Probability distributions EBP038A05: 2020-2021 Assignments plus solutions

Nicky D. van Foreest, Ruben van Beesten
Joost Doornbos, Wietze Koops, Mikael Makonnen, Zexuan Yuan

March 15, 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 $f_{X,Y}(x,y) = 2I_{x \leq y}$.

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 ith period, we receive a contribution X_i uniform on the set $\{1,2,\dots,10\}$ and a claim C_i uniform on $\{0,1,\dots 8\}$.

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.

Python Code

```python
import numpy as np

np.random.seed(3)

num = 10

a = np.random.uniform(size=num)
b = np.random.uniform(size=num)
success = np.abs(a - b) < 0.25
print(a)
print(b)
print(success)
print(success.mean(), success.var())
```

R Code

```r
set.seed(3)

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.

Python Code

```python
1 import numpy as np
2 from scipy.stats import expon
3 from scipy.integrate import quad
4
5 labda = 3
```

```python

X = expon(scale = 1 / labda).rvs(1000)
# print(X)
print(X.mean())

success = (X > 1) * (X < 4)
# print(success)
print(success.mean(), success.std())


def F(x):  # CDF
    return 1 - np.exp(-labda * x)


def f(x):  # density
    return labda * np.exp(-labda * x)


print(F(4) - F(1))

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

R Code

```r
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 λ. 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. It is essential that you include your arguments.

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 ASSIGNMENT 2

2.1 *Have you read well?*

Ex 2.1. Example 7.2.2. Write down the integral to compute $\mathsf{E}\left[(X-Y)^2\right]$. You don't have to solve the integral.

Ex 2.2. Give a brief example of a situation where it might be more convenient to employ the correlation instead of the covariance and explain why.

Ex 2.3. In queueing theory the concept of squared coefficient of variance SCV of a rv X is very important. It is defined as $C = \mathsf{V}[X]/(\mathsf{E}[X])^2$. Is the SCV of X equal to Corr(X,X)? Can it happen that $C > 1$?

Ex 2.4. Using the definition of Covariance (Definition 7.3.1) derive the expression $\mathsf{Cov}[X,Y] = \mathsf{E}[XY] - \mathsf{E}[X]\,\mathsf{E}[Y]$. Use this finding to show why independence of X and Y implies their uncorrelatedness (note that the converse does not hold).

Ex 2.5. Let U,V be two r.v.s and let $a,b \in \mathbb{R}$. Express $\mathsf{Cov}[a(U+V),b(U-V)]$ in terms of $\mathsf{V}[U]$, $\mathsf{V}[V]$ and $\mathsf{Cov}[U,V]$ (by using the expression obtained in the previous question).

Ex 2.6. Prove the key properties of covariance 1 to 5 on page 327 of the book (page 338 pdf).

Ex 2.7. Come up with a short illustrative example in which the random vector $\mathbf{X} = (X_1,\dots,X_6)$ follows a Multinomial Distribution with parameters $n = 10$ and $\mathbf{p} = (\frac{1}{6},\dots,\frac{1}{6}) \in \mathbb{R}^6$.

Ex 2.8. Is the following claim correct? If the r.v.s X,Y are both normally distributed, then (X,Y) follows a Bivariate Normal distribution.

Ex 2.9. Let (X,Y) follow a Bivariate Normal distribution, with X and Y marginally following $\mathcal{N}(\mu,\sigma^2)$ and with correlation ρ between X and Y.

1. Use the definition of a Multivariate Normal Distribution to show that $(X+Y,X-Y)$ is also Bivariate Normal.
2. Find the marginal distributions of $X+Y$ and $X-Y$.
3. Compute $\mathsf{Cov}[X+Y,X-Y]$ there was a typo here. Then, write down the expression for the joint PDF of $(X+Y,X-Y)$.

Ex 2.10. Let X,Y,Z be i.i.d. $\mathcal{N}(0,1)$. Determine whether or not the random vector

$$\mathbf{W} = (X+2Y, 3X+4Z, 5Y+6Z, 2X-4Y+Z, X-9Z, 12X+\sqrt{3}Y-\pi Z)$$

is Multivariate Normal. (Explain in words, don't do a lot of tedious math here!)

2.2 *Exercises at about exam level*

Ex 2.11. Take $X \sim \text{Unif}(\{-2,-1,1,2\})$ and $Y = X^2$. What is the correlation coefficient of X and Y? If we would consider another distribution for X, would that change the correlation?

Ex 2.12. We have a machine that consists of two components. The machine works as long as both components have not failed (in other words, the machine fails when one of the two components fails). Let X_i be the lifetime of component i.

1. What is the interpretation of $\min\{X_1, X_2\}$?
2. If X_1, X_2 iid $\sim \text{Exp}(10)$ (in hours), what is the probability that the machine is still 'up' (i.e., not failed) at time T?
3. Use the previous result to determine the distribution of $\min\{X_1, X_2\}$.
4. What is the expected time until the machine fails?

Ex 2.13. We have two r.v.s X and Y with the joint PDF $f_{X,Y}(x,y) = \frac{6}{7}(x+y)^2$ for $x, y \in (0,1)$ and 0 else. Also we consider the two r.v.s U and V with the joint PDF $f_{U,V}(u,v) = 2$ for $u, v \in [0,1], u+v \le 1$ and 0 else.

1. Compute $\mathsf{P}\{X+Y>1\}$.
2. Compute $\mathsf{Cov}[U,V]$.

(Hint: first draw the area over which you want to integrate, if this does not help check out the discussion board post on exercise 7.13a from the first Tutorial)

2.3 *Coding skills*

VERIFY THE ANSWERS OF BH.5.6.5 Read this example of BH first. We chop up the exercise in many small exercises..

For the python code below, run it for a small number of samples; here I choose `samples=2`. Read the print statements, and use that to answer the questions below.

Python Code

```python
import numpy as np
from scipy.stats import expon

np.random.seed(10)

labda = 6
num = 3
samples = 2

```

```
10 X = expon(scale=labda).rvs((samples, num))
11 print(X)
12 T = np.sort(X, axis=1)
13 print(T)
14 print(T.mean(axis=0))
15
16 expected = np.array([labda / (num - j) for j in range(num)])
17 print(expected)
18 print(expected.cumsum())
```

R Code

```
1 set.seed(10)
2
3 labda = 6
4 num = 3
5 samples = 2
6
7 X = matrix(rexp(samples * num, rate = 1 / labda), nrow = samples, ncol = num)
8 print(X)
9 bigT = X
10 for (i in 1:samples) {
11   bigT[i,] = sort(bigT[i,])
12 }
13 print(bigT)
14 print(colMeans(bigT))
15
16 expected = rep(0, num)
17 for (j in 1:num) {
18   expected[j] = labda / (num - (j - 1))
19 }
20 print(expected)
21 print(cumsum(expected))
```

Ex 2.14. In line P.11[1] we print the value of `X` in line P.10, R.7 and R.8, respectively. What is the meaning of `X`?

Ex 2.15. What is the meaning of `T` in line P.12 (R.11)?

Ex 2.16. What do we print in line P.14, R.14?

Ex 2.17. What is meaning of the variable `expected`?

Ex 2.18. What is the `cumsum` of `expected`?

[1] Line P.x refers to line x of the Python code. Line R.x refers to line x of the R code.

Ex 2.19. Now that you understand what is going on, rerun the simulation for a larger number of samples, e.g., 1000, and discuss the results briefly.

ON BH.7.48 Read this exercise first and solve it. Then consider the code below.

Python Code

```
import numpy as np

np.random.seed(3)


def find_number_of_maxima(X):
    num_max = 0
    M = -np.infty
    for x in X:
        if x > M:
            num_max += 1
            M = x
    return num_max


num = 10
X = np.random.uniform(size=num)
print(X)

print(find_number_of_maxima(X))

samples = 100
Y = np.zeros(samples)
for i in range(samples):
    X = np.random.uniform(size=num)
    Y[i] = find_number_of_maxima(X)

print(Y.mean(), Y.var(), Y.std())
```

R Code

```
set.seed(3)

find_number_of_maxima = function(X) {
  num_max = 0
  M = -Inf
  for (x in X) {
    if(x > M) {
```

```
8        num_max = num_max + 1
9        M = x
10      }
11    }
12    return(num_max)
13  }
14
15
16  num = 10
17  X = runif(num, min = 0, max = 1)
18  print(X)
19
20  print(find_number_of_maxima(X))
21
22  samples = 100
23  Y = rep(0, samples)
24  for (i in 1:samples) {
25    X = runif(num, min = 0, max = 1)
26    Y[i] = find_number_of_maxima(X)
27  }
28
29  print(mean(Y))
30  print(var(Y))
31  print(sd(Y))
```

Ex 2.20. Explain how the small function in lines P.6 to P.13 (R.4-R.12) works. (You should know that `x += 1` is an extremely useful abbreviation of the code `x = x + 1`).

Ex 2.21. Explain the code in lines P.25 and P.26 (R.25, R.26).

WHY IS THE EXPONENTIAL DISTRIBUTION SO IMPORTANT? At the Paris metro, a train arrives every 3 minutes on a platform. Suppose that 50 people arrive between the departure of a train and an arrival. It seems entirely reasonable to me to model the arrival times of the individual people as distributed on the interval $[0,3]$. What is the distribution of the inter-arrival times of these people? It turns out to be exponential!

You might want to compare your final result to Figure BH.13.1 (It is not forbidden to read the book beyond what you have to do for this course!). In this exercise we use simulation to see that clustering of arrival times.

Python Code

```python
1  import numpy as np
2
3  np.random.seed(3)
```

```
4
5
6 num = 5 # small sample at first, for checking.
7 start, end = 0, 3
8 labda = num / (end - start)  # per minute
9 print(1 / labda)
10
11 A = np.sort(np.random.uniform(start, end, size=num))
12 print(A)
13 print(A[1:])
14 print(A[:-1])
15 X = A[1:] - A[:-1]
16 print(X)
17
18 print(X.mean(), X.std())
```

R Code

```
1 set.seed(3)
2
3
4 num = 5
5 start = 0
6 end = 3
7 labda = num / (end - start)
8 print(1 / labda)
9
10 A = sort(runif(num, min = start, max = end))
11 print(A)
12 print(A[-1])
13 print(A[-length(A)])
14 X = A[-1] - A[-length(A)]
15 print(X)
16
17 print(mean(X))
18 print(sd(X))
```

Ex 2.22. Explain the result of line P.12 (R.13)

Ex 2.23. Compare the result of line P.13 and P.14 (R.12, R.13); explain what is `A[1:]` (`A[-1]`)

Ex 2.24. Compare the result of line P.12 and P.14 (R.11 and R.13); explain what is `A[:-1]` (`A[-length(A)]`).

Ex 2.25. Explain what is `X` in P.15 (R.14)

Ex 2.26. Why do we compare $1/\lambda$ and `X.mean()`?

Ex 2.27. Recall that $\mathsf{E}[X] = \sigma(X)$ when $X \sim \text{Exp}(\lambda)$. Hence, what do you expect to see for `X.std()`?

Ex 2.28. Run the code for a larger sample, e.g. 50, and discuss (very briefly) your results.

2.4 *Challenges*

This exercise will given an example of how probability theory can pop up in OR problems, in particular in linear programs. It introduces you to the concept of *recourse models*, which you will learn about in the master course Optimization Under Uncertainty. Disclaimer: the story is quite lengthy, but the concepts introduced and questions asked are in fact not very hard. We just added the story to make things more intuitive.

WE CONSIDER A pastry shop that only sells one product: chocolate muffins. Every morning at 5:00 a.m., the shop owner bakes a stock of fresh muffins, which he sells during the rest of the day. Making one muffin comes at a cost of $c = \$1$ per unit. Any leftover muffins must be discarded at the end of the day, so every morning he starts with an empty stock of muffins.

The owner has one question for you: determine the amount x of muffins that he should make in the morning to minimize his production cost. Note that the owner never wants to disappoint any customer, i.e., he requires that $x \geq d$, where d is the daily demand for muffins.

The problem can be formulated as a linear program (LP):

$$\min_{x\geq 0}\{cx \,:\, x \geq d\}. \tag{2.1}$$

For simplicity, we ignore the fact that x should be integer-valued.

Ex 2.29. Determine the optimal value x^* for x and the corresponding objective value in case d is deterministic.

Of course, in practice d is not deterministic. Instead, d is a random variable with some distribution. However, note that the LP above is ill-defined if d is a random variable. We cannot guarantee that $x \geq d$ if we do not know the value of d.

You explained the issue to the shop owner and he replies: “Of course, you’re right! You know, whenever I’ve run out of muffins and a customer asks for one, I make one on the spot. I never disappoint a customer, you know! It does cost me 50% more money to produce them on the spot, though, you know.”

Mathematically speaking, the shop owner just gave you all the (mathematical) ingredients to build a so-called *recourse model*. We introduce a *recourse variable* y in our model, representing the amount of muffins produced on the spot. Production comes at a unit cost of $q = 1.5c = \$1.5$. Assuming that we know the distribution of d, we can then minimize the *expected total cost*:

$$\min_{x\geq 0}\left\{cx + \mathsf{E}[v(d,x)]\right\}, \tag{2.2}$$

where $v(d,x)$ is the optimal value of another LP, namely the *recourse problem*:

$$v(d,x) := \min_{y\geq 0}\{qy \ : \ x+y \geq d\}, \tag{2.3}$$

for given values of d and x. The recourse problem can easily be solved explicitly: we get $y = d - x$ if $d \geq x$ and $y = 0$ if $d < x$. So we obtain

$$v(d,x) = q(d-x)^+, \tag{2.4}$$

where the operator $(\cdot)^+$ represents the *positive value* operator, defined as

$$(s)^+ = \begin{cases} s & \text{if } s \geq 0, \\ 0 & \text{if } s < 0. \end{cases} \tag{2.5}$$

Ex 2.30. To get some more insight into the model, suppose (for now) that $d \sim U\{10,20\}$. Solve the model, i.e., find the optimal amount x^*. *Hint: First, compute the value of* $\mathsf{E}[v(d,x)]$ *as a function of x. Then find the optimal value of x.*

Ex 2.31. What is the expected recourse cost (expected cost of on-the-spot production) at the optimal solution x^*, i.e., compute $\mathsf{E}[v(d,x^*)]$?

To solve the model correctly, we need the true distribution of d. We learn the following from the shop owner: "My granddaughter, who's always running around in my shop, is a bit data-crazy, you know, so she's been collecting some data. I remember her saying that 'the demand from male and female customers are both approximately normally distributed, with mean values both equal to 10 and standard deviations of 5'. She also mentioned something about correlation, but I don't remember exactly, you know. It was either almost 1 or almost −1. I hope this helps!"

Mathematically, we've learned that $d = d_m + d_f$, with $(d_m, d_f) \sim \mathcal{N}(\mu, \Sigma)$, where $\mu = (\mu_m, \mu_f) = (10, 10)$ and $\Sigma_{11} = \sigma_m^2 = \Sigma_{22} = \sigma_f^2 = 5^2 = 25$. Finally, $\Sigma_{12} = \Sigma_{21} = \mathsf{Cov}\left[d_m, d_f\right] = \rho\sigma_m\sigma_f = 25\rho$. Also, we know that either $\rho \approx 1$ or $\rho \approx -1$.

Ex 2.32. Calculate x^* and the corresponding objective value for the case $\rho = -1$. (Do not read $\rho = 1$, this case is not simple.)

Ex 2.33. Consider the two extreme cases $\rho = 1$ and $\rho = -1$. In which case will the shop owner have lower expected total costs? Provide a short, intuitive explanation. *Hint: you don't have to compute x^* for the case where $\rho = 1$ (this is not easy!).*

3 ASSIGNMENT 3

3.1 *Have you read well?*

Ex 3.1. Let X have the discrete uniform distribution on the set $\{0,1,2,3,4,5\}$. Derive the PMF and the CDF of $Z = 3X$. Explicitly specify the domain.

Ex 3.2. Let $X \sim \text{Unif}(0,5)$. Using the one dimensional change of variables theorem (BH.8.1.1), derive the PDF and the CDF of $Z = 3X$. Explicitly specify the domain.

Ex 3.3. Let $X \sim \text{Norm}(\mu, \sigma^2)$. Using the one dimensional change of variables theorem BH.8.1.1, show that $Z = \frac{X-\mu}{\sigma}$ follows the standard normal distribution.

Ex 3.4. Let $X \sim \text{Exp}(1)$. Derive the PDF of e^{-X}.

Ex 3.5. Let X, Y be i.i.d. standard normal. Using the n-dimensional change of variables theorem, derive the joint PDF of $(X+Y, X-Y)$.

Check your final answer using BH.7.5.8.

Ex 3.6. Specify the domain of the new random variable for the following transformations; this important aspect of the change of variables is often overlooked. Let U, V, W, X, X_1, X_2, Y and Z be r.v.s and let a, b and c be arbitrary constants.

1. $Z = Y^4$ for $Y \in (-\infty, \infty)$;
2. $Y = X^3 + a$ for $X \in (0,1)$;
3. $U = |V| + b$ for $V \in (-\infty, \infty)$;
4. $Y = e^{X^3}$ for $X \in (-\infty, \infty)$;
5. $V = U I_{U \leq c}$ for $U \in (-\infty, \infty)$;
6. $Y = \sin(X)$ for $X \in (-\infty, \infty)$;
7. $Y = \frac{X_1}{X_1+X_2}$ for $X_1 \in (0, \infty)$ and $X_2 \in (0, \infty)$;
8. $Z = \log(UV)$ for $U \in (0, \infty)$ and $V \in (0, \infty)$.

Ex 3.7. To find the distribution of a convolution through the change of variables formula, we seem to need to add a 'redundant' equality? But why is that? What would be the problem if we do not add this? Explain in your own words.

Ex 3.8. When adding a different equality, we need to be careful to not create a functional relationship between our two new variables U, V, for example $U = X+Y$ and $V = \sin(X+Y)$, or $U = \frac{X}{Y}$ and $V = \frac{Y}{X}$ for conforming X, Y. What would happen to the determinant of the Jacobian matrix if we did? Why would this happen? Explain in your own words.

Ex 3.9. In this exercise, we combine what we learned in BH.8.1.4 and BH.8.1.9. Let S be the sum of two i.i.d. chi-square distributed variables. Using just these two examples, show that $S \sim \text{Exp}(1/2)$.

Ex 3.10. A student has obtained an i.i.d. random sample of size 2 from a Cauchy distribution. Let the r.v.s X and Y model the values of the first and second sample. Since s/he does not know what the mean of a Cauchy distribution is, s/he wants to average the sample to obtain what she thinks is a good estimate of the true mean.

To find the distribution of this sample mean, we need to find an expression for $f_W(w)$, where $W = \frac{X+Y}{2}$.

1. Find an expression for $f_W(w)$ in the form of an integral, but do not solve it.
2. It turns out that if we solve the integral, we get that $f_W(w) = f_X(w)$. The distribution of our sample mean is still Cauchy; we did not obtain a better estimate of the Cauchy mean by calculating the sample mean!

   Explain (in your own words) why this makes sense.

3.2 *About exam level*

Ex 3.11. Let X, Y iid $\sim \text{Unif}([0,1])$.

1. Write a computer program in python or R to estimate $\mathsf{P}\{X+Y \leq 1, XY \leq 2/9\}$.

Ex 3.12. Let X, Y be continuous r.v.s with CDF $F_{X,Y}(x,y) = (x-1)^2(y-2)/8$ for $x \in (1,3)$.

a. Explain that $y \in (2,4)$ for F to be a proper CDF.

b. What is $F(3,7)$?

c. Determine the PDF.

d. Compute $\mathsf{P}\{2 < X < 3\}$

e. Compute $\mathsf{P}\{2 < X < 3, 2 < Y < 3\}$.

f. Compute $\mathsf{P}\{Y < 2X\}$.

g. Compute $\mathsf{P}\{Y \leq 2X\}$.

h. Compute $\mathsf{P}\{Y < 2X, Y + 2X > 6\}$.

Ex 3.13. Consider the general case where we are given the relationship $U = V^4$ between the random variables U and V for $V \in (-3,2)$.

Explain why we cannot simply invoke the change of variables theorem.

Now imagine V following a uniform distribution on the given interval. Consider the given transformation on the intervals $(-3,0)$ and $(0,2)$ separately. Explain why this allows you to employ the change of variables theorem and find the distribution of U on these intervals. Finally combine these results (using indicator functions) and state the PDF of U (remember to adjust the domain for the indicator functions according to the transformations).

Ex 3.14. Let $U \sim \text{Unif}(0,\pi)$. Use BH.8.1.9 to show that $X = \tan(U)$ has the Cauchy distribution. Compare this exercise to BH.8.1.5.

3.3 *Coding skills*

PING PONG BALLS How many ping pong balls fit into an Airbus Beluga? One way to answer this is as follows. According to this wiki-page the cargo volume V of this airplane is $1500\,\text{m}^3$. But this number is based on the physical dimensions that is available to store containers, tanks, and so on. So, I estimate the volume as about twice that amount, i.e., $V = 2500\,\text{m}^3$. The volume of a ping pong ball is $v = 4\pi r^3/3 = 33.49333333333333\,\text{cm}^3$ with $r = 2$ cm. A plain division gives 74.6268656716418 ping pong balls. Note, I left out the 10^6 conversion from meters to cm, and I do not take into the sphere packing factor. Besides that, I hope you agree with me that providing an result with the precision as given here is plain ridiculous. (But from reason incomprehensible to me, even professional econometricians like to report results with 10 digits or more, without questioning the precision.)

However, I know that the volumes of an air plane and a ping pong ball is an estimate, rather than a precise number as assumed above. It seems to be better to approximate V and v as rvs. Let's assume that

$$V \sim N(2500, 500^2), \qquad v \sim N(33.5, 0.5^2),$$

where the variances express my trust in my guess work. What is now the mean of $N = V/v$ and its std? In fact, finding the closed form expression for the distribution of N is not entirely simple. However, with simulation it's easy to get an estimate.

Ex 3.15. How does this exercise relate to BH.8.11 and BH.8.12? What is similar, what are crucial differences?

Ex 3.16. Use the documentation of the `norm` (`rnorm`) function of python (R) to explain why we set the scale as we do. Relate this to location-scale discussion in BH.

Ex 3.17. Explain lines 8 and 13 of the python code or lines 5 and 8 of the R code. (Optional: There is a conceptual difference between the two languages here. If you are interested in both languages, also comment on the difference)

Ex 3.18. Use the code below to provide the estimates.

Ex 3.19. Contrary to BH.7.1.25 if you run the code below, you'll see that $\mathsf{E}[N] < \infty$, and, in fact, very near to the deterministic answer. But isn't this strange? We divide two normal random variables, just like BH.7.1.25, but there the expectation is infinite. Comment on the difference.

The numerical results suggest the interesting guess $\mathsf{V}[N] \approx \mathsf{V}[V] * \mathsf{V}[v]$, but is this true more generally? In Section 3.4 we study this problem in more detail.

Python Code

```python
import numpy as np
from scipy.stats import norm

num = 500

np.random.seed(3)

V = norm(loc=2500, scale=500)
v = norm(loc=33.5, scale=0.5)

print(V.mean(), V.std()) # just a check

N = V.rvs(num) / v.rvs(num)
print(N.mean(), N.std())

print(2500/33.5)
print(np.sqrt(500*0.5))
```

R Code

```r
num <- 500

set.seed(3)

V = rnorm(num, 2500, 500)
v = rnorm(num, 33.5, 0.5)

N = V / v
paste(mean(N), sd(N))

2500/33.5
sqrt(500*0.5)
```

SUMS OF RVS We start from BH.8.27 (which you have to read now). We are interested in the difference between the distribution of $X+Y+Z$ and the normal distribution. But why the normal distribution? As it turns out, the central limit law, see BH.10, states that the distribution of sums of r.v.s converge to the normal distribution (in a specific sense)

Here some code to simulate.

Python Code

```python
import numpy as np
from scipy.stats import norm

import matplotlib.pylab as plt
import seaborn as sns

sns.set()

np.random.seed(3)

k = 3
Zexact = norm(loc=k / 2, scale=np.sqrt(k / 12))
X = np.arange(0, 3, 0.1)

XYZ = np.random.uniform(size=(4000, k))
# print(XYZ)  # if you want to see it.
Z = XYZ.sum(axis=1)
sns.distplot(Z)
plt.plot(X, Zexact.pdf(X))
plt.show()
```

R Code

```r
set.seed(3)

k = 3
X <- seq(0, 3, by = 0.1)
Zexact <- dnorm(X, mean = k / 2, sd = sqrt(k / 12))


XYZ <- matrix(NA, 4000, k)
for (i in 1:k) {
  XYZ[,i] <- runif(4000, min = 0, max = 1)
}
Z <- rowSums(XYZ)

par()
hist(Z, prob = TRUE, breaks = 31)
lines(X, Zexact, type = "l", col = "orange")
lines(density(Z), col = "blue")
```

Ex 3.20. What is the shape of XYZ in the code above, i.e., how many rows and columns does it have? If you don't know, run the code, and print it.

Ex 3.21. What is the shape (rows and columns) of Z?

Ex 3.22. Explain the values for `loc` and `shape` in `Zexact`. (Read the documentation of scipy.stats.norm on the web is necessary.) To which definition in BH does this loc-scale transformation relate?

Ex 3.23. Change the seed to your student id, or any other number you like, run the code, and include the graph produced by your simulation. Explain what you see.

Now we do an exact computation.

Python Code

```python
import numpy as np
from scipy.stats import norm

import matplotlib.pylab as plt
import seaborn as sns

sns.set()

N = 200
x = np.linspace(0, 2, 2 * N)
fx = np.ones(N) / N
f2 = np.convolve(fx, fx)
f3 = np.convolve(f2, fx)

k = 3

x = np.linspace(0, k, len(f3))
Zexact = norm(loc=k / 2, scale=np.sqrt(k / 12))


plt.plot(x, N * f3, label="conv")
plt.plot(x, Zexact.pdf(x))
plt.legend()
plt.show()
```

R Code

```r
N = 200
x = seq(0, 2, length.out = 2 * N)
fx = rep(1, N) / N
f2 = convolve(fx, fx, type = "open")
f3 = convolve(f2, fx, type = "open")

```

```
k = 3

x = seq(0, k, length.out = length(f3))
Zexact = dnorm(x, mean = k/2, sd = sqrt(k / 12))

par()
plot(x, N * f3, col = "blue", type = "l", ylim = c(0, 0.8))
lines(x, Zexact, type = "l", col = "orange")
legend("topright", legend = "conv", bty = "n",
       lwd = 2, cex = 1.2, col = "blue", lty = 1)
```

Ex 3.24. Read the documentation of np.convolve . Why is it called like this?

Ex 3.25. In the code, what is `f2`?

Ex 3.26. What is `f3`?

Ex 3.27. Why do we set `k=3`?

Ex 3.28. A bit harder, why do we plot `N*f3`, i.e., why do have to multiply with `N`? Relate this to the meaning of $f(x)\mathrm{d}x$, where f the density of some random variable. (To understand why is very important. Think hard, and read the solution when it becomes available.)

Ex 3.29. Yet a tiny bit harder, consider `f4 = np.convolve(f3, fx)` and `g4 = np.convolve(f2, f2)`. Why are they, numerically speaking, equal?

Ex 3.30. When you would compute the maximum of `np.abs(f4 -g4)` you would see that this is about 10^{-10}, or so. Hence, a small number. This is not equal to 0, but we know that this is due to rounding effects.

How can we use the function `np.isclose()` to get around this problem? (You should memorize from this question that you should take care when testing on whether floating point numbers are the same or not.)

3.4 *Challenges*

This challenge is a continuation of the Beluga case of Section 5.3, and we discuss some ways to check whether $\mathsf{V}[N] \approx \mathsf{V}[V]\,\mathsf{V}[v]$ holds in general, and then we try to find a better approximation. We chopped up the challenge into many exercises, to help you organize the ideas.

Recall that in Section 5.3 we have been a bit sloppy about the units, measuring the volumes of the airplane in m^3 and a ping pong ball in cm^3, so actually N is in millions of ping pong balls. Note that using different units can easily lead to confusion; as a take-away , choose one unit.

One way to check the correctness of $\mathsf{V}[N] \approx \mathsf{V}[V]\,\mathsf{V}[v]$ is to change the scale. In fact, memorize that changing scale is an easy way to check laws.

Ex 3.31. Suppose we instead measure the size of a ping pong ball in meters and the size of the airplane in hectometers. Explain that N is still in millions of ping pong balls. What happens to $\mathsf{V}[N]$ and what happens to $\mathsf{V}[V]\,\mathsf{V}[v]$ (theoretically)?

Another way to check a statement is to consider some extreme cases.

Ex 3.32. Suppose that we would know the size of a ping pong ball very accurately, i.e. we consider the extreme case where $\mathsf{V}[v] \to 0$. Explain that the approximation is not a good approximation in this limit.

Ex 3.33. Which of these two checks convinces you most that something is wrong with this approximation, and why?

We now turn to the task of trying to find a good approximation.

Ex 3.34. Assume that X and Y are independent. Show that

$$\mathsf{V}[XY] = \mathsf{V}[X]\,\mathsf{V}[Y] + \mathsf{V}[X]\,\mathsf{E}[Y]^2 + \mathsf{E}[X]^2\,\mathsf{V}[Y].$$

Ex 3.35. Assume in addition that we know at least one of X and Y quite precisely. Argue that the following is then a good approximation:

$$\mathsf{V}[XY] \approx \mathsf{V}[X]\,\mathsf{E}[Y]^2 + \mathsf{E}[X]^2\,\mathsf{V}[Y].$$

So far we have only considered the variance of a product, but we would like to know the variance of a ratio. For this we can use Taylor expansions to make accurate approximations.

Ex 3.36. Find the first order Taylor expansion of $\frac{1}{Z}$ around $a = \mathsf{E}[Z]$. By taking the expectation and the variance of this expansion, show that

$$\mathsf{E}\left[\frac{1}{Z}\right] \approx \frac{1}{\mathsf{E}[Z]}, \qquad \mathsf{V}\left[\frac{1}{Z}\right] \approx \frac{\mathsf{V}[Z]}{\mathsf{E}[Z]^4}.$$

Ex 3.37. Combine all of the above to derive the following approximation for the variance of the ratio of two independent random variables X and Z:

$$\mathsf{V}\left[\frac{X}{Z}\right] \approx \frac{\mathsf{V}[X]}{\mathsf{E}[Z]^2} + \mathsf{E}[X]^2\,\frac{\mathsf{V}[Z]}{\mathsf{E}[Z]^4}.$$

Ex 3.38. Check this approximation in the ways of the first two exercises.

After doing all this work, we would of course like to know how well this approximation does. When comparing the approximation to the sample standard deviation found in [3.15] for `num=500`, the result may be a bit disappointing. However, this is just because the sample standard deviation is also an estimate of the actual standard deviation of N, so by chance the result may be closer to $\mathsf{V}[V]\,\mathsf{V}[v]$ than to our new approximation.

In Chapter 10, you will learn something about the distribution of the sample variance. For now, just increase `num`. We know this decreases the variance of the sample mean and it also decreases the variance of the sample variance, so we get a more accurate estimate.

Ex 3.39. Use the result of the previous exercise to compute an approximation for $\mathsf{V}[N] = \mathsf{V}[V/v]$. Also use the code with a (much) higher value of `num`, to show that the approximation $\mathsf{V}[N] \approx \mathsf{V}[V]\,\mathsf{V}[v]$ is likely to be worse, even in the setting of [3.15] where it was quite good.

The following two exercises are really optional, but I found them very neat and insightful.

Ex 3.40. Recall that for a non-negative random variable X with finite variance, we define the squared coefficient of variation as $\mathrm{SCV}(X) = \mathsf{V}[X]/\mathsf{E}[X]^2$. Using the SCV, show that the approximations of [3.35] and [3.36] can be rewritten in the following neat way:

$$\begin{aligned} &\mathrm{SCV}(XY) \approx \mathrm{SCV}(X) + \mathrm{SCV}(Y). \\ &\mathrm{SCV}(1/Z) \approx \mathrm{SCV}(Z). \end{aligned}$$

In BH.10, you will learn Jensen's inequality, which implies that $\mathsf{E}\left[\frac{1}{Z}\right] \geq \frac{1}{\mathsf{E}[Z]}$ for all positive random variables Z. In the following exercise, we reflect on this by finding a more accurate approximation based on the second order Taylor expansion.

Ex 3.41. Find the second order Taylor expansion of $\frac{1}{Z}$ around $a = \mathsf{E}[Z]$. By taking the expectation, show that

$$\mathsf{E}\left[\frac{1}{Z}\right] \approx \frac{1}{\mathsf{E}[Z]} + \frac{2\mathsf{V}[Z]}{\mathsf{E}[Z]^3}.$$

Note that this is always at least $\dfrac{1}{\mathsf{E}[Z]}$.

4 ASSIGNMENT 4

4.1 *Have you read well?*

Ex 4.1. Correct? Let T be the sum of two i.i.d. Unif(0,1) r.v.s. Then there exist a, b such that $T \sim \text{Beta}(a, b)$. (You don't need to derive the distribution of T.)

Ex 4.2. Show that $\beta(1, b) = 1/b$ by integrating the PDF of the beta distribution for $a = 1$. (Do not use the results of BH 8.5 for this exercise.)

Ex 4.3. Let $a, b > 1$. Show that the PDF of the beta distribution attains a maximum at $x = \frac{a-1}{a+b-2}$. Explicitly indicate where the assumption that $a, b > 1$ is used.

Ex 4.4. Explain in your own words:

1. What is a prior?
2. What is a conjugate prior?

Ex 4.5.

1. Look up on the web: what is the conjugate prior of the multinomial distribution? Give a name and a formula.
2. Explain why the Beta distribution is a special case of this distribution.

Ex 4.6. You make a test with n multiple choice questions and you give the correct answer to each question independently with probability p. The teacher's prior belief about p is reflected by a uniform distribution: $p \sim \text{Unif}(0,1)$. Let X be the number of correct answers you give. What is the teacher's posterior distribution $p|X = k$? (You don't have to do a lot of math here; simply use a result from the book.)

Ex 4.7. You find a coin on the street. Initially, you are rather confident that this should be (approximately) a fair coin. This is reflected in your prior belief of the probability p of heads: $p \sim \text{Beta}(10, 10)$. Your friend is a bit more skeptical and assumes a uniform prior: $p \sim \text{Unif}(0,1)$. You toss the coin 1000 times, and it comes up heads 900 times.

1. Determine your posterior distribution. (Again, use a result from the book)
2. Determine your friend's posterior distribution.
3. Compare the means of your posterior distribution and your friend's posterior distribution. Comment on the effect of the prior distribution if you have a lot of data.

Ex 4.8. Consider the chi-square distribution from BH Example 8.1.4.

Starting from the expression $f_Y(y) = \varphi\left(\sqrt{y}\right) y^{-1/2}$ in this example, show that this chi-square distribution is a special case of the Gamma distribution and specify the corresponding values of the parameters a and λ.

Ex 4.9. Correct? The sum of any two Gamma distributions is again Gamma.

Ex 4.10. Prove by induction that $\Gamma(n) = (n-1)!$ if n is a positive integer.

Ex 4.11. Correct? The Poisson distribution is the conjugate prior of the Gamma.

Ex 4.12. Let $X \sim \text{Gamma}(4,2)$ and $Y \sim \text{Gamma}(7,2)$ be independent r.v.s. What is the distribution of $X+Y$? What is the distribution of $\frac{X}{X+Y}$?

Ex 4.13. (This question is about Section 8.6 (order statistics). If you can answer this question, then you basically know everything you need to know about order statistics for the purpose of this course.)
Let $X_1, X_2, \ldots, X_9$ be a collection of random variables. Fill in the gaps (with just one word each time):

1. $X_{(1)}$ denotes the ... of $X_1, X_2, \ldots, X_9$.
2. $X_{(9)}$ denotes the ... of $X_1, X_2, \ldots, X_9$.
3. $X_{(5)}$ denotes the ... of $X_1, X_2, \ldots, X_9$.

4.2 *Exercises at about exam level*

You walk into a bar and you find two people, Amy and Bob, playing a game of darts. Their game consists of several rounds, called *legs*, and the first person to win 7 legs wins the match. You have never seen Amy or Bob play before, so you don't know their strength. Denoting by p the probability that Amy wins a leg, your prior belief can be modeled by a uniform distribution: $p \sim \text{Unif}(0,1)$. (Note: we assume that p remains constant during the entire match; even though your *beliefs* about p might change.)

Denoting by A a leg won by Amy and by B a leg won by Bob, the result of the first 10 legs is: $AAABBAABAB$. Your friend Charles is very confident that Amy will win the match and he offers you a bet: if Amy wins the match, you must pay €10 to Charles; if Bob wins the match, he must pay you €300. You are tempted to take the bet, but you want to do some calculations first.

Ex 4.14. Is the order in which Amy and Bob won their respective legs relevant for your posterior probability that Bob will win the match?

Ex 4.15. Let A_n denote the number of legs that Amy won out of a total of n legs. Express the result of the first 10 legs in terms of A_n

Ex 4.16. What is the distribution of $A_n|p$ (i.e., the distribution of A_n given the value of p)?

Ex 4.17. Find the posterior distribution of p after observing $A_n = k$.

Ex 4.18. According to your posterior belief about p, what is the probability that Bob wins the match? Express your answer in terms of beta functions. (Hint: Use the law of total probability.)

Ex 4.19. Using the expression

$$\beta(a,b) = \frac{(a-1)!(b-1)!}{(a+b-1)!} \tag{4.1}$$

for every positive integers a, b, compute the probabilty from the previous question as a number.

Ex 4.20. Assuming that you want to maximize your expected outcome, should you take the bet?

4.3 *Coding skills*

We simulate the post office part of BH.8.36. Read it now, i.e., before reading the text below. Then read the code below. In the questions we ask you to explain what the code does. There are lots of print statements that have been commented out, but we left them in for you to include while experimenting with the code to see how the code works. (I often use print statements of intermediate results when writing a program, just to see whether I am still on track. Once I checked, I remove them, because they clutter the looks of the code.)

Python Code

```python
import numpy as np
from scipy.stats import expon

np.random.seed(3)

labdas = np.array([3, 4])

N = 10

T1 = expon(scale=1 / labdas[0]).rvs(N)
# print(T1.mean())
T2 = expon(scale=1 / labdas[1]).rvs(N)
T = np.zeros([2, N])
T[0, :] = T1
T[1, :] = T2
# print(T)
server = np.argmin(T, axis=0)
# print(server)

# BH.8.36.b
print((server == 1).mean())

# BH.8.36.c
```

```
# print(labdas[server])
S = expon(scale=1).rvs(N) / labdas[server]
# print(S)

D = np.min(T, axis=0) + S
# print(D)

print(D.mean(), D.std())
```

R Code

```r
set.seed(3)

labdas = 3:4

N = 10

T1 = rexp(N, rate = labdas[1])
T2 = rexp(N, rate = labdas[2])


bigT = matrix(0, nrow = 2, ncol = N)
bigT[1,] = T1
bigT[2,] = T2

server = rep(0, ncol(bigT))

for (i in 1:ncol(bigT)) {
  server[i] = which.min(bigT[,i])
}

print(mean((server == 1)))

S = rexp(N, rate = 1) / labdas[server]
print(S)

D = apply(bigT, MARGIN = 2, FUN = min) + S

print(mean(D))
print(sd(D))
```

Ex 4.21. Explain why `T1` corresponds to a number of service times of the first server.

Ex 4.22. To what do the rows of `T` (`bigT`) correspond?

Ex 4.23. What is the content of `server`? Why do we compute this?

Ex 4.24. Explain how we use the fundamental bridge in line P.21 (R.21) to answer BH.8.36.b.

Ex 4.25. Alice is taken into service by the server that finishes first. We need to simulate the service time that Alice needs at that server. Explain how we do this in line P.25 (R.23). Hint, reread BH.5.5 on the exponential. BTW, this is a good time to reread BH.5.3.

Ex 4.26. Why is `np.min(T, axis=0)` in the python code (`apply(bigT, MARGIN=2, FUN=min)` in R) the time Alice spends waiting in queue, i.e., the time Alice spends in the post office before her service starts?

Ex 4.27. Why is `D` the departure time of Alice, i.e., the time Alice spends in the post office?

Ex 4.28. Set `N` to 1000 or so, or any other large number to your liking, but not so large that your computer will keep simulating for a month... Compare the values of the simulation to the theoretical result that you have to compute in BH.8.36.c.

Ex 4.29. Run the code for $\lambda_1 = 1$ fixed, and take λ_2 equal to $0.5, 1, 1.5$ and 2 successively. Compute the mean time waiting time and mean sojourn time of Alice, and put your results in a table. Compare the results of the simulation to the theoretical values.

There is an important lesson to learn here. With simulation it is, often, reasonably simple to get numerical answers, but it requires many simulations to see a pattern in the numbers. For patterns we can better use theory, as theory gives us formulas that show how the output of some model depends on the input.

4.4 *Challenges*

In this exercise, we discuss Benford's law. Recall that the first step towards this law was taken in Lecture 5, Exercise 3. In this exercise, we showed that if X, Y are i.i.d. uniform on $[1, 10)$, that then the density of $Z = XY$ is given by

$$f_Z(z) = \frac{\log(\min\{10, z\}) - \log(\max\{1, z/10\})}{81} I_{1 \leq z \leq 100}.$$

Note that log denotes the natural logarithm.
Benford's law is a statement about the distribution of the first digit of the product of sufficiently many variables that are i.i.d. uniform on $[1, 10)$. We first consider the first digit of the product of two such variables, i.e. the first digit of Z.

Ex 4.30. Let K be the first digit of Z. Show that the PMF of K is given by

$$P(K = k) = \frac{9k\log(k) - 9(k+1)\log(k+1) + 9 + 10\log(10)}{81}$$

for $k \in \{1, 2, \ldots, 9\}$.

Ex 4.31. Check that $\sum_{k=1}^{9} P(K=k)=1$. This can be done nicely by recognizing a *telescoping sum*: many terms cancel because they appear once with a minus and once with a plus.

Another way to derive the first digit of Z is to first divide Z by 10 if $Z \geq 10$. This yields a random variable W with support $[1,10)$. Clearly, the division doesn't affect the first digit. The next exercise asks to derive the resulting density. This can be a bit tricky; you should check your answer by verifying that the distribution of the first digit of W matches the distribution of the first digit of Z.

Ex 4.32. Let $W = Z$ if $1 \leq Z < 10$ and $W = \frac{Z}{10}$ if $10 \leq Z < 100$. Derive the density of W.

We now turn to the product of more than two (independent) random variables. It would be very tedious to do this analytically, so we will instead use some code. However, to do this we have to approximate the continuous uniform variable by a discrete random variable. We use the discrete uniform distribution on $\{1+0.5\cdot\frac{9}{s}, 1+1.5\cdot\frac{9}{s}, 1+2.5\cdot\frac{9}{s}, \ldots, 1+(s-0.5)\cdot\frac{9}{s}\}$; in total this set has s elements. However, a product of two elements from this set may not again be an element of this set. To solve this, we identify all elements of the interval $\left(1+k\cdot\frac{9}{s}, 1+(k+1)\cdot\frac{9}{s}\right)$ with $1+(k+0.5)\cdot\frac{9}{s}$. We now use a loop to approximate the distribution of the product of $p+1$ random variables by looking at all possible values of the product of p random variables and one additional uniformly distributed random variable. Note that in the code, s is called `steps` and p is called `p_idx`.

Executing the code may take a while. If it takes more than 1 minute, you may decrease `steps`, but please do note that you did so.

Python Code

```python
import math

steps = 900
products = 15
p_unif = [1.0/steps] * steps
p_mat = [p_unif]

for p_idx in range(1, products):
    p_vec = [0] * steps
        for s1 in range(steps):
            for s2 in range(steps):
                product = (1 + (s1 + 0.5)*9/steps) * (1 + (s2 + 0.5)*9/steps)
                prod_probability = p_mat[p_idx - 1][s1] * 1/steps

                if product > 10:
                    product = product/10

                prod_idx = math.floor((product-1)/9 * steps)
                p_vec[prod_idx] += prod_probability
```

```python

    p_mat.append(p_vec)


p_digits = []
for p_idx in range(products):
    vec = []
    for digit in range(1, 10):
        pd = sum(p_mat[p_idx][((digit-1)*steps//9):(digit*steps//9)])
        vec.append(round(pd, 6))
    p_digits.append(vec)

print(p_digits)
```

R Code

```r
steps <- 900
products <- 15
p_unif <- rep(1/steps, steps)
p_mat <- matrix(0, nrow = steps, ncol = products)
p_mat[, 1] <- p_unif

for (p_idx in 2:products) {
  p_vec <- rep(0, steps)
  for (s1 in 1:steps) {
    for (s2 in 1:steps) {
      product <- (1 + (s1 - 0.5)*9/steps) * (1 + (s2 - 0.5)*9/steps)
      prod_probability <- p_mat[s1, p_idx - 1] *  1/steps

      if (product > 10) {
        product <- product/10
      }

      prod_idx <- ceiling((product-1)/9 * steps)
      p_vec[prod_idx] <- p_vec[prod_idx] + prod_probability
    }
  }
  p_mat[, p_idx] = p_vec
}

p_digits <- matrix(0, nrow = 9, ncol = products)
for (p_idx in 1:products) {
  for (digit in 1:9) {
    pd = sum(p_mat[((digit-1)*(steps/9)+1):(digit*(steps/9)), p_idx])
```

```
29     p_digits[digit, p_idx] = round(pd, 6)
30   }
31 }
32 p_digits
```

Ex 4.33. Explain line P.13 (R.12) of the code.

Ex 4.34. Briefly comment on the results for $p = 2$ compared the exact result derived in the first exercise. Why is it important to make this comparison?

When looking at the results for larger p, it seems that the probabilities converge. The limit random variable B then satisfies the property that the first digit of B and the first digit of BU (where $U \sim \text{Unif}(1,10)$) are identically distributed. Proving this is quite challenging (even for the challenge). In addition, we first need to know what the distribution of B is.

To guess the distribution of the first digit of B, we look at the results of our code and try some transformations to see if this yields familiar numbers. It turns out that the first digit M of B has the following distribution:

$$P(M = k) = \log_{10}\left(\frac{k+1}{k}\right),$$

for $k \in \{1, 2, \ldots, 9\}$.

Ex 4.35. Briefly comment on these exact values of $P(M = k)$ compared to the values for $p = 15$ that result from the code. Give two reasons why the code results are not exact. Which reason do you think is the most important?

Besides the theoretical aspects covered in this challenge, Benford's law states that the first digit of numbers of naturally occurring sets that span several orders of magnitude, such as vote counts by county (or municipality), transaction sizes, etc., approximately follow this distribution. Initially this was just seen as an interesting curiosity of no practical value, but recently it has been used in fraud detection. If you're interested, you might check out this YouTube video by Numberphile: https://www.youtube.com/watch?v=XXjlR2OK1kM&ab_channel=Numberphile

5 ASSIGNMENT 5

5.1 *Have you read well?*

Ex 5.1. Compute the expected outcome of a die throw (with a 6-sided die), given that the outcome is even. Introduce proper notation for random variables and events.

Ex 5.2. Consider a casino where, for any $a > 0$, it is possible to pay a euro and get a chance of $\frac{1}{5}$ on receiving $4a$ euro and a chance of $\frac{4}{5}$ of receiving nothing. Adam enters the casino with b euros, and bets half of his money on this gamble. Let X be the amount of money he has after the gamble. After that, he again bets half of the money he then has (i.e. half of X) on this gamble. Let Y be the amount of money he has after the second gamble.

1. Compute $\mathsf{E}[X]$.
2. Compute $\mathsf{E}[Y|X]$.
3. Compute $\mathsf{E}[Y]$.

Explicitly mention the laws/rules you use.

Ex 5.3. Let $N \sim \text{Pois}(\lambda)$, and let $X|N \sim \text{Bin}(N, p)$, where $p \in (0,1)$ and $\lambda > 0$ are known constants. Compute $\mathsf{E}[X]$ using Adam's law. Check your answer using the chicken-egg story; with this story you can also obtain the distribution of X.

Ex 5.4. Correct? If A is an event and I_A is its indicator, then for all random variables X we have $\mathsf{E}[X \mid A] = \mathsf{E}[X \mid I_A]$.

Ex 5.5. Correct? If X and Y are independent, then $\mathsf{V}[\mathsf{E}[Y \mid X]] = 0$.

Ex 5.6. Let $X \sim \text{Exp}(\lambda)$, and let a be a constant.

1. Compute $\mathsf{E}[X \mid X \geq a]$ using an integral and an indicator.
2. Explain the answer using a property of the exponential distribution.

Ex 5.7. A hat contains 9 fair coins and one coin that lands heads with probability 0.8. You pick a coin from the hat uniformly at random and toss it 10 times. Let A be the event that you pick a fair coin, and let X be the number of heads. Let B be the event that the first four tosses all show heads.

1. Compute $\mathsf{E}[X \mid A]$.
2. Compute $\mathsf{E}[X \mid A^c]$.
3. Compute $\mathsf{E}[X]$.
4. Compute $\mathsf{P}\{B\}$.

5. Compute $\mathsf{P}\{A \mid B\}$.

6. Compute $\mathsf{E}[X \mid B]$.

7. Compute $\mathsf{E}[X \mid B^c]$. *Hint*: it is not necessary to compute $\mathsf{P}\{A \mid B^c\}$.

Ex 5.8. Consider random variables $X, Y \in [0,1]^2$ with joint PDF $f_{X,Y}(x,y) = 2 I_{x \leq y}$. Determine $\mathsf{E}[Y \mid X]$ and $\mathsf{E}[X \mid Y]$.

Ex 5.9. Prove that $\mathsf{E}[X \mid X \geq a] > \mathsf{E}[X]$ for any a with $0 < \mathsf{P}\{X \geq a\} < 1$.

Ex 5.10. Let $N \sim \text{Pois}(\lambda)$ and let $X|N \sim \text{Bin}(N, p)$, where $p \in (0,1)$ and $\lambda > 0$ are known constants. Find $\mathsf{E}[N \mid X]$.

5.2 *Exercises at about exam level*

We derive Eve's law in a slightly different way then in BH.

Define $\hat{X} = \mathsf{E}[X \mid Y]$ as an *estimator* of X and $\tilde{X} = X - \hat{X}$ as the estimation error.

Ex 5.11. Show that $\mathsf{E}\left[\tilde{X} \,\middle|\, Y\right] = 0$.

Ex 5.12. Prove that $\mathsf{E}\left[\tilde{X}\right] = 0$. What does it mean that $\mathsf{E}\left[\tilde{X}\right] = 0$?

Ex 5.13. Prove that $\mathsf{E}\left[\tilde{X}\hat{X}\right] = 0$.

Ex 5.14. Show that $\mathsf{Cov}\left[\hat{X}, \tilde{X}\right] = 0$. Conclude that

$$\mathsf{V}[X] = \mathsf{V}\left[\hat{X} + \tilde{X}\right] = \mathsf{V}\left[\hat{X}\right] + \mathsf{V}\left[\tilde{X}\right]. \tag{5.1}$$

Ex 5.15. Prove that $\mathsf{V}\left[\tilde{X}\right] = \mathsf{E}[\mathsf{V}[X \mid Y]]$. Conclude Eve's law.

Ex 5.16. Prove that

$$\mathsf{E}\left[(Y - \mathsf{E}[Y|X] - h(X))^2\right] = \mathsf{E}\left[(Y - \mathsf{E}[Y|X])^2\right] + \mathsf{E}\left[(h(X))^2\right] \tag{5.2}$$

for all random variables X, Y and all functions h.

Explain why this result implies that $\mathsf{E}[Y|X]$ is the best predictor of Y based on X.

5.3 *Coding skills*

THE MYSTERY BOX We use simulation to solve BH.9.7. Read it now, i.e., before reading the text below, then read the code below. Note how short this code is; amazing, isn't it?

Python Code

```python
import numpy as np
from scipy.stats import uniform
import matplotlib.pyplot as plt
```

```python

np.random.seed(3)


N = 1000
a, b = 0, 1000_000
V = uniform(a, b).rvs(N)

x_range = np.linspace(b / 5, b / 2, num=50)
y = np.zeros_like(x_range)

for i, b in enumerate(x_range):
    payoff = (V - b) * (b >= V / 4)
    y[i] = payoff.mean()


plt.plot(x_range, y)
plt.show()
```

R Code

```r
set.seed(3)

N = 1000
a = 0
b = 1000000
V = runif(N, min = a, max = b)

x_range = seq(b / 5, b / 2, length.out = 50)
y = rep(0, length(x_range))


i = 1
for (b in x_range) {
  payoff = (V - b) * (b >= V / 4)
  y[i] = mean(payoff)
  i = i + 1
}

plot(x_range, y, type = "l", col = "blue")
```

Ex 5.17. For the python code use the scipy documentation to explain why $V \sim \mathrm{Unif}([0, 10^6])$. For R, explain the save fo `runif`.

Ex 5.18. What are the smallest and the largest value of `x_range`?

Ex 5.19. Run the code above and make a graph. Include the graph in your report, and explain what you see in the graph. For instance, is there a maximum? If so, can you explain where the maximum occurs? Can you explain how the maximum should be?

Ex 5.20. Suppose after seeing the graph of the payoffs, and this graph would only increase, or decrease, how would you change `x_range`? Do you expect to see a maximum?

Ex 5.21. For N small, e.g. N=10, you can get quite strange values. Why is that?

Ex 5.22. Change the acceptance threshold from to $V/4$ to $V/5$ (or $V/6$, or some other value you like), and make a graph of the payoffs. Include the graph in your report.

Ex 5.23. Change the payoff function to e.g $\sqrt{V-b}$, or some weird function that you like particularly such as $\sin|V-b|$ (any non-trivial function goes). Make a graph of the mean and std of the payoff. Can you explain your graph?

KELLY MAKES BETS This simulation exercise is based on BH.9.25. Please read the exercise first, and then the code below.

Python Code

```python
import numpy as np
from scipy.stats import bernoulli

np.random.seed(3)

n = 5
num = 10

p = 0.5
S = bernoulli(p).rvs([num, n]) * 2 - 1
# print(S)

x = np.zeros([num, n])
x[:, 0] = 100
# print(x)
f = 0.25

for i in range(1, n):
    x[:, i] = x[:, i - 1] * (1 + f * S[:, i])

# print(x)
print(x.mean(axis=0), x.std(axis=0))
```

R Code

```
1 set.seed(3)
2
3 n = 5
4 num = 10
5
6 p = 0.5
7 S = matrix(rbinom(n * num, size = 1, prob = p), num, n) * 2 - 1
8
9 x = matrix(0, num, n)
10 x[, 1] = 100
11 f = 0.25
12
13 for (i in 2:n) {
14   x[, i] = x[, i - 1] * (1 + f * S[, i])
15 }
16
17 print(colMeans(x))
18 print(apply(x, MARGIN = 2, sd))
```

Ex 5.24. The documentation of `bernoulli` (`rbinom`) tells us that we get an array (matrix) with `num` rows and n columns with 0 and 1s. Explain that by multiplying with 2 and subtracting 1 we get an array with 1s and -1s.

Ex 5.25. The jth columns of `X` corresponds to the jth bet. Explain how line P.19 (r.14) works.

Ex 5.26. Choose some p and f to your liking, run an experiment, and compare the values of the experiment(simulation) to the theoretical values, i.e., what you get when you solve problem BH.9.25.

5.4 *Challenges*

Consider the setting of BH Exercise 9.25, which you also studied in the coding section. We use the notation from that exercise. In this exercise we will discuss how to set f, the betting fraction. In particular, we will discuss the *Kelly criterion*, which states that the betting fraction should be $f = 2p-1$ if $p > \frac{1}{2}$ is the winning probability.

We discuss its relationship to expected utility theory, which you will also study in Introduction to Mathematical Economics. Expected utility theory states that bets should be chosen to maximize expected utility. So we solve $\max_{0\leq f\leq 1} \mathsf{E}[U(X_{n+1})|X_n]$.

Ex 5.27. Show that solving the maximization problem for the utility function $U(x) = \log(x)$ yields the betting fractions from the Kelly criterion, $f = 2p-1$ if $p > \frac{1}{2}$ and $f = 0$ if $p \leq \frac{1}{2}$.

Other people may have a different utility function, which yields a different betting fraction.

Ex 5.28. Calculate the utility maximizing betting fraction f if $U(x) = \sqrt{x}$.

Note that for both of these utility functions, the betting fraction f does not depend on the wealth X_n before the gamble, but in general f does depend on X_n.

Now that we have two different betting fractions, we compare them. For that, we first need the following result:

Ex 5.29. Assume that f does not depend on X_n. Let $x_0 = 1$. Show that there exist constants a, b such that $\log(X_n) = aW + b$ and $W \sim \text{Bin}(n, p)$, and determine a and b in terms of f.

Theorem 10.3.6. states that (for sufficiently large n) we can approximate a random variable with the binomial distribution $W \sim \text{Bin}(n, p)$ by a random variable with the normal distribution $\text{Norm}(np, np(1-p))$. While you will only learn about the proof of this next week, we are already going to use this approximation here.

Ex 5.30. Two people (Carl and Daria) participate in n rounds of this betting game. Their games are independent. Carl's initial wealth is $x_0 = 1$ and Daria's initial wealth is $y_0 = 1$. We denote Carl's wealth after n rounds by X_n and Daria's wealth after n rounds by Y_n. Carl chooses f according to the Kelly criterion, i.e. $f = 2p - 1$. Daria chooses f to be the utility maximizing betting fraction for $U(x) = \sqrt{x}$. Use the previously mentioned normal approximation to derive an approximation for the difference $\log(Y_n) - \log(X_n)$.

Kelly's criterion does not mention utility functions, it just recommends to set $f = 2p - 1$ regardless of one's utility function. The next exercise is meant to give some insight why.

Ex 5.31. Use `pnorm` in R, or `norm.cdf` in Python, to approximate $P(X_n > Y_n)$ for some chosen values for n and p. What do you think that happens if $n \to \infty$ for a fixed p? Explain why this is an argument to use the Kelly criterion regardless of one's utility function. Also, explain why maximizing utility suggests a different f in spite of this result.

6 HINTS

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

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

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

$$P(N = n+1) = \frac{\lambda}{1+n} P(N = n). \tag{6.1}$$

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

$$\lambda = \frac{\mathsf{P}\{Y = 1\}}{(1-p)\mathsf{P}\{Y = 0\}}. \tag{6.2}$$

h.3.9. Let X, Y be i.i.d. standard normal. Since the square of a standard normal r.v. is chi-square distributed, we can write S as $S = X^2 + Y^2$ (here we use BH.8.1.4).

h.3.13. What is the domain of V on each of the intervals $(-3, 0)$ and $[0, 2)$? For the final part, combining the results into one PDF: Use LOTP, conditioning on $U \geq 0$.

h.5.10. For a smart argument, use the chicken-egg story. Recall that the number of hatched eggs and the number of unhatched eggs are independent (since $N \sim \text{Pois}(\lambda)$); i.e. $N - X$ and X are independent.

h.5.12. Use [5.11]

h.5.13. Use [5.12] and the definitions.

7 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.

$$f_X(x)=\int_0^1 f_{X,Y}(x,y)\,\mathrm{d}y = 2\int_0^1 I_{x\leq y}\,\mathrm{d}y = 2\int_x^1 \mathrm{d}y = 2(1-x) \tag{7.1}$$

$$f_Y(y)=\int_0^1 f_{X,Y}(x,y)\,\mathrm{d}x = 2\int_0^1 I_{x\leq y}\,\mathrm{d}x = 2\int_0^y \mathrm{d}y = 2y. \tag{7.2}$$

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

$$\begin{aligned}
F_{X,Y}(x,y) &= \int_0^x\int_0^y f_{X,Y}(u,v)\,\mathrm{d}v\,\mathrm{d}u && (7.3)\\
&= 2\int_0^x\int_0^y I_{u\leq v}\,\mathrm{d}v\,\mathrm{d}u && (7.4)\\
&= 2\int_0^x\int I_{u\leq v}\,I_{0\leq v\leq y}\,\mathrm{d}v\,\mathrm{d}u && (7.5)\\
&= 2\int_0^x\int I_{u\leq v\leq y}\,\mathrm{d}v\,\mathrm{d}u && (7.6)\\
&= 2\int_0^x [y-u]^+\,\mathrm{d}u, && (7.7)
\end{aligned}$$

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

$$2\int_0^x [y-u]^+\,\mathrm{d}u = 2\int_0^x (y-u)\,\mathrm{d}u = 2yx-x^2, \tag{7.8}$$

while if $y\leq x$,

$$2\int_0^x [y-u]^+\,\mathrm{d}u = 2\int_0^y (y-u)\,\mathrm{d}u = 2y^2-y^2=y^2 \tag{7.9}$$

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.

$$\frac{F_{X,Y}(x,y)}{F_X(x)} = \frac{\mathsf{P}\{X \le x, Y \le y\}}{\mathsf{P}\{X \le x\}} \tag{7.10}$$

In the notes we define the conditional CDF as the function $F_{X|Y}(x|y) = \mathsf{P}\{X \le x | Y = y\}$. 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. $X \le x$. 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 $\{Y \le y\}$. 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 ith 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.

$$\mathsf{P}\{I_0 <= 0\} = \mathsf{P}\{2 + X_1 - C_1 < 0\} = \frac{1}{10}\sum_{i=1}^{10} \mathsf{P}\{C_1 > 2 + i\} = \frac{1}{10}\sum_{i=1}^{5} \mathsf{P}\{C_1 > 2 + i\} \tag{7.11}$$

$$= \frac{1}{10}\sum_{i=1}^{5} \frac{6-i}{9}. \tag{7.12}$$

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 Code

```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
```

```python
            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.2.1. We have

$$\mathsf{E}\left[(X-Y)^2\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-y)^2 f_{X,Y}(x,y)\,dx\,dy \tag{7.13}$$
$$= \int_0^1\int_0^1 (x-y)^2\,dx\,dy \tag{7.14}$$

s.2.2. Any situation in which the units of measurement might be distracting. Correlation is usually easier to interpret.

s.2.3. Answers: no and yes.

We have

$$C = \frac{\mathsf{V}[X]}{(\mathsf{E}[X])^2}, \tag{7.15}$$

which does not equal

$$\mathrm{Corr}(X,X) = \frac{\mathsf{Cov}[X,X]}{\sqrt{\mathsf{V}[X]\,\mathsf{V}[X]}} = 1 \tag{7.16}$$

in general (for instance, consider a degenerate random variable $X \equiv 1$). Next, consider a $N(1,100)$ random variable. Then,

$$C = 100/(1^2) = 100 > 1. \tag{7.17}$$

s.2.4. We have

$$\begin{aligned}
\mathsf{Cov}[X,Y] &= \mathsf{E}[(X-\mathsf{E}[X])(Y-\mathsf{E}[Y])] && (7.18)\\
&= \mathsf{E}[XY - X\,\mathsf{E}[Y] - Y\,\mathsf{E}[X] + \mathsf{E}[X]\,\mathsf{E}[Y]] && (7.19)\\
&= \mathsf{E}[XY] - \mathsf{E}[X]\,\mathsf{E}[Y] - \mathsf{E}[Y]\,\mathsf{E}[X] + \mathsf{E}[X]\,\mathsf{E}[Y] && (7.20)\\
&= \mathsf{E}[XY] - \mathsf{E}[X]\,\mathsf{E}[Y]. && (7.21)
\end{aligned}$$

s.2.5. By linearity of the covariance wea have

$$\begin{aligned}
\mathsf{Cov}[a(U+V), b(U-V)] &= a\Big(\mathsf{Cov}[U, b(U-V)] + \mathsf{Cov}[V, b(U-V)]\Big) && (7.22)\\
&= a\Big(b\big(\mathsf{Cov}[U,U] - \mathsf{Cov}[U,V]\big) + b\big(\mathsf{Cov}[V,U] - \mathsf{Cov}[V,V]\Big) && (7.23)\\
&= a\Big(b\big(\mathsf{Cov}[U,U] - \mathsf{Cov}[U,V]\big) + b\big(\mathsf{Cov}[V,U] - \mathsf{Cov}[V,V]\Big) && (7.24)\\
&= ab\Big(\mathsf{V}[U] - \mathsf{Cov}[U,V] + \mathsf{Cov}[V,U] - \mathsf{V}[V]\Big) && (7.25)\\
&= ab\Big(\mathsf{V}[U] - \mathsf{V}[V]\Big). && (7.26)
\end{aligned}$$

s.2.6. 1. We have

$$\mathsf{Cov}[X,X] = \mathsf{E}[XX] - \mathsf{E}[X]\,\mathsf{E}[X] = \mathsf{E}\left[X^2\right] - \mathsf{E}[X]^2 = \mathsf{V}[X]. \tag{7.27}$$

2. We have

$$\mathsf{Cov}[X,Y] = \mathsf{E}[XY] - \mathsf{E}[X]\,\mathsf{E}[Y] = \mathsf{E}[YX] - \mathsf{E}[Y]\,\mathsf{E}[X] = \mathsf{Cov}[Y,X]. \tag{7.28}$$

3. We have

$$\mathsf{Cov}[X,c] = \mathsf{E}[Xc] - \mathsf{E}[X]\,\mathsf{E}[c] = c\,\mathsf{E}[X] - c\,\mathsf{E}[X] = 0. \tag{7.29}$$

4. We have

$$\mathsf{Cov}[aX,Y] = \mathsf{E}[aXY] - \mathsf{E}[aX]\,\mathsf{E}[Y] = a\big(\mathsf{E}[XY] - \mathsf{E}[X]\,\mathsf{E}[Y]\big) = a\,\mathsf{Cov}[X,Y]. \tag{7.30}$$

5. We have

$$\begin{aligned}\mathsf{Cov}[X+Y,Z] &= \mathsf{E}[(X+Y)Z] - \mathsf{E}[X+Y]\,\mathsf{E}[Z] && (7.31)\\ &= \mathsf{E}[XZ+YZ] - \big(\mathsf{E}[X]+\mathsf{E}[Y]\big)\mathsf{E}[Z] && (7.32)\\ &= \mathsf{E}[XZ] - \mathsf{E}[X]\,\mathsf{E}[Z] + \mathsf{E}[YZ] - \mathsf{E}[Y]\,\mathsf{E}[Z] && (7.33)\\ &= \mathsf{Cov}[X,Z] + \mathsf{Cov}[Y,Z]. && (7.34)\end{aligned}$$

s.2.7. We throw 10 fair dice. X_i denotes the number of dice that show the number i, $i = 1, \dots, 6$.

s.2.8. No, this does not always hold. It does hold when X and Y are independent, though.

s.2.9. In hindsight, this question was more an exam-level question.

1. Since (X,Y) are bivariate normally distributed, every linear combination of X and Y is normally distributed. Note that every linear combination of $(X+Y)$ and $(X-Y)$ can be written as a linear combination of X and Y. Hence, every linear combination of $(X+Y)$ and $(X-Y)$ is normally distributed. Hence, $(X+Y, X-Y)$ is bivariate normally distributed.

2. By the story above, both X and Y are normally distributed. We have

$$\mathsf{E}[X+Y] = \mathsf{E}[X] + \mathsf{E}[Y] = \mu + \mu = 2\mu, \tag{7.35}$$

and

$$\mathsf{E}[X-Y] = \mathsf{E}[X] - \mathsf{E}[Y] = \mu - \mu = 0. \tag{7.36}$$

Moreover,

$$\mathsf{V}[X+Y] = \mathsf{V}[X] + \mathsf{V}[Y] + 2\,\mathsf{Cov}[X,Y] = 2\sigma^2 + 2\rho\sigma^2 = 2(1+\rho)\sigma^2. \tag{7.37}$$

Simlarly,

$$\begin{aligned}\mathsf{V}[X-Y] &= \mathsf{V}[X] + \mathsf{V}[-Y] + 2\,\mathsf{Cov}[X,-Y] = \mathsf{V}[X] + \mathsf{V}[Y] - 2\,\mathsf{Cov}[X,Y] && (7.38)\\ &= 2\sigma^2 - 2\rho\sigma^2 = 2(1-\rho)\sigma^2. && (7.39)\end{aligned}$$

So we have found that $X+Y \sim N(2\mu, 2(1+\rho)\sigma^2)$ and $X-Y \sim N(0, 2(1-\rho)\sigma^2$.

3. We have

$$\begin{aligned}\mathsf{Cov}[X+Y,X-Y] &= \mathsf{Cov}[X,X] - \mathsf{Cov}[X,Y] + \mathsf{Cov}[Y,X] - \mathsf{Cov}[Y,Y] && (7.40)\\ &= \mathsf{V}[X] - \mathsf{V}[Y] = \sigma^2 - \sigma^2 = 0. && (7.41)\end{aligned}$$

Write $U = X+Y$, $V = X-Y$. Plugging all the parameters into the formula for the joint pdf of a bivariate normal distribution (see `https://en.wikipedia.org/wiki/Multivariate_normal_distribution#Bivariate_case`), we obtain

$$f_{U,V}(u,v) = \frac{1}{2\pi\sqrt{2(1+\rho)\sigma^2 2(1-\rho)\sigma^2}} \exp\left(-\frac{1}{2}\left[\frac{(u-2\mu)^2}{2(1+\rho)\sigma^2} + \frac{v^2}{2(1-\rho)\sigma^2}\right]\right). \tag{7.42}$$

s.2.10. Since X, Y, Z are independent normally distributed variables, (X, Y, Z) is multivariate normally distributed. Hence, every linear combination of X, Y, Z is normally distributed. Note that every linear combination of the elements of W can be written as a linear combination of X, Y, Z. Hence, every linear combination of the elements of W is normally distributed. Hence, W is multivariate normally distributed.

s.2.11. We have

$$\mathsf{Cov}[X,Y] = \mathsf{Cov}\left[X, X^2\right] = \mathsf{E}\left[XX^2\right] - \mathsf{E}[X]\,\mathsf{E}\left[X^2\right] = 0 - 0 \cdot 2.5 = 0. \tag{7.43}$$

Hence, $\mathrm{Corr}(X,Y) = 0$.
Yes, for instance, take $X \sim \mathrm{Unif}(\{0,1\})$. Then,

$$\mathsf{Cov}[X,Y] = \mathsf{E}\left[XX^2\right] - \mathsf{E}[X]\,\mathsf{E}\left[X^2\right] = 0.5 - 0.5 \cdot 0.5 = 0.25. \tag{7.44}$$

s.2.12.

1. The interpretation is: the time until the first component fails. That is, the time until the machine stops working.
2. Let $\lambda = 10$. We have

$$\begin{aligned}
\mathsf{P}\{\text{machine not failed at time } T\} &= \mathsf{P}\{\min\{X_1, X_2\} > T\} && (7.45)\\
&= \mathsf{P}\{X_1 > T, X_2 > T\} && (7.46)\\
&= \mathsf{P}\{X_1 > T\}\,\mathsf{P}\{X_2 > T\} && (7.47)\\
&= e^{-\lambda T} \cdot e^{-\lambda T} && (7.48)\\
&= e^{-(2\lambda)T} && (7.49)\\
&= e^{-20T} && (7.50)\\
& && (7.51)
\end{aligned}$$

3. Note that

$$\mathsf{P}\{\min\{X_1, X_2\} \le T\} = 1 - \mathsf{P}\{\min\{X_1, X_2\} > T\} = 1 - e^{-20T}. \tag{7.52}$$

   Note that this is the cdf of an exponential distribution with parameter 20. Hence, $\min\{X_1, X_2\} \sim \exp(20)$.

4. The expected time until the machine fails is

$$\mathsf{E}[\min\{X_1, X_2\}] = 1/20, \tag{7.53}$$

   i.e., 3 minutes. Apparently, the machine is not very robust.

s.2.13. 1. We have

$$\begin{aligned}
\mathsf{P}\{X+Y>1\} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I_{X+Y>1} f_{X,Y}(x,y)\,dy\,dx && (7.54)\\
&= \int_0^1\int_{1-x}^{1}\frac{6}{7}(x+y)^2\,dy\,dx && (7.55)\\
&= \frac{6}{7}\int_0^1\left[\frac{1}{3}(x+y)^3\right]_{y=1-x}^{1}dx && (7.56)\\
&= \frac{2}{7}\int_0^1\Big((x+1)^3-(x+1-x)^3\Big)dx && (7.57)\\
&= \frac{2}{7}\int_0^1\Big((x+1)^3-1\Big)dx && (7.58)\\
&= \frac{2}{7}\left[\frac{1}{4}(x+1)^4-x\right]_{x=0}^{1} && (7.59)\\
&= \frac{1}{14}\left[(x+1)^4-4x\right]_{x=0}^{1} && (7.60)\\
&= \frac{1}{14}\Big(\big((1+1)^4-4\big)-\big((0+1)^4-0\big)\Big) && (7.61)\\
&= \frac{1}{14}\Big(16-4-1\Big) && (7.62)\\
&= \frac{11}{14}. && (7.63)
\end{aligned}$$

2. We have

$$\mathsf{Cov}[U,V] = \mathsf{E}[UV]-\mathsf{E}[U]\,\mathsf{E}[V]. \qquad (7.64)$$

First, we compute

$$\begin{aligned}
\mathsf{E}[UV] &= \int_0^1\int_0^{1-u} 2uv\,dv\,du && (7.65)\\
&= \int_0^1 [uv^2]_{v=0}^{1-u}\,du && (7.66)\\
&= \int_0^1 \big(u(1-u)^2-0\big)du && (7.67)\\
&= \int_0^1 u(1-2u+u^2)\,du && (7.68)\\
&= \int_0^1 (u-2u^2+u^3)\,du && (7.69)\\
&= \left[\frac{1}{2}u^2-\frac{2}{3}u^3+\frac{1}{4}u^4\right]_{u=0}^{1} && (7.70)\\
&= \frac{1}{2}-\frac{2}{3}+\frac{1}{4} && (7.71)\\
&= \frac{1}{12}. && (7.72)
\end{aligned}$$

Next,

$$\begin{aligned}
\mathsf{E}[U] &= \int_0^1 \int_0^{1-u} 2u\,dv\,du && (7.73)\\
&= \int_0^1 2u \int_0^{1-u} 1\,dv\,du && (7.74)\\
&= \int_0^1 2u(1-u)\,du && (7.75)\\
&= 2\int_0^1 (u-u^2)\,du && (7.76)\\
&= 2\Big[\frac{1}{2}u^2 - \frac{1}{3}u^3\Big]_{u=0}^1 && (7.77)\\
&= 2\Big(\frac{1}{2} - \frac{1}{3}\Big) && (7.78)\\
&= \frac{1}{3} && (7.79)
\end{aligned}$$

By symmetry, $\mathsf{E}[V] = \frac{1}{3}$. Hence,

$$\begin{aligned}
\mathsf{Cov}[U,V] &= \mathsf{E}[UV] - \mathsf{E}[U]\,\mathsf{E}[V] && (7.80)\\
&= \frac{1}{12} - \frac{1}{3}\frac{1}{3} && (7.81)\\
&= \frac{1}{12} - \frac{1}{9} && (7.82)\\
&= -\frac{1}{36}. && (7.83)
\end{aligned}$$

s.2.14. X is a matrix of i.i.d. draws from an exponential distribution with parameter λ.

s.2.15. T is a sorted version of X, where we sort each row increasingly.

s.2.16. We print the mean value of each column of T.

s.2.17. This is an array with expected values of the ith order statistic $X_{(i)}$ (see B.H.5.6.5 afor a proof of this result).

s.2.18. The cumsum is the cumulative sum up to and including the current index. So the final entry indicates the expected value of the sum of all three entries of `expected`.

s.2.19. The result of `print(T.mean(axis=0))` should be close to that of `print(expected.cumsum())`.

s.2.20. It iterates through the elements of X and checks how often the current value is larger than any of the previous values.

s.2.21. We draw a sample of a $U[0,1]$ distribution of size `num` and compute the corresponding number of maxima (or “records”).

s.2.22. This are the arrival times of 5 passengers within the time interval of 3 minutes (sorted increasingly).

s.2.23. `A[1:]` is an array of all elements of `A` except the first one.

s.2.24. `A[:-1]` is an array of all elements of `A` except the last one.

s.2.25. `X` consists of the interarrival times.

s.2.26. $1/\lambda$ is the expected interarrival time. `X.mean()` is the sample average of the interarrival times.

s.2.27. For `X.std()` we expect to see $1/\lambda = 0.6$ too (if X is indeed exponentially distributed with parameter λ).

s.2.28. For a sample of size 50, we expect an average interarrival time of 0.06 and an equal standard deviation if the distribution of the interarrival times is indeed exponential. We indeed observe a sample mean and sample average that are very close to this value.

s.2.29. If d is deterministic and known then $x^* = d$, since cost in increasing in x.

s.2.30. We know that at least 10 muffins will be needed so we stock at least 10 muffins. If we stock less, then we know for sure that we need to bake an additional muffin on the spot which increases cost. Also, we never need more than 20 muffins. So $10 \le x^* \le 20$. Note that

$$\begin{aligned} E[v(d,x)] &= E\left[q(d-x)^+\right] = q\int_{10}^{20}(d-x)^+\tfrac{1}{10}\mathrm{d}d = q\int_x^{20}(d-x)\tfrac{1}{10}\mathrm{d}d \\ &= \tfrac{q}{10}\left[\tfrac{1}{2}(d-x)^2\right]_x^{20} = \tfrac{q}{20}(20-x)^2. \end{aligned}$$

Hence, $cx + E[v(d,x)] = cx + \frac{q}{20}(20-x)^2 = x + 0.075(20-x)^2 = 30 - 2x + 0.075x^2$.
Setting the derivative $-2 + 0.15x$ to 0 yields $x^* = \frac{2}{0.15} = \frac{40}{3}$.

s.2.31. $E[v(d,x^*)] = \frac{q}{20}(20-x^*)^2 = \frac{\$1.5}{20}\left(\frac{20}{3}\right)^2 = \$\frac{10}{3}$.

s.2.32. The sum $d_m + d_f$ is again normally distributed with mean $\mathsf{E}\left[d_m + d_f\right] = \mathsf{E}[d_m] + \mathsf{E}\left[d_f\right] = 20$ and variance $\mathsf{V}\left[d_m + d_f\right] = \mathsf{V}[d_m] + 2\,\mathsf{Cov}\left[d_m, d_f\right] + \mathsf{V}\left[d_f\right] = 25(2+2\rho) = 0$.
So actually, demand is deterministic; so $x^* = 20$, with cost $cx^* + 0 = \$20$.

s.2.33. For $\rho = -1$, the expected total cost will be lower, since in this case d is actually deterministic and hence in the optimal policy, all muffins will be produced at cost \$1.
For $\rho = 1$, the expected number of muffins is still the same but either some muffins will be wasted or some muffins will be produced at cost \$1.5 instead.

s.3.1.

$$X \in \{0,\dots,5\} \implies Z \in \{0,3,6,9,12,15\}, \quad \text{and not in } \{0,\dots,15\}, \tag{7.84}$$
$$z = g(x) = 3x, \tag{7.85}$$
$$p_Z(z) = \sum_{x:g(x)=z} p_X(x) = \frac{1}{6} I_{z\in\{0,3,6,9,12,15\}}, \tag{7.86}$$
$$F_Z(z) = \frac{1}{6}\sum_{x=0}^{z} I_{x\in\{0,3,6,9,12,15\}}. \tag{7.87}$$

s.3.2.

$$X \in [0,5] \implies Z \in [0,15], \tag{7.88}$$
$$z = 3x = g(x) \implies x = z/3, \tag{7.89}$$
$$f_Z(z) = f_X(x)\frac{\mathrm{d}x}{\mathrm{d}z}, \tag{7.90}$$
$$\frac{\mathrm{d}z}{\mathrm{d}x} = 3, \tag{7.91}$$
$$f_Z(z) = f_X(z/3)\frac{1}{3}. \tag{7.92}$$

$F_Z(u) = 1$ for $u \geq 15$ and $F_Z(u) = 0$ for $u \leq 0$. When $0 \leq u \leq 15$,

$$F_Z(u) = \int_0^u f_X(z/3)\frac{1}{3}\,\mathrm{d}z = \frac{1}{5}\int_0^u I_{0\leq z/3\leq 5}\frac{1}{3}\,\mathrm{d}z \tag{7.93}$$
$$= \frac{1}{5}\int_0^u I_{0\leq z\leq 15}\frac{1}{3}\,\mathrm{d}z = \frac{u}{15}. \tag{7.94}$$

s.3.3.

$$z = g(x) = (x-\mu)/\sigma, \implies x = \sigma z + \mu \tag{7.95}$$
$$f_Z(z) = f_X(x)\frac{\mathrm{d}x}{\mathrm{d}z}, \tag{7.96}$$
$$\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{1}{\sigma}, \tag{7.97}$$
$$f_Z(z) = f_X(x)\sigma = \sigma f_X(\sigma z + \mu) \tag{7.98}$$

and now using the density of $X \sim \text{Norm}(\mu,\sigma)$,

$$f_Z(z) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(\sigma z+\mu-\mu)^2/2\sigma^2}\sigma = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}. \tag{7.99}$$

s.3.4.

$$z = g(x) = e^{-x} \implies x = -\log z, \tag{7.100}$$
$$x \in (0,\infty) \implies z \in (0,1), \tag{7.101}$$
$$f_Z(z) = f_X(x)\frac{\mathrm{d}x}{\mathrm{d}z}, \tag{7.102}$$
$$\frac{\mathrm{d}z}{\mathrm{d}x} = -e^{-x}, \quad \text{Don't forget to take the abs value next,} \tag{7.103}$$
$$f_Z(z) = f_X(x)e^{x} = e^{-x}e^{x} = 1 I_{0<z<1}, \tag{7.104}$$

where we include the domain of Z in the last equality.

s.3.5.

$$(u,v) = (x+y, x-y) = g(x,y) \implies (x,y) = ((u+v)/2, (u-v)/2), \tag{7.105}$$

$$\frac{\partial(u,v)}{\partial(x,y)} = \begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} = -2 \implies |-2| = 2, \tag{7.106}$$

$$f_{U,V}(u,v) = f_{X,Y}(x,y)\frac{\partial(x,y)}{\partial(u,v)} = f_{X,Y}((u+v)/2, (u-v)/2)/2 \tag{7.107}$$

$$= \frac{1}{4\pi} e^{-((u+v)/2)^2/2} e^{-((u-v)/2)^2/2} \tag{7.108}$$

$$= \frac{1}{4\pi} e^{-u^2/4 - v^2/4}, \tag{7.109}$$

where we work out the squares and simplify. Hence, U and V are independent and normally distributed with mean 0 and $\sigma = \sqrt{2}$. This is in line with our earlier definition of a multi-variate normal distribution.

s.3.6.

1. $Z = Y^4 \in [0,\infty)$ for $Y \in (-\infty,\infty)$;
2. $Y = X^3 + a \in (a, a+1)$ for $X \in (0,1)$;
3. $U = |V| + b \in [b,\infty)$ for $V \in (-\infty,\infty)$;
4. $Y = e^{X^3} \in (0,\infty)$ for $X \in (-\infty,\infty)$;
5. $V = U I_{U \leq c} \in (-\infty, c]$ for $U \in (-\infty,\infty)$;
6. $Y = \sin(X) \in [-1,1]$ for $X \in (-\infty,\infty)$;
7. $Y = \frac{X_1}{X_1+X_2} \in (0,1)$ for $X_1 \in (0,\infty)$ and $X_2 \in (0,\infty)$;
8. $Z = \log(UV) \in (-\infty,\infty)$ for $U \in (0,\infty)$ and $V \in (0,\infty)$.

s.3.7. If we would not add this extra variable, we cannot use the change of variables theorem. We also need a function to deal with the scaling. In the change of variables theorem, this is the Jacobian.

There is also another problem. Consider the function $g(x,y)$ that maps $\mathbb{R}^2$ to $\mathbb{R}$. The inverse set $\{(x,y) : g(x,y) = z\}$ can be quite complicated, while the set $\{y : g(x,y) = z\}$ for a fixed x is hopefully just one point. Hence, the mapping $(x,y) \to (x, g(x,y))$ is, at least locally, one-to-one.

It is possible to deal with the more general problem, but this requires much more theory than we need for this course.

s.3.8. When the variables become dependent, the Jacobian becomes zero. For instance, in the latter case,

$$\frac{\partial(u,v)}{\partial(x,y)} = \begin{vmatrix} 1/y & -x/y^2 \\ -y/x^2 & 1/x \end{vmatrix} = \frac{1}{xy} - \frac{x}{y^2}\frac{y}{x^2} = 0. \tag{7.110}$$

Moreover, the function g is not locally one-to-one.

s.3.9. From BH.8.1.4: Z chi-square $\implies X = \sqrt{Z} \sim \text{Norm}(0,1)$. Then, from BH.8.1.9,

$$X^2 + Y^2 = (\sqrt{2T}\cos U)^2 + (\sqrt{2T}\sin U)^2 = 2T\left(\cos^2 U + \sin^2 U\right) = 2T \sim \text{Exp}(1/2), \tag{7.111}$$

when $X, Y \sim \text{Norm}(0,1)$.

s.3.10. Take $g(x,y) = (x,w) = (x,(x+y)/2)$. Then, $y = 2w - x$.

$$\frac{\partial(x,w)}{\partial(x,y)} = \begin{vmatrix} 1 & 0 \\ 1/2 & 1/2 \end{vmatrix} = 1/2, \tag{7.112}$$

$$f_{X,W}(x,w) = f_{X,Y}(x,y)\frac{\partial(x,y)}{\partial(x,w)} = \frac{1}{\pi(1+x^2)}\frac{1}{\pi(1+(2w-x)^2)}2, \tag{7.113}$$

$$f_W(w) = \int_{-\infty}^{\infty} f_{X,W}(x,w)\,\mathrm{d}x = \frac{2}{\pi^2}\int_{-\infty}^{\infty} \frac{1}{1+x^2}\frac{1}{1+(2w-x)^2}\,\mathrm{d}x. \tag{7.114}$$

The expectation of a Cauchy distributed r.v. X is not well-defined because $\mathsf{E}[|X|] = \infty$. As a consequence, taking the average of some outcomes (i.e. a sample) will also not give a sensible answer.

s.3.11. The idea is this. We generate a bunch of uniform random deviates (simulated values). Then we count how often the set $\{x + y \leq 1, xy \leq 2/9\}$ is hit.

Here is the code.

Python Code

```python
import numpy as np

np.random.seed(3)

num = 100

X = np.random.uniform(0, 1, num)
Y = np.random.uniform(0, 1, num)
U = X + Y
V = X * Y
success = (U <= 1) * (V <= 2 / 9)
print(sum(success) / num)
```

s.3.12. a.

$$
\begin{aligned}
F \geq 0 &\implies 2 < y &\quad (7.115)\\
F \leq 1 &\implies F(3, y) \leq 1 \implies F(3,4) = 1 &\quad (7.116)
\end{aligned}
$$

b. $F(3,7) = 1$.

c. $f(x, y) = \partial_x \partial_y F(x, y) = (x-1)/4$ for $x \in (1,3), y \in (2,4)$ and 0 elsewhere.

d.

$$
\begin{aligned}
\mathsf{P}\{2 < X < 3\} &= F_X(3) - F_X(2) &\quad (7.117)\\
&= F_{X,Y}(3,4) - F_{X,Y}(2,4) = 1 - 1 \cdot 2/8 = 3/4. &\quad (7.118)
\end{aligned}
$$

e. Make a drawing of the rectangle $[2,3] \times [2,4]$. Then check what parts of this are covered by $F_{X,Y}$.

$$
\mathsf{P}\{2 < X < 3, 2 < Y < 3\} = F_{X,Y}(3,3) - F_{X,Y}(2,3) - F_{X,Y}(3,2) + F_{X,Y}(2,2). \quad (7.119)
$$

The rest is just number plugging.

f. Use the fundamental bridge and c.

$$
\begin{aligned}
\mathsf{P}\{Y < 2X\} &= \mathsf{E}[I_{Y<2X}] &\quad (7.120)\\
&= \iint I_{y<2x} f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y &\quad (7.121)\\
&= \frac{1}{4} \iint I_{y<2x} I_{2<y<4} I_{1<x<3} (x-1)\,\mathrm{d}x\,\mathrm{d}y &\quad (7.122)\\
&= \frac{1}{4} \int_1^3 (x-1) \int I_{2<y<\min\{2x,4\}}\,\mathrm{d}y\,\mathrm{d}x &\quad (7.123)\\
&= \frac{1}{4} \int_1^3 (x-1)(\min\{2x,4\} - 2)\,\mathrm{d}x &\quad (7.124)\\
&= \frac{1}{4} \int_1^2 (x-1)(2x-2)\,\mathrm{d}x + \frac{1}{4} \int_2^3 (x-1)(4-2)\,\mathrm{d}x. &\quad (7.125)
\end{aligned}
$$

Finishing the computation must be easy for you now (and if not, practice real hard).

g. As X, Y continuous, the answer is equal to that of f.

h. Similar to f. but a bit more involved.

$$
\begin{aligned}
\mathsf{P}\{Y < 2X, Y + 2X > 6\} &= \mathsf{E}\left[I_{Y<2X, Y>6-2X}\right] &\quad (7.126)\\
&= \iint I_{y<2x, y>6-2x} f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y &\quad (7.127)\\
&= \frac{1}{4} \iint I_{y<2x, y>6-2x} I_{2<y<4} I_{1<x<3} (x-1)\,\mathrm{d}x\,\mathrm{d}y &\quad (7.128)\\
&= \frac{1}{4} \int_1^3 (x-1) \int I_{\max\{2,6-2x\}<y<\min\{2x,4\}}\,\mathrm{d}y\,\mathrm{d}x &\quad (7.129)\\
&= \frac{1}{4} \int_1^3 (x-1)[\min\{2x,4\} - \max\{2, 6-2x\}]^+\,\mathrm{d}x, &\quad (7.130)
\end{aligned}
$$

where we need the $[\cdot]^+$ to ensure the positivity of $\min\{2x,4\}-\max\{2,6-2x\}$. To see this, make a graph of the function $\min\{2x,4\}-\max\{2,6-2x\}$. Also, from this graph,

$$= \frac{1}{4}\int_{3/2}^{2}(x-1)(2x-6+2x)\,\mathrm{d}x + \frac{1}{4}\int_{2}^{3}(x-1)(4-2)\,\mathrm{d}x. \tag{7.131}$$

The rest is for you.

s.3.13. The function $g(x)=x^4$ is not one-to-one on $\mathbb{R}$. It is, however, locally, one-to-one, around the roots of U. (In this course we don't deal with complex numbers, for your interest, it can be proven that the equation x^4-y has, in general, four roots in the complex plane.)

We need to be bit careful with applying the change of variables formula, but we are OK if we apply it locally around the roots $U^{1/4}$ and $-U^{1/4}$. However, mind that we also should take care of the domain of V, so it might be that these roots don't lie in the domain of V.

With all this, let's first tackle the Jacobian, and then get the domain right with indicators.

$$u = g(v) = v^4 \implies v = \pm u^{1/4}, \tag{7.132}$$

$$f_U(u)\,\mathrm{d}u = f_V(v)\,\mathrm{d}v \implies f_U(u) = f_V(v)\frac{\mathrm{d}v}{\mathrm{d}u}, \tag{7.133}$$

$$\frac{\mathrm{d}u}{\mathrm{d}v} = 4v^3 = 4u^{3/4}I_{v\geq 0} - 4u^{3/4}I_{v<0}, \tag{7.134}$$

$$f_U(u) = \frac{f_V(-u^{1/4})}{4(-u)^{3/4}}I_{-u^{1/4}\in(-3,0)} + \frac{f_V(u^{1/4})}{4(u)^{3/4}}I_{u^{1/4}\in[0,2)} \tag{7.135}$$

$$= \frac{f_V(-u^{1/4})}{4(-u)^{3/4}}I_{u\in(0,81)} + \frac{f_V(u^{1/4})}{4(u)^{3/4}}I_{u\in[0,16)}. \tag{7.136}$$

If V has the uniform distribution, then $f_V(v)=\frac{1}{5}$ for $v\in(-3,2)$, so

$$f_U(u) = \frac{1}{20(-u)^{3/4}}I_{u\in(0,81)} + \frac{1}{20(u)^{3/4}}I_{u\in[0,16)}. \tag{7.137}$$

s.3.14. Here is a direct approach.

$$x = \tan u = g(u) \implies u = \arctan x \tag{7.138}$$

$$\frac{\mathrm{d}x}{\mathrm{d}u} = \frac{1}{\cos^2 u} = \frac{\sin^2 u + \cos^2 u}{\cos^2 u} = \tan^2 u + 1 = x^2 + 1, \tag{7.139}$$

$$f_X(x) = f_U(u)\frac{\mathrm{d}u}{\mathrm{d}x} = \frac{1}{\pi}I_{u\in(0,\pi)}\frac{1}{1+x^2} \tag{7.140}$$

$$= \frac{1}{\pi(1+x^2)}I_{\arctan x\in(0,\pi)} = \frac{1}{\pi(1+x^2)}. \tag{7.141}$$

In the last equation we just shifted the tan from $(-\pi/2,\pi/2]$ to the interval $(0,\pi)$. The tan has also a proper inverse in $(0,\pi)$ (make a drawing of tan to see this), hence all is well-defined.

s.3.19. If we divide two $\sim \text{Norm}(0,1)$ r.v.s we obtain a Cauchy distributed r.v. But in our Beluga case, the normally distributed r.v.s have positive expectation.

s.3.28. Suppose we chop up the area under some arbitrary function g in blocks of height $g(x)$ and length Δx. Then the area of such a block is $g(x)\Delta x$.

In our case, we chop up the interval in parts with length $\Delta x = 1/N$. The elements of $f3$ are such that $f3[i] = f_3(x_i)\Delta x$, where x_i lies in the ith interval and f_3 is the density of the sum of the three r.v.s. But then, $f_3(x_i) = f3[i]/\Delta x = Nf3[i]$.

Forgetting to scale with $\Delta x = 1/N$ is a common error when dealing with densities. Hence, recall that, notationally, $f(x)\,\mathrm{d}x$ means a block of height $f(x)$ and length $\mathrm{d}x$. Don't forget to deal with the $\mathrm{d}x$!

s.3.29. Take four r.v.s U, V, X, Y. Then

$$(U+V+X)+Y = (U+V)+(X+Y). \tag{7.142}$$

Thus the density of $(U+V+X)+Y$ must be the same as the density of $(U+V)+(X+Y)$.

s.4.1. Incorrect: The support of T is $(0,2)$ whereas the support of any beta distribution is $(0,1)$. Hence, T does not have a beta distribution for some a, b.
Also see page 378 of the book for the distribution of the sum of two uniform distributions. This might help your intuition for this solution.

s.4.2. We use that the PDF integrates to 1:

$$1 = \int_0^1 \frac{1}{\beta(1,b)}(1-x)^{b-1}\,\mathrm{d}x = \frac{1}{\beta(1,b)}\left[-\frac{1}{b}(1-x)^b\right]_0^1 = \frac{1}{\beta(1,b)b}.$$

Hence, $\beta(1,b) = \frac{1}{b}$.

s.4.3. The scaling factor $\beta(a,b)$ is a positive constant, so we may as well leave it out and maximize $x^{a-1}(1-x)^{b-1}$. Note that its derivative (to x) is given by

$$\begin{aligned}\frac{d}{dx}x^{a-1}(1-x)^{b-1} &= ((a-1)(1-x)-(b-1)x)x^{a-2}(1-x)^{b-2}\\ &= ((a-1)-(a+b-2)x)x^{a-2}(1-x)^{b-2}.\end{aligned}$$

Setting this to zero yields $x = \frac{a-1}{a+b-2}$ as the only candidate for an interior optimum. Since $a, b > 1$, we have $0 < x < 1$. If $a, b > 1$, then the PDF converges to 0 as $x \to 0$ or $x \to 1$, so then we conclude that $x = \frac{a-1}{a+b-2}$ indeed yields a maximum. **Think about this last sentence; most groups did not use the information that $a, b > 1$ correctly.**

s.4.4. A prior is a distribution reflecting one's information or belief about a parameter before updating it with information. **Note: almost nobody had a completely satisfactory answer here. Compare your solution to the definition above. If they are different, try to understand how exactly your solution was different and determine which definition is better.**
A conjugate prior is a prior distribution such that the posterior distribution is in the same family of distributions.

s.4.5. Dirichlet distribution. The Beta distribution is a special case of the Dirichlet distribution, because binomial is a special case of multinomial. Of course, this can also be shown directly using the formula.

s.4.6. The prior is $p \sim \text{Beta}(1,1)$. The posterior is $p|X = k \sim \text{Beta}(1+k, 1+n-k)$.

s.4.7. Let X denote the number of heads.

1. Your posterior is $p|X = 900 \sim \text{Beta}(910,110)$.
2. Your friend's posterior is $p|X = 900 \sim \text{Beta}(901,101)$.
3. The mean of your posterior is $\frac{910}{910+110} = \frac{91}{102} \approx 0.892$; the mean of your friend's posterior is $\frac{901}{901+101} = \frac{901}{1002} \approx 0.899$. The difference is small, so the effect of the prior distribution is small if you have a lot of data. This effect is known as *washing out the prior*.

s.4.8. We fill in $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ to find

$$f_Y(y) = \varphi\left(\sqrt{y}\right) y^{-1/2} = \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-(\sqrt{y})^2/2} = \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2},$$

so $\alpha = \frac{1}{2}$ and $\lambda = \frac{1}{2}$.

s.4.9. Incorrect: the scale parameters λ need to be the same **and** both random variables need to be independent.

s.4.10. The base case is $n = 1$. We have $\Gamma(1) = \int_0^\infty e^{-x}\,\mathrm{d}x = 1 = 0!$, so the statement holds for $n = 1$. Now let $k \in \mathbb{N}$ be arbitrary and assume that the statement holds for $n = k$, i.e. that $\Gamma(k) = (k-1)!$. Then

$$\Gamma(k+1) = k\Gamma(k) = k(k-1)! = k! = ((k+1)-1)!, \tag{7.143}$$

so the statement also holds for $n = k+1$. By mathematical induction, we conclude that $\Gamma(n) = (n-1)!$ for all positive integers n.

s.4.11. Incorrect: It is the other way around, the Gamma distribution is the conjugate prior of the Poisson distribution. This statement doesn't make much sense, for example one would need to say for which parameter of the Gamma distribution it is the prior. In addition, the parameters of the Gamma distribution can be any positive real number, so the conjugate prior of (either parameter) of the Gamma distribution is a continuous distribution, so in particular not the Poisson distribution.

s.4.12. $X + Y \sim \text{Gamm}(11,2)$ and $\frac{X}{X+Y} \sim \text{Beta}(4,7)$.

s.4.13.

1. Minimum
2. Maximum

3. Median

s.4.14. No. The only relevant information is the amount of legs won by each player.

s.4.15. Our current information can be represented as: $A_{10} = 6$.

s.4.16. We have $A_n \sim \text{Bin}(n, p)$.

s.4.17. Let f_0 denote the prior distribution of p. Then for the posterior pdf we find by Bayes' theorem:

$$f_1(p|A_n = k) = \frac{\mathsf{P}\{A_n = k \mid p\} f_0(p)}{\mathsf{P}\{A_n = k\}} \tag{7.144}$$

$$= \frac{\binom{n}{k} p^k (1-p)^{n-k} \cdot 1}{\mathsf{P}\{A_n = k\}} \tag{7.145}$$

$$\propto p^k (1-p)^{n-k}, \tag{7.146}$$

in which we recognize the pdf of a $\text{Beta}(k+1, n-k+1)$ distribution (up to a normalizing constant). Hence, $p|A_n = k \sim \text{Beta}(k+1, n-k+1)$.

s.4.18. Important: we have already observed 10 legs with an outcome, with which we have updated our belief. Hence, we should use the *posterior* distribution given $A_{10} = 6$ in this exercise! (Most groups did this incorrectly.) Think about it. Suppose instead we had observed 1000 legs and Amy had won 990 of them (i.e., $A_{1000} = 990$). Wouldn't we use this information if someone offered us a bet?
Note that Bob should wins the match if and only if he wins the next three legs. Let W_k be short-hand notation for the event "Bob wins the kth leg". Then, observing that W_{11}, W_{12}, W_{13} are independent, and using the LOTP in the fourth step, we obtain

$$\mathsf{P}\{\text{Bob wins the match} \mid A_{10} = 6\} = \mathsf{P}\{W_{11} \cap W_{12} \cap W_{13} \mid A_{10} = 6\} \tag{7.147}$$

$$= \mathsf{P}\{W_{11} \mid A_{10} = 6\}\, \mathsf{P}\{W_{12} \mid A_{10} = 6\}\, \mathsf{P}\{W_{13} \mid A_{10} = 6\} \tag{7.148}$$

$$= \mathsf{P}\{W_{11} \mid A_{10} = 6\}^3 \tag{7.149}$$

$$= \int_0^1 \mathsf{P}\{I_{11} \mid p, A_{10} = 6\}^3 f_1(p|A_{10} = 6) dp \tag{7.150}$$

$$= \int_0^1 (1-p)^3 \cdot \frac{p^6 (1-p)^4}{\beta(7,5)} dp \tag{7.151}$$

$$= \frac{\beta(7,8)}{\beta(7,5)} \int_0^1 \frac{p^6 (1-p)^7}{\beta(7,8)} dp \tag{7.152}$$

$$= \frac{\beta(7,8)}{\beta(7,5)}. \tag{7.153}$$

(Note that we very explicitly do all the steps here. It might be more intuitively clear if you skip the first few steps and write

$$\mathsf{P}\{\text{Bob wins the match} \mid A_{10} = 6\} = \int_0^1 (1-p)^3 f_1(p|A_{10} = 6) dp, \tag{7.154}$$

and work from there).

s.4.19. We have

$$\begin{aligned}
\mathsf{P}\{\text{Bob wins the match} \mid A_{10}=6\} &= \frac{\beta(7,8)}{\beta(7,5)} && (7.155)\\
&= \left(\frac{6!7!}{14!}\right) \Big/ \left(\frac{6!4!}{11!}\right) && (7.156)\\
&= \frac{7!/4!}{14!/11!} && (7.157)\\
&= \frac{7\cdot 6\cdot 5}{14\cdot 13\cdot 12} && (7.158)\\
&= 5/52 && (7.159)\\
&= 0.0962. && (7.160)
\end{aligned}$$

s.4.20. Our expected profit when taking the bet is

$$\begin{aligned}
&300\cdot \mathsf{P}\{\text{Bob wins the match} \mid A_{10}=6\} - 10\cdot \mathsf{P}\{\text{Amy wins the match} \mid A_{10}=6\} && (7.161)\\
&= 300\cdot\frac{5}{52} - 10\cdot(1-\frac{5}{52}) && (7.162)\\
&= 19.808. && (7.163)
\end{aligned}$$

So we expect to make a profit of €19.81. Hence, you should take the bet.

s.5.1. Let X be the outcome of the die throw (note that X is a random variable) and let A be the event that the outcome is even. Then

$$\mathsf{E}[X \mid A] = 2\,\mathsf{P}\{X=2|A\} + 4\,\mathsf{P}\{X=4|A\} + 6\,\mathsf{P}\{X=6|A\} = \tfrac{1}{3}\cdot(2+4+6) = 4.$$

We conclude that $\mathsf{E}[X \mid A] = 4$.

s.5.2. 1. Since Adam keeps $b/2$ and does the gamble with $a = b/2$, we have

$$\mathsf{E}[X] = b/2 + \tfrac{1}{5}\cdot 4(b/2) + \tfrac{4}{5}\cdot 0 = 0.9b.$$

2. The computation is the same as in part 1., but with X instead of b:

$$\mathsf{E}[Y|X] = X/2 + \tfrac{1}{5}\cdot 4(X/2) + \tfrac{4}{5}\cdot 0 = 0.9X.$$

Note that the result is a random variable.
3. Using Adam's law (and linearity of expectation), we conclude that:

$$\mathsf{E}[Y] = \mathsf{E}[\,\mathsf{E}[Y|X]] = \mathsf{E}[0.9X] = 0.9\,\mathsf{E}[X] = 0.81b.$$

In general, if Adam would do this n times, the expected amount of money he has after n such gambles would be $0.9^n b$. This would be very difficult to show without Adam's law!

s.5.3. We have $\mathsf{E}[X|N] = Np$, so using Adam's law (and linearity of expectation), we conclude that $\mathsf{E}[X] = \mathsf{E}[\,\mathsf{E}[X|N]] = \mathsf{E}[Np] = \mathsf{E}[N]p = \lambda p$.

This is in accordance with $X \sim \text{Pois}(\lambda p)$, which was shown in the chicken-egg story.

s.5.4. Incorrect: $\mathsf{E}[X \mid A]$ is a number since A is an event, whereas $\mathsf{E}[X \mid I_A]$ is a random variable since I_A is a random variable. A correct statement is $\mathsf{E}[X \mid A] = \mathsf{E}[X \mid I_A = 1]$.

s.5.5. Correct, if X and Y are independent, then $\mathsf{E}[Y \mid X] = \mathsf{E}[Y]$ which is constant (formally, a degenerate random variable).

Since the variance of a constant is 0, we conclude that $\mathsf{V}[\mathsf{E}[Y \mid X]] = 0$.

s.5.6. 1. We compute $\mathsf{E}[X \mid X \geq a]$ as follows:

$$\begin{aligned}
\mathsf{E}[X \mid X \geq a] &= \int_0^\infty y f(y|A)\,\mathrm{d}y \\
&= \int_0^\infty y \frac{\lambda e^{-\lambda y} I_{y \geq a}}{e^{-\lambda a}}\,\mathrm{d}y \\
&= \lambda \int_a^\infty y e^{-\lambda(y-a)}\,\mathrm{d}y \\
&= -y e^{-\lambda(y-a)}\Big|_a^\infty + \int_a^\infty e^{-\lambda(y-a)}\,\mathrm{d}y \\
&= a - \frac{1}{\lambda} e^{-\lambda(y-a)}\Big|_a^\infty = a + \frac{1}{\lambda}.
\end{aligned}$$

2. The result also follows from the memoryless property, which states that conditional on the event that $X \geq a$, we have that $X - a | X \geq a \sim \text{Exp}(\lambda)$.

s.5.7. 1. Note that $X \mid A \sim \text{Bin}(10, 0.5)$, so $\mathsf{E}[X \mid A] = 10 \cdot 0.5 = 5$.
2. Note that $X \mid A^c \sim \text{Bin}(10, 0.8)$, so $\mathsf{E}[X \mid A^c] = 10 \cdot 0.8 = 8$.
3. By LOTE we have $\mathsf{E}[X] = \mathsf{P}\{A\}\,\mathsf{E}[X \mid A] + \mathsf{P}\{A^c\}\,\mathsf{E}[X \mid A^c] = 0.9 \cdot 5 + 0.1 \cdot 8 = 5.3$.
4. Note that $\mathsf{P}\{B \mid A\} = 0.5^4$ and $\mathsf{P}\{B \mid A^c\} = 0.8^4$. By LOTP we have

$$\mathsf{P}\{B\} = \mathsf{P}\{A\}\,\mathsf{P}\{B \mid A\} + \mathsf{P}\left\{A^c\right\}\mathsf{P}\left\{B \,\middle|\, A^c\right\} = 0.9 \cdot 5 + 0.1 \cdot 8 = 0.09721.$$

5. By Bayes' rule $\mathsf{P}\{A \mid B\} = \frac{\mathsf{P}\{B \mid A\}\mathsf{P}\{A\}}{\mathsf{P}\{B\}} \approx 0.57864$.
6. Note that $\mathsf{E}[X \mid A, B] = 4 + 6 \cdot 0.5 = 7$ and $\mathsf{E}[X \mid A^c, B] = 4 + 6 \cdot 0.8 = 8.8$. By LOTP with extra conditioning we have

$$\mathsf{P}\{X \mid B\} = \mathsf{P}\{A \mid B\}\,\mathsf{E}[X \mid A, B] + \mathsf{P}\left\{A^c \,\middle|\, B\right\}\mathsf{E}\left[X \,\middle|\, A^c, B\right] \approx 7.75844.$$

7. By LOTE we have $\mathsf{P}\{B\}\,\mathsf{E}[X \mid B] + \mathsf{P}\{B^c\}\,\mathsf{E}[X \mid B^c] = \mathsf{E}[X] = 5.3$. We know $\mathsf{P}\{B\}$ and $\mathsf{E}[X \mid B]$, so solving this for $\mathsf{E}[X \mid B^c]$ yields $\mathsf{E}[X \mid B^c] \approx 5.035$.

s.5.8. The marginal density of X is given by $f_X(x) = 2(1-x)$.

So the conditional density is given by $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{I_{x \leq y}}{1-x}$. Hence,

$$\mathsf{E}[Y \mid X = x] = \int_0^1 y \frac{I_{x \leq y}}{1-x}\,\mathrm{d}y = \frac{1}{1-x}\int_x^1 y\,\mathrm{d}y = \frac{1}{1-x}\left[\tfrac{1}{2}y^2\right]_x^1 = \frac{\frac{1}{2}\left(1-x^2\right)}{1-x} = \tfrac{1}{2}(1+x).$$

We conclude that $\mathsf{E}[Y \mid X] = \frac{1}{2}(1+X)$.

The marginal density of Y is given by $f_Y(y) = 2y$.
So the conditional density is given by $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{I_{x\le y}}{y}$. So

$$\mathsf{E}[X \,|\, Y = y] = \int_0^1 x \frac{I_{x\le y}}{y}\,\mathrm{d}x = \frac{1}{y}\int_0^y x\,\mathrm{d}x = \tfrac{1}{2}y.$$

We conclude that $\mathsf{E}[X \,|\, Y] = \frac{1}{2}Y$.

s.5.9. Note that $\mathsf{E}[X \,|\, X \ge a] \ge a > \mathsf{E}[X \,|\, X < a]$. By LOTE:

$$\begin{aligned}\mathsf{E}[X] &= \mathsf{P}\{X \ge a\}\,\mathsf{E}[X \,|\, X \ge a] + \mathsf{P}\{X < a\}\,\mathsf{E}[X \,|\, X < a] \\ &< \mathsf{P}\{X \ge a\}\,\mathsf{E}[X \,|\, X \ge a] + \mathsf{P}\{X < a\}\,\mathsf{E}[X \,|\, X \ge a] \\ &= \mathsf{E}[X \,|\, X \ge a],\end{aligned}$$

where the inequality is strict since $\mathsf{P}\{X < a\} > 0$.

s.5.10. With the hint,

$$\mathsf{E}[N \,|\, X] = \mathsf{E}[N - X \,|\, X] + \mathsf{E}[X \,|\, X] = \mathsf{E}[N - X] + X = \lambda(1-p) + X.$$

As a check, $\mathsf{E}[\mathsf{E}[N \,|\, X]] = \mathsf{E}[\lambda(1-p) + X] = \lambda(1-p) + \lambda p = \lambda = \mathsf{E}[N]$.
Here is straightforward computation. You should check each and every step as they are based on pattern recognition.

$$\begin{aligned}\mathsf{E}[N \,|\, X = k] &= \sum_{n=k}^{\infty} n\,\mathsf{P}\{N = n \,|\, X = k\} && (7.164)\\ &= \frac{1}{\mathsf{P}\{X = k\}} \sum_{n=k}^{\infty} n e^{-\lambda}\frac{\lambda^n}{n!}\binom{n}{k} p^k (1-p)^{n-k} && (7.165)\\ &= \frac{1}{\mathsf{P}\{X = k\}} \sum_{n=k}^{\infty} n e^{-\lambda}\frac{1}{n!}\frac{n!}{k!(n-k)!}(\lambda p)^k (\lambda(1-p))^{n-k} && (7.166)\\ &= \frac{e^{-\lambda p}(\lambda p)^k/k!}{\mathsf{P}\{X = k\}} \sum_{n=k}^{\infty} n e^{-\lambda(1-p)}\frac{1}{(n-k)!}(\lambda(1-p))^{n-k} && (7.167)\\ &= \sum_{n=k}^{\infty} n e^{-\lambda(1-p)}\frac{1}{(n-k)!}(\lambda(1-p))^{n-k} && (7.168)\\ &= \sum_{n=0}^{\infty} (n+k) e^{-\lambda(1-p)}\frac{1}{n!}(\lambda(1-p))^{n} && (7.169)\\ &= k + \sum_{n=0}^{\infty} n e^{-\lambda(1-p)}\frac{1}{n!}(\lambda(1-p))^{n} && (7.170)\\ &= k + \lambda(1-p). && (7.171)\end{aligned}$$

Hence, $\mathsf{E}[N \,|\, X] = \lambda(1-p) + X$. Since $\mathsf{E}[X] = \lambda p$, we get $\mathsf{E}[N] = \lambda$ with Adam's law, as above.

s.5.11.

$$\mathsf{E}\left[\tilde{X}\,\middle|\,Y\right] = \mathsf{E}\left[X - \hat{X}\,\middle|\,Y\right] = \mathsf{E}[X\,|\,Y] - \mathsf{E}[\mathsf{E}[X\,|\,Y]\,|\,Y] \tag{7.172}$$

$$= \mathsf{E}[X\,|\,Y] - \mathsf{E}[X\,|\,Y]\,\mathsf{E}[1\,|\,Y] \tag{7.173}$$

$$= \mathsf{E}[X\,|\,Y] - \mathsf{E}[X\,|\,Y]1 = 0 \tag{7.174}$$

s.5.12.

$$\mathsf{E}\left[\tilde{X}\right] = \mathsf{E}\left[\mathsf{E}\left[\tilde{X}\,\middle|\,Y\right]\right] = \mathsf{E}[0\,|\,Y] = 0. \tag{7.175}$$

This means that the estimation error $\tilde{X}$ does not have bias.

s.5.13.

$$\mathsf{E}\left[\tilde{X}\hat{X}\right] = \mathsf{E}\left[\mathsf{E}\left[\tilde{X}\hat{X}\,\middle|\,Y\right]\right] \tag{7.176}$$

$$= \mathsf{E}\left[\mathsf{E}\left[\tilde{X}\,\mathsf{E}[X\,|\,Y]\,\middle|\,Y\right]\right] \tag{7.177}$$

$$= \mathsf{E}\left[\mathsf{E}[X\,|\,Y]\,\mathsf{E}\left[\tilde{X}\,\middle|\,Y\right]\right] \tag{7.178}$$

$$= \mathsf{E}[\mathsf{E}[X\,|\,Y]0\,|\,Y] = 0 \tag{7.179}$$

s.5.14. Using the previous exercises,

$$\mathsf{Cov}\left[\hat{X},\tilde{X}\right] = \mathsf{E}\left[\hat{X}\tilde{X}\right] - \mathsf{E}\left[\hat{X}\right]\mathsf{E}\left[\tilde{X}\right] = 0 - \mathsf{E}\left[\hat{X}\right]0 = 0. \tag{7.180}$$

Next, from the definition of $\tilde{X} = X - \hat{X} \implies X = \tilde{X} + \hat{X}$. The variance of the sum is the sum of the variances since $\hat{X}$ and $\tilde{X}$ are uncorrelated.

s.5.15. Since $\mathsf{E}\left[\tilde{X}\right] = 0$,

$$\mathsf{V}\left[\tilde{X}\right] = \mathsf{E}\left[\tilde{X}^2\right] \tag{7.181}$$

$$= \mathsf{E}\left[\mathsf{E}\left[\tilde{X}^2\,\middle|\,Y\right]\right] \tag{7.182}$$

$$= \mathsf{E}\left[\mathsf{E}\left[(X - \hat{X})^2\,\middle|\,Y\right]\right] \tag{7.183}$$

$$= \mathsf{E}\left[\mathsf{E}\left[(X - \mathsf{E}[X\,|\,Y])^2\,\middle|\,Y\right]\right] \tag{7.184}$$

$$= \mathsf{E}[\mathsf{V}[X\,|\,Y]], \tag{7.185}$$

where the last equation follow from the definition of $\mathsf{V}[X\,|\,Y]$.

For Eve's law, use the above and the previous exercise to see that

$$\mathsf{V}[X] = \mathsf{V}\left[\hat{X}\right] + \mathsf{V}\left[\tilde{X}\right] = \mathsf{V}[\mathsf{E}[X\,|\,Y]] + \mathsf{E}[\mathsf{V}[X\,|\,Y]]. \tag{7.186}$$

s.5.16. By the linearity of expectation and BH Theorem 9.3.9:

$$\begin{aligned}\mathsf{E}\left[(Y - \mathsf{E}[Y|X] - h(X))^2\right] &= \mathsf{E}\left[(Y - \mathsf{E}[Y|X])^2 - 2(Y - \mathsf{E}[Y|X])h(X) + (h(X))^2\right] \\ &= \mathsf{E}\left[(Y - \mathsf{E}[Y|X])^2\right] - \mathsf{E}[2(Y - \mathsf{E}[Y|X])h(X)] + \mathsf{E}\left[(h(X))^2\right] \\ &= \mathsf{E}\left[(Y - \mathsf{E}[Y|X])^2\right] + \mathsf{E}\left[(h(X))^2\right].\end{aligned}$$

Since $\mathsf{E}\left[(h(X))^2\right] \geq 0$, we conclude that $\mathsf{E}\left[(Y - \mathsf{E}[Y|X] - h(X))^2\right] \geq \mathsf{E}\left[(Y - \mathsf{E}[Y|X])^2\right]$ for any function h, so $\mathsf{E}[Y|X]$ is the predictor of Y based on X with the lowest mean squared error, i.e. the best predictor of Y based on X.