# Probability distributions EBP038A05: 2020-2021 Assignments

Nicky D. van Foreest and student assistents

February 16, 2021

## CONTENTS

## GENERAL INFORMATION

Here we just provide the exercises of the assignments. For information with respect to grading we refer to the course manual.

Each assignment contains several sections. The first section is meant to help you read the book well and become familiar with definitions and concepts of probability theory. These questions are mostly simple checks, not at exam level, but lower. The second section contains some exercises at about the exam level to get you started. Here you have to derive and explain a solution, in mathematical notation. Most of the selected exercises of the book are also at about (or just a bit above) exam level. The third section is about coding skills. We explain the rationale presently. The final section with challenges is for those students that like a challenge; the problems are above exam level.

You have to get used to programming and checking your work with computers, for instance by using simulation. The coding exercises address this skill. You should know that much of programming is ‘monkey see, monkey do’. This means that you take code of others, try to understand it, and then adapt it to your needs. For this reason we include the code to answer the question. The idea is that you copy the code, you run it and include the numerical results in your report. You should be able to explain how the code works. For this reason we include questions in which you have explain how the most salient parts of the code works.

We include python and R code, and leave the choice to you what to use. In the exam we will also include both languages in the same problem, so you can stay with the language you like. You should know, however, that many of you will need to learn multiple languages later in life. For instance, when you have to access databases to obtain data about customers, patients, clients, suppliers, inventory, demand, lifetimes (whatever), you often have to use `sql`. Once you have the raw data, you process it with `R` or python to do statistics or make plots. (While I (= NvF) worked at a bank, I used Fortran for numerical work, AWK for string parsing and making tables, excel, SAS to access the database, and matlab for other numerical work, all next to each other. I got tired of this, so I went to using python as it did all of this stuff, but then within one language.) For your interest, based on the statistics here or here, python scores (much) higher than R in popularity; if you opt for a business career, the probability you have to use python is simply higher than to have to use R.

You should become familiar with look up documentation on coding on the web, no matter your programming language of choice. Invest time in understanding the, at times, rather technical and terse, explanations. Once you are used to it, the core documentation is faster to read, i.e., less clutter. In the long run, it pays off.

The rules:

1. For each assigment you have to turn in a pdf document typeset in LaTeX. Include a title, group number, student names and ids, and date.

2. We expect brief answers, just a sentence or so, or a number plus some short explanation. The idea of the assignment is to help you studying, not to turn you in a writer.

3. When you have to turn in a graph, provide decent labels and a legend, ensure the axes have labels too.

# 1 ASSIGNMENT 1

## 1.1 *Have you read well?*

**Ex 1.1.** In your own words, explain what is

1. a joint PMF, PDF, CDF;

2. a conditional PMF, PDF, CDF;

3. a marginal PMF, PDF, CDF.

**Ex 1.2.** We have two r.v.s $X, Y \in [0,1]^2$ (here $[0,1]^2 = [0,1] \times [0,1]$) with the joint PDF $f_{X,Y}(x,y) = 2 I_{x \leq y}$.

1. Are $X$ and $Y$ independent?

2. Compute $F_{X,Y}(x,y)$.

**Ex 1.3.** Correct (that is, is the following claim correct?)? We have two continuous r.v.s $X,Y$. Even though the joint CDF factors into the product of the marginals, i.e., $F_{X,Y}(x,y) = F_X(x)F_Y(y)$, it is still possible in general that the joint PDF does not factor into a product of marginals PDFs of $X$ and $Y$, i.e., $f_{X,Y}(x,y) \neq f_X(x)f_Y(y)$.

**Ex 1.4.** Consider $F_{X,Y}(x,y)/F_X(x)$. Write this expression as a conditional probability. Is this equal to the conditional CDF of $X$ and $Y$?

**Ex 1.5.** Let $X$ be uniformly distributed on the set $\{0,1,2\}$ and let $Y \sim \text{Bern}(1/4)$; $X$ and $Y$ are independent.

1. Present a contingency table for the $X$ and $Y$.
2. What is the interpretation of the column sums the table?
3. What is the interpretation of the row sums of the table?
4. Suppose you change some of the entries in the table. Are $X$ and $Y$ still independent?

**Ex 1.6.** Apply the chicken-egg story. A machine makes items on a day. Some items, independent of the other items, are failed (i.e., do not meet the quality requirements). What is $N$, what is $p$, what are the 'eggs' in this context, and what is the meaning of 'hatching'? What type of 'hatching' do we have here?

**Ex 1.7.** Correct? We have two r.v.s $X$ and $Y$ on $\mathbb{R}^+$. It is given that $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for $x,y \leq 1/3$. Then $X$ and $Y$ are necessarily independent.

**Ex 1.8.** I select a random guy from the street, his height $X \sim \text{Norm}(1.8,0.1)$, and I select a random woman from the street, her height is $Y \sim \text{Norm}(1.7,0.08)$. I claim that since I selected the man and the woman independently, their heights are independent. Briefly comment on this claim.

**Ex 1.9.** Correct? For any two r.v.s $X$ and $Y$ on $\mathbb{R}^+$ with marginals $F_X$ and $F_Y$, it holds that $\mathsf{P}\{X \leq x, Y \leq y\} = F_X(x)F_Y(y)$.

**Ex 1.10.** Theorem 7.1.11. What is the meaning of the notation $X|N = n$?

**Ex 1.11.** Correct? $X,Y$ are two discrete r.v.s with CDF $F_{X,Y}$. We can compute the PDF as $\partial_x\partial_y F_{X,Y}(x,y)$.

### 1.2 *Exercise at about exam level*

**Ex 1.12.** This is about the simplest model for an insurance company that I can think of. We start with an initial capital $I_0 = 2$. The company receives claims and contributions every period, a week say. In the $i$th period, we receive a contribution $X_i$ uniform on the set $\{1,2,\ldots,10\}$ and a claim $C_i$ uniform on $\{0,1,\ldots 8\}$.

1. What is the meaning of $I_1 = I_0 + X_1 - C_1$?

2. What is the meaning of $I_2 = I_1 + X_2 - C_2$?
3. What is the interpretation of $I_1' = \max\{I_0 - C_1, 0\} + X_1$?
4. What is the interpretation of $I_2' = \max\{I_1' - C_2, 0\} + X_2$?
5. What is the interpretation of $\bar{I}_n = \min\{I_i : 0 \leq i \leq n\}$?
6. What is $\mathsf{P}\{I_1 < 0\}$?
7. What is $\mathsf{P}\{I_1' < 0\}$?
8. What is $\mathsf{P}\{I_2 < 0\}$?
9. What is $\mathsf{P}\{I_2' < 0\}$?
10. Provide an interpretation in terms of the inventory of rice, say, at a supermarket for $I_1$ and $I_1'$.

### 1.3 *Coding skills*

**Ex 1.13.** Use simulation to estimate the answer of BH.7.1. Run the code below and explain line 9 of python code or line 7 of the R code.

Then run the code for a larger sample, e.g, `num=1000` or so, but remove the prints of `a`, `b`, and `succes`, because that will fill your screen with numbers you don't need. Only for small simulations such output is handy so that you can check the code.

Compare the value of the simulation to the exact value.

```
1  import numpy as np
2
3  np.random.seed(3)
4
5  num = 10
6
7  a = np.random.uniform(size=num)
8  b = np.random.uniform(size=num)
9  success = np.abs(a - b) < 0.25
10 print(a)
11 print(b)
12 print(success)
13 print(success.mean(), success.var())
```

```
1  set.seed(3)
2
```

```
num <- 10

a <- runif(num)
b <- runif(num)
success <- abs(a-b) < 0.25
a
b
success
paste(mean(success), var(success))
```

Challenge (not obligatory): If you like, you can include a plot of the region (in time) in which Alice and Bob meet, and put marks on the points of the simulation that were 'successful'.

**Ex 1.14.** Let $X \sim \text{Exp}(3)$. Find a simple expression for $\mathsf{P}\{1 < X \leq 4\}$ and compute the value. Then use simulation to check this value. Finally, use numerical integration to compute this value. What are the numbers? Explain lines 11, 21 and 26 of the python code or lines 7, 17 and 23 of the R code.

```
1  import numpy as np
2  from scipy.stats import expon
3  from scipy.integrate import quad
4
5  labda = 3
6
7  X = expon(scale = 1 / labda).rvs(1000)
8  # print(X)
9  print(X.mean())
10
11 success = (X > 1) * (X < 4)
12 # print(success)
13 print(success.mean(), success.std())
14
15
16 def F(x):  # CDF
17     return 1 - np.exp(-labda * x)
18
19
20 def f(x):  # density
21     return labda * np.exp(-labda * x)
22
23
24 print(F(4) - F(1))
```

```

I = quad(f, 1, 4)
print(I)
```

```
labda <- 3

X <- rexp(1000, rate = labda)
# X
mean(X)

success <- (X > 1) * (X < 4)
# print(success)
paste(mean(success), sd(success))


CDF <- function(x) {  # CDF
  return(1 - exp(-labda * x))
}

f <- function(x) {    # density
  return(labda * exp(-labda * x))
}


CDF(4) - CDF(1)

I = integrate(f, 1, 4)
I
```

### 1.4 *Challenges, optional*

You are free to chose one of these problems, but of course you can do both if you like.

A UNIQUENESS PROPERTY OF THE POISSON DISTRIBUTION Consider again the chicken-egg story (BH 7.1.9): A chicken lays a random number of eggs $N$ an each egg independently hatches with probability $p$ and fails to hatch with probability $q = 1-p$. Formally, $X|N \sim \text{Bin}(N,p)$. Assume also that $X|N \sim \text{Bin}(N,p)$ and that $N-X$ is independent of $X$. For $N \sim \text{Pois}(\lambda)$ it is shown in BH 7.1.9 that $X$ and $Y$ are independent. This exercise asks for the converse: showing that the independence of $X$ and $Y$ implies that $N \sim \text{Pois}(\lambda)$ for some $\lambda$. Hence, the Poisson distribution is quite special: it is the only distribution for which the number of hatched eggs doesn't tell you anything about the number of unhatched eggs.

Let $0 < p < 1$. Let $N$ be an r.v. taking non-negative integer values with $P(N > 0) > 0$. Assume also that $X|N \sim \text{Bin}(N,p)$ and that $N - X$ is independent of $X$.

**Ex 1.15.** Use the assumption that $\mathsf{P}\{N > 0\} > 0$ to prove that $N$ has support $\mathbb{N}$, i.e. $\mathsf{P}\{N = n\} > 0$ for all $n \in \mathbb{N}$. Note: $0 \in \mathbb{N}$.

**Ex 1.16.** Write $Y = N - X$. Prove that

$$\mathsf{P}\{X = x\}\,\mathsf{P}\{Y = y\} = \binom{x+y}{x} p^x (1-p)^y\,\mathsf{P}\{N = x + y\}. \tag{1.1}$$

**Ex 1.17.** Prove that $N$ is Poisson distributed.

IMPROPER INTEGRALS AND THE CAUCHY DISTRIBUTION This problem challenges your integration skills and lets you think about the subtleties of integrating a function over an infinite domain. (Such integrals are called improper integrals.)

Assume that $X$ has the Cauchy distribution. Recall that $\mathsf{E}[X]$ does not exist (hence, it is not automatic that the expectation of a some arbitrary r.v. exists).

**Ex 1.18.** Why does $\mathsf{E}\left[\frac{|X|}{X^2+1}\right]$ exist? Find its value.

**Ex 1.19.** Explain why the previous exercise implies that $\mathsf{E}\left[\frac{X}{X^2+1}\right]$ exists. Then find its value.

# 2 HINTS

**h.1.8.** From this exercise you should memorize this: **independence is a property of the joint CDF, not of the rvs**.

**h.1.17.** Use the relation of the previous exercise to show that

$$P(N=n+1)=\frac{\lambda}{1+n}P(N=n). \tag{2.1}$$

*Bigger hint:* Fill in $y=0$ in the LHS and RHS of (1.1); call this expression 1. Then fill in $y=1$ to a obtain a second expression. Divide these two expressions and note that $\mathsf{P}\{X=x\}$ cancels. Finally, define

$$\lambda=\frac{\mathsf{P}\{Y=1\}}{(1-p)\mathsf{P}\{Y=0\}}. \tag{2.2}$$

## 3 SOLUTIONS

Compare your answers very carefully against ours. You should spend time thinking about the definition and notation we use. For instance, there is conceptual huge difference between $X$ and $x$. More generally, good notation and good understanding correlate (positively).

**s.1.1.** Check the definitions of the book.

Mistake: To say that $\mathsf{P}\{X = x\}$ is the PMF for a continuous random variable is wrong, because $\mathsf{P}\{X = x\} = 0$ when $X$ is continuous.

Why is $\mathsf{P}\{1 < x \leq 4\}$ wrong notation? hint: $X$ should be a capital. What is the difference between $X$ and $x$?

**s.1.2.**

$$f_X(x) = \int_0^1 f_{X,Y}(x,y)\,\mathrm{d}y = 2\int_0^1 I_{x\leq y}\,\mathrm{d}y = 2\int_x^1 \mathrm{d}y = 2(1-x) \tag{3.1}$$

$$f_Y(y) = \int_0^1 f_{X,Y}(x,y)\,\mathrm{d}x = 2\int_0^1 I_{x\leq y}\,\mathrm{d}x = 2\int_0^y \mathrm{d}y = 2y. \tag{3.2}$$

But $f_{X,Y}(x,y) \neq f_X(x) f_Y(y)$, hence $X, Y$ are dependent.

$$\begin{aligned} F_{X,Y}(x,y) &= \int_0^x \int_0^y f_{X,Y}(u,v)\,\mathrm{d}v\,\mathrm{d}u && (3.3)\\ &= 2\int_0^x \int_0^y I_{u\leq v}\,\mathrm{d}v\,\mathrm{d}u && (3.4)\\ &= 2\int_0^x \int I_{u\leq v} I_{0\leq v\leq y}\,\mathrm{d}v\,\mathrm{d}u && (3.5)\\ &= 2\int_0^x \int I_{u\leq v\leq y}\,\mathrm{d}v\,\mathrm{d}u && (3.6)\\ &= 2\int_0^x [y-u]^+\,\mathrm{d}u, && (3.7) \end{aligned}$$

because $u \geq y \implies I_{u\leq v\leq y} = 0$. Now, if $y > x$,

$$2\int_0^x [y-u]^+\,\mathrm{d}u = 2\int_0^x (y-u)\,\mathrm{d}u = 2yx - x^2, \tag{3.8}$$

while if $y \leq x$,

$$2\int_0^x [y-u]^+\,\mathrm{d}u = 2\int_0^y (y-u)\,\mathrm{d}u = 2y^2 - y^2 = y^2 \tag{3.9}$$

Make a drawing of the support of $f_{X,Y}$ to help to understand this better.

**s.1.3.**

$$\partial_x \partial_y F_{X,Y}(x,y) = \partial_x \partial_y F_X(x) F_Y(y) = \partial_x F_X(x) \partial_y F_Y(y) = f_X(x) f_Y(y).$$

**s.1.4.**

$$\frac{F_{X,Y}(x,y)}{F_X(x)} = \frac{\mathsf{P}\{X \leq x, Y \leq y\}}{\mathsf{P}\{X \leq x\}} \tag{3.10}$$

In the notes we define the conditional CDF as the function $F_{X|Y}(x|y) = \mathsf{P}\{X \leq x|Y = y\}$. This is not the same as the function above.

Mistake: $F_{X,Y}(x,y) \neq \mathsf{P}\{X = x, Y = y\}$. If you wrote this, recheck BH. for the conditional CDF, you do not condition on e.g. $X \leq x$. Compare your answer to what is written in the notes or the solution manual. Good notation and good understanding are positively correlated :).

**s.1.5.** $\mathsf{P}\{X = 0, Y = 0\} = 1/3 \cdot 3/4$, $\mathsf{P}\{X = 0, Y = 1\} = 1/3 \cdot 1/4$, and so on.

If we have one column with $Y = 0$ and the other with $Y = 1$, then the sum over the columns are $\mathsf{P}\{Y = 0\}$ and $\mathsf{P}\{Y = 1\}$. The row sum for row $i$ are $\mathsf{P}\{X = i\}$.

Changing the values will (most of the time) make $X$ and $Y$ dependent. But, what if we changes the values such that $\mathsf{P}\{X = 0, Y = 0\} = 1$? Are $X$ and $Y$ then again independent? Check the conditions again.

**s.1.6.** The number of produced items (laid eggs) is $N$. The probability of hatching is $p$, that is, an item is ok. The hatched eggs are the good items.

**s.1.7.** For $X, Y$ to be independent, it is necessary that $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for all $x, y$, not just one particular choice. (This is an example that satisfying a necessary condition is not necessarily sufficient.)

**s.1.8.** Many answers are possible here, depending on extra assumptions you make. Here is one. Suppose, just by change, the fraction of taller guys in the street is a bit higher the population fraction. Assuming that taller (shorter) people prefer taller (shorter) spouses, there must be dependence between the height of the men and the woman. This is because when selecting a man, I can also select his wife.

Mistake: $\mathsf{P}\{Y\}$ is wrong notation. This is wrong because we can only compute the probability of an event, such as $\{Y \leq y\}$. But $Y$ itself is not an event.

**s.1.9.** Only when $X, Y$ are independent.

Mistake: independence of $X$ and $Y$ is not the same as the linear independence. Don't confuse these two types of dependene.

**s.1.10.** Given $N = n$, the random variable $X$ has a certain distribution, binomial for instance.

**s.1.11.** This claim is incorrect, because $X, Y$ are discrete, hence they have a PMF, not a PDF.

Mistake: Someone said that $\partial_x \partial_y$ is not correct notation; however, it is correct! It's a (much used) abbreviation of the much heaver $\partial^2/\partial x \partial y$. Next, the derivative of the PMF is not well-defined (at least, not within this course. If you object, ok, but then show that you passed a decent course on measure theory.)

**s.1.12.** This question tests your modeling skills too.

In hindsight, the questions have to reorganized a bit. The capital at the end of the $i$th week is $I_i = I_{i-1} + X_i - C_i$.

Suppose claims arrive at the beginning of the week, and contributions arrive at the end of the week (people prefer to send in their claims early, but they prefer to pay their contribution as late as possible). If we don't have sufficient money in cash, then we cannot pay a claim. Thus, $\max\{I_0 - C_1\}$ is our capital just before the contribution arrives. Hence, $I'_1$ is our capital at the end of week 1 under the assumption that we never pay out more than we have in cash. Likewise for $I'_2$

$\bar{I}_n$ is the lowest capital we have seen for the first $n$ weeks.

In the supermarket setting, $I_i$ is our inventory is we can be temporarily out of stock, but as soon as new deliveries—so called replenishments—arrive then we serve the waiting customers immediately. The model with $I'$ corresponds to a setting is which we consider unmet demand as lost.

$$\mathsf{P}\{I_0 <= 0\} = \mathsf{P}\{2 + X_1 - C_1 < 0\} = \frac{1}{10}\sum_{i=1}^{10} \mathsf{P}\{C_1 > 2 + i\} = \frac{1}{10}\sum_{i=1}^{5} \mathsf{P}\{C_1 > 2 + i\} \tag{3.11}$$

$$= \frac{1}{10}\sum_{i=1}^{5} \frac{6-i}{9}. \tag{3.12}$$

When grading, I realized that questions 8 was not quite reasonable to ask as an exam question. We graded this leniently. As I find it too boring to compute these probabilities by hand, here is the python code. The ideas in the code are highly interesting and useful. The main data structure here is a dictionary, one of the most used data structure in python. I don't have the R code yet, so if you take the (unwise) decision to stick to only R, you have to wait a bit until somebody sends me the R code for this problem.

```python
C = {}
for i in range(0, 9):
    C[i] = 1 / 9

X = {}
for i in range(1, 11):
    X[i] = 1 / 10


I0 = 2

I1 = {}
for k, p in X.items():
    for l, q in C.items():
        i = I0 + k - l
        I1[i] = I1.get(i, 0) + p * q

print("I1, ", sum(I1.values()))  # check


# compute P(I1<0):
P = sum(r for i, r in I1.items() if i < 0)
print(P)


I2 = {}
for i, r in I1.items():
    for k, p in X.items():
        for l, q in C.items():
            j = i + k - l
            I2[j] = I2.get(j, 0) + r * p * q

print("I2 ", sum(I2.values()))  # just a check

# compute P(I2<0):
P = sum(r for i, r in I2.items() if i < 0)
```

Interestingly, $I_i' \geq 1$. (This is so simple to see that I first did it wrong.)

Mistake: note that $X_i$ and $C_i$ are discrete r.v.s, not continuous. The sum of two uniform random variables is not uniform. For example, think of the sum of two die throws. Is getting 2 just as likely as getting 7?

**s.1.14.** Mistakes: Simulation and numerical integration are not the same. Formulate your answers precisely: it is not simulation that yields exactly the same value!

**s.1.15.** In this exercise we want to prove that $N$ is Poisson distributed. So you cannot assume this in your solution.

**s.1.16.**

**s.1.17.**

**s.1.18.** It is essential that you include your arguments.

**s.1.19.**