Statstical Methods in Data Science II

SDS II - Homework 1

Luca Mazzucco: 1997610

A.Y. 2023-2024

A) Simulation:

Joint probability distribution:

```
##      1    2    3
## 1 0.06 0.17 0.10
## 2 0.10 0.12 0.11
## 3 0.14 0.02 0.18
```

In order to check that joint probability distribution J is a probability distribution:

- **Non-negativity**: all probabilities in the distribution must be non-negative
- **Summation to 1**: sum of all probabilities over all the possible outcomes equal to 1

```
P_y <- rowSums(J) # marginal of Y
P_z <- colSums(J) # marginal of Z

c(sum(P_y) == 1, sum(P_z) == 1)  # marginal of Y, marginal of Z
```

```
## [1] TRUE TRUE
```

From the joint distribution J can be derived *six conditional distributions*:

One for each event Y given Z (z={1,2,3}).

```
##         1         2         3
## 0.2000000 0.3333333 0.4666667
```

```
##          1          2          3
## 0.54838710 0.38709677 0.06451613
```

```
##         1         2         3
## 0.2564103 0.2820513 0.4615385
```

And same for Z given Y (y={1,2,3})

```
##         1         2         3
## 0.1818182 0.5151515 0.3030303
```

```
##         1         2         3
## 0.3030303 0.3636364 0.3333333
```

```
##          1          2          3
## 0.41176471 0.05882353 0.52941176
```

Make sure they are probability distributions:

```
c(sum(J[, 1]/P_z[1]), sum(J[, 2]/P_z[2]), sum(J[, 3]/P_z[3])) # each distrib sum up to one
```

```
## [1] 1 1 1
```

```
c(sum(J[1, ]/P_y[1]), sum(J[2, ]/P_y[2]), sum(J[3, ]/P_y[3]))
```

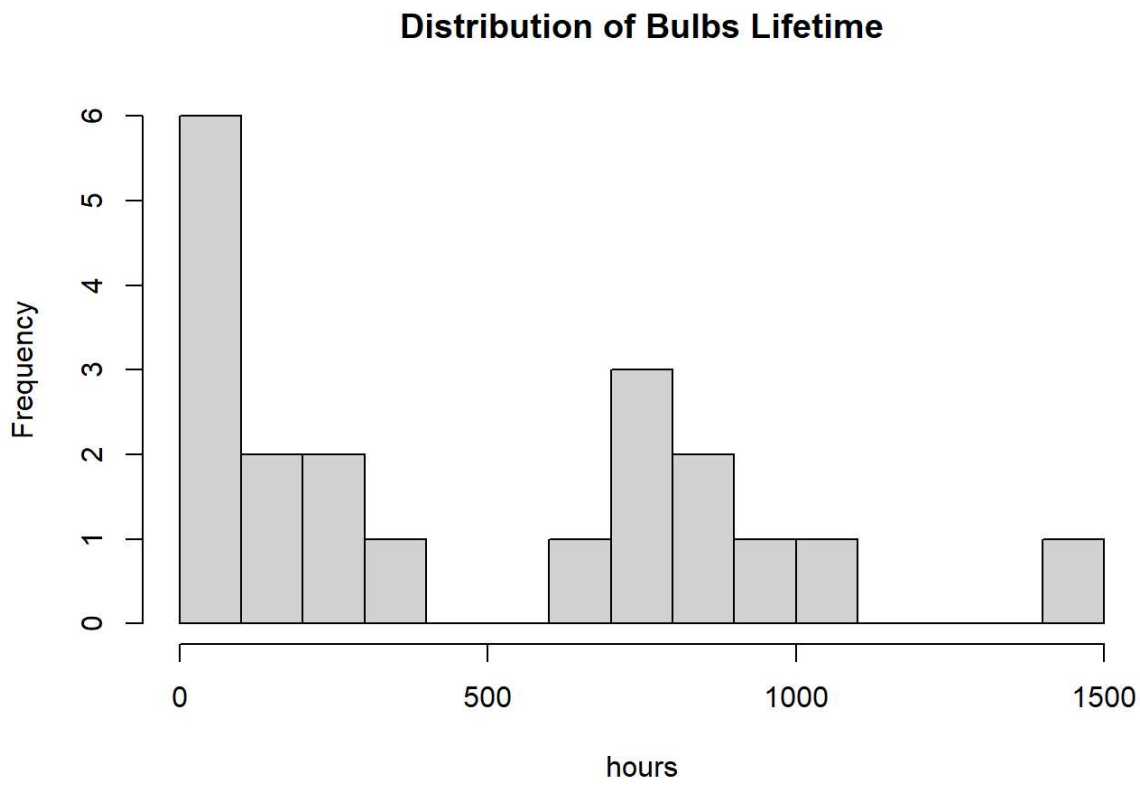
```
## [1] 1 1 1
```

Simulate from this J distribution:

```
events_y <- 1:3
events_z <- 1:3

n_simulations <- 100
simulated_data <- numeric(n_simulations)

for (i in 1:n_simulations) {
  y <- sample(events_y, 1, prob = rowSums(J))
  z <- sample(events_z, 1, prob = J[y, ])
  simulated_data[i] <- z}

# simulated_data
```

B) **Bulbs lifetime**: a conjugate Bayesian analysis of exp data:

```
y_obs <- c(1, 13, 27, 43, 73, 75, 154, 196, 220, 297, 344, 610, 734,    #Bulbs lifetime data (in hours)
                                  783, 796, 845, 859, 992, 1066, 1471)
```



Main ingredients of the **Bayesian model** are:

- **Assumption** on the **statistical model**: that represent our beliefs about the probability distribution of the observed data Y, given a specific parameter θ.
  Exponential model → $Y_i|\theta \sim \text{Exp}(\theta)$
- **Prior distribution -** π(θ) of the unknown parameter of interest → $\theta \sim Gamma(r, s)$
  Here used the gamma distribution, which is a conjugate prior distribution for the exponential model.
- **Posterior distribution** π(θ|y) → what we are really interested, it's output of the model, used to make inference
  Updated based on the observed data through the **Bayes rule**

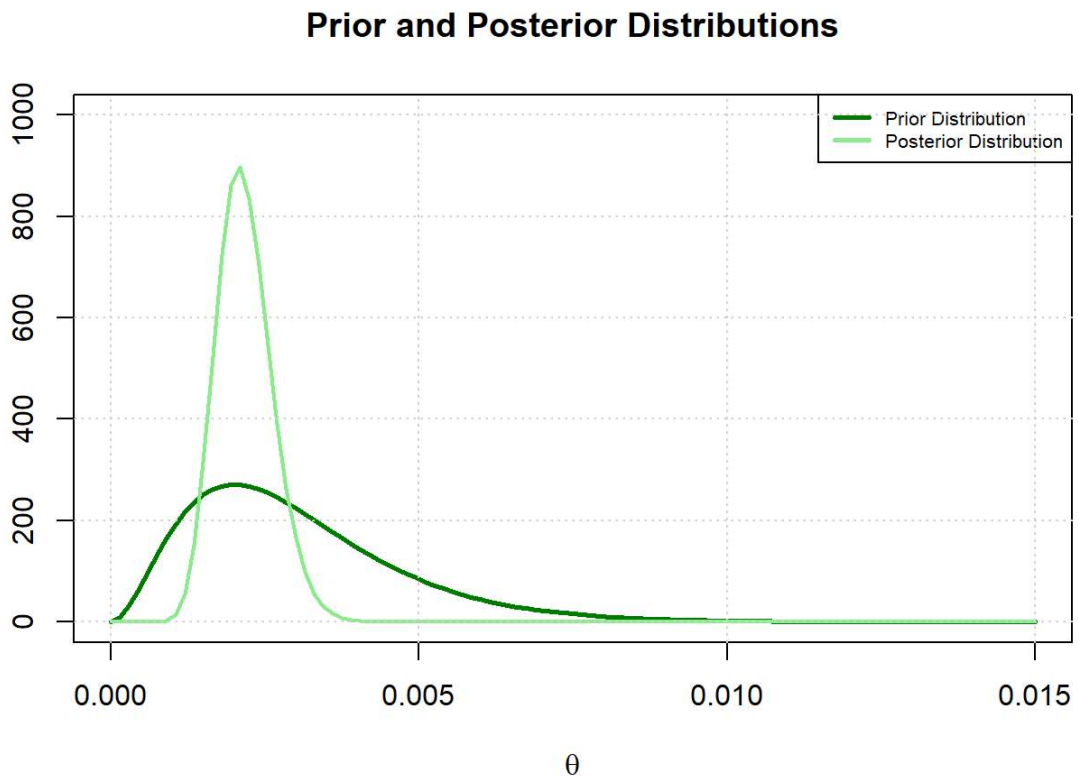
Choose a conjugate prior distribution π(θ):

```
mean = 0.003
std = 0.00173

r_prior <- mean/std^2
s_prior <- mean^2/std^2
```

```
## Prior parameters of the gamma distribution: r = 1002.372 s = 3.007117
```

```
sum_y <- sum(y_obs)
n <- length(y_obs)

r_post <- r_prior + sum_y
s_post <- s_prior + n
```

```
## Post parameters of the gamma distribution: r = 10601.37 s = 23.00712
```



Point estimation analysis on θ:

```
## posterior median = 0.002138842
```

```
## posterior mean = 0.002170202
```

```
## posterior var = 0.0000002047095
```

```
## prior_mean = 0.003
## prior_var = 0.0000029929
```

Variance from Posterior distribution is much lower than Prior variance. Uncertainty in Prior was reduced thanks to the observation of the data.
This explains the Prior distribution is vaguer than the Posterior, because it captures a wider uncertainty
Also the mean has been lowered, moving from 0.003 to 0.0021 in Posterior distribution.

Relevant information learnt about the average lifetime of the bulb (ψ = 1/θ):

```
## Posterior statistics on bulbs lifetime:
```

```
## median = 467.5428
```

```
## mean = 460.7866 → hours average lifetime
```

```
## mean of observated data = 479.95
```

The posterior mean shows average lifetime is equal to 460.78 hours and corresponds quite well to the average of the observed data.

Comparing the Prior and Posterior distributions, the average lifetime of bulbs increased:

```
## prior_mean = 333.3333 hours
```

Finding the probability that the **average bulb lifetime exceeds 550 hours**.
Using the cumulative distribution function from the posterior gamma distribution learned before:

```
avg_value <- pgamma(1/550, shape = s_post, rate = r_post)  # ψ = 1/ϑ
```

```
## the probability of a bulb with lifetime greater than 550 is : 0.2254117
```

C) Exchangeability

Given a sequence of random variables $X_1, \ldots, X_n, \ldots$ it is **exchangeable** if
for any k-tuple $(n_1, \ldots, n_k)$ and any permutation$\sigma = (\sigma_1, \ldots, \sigma_k)$ of the first k integers, the following holds:

$$(X_{n_1}, \ldots, X_{n_k}) \stackrel{d}{=} (X_{n_{\sigma_1}}, \ldots, X_{n_{\sigma_k}})$$

From the **De Finetti theorem**:

If $X_1, \ldots, X_n, \ldots$ is an exchangeable process of binary random variables, there exists a distribution π on [0, 1] such that

$$\Pr(X_1 = x_1, \ldots, X_n = x_n) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \pi(\theta) d\theta$$

Where the random variables are conditionally independent and identically distributed.