Applied Statistics – Exercise 3

Goal

To get confident with conditional probability and discrete random variables. To make first visualizations of discrete distributions in R.

Problems

T=Theoretical Exercise, R=R Exercise

1. (T)

We are at a train station, waiting for a train. Suppose that the probability of snow is 0.1. If it is snowing, then the probability that the train is delayed is 0.6, otherwise, it is 0.2. Given that the train is delayed, what is the probability that it is snowing? Define appropriate events, and compute the conditional probability.

```
The text describes two events that we define as follows:
S : It snows
D : The train is delayed

From the text, we are given the following probabilities related to those events:
P(S) = 0.1 (The probability that it snows)
P(D|S) = 0.7 (The conditional probability that the train is delayed if it snows)
P(D|$S^C$) = 0.2 (The conditional probability that the train is delayed if it is now snowing)

We are now searching for another conditional probability, namely the probability that it snows given tha
```

$$P(S|D) = \frac{P(S \cap D)}{P(D)} = \frac{P(D|S) \cdot P(S)}{P(D|S) \cdot P(S) + P(D|S^C) \cdot P(S^C)} = \frac{0.7 \cdot 0.1}{0.7 \cdot 0.1 + 0.2 \cdot (1 - 0.1)} = \frac{0.07}{0.25} = 0.28 = 28\%$$

```
We have therefore found the probability that it snows given that the train is delayed to be 28%
```

2. (T)

Consider the following game. A coin is tossed repeatedly until we get heads. For a single toss the probability of heads is p, and tosses are independent. You are to guess if the number of tosses needed to get the first head is even or odd: if your guess is right, you win. Should you pick “even” or “odd” as your guess?

Hint: You can use of the following in your solution. If $0 \leq a < 1$, then

$$\sum_{k \geq 0} a^k = \frac{1}{1-a}$$

```
The descibed game is an example of a geometric distribution, in the discrete random variable X: Tosses r
The PMF (the function that assigns a probability to each value X can take, in this case the a countably
```

$$PMF : P(X = k) = p \cdot (p - 1)^{k-1}$$

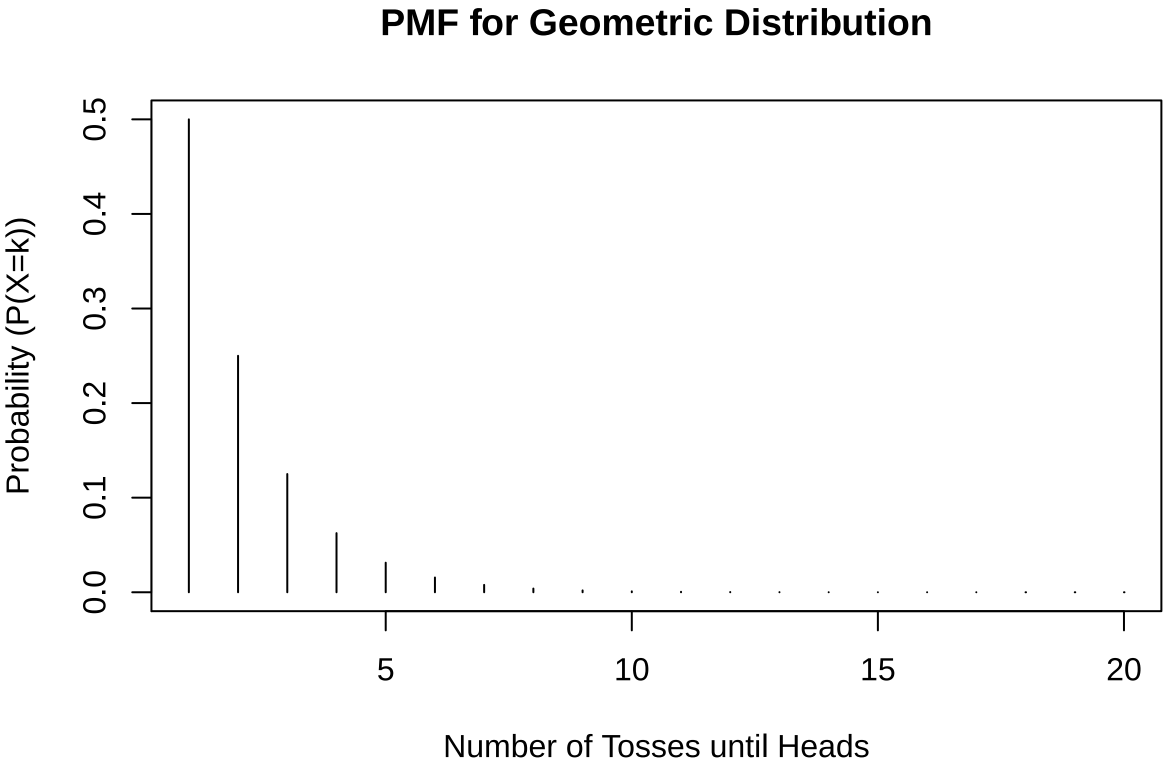
In example to get heads on the first throws is P(X=1)=1/2 or on the second throw P(X=2)=1/4

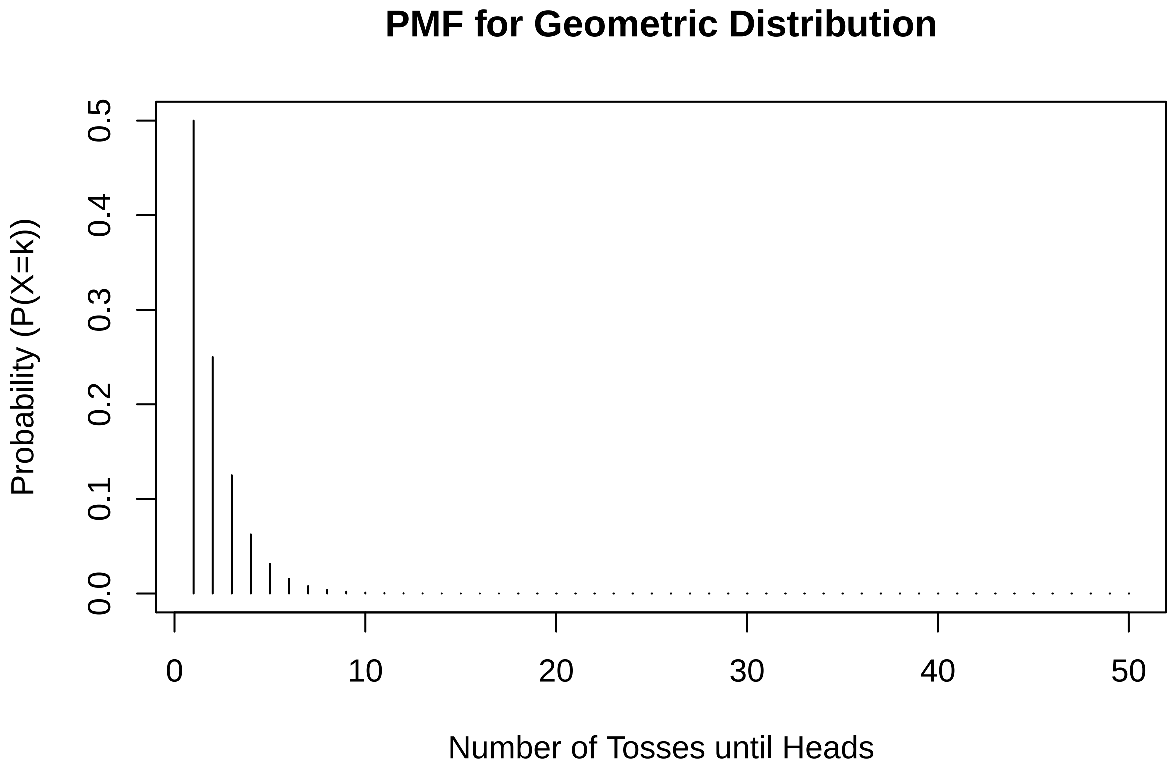
```
compute_geometric <- function(k, prob) {
  return(dgeom(k-1, prob=prob))
}
for (i in 1:5){
  print(compute_geometric(i,prob=.5))
}
```

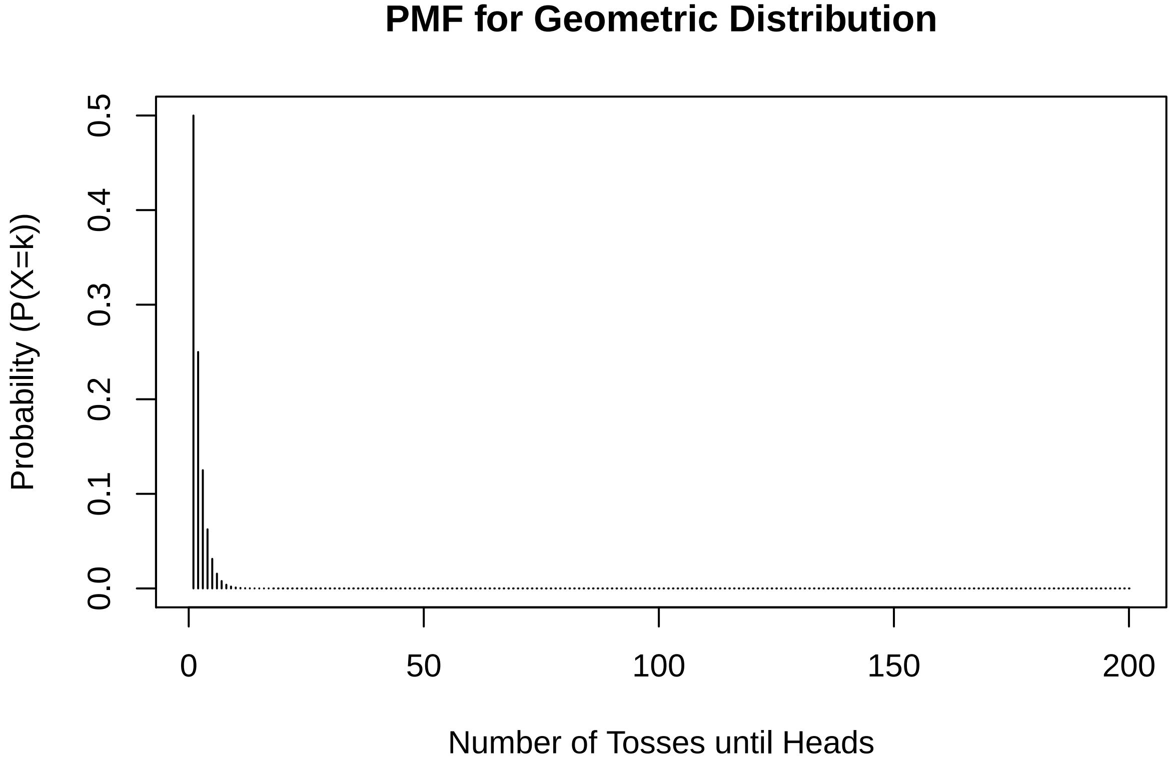
```
## [1] 0.5
## [1] 0.25
## [1] 0.125
## [1] 0.0625
## [1] 0.03125
```

We can also plot the probability mass function on an arbitary range (since the range of the random variable is countably infinite, we could make the range of the x-axis infinitely large, however this wouldn't make sense, because the graph quickly converges to 0 for large x)

```
plot_pmf_geometric <- function(range, prob) {
  plot(range, dgeom(range-1, prob=prob),
       main='PMF for Geometric Distribution',
       type='h',
       xlab='Number of Tosses until Heads',
       ylab='Probability (P(X=k))')
}
for (i in c(20, 50, 200)) {
  plot_pmf_geometric(1:i, .5)
}
```







The question is now, which of the following series is more likely:
{P(k): k is odd} or {P(k): k is even}
The series whose sum of all elements is higher, is more likely to occur and should thus be chosen in th

Odd: $P(X=1)+P(X=3)+P(X=5)+... = \frac{1}{2}+\frac{1}{8}+\frac{1}{32}+...$Even: $P(X=2)+P(X=4)+P(X=6)+... = \frac{1}{4}+\frac{1}{16}+\frac{1}{64}+...$

This reminds us of the geometric sum:

$$\frac{1}{2}+\frac{1}{4}+\frac{1}{8}... = 1$$

When subtracting all probabilites we get from our even series, we get a total probability of 2/3, while we get a total probability of 1/3 for our even series (when subtracting the probabilities from our odd series). We can therefore conclude that it is better to pick "odd".

3. (T)

A fair die is thrown until the sum of the results of the throws exceeds 6. Let the random variable X be the number of throws needed for this. Find the probability mass function of X.

The exercise descibes a random experiment of continuous dice rolls with the discrete random variable X:

$$p(X=2) = \frac{22}{36}p(X=3) = \frac{70}{216}p(X=4) = p(X=5) = p(X=6) = 6\frac{1}{6}^6 = \frac{1}{776}p(X=7) = \frac{1}{6}^7 = \frac{1}{279936}$$

For all other a, we get P(X=a) = 0

4. (R)

Consider you have two fair coins that you toss simultaneously (fair coins have a 0.5 probability of heads). You repeat the trial 15 times. Let X be the random variable indicating the number of cases, where both coins come with heads up. In the following exercises you can use the `dbinom` and `pbinom` functions.

a) What is the probability $P(X=5)$?

b) What is the probability $P(X \leq 5)$?

```
# probability of getting exactly five times {H,H} in 15 double coin tosses
dbinom(5, size=15, prob=.25)
```

```
## [1] 0.165146
```

```
# probability of getting at most five times (H,H) in 15 double coin tosses
pbinom(5, size=15, prob=.25)
```

```
## [1] 0.8516319
```

```
sum(dbinom(0:5, size=15, prob=.25)) # converting PMF to CDF
```

```
## [1] 0.8516319
```

5. (R)

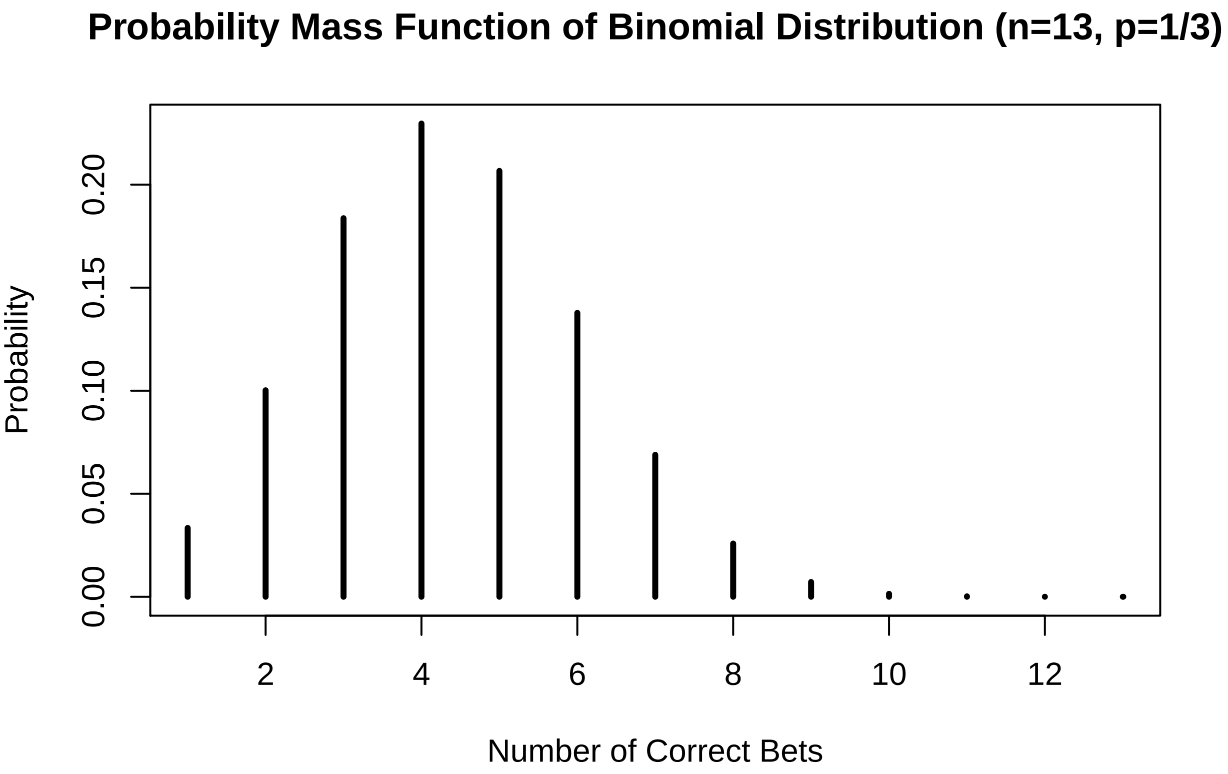
Imagine a football betting setting where there are 13 football games. Each game can have three possible outcomes: home team wins (1), the teams play even (E) or the visitor team wins (2). Model the outcome of each game as a random process where each of the three outcomes are equally probable and independent from other games. Let the random variable X characterise the number of correct guesses for the 13 outcomes in one betting.

a) Write down the analytic forms for the probability mass function of X.

b) Illustrate the probability mass function by plotting it in a figure.

c) What is the probability that one gets all the 13 outcomes right?

```
range <- 1:13
pmf <- dbinom(range, size=13, prob=1/3) # a vector holding the probabilities of the number of correct g
plot(range, pmf, main='Probability Mass Function of Binomial Distribution (n=13, p=1/3)', xlab='Number
```



```
# probability of getting all bets correct is P(X=13)
print(dbinom(13, size=13, prob=1/3)) # or from variable pmf[13]
```

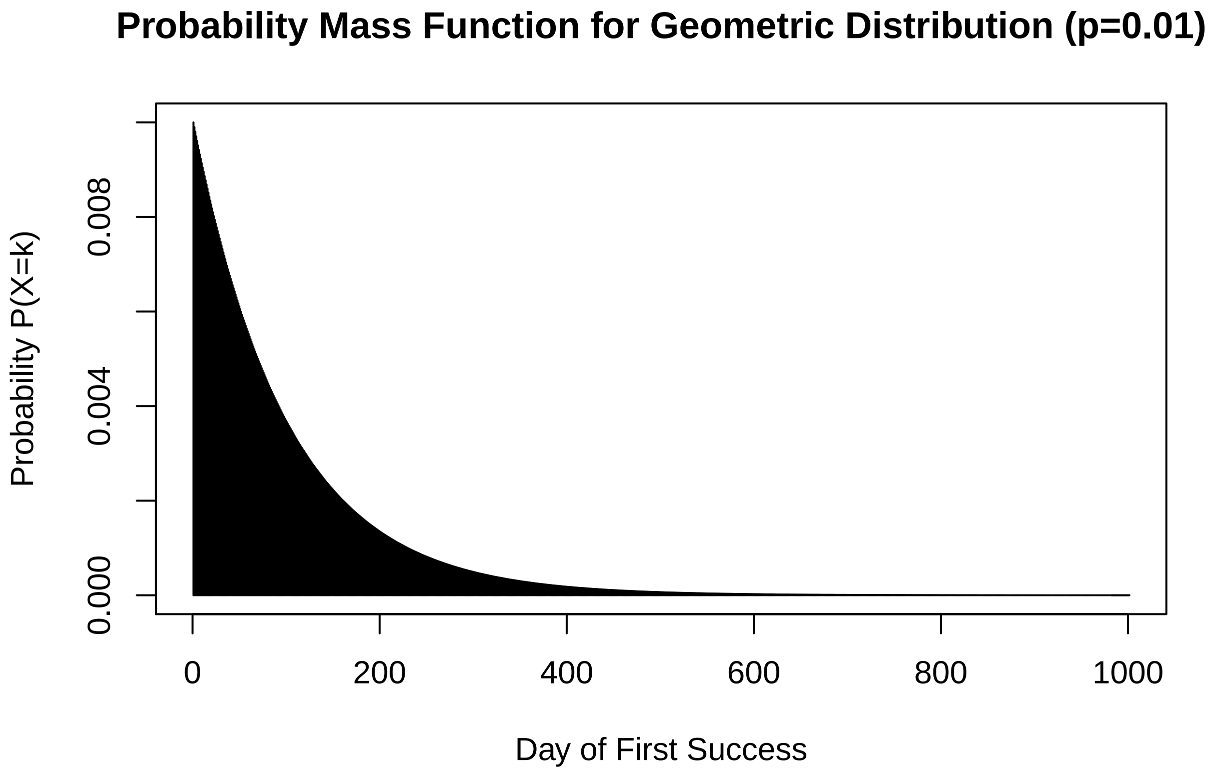
```
## [1] 6.272255e-07
```

6. (R)

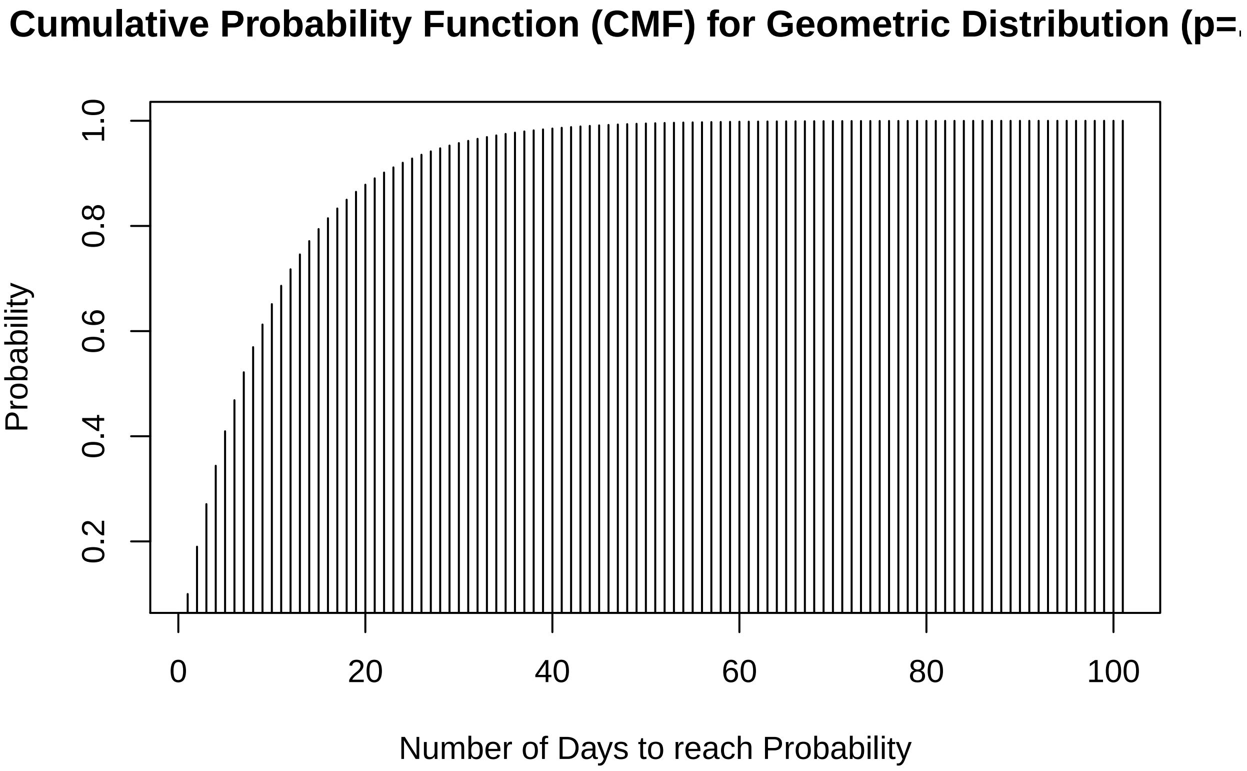
You are a collector of soccer players cards. There is just one card missing from your collection. Every day you buy one, and with the probability 1/100 it is the one you are missing. Each purchase is independent from the others. Model the number of days it takes to find the missing card by the random variable $X \sim Geo(1/100)$.

a) Plot the distribution function of X.

```
plot(1:1001, dgeom(0:1000, prob=.01), # we need to adjust the ranges because r interprets the k in geom
     main='Probability Mass Function for Geometric Distribution (p=0.01)',
     xlab="Day of First Success",
     ylab='Probability P(X=k)',
     type='h')
```



```
plot(1:101, pgeom(0:100, prob=.1),
     main = 'Cumulative Probability Function (CMF) for Geometric Distribution (p=.01)',
     xlab = 'Number of Days to reach Probability',
     ylab = 'Probability',
     type='h')
```



b) How many cards do you have to buy so that the chance of finding the missing card is at least 0.5? How about at least 0.95? (play with different ranges for k).

```
number.of.cards <- function(k) {
  return (sum(pgeom(0:100, prob=.1) <= k))
}

print(number.of.cards(.5))
```

```
## [1] 6
```

```
print(number.of.cards(.95))
```

```
## [1] 28
```

c) Assume you have tried for 20 days, but you have not won yet. For how many days do you need to try further so that you have at least a 0.5 chance of winning?

Since the events, in this case finding the missing player in one of the packs, are independent from one another, it doesn't matter, for how many days in a row we have not drawn the card. It does not increase our chances of getting him from now on. Therefore the probability of getting the card with a change of at least 50% is equal to our computation for b), so we would have to draw cards for at least 6 more days.