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.

Solution:

Given the following events:

$$S = \text{“Will snow”};\ D = \text{“Train will be delayed”}$$

We are also given the following probabilities:

$$P(S) = 0.1;\ P(D|S) = 0.6;\ P(D \mid S^c) = 0.2$$

We need to find a $P(S|D)$ according to Bayes' theorem:

$$P(S|D) = \frac{P(D|S) \cdot P(S)}{P(D)}$$

We already know $P(D|S)$ and $P(S)$, so the only unknown is $P(D)$.

$$P(D) = P(D \cap S) + P(D \cap S^c) = P(D|S) \cdot P(S) + P(D \mid S^c) \cdot P(S^c) = 0.6 \cdot 0.1 + 0.2 \cdot 0.9 = 0.24$$

Then, let's plug everything to the Bayes' theorem:

$$P(S|D) = \frac{P(D|S) \cdot P(S)}{P(D)} = \frac{0.6 \cdot 0.1}{0.24} = 0.25$$

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

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

Solution

Let's first define a random variable X:

$$X = \text{“Number of tosses until we get the heads first ”}$$

The probability mass function of this event is:

$$P_X(k) = p \cdot (1-p)^{k-1}$$

After that let's define a Bernoulli random variable Y:

$$Y = \text{“0 if number of tosses until first heads is even and 1 if uneven”}$$

The probability mass function in this case:

$$P_Y(0) = \sum_{k \geq 0} P_X(2 \cdot k + 2) = \sum_{k \geq 0} p \cdot (1-p)^{2 \cdot k + 1}$$

$$P_Y(1) = \sum_{k \geq 0} P_X(2 \cdot k + 1) = \sum_{k \geq 0} p \cdot (1-p)^{2 \cdot k}$$

To reveal whether we need to pick “even” or “odd” as the guess, we need to calculate either $P_Y(0)$ or $P_Y(1)$, as both of them are complementary. Here I calculate the $P_Y(1)$ as it is better corresponds with a geven hint:

$$P_Y(1) = \sum_{k \geq 0} p \cdot (1-p)^{2 \cdot k} = p \cdot \sum_{k \geq 0} ((1-p)^2)^k \overset{\text{Using hint (if p > 0)}}{=} p \cdot \left(\frac{1}{1-(1-p)^2}\right) = \frac{1}{2-p}$$

It is seen that $P_Y(1)$ is always greater or equal than 0.5. The only special case is when p takes values in the interval $[0, 1]$:

1) If $p > 0$, we use the hint and the values we get from $P_Y(1)$ are strictly increasing in the range $(\frac{1}{2}, 1]$. It is always better to bet for an odd number.
2) If $p = 0$, we will never get heads, so it does not matter.

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.

Solution

Firstly, let's define a random variable X:

$$X = \text{“Number of tosses until the sum of the results of the throws exceeds 6”}$$

Since this is a fair dice, the random variable X lies in the interval from 2 to 7. We cannot get number bigger than 6 in one toss, as well as if we get ones over and over, after 7 throws we will eventually get 7 in sum. So,

$$X \in [2; 7]$$

For a given number of throws n,

$$F(n) = P(X \leq n)$$

. So, we should find the probability of an event A in the experiment of throwing the die n times. Considering that,

$$F(n) = P(A) = \frac{\text{number of outcomes in the event A}}{\text{total number of outcomes in the sample space}}$$

.

Therefore, the probability mass function in this case is:

$$P(X = 2) = \frac{22}{36}$$
$$P(X = 3) = \frac{70}{216}$$
$$P(X = 4) = \frac{21}{1296}$$
$$P(X = 5) = \frac{15}{7776}$$
$$P(X = 6) = 6\frac{1}{6}^6 = \frac{1}{776}$$
$$P(X = 7) = \frac{1}{6}^7 = \frac{1}{279936}$$

For all other A the

$$P(X = A) = 0$$

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 $P(X = 5)$?

Solution

X is a binomial distribution with parameters $n = 15$, number of trials, and the probability to get two heads while tossing two fair coins, or the probability of success of a single trial is $p = \frac{1}{4}$. We can use the formula for the probability mass function here, which is:

$$P(X = 5) = \binom{15}{5} \cdot p^5 \cdot (1-p)^{15-5} \approx 0.165$$

In R I calculated this using simply *dbinom*:

```
cat(dbinom(5, 15, 1/4)) # cat to display the result
```

```
## 0.165146
```

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

Solution

Here we can use another way to find the probability - the cumulative distribution function of a binomial in R:

```
cat(pbinom(5, 15, 1/4))
```

```
## 0.8516319
```

However, we can also simply use the probability mass function, as we did before:

$$P(X \leq 5) = \sum_{k=0}^{5} P(X = k) \approx 0.857$$

```r
cat(sum(dbinom(0:5, 15, 1/4)))
```

```
## 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.

Solution

X is a number of correct quesses with parameters $n = 13$ as a number of outcomes, and the equal outcomes probabilities $p = \frac{1}{3}$.

The probability mass finction in this case is:

$$P(X = n) = \binom{13}{n} \cdot p^n \cdot (1-p)^{13-n}$$

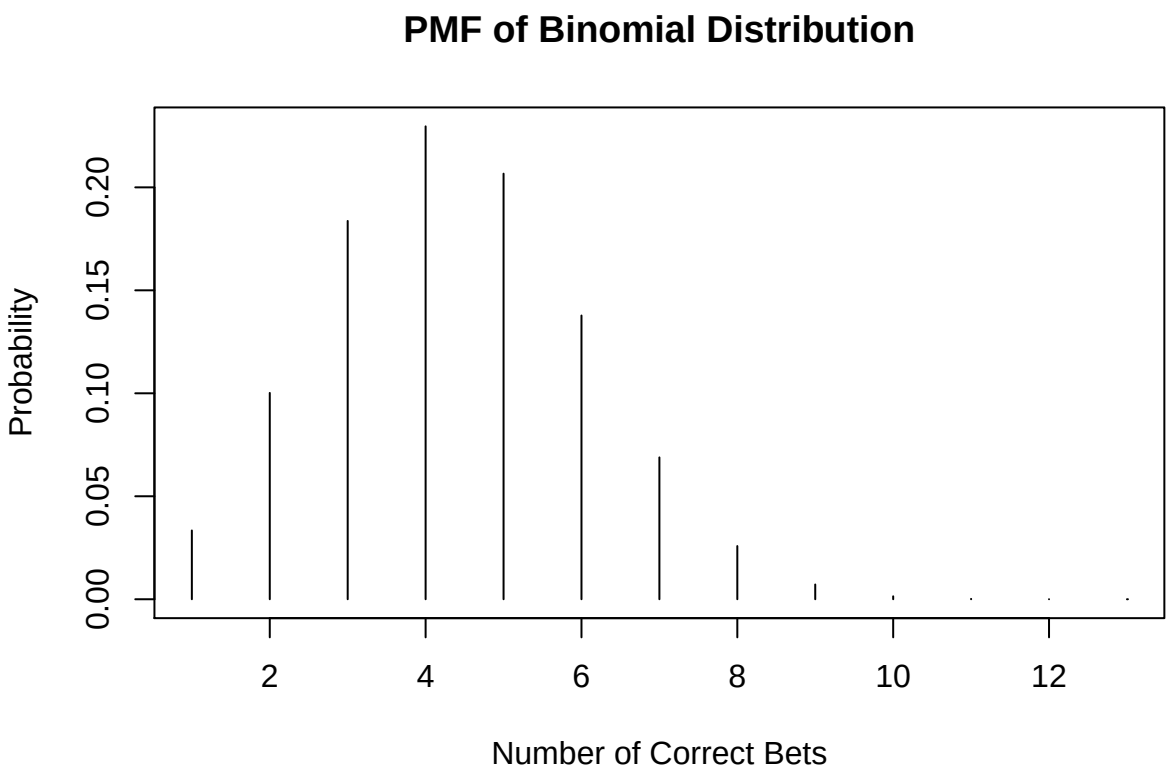
b) Illustrate the probability mass function by plotting it in a figure.

Solution

I illustrate the probability mass function as a histogram in this case, simply using *dbinom*.

```r
p <- 1/3
n <- 13

plot(1:n, dbinom(1:n, n, p), main = 'PMF of Binomial Distribution',
     xlab = 'Number of Correct Bets', ylab = 'Probability', type = 'h')
```



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

Solution

To find the answer we can just fit *dbinom* with 13 successes.

```
cat(dbinom(13, n, p))
```

```
## 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.

Solution

While we know the probability mass function as:

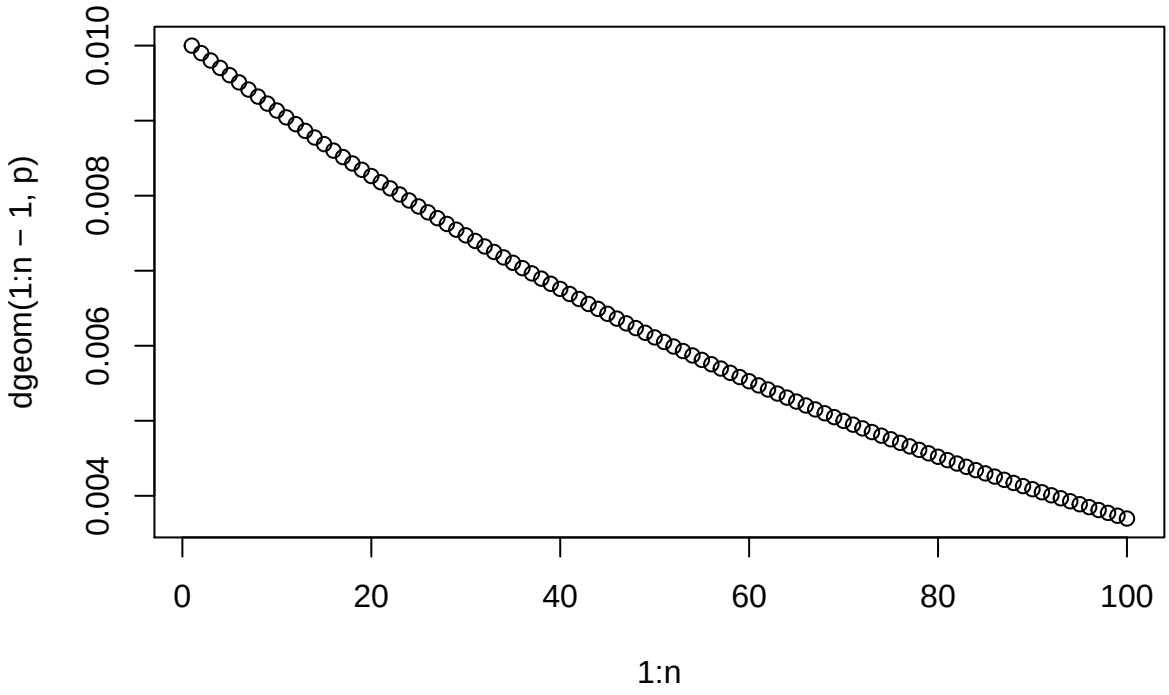
$$P(x) = p \cdot (1-p)^{x-1}$$

In R it is calculated in a different way:

$$P(x) = p \cdot (1-p)^{x}$$

Instead of the total number of *trials* in which the last is a success, we can perceive it as the number of *failures* before getting a success. The number of failures here is derived by subtracting 1 to the number of trials.

```r
p <- 1/100
n <- 100
plot(1:n, dgeom(1:n - 1, p))
```



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).

Solution

Instead of playing with different values, I just get that numbers directly with *qgeom* function. It is kind of the approximate inverse of the cumulative distribution function: With a given a probability p it reurns the smallest value such that *pgeom* of that value is greater than p.

```r
# for 0.5
p <- 1/100
cat("For 0.5: Number of failures:", qgeom(0.5, p), "; Number of trials:", qgeom(0.5, p) + 1)
```

```
## For 0.5: Number of failures: 68 ; Number of trials: 69
```

```r
cat(sprintf("Probability of success after %d trials: %f", 69, pgeom(68, p)))
```

```
## Probability of success after 69 trials: 0.500163
```

```r
# for 0.95
p <- 1/100
cat("For 0.95: Number of failures:", qgeom(0.95, p), "; Number of trials:", qgeom(0.95, p) + 1)
```

```
## For 0.95: Number of failures: 298 ; Number of trials: 299
```

```r
cat(sprintf("Probability of success after %d trials: %f", 299, pgeom(298, p)))
```

```
## Probability of success after 299 trials: 0.950464
```

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?

Solution

The geometric distribution has the memoryless property. Hence:

$$P(X \geq k + x \mid X > x) = P(X \geq k)$$

The probability $P(X \geq x + 20 \mid X > 20)$ is the same as $P(X \geq x)$. Following that, we can use the same procedure as in section *b)*. So, we still need 69 trials in order to get a 0.5 chance of winning.