Probability and Stochastic Processes

with a Twist of GNU Octave towards Queueing

Tommi Sottinen

tommi.sottinen@iki.fi
www.uva.fi/~tsottine/psp/psp.pdf

March 14, 2018
Revised February 6, 2019

Contents

Preface

Nobody reads the preface! At least the students don't. Therefore, I guess you are a teacher who is contemplating on using these notes in your own course. You are welcome to do so! Also, the LaTeX source code for these notes are available from the author upon request. You are also allowed to make any changes to the notes. I only hope you will give me credit somewhere in your derived notes. If you forget to give credit, I will forgive you.

These lecture notes are for the master's level course STAT 3120 "Probability and Stochastic Processes" lectured for the first time in spring 2017 at the University of Vaasa, and updated for spring 2018. This is a 5 ECTS credit course with approximately 40 hours of lectures and 20 hours of exercises. One hour is 45 minutes. One lecture in these notes is supposed to mean approximately one lecture session of 2 hours (2 times 45 minutes) in class. This will probably not happen in practice.

The focus of these notes is to prepare the students for queueing theory. In that sense these lectures goes like a train towards that final station, but as a local train that stops at many stations on the track. The students are assumed to have some basic knowledge of probability theory and to know at least the elements of computer programming, preferably with Matlab or Octave. In Part I of these notes we recall some basic facts of probability and random variables with the emphasis of the so-called conditioning trick that is fundamental in the analysis of Markov chains. In Part II we introduce some random variables that are useful in queuing theory. Part III is a rather standard, but concise, introduction Markov chains. Finally, Part IV is an introduction to the very basics of queueing theory. After that, there is an appendix with a list of exam questions.

I have tried to argue the validity of each claim I make in these notes (with the notable exception of the Lévy's continuity theorem) with a less-than-rigorous proof. It is my sincere hope that this will help the students' understanding and not to confuse them too much.

In these lecture notes I have originally used GNU Octave version 4.0.0 (later 4.2.2) installed on a laptop running Ubuntu 16.04 (later 18.04), but other versions and other OS's should work just fine. The only inconvenience I have encountered is the different plotting systems especially as producing PDF's (and I don't mean probability distribution functions) is concerned.

Civitanova Marche, Helsinki, and Vaasa
March 14, 2018 T. S.

These lecture notes have been revised February 6, 2019. Some typos have been corrected and some muddled thinking has been clarified. The gamma distribution has been changed into the Erlang distribution.

Vaasa
February 6, 2019 T. S.

The Focus Problem

After finishing these lectures, you should be able so solve the following problem; and not only to solve, but understand the assumptions and limitations of the solution, and to apply the solution method to related and seemingly unrelated problems:

> You are queuing in the Ministry of Love to get access to Room 101. There are 5 clerks serving the customers, and a single queue using the fist-in-first-out queuing policy. At the moment all the clerks are a busy and there are 12 customers in line in front of you. You have been waiting for 20 minutes. During that time you have observed the service times of 3 customers. They were 1 minute, 5 minutes and 18 minutes.
>
> (i) What is the probability that your total waiting time plus service time in the Ministry of Love will exceed 1 hour?
>
> (ii) Suppose you have now waited 25 minutes and there are 11 customers in front of you and you have recorded an extra service time which was 10 seconds. What is now the probability that your total waiting time plus service time in the Ministry of Love will exceed 1 hour?

Actually, we will not solve The Focus Problem completely. We will only give a relatively reasonable solution. A complete solution would be a nice Master's Thesis, which I would be happy to supervise.

Bibliography

[1] EATON, J. (1996–2018) *GNU Octave Manual* (Version 4.4.1 in February 3, 2019) www.gnu.org/software/octave/doc/interpreter/.

[2] GUTTORP, P. (1995) *Stochastic Modeling of Scientific Data* Chapman and Hall/CRC.

[3] ROSS, S.M. (2010) *Introduction to Probability Models. 10th ed.* Academic Press.

Part I

Conditioning Tricks

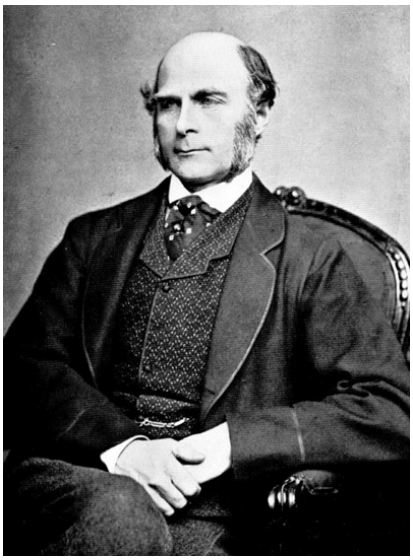
Lecture 1

Conditioning Tricks of Means

Sir Francis Galton (1822–1911) was an English polymath. He created the statistical concept of correlation, and was the first to apply statistical methods to the study of inheritance of intelligence. He is considered the founder of psychometrics. He was a pioneer in eugenics, coining the term itself and the phrase "nature versus nurture". He devised a method for classifying fingerprints that proved useful in forensic science.

In Galton's time there was concern amongst the Victorians that aristocratic surnames were becoming extinct. Galton originally posed the question regarding the probability of such an event in an 1873 issue of *The Educational Times*, and the Reverend **Henry William Watson** (1827–1903) replied with a solution. Together, they then wrote an 1874 paper entitled *On the probability of the extinction of families*. Galton and Watson appear to have derived their process independently of the earlier work by **Irénée-Jules Bienaymé** (1796–1878). The solution is called the **Galton–Watson branching process**, or the **Galton–Watson–Bienaymé branching process**, or simply the **branching process**.

While many of the Galton's viewpoints can be seen today as antiquated, or even offensive, the Galton–Watson branching process is still a central probabilistic tool applied in various fields of science.



Sir Francis Galton (1822–1911)

Example 1.1 of this lecture (and the following three lectures) deals with a branching processes. More precisely, we are interested in the number of (male) offspring in a given generation of a single forefather. The eventual goal will be to determine the probability of the ultimate extinction of all family lines of the said forefather. This problem will be solved much later in Lecture 4. In this lecture, we confine ourselves in analyzing the mean of the distribution of a given generation in the family tree. Indeed, understanding the mean is the first thing to do in understanding a random phenomenon.

This first lecture, and indeed the first part of this book, is called "Conditioning Tricks" because conditioning is the key trick we need to analyze stochastic processes, and life, the universe, and all such things.

1.1 Example (Persé–Pasquale Noble Family Tree, I)
The most noble family of Persé–Pasquale is worried of their continuing existence. At the moment there is only one male descendant of this most noble line. According to the family records, the males of the noble family of Persé–Pasquale have sired male children as follows

Number of male children	Frequency
0	503
1	62
2	859
More than 2	0

(a) What is the probability that the 6^{th} generation has more than 10 male descendants?
(b) What is the average number of descendants in the 6^{th} generation?
(c) What is the variance of the number of descendants in the 6^{th} generation?
(d) What is the probability that the Persé–Pasquale family will be ultimately extinct?

The **offspring distribution** $\mathbf{p} = [p_x]_{x=0}^{\infty}$, i.e., the distribution of the number of (male) children sired by a given male in the Persé–Pasquale family, can be estimated from the given data in a naïve way by using the **method of relative frequencies**. Since we have in total

$$503 + 62 + 859 + 0 \quad = \quad 1424$$

observations, we obtain the probabilities

$$\begin{aligned} p_0 &= 503/1424 &&= 0.353230, \\ p_1 &= 62/1424 &&= 0.043539, \\ p_2 &= 859/1424 &&= 0.603230, \\ p_x &= 0/1424 &&= 0.000000 \text{ for } x \geq 3. \end{aligned}$$

Since there is no data, or any other reason, to assume that there is any particular dependence structure between the number of male children of the different males in the Persé–Pasquale family tree, we assume that they are independent. Also, we assume that all the males in the Persé–Pasquale family have the same offspring distribution, that is given above. Let X_n denote the number of male descendants in the n^{th} generation of the Persé–Pasquale family tree. Let $\xi_{n,i}$ denote the number of male children sired by the i^{th} (male) member of the $(n-1)^{\text{th}}$ generation of the Persé–Pasquale family. Then

$$X_n \quad = \quad \xi_{n,1} + \xi_{n,2} + \cdots + \xi_{n,X_{n-1}}.$$

Suppose, for example, that the current only descendant ($X_0 = 1$) has 2 children. Then $X_1 = \xi_{1,1} = 2$. Suppose the first child has no children and the second child has three children. Then $X_2 = \xi_{2,1} + \xi_{2,2} = 0 + 3 = 3$.

We have just developed a mathematical model that is called the **branching process**.

1.2 Definition (Branching Process)
A stochastic process X_n, $n \in \mathbb{N}$, given by the initial condition $X_0 = 1$ and the recursion

$$X_n \quad = \quad \xi_{n,1} + \xi_{n,2} + \cdots + \xi_{n,X_{n-1}},$$

is called a **branching process**. The summands $\xi_{n,i}$ are independent and identically distributed with **offspring distribution**

$$p_x \quad = \quad \mathbb{P}[\xi_{n,i} = x], \qquad x \in \mathbb{N}.$$

1.3 Remark
On the first sight, things look very rosy for the continuing existence of the Persé–Pasquale family. Each generation is looking for a healthy 25 % increase, since the mean male reproductive rate, or the mean of the offspring distribution is

$$\mu \quad = \quad \sum_x x\, p_x \quad = \quad 1.2500.$$

After some analysis, we will see later in Lecture 4 that the things are not so rosy after all!

In theory, the distribution of the n^{th} generation X_n is completely determined by the offspring distribution $\mathbf{p} = [p_x]_{x \in \mathbb{N}}$. In practice, we will see later in Lecture 3 that the distribution of X_n is complicated, and even a numerical implementation of it is somewhat nontrivial. Indeed, it can be said that the solution to the question (a) of Example 1.1 is very complicated and fragile, while the solution to the questions (b) and (c) are easy and robust. Problem (d) will also turn out to be relatively easy and robust, once we develop the analytical tools for it in Lecture 4.

Law of Total Probability

Almost all **conditioning tricks** in probability come from the following observation: Suppose we are interested in the probability $\mathbb{P}[A]$ of some event A. The probability $\mathbb{P}[A]$ may be difficult to calculate. However, sometimes the **conditional probabilities**

$$\mathbb{P}[A|B_k] \quad = \quad \frac{\mathbb{P}[A\,,\, B_k]}{\mathbb{P}[B_k]}$$

may be relatively easy to calculate for some **alternatives** B_k, $k \in \mathbb{N}$. (Alternative means that exactly one of the events B_k, $k \in \mathbb{N}$, will happen.) Suppose further that the probabilities $\mathbb{P}[B_k]$, $k \in \mathbb{N}$, for the alternatives are easy to calculate. Then the probability $\mathbb{P}[A]$ is relatively easy to calculate. Indeed, we have the following basic result:

1.4 Lemma (Law of Total Probability)
Let B_k, $k \in \mathbb{N}$, be such events that precisely one of them will happen. Then

$$\mathbb{P}[A] \;=\; \sum_{k\in\mathbb{N}} \mathbb{P}[B_k]\,\mathbb{P}[A|B_k].$$

To see how the conditioning formula above is true, simply think like this: In order for the event A to happen, first one (any one) of the alternative events B_k must have happened, and then A happens, given that the alternative B_k has happened.

1.5 Remark

(i) If X and Y are both discrete random variables, then the law of total probability gives us

$$\mathbb{P}[X = x] \;=\; \sum_{y} \mathbb{P}[Y = y]\,\mathbb{P}[X = x|Y = y].$$

(ii) If X and Y are both continuous random variables, then the law of total probability gives us (after approximating integrals with Riemann sums and passing to the limit)

$$f_X(x) \;=\; \int_{-\infty}^{\infty} f_Y(y) f_{X|Y}(x|y)\,\mathrm{d}y,$$

where $f_X(x)$ and $f_Y(y)$ are the density functions of X and Y, and $f_{X|Y}(x|y)$ is the conditional density function of X given Y.

(iii) For general random variables X and Y, one can use the **Leibniz–Stieltjes formalism** and write the law of total probability as

$$\mathbb{P}[X \in \mathrm{d}x] \;=\; \int_{-\infty}^{\infty} \mathbb{P}[X \in \mathrm{d}x \,|\, Y = y]\,\mathbb{P}[Y \in \mathrm{d}y].$$

Here the integration is with respect to the variable y and $\mathrm{d}x$ denotes an infinitesimal interval around x, and, formally,

$$\mathbb{P}[X \in \mathrm{d}x \,|\, Y = y] \;=\; \frac{\mathbb{P}[X \in \mathrm{d}x\,,\, Y \in \mathrm{d}y]}{\mathbb{P}[Y \in \mathrm{d}y]}.$$

It is possible to use Lemma 1.4 to construct the probability distribution of the n^{th} generation of the branching process. This is, unfortunately, rather technical. Therefore we postpone the construction to Lecture 3. In this lecture we confine ourselves in understanding the mean the n^{th} generation distribution.

Adam's Law

Motto: The distribution is complicated. The mean is easy.

Let X and Y be random variables. The **conditional expectation** or **conditional mean**

$$\mathbb{E}[X \,|\, Y]$$

is a random variable whose value is known if the value of Y is known. In other words $\mathbb{E}[X|Y] = g(Y)$, where g is some function depending on the joint distribution of X and Y. If the value of Y is known, to be y, say, then $\mathbb{E}[X|Y = y] = g(y)$.

1.6 Remark

(i) If X and Y are both discrete random variables, then

$$\mathbb{E}[X \,|\, Y = y] \;=\; \sum_x x\, \mathbb{P}[X = x \,|\, Y = y].$$

(ii) If X and Y are both continuous random variables, then

$$\mathbb{E}[X \,|\, Y = y] \;=\; \int_{-\infty}^{\infty} x\, f_{X|Y}(x|y)\,\mathrm{d}x,$$

where $f_{X|Y}(x|y)$ is the conditional density function of X given Y.

(iii) For general random variables X and Y, the **Leibniz–Stieltjes formalism** gives the formula

$$\mathbb{E}[X|Y = y] \;=\; \int_{-\infty}^{\infty} x\, \mathbb{P}[X \in \mathrm{d}x \,|\, Y = y],$$

where the integration is with respect to the variable x.

Suppose then, for simplicity, that X and Y are both discrete random variables. Then, by using the law of total probability of Lemma 1.4 and Remark 1.5(i), we obtain

$$\mathbb{P}[X = x] = \sum_y \mathbb{P}[X = x \,|\, Y = y]\,\mathbb{P}[Y = y].$$

Consequently,

$$\begin{aligned}
\mathbb{E}[X] &= \sum_x x\,\mathbb{P}[X=x] \\
&= \sum_x x\left(\sum_y \mathbb{P}[X=x\,|\,Y=y]\,\mathbb{P}[Y=y]\right) \\
&= \sum_y \left(\sum_x x\,\mathbb{P}[X=x\,|\,Y=y]\right)\mathbb{P}[Y=y] \\
&= \sum_y \mathbb{E}[X\,|\,Y=y]\,\mathbb{P}[Y=y] \\
&= \mathbb{E}\Big[\mathbb{E}[X|Y]\Big].
\end{aligned}$$

Thus, we have shown (for discrete random variables) the following **law of total expectation**, a.k.a. the **Adam's law**:

1.7 Lemma (Adam's Law)
Let X and Y be random variables. Then

$$\mathbb{E}[X] = \mathbb{E}\Big[\mathbb{E}[X\,|\,Y]\Big].$$

Adam's law is all we need in order to calculate the expectations of branching processes. Let us denote by μ the mean of the offspring distribution, i.e.,

$$\mu = \sum_x x\,p_x,$$

where $p_x = \mathbb{P}[\xi_{n,i} = x]$. Now, the expectation is always linear. Thus a naïve first try would be to calculate

$$\begin{aligned}
\mathbb{E}[X_n] &= \mathbb{E}\big[\xi_{n,1} + \xi_{n,2} + \cdots + \xi_{n,X_{n-1}}\big] \\
&= \mathbb{E}\big[\xi_{n,1}\big] + \mathbb{E}\big[\xi_{n,2}\big] + \cdots + \mathbb{E}\big[\xi_{n,X_{n-1}}\big] \\
&= X_{n-1}\,\mu.
\end{aligned}$$

Unfortunately, the calculation above is not completely correct. The problem is that the number of summands X_{n-1} is random. This problem is easily corrected by using a **conditioning trick**: Suppose the value of X_{n-1} is known (and thus non-random). Then we can use the linearity of the conditional expectation and we obtain

$$\begin{aligned}
\mathbb{E}[X_n|X_{n-1}] &= \mathbb{E}\big[\xi_{n,1} + \xi_{n,2} + \cdots + \xi_{n,X_{n-1}}\,|\,X_{n-1}\big] \\
&= \mathbb{E}\big[\xi_{n,1}\,\big|\,X_{n-1}\big] + \mathbb{E}\big[\xi_{n,2}\,\big|\,X_{n-1}\big] + \cdots + \mathbb{E}\big[\xi_{n,X_{n-1}}\,\big|\,X_{n-1}\big] \\
&= \mathbb{E}\big[\xi_{n,1}\big] + \mathbb{E}\big[\xi_{n,2}\big] + \cdots + \mathbb{E}\big[\xi_{n,X_{n-1}}\big] \\
&= X_{n-1}\,\mu,
\end{aligned}$$

since the random summands $\xi_{n,i}$ are independent of the size of the previous generation X_{n-1}. Now, by using the Adam's law, we obtain the recursive formula

$$\begin{aligned}\mathbb{E}[X_n] &= \mathbb{E}[\mathbb{E}[X_n|X_{n-1}]] \\ &= \mathbb{E}[X_{n-1}]\mu.\end{aligned}$$

This recursion is straightforward to solve. Indeed, by working backwards, we see that

$$\begin{aligned}\mathbb{E}[X_n] &= \mu\mathbb{E}[X_{n-1}] \\ &= \mu^2\mathbb{E}[X_{n-2}] \\ &= \mu^3\mathbb{E}[X_{n-3}] \\ &= \cdots \\ &= \mu^n\mathbb{E}[X_0].\end{aligned}$$

Since $\mathbb{E}[X_0] = 1$, we obtain the following:

1.8 Proposition (Branching Means)
The mean of the n^{th} generation distribution of a branching process with offspring distribution $\mathbf{p}$ is

$$\mathbb{E}[X_n] = \mu^n,$$

where μ is the mean of the offspring distribution.

Below is an Octave script file that solves part (b) of Example 1.1

```
################################################################################
## FILE: perse_pasquale_b.m
##
## Mean of the offpring distribution of the Perse–Pasquale family tree.
################################################################################

## data is the frequencies.
data = [503 62 859];

## Offspring distribution is the relative frequencies.  Note that Octave starts
## indexing with 1.  So p(1) is the probability of 0 offspring.
p = data/sum(data);

## The mean is calculated by using dot product with the row vector [0 1 2 ...].
x = 0:(length(p)-1);
mu = x*p';

## Solution to Example 1.1 part (b)
n = 6;
sol_b = mu^n
```

www.uva.fi/~tsottine/psp/perse_pasquale_b.m

Let us explain the workings of the m-file `perse_pasquale_b.m`.

1.9 Remark (Get (Used to) Octave)
We explain the workings of an Octave m-file now, but not later. In later lectures we assume a good working knowledge of Octave. So pay attention! If you are an experienced Matlab user, you should have no problem with Octave (or vice versa). If you are new to Octave or Matlab, you should visit the page https://www.gnu.org/software/octave/. You should also install Octave on your own computer. It's free! If you want to try it without installing, go to the page https://octave-online.net/.

1.10 Remark (Calling Octave m-files)
To execute the (script) m-file `perse_pasquale_b.m`, simply write its name `perse_paquale_b` in the Octave prompt. Note that you should not include the file extension `.m` and also that you should have the file in your working directory (or in your search path).

So, let's get to the contents of the m-file `perse_pasquale_b.m` at hand.

The lines 1–5 all begin with the comment symbol `#`. This means that Octave does not try to understand what is written there. (In Matlab the comment symbol is `%`. This works also with Octave.)

1.11 Remark (Comment Style)
The reason for repeating the comment sign (`##`) is a matter of **style and convenience**. A singe comment sign (`#`) is typically used for debugging and experimenting with different parameters.

The first comment block of lines 1–5 (lines 2–4, actually) is also printed out if the user types

```
help perse_pasquale_b
```

in the Octave console.

The empty line 6 terminates the help block, and makes the code easier to read. Otherwise it does nothing.

The comment line 7 is there for the human readers' convenience and help.

Line 8 assigns the row vector of the data to the variable `data`. The semicolon (`;`) in the end of line 8 (and later in the end of most lines) prevents Octave from printing out the result (the contents of the variable `data` in this case).

Line 9 does nothing, and is there for the human readers' convenience; Octave does not care about this line.

In lines 10–11 we have comments to help the human reader of the code to better understand what is going on; Octave itself does not care about these lines.

In line 12 the variable `p` is set to be the row vector

$$p_x \quad = \quad \frac{data_x}{\sum_y data_y}, \qquad x = 1, 2, 3.$$

The empty line 13 is there simply to make the code easier to read. Ditto for the comment line 14. Octave does not care about these lines.

Line 15 sets `x` to be the row vector $[0\,1\,\cdots\,(k-1)]$, where $k-1$ is the largest possible number of offspring (`2` in our case).

Line 16 calculates the inner product of `x` and `p`. The apostrophe (`'`) after `p` makes `p'` a column vector. Consequently, what is calculated in line 16 is

$$\mathbf{x}\mathbf{p}' \quad = \quad [x_1\, x_2\, \cdots\, x_k] \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_k \end{bmatrix} \quad = \quad \sum_{i=1}^{k} x_i p_i,$$

which is just the mean of $\mathbf{p}$, since $x_i = i - 1$.

Lines 17–18 should be self-explanatory: they do nothing.

In line 19 we define `n` to be the number of generations we are interested in. It is a good idea **not** to set $n = 6$ in the final line 20, i.e., **not** to have lines 19–20 combined with a single line having `sol_b = mu^6`. Indeed, sometimes we wish to change the parameters, and it is easier to do so, if they are defined separately. We could have (should have) also defined `n` somewhere in the beginning of the file. Indeed, it is a good idea to define parameters separately in the beginning of an m-file, so that it is easier to experiment with them.

Finally, the solution to the problem (b) of Example 1.1 is calculated in line 20. Note the missing semicolon in the end of the line. This forces the output of the variable `sol_b` to be printed out when the m-file `perse_pasquale_b.m` is executed.

1.12 Example (Persé–Pasquale Family Tree, I, Solution (b))
Running the m-file `perse_pasquale_b.m` in the Octave console gives us the solution $\mathbb{E}[X_6] = 3.8147$.

Exercises

1.1 Exercise
Consider Example 1.1. Calculate the expectations of the sizes of the 1^{st}, 2^{nd}, 3^{rd} and $1\,000^{\text{th}}$ generations.

1.2 Exercise
Calculate the expectations of the $1\,000^{\text{th}}$ generation of a branching process with with offspring distributions

(a) $[0.5000 \quad 0.5000]$,
(b) $[0.3333 \quad 0.3333 \quad 0.3333]$,
(c) $[0.3000 \quad 0.5000 \quad 0.2000]$,
(d) $[0.2000 \quad 0.5000 \quad 0.3000]$.

1.3 Exercise (Flatus Lake, I)

(a) The Flatus bacteria reproduces by splitting. Every minute it splits with probability 0.1 into two bacteria. Luckily, the Flatus bacteria does not live very long: with probability 0.2, it dies within any given minute. A lake is contaminated by approximately 1 000 Flatus bacteria today at 8 a.m. How many Flatus bacteria are there in average living in the lake at 9 a.m. today?

(b) The Virilus Flatus bacteria is a nasty mutation of the Flatus bacteria. Otherwise it is completely similar to its weaker cousin, except that it splits withing a given minute with probability 0.2 and dies with probability 0.1. A lake is contaminated by approximately 1 000 Virilus Flatus bacteria today at 8 a.m. How many Flatus bacteria are there in average living in the lake at 9 a.m. today?

1.4 Exercise
Show the Adam's law of Lemma 1.7 for the case where one or both of the random variables X and Y is continuous.

1.5 Exercise
Let X and Y be independent random variables. Show that then the conditional expectation $\mathbb{E}[X|Y]$ is just the (unconditional) expectation $\mathbb{E}[X]$.

Lecture 2

Conditioning Tricks for Variances

Rev. Henry William Watson (1827–19103) was an English mathematician and an ordained priest at Cambridge Apostle.

Watson wrote a number of mathematics and physics books that were influential in his time, but not considered classics today.

Watson was highly appreciated in his time. For example, he was elected a fellow of the Royal Society in 1881. He was given an honorary D.Sc. by Cambridge in 1883. He was nominated by the Senate of Cambridge University to represent it as a governor on the King Edward's Foundation in Birmingham. He was bailiff of King Edward's School for three years. Nowadays, Watson is only remembered by his work on branching processes with Galton.

Rev. Henry William Watson (1827–1903)

Example 2.1 of this lecture is a continuation of Example 1.1 of Lecture 1. We will solve part (c) concerning the variance of the generations of the branching process.

2.1 Example (Persé–Pasquale Noble Family Tree, II)
The most noble family of Persé–Pasquale is worried of their continuing existence. At the moment there is only one male descendant of this most noble line. According to the family records, the males of the noble family of Persé–Pasquale have sired male children as follows

Number of male children	Frequency
0	503
1	62
2	859
More than 2	0

(a) What is the probability that the 6^{th} generation has more than 10 male descendants?
(b) What is the average number of descendants in the 6^{th} generation?
(c) What is the variance of the number of descendants in the 6^{th} generation?
(d) What is the probability that the Persé–Pasquale family will be ultimately extinct?

Conditional Variance

Let X and Y be random variables. The **conditional variance** is defined via the conditional expectation as

$$\mathbb{V}[X|Y] \;=\; \mathbb{E}\Big[\big(X-\mathbb{E}[X|Y]\big)^2\,\Big|\,Y\Big].$$

Another way of thinking conditional variance is to understand it as variance, where the expectation is replaced everywhere with the conditional expectation.

2.2 Remark (Short-Hands and Prediction)
The definition and consequently calculations involving conditional variances can be a bit of an eyeful. For the untrained eye, it can be a good idea to use short-hand notations. For example, denoting $\hat{X} = \mathbb{E}[X|Y]$ can be helpful. This is a good notation also for the reason that the conditional expectation $\mathbb{E}[X|Y]$ is indeed the best possible estimate for X after observing Y. Similarly, it may be a good idea to write $\hat{\mathbb{E}}[\,\cdot\,]$ for $\mathbb{E}[\,\cdot\,|Y]$ and $\hat{\mathbb{V}}[\,\cdot\,]$ for $\mathbb{V}[\,\cdot\,|Y]$. With this notation we have also that $\hat{\mathbb{E}}[X] = \hat{X}$.

Like the conditional expectation $\mathbb{E}[X|Y]$, the conditional variance $\mathbb{V}[X|Y]$ is a random variable whose value is known, if the value of the conditioning variable Y is known. In other words, $\mathbb{V}[X|Y=y] = v(y)$ for some function $v(y)$ depending on the joint distribution of X and Y.

Steiner's Translation Formula

In calculating variances (and conditional variances) the so-called Steiner's translation formula (known as **König–Huygens formula** for the French) is most useful. For uncoditional variance the translation formula states that

$$\mathbb{V}[X] \;=\; \mathbb{E}[X^2]-\mathbb{E}[X]^2.$$

To see why this formula is true, we can simply use the definition of the variance, simple algebra, and the linearity of expectation:

$$\begin{aligned}
\mathbb{V}[X] &= \mathbb{E}\big[(X-\mathbb{E}[X])^2\big] \\
&= \mathbb{E}\big[X^2-2X\mathbb{E}[X]+\mathbb{E}[X]^2\big] \\
&= \mathbb{E}\big[X^2\big]+\mathbb{E}\big[-2\mathbb{E}[X]\;X\big]+\mathbb{E}\big[\mathbb{E}[X]^2\big] \\
&= \mathbb{E}[X^2]-2\mathbb{E}[X]\mathbb{E}[X]+\mathbb{E}[X]^2 \\
&= \mathbb{E}[X^2]-\mathbb{E}[X]^2.
\end{aligned}$$

Now, since conditional expectation shares all the properties of expectation (if the condition is kept fixed), we see that the calculations above extend immediately to the conditional variance. Thus, we have the following lemma.

2.3 Lemma (Steiner's Translation Formula)
Let X and Y be random variables. Then

$$\mathbb{V}[X|Y] \;=\; \mathbb{E}[X^2|Y]-\mathbb{E}[X|Y]^2.$$

Eve's Law

Motto: The distribution is complicated. The variance is relatively easy.

The following **law of total variance** is also called the Eve's law, because of the operators $\mathbb{E}$ and $\mathbb{V}$. This is also the reason why the law of total expectation is called Adam's law: it comes first and is simpler.

2.4 Lemma (Eve's Law)
Let X and Y be random variables. Then

$$\mathbb{V}[X] \;=\; \mathbb{E}\Big[\mathbb{V}[X|Y]\Big]+\mathbb{V}\Big[\mathbb{E}[X|Y]\Big].$$

To show why Lemma 2.4 is indeed true, let us start by recalling the Steiner's translation formula. To avoid headache, we write the formula by using the short-hands introduced in Remark 2.2:

$$\begin{aligned}\hat{\mathbb{V}}[X] &= \hat{\mathbb{E}}[X^2]-\hat{\mathbb{E}}[X]^2\\ &= \hat{\mathbb{E}}[X^2]-\hat{X}^2.\end{aligned}$$

Now, by the unconditional Steiner's translation formula and the Adam's law we have

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}[X^2]-\mathbb{E}[X]^2\\ &= \mathbb{E}[\hat{\mathbb{E}}[X^2]]-\mathbb{E}[\hat{X}]^2.\end{aligned}$$

But by Steiner's translation formula this is

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}\big[\hat{\mathbb{V}}[X]+\hat{X}^2\big]-\mathbb{E}\big[\hat{X}\big]^2\\ &= \mathbb{E}\big[\hat{\mathbb{V}}[X]\big]+\mathbb{E}\big[\hat{X}^2\big]-\mathbb{E}\big[\hat{X}\big]^2,\end{aligned}$$

which is the Eve's law, since

$$\mathbb{E}[\hat{X}^2]-\mathbb{E}[\hat{X}]^2 \;=\; \hat{\mathbb{V}}[\hat{X}].$$

Indeed, by using our short-hand notation, the Eve's law is just

$$\mathbb{V}[X] \;=\; \mathbb{E}[\hat{\mathbb{V}}[X]]+\mathbb{V}[\hat{X}].$$

2.5 Remark
In a very formal way the Eve's law can be written as

$$\mathbb{V} \;=\; \mathbb{E}\hat{\mathbb{V}} + \mathbb{V}\hat{\mathbb{E}}.$$

2.6 Remark
Let us analyze the Eve's law a bit. Suppose we are interested in the variability of X around its mean $\mathbb{E}[X]$. This is of course nothing but $\mathbb{V}[X]$. Suppose then that we want to "explain" or predict X with some other random variable Y. The best prediction for X is then nothing but $\hat{X} = \hat{\mathbb{E}}[X] = \mathbb{E}[X|Y]$. Then the second term

$$\mathbb{V}[\mathbb{E}[X|Y]] \;=\; \mathbb{V}[\hat{\mathbb{E}}[X]] \;=\; \mathbb{V}[\hat{X}]$$

is the part of the variability of X that is explained by the variability of Y, while the first term

$$\mathbb{E}[\mathbb{V}[X|Y]] \;=\; \mathbb{E}[\hat{\mathbb{V}}[X]]$$

is the part of the variability of X that cannot be explained by Y.

Now we are ready to solve part (c) of Example 1.1 (or Example 2.1, which is the same). Let X_n be, as before, the size of the n^{th} generation. Obviously, the key **conditioning trick** is to calculate the conditional variance $\mathbb{V}[X_n|X_{n-1}]$ first. Since the summands in the n^{th} generation are all independent of each others and of the size of the $(n-1)^{\text{th}}$ generation, the conditional variance is linear:

$$\begin{aligned}
\mathbb{V}[X_n\,|X_{n-1}] &= \mathbb{V}\left[\xi_{n,1} + \xi_{n,2} + \cdots + \xi_{n,X_{n-1}}\,|X_{n-1}\right] \\
&= \mathbb{V}\left[\xi_{n,1}\,|X_{n-1}\right] + \mathbb{V}\left[\xi_{n,2}\,|X_{n-1}\right] + \cdots + \mathbb{V}\left[\xi_{n,X_{n-1}}\,|X_{n-1}\right] \\
&= X_{n-1}\sigma^2,
\end{aligned}$$

Here σ^2 is the offspring variance

$$\sigma^2 \;=\; \sum_x (x-\mu)^2\, p_x$$

and $\mathbf{p}$ is the offspring distribution and μ is the offspring mean. Once we recall from Lecture 1 that $\mathbb{E}[X_n|X_{n-1}] = X_{n-1}\mu$, we are ready to use the Eve's formula:

$$\begin{aligned}
\mathbb{V}[X_n] &= \mathbb{E}\Big[\mathbb{V}[X_n|X_{n-1}]\Big] + \mathbb{V}\Big[\mathbb{E}[X_n|X_{n-1}]\Big] \\
&= \mathbb{E}\left[X_{n-1}\sigma^2\right] + \mathbb{V}\left[X_{n-1}\mu\right] \\
&= \sigma^2\mu^{n-1} + \mu^2\mathbb{V}[X_{n-1}].
\end{aligned}$$

Now, we have to solve this nasty looking recursion. To ease our eyes during the work, we denote $V_n = \mathbb{V}[X_n]$. So, with this notation, we have the recursion

$$V_n = \sigma^2\mu^{n-1} + \mu^2 V_{n-1}$$

to solve. Working out this recursion backwards a couple of times, the general picture of what is happening becomes clear pretty soon:

$$\begin{aligned}
V_n &= \sigma^2\mu^{n-1} + \mu^2\left(\sigma^2\mu^{n-2} + \mu^2 V_{n-2}\right) \\
&= \sigma^2(\mu^{n-1} + \mu^n) + \mu^4 V_{n-2} \\
&= \sigma^2(\mu^{n-1} + \mu^n) + \mu^4\left(\sigma^2\mu^{n-3} + \mu^2 V_{n-3}\right) \\
&= \sigma^2(\mu^{n-1} + \mu^n + \mu^{n+1}) + \mu^6 V_{n-3} \\
&= \cdots \\
&= \sigma^2(\mu^{n-1} + \mu^n + \cdots + \mu^{2n-2}) + \mu^{2n} V_0 \\
&= \sigma^2(\mu^{n-1} + \mu^n + \cdots + \mu^{2n-2}) \\
&= \sigma^2\mu^{n-1}\left(1 + \mu + \cdots + \mu^{n-1}\right).
\end{aligned}$$

Now, we see that we are dealing with a **geometric series**:

$$1 + \mu + \cdots + \mu^{n-1} = \sum_{k=0}^{n-1} \mu^k = \begin{cases} \frac{1-\mu^n}{1-\mu}, & \text{if } \mu \neq 1, \\ n, & \text{if } \mu = 1, \end{cases}.$$

Consequently, we have obtained the following:

2.7 Proposition (Branching Variances)
The variance of the n^{th} generation distribution of a branching process with offspring distribution $\mathbf{p}$ is

$$\mathbb{V}[X_n] = \begin{cases} \sigma^2\mu^{n-1}\frac{1-\mu^n}{1-\mu}, & \text{if } \mu \neq 1, \\ n\sigma^2, & \text{if } \mu = 1, \end{cases}$$

where μ and σ are the mean and variance of the offspring distribution.

Below is an Octave script file that solves part (c) of Example 2.1

```
1  ##############################################################################
2  ## FILE: perse_pasquale_c.m
3  ##
4  ## Variance of the offpring distribution of the Perse–Pasquale family tree.
5  ##############################################################################
6
7  ## data is the frequencies.
8  data = [503 62 859];
9
10 ## Offspring distribution is the relative frequencies.  Note that Octave starts
```

```
11 ## indexing with 1.  So p(1) is the probability of 0 offspring.
12 p = data/sum(data);
13
14 ## The mean and the variance are calculated by using dot product with the row
15 ## vector [0 1 2 ...].
16 x = 0:(length(p)-1);
17 mu = x*p';
18 sigma_sq = (x-mu).^2 * p';
19
20 ## Solution to Example 1.1 part (c)
21 n = 6;
22 if (mu != 1)
23     sol_c = sigma_sq*mu^(n-1)*( (1-mu^n)/(1-mu) )
24 else
25     sol_c = n*sigma_sq
26 endif
```

www.uva.fi/~tsottine/psp/perse_pasquale_c.m

The code is a slight modification of the m-file `perse_pasquale_b.m`. It should be self-explanatory, except maybe for the line 18. There one should note the **pointwise** power: if `x` is a matrix, then `x^2` would be the matrix product, which in this case would not make any sense. Everything is a matrix in Octave, and Octave tries to understand sums, products and powers in terms of matrix algebra. In order to apply the operations pointwise, we must use the **dot-notation**.

2.8 Example (Persé–Pasquale Family Tree, I, Solution (c))
Running the m-file `perse_pasquale_c.m` in the Octave console gives us the solution $\mathbb{V}[X_6] = 30.716$.

Exercises

2.1 Exercise
Consider Example 1.1. Calculate the expectations and variances of the sizes of the 1^{st}, 2^{nd}, 3^{rd} and $1\,000^{\text{th}}$ generations.

2.2 Exercise
Calculate the expectations and variances of the $1\,000^{\text{th}}$ generation of a branching process with with offspring distributions

(a) [0.5000 0.5000],
(b) [0.3333 0.3333 0.3333],
(c) [0.3000 0.5000 0.2000],
(d) [0.2000 0.5000 0.3000].

2.3 Exercise
Let X and Y be independent random variables. Show that then the conditional variance $\mathbb{V}[X|Y]$ is just the (unconditional) variance $\mathbb{V}[X]$.

2.4 Exercise (Toilets for The Gods Themselves)
In an alternate universe there are three genders: *rationals*, *emotionals* and *parentals*. Naturally, all the different genders have separate public toilets. In addition to the three established genders, there are also *cissies* who do not identify to any of them. Naturally, they also have separate public toilets.

The following frequency data has been observed on how long the different genders (and cissies) spend on public toilets

Time (in minutes)	Rationals	Emotionals	Parentals	Cisses
0 – 1	129	131	16	2
1 – 2	198	102	8	
2 – 3	18	18	30	
3 – 4	15	19	2	
4 – 5	2	9		
5 – 6				
6 – 7			7	
7 – 8	7	6		1
8 – 9	3	2		
9 –	1	1		

It is said that rationals are the quickies in toilet and emotionals are the slowest. It is also said that this is a stupid gender stereotype since the variation inside genders are greater than the variation between the genders. Given the data above, is that true? What does that mean that "variations inside are greater than variations between"?

Lecture 3

Conditioning Tricks for Distributions

Irénée-Jules Bienaymé (1796–1878), was a French statistician. He contributed to the fields of probability and statistics, and to their application to finance, demography and social sciences. He formulated the Bienaymé–Chebyshev inequality (more commonly known as Chebyshev inequality or Markov inequality) and the Bienaymé formula for the variance of a sum of uncorrelated random variables.

Bienaymé was the first one to formulate and solve the extinction problem of families, already in 1845. Unfortunately, his work fell into obscurity. Thus the work of Galton and Watson that appeared some 30 years later has received unjustified prestige. In addition for not receiving due credit of his work by the posterity, Bienaymé's life was also marked by bad luck. He attended the École Polytechnique in 1815. Unfortunately that year's class was excluded in the following year by Louis XVIII because of their sympathy for Bonapartists. Later Bienaymé was an inspector general in the Finance Ministry, but was removed in 1848 for his lack of support for the new Republican regime. He then became professor of probability at the Sorbonne, but lost his position in 1851.

Irénée–Jules Bienaymé (1796–1878)

We continue with the remaining problems of Example 1.1 and/or 2.1, recalled as Example 3.1 below. In this lecture, not only we discuss some interesting mathematics, but we also show how adequate programming skills are necessary in 21^{st} century mathematics.

We derive the distribution of the n^{th} generation of a branching process formally by using convolutions. The resulting formula, Proposition 3.5, does not appear, as far as the author knows, in any other textbook that deals with branching processes. The reason is obvious: the formula is mostly useless in practice, unless one is willing to do some computer programming. And even with computers, the naïve algorithm we implement here is extremely slow and unreliable. Indeed, for example for an offspring distribution with length 4, the calculation of the 6^{th} generation distribution seems practically impossible with the naïve recursive algorithm. At least the author's laptop tried to calculate it for some 2 hours and then crashed, probably because the recursion grew too deep and wide.

3.1 Example (Persé–Pasquale Noble Family Tree, III)
The most noble family of Persé–Pasquale is worried of their continuing existence. At the moment there is only one male descendant of this most noble line. According to family records, the males of the noble family of Persé–Pasquale have sired male children as follows

Number of male children	Frequency
0	503
1	62
2	859
More than 2	0

(a) What is the probability that the 6^{th} generation has more than 10 male descendants?
(b) What is the average number of descendants in the 6^{th} generation?
(c) What is the variance of the number of descendants in the 6^{th} generation?
(d) What is the probability that the Persé–Pasquale family will be ultimately extinct?

Distribution of Sums

In order to answer to the question (a) of Example 2.1, we have to understand the distribution of random sums of random variables. Let us start this quest in a gentle way by considering the sum of two discrete random variables. It turns out that the notion of convolution is precisely what we need.

3.2 Definition (Discrete Convolution)
Let $\mathbf{p}$ and $\mathbf{q}$ be two (probability) vectors. Their (discrete) **convolution** is the vector $\mathbf{p}*\mathbf{q}$ defined as

$$(\mathbf{p}*\mathbf{q})_x = \sum_y p_{x-y} q_y .$$

The (discrete) **convolution power** is defined recursively as

$$\begin{aligned} \mathbf{p}^{*1} &= \mathbf{p}, \\ \mathbf{p}^{*n} &= \mathbf{p} * \mathbf{p}^{*(n-1)}. \end{aligned}$$

Historically, the notion of convolution does not originate from the probability theory. However, it is very well suited for it. Indeed, suppose that X and Y are both dicrete and

mutually independent random variables with probability distribution functions $\mathbf{p}$ and $\mathbf{q}$, respectively. Then, by using a **conditioning trick**, we note that

$$\begin{aligned}\mathbb{P}[X+Y=x] &= \sum_y \mathbb{P}[X+Y=x\,,\,Y=y]\\ &= \sum_y \mathbb{P}[X=x-y\,,\,Y=y]\\ &= \sum_y \mathbb{P}[X=x-y]\,\mathbb{P}[Y=y\,|\,X=x-y]\\ &= \sum_y \mathbb{P}[X=x-y]\,\mathbb{P}[Y=y].\end{aligned}$$

Thus, we have shown the following:

3.3 Lemma (Distribution of Independent Sums)
Let X and Y be two independent discrete random variables with probability distribution functions $\mathbf{p}$ and $\mathbf{q}$. Then their sum $X+Y$ has the probability distribution function $\mathbf{p}*\mathbf{q}$.

Distribution of Random Sums

The generation X_n of a branching process is a **random sum** of independent identically distributed random variables $\xi_{n,1},\ldots,\xi_{n,X_{n-1}}$, where the independent summands $\xi_{n,j}$ have the common offspring distribution $\mathbf{p}$. Lemma 3.3 can be used for **non-random sums**. Indeed, if X_{n-1} is fixed to be some y, then

$$\begin{aligned}p^n_{x|y} &= \mathbb{P}[X_n=x\,|\,X_{n-1}=y]\\ &= \mathbb{P}[\xi_{n,1}+\cdots+\xi_{n,X_{n-1}}=x\,|\,X_{n-1}=y]\\ &= \mathbb{P}[\xi_{n,1}+\cdots+\xi_{n,y}=x\,|\,X_{n-1}=y]\\ &= \mathbb{P}[\xi_{n,1}+\cdots+\xi_{n,y}=x]\\ &= p^{*y}_x.\end{aligned}$$

3.4 Remark

(i) The notation p^{*y}_x should be understood as $(p^{*y})_x$, i.e., p^{*y}_x is the x^{th} coordinate of the convoluted vector $\mathbf{p}^{*y}=\mathbf{p}*\mathbf{p}*\cdots*\mathbf{p}$ (y times).

(ii) There is no typo in the formula $p^n_{x|y}=p^{*y}_x$. The **conditional** probability $p^n_{x|y}$ is indeed independent of n.

Now, we only need to **uncondition** by using the law of total probability to get rid of the **conditioning trick** $\{X_{n-1} = y\}$:

$$\begin{aligned} p_x^n &= \mathbb{P}[X_n = x] \\ &= \sum_y \mathbb{P}[X_n = x \,|\, X_{n-1} = y]\,\mathbb{P}[X_{n-1} = y] \\ &= \sum_y p_{x|y}^n \; p_y^{n-1} \\ &= \sum_y p_x^{*y} \; p_y^{n-1}. \end{aligned}$$

Finally, we note that we must interpret

$$\begin{aligned} p_x^{*0} &= p_{x|0}^n \\ &= \mathbb{P}[X_n = x \,|\, X_{n-1} = 0] \\ &= \begin{cases} 1, & \text{if } x = 0, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Indeed, an extinct family line remains extinct.

We have obtained the following:

3.5 Proposition (Branching distributions)
The n^{th} generation distribution of a branching process with offspring distribution $\mathbf{p}$ is given by the recursion

$$\begin{aligned} p_x^1 &= p_x, \\ p_x^n &= \sum_y p_x^{*y} \; p_y^{n-1}, \quad \text{for} \quad n > 1, \end{aligned}$$

with the convention that

$$p_x^{*0} = \begin{cases} 1, & \text{if } x = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Computing Random Sums

The result of Proposition 3.5 is very difficult analyze. However, it is not **too difficult** to implement it for numerical calculations, if the offspring distribution $\mathbf{p}$ is a vector of **finite** length. This is done by using the following Octave functions.

```
1 function pn = conv_power(p,n)
2 ## Function pn = conv_power(p, n) returns the n:th convolution power of the
3 ## vector p.
```

```
4 ##
5 ## See also: conv
6
7     if (n == 1)
8         pn = p;
9     else
10        pn = conv( p, conv_power(p,n-1) );
11    endif
12 endfunction
```

www.uva.fi/~tsottine/psp/conv_power.m

The function `conv_power` should be self-explanatory. It is a simple **recursive** code that uses the Octave's built-in function `conv`. The recursion call, i.e., **the beef** of the function, is in line 10. Lines 7–8 are the end of the recursion that prevent the code from entering an infinite loop (if the function is called properly).

```
1  function prob = cond_branching_pdf(x, p, y)
2  ## Function prob = cond_branching_pdf(x, p, y) returns the probability that the
3  ## n:th generation of a branching process with offspring distribution p has
4  ## exactly x offspring, given that the (n-1):th generation has exactly y
5  ## offspring.
6  ##
7  ## Uses function conv_power.
8
9     ## Maximum offspring size
10    k = length(p)-1;
11
12    if (x > k*y)
13        prob = 0;
14    elseif (y == 0)
15        if (x == 0)
16            prob = 1;
17        else
18            prob = 0;
19        endif
20    elseif (y == 1)
21        prob = p(x+1);
22    else
23        prob = conv_power(p,y)(x+1);
24    endif
25 endfunction
```

www.uva.fi/~tsottine/psp/cond_branching_pdf.m

The function `cond_branching_pdf` is implicitly recursive via the call to the recursive function `conv_power` in line 23. The function `cond_branching_pdf` checks many obvious cases before resorting into the recursion call. The reason for this is that, first, the recursion call is time-consuming, and second, we want to be sure that the function `conv_power` is called with proper arguments.

```
1 function prob = branching_pdf(x, p, n)
2 ## Function prob = branching_pdf(x, p, n) returns the probability that the n:th
3 ## generation of a branching process with offspring distribution p has exactly
4 ## x offspring.
```

```
5  ##
6  ## Uses functions  cond_brancing_pdf and  conv_power.
7
8      ## Maximum offspring size
9      k = length(p)-1;
10     ## Maximum generation n size.
11     kn = k^n;
12
13     if (x > kn)
14         prob = 0;
15     elseif (n == 1)
16         prob = p(x+1);
17     else
18         prob = 0;
19         for y=0:kn
20             prob = prob + cond_branching_pdf(x,p,y)*branching_pdf(y,p,n-1);
21         endfor
22     endif
23 endfunction
```

www.uva.fi/~tsottine/psp/branching_pdf.m

The function `branching_pdf` works pretty much the same way as the function `cond_branching_pdf`. The main difference is the `for` loop in lines 19–21, where we calculate the sum

$$\sum_y p_x^{*y} p_y^{n-1}.$$

By running the code below, we get the solution to the part (a) of Example 3.1 (which is the same as Example 1.1 and Example 2.1).

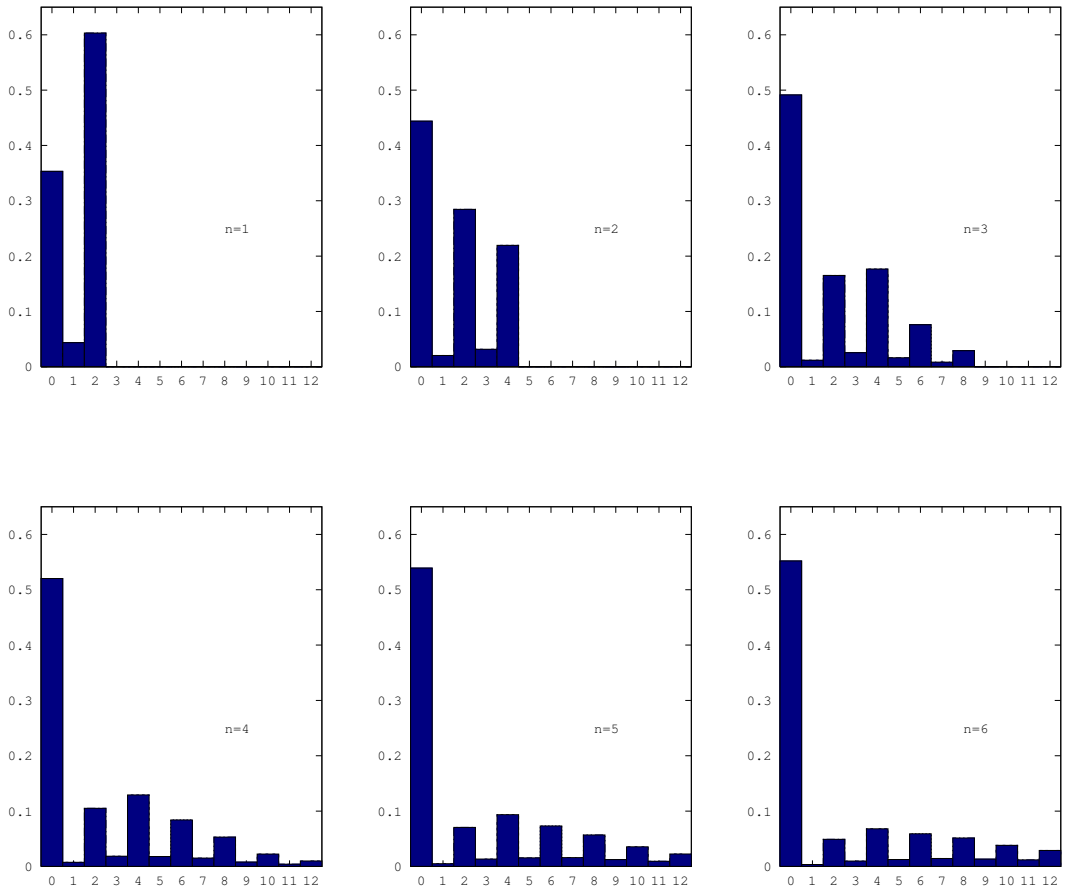
```
1  ##############################################################################
2  ## FILE: perse_pasquale_a.m
3  ##
4  ## Probability of more than 10 descendants in the 6:th generation of the
5  ## Perse–Pasquale family tree.
6  ##
7  ## N.B.  The calculations will take a while.
8  ##
9  ## Uses functions  conv_power,  cond_branching_pdf,  branching_pdf
10 ##############################################################################
11
12 ## data is the frequencies.
13 data = [503 62 859];
14
15 ## Offspring distribution is the relative frequencies.  Note that Octave starts
16 ## indexing with 1.  So p(1) is the probability of 0 offspring.
17 p = data/sum(data);
18
19 ## Solution to Example 1.1 part (a)
20 n = 6;
21 K = 10;
22
23 prob = 0;
24 for x=0:K
```

```
25      prob = prob + branching_pdf(x, p, n);
26  endfor
27
28  sol_a = 1-prob;
```

www.uva.fi/~tsottine/psp/perse_pasquale_a.m

3.6 Example (Persé–Pasquale Family Tree, II, Solution (a))
Running the m-file `perse_pasquale_a.m` in the Octave console gives us (after some 8 minutes) the solution $\mathbb{P}[X_6 > 10] = 0.12895$.

Let us end this lecture by visualizing the first six generations of the Persé-Pasquale family. We note that it took 12 minutes to calculate the plot by using the code below. So, our algorithm is not very good or useful.



Probability mass functions of the first six generations of Example 2.1.

The graphs above where generated by the m-file listed below. The code should be relatively self-explanatory, so we do not explain it here.

```
################################################################################
## FILE: perse_pasquale_pmfs.m
##
## Plots distributions of generations of a branching process.
################################################################################

################################################################################
## Frequencies and the offspring distribution.
################################################################################

data = [503 62 859];
p = data/sum(data);

################################################################################
## Calculations.
################################################################################

max_off = 12;
x = 0:max_off;
Px = zeros(6, max_off+1);
for x0 = x
    Px(1, x0+1) = branching_pdf(x0, p, 1);
    Px(2, x0+1) = branching_pdf(x0, p, 2);
    Px(3, x0+1) = branching_pdf(x0, p, 3);
    Px(4, x0+1) = branching_pdf(x0, p, 4);
    Px(5, x0+1) = branching_pdf(x0, p, 5);
    Px(6, x0+1) = branching_pdf(x0, p, 6);
endfor

################################################################################
## Plotting (bar plots).
################################################################################

## Width of the bar in the bar plot.
w = 1;

## Plotting window [x1, x2, y1, y2].
y_max = 0.65;
plotlims = [-0.5, max_off+0.5, 0, y_max];

subplot(2,3,1);                         # 2 rows, 2 columns, 1st plot.
    bar(x,Px(1,:), w);
    text(max_off-4, 0.25, 'n=1');
    axis(plotlims);
subplot(2,3,2);                         # 2 rows, 2 columns, 2nd plot.
    bar(x, Px(2,:), w);
    text(max_off-4, 0.25, 'n=2');
    axis(plotlims);
subplot(2,3,3);                         # 2 rows, 2 columns, 3rd plot.
    bar(x, Px(3,:), w);
    text(max_off-4, 0.25, 'n=3');
    axis(plotlims);
subplot(2,3,4);                         # 2 rows, 2 columns, 4th plot.
```

```
54      bar(x, Px(4,:), w);
55      text(max_off-4, 0.25, 'n=4');
56      axis(plotlims);
57  subplot(2,3,5);                          # 2 rows, 2 columns, 5th plot.
58      bar(x, Px(5,:), w);
59      text(max_off-4, 0.25, 'n=5');
60      axis(plotlims);
61  subplot(2,3,6);                          # 2 rows, 2 columns, 6th plot.
62      bar(x, Px(6,:), w);
63      text(max_off-4, 0.25, 'n=6');
64      axis(plotlims);
```

www.uva.fi/~tsottine/psp/perse_pasquale_pmfs.m

Exercises

3.1 Exercise
Calculate the distributions of the 1^{st}, 2^{nd} and 3^{rd} generations of the branching process with offspring distributions

(a) [0.5000 0.0000 0.5000],
(b) [0.3333 0.3333 0.3333],
(c) [0.3000 0.5000 0.2000],
(d) [0.2000 0.5000 0.3000].

3.2 Exercise (Flatus Lake, II)
Consider Exercise 1.3

(a) Calculate the distribution of the n^{th} generation offspring of a single Flatus bacteria for $n = 1, 2, 3$.
(b) Calculate the distribution of the n^{th} generation offspring of a single Virilus Flatus bacteria for $n = 1, 2, 3$.

3.3 Exercise
Visualize the distributions of the 1^{st}, 2^{nd}, 3^{rd}, 4^{th}, 5^{th}, 6^{th}, 7^{th} and 8^{th} generations of the branching process with offspring distributions

(a) [0.7500 0.2500],
(b) [0.5000 0.5000],
(c) [0.2500 0.7500],
(d) [0.0001 0.9999].

3.4 Exercise
Suppose a branching process has the offspring distribution $[0.1 \;\; 0.0 \;\; 0.2 \;\; 0.0 \;\; 0.4 \;\; 0.3]$.

(a) How many descendants can there be in generation 127?
(b) What are the possible number of descendants in generation 6?
(c) What are the possible number of descendants in generation 7?

3.5 Exercise
Calculate the 7^{th} generation distribution for a branching process with offspring distribution $[0.15 \;\; 0.45 \;\; 0.30 \;\; 0.00 \;\; 0.00 \;\; 0.10]$.

N.B.: As far as the author knows, this problem is way beyond the capabilities of the algorithm used in the Octave function `branching_pdf` .

Lecture 4

Analytical Tools

Brook Taylor (1685–1731) was an English mathematician who is best known for Taylor's theorem and the Taylor series. His work *Methodus Incrementorum Directa et Inversa* (1715) added a new branch to higher mathematics, now called the "calculus of finite differences". Among other ingenious applications, he used it to determine the form of movement of a vibrating string, by him first successfully reduced to mechanical principles. The same work contained the celebrated formula known as Taylor's formula, the importance of which remained unrecognized until 1772, when **Joseph-Louise Lagrange** realized its powers and termed it "the main foundation of differential calculus".

In probability we are often interested in calculating convolutions, since they are the sums of independent random variables. Unfortunately, the convolution is a complicated operation. However, by using the Taylor series "backwards" we obtain the so-called probability generating function, for which the calculation of convolutions becomes easy. Unfortunately, the probability generating functions only work for $\mathbb{N}$-valued random variables. Fortunately, for general random variables there are related transformations: the moment generating functions and the characteristic functions.

Brook Taylor (1685–1731)

Finally, we are ready to turn into the final problem (d) of the Example 1.1 recalled below as Example 4.1. This problem is, as far as the author knows, very difficult to solve without using analytical tools (probability generating functions, moment generating functions, or characteristic functions). With analytical tools the problem is, in theory, quite simple. However, in practice, to compute the explicit solution one typically has to resort into numerical methods.

In this lecture, we will solve the final problem (d) of Example 4.1 by using an analytical tool called the probability generating function. Then, we will briefly introduce the related analytical tools: moment generating functions and characteristic functions, although they are not needed for the solution. Note that any of of these three analytical tools could have been used to solve the problem, and in practice any of them would work just as easily.

4.1 Example (Persé–Pasquale Noble Family Tree, IV)
The most noble family of Persé–Pasquale is worried of their continuing existence. At the moment there is only one male descendants of this most noble line. According to family records, the males of the noble family of Persé–Pasquale have sired male children as follows

Number of male children	Frequency
0	503
1	62
2	859
More than 2	0

(a) What is the probability that the 6^{th} generation has more than 10 male descendants?
(b) What is the average number of descendants in the 6^{th} generation?
(c) What is the variance of the number of descendants in the 6^{th} generation?
(d) What is the probability that the Persé–Pasquale family will be ultimately extinct?

The extinction problem for the branching process X_n, $n \in \mathbb{N}$, with offspring distribution $\mathbf{p}$ is to calculate the **ultimate extinction probability**

$$\begin{aligned}\rho &= \mathbb{P}[\text{ultimate extinction}] \\ &= \mathbb{P}\Big[X_n = 0 \text{ for some } n\Big] \\ &= \mathbb{P}\Big[X_n = 0 \text{ eventually }\Big] \\ &= \mathbb{P}\left[\bigcup_n \{X_n = 0\}\right].\end{aligned}$$

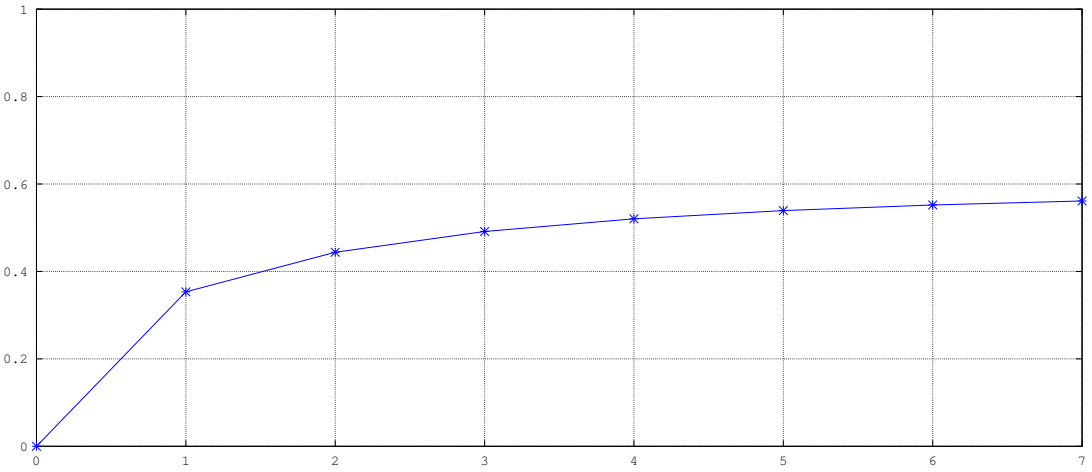
Denote

$$\begin{aligned}\rho_n &= \mathbb{P}\Big[\text{extinction of the } n^{\text{th}} \text{ generation}\Big] \\ &= \mathbb{P}[X_k = 0 \text{ for some } k \le n] \\ &= \mathbb{P}\left[\bigcup_{k\le n} \{X_k = 0\}\right] \\ &= \mathbb{P}[X_n = 0].\end{aligned}$$

A simple-minded way to calculate the ultimate extinction probability would be to note that, because of the **monotonic continuity of probability**,

$$\rho = \lim_{n\to\infty} \rho_n.$$

Now $\rho_n = p_0^n$, which can in principle be calculated by using Proposition 3.5. Below is an illustration of p_0^n for $n = 0, 1, 2, 3, 4, 5, 6, 7$. (With $n = 8$ my Octave calculated 4 hours and then crashed.)



Finite generations extinction probabilities for Example 3.1.

The plot above was generated by the m-file below.

```
##############################################################################
## FILE: perse_pasquale_extictions.m
##
## Plots extinction probabilites within given generations.
##############################################################################

##############################################################################
## Data and parameters, and initializations
##############################################################################

data = [503 62 859];                          ## Frequencies
p = data/sum(data);                           ## Offspring distribution
nmax = 7;                                     ## Maximum number of generations
n = 1:nmax;
probn = zeros(1,nmax);                        ## Initialization for speeding up

##############################################################################
## Calculations (that take some sweet time)
##############################################################################

for n0 = n
    probn(n0) = branching_pdf(0, p, n0);
endfor

##############################################################################
## Plotting
##############################################################################

plot([0 n], [0 probn], 'marker', '*')
plotlims = [0, nmax, 0, 1];
axis(plotlims);
```

```
32 grid on;
```

www.uva.fi/~tsottine/psp/perse_pasquale_extinctions.m

We see some possible convergence in the plot for p_0^n, $n = 0,1,2,3,4,5,6,7$. It seems that the probability of ultimate extinction is something like little less than 0.6. However, in order to be quite certain we have hit the neighborhood of the true ultimate extinction probability $\rho = \lim p_0^n$, we have to calculate p_0^n for large values of n. Unfortunately, this is not practical. Indeed, it already took 1 hour to calculate p_0^7. Calculating something like p_0^{50}, say, is not possible in practice. At least not with the recursive algorithm implemented in the Octave function `branching_pdf`. A practical solution is given by using an analytical tool called the **probability generating function**. We introduce the probability generating functions later after we have discussed the **Taylor's expansion**.

4.2 Remark (Almost Sure Extinction)
Before going into any further analysis, let us try to see if there is an easy solution. Let μ be the average number of males sired by a given noble of the Persé–Pasquale family, i.e, $\mu = \mathbb{E}[\xi_{n,i}]$. Suppose $\mu < 1$. Then the extinction happens for sure. Indeed, since the expected number of descendants in generation n is μ^n, we see that $\mathbb{E}[X_n] \to 0$. Now we recall the **Markov's inequality** for non-negative random variables X:

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}, \qquad \text{for all} \quad a > 0.$$

Consequently

$$\begin{aligned} \mathbb{P}[X_n > 0] &= \mathbb{P}[X_n \geq 1] \\ &\leq \mathbb{E}[X_n] \\ &= \mu^n. \end{aligned}$$

Therefore $\mathbb{P}[X_n > 0] \to 0$, which means that the extinction will happen eventually. If $\mu = 1$, it is also possible (but more technical) to argue that the extinction will happen eventually, unless there is no randomness in the branching process.

For Persé–Pasquale family we have $\mu = 1.25$. Consequently, the extinction does not appear to be certain. We have to analyze further.

Taylor Approximation

The key idea in the analytical tools presented here comes from the Taylor's polynomial approximation of smooth functions. (We note that there are better versions of the Taylor's approximation theorem that do not require the existence of the $(n+1)^{\text{th}}$ derivative as Lemma 4.3 below does.)

4.3 Lemma (Taylor's Approximation)
Let $f(x)$ be a function that has $n+1$ continuous derivatives at point a. Then we can approximate

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \varepsilon_n(x)(x-a)^n,$$

where $\varepsilon_n(x) \to 0$ as $x \to a$.

We assume that the reader is familiar with the Taylor's approximation. However, just in case the reader has forgotten, we briefly explain why it is true by using the case $n = 2$ as an example; the case for general n is then easy to see.

By the **fundamental theorem of calculus** we have

$$f(x) = f(a) + \int_a^x f'(y)\,\mathrm{d}y.$$

Now, let us use the fundamental theorem of calculus twice more. We obtain

$$\begin{aligned} f(x) &= f(a) + \int_a^x f'(y)\,\mathrm{d}y \\ &= f(a) + \int_a^x \left[f'(a) + \int_a^y f''(z)\,\mathrm{d}z \right] \mathrm{d}y \\ &= f(a) + \int_a^x \left[f'(a) + \int_a^y \left[f''(a) + \int_a^z f'''(v)\,\mathrm{d}v \right] \mathrm{d}z \right] \mathrm{d}y. \end{aligned}$$

Then, by using the linearity of the integral, we obtain

$$\begin{aligned} f(x) &= f(a) + \int_a^x \left[f'(a) + \int_a^y \left[f''(a) + \int_a^z f'''(v)\,\mathrm{d}v \right] \mathrm{d}z \right] \mathrm{d}y \\ &= f(a) + \int_a^x f'(a)\,\mathrm{d}y + \int_a^x \int_a^y f''(a)\,\mathrm{d}z\mathrm{d}y + \int_a^x \int_a^y \int_a^z f'''(v)\,\mathrm{d}v\mathrm{d}z\mathrm{d}y \\ &= f(a) + f'(a)\int_a^x \mathrm{d}z + f''(a)\int_a^x \int_a^y \mathrm{d}v\mathrm{d}z + \int_a^x \int_a^y \int_a^z f'''(v)\,\mathrm{d}v\mathrm{d}z\mathrm{d}y \\ &= f(a) + f'(a)(x-a) + f''(a)\frac{(x-a)^2}{2} + \int_a^x \int_a^y \int_a^z f'''(v)\,\mathrm{d}v\mathrm{d}z\mathrm{d}y. \end{aligned}$$

Thus, in order to see that Lemma 4.3 is true for $n = 2$, it remains to show that

$$\int_a^x \int_a^y \int_a^z f'''(v)\,\mathrm{d}v\mathrm{d}z\mathrm{d}y = \varepsilon_2(x)(x-a)^2,$$

where $\varepsilon_2(x) \to 0$ as $x \to a$. Now, since $f'''(v)$ is continuous, it is bounded around a with a number C, say. Therefore,

$$\begin{aligned}\left|\int_a^x\int_a^y\int_a^z f'''(v)\,\mathrm{d}v\mathrm{d}z\mathrm{d}y\right| &\leq \int_a^x\int_a^y\int_a^z C\,\mathrm{d}v\mathrm{d}z\mathrm{d}y\\ &= C\frac{(x-a)^3}{3!}\\ &= C\frac{(x-a)}{3!}(x-a)^2.\end{aligned}$$

This shows Lemma 4.3 for the case $n = 2$. The case for general n can be seen easily by iterating the arguments above, although the formulas become quite messy.

4.4 Remark (Analytic Functions)
Sometimes a function $f(x)$ can be expressed as its Taylor expansion, or more precisely, as its **Taylor series** (around point $a = 0$) as

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!}x^k.$$

Such functions are called **analytic**. Examples include of course all polynomials and the following common functions:

$$\begin{aligned}\mathrm{e}^x &= \sum_{k=0}^{\infty}\frac{1}{k!}x^k,\\ \ln(1+x) &= \sum_{k=1}^{\infty}\frac{(-1)^{k-1}}{k}x^k, \quad \text{for } |x|<1,\\ \sin x &= \sum_{k=0}^{\infty}\frac{(-1)^k}{(2k+1)!}x^{2k+1},\\ \cos x &= \sum_{k=0}^{\infty}\frac{(-1)^k}{(2k)!}x^{2k}.\end{aligned}$$

Actually, almost all of the functions one learns at school are analytic.

4.5 Remark
In Remark 4.4 above, $f^{(k)}(x)$ denotes the k^{th} derivative, and, as always, the 0^{th} derivative is the function itself: $f^{(0)}(x) = f(x)$.

Probability Generating Functions

The probability generating function can be thought of being the inverse of a Taylor expansion, where the coefficients, that is, the derivatives at point $a = 0$, correspond to the probabilities.

4.6 Definition (Probability Generating Function)
The **probability generating function** $G(\theta)$ of an $\mathbb{N}$-valued random variable X is

$$G(\theta) = \sum_x \mathbb{P}[X = x]\theta^x.$$

4.7 Remark
In Definition 4.6 we have, as always, $0^\theta = 0$ for $\theta \neq 0$ and $0^0 = 1$.

4.8 Remark
Definition 4.6 can also be expressed as

$$G(\theta) = \mathbb{E}\left[\theta^X\right],$$

and this, taken as the definition, in principle, works for more general random variables than the $\mathbb{N}$-valued ones. So, the probability generating function is closely related to the **Mellin transform** used in analysis.

The name "probability generating function" comes from Lemma 4.9 below, which also further elaborates how the probability generating function is a kind of an inverse of Taylor series. Also, it follows from Lemma 4.9 that the probability mass function $\mathbb{P}[X = x]$ can be recovered from the probability generating function $G(\theta)$, and vice versa. To see why Lemma 4.9 below is true, one only needs to derivate. Indeed, after differentiation, all the powers less than x vanish and all the powers greater that x vanish when evaluated at $x = 0$. Only the power x remains. (Try it!)

4.9 Lemma (Probability Generating Function)
Let X be an $\mathbb{N}$-valued random variable with probability generating function $G(\theta)$. Then

$$\mathbb{P}[X = x] = \frac{G^{(x)}(0)}{x!}.$$

4.10 Remark
In Lemma 4.9 above, as always, $0! = 1$. Also note that $G^{(0)}(0) = G(0)$, by definition of the 0^{th} derivative, and

$$G(0) = \mathbb{P}[X = 0],$$

since

$$G(0) = \sum_x \mathbb{P}[X = x]0^x,$$

and $\theta^x = 0$ for $x \neq 0$ and $0^0 = 1$.

Probability generating functions are useful, because they transform the nasty business of taking convolutions into the simple operation of multiplication. Indeed, suppose X and Y are independent $\mathbb{N}$-valued random variables with probability generating functions G_X and G_Y, respectively. Then

$$\begin{aligned} G_{X+Y}(\theta) &= \mathbb{E}\left[\theta^{X+Y}\right] \\ &= \mathbb{E}\left[\theta^X \theta^Y\right] \\ &= \mathbb{E}\left[\theta^X\right]\mathbb{E}\left[\theta^Y\right] \\ &= G_X(\theta)G_Y(\theta). \end{aligned}$$

By iterating the argument above, we obtain the following lemma.

4.11 Lemma (Probability Generating Functions for Independent Sums)
Let $X_1, X_2, \ldots, X_n$ be independent $\mathbb{N}$-valued random variables with probability generating functions $G_{X_1}(\theta)$, $G_{X_2}(\theta)$, $\ldots$, $G_{X_n}(\theta)$. Then the sum

$$S = X_1 + X_2 + \cdots + X_n$$

has the probability generating function

$$G_S(\theta) = G_{X_1}(\theta)G_{X_2}(\theta)\cdots G_{X_n}(\theta).$$

Probability generating functions are particularly useful for dealing with independent random sums, as the following proposition shows.

4.12 Lemma (Probability Generating Functions for Independent Random Sums) Suppose $X_1, X_2, \ldots$, are independent and identically distributed $\mathbb{N}$-valued random variables. Suppose that N is an independent $\mathbb{N}$-valued random variable. Let

$$S = X_1 + X_2 + \cdots + X_N.$$

Then the probability generating function of S is

$$G_S(\theta) = G_N(G_X(\theta)),$$

where $G_X(\theta)$ is the common probability generating function of the summands X_i.

Lemma 4.12 follows from the following chain or arguments, where the key **conditioning trick** is to condition on the number of summands N and then use independence:

$$\begin{aligned} G_S(\theta) &= \mathbb{E}\left[\theta^{X_1+\cdots+X_N}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\theta^{X_1+\cdots+X_N}|N\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\theta^{X_1}|N\right]\cdots\mathbb{E}\left[\theta^{X_N}|N\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\theta^{X_1}\right]\cdots\mathbb{E}\left[\theta^{X_N}\right]\right] \\ &= \mathbb{E}\left[G_X(\theta)^N\right] \\ &= G_N(G_X(\theta)). \end{aligned}$$

Lemma 4.12 is useful for branching processes. Indeed, let X_n, $n \in \mathbb{N}$, be a branching process with offspring distribution $\mathbf{p}$. Let $G_n(\theta)$ be the probability generating function of X_n and let $G(\theta)$ be the probability generating function of the offspring distribution. Obviously $G_1(\theta) = G(\theta)$. For $G_2(\theta)$ we note that $X_2 = \xi_{2,1} + \xi_{2,2} + \cdots + \xi_{2,X_1}$, where the $\xi_{i,j}$'s are independent and identically distributed with probability generating function $G(\theta)$. Consequently, by using Lemma 4.12 with $N = X_1$, we have

$$\begin{aligned} G_2(\theta) &= G_1(G(\theta)) \\ &= G(G(\theta)) \\ &= G \circ G(\theta). \end{aligned}$$

The last line here is just the definition of **composition** for functions. The same argument for $X_n = \xi_{n,1} + \xi_{n,2} + \cdots + \xi_{n,X_{n-1}}$ yields

$$G_n(\theta) = G_{n-1}(G(\theta)).$$

Now iterating the recursion above backwards with n, we obtain

$$\begin{aligned} G_n(\theta) &= G_{n-2}(G(G(\theta)) \\ &= G_{n-2}(G \circ G(\theta)) \\ &= G_{n-3}(G \circ G \circ G(\theta)) \\ &= \cdots \\ &= \underbrace{G \circ \cdots \circ G}_{n \text{ times}}(\theta). \end{aligned}$$

Since the last line above is (by definition) the n^{th} **composition power** $G^{\circ n}(\theta)$ of the function $G(\theta)$, we have found out the following truth.

4.13 Proposition (Branching probability generating functions)
Let X_n, $n \in \mathbb{N}$, be a branching process with offspring distribution $\mathbf{p}$. Let $G(\theta)$ be the probability generating function of the offspring distribution. Then the probability generating function of the n^{th} generation is $G^{\circ n}(\theta)$.

Proposition 4.13 provides a new way to calculate the distribution of the n^{th} generation of the branching process. Indeed,

$$\begin{aligned} p_x^n &= \frac{1}{x!} G_n^{(x)}(0) \\ &= \frac{1}{x!} (G^{\circ n})^{(x)}(0). \end{aligned}$$

Unfortunately, this formula is also recursive and, moreover, includes differentiation, which may turn out to be very unstable and time-consuming to compute with Octave. For the n^{th} generation extinction probabilities we have the formula

$$\begin{aligned} p_0^n &= G_n(0) \\ &= G^{\circ n}(0). \end{aligned}$$

This formula does not look very attractive, either.

Let us get back to Example 4.1. To solve it, we use once again a **conditioning trick**. Suppose that at generation 1 we have $X_1 = x$, say. Then in order for the extinction to happen, each one of these x lines must go extinct individually. By independence and symmetry, the probability for this is ρ^x. Consequently, the law of total probability gives us

$$\begin{aligned} \rho &= \mathbb{P}[X_n = 0 \text{ for some } n] \\ &= \sum_x \mathbb{P}[X_1 = x] \mathbb{P}[X_n = 0 \text{ for some } n \,|\, X_1 = x] \\ &= \sum_x p_x \rho^x. \end{aligned}$$

Now, by noticing the probability generating function in the equation above, we obtain the following.

4.14 Theorem (Branching Extinction)
The probability of ultimate **extinction** for a branching process is given by the smallest positive root of the equation

$$\rho = G(\rho),$$

where $G(\theta)$ is the probability generating function of the offspring distribution.

Now, solving the equation of Theorem 4.14 can sometimes be very tedious and most often analytically simply impossible. In Example 4.1 the offspring distribution has non-zero values for only 0, 1, and 2: $\mathbf{p} = [p_0 \;\; p_1 \;\; p_2]$, which means that we are left with a quadratic equation. This we can solve easily analytically.

4.15 Example (Persé-Pasquale Family Tree, IV, Solution (d))
We need to solve, for $\mathbf{p} = [0.353230 \;\; 0.043539 \;\; 0.603230]$,

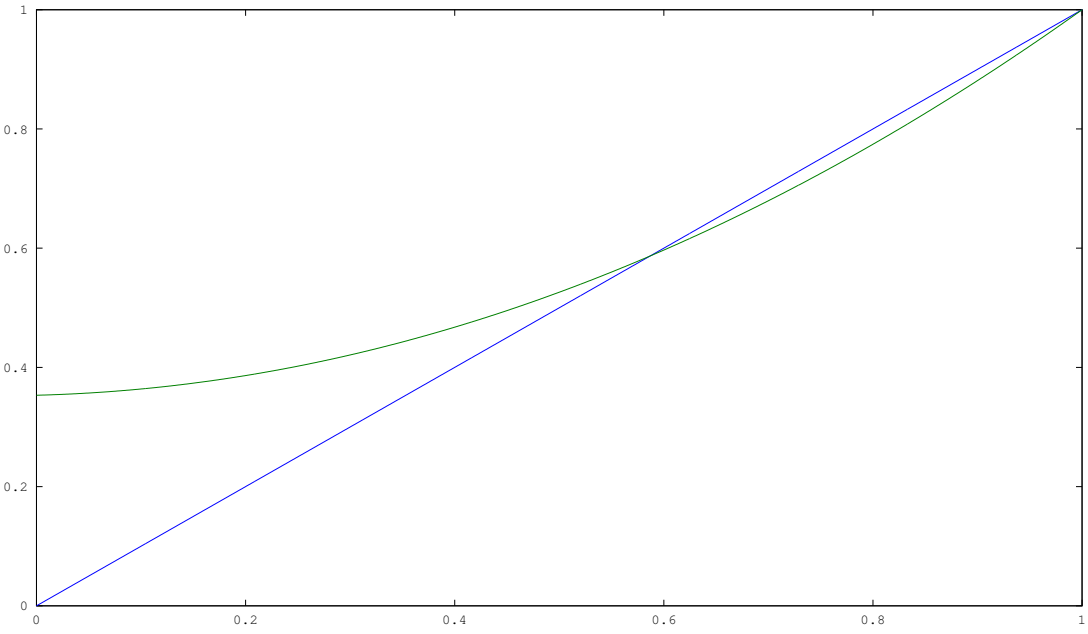
$$\rho \;\; = \;\; p_0 + p_1\rho + p_2\rho^2.$$

In standard form this quadratic equation reads

$$p_2\rho^2 + (p_1 - 1)\rho + p_0 \;\; = \;\; 0.$$

Therefore, the smallest positive root is obtained by taking "$-$" in the "$\pm$" in the quadratic formula:

$$\rho \;\; = \;\; \frac{1 - p_1 - \sqrt{(p_1-1)^2 - 4p_2p_0}}{2p_2} \;\; = \;\; 0.58556$$



Extinction probability for Example 3.1.

Note that even though there is a huge 25 % growth for each generation, the family will more likely than not go extinct eventually.

Let us note that the solution for the equation $\rho = G(\rho)$ was easy, since we had $p_x = 0$ for $x > 2$. This lead into a quadratic equation, for which we have a simple analytical

solution. For higher degree equations, the analytical solutions are very difficult to come by, or outright impossible. Therefore, one must resort to numerical methods. If the offspring distribution is finite, then the resulting equation is a polynomial equation. With Octave such equations can be solved by using the function `roots`. If the offspring distribution is not finite, then one has to use more general root-finding methods. In this case Octave function `fzero` might be useful.

Moment Generating Functions

Probability generating functions work well for $\mathbb{N}$-valued random variables. For general real valued random variables we must use other tools. Indeed, e.g., the sum

$$G(\theta) \quad = \quad \sum_x \mathbb{P}[X = x]\,\theta^x$$

would be identically zero for continuous distributions X. The definition

$$G(\theta) \quad = \quad \mathbb{E}\left[\theta^X\right]$$

would work better, but then there is a problem of understanding the power θ^X for non-integer X. This can be done by using the exponential function, or even better, by using a change of variables. This gives us the following definition.

4.16 Definition (Moment Generating Function)
Let X be a random variable. Its **moment generating function** is

$$M(\theta) \quad = \quad \mathbb{E}\left[\mathrm{e}^{\theta X}\right].$$

If X is $\mathbb{N}$-valued, then the probability and moment generating functions are connected by a simple change of variables:

$$M(\theta) \quad = \quad G(\mathrm{e}^{\theta}).$$

If X has continuous distribution with probability density function $f(x)$, then its probability generating function is

$$M(\theta) \quad = \quad \int_{-\infty}^{\infty} \mathrm{e}^{\theta x} f(x)\,\mathrm{d}x.$$

So, in this case the moment generating function is basically the **Laplace transform** of the density function $f(x)$.

The name "moment generating" function comes from the following. Recall that the n^{th} **moment** of the random variable X is

$$m_n \quad = \quad \mathbb{E}[X^n].$$

Now, expand the exponential function as Taylor's series

$$e^{\theta X} = 1 + \theta X + \frac{\theta^2}{2!}X^2 + \frac{\theta^3}{3!}X^3 + \cdots + \frac{\theta^n}{n!}X^n + \cdots .$$

By taking expectations on both sides, we see that

$$M(\theta) = 1 + \theta m_1 + \frac{\theta^2}{2!}m_2 + \frac{\theta^3}{3!}m_3 + \cdots + \frac{\theta^n}{n!}m_n + \cdots .$$

By differentiating both sides we obtain the following.

4.17 Lemma (Moment Generating Function)
The n^{th} moment of the random variable X with moment generating function $M(\theta)$ can be calculated as

$$m_n = M^{(n)}(0).$$

Similarly to probability generating functions, also moment generating functions determine the distribution uniquely. Unfortunately, this is not so easy to see as in the case of probability generating functions. One would hope that the moments would define the distribution uniquely, but unfortunately that would not be completely true. We state the following important lemma without any proof or idea of its validity.

4.18 Lemma (Moment Generating Functions Define Distributions)
Let two random variables have the same moment generating function, then their distributions are the same.

Finally, let us point out that the results on sums of random variables in terms of the probability generating functions translate in the obvious way to moment generating functions.

Characteristic Functions

There is a small problem with moment generating functions: not all random variables admit one. The problem is that in order for the expectation $\mathbb{E}[e^{\theta X}]$ to exist, the random variable X must have **light tails**, i.e., the probabilities $\mathbb{P}[|X| \geq x]$ must converge to zero exponentially fast. This is not true for all random variables. An example is the standard **Cauchy distribution** that is a continuous distribution with probability density function

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

Indeed, for the Cauchy distribution, none of the moments m_n, $n \in \mathbb{N}$, exist. For $n = 1$ a simple-minded one would like to assign, by symmetry, that $\mathbb{E}[X] = 0$. Unfortunately, we need more for the expectation: we need $\mathbb{E}[|X|] < \infty$ for $\mathbb{E}[X]$ to make sense in, say, the **law of large numbers**. But for the standard Cauchy distributed random variable X, we have

$$\mathbb{E}[|X|] \;=\; \int_{-\infty}^{\infty} |x| f(x)\,\mathrm{d}x \;=\; \frac{2}{\pi}\int_0^{\infty} \frac{x}{1+x^2}\,\mathrm{d}x \;=\; \infty.$$

The analytical tool for the fat tailed random variables is provided by complex analysis. Let $z = x + \mathrm{i}y \in \mathbb{C}$. Recall the **complex exponential**

$$\mathrm{e}^z \;=\; \mathrm{e}^x(\cos y + \mathrm{i}\sin y).$$

Now, while the real exponential $\mathrm{e}^{\theta x}$ can easily be very large, the imaginary exponential $\mathrm{e}^{\mathrm{i}\theta x}$ is bounded for all $x \in \mathbb{R}$. Indeed,

$$\left|\mathrm{e}^{\mathrm{i}\theta x}\right| \;=\; \left|\cos(\theta x) + \mathrm{i}\sin(\theta x)\right| \;=\; \sqrt{\cos^2(\theta x) + \sin^2(\theta x)} \;=\; 1.$$

This suggests the following definition.

4.19 Definition (Characteristic Function)
Let X be a random variable. Its **characteristic function** is

$$\varphi(\theta) \;=\; \mathbb{E}\left[\mathrm{e}^{\mathrm{i}\theta X}\right] \;=\; \mathbb{E}[\cos(\theta X)] + \mathrm{i}\,\mathbb{E}[\sin(\theta X)].$$

The connection between characteristic function and moment generating function, when the latter exists, is a simple one:

$$\varphi(-\mathrm{i}\theta) \;=\; M(\theta).$$

Actually, this is basically the same connection that one has between the **Laplace transform** and the **Fourier transform**. Indeed, the moment generating function is, basically, the Laplace transform, and the characteristic function is, basically, the Fourier transform.

The characteristic function exists for all random variables. Also, the characteristic function defines the distribution uniquely. Moreover, the characteristic functions characterize the convergence in distribution (hence the name, the author guesses). This last statement is called the **Lévy's continuity theorem**. To emphasize the importance of the last statement, me make it a lemma. Before that we recall the notion of convergence in distribution.

4.20 Definition (Convergence in Distribution)
Let $X_1, X_2, \ldots$ be random variables with **cumulative distribution functions**

$$F_n(x) \;=\; \mathbb{P}[X_n \le x].$$

Let X be a random variable with cumulative distribution function $F(x)$. Then

$$X_n \xrightarrow{d} X,$$

i.e., X_n converges to X **in distribution** if

$$F_n(x) \to F(x)$$

for all x for which $F(x)$ is continuous.

The notion of convergence in distribution is a bit complicated. However, it is precisely the notion one needs for the **central limit theorem**, which in turn is relatively easy to prove once one knows the following lemma.

4.21 Lemma (Lévy's Continuity Theorem)
Let $X_1, X_2, \ldots$ be random variables with characteristic functions $\varphi_{X_1}(\theta), \varphi_{X_2}(\theta), \ldots$. Let X be a random variable with characteristic function $\varphi_X(\theta)$. Then

$$X_n \xrightarrow{d} X \quad \text{if and only if} \quad \varphi_{X_n}(\theta) \to \varphi_X(\theta).$$

Finally, Let us point out that the results on sums of random variables in terms of the probability generating functions translate (and generalize!) in the obvious way to characteristic functions.

Exercises

4.1 Exercise
Calculate the ultimate extinction probabilities for the branching processes having offspring distributions

(a) [0.5000 0.5000],
(b) [0.3333 0.3333 0.3333],
(c) [0.3000 0.5000 0.2000],
(d) [0.2000 0.5000 0.3000].

4.2 Exercise
Calculate (numerically) the ultimate extinction probabilities for the branching processes having offspring distributions

(a) [0.5000 0.0000 0.0000 0.5000],
(b) [0.2500 0.2500 0.2500 0.2500],
(c) [0.4000 0.3000 0.2000 0.1000],
(d) [0.1000 0.2000 0.3000 0.4000].

4.3 Exercise (The Ultimate Instability of Ecology)
Suppose the branching process is a reasonable model for the sizes of animal populations. What does this say about the stability in ecology?

4.4 Exercise
Consider a branching process with **Poisson** offspring distribution with parameter $\lambda > 0$. That is,

$$p_x = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2 \ldots.$$

Visualize the ultimate extinction probability as a function of λ.

4.5 Exercise
Formulate, when possible, the analogs of the results 4.9–4.14 given for probability generating functions for moment generating functions and characteristic functions.

4.6 Exercise
(a) Give an example of a random variable that has moment generating function, but does not have probability generating function.
(b) Give an example of a random variable that has probability generating function, but does not have moment generating function.
(c) Give an example of a random variable that has all moments, but does not have moment generating function.
(d) Calculate the characteristic functions of your examples.

Part II

Some Interesting Probability Distributions

Lecture 5

Binomial Distribution

The binomial distribution is one of the two natural sampling distributions, the other being the hypergeometric distribution. The origins of the binomial distribution is shrouded in history. In its symmetric form, i.e., when the individual success probability is half, the distribution is probably older than any writing system.

Maybe the first one to study the non-symmetric binomial distribution was the Swiss mathematician **Jacob Bernoulli** (1654–1705). Jacob was the oldest in the long and prestigious Bernoulli dynasty of scientific geniuses.

In his posthumously published book *Ars Conjectandi* (1713), Jacob Bernoulli was the first one to publish a version of the law of large numbers. In the same work, the first relatively rigorous study of the binomial distribution was presented. *Ars Conjectandi*, meaning the "Art of Guessing", was one of the first textbooks ever on probability theory, and certainly the most influential one ever. Finally, we note that the word "stochastics" for probability is a (very dry) joke on the name of the Jacob's book, being a some sort of a translation of the Latin word "Conjectandi" into Classical Greek.

Jacob Bernoulli (1654–1705)

The key example 5.1 below and its follow-up example 5.11 of this lecture deal with a **queuing system** in a so-called **stationary state**. We will learn more about queuing systems and stationary states of stochastic processes later. For now, we just give a taste.

5.1 Example (Quality of Service for a Link, I)
There are 5 devices in a teletraffic system sharing a common link. Each device is idle with probability 90 %. When they transmit, they do it with a constant rate of 3 Mb/s. How big should the link capacity be so that the probability that the link can serve all the devices at any given time is at least 99 %?

5.2 Remark
Using numbers in analysis is just silly! One forgets where the numbers came from, and moreover one would be unnecessarily specific. Therefore, we use symbols instead of numbers.

Let S be the workload of the link and let c be its capacity **per device**. Let α be the **quality-of-service** parameter, i.e., the probability that all demands are met. The question is then to find such a c that

$$\mathbb{P}[S > nc] \quad \leq \quad 1-\alpha,$$

where n is the number of devices sharing the link.

To answer the question of Example 5.1, we must develop a stochastic model for the distribution of the workload S. With the minimal data given in Example 5.1 we a forced to make lots of **independence** and **stationarity** assumptions. Indeed, there is no data to justify any kind of special dependence structure. Also, as far as the author knows, there is no "universally accepted" reason for any specific dependence structure, either. Also, we need to make the stationarity assumption that the workload is statistically the same all over the time. This means, in particular, that there are no pre-known "busy hours".

We assume that each device will **independently** demand service with the **same** probability $p = 1-90\% = 0.1$ **at any given time**. Taking 3 Mb/s to be the unit, we can model the demanded workload S as the random variable

$$S \quad = \quad X_1 + X_2 + X_3 + X_4 + X_5,$$

where the summands X_i are **independent and identically distributed** with distribution

$$X_i \quad = \quad \begin{cases} 1, & \text{with probability } p, \\ 0, & \text{with probability } 1-p. \end{cases}$$

We recall, that the distribution of the summands X_i has a name.

5.3 Definition (Bernoulli Distribution)
Let $p \in (0,1)$. A random variable X which takes value 1 with probability p and 0 with probability $1-p$ is called **Bernoulli** distributed.

Bernoulli distribution is $\mathbb{N}$-valued and its probability generating function is

$$\begin{aligned} G(\theta) &= \mathbb{E}\left[\theta^X\right] \\ &= \mathbb{P}[X=0]\theta^0 + \mathbb{P}[X=1]\theta^1 \\ &= (1-p) + p\theta. \end{aligned}$$

The mean of the Bernoulli distribution is

$$\begin{aligned}\mathbb{E}[X] &= 0 \times \mathbb{P}[X=0] + 1 \times \mathbb{P}[X=1] \\ &= \mathbb{P}[X=1] \\ &= p,\end{aligned}$$

and the variance of the Bernoulli distribution is

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= \left(0^2 \times \mathbb{P}[X=0] + 1^2 \times \mathbb{P}[X=1]\right) - p^2, \\ &= p - p^2 \\ &= p(1-p).\end{aligned}$$

5.4 Remark
From the formula of the variance we see that the variability of the Bernoulli distribution is at its highest in the symmetric case $p = 1/2$.

Qualitative Approach to Binomial Distribution

The binomial model is precisely the model that fits Example 5.1.

5.5 Definition (Binomial Distribution)
Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables each having Bernoulli distribution with parameter p. Then their sum

$$S = X_1 + X_2 + \cdots + X_n$$

is **binomially distributed** with parameters n and p.

We calculate the probability mass function of the binomial distribution in the next section. Before that we state some easy properties that follow immediately from Definition 5.5 and the properties of the Bernoulli distribution.

5.6 Proposition (Properties of Binomial Distribution)
Let S be binomially distributed with parameters n and p. Then

(i) its probability generating function is

$$G_{n,p}(\theta) = ((1-p) + p\theta)^n,$$

(ii) its mean is

$$\mathbb{E}[S] \quad = \quad np,$$

(iii) and its variance is

$$\mathbb{V}[S] \quad = \quad np(1-p).$$

Let us first see the case for the probability generating function. Now, the Bernoulli distribution has probability generating function

$$G_p(\theta) \quad = \quad (1-p) + p\theta.$$

Since the binomial distribution is a sum of n independent identically distributed Bernoulli variables, its probability generating function is

$$G_{n,p}(\theta) \quad = \quad G_p(\theta)^n$$

Formula 5.6(i) follows from this.

Formulas 5.6(ii) and 5.6(iii) can be derived from formula 5.6(i) by differentiation. There is, however, an easier way: since the Bernoulli summands in Definition 5.5 are independent, the variance is linear. The expectation is always linear. Consequently we have

$$\begin{aligned}
\mathbb{E}[S] &= \mathbb{E}[X_1 + X_2 + \cdots + X_n] \\
&= \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n] \\
&= p + p + \cdots + p \\
&= np,
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}[S] &= \mathbb{V}[X_1 + X_2 + \cdots + X_n] \\
&= \mathbb{V}[X_1] + \mathbb{V}[X_2] + \cdots + \mathbb{V}[X_n] \\
&= p(1-p) + p(1-p) + \cdots + p(1-p) \\
&= np(1-p),
\end{aligned}$$

Quantitative Approach to Binomial Distribution

The qualitative definition 5.5 is pretty close to a quantitative definition. We only need to calculate the distribution of the sum S of n independent identically distributed Bernoulli random variables. One approach would be to use the law of total probability of Lemma 1.4. Another approach would be to use probability generating functions. We take the latter approach.

Recall that the probability generating function of a binomially distributed random variable with parameters n and p is

$$G_{n,p}(\theta) \quad = \quad \big((1-p) + p\theta\big)^n.$$

So, to calculate the probabilities $\mathbb{P}[S=s]$, $s\in\mathbb{N}$, for a binomially distributed random variable, we only need to differentiate. We obtain

$$\begin{aligned}\mathbb{P}[S=s] &= \frac{1}{s!}G^{(s)}_{n,p}(0)\\ &= \frac{1}{s!}\left[\frac{\mathrm{d}^s}{\mathrm{d}\theta^s}\Big[\big((1-p)+p\theta\big)^n\Big]\right]_{\theta=0}\\ &= \frac{n(n-1)\cdots(n-s+1)}{s!}p^s\Big[\big((1-p)+p\theta\big)^{n-s}\Big]_{\theta=0}\\ &= \binom{n}{s}p^s(1-p)^{n-s}.\end{aligned}$$

So, we have found the binomial distribution:

5.7 Definition (Binomial Distribution)
A random variable S has the **binomial distribution** with parameters $n\in\mathbb{N}$ and $p\in(0,1)$ if it has the probability mass function

$$\mathbb{P}[S=s] = \binom{n}{s}p^s(1-p)^{n-s} \qquad s=0,1,2,\ldots,n.$$

5.8 Remark
There is, of course, the more traditional, **combinatorial**, way to deduce the binomial distribution. It goes like this: consider the **sequence** $X_1,X_2,\ldots,X_n$. Each of the elements in the sequence is either 0 or 1. Suppose the sum $S=X_1+X_2+\cdots+X_n$ takes the value s. One way this can happen is that the first s of the summands take the value 1 and the rest $(n-s)$ take the value 0. The probability for this to happen is, by **independence**,

$$p^s\,(1-p)^{n-s}.$$

Now, this is just one way how the event $\{S=s\}$ can happen. Another way would be that, first $X_1=0$, and then $X_2=X_3=\cdots X_{s+1}=1$, and then $X_t=0$ for the rest $t>s+1$. The probability of this to happen is, by **independence and symmetry**, the same as before:

$$(1-p)\,p^s\,(1-p)^{n-s-1} = p^s\,(1-p)^{n-s}.$$

In general, all the ways the s number of 1's are scattered in the sequence of 0's and 1's of length n, have the same probability

$$p^s(1-p)^{n-s}.$$

The number of ways this scattering can happen is given by the **binomial coefficients**

$$\binom{n}{s} = \frac{n!}{(n-s)!\,s!}.$$

Definition 5.7 follows from this.

5.9 Remark (Sampling with or without Replacement)
The binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N. If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is a **hypergeometric distribution**, not a binomial one. However, for N much larger than n, the binomial distribution remains a good approximation, and is widely used.

Let us then get back to Example 5.1. We denote by $F_{n,p}(s)$ the **cumulative distribution function** of the binomial distribution having parameters n and p:

$$F_{n,p}(s) \quad = \quad \mathbb{P}[S \leq s] \quad = \quad \sum_{y=0}^{s} \binom{n}{y} p^y (1-p)^{n-y}.$$

Then, the problem of Example 5.1 becomes of solving c from the inequality

$$1 - F_{5,0.1}(5c) \quad \leq \quad 0.01.$$

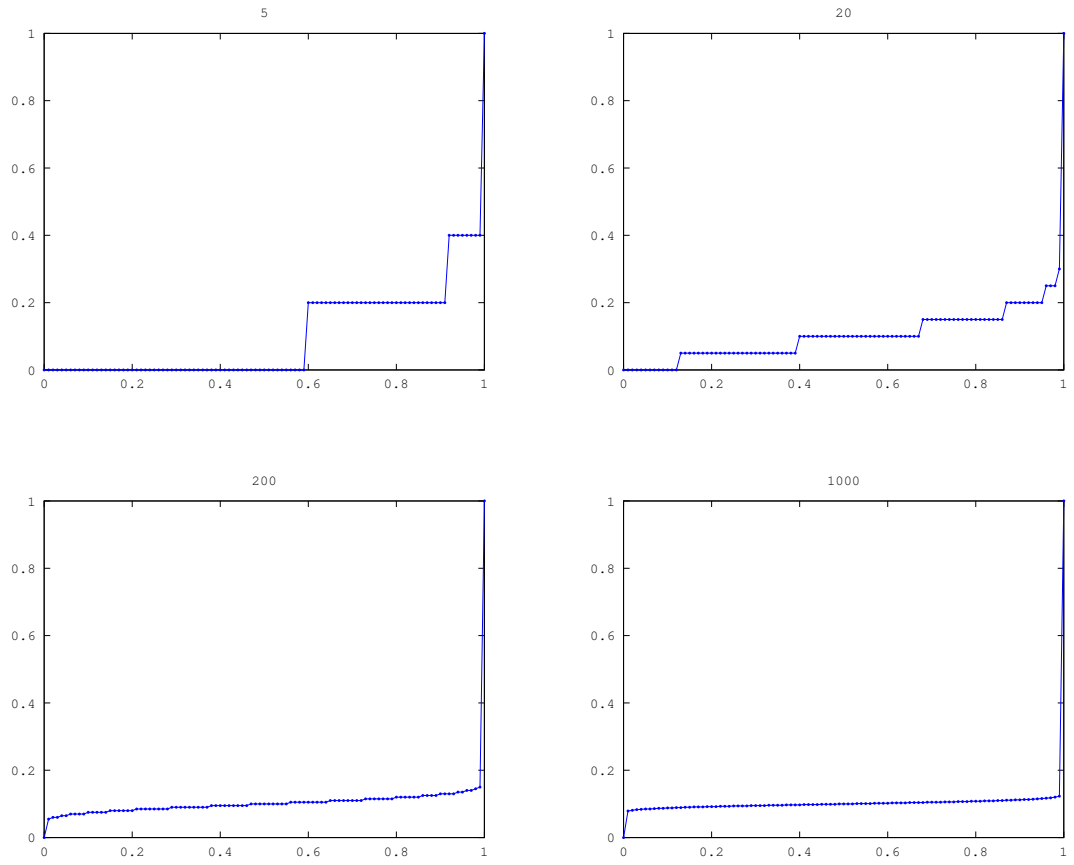
Since $F_{n,p}(s)$ is increasing, it has a **generalized inverse**. Actually, every cumulative distribution function has a generalized inverse: they are called the **quantile functions**. So, after a little bit of algebra, the inequality above becomes

$$c \quad \geq \quad \frac{F^{-1}_{5,0.1}(0.99)}{5}.$$

Unfortunately, neither $F_{n,p}(s)$ nor its inverse $F^{-1}_{n,p}(s)$ admit reasonable closed-form expressions. There are, however, quite fast algorithms for calculating them. In Octave the functions $F_{n,p}(s)$ and $F^{-1}_{n,p}(s)$ are implemented as `binocdf` and `binoinv`, respectively.

5.10 Example (Quality of Service for a Link, I, Solution)
Typing `binoinv(0.99, 5, 0.1)/5` in Octave console gives $c = 0.40000$. This means that the link capacity should be at least $5 \times 0.4 \times 3$ Mb/s$= 6$ Mb/s.

Example 5.1 and its solution 5.10 investigated the problem from the link provider's point of view. From the clients' point of view the solution does not make sense! To see this, look at the first graph below: for a 59% quality-of-service there is no need for any kind of link!



Per device link size requirements c on the y-axes plotted against the quality-of-service parameter α on the x-axes for fixed $p = 0.1$ and different n.

The picture above was generated by the following code.

```
################################################################################
## FILE: qos_link_i.m
##
## Visualizes the quality-of-service in a binomial traffic model from the
## system point of view.
################################################################################

## Parameters for binomial distributions
n = [5 20 200 1000];
p = 0.1;

## Plotting parameters
N = 101;                                        ## Grid size.
alpha = linspace(0,1,N);                        ## The grid.

## Plotting
for i=1:length(n)
    subplot(length(n)/2,length(n)/2,i)
        for k = 1:N
```

```
20              c(k) = binoinv(alpha(k), n(i), p)/n(i);
21          endfor
22          plot(alpha,c, 'marker', '.')
23          title(n(i))
24  endfor
```

www.uva.fi/~tsottine/psp/qos_link_i.m

Binomial Palm Distribution

Let us reformulate Example 5.1 to express the clients' point of view. This is not the link provider's point for view, no matter how much the link company's sales representative would like to tell the clients that it is!

5.11 Example (Quality of Service for a Link, II)
There are 5 devices in a teletraffic system sharing a common link. Each device is idle with probability 90 %. When they transmit, they do it with a constant rate of 3 Mb/s. How big should the link capacity be so that the probability that a given device can transmit is at least 99 %.

The insider's, or client's point of view is described by the so-called Palm distribution named after the Swedish teletrafficist **Conrad "Conny" Palm** (1907–1951). We give an informal, and rather vague, definition below. After that we explain what it means in the context of examples 5.1 and 5.11.

5.12 Definition (Palm Probability)
Consider a queuing system where the clients are statistically interchangeable. Let $\mathbb{P}$ be the probability that models the outsiders point of view of the system. The **Palm probability** $\mathbb{P}^*$ is the clients point of view.

Let us consider the clients (devices) of Example 5.11. Because of **symmetry**, i.e., **statistical interchangeability** of the devices, we can take any device. We take X_1. Let S be as before, the total number of devices transmitting at a given time. Now, what does the device (or client) X_1 see? If X_1 is not in the system, it does not see anything. Therefore, X_1 cannot see the event $\{S = 0\}$, the empty system. If X_1 is in the system, then $\{S \geq 1\}$, since there is at least one client in the system, the client X_1. What X_1 sees, is the Palm probability

$$\mathbb{P}^*[S = s] \quad = \quad \mathbb{P}[S = s \,|\, X_1 = 1].$$

Therefore, the distribution of the system S under the Palm probability $\mathbb{P}^*$ is, by the **independence** of the devices $X_1, X_2, \ldots, X_n$,

$$\begin{aligned}\mathbb{P}^*[S=s] &= \mathbb{P}[X_1+X_2+\cdots+X_n=s\,|\,X_1=1]\\ &= \mathbb{P}[1+X_2+\cdots+X_n=s\,|\,X_1=1]\\ &= \mathbb{P}[X_2+\cdots+X_n+1=s]\\ &= \mathbb{P}[S^*=s],\end{aligned}$$

where

$$S^*-1 = X_2+X_3+\cdots+X_n$$

is binomially distributed with parameters $n-1$ and p.

We can formalize our discovery as the following proposition.

5.13 Proposition (Binomial Palm Distribution)
The Palm distribution for S that is binomially distributed with parameters n and p is the distribution of S^*, where S^*-1 is binomially distributed with parameters $n-1$ and p.

For the system, or for the link provider, the relevant quality-of-service requirement was

$$\mathbb{P}[S>nc] \leq 1-\alpha,$$

since the link provider looks the queueing system "from the outside". For the client, the relevant quality-of-service requirement is

$$\mathbb{P}^*[S>nc] \leq 1-\alpha,$$

since the client looks the queueing system "from the inside".

By Proposition 5.13, we can rewrite the clients' quality-of-service requirement as

$$\mathbb{P}[S^*>nc] \leq 1-\alpha.$$

Also, by Proposition 5.13, we have

$$\mathbb{P}[S^*>nc] = 1-F_{n-1,p}(nc-1),$$

where $F_{n-1,p}(s)$ is the cumulative distribution function of the binomial distribution with parameters $n-1$ and p. Consequently, after a little bit of algebra, we obtain the **client's quality-of-service requirement**

$$c \geq \frac{F^{-1}_{n-1,p}(\alpha)+1}{n}.$$

Note that this is different from the **system's quality-of-service requirement**

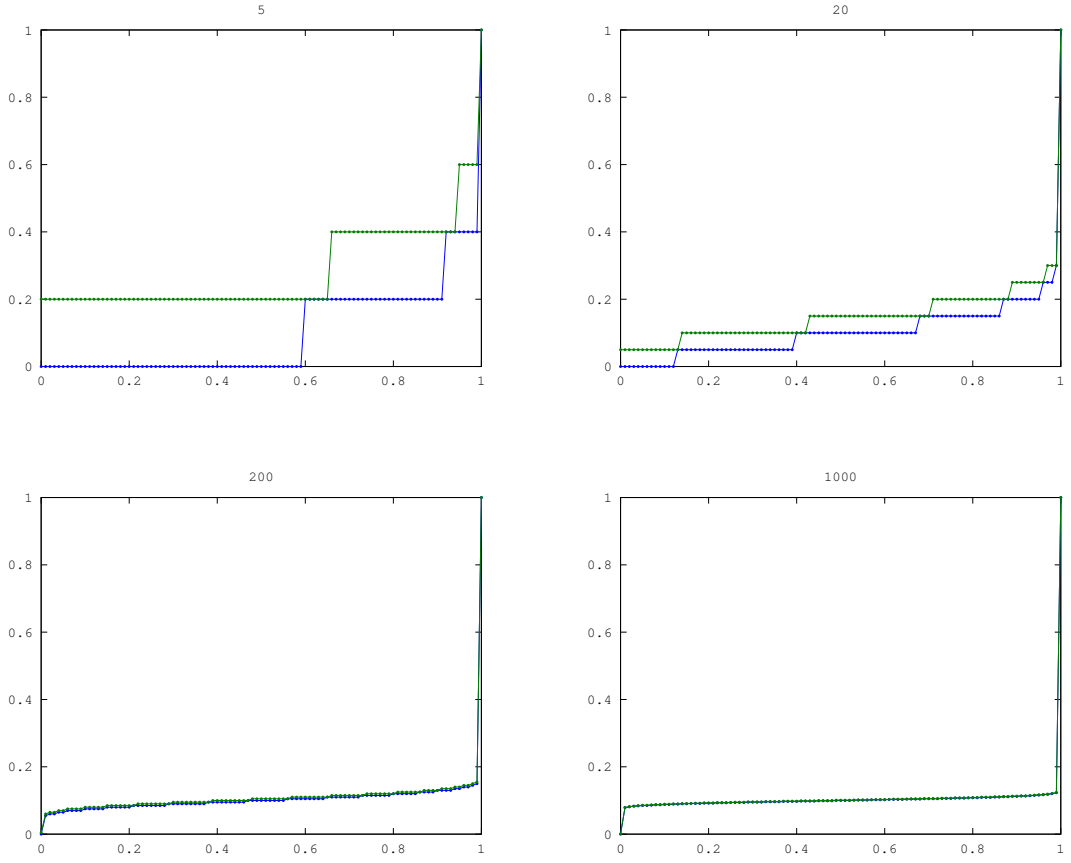
$$c \geq \frac{F^{-1}_{n,p}(\alpha)}{n}.$$

Now we are ready to solve Example 5.11, and give the client's point of view:

5.14 Example (Quality of Service for a Link, II, Solution)
Typing `(binoinv(0.99, 5-1, 0.1)+1)/5` in Octave console give us solution $c = 0.60000$. This means that the link capacity should be at least $5 \times 0.6 \times 3$ Mb/s$= 9$ Mb/s.

So, for the link provider 99% quality-of-service means a 6 Mb/s link, while for the client 99% quality-of-service means a 9 Mb/s link. Quite a difference!

Finally, to compare the outsider's and insider's points of view, we plot the quality-of-service against the link capacity per device for both of the viewpoints with values $n = 5$, 20, 200, 1 000 for the number of clients.



Per device link size requirements c, with inside (green) and outside (blue) views on the y-axes plotted against the quality-of-service parameter α on the x-axes for fixed $p = 0.1$ and different n.

5.15 Remark (Quality-of-Service Folklore)
From the picture above we can read, e.g., that

(i) The insider needs more capacity for the same quality-of-service than the outsider.
(ii) When n is big there is virtually no difference between what the insider and the outsider see.
(iii) High quality is always extremely expensive.
(iv) If n is large good quality is cheap.

The picture above was generated by the code below.

```
################################################################################
## FILE: qos_link_ii.m
##
## Visualizes the quality-of-service in a binomial traffic model from the
## system point of view.
################################################################################

## Parameters for binomial distributions.
n = [5 20 200 1000];
p = 0.1;

## Plotting parameters
N = 101;                                                ## Grid size.
alpha = linspace(0,1,N);                                ## The grid.

## Plotting
for i=1:length(n)
    subplot(length(n)/2,length(n)/2,i)
        for k = 1:N
            cout(k) = binoinv(alpha(k), n(i), p)/n(i);
            cin(k) = ( binoinv(alpha(k), n(i)-1, p) + 1 )/n(i);
        endfor
        plot(alpha, cout, 'marker', '.', alpha, cin, 'marker', '.')
        title(n(i))
endfor
```

www.uva.fi/~tsottine/psp/qos_link_ii.m

Exercises

5.1 Exercise
Answer to the question of Example 5.1 when the activity parameter p is

(a) 0.05,
(b) 0.1,
(c) 0.2,
(d) 0.95.

5.2 Exercise
Answer to the question of Example 5.11 when the activity parameter p is

(a) 0.05,
(b) 0.1,
(c) 0.2,
(d) 0.95.

5.3 Exercise
Consider Example 5.1. Let $n = 20$ and $\alpha = 0.95$. Visualize the connection between the activity parameter p and the per device capacity c.

5.4 Exercise
Consider Example 5.11. Let $n = 20$ and $\alpha = 0.95$. Visualize the connection between the activity parameter p and the per device capacity c.

5.5 Exercise (Poisson-Binomial Distribution)
Let X_i be Bernoulli random variables with different parameters p_i. The distribution of the sum

$$S = X_1 + X_2 + \cdots + X_n$$

is called the **Poisson-binomial distribution**.

(a) Find out the probability generating function of the Poisson-binomial distribution.
(b) Express, somehow, the probability distribution function of the Poisson-binomial distribution.

5.6 Exercise (Generalized Poisson-Binomial Distribution)
Let X_i be Bernoulli random variables with different parameters p_i and let a_i be positive real numbers. Set

$$S = a_1X_1 + a_2X_2 + \cdots + a_nX_n.$$

Let us call the distribution of S the **generalized Poisson-binomial distribution**.

(a) How is the generalized Poisson-binomial distribution related to Example 5.1?
(b) Calculate the moment generating function of the generalized Poisson-binomial distribution.
(c) What is the probability generating function of the generalized Poisson-binomial distribution?
(d) Can you express somehow the probability distribution function of the generalized Poisson-binomial distribution?

Lecture 6

Poisson Distribution

The Poisson distribution is named after the French mathematician **Siméon Denis Poisson** (1781–1840) who introduced the distribution in 1837 in his work *Recherches sur la probabilité des jugements en matiére criminelle et en matiére civile* ("Research on the Probability of Judgments in Criminal and Civil Matters"). In his work the Poisson distribution describes the probability that a random event will occur in a time and/or space interval under the condition that the probability of any single event occurring is very small p, but the number of trials is very large N. So, for Siméon Denis Poisson, the Poisson(λ) distribution was a limit of binomial(N, p) distributions in the sense of the Law of Small Numbers 6.10: $p \to 0$ and $N \to \infty$, but $pN \to \lambda > 0$.

Another pioneer of the Poisson distribution was the Polish–German economist–statistician **Ladislaus Bortkiewicz** (1868–1931) who coined the term "Law of Small Numbers" in his 1898 investigation of the number of soldiers in the Prussian army killed accidentally by horse kick. Some have suggested that the Poisson distribution should be renamed the "Bortkiewicz distribution".

Siméon Denis Poisson (1781–1840)

The Poisson distribution is, in some sense, the uniform distribution on the natural numbers $\mathbb{N} = \{0, 1, 2, \ldots\}$. Indeed, the Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently. In the key example of this lecture, Example 6.1 below, the events are scattered in space, not in time.

6.1 Example (Malus Particles)
The Lake Diarrhea has, on average, 7 Malus particles per one liter. Magnus Flatus lives on the shore of the Lake Diarrhea. He drinks daily 2 liters of water from the Lake Diarrhea. The lethal daily intake of Malus particles is 30. What is the probability that Magnus Flatus will have a lethal intake of Malus particles in a given day?

Qualitative Approach to Poisson Distribution

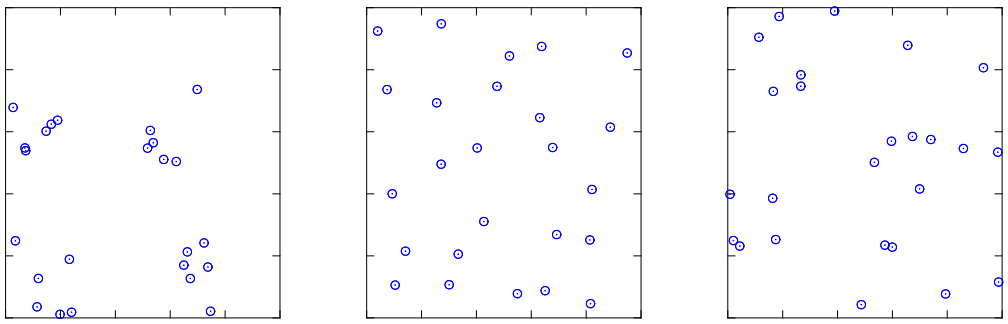
To answer the question of Example 6.1 we need to know the distribution of the random variable X that denotes the number of Malus particles in a 2 liter sample from the Lake Diarrhea. To fix the distribution of X we have to assume something about the distribution of the Malus particles in the lake. We know the average of the Malus particles: 7 per liter. Without any additional information, it is natural to assume that the particles are **independently and homogeneously scattered** in the lake. This means that knowledge of the amount of Malus particles in one sample does not help in predicting the amount of Malus particles in another sample (independence) and that samples taken from different parts of the lake are statistically the same (homogeneity). This leads us to the qualitative definition of the Poisson distribution, or actually the Poisson point process:

6.2 Definition (Poisson Point Process)
Let $\mathscr{A}$ be a collection of subsets of the Euclidean space $\mathbb{R}^d$ and let $\mathrm{vol}(A)$ denote the volume of the set A of $\mathbb{R}^d$. The family $X(A)$, $A \in \mathscr{A}$, is a **Poisson point process** with parameter $\lambda > 0$ if

(i) $X(A)$ takes values in $\mathbb{N} = \{0, 1, 2, \ldots\}$.
(ii) The distribution of $X(A)$ depends only on $\lambda \mathrm{vol}(A)$.
(iii) If A and B are disjoint, then $X(A)$ and $X(B)$ are independent.
(iv) $\mathbb{E}[X(A)] = \lambda \mathrm{vol}(A)$ for each A in $\mathscr{A}$.

In Example 6.1, $\mathscr{A}$ is the collection of all possible water samples from Lake Diarrhea, $X(A)$ is the number of Malus particles in the sample A, and Lake Diarrhea is a subset of $\mathbb{R}^3$.

6.3 Remark
For the untrained eye, it is not easy to fathom independent homogeneous scattering. The following pictures are here to train your eye.



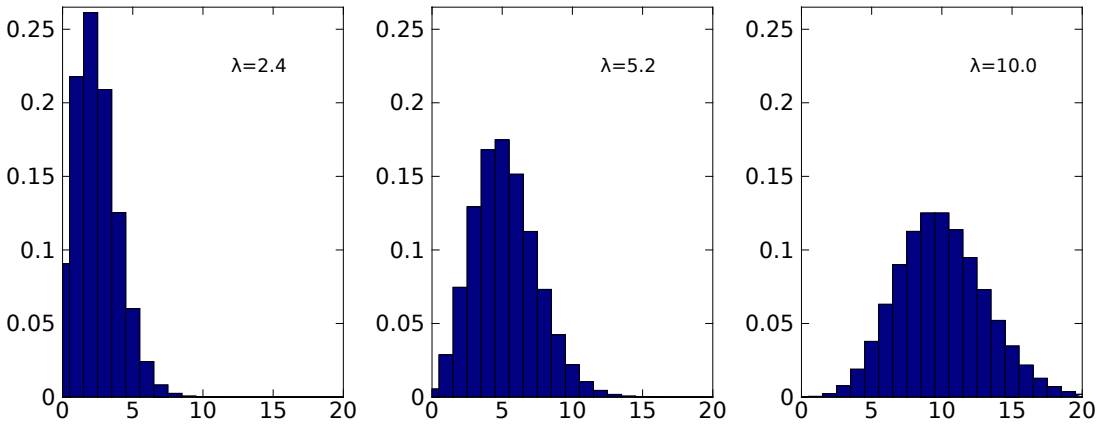
Samples of point processes. See `point_processes.m` to see which one is the Poisson point process.

Quantitative Approach to Poisson Distribution

The qualitative definition 6.2 does not, yet, allow calculations of probabilities, although some expectations can be calculated. Indeed, we already did calculate some expectations in Exercise 6.2. However, it turns out that the qualitative definition 6.2 actually fixes the distributions completely as Theorem 6.5 will show. Before that, let us recall the Poisson distribution.

6.4 Definition (Poisson Distribution)
A random variable X has the **Poisson distribution** with parameter $\lambda > 0$ if it has the probability mass function

$$\mathbb{P}[X = x] \quad = \quad \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots.$$



Probability mass functions of Poisson distribution with different parameters λ.

The plots above were generated with Octave with the following code:

```
##############################################################################
## FILE: poisson_pmfs.m
##
## Plots some probability mass functions (pmf) of Poisson distributions.
##############################################################################

## Data for plots.
lambda = [2.4, 5.2, 10.0];              ## Parameters for the plots.
x = 0:20;                               ## The x's for Px=prob(x) to be plotted.
w = 1;                                  ## Width of the bar in the bar plot.
Px(1,:) = poisspdf(x, lambda(1));       ## 1st row for lambda(1).
Px(2,:) = poisspdf(x, lambda(2));       ## 2nd row for lambda(2).
Px(3,:) = poisspdf(x, lambda(3));       ## 3rd row for lambda(3).

## Plotting (bar plots).
plotlims = [0, 20, 0, 0.265];           ## Plotting window [x1, x2, y1, y2].
```

```
subplot(1,3,1);                          ## 1 row, 3 columns, 1st plot.
    bar(x, Px(1,:), w);
    text(12, 0.225, '\lambda=2.4');
    axis(plotlims);
subplot(1,3,2);                          ## 1 row, 3 columns, 2nd plot.
    bar(x, Px(2,:), w);
    text(12, 0.225, '\lambda=5.2');
    axis(plotlims);
subplot(1,3,3);                          ## 1 row, 3 columns, 3rd plot.
    bar(x, Px(3,:), w);
    text(12, 0.225, '\lambda=10.0');
    axis(plotlims);
```

www.uva.fi/~tsottine/psp/poisson_pmfs.m

6.5 Theorem (Poisson Point Process is Poisson Distributed)
For the Poisson point process $X(A)$, $A \in \mathscr{A}$, of Definition 6.2 it must hold true that

$$\mathbb{P}[X(A)=x] \quad = \quad \mathrm{e}^{-\lambda \mathrm{vol}(A)} \frac{(\lambda \mathrm{vol}(A))^x}{x!} \qquad \text{for all } A \text{ in } \mathscr{A} \text{ and } x = 0,1,\ldots,$$

where $\mathrm{vol}(A)$ is the volume of the set A.

Let us argue how the Poisson distribution arises from the Poisson point process, i.e, let us argue why Theorem 6.5 holds. We use the method of probability generating functions. Let

$$G\big(\theta; \mathrm{vol}(A)\big) \quad = \quad \sum_x \mathbb{P}[X(A)=x]\, \theta^x,$$

be the probability generating function of the random variable $X(A)$. The **key trick** is to split the set A into two parts, A_v and A_w with volumes v and w. Then the volume of A is $v+w$ and by the independence assumption we obtain the functional equation

$$G(\theta; v+w) \quad = \quad G(\theta; v) G(\theta; w).$$

Denote $g(\theta; v) = \log G(\theta; v)$. Then from the above we obtain the **Cauchy's functional equation**

$$g(\theta; v+w) \quad = \quad g(\theta; v) + g(\theta; w).$$

Here we keep the parameter θ fixed and consider v and w as variables. So, $g(\theta; v)$ is **additive** in v. Since $g(\theta; v)$ is also increasing in v, it follows that $g(\theta; v) = v\psi(\theta)$ for some $\psi(\theta)$. So, $G(\theta; v) = e^{v\psi(\theta)}$. Since $G(\theta; v)$ is a probability generating function we

must have,

$$\begin{aligned} G(1;\nu) &= \sum_x \mathbb{P}[X(A_\nu)=x]\,1^x, \\ &= 1 \\ G'(1;\nu) &= \sum_x x\mathbb{P}[X(A_\nu)=x]\,1^x \\ &= \mathbb{E}[X(A_\nu)] \\ &= \lambda\nu. \end{aligned}$$

Thus, we must have $\psi(\theta)=\lambda(\theta-1)$. So,

$$G(\theta;\nu) = \mathrm{e}^{\lambda\nu(\theta-1)}.$$

Since probability generating functions determine probabilities, the claim follows from Proposition 6.6(i) below, the proof of which is left as Exercise 6.3.

6.6 Proposition (Properties of Poisson Distribution)
Let X be Poisson-distributed with parameter λ. Then,

(i) $G_X(\theta)=\mathrm{e}^{\lambda(\theta-1)}$,
(ii) $\mathbb{E}[X]=\lambda$,
(iii) $\mathbb{V}[X]=\lambda$.

6.7 Example (Malus Particles, Solution)
Now we know that the number of Malus particles Magnus Flatus consumes daily has the distribution

$$\mathbb{P}[X=x] = \mathrm{e}^{-14}\frac{14^x}{x!}.$$

The probability in question is

$$\begin{aligned} \mathbb{P}[X\geq 30] &= 1-\mathbb{P}[X\leq 29] \\ &= 1-\sum_{x=0}^{29}\mathrm{e}^{-14}\frac{\mathrm{e}^{14x}}{x!}. \end{aligned}$$

Since we do not want to calculate all the 30 terms of the sum above by hand, we use Octave. The simplest way of doing this is to call the Octave function `poisscdf`. So, typing `1-poisscdf(29,14)` we get the answer 0.01358 %.

Sums of Independent Poisson Distributions

From the qualitative approach to the Poisson process it is **intuitively** clear that if X_1 and X_2 are **independent** Poisson distributed with parameters λ_1 and λ_2, respectively, then their sum X_1+X_2 is also Poisson distributed with parameter $\lambda_1+\lambda_2$. **Rigorously** this can be seen by comparing the probability generating functions:

$$\begin{aligned} G_{X_1+X_2}(\theta) &= G_{X_1}(\theta)G_{X_2}(\theta) \\ &= \mathrm{e}^{\lambda_1(\theta-1)}\mathrm{e}^{\lambda_2(\theta-1)} \\ &= \mathrm{e}^{(\lambda_1+\lambda_2)(\theta-1)}, \end{aligned}$$

which is the probability generating function of a Poisson random variable with parameter $\lambda_1+\lambda_2$.

Repeating the arguments above for n summands, we obtain the following:

6.8 Proposition (Poisson Sum)
Let $X_1, X_2, \ldots, X_n$ be independent Poisson distributed random variables with parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$, respectively. Then their sum $X_1+X_2+\cdots+X_n$ is Poisson distributed with parameter $\lambda_1+\lambda_2+\cdots+\lambda_n$.

Let us then consider a reverse of Proposition 6.8. Suppose X_1 and X_2 are **independent** Poisson distributed random variables with parameters λ_1 and λ_2, respectively. Suppose further that we know the value of their sum: $X_1+X_2=x$. What can we say about X_1? **Intuitively**, we can argue as follows: each point of the Poisson point process X_1+X_2 comes independently from either X_1 or X_2. The relative contribution of X_1 to the points is $\lambda_1/(\lambda_1+\lambda_2)$. So, this is the probability of success, if success means that the point comes from the random variable X_1. Since these successes are independent we arrive at the binomial distribution: X_1 is binomially distributed with parameters x and $\lambda_1/(\lambda_1+\lambda_2)$. **Rigorously**, the educated guess above is seen to be true from the following calculations: First, simple use of definitions yield

$$\begin{aligned} \mathbb{P}[X_1=x_1 \mid X_1+X_2=x] &= \frac{\mathbb{P}[X_1=x_1, X_2=x-x_1]}{\mathbb{P}[X_1+X_2=x]} \\ &= \frac{\mathbb{P}[X_1=x_1]\mathbb{P}[X_2=x-x_1]}{\mathbb{P}[X_1+X_2=x]} \\ &= \mathrm{e}^{-\lambda_1}\frac{\lambda_1^{x_1}}{x_1!}\mathrm{e}^{-\lambda_2}\frac{\lambda_2^{x-x_1}}{(x-x_1)!}\Big/\mathrm{e}^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^x}{x!} \end{aligned}$$

Then rearranging the terms in the result above yields

$$\mathbb{P}[X_1=x_1 \mid X_1+X_2=x] = \binom{x}{x_1}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{x_1}\left(1-\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{x-x_1}.$$

So, we see that X_1 given $X_1 + X_2 = x$ is binomially distributed with parameters x and $\lambda_1/(\lambda_1 + \lambda_2)$. Repeating the arguments above for n summands, we obtain the following:

6.9 Proposition (Reverse Poisson Sum)
Let $X_1, X_2, \ldots, X_n$ be independent Poisson distributed random variables with parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$, respectively. Let $X = X_1 + X_2 + \cdots + X_n$ and $\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$. Then, conditionally on $X = x$ the random variables X_k, $k = 1, 2, \ldots, n$, are binomially distributed with parameters x and λ_k/λ.

Law of Small Numbers

Proposition 6.9 gave one important connection between the Poisson and the binomial distributions. There is another important connection between these distributions. This connection, the law of small numbers 6.10 below, is why Siméon Denis Poisson introduced the Poisson distribution.

6.10 Theorem (Law of Small Numbers)
Let X_n, $n \in \mathbb{N}$, be binomially distributed with parameters n and p_n. Suppose that $np_n \to \lambda$ as $n \to \infty$. Then the distribution of X_n converges to the Poisson distribution with parameter λ, i.e.,

$$\mathbb{P}[X_n = x] \quad \to \quad \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!} \qquad \text{for all } x = 0, 1, 2, \ldots,$$

whenever $np_n \to \lambda$.

Theorem 6.10 follows easily by using Levy's continuity theorem (Lemma 4.21). To apply it we should strictly speaking use the characteristic functions. We will use the more familiar probability generating functions instead. The proper proof with characteristic functions follows by replacing θ with $\mathrm{e}^{\mathrm{i}\theta}$ in the calculations below.

In the language of probability generating functions, to prove Theorem 6.10, we need to show that

$$\lim_{n\to\infty} G_n(\theta) \quad = \quad G(\theta),$$

where $G_n(\theta)$ is the probability generating function of a binomial distribution with parameters n and p_n and $G(\theta)$ is the probability generating function of a Poisson distribution with parameter λ, where

$$\lambda \quad = \quad \lim_{n\to\infty} np_n.$$

By plugging in the actual forms of the probability generating functions in question, we are left with the task of showing that

$$\lim_{n\to\infty}\Big((1-p_n)+p_n\theta\Big)^n \quad = \quad \mathrm{e}^{\lambda(\theta-1)}.$$

Now, recall that

$$\mathrm{e}^x \quad = \quad \lim_{n\to\infty}\left(1+\frac{x_n}{n}\right)^n,$$

if $x_n \to x$. Consequently, we get

$$\begin{aligned}\lim_{n\to\infty}\Big((1-p_n)+p_n\theta\Big)^n &= \lim_{n\to\infty}\Big(1+p_n(\theta-1)\Big)^n \\ &= \lim_{n\to\infty}\left(1+\frac{np_n(\theta-1)}{n}\right)^n \\ &= \mathrm{e}^{\lambda(\theta-1)},\end{aligned}$$

which proves the law of small numbers.

Finally, let us remark that the **law of small numbers** is closely related to a more famous limit result: the **central limit theorem** that gives a Gaussian limit. The reader is invited to contemplate the differences and similarities of the Poisson and the Gaussian approximation of binomial distributions.

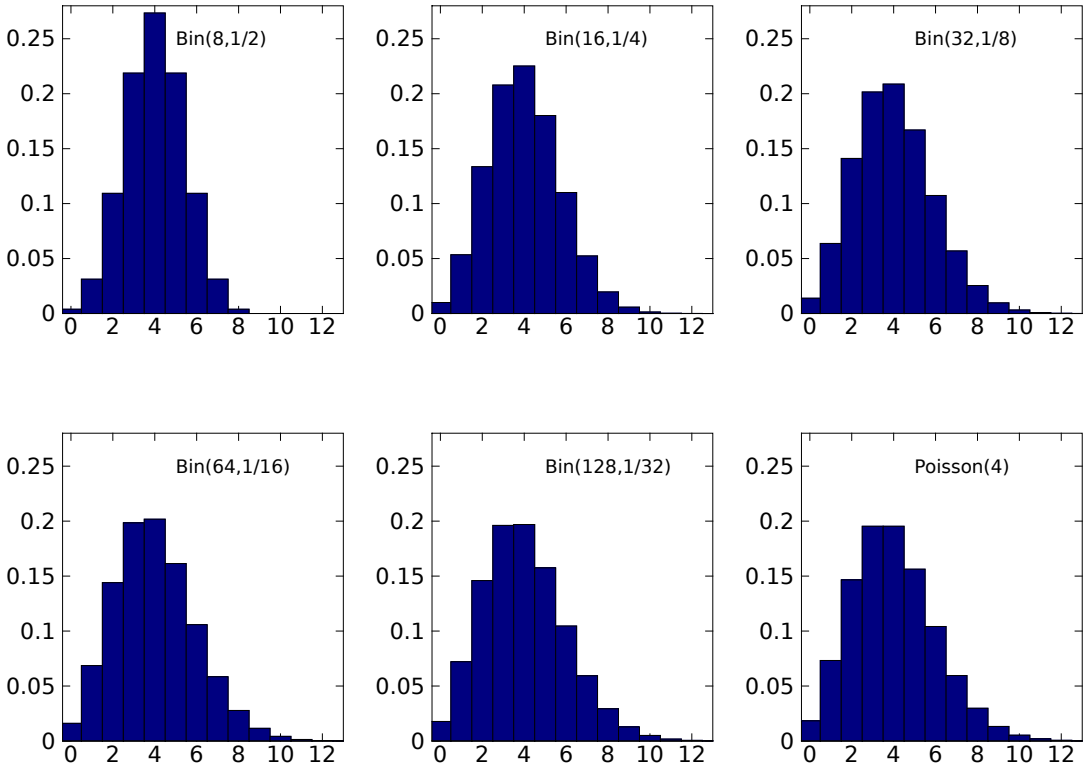


Illustration of how the Binomial becomes Poisson by the law of small numbers.

The plots above were generated with Octave with the following code:

```
################################################################################
## FILE: poisson_binomial.m
##
## An illustration of the Law of Small Numbers, i.e., how binomial
## distributions approximate the Poisson distribution, or vice versa.
##
## This is quick and very dirty coding.  No-one should learn this style!
################################################################################

lambda = 4;                         ## Limit lambda=p*n is fixed.
w = 1;                              ## The width of the column for the bar plot.
plotlims = [-0.4, 13, 0, 0.28];     ## Plotting window [x1, x2, y1, y2].

## The 2x3 subplots.
x = 0:8;
n = 8;
P = binopdf(x,n,lambda/n);
subplot(2,3,1)
    bar(x, P, w);
    text(5, 0.25, 'Bin(8,1/2)');
    axis(plotlims);

x = 0:15;

n = 16;
P = binopdf(x,n,lambda/n);
subplot(2,3,2)
    bar(x, P, w);
    text(5, 0.25, 'Bin(16,1/4)');
    axis(plotlims);

n = 32;
P = binopdf(x,n,lambda/n);
subplot(2,3,3)
    bar(x, P, w);
    text(5, 0.25, 'Bin(32,1/8)');
    axis(plotlims);

n = 64;
P = binopdf(x,n,lambda/n);
subplot(2,3,4)
    bar(x, P, w);
    text(5, 0.25, 'Bin(64,1/16)');
    axis(plotlims);

n = 256;
P = binopdf(x,n,lambda/n);
subplot(2,3,5)
    bar(x, P, w);
    text(5, 0.25, 'Bin(128,1/32)');
    axis(plotlims);

P = poisspdf(x,lambda);
subplot(2,3,6);
```

```
55      bar(x, P, w);
56      text(5, 0.25, 'Poisson(4)');
57      axis(plotlims);
```

www.uva.fi/~tsottine/psp/poisson_binomial.m

Exercises

6.1 Exercise
Let X be Poisson distributed with parameter 1.5. Calculate

(a) $\mathbb{P}[X = 1.5]$
(b) $\mathbb{P}[X = 0 \text{ or } X = 10]$
(c) $\mathbb{P}[X > 1.5]$,
(d) $\mathbb{P}[1 < X \leq 10]$

6.2 Exercise
Consider Example 6.1.

(a) How many Malus particles, on average, would Magnus Flatus get daily if he would drink only 1 liter per day?
(b) Suppose Magnus Flatus wants to get, on average, only 10 Malus particles per day. How many liters can he drink from the Lake Diarrhea daily?
(c) Suppose Magnus Flatus wants to get, on average, only 30 Malus particles per year. How many liters can he drink from the Lake Diarrhea daily?

6.3 Exercise
Prove Proposition 6.6.

6.4 Exercise
Let X_1, X_2 and X_3 be independent Poisson distributed random variables with parameters 2.5, 5.0 and 10.0, respectively. Calculate the probabilities

(a) $\mathbb{P}[X_1 + X_2 = 0]$
(b) $\mathbb{P}[X_1 + X_2 + X_3 = 1]$
(c) $\mathbb{P}[X_1 + X_2 \leq 10]$
(d) $\mathbb{P}[X_1 = 0 \,|\, X_1 + X_2 + X_3 = 1]$
(e) $\mathbb{P}[X_1 = 3 \,|\, X_2 + X_2 + X_3 = 3]$
(f) $\mathbb{P}[X_1 \leq 1 \,|\, X_1 + X_2 + X_3 = 1]$

6.5 Exercise
Let X_1 be Poisson distributed with parameter 2 and let X_2 be Poisson distributed with parameter 5. Suppose $X_1 + X_2 = 10$. What is the probability that $X_1 > X_2$?

Lecture 7

Exponential Distribution

The origins of the exponential distribution are shrouded in history. One of the first persons to study the exponential distribution was the English mathematician and founder of mathematical statistics **Karl Pearson** (1857–1936) in his 1895 work *Contributions to the mathematical theory of evolution, II: Skew variation in homogeneous material*. In this work the exponential distribution is a special case of the gamma distribution, which itself is a Pearson distribution of type III. Pearson was interested modeling distributions with skew in general, not on the special properties of the exponential distribution as such.

Later, the exponential distribution became the central distribution in, e.g., queueing theory and insurance mathematics. This is mainly due to the memoryless property of the exponential distribution: if one has to wait an exponentially distributed time then the remaining waiting time always has the same distribution, no matter how long one has already waited. This property makes the exponential distribution the quintessential waiting-time distribution. Perhaps the first one to use the exponential distribution as the universal waiting-time distribution was the Danish mathematician and founder of traffic engineering and queueing theory **Agner Krarup Erlang** (1878–1929) in his work 1909 work *The Theory of Probabilities and Telephone Conversations*.

Karl Pearson (1857–1936)

The exponential distribution is, in some sense, the uniform distribution on the positive reals $\mathbb{R}_+ = \{x \in \mathbb{R}; x \geq 0\} = [0, \infty)$. Indeed, the exponential distribution is the natural choice for the (remaining or total) waiting-time if one does not have the information on how long the customer has already waited. The key example of this lecture, Example 7.1 below, illustrates this point.

7.1 Example (Waiting in Line, I)
Lady Candida needs to powder her nose. She finds the powder room occupied. From past experience, Lady Candida knows that the average time a lady spends in powdering her nose is 10 minutes. Lady Candida has waited for 5 minutes. The powder room is still occupied. What is the probability that Lady Candida still has to wait for at least 5 minutes?

Qualitative Approach to Exponential Distribution

To solve Example 7.1 we need to know the distribution of the random variable T that denotes the remaining waiting time. Without any additional information on the time that has already been spent in waiting it is natural to assume that the distribution of the remaining waiting time is always the same and depends only on the mean waiting time parameter, which we denote by $1/\lambda$. The reason for choosing $1/\lambda$ instead of simply λ, is to make the parametrization consistent with the Poisson distribution. The connection between the Poisson and the Exponential distribution will be explained much later when we study the **Poisson process** in Lecture 12.

7.2 Definition (Exponential Waiting Time)
The **waiting time** T has **exponential distribution** with parameter $\lambda > 0$ if

(i) $\mathbb{P}[T > t+s | T > s] = \mathbb{P}[T > t]$ for all $t, s > 0$.
(ii) $\mathbb{E}[T] = \frac{1}{\lambda}$.

So, Example 7.1 has been reduced to calculating $\mathbb{P}[T > 5]$, where T has exponential distribution with parameter $1/10$. Indeed, by 7.2(i) $\mathbb{P}[T > 5+5 | T > 5] = \mathbb{P}[T > 5]$ and by 7.2(ii) $\lambda = 1/10$.

7.3 Remark (Light or Heavy Tails)
The exponential distribution is a **mathematically convenient compromise** between ultra light tails and heavy tails:

Ultra Light Tails The person being served in front of you in a queue has already taken a lot of time. Therefore it is reasonable to assume that she will be finished very soon.

Heavy Tails The person being served in front of you in a queue has already taken a lot of time. Therefore it is reasonable to assume that there are some complications so that she will take still a very long time to finish her business.

Quantitative Approach to Exponential Distribution

The **qualitative definition** 7.2 does not allow calculations. Moreover, it may turn out that the qualitative definition does not even determine the distribution uniquely, or even worse: there are no distributions satisfying the assumptions of Definition 7.2. Luckily, it turns out that the quantitative definition is equivalent to the following **quantitative definition**. For

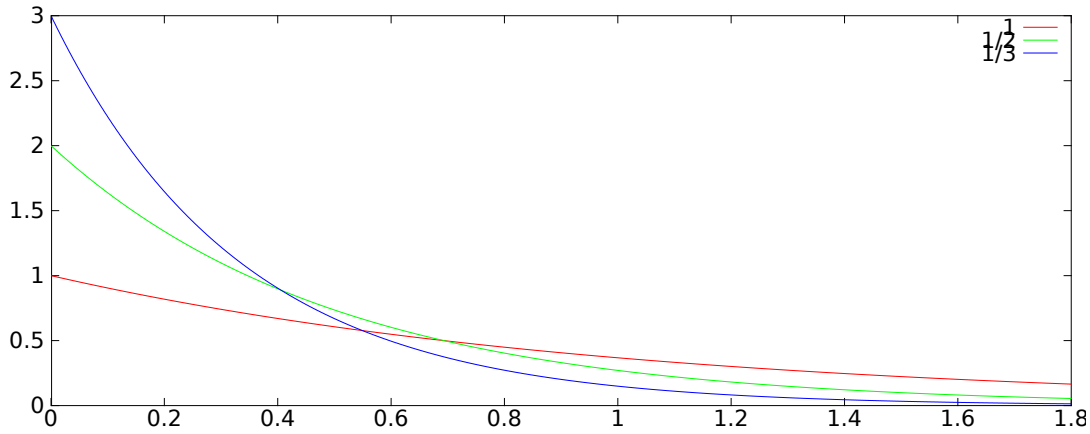
the definition we recall that a random variable X is **continuous** if its **cumulative distribution function** $F(x) = \mathbb{P}[X \leq x]$ is differentiable. We call the derivative $f(x) = F'(x)$ the (probability) **density function** of X. Note that for continuous random variables

$$\mathbb{P}[a \leq X \leq b] \quad = \quad \int_a^b f(x)\,\mathrm{d}x,$$

and, in particular $\mathbb{P}[X = x] = 0$ for any fixed value x.

7.4 Definition (Exponential Distribution)
A positive random variable T has the **exponential distribution** with parameter $\lambda > 0$ if it has the probability density function

$$f_T(t) \quad = \quad \lambda \mathrm{e}^{-\lambda t}, \qquad t \geq 0.$$



Probability density functions of exponential distributions with different parameters λ.

The picture above was generated with the Octave code listed below:

```
1  ##############################################################################
2  ## FILE: exponential_pdfs.m
3  ##
4  ## Plots some probability density functions (pdf) of Exponential distributions.
5  ##############################################################################
6
7  ## Data for plots.
8  lambda = [1, 1/2, 1/3];              ## Means (1/the prameters) for the plots.
9  x = linspace(0, 1.8, 5000);          ## The grid for Px=prob(x) plotted.
10 Px(1,:) = exppdf(x, lambda(1));      ## 1st row for 1/lambda(1).
11 Px(2,:) = exppdf(x, lambda(2));      ## 2nd row for 2/lambda(2).
12 Px(3,:) = exppdf(x, lambda(3));      ## 3rd row for 3/lambda(3).
13
14 ## Plotting (standard plot).
15 hold on;
16 plot(x, Px(1,:), '1;1;','linewidth', 2);
```

```
17 plot(x, Px(2,:), '2;1/2;','linewidth', 2);
18 plot(x, Px(3,:), '3;1/3;','linewidth', 2);
19 axis([0.001, 1.8, 0, 3]);
```

www.uva.fi/~tsottine/psp/exponential_pdfs.m

7.5 Theorem
Definitions 7.2 and 7.4 are the same.

Let us argue that Definition 7.4 and Definition 7.2 are indeed the same. The argument has two sides: first we have to show that the distribution given by Definition 7.4 satisfies the assumptions of Definition 7.2 (this is easy), and second we have to show that a random variable given by Definition 7.2 is necessary of the form given by Definition 7.2 (this is the tricky part).

Let us start with the easy part. Suppose the random variable T has the density function $f_T(t)$ given by Definition 7.4. First, we note that

$$\begin{aligned}\mathbb{P}[T > t+s \mid T > s] &= \frac{\mathbb{P}[T > t+s, T > s]}{\mathbb{P}[T > s]} \\ &= \frac{\mathbb{P}[T > t+s]}{\mathbb{P}[T > s]}.\end{aligned}$$

Consequently,

$$\begin{aligned}\mathbb{P}[T > t+s \mid T > s] &= \frac{\int_{t+s}^{\infty} \lambda e^{-\lambda u}\,du}{\int_{s}^{\infty} \lambda e^{-\lambda u}\,du} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} \\ &= e^{-\lambda t} \\ &= \mathbb{P}[T > t].\end{aligned}$$

Thus assumption (i) of Definition 7.2 holds. Let us then check the assumption (ii) of Definition 7.2, i.e., let us calculate the expectation of the random variable T. Straightforward

integration by parts yields

$$\begin{aligned}
\mathbb{E}[T] &= \int_0^\infty t\lambda e^{-\lambda t}\,dt \\
&= \lambda\left[\frac{-te^{-\lambda t}}{\lambda}\Big|_0^\infty - \int_0^\infty \frac{-1}{\lambda}e^{-\lambda t}\,dt\right] \\
&= \lambda\left[0+\frac{1}{\lambda}\int_0^\infty e^{-\lambda t}\,dt\right] \\
&= \lambda\left[0+\frac{1}{\lambda}\frac{-e^{-\lambda t}}{\lambda}\Big|_0^\infty\right] \\
&= \lambda\frac{1}{\lambda^2} \\
&= \frac{1}{\lambda}.
\end{aligned}$$

Let us then consider the tricky part. Suppose that T satisfies the assumptions of Definition 7.2. The trick is now to develop assumption (i) of Definition 7.2 into a Cauchy's functional equation and then treat assumption (ii) of Definition 7.2 merely as a **normalizing constant**. Denote

$$\bar{F}_T(t) = \mathbb{P}[T > t],$$

i.e., $\bar{F}_T(t)$ is the **complementary cumulative distribution function** if T. Now, clearly

$$\mathbb{P}[T > t+s, T > s] = \mathbb{P}[T > t+s].$$

Consequently, by Definition 7.2(i)

$$\frac{\bar{F}_T(t+s)}{\bar{F}_T(s)} = \bar{F}_T(t),$$

Thus, by multiplying both sides with $\bar{F}_T(s)$ and then taking logarithms we obtain the **Cauchy's functional equation**

$$\bar{f}(t+s) = \bar{f}(t)+\bar{f}(s),$$

where we have denoted $\bar{f}(t) = \log\bar{F}(t)$. Since $\bar{f}(t)$ is decreasing, the only possible solution to the Cauchy's functional equation is the linear function $\bar{f}(t) = ct$ for some constant c. Consequently, $\bar{F}(t) = e^{ct}$. Thus, for the density function we have

$$\begin{aligned}
f_T(t) &= -\frac{d}{dt}\bar{F}(t) \\
&= -ce^{ct}.
\end{aligned}$$

Finally, by the normalizing assumption, i.e., Definition7.2(ii), we must have $\lambda = -c$.

We have shown that the definitions 7.2 and 7.4 are indeed the same.

Now we are ready to solve Example 7.1

7.6 Example (Waiting in Line, I, Solution)
Let T be Lady Candida's remaining waiting time. Then

$$\begin{aligned}\mathbb{P}[T > 5] &= \int_5^\infty f_T(t)\,\mathrm{d}t \\ &= \int_5^\infty \frac{1}{10} e^{-\frac{1}{10}t}\,\mathrm{d}t \\ &= \mathrm{e}^{-\frac{1}{10}\times 5} \\ &= 0.60653.\end{aligned}$$

So, Lady Candida still has to wait for at least 5 minutes with probability 60.653%.

Sums of Independent Exponential Distribution: Erlang Distribution

A very important problem is to know the distribution of a sum of independent exponential distributions. Indeed, consider the following Example 7.7 that is an extension of the Lady Candida's waiting problem Example 7.1:

7.7 Example (Waiting in Line, II)
Lady Candida needs to powder her nose. She finds the powder room occupied and 8 ladies waiting in line in front of her. From past experience, Lady Candida knows that the average time a lady spends in powdering her nose is 10 minutes. Lady Candida has waited for 45 minutes. The powder room is still occupied and now there are 6 ladies waiting in line in front of her. What is the probability that Lady Candida still has to wait for at least 35 minutes?

After a bit of contemplating the memoryless random variables, one sees that Lady Candida's remaining waiting time is $S = T_1 + T_2 + \cdots + T_7$, where $T_1, \ldots, T_7$ are each independent exponentially distributed random variables with the same mean of 10 minutes. Indeed, there are 6 ladies in the line in front of Lady Candida, and a 7^{th} one in the powder room. Since there is no reason to assume any kind of specific dependence structure for the individual (remaining) waiting times $T_1, T_2, \ldots, T_7$, it is natural to assume that they are independent.

Let us generalize the problem above slightly: we replace 7 by n and average of 10 minutes by $1/\lambda$. So, we need to find the distribution of $S = T_1 + T_2 + \cdots + T_n$, where the summands $T_1, T_2, \ldots, T_n$ are independent exponentially distributed with common parameter λ.

Recall, that if X_1 and X_2 are independent random variables with probability density functions f_{X_1} and f_{X_2}, respectively, then their sum $S_2 = X_1 + X_2$ has the probability density function given by the **continuous convolution**

$$\begin{aligned} f_{S_2}(s) &= (f_{X_1} * f_{X_2})(s) \\ &= \int_{-\infty}^{\infty} f_{X_1}(s-x) f_{X_2}(x)\,\mathrm{d}x. \end{aligned}$$

Indeed, by a **conditioning trick**, and by using **Leibniz formalism**, we see that

$$\begin{aligned} f_{S_2}(s)\,\mathrm{d}s &= \mathbb{P}[S_2 \in \mathrm{d}s] \\ &= \mathbb{P}[X_1 + X_2 \in \mathrm{d}s] \\ &= \int_{x=-\infty}^{\infty} \mathbb{P}[X_2 \in \mathrm{d}x\,,\, X_1 + X_2 \in \mathrm{d}s] \\ &= \int_{x=-\infty}^{\infty} \mathbb{P}[X_2 \in \mathrm{d}x\,,\, X_1 + x \in \mathrm{d}s] \\ &= \int_{x=-\infty}^{\infty} \mathbb{P}[X_1 \in \mathrm{d}s - x]\mathbb{P}[X_2 \in \mathrm{d}x] \\ &= \int_{x=-\infty}^{\infty} f_{X_1}(s-x)\mathrm{d}s f_{X_2}(x)\,\mathrm{d}x \\ &= \left(\int_{x=-\infty}^{\infty} f_{X_1}(s-x) f_{X_2}(x)\,\mathrm{d}x \right) \mathrm{d}s. \end{aligned}$$

By iteration, we see that the sum of $S_n = X_1 + X_2 + \cdots + X_n$ of n independent random variables $X_1, X_2, \ldots, X_n$ with common distribution f_X has the distribution given by the **continuous convolution power**

$$f_{S_n}(s) = f_X^{*n}(s)$$

defined by the recursion

$$\begin{aligned} f_{S_2}(s) &= (f_X * f_X)(s) \\ &= \int_{-\infty}^{\infty} f_X(s-x) f_X(x)\,\mathrm{d}x, \\ f_{S_n}(s) &= (f_{S_{n-1}} * f_X)(s) \\ &= \int_{-\infty}^{\infty} f_{S_{n-1}}(s-x) f_X(x)\,\mathrm{d}x. \end{aligned}$$

Let us then calculate the convolution powers for the exponential distribution. For the

case $S_2 = T_1 + T_2$ we obtain

$$\begin{aligned} f_{S_2}(s) &= \int_{-\infty}^{\infty} f_T(s-t) f_T(t)\,\mathrm{d}t \\ &= \int_0^s \lambda \mathrm{e}^{-\lambda(s-t)} \lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t \\ &= \lambda^2 \int_0^s \mathrm{e}^{-\lambda(s-t)-\lambda t}\,\mathrm{d}t \\ &= \lambda^2 \mathrm{e}^{-\lambda s} \int_0^s \mathrm{d}t \\ &= \lambda^2 \mathrm{e}^{-\lambda s} s. \end{aligned}$$

Consider then the case $S_3 = T_1 + T_2 + T_3$:

$$\begin{aligned} f_{S_3}(s) &= \int_{-\infty}^{\infty} f_{S_2}(s-t) f_T(t)\,\mathrm{d}t \\ &= \int_0^s \lambda^2 (s-t) \mathrm{e}^{-\lambda(s-t)} \lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t \\ &= \lambda^3 \int_0^s (s-t) \mathrm{e}^{-\lambda(s-t)-\lambda t}\,\mathrm{d}t \\ &= \lambda^3 \mathrm{e}^{-\lambda s} \int_0^s (s-t)\,\mathrm{d}t \\ &= \lambda^3 \mathrm{e}^{-\lambda s} \frac{s^2}{2}. \end{aligned}$$

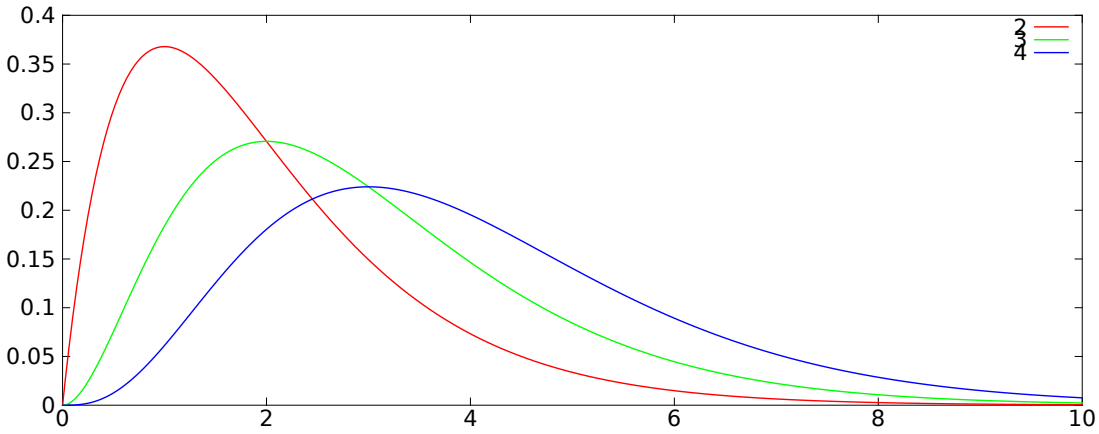
For the cases S_4, S_5 and S_6, similar calculations yield

$$\begin{aligned} f_{S_4}(s) &= \lambda^4 \mathrm{e}^{-\lambda s} \frac{s^3}{2 \times 3}, \\ f_{S_5}(s) &= \lambda^5 \mathrm{e}^{-\lambda s} \frac{s^4}{2 \times 3 \times 4}, \\ f_{S_6}(s) &= \lambda^6 \mathrm{e}^{-\lambda s} \frac{s^5}{2 \times 3 \times 4 \times 5}. \end{aligned}$$

(Don't believe me! Do the calculations yourself!) Now, because of the calculations above, the case for general n is pretty obvious:

7.8 Proposition (Exponential Sum)
Let $T_1, T_2, \ldots, T_n$ be independent exponentially distributed random variables with common parameter λ. Then their sum $S = T_1 + T_2 + \cdots + T_n$ has the **Erlang distribution** with parameters n and λ, i.e., S has the probability density function

$$f_S(s) = \frac{\lambda^n}{(n-1)!} s^{n-1} \mathrm{e}^{-\lambda s}, \qquad s \geq 0.$$



Probability density functions of Erlang distributions with $\lambda = 1$ and $n = 2, 3, 4$.

The picture above was generated with the Octave code listed below:

```
1  ##############################################################################
2  ## FILE: erlang_pdfs.m
3  ##
4  ## Plots some probability density functions (pdf) of Erlang distributions.
5  ##############################################################################
6
7  lambda = 1;                                  ## Parameter lambda is fixed.
8  n = [2, 3, 4];                               ## Parameter n varies.
9  x = linspace(0, 10, 5000);                   ## The x-grid for Px=prob(x).
10
11 ## Gamma (or Erlang) densities for the parameters n and lambda.
12 Px(1,:) = gampdf(x, n(1), lambda);
13 Px(2,:) = gampdf(x, n(2), lambda);
14 Px(3,:) = gampdf(x, n(3), lambda);
15
16 ## Plotting.
17 hold on;                                     ## Plot all on the same pic.
18 plot(x, Px(1,:), '1;2;','linewidth', 4);
19 plot(x, Px(2,:), '2;3;','linewidth', 4);
20 plot(x, Px(3,:), '3;4;','linewidth', 4);
```

www.uva.fi/~tsottine/psp/erlang_pdfs.m

Now we are ready almost to solve Exercise 7.7. We just have to calculate the integral

$$\mathbb{P}[S > s] \quad = \quad \int_s^\infty \frac{\lambda^n}{(n-1)!} u^{n-1} \mathrm{e}^{-\lambda u} \, \mathrm{d}u.$$

for $n = 7$, $\lambda = 1/10$ and $s = 35$. This integral can be calculated by repeated use of integration by parts to reduce the parameter n. Indeed, for $n = 1$ we have simply

$$\int_s^\infty \lambda \mathrm{e}^{-\lambda u} \, \mathrm{d}u \quad = \quad \mathrm{e}^{-\lambda s}.$$

For $n = 2$, we have, by using integration by parts, and by using the result above, that

$$\begin{aligned}
\int_s^\infty \lambda^2 u e^{-\lambda u}\,du &= -\int_s^\infty \lambda u\,(-\lambda)e^{-\lambda u}\,du \\
&= -\int_s^\infty \lambda u\,\frac{d}{du}\left[e^{-\lambda u}\right]du \\
&= -\left[\lambda u e^{-\lambda u}\Big|_s^\infty - \int_s^\infty \frac{d}{du}[\lambda u]e^{-\lambda u}\,du\right] \\
&= \lambda s e^{-\lambda s} + \int_s^\infty \lambda e^{-\lambda u}\,du \\
&= \lambda s e^{-\lambda s} + e^{-\lambda s} \\
&= e^{-\lambda s}[1+\lambda s].
\end{aligned}$$

For $n = 3$, we have, by using integration by parts, and by using the calculations above, that

$$\begin{aligned}
\int_s^\infty \frac{\lambda^3}{2}u^2 e^{-\lambda u}\,du &= -\int_s^\infty \frac{\lambda^2}{2}u^2\,(-\lambda)e^{-\lambda u}\,du \\
&= -\int_s^\infty \frac{\lambda^2}{2}u^2\,\frac{d}{du}\left[e^{-\lambda u}\right]du \\
&= -\left[\frac{\lambda^2}{2}u^2 e^{-\lambda u}\Big|_s^\infty - \int_s^\infty \frac{d}{du}\left[\frac{\lambda^2}{2}u^2\right]e^{-\lambda u}\,du\right] \\
&= \frac{\lambda^2}{2}s^2 e^{-\lambda s} + \int_s^\infty \lambda^2 u e^{-\lambda u}\,du \\
&= \frac{\lambda^2}{2}s^2 e^{-\lambda s} + e^{-\lambda s}[1+\lambda s] \\
&= e^{-\lambda s}\left[1+\lambda s+\frac{1}{2}(\lambda s)^2\right].
\end{aligned}$$

Now the pattern is clear (and can be proven by using the induction argument): we have

$$\mathbb{P}[S > s] = e^{-\lambda s}\sum_{k=0}^{n-1}\frac{1}{k!}(\lambda s)^k.$$

7.9 Remark (Gamma Functions)
The **gamma function**

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$$

and the (lower) **incomplete gamma function**

$$\gamma(z,x) = \int_0^x t^{z-1}e^{-t}\,dt$$

are built in numerically in many mathematical software. The probabilities $\mathbb{P}[S > s]$ can be expressed in terms of these gamma functions. Indeed, after a little bit of algebra (Exercise 7.5) one sees that if S has Erlang distribution with parameters n and λ, then

$$\mathbb{P}[S > s] \quad = \quad 1 - \frac{\gamma(n, \lambda s)}{\Gamma(n)}.$$

7.10 Example (Waiting in Line, II, Solution)
Let S be Lady Candida's remaining waiting time. By no-memory and independence assumption $S = T_1 + \cdots + T_7$ where $T_1, \ldots, T_7$ are independent exponential random variables with common parameter $\lambda = 1/10$. Consequently S has Erlang distribution with parameters $n = 7$ and $\lambda = 1/10$. Thus

$$\mathbb{P}[S > 35] \quad = \quad \mathrm{e}^{-3.5} \sum_{k=0}^{6} \frac{1}{k!} 3.5^k.$$

This probability can be calculated by using the following code:

```
p = 0;
for k = 0:6
    p = p + 1/factorial(k)*(3.5)^k;
endfor
p = e^(-3.5)*p
```

So, We obtain the result 0.93471. So, the probability that Lady Candida still has to wait at least 35 minutes is 93.471%.

Exercises

7.1 Exercise
Let T be exponentially distributed with parameter 5. Calculate

(a) $\mathbb{P}[T \leq 3]$
(b) $\mathbb{P}[T \geq 3]$
(c) $\mathbb{P}[2 \leq T \leq 3]$
(d) $\mathbb{P}[T \geq 3 \,|\, T \geq 1]$
(e) $\mathbb{P}[T \leq 3 \,|\, T \geq 1]$
(f) $\mathbb{P}[2 \leq T \leq 3 \,|\, T \geq 1]$

7.2 Exercise
Let T be exponentially distributed with parameter λ. Calculate its moment generating function, and by using it how that $\mathbb{E}[T] = \frac{1}{\lambda}$ and $\mathbb{V}[T] = \frac{1}{\lambda^2}$

7.3 Exercise
Let S be Erlang distributed with parameters $n = 3$ and $\lambda = 5$. Calculate the probabilites

(a) $\mathbb{P}[S \geq 9]$
(b) $\mathbb{P}[S \leq 9]$
(c) $\mathbb{P}[6 \leq S \leq 9]$
(d) $\mathbb{P}[6 \leq S \leq 9 \text{ or } S \geq 21]$

7.4 Exercise
Show that the gamma function $\Gamma(z)$ is a generalization of the factorial, i.e.,

$$\Gamma(n) = (n-1)!$$

if $n \in \mathbb{N}$.

7.5 Exercise
Let S be Erlang distributed with parameters n and λ. Show that

$$\mathbb{P}[S > s] \quad = \quad 1 - \frac{\gamma(n, \lambda s)}{\Gamma(n)}.$$

Lecture 8

Gaussian Distribution

The normal distribution, or the Gaussian distribution is named after the German mathematician **Johann Carl Friedrich Gauss** (1777–1855) who certainly was *Princeps mathematicorum*, but he does not deserve to have the normal distribution named after him. Gauss did some remarkable work with the normal distribution in his 1821–1826 essays *Theoria combinationis observationum erroribus minimis obnoxiae* about measuring errors, but the true reason for the importance of Gaussian distribution is due to the central limit theorem, and Gauss did not study that.

The first glimpse of the central limit theorem was given by the French mathematician **Abraham de Moivre** (1667–1754) who showed that the binomial distribution can be approximated by the Gaussian distribution. The first one to really discover the central limit theorem was the French mathematician **Pierre-Simon Laplace** (1749–1827) in his book *Théorie Analytique des Probabilités* published in 1812. The final word in its modern generality for the central limit theorem was given by the Finnish mathematician **Jarl Waldemar Lindeberg** (1876-1932) and the Croatian-American mathematician **William Feller** (1906–1970).

Pierre-Simon, marquis de Laplace (1749–1827)

The Gaussian distribution, or the normal distribution, is arguably the most important distribution in all probability. The reason is of course the celebrated central limit theorem. The key Example 8.1 of this section is an application of the central limit theorem. Indeed, it is very difficult to see how one could solve Example 8.1 without resorting to the central limit theorem.

8.1 Example (Bleedingheart Charity)
Ms. Bleedingheart is collecting donations for Good Cause Charity. She wants to raise 100 000 euros. She has already contacted 6 donors who have contributed 950, 800, 1 000, 1 100, 850 and 1 000 euros, respectively. How many donors Ms. Bleedingheart needs still to contact in order to have at least 95% probability of reaching her goal of 100 000 euros?

Gaussian Distribution Quantitatively

We will introduce the Gaussian distribution here **quantitatively**. The **qualitative** approach will be served by the **central limit theorem** in the next section.

8.2 Definition (Gaussian Distribution)

(i) A continuous random variable X has **standard Gaussian distribution** if it has the density function

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \qquad x \in \mathbb{R}.$$

In this case we denote $X \sim N(0,1)$. We also denote

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}z^2}\, dz, \qquad x \in \mathbb{R}$$

for the cumulative distribution function of the standard Gaussian random variable.

(ii) A continuous random variable Y has **Gaussian distribution** with parameters μ and σ^2 if there exist a random variable $X \sim N(0,1)$ such that $Y = \sigma X + \mu$. In this case we denote $Y \sim N(\mu, \sigma^2)$. We also denote by $\phi(x;\mu,\sigma^2)$ and $\Phi(x;\mu,\sigma^2)$ the probability density function and the cumulative distribution function of Y.

The parameters μ and σ^2 have the usual interpretation. This is the message of the following proposition, the proof of which is left as Exercise 8.3.

8.3 Proposition
Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{E}[X] = \mu$ and $\mathbb{V}[X] = \sigma^2$.

The density function of the Gaussian random variable $Y \sim N(\mu, \sigma^2)$ can be calculated by changing the variable. Indeed, let us start with the cumulative distribution function:

$$\begin{aligned} \Phi(y;\mu,\sigma^2) &= \mathbb{P}[Y \le y] \\ &= \mathbb{P}[\mu + \sigma X \le y] \\ &= \mathbb{P}\left[X \le \frac{y-\mu}{\sigma}\right] \\ &= \Phi\left(\frac{y-\mu}{\sigma}\right). \end{aligned}$$

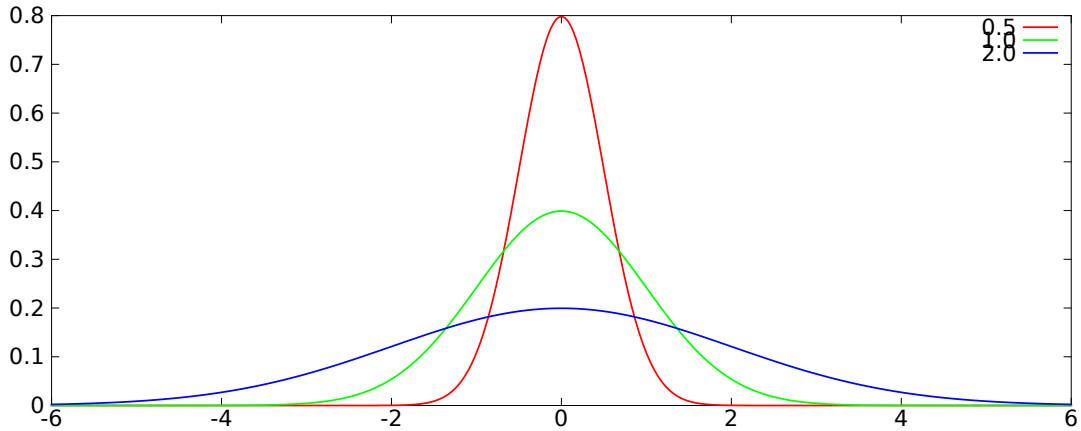
Consequently, by taking the derivatives, we obtain

$$\begin{aligned}\phi(y;\mu,\sigma^2) &= \frac{\mathrm{d}}{\mathrm{d}y}\Phi\left(\frac{y-\mu}{\sigma}\right)\\ &= \frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}.\end{aligned}$$

Thus we have shown the following:

8.4 Proposition
$Y \sim \mathrm{N}(\mu,\sigma^2)$ if and only if it is a continuous random variable with density function

$$\phi(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, \qquad y \in \mathbb{R}.$$



Probability density functions $\phi(x;0,\sigma^2)$ of the $\mathrm{N}(0,\sigma^2)$ distributions with different parameters σ.

The picture above was generated with the Octave code listed below:

```
1  ##############################################################################
2  ## FILE: gauss_pdfs.m
3  ##
4  ## Plots some probability density functions (pdf) of Gaussian distributions.
5  ##############################################################################
6
7  ## Data for plots.
8  sigma = [0.5, 1, 2];                    ## sigma varies.
9  y = linspace(-6, 6, 5000);              ## The grid for Py=prob(y) plotted.
10 Py(1,:) = normpdf(y, 0, sigma(1));      ## 1st row for sigma(1).
11 Py(2,:) = normpdf(y, 0, sigma(2));      ## 2nd row for sigma(2).
12 Py(3,:) = normpdf(y, 0, sigma(3));      ## 3rd row for sigma(3).
13
14 ## Plotting (standard plot).
```

```
15 hold on;
16 plot(y, Py(1,:), '1;0.5;','linewidth', 4);
17 plot(y, Py(2,:), '2;1.0;','linewidth', 4);
18 plot(y, Py(3,:), '3;2.0;','linewidth', 4);
```

www.uva.fi/~tsottine/psp/gauss_pdfs.m

Unfortunately, there is no way of calculating the integral $\Phi(x) = \int_{-\infty}^{x} \phi(z)\,dz$ analytically. Thus, in computing probabilities for Gaussian random variables one has to resort to numerical methods. Luckily, the function $\Phi(x)$ or its relative the error function $\mathrm{erf}(x)$ are implemented in all mathematical software and programming languages worth their salt. In Octave, there is the function `stdnormal_cdf(x)` that is the same as $\Phi(x)$. Also, there is the function `stdnormal_pdf(x)` that is the same as $\phi(x)$. Moreover there are functions `normcdf(y,m,s)` and `normpdf(y,m,s)` that work just like the functions $\Phi(y;m,s^2)$ and $\phi(y;m,s^2)$.

Fun Fact

There is the **fun fact** that the standard Gaussian density is the only probability density function satisfying the differential equation

$$\phi'(x) \;=\; -x\phi(x).$$

Indeed, it is easy to check that the function $\phi(x)$ defined by Definition 8.2(i) satisfies the fun fact and the uniqueness follows from the general theory of ordinary differential equations. What makes this fact even funnier, is that the same differential equation for the **characteristic function** also determines the Gaussian distribution. Indeed, let $X \sim \mathrm{N}(0,1)$ and let $\varphi(\theta) = \varphi_X(\theta)$ be its characteristic function. Then, on the one hand, by symmetry,

$$\begin{aligned}\varphi(\theta) &= \int_{-\infty}^{\infty} \cos(\theta x)\phi(x)\,dx + \mathrm{i}\int_{-\infty}^{\infty} \sin(\theta x)\phi(x)\,dx \\ &= \int_{-\infty}^{\infty} \cos(\theta x)\phi(x)\,dx.\end{aligned}$$

On the other, by changing the order of differentiation and integration

$$\begin{aligned}\varphi'(\theta) &= \frac{\mathrm{d}}{\mathrm{d}\theta}\int_{-\infty}^{\infty} \cos(\theta x)\phi(x)\,dx \\ &= \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta}\left[\cos(\theta x)\phi(x)\right] dx \\ &= \int_{-\infty}^{\infty} \sin(\theta x)\; x\phi(x)\; dx\end{aligned}$$

Then, by the fun fact and by the integration by parts,

$$\begin{aligned}\varphi'(\theta) &= -\int_{-\infty}^{\infty} \sin(\theta x)\, \phi'(x)\, \mathrm{d}x \\ &= -\theta \int_{-\infty}^{\infty} \cos(\theta x)\phi(x)\, \mathrm{d}x \\ &= -\theta\varphi(\theta).\end{aligned}$$

So, we see that we have the fun fact differential equation for the Gaussian characteristic function also:

$$\varphi'(\theta) = -\theta\varphi(\theta).$$

We note that since $\varphi(0) = 1$ for any characteristic function, the only solution to the fun fact differential equation for the characteristic function is

$$\varphi(\theta) = \mathrm{e}^{-\frac{1}{2}\theta^2}.$$

Finally, since for any random variable Y and any numbers a and b we have

$$\begin{aligned}\varphi_{a+bY}(\theta) &= \mathbb{E}\left[\mathrm{e}^{\mathrm{i}\theta(a+bY)}\right] \\ &= \mathbb{E}\left[\mathrm{e}^{\mathrm{i}\theta a}\mathrm{e}^{\mathrm{i}\theta bY}\right] \\ &= \mathrm{e}^{\mathrm{i}a\theta}\mathbb{E}\left[\mathrm{e}^{\mathrm{i}(b\theta)Y}\right] \\ &= \mathrm{e}^{\mathrm{i}a\theta}\varphi_Y(b\theta),\end{aligned}$$

we see that we have the following result:

8.5 Proposition
Let $Y \sim \mathrm{N}(\mu, \sigma^2)$. Then

$$\varphi_Y(\theta) = \mathrm{e}^{\mathrm{i}\mu\theta - \frac{1}{2}\sigma^2\theta^2}.$$

Central Limit Theorem

One way of expressing the central limit theorem is to state that if a random variable can be thought to be a sum of many small independent components that are similar, then the distribution of the random variable is close to Gaussian. Theorem 8.6 states the same thing in more technical language.

8.6 Theorem (Central Limit Theorem or Normal Approximation)
Let $X_1, X_2, \ldots$ be independent identically distributed random variables with common expectation μ and variance σ^2. Denote

$$S_n \quad = \quad \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{X_k - \mu}{\sigma}.$$

Then S_n is **asymptotically normal** in the sense that

$$\lim_{n\to\infty} \mathbb{P}[S_n \leq s] \quad = \quad \Phi(s)$$

for all $s \in \mathbb{R}$.

There are many proofs for the central limit theorem, but none is as simple as the complex-analytic proof that uses **characteristic functions**. Basically, all that is needed is **Levy's continuity theorem** of Lemma 4.21 and Proposition 8.5. We give this proof below.

Let us first note that we may take, without any loss of generality, that $\mu = 0$ and $\sigma = 1$. In other words, we can consider the standardized sum

$$S_n \quad = \quad \frac{1}{\sqrt{n}} [X_1 + X_2 + \cdots + X_n],$$

where the summands X_k are normalized so that $\mathbb{E}[X_k] = 0$ and $\mathbb{V}[X_k] = 1$. Then all we have to do is to show that

$$\varphi_{S_n}(\theta) \quad \to \quad \mathrm{e}^{-\frac{1}{2}\theta^2}.$$

Now,

$$\begin{aligned} \mathrm{e}^{\mathrm{i}\theta S_n} &= \mathrm{e}^{\mathrm{i}\theta\left(\frac{1}{\sqrt{n}}\sum_{k=1}^{n} X_k\right)} \\ &= \prod_{k=1}^{n} \mathrm{e}^{\mathrm{i}\frac{\theta}{\sqrt{n}} X_k}. \end{aligned}$$

Consequently, we always have

$$\varphi_{S_n}(\theta) \quad = \quad \mathbb{E}\left[\prod_{k=1}^{n} \mathrm{e}^{\mathrm{i}\frac{\theta}{\sqrt{n}} X_k}\right].$$

Since the summands X_k are **independent and identically distributed**, we have

$$\begin{aligned} \varphi_{S_n}(\theta) &= \prod_{k=1}^{n} \mathbb{E}\left[\mathrm{e}^{\mathrm{i}\frac{\theta}{\sqrt{n}} X_k}\right] \\ &= \varphi_{X_1}\left(\frac{\theta}{\sqrt{n}}\right)^n \end{aligned}$$

Next, we use the following second-order **Taylor approximation** for characteristic functions. First, we note that

$$\begin{aligned}\varphi'(\theta) &= \mathrm{i}\mathbb{E}\left[X\mathrm{e}^{\mathrm{i}\theta X}\right],\\ \varphi''(\theta) &= -\mathbb{E}\left[X^2\mathrm{e}^{\mathrm{i}\theta X}\right].\end{aligned}$$

Therefore, for a random variable with mean zero and variance one, we have

$$\begin{aligned}\varphi(0) &= 1,\\ \varphi'(0) &= 0,\\ \varphi''(0) &= -1,\end{aligned}$$

and the Taylor approximation becomes

$$\varphi(\theta) = 1-\frac{1}{2}\theta^2+\varepsilon(\theta)\theta^2,$$

where $\varepsilon(\theta)\to 0$ as $\theta\to 0$. Replacing θ with $\theta/\sqrt{n}$ we obtain:

$$\begin{aligned}\varphi_{S_n}(\theta) &= \varphi_{X_1}\left(\frac{\theta}{\sqrt{n}}\right)^n\\ &= \left(1+\frac{-\frac{1}{2}\theta^2}{n}+\varepsilon\left(\frac{\theta}{\sqrt{n}}\right)\frac{\theta^2}{n}\right)^n\\ &\to \mathrm{e}^{-\frac{1}{2}\theta^2},\end{aligned}$$

where the last line follows from the very definition of the constant e.

8.7 Remark (Is Central Limit Theorem Obsolete?)
Traditionally one has used the central limit theorem to ease the calculations of e.g. binomial distributions. Nowadays, with computers, such usage is mostly pointless. Indeed, if one knows the distributions of the independent summands, the distribution of their sum is simply a convolution. Of course, calculating convolutions by hand is very tedious and often there are no analytical solutions. However, computers can calculate such things numerically very fast, at least sometimes. Thus, in the 21st century the central limit theorems is mainly useful in the case where the distribution of the summands is unknown. But this is the case of Example 8.1, which we are now, finally, ready to solve.

8.8 Example (Bleedingheart Charity, Solution)
Let us model each single donation as an independent random variable X_k. So, if there are n donors, then the total amount donated is $S_n=\sum_{k=1}^n X_k$. From the 6 donations we

estimate the average donation μ and its variance σ^2 as

$$\begin{aligned}\hat{\mu} &= \frac{1}{6}(950+800+1\,000+1\,100+850+1000)\\ &= 950,\\ \hat{\sigma}^2 &= \frac{1}{5}\left(0^2+150^2+50^2+150^2+100^2+50^2\right)\\ &= 12\,000.\end{aligned}$$

Now, Ms. Bleedingheart has already raised 5 700 euros, so we are looking for n such that $\mathbb{P}[S_n \geq 94\,300] \geq 0.95$. By the central limit theorem we assume that S_n is approximately $\mathrm{N}(n\hat{\mu}, n\hat{\sigma}^2)$ distributed, for big n. In other words,

$$\begin{aligned}\mathbb{P}[S_n \geq 94\,300] &= \mathbb{P}\left[\frac{S_n - n\hat{\mu}}{\sqrt{n}\hat{\sigma}} \geq \frac{94\,300 - n\times 950}{\sqrt{n}\times 109.5}\right]\\ &\approx 1-\Phi\left(\frac{94\,300-n\times 950}{\sqrt{n}\times 109.5}\right).\end{aligned}$$

So, by rearranging the equation and then taking the standard Gaussian quantile function $\Phi^{-1}(q)$ on both sides we obtain the criterion

$$\Phi^{-1}(5\%) \approx \frac{94\,300-n\times 950}{\sqrt{n}\times 109.5}.$$

Octave's function `stdnormal_inv(x)` that means $\Phi^{-1}(x)$ tells us that $\Phi^{-1}(0.05) = -1.6449$. Thus we get the criterion

$$-1.6449 \approx \frac{94\,300-n\times 950}{\sqrt{n}\times 109.5}.$$

This can be developed into a quadratic equation in n and thus solved analytically. However, it is easier simply to check for different values of n in Octave. We obtain that Ms. Bleedingheart should still contact $n \approx 102$ donors.

Stirling's Approximation

The central limit theorem of Theorem 8.6 is closely related to the law of small numbers of Theorem 6.10. Let us investigate one aspect of their interconnection by giving a less-than-rigorous proof of the extremely useful **Stirling's approximation** of the factorial.

Let S be a Poisson random variable with mean n. We know that S can be thought to be an exact sum of many independent identically distributed Poisson random variables or an approximate sum of many independent identically distributed binomial random variables.

Thus, on the one hand, we can apply the central limit theorem:

$$\begin{aligned}\mathbb{P}[S=n] &= \mathbb{P}[n-1<S\leq n]\\ &= \mathbb{P}\left[-\frac{1}{\sqrt{n}}<\frac{S-n}{\sqrt{n}}\leq 0\right]\\ &\approx \int_{-\frac{1}{\sqrt{n}}}^{0}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}x^2}\,\mathrm{d}x\\ &\approx \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{n}}.\end{aligned}$$

On the other hand, we can calculate explicitly

$$\mathbb{P}[S=n] = \mathrm{e}^{-n}\frac{n^n}{n!}.$$

By equating the exact solution and the approximate one, we obtain the following approximation for the factorial.

8.9 Lemma (Stirling's Approximation)
Let n be big. Then

$$n! \approx \sqrt{2\pi}\,n^{n+\frac{1}{2}}\,\mathrm{e}^{-n}.$$

Exercises

8.1 Exercise
Let $Y\sim \mathrm{N}(-1,0.6)$. Calculate

(a) $\mathbb{P}[Y\leq 0]$
(b) $\mathbb{P}[Y\geq 0]$
(c) $\mathbb{P}[-2\leq Y\leq 0]$
(d) $\mathbb{P}[Y\leq -3 \text{ or } 0\leq Y\leq 1]$

8.2 Exercise
Consider Example 8.1.

(a) Suppose Ms. Bleedingheart receives additional 3 donations of 945, 950 and 955 euros to the 6 donations she already has received. How many donations she still needs to have 95% chance of meeting her goal 100 000 euros?

(b) Suppose Ms. Bleedingheart receives additional 3 donations of 5, 10 and 3 000 euros to the 6 donations she already has received. Will Ms. Bleedingheart need more ore less additional donors to meet her goal than in the case (a) above?

8.3 Exercise
Prove Proposition 8.3.

8.4 Exercise
Let $Y_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $Y_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ be independent. Calculate the density function of the random variable $Y_1 + Y_2$.

8.5 Exercise
Write an Octave program that solves problems of the Bleedingheart Charity type of Example 8.1. The input parameters of the problem should be the goal (100 000 euros in the example), vector of received donations ($[950, 800, 1\ 000, 1\ 100, 850, 1000]$ in the example) and the level of certainty required (95% in the example). The output is n, the number of additional donors required.

8.6 Exercise
(i) Use Stirling's approximation to the **binomial coefficients**

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

where $k << n$, and both k and n are large.

(ii) Implement the approximation of part (i) with Octave, and compare it with the exact calculations.

Part III

Stochastic Processes

Lecture 9

Markov Chains as Matrices

The Markov chains are named after the Russian mathematician **Andrey Andreyevich Markov** (1856–1922) who introduced the concept in 1906. Markov's motivation was apparently to show that the law of large numbers can hold for dependent random variables. Indeed, the word "chain" was used by Markov to suggest a sequence of pairwise dependent variables. By proving the law of large numbers for dependent random variables, i.e., to Markov chains, Markov won a theological argument on the existence of free will. Indeed, **Pavel Nekrasov** and the Moscow School of Mathematics argued that the law of large numbers must imply independence which was interpreted as free will. Markov showed that Nekrasov was wrong. Later, in 1913, Markov applied his chains to analyze the 20 000 first words of **Alexander Pushkin's** novel in verse *Eugene Onegin*.

More practical applications of Markov chains followed soon. Already in 1917 the Danish mathematician **Agner Krarup Erlang** used Markov chains to obtain formulas for call loss probabilities and waiting time distributions in telephone networks. Later Markov chains have found applications in all areas of science and art. It is virtually impossible to underestimate their importance.

Andrey Markov (1856–1922)

Mathematical models are always simplifications of the reality. In probabilistic modeling the most natural simplification is independence. Indeed, random variables can be dependent in infinitely many different ways, but independent only in one way. Unfortunately, sometimes the independence assumption is simply silly. Indeed, it would be completely nuts to argue that the daily (mean) temperatures are independent, since warm days are typically followed by warm ones, and cold days by cold ones. So, one must often assume some kind on dependence. Markov chains assume a weak form of dependence that can be stated as **the future depends on the past only through the present**. In the case of daily temperatures that would mean that the probability distribution of tomorrow's temperature depends only on today's temperature and not on the temperatures yesterday, the day before yesterday, and so on. This kind of modeling is not completely silly, as assuming independence would be, but it is still a simplification. It is well-known that there is long-range dependence in weather: temperatures hundred years ago still have effect on

tomorrows temperature. Luckily this effect is tiny, so Markovian modeling is practical for many purposes.

The key Example 9.1 of this lecture illustrates how "natural" assumptions lead into a Markovian model. In the example it is clear that we have to make assumptions, since the only statistical data we have is a single number: 0.3.

9.1 Example (Bonus–Malus Insurance)
Most automobile insurance premiums are determined by a bonus–malus system, i.e., the policyholder will get a lower or higher premium the next year depending on the number of claims she has made on the previous years. Suppose the upgrades or downgrades for the policyholder are determined by the following table. The fist column gives the discount to annual premium, the second column gives the years with no claims needed for upgrade and the third, fourth and fifth column give the next years discount if there has been 1, 2 or more than 2 claims in the current year.

Discount	**Years**	**1 claim**	**2 claims**	>2 **claims**
0%	3	0%	0%	0%
20%	1	0%	0%	0%
40%	2	20%	0%	0%
60%	N/A	40%	20%	0%

Suppose that an average policyholder has 0.3 claims per year. What is the probability that an average new policyholder that starts with 0% discount will have 60% discount after 10 years and continues to have the 60% discount for the successive 5 years?

Markovian Modeling

A **stochastic process** is a family of random variables indexed by time. In this part of the lectures we consider **discrete time**, i.e., each time point has a previous and a next time point. Consequently, stochastic processes are sequences of random variables X_n, $n \in \mathbb{N}$, where n is the time index. Time $n = 0$ is typically interpreted as "now". When we are using Octave, we will usually assume that the time "now" is $n = 1$, since Octave starts indexing with 1.

The **state-space**, denoted by $\mathbb{S}$, of a stochastic process X_n, $n \in \mathbb{N}$, is the space of all possible values of the random variables X_n. In this part we assume that also the state-space is discrete. This means that we can write, e.g., $\mathbb{S} = \{s_0, s_1, s_2 \ldots\}$. In practice the state-space is usually either $\mathbb{Z}$, $\mathbb{N}$ or finite.

9.2 Remark (Practical Modeling)
To give a **complete probabilistic description** of a discrete-time discrete-state stochastic process one has to determine the all the **joint probabilities**

$$\mathbb{P}\big[X_0 = s_{i_0}, X_1 = s_{i_1}, X_2 = s_{i_2}, \ldots, X_n = s_{i_n}\big].$$

for all $n \in \mathbb{N}$ and $s_{i_0}, \ldots, s_{i_n} \in \mathbb{S}$. This is obviously impractical! To make things practical, one has to make simplifying assumptions. The **Markovian assumption**, i.e., **the future is independent of the past given the present** , leads to the following formal definition.

9.3 Definition (Markov Chain)
A discrete-time stochastic process X_n, $n \in \mathbb{N}$, with discrete state-space $\mathbb{S}$ is a **Markov chain** if it satisfies the **Markovian assumption**

$$\mathbb{P}\big[X_{n+1} = s_{i_{n+1}} \,\big|\, X_0 = s_{i_0}, X_1 = s_{i_1}, \ldots, X_n = s_{i_n}\big] \;=\; \mathbb{P}\big[X_{n+1} = s_{i_{n+1}} \,\big|\, X_n = s_{i_n}\big].$$

If $\mathbb{P}[X_{n+1} = s_j | X_n = s_i]$ is independent of n, then the Markov chain is **time-homogeneous**. In this case the matrix $\mathbf{P} = [P_{ij}]$, where

$$P_{ij} \;=\; \mathbb{P}\big[X_{n+1} = s_j \,\big|\, X_n = s_i\big],$$

is the **transition probability matrix** of the time-homogeneous Markov chain X_n, $n \in \mathbb{N}$.

9.4 Remark (Homogeneity Assumption)
In what follows we will almost always assume that our Markov chains are time-homogeneous and do not bother stating it out explicitly. Let us note, however, that the assumption of time-homogeneity is sometimes very silly. Indeed, consider e.g. daily temperatures. It makes huge difference whether n denotes a summer day or a winter day.

To give a complete probabilistic description of a Markov chain one only needs its transition probability matrix $\mathbf{P} = [P_{ij}]$ and an **initial distribution** or **initial probability** $\mathbf{p} = [p_i]$, where

$$p_i \;=\; \mathbb{P}[X_0 = s_i].$$

Indeed, by repetitive use of the **product rule** $\mathbb{P}[A,B] = \mathbb{P}[A]\mathbb{P}[B|A]$, we can write the complete probabilistic description as

$$\begin{aligned}
&\mathbb{P}\big[X_0 = s_{i_0}, X_1 = s_{i_1}, X_2 = s_{i_2}, \ldots, X_n = s_{i_n}\big] \\
&\quad= \mathbb{P}\big[X_0 = s_{i_0}\big]\,\mathbb{P}\big[X_1 = s_{i_1} \,\big|\, X_0 = s_{i_0}\big]\,\mathbb{P}\big[X_2 = s_{i_2} \,\big|\, X_0 = s_{i_0}, X_1 = s_{i_1}\big]\cdots \\
&\qquad\cdots\mathbb{P}\big[X_n = s_{i_n} \,\big|\, X_0 = s_{i_0}, X_1 = s_{i_1}, \ldots, X_{n-1} = s_{i_{n-1}}\big].
\end{aligned}$$

By the Markovian assumption we obtain

$$\begin{aligned}
&\mathbb{P}\big[X_0 = s_{i_0}, X_1 = s_{i_1}, X_2 = s_{i_2}, \ldots, X_n = s_{i_n}\big] \\
&\quad = \mathbb{P}\big[X_0 = s_{i_0}\big]\,\mathbb{P}\big[X_1 = s_{i_1} \,\big|\, X_0 = s_{i_0}\big]\,\mathbb{P}\big[X_2 = s_{i_2} \,\big|\, X_1 = s_{i_1}\big]\cdots \\
&\qquad \cdots \mathbb{P}\big[X_n = s_{i_n} \,|\, X_{n-1} = s_{i_{n-1}}\big],
\end{aligned}$$

and finally, by the time-homogeneity, we obtain

$$\begin{aligned}
&\mathbb{P}\big[X_0 = s_{i_0}, X_1 = s_{i_1}, X_2 = s_{i_2}, \ldots, X_n = s_{i_n}\big] \\
&\quad = \mathbb{P}\big[X_0 = s_{i_0}\big]\,\mathbb{P}\big[X_1 = s_{i_1} \,\big|\, X_0 = s_{i_0}\big]\,\mathbb{P}\big[X_1 = s_{i_2} \,\big|\, X_0 = s_{i_1}\big]\cdots \\
&\qquad \cdots \mathbb{P}\big[X_1 = s_{i_n} \,|\, X_0 = s_{i_{n-1}}\big] \\
&\quad = p_{i_0} P_{i_0,i_1} P_{i_1,i_2} \cdots P_{i_{n-1},i_n}.
\end{aligned}$$

9.5 Example (Branching Process)
Recall the branching process, X_n, $n \in \mathbb{N}$, with offspring distribution $\mathbf{q}$. This is a Markov chain with state space $\mathbb{S} = \mathbb{N}$ and initial distribution $\mathbb{P}[X_0 = 1] = 1$. We know that its transition probabilities are given by the convolution power

$$P_{ij} = q_j^{*i}.$$

These transition probabilities can also be calculated as follows:

$$\begin{aligned}
P_{i,0} &= q_0^i, \\
P_{i,1} &= \binom{i}{1} q_1 q_0^{i-1}, \\
P_{i,2} &= \binom{i}{1} q_2 q_0^{i-2} + \binom{i}{2} q_1^2 q_0^{i-2}, \\
P_{i,3} &= \binom{i}{1} q_3 q_0^{i-1} + \binom{i}{2} q_2 q_1 q_0^{i-2} + \binom{i}{3} q_1^3 q_0^{i-3}, \\
P_{i,4} &= \binom{i}{1} q_4 q_0^{i-1} + \binom{i}{2} q_3 q_1 q_0^{i-2} + \binom{i}{2} q_2^2 q_0^{i-2} + \binom{i}{3} q_2 q_1^2 q_0^{i-3} + \binom{i}{4} q_1^4 q_0^{i-4}, \\
&\vdots
\end{aligned}$$

The general formula should be clear, but difficult to write.

The probabilistic nature of a Markov chain is completely determined by its initial distribution $\mathbf{p}$ and transition probability matrix $\mathbf{P}$. The reverse is also true, in a certain sense. Of course not all (possibly ∞-dimensional) vectors $\mathbf{p}$ and matrices $\mathbf{P}$ correspond to a Markov chain. Obviously, the vector $\mathbf{p}$ has to be a **probability vector**, i.e.

$$p_i \geq 0 \quad \text{for all} \quad s_i \in \mathbb{S} \qquad \text{and} \qquad \sum_{i;\, s_i \in \mathbb{S}} p_i = 1.$$

A similar condition is needed for the matrix $\mathbf{P}$:

9.6 Definition (Stochastic Matrix)
A (possibly $\infty \times \infty$) square matrix $\mathbf{P} = [P_{ij}]$ is a **stochastic matrix** if each of its rows are probability vectors, i.e.,

(i) $P_{ij} \geq 0$ for all i, j.
(ii) $\sum_j P_{ij} = 1$ for all i.

Let $\mathbb{I}$ be an index set that is either finite, $\mathbb{N}$ or $\mathbb{Z}$. Now we can generate realizations of a Markov chain as with state-space $\mathbb{S} = \{s_i\,;\, i \in \mathbb{I}\}$ having initial distribution given by the probability vector $\mathbf{p} = [p_i]_{i\in\mathbb{I}}$ and transition probabilities given by the stochastic matrix $\mathbf{P} = [P_{ij}]_{i,j\in\mathbb{I}}$ as follows: First we need to generate random variable X with $\mathbb{P}[X = s_i] = p_i$. This can be done by transforming a uniform random variable as follows.

9.7 Algorithm (Generation of Random Variables)
Generate a realization u of a uniformly on $[0,1]$ distributed random variable U. Partition the interval $[0,1]$ into subintervals $[a_i, b_i)$, $i \in \mathbb{I}$, such that $|b_i - a_i| = p_i$. Set $x = s_i$ if $a_i \leq u < b_i$. Then x is a realization of a random variable X having distribution $\mathbf{p}$.

Then one can generate the Markov chain with Algorithm 9.8 below.

9.8 Algorithm (Generation of Markov Chains)
(i) Generate a realization x_0 of X_0 from the distribution $\mathbf{p} = [p_i]_{i\in\mathbb{I}}$.
(ii) If realizations $x_0, x_1, \ldots, x_n$ for $X_0, \ldots, X_n$ are already generated, generate a new realization for X_{n+1} from the distribution $[P_{x_n,j}]_{j\in\mathbb{I}}$.

In practice it is difficult to implement the procedure aboves, if the index set $\mathbb{I}$ or, equivalently, the state-space $\mathbb{S}$ is infinite. In the finite case it is, however, quite easy. Indeed, the Octave functions `rand_pmf` and `rand_mc` introduced in the last section of this lecture generate random variables and Markov chains with finite state-space.

Let us end this section by collecting what we know about Markov chains and matrices.

9.9 Theorem (Markov Chains as Matrices)
The complete probabilistic nature of a Markov chain is determined by its initial distribution $\mathbf{p}$ and transition probability matrix $\mathbf{P}$. Conversely, for any probability vector $\mathbf{p}$ and stochastic matrix $\mathbf{P}$ of comparable dimensions there exists a Markov chain initial distribution $\mathbf{p}$ and transition probability matrix $\mathbf{P}$. This Markov chain can be generated by Algorithm 9.8.

Chapman–Kolmogorov Equations

To answer the question posed in Example 9.1 we need to know **n-step transition probabilities**

$$P_{ij}^n \quad = \quad \mathbb{P}\big[X_n = s_j \,\big|\, X_0 = s_i\big].$$

To find these, the key tool is the **law of total probability** or the **conditioning trick**. The law of the total probability states that for any event A we have

$$\mathbb{P}[A] \quad = \quad \sum_k \mathbb{P}[A, B_k],$$

if B_k's are alternatives, i.e., precisely one of them occurs. The conditioning trick is the law of total probability combined with the **product rule** of conditional probability:

$$\mathbb{P}[A] \quad = \quad \sum_k \mathbb{P}[B_k]\,\mathbb{P}[A\,|\,B_k].$$

Let us start wit the 2-step transitions. Suppose the Markov chain X_n, $n \in \mathbb{N}$, starts at the state s_i at time 0 and ends in the state s_j at time 2. Now, at time 1 it can be in **any** state, but surely it is in **some** state. Moreover, it is in **precisely one state**. Thus the probability of going in two steps from state s_i to state s_j is the sum of probabilities of going from the fixed state s_i to some (i.e. any) state s_k and then from the state s_k to the fixed state s_j. With compact mathematical notation this means that

$$P_{ij}^2 \quad = \quad \sum_k P_{ik} P_{kj}.$$

So, we see the suggestive notation P_{ij}^2 indeed is the **matrix multiplication**, i.e., the 2-step transition probability matrix $\mathbf{P}^2$ is indeed $\mathbf{P} \cdot \mathbf{P}$.

Let us then consider the general case of $n+m$ transitions. Suppose we already know the n-step and m-step transition probabilities P_{ik}^n and P_{kj}^m for **all** states s_k. Now the trick is to **condition** the Markov chain to be in a state s_k after n steps. Then the probability of going from s_i to s_j in $n+m$ steps, conditioned on being at a state s_k after n steps, is obviously $P_{ik}^n P_{kj}^m$. The final step is to **uncondition** the assumption that the chain is in state s_k after n steps. This is done by summing over all the possible states s_k the Markov chain can be in after n steps. This leads to the following equations by Chapman and Kolmogorov.

9.10 Theorem (Chapman–Kolmogorov Equations)
The multi-step transition probabilities of a Markov chain satisfy the **Chapman–Kolmogorov equations**

$$P_{ij}^{n+m} \quad = \quad \sum_k P_{ik}^n P_{kj}^m.$$

In particular, the **n-step transition probability matrix $\mathbf{P}^n$** is the n^{th} matrix power of the 1-step transition probability matrix $\mathbf{P}$.

9.11 Remark
Note that the notation P_{ij}^n has to be understood as $(P^n)_{ij}$, **not** as $(P_{ij})^n$. Indeed, consider the two-state Markov chain with transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{bmatrix}.$$

Then, e.g.,

$$P_{11}^2 \;=\; P_{11}P_{11} + P_{12}P_{21} \;=\; 0.9 \times 0.9 + 0.1 \times 0.4 \;=\; 0.85,$$

while $(P_{11})^2 = 0.9^2 = 0.81$ is the probability that the Markov chain stays in the state 1 for the entire length of 2 steps.

Theorem 9.10 tells us how to calculate the conditional probabilities. Let us then consider the unconditional probabilities. Denote by $\mathbf{p}^n$ by the n-time distribution of a Markov chain with initial distribution $\mathbf{p}$ transition probability matrix $\mathbf{P}$, i.e., $p_i^n = \mathbb{P}[X_n = s_i]$. Then, by the **law of total probability** or by the **conditioning trick**,

$$\begin{aligned} p_j^n &= \mathbb{P}\big[X_n = s_j\big] \\ &= \sum_i \mathbb{P}\big[X_0 = s_i\big]\mathbb{P}\big[X_n = s_j \,\big|\, X_0 = s_i\big] \\ &= \sum_i p_i P_{ij}^n. \end{aligned}$$

So, we have shown the following:

9.12 Theorem (Unconditional Distribution of Markov Chain)
Let X_n, $n \in \mathbb{N}$, be a Markov chain with initial distribution $\mathbf{p}$ and transition probability matrix $\mathbf{P}$. Then the distribution of the Markov chain at time n is

$$\mathbf{p}^n \;=\; \mathbf{p}\,\mathbf{P}^n.$$

Now we are almost ready to solve Example 9.1. We still have an apparently huge problem at hand. Indeed, suppose we want to model the policyholders yearly premium by a stochastic process X_n, $n \in \mathbb{N}$, with state-space $\mathbb{S} = \{0\%, 20\%, 40\%, 60\%\}$. The huge problem is that X_n, $n \in \mathbb{N}$, is not a Markov chain! Indeed, the transition possibilities from state 0% to state 20% depend on how many years the policyholder has spent in the state 0%. So the future is dependent on the (relatively) distant past. Fortunately, this problem can be solved by **enlarging the state-space**. In the solution of Example 9.1 we show how to do this. In general, the method of enlarging the state-space works always, i.e., **every**

stochastic process is Markovian under suitably enlarged state-space. In theory this is good news: Markovian modeling is enough for all purposes. In practice the news are not so good. Enlarging the state-space is often impractical.

9.13 Example (Bonus–Malus Insurance, Solution)
Let X_n, $n \in \mathbb{N}$, denote the state of the policyholder at year n. To make X_n, $n \in \mathbb{N}$, a Markov chain we consider the following enlarged state-space:

1: First year with no claims at premium 0%.
2: Second year with no claims at premium 0%.
3: Third year with no claims at premium 0%.
4: First year with no claims at premium 20%.
5: First year with no claims at premium 40%.
6: Second year with no claims at premium 40%.
7: Premium 60%.

Assume then that the probabilities of 0, 1, 2 or more claims on each year are independent of the previous years' claims. (Indeed, we have no data to assume otherwise!) Let us denote these probabilities by a_0, a_1, a_2 and $a_{>2}$. Then X_n, $n \in \mathbb{N}$, is a Markov chain with initial distribution $\mathbb{P}[X_0 = 1] = 1$ and transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 1-a_0 & a_0 & 0 & 0 & 0 & 0 & 0 \\ 1-a_0 & 0 & a_0 & 0 & 0 & 0 & 0 \\ 1-a_0 & 0 & 0 & a_0 & 0 & 0 & 0 \\ 1-a_0 & 0 & 0 & 0 & a_0 & 0 & 0 \\ a_2+a_{>2} & 0 & 0 & a_1 & 0 & a_0 & 0 \\ a_2+a_{>2} & 0 & 0 & a_1 & 0 & 0 & a_0 \\ a_{>2} & 0 & 0 & a_2 & a_1 & 0 & a_0 \end{bmatrix}$$

Next we have to somehow determine the probabilities a_0, a_1, a_2 and $a_{>2}$. Again, with the almost compete lack of data, we have to assume. We **assume** that the claims occur "completely randomly". This means that the yearly claims are Poisson distributed. Indeed, the Poisson distribution is the "natural" distribution on the set of natural numbers. Since the only data point we have is the average number of claims, 0.3, this leads to probabilities

$$a_k = \mathrm{e}^{-0.3}\frac{0.3^k}{k!}, \qquad k = 0, 1, \ldots.$$

So, in particular,

$$a_0 = 0.741, \quad a_1 = 0.222, \quad a_2 = 0.033 \quad \text{and} \quad a_{>2} = 0.004.$$

Plugging these numbers into the symbolic transition probability matrix we obtain the numeric transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0.259 & 0.741 & 0 & 0 & 0 & 0 & 0 \\ 0.259 & 0 & 0.741 & 0 & 0 & 0 & 0 \\ 0.259 & 0 & 0 & 0.741 & 0 & 0 & 0 \\ 0.259 & 0 & 0 & 0 & 0.741 & 0 & 0 \\ 0.037 & 0 & 0 & 0.222 & 0 & 0.741 & 0 \\ 0.037 & 0 & 0 & 0.222 & 0 & 0 & 0.741 \\ 0.004 & 0 & 0 & 0.033 & 0.222 & 0 & 0.741 \end{bmatrix}$$

Now, the probability that the new policyholder will have 60% discount in 10 years and continue to have it for the successive 5 years is

$$P_{1,7}^{10} \, (P_{7,7})^5 \quad = \quad 5{,}68\%.$$

(This number was calculated with Octave. I strongly recommend **not** to calculate it by hand!)

Simulating Markov Chains

Consider a Markov Chain X_n, $n \in \mathbb{N}$, with a finite state-space $\mathbb{S} = \{s_1, \ldots, s_K\}$.

The following function `rand_pmf` simulates the initial distribution of the Markov chain.

```
function x = rand_pmf(p, s)
## Function  x = rand_pmf(p, s)  returns a random sample  x  of a simple random
## X having probability mass function P[X=s(k)] = p(k).   If s is omitted, then
## it is assumed that p(k) = P[X=k].

    K = length(p);                          ## The size of the state-space.
    U = rand(1);                            ## Uniform [0,1] random sample.
    if nargin == 1                          ## Set state-space if not given.
        s = 1:K;
    endif
    cp = cumsum(p);                         ## Cumulative prob. mass vector.
    for k=1:K                               ## Find sample. Exit when found.
        if U <= cp(k)
            x = s(k);
            return;
        endif
    endfor
    x = s(K);                               ## Paranoia.
endfunction
```

www.uva.fi/~tsottine/psp/rand_pmf.m

The following function `rand_mc` simulates finite state-space Markov chains. It needs the function `rand_pmf` to work.

```
function x = rand_mc(p, P, N, s)
## Function  x = rand_mc(p, P, N, s)  returns a random sample x  of length N of
## a Markov chain with initial distribution p,  transition probability matrix P
## and finite state-space s.  If s is omitted, then it is assumed to be
## {1,2,..,length(p)}.
##
## REQUIRES: rand_pmf.

    K = length(p);                          ## The size of the state-space.
    if nargin == 3                          ## Set state-space, if not given.
        s = 1:K;
    endif
    x = zeros(1,N);                         ## Initialize output vector.
    x(1) = rand_pmf(p,s);                   ## Get starting variable.
    for n=2:N                               ## Loop for remainding variables.
        for j=1:K                           ## Find the index k of s(k).
            if x(n-1) == s(j)
                k = j;
            endif
        endfor
        x(n) = rand_pmf(P(k,:),s);          ## Get transitions.
    endfor
endfunction
```

www.uva.fi/~tsottine/psp/rand_mc.m

The following script `exa_rand_mc` illustrates the simulation function `rand_mc`. The Markov chain simulated is a symmetric **random walk** with reflecting boundaries.

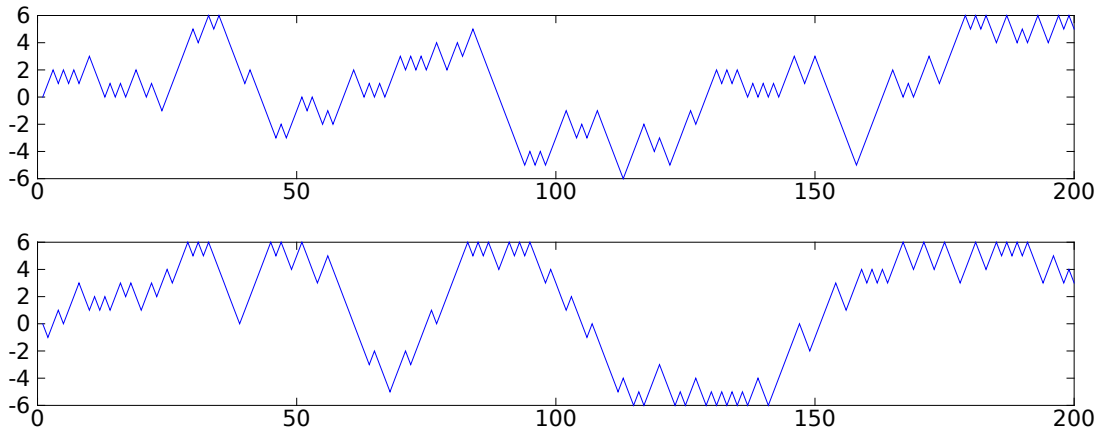
```
##############################################################################
## FILE: exa_simu_mc.m
##
## An illustration of how the function rand_mc works.
##
## REQUIRES: rand_pmf, rand_mc.
##############################################################################

rand("state", 210873);                          ## Fix randomness.
M = 2;                                          ## Number of samples.
N = 200;                                        ## Lenght of each sample.
s = [-6 -5 -4 -3 -2 -1  0  1  2  3  4  5  6];   ## State-space.
p = [ 0  0  0  0  0  0  1  0  0  0  0  0  0];   ## MC starts at 0.

## Transition probabilites: a random walk with reflecting boundaries.
P = [ 0    1    0    0    0    0    0    0    0    0    0    0    0 ;  ## -6
      0.5  0    0.5  0    0    0    0    0    0    0    0    0    0 ;  ## -5
      0    0.5  0    0.5  0    0    0    0    0    0    0    0    0 ;  ## -4
      0    0    0.5  0    0.5  0    0    0    0    0    0    0    0 ;  ## -3
      0    0    0    0.5  0    0.5  0    0    0    0    0    0    0 ;  ## -2
      0    0    0    0    0.5  0    0.5  0    0    0    0    0    0 ;  ## -1
      0    0    0    0    0    0.5  0    0.5  0    0    0    0    0 ;  ##  0
      0    0    0    0    0    0    0.5  0    0.5  0    0    0    0 ;  ##  1
      0    0    0    0    0    0    0    0.5  0    0.5  0    0    0 ;  ##  2
      0    0    0    0    0    0    0    0    0.5  0    0.5  0    0 ;  ##  3
      0    0    0    0    0    0    0    0    0    0.5  0    0.5  0 ;  ##  4
```

```
27        0     0     0     0     0     0     0     0     0     0     0.5   0     0.5;  ##  5
28        0     0     0     0     0     0     0     0     0     0     0     1     0   ]; ##  6
29 ##    -6    -5    -4    -3    -2    -1     0     1     2     3     4     5     6
30
31 x = zeros(M,N);                                    ## Initialize the output.
32 for m=1:M                                          ## M samples of lenght N each.
33     x(m,:) = rand_mc(p,P,N,s);
34 endfor
35
36 for m=1:M                                          ## Plot the samples vertically.
37     subplot(M,1,m)
38     plot(x(m,:));
39 endfor
```

www.uva.fi/~tsottine/psp/exa_rand_mc.m



Realizations of a random walk with reflecting boundary.

Exercises

9.1 Exercise

Which of the following are stochastic matrices?

(a)
$$\begin{bmatrix} 0.2 & 0.6 \\ 0.1 & 0.9 \end{bmatrix}$$

(b)
$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

(c)
$$\begin{bmatrix} 0.9 & 0.4 & -0.3 \\ 0.9 & 0.1 & 0 \\ 0 & 0.2 & 0.8 \end{bmatrix}$$

(d)
$$[1]$$

For the stochastic matrices calculate the corresponding 7-step transition probability matrices.

9.2 Exercise
Consider the Markov Chain X_n, n$\in\mathbb{N}$, with state-space $\mathbb{S}=\{1,2,3\}$ and transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

and initial probability vector (for X_0)

$$\mathbf{p} = [0.7\ 0\ 0.3].$$

Calculate the probabilities

(a) $\mathbb{P}[X_1 = 1]$
(b) $\mathbb{P}[X_3 = 1]$
(c) $\mathbb{P}[X_6 = 2]$
(d) $\mathbb{P}[X_3 = X_2 = 3]$

9.3 Exercise (Random Walk)
Let X_n, $n \in \mathbb{N}$, be the symmetric random walk, i.e., the Markov chain with transition probabilities

$$P_{i,i+1} = \frac{1}{2} = P_{i,i-1}.$$

Assume $X_0 = 1$. Calculate the probabilities

(a) $\mathbb{P}[X_4 = 2]$
(b) $\mathbb{P}[X_6 = -2]$
(c) $\mathbb{P}[X_8 = 0]$
(d) $\mathbb{P}[X_{1973} = 0]$

9.4 Exercise
Consider Example 9.1.

(a) How fast can a new policyholder raise to the level of 60% discount and what is the probability that this happens?
(b) What is the probability that a new policyholder remains in the level of 0% discount for 10 years?
(c) What is the probability that a new policyholder will not reach the level of 60% discount on any year in 25 years?

(d) What is the distribution of discounts for policyholders that have had the policy for 20 years?

Hint for (b) and (b): Let X_n, $n \in \mathbb{N}$, be a Markov chain with state-space $\mathbb{S}$ and transition probability matrix $\mathbf{P}$. Let $\mathscr{S} \subset \mathbb{S}$. Consider a new **killed Markov chain** Y_n, $n \in \mathbb{N}$, that is a constructed from the Markov chain X_n, $n \in \mathbb{N}$, by **killing** it to the set of states $\mathscr{S}$ in the following way: The state-space of Y_n, $n \in \mathbb{N}$, is $\{s_i, \dagger; s_i \in \mathbb{S}\}$ and the transition probability matrix of Y_n, $\in \mathbb{N}$, is

$$\begin{aligned} Q_{ij} &= P_{ij} && \text{if} \quad s_i, s_j \notin \mathscr{S}, \\ Q_{i,\dagger} &= \sum_{j : s_j \in \mathscr{S}} P_{ij} && \text{if} \quad s_j \notin \mathscr{S}, \\ Q_{\dagger,\dagger} &= 1. \end{aligned}$$

Then Y_n will never leave the state S corresponding to the set of states $\mathscr{S}$, if it enters it. Consequently, the probability that the Markov chain will enter the set of states $\mathscr{S}$ any time before (or including) time n is $Q^n_{i,\dagger}$, where s_i is the initial state of the Markov chain X_n, $n \in \mathbb{N}$.

9.5 Exercise
Consider Example 9.1. Suppose the insurance company has the following yearly data on the number of claims made by the policyholders:

Claims	Policyholders	Claims	Policyholders
0	4 905	4	7
1	1 120	5	0
2	114	6	0
3	0	7	1

Answer to the question of Example 9.1 with this additional information

(a) as a believer of the Poisson type independent claim assumption,
(b) by believing that the data speaks for itself.

9.6 Exercise
Consider a time-homogeneous stochastic process X_n, $n \in \mathbb{N}$, with state-space $\mathbb{S} = \{1, 2\}$

and the following non-Markovian 2-step dependence structure

$$\begin{aligned}
\mathbb{P}\big[X_n = 1 \,\big|\, X_{n-1} = 1, X_{n-2} = 1\big] &= 0.91,\\
\mathbb{P}\big[X_n = 1 \,\big|\, X_{n-1} = 1, X_{n-2} = 2\big] &= 0.07,\\
\mathbb{P}\big[X_n = 1 \,\big|\, X_{n-1} = 2, X_{n-2} = 1\big] &= 0.02,\\
\mathbb{P}\big[X_n = 1 \,\big|\, X_{n-1} = 2, X_{n-2} = 2\big] &= 0.00,
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}\big[X_n = 2 \,\big|\, X_{n-1} = 1, X_{n-2} = 1\big] &= 0.09,\\
\mathbb{P}\big[X_n = 2 \,\big|\, X_{n-1} = 1, X_{n-2} = 2\big] &= 0.93,\\
\mathbb{P}\big[X_n = 2 \,\big|\, X_{n-1} = 2, X_{n-2} = 1\big] &= 0.98,\\
\mathbb{P}\big[X_n = 2 \,\big|\, X_{n-1} = 2, X_{n-2} = 2\big] &= 1.00.
\end{aligned}$$

Model X_n, $n \in \mathbb{N}$, as a Markov chain by enlarging its state-space and calculate the probability

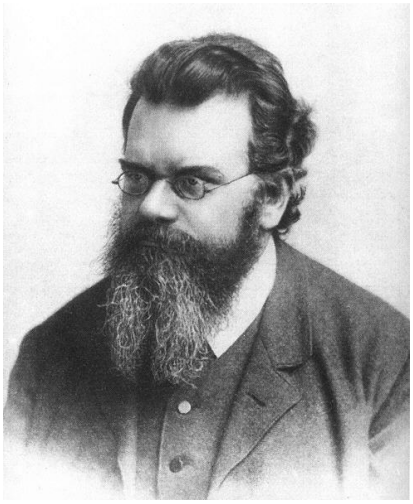
$$\mathbb{P}\big[X_7 = X_6 = 1 \,\big|\, X_3 = 1\big].$$

Lecture 10

Classification of Markovian States

This Lecture is preparation for the next Lecture 11 that deals with the long-term behavior of Markov chains. The key result, or rather a concept, in the long-term behavior is *ergodicity* which is a generalization of the *law of large numbers*. This generalization was a major motivation for **Andrey Adreyevich Markov** himself for winning his theological argument concerning the free will against **Pavel Nekrasov** and the Moscow School of Mathematics.

On a more practical level the concept of ergodicity was first introduced by the Austrian physicist and philosopher **Ludwig Bolzmann** (1844–1906) in the context of statistical mechanics. Indeed, it was he who coined the term *ergodic* from the Greek words of *ergon* (work) and *odos* (path). In the context of statistical mechanics or thermodynamics the ergodicity, or the *ergodic hypothesis* means that over long periods of time, the time spent by a system in some region of the phase-space with the same energy is proportional to the volume of this region. In the language of stochastic processes this means that the time-averages and the probability averages are the same, as for the probabilists the phase-space is the probability space.



Ludwig Boltzmann (1844–1906)

We analyze here the states of a Markov chain and present concepts that are needed to give conditions under which the long-time behavior of a Markov chain is "nice". The key Example 10.1 below is chosen so that it should be as "un-nice" as reasonable. In some sense its long-time behavior is reasonable but it does not fit into the nice ergodic theory.

10.1 Example (Confused Ant)
An ant is dropped in an infinitely long corridor where the ant can only take steps left or right. The corridor has one wall in the middle (whatever "middle" means in an infinitely long corridor). The wall is sticky on the left side: if the ant hits the wall on the left, it will get stuck. On the right side the wall makes the ant bounce off. Naturally, the ant gets confused and starts to take steps left and right completely randomly. What will happen to the ant eventually?

Communication Classes

Two given states of a Markov chain communicate, if starting from any one of them one can eventually reach the other, and vice versa.

10.2 Definition (Accessibility and Communication)
Let X_n, $n \in \mathbb{N}$, be a Markov chain with (discrete) state-space $\mathbb{S} = \{s_k; k \in \mathbb{I}\}$ and transition probability matrix $\mathbf{P} = [P_{ij}]_{i,j\in\mathbb{I}}$. A state s_j is **accessible** from the state s_i if $P_{ij}^n > 0$ for some $n \in \mathbb{N}$. In this case we denote $i \to j$. If both $i \to j$ and $j \to i$ hold, we say that the states s_i and s_j **communicate** and denote $i \leftrightarrow j$.

10.3 Remark (Equivalence Relation)
The communication relation $\leftrightarrow$ is, as the symbol suggests, an **equivalence relation**, i.e., it satisfies

(i) $i \leftrightarrow i$ (reflexivity),
(ii) if $i \leftrightarrow j$, then $j \leftrightarrow i$ (symmetry),
(iii) if $i \leftrightarrow k$ and $k \leftrightarrow j$, then $i \leftrightarrow j$ (transitivity).

Indeed, reflexivity and symmetry are obvious. To see transitivity, suppose that $P_{ik}^n > 0$ and $P_{kj}^m > 0$. Then, $P_{ij}^{n+m} \geq P_{ik}^n P_{kj}^m > 0$, and similarly for $P_{ji}^{n+m} > 0$ (with some different n and m). Since any equivalence relation on any set will split the set into **equivalence classes** we shall speak of **communication classes**, or **classes** for short.

As a first example, consider a Markov chain with transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0.2 & 0.8 & 0 \\ 0.7 & 0.3 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

It is obvious that this chain has two communication classes: $\{s_1, s_2\}$ and $\{s_3\}$. Indeed, if the Markov enters the states s_1 or s_2, it will never reach the state s_3: $\{s_1, s_2\}$ is an **absorbing class**. In the same way the class $\{s_3\}$, or the state s_3 is absorbing: if the Markov chain ever enters the **absorbing state** s_3 it will never leave it. In this example the classes $\{s_1, s_2\}$ and $\{s_3\}$ are actually **isolated**, i.e., the Markov chain starting from the one class will never reach the other class, and vice versa.

As a second example, consider a Markov chain with transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0.2 & 0.7 & 0.1 \\ 0.3 & 0.3 & 0.4 \\ 0 & 0 & 1 \end{bmatrix}.$$

Again, this chain has two communication classes: $\{s_1, s_2\}$ and $\{s_3\}$. Indeed, s_1 and s_2 communicate: one can reach state s_1 from state s_2 with a single step, and vice versa. One can also reach the state s_3 from both of the states s_1 and s_2. So, the classes $\{s_1, s_2\}$ and $\{s_3\}$ are not **isolated**. But the reaching goes only one way: if the Markov chain ever enters the **absorbing state** s_3 it will never leave it.

As a third example, consider a Markov chain with the transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0.7 & 0.1 & 0.2 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

This Markov chain has three communication classes $\{s_1\}$, $\{s_2, s_3, s_4\}$ and $\{s_5\}$. Indeed, the state 1 communicates only with itself, since if you leave it, you never come back. In a quite opposite manner the state s_5 only communicates with itself, since it is **absorbing**: once you enter, you never leave. States s_2, s_3 and s_4 form a communication class $\{s_2, s_3, s_4\}$. To see that s_2, s_3 and s_4 communicate, it is enough to find a possible "round-trip" around the states. One possible round-trip is $s_2 \to s_4 \to s_2 \to s_3 \to s_4 \to s_2$.

For the log-term behavior of Markov chains many communications classes are a nuisance. Therefore, there is a name for Markov chains with only one communication class.

10.4 Definition (Irreducibility)
If a Markov chain has only one communication class it is called **irreducible**.

To see how irritating many communications classes can be in the long time behavior, consider the Markov chain with state space $\mathbb{S} = \{s_1, s_2, s_3\}$ and transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0.2 & 0.8 & 0 \\ 0.7 & 0.3 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Suppose that we are interested in the **limiting probabilites**

$$\tilde{\pi}_j = \lim_{n \to \infty} P_{ij}^n.$$

To understand what is goin on, let us calculate the $n = 10\,000$ step transition probability matrix:

$$\mathbf{P}^{10\,000} = \begin{bmatrix} 0.46667 & 0.53333 & 0.00000 \\ 0.46667 & 0.53333 & 0.00000 \\ 0.0000 & 0.00000 & 1.00000 \end{bmatrix}.$$

So, if $i \in \{1,2\}$, then we have

$$\begin{aligned} \tilde{\pi}_1 &= 0.46667, \\ \tilde{\pi}_2 &= 0.53333, \\ \tilde{\pi}_3 &= 0.00000, \end{aligned}$$

while for $i = 3$, we have

$$\begin{aligned} \tilde{\pi}_1 &= 0.00000, \\ \tilde{\pi}_2 &= 0.00000, \\ \tilde{\pi}_3 &= 1.00000. \end{aligned}$$

Consequently, the limiting distribution $\tilde{\boldsymbol{\pi}}$ depends depends on the initial state of the process, which is not desirable.

In general, it may be difficult to determine the communication classes of a Markov chain by following a **finite** algorithm. In principle, one could just calculate P_{ij}^n and P_{ji}^n for different values of n to check if s_i and s_j communicate. Unfortunately, this can only give a positive answer if we find that $P_{ij}^n > 0$ and $P_{ji}^n > 0$ for some (different) n along the way. But negative answer is not possible, unless one checks all the infinite possibilities for all $n \in \mathbb{N}$. This **infinite** algorithm is, at least for the current computers, impossible. In the case of finite state-space Markov chains the communications can be checked, however. The key observation is that for a finite-state Markov chain with K states it is enough to check the n-step transitions upto $\mathbf{P}^n$ for $n \leq K-1$. Indeed, if one cannot find a path from any state to any other state in a collection of K states with at most $K-1$ steps, then there is no such path! This is a simple **counting argument**.

The Octave function `comm_mc` listed below checks whether two states s_i and s_j of a finite state-space Markov chain, determined by its transition probability matrix $\mathbf{P}$, communicate.

```
function bool = comm_mc(P, i, j)
## Function bool = comm_mc(P, i, j) returns 0 if the states i and j of a Markov
## chain with transition probability matrix P communicate, and 0 otherwise.

    if acc_mc(P,i,j) && acc_mc(P,j,i)      ## Use the auxiliary function below.
        bool = 1;
    else
        bool = 0;
    endif
endfunction

## Auxiliary function that determines if j is accessible from i by checking all
## the possible paths of lenght 0,1,...,rows(N)-1.

function bool = acc_mc(P, i, j)
    bool = 0;                              ## Assume no access.
    K = rows(P);                           ## Number of states.
    for n=0:(K-1)                          ## Check the access paths upto K-1.
        if (P^n)(i,j) > 0
            bool = 1;                      ## Access found.  Exit the function.
            return;
```

```
22          endif
23      endfor
24  endfunction
```

www.uva.fi/~tsottine/psp/comm_mc.m

Transience and Recurrence, and Positive Recurrence

Let T_{ij} be the **random time** it takes for a Markov chain to reach the state s_j when it starts from the state s_i. We denote by $T_i = T_{ii}$, for the **return time** of the state s_i. We also denote the **mean return time** by

$$m_i \quad = \quad \mathbb{E}[T_i].$$

10.5 Definition (Transience, Null Recurrence and Positive Recurrence)
Let X_n, $n \in \mathbb{N}$ be a Markov chain with state space $\mathbb{S}$. A state $s_i \in \mathbb{S}$ is **transient** if

$$\mathbb{P}[T_i = \infty] \quad > \quad 0.$$

A state $s_i \in \mathbb{S}$ of is **recurrent** if

$$\mathbb{P}[T_i < \infty] \quad = \quad 1.$$

A state $s_i \in \mathbb{S}$ is **positive recurrent** it is recurrent and

$$m_i \quad < \quad \infty.$$

A state $s_i \in \mathbb{S}$ that is recurrent but not positive recurrent is **null recurrent**.

So, a transient state is a state that occurs only finitely many times and a recurrent state is a state that occurs infinitely many times, and a positive recurrent state occurs frequently enough to admit finite average. For the long-run behavior, the positive recurrence is desirable. Indeed, the long-time proportions $\bar{\pi}_i$ for a state s_i satisfy

$$\bar{\pi}_i \quad = \quad \frac{1}{m_i}.$$

Thus, $\bar{\pi}_i = 0$ for states that are either transient or null recurrent.

From Definition 10.5 it may seem that all states are either recurrent or transient. This is not true for general stochastic processes. Indeed, it may happen that a state will occur infinitely or finitely many times depending on the particular realization of the process. In the case of time-homogeneous Markov chains this is not possible, however. Indeed, let f_i denote the probability that a time-homogeneous Markov chain, starting from the state s_i (or having entered the state s_i) will return to the state s_i. Then s_i is transient if $f_i < 1$

and recurrent if $f_i = 1$. Indeed, suppose $f_i < 1$. Then for each time the Markov chain enters the state s_i there is the positive probability $1-f_i$ that the process will never re-enter the state s_i. Consequently, starting with the state s_i, the probability that the process will re-visit the state exactly $n-1$ times is $f_i^{n-1}(1-f_i)$. This means that the number of times a transient state s_i is re-visited has **geometric distribution** with mean $1/(1-f_i)$. Since this is true for **all** of the possible realizations, it follows that if $f_i < 1$, the state s_i is transient. If, however, $f_i = 1$, then the Markov chain will re-enter the state s_i infinitely often. Indeed, each time the Markov chain hits the state s_i it will **regenerate**: the chain X_n, $n \geq T_i$, and the original chain X_n, $n \geq 0$, are probabilistically the same (assuming that $X_0 = s_i$).

Let us then look for quantitative criteria for transience and recurrence. Let

$$I_n(i) \;=\; \begin{cases} 1, & \text{if } X_n = s_i \\ 0, & \text{if } X_n \neq s_i \end{cases}$$

i.e., $I_n(i)$ is the **indicator** of the event $\{X_n = s_i\}$. Then $\sum_{n=0}^{\infty} I_n(i)$ is the total number of times the Markov chain visits the state s_i. Now,

$$\begin{aligned} \mathbb{E}\left[\sum_{n=0}^{\infty} I_n(i) \,\middle|\, X_0 = s_i\right] &= \sum_{n=0}^{\infty} \mathbb{E}\big[I_n(i) \,\big|\, X_0 = s_i\big] \\ &= \sum_{n=0}^{\infty} \mathbb{P}\big[X_n = s_i \,\big|\, X_0 = s_i\big] \\ &= \sum_{n=0}^{\infty} P_{ii}^n. \end{aligned}$$

This sum is finite for transient states and infinite for recurrent states. Let us then consider the positive recurrent states. We see, just as above, that the average number of visits to the starting state s_i that occur before the time N is $\sum_{n=0}^{N} P_{ii}^n$. Consequently, the average number of times the state s_i is visited **per unit time** is given by the **Cesàro mean**

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} P_{ii}^n.$$

This is also the long-time proportion $\bar{\pi}_i$ the Markov chain spends at the state s_i, and $\bar{\pi}_i = 1/m_i$. Consequently, the Cesàro limit above must be strictly positive for positive recurrent states. We have argued the following definite criteria for the transience and (null or positive) recurrence in terms of the transition probability matrix $\mathbf{P}$.

10.6 Theorem (Criteria for Transience and Recurrence)
A state s_i of a Markov chain with transition probability matrix $\mathbf{P}$ is

$$\begin{aligned}
\text{transient} \quad &\text{if and only if} \quad \sum_{n=0}^{\infty} P_{ii}^n < \infty \\
\text{null recurrent} \quad &\text{if and only if} \quad \sum_{n=0}^{\infty} P_{ii}^n = \infty \quad \text{and} \quad \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} P_{ii}^n = 0 \\
\text{positive recurrent} \quad &\text{if and only if} \quad \sum_{n=0}^{\infty} P_{ii}^n = \infty \quad \text{and} \quad \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} P_{ii}^n > 0
\end{aligned}$$

Moreover, the **mean recurrence time** m_i of any state $s_i \in \mathbb{S}$ satisfies the Cesàro limit

$$\frac{1}{m_i} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} P_{ii}^n.$$

Theorem 10.6 tells us that what determines the transience, null recurrence and positive recurrence is how fast the n-step probabilities P_{ii}^n, $n \in \mathbb{N}$, decrease. The slower they decrease the more there is recurrent behavior.

10.7 Corollary (Transience and Recurrence Are Class Properties)
If s_i is transient and s_i communicates with s_j, then s_j is also transient. Ditto for null recurrence and positive recurrence.

Let us argue why the Corollary 10.7 above is true. Suppose that s_i is recurrent and communicates with s_j. Then, for some n and m we have $P_{ij}^n > 0$ and $P_{ji}^m > 0$. Obviously, for **any** N we have $P_{jj}^{n+m+N} \geq P_{ji}^n P_{ii}^N P_{ij}^m$. Consequently,

$$\begin{aligned}
\sum_{N=0}^{\infty} P_{jj}^{N+n+m} &\geq \sum_{N=0}^{\infty} P_{ji}^n P_{ii}^N P_{ij}^m \\
&= P_{ji}^n P_{ij}^m \sum_{N=0}^{\infty} P_{ii}^N \\
&= \infty,
\end{aligned}$$

which shows the s_j is recurrent. The transient case is now obvious, since a state cannot be both transient and recurrent, and it must be one or the other. Let us consider then the most difficult case: positive recurrence. Suppose that the states s_i and s_j communicate and that s_i is positive recurrent. Let m_i be the mean recurrence time of s_i and let m_j be the mean

recurrence time of s_j. By assumption, $m_i < \infty$. We need to show that also $m_j < \infty$. Let m_{ji} be the average time it takes for the Markov chain to reach the state s_i starting from the state s_j. Now, let n be the shortest number of steps so that s_j can be reached from s_i by n steps. We use a **conditioning trick** now. Let A be the event that, starting from the state s_i the Markov chain reaches the state s_j exactly with n steps without visiting the state s_i in doing so, i.e.,

$$A = \{X_1 \neq s_i, X_2 \neq s_i, \ldots X_{n-1} \neq s_i, X_n = s_j\}.$$

Then by conditioning on A we see that

$$\begin{aligned} m_i &= \mathbb{E}[T_i] \\ &= \mathbb{E}[T_i|A]\mathbb{P}[A] + \mathbb{E}[T_i|A^c]\mathbb{P}[A^c] \\ &\geq \mathbb{E}[T_i|A]\mathbb{P}[A] \\ &= (n + m_{ji})\mathbb{P}[A]. \end{aligned}$$

This shows that $m_{ji} < \infty$. Let us then show that $m_{ij} < \infty$. Let $X_0 = s_i$ and denote by $T_i(r)$ be the length of the r^{th} **excursion** of the Markov chain around the state s_i, i.e., the n^{th} revisit of the Markov chain to the state s_i is at the random time $T_i(1)+T_i(2)+\cdots+T_i(n)$. Note that $T_i(r)$, $r \in \mathbb{N}$, are independent and they have the same distribution as T_i (which is the first excursion). Let N_{ji} denote the number of revisits the Markov chain takes to the state s_j before it visits the state s_i. This is geometrically distributed random variable with some parameter p that we are not interested in. The main point is that $\mathbb{E}[N_{ji}] < \infty$ and that

$$T_{ij} \leq \sum_{r=1}^{N_{ji}} T_i(r).$$

Consequently,

$$\begin{aligned} m_{ij} &= \mathbb{E}\left[\sum_{r=1}^{N_{ji}} T_i(r)\right] \\ &= \mathbb{E}\big[N_{ij}\big]\mathbb{E}\big[T_i(r)\big] \\ &= \mathbb{E}\big[N_{ij}\big]m_i \\ &< \infty. \end{aligned}$$

The claim follows now by noting the following obvious **subadditivity**:

$$m_j \leq m_{ji} + m_{ij}.$$

Finally, let us consider transience, null recurrence and positive recurrence in **finite state-space** Markov chains. First note that by a simple **counting argument** we see that a finite state-space Markov chain must have at least one recurrent class. Indeed, suppose all classes are transient. This means that the Markov chain visits each of its states only

finitely many times. But since there are finitely many states, the Markov chain will visit any and all of its states only finitely many times. Since there are infinitely many time points, this is obviously absurd! So, there must be at least one recurrent state. Also with finite state-spaces the concept of positive recurrence is irrelevant: null recurrence cannot occur. Indeed, for null recurrent states s_j we have for all starting states s_i that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} P_{ij}^n \quad = \quad 0.$$

Then, by summing over all the finite number of states we obtain from this that

$$\lim_{n\to\infty} \frac{1}{N} \sum_{j} \sum_{n=1}^{N} P_{ij}^n \quad = \quad 0.$$

But this implies the following absurdity:

$$\begin{aligned} 0 \quad &= \quad \lim_{N\to\infty} \frac{1}{N} \sum_{j} \sum_{n=1}^{N} P_{ij}^n \\ &= \quad \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \sum_{j} P_{ij}^n \\ &= \quad 1, \end{aligned}$$

since $\sum_j P_{ij}^n = 1$. Thus the recurrent state s_i cannot be null recurrent.

For infinite state-space, null recurrence and all transience can occur. Indeed, for the **non-symmetric random walk** $P_{i,i+1} = p = 1 - P_{i,i-1}$, $p \neq 1/2$, all states $i \in \mathbb{S} = \mathbb{Z}$ are transient and for the **symmetric random walk** $P_{i,i+1} = 1/2 = P_{i,i-1}$ all states $i \in \mathbb{S} = \mathbb{Z}$ are null recurrent. (Exercise 10.6)

Periods

Finally, let us end our classification of states by discussing periods. Periods are messy! They can make the long-term behavior of a Markov chain unclear in the sense that limiting probabilities and long-term probabilities differ. This is why we would like our Markov chains to be aperiodic. Of course, this is not always possible.

10.8 Definition (Periods)
A state $s_i \in \mathbb{S}$ of a Markov chain X_n, $n \in \mathbb{N}$, with transition probability matrix $\mathbf{P}$ has **period** d_i if $P_{ii}^n = 0$ whenever n is not divisible by d_i and d_i is the largest number with this property. If $d_i = 1$, then the state s_i is **aperiodic**. If all the states of the Markov chains are aperiodic, we say that the Markov chain is aperiodic.

Definition 10.8 is maybe difficult to understand, but in practice checking the periods is often relatively easy. For example, for the Markov chain with transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

it is relatively easy to see that $d_1 = d_2 = d_3 = 3$ as the Markov chain will alternate around the three states like $s_1 \to s_3 \to s_2 \to s_1 \to s_3 \to s_2 \to s_1 \to \cdots$. It is also clear in this case, by symmetry, that each of the state is equally probable in the long run no matter what the initial distribution is, that is the time the Markov chain spends in the long run in any of the states s_1, s_2 or s_3 is $1/3$. However, the limiting probabilities $\lim_{n\to\infty} P_{ij}^n$ do not exist.

Periodicity is also a class property. Indeed, suppose n and m are such that $P_{ij}^n > 0$ and $P_{ji}^m > 0$. Let k be such that $P_{jj}^k > 0$. Let d_i and d_j be the periods of the communicating states s_i and s_j, respectively. Then, $P_{ii}^{n+m} \geq P_{ij}^n P_{ji}^m > 0$ and $P_{ii}^{n+m+k} \geq P_{ij}^n P_{jj}^k P_{ji}^m > 0$. But this means that d_i must divide $n+m$ and $n+m+k$. Consequently, d_i must also divide their difference k for any such k that $P_{jj}^k > 0$. This means that d_i divides d_j. Changing the roles of i and j we see that also d_j divides d_i. Therefore, $d_i = d_j$, and we have shown the following.

10.9 Proposition (Period is Communication Class Property)
If s_i has period d and if s_i communicates with s_j, then also s_j has period d.

Ergodicity

Let us give the name **ergodic** to the Markov chains that behave nicely in the long run. It should be noted that there is a whole theory in mathematics and physics, called the **ergodic theory** that deals with the question on when **time averages** and **state-space averages** are the same. For unfortunate historical reasons, the definition below for a Markov chain to be ergodic is unnecessarily strict for for the time and state-space averages to coincide.

10.10 Definition (Ergodic Markov Chain)
An irreducible aperiodic positive recurrent Markov chain is called **ergodic**.

After all this analysis of different states, let us analyze the confused ant of Example 10.1.

10.11 Example (Confused Ant, Solution)
Let us call the position of the wall 0. Steps left then lead to states $-1,-2,\ldots$ and steps right lead to states $1,2,\ldots$. Since the wall is different from the left and from the right we split the position 0 into two states $0-$ and $0+$, denoting the left and right side of the wall. The transition probabilities for the positions X_n of the confused ant at time n are partly given by the random walk:

$$P_{i,i-1} \;=\; \frac{1}{2} \;=\; P_{i,i+1}$$

if $i-1,i,i+1 \notin \{0-,0+\}$. For the left wall $0-$ we have $P_{i,0-}=0$, if $i\neq -1$, $P_{-1,0-}=1/2$ and $P_{0-,0-}=1$. For the right wall we have $P_{i,0+}=0$, if $i\neq 1$, $P_{1,0+}=1/2$ and $P_{0+,1}=1$.

There are three communication classes: $\{\ldots,-2,-1\}$, $\{0-\}$ and $\{0+,1,2\ldots\}$.

The class $\{\ldots,-2,-1\}$ is transient: if the ant is dropped in this class it will eventually get stuck to the absorbing state $0-$.

The class $\{0-\}$, being absorbing, is positive recurrent.

Finally, the class $\{0+,1,2,\ldots\}$ is null recurrent. It is enough to consider any state in the class. Let us choose $0+$. Now, by bouncing, $P^{2n+1}_{0+,0+}=0$ for all $n\geq 0$. So, we consider the even steps. For them, due to **Stirling's formula**,

$$P^{2n}_{0+,0+} \;\approx\; \binom{2n}{n}2^{-2n} \;\approx\; \frac{1}{\sqrt{\pi n}}.$$

Therefore,

$$\sum_{n=0}^{\infty} P^n_{0+,0+} \;\approx\; \sum_{n=1}^{\infty}\frac{1}{\sqrt{\pi n}} \;=\; \infty$$

and, consequently, the class $\{0+,1,2,\ldots\}$ is recurrent. Also,

$$\frac{1}{N}\sum_{n=0}^{N} P^n_{0+,0+} \;\approx\; \frac{1}{N}\sum_{n=1}^{N}\frac{1}{\sqrt{\pi n}} \;\approx\; 0.$$

Consequently, the class $\{0+,1,2,\ldots\}$ is null recurrent.

Exercises

10.1 Exercise
Find out the communication classes of the following Markov chains.

(a)
$$\begin{bmatrix} 0.1 & 0.9 & 0 \\ 0 & 0.2 & 0.8 \\ 0 & 0 & 1 \end{bmatrix}$$

(b)
$$\begin{bmatrix} 0.1 & 0.7 & 0.1 & 0.1 \\ 0.2 & 0.2 & 0.2 & 0.4 \\ 0.1 & 0.1 & 0.1 & 0.7 \\ 0.3 & 0.3 & 0.1 & 0.3 \end{bmatrix}$$

(c)
$$\begin{bmatrix} 0.1 & 0.9 & 0 \\ 0 & 0.2 & 0.8 \\ 1 & 0 & 0 \end{bmatrix}$$

(d)
$$\begin{bmatrix} 0.1 & 0 & 0.9 & 0 \\ 0 & 1 & 0 & 0 \\ 0.8 & 0 & 0.2 & 0 \\ 0 & 0.1 & 0 & 0.9 \end{bmatrix}$$

10.2 Exercise
Find out which states of the following Markov chains are transient and which are recurrent (null or positive).

(a)
$$\begin{bmatrix} 0.1 & 0.9 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

(b)
$$\begin{bmatrix} 0.1 & 0.7 & 0.1 & 0.1 \\ 0.2 & 0.2 & 0.2 & 0.4 \\ 0 & 0 & 1 & 0 \\ 0.3 & 0.3 & 0.1 & 0.3 \end{bmatrix}$$

(c)
$$\begin{bmatrix} 0.1 & 0 & 0.9 \\ 0 & 0.2 & 0.8 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

(d)
$$\begin{bmatrix} 0.1 & 0 & 0.9 & 0 \\ 0.8 & 0 & 0.2 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0.9 \end{bmatrix}$$

10.3 Exercise
Find out which states of the following Markov chains are transient and which are recurrent (null or positive). Also, find out the period of each state.

(a)
$$\begin{bmatrix} 0.1 & 0.9 & 0 \\ 0 & 0.2 & 0.8 \\ 1 & 0 & 0 \end{bmatrix}$$

(b)
$$\begin{bmatrix} 0.1 & 0.7 & 0.1 & 0.1 \\ 0 & 0 & 1 & 0 \\ 0 & 0.6 & 0 & 0.4 \\ 0.3 & 0.3 & 0.1 & 0.3 \end{bmatrix}$$

(c)
$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0.2 & 0.8 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

(d)
$$\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0.8 & 0 & 0.2 & 0 \\ 0 & 0.3 & 0 & 0.7 \\ 0 & 0 & 0.1 & 0.9 \end{bmatrix}$$

10.4 Exercise (Random Walk)
Consider the random walk with absorbing boundary at $i = -10$ and a reflecting boundary at $i = 10$. This means that $P_{i,i+1} = 1/2 = P_{i-1,i}$ except that $P_{-10,-10} = 1$ and $P_{10,9} = 1$. Classify each state of the process on whether it is transient, null recurrent, positive recurrent, absorbing and what is its period.

10.5 Exercise (Martingale Strategy)
John Player has found a special and clever way of playing roulette. He will bet 100 euros on red. If the outcome is red, then he has won 100 euros. If the outcome is black, then having lost the 100 euros, he will continue and bet 200 euros on red. If the outcome is now red, he has won 100. If the outcome is again black, then he will bet 400 euros on red, and so on. This strategy of always doubling your bet until you win is called a **martingale strategy**. With this strategy the player is sure to win his initial bet eventually. What is wrong with this strategy, i.e., why are the casinos not worried about martingale-players?

10.6 Exercise (Symmetry and Transience of Random Walks)
Consider the **random walk** X_n, $n \in \mathbb{N}$, where $X = 0$ and $P_{i,i+1} = p = 1 - P_{i,i-1}$, $i \in \mathbb{S} = \mathbb{Z}$. Show that

(a) if the random walk is **symmetric**, i.e., $p = 1/2$, then every state $i \in \mathbb{Z}$ is null recurrent,
(b) if the random walk is **non-symmetric**, i.e., $p \neq 1/2$, then every state $i \in \mathbb{Z}$ is transient.

Lecture 11

Markovian Long Run and Ergodicity

A stochastic process is said to be *ergodic* if its statistical properties can be deduced from a single, sufficiently long, random sample of the process. The reasoning is that any collection of random samples from a process must represent the average statistical properties of the entire process. In other words, regardless of what the individual samples are, a birds-eye view of the collection of samples must represent the whole process. Conversely, a process that is not ergodic is a process that changes erratically at an inconsistent rate. There is a whole separate field of science called the ergodic theory that deals with dynamical systems and their ergodicity. Arguable the most celebrated result in this theory is the Ergodic Theorem due to the American mathematician **George David Birkhoff** (1884–1944).

A Markov chain is ergodic if it is irreducible, aperiodic and positive recurrent. For Markov chains this is actually (somewhat confusingly) the definition of ergodicity. Note, however, that the definition of ergodicity can be stated for general stochastic processes. It is just the Markov case where one obtains the nice and relatively simple characterization given above.

George David Birkhoff (1884–1944)

11.1 Example (Brutopian Caste System)
The Kingdom of Brutopia recently divided its population into three equally large castes: the Nobility, the Burghers and the Serfs. The mobility between the castes will be (somehow enforced to be)

	Noble	**Burgher**	**Serf**
Noble	95%	5%	0%
Burgher	1%	61%	38%
Serf	0%	2%	98%

What will happen to the castes in the long run?

The key Example 11.1 of this lecture deals in calculating the long-term behavior of an ergodic Markov chain. It deals with a finite state-space Markov chain, which makes the situation relatively simple, as there are not so many ways things can go bad.

Long-Run, Limiting, and Stationary Probabilities

What happens in the long-run in a Markov chain (if anything) is not clear. Indeed, it is not clear what the "long-run" means. Definition 11.2 below gives three reasonable interpretations: long-run proportions, limiting probabilities and stationary probabilities.

Recall that X_n, $n \in \mathbb{N}$, is a time-homogeneous discrete-time Markov chain with state-space $\mathbb{S} = \{s_i; i \in \mathbb{I}\}$, where the index set $\mathbb{I}$ is discrete. The initial distribution of the Markov chain is denoted by $\mathbf{p} = [p_i]_{i\in\mathbb{I}}$, and the transition probability matrix of the Markov chain is denoted by $\mathbf{P} = [P_{ij}]_{i,j\in\mathbb{I}}$.

Let $V_n(j)$ be the number of times the Markov chain visits state s_j before time n. Note that in the notation of Lecture 10 this means that

$$V_n(j) \quad = \quad \sum_{k=0}^{n-1} I_k(j).$$

11.2 Definition (Long-Run, Limiting and Stationary Probabilities)

(i) If the limits

$$\bar{\pi}_j \quad = \quad \lim_{n\to\infty} \frac{1}{n} V_n(j)$$

exist and are independent of the initial distribution $\mathbf{p}$, then the limit $\bar{\boldsymbol{\pi}} = [\bar{\pi}_j]_{j\in\mathbb{I}}$ is called the **long-run probability** of the Markov chain.

(ii) If the limits

$$\tilde{\pi}_j \quad = \quad \lim_{n\to\infty} \sum_{i\in\mathbb{I}} p_i P_{ij}^n$$

exist and are independent of the initial distribution $\mathbf{p}$, then $\tilde{\boldsymbol{\pi}} = [\tilde{\pi}_j]_{j\in\mathbb{I}}$ is called the **limiting probability** of the Markov chain.

(iii) If the **(full) balance equation**

$$\boldsymbol{\pi}\mathbf{P} \quad = \quad \boldsymbol{\pi}$$

admits a solution $\boldsymbol{\pi} = [\pi_j]_{j\in\mathbb{I}}$ that is a probability vector, then the solution is called the **stationary probability** of the Markov chain.

The long-run and limiting probabilities are self-explanatory. The stationary probability is less so. So, let us explain why the stationary probability is named so and where the

balance equation comes from. Suppose that at some point of time (say n) the Markov chain has distribution $\mathbf{q}_n$. Then, at the next point $(n+1)$ the Markov chain has distribution $\mathbf{q}_{n+1} = \mathbf{q}_n\mathbf{P}$. Now, if the balance equation holds for $\mathbf{q}_n$, i.e., $\mathbf{q}_n\mathbf{P} = \mathbf{q}_n$, then $\mathbf{q}_n = \mathbf{q}_{n+1}$, and consequently $\mathbf{q}_m = \mathbf{q}_n$ for all times $m \geq n$. The interpretation of this is that at time n the Markov chain has reached its **stationary state**. In other words, the stationary probability $\boldsymbol{\pi}$ can be characterized as such initial probability that the Markov X_n, $n \in \mathbb{N}$, is stationary, i.e., for each $\ell \in \mathbb{I}$, the unconditional probabilities

$$\mathbb{P}[X_n = s_\ell] \quad = \quad \sum_{k\in\mathbb{I}} \mathbb{P}[X_n = s_\ell | X_0 = s_k]\,\mathbb{P}[X_0 = s_k]$$

do not depend on the time n, if $\mathbb{P}[X_0 = s_k] = \pi_k$ for all $k \in \mathbb{I}$.

11.3 Remark (Flux In = Flux Out Principle)
Another way of looking at the balance equation is to note that since $\sum_{i\in\mathbb{I}} P_{ji} = 1$, we can write the balance equation (actually, a set of equations)

$$\pi_j \quad = \quad \sum_{i\in\mathbb{I}} \pi_i P_{ij} \qquad \text{for all } j \in \mathbb{I}$$

as

$$\sum_{i\in\mathbb{I}} \pi_j P_{ji} \quad = \quad \sum_{i\in\mathbb{I}} \pi_i P_{ij} \qquad \text{for all } j \in \mathbb{I}.$$

The interpretation of the equation above is that **for all states s_j the rate at which the chain leaves the state s_j is the same as the rate at which the chain enters the state s_j**. This interpretation of the balance equation is extremely useful in the continuous time Markov chain setting and in the queueing theory.

Let us then discuss the connection between the long-run, limiting and stationary probabilities.

The limiting probability is a fragile notion in the sense that it may fail to exist even though the both long-run and the stationary probabilities exist, and are the same. To see this simply consider the very simple alternating Markov chain

$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

It is not difficult to see that the long-run probabilities are $\bar{\boldsymbol{\pi}} = [0.5 \;\; 0.5]$ and that they also satisfy the balance equation. Consequently, in this case $\bar{\boldsymbol{\pi}} = \boldsymbol{\pi} = [0.5 \;\; 0.5]$. The limiting probability $\tilde{\boldsymbol{\pi}}$ does not exists, however. Indeed, suppose the initial distribution is $\mathbf{p} = [1 \;\; 0]$. Then it is easy to see that $\mathbf{p}^{2n} = [1 \;\; 0]$ and $\mathbf{p}^{2n+1} = [0 \;\; 1]$. In this counterexample it is clear that it is the period (or alternating nature) of the Markov chain that messed up the limiting probabilities. The long-run probabilities or the stationary probabilities were not affected by the period.

The existence of the limiting probability implies the existence of the stationary probability. Indeed, suppose the limit $\tilde{\boldsymbol{\pi}} = \lim_{n\to\infty} \mathbf{p}\mathbf{P}^n$ exist and is independent of the initial probability $\mathbf{p}$. Then

$$\begin{aligned}\tilde{\boldsymbol{\pi}} &= \lim_{n\to\infty} \mathbf{p}\mathbf{P}^n \\ &= \lim_{n\to\infty} \mathbf{p}\mathbf{P}^{n+1} \\ &= \lim_{n\to\infty} \mathbf{p}\mathbf{P}^n\mathbf{P} \\ &= (\lim_{n\to\infty} \mathbf{p}\mathbf{P}^n)\mathbf{P} \\ &= \tilde{\boldsymbol{\pi}}\mathbf{P}.\end{aligned}$$

So, we see that the limiting probability $\tilde{\boldsymbol{\pi}}$ solves the balance equation.

The existence of the stationary probability does not imply the long-run probability Indeed, consider the Markov chain

$$\mathbf{P} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Then any probability distribution satisfies the balance equation. However, in this case the long-run probability is simply the initial distribution. So, it certainly is not independent of it. The problem here is that the Markov chain is not irreducible: the states s_1 and s_2 do not communicate.

Law of Large Numbers

Let us explain the role of the positive recurrence in the long-time behavior of Markov chains.

Recall that $m_i = \mathbb{E}[T_i]$ is the mean recurrence time for the state s_i. If the state s_i is null recurrent or transient, then $m_i = \infty$.

The following law of large numbers for Markov chains probably dates back to the theological dispute on the free will between Andrey Markov and Pavel Nekrasov.

11.4 Theorem (Law of Large Numbers)
Let X_n, $n \in \mathbb{N}$, be an **irreducible** Markov chain. Then the long-run probability $\bar{\boldsymbol{\pi}}$ exists and

$$\bar{\pi}_j = \frac{1}{m_j}.$$

Let us argue why the Law of Large Numbers for Markov chains 11.4 above is true.

For transient and null recurrent states Theorem 11.4 is clear. Indeed, in this case we obviously have $\bar{\pi}_j = 1/m_j = 0$.

Let us then consider the positive recurrent states.

Naturally, we will use the classical law of large numbers for independent and identically distributed random variables.

Let $T_j(r)$ be the length of the r^{th} **excursion**, i.e., the time it takes for the Markov chain to return to the state s_j the r^{th} time after the $(r-1)^{\text{th}}$ visit to the state s_j. Then it follows from the Markovianity that the random variables $T_j(r)$, $r \in \mathbb{N}$, are independent and identically distributed and their distribution is the same as that of the (first) return time T_j. Consequently, the classical law of the large numbers state that

$$\frac{1}{V}\sum_{r=1}^{V} T_j(r) \quad \to \quad m_j \qquad \text{as } V \to \infty.$$

Note now that $\sum_{r=1}^{V} T_j(r)$ is the time the V^{th} revisit to the state s_j occurs. Since $V_n(j)$ is the number of times s_j is visited before time n, we have the double estimate

$$\sum_{r=1}^{V_n(j)} T_j(r) \quad \leq \quad n \quad \leq \quad \sum_{r=1}^{V_n(j)+1} T_j(r).$$

This estimate looks complicated, but if you meditate on it, perhaps over a nice cup of tea, you will eventually realize that it is actually quite trivial.

Now, since s_j is recurrent, $V_n(j) \to \infty$, and the claim of Theorem 11.4 follows by flipping x to $1/x$ in the double estimate

$$\frac{1}{V_n(j)}\sum_{r=1}^{V_n(j)} T_j(r) \quad \leq \quad \frac{n}{V_n(j)} \quad \leq \quad \frac{1}{V_n(j)}\sum_{r=1}^{V_n(j)+1} T_j(r)$$

and the **hamburger principle**.

It is important to note here that **only** the irreducibility of the Markov chain was needed in the arguments above. So, Theorem 11.4 is quite general: the Markov chain can be periodic and/or transient. Finally, recall from Lecture 10 that the reciprocal mean recurrence time $1/m_j$ is the Cesàro mean of the n-step transition probabilities. Consequently,

$$\bar{\pi}_j \quad = \quad \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} P_{jj}^n.$$

Ergodic Theorem

Recall that a Markov chain is **ergodic** if it is irreducible, aperiodic and positive recurrent. For irreducible Markov chains, Theorem 11.4 implies that the long-run probabilities exist and they satisfy

$$\bar{\pi}_j \quad = \quad \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} P_{jj}^n.$$

Now, it can be shown that if (and only if) the Markov chain is **aperiodic**, then the Cesàro limit above actually converges as a normal limit. In other words, the limiting probabilities

$$\tilde{\pi}_j \quad = \quad \lim_{N\to\infty} P_{jj}^N$$

exist. Since the existence of the limiting probability implies the existence of the stationary probability, we have the following ergodic theorem.

11.5 Theorem (Ergodic Theorem)
Let X_n, $n \in \mathbb{N}$, be **ergodic** Markov chain. Then the long-run probability $\bar{\boldsymbol{\pi}}$, the limiting probability $\tilde{\boldsymbol{\pi}}$ and the stationary probability $\boldsymbol{\pi}$ all exist, and they all are the same. Moreover, the **time-averages are state-space averages** in the sense that for all (bounded) functions $f : \mathbb{S} \to \mathbb{R}$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) \quad = \quad \sum_{i\in\mathbb{I}} f(s_i)\,\pi_i.$$

Solving Balance Equations

The balance equation $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$, or the linear system

$$\pi_j \quad = \quad \sum_{i\in\mathbb{I}} \pi_i P_{ij}, \quad j \in \mathbb{I},$$

can be solved in many ways. One should note however, that we insist that $\boldsymbol{\pi}$ is a **probability** vector. Indeed, without this additional assumption a zero vector would always be a solution. Also, if a vector $\mathbf{v}$ is a solution, then the vector $c\mathbf{v}$ would also be a solution for any scalar constant c. Thus, the assumption that $\boldsymbol{\pi}$ is a probability vector is essential in making the solution unique.

Adding the requirement $\sum_{i\in\mathbb{I}} \pi_i = 1$ to the balance equation is quite simple. Adding the non-negativity requirement $\pi_i \geq 0$ is less so. Thus, the usual strategy is just to omit the non-negativity requirement, and hope for the best. In other words, one solves the system

$$\begin{aligned} \pi_j &= \sum_{i\in\mathbb{I}} \pi_i P_{ij}, \quad j \in \mathbb{I}, \\ 1 &= \sum_{i\in\mathbb{I}} \pi_i, \end{aligned}$$

and hopes to get a solution with non-negative π_j's. For Octave implementation, we note that the system above can be written as

$$\begin{aligned} \boldsymbol{\pi}(\mathbf{P}-\mathbf{I}) &= \mathbf{0}, \\ \boldsymbol{\pi}\mathbf{1}' &= 1, \end{aligned}$$

or even more compactly as

$$\boldsymbol{\pi}\begin{bmatrix}\mathbf{P}-\mathbf{I} & \mathbf{1}'\end{bmatrix} \;=\; [\mathbf{0}\;\; 1].$$

Here $\mathbf{I}$ is the identity matrix; $\mathbf{0}$ and $\mathbf{1}$ are row vectors of zeros and ones, respectively. Now, Octave's **right division** operator `/` solves linear systems from the "right": the solution of the system `x*A=b` is given by `x=b/A`. Writing the Octave code that solves the balance equation by solving a linear system should now be quite obvious.

11.6 Remark
Instead of solving the balance equation as a linear system, one can solve the corresponding left **eigenvalue** problem

$$\boldsymbol{\pi}\mathbf{P} \;=\; \lambda\boldsymbol{\pi}.$$

The stationary distribution $\boldsymbol{\pi}$ is the left eigenvector corresponding to the eigenvalue $\lambda = 1$. This method of solving will typically involve normalizing the resulting eigenvector to get a probability vector. This is the case with Octave's function `eig`, for example. Type `help eig` on the Octave console to see how `eig` works, and how to use it to solve the balance equation.

If the Markov chain is ergodic, one can solve the balance equation by using the limiting probability. This is how we solve Example 11.1 below. In the solution we take $n = 10\,000$ just for the fun of the overkill. Indeed, there is no visible change in the n-step transition matrices after $n = 400$.

11.7 Example (Brutopian Caste System, Solution)
The Markov chain modeling the castes of Brutopia has transition probability matrix

$$\mathbf{P} \;=\; \begin{bmatrix} 0.95 & 0.05 & 0.00 \\ 0.01 & 0.61 & 0.38 \\ 0.00 & 0.02 & 0.98 \end{bmatrix}.$$

Obviously the Markov chain is ergodic. Thus the limiting, long-run and stationary distributions exist and they are all the same. Therefore, we can calculate, e.g.,

$$\mathbf{P}^{10\,000} \;=\; \begin{bmatrix} 0.0099 & 0.0495 & 0.9406 \\ 0.0099 & 0.0495 & 0.9406 \\ 0.0099 & 0.0495 & 0.9406 \end{bmatrix}.$$

This means that in the long run (or eventually) the Brutopian population will consist of 1% nobles, 5% burghers and 94% serfs.

Exercises

11.1 Exercise
Suppose that the probability whether it rains tomorrow depends only on whether it has rained today. Let X_n, $n \in \mathbb{N}$, be the Markov chain modeling the weather: $X_n = 0$ if it rains at day n and $X_n = 1$ if it does not rain at day n. Let

$$\mathbf{P} = \begin{bmatrix} 0.95 & 0.05 \\ 0.30 & 0.70 \end{bmatrix}$$

be the transition probability matrix of X_n, $n \in \mathbb{N}$.

(a) Suppose that on Monday it rains with probability 0.25. What is the probability that it rains on Wednesday?
(b) In the long run, how may rainy and non-rainy days would you expect in this model?

11.2 Exercise (Skewed Random Walk with Absorbing Boundaries)
Consider a skewed random walk, i.e., a Markov chain with state-space $\mathbb{S} = \mathbb{Z}$ with transition probabilities

$$P_{i,i+1} = p = 1 - P_{i,i-1},$$

if $i \notin \{l, u\}$, and

$$P_{l,l} = 1 = P_{u,u}.$$

What will happen to the Markov chain in the long run?

11.3 Exercise (Cesàro Mean)
Consider a sequence a_n, $n \in \mathbb{N}$, and its Cesàro mean $\bar{a}_n = \frac{1}{n}\sum_{k=1}^{n} a_k$, $n \in \mathbb{N}$.

(a) Suppose $a_n \to a$. Show that then $\bar{a}_n \to a$, but the converse does not hold.
(b) How is part (a) related to Markov chains?

11.4 Exercise (Gambler's Ruin)
Mr. S. and Ms. L. are playing a coin-tossing game with a fair coin. If the coin lands on heads, Ms. L. will give Mr. S. one Euro. If the coin lands on tails, Mr. S. will give Ms. L.

one Euro. Mr. S. has capital of 500 Euros and Ms. L. has capital of 50 Euros. The game is played until either Mr. S. or Ms. L. loses his/her capital.

(a) What is the probability that Ms. L. wins the game?
(b) Suppose the coin is not fair. Let p be the probability that Ms. L. wins and individual toss. What should p be so that the game is fair, i.e., the probability for Ms. L. to win the game is $1/2$?

Hint: Develop a recurrence equation for losing one's capital by conditioning on what happens after the first coin toss, or just use Google.

11.5 Exercise
Write an Octave function that solves the balance equation $\boldsymbol{\pi}\mathbf{P} = \mathbf{P}$ for a Markov chain by using

(a) limiting probabilities,
(b) by solving the equation as a linear system,
(c) by calculating the eigenvalue decomposition.

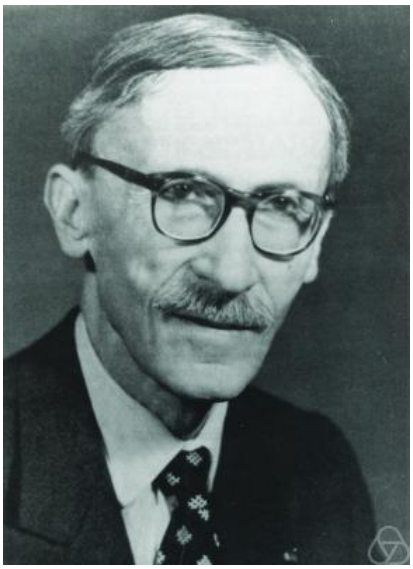
Lecture 12

Poisson Process

The Poisson process is one of the most widely-used counting processes. It is usually used in scenarios where we are counting the occurrences of certain events, or arrivals, that appear to happen at a constant rate or intensity, but completely at random. Without any special knowledge of the randomness in the occurrences of the events, it is natural to assume that they are independently and homogeneously scattered in time. Therefore, the Poisson process is the cumulative Poisson point process on the positive real line. Since the points of the Poisson point process are homogeneously and independently scattered, this means that the Poisson process has stationary and independent increments.

Processes with stationary and independent increments are called Lévy processes in honor of the French mathematician **Paul Pierre Lévy** (1886–1971). If a Lévy process is continuous, then it can be shown that it is the Brownian motion. If a Lévy process is a counting process, then we shall show in this lecture that it is the Poisson process. It can be shown that all other Lévy processes can be constructed from these two processes by using superpositioning.

The Poisson process is also a continuous-time Markov chain with discrete state-space. Indeed, all Lévy processes are continuous-time Markov processes.



Paul Lévy (1886–1971)

12.1 Example (The Big Machine)
Technician Genus Paperus oversees The Big Machine. The Big Machine has, on average, a nasty failure in every 4 hours. To prevent a nasty failure, Genus Paperus must press a red reset button. One nasty failure is not too bad, but two nasty failures without reseting will lose Genus Paperus his job, and 5 nasty failures without reseting will lose him his life as The Big Machine will explode. Now, Genus Paperus is a lazy bastard. So, in his 8 hours shift he waits for 4 hours for possible nasty failures, presses the button if needed, and then goes to sleep. What is the probability that Genus Paperus will lose his job in a given shift?

Qualitative Approach to Poisson Process

A stochastic process $N(t),\ t \geq 0$, is said to be a **counting process** if $N(t)$ represents the total number of "events" or "arrivals" that occur by time t. In a more technical language, a counting process is a process that satisfies

(i) $N(t) \in \mathbb{N}$,
(ii) if $s \leq t$, then $N(s) \leq N(t)$,

The Poisson process is a counting process where the arrival points are independently and homogeneously scattered. This means that the Poisson process is a cumulative Poisson point process on $\mathbb{R}_+$. Yet another way of saying what we just said is: If $X(A)$, $A \subset \mathbb{R}_+$, is a Poisson point process, then $N(t) = X([0,t])$, $t \geq 0$, is a Poisson process. And now, we say what we just said three times, a fourth time in the definition below:

12.2 Definition (Poisson Process)
A continuous-time stochastic process $N(t),\ t \geq 0$, is the **Poisson process** with rate λ, if it is a counting process with independent and stationary increments with $\mathbb{E}[N(t)] = \lambda t$.

Continuous-time Markov chains can be defined pretty much the same way as discrete-time Markov chains:

12.3 Definition (Continuous-Time Markov Chain)
A continuous-time stochastic process $X(t),\ t \geq 0$, with discrete state-space $\mathbb{S} = \{s_i\,;\, i \in \mathbb{I}\}$ is a **Markov chain** if

$$\mathbb{P}\big[X(t) = s_j \,|\, X(t_1) = s_{i_1}, X(t_2) = s_{i_2}, \ldots, X(t_n) = s_{i_n}\big] \;=\; \mathbb{P}\big[X(t) = s_j \,|\, X(t_n) = s_{i_n}\big]$$

for all $t_1 \leq t_2 \leq \cdots \leq t_n \leq t$ and $s_j, s_{i_1}, s_{i_2}, \ldots, s_{i_n} \in \mathbb{S}$. If for all $s \leq t$ and $s_i, s_j \in \mathbb{S}$,

$$\mathbb{P}\big[X(t) = s_j \,|\, X(s) = s_i\big] \;=\; \mathbb{P}\big[X(t-s) = s_j \,|\, X(0) = s_i\big],$$

we say that the Markov chain $X(t),\ t \geq 0$, is **time-homogeneous**.

The Poisson process is a time-homogeneous continuous-time Markov chain. Indeed, we only need to "force" the independent and stationary increments $X(t_k) - X(t_{k-1}) = s_{i_k} - s_{i_{k-1}}$ into Definition 12.3 in the following way. Let $N(t),\ t \geq 0$, be a Poisson process. Let $t_1 \leq t_2 \leq \cdots \leq t_n \leq t$ and $j, i_1, \ldots, i_n \in \mathbb{N}$. Denote

$$\Delta N(t_k) = N(t_k) - N(t_{k-1})$$

and

$$\Delta j_k = j_k - j_{k-1}.$$

Let $t_0 = 0$ and $j_0 = 0$. Then, since $N(0) = 0$, we have, by using the independence and stationarity of the increments $\Delta N(t_k)$, $k \le n$, that

$$\begin{aligned}
&\mathbb{P}[N(t) = j \,|\, N(t_1) = j_1, \ldots, N(t_n) = j_n] \\
&\quad = \mathbb{P}[N(t) = j \,|\, \Delta N(t_1) = j_1, \Delta N(t_2) = \Delta j_2, \ldots, \Delta N(t_n) = \Delta j_n] \\
&\quad = \mathbb{P}[N(t) - N(t_n) = j - j_n \,|\, \Delta N(t_1) = j_1, \Delta N(t_2) = \Delta j_2, \ldots, \Delta N(t_n) = \Delta j_n] \\
&\quad = \mathbb{P}[N(t) - N(t_n) = j - j_n] \\
&\quad = \mathbb{P}[N(t - t_n) = j - j_n].
\end{aligned}$$

Thus, we have obtained the following result:

12.4 Proposition (Poisson Process is Markovian)
The Poisson process is a continuous-time time-homogeneous Markov chain with state-space $\mathbb{S} = \mathbb{N}$.

Quantitative Approach to Poisson Process

Since the Poisson process $N(t)$, $t \ge 0$, with rate λ is the cumulative Poisson point process on the positive real line $\mathbb{R}_+$, i.e.,

$$N(t) \quad = \quad X([0, t]),$$

where $X(A)$, $A \subset \mathbb{R}_+$, is a Poisson point process with parameter λ, the results of Lecture 6 imply that:

12.5 Theorem (Poisson Process has Poisson Distribution)
Let $N(t)$, $t \ge 0$, be a Poisson process with rate $\lambda > 0$. Then

$$\mathbb{P}[N(t) = n] \quad = \quad e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \qquad \text{for } n = 0, 1, 2, \ldots,$$

i.e, $N(t)$ has the Poisson distribution with parameter λt.

Combining Theorem 12.5 with the stationary and independent increments makes it possible to calculate virtually everything related to the Poisson process. For example, suppose we want to calculate

$$\mathbb{P}\big[N(10) = 7\,,\, N(5) = 7\,,\, N(4) = 6 \,\big|\, N(1) = 1\,,\, N(3) = 2\big]$$

for a Poisson process $N(t)$ with rate 0.2. By "forcing" independent increments to the probability in question, we obtain

$$\begin{aligned}
&\mathbb{P}\big[N(10)=7\,,N(5)=7\,,N(4)=6\,\big|\,N(1)=1\,,N(3)=2\big]\\
&\quad=\ \mathbb{P}\big[N(10)-N(5)=0\,,N(5)-N(4)=1\,,N(4)-N(3)=4\,\big|\,N(3)-N(1)=1\big]\\
&\quad=\ \mathbb{P}[N(10)-N(5)=0]\,\mathbb{P}[N(5)-N(4)=1]\,\mathbb{P}[N(4)-N(3)=4].
\end{aligned}$$

By using the stationarity of the increments and by plugging in the Poisson probabilities, we obtain

$$\begin{aligned}
&\mathbb{P}\big[N(10)=7\,,N(5)=7\,,N(4)=6\,\big|\,N(1)=1\,,N(3)=2\big]\\
&\quad=\ \mathbb{P}[N(5)=0]\,\mathbb{P}[N(1)=1]\,\mathbb{P}[N(1)=4]\\
&\quad=\ \mathrm{e}^{-0.2\times5}\ \times\ \mathrm{e}^{-0.2}\times0.2\ \times\ \mathrm{e}^{-0.2}\times\frac{0.2^4}{4!}\\
&\quad=\ 3.2880\times10^{-6}.
\end{aligned}$$

12.6 Example (The Big Machine, Solution)
Let $N(t)$, $t\geq0$, be the process that counts the nasty failure of The Big Machine during Genus Paperus's shift. Let us measure the time by hours. Then the question asked is

$$\mathbb{P}[N(8)-N(4)\geq2\,|\,N(4)=0]\,,$$

where $N(t)$, $t\geq0$, is a counting process with rate $1/4=0.25$. A reasonable assumption is that $N(t)$, $t\geq0$ is a Poisson process. Then we can calculate

$$\begin{aligned}
&\mathbb{P}[N(8)-N(4)\geq2\,|\,N(4)=0]\\
&\quad=\ \mathbb{P}[N(4)\geq2]\\
&\quad=\ 1-\mathbb{P}[N(4)\leq1]\\
&\quad=\ 1-\big(\mathbb{P}[N(4)=0]+\mathbb{P}[N(4)=1]\big)\\
&\quad=\ 1-\big(\mathrm{e}^{-0.25\times4}+\mathrm{e}^{-0.25\times4}\times(0.25\times4)\big)\\
&\quad=\ 26.424\ \%.
\end{aligned}$$

Let us then consider the points, or arrivals, of the Poisson process. Let S_n be the time the n^{th} arrival occurs, i.e.,

$$S_n\ =\ \min\{t\geq0\,;\,N(t)=n\}\,.$$

Let $T_n=S_{n+1}-S_n$ be the interarrival time between the $(n+1)^{\text{th}}$ and the n^{th} arrival. Since the Poisson process has stationary and independent increments, the interarrival times are identical in distribution and independent. Also, since the Poisson process is Markovian, the remaining time it spends in the current state is independent from the time it has already spent in the current state. This implies that the interarrival times have the no-memory property. Consequently, we have the following result:

12.7 Proposition (Poisson Arrivals)
Let $N(t)$, $t \geq 0$, be a Poisson process with rate λ. Then its interarrival times T_n are independent and exponentially distributed with parameter λ. The arrival time S_n is Erlang distributed with parameters n and λ.

12.8 Remark (Simulating Poisson Process)
Proposition 12.7 is very useful in simulation. Indeed, to simulate a Poisson process one only needs to simulate independent exponentially distributed random variables. Another way to simulate the Poisson process is to use the law of small numbers and approximate the Poisson process by a **Bernoulli process**. A Bernoulli process B_n, $n \in \mathbb{N}$, is a discrete-time Markov chain, where $B_0 = 0$ and at each time point n, B_n jumps up with independent probability p, and stays in its previous state with probability $1-p$. If the time-step Δt is small and $p = \lambda \Delta t$, then the Bernoulli process is close to the Poisson process.

Proposition 12.9 below states that Poisson processes remain Poisson processes under merging and splitting. This result is very convenient in queuing systems. Indeed, it says that if the input streams to a queueing system are independent Poisson processes, then the total input stream is also a Poisson process. Also, if a Poisson stream is split randomly to different queues, then the individual input streams to the queues are independent Poisson processes.

12.9 Proposition (Merging and Splitting)
(i) Suppose that $N_1(t)$ and $N_2(t)$ are two independent Poisson processes with respective rates λ_1 and λ_2. Then the **merged** process $N_1(t)+N_2(t)$ is a Poisson process with rate $\lambda_1+\lambda_2$.
(ii) Suppose that $N(t)$ is a Poisson process with rate λ and that each arrival is marked with probability p independent of all other arrivals. Let $N_1(t)$ and $N_2(t)$ denote respectively the **split** processes, i.e., number of marked and unmarked arrivals in $[0,t]$. Then $N_1(t)$ and $N_2(t)$ are independent Poisson processes with respective rates λp and $\lambda(1-p)$.

The validity of Proposition 12.9 can be seen by noticing that the property of independent and stationary increments is preserved under independent merging and splitting.

Continuous-Time Markov Chains

We end this lecture by briefly discussing general continuous-time time-homogeneous Markov chains $X(t)$, $t \geq 0$, having discrete state space $\mathbb{S} = \{s_i\,;\, i \in \mathbb{I}\}$, where $\mathbb{I}$ is either $\mathbb{N}$, $\mathbb{Z}$, or finite. The discussion here is quite informal, and we omit all the nasty details. Nevertheless, this discussion should be valuable in building intuition for Markovian queueing systems.

Suppose the Markov chain is in state s_i at time t. Then, by the Markov property, the remaining **sojourn time** in the state s_i of the chain is independent of the time already spent in the state. Consequently, the time a Markov chain spends in a state s_i is exponentially distributed with some state-dependent parameter ν_i. When the Markov chain decides to leave the state s_i, it will jump to the state s_j with probability P°_{ij}. The matrix $\mathbf{P}^\circ$ is the transition probability matrix of the **skeleton** of the chain $X(t)$, $t \geq 0$, that looks at the continuous time process only at the times the state changes. Obviously we have $P^\circ_{ii} = 0$, but otherwise $\mathbf{P}^\circ$ can be any stochastic matrix. So, a continuous-time Markov chain is completely determined by the **sojourn intensities** $\boldsymbol{\nu}$ and and the **skeletal transition probabilities** $\mathbf{P}^\circ$. From the philosophical point of view this is good news: continuous-time Markov chains have a clear structure. From the simulation point of view this is also good news: it is pretty obvious how to simulate continuous-time Markov chains. From the theoretical-technical point of view the news are not so good. The structure is pretty difficult to analyze.

Let us then try to analyze continuous-time Markov chains from the theoretical-technical point of view. Since the Markov chain is time-homogeneous (by assumption), the thing to analyze is the family $\mathbf{P}(t)$, $t \geq 0$, of transition probability matrices, where

$$P_{ij}(t) \quad = \quad \mathbb{P}\big[X(t+t_0) = s_j \,\big|\, X(t_0) = s_i\big].$$

This family of matrices is very inconvenient as such. Fortunately, just like in the discrete-time case, we can use the **law of total probability** to obtain the **Chapman–Kolmogorov equations**

$$\mathbf{P}(t+s) \quad = \quad \mathbf{P}(t)\mathbf{P}(s).$$

Then, under some very mild regularity conditions, the continuous-time Chapman–Kolmogorov equations work just like the **Cauchy's functional equation**: by setting

$$Q_{ij} \quad = \quad P'_{ij}(0),$$

the solution of the equation $\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s)$ is given by

$$\mathbf{P}(t) \quad = \quad \mathrm{e}^{t\mathbf{Q}},$$

where we have used the **matrix exponential**

$$\mathrm{e}^{\mathbf{A}} \quad = \quad \sum_{n=0}^{\infty} \frac{1}{n!}\mathbf{A}^n.$$

So far this does not look very promising. We have only found a very complicated way to express the transition probabilities $\mathbf{P}(t)$, $t \geq 0$. However, it turns out that the matrix $\mathbf{Q}$, also called the **infinitesimal generator** of the transition probabilities $\mathbf{P}(t)$, $t \geq 0$, has a very clear probabilistic interpretation. Indeed,

$$\begin{aligned} Q_{ij} &= \text{the rate (or flux) at which the transition } s_i \to s_j \text{ occurs if } i \neq j, \\ -Q_{ii} &= \text{the rate (or flux) at which the state } s_i \text{ is stayed.} \end{aligned}$$

From the practical point of view this means that, in continuous-time Markovian modeling, one should model the infinitesimal generator $\mathbf{Q}$, not the transition probabilities $\mathbf{P}(t)$, $t \geq 0$.

12.10 Remark
We note that $\sum_j Q_{ij} = 0$ (total flux) and $Q_{ii} = -\sum_{j \neq i} Q_{ij}$, (flux-to-stay is negative flux-to-leave).

Let us then have three examples without any proofs or explanations whatsoever.

As a first example, consider the Poisson process with rate λ. It has the infinitesimal generator

$$\mathbf{Q} = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & 0 & \cdots \\ 0 & -\lambda & \lambda & 0 & 0 & 0 & \cdots \\ 0 & 0 & -\lambda & \lambda & 0 & 0 & \cdots \\ 0 & 0 & 0 & -\lambda & \lambda & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}.$$

As a second example, consider a single-server queue, where customers arrive at Poisson rate λ and they are served with exponential service times with parameter μ. So, we have input rate λ and output rate μ. The continuous-time Markov chain denoting the number of customers in the system has infinitesimal generator

$$\mathbf{Q} = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & 0 & \cdots \\ \mu & -(\lambda+\mu) & \lambda & 0 & 0 & 0 & \cdots \\ 0 & \mu & -(\lambda+\mu) & \lambda & 0 & 0 & \cdots \\ 0 & 0 & \mu & -(\lambda+\mu) & \lambda & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}.$$

As a third example, consider a queue with two servers, where customers arrive at Poisson rate λ and they are served with exponential service times with parameter μ. So, we have input rate λ and output rate μ if there is only one customer and output rate 2μ if there are two or more customers. The continuous-time Markov chain denoting the number of customers in the system has infinitesimal generator

$$\mathbf{Q} = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & 0 & \cdots \\ \mu & -(\lambda+\mu) & \lambda & 0 & 0 & 0 & \cdots \\ 0 & 2\mu & -(\lambda+2\mu) & \lambda & 0 & 0 & \cdots \\ 0 & 0 & 2\mu & -(\lambda+2\mu) & \lambda & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}.$$

12.11 Remark (Flux In = Flux Out Principle)
The infinitesimal generator $\mathbf{Q}$ is useful in finding the stationary or limiting probabilities

$$\pi_i \quad = \quad \lim_{t\to\infty} \mathbb{P}[X(t) = s_i],$$

if they exist. The idea is the **flux in = flux out principle** for the state s_i:

$$\sum_{j\neq i} \pi_j Q_{ji} \quad = \quad \sum_{j\neq i} \pi_i Q_{ij}.$$

The limiting probabilities exist if the continuous-time Markov chain is **irreducible and positive recurrent**. There is no problem with the periods, as they are impossible in continuous-time time-homogeneous models.

Finally, we note that the **(full) balance equations**, or the flux in = flux out principle, can be written by using the matrix notation shortly as

$$\boldsymbol{\pi}\mathbf{Q} \quad = \quad \mathbf{0}.$$

Indeed, since $Q_{ii} = -\sum_{j\neq i} Q_{ij}$, we have for all $i \in \mathbb{I}$ that

$$\begin{aligned}\sum_j \pi_j Q_{ji} &= \sum_{j\neq i} \pi_j Q_{ji} + \pi_i Q_{ii} \\ &= \sum_{j\neq i} \pi_i Q_{ij} - \pi_i \sum_{j\neq i} Q_{ij} \\ &= 0.\end{aligned}$$

So, the flux in = flux out principle can also be stated as a principle of zero total flux.

Exercises

12.1 Exercise
Let $N(t)$, $t \geq 0$, be a Poisson process with intensity 0.5. Calculate

(a) $\mathbb{P}[N(1) = 0, N(3) = 0]$,
(b) $\mathbb{P}[N(3) = 5, N(2) = 1 | N(1) = 0]$,
(c) $\mathbb{P}[N(3) > 2 | N(1) = 1]$,
(d) $\mathbb{P}[N(5) \leq 4 | N(1) = 2]$.

12.2 Exercise
Consider Example 12.1. What is the probability that technician Genus Paperus will die during a given shift in The Big Machine's explosion.

12.3 Exercise
People immigrate into the Kingdom of Brutopia at a Poisson rate 1 per day.

(a) What is the expected time until the tenth immigrant arrives?
(b) What is the probability that the elapsed time between the tenth and the eleventh immigrant exceeds two days?

12.4 Exercise (Simulation of Poisson and Bernoulli Processes)
Make an Octave function that simulates the Poisson process with rate λ on the interval $[0, 1]$

(a) exactly by using the exponential interarrival times,
(b) approximately by using the law of small numbers.

12.5 Exercise
A Markovian queue is fed customers with rate λ. There are some servers, that each feed out customers with rate μ. Let $X(t)$ denote the number of customers in the system at time t. What is the infinitesimal generator of the Markov chain $X(t)$, $t \geq 0$, when

(a) there are three servers,
(b) there are m servers,
(c) there are infinite number of servers?

Part IV

Queueing

Lecture 13

Little, Palm, and PASTA

Queueing theory is the mathematical study of waiting lines, or queues. It has its origins in the research by the Danish mathematician **Agner Krarup Erlang** (1878–1929) who studied the Copenhagen telephone exchange. Other pioneers of queueing theory include the Swedish statistician **Conrad** "**Conny**" **Palm** (1907–1951), who studied how the system's and the client's point of view of a queueing system differ. In his honor, the client's point of view is now called the Palm distribution. Maybe the most celebrated result of queueing theory is the Little's law named after the American operations researcher **John Little** (1928–). His law

$$\ell = \lambda w$$

combines in a neat way the average load, the input rate and the average waiting time of a queueing system.

John Little (1928–)

13.1 Example (Don Guido's Wine Cellar)
Don Guido has all the time on average 150 cases of wine his basement. Don Guido consumes 25 cases per year. How long does Don Guido hold each case?

Palm and PASTA

Consider $X(t)$, the state of a queuing system at time t. For definiteness, you can think of $X(t)$ as the total number of clients in the queueing system, so that the state-space is $\mathbb{N}$. If you are more bold, you can think of $X(t)$ as a vector that gives the number of clients in different queues and servers of the system, so that the state-space is $\mathbb{N}^m$, where m is the total number of queues and servers in the system.

Let

$$\pi_x \quad = \quad \lim_{t\to\infty} \mathbb{P}[X(t) = x], \quad x \in \mathbb{S},$$

and **assume** that these limiting system probabilities exist.

13.2 Remark (Stationarity and Full Balance)
If the system is Markovian, then $\boldsymbol{\pi}$ is also the stationary distribution of the system solving the **(full) balance equation** $\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$, where $\mathbf{Q}$ is the **infinitesimal generator** of the continuous-time discrete-state Markov chain $X(t)$, $t \geq 0$.

Another way of looking at the limiting probability π_X is to note that π_x is also the long-run proportion of time that the system is in state x. Yet another way of looking at the probability π_x is by considering an outsider, who observes the system at an independent random time:

$$\pi_x \quad = \quad \text{probability that an outsider finds the system at state } x.$$

Thus, the probabilities $\boldsymbol{\pi} = [\pi_x]_{x\in\mathbb{S}}$ are the **outsider's view** or the **system's view** of the queuing system.

Let us then consider the **insider's view** or the **client's view**. For a client, the obvious point of interest is the **Palm probability**

$$\pi^*_x \quad = \quad \text{probability that the arriving client finds the system at state } x.$$

13.3 Remark (Inside and Outside)
In general, $\boldsymbol{\pi}$ and $\boldsymbol{\pi}^*$ can be different. Indeed, consider a queue, where all the client are served with a deterministic constant time: 1 second, say. Suppose that the interarrival times of the clients are strictly greater than 1 second. This means that every client finds the queue empty: $\pi^*_0 = 1$. However, $\pi_0 < 1$, as they may be some clients in the system sometimes. This is very nice for the client: no waiting in queue, ever. Unfortunately, typically the case is quite the opposite. Indeed, consider the following queue: it takes, as in the previous example, exactly 1 second to serve a client. The customers arrive, on average only once in a million years. But when they arrive, they arrive in batches of two (the other after a millisecond after the one). So, from the system's point of view the queue is virtually always empty, while from the client's point of view the queue is busy with probability $1/2$.

If the queueing system if fed by a Poisson process, then the Palm and the system probabilities are the same. This **arrival theorem** is called **Poisson Arrivals See Time-Averages**, or **PASTA** for short. Consider a time interval of length T and a smaller time interval inside it of length t. We assume that there is a single Poisson arrival on the interval and show that the probability that the arrival takes place in the shorter interval is t/T. This will

show that the Poisson arrival sees the system in the same way as the random independent outsider. In a more technical language, this means that we have to show that

$$\mathbb{P}[N(t_0)=0, N(t+t_0)=1 \,|\, N(T)=1] \;=\; \frac{t}{T},$$

for all t and t_0 such that $t+t_0<T$. (Here we chose the interval of length T to be $[0,T]$, which we can do because of the stationary of the increments). Now, by the definition of the conditional expectation and by the independence and stationarity of the increments,

$$\begin{aligned}
&\mathbb{P}[N(t_0)=0\,,\, N(t+t_0)=1 \mid N(T)=1] \\
&\quad= \mathbb{P}[N(t_0)=0\,,\, N(t+t_0)=1\,,\, N(T)=1] \,\big/\, \mathbb{P}[N(T)=1] \\
&\quad= \mathbb{P}[N(t_0)-N(0)=0\,,\, N(t+t_0)-N(t_0)=1\,,\, N(T)-N(t+t_0)=0] \,/\, \mathbb{P}[N(T)=1] \\
&\quad= \mathbb{P}[N(t_0)=0]\mathbb{P}[N(t)=1]\mathbb{P}[N(T-t-t_0)=0] \,/\, \mathbb{P}[N(T)=1].
\end{aligned}$$

Hence, by plugging in the Poisson probabilities, we have that

$$\begin{aligned}
&\mathbb{P}[N(t_0)=0, N(t+t_0)=1 \,|\, N(T)=1] \\
&\quad= \mathrm{e}^{-\lambda t_0}\,\lambda t \mathrm{e}^{-\lambda t}\,\mathrm{e}^{-\lambda(T-t-t_0)} \;\big/\; \lambda T \mathrm{e}^{-\lambda T} \\
&\quad= t \,\big/\, T.
\end{aligned}$$

We have shown the following **arrival theorem**:

13.4 Theorem (PASTA)
Consider a queueing system that is fed by a Poisson process. Then the system probability π and the Palm probability π^* are the same.

13.5 Remark (Lack of Anticipation)
The insider's and the outsider's view can agree even for non-Poisson arrivals. Indeed, it can be shown that the PASTA property is true under the so-called **lack of anticipation assumption**: the future arrivals are independent of the past states of the system.

13.6 Remark (Waiting Paradox)
PASTA implies the following counter-intuitive waiting paradox:

> Suppose buses come to a bus stop according to Poisson process with 12 minutes intervals on average. You arrive to the bus stop completely randomly. Then the average waiting time for the bus to come for you is 12 minutes.

One would expect that the average waiting time is $12/2 = 6$ minutes. This is, however, the best possible average waiting time which is attained only in the non-random case, where the buses come **exactly** in 12 minute intervals. The more there is randomness in the interarrival times, the more there is average waiting time for the random client at the bus stop. Actually, the Poisson case is by far not the worst. The average waiting time can even be infinite, even though the average interarrival times are finite. The intuitive reason for this is that the client comes, with high probability, during an interval that is much longer than average. Actually, it can be shown that your average waiting time is given by the so-called mean forward recurrence time

$$w^* \quad = \quad \frac{\mathbb{E}[T^2]}{2\mathbb{E}[T]},$$

where T is the interarrival time of the buses. If T is fixed length $1/\lambda$, then $w^* = 1/(2\lambda)$. If T is the Poisson interarrival, i.e., T is exponentially distributed with mean $1/\lambda$, then $w^* = 1/\lambda$. If T has heavy tails in the sense that $\mathbb{E}[T^2] = \infty$, while $\mathbb{E}[T] = 1/\lambda$, then $w^* = \infty$.

Little's Law

Consider a queueing system where clients arrive as a point process (i.e., one-by-one), and after some time leave as a point process (i.e., one-by-one). Let S_n denote the **arrival time** of the n^{th} client. We assume that $S_n \to \infty$. The **arrival process** of the system is

$$N(t) \quad = \quad \max\{n\,;\, S_n \le t\},$$

i.e., $N(t)$ counts how many of the arrival points S_n occur in the interval $[0,t]$. Once the n^{th} client enters the system at time S_n, it will stay in the system a random time W_n. Then, the client leaves at time $S_n^\dagger = S_n + W_n$. Let $N^\dagger(t)$ denote the **departure process**, i.e.,

$$N^\dagger(t) \quad = \quad \max\left\{n\,;\, S_n^\dagger \le t\right\}.$$

The n^{th} client is in the system if and only if $S_n \le t \le S_n^\dagger$. Recall that the **indicator** $\mathbf{1}_A$ for an event A is the random variable that takes value 1 if A happens, and value 0 otherwise. Then, the informal definition

$$\begin{aligned} L(t) \quad &= \quad \textbf{load} \text{ of the system at time } t \\ &= \quad \text{the number of clients in the system at time } t \end{aligned}$$

can be written more formally as

$$\begin{aligned} L(t) &= \sum_{n=1}^{\infty} \mathbf{1}_{\left\{S_n \le t \le S_n^{\dagger}\right\}} \\ &= \sum_{n\,;\,S_n \le t} \mathbf{1}_{\left\{W_n > t - S_n\right\}} \\ &= \sum_{n=1}^{N(t)} \mathbf{1}_{\left\{W_n > t - S_n\right\}}. \end{aligned}$$

Informally, we define

$$\begin{aligned} \lambda &= \text{arrival rate} \\ w &= \text{average waiting time} \\ \ell &= \text{average load.} \end{aligned}$$

The formal definitions are

$$\begin{aligned} \lambda &= \lim_{t\to\infty} \frac{N(t)}{t}, \\ w &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} W_k, \\ \ell &= \lim_{t\to\infty} \frac{1}{t} \int_0^t L(s)\,\mathrm{d}s, \end{aligned}$$

with the **assumption** that these limits exist. We further assume that

$$\lambda = \lambda^{\dagger} = \lim_{t\to\infty} \frac{N^{\dagger}(t)}{t},$$

i.e., arrival rate = departure rate.

13.7 Remark
The assumption arrival rate = departure rate is actually unnecessary. It follows from the existence of the parameters λ, w and ℓ. In any case, it is clear that for Little's law to hold true, we must have the equality of arrival and departure rates. Indeed, departure rate cannot exceed arrival rate for obvious reasons; and if the departure rate is less than the arrival rate, the system would blow up giving infinite load.

Let us then see how the characteristics λ, w and ℓ are connected by the Little's law. We first note that the area under the path of $L(s)$ from 0 to t, i.e, $\int_0^t L(s)\,ds$, is simply the sum of whole and partial waiting times (e.g., rectangles of height 1 and lengths W_k). If the system is empty at time t, then we have exactly

$$\int_0^t L(s)\,\mathrm{d}s = \sum_{k=1}^{N(t)} W_k,$$

and in general we have

$$\sum_{k=1}^{N^\dagger(t)} W_k \quad \leq \quad \int_0^t L(s)\,\mathrm{d}s \quad \leq \quad \sum_{k=1}^{N(t)} W_k.$$

From this we obtain

$$\frac{N^\dagger(t)}{t}\,\frac{1}{N^\dagger(t)}\sum_{k=1}^{N^\dagger(t)} W_k \quad \leq \quad \frac{1}{t}\int_0^t L(s)\,\mathrm{d}s \quad \leq \quad \frac{N(t)}{t}\,\frac{1}{N(t)}\sum_{k=1}^{N(t)} W_k.$$

Now, by letting $t \to \infty$, and by using the **hamburger principle** we see, after deep contemplation of all that was discussed, that:

13.8 Theorem (Little's Law)
If both λ and w exist and are finite, then $\lambda^\dagger$ and ℓ exist; $\lambda^\dagger = \lambda$ and

$$\ell \quad = \quad \lambda w.$$

Assuming some kind of weak stability or stationarity, the solution to Don Guido's wine cellar problem can now be given by using the Little's law.

13.9 Example (Don Guido's Wine Cellar, Solution)
We note that if

$$\begin{aligned} \lambda &= \text{rate at which Don Guido consumes/buys cases,} \\ w &= \text{the waiting time of a case in the cellar,} \\ \ell &= \text{the number of cases in the cellar,} \end{aligned}$$

then the answer is given by the Little's law as

$$w \quad = \quad \frac{\ell}{\lambda} \quad = \quad \frac{150}{25/\text{years}} \quad = \quad 6 \text{ years}.$$

Exercises

13.1 Exercise
The Palm probabilities $\boldsymbol{\pi}^*$ are also called **Palm's birth probabilities** since they the what the arriving client, a.k.a. the birth, sees. The **Palm's death probabilities** $\boldsymbol{\pi}^\dagger$ are what the leaving client, a.k.a, the death, sees. They are defined as

$$\pi_n^\dagger \quad = \quad \text{proportion of clients leaving behind } n \text{ in the system when they depart.}$$

(a) Give an example of a queueing system where all the probabilities $\boldsymbol{\pi}$, $\boldsymbol{\pi}^*$ and $\boldsymbol{\pi}^\dagger$ are different.
(b) Suppose that the clients come and leave only one-by-one, i.e., there are no batch arrivals or departures. Argue that $\boldsymbol{\pi}^* = \boldsymbol{\pi}^\dagger$.

13.2 Exercise

(a) A fast food hamburger restaurant uses 2 250 kilograms of hamburger mince each week. The manager of the restaurant wants to ensure that the meat is always fresh i.e. the meat should be no more than two months old on average when used. How much hamburger mince should be kept in the refrigerator as inventory?
(b) A stable queuing system is fed clients with intensity 18 per minute. There are, on average, 1 500 clients in the system. What is the average time a client spends in the system?

13.3 Exercise

(a) It takes 120 days on average to sell a house. You observe from monitoring the classified ads that over the past year the number of houses for sale has ranged from 20to 30 at any point in time, with an average of 25. What can you say about the number of transactions in the past year?
(b) A queue depth meter shows an average of nine jobs waiting to be serviced. Another meter shows a mean throughput of 50 per second. What is the mean response time?

13.4 Exercise

The world population is approximately 7.5 (American) billions and the average life expectancy of a baby born today is 71.4 years (or was at 25th of February, 2017).

(a) Calculate, by using the Little's law, the average number of people that die each day.
(b) Explain why the number you got from the Little's law is probably wrong.

Lecture 14

Markovian Queues

A Markovian queue has the Poisson process as arrivals, i.e., the interarrival times are independent and exponentially distributed. The service times are also independent and exponentially distributed. The Markovian queues are very convenient: When the input rate is strictly less than the output rate, they satisfy the assumptions of the Little's law. They also satisfy the PASTA property. Also, the stationary distribution of a Markovian queue can be calculated by using a "flux in is flux out" principle. Finally, it can be shown that the output process of a Markovian queue is the "same" Poisson process as the input process.

A great pioneer of queuing theory was the Danish mathematician, statistician and engineer **Agner Krarup Erlang** (1878–1929), who studied the Copenhagen telephone exchange in his 1909 work *The Theory of Probabilities and Telephone Conversations* he introduced the Poisson process as the arrival process, and in doing so introduced the Markovian queue. The teletraffic unit of offered load is named erlang in Erlang's honor. Also, the distribution arising as a sum of independent identically distributied exponentals is named in his honor.

Agner Krarup Erlang (1878–1929)

14.1 Example (Mini-Market Queue, I)
Clients come to a local mini-market, on average, in 2 minute intervals. The clerk serves the clients with average rate of 3 clients per minute. The local mini-market is quite small: it can have only 6 clients in queue. If there are 6 clients in queue, the 7^{th} client has to wait on the street. This would be a great embarrassment for the manager of the local mini-market. You are going to shop in the mini-market. What is the probability that you will cause a great embarrassment to the manager?

The problem of Example 14.1 can be modeled as a Markovian queue with unlimited queueing capacity (the M/M/1 queue in the short Kendall's notation), assuming that the queue can build on the street. If this is not possible, then we have to model the problem with a finite-capacity queue (M/M/1/K queue, with $K = 6$, in the semi-short Kendall's notation).

M/M/1 Queue

In the short **Kendall's notation**, M/M/1 queue denotes a queue where the arrival process is **M**arkovian, or **M**emoryless, the service times are **M**emoryless, and there is **1** one server. In this short Kendall's notation the following standing assumptions are in force: The inter-arrivals and the service times are all mutually independent, and the infinite-capacity queue has FIFO (**F**irst-**I**n-**F**irst-**O**ut) policy. A more formal definition of the M/M/1 queue is given below.

14.2 Definition (M/M/1 Queue)
The **M/M/1 queue** is a single-server queue with unlimited capacity, where the queue is fed by a Poisson process with rate λ and the service times are independent and exponentially distributed with rate μ.

Let $L(t)$, $t \geq 0$, denote the **workload**, or **load**, of a Markovian queue, i.e., the number of clients either waiting in the queue or being served. Then from the probabilistic point of view $L(t)$, $t \geq 0$, for the $M/M/1$ queue is a constant-rate **birth-and-death process**. The "births" are the arrivals to the system and the "deaths" are departures from the system. The arrivals have rate λ and the departures have rate μ. If $\mu > \lambda$, then the system will eventually explode in the sense that $L(t) \to \infty$. The same is true for $\mu = \lambda$, although this is not immediately obvious. We denote

$$\rho \quad = \quad \frac{\lambda}{\mu}.$$

So, ρ is the **birth-to-death ratio**: there are, on average, ρ arrivals (births) for every departure (deaths). In the queueing slang, ρ is also called **utilization** of the buffer (or the queue). So, in order for the queue not to explode, its utilization must be strictly less than one. If $\mu < \lambda$, i.e., $\rho < 1$, then, as $t \to \infty$, the system will reach a **stationary state** meaning that the limiting probabilities

$$\pi_n \quad = \quad \lim_{t\to\infty} \mathbb{P}[L(t) = n], \quad n \in \mathbb{N},$$

exist. These limiting probabilities are also the long-run proportions of time the system has n clients in it (outsider's view). By **PASTA**, they are also the probabilities that an arriving customer finds n clients in the system (insider's view): $\pi_n = \pi_n^*$.

The line-of-attack in solving the stationary distribution $\boldsymbol{\pi}$ is to use the **flux in = flux out** principle. For a non-empty state $n \geq 1$ the influx is $\lambda\pi_{n-1} + \mu\pi_{n+1}$ (λ arrivals from state $n-1$ and μ departures from state $n+1$). The outflux is $(\lambda+\mu)\pi_n$ (λ arrivals and μ departures from state n). So, for $n \geq 1$ we have the **full balance equation**

$$\lambda\pi_{n-1} + \mu\pi_{n+1} \quad = \quad (\lambda+\mu)\pi_n.$$

The full balance equation can be solved by solving the **detailed balance equation**

$$\mu\pi_n \quad = \quad \lambda\pi_{n-1}.$$

The detailed balance equation has a clear probabilistic interpretation. It means that the flux from the state n to the state $n-1$ (the left-hand-side) is equal to the flux from the state $n-1$ to the state n (the right-hand-side).

14.3 Remark (Detailed and Full Balance)
In general, the detailed balance equation (a.k.a. the **local balance equation**) is more restrictive than the full balance equation: detailed balance implies full balance, but not vice versa. In our queuing models they are the same, however. The reason for this is that our queuing models are **time-reversible**. We do not go into the details of reversed time in these lectures.

To solve the detailed balance equation $\mu\pi_n = \lambda\pi_{n-1}$, we divide it by μ on the both sides and recall that $\lambda/\mu = \rho$. We obtain the equation

$$\pi_n = \rho\pi_{n-1}.$$

This recursion is easy to solve. Indeed, working backwards a couple of times we obtain

$$\begin{aligned}\pi_n &= \rho\pi_{n-1}\\ &= \rho\rho\pi_{n-2}\\ &= \rho^2\pi_{n+2}\\ &= \rho^2\rho\pi_{n-3}\\ &= \rho^3\pi_{n-3}.\end{aligned}$$

So, we see that π_n is given by

$$\pi_n = \rho^n\pi_0.$$

The "initial condition" π_0 is given by the fact that $\boldsymbol{\pi}$ is a probability distribution, i.e., $\sum_n \pi_n = 1$. Recognizing the **geometric series**, we see that

$$\sum_{n=0}^{\infty}\rho^n\pi_0 = \frac{\pi_0}{1-\rho}.$$

Therefore,

$$\pi_0 = 1-\rho$$

(which is, by the way, the probability that the system is empty). So, we have obtained the following:

14.4 Proposition (M/M/1 Queue Stationary Distribution)
For the M/M/1 queue with arrival rate λ, service rate μ and utilization $\rho = \lambda/\mu < 1$ the stationary probabilities are

$$\pi_n = \rho^n(1-\rho), \qquad n \in \mathbb{N}.$$

14.5 Example (Mini-Market, I, Solution)
Let L be the stationary load of the mini-market's queuing system. The probability in question is

$$\begin{aligned}\mathbb{P}[\text{Queue length is } > 6] &= \mathbb{P}[L > 7] \\ &= 1 - \mathbb{P}[L \le 6] \\ &= 1 - \sum_{n=0}^{6} \rho^n (1-\rho),\end{aligned}$$

where

$$\rho = \frac{1/2}{3} = 0.16667.$$

We obtain that the probability of causing a great embarrassment is 3.6×10^{-4} %, which is virtually zero.

We used the interpretation that the customer being served is not queueing. If the interpretation is that being served is also queueing, then the probability is

$$\begin{aligned}\mathbb{P}[L > 6] &= 1 - \mathbb{P}[L \le 5] \\ &= 1 - \sum_{n=0}^{5} \rho^n (1-\rho) \\ &= 2.1 \times 10^{-3} \ \%,\end{aligned}$$

which is still incredibly small.

Before going into more complicated models than the M/M/1 queue, let us see how Proposition 14.4 combines with the Little's law. In the Little's law

$$\ell = \lambda w$$

for an M/M/1 queue with input rate λ and output rate μ, the λ is simply λ. For the load ℓ we have

$$\begin{aligned}\ell &= \sum_{n=0}^{\infty} n\pi_n \\ &= \sum_{n=0}^{\infty} n\rho^n (1-\rho) \\ &= \frac{\rho}{1-\rho},\end{aligned}$$

where we used the well-known **power series formula**

$$\sum_{n=0}^{\infty} n x^n = \frac{x}{(1-x)^2}.$$

The waiting time w is then given by the Little's law as

$$\begin{aligned} w &= \frac{\ell}{\lambda} \\ &= \frac{\rho/(1-\rho)}{\lambda}. \end{aligned}$$

So, after a little bit of extremely simple algebra we obtain the following:

14.6 Proposition (M/M/1 characteristics)
For the M/M/1 queue with arrival rate λ, service rate μ and utilization $\rho = \lambda/\mu < 1$ the average load ℓ and the average waiting time w are given by

$$\begin{aligned} \ell &= \frac{\rho}{1-\rho}, \\ w &= \frac{1/\mu}{1-\rho}. \end{aligned}$$

14.7 Remark (Power Series)
In probability theory, and in life in general, one is often required to calculate power series of type

$$\sum_{n=0}^{\infty} p(n)x^n,$$

where $p(n)$ is a polynomial of n, i.e., of the form

$$p(n) = a_d n^d + a_{d-1} n^{d-1} + \cdots + a_2 n^2 + a_1 n + a_0.$$

These can be calculated by using the following (relatively general) facts in a reasonably clever way

(i)

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x},$$

(ii)

$$\frac{\mathrm{d}^k}{\mathrm{d}x^k}\left[\sum_{n=0}^{\infty} p(n)x^n\right] = \sum_{n=0}^{\infty} \frac{\mathrm{d}^k}{\mathrm{d}x^k}[p(n)x^n]$$

M/M/1/K Queue

Having understood the infinite-capacity M/M/1 queue, let us then consider the same queue, but with finite capacity K, i.e., in the semi-short Kendall's notation, the M/M/1/K queue.

14.8 Definition (M/M/1/K Queue)
The **M/M/1/K queue** is a single-server queue with capacity K, where the queue is fed by a Poisson process with rate λ and the service times are independent and exponentially distributed with rate μ. If there are K clients in the system, the next arriving client will be thrown out (i.e., denied service).

The **full balance equation** for the M/M/1/K queue is precisely the same as for the M/M/1 queue:

$$\lambda\pi_{n-1} + \mu\pi_{n+1} \quad = \quad (\lambda+\mu)\pi_n.$$

As before, denoting $\rho = \lambda/\mu$, we obtain after some clever algebra (or by using a **detailed balance equation** just as in the case of the M/M/1 queue) that

$$\pi_n \quad = \quad \rho^n \pi_0.$$

Now, the difference to the M/M/1 queue is in the probability π_0. Indeed, we have $\sum_{n=0}^{K} \pi_n = 1$, which is a **(incomplete) geometric series**. We obtain

$$\pi_0 \quad = \quad \frac{1-\rho}{1-\rho^{K+1}}.$$

Consequently, we have the following:

14.9 Proposition (M/M/1/K Queue Stationary Distribution)
For the M/M/1/K queue with arrival rate λ, service rate μ and utilization $\rho = \lambda/\mu$ the stationary probabilities are

$$\pi_n \quad = \quad \rho^n \frac{1-\rho}{1-\rho^{K+1}}, \qquad n = 0, 1, \ldots, K.$$

14.10 Remark (Stability through Denial of Access)
In the finite-capacity queue M/M/1/K, there is no need to assume that $\rho < 1$. Indeed, the balance is guaranteed by the fact that the state-space of the system is finite.

M/M/c Queue

Let us then consider a queue with many servers. The queue itself has infinite capacity.

14.11 Definition (M/M/c Queue)
The **M/M/c queue** is an unlimited queue which is fed by a Poisson process with rate λ and there are c servers, each having independent and exponentially distributed service times with rate μ. The customers wait in a single queue.

The output rate of the M/M/c queue is, **intuitively**, $c\mu$. **Rigorously**, the output rate of $c\mu$ follows from the following lemma.

14.12 Lemma (Minimum of Exponentials)
Let $T_1, \ldots, T_c$ be independent identically distributed exponential random variables with parameter μ. Then $\min(T_1, \ldots, T_c)$ is exponentially distributed with parameter $c\mu$.

To see why Lemma 14.12 is true, let us consider the case $c = 2$. The general case is then pretty obvious. It turns out that it is convenient to work with the **complementary cumulative distribution functions**, since

$$\mathbb{P}[\min(T_1, T_2) > t] \quad = \quad \mathbb{P}[T_1 > t\,,\, T_2 > t].$$

Now, by the independence,

$$\mathbb{P}[\min(T_1, T_2) > t] \quad = \quad \mathbb{P}[T_1 > t]\,\mathbb{P}[T_2 > t].$$

By plugging in the complementary cumulative distribution function of the exponential distribution with parameter μ, we see that

$$\begin{aligned} \mathbb{P}[\min(T_1, T_2) > t] &= \mathbb{P}[T_1 > t]\,\mathbb{P}[T_2 > t] \\ &= \mathrm{e}^{-\mu t}\,\mathrm{e}^{-\mu t} \\ &= \mathrm{e}^{-(2\mu)t}. \end{aligned}$$

This means that $\min(T_1, T_2)$ is exponentially distributed with parameter 2μ. The general case for $c \geq 2$ can be shown by iterating the arguments above. Thus, the claim of Lemma 14.12 should be obvious now.

In order to have balance (full or detailed, they are the same in this model), we assume that

$$\rho \quad = \quad \frac{\lambda}{c\mu} \quad < \quad 1.$$

The M/M/c queue is actually a **birth-and-death queue** with parameters

$$\begin{aligned} \lambda_n &= \lambda, \\ \mu_n &= \min(n,c)\mu. \end{aligned}$$

The input rate (or birth rate) is **constant** λ. The output rate (or death rate) is **state-dependent** $\mu_n = \min(n,c)\mu$, since at most n of the c servers can be active, and by Lemma 14.12 the minimum of n independent exponentials, each having rate μ, is exponential with rate $n\mu$.

Let us then consider the balance equations. The **detailed balance equation** is

$$\lambda\pi_{n-1} = \min(n,c)\mu\pi_n$$

with the normalization (coming from the **(incomplete) geometric series**)

$$\frac{1}{\pi_0} = \sum_{k=0}^{c-1} \frac{(c\rho)^k}{k!} + \frac{(c\rho)^c}{c!}\frac{1}{1-\rho},$$

which is unfortunately as simple as it gets.

14.13 Proposition (M/M/c Queue Stationary Distribution)
For the M/M/c queue with arrival rate λ, c server with each having service rate μ and utilization $\rho = \lambda/(c\mu) < 1$ the stationary probabilities are

$$\begin{aligned} \pi_0 &= \left[\sum_{k=0}^{c-1} \frac{(c\rho)^k}{k!} + \frac{(c\rho)^c}{c!}\frac{1}{1-\rho}\right]^{-1}, \\ \pi_n &= \pi_0 \frac{(c\rho)^n}{n!}, \qquad \text{for } n = 1,\ldots,c-1, \\ \pi_n &= \pi_0 \frac{(c\rho)^c}{c!}, \qquad \text{for } n = c, c+1, \ldots. \end{aligned}$$

Birth-and-Death Queues

Let us then consider briefly the general birth-and-death queue. This gives us a framework that can be applied to, e.g., queues with finite capacity or/and many servers. Consequently, the results of the previous sections of this lecture follow as corollaries of the results given in this section.

14.14 Definition (Birth-and-Death Queue)
A Markovian queue with state-dependent arrival rates $\boldsymbol{\lambda} = [\lambda_n]_{n\in\mathbb{N}}$ and service rates $\boldsymbol{\mu} = [\mu_n]_{n\in\mathbb{N}}$ is called the **birth-and-death queue**.

The stationary probabilities for the birth-and-death queues, the **balance of birth and death** if you will, can be found in the normal way by using the **flux in = flux out principle**. The flux-in to the state n is $\lambda_{n-1}\pi_{n-1}$ births from the state $n-1$ plus the $\mu_{n+1}\pi_{n+1}$ deaths from the state $n+1$. The flux-out from the state n is $(\lambda_n+\mu_n)\pi_n$: $\lambda_n\pi_n$ births to the state $n+1$ plus $\mu_n\pi_n$ deaths to the state $n-1$. Consequently, the **full balance equation** for the birth-and-death queues takes the form

$$\lambda_{n-1}\pi_{n-1}+\mu_{n+1}\pi_{n+1} \quad = \quad (\lambda_n+\mu_n)\pi_n.$$

This birth-and-death balance can be written in a more simple form as the **detailed balance equation**

$$\lambda_n\pi_n \quad = \quad \mu_{n+1}\pi_{n+1}.$$

The interpretation of this form is analogous to the M/M/1 case: The left-hand-side is the flux from the state n to the state $n+1$ and the right-hand-side is the flux from the state $n+1$ to the state n.

Denoting

$$\rho_n \quad = \quad \frac{\lambda_n}{\mu_{n+1}},$$

we can write the detailed balance equation above in a form of a simple recursion

$$\pi_{n+1} \quad = \quad \rho_n\pi_n,$$

This recursion cannot be solved in closed from in this generality. Thus, we have to do with the following theorem.

14.15 Theorem (Stationary Birth-And-Death Queue)
The stationary distribution, when it exists, of a birth-and-death queue with parameters $\boldsymbol{\lambda}=[\lambda_n]_{n\in\mathbb{N}}$ and $\boldsymbol{\mu}=[\mu_n]_{n\in\mathbb{N}}$ is given by

$$\pi_n \quad = \quad \frac{\prod_{k=0}^{n-1}\rho_k}{\sum_{m=0}^{\infty}\prod_{k=0}^{m-1}\rho_k}, \qquad n\in\mathbb{N},$$

where

$$\rho_k \quad = \quad \frac{\lambda_k}{\mu_{k+1}}.$$

A sufficient condition for the existence of the stationary distribution is $\rho_k\le\rho<1$ for all $k\in\mathbb{N}$.

14.16 Remark (Empty Sums and Products)
In the above and below, we use the convenient standard interpretations that empty sum is 0 and empty product is 1.

The characteristics λ, ℓ and w in the Little's law can also be given for the general birth-and-death queue in terms of the parameters $\boldsymbol{\lambda} = [\lambda_n]_{n\in\mathbb{N}}$ and $\boldsymbol{\mu} = [\mu_n]_{n\in\mathbb{N}}$. The input rate for the birth-and-death queue is

$$\begin{aligned}\lambda &= \sum_{n=0}^{\infty} \lambda_n \pi_n \\ &= \frac{\sum_{n=0}^{\infty} \lambda_n \prod_{k=0}^{n-1} \rho_k}{\sum_{m=0}^{\infty} \prod_{k=0}^{m-1} \rho_k}.\end{aligned}$$

The average load is

$$\begin{aligned}\ell &= \sum_{n=0}^{\infty} n \pi_n \\ &= \frac{\sum_{n=0}^{\infty} n \prod_{k=0}^{n-1} \rho_k}{\sum_{m=0}^{\infty} \prod_{k=0}^{m-1} \rho_k},\end{aligned}$$

and the average waiting time is

$$\begin{aligned}w &= \frac{\ell}{\lambda} \\ &= \frac{\sum_{n=0}^{\infty} n \prod_{k=0}^{n-1} \rho_k}{\sum_{m=0}^{\infty} \prod_{k=0}^{m-1} \rho_k} \Bigg/ \frac{\sum_{n=0}^{\infty} \lambda_n \prod_{k=0}^{n-1} \rho_k}{\sum_{m=0}^{\infty} \prod_{k=0}^{m-1} \rho_k} \\ &= \frac{\sum_{n=0}^{\infty} n \prod_{k=0}^{n-1} \rho_k}{\sum_{n=0}^{\infty} \lambda_n}.\end{aligned}$$

Unfortunately, in the general case of birth-and-death process the formulas above cannot be simplified, at least not much. Fortunately, the formulas above are easy enough to implement in any reasonable programming language or mathematical software (probably even with unreasonable ones like Excel; I have not tried, for I am not a masochist).

Exercises

14.1 Exercise
A single-server Markovian queue with unlimited capacity is fed by 3 customers per minute on average, and the service times are, on average, 10 seconds.

(a) What is the probability that a customer arriving to the system finds it empty?
(b) What is the average time a customer spends in the system?

14.2 Exercise
A single-server Markovian queue with unlimited capacity is fed by 5 customers per minute on average, and it can server, on average, 8 customers per minute.

(a) What is the probability that a customer arriving to the system finds its queue empty?
(b) How many customers are the, on average, in the queue?

14.3 Exercise
A Markovian queue with capacity 12 customers is fed by 10 customers per minute on average, and the service times are, on average, 5 seconds. If the queue is full, then the arriving customer will be denied service.

(a) What is the probability that a customer arriving to the system will be denied service?
(b) What is the probability that a customer arriving to the system will get service without waiting?

14.4 Exercise
The local hardware store has two clerks. The clerks can either work together or separately. If they work together, they can serve at the rate of 4 customers per hour. If they work separately, they can serve each at the rate of 2 customers per hour.

(a) The manager of the local hardware store wants to maximize the service rate. Should the clerks work together or separately?
(b) The manager wants to minimize the probability that an arriving customer has to wait in line. Should the clerks work together or separately?

14.5 Exercise (Focus Problem)
You are queuing in the Ministry of Love to get access to Room 101. There are 5 clerks serving the customers, and a single queue using the fist-in-first-out queuing policy. At the moment all the clerks are a busy and there are 12 customers in line in front of you. You have been waiting for 20 minutes. During that time you have observed the service times of 3 customers. They were 1 minute, 5 minutes and 18 minutes.

(i) What is the probability that your total waiting time plus service time in the Ministry of Love will exceed 1 hour?

(ii) Suppose you have now waited 25 minutes and there are 11 customers in front of you and you have recorded an extra service time which was 10 seconds. What is now the probability that your total waiting time plus service time in the Ministry of Love will exceed 1 hour?

Part V

Appendix

Appendix A

Exam Questions

There will be four questions in the exam that will last 3 hours. The questions are chosen randomly from the following list of problems by using the Sottinen$(n, p, q, \rho, \lambda, \mu, \ddot{o})$ distribution. Each problem and each part has equal weight in grading. For probabilistic reasons there is some slight repetition in the problems. The exam will be closed-book and only pocket calculators will be allowed, so no Octave in the exam.

Conditioning Tricks

A.1 Problem
Consider a branching process with offspring distribution

$$\mathbf{p} = [0.30 \quad 0.15 \quad 0.50 \quad 0.05].$$

Calculate

(a) The mean of the 7^{th} generation,
(b) The variance of the 7^{th} generation,

A.2 Problem
Consider a branching process with offspring distribution

$$\mathbf{p} = [0.30 \quad 0.19 \quad 0.50 \quad 0.00 \quad 0.01].$$

Calculate

(a) The mean of the 6^{th} generation,
(b) The variance of the 6^{th} generation,

A.3 Problem
Explain briefly

(a) the Adam's law,
(b) the Eve's law.

A.4 Problem
Explain briefly

(a) the law of total probability,
(b) the law of total variance.

A.5 Problem
Consider a branching process with offspring distribution

$$\mathbf{p} = [0.25 \quad 0.00 \quad 0.75].$$

Calculate the distributions of

(a) the second generation,
(b) the third generation.

A.6 Problem
Consider a branching process with offspring distribution

$$\mathbf{p} = [0.80 \quad 0.00 \quad 0.20].$$

Calculate the distributions of

(a) the second generation,
(b) the third generation.

A.7 Problem
Calculate the probability generating function of $\mathbb{N}$-valued random variables with probability mass functions

(a)

$$\mathbb{P}[X=x] \;=\; \begin{cases} 0.20 & \text{if} \quad x=0, \\ 0.00 & \text{if} \quad x=1, \\ 0.80 & \text{if} \quad x=2. \end{cases}$$

(b)

$$\mathbb{P}[X=x] \;=\; \mathrm{e}^{-3}\frac{3^x}{x!}, \quad x \in \mathbb{N}.$$

A.8 Problem
Calculate the probability generating function of $\mathbb{N}$-valued random variables with probability mass functions

(a)

$$\mathbb{P}[X=x] \;=\; \begin{cases} 0.90 & \text{if} \quad x=0, \\ 0.05 & \text{if} \quad x=1, \\ 0.00 & \text{if} \quad x=2, \\ 0.04 & \text{if} \quad x=3, \\ 0.01 & \text{if} \quad x=4. \end{cases}$$

(b)

$$\mathbb{P}[X=x] \;=\; \binom{4}{x} 0.1^x \times 0.9^{4-x}, \quad x=0,1,2,3,4.$$

A.9 Problem
Calculate the ultimate extinction probabilities for the branching processes having offspring distributions

(a) $\mathbf{p} \;=\; [0.20 \;\; 0.30 \;\; 0.50]$,
(b) $\mathbf{p} \;=\; [0.75 \;\; 0.15 \;\; 0.05 \;\; 0.00 \;\; 0.05]$.

A.10 Problem
Calculate the ultimate extinction probabilities for the branching processes having offspring distributions

(a) $\mathbf{p} = [0.20 \;\; 0.10 \;\; 0.70]$,
(b) $\mathbf{p} = [0.00 \;\; 0.85 \;\; 0.05 \;\; 0.00 \;\; 0.10]$.

Some Interesting Probability Distributions

A.11 Problem
Let X be binomially distributed with parameters 3 and 0.2. Let Y be binomially distributed with parameters 2 and 0.2. Let X and Y be independent. Calculate the probabilities

(a) $\mathbb{P}[X + Y = 3]$,
(b) $\mathbb{P}[Y = 2 \mid X = 0]$.

A.12 Problem
Let X be binomially distributed with parameters 6 and 0.5. Let Y be binomially distributed with parameters 3 and 0.5. Let X and Y be independent. Calculate the probabilities

(a) $\mathbb{P}[X + Y = 5]$,
(b) $\mathbb{P}[Y = 3 \mid X = 0]$.

A.13 Problem
There are 3 clients in a teletraffic system sharing a common link. Each client is idle with probability 50 %. When the clients transmit, they transmit with a constant rate of 1 Mb/s. How big should the link capacity be to provide 95 % quality-of-service

(a) from the system's point of view,
(b) from the clients' point of view?

A.14 Problem
There are 4 clients in a teletraffic system sharing a common link. Each client is idle with probability 90 %. When the clients transmit, they transmit with a constant rate of 3 Mb/s. How big should the link capacity be to provide 95 % quality-of-service

(a) from the system's point of view,
(b) from the clients' point of view?

A.15 Problem
Let X be Poisson distributed with parameter 2. Let Y be Poisson distributed with parameter 4. Suppose that X and Y are independent. Calculate

(a) $\mathbb{P}[X+Y=2]$,
(b) $\mathbb{P}[Y=0 \mid X+Y=1]$.

A.16 Problem
Let X and Y be independent and Poisson distributed with parameter 3. Calculate

(a) $\mathbb{P}[X+Y=1]$,
(b) $\mathbb{P}[Y=0 \mid X+Y=1]$.

A.17 Problem
The Lake Diarrhea has, on average, 0.1 Malus particles per one liter. Magnus Flatus lives on the shore of the Lake Diarrhea. He drinks daily 2 liters of water from the Lake Diarrhea. The lethal daily intake of Malus particles is 3.

(a) What is the probability that Magnus Flatus will have a lethal intake of Malus particles in a given day?
(b) What is the probability that Magnus Flatus will have a lethal intake of Malus particles during a given year?

A.18 Problem
The Lake Diarrhea has, on average, 0.05 Malus particles per one liter. Magnus Flatus lives on the shore of the Lake Diarrhea. He drinks daily 1 liter of water from the Lake Diarrhea. The lethal daily intake of Malus particles is 4.

(a) What is the probability that Magnus Flatus will have a lethal intake of Malus particles in a given day?
(b) What is the probability that Magnus Flatus will have a lethal intake of Malus particles during a 40 year period?

A.19 Problem
Let T be exponentially distributed with parameter 1.5. Calculate

(a) $\mathbb{P}[T \leq 1]$
(b) $\mathbb{P}[1 \leq T \leq 2 \mid T \geq 1]$.

A.20 Problem
Let T be exponentially distributed with mean 0.667. Calculate

(a) $\mathbb{P}[T \leq 1]$
(b) $\mathbb{P}[0.5 \leq T \leq 1 \mid T \geq 0.2]$.

A.21 Problem
Let T_1 and T_2 be independent exponentially distributed random variables with parameter 2. Calculate

(a) $\mathbb{P}[0.2 \leq T_1 + T_2 \leq 0.4]$
(b) $\mathbb{P}[0.5 \leq T_1 \leq 1 \mid T_2 \geq 0.5]$.

A.22 Problem
Let T_1 and T_2 be independent exponentially distributed random variables with mean $1/2$. Calculate

(a) $\mathbb{P}[0.2 \leq T_1 + T_2 \leq 0.8]$
(b) $\mathbb{P}[T_1 \leq T_2]$.

A.23 Problem
Explain briefly what is

(a) the exponential distribution,
(b) the Erlang distribution.

A.24 Problem
Explain briefly

(a) how the Erlang distribution is related to the exponential distribution,
(b) and where the Erlang distribution gets its name?

A.25 Problem
Explain briefly

(a) the law of small numbers,
(b) the central limit theorem.

A.26 Problem
Let $X_1, X_2, \ldots, X_{60}$ be independent Bernoulli trials each having success probability $p = 1/4$. Let $S = X_1 + \cdots + X_{60}$.

(a) What is the law of small numbers approximation of S?
(b) What is the central limit theorem approximation of S?

Stochastic Processes

A.27 Problem
Consider a time-homogeneous Markov chain X_n, $n \in \mathbb{N}$, with state space $\mathbb{S} = \{0, 1, 2, 3\}$ and transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0.95 & 0.00 & 0.05 & 0.00 \\ 0.30 & 0.10 & 0.00 & 0.60 \\ 0.50 & 0.45 & 0.00 & 0.05 \\ 0.20 & 0.10 & 0.10 & 0.60 \end{bmatrix}.$$

Calculate the transition probabilities

(a) $\mathbb{P}[X_1 = 2 \,|\, X_0 = 1]$,
(b) $\mathbb{P}[X_8 = 3 \,|\, X_6 = 1]$.

A.28 Problem
Consider a time-homogeneous Markov chain X_n, $n \in \mathbb{N}$, with state space $\mathbb{S} = \{0,1,2,3\}$ and transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0.90 & 0.00 & 0.10 & 0.00 \\ 0.30 & 0.60 & 0.00 & 0.10 \\ 0.50 & 0.40 & 0.00 & 0.10 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}.$$

Calculate the transition probabilities

(a) $\mathbb{P}[X_1 = 1 \text{ or } X_1 = 2 \,|\, X_0 = 1]$,
(b) $\mathbb{P}[X_2 = 0 \,|\, X_0 = 0]$.

A.29 Problem
Consider a time-homogeneous Markov chain X_n, $n \in \mathbb{N}$, with state space $\mathbb{S} = \{0,1,2\}$ and transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0.95 & 0.00 & 0.05 \\ 0.30 & 0.70 & 0.00 \\ 0.50 & 0.50 & 0.00 \end{bmatrix}.$$

Suppose the initial distribution is

$$\mathbf{p} = [0.10 \;\; 0.80 \;\; 0.10].$$

Calculate the probabilities

(a) $\mathbb{P}[X_0 = 0, X_1 = 2]$,
(b) $\mathbb{P}[X_3 = X_2 = X_1]$.

A.30 Problem
Consider a time-homogeneous Markov chain X_n, $n \in \mathbb{N}$, with state space $\mathbb{S} = \{0,1,2\}$ and transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0.95 & 0.00 & 0.05 \\ 0.30 & 0.70 & 0.00 \\ 0.50 & 0.00 & 0.50 \end{bmatrix}.$$

Suppose the initial distribution is

$$\mathbf{p} = [0.20 \;\; 0.00 \;\; 0.80].$$

Calculate the probabilities

(a) $\mathbb{P}[X_1 = 0]$,
(b) $\mathbb{P}[X_2 = X_1 = X_0]$.

A.31 Problem
Find out which states of the following Markov chains are transient and which are recurrent (null or positive). Also, find out the period of each state.

(a)
$$\begin{bmatrix} 0.1 & 0.9 & 0.0 \\ 0.0 & 0.2 & 0.8 \\ 1.0 & 0.0 & 0.0 \end{bmatrix}$$

(b)
$$\begin{bmatrix} 0.0 & 0.0 & 1.0 \\ 0.0 & 0.2 & 0.8 \\ 0.5 & 0.0 & 0.5 \end{bmatrix}$$

A.32 Problem
Find out which states of the following Markov chains are transient and which are recurrent (null or positive). Also, find out the period of each state.

(a)
$$\begin{bmatrix} 0.1 & 0.9 & 0.0 \\ 0.0 & 0.0 & 1.0 \\ 0.5 & 0.0 & 0.5 \end{bmatrix}$$

(b)
$$\begin{bmatrix} 0.0 & 1.0 & 0.0 \\ 1.0 & 0.0 & 0.0 \\ 0.1 & 0.0 & 0.9 \end{bmatrix}$$

A.33 Problem
Suppose that the probability whether it rains tomorrow depends only on whether it has rained today. Let X_n, $n \in \mathbb{N}$, be the Markov chain modeling the weather: $X_n = 0$ if it rains at day n and $X_n = 1$ if it does not rain at day n. Let

$$\mathbf{P} = \begin{bmatrix} 0.95 & 0.05 \\ 0.30 & 0.70 \end{bmatrix}$$

be the transition probability matrix of X_n, $n \in \mathbb{N}$.

(a) Suppose that on Monday it rains with probability 0.25. What is the probability that it rains on Wednesday?
(b) In the long run, how may rainy and non-rainy days would you expect in this model?

A.34 Problem
Suppose that the probability whether it rains tomorrow depends only on whether it has rained today. Let X_n, $n \in \mathbb{N}$, be the Markov chain modeling the weather: $X_n = 0$ if it rains at day n and $X_n = 1$ if it does not rain at day n. Let

$$\mathbf{P} = \begin{bmatrix} 0.90 & 0.10 \\ 0.50 & 0.50 \end{bmatrix}$$

be the transition probability matrix of X_n, $n \in \mathbb{N}$.

(a) Suppose that on Monday it rains. What is the probability that it rains on Thursday?
(b) In the long run, how may rainy and non-rainy days would you expect in this model?

A.35 Problem
Explain briefly

(a) what is Cesàro mean,
(b) and how is it related to the long-run behavior of Markov chains.

A.36 Problem
Explain briefly

(a) what ergodicity means in connection to Markov chains,
(b) and where aperiodicity is needed for the ergodic theorem.

A.37 Problem
Explain briefly

(a) what is the balance equation,
(b) and how it is related to limiting probabilities.

A.38 Problem
Explain briefly

(a) what is the balance equation,
(b) and how it is related to eigenvalues and eigenvectors.

A.39 Problem
Mr. S. and Ms. L. are playing a coin-tossing game with a fair coin. If the coin lands on heads, Ms. L. will give Mr. S. one euro. If the coin lands on tails, Mr. S. will give Ms. L. one Euro. Mr. S. has capital of 500 euros and Ms. L. has capital of 50 euros. The game is played until either Mr. S. or Ms. L. loses his/her capital.

(a) What is the probability that Mr. S. wins the game?
(b) Suppose the coin is not fair. Let p be the probability that Mr. S. wins. What should p be so that the game is fair, i.e., the probability for Mr. S. to win is $1/2$?

A.40 Problem
A symmetric random walk X_n, $n \in \mathbb{N}$, with state-space $\mathbb{S} = \mathbb{Z}$ start at point $X_0 = c$. Let $l < c < u$

(a) What is the probability that the random walk hits the boundary u before it hits the boundary l?
(b) What is the probability that the random walk does not hit either of the boundaries l or u?

A.41 Problem
Let $N(t)$, $t \geq 0$, be a Poisson process with intensity 2. Calculate

(a) $\mathbb{P}[N(3) = 5 \,|\, N(1) = 4]$,
(b) $\mathbb{P}[N(3) = 5\,, N(2) = 1\,, N(1) = 1]$.

A.42 Problem
Let $N(t)$, $t \geq 0$, be a Poisson process with intensity 3. Calculate

(a) $\mathbb{P}[N(1) = 0]$,
(b) $\mathbb{P}[N(3) = 5\,, N(2) = 1 \,|\, N(1) = 0]$.

A.43 Problem
Explain briefly how

(a) the Poisson process and the exponential distribution are connected,
(b) and the Poisson process and the Erlang distribution are connected.

A.44 Problem
Explain briefly how one can construct a Poisson process by using

(a) independent exponentially distributed random variables,
(b) or a biased coin with the law of small numbers.

Queueing

A.45 Problem
(a) A fast food hamburger restaurant uses 3 500 kilograms of hamburger mince each week. The manager of the restaurant wants to ensure that the meat is always fresh i.e. the meat should be no more than two days old on average when used. How much hamburger mince should be kept in the refrigerator as inventory?
(b) The Acme Australia insurance company processes 12 000 insurance claims per year. At any point in time, there are 480 insurance claims in head office in various phases of processing. Assuming that the office works 50 weeks per year, find out the average processing time of an insurance claim in days.

A.46 Problem
(a) Don Guido has on average 150 cases of wine his basement. Don Guido consumes 25 cases per year. How long does Don Guido hold each case?
(b) A hardware vendor manufactures 300 million euros worth of PCs per year. On average, the company has 45 million euros in accounts receivable. How much time elapses between invoicing and payment?

A.47 Problem
Explain briefly the following queueing concepts:

(a) Palm probability.
(b) PASTA.

A.48 Problem
Explain briefly the following queueing concepts:

(a) Little's law.
(b) Waiting paradox.

A.49 Problem
A Markovian queue with unlimited capacity is fed by 2 customers per minute on average, and the service times are, on average, 10 seconds.

(a) What is the probability that a customer arriving to the system finds it empty?
(b) What is the average time a customer spends in the system?

A.50 Problem
A Markovian queue with unlimited capacity is fed by 3 customers per minute on average, and the service times are, on average, 15 seconds.

(a) What is the probability that a customer arriving to the system finds it busy?
(b) What is the average time a customer spends queueing in the system?

A.51 Problem
A Markovian queue is fed clients with intensity 10 per minute.

(a) The clients are served with intensity 20 per minute. What is the average workload of the system?
(b) There are, on average, 1 500 clients in the system. What is the average time a client spends in the system?

A.52 Problem
A Markovian queue serves client with intensity 50 per minute.

(a) There are on average 2 000 clients in the system. What is the rate of arrival of new clients?
(b) On average, 40 clients come to the system each minute. What is the average number of clients in the system?

A.53 Problem
A Markovian queue with capacity 10 customers is fed by 8 customers per minute on average, and the service times are, on average, 5 seconds. If the queue is full, then the arriving customer will be denied service.

(a) What is the probability that a customer arriving to the system will be denied service?
(b) What is the probability that a customer arriving to the system will get service without waiting?

A.54 Problem
A Markovian queue with capacity 20 customers is fed by 10 customers per minute on average, and the service times are, on average, 2 seconds. If the queue is full, then the arriving customer will be denied service.

(a) What is the proportion of customers arriving to the system that will be denied service?
(b) What is the probability that a customer arriving to the system will have to wait to be served?

Index