Probability distributions EBP038A05: 2020-2021 Assignments

Nicky D. van Foreest and student assistents

January 12, 2021

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

The assignments contain 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. Most of the selected exercises of the book are also at about (or just a bit above) exam level. The section with challenges is for those students that like a challenge; the problems are above exam level. The final section is about coding skills. We explain the rationale next.

You have to get used to programming and checking your work with computers, for instance by using simulation. These 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 in the language you like. You should know, however, that many of you will need both languages later in life.

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. 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 and Y, both $\sim \mathrm{U}([0,1])$, and $Y > X$.

1. Why is the joint PDF $f_{X,Y}(x,y) = I_{x \le y}$?
2. Are X and Y independent?
3. Compute $F_{X,Y}(x,y)$.

Ex 1.3. Claim (that is, is the following clam correct?): We have two continuous r.v.s. X, Y. Claim: even though the joint CDF factors into the product of the marginal it is still possible in general that the joint PDF does not factors into a product of marginals PDFs of X and Y.

Ex 1.4. Express Bayes' formula for two rvs X and Y in terms of the joint CDF, i.e., provide a formula.

Ex 1.5. What is a contingency table?

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. Apply the chicken-egg story. Families enter a zoo in a given hour. Some families have one child, other two, and so on. What are the 'eggs' in this context, and what is the meaning of 'hatching'?

Ex 1.8. Claim: We have two rvs X and Y on $\mathbb{R}^+$. It is given that $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for $x, y \le 1/3$. Then X and Y are necessarily independent.

Ex 1.9. 'The man on the street' says that two throws of a die are independent, but does not mention the joint CDF. What do you think of this claim of independence? (Hint, from this exercise you should memorize this: **independence is a property of the joint CDF, not of the rvs**.)

Ex 1.10. I select a random guy from the street, his height $X \sim \mathrm{N}(1.8, 0.1)$, and I select a random woman from the street, her height is $Y \sim \mathrm{N}(1.7, 0.08)$. I claim that since I selected the man and the woman independently, their heights are independent. Comment on this claim.

Ex 1.11. Correct? For any two rvs X and Y on $\mathbb{R}^+$ with marginals F_X and F_Y. Then $\mathsf{P}\{X \le x, Y \le y\} = F_X(x)F_Y(x)$.

Ex 1.12. Theorem 7.1.11. What is the meaning of the notation $X|N = n$?

Ex 1.13. Correct? X, Y two discrete rvs with CDF $F_{X,Y}$. We can compute the PDF as $\partial_x \partial_y F_{X,Y}(x,y)$.

1.2 *Exercises at about exam level*

Ex 1.14. We throw an unbiased die with six sides; the result of the ith throw is X_i.

1. What is the sample space of the two throws (X_1, X_2)?
2. What is the joint CDF?
3. What is the joint PMF?
4. Marginalize out X_2 to show that $\mathsf{P}\{X_1 = 5\} = 1/6$.
5. Use the fundamental bridge and indicators to compute $\mathsf{P}\{X_1 > X_2\}$.
6. Use the fundamental bridge and indicators to compute $\mathsf{P}\{|X_1 - X_2| < 1\} = 1/6$.
7. Use the fundamental bridge and indicators to compute $\mathsf{P}\{|X_1 - X_2| \leq 1\}$.
8. How would you use simulation to estimate $\mathsf{P}\{|X_1 - X_2| \leq 1\}$?

Ex 1.15. We select a random married couple (a man and a woman). His height is $X \sim \mathrm{N}(1.8, 0.1)$, her height is $Y \sim \mathrm{N}(1.7, 0.08)$ in meters.

1. What is the sample space of (X, Y)?
2. If your answer to question 1 is correct, you must have noticed that the potentially the height of the man and the woman can be negative. Is this a problem for this model?
3. What is the joint CDF?
4. What is the joint PDF?
5. Marginalize out Y to show that $X \sim \mathbb{N}1.8, 0.1$.
6. Use the fundamental bridge and indicators to write $\mathsf{P}\{X > Y\}$ as an integral.
7. Use the fundamental bridge and indicators to write $\mathsf{P}\{|X - Y| < 0.1\}$ as an integral. You don't have to solve the integral.

Ex 1.16. Take $X \sim \mathrm{U}(\{-2, -1, 1, 2\})$ and $\eta = X^2$. What is the correlation coefficient of X and η? If we would consider another distribution for X, would that change the correlation?

Ex 1.17. 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 $\{1, 2, \ldots, 10\}$ and a claim C_i uniform on $\{0, 1, \ldots 8\}$.

1. What is the interpretation of $\bar{I}_n = \min\{I_i : 0 \leq i \leq n\}$?
2. What is the meaning of $I_1 = I_0 + X_1 - C_1$?
3. What is the meaning of $I_2 = I_1 + X_2 - C_2$?
4. What is the interpretation of $I'_1 = \max\{I_0 - C_1, 0\} + X_1$?
5. What is the interpretation of $I'_2 = \max\{I'_1 - C_2, 0\} + X_2$?
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'.

11. Provide also an interpretation in terms of a degradation and repair process of an item. (if you find this difficult, search a bit on the web on reliability theory.) Comment on how good you think this model is to analyze such degradation and repair processes.

Ex 1.18. We have a machine that consists of two components. The machine works as long as not both components have failed. Let X_i be the lifetime of component i.

1. What is the interpretation of $\min\{X_1, X_2\}$?

2. What is the interpretation of $\max\{X_1, X_2\}$?

3. If X_1, X_2 iid $\sim$ Exp(10) (in hours), what is the expected time until the machine fails?

4. If X_1, X_2 iid $\sim$ Exp(10) (in hours), what is the probability that the machine is still 'up' (i.e., not failed) at time $T = 50$?

Ex 1.19. Let X be the result of the throw of a coin. It is given that $\mathsf{P}\{X = H\} = p = 1 - \mathsf{P}\{X = T\}$. When $X = H$, we choose a fair die with 4 sides with values $1, 2, 3, 4$, when $X = T$ we choose a fair die with 6 sides with values $1, \ldots, 6$. Let Y_i be the value of the ith throw with the die.

1. What is the PMF of X and Y_1?

2. Marginalize the answer of part a to show that $\mathsf{P}\{X = H\} = p$.

3. What is $\mathsf{P}\{Y_1 = 1\}$?

4. What is $\mathsf{P}\{X = H \mid Y_1 = 1\}$?

5. What is $\mathsf{P}\{X = H \mid Y_1 = 1, Y_2 = 2\}$?

6. What is $\mathsf{P}\{X = H \mid Y_1 = Y_2 = \cdots = Y_n = 1\}$?

Ex 1.20. Check first BH 7.2.3. When X, Y iid $\sim$ N(0,1), then $X - Y \sim$ N(0,2). However, when X, Y iid $\sim$ P(λ), then prove first that $X + Y \sim$ P(2λ), but note that $X - Y$ is not $\sim$ P(0). Explain this difference between the Poisson and normal distribution.

Ex 1.21. Assume that X has the Cauchy distribution.

1. Does $\mathsf{E}\left[\frac{X}{X^2+1}\right]$ exist? If so, find its value.

2. Does $\mathsf{E}\left[\frac{|X|}{X^2+1}\right]$ exist? If so, find its value.

1.3 *Challenges*

Ex 1.22. Consider the 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 $q = 1 - p$. Let X be the number of eggs that hatch and let Y be the number of eggs that do not hatch, so $X + Y = N$. For $N \sim \mathrm{P}(\lambda)$ it is shown in BH 7.1.9 that X and Y are independent. This exercise asks for the converse. Assume that X and Y are independent. Prove that there exists a $\lambda > 0$ such that $N \sim \mathrm{P}(\lambda)$.

1.4 *Coding skills*

Ex 1.23. Use simulation to estimate the answer of BH.7.1. Run the code below and explain line 9 of python code or line of the R code. Compare the value of the simulation to the exact value.

```python
import numpy as np

np.random.seed(3)

num = 1000

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

```r
a <- 3
```

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 1.24. 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. Explain lines lines 11, 21 and 26 of the python code.

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

labda = 3

X = expon(scale=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)
```

Ex 1.25. How many ping pong balls fit into an Airbus Beluga? One way to answer this is as follows. According to this wikipage the cargo volume V of this airplane is 1500 m^3. But V 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 m^3. The volume of a ping pong ball is $v = 4\pi r^3/3 = 33.49333333333333 cm^3$ 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. (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 volume of the 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

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

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 entirely simple. However, with simulation it's easy to get an estimate. Use the code to provide these estimates, and explain line 11 of the python code.

Contrary to BH.7.1.25 we get a fine mean, i.e, $\mathsf{E}[N] < \infty$. But isn't this strange? Here we divide two normal random variables, and in BH.7.1.25 also two normal rvs are divided. Comment on the difference.

The numerical results suggest the interesting guess $\mathsf{V}[N] \approx \mathsf{V}[V] * \mathsf{V}[v]$. For the moment I do not completely understand why this is so, or whether it is true in general.

```
1  import numpy as np
2  from scipy.stats import norm
3
4  num = 500
5
6  np.random.seed(3)
7
8  V = norm(2500, 500)
9  v = norm(33.5, 0.5)
10
11 N = V.rvs(num) / v.rvs(num)
12 print(N.mean(), N.std())
13
14 print(2500/33.5)
15 print(np.sqrt(500*0.5))
```

2 ASSIGNMENT 2

2.1 *Have you read well?*

Ex 2.1. What is the difference between 1D LOTUS and 2D LOTUS?

Ex 2.2. 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 2.3. In queueing theory the concept of squared coefficient of variance SCV of a rv X is very important. It is defined as $C = \mathbb{V}[X]/(\mathsf{E}[X])^2$. Is the SCV of X equal to $\mathrm{Corr}(X,X)$? Can it happen that $C > 1$?

2.2 *Exercises at about exam level*

Ex 2.4. Derive the results of BH 7.3.6 without smart tricks. Thus, you have to use the fundamental bridge to show that

$$\mathsf{E}[ML] = \mathsf{E}[X]\,\mathsf{E}[Y] = 1, \qquad \mathsf{E}[M] = 3/2, \qquad \mathsf{E}[L] = 1/2,$$
$$\mathsf{E}\left[L^2\right] = 1/4, \qquad \mathsf{E}\left[M^2\right] = 2\,\mathsf{E}\left[X^2\right] - \mathsf{E}\left[L^2\right] = 7/2$$
$$\mathsf{V}[M] = \mathsf{E}\left[M^2\right] - (\mathsf{E}[M])^2, \qquad \mathsf{V}[L] = \mathsf{E}\left[L^2\right] - (\mathsf{E}[L])^2.$$

You can use the document 'Memoryless exercursions' to see how to solve these problems.

2.3 *Coding skills*

TODO 2.1. Maximum of independent r.v.s, BH.5.6.5.

1. Make 1d array of uniform 0,1 random rvs
2. compute mean and variance
3. Why include a seed
4. include seed
5. Make [n, p] matrix, n samples along rows
6. Sort along axis 1
7. Sort along axis 0
8. Compute mean and std along axis 0
9. Make large number of data, with e.g sample-no = 1000, and redo the above
10. Change to exponential distribution
11. Show how to use the online documentation for np.random.exponential
12. Show the effect of the scale parameter
13. Compare mean to theoretical value
14. Change to geometric distribution

TODO 2.2. Let $X \sim \text{Exp}(\lambda)$ and $Y \sim N(\mu, \sigma)$, independent of X. So, draw X first and let the outcome be x; then draw $Y \sim N(x, \sigma)$. Take $\lambda = 4$, $\mu = 5$, $\sigma = 3$.

1. Make a 3D plot of $F_{X,Y}$.
2. Make a 3D plot of $f_{X,Y}$.
3. Plot f_X, i.e., plot $\mu e^{-\mu t}$. Then use simulation to marginalize out Y from $f_{X,Y}$ to obtain $\hat{f}_X$; we write $\hat{f}_X$ because it has been obtained from simulation. Plot $\hat{f}_X$ in the same figure as f_X, and compare the result.
4. Use simulation to estimate $f_{X|Y}$. Plot this in the same graph for various values of $X = x$.
5. Make a 3D plot of f

TODO 2.3. Let $X \sim \text{Exp}(\lambda)$ and $Y|X \sim N(X, \sigma)$. So, draw X first and let the outcome be x; then draw $Y \sim N(x, \sigma)$. Take $\lambda = 5$, $\sigma = 3$.

1. Use simulation to estimate $\mathsf{E}[Y]$.
2. Make a 3D plot of $F_{X,Y}$.
3. Make a 3D plot of $f_{X,Y}$.
4. Plot f_X, i.e., plot $\lambda e^{-\lambda t}$. Then use simulation to marginalize out Y from $f_{X,Y}$ to obtain $\hat{f}_X$. Plot $\hat{f}_X$ in the same figure as f_X, and compare the result.
5. Use simulation to estimate $f_{X|Y}$. Plot this in the same graph for various values of $X = x$.
6. Make a 3D plot of f

TODO 2.4. On BH.7.32

1. Solve this problem and explain your solution.
2. Let $L = \min\{\}X, Y\}$ and $M = \max\{\}X, Y\}$. Use the memoryless property to explain that $\mathsf{E}[M] = \mathsf{E}[L] + 1/\lambda$.
3. Let $L_n = \min\{\}X_i : i = 1, \ldots, n\}$ where $X_i \sim \text{Exp}(()\lambda)$, and likewise $M_n = \max\{\}X_i\}$. Use the memoryless property to explain that $\mathsf{E}[M_n] = \mathsf{E}[L_{n-1}] + \mathsf{E}[M_{n-1}]$.
4. What is $\mathsf{E}[M_n]$ for $n = 5$ and $\lambda = 4$?
5. Explain the idea behind code below. In particular, why do we make a matrix of exponentially distributed random variables? Why is the `sort` along `axis=1`? Why is the mean along `axis=0`? Look up the meaning of `cumsum` in the `numpy` docs. Why is there a `cumsum` in the computation of `times`?
6. Run the code, and include your output

```
import numpy as np

np.random.seed(10)

labda = 4
num = 5
samples = 400

X = np.random.exponential(1 / labda, size=(samples, num))
print(X)
X.sort(axis=1)
print(X.mean(axis=0))
print(X)

times = np.array([1 / ((num - j) * labda) for j in range(num)])
times = times.cumsum()
print(times)
```

TODO 2.5. On BH.7.48

1. Solve the problem and explain your solution.
2. Below is the python code to estimate the mean and variance by means of simulation. Explain how the algorithm works, in particular, how does `find_maxima` work?.
3. Replace the seed for the random number generator with your student number (without the "s" of course). Run this code, and include the numerical results in your report. If you prefer to use `R`, that is ok too, but then port the ideas of the code below to `R`.

```
import numpy as np

np.random.seed(3)

num = 10

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


```

```
def find_maxima(X):
    Xstar = np.zeros_like(X)
    M = -np.infty
    for i, x in enumerate(X):
        if x > M:
            Xstar[i] = 1
            M = x
    return Xstar


Xstar = find_maxima(X)
print(Xstar)

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

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

2.4 *TODO Applications*

Use Story 13.4.2 to generate exponentially distributed inter-arrival times. (It is not forbidden to use results we have not discussed yet.)

2.5 *Why is the Exponential Distribution so important?*

The exponential distribution plays a very important role in probability theory, but why? This assignment is meant to answer this question.

2.5.1 *Train arrivals*

Suppose a train departs between 10:00 and 10:15 minutes, and that 250 people will take this train. As a very simple model, let's suppose that each person arrives uniformly distributed on the interval $[0,15]$. What is the distribution of the inter-arrival times of these people? In the next couple of exercises we will use simulation to show that the exponential distribution is a very reasonable model.

The basic idea behind the simulation is as follows. First we use a random number generator to generate an array A of arrival times. Second, we compute the inter-arrival time

$$X_i = A_i - A_{i-1}, \quad i = 1,2,\ldots, \tag{2.1}$$

between the arrivals. Third, we compute the empirical distribution F_e of X. Fourth, we plot $F_e(x)$ and the theoretical distribution $F(x) = 1 - e^{-\lambda x}$ for a proper λ. Once all steps are done, hopefully (the graphs of) $F_e(x)$ and $F(x)$ are (very) similar, so that we can conclude that the inter-arrival times are approximately exponentially distributed.

For the λ, observe that all arrivals occurred between $m = \min\{A\}$ and $M = \max\{A\}$. Thus, as a simple estimate take $\lambda = n/(M-m)$. This is easy, because when you change the size of A, or the distribution, then this estimate of λ scales in the right way.

One of the nice, and bad, things of `R` is that many algorithms are included. This is handy when you know what you do, but it leaves you clueless if you have to do something new. More generally, using standard functions does not help you develop algorithmic skills. However, this is very important if you plan to use large amounts of data in your later career, as actuary, consultant, banker, financial quant, whatever. For this reason we discuss here how to make an empirical distribution function, even though you can just invoke the `ecdf` function (empirical cumulative distribution function) in `R` to compute it for you.

As a concrete, but simple example, suppose we are given the following set of ages of people $X = (20,25,18,18,19)$. The empirical distribution function is defined as

$$F_e(x) = n^{-1} \sum_{i=1}^{n} I_{X_i \le x}.$$

To compute this efficiently, we first sort the ages: $(18,18,19,20,25)$. Next, we give a count number to each individual: $(1,2,3,4,5)$, and divide this count number by $n = 5$, since there are 5 persons. Finally, we plot the ages along the x axis, and $(1/5,2/5,3/5,4/5,5/5)$ along the y axis. When we plot this, we see that $F_e(18) = 2/5$. To get things really correct, we should remove the double counts, such as the 18, but we skip this here.

You can choose between python or R to make the simulations and the plots. We include sample code for each to help you get started; it's up to you what you like to use. As an advice: learn both. `R` is handy for data analysis, but it is not used much besides academia; in business, machine learning, programming, python is much more common.

First we need to load some libraries. Check the web on what they do.

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3 import seaborn as sns
```

Now we set a theme for `seaborn` to make nice graphs, and we set a seed for the random number generator of `numpy` so that we get the same random numbers every time we do a run. This helps to find bugs. (If you get different numbers each and every time, checking whether the results are correct becomes very tedious very rapidly.) We also set the labels on the axis.

```
np.random.seed(3)
sns.set_theme()

plt.xlabel('x', fontsize=16)
plt.ylabel('y', fontsize=16)
```

Here is the algorithm to compute the `ecdf`.

```
def ecdf(data):
    x = np.sort(data)
    n = x.size
    y = np.arange(1, n + 1) / n
    return (x, y)
```

Now we compute the arrival times and sort them

```
num = 250
A = np.sort(np.random.uniform(0, 15, size=num))
labda = A.size / (A.max() - A.min())
```

Finally, compute the inter-arrival times and plot the `ecdf`.

```
X = A[1:] - A[:-1]
x, y = ecdf(X)
plt.scatter(x=x, y=y)
plt.plot(x, 1 - np.exp(-labda * x))
# plt.show()
```

2.6 *Empirical distribution functions*

Ex 2.5. Make a graph of the empirical and theoretical distrribution for $n = 250$, i.e. the size of A is 250, of X and the theoretical distribution. Explain what we can see in this graph.

Ex 2.6. Make a graph of F_e and F for $n = 10$. Explain.

Ex 2.7. Make a graph of $|F_e - F|$ for $n = 250$. What do you see?

Ex 2.8. Make a graph of $\log(1 - F_e)/\lambda$ for $n = 250$. What do you see? Why did we take transform of F_e?

Ex 2.9. Take the arrival times as normally distributed with mean $\mu = 7.5$ and $\sigma = 3$, and make the graphs. Explain.

Normally distributed, early and late arrivals, i.e., $\mu = 7.5$ and $\sigma = 5$, and then merge the ones that are in time with the ones that are late and early.

2.7 *Measuring inter arrival-times*

We have a device to measure the time between two arrivals, of jobs for instance, or customers in a shop, or particles in radio-active decay. After a measurement, the device has to recharge so it cannot measure arrivals that occur within 10 seconds from each other. Also, if there is no arrival within 50, it resets itself. Hence, the device cannot easily measure inter-arrival times longer than 1 minute. Assume that the inter-arrival times are exponentially distributed with some unknown λ. Let $\bar{x} = \sum_{n=1}^{N} x_n/N$ be the sample mean of N measured inter-arrival times.

1. Explain that, if $\lambda \ll 1$ minute, $\hat{\lambda} = \bar{x} - 10$ is a reasonable estimator for λ.
2. How would you obtain a reasonable estimate of λ when $\lambda \gg 1$ minute?

3 ASSIGNMENT 3

Topics of chapter 8.

3.1 *Have you read well?*

Ex 3.1. Explain in your own words:

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

Ex 3.2. Look up on the web: what is the conjugate prior of the multinomial distribution?

4 THE BAKER OF POINCARE

Henri Poincaré was a French mathematician who taught at the Sorbonne around 1900. The following anecdote about him is probably fabricated, but it makes an interesting probability problem. Supposedly Poincaré suspected that his local bakery was selling loaves of bread that were lighter than the advertised weight of 1 kg, so every day for a year he bought a loaf of bread, brought it home and weighed it. At the end of the year, he plotted the distribution of his measurements and showed that it fit a normal distribution with mean 950 g and standard deviation 50 g. He brought this evidence to the bread police, who gave the baker a warning. For the next year, Poincaré continued the practice of weighing his bread every day. At the end of the year, he found that the average weight was 1000 g, just as it should be, but again he complained to the bread police, and this time they fined the baker.

Why? Because the shape of the distribution was asymmetric. Unlike the normal distribution, it was skewed to the right, which is consistent with the hypothesis that the baker was still making 950 g loaves, but deliberately giving Poincaré the heavier ones. Exercise 5.6 Write a program that simulates a baker who chooses n loaves from a distribution with mean 950 g and standard deviation 50 g, and gives the heaviest one to Poincaré. What value of n yields a distribution with mean 1000 g? What is the standard deviation?

5 ASSIGNMENT 4

5.1 *Have you read well?*

5.2 *Coding skills*

Ex 5.1. Bekijk het maximale verschil van de verdeling van de som van 3 uniformen, en de normale verdeling.

5.2.1 *Bayesian priors, Testing for rare deceases, Making the plot of Exercise 7.86*

In line with Exercise 8.33, we are now going to analyze the effect on $\mathsf{P}\{D \mid T\}$ when the sensitivity is not known exactly. So, why is this interesting? In Example 2.3.9 the sensitivity is given, but in fact, in 'real' experiments, this is not always known as accurately as assumed in this example. For example, in this paper: False-positive COVID-19 results: hidden problems and costs it is claimed that 'The current rate of operational false-positive swab tests in the UK is unknown; preliminary estimates show it could be somewhere between 0·8\Hence, even though it is claimed that PCR tests 'have analytical sensitivity and specificity of greater than 95\Simply put, the specificity and sensitivity are not precisely known, hence this must affect $\mathsf{P}\{D \mid T\}$.

To help you, we show how to make one graph. Then we ask you to make a few on your own, and comment on them.

5.2.2 *Redoing the computation of the Example 2.3.9*

I write `p_D_g_T$` for $\mathsf{P}\{D \mid T\}$. Here is how this can be implemented in python.

```
sensitivity = 0.95
specificity = 0.95
p_D = 0.01

p_T = sensitivity * p_D + (1-specificity)*(1-p_D)
p_D_g_T  = sensitivity * p_D/p_T
p_D_g_T
```

1. Make a plot of $\mathsf{P}\{D \mid T\}$ in which you vary the sensitivity from 0.9 to 0.99. Explain what you see.
2. Make a plot of $\mathsf{P}\{D \mid T\}$ in which you vary the specificity from 0.9 to 0.99.
3. Make a plot of $\mathsf{P}\{D \mid T\}$ in which you vary $\mathsf{P}\{D\}$ from 0.01 to 0.5. Explain what you see.
4.

5.3 *Compound Poisson distribution, hitting times, and overshoot distribution*

6 ASSIGNMENT 5

6.1 *Have you read well?*

7 BAYES' BILLIARDS

Take $n = 100$ samples.

7.1 *One coin*

1. Take success probability $p = 1/2$.
2. Make matrix with n rows, n columns. Each row is an experiment of n throws of the coin.
3. Plot the histogram of the number of heads.

7.2 *Three coins*

1. Take three coins with success probabilities $p = 1/4, 1/2, 3/4$.
2. Make a simulation for each coin.
3. If we select a coin with probability 1/3, the total histogram is the 1/3 times the sum of the histograms of each of the coins. That is, $\mathsf{P}\{X = k\} = \mathsf{P}\{X = k|C = i\}\,\mathsf{P}\{C = i\}$, where C is one of chosen coins; here we take $\mathsf{P}\{C = i\} = 1/3$.

7.3 *five coins*

1. Select with uniform probability one out of five coins with success probabilities $p = i/5$, $i = 1,\ldots,5$.
2. Make a simulation for each coin.
3. Make the histogram $\mathsf{P}\{X = k\} = \mathsf{P}\{X = k|C = i\}\,\mathsf{P}\{C = i\}$, where $\mathsf{P}\{C = i\} = 1/5$.

8 ASSIGNMENT 6

8.1 *Have you read well?*