Statistique – Measures of dispersion

A la fin de ce chapitre, vous devez être capable de :
- calculer la médiane et les quartiles d'une série définie par effectifs ou fréquences ;
- calculer des effectifs cumulés, des fréquences cumulées ;
- représenter une série statistique par une courbe des fréquences cumulées ;
- utiliser un logiciel ou une calculatrice pour étudier une série statistique.

9.1 Some simple examples

Compute the median and quartiles for each of the following sets of statistical data :

1. 11 ; 24 ; 32 ; 41 ; 49 ; 66 ; 85.

2. 12 ; 25 ; 8 ; 13 ; 15 ; 5.

3.

Value	2	3	4	5	6	7
Frequency	12	23	14	42	17	21

4.

Value	2	3	4	5	6	7
Frequency	0.25	0.1	0.1	0.2	0.15	0.2

9.2 The cumulative frequencies polygon

In a class, the marks of a test are as shown below :

3 - 12 - 10 - 5 - 7 - 9 - 7 - 12 - 18 - 9 - 8 - 10 - 15 - 3 - 15 - 7 - 12 - 9 - 10 - 5

1. **a.** Find the median and the quartiles and explain your method.

b. Interpret the median and the quartiles in the context of this exercise.

2. Approximate values of the median and quartiles can also be found using the cumulative frequencies polygon.

a. Fill out a table with the relative frequencies distribution and the cumulative relative frequencies.

b. Draw the cumulative frequencies polygon (units 1 cm in abscissa, 10cm in ordinate).

c. Draw the line $y = 0.5$ and find the abscissa of its intersection with the frequency polygon. This is an approximate value of the median. Compare it to the value you found in the previous question.

d. Adapt the previous method to find approximate values of the two quartiles.

9.3 CO_2 Emissions in 2009

The file CO2Emissions.ods in the folder `PartageEleves/2ndeEuro/` gives the total CO_2 emissions of some countries in 2009, in Gigagrams. The countries are sorted by alphabetical order.

1. Compute the positions (not the values) of the median and the two quartiles.
2. Can you find the values of the median and the quartiles with the countries sorted that way ?
3. Select the cells B2 to C42, then use the Data/Sort command to sort the countries by CO_2 emissions, in ascending order.
4. What country had the lowest CO_2 emissions in 2009 ?
5. What country had the highest CO_2 emissions in 2009 ?
6. Find the median and quartiles of this statistical series with the usual method.
7. a. In the cells C43 to C45, enter the formulas `=MEDIANE(C2:C42)`, `=QUARTILE(C2:C42;1)` and `=QUARTILE(C2:C42;3)`.
   b. Explain these three formulas.
   c. Do you get the same results as with the usual method ? If not, which ones are correct ?

9.4 Quality control

A machine is producing cylindric bars of steel for reinforced concrete, with a theoretical diameter of 25mm. The quality of the production is controlled by extracting a sample of 200 bars from the production. The diameters of these bars are given below, in inches to 3DP.

Diameter	0.949	0.957	0.965	0.972	0.980	0.988	0.996	1.004	1.012	1.02
Frequency	2	8	26	48	38	28	20	16	10	4

Part A – Computation of the main statistical measures

1. Explain the meaning of the value 38 in the table.
2. Find out the population, the size of the population, the data and the type of data of this statistical series.
3. Give the relative frequency distribution.
4. Draw the frequencies polygon (scale : 1 cm for 0.2 unit on x-axis and 1 cm for 1 unit on y-axis).
5. Compute the mean of the data :
   a. using the absolute frequencies ;
   b. using the relative frequencies.
6. Compute the median, the first and the third quartile of the data.
7. Interpret the average and the median in the context of this exercise.

Partie B – Analysis of the data

8. The machine is considered to be working correctly if :
   - the range of the sample is less then 10% of the mean m ;
   - the difference between the mean and the median is less than 0.2 ;

- 95% of the bars have a diameter in the interval $[m - 0.8, m + 0,8]$.

According to the values in the sample, can we consider that the machine is working correctly ?

9. A supervisor notices that the diameters were incorrectly measured and are in fact 0.04 inches more than the recorded values.

a. What is the true mean of the sample ? Explain ?

b. What is the true range of the sample ? Explain ?

c. What is the true median of the sample ? Explain ?

9.5 Comparing two classes

Two classes were given the same math test. Here are the results, in two separate tables with the marks in the first row and the frequencies in the second row :

Class # 1																				
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	0	0	0	0	0	0	1	2	1	3	7	7	4	2	1	1	1	0	0	0

Class # 2																				
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	0	0	1	0	1	1	2	0	4	5	5	3	0	1	0	0	0	2	3	2

The two teachers, talking about their classes, say that they managed the test in the same way, as the two means are equal.
Use all the statistical measures that you know, which may be found using a calculator, to comment on that.

9.6 Legal marital age

In every country, there is a legal minimum age to be married. For example, in France, since 2006, no-one can be married before the age of 18. In some countries, the legal age is different for boys and girls. The table below gives the number of countries for different legal ages for girls.

Legal age	12	13	14	15	16	17	18	19	20	21
Number of countries	1	1	1	12	16	14	120	3	6	16

1. Count the number of countries involved in this study.

2. Compute the average legal marital age in these countries. Compare this value to the legal age in France.

3. Find out the median and the quartiles of the legal marital age in this data. Interpret these values with three precise sentences.

4. The quartiles are not very useful with this data. Explain why.

5. Instead of the quartiles, we can use another set of statistical indicators called the deciles. Here are their definitions :

- the first decile, D_1, is a value such that at least 10% of the data are less than D_1, and at least 90% of the data are more than D_1 ;
- the ninth decile, D_9, is a value such that at least 90% of the data are less than D_9, and at least 10% of the data are more than D_9.

Find out the values of D_1 and D_9 for this data.

9.7 Median and quartiles with grouped values

In some situations, there are so many different values possible in a statistical study that it is needed to group the values in intervals. As an example, the table below gives the employment size of firms in the U.S. in 2008, including small businesses.

Employment size	Frequency
Firms with 1 to 4 employees	3,617,764
Firms with 5 to 9 employees	1,044,065
Firms with 10 to 19 employees	633,141
Firms with 20 to 99 employees	526,307
Firms with 100 to 499 employees	90,386
Firms with 500 to 749 employees	6,060
Firms with 750 to 999 employees	3,038
Firms with 1,000 to 1,499 employees	3,044
Firms with 1,500 to 1,999 employees	1,533
Firms with 2,000 to 2,499 employees	904
Firms with 2,500 to 4,999 employees	1,934
Firms with 5,000 to 9,999 employees	975
Firms with 10,000 employees or more	981

1. Explain the meaning of the number 90,386 in the row "Firms with 100 to 499 employees".
2. Compute the total number of firms covered by this study.
3. Use the middle of each interval to compute the mean number of employees.
4. Use the usual method to find the positions of the median and the quartiles.
5. What can we deduce about the median and quartiles from the previous answers? Can we find their exact values?
6. An approximate value of the median can be found with the formula below, where
   - L_m is the lower boundary of the median class;
   - c is the size of the median class;
   - F_{m-1} is the cumulative frequency of the class before median class;
   - f_m is the frequency of the median class;
   - n is the total number of the data.

   $$Med = L_m + \left[\frac{\frac{n}{2} - F_{m-1}}{f_m}\right] \times c$$

   Use the formula with this statistical series.

9.8 True of false?

Are the following sentences true of false? Each time it's false, give a counter-example.

1. In a statistical series, exactly half of the values are less than or equal to the median.
2. In a statistical series, exactly half of the values are less than or equal to the mean.
3. The median of a statistical series is always less then its mean.
4. The median of a statistical series may be different from its mean.
5. If the median of a statistical series is equal to 42, then the third quartile is greater than or equal to 42.

6. In a statistical series, there are as many values lower than the first quartile as there are values greater than the third quartile.
7. If the minimum in a statistical series is equal to 5, then all the values in the series are greater than or equal to 5.
8. If the minimum in a statistical series is equal to 5, then there is at least one value in the series greater than 5.
9. If the quartiles of a statistical series are equal to 7.1 and 11.9, then at least half of the values are between 7.1 et 11.9.
10. If more than half the values of a series are between 7.1 et 11.9, then the quartiles are equal to 7.1 et 11.9.

9.9 The median

Below is an excerpt from the book *Statistical Calculation For Beginners*, by E.G. Chambers, Cambridge University Press.

> Another form of average that is sometimes used for convenience is the *median*. This, as its name implies, is the middle observation and it is easily found in ungrouped data by ranking the sample of observations in the order of their size and finding the central observation, if N [the number of observations] is odd, or the mean of the two central observations, if N is even.
>
> If N is odd, the median will be the $\frac{N+1}{2}$th observation. If N is even, it will be the mean of the $\frac{N}{2}$th and the $\left(\frac{N}{2}+1\right)$th observations. For example, the median of 2, 4, 6, 8, 10, 12 is the mean of the 3rd and 4th readings, i.e. $(6+8)/2 = 7$.
>
> The median is representative of a set of observations in the sense that there are exactly as many observations greater than it as there are less. If the distribution is perfectly symmetrical, the median is equal to the mean.

1. What is the median of a set of observations and how is it found ?
2. Explain the difference between the median when N is odd and when N is even.
3. Consider the set of numbers 2, 4, 6, 8, 10, 12 given as an example in the text.
   a. Check the median given in the text, then compute the arithmetic mean.
   b. What would be the median if the set had included an extra value equal to 14 ?
   c. What would be the median if the set had included an extra value equal to 200 ? What would be the arithmetic mean in that case ?
   d. What property of the median, that the arithmetic mean doesn't share, is highlighted by the answers to the previous questions ?
4. In what precise way is the median representative of a set of observations ? Illustrate this property with an example.
5. a. Give an example of a set of 5 observations whose median is equal to the mean.
   b. Give an example of a set of 5 observations whose median is not equal to the mean.

9.10 An algorithm to check the median

```
begin
    Input : S, a statistical series ;
    n, the number of values in the series ;
    m, a possible value for the median ;
    l takes the value 0 ;
    g takes the value 0 ;
    for every term t in the series S do
        if t ⩽ m then
            l takes the value l + 1 ;
        else
            g takes the value g + 1 ;
        end if
    end for
    if l ⩾ n/2 and g ⩾ n/2 then
        Output : 'm is a median of this series." ;
    else
        Output : 'm is not a median of this series." ;
    end if
end
```

1. After applying this algorithm to a certain statistical series such that $n = 13$, we get, at the end of the **for** loop, the values $l = 2$ and $m = 11$. What will be the output of the algorithm ?
2. In this question, the statistical series entered is $S = 2, 5, 3, 7, 2, 9, 8, 9, 11, 3$.
   a. What is the value of n ?
   b. Apply the algorithm with $m = 2$, showing the successive values of l and g in a table with two rows. What is the output ?
   c. Apply the algorithm with $m = 6$, showing the successive values of l and g in a table with two rows. What is the output ?
   d. Are there other values of m for which the output will be the same as in the previous question ?
3. How would you modify the algorithm to check the value of the first quartile, instead of the median ?

Homework #6 – Carbon emissions

The exhaust gases of a car containe carbone dioxide (CO_2), which is not toxic, and carbone monoxide (CO), which is toxic. During a pollution control, the emissions of dioxide and monxide are measured and noted respectively T_1 for CO_2 and T_2 for CO. Then, the compound rate T is computed :

$$T = \frac{15T_2}{T_1 + T_2}$$

Regulations require that T be less than 4.5.

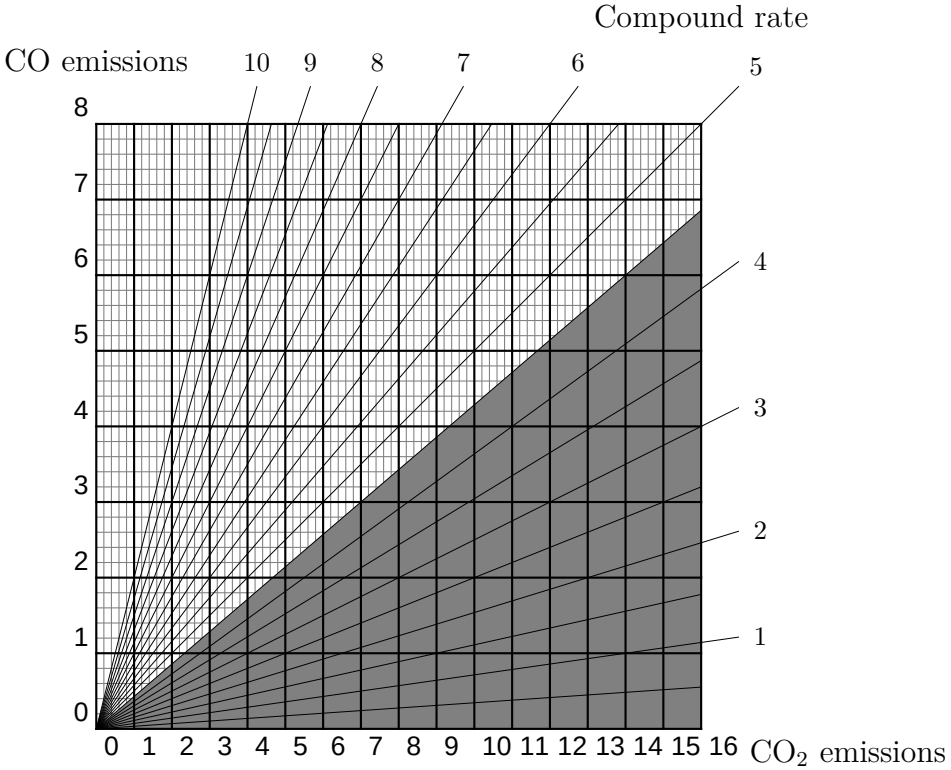
1. Copy and fill the table below, giving the emission of six cars labelled A to F :

Cars	A	B	C	D	E	F
T_1	12.5	6.5	10	7		11
T_2	3	5		4.5	2	1.5
T			3		3	

2. Which cars do not meet the regulations ?

The control staff use a graph (reproduced below), called an abacus.

- The CO_2 emissions are shown on the x-axis.
- The CO emissions are shown on the x-axis.
- The compound rate T associated to the couple $(T_1; T_2)$ is shown as a set of half-lines starting from O and labelled from $T = 0.5$ to $T = 10$.



3. The point such that $T_1 = 7$ and $T_2 = 3$ is on the half-line labelled $T = 4.5$. Check that this is correct.

4. Use the abacus to read the compound rate in the two situations below, and check the values with the formula.

a. Car G : $T_1 = 6$ and $T_2 = 2$;

b. Car H : $T_1 = 10$ and $T_2 = 5$.

Do these cars meet the regulations ?

5. What is the grey region on the abacus ?

6. Prove that $T = 4.5$ if and only if $T_2 = \frac{3}{7}T_1$. Then, check that the slope of half-line labelled $T = 4.5$ is indeed $\frac{3}{7}$.

7. What is the slope of the half-line labelles $T = 2$?

8. A car such that $T_1 = 8$ meets the regulations. What can you say about the value of T_2 for that car ?

9. A car such that $T_2 = 5$ doesn't meet the regulations. What can you say about the value of T_1 for that car ?