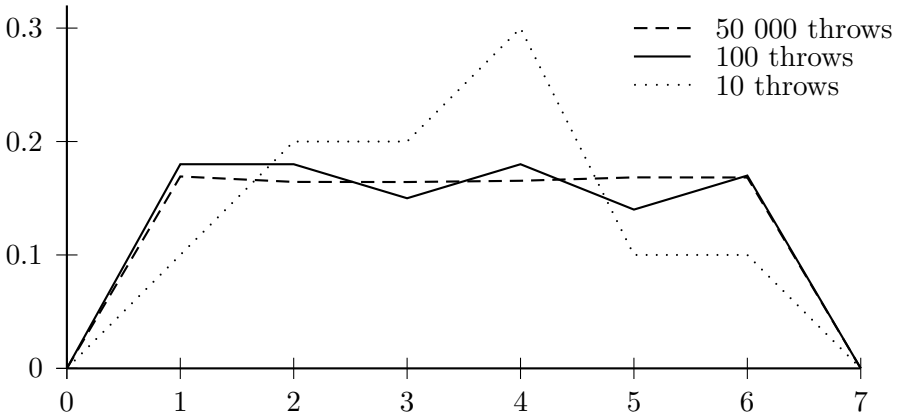
Sampling and simulating

A la fin de ce chapitre, vous devez être capable de :
- concevoir, mettre en œuvre et exploiter des simulations de situations concrètes à l'aide du tableur ou d'une calculatrice ;
- exploiter et faire une analyse critique d'un résultat d'échantillonnage.

16.1 Fluctuations when throwing a die

The frequency polygons below were obtained by forcing the pupils in a class to throw a a fair die 10 times, 100 times, and 50000 times.



1. Read on each graph the approximate frequency of the side 4.
2. Draw on the graph a red horizontal line representing the theoretical probability of getting one specific side.
3. What do you notice about the distance between each graph and the red line ?
4. Write a sentence about the phenomenon showcased by this set of frequency polygons.

16.2 The US 2008 election

In 2008, the American electors had to choose between the Republican John McCain and the Democrat Barack Obama. Surveys were organised by both parties to **estimate** the proportion of electors who wanted to vote for each candidate. As it's impossible to gather the opinions of all the electors, surveys are carried over small parts of the population, called **samples**. We will consider that samples are built randomly.

Part A – Sampling fluctuation

1. Over a sample of 900 electors, 497 declared that they wanted to vote for Obama. Compute the percentage of potential Obama electors in this sample.
2. Ten other surveys were organized over the same period. The size of each sample and the number of potential Obama electors are given in the table below.

Survey	1	2	3	4	5	6	7	8	9	10
Size	895	873	900	885	899	842	878	900	897	892
Obama electors	462	493	501	437	467	447	468	495	488	478

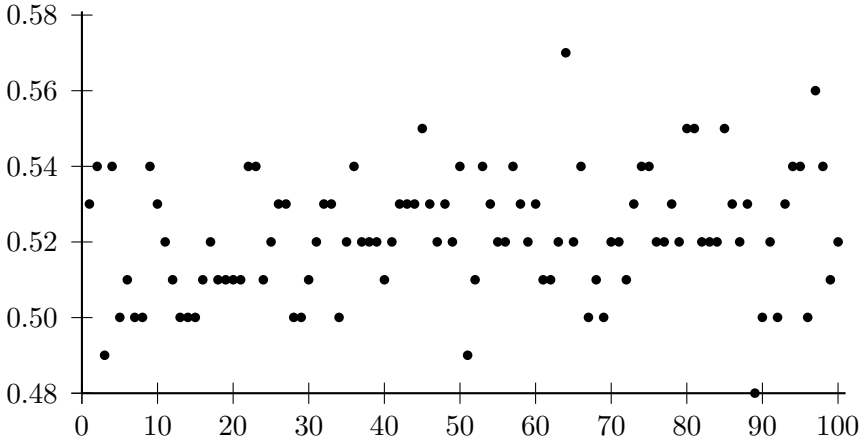
   a. Compute the percentage of potential Obama electors in each sample. Round the answers to 2DP.

b. If McCain had only known about the 4th survey, what could he have deduced ?

c. Can you deduce from these surveys the actual percentage of Obama voters ?

Part B – Margin of error

The chances that a sample will yield the true value in the whole population are very small. Furthermore, there may be important differences between the percentages in different samples. This phenomenon is known as **sampling fluctuation**.
To illustrate this, the results of one hundred surveys were collected, each one over a population of 900 people. The scatterplot below shows the percentage of potential Obama electors in each survey.



1. It turns out that, on Election day, Obama won with 53% of the votes. On the scatterplot, show the proportion p of Obama electors in the whole population with a horizontal red line. How many simulated surveys gave that exact value ?

2. **a.** The value $m = \frac{1}{\sqrt{n}}$, where n is the size of a sample, is called the **margin of error at 95% confidence** for that sample. Compute this value to 3DP.

b. On the graph, show the values $p - m$ and $p + m$ with two horizontal blue lines.

c. How many surveys gave a percentage included in the interval $[p - m; p + m]$, called **fluctuation interval at 95% confidence** ?

d. Is the answer to the previous question consistent with the name of the interval ?

16.3 The French lottery and odd numbers

The principles of the French National Lottery (Loto) are fairly simple. Each player picks six numbers (plus one, that we won't consider in this exercise) between 1 and 49. On lottery day, 6 over 49 balls with numbers from 1 to 49 are randomly drawn from a machine. The balls are not put back in the machine, so the same number cannot appear twice in a drawing. The order in which the balls are drawn is irrelevant.
Among the numbers from 1 to 49, there are 25 odd numbers and 24 even numbers.

Part A – Drawing a single number

In this part, we consider the random experiment that consists in drawing a single ball from the 49 in the machine.

1. What is the probability of the drawn number being odd ? Give the result as an irreducible fraction and as an approximate value to 2DP.

2. Fifty samples, each made of $n = 100$ independant drawings of a ball were simulated with a computer. For each sample, the proportion of odd numbers was computed. The results of these fifty samples of size 100 are given below.

0.44	0.52	0.50	0.44	0.51	0.41	0.44	0.40	0.57	0.50
0.51	0.43	0.59	0.46	0.55	0.35	0.55	0.43	0.53	0.53
0.45	0.42	0.47	0.48	0.50	0.45	0.48	0.47	0.46	0.57
0.52	0.55	0.53	0.46	0.45	0.44	0.45	0.48	0.51	0.46
0.55	0.48	0.43	0.51	0.49	0.38	0.52	0.40	0.50	0.46

a. How many samples showed a proportion equal to the theoretical value to 2DP ?

b. Compute the fluctuation interval at 95% confidence.

c. How many samples showed a proportion inside the margin of error ?

d. Can you find a margin of error at 98% confidence ?

Part B – Drawing six numbers

In this second part, we consider the random experiment that consists of drawing successively six balls, without putting them back in the machine. It can be proven that in each drawing of six numbers, there is an average of 3.0612 odd numbers, so a proportion $q = \frac{3.0612}{6} \approx 0.51$, or approximately 51%.
Fifty samples, each made of $n = 100$ independant drawings of six succesive balls were simulated with a computer. For each sample, the proportion of odd numbers was computed. The results of these fifty samples of size 100 are given below.

0.527	0.475	0.500	0.522	0.558	0.518	0.510	0.518	0.550	0.607
0.515	0.468	0.517	0.485	0.498	0.473	0.505	0.507	0.492	0.498
0.508	0.408	0.563	0.612	0.542	0.497	0.508	0.498	0.500	0.535
0.498	0.508	0.525	0.478	0.517	0.528	0.492	0.487	0.535	0.523
0.512	0.543	0.522	0.482	0.530	0.478	0.508	0.532	0.528	0.527

1. Compute the fluctuation interval at 95% confidence.

2. How many samples showed a proportion inside the interval ?

Part C – Probabilities on the number of odd numbers

The table below shows the probabilities of drawing k odd numbers among the six, for k from 0 to 6. Values have been rounded to 3DP.

Odd numbers	0	1	2	3	4	5	6
Probability	0.010	0.076	0.228	0.333	0.250	0.091	0.013

For each of the following sentences, say if it's true or false. Justify each answer with a computation or an explanation.

1. There are more chances to draw 4 odd numbers or more than 2 odd numbers or less.

2. There are more than 90% chances to draw at least 2 odd numbers.

3. There are as many chances to draw exactly 3 odd numbers than exactly 3 even numbers.

4. There are 50% chances to draw as many odd numbers as even numbers.

5. There are more chances to draw no even number than to draw no odd number.

6. There are as many chances to draw at least 3 odd numbers than at least 3 even numbers.

7. It's a good strategy to play only odd numbers.

16.4 A biaised four-sided die

A role-playing enthusiast has bought a new die with four sides. She notices that there is a dent on the number four vertex and fears that it may make the die biased.

1. What should be the probability p of getting a four if the die was really balanced ?
2. She throws the die 50 times and gets 11 times the number four.
   a. Compute the proportion of occurences of the number four in the sample.
   b. Compare the proportion in the sample to the value probability you gave in question 1. What do you conclude about the die ?
   c. Compute the margin of error and the fluctuation interval at 95% confidence for the probability p and a sample of 50 throws.
   d. Is your previous conclusion still the same ?
3. She still isn't convinced and throws the die 250 times. She gets 55 times the number four. Answer the previous questions with this new sample.
4. While she's satisfied with the results of her expriment, a friend tells her that 250 throws are not enough to decide if the die is biased. She then throws the die 2000 times and counts 440 occurences of the number four.
   Answer the previous questions with this sample.
5. What do you notice about the margin of error when the size of the sample increases ? What impact can it have on a test like this ?

16.5 Male-female parity or not ?

Two companies A and B are hiring people in a region where there are as many men as women. By law, they are bound to male-female parity. In company A, there are 100 employees and 43 of them are women. In company B there are $2,500$ employees with $1,150$ women.

1. a. Compute the proportion of women for each company.
   b. What do you think of the way each company respect parity ?
2. a. If parity was respected, what should be the proportion of women ?
   b. Compute the fluctuation interval at 95% confidence for each company.
   c. Does the previous result confirm your answer to question 1.b ? Explain.

16.6 Using margins of error to make decisions

Part A – Accepting or rejecting an assumption

It is known that in the French population, 26% are allergic to pollen. The sanitary services in a city suspect that the proportion is more important in their town thand elsewhere in France. To check if this is true, they study a sample of 400 people, and observe that 130 suffer from that allergy.

1. Compute the fluctuation interval at 95% confidence.
2. What is the frequency of allergic individuals in this sample ?
3. Does this result confirm the suspicions of the sanitary services ?

Part B – A car factory

In a car factory, a control is done for flaws of the type "grainy spots on the hood". Normally, 20% of the vehicles present this kind of flaws. While controlling a random sample of 50 vehicles produced in the same week, it is seen that 13 vehicles have it. Should it be a matter of concern ?

Part C – Parity in French Region councils

After the 2004 regional elections in France, the repartition between women and men in four regional councils was as follows. We consider that these councils are random samples of the local politician population in each region.

	Men	Women	Total
Burgundy	32	25	57
Brittany	38	47	85
Rhône-Alpes	81	76	157
Île-de-France	103	106	209

1. Supposing that parity between men and women is real in a regional council, what should be the percentage of women in that council ?
2. Compute the fluctuation interval at 95% confidence for the proportion of women in each council.
3. What do you think of the parity between men and women in the local politician population of each of these regions ?

Part D – Rodrigo Partida's case

In 1970, the Mexican-American Rodrigo Partida was sentenced to eight years of prison. He appealed to the judgment contending that he was denied due process and equal protection of law because the grand jury of Hidalgo County, Texas, which indicted him, was unconstitutionally underrepresented by Mexican-Americans. He introduced evidence that in 1970, the total population of Hidalgo County was 181,535 persons of which 143,611, or approximately 79.2% were persons of Spanish language or Spanish surname. Next, he presented evidence showing the composition of the grand jury lists over a period of ten years prior to and including the term of court in which the indictment against him was returned. Of the 870 persons selected for grand jury duty, only 39.0% were Mexican-Americans. If you were a judge in the court of appeals, how would you react to these allegations ?

16.7 Lime or orange Tic Tac

In this exercise, we will try to answer to an existential question :

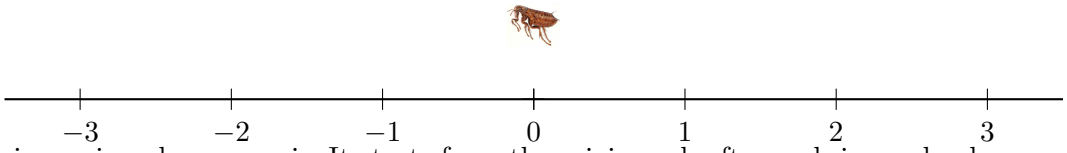
Is there the same proportion of each flavour in a box of lime or orange Tic Tac ?

To do so, each pupil in the class will be given a box of candies and use it as a sample of the whole lime or orange Tic Tac production. To avoid biasing the experiment, it's important not to eat a single candy before the end of the exercise.

1. Count the number of candies in your box. What does it tell you about your sample ?
2. Assume that the proportions of each flavour are the same. Compute the fluctuation interval at 95% confidence for the proportion of lime candies.
3. Count the number of lime candies in your box and compute the observed proportion.
4. What can you conclude from your sample ?
5. How many pupils in the class rejected the hypothesis that the proportions of each flavour are the same ?
6. Put the candies back in the box and/or eat them.

16.8 Write an algorithm where the input are a proportion to test and a sample size, and ouput are the boundaries of the fluctuation interval. Implement it with your calculator.

16.9 Random walks on an axis



A flea is moving along an axis. It starts from the origin and, after each jump, lands one unit to the right or one unit to the left, randomly and with the same probability.
A sequence of jumps is called a *walk*. For example, if the flea is always jumping to the right, the walk will be noted RRRR. If it alternates between right and left, the walk will be noted RLRL.

Part A – Simulations of 4-jumps walks

The "Random" or "Alea" function on your calculator delivers a random decimal number between 0 and 1.

1. Devise a method to simulate a 4-jumps walk using the "Random" function.
2. Simulate 25 walks and note the final position of the flea at the end of each walk.
3. What are the possible final positions on the axis ? Explain why some are impossible.
4. Count the number of walks for each final position and show the counts in a table.
5. Add a row to the previous table with the absolute frequencies for the whole class.
6. Compute the relative frequencies for the whole class.
7. Compute the average final position of the flea at the end of a 4-jumps walk.

Part B – An algorithm

A random walk can be described by the algorithm shown on the right-hand side, where the alea function delivers a random number in the interval $[0, 1[$. Parts of the algorithm have been omitted on purpose.

```
begin
    0 → x ;
    1 → i ;
    while i ⩽ 4 do
        if alea < 0.5 then
            ..... → x ;
        else
            ..... → x ;
        end if
        i + 1 → i ;
    end while
    Output: x ;
end
```

1. Explain the functions of the integers x and i in this algorithm.
2. Fill the two incomplete lines.
3. Here are the results of applying the algorithm once. What is the final position of the flea at the end of this walk ?

i		1	2	3	4
alea		0.37	0.01	0.93	0.11
x	0	1	2	1	2

4. Apply the algorithm to create five new walks, using the random function of your calculator and displaying all the steps of the algorithm like in the example of the previous question.
5. How would you change the algorithm to simulate a 30-jumps walk ?

Part C – Probabilistic study

In this part, we will use probabilities to study the situation and compare the theoretical results to the frequencies we found in part A.

1. Draw a tree to show all the possible 4-jumps walks. At the end of each branch, write the final position of the flea.

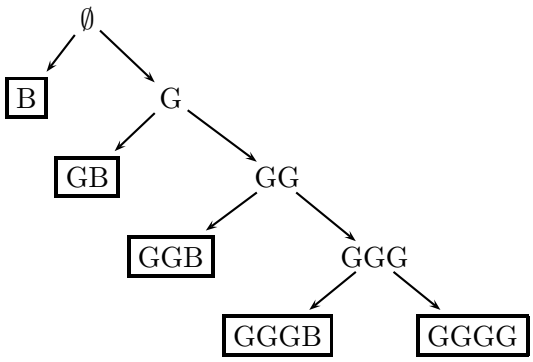
2. Use the tree to compute the probability of each final position and give the results in a probability table.
3. Compute the margin of error at 95% confidence for your sample of 25 random walks.
4. For each probability, count in the class how many samples of 25 walks gave a frequency within the margin of error.

16.10 A birth policy

A government has decided to impose a strict birth policy. Births in a family must stop as soon as a boy is born or after the birth of the fourth child.

We consider in this exercise that the probabilities of giving birth to a girl or a boy are equal and that each birth is independant from the previous births in the same family.

This birth policy can be represented as a tree, where the possible families are boxed.



Part A – Simulation and statistical approach

1. Discuss in the class to conjecture a value for the percentage of girls generated by this policy.
2. Devise a method to simulate the composition of a family with the calculator.
3. Simulate and write down the composition of 100 families. Count the number of children per family and show the results in a table with absolute and relative frequencies.
4. Compute the arithmetic mean m_4 and the median d_4 for the number of children per family in your sample of 100 families.
5. Compute the arithmetic mean M_4 and the median D_4 for all the families in the class.

Part B – An algorithm

This process can be described as an algorithm. The output is then a list of digits, with 0 representing a girl and 1 representing a boy.

1. Explain the functions of the whole numbers x and i in this algorithm.
2. Explain the condition "$x \neq 1$ and $i \leqslant 4$". Does it ensure that the algorithm will always stop ?
3. What is the function of the list L in this algorithm ?
4. Explain the notation $L(i)$.
5. Here are the results of applying the algorithm once. Apply the algorithm to get 5 families, displaying all the steps of the algorithm like in the example.

```
begin
    Clear list L ; 0 → x ; 1 → i ;
    while x ≠ 1 and i ⩽ 4 do
        if alea < 0.5 then
            0 → x ;
        else
            1 → x ;
        end if
        x → L(i) ;
        i + 1 → i ;
    end while
    Output: L ;
end
```

i		1	2	3
alea		0.37	0.01	0.93
x	0	0	0	1
L	()	(0)	(0, 0)	(0, 0, 1)

Part C – The percentage of girls

The aim of this part is to study the percentage of girls g induced by this birth policy, and therefore check the answer to the first question of part A. To do so, we will first use the simulations of part A, and then the probabilities of part C.

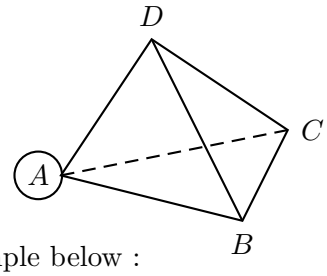
1. Use the value conjectured by the class at the beginning of the exercise to compute the fluctuation interval at 95% confidence in a sample of 100 families.
2. Compute the percentage of girls in your 100 simulated families. According to this result, would you reject the hypothesis formulated by the class ?
3. Answer the previous questions with the sample made of all the families simulated in the class. Is the conclusion the same ?

Part D – Probabilistic study

1. Copy the tree at the beginning of the exercise and add the probabilities.
2. Compute the probability of each type of family.
3. Show in a table the possible numbers of children and their probabilities. Are these probabilities consistent with the frequencies found at the end of part A ?
4. Use the table to compute the expected value for the number of children in a family.
5. Compute the expected values of the numbers of girls and boys in a family. Deduce the theoretical proportion of girls. Was your initial hypothesis correct ?

16.11 Random walks on a tetrahedron

An ant is walking on the edges of a tetrahedron $ABCD$, starting from vertex A. When it gets to a vertex, it chooses randomly the next edge it will walk on. The aim of this exercise is to study the time it will take for the ant to go back to vertex A, assuming that it walks along one edge in exactly 1 minute.



A walk will be noted as a succession of vertices, as in the example below :

$$A \to D \to B \to C \to A.$$

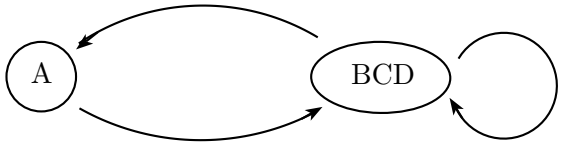
A walk will always start from A and stop as soon as the ant comes back to A.

Part A – Simulations

1. Devise a method to simulate a random walk.
2. Simulate 25 random walks and count the duration of each one. Gather the data in a table with the absolute frequency of each duration (from 1 to 20 minutes).
3. Explain the value in the column for 1 minute.
4. Is the duration necessarily less than 20 minutes ?
5. Find out the minimum, maximum, range, mean and median of this data.
6. Carry out the previous computations for all the simulated walks in the class.

Part B – Probabilistic study

We can illustrate the situation with a probabilistic graph. The vertices B, C, D, for which the walk doesn't end, have been gathered as a single vertex noted BCD. From vertex A, the only possibility is to go to BCD, while from BCD it's possible to go to A or stay in BCD.



1. Compute the probabilities to go to A and to stay in BCD when you are in BCD. Write theses probabilities on the edges of the graph.
2. Build a probability tree to illustrate a four-steps random walk.
3. Compute the probabilities of a 2-minutes, a 3-minutes and a 4-minutes walk.
4. Without adding a level to the tree, conjecture a value for the probability of a 5-minutes walk. Deduce the probability of a walk lasting 5 minutes or less.

16.12 Estimation in the 2008 US election

In this exercise, we look again at the US 2008 election. We will introduce a better method of estimation, based on the concept of margin of error. Instead of a simple point estimate, we will build for each sample a *confidence interval* whose diameter depends on the margin of error we allow.

We still note p the percentage of Obama electors in the whole population (so $p = 0.53$). Now, consider a sample fo size n yielding a point estimate f of p. We've seen in the previous part that the margin of error at 95% confidence is $m = \frac{1}{\sqrt{n}}$. Indeed, the probability of the point estimate f being in the interval $\left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}}\right]$ is approximately equal to 95%.

1. Translate the fact that f belongs to that interval with two inequalities.
2. Prove that the fact that f belongs to that interval is equivalent to the fact that p belongs to the interval $\left[f - \frac{1}{\sqrt{n}}, f + \frac{1}{\sqrt{n}}\right]$.

The interval $\left[f - \frac{1}{\sqrt{n}}, f + \frac{1}{\sqrt{n}}\right]$ is called a 95% *confidence interval*. Intuitively, this means that, knowing f and not p, we have a 5% risk of being wrong if we consider that p is in the interval. But, as p is fixed, it's not really correct to talk about probability. Once the confidence interval is determined, p is either in it or not !

3. Find the 95% confidence intervals for the surveys of exercise 1 part A.
4. How many surveys gave a confidence interval including the real value ?

16.13 The referendum on the European constitution

1. The French referendum on the Treaty establishing a Constitution for Europe was held on 29 May 2005 to decide whether France should ratify the proposed Constitution of the European Union. The question put to voters was : "Do you approve the bill authorising the ratification of the treaty establishing a Constitution for Europe ?"
   Below are given the results of some surveys carried out before the referendum.

Dates	Institute	Size	Proportion of « no »
18 and 19 March 2005	Ipsos	860	0.52
25 and 26 March 2005	Ipsos	944	0.54
1er and 2 April 2005	Ipsos	947	0.52
16 and 17 March 2005	CSA	802	0.51
23 March 2005	CSA	856	0.55
1 and 2 April 2005	Louis Harris	1004	0.54
31 March and 1 April 2005	IFOP	868	0.55
24 March 2005	IFOP	817	0.53

a. Find the 95% confidence interval for each survey.

b. The result was a victory for the "No" campaign, with 54.67%. A commentator then said that not many surveys had anticipated such a decisive result. What do you think of that opinion ?

2. The United Kingdom referendum was expected to take place in 2006. Following the rejection of the Constitution by voters in France in May 2005 and in the Netherlands in June 2005, the referendum was postponed indefinitely.
ICM research asked 1,000 voters in the third week of May 2005 “If there were a referendum tomorrow, would you vote for Britain to sign up to the European Constitution or not ?” : 57% said no. Find the 95% confidence interval for this survey. If you were a politician, what would you deduce from this ?

Homework #11 – The Euler Line

The aim of this homework is to study the famous Euler line on an example..
Consider an orthonormal coordinate graph $(O;\vec{i},\vec{j})$ with the points $A(-3,1)$, $B(5,1)$ and $C(-2,8)$.
We call A' the midpoint of $[BC]$ and B' the midpoint of $[AC]$.

1. For the following questions, you will need to find information on the internet or in maths books.

a. Who was Euler ? When did he live ?

b. What is the Euler line of a triangle ?

c. Draw the triangle ABC and its Euler line.

2. **a.** Compute the coordinates of A' and B'.

b. Find out equations of the medians through A and B in triangle ABC.

c. Deduce the coordinates of G, the centroid of ABC.

3. Consider the point $R(1,4)$.

a. Compute the lengths RA, RB and RC.

b. What is the point R in the triangle ABC ?

We now admit that two lines Δ and Δ' with respective slope-intercept equations $y = mx + p$ and $y = m'x + p'$ are perpendicular if and only if $m \times m' = -1$.

4. **a.** What is the slope of any line perpendicular to (BC).

b. Deduce the slope-intercept equation of the altitude through A in ABC.

c. Find out the simplest equation of the altitude through C in triangle ABC.

d. Deduce the coordinates of H, the orthocenter of the triangle ABC.

5. Use vectors to check that the points G, R and H are collinear.

6. The Euler line passes through a fourth important point in the triangle. Find the name and definition of this point.