Game 2: 462.430625
 4:
 5: a = 9 , b = 41
 6: Expected return for Game 1: -6.40737
 7: Expected return for Game 2: 29.460324
 8:
 9: a = 3 , b = 29
10: Expected return for Game 1: -51.245386
11: Expected return for Game 2: -649.866488
12:
```

Figure 2: Expected values of Games 1 and 2, for each of the priors and their resulting posteriors; $\mathcal{B}e(10,36)$, $\mathcal{B}e(9,41)$, $\mathcal{B}e(3,19)$.

2): Given θ_i generate x_i from $\mathcal{B}(52,\theta_i)$.

We estimate $E[X]$ as $\bar{x}$. The expected returns are presented in lines 9-11 in Figure 2 below. I would prefer to abstain from both games because I should expect to lose money.[3]

Question 6

The calculations of expected returns are highly sensitive to the choice of prior, and impact the ultimate decision on whether to play the games. The $\mathcal{B}e(8,8)$ prior would have had me optimistic about playing either game, while the $\mathcal{B}e(7,13)$ prior would have resulted in both games seeming fairly risky. While the expected return is positive for Game 2 with prior $\mathcal{B}e(7,13)$, the return is only a small fraction of the stake. The $\mathcal{B}e(1,1)$ resulted in my belief that both games are too risky.

[3]Source for esmitating expected returns in appendix: `src/games.r`

```r
df <- data.frame(a = c(3 , 10, 9),
                 b = c(29, 36, 41))
library(ggplot2)
num_samples <- 10000000
monte_carlo_p8 <- function(alpha, beta) {
    theta_samples <- rbeta(num_samples, alpha, beta)
    W_samples <- rbinom(num_samples, 52, theta_samples)
    probability <- mean(W_samples >= 8)
    cat("Estimated probability:", probability,"\n")
    cat("Num samples:", num_samples, "\n")
}

for (i in 1:nrow(df)) {
  pair <- df[i,]
  monte_carlo_p8(pair$a, pair$b)
}

alpha <- 3
beta <- 29
theta_samples <- rbeta(num_samples, alpha, beta)
W_samples <- rbinom(num_samples, 52, theta_samples)
probability <- mean(W_samples >= 8)
cat("Estimated probability:", probability,"\n")
cat("Num samples:", num_samples, "\n")
df <- data.frame(W=W_samples)

ggplot(df, aes(x = W)) +
  geom_bar(aes(y = ..count.. / sum(..count..)),
           fill = "skyblue", color = "black") +
  labs(title = "PMF of W_samples",
       x = "Number of Wins (W)",
       y = "Probability") +
  theme_minimal()

ggsave(argv[1])
```

```r
library(gsl)
library(ggplot2)

beta_entropy <- function(params) {
  a <- params[1]
  b <- params[2]

  if (a <= 0 || b <= 0) {
    stop("Parameters ’a’ and ’b’ must be greater than zero.")
  }

  psi_sum <- psi(a) + psi(b)
  entropy <- log(beta(a, b)) - (a - 1) * psi(a) - (b - 1) * psi(b) + (a + b - 2) * psi_sum

  cat(a,b,a+b,entropy,"\n")
  return(entropy)
}

pairs <- expand.grid(a= 1:10, b=1:10)
pairs$entropy <- apply(pairs, 1, beta_entropy)

ggplot(pairs, aes(x = a + b, y = entropy)) +
  geom_point() +
  labs(x = "a + b", y = "Entropy") +
  ggtitle("Entropy of Beta Distribution vs. a + b")

ggsave(argv[1])
```

```r
library(latex2exp) # for TeX
library(ggplot2)
library(gsl) # for psi

generate_beta_density <- function(a, b) {
  p <- seq(0, 1, length.out = 1000)
  density <- dbeta(p, a, b)
  data.frame(p = p, density = density)
}

beta_entropy <- function(p) {
  a <- p[1]
  b <- p[2]
  psi_sum <- psi(a) + psi(b)
  entropy <- log(beta(a, b)) - (a - 1) * psi(a) - (b - 1) * psi(b) + (a + b - 2) * psi_sum
  return(entropy)
}

hypers <- rbind(c(1, 1), c(8, 8), c(7, 13), c(3, 29), c(10, 36), c(9, 41))
beta_entropies <- apply(hypers, 1, beta_entropy)

df_all <- NULL
for (i in 1:nrow(hypers)) {
  df <- generate_beta_density(hypers[i, 1], hypers[i, 2])
  df$group <- paste("a=", hypers[i, 1], ", b=", hypers[i, 2], "",sep="")
  df_all <- rbind(df_all, df)
}

ggplot(df_all, aes(x = p, y = density, color = group)) +
  geom_line() +
  labs(x = TeX("$\\theta$"), y = TeX("$p(\\theta; a, b)$"), color = "Parameters") +
  theme_minimal() +
  theme(legend.position = "top")

ggsave(argv[1])
```

```r
library(MASS)
library(gsl)

a <- as.integer(argv[1])
b <- as.integer(argv[2])
df <- data.frame(a = c(1, 8, 7,  3 ),
                 b = c(1, 8, 13, 29))

beta.mean <- function(a,b) {
  return(a/(a+b))
}
beta.mode <- function(a,b) {
  return((a-1)/(a+b-2))
}
beta.var  <- function(a,b) {
  return((a*b)/((a+b)*(a+b)*(a+b+1)))
}

beta.entropy <- function(a,b) {
  psi_sum <- psi(a) + psi(b)
  entropy <- log(beta(a, b)) - (a - 1) * psi(a) - (b - 1) * psi(b) + (a + b - 2) * psi_sum
  return(entropy)
}

summarise_beta <- function(params){
  a <- params[1]
  b <- params[2]
  cat("a =",a,", b = ", b,"\n")
  conf_interval <- qbeta(c(0.025, 0.975),a,b)

  cat("Mean:",     beta.mean(a,b), "\n")
  cat("Mode:",     beta.mode(a,b), "\n")
  cat("Variance:", beta.var(a,b),  "\n")
  # cat("Entropy:", beta.entropy(a,b),  "\n")
  cat("95% Confidence Interval: |", conf_interval[1], "-", conf_interval[2], "| = ", conf_interval[2] - conf_interval[1], "\n\n")
}

for (i in 1:nrow(df)) {
  pair <- df[i,]
  summarise_beta(c(pair$a, pair$b))
}
```

```r
library(gsl)
library(ggplot2)

beta.var  <- function(pair) {
  a <- pair[1]
  b <- pair[2]
  return((a*b)/((a+b)*(a+b)*(a+b+1)))
}

pairs <- expand.grid(a= 1:10, b=1:10)
pairs$entropy <- apply(pairs, 1, beta.var)

ggplot(pairs, aes(x = a + b, y = entropy)) +
  geom_point() +
  labs(x = "a + b", y = "Variance") +
  ggtitle("Variance of Beta Distribution vs. a + b")

ggsave(argv[1])
```

```r
library(stats)

# Parameters
alpha <- 3
beta <- 29

df <- data.frame(a = c(10, 9,  3 ),
                 b = c(36, 41, 29))
apply(df, 1, function(row) {
  num_samples <- 10000000
  num_games <- 52
  alpha <- row[1]
  beta <- row[2]

  # Simulate Game 1
  theta_samples <- rbeta(num_samples, alpha, beta)
  W_samples <- rbinom(num_samples, num_games, theta_samples)
  game1_return <-  (10 * W_samples) - 100

  # Simulate Game 2
  game2_return <- (10 * W_samples^2) - 1000

  # Calculate expected return for each game
  expected_return_game1 <- mean(game1_return)
  expected_return_game2 <- mean(game2_return)

  cat(paste("a =", row[1], ", b =", row[2]), "\n")
  cat(paste("Expected return for Game 1:", expected_return_game1), "\n")
  cat(paste("Expected return for Game 2:", expected_return_game2),"\n")
  cat("\n")
})
```

```r
library(stats)

# Define the functions for p(W >= 8 | theta) and p(theta)
p_W_given_theta <- function(theta) {
  1 - pbinom(7, 52, theta)
}

p_theta <- function(theta) {
  dbeta(theta, 3, 29)
}

# Compute the joint probability by integrating the product of the two functions
joint_probability <- function(theta) {
  p_W_given_theta(theta) * p_theta(theta)
}

# Integrate the joint probability function numerically
result <- integrate(joint_probability, lower = 0, upper = 1)

# The result$value contains the estimated probability
print(paste("Estimated probability:", result$value))
```