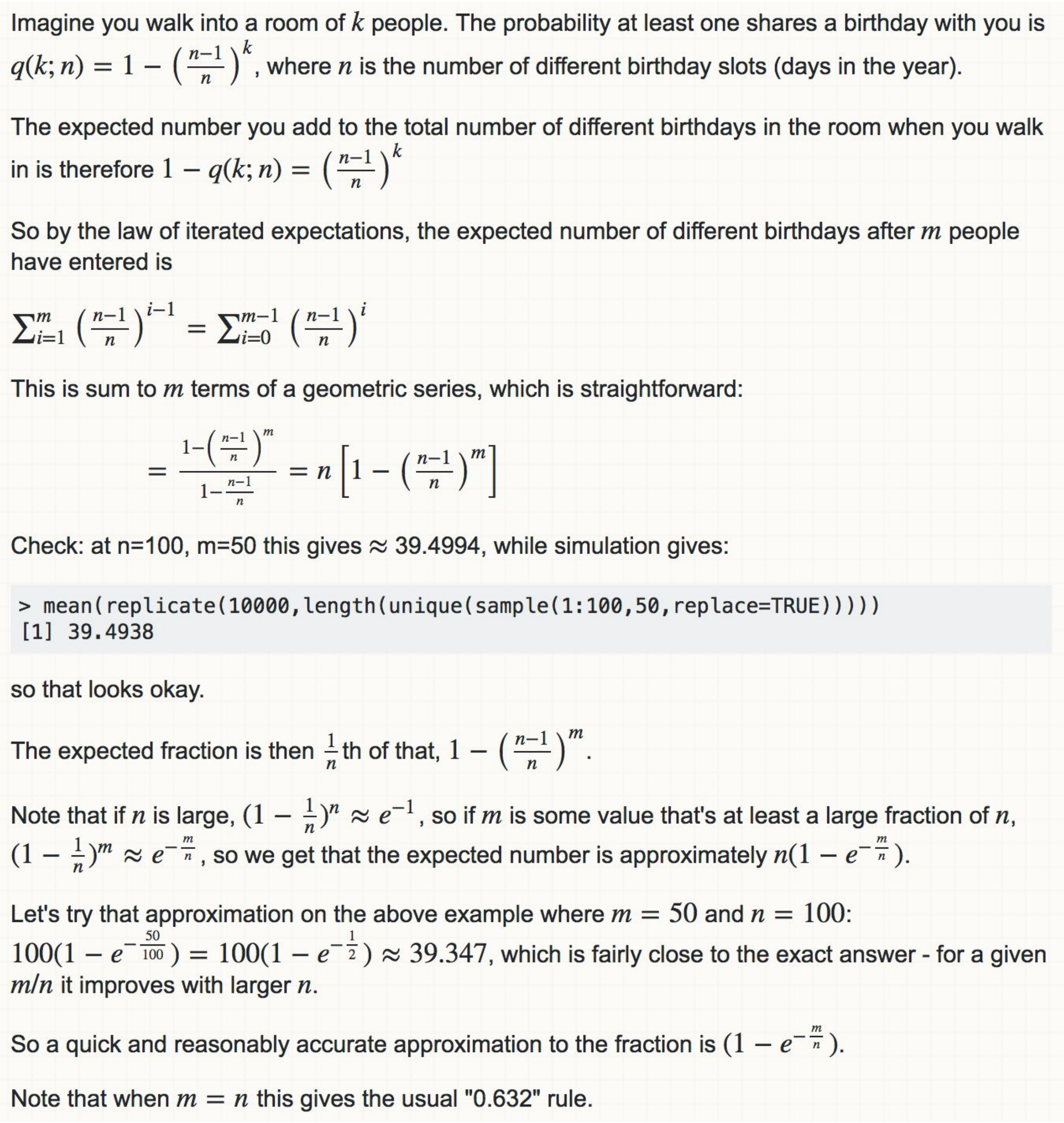
Week 4 Homework Solutions

1. No standard answers. Results may vary based on model tuning.
2. When number of times to draw sample (m) equals number of records in original data (n), this fraction is 0.632. Below is the derivation:



Imagine you walk into a room of k people. The probability at least one shares a birthday with you is $q(k;n) = 1 - \left(\frac{n-1}{n}\right)^k$, where n is the number of different birthday slots (days in the year).

The expected number you add to the total number of different birthdays in the room when you walk in is therefore $1 - q(k;n) = \left(\frac{n-1}{n}\right)^k$

So by the law of iterated expectations, the expected number of different birthdays after m people have entered is

$$\sum_{i=1}^{m} \left(\frac{n-1}{n}\right)^{i-1} = \sum_{i=0}^{m-1} \left(\frac{n-1}{n}\right)^{i}$$

This is sum to m terms of a geometric series, which is straightforward:

$$= \frac{1-\left(\frac{n-1}{n}\right)^m}{1-\frac{n-1}{n}} = n\left[1 - \left(\frac{n-1}{n}\right)^m\right]$$

Check: at n=100, m=50 this gives ≈ 39.4994, while simulation gives:

```
> mean(replicate(10000,length(unique(sample(1:100,50,replace=TRUE)))))
[1] 39.4938
```

so that looks okay.

The expected fraction is then $\frac{1}{n}$th of that, $1 - \left(\frac{n-1}{n}\right)^m$.

Note that if n is large, $(1-\frac{1}{n})^n \approx e^{-1}$, so if m is some value that's at least a large fraction of n, $(1-\frac{1}{n})^m \approx e^{-\frac{m}{n}}$, so we get that the expected number is approximately $n(1-e^{-\frac{m}{n}})$.

Let's try that approximation on the above example where $m = 50$ and $n = 100$: $100(1-e^{-\frac{50}{100}}) = 100(1-e^{-\frac{1}{2}}) \approx 39.347$, which is fairly close to the exact answer - for a given m/n it improves with larger n.

So a quick and reasonably accurate approximation to the fraction is $(1-e^{-\frac{m}{n}})$.

Note that when $m = n$ this gives the usual "0.632" rule.