TMA4300 - Exercise 1

Stochastic simulation

Marcus A. Engebretsen and Gina Magnussen

Problem A

Probability integral transform, rejection sampling and bivariate techniques

1. Sampling from g(x) - Probability integral transform

The probability density function

$$g(x) = \begin{cases} cx^{\alpha-1}, & 0 < x < 1 \\ ce^{-x} & 1 \leq x \\ 0 & \text{otherwise,} \end{cases}$$

is given with c as a normalising constant and $\alpha \in (0,1)$.

a)

The cumulative distribution function $G_X(x)$ is then found by integrating over the different domains and taking into account that the cdf must be continuous,

$$G_X(x) = \begin{cases} \frac{c}{\alpha} & 0 < x < 1, \\ \frac{c}{\alpha} + ce^{-1} - ce^{-x} & 1 \leq x \\ 0 & \text{else} \end{cases}$$

By using the property that the area under a pdf should integrate to 1, we find that $c = \frac{e\alpha}{e+\alpha}$

The probability integral transform, setting $u = G_X(x)$ and solving for x, gives

$$x = G^{-1}(u) = \left(u\frac{e+\alpha}{e}\right)^{\frac{1}{\alpha}}, \qquad 0 < x < 1$$
$$x = G^{-1}(u) = -\log\left[\left(\frac{e+\alpha}{e\alpha}\right) - \frac{u}{c}\right] = -\log\left[\frac{1}{c}(1-u)\right] \qquad 1 \leq x$$

b)

R function

R plot: gsample

2. Gamma distribution with $\alpha \in (0,1)$, $\beta = 1$ generated by rejection sampling

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)}x^{\alpha-1}e^{-x}, & 0 < x \\ 0 & \text{else} \end{cases}$$

with $\alpha \in (0,1)$ and $\beta = 1$.

a)

The acceptance probability R:

$$\frac{1}{\gamma}\frac{f(x)}{g(x)}$$

$$\gamma_1 = \frac{f(x)}{g(x)} = \begin{cases} \frac{1}{\Gamma(\alpha)}\frac{e^{-x}}{c} & 0 < x < 1 \\ \frac{1}{\Gamma(\alpha)}\frac{x^{\alpha-1}}{c} & x \geq 1 \end{cases}$$

$$\frac{\partial \gamma_1}{\partial x} = \begin{cases} -\frac{1}{\Gamma(\alpha)}\frac{e^{-x}}{c} & 0 < x < 1 \\ \frac{\alpha-1}{\Gamma(\alpha)}\frac{x^{\alpha-2}}{c} & x \geq 1 \end{cases}$$

Solving $\frac{\partial \gamma_1}{\partial x} = 0$ for x and inserting into the expression for R will now maximize the acceptance probability given the constraint that $f(x) < \frac{1}{\gamma}g(x)$. CHECK CONSTANTS!!

$$\gamma_1 = \begin{cases} \frac{1}{\Gamma(\alpha)}\frac{1}{c} & 0 < x < 1 \\ \frac{1}{\Gamma(\alpha)}\frac{1}{c} & x \geq 1 \end{cases}$$

which in turn gives an acceptance probability

$$R = \begin{cases} e^{-x} & 0 < x < 1 \\ x^{\alpha-1} & x \geq 1 \end{cases}$$

b) R function

3. Ratio of uniforms - gamma, $\alpha > 1, \beta = 1$.

Consider now the same distribution, but with parameters $\alpha > 1$ and $\beta = 1$. The ratio of uniforms method is used to simulate samples from this distribution. With C_f, f^*, a, b_+ and b_- as given in formula (3) and (4) in the exercise description, we find that

Need to find maximum of $f^*(x)$ to find the bounds for the area C_f. Since $f^*(x)$ is a concave function, we solve

$$\frac{df^*(x)}{dx} = (\alpha-1)x^{\alpha-2} - x^{\alpha-1}e^{-x} = 0 \implies x = \alpha - 1 \frac{d(x^2 f^*(x))}{dx} = 0 \implies x = \alpha + 1$$

$$\begin{aligned} f^*(\alpha-1) &= (\alpha-1)^{\alpha-1}e^{1-\alpha} \\ a &= \sqrt{\sup_x f^*(x)} = \sqrt{f^*(\alpha-1)} = \sqrt{(\alpha-1)^{\alpha-1}e^{1-\alpha}} \\ b_+ &= \sqrt{\sup_{x\geq 0} x^2 f^*(x)} = \sqrt{(\alpha+1)^2 f^*(\alpha+1)} = (\alpha+1)^{\frac{\alpha+1}{2}}e^{-\frac{\alpha+1}{2}} \\ b_- &= (\sqrt{\sup_{x\leq 0} x^2 f^*(x)?}) = 0 \end{aligned}$$

$$C_f = [0,a]\times[b_-,b_+] = \left[0, \sqrt{(\alpha-1)^{\alpha-1}e^{1-\alpha}}\right] \times \left[0, (\alpha+1)^{\frac{\alpha+1}{2}}e^{-\frac{\alpha+1}{2}}\right]$$

Ratio of uniforms:

$$x_1 = \frac{u_2}{u_1} \qquad\qquad x_2 = u_1$$

Problem B

The Dirichlet distribution: Simulating using known relations

$$
\begin{aligned}
f_z(z,\alpha) = dz_1...dz_k &\propto (z_1^{\alpha_1-1})e^{-z_1}...(z_k^{\alpha_k-1})e^{-z_k}dz_1...dz_k \\
&= z_1^{\alpha_1-1}...z_k^{\alpha_k-1}e^{-(z_1+...+z_k)}dz_1...dz_k \\
&= z_1^{\alpha_1-1}...z_k^{\alpha_k-1}e^{-v}dz_1...dz_k \\
\text{where } \ v = -(z_1+...+z_k)
\end{aligned}
$$

Change of variables

$$z_i = x_i \cdot v \implies dz_i = dx_i \cdot v + x_i dv$$

Using $\sum_{i=1}^{k} x_i = 1$ and $dx_1 + ... + dx_k = 0$ Define $w = dz_1 + ... + dz_{k-1} = [dx_1 + ... + dx_{k-1}]v + [x_1 + ... + x_k]dv$

Then

$$
\begin{aligned}
dz_k &= dx_k v + x_k v \\
&= -[dx_1 + ... + dx_{k-1}]v + [1 - [x_1 + ... x_{k-1}]]dv \\
&= dv - ([dx_1 + ... + dx_{k-1}]v + [x_1 + ... x_{k-1}]dv) \\
&= dv - w
\end{aligned}
$$

Using exterior algebra:

$$
\begin{aligned}
dz_1 \wedge ... \wedge dz_{k-1} \wedge dz_k &= (dz_1 \wedge ... \wedge dz_{k-1}) \wedge (dv - w) \\
&= ... \\
&= v^{k-1} dx_1 \wedge ... \wedge dx_{k-1} \wedge dv
\end{aligned}
$$

Filling into the expression gives:

$$
\begin{aligned}
f_z(z,\alpha)dz_1...dz_k &\propto (x_1 v)^{\alpha_1-1}...(x_{k-1}v)^{\alpha_{k-1}-1}(v(1-[x_1+...+x_{k-1}]))^{\alpha_k-1}e^{-v^{k-1}}dx_1 \wedge ... \wedge dx_{k-1} \wedge dv \\
&= v^{\alpha_1+...+\alpha_{k-1}}e^{-v}dv \left(x_1^{\alpha_1-1}...x_{k-1}^{\alpha_{k-1}-1}\right)\left(1-\sum_{i=1}^{k-1} x_i\right)^{\alpha_{k-1}} dx_1...dx_{k-1}dv
\end{aligned}
$$

```r
# -------------------- Sample Dirichlet ----------------------#
n <- 10000
K <- 3
alpha <- 1:K


zSample <- matrix(0,n,K)
dirSample <- matrix(0,n,K)
dirMean <- matrix(0,K,1)

for (i in 1:n) {
  for (j in 1:K){
    zSample[i, j] <- sampleGamma(alpha[j], 1, 1)
  }
  for (j in 1:K){
    dirSample[i, j] <- zSample[i,j]/(sum(zSample[i,]))
  }
}
```

```
for (j in 1:K) {
  dirMean[j] <- mean(dirSample[j,])
}

# Verify marginal distributions are beta
xUni <- seq(0,1,1/n)
betaSamples_1 <- dbeta(xUni, alpha[1], sum(alpha)-alpha[1])
betaSamples_2 <- dbeta(xUni, alpha[2], sum(alpha)-alpha[2])
betaSamples_3 <- dbeta(xUni, alpha[3], sum(alpha)-alpha[3])


# Plotting
# par(mfrow=c(1,3))
truehist(dirSample[,1])
lines(xUni, betaSamples_1, col = "red", lwd = 2)

truehist(dirSample[,2])
lines(xUni, betaSamples_2, col = "red", lwd = 2)

truehist(dirSample[,3])
lines(xUni, betaSamples_3, col = "red", lwd = 2)
```

Problem C

A toy Bayesian model: Birthdays

The probability that two or more students has their birthday on the same day can be simulated as follows: Randomly assign a birth date to the 35 in a given NTNU class, and check for duplicate dates. If so, count this event as a success. If not, assign dates randomly again and to the same check. Do this many times. The estimate is then the number of successes divided by the number of trials.

1: Independent birthdays, equally likely

a)

```
# Birthdays
sim <- 10000
stud <- 35
count <- 0

for (i in 1:sim){
  bdays <- round(runif(stud)*365)
  if (sum(duplicated(bdays))>=1){
    count <- count + 1
  }
}

prob <- count/sim
print(prob)
```

b)

Exact calculation: The probability that two or more students has their birthday on the same day, is the complement of the event that no students has their birthday on the same day. Thus

$$\begin{aligned} P(\text{no. of students with same birthday} \geq 2) &= 1 - P(\text{no students with same birthday}) \\ &= 1 - \frac{365!}{330!\, 365^{35}} = 0.8144 \end{aligned}$$

The difference between the simulated and exact answer is of order 10^{-3} or better with enough simulations, meaning the simulated probability is good.

2: Bayesian model

a)

The posterior distribution of (q_1, q_2, q_3, q_4), the distribution of the probabilities of being born in a specific season given observations, is

$$\begin{aligned} p(q_{1:4}|x_{1:4}) &= \frac{p(x_{1:4}|q_{1:4})P(q_{1:4})}{\int p(x_{1:4}|q_{1:4})P(q_{1:4})dq_{1:4}} \\ &\propto p(x_{1:4}|q_{1:4})P(q_{1:4}) \\ &\propto \prod_{i=1}^{k} q_i^{x_i} \cdot \prod_{i=1}^{k} q_i^{\alpha_i - 1} \\ &\propto \prod_{i=1}^{k} q_i^{x_i + \alpha_i - 1} \\ &\propto \prod_{i=1}^{k} q_i^{\bar{\alpha_i} - 1} \end{aligned}$$

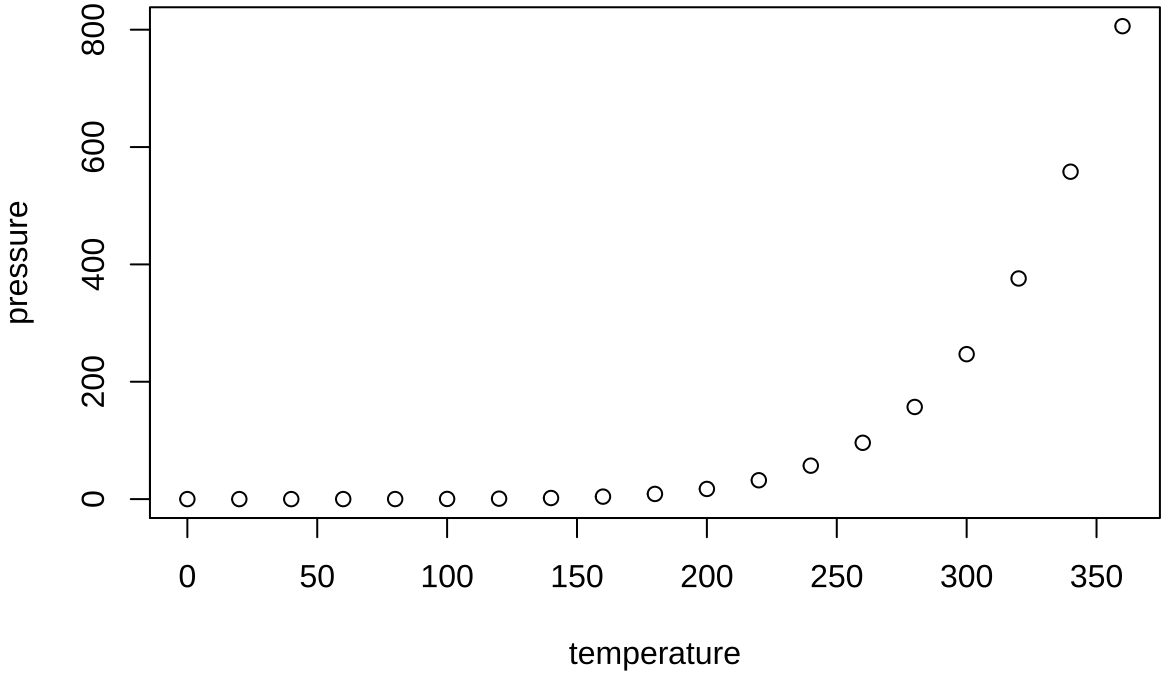
where $(x_{1:4}) = (x_1, x_2, x_3, x_4)$ and similary for q. Thus, the posterior distribution of (q_1, q_2, q_3, q_4) is a Dirichlet distribution with parameters $\bar{\alpha} i = x_i + \alpha_i$, $i = 1, ..., 4$.

b)

When the joint posterior distribution of $q_{1:4}$ is Dirichlet, the marginal distribution for each season is $q_i \sim beta\left(\alpha_i, \sum_{k=1}^{4} \alpha_k - \alpha_i\right)$.

These marginal distributions are plotted by simulating several Dirichlet distributions $x = (x_1, ..., x_K)$, extracting for instance all x_1s, which are beta distributed with the same parameters, and making a histogram of these samples. The built-in function in R is used to verify that the simulation of the beta distributions, and thus also the Dirichlet distribution. Another option is to simulate the marginal beta distributions by using the relation between the gamma and the beta distribution, since we in the previous tasks have shown this sample generator to be reasonably correct.

%% sampleBeta.R



3: Sample/Estimate p

Given the probabilities q_i, of being born within a specific season we want to find the probability of p. This is here done by simulation. Denote

$$N_j = \text{no. of students born in season j}, \;\; j \in \{1 = \text{spring}, ..., 4 = \text{winter}\}$$

and set $m = 35$ again.

a) Sampling from posterior distribution $P(q_1, q_2, q_3, q_4|x_1, x_2, x_3, x_4)$ and $P(N_1, N_2, N_3, N_4|q_1, q_2, q_3, q_4))$

Sampling from posterior distribution $p(q_1, q_2, q_3, q_4|x_1, x_2, x_3, x_4)$

```
# -------------- Multinomial sampling -----------------#
# Input:
# p : Sorted list of probabilities summing to unity.
# N : Number of draws
# Output:
# out: vector of draws corresponding to each interval

sampleMultinomial <- function(p, N) {

  len_p <- length(p)

  # Compute cumulative sum
  cumSum <- matrix(0,len_p, 1)
  cumSum[1] <- p[1]
  for (i in 2:len_p) {
    cumSum[i] <- cumSum[i-1] + p[i]
  }
  u <- runif(N)

  # Increment output draws
  out <- matrix(0, len_p, 1)
  for (i in 1:N) {
    for (j in 1:len_p) {
      if (u[i] < cumSum[j]) {
        out[j] <- out[j] + 1
        break
      }
    }
  }
  return (out)
}
```

As in C.2, the distribution of $(N_1, N_2, N_3, N_4 | q_1, q_2, q_3, q_4)$ will be multinomial with parameters (m, q_1, q_2, q_3, q_4), but with $m = 35$.

$(q_1, q_2, q_3, q_4) \sim Dirichlet(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$
$(N_1, N_2, N_3, N_4) \sim multinomial(m, q_1, q_2, q_3, q_4)$

How to sample from a multinomial distribution: Divide the unit interval $[0, 1]$ into four subsequent intervals $[0, q_1] \cup (q_1, q_1 + q_2] \cup (q_1 + q_2, q_1 + q_2 + q_3] \cup (q_1 + q_2 + q_3, q_1 + q_2 + q_3 + q_4]$ with length q_1, q_2, q_3, q_4 respectively. Then sample $u \sim Unif[0, 1]$ and increment N_j in the corresponding interval, in this case season, for $j = 1, ..., 4$. That is, if u is between q_1 and $q_1 + q_2$, then N_2 is incremented. Do this $m = 35$ times.

b)

A formula for

$$p = f(N_1, N_2, N_3, N_4)$$

can be found by seeing that

$$\begin{aligned} p &= P\left[(Y_1 \cup Y_2 \cup Y_3 \cup Y_4) = 1\right] \\ &= 1 - P\left[(Y_1 \cup Y_2 \cup Y_3 \cup Y_4) = 0\right] \end{aligned}$$

where Y_j is the event that two or more students have the same birthday in season j.

Since birthdays between different students are independent of each other, we can assume that the event of no students having the same birthday in the spring is independent of the event that no students have the same birthday in the winter, since the events are disjoint. So

$$p = 1 - [P(Y_1 = 0) + P(Y_2 = 0) + P(Y_3 = 0) + P(Y_4 = 0)]$$

Now, since the outcome of Y_i is binary, we can render the above equation to a function of N_i using that K_i is the number of days in each season. This gives

$$\begin{aligned} P(Y_i = 0) &= \frac{K_i(K_i - 1)(K_i - 2)...(K_i - (N_i - 1))}{K_i!} \\ &= \binom{K_i}{N_i} \cdot \frac{N_i!}{K_i^{N_i}} \end{aligned}$$

and so

$$p = 1 - \left[\sum_{i=1}^{4} \binom{K_i}{N_i} \cdot \frac{N_i!}{K_i^{N_i}}\right]$$

Now p can be calculated by sampling from $(N_1, N_2, N_3, N_4 | q_1, q_2, q_3, q_4)$, the multinomial distribution, calculating p by using the formula above, repeating this n times and then calculate the mean.

Confidence interval of posterior mean of p: Sort all the n found values for p, and then remove the 2.5% highest and lowest values.

Posterior mean is larger than the probability in C.1. This is reasonable because the probabilities q determine how likely it is to have a birthday within one season. If one q_i is larger than another, there are more chances of birthdays colliding within that season, and thus the probability of two or more people having the same birthday increases.

```r
# ----------------------------- Sample posterior means, 95 % confidence -----------------------------#
findIndex <- function(vec){
  for (i in 1:length(vec)) {
    if (vec[i] == 1) {
      return (i)
    }
  }
  return(-1)
}

calcP <- function(weights, N, K) {

  # Choose random season
  u <- runif(1)
  p <- c(0.25, 0.5, 0.75, 1.0)
  out <- sampleMultinomial(p, 1)
  index <- findIndex(out)
  temp <- weights[index] * factorial(K[index]) / factorial(K[index]-N[index]) * 1 / K[index]^N[index]
  return (1 - temp)
}

samplePosterior <- function(p, students, draws, daysInSeason) {

  post <- matrix(0, draws, 1)
  out <- matrix(1, length(p),1)

  # For efficiency
  #p <- sort(p, ndex.return = FALSE)

  for (i in 1:draws){
    out <- sampleMultinomial(p, students)
    post[i] <- calcP(p, out, daysInSeason)
  }

  return (post)
}

computePosteriorMeans <- function(p, students, draws, daysInSeason) {

  post <- samplePosterior(p, students, draws, daysInSeason)

  return (post)
}
```

```
students <- 35
draws <- 100000

daysInSeason <- matrix(c(92, 92, 91, 90))
p <- matrix(c(92 / 365, 92 / 365, 91 / 365, 92 / 365))

means <- computePosteriorMeans(p, students, draws, daysInSeason)
means_sorted <- sort(means)
truehist(means_sorted[2500:97500])
```

c)

Assume now a Dirichlet prior with $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 20$.

Importance sampling

Comment and explain result: Small equal alphas won't affect the distribution to much. Basicly implying that the observations will play the more important role in terms of the resulting posterior mean probability. Bigger alphas with all being equal will give uniformly distributed birthdays between seasons, thereby pushing the q_i closer to 0.25. In other words, the observations will become less important as alpha increases and the result should resemple the probability obtained in C1 more.

When on the other hand alphas are tuned to different values the result will vary differently. Consider the case where a single alpha value is bigger than the others. When this happens the pdf will get weighted towards the large alpha value thereby increasing the chance of colliding birthdays, implying an increase in p.

Since we don't have any simulation result this is all based on intuition and discussion with fellow students and should not be consider as a tested result, but rather a theory that our group has regarding the behaviour of the posterior mean with respect to a varying alpha.