Jens Andreas Teigland Holck

The Birthday Paradox Extended

The birthday paradox is a well-known statistical problem which concerns the probability that in a set of n randomly chosen, some pair of them will have the same birthday. The name itself is actually somewhat misleading as the solution is not a paradox at all, but rather counter-intuitive. In a group of 23 people or more, there is a more than a 50% chance that 2 people or more will share a birthday. How could this be?

Let us first assume for simplicity that there are 365 days in every year, thus ignoring leap ears and that each day of the year is equally likely for a birthday. We define A as the event that there are at least two people born on the same day. It is easier to determine A^c and then use the relation $P(A) + P(A^c) = 1$. The probability that exactly two people does not have the same birthday is given by 365/365 · 364/365. For exactly three people there are 363 birthdays left that is different from the first two so that the probability that exactly three people do not have the same birthday is given by 365/365 · 364/365 · 363/365. For n people we will then have

$$P(A^c) = \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \ldots \cdot \frac{365-(n-1)}{365} = \frac{365!}{(365-n)! \cdot 365^n}$$

thus $$P(A) = 1 - \frac{365!}{(365-n)! \cdot 365^n}$$

Setting $n = 23$, we find that $P(A) \approx 50.7\%$.

In reality though, things are a bit more complicated. In Norway, the most common birthday month is April and the least common is November. We will now dismiss some of our earlier assumptions, namely that all days of the year are equally likely for a birthday, i.e. different days of the year have different probabilities. We divide the year into four seasons with 92 days in spring and summer, 91 days in autumn and 90 days in winter. We denote the probability of being born in spring, summer, autumn and winter by *q1*, *q2*, *q3* and *q4*, respectively. These probabilities will obviously sum to 1. We adopt a Bayesian model and assume a Dirichlet prior distribution for (*q1*, *q2*, *q3*, *q4*) with *α1*=*α2*=*α3*=*α4* = 0.5. To obtain information about these probabilities we randomly select *m* = 200 people and count the number of these *m* people that are born in spring (*x1*), summer (*x2*), autumn (*x3*) and winter (*x4*). We assume (*x1*, *x2*, *x3*, *x4*)|(*q1*, *q2*, *q3*, *q4*) to have a multinomial distribution with parameters (*m*, *q1*, *q2*, *q3*, *q4*). Assume we observed *x1* = 55, *x2* = 57, *x3* = 48 and *x4* = 40.

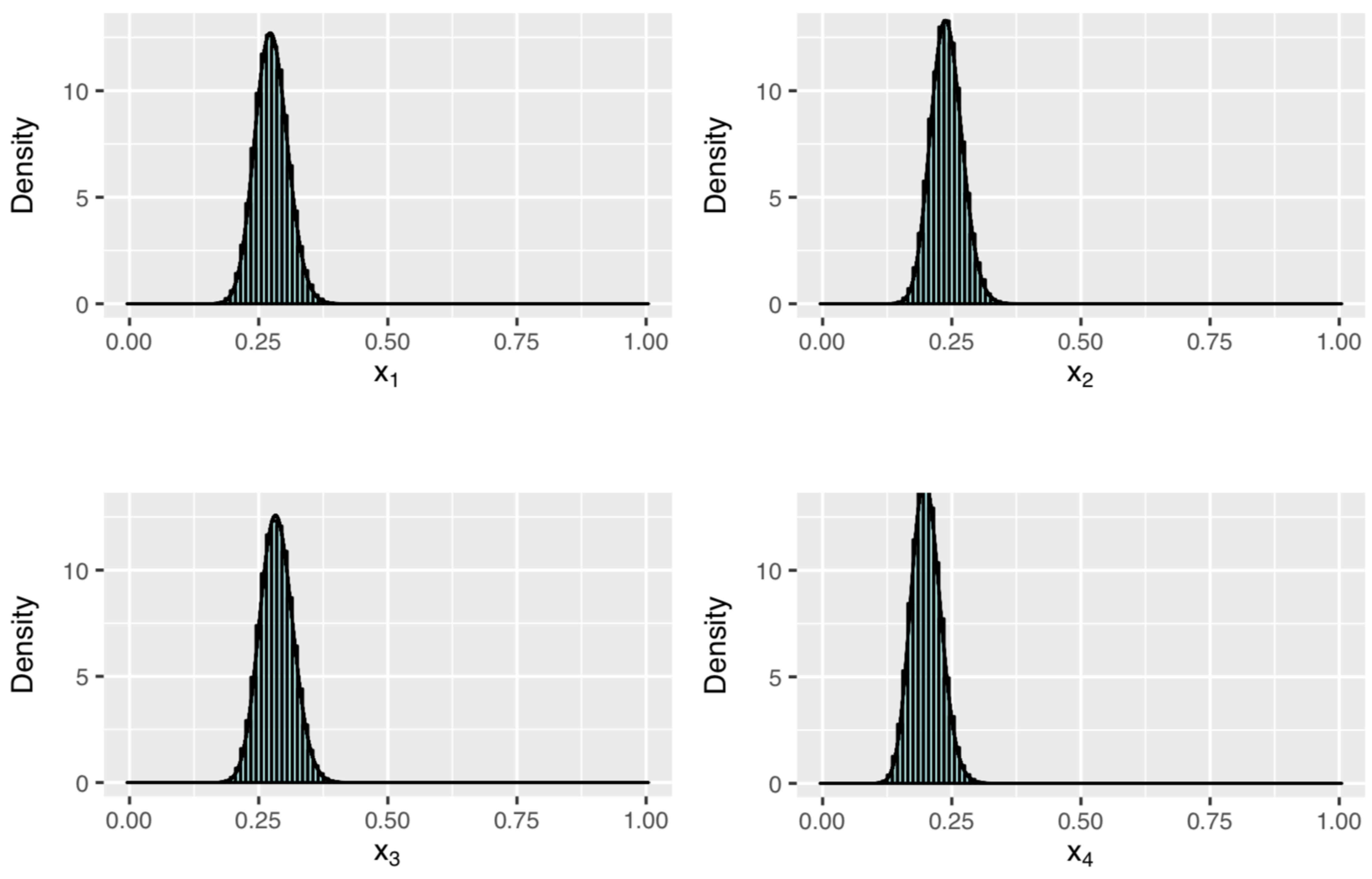
We first show that the posterior distribution of (*q*1, *q*2, *q*3, *q*4) is a Dirichlet distribution. It is enough to show that

$$f(q|x) \propto f(x|q) \cdot f(q).$$

Inserting for the multinomial distribution of *f*(x|*q*) and the Dirichlet prior distribution for (*q*1, *q*2, *q*3, *q*4), we end up with

$$\begin{aligned} f(q|x) &\propto \cdot q_1^{x_1} \cdot \ldots \cdot q_K^{x_K} \cdot f(q) \\ &\propto q_1^{x_1} \cdot \ldots \cdot q_K^{x_K} \cdot \cdot q_1^{\alpha_1 - 1} \cdot \ldots \cdot q_{K-1}^{\alpha_{K-1}-1} \cdot \left(1 - \sum_{k=1}^{K-1} q_k\right)^{\alpha_K - 1} \\ &\propto q_1^{x1+\alpha_1-1} \cdot \ldots \cdot q_{K-1}^{x_{K-1}+\alpha_{K-1}-1} \cdot \left(1 - \sum_{k=1}^{K-1} q_k\right)^{x_K+\alpha_K-1} \qquad \sim Dirichlet(x+\alpha) \end{aligned}$$

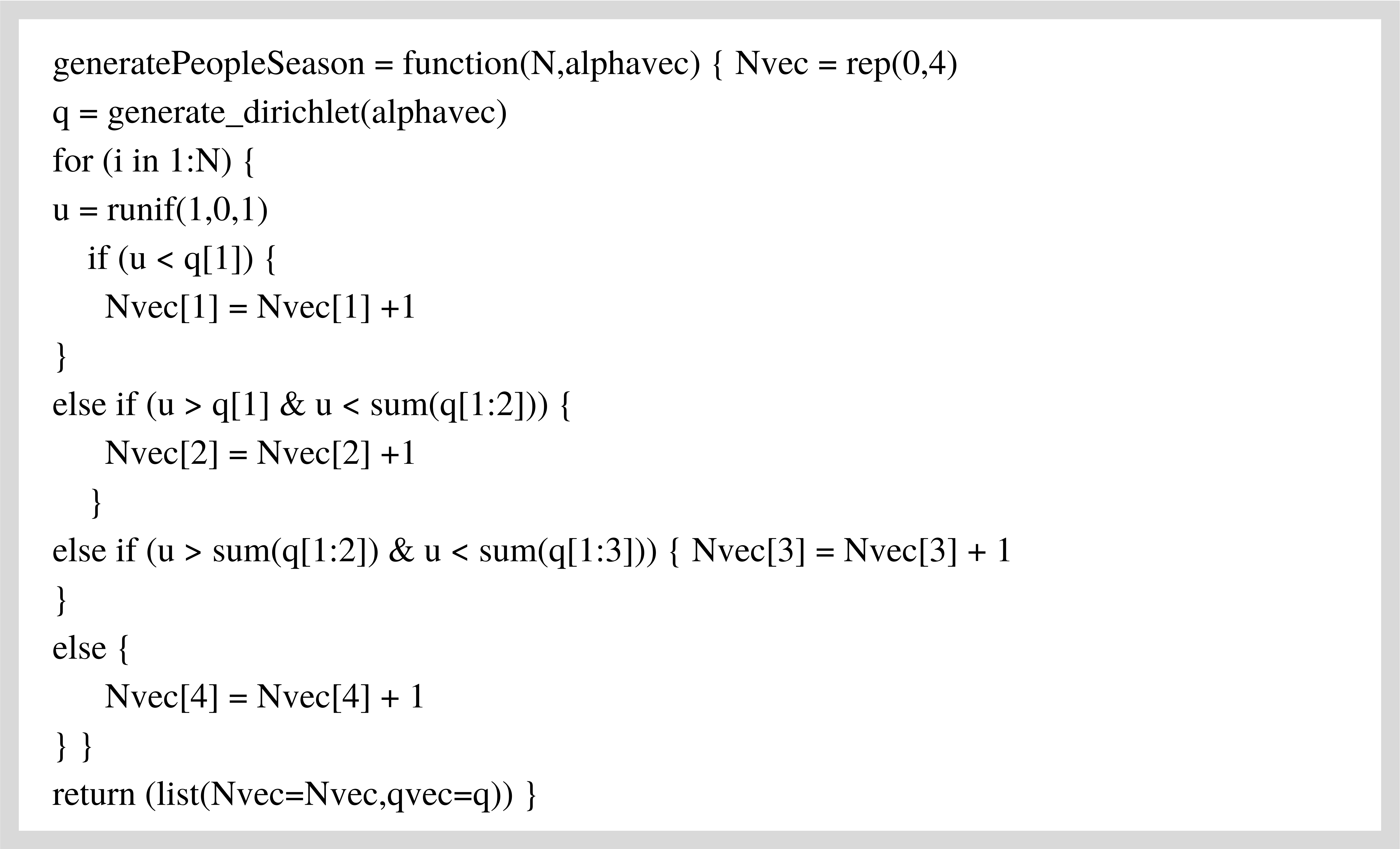
We plot the marginal posterior distribution of each qi. Each qi will now be Beta distributed, $q_i \sim Beta(\alpha_i, \sum_{k=1}^{4} \alpha_k - \alpha_i)$. Figure 6 shows a histogram of the samples for each *qi* together with the pdf of the beta distribution.



Assuming that the probability of having birthday at a specific day of the year is determined through the probabilities ($q1$,$q2$,$q3$,$q4$), we now want to find the probability, p, of two or more people having birthday on the same day. However, it is not easy to directly determine the posterior distribution of p, so we instead estimate p by simulation.

Let $N1$ denote the number of people with birthday in the spring, and let $N2$, $N3$ and $N4$ correspondingly denote the number of people with birthday in the summer, autumn and winter, respectively.

The following function returns a sample from the posterior distribution of ($q1$, $q2$, $q3$, $q4$) and a correspoding sample of ($N1$, $N2$, $N3$, $N4$) given these probabilities.



```
generatePeopleSeason = function(N,alphavec) { Nvec = rep(0,4)
q = generate_dirichlet(alphavec)
for (i in 1:N) {
u = runif(1,0,1)
  if (u < q[1]) {
     Nvec[1] = Nvec[1] +1
}
else if (u > q[1] & u < sum(q[1:2])) {
     Nvec[2] = Nvec[2] +1
   }
else if (u > sum(q[1:2]) & u < sum(q[1:3])) { Nvec[3] = Nvec[3] + 1
}
else {
     Nvec[4] = Nvec[4] + 1
} }
return (list(Nvec=Nvec,qvec=q)) }
```

We now find a formula for *p* as function of (*N*1,*N*2,*N*3,*N*4) and use this to estimate the posterior mean of *p*. The approach is similar to as before. We first define *D*i as the number of days in season *i* for *i* = 1, 2, 3, 4, where *i*=1 is spring, *i*=2 is summer *i*=3 is autumn and *i*=4 is winter. Similarily *A*i is the event that *A* occurs in season *i*. Then

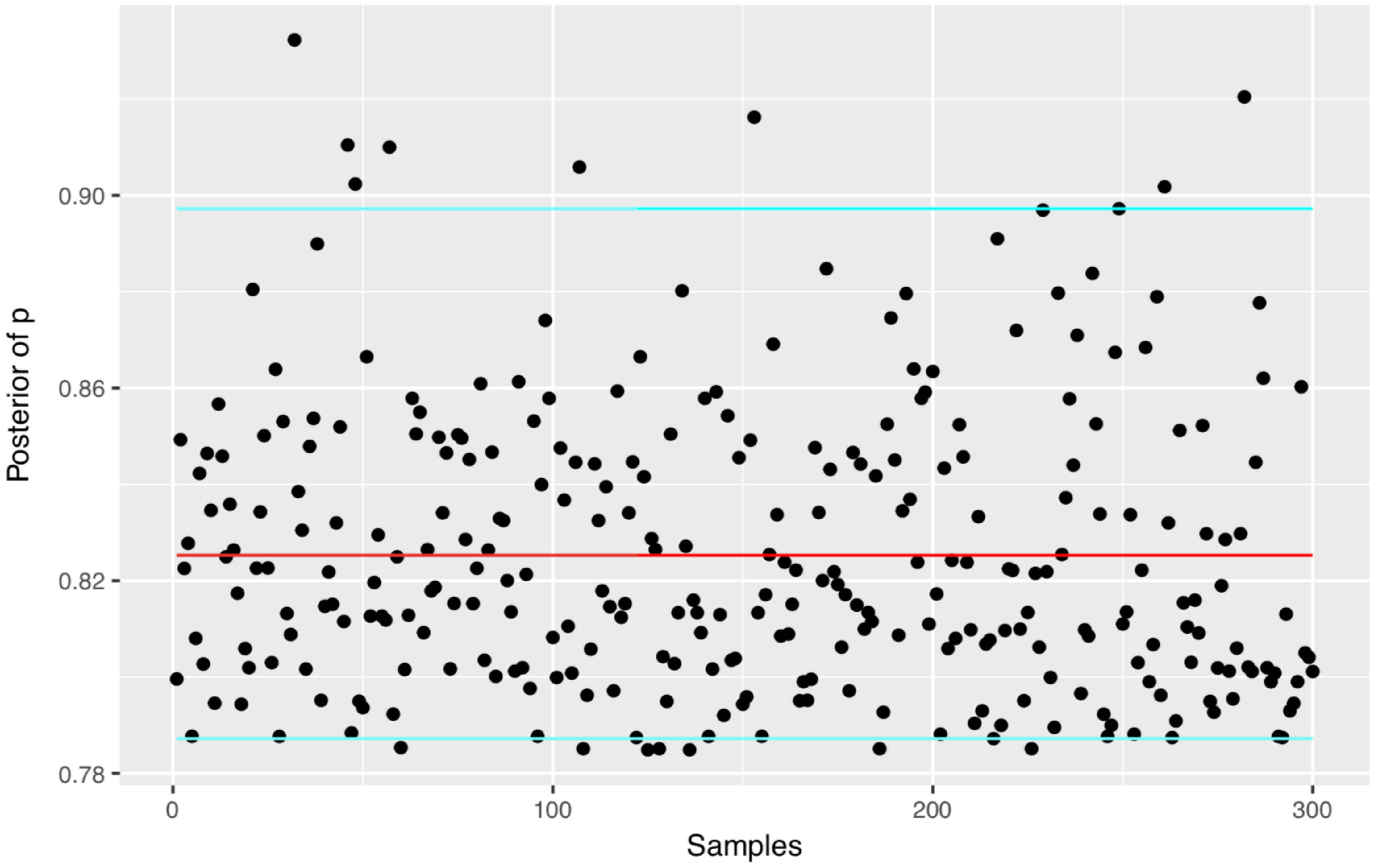
$$P(A) = 1 - \prod_{i=1}^{4} P(A_i^c) = 1 - \prod_{i=1}^{4} \frac{D_i!}{(D_i - N_i)!D_i^{N_i}}$$

The following function estimates the posterior mean of *p*. Note that the function is implemented on a logarithmic scale to aviod computational problems.

```
estimate_p = function(Nvec) {
days = c(91,92,92,90)
Dummyvec = seq(1,92,1)
logvec = log(Dummyvec)
logPc = 0for (i in 1:4) {
logPc = sum(logvec[1:days[i]]) - sum(logvec[1:(days[i]-Nvec[i])]) - Nvec[i]*log(days[i]) +
logPc
}
Pc = exp(logPc)
return (1-Pc)
}
```

Combining with the following function, we get an estimate for the posterior mean of p, in addition to a 95% confidence interval.

```
confidenceinterval = function(n,N,alphavec){
  #Assumes that alphavec contains correct parameters for the posterier dirichlet
pcut = round(n*0.025) p = rep(0,n)
x = seq(1,n,1)
w = rep(0,n)for (i in 1:n) {
Nvec = generatePeopleSeason(N,alphavec)$Nvec p[i] = estimate_p(Nvec)
}
pplot = p
p = sort(p)
lower = p[pcut]
upper = p[n-pcut]
meanP = mean(p)
cInf = c(lower,upper)
return (list(pplot=pplot,meanP=meanP,cInf = cInf))
}
```



The figure shows 300 samples versus the posterior of *p*. The red line indicates the mean value and the lower and upper blue lines shows the 95% confidence interval.

The posterior mean is
0.8252731
and the 95% confidence interval is
0.7872153 0.8972443
Changing the number of samples to n = 10000. The posterior mean of p becomes
0.8198681
and the 95% confidence interval is
0.7874379 0.8801871
The posterior mean of *p* is larger then the probability, which fits well with our intuition. This is very reasonable as we know have prior knowledge of which season the people are born in.