Chapter 9

Statistics

CHALLENGE 9

EVALUATION 9

CHALLENGE 9

1. Give an example of a statistical survey that is

a) a poll. ***Varied answers***

b) a census. ***Varied answers***

c) a study. ***Varied answers***

2. Several variables from the students in your class are to be studied. Give an example of a variable that is

a) qualitative. ***Varied answers***

b) quantitative and discrete. ***Varied answers***

c) quantitative and continuous. ***Varied answers***

3. Explain the following sampling techniques using a statistical survey of your choice.

a) Random sampling. ***Varied answers***

b) Systematic sampling. ***Varied answers***

c) Stratified sampling. ***Varied answers***

d) Cluster sampling. ***Varied answers***

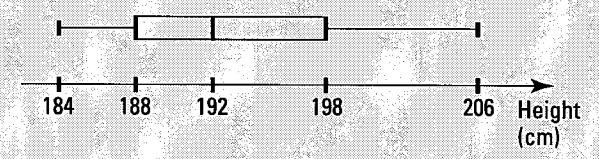
4. A bias is any error occurring in a statistical survey. Give several examples of biases that can occur in a statistical survey.

Varied answers

5. A teacher informs his students about the following characteristics of their last test. Mean: 72; median: 68; mode: 70. Interpret these 3 results.

- ***On average, the students scored 72 (in other words, if every student had the same mark, it would be 72).***
- ***Approximately half the class got a mark less than 68.***
- ***The most frequent mark on the test was 70.***

6. The box-and-whisker plot on the right illustrates the height (cm) of a basketball team's players. What is this graph telling us?



Varied answers

Approximately 25% of the players are less than 188 cm tall.

Approximately 50% of the players are less than 192 cm tall.

Approximately 75% of the players are less than 198 cm tall.

All players are between 184 and 206 cm tall.

9.1 Statistical surveys

Activity 1 Census – Poll – Study

The Lakeside school has a population of 400 students. Every student in this high school is surveyed to determine their mother tongue, number of siblings and height (in cm).

a) Is this survey a census, a poll or a study? Justify your answer.

A census because the entire student population is surveyed.

b) Give a few possible values for each of the following variables.

1. Mother tongue. ***French, English, Spanish, Italian...***
2. Number of siblings. ***0, 1, 2, 3...***
3. Height (in cm). ***Any real number between 140 and 210 cm.***

c) Indicate the nature of each variable (qualitative or quantitative).

1. Mother tongue. ***qualitative***
2. Number of siblings. ***quantitative***
3. Height (in cm). ***quantitative***

d) To determine the school's most popular meal, the cafeteria staff decides to ask 10 students chosen at random in each of the 5 levels.

1. Is the cafeteria staff's survey a census, a poll or a study? Justify your answer.

   A poll, since only part of the school's population is surveyed.

2. Which of the following samples seems more representative of the school's population? Asking 10 students from each level or asking 50 secondary 1 students? Justify your answer.

   Since meal preference can change with age, choosing 10 students at each level would give a more representative sample than choosing 50 secondary 1 students.

e) In order to gauge the effectiveness of the new mathematics program, the Minister of Education asks the school's mathematics department head. Is this a census, a poll or a study? Justify your answer.

A study because the Minister is asking an expert in this field.

STATISTICAL SURVEYS

- A **population** is a set of persons, objects... that are being considered in a statistical survey. A **sample** is a subset of this population.

  Ex.: From the population of Brentwood high school, the students in Nancy's class are chosen as a sample.

- The total number of elements in the population or the sample is called the **size**. The size of a population is represented by N and the sample size is represented by n.

  The **survey rate** is equal to $\frac{n}{N}$.

  Ex.: From a population size of 10 000, a sample size of 100 is chosen. The survey rate is therefore equal to $\frac{100}{10\,000}$ or 1%.

- The characteristic of a population's individuals that we wish to study is called the **variable**. The variable's **modalities** are the variable's possible values.

  Ex.: In the population of the Marie Curie School, we study
  - the variable "number of siblings" which has values that could be: 0, 1, 2...
  - the variable "eye colour" which has values that could be: blue, green, brown, black...

- **Two types of variables** are considered.
  - A variable is **quantitative** if it expresses a quantity and takes numerical values.
    Ex.: The variable "number of siblings" which has the numerical values: 0, 1, 2...
  - A variable is **qualitative** if it expresses a quality and does not take numerical values.
    Ex.: The variable "eye colour" that has non-numerical values such as blue, green, brown, black...

- A **census** is a statistical survey where every element of the population is surveyed.

- A **poll** is a statistical survey where a sample is studied to infer information about the population from which a sample is taken.

  A sample is considered **representative** when it has the same characteristics as the population from which it is taken and gives a good general idea of the population. If the sample is not representative, it is considered **biased**.

- A **study** is a statistical survey where experts in the field being studied are surveyed.

1. For each of the following statistical surveys, indicate
1. the population being studied.
2. the variable being studied and its possible values.
3. the type of variable (qualitative or quantitative).

a) We record the number of children living in each apartment of a building.

1. ***The set of all the apartments in the building.***
2. ***The number of children: 0, 1, 2, 3, 4...***
3. ***Quantitative variable.***

b) At the exit of a movie theatre, people are asked their opinion on the movie they just saw.

1. ***The set of all people.***
2. ***The opinion of the movie: excellent, good, average, mediocre, bad.***
3. ***Qualitative variable.***

c) A school's students are asked how long (in minutes) it takes them to school in the morning.

1. ***The school's student population.***
2. ***The duration of their route. Any real number located, for example, in the interval]0, 120].***
3. ***Quantitative variable.***

d) A school's secondary 3 students are surveyed to determine their favorite subject.

1. ***All sec 3 students of that school.***
2. ***Favorite subject: math, physical education, history...***
3. ***Qualitative variable.***

e) The number of passengers per car is recorded at the entrance of the Champlain bridge.

1. ***All the cars taking the Champlain bridge.***
2. ***The number of passengers: 1, 2, 3, 4, 5...***
3. ***Quantitative variable.***

2. Indicate if the given statistical survey is a census, a poll or a study.

a) A tour guide is asked about the history of a site on the tour. ***Study***

b) A tour coordinator asks some of the tourists about their level of satisfaction with the tour. ***Poll***

c) The secondary 3 student council representative asks all secondary 3 students to determine the choice of the next extra-curricular activity. ***Census***

d) A doctor is asked about the effects of a certain medicine on his patients. ***Study***

e) A journalist stands on a street corner and asks people walking by about their opinion on a new governmental bill proposition. ***Poll***

3. True or false?

a) A census

1. gives more precise information than a poll. ***True***
2. is less expensive to carry out than a poll. ***False***
3. is a more lengthy survey than a poll. ***True***

b) A poll

1. gives quicker results than a census. ***True***
2. enables you to make decisions with as much certainty as a census. ***False***
3. is necessary when the population being studied is too large. ***True***

4. For each of the following situations, indicate if it is preferable to do a census or a poll.

a) We want to know the average age of a group of 30 students. ***Census***

b) We want to verify the quality of the tin cans being manufactured in a factory. ***Poll***

c) We want to create an electoral list for an upcoming municipal election. ***Census***

d) We want to evaluate the popularity of the political party currently in office. ***Poll***

ACTIVITY 2 Quantitative variables

We consider a high school's entire student population. The following quantitative variables are being studied:

1. number of spoken languages
2. number of siblings
3. distance from home to school
4. time required to travel to school

a) Which variables can only take integer values?

Number of spoken languages, number of siblings

b) Which variables could take any real number as values?

Distance from home to school, time required to travel to school

DISCRETE QUANTITATIVE VARIABLES – CONTINUOUS QUANTITATIVE VARIABLES

There are two types of quantitative variables.

- A quantitative variable is **discrete** when it can only take values that are separate from each other and are generally **integers**.

  Ex.: The variable "number of goals scored" in a hockey game is a discrete quantitative variable. The possible values are: 0, 1, 2, 3...

- A quantitative variable is **continuous** when its values can be any real number within an interval.

  Ex.: The height of the players on a hockey team is a continuous quantitative variable. The variable "height" can be any real value within the interval [160, 210].

5. Among the following quantitative variables, indicate if they are discrete (D) or continuous (C).

a) A person's weight (in kg). ***C***

b) The number of mistakes on a spelling test. ***D***

c) The temperature (in °C). ***C***

d) The number of cars stopped at a red light. ***D***

e) The number of participants at a convention. ***D***

f) The lifespan of a light bulb. ***C***

6. Name a few quantitative variables that are

a) discrete. ***Varied answers***

b) continuous. ***Varied answers***

9.2 Statistical tables and diagrams

Activity 1 Representing qualitative variables

The hair colour of the 24 students in a class is recorded.

bl	br	bl	B	br	B	br	bl	br	br
B	R	br	br	B	br	br	bl	bl	br
br	bl	br	R						

bl: blond; br: brown; B: black; R: red

a) Complete the condensed frequency table on the right indicating the distribution of the students according to their hair colour.

b) 1. How many students have brown hair? ***12***

2. What percentage of the students have black hair? ***16.7 %***

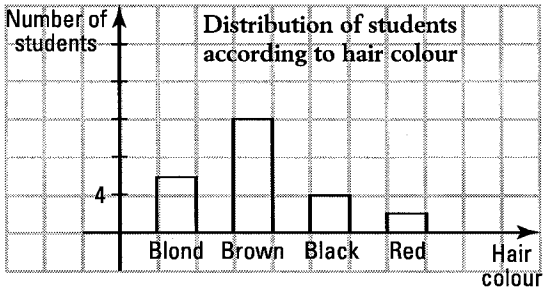
c) What is the most predominant hair colour in this class? ***Brown***

d) A qualitative variable is illustrated by a bar graph or a circle graph (pie chart).

1. Complete the bar graph.

Distribution of students according to hair colour

Coulour	Frequency	Relative frequency (%)
Blond	6	25
Brown	***12***	***50***
Black	***4***	***16.7***
Red	***2***	***8.3***
Total	24	100

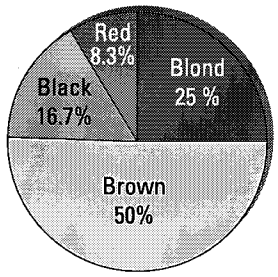


2. Complete the circle graph after calculating the angle allocated to each sector.

Distribution of students according to hair colour

Colour	Frequency	Angle
blond	6	90°
brown	12	***180°***
black	4	***60°***
red	2	***30°***
Total	24	360°

Distribution of students according to hair colour



ACTIVITY 2 Using a bar graph to represent a discrete quantitative variable

The number of people living in each apartment of a building is recorded.

2	4	3	1	0	2	4	5	4	6
4	6	2	5	3	4	2	3	1	3
3	2	1	3	5					

a) Group the data in a condensed frequency table.

Apartment distribution according to the number of residents

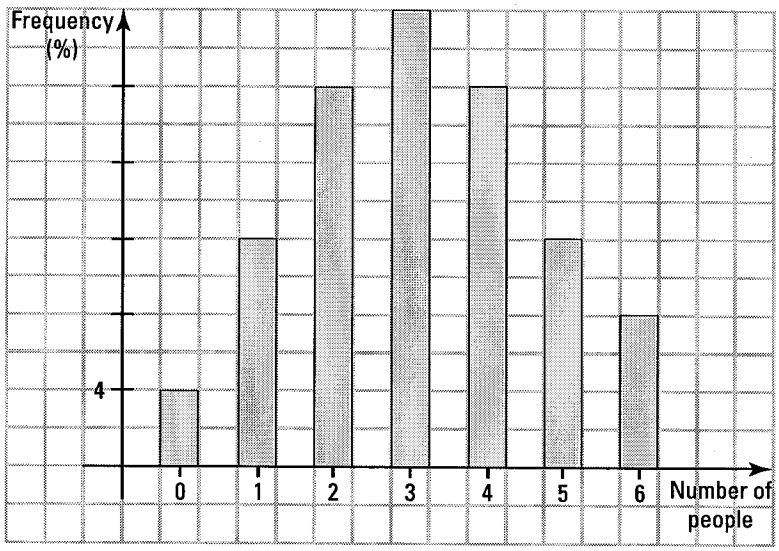
Number of people	Frequency	Relative frequency (%)
0	*1*	*4*
1	*3*	*12*
2	*5*	*20*
3	*6*	*24*
4	*5*	*20*
5	*3*	*12*
6	*2*	*8*
Total	*25*	*100*

b) What percentage of apartments have

1. 2 residents? *20 %*
2. less than 2 residents? *16 %*
3. 3 residents or less? *60 %*
4. more than 4 residents? *20 %*
5. 3 or more residents? *64 %*

c) Draw the bar graph for this situation.

Apartment distribution according to the number of residents



d) What is the number of people

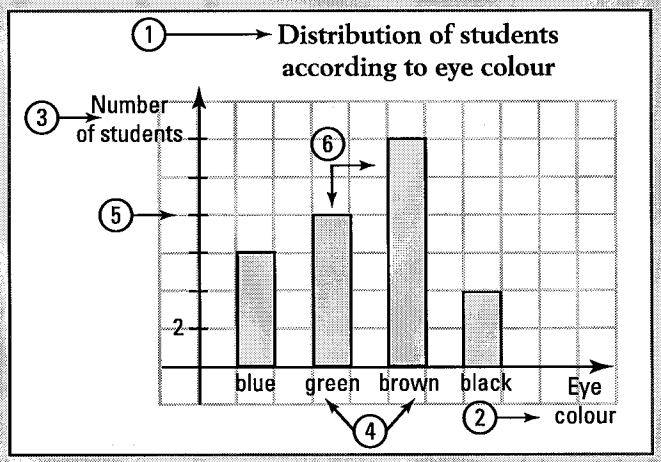
1. most observed? *3 people*
2. least observed? *0 people*

DIAGRAMS

- A bar graph can be used to illustrate a qualitative variable or a discrete quantitative variable.

Distribution of students according to eye colour

Eye color	Number of students
blue	6
green	8
brown	12
black	4
Total	30



The main elements are

① the title

② the label of horizontal axis: the variable "eye colour".

③ the label of vertical axis: the frequencies "number of students".

④ the label of bars (the variable's modalities): blue, green, ...

⑤ the scale of vertical axis: the scale is chosen taking the frequencies into consideration.

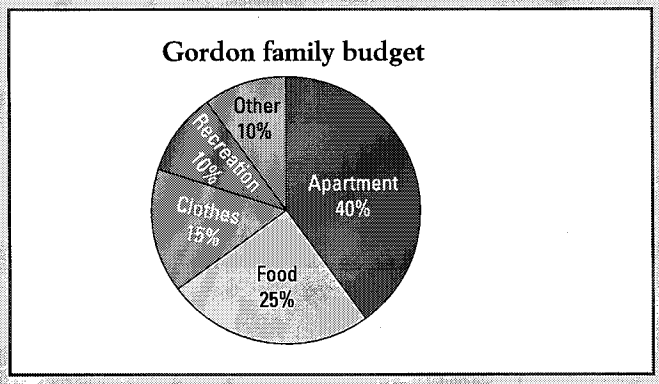
⑥ the bars with uniform width and equally spaced out. The height of each bar is proportional to the frequency.

In a bar graph, the bars can be represented **vertically** or **horizontally**.

- A circle graph can be used to illustrate a **qualitative** variable. Each sector represents a part of a whole.

Gordon family budget

Sector	Relative frequency (%)	Angle
Apartment	40	144°
Food	25	90°
Clothes	15	54°
Recreation	10	36°
Other	10	36°
Total	100	360°

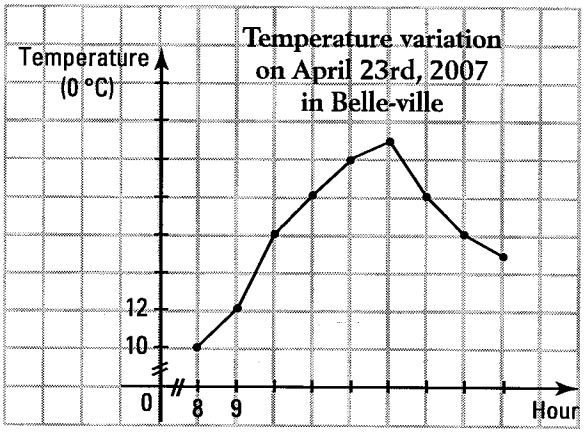


1. Calculate the angle allocated to each sector by multiplying the sector's relative frequency by 360°. Thus, the measure of the angle for the sector "apartment" is $40\ \% \times 360° = 144°$.
2. Draw a circle and construct each sector using a protractor.
3. Clearly identify each sector with its allocated percentage.
4. Give a title to the diagram.

ACTIVITY 3 Representing a continuous quantitative variable changing over time

a) Use a broken line graph to represent the temperature variation recorded hourly on April 23rd, 2007 in Belle-ville.

Hour	8	9	10	11	12	13	14	15	16
Temperature	10	12	16	18	20	21	18	16	15



b) 1. At what time of day was the maximum temperature recorded? *at 1 p.m.*

2. What was the maximum temperature? *21°*

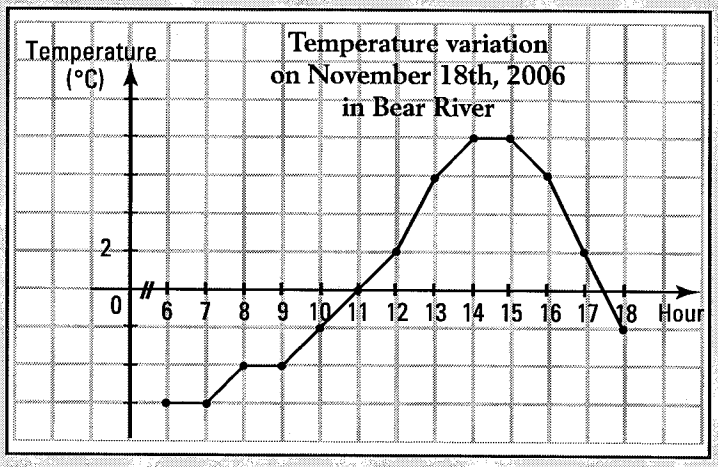
c) Between which consecutive hours did we observe the greatest temperature variation?
Between 9 and 10 a.m.

d) Describe the temperature variation on this day.
The temperature consistently climbed until 1 p.m. and then consistently dropped until 4 p.m.

BROKEN LINE GRAPHS

- A broken line graph is used to represent a variable that is evolving continuously: temperature variation, plant growth, stock value...

The broken line graph on the right has several consecutive segments joining a succession of points. Each point indicates the temperature recorded for each hour.



ACTIVITY 4 Grouping data into classes – Histograms

Here are the results of 30 students on a mathematics test.

78	60	80	88	78	86	77	60	64	85	70	88	77	45	47
93	56	74	50	83	97	70	94	67	77	84	62	72	82	57

a) Since there are a lot of data that are for the most part distinct, we group them into classes. Group the data into 6 consecutive classes with an amplitude of 10.

Distribution of students according to their test result

Class #	Classes	Tally	Frequency	Relative frequency%
1	[40 – 50[	||	*2*	*6.7*
2	*[50 – 60[*	|||	*3*	*10*
3	*[60 – 70[*	𝍸	*5*	*16.7*
4	*[70 – 80[*	𝍸 ||||	*9*	*30*
5	*[80 – 90[*	𝍸 |||	*8*	*26.7*
6	*[90 – 100[*	|||	*3*	*10*
	Total		*30*	*100.1*

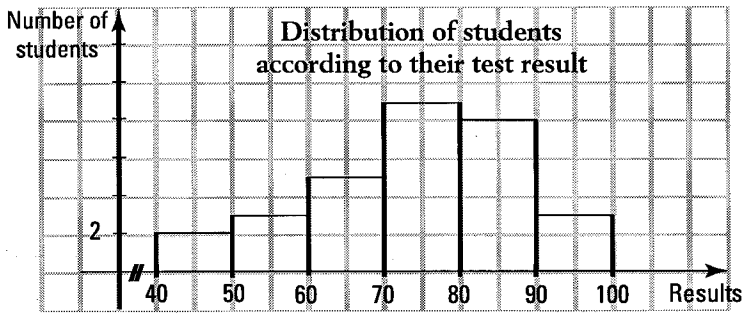
Each class is **closed** on the **left** and **open** on the **right**... The result of 50 therefore belongs to the 2nd class.

Because of rounding, the sum of the percentages is close to 100%

b) What percentage of students received a result

1. greater than or equal to 60 and less than 70? *16.7 %*
2. less than 60? *16.7 %*
3. greater than 80? *36.7 %*

c) We use a **histogram** to represent data grouped in classes. Complete the following histogram.



Since the classes are adjacent, the rectangles are adjacent.

d) In which class do we find

1. the most students? *[70 – 80[*
2. the least students? *[40 – 50[*

ACTIVITY 5 Representing a continuous variable with a histogram

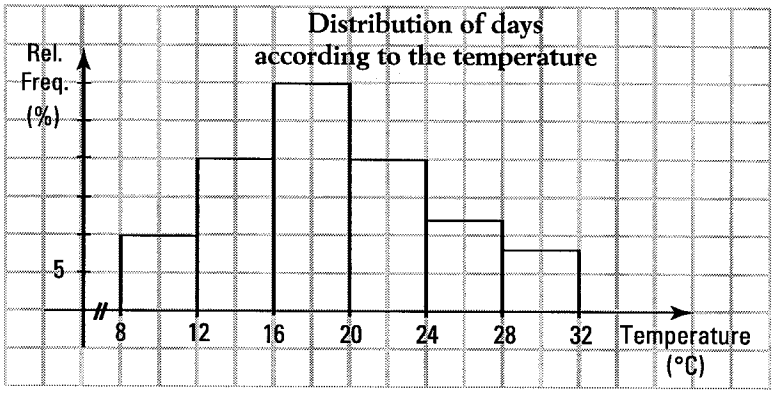
In the Spring, the temperature was recorded at noon for 50 consecutive days. The variable "temperature" being a continuous quantitative variable, the data being numerous and for the most part distinct, the data was therefore grouped into classes in the table below.

Temperature (°C)	Number of days	Relative frequency (%)
[8 – 12[	5	***10***
[12 – 16[	10	***20***
[16 – 20[	15	***30***
[20 – 24[	10	***20***
[24 – 28[	6	***12***
[28 – 32[	4	***8***
	50	***100***

a) Explain why the variable being studied is continuous.

The temperature is a variable with real number values.

b) Construct a histogram to represent this situation.



c) What is, approximately, the percentage of days where we observed a temperature

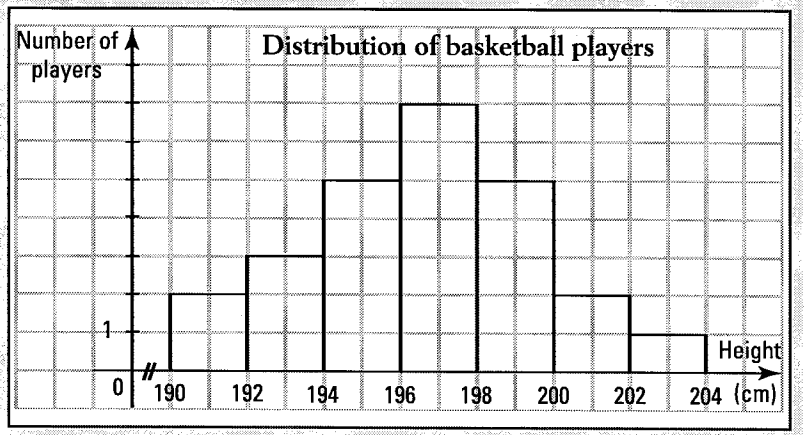
1. less than 14 °C? ***20%***
2. greater than 26 °C? ***14%***

HISTOGRAMS

- A histogram is used to represent data that is grouped in classes.

  Ex.: The histogram on the right shows the distribution of basketball players according to their height (in cm).

  Note that:
  - the players' heights vary between 190 cm and 204 cm.
  - the class [196-198[contains the most players.



1. The number of passengers per car are recorded at a border crossing.

1	2	3	2	1	3	4	2	1	4
5	3	2	1	4	3	6	1	2	1
1	5	4	3	2	6	1	2	3	4
4	5	3	2	1	3	1	2	4	1
2	1	1	3	4	1	2	1	3	2

a) Identify

1. the population being studied. ***The set of vehicles crossing the border***
2. the variable being studied. ***The number of passengers***
3. the type of variable. ***Discrete quantitative variable***

b) Group the data in a condensed frequency table.

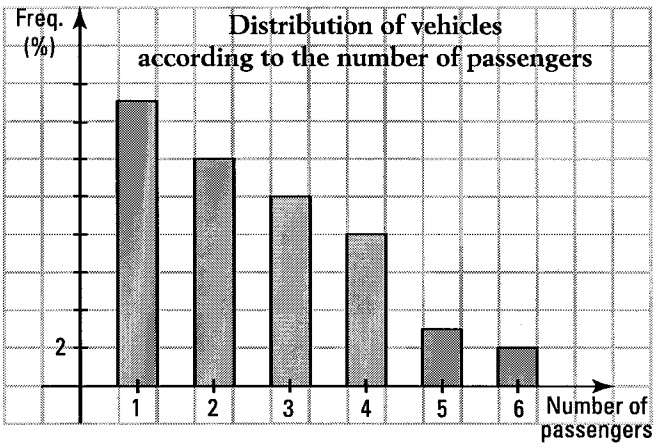
Distribution of vehicles according to the number of passengers

Number of passengers	Frequency	Relative frequency (%)
1	15	30
2	12	24
3	10	20
4	8	16
5	3	6
6	2	4
	50	100

c) In what percentage of vehicles do we observe

1. 3 passengers? ***20%***
2. less than 3 passengers? ***54%***
3. more than 4 passengers? ***10%***
4. at most 4 passengers? ***90%***
5. at least 4 passengers? ***26%***
6. at least one passenger? ***100%***

d) Construct a bar graph to represent this situation.



e) What number of passengers was

1. observed the most often? ***1*** 2. observed the least often? ***6***

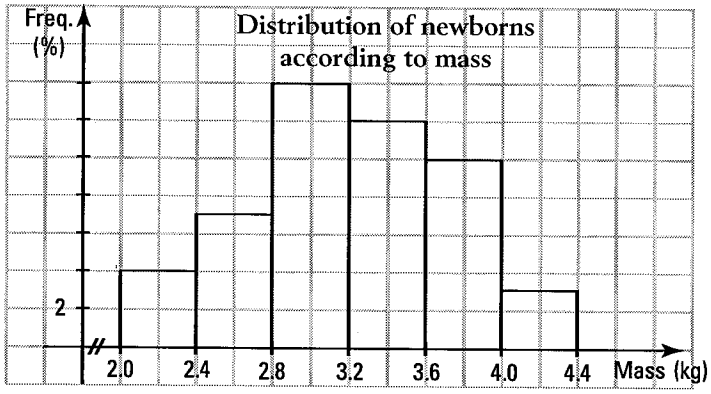
2. The mass (in kg) of forty newborns is recorded at birth.

2.2	2.8	3.2	3.6	2.2	2.9	3.2	3.7	2.3	2.9
2.5	3.0	3.3	3.8	2.5	3.0	3.3	3.9	2.5	3.0
2.6	3.1	3.4	3.9	2.6	3.1	3.5	3.9	2.7	3.1
3.3	3.4	3.5	3.7	3.9	4.1	2.8	3.1	3.6	4.3
2.5	2.3	2.9	3.5	3.4	4.0	2.9	3.9	3.5	3.1

a) Group the data into 6 adjacent classes with an amplitude of 0.4 kg knowing that the lower limit of the 1st class is 2 kg.

N°	Classes	Tally	Frequency	Freq. (%)
1	*[2.0 – 2.4[*	||||	*4*	*8*
2	*[2.4 – 2.8[*	𝍸 ||	*7*	*14*
3	*[2.8 – 3.2[*	𝍸 𝍸 ||||	*14*	*28*
4	*[3.2 – 3.6[*	𝍸 𝍸 ||	*12*	*24*
5	*[3.6 – 4.0[*	𝍸 𝍸	*10*	*20*
6	*[4.0 – 4.4[*	|||	*3*	*6*
			50	*100*

b) Construct a histogram illustrating this situation.



c) What percentage of newborns have a mass situated in the

1. 1st class? *8%*
2. the last two classes? *26%*

d) What percentage of newborns have a mass of

1. less than 2.8 kg? *22%*
2. at least 3.2 kg? *50%*

e) In which class do we find the most newborns?

The 3rd class, which is the class [2.8–3.2[

f) Are the following statements true or false?

1. More than half of the newborns have a mass situated in the 3rd or 4th classes.

   True

2. There are as many newborns with a mass less than 3.2 kg as newborns with a mass greater than or equal to 3.2 kg.

   True

9.3 Sampling techniques

ACTIVITY 1 Random sample

In the table below, we give the age and mother tongue of 50 people vacationing at a resort (F: Francophone; A: Anglophone; X: other).

N°	Age	Language	N°	Age	Language	N°	Age	Language	N°	Age	Language	N°	Age	Language
01	15	F	11	16	F	21	17	F	31	16	F	41	16	X
02	14	A	12	15	F	22	15	X	32	15	A	42	17	F
03	16	F	13	16	A	23	16	F	33	16	F	43	14	X
04	14	F	14	14	A	24	14	A	34	15	F	44	16	F
05	16	X	15	16	X	25	16	F	35	16	F	45	15	A
06	17	F	16	15	A	26	15	X	36	17	A	46	16	X
07	17	A	17	16	F	27	16	A	37	15	F	47	16	F
08	15	F	18	15	F	28	14	F	38	17	F	48	15	F
09	16	F	19	17	A	29	15	F	39	16	A	49	15	A
10	17	X	20	17	F	30	16	F	40	17	F	50	16	A

a) Using a condensed frequency table, represent

1. the distribution of the vacationers according to age.

Age	Freq.	Freq. (%)
14	*6*	*12*
15	*14*	*28*
16	*20*	*40*
17	*10*	*20*
Total	*50*	*100*

2. the distribution of the vacationers according to mother tongue.

Language	Freq.	Freq. (%)
Francophone	*28*	*56*
Anglophone	*14*	*28*
Other	*8*	*16*
Total	*50*	*100*

b) A sample is random when we randomly choose the individuals belonging to the sample and where each individual is equally likely to be chosen. Indicate a technique that will enable you to choose a random sample with a size of 10 from this population of 50 vacationers.

For example, I place 50 papers numbered 1 to 50 in a bag. I randomly choose 10 papers successively and without replacement. The 10 drawn numbers enable me to create my sample, since each person is numbered.

c) 1. Choose a random sample of 10 from this population of 50 vacationers.
Varied answers

2. Compare the percentage of francophones in your sample to the percentage of francophones in the population. ***Varied answers***

3. Compare the percentage of 15 year olds in your sample to the percentage of 15 year olds in the population. ***Varied answers***

ACTIVITY 2 Systematic sampling

Consider the data (from Actvitiy1) indicating the age and mother tongue of 50 people vacationing at a resort.

a) We now present a technique for choosing a systematic sample of size $n = 8$ from a population of size $N = 50$.

The polling step, represented by "p", is the closest integer to the ratio $\frac{N}{n}$.

1. Calculate the polling step. $p = \frac{N}{n} = \frac{50}{18} = 6{,}25 \simeq 6$
2. Randomly choose an integer between 1 and p. Let k designate your chosen number ***Varied anwers***
3. On the list of vacationers, the first person chosen for the sample will be in the position k, the next will have the position $k + p$... and so forth. We therefore obtain a systematic sample..

N°	List of vacationers
01	
02	
⋮	
k	
⋮	
$k + p$	
⋮	
$k + 2p$	
⋮	

b) 1. Choose a systematic sample of size 8 from this population.
Varied answers

2. Compare the percentage of francophones in your sample to the percentage of francophones in the population.
Varied answers

3. Compare the percentage of 15 year olds in your sample to the percentage of 15 year olds in the population.
Varied answers

SIMPLE RANDOM SAMPLING – SYSTEMATIC SAMPLING

There are several sampling techniques for drawing a sample of size n from a population of size N.

- **Simple random sampling**: method consisting of randomly choosing the individuals belonging to the sample according to a technique whereby each individual is equally likely to be chosen. (see activity 1)
- **Systematic sampling**: method requiring a list of the entire population where the individuals are numbered 1 to N.
  - Calculate the polling step "p" which is equal to the nearest integer to the ratio $\frac{N}{n}$.
  - Randomly choose a number, designated by k, located between 1 and p.
  - Systematically choose the n individuals belonging to the sample. The first individual is number k, the second is number $k + p$, the third is $k + 2p$... and so forth until the n individuals belonging to the sample have been selected. (see activity 2)

N°	Individual
1	
⋮	
k	
⋮	
$k + p$	
⋮	
$k + 2p$	
⋮	

Activity 3 Stratified sample

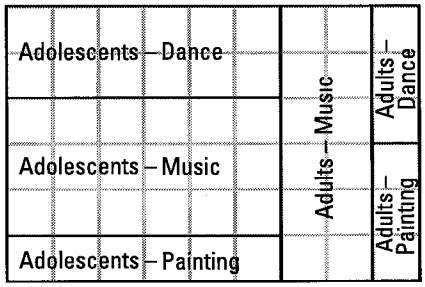
The table on the right shows the distribution of the 600 people registered for classes given at the community centre.

Course \ Person	Adolescents	Adults	Total
Dance	120	60	180
Music	180	120	300
Painting	60	60	120
Total	360	240	600

The population of 600 people has been subdivided into 6 groups called strata. For example, there are 60 individuals of this population who are grouped in the stratum "Adults – Dance".

Choosing a stratified sample of size $n = 40$ consists of randomly choosing from each of the 6 strata a number of individuals proportional to the weight of the stratum in the population.

Population divided into 6 strata



a) 1. Indicate the weight of the stratum "Adults – Dance" in the population by establishing a ratio or a percentage. Interpret your result.

$\frac{60}{600}$ or 10%. 10% of the population are adults registered in the Dance course.

2. Let x represent the number of individuals that must be chosen from the stratum "Adults – Dance" to be part of the sample. Write a proportion which enables you to determine x if we want a stratified sample, and calculate x.

$\frac{x}{40} = \frac{60}{600}$ $x = 4$

b) Determine the number of individuals in a stratified sample of size $n = 40$ that are

1. adolescents. *24* 2. in music. *20* 3. adolescents in music. *12*

c) Complete the table on the right by indicating the number of individuals that we must choose per stratum when randomly choosing a stratified sample of size $n = 40$.

	Adolescents	Adults	Total
Dance	*8*	*4*	*12*
Music	*12*	*8*	*20*
Painting	*4*	*4*	*8*
Total	*24*	*16*	*40*

d) In a stratified population, is it plausible that there is little variation between individuals within a stratum, but that there is a lot of variation from one stratum to another?

Yes

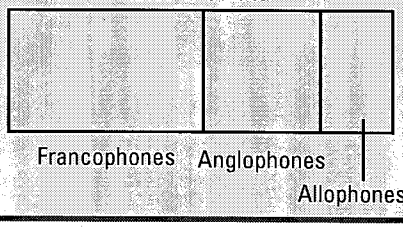
STRATIFIED SAMPLING

- Stratified sampling is a method that requires knowledge of certain characteristics of the population obtained by a preceding census. The population is then divided into groups, called **strata**, which are relatively homogeneous. We **randomly** choose from each stratum a certain number of individuals that will be part of the sample. The number of individuals chosen from each stratum is **proportional** to the weighting of that stratum in the population. We therefore form a sample representing the same characteristics in the same proportions as the population. (see activity 3)

  Ex.: The population of a town is 50% francophone, 30% anglophone and 20% allophone. To create a stratified sample of 1200 people, we must choose 600 francophones, 360 anglophones and 240 allophones.

  $\left(\frac{600}{1200} = 50\%; \quad \frac{360}{1200} = 30\%; \quad \frac{240}{1200} = 20\%\right)$.

Population of the town divided into 3 strata



ACTIVITY 4 Cluster sampling method

A school board wants to study the impact of the new mathematics program on the secondary 3 students.

In this school board, there are a total of 100 secondary 3 groups (clusters) of 30 students.

We want to look at the results of a common exam taken by all of the students in this school board. We want to form a sample with a size of 150.

a) Is it plausible to say that there exists a lot of variation within each group but little variation from one group to another? ***Yes***

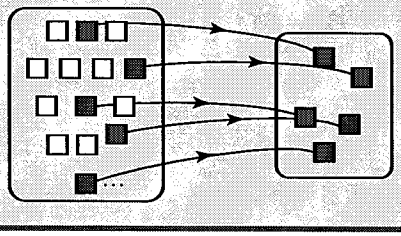
b) Cluster sampling consists of randomly choosing a certain number of clusters (groups) and choosing every individual within those clusters to form the sample. Explain how to proceed if you want to use the cluster sampling method to form a sample with a size of 150.

Number the 100 clusters. Once the clusters are numbered, randomly choose 5 clusters of 30 students. The 150 students chosen will form the sample.

CLUSTER SAMPLING

- A population is made of geographic groupings called **clusters**. When there is a lot of variation within each cluster, but little variation between the clusters, we use **cluster** sampling which consists of choosing a few clusters at **random**. All students within the chosen clusters will form the sample.

  Ex.: A population of size N = 2000 is made of 100 clusters of 20 individuals. We **randomly** choose 5 clusters to form a sample of size $n = 100$.



1. In each of the following cases, indicate the sampling method being used.

a) To gather opinion on the construction of a new park, the mayor of a town wants to conduct a poll. He chooses, from a table of random numbers, 500 residents from the electoral list.
Random sampling

b) A school has a population of 750 students, of which 300 are in secondary 3, 250 in secondary 4 and 200 in secondary 5. To conduct a poll, the student council forms a sample consisting of 12 secondary 3 students, 10 secondary 4 students and 8 secondary 5 students.
Stratified sampling

c) To determine the quality of the cans coming from an assembly line, we remove one can for every 1000 cans that pass through the assembly line. ***Systematic sampling***

2. Explain why the sampling method is inappropriate and suggest a better one.

a) To determine the quality of televisions coming from an assembly line, we inspect the 100 televisions made between 10 and 11 a.m.
Since the assembly line can malfunction at any time of the day, it is preferable to systematically choose a television for every p televisions passing through the assembly line to ensure quality.

b) To determine the favorite leisure activity of a school's students, all of the secondary 5 students are surveyed.
Since preference in leisure activities can change with age, it is preferable to form a stratified sample, with each stratum being a school level.

3. In a school, 60% of the students are of Quebec origin, 20% of European origin, 15% of Asian origin and 5% of African origin. We want to form a sample of 40 students to determine the most popular physical activity.

a) What would be the best sampling method? Justify your answer.
Since ethnic origin can influence one's favorite physical activity, it would be recommended to choose a stratified sample so it could be as representative as possible of the school's population.

b) Indicate the composition of a sample that is representative of the studied population.
24 students of Quebec origin, 8 of European origin, 6 of Asian origin and 2 of African origin.

4. We want to form a sample of 50 students from a school with a population of 495 students. Explain how to proceed using the following sampling techniques:

a) random. ***Assign a number to each student and randomly choose 50 numbers successively without replacement.***

b) systematic. ***Polling step*** $p = \frac{495}{50} \approx 10$. ***Randomly choose a number k, between 1 and 10. Choose the students numbered k, k + 10, k + 20,... until the sample of 50 students has been formed.***

c) stratified. ***The students can be classified by school level (strata) and chosen at random from each stratum based on the weight given to that stratum in the population.***

5. For each of the following studies, choose an appropriate sampling technique for the targeted population.

a) A record company executive wants to determine the music preferences of Quebecers.

The sample must be stratified to get a profile of the population that takes age, gender, mother tongue ... into consideration.

b) The president of a car manufacturing company is launching the design of a new sports car and wants to know what colour is most popular.

The sample is random. Choose a random sample of n sports cars and observe its colour.

c) We want to control the quality of strawberries coming from a farm during the harvesting season.

The sampling is systematic depending on the harvesting days. This will give us a sample with different times within the harvesting season and therefore give a more accurate picture of this farm's strawberry production.

d) We want to know the weekly amount of money spent on groceries in Quebec households.

A stratified sample which gives a sample that takes into consideration the income and number of people in the household.

6. The table on the right gives the distribution of the tourists on a cruise, according to gender and mother tongue. We randomly choose 50 tourists such that the sample reflects the given percentages.

	Men	Women
French	20%	18%
English	18%	16%
Other	16%	12%

a) What is the sampling technique being used?

Stratified sampling technique.

b) Indicate how many tourists in this sample are

1. men. ***27***
2. Francophone. ***19***
3. Anglophone women ***8***
4. Francophone men. ***10***

7. In order to measure certain characteristics (height, weight, ...) of hockey players on Collegiate teams, 4 teams were chosen at random and every player on those 4 teams were part of the sample. The league has 15 teams of 20 players. What sampling technique was used?

Cluster sampling technique.

8. Complete the following by the appropriate term: "little" or "a lot of".

a) In stratified sampling, we generally observe ***a lot of*** variation from one stratum to the other, but ***little*** variation within each stratum.

b) In cluster sampling, we generally observe ***little*** variation from one cluster to the other, but ***a lot of*** variation within each cluster.

9. The following table gives the distribution of a school's students.

We want to form a sample of 80 students; this sample must be representative of the strata identified in the table.

	Number of girls	Number of boys
First cycle	70	90
Second cycle	100	140

How many boys in the 1st cycle should be in this sample? *18*

10. The following table shows the distribution of 2000 spectators at a concert according to age.

Age	Women	Men	Total
[18, 30[	180	190	370
[30, 45[	310	280	590
[45, 60[	330	380	710
[60, 80[	170	160	330
Total	990	1010	2000

We want to form a sample of 200 people. This sample must be representative of the strata mentioned in the table.

How many women aged 45 to 60 years must be in this sample? *33*

11. Associate the appropriate sampling technique to the information gathering in each of the following cases.

Context of the statistical survey	Sampling techniques
a) A poll is conducted in a school of 840 students on the meal choices to be included on the cafeteria's menu. We randomly ask students from secondary 1 to secondary 5 respecting the proportion of each level within the school.	1) Random sampling 2) Stratified sampling
b) A music camp has 400 campers divided into 12 groups according to their instrument. A questionnaire to determine their preference in recreational activities is distributed to 2 groups chosen at random.	3) Systematic sampling 4) Cluster sampling
c) A beauty products company contacts 600 people to ask them about their favorite soap. The first name on each page of the telephone book is chosen.	

9.4 Sources of bias

Bias is any error involved in a statistical survey. There are several sources of bias. The following activities indicate the different sources of bias that can occur in a statistical survey

Activity 1 Sources of bias in choosing the sample

In each of the following situations, indicate why the sample is a source of bias.

a) To determine the recreational activities of Montreal's retired community, the retirees of a senior citizen's home in Montreal are polled.

Only senior citizens living in a senior's residence are considered. They do not take into consideration the younger retirees who are more likely to have a more active lifestyle.

b) To determine the percentage of Quebecers in favor of increased subsidies to the arts and culture sectors, we survey people at the exit of a theatre showing a play.

The chosen sample is composed of people already interested in the cultural domain since they have just seen a play. The percentage of Quebecers in favor of subsidizing the arts would therefore surely be inflated with this sample.

c) To determine the percentage of Montrealers in favor of being allowed to turn right on a red light, we ask people who are stopped at a red light.

The sample is composed only of drivers. We forgot to survey the pedestrians. Pedestrians, who tend not to have cars, would probably be more against the right to turn right on a red light.

Activity 2 Sources of bias in the questionnaire

For each of the following survey questions,

1. explain why it is a source of bias.
2. reformulate the question.

a) Do you agree with rich people benefiting from social programs?

1. ***The social programs are not specified (some are more essential than others) and wealth is not measured.***
2. ***Do you agree with cutting family allowances to homes having a total annual revenue of more than $80 000?***

b) Do you go the cinema often? Yes ☐ No ☐

1. ***Because the frequency is not specified, it is impossible to quantify "often" in the same way for everyone.***
2. ***Do you go the cinema, on average***

never or once per month ☐ 2 or 3 times per month ☐ 4 or more times per month ☐

c) Do you agree with getting less French homework, but more math homework?
Yes ☐ No ☐

1. *A student who wants less homework in both subjects cannot answer the question. We need to formulate two questions instead of one.*

2. *Do you agree with 1. Getting less French homework? Yes ☐ No ☐*
2. Getting less math homework? Yes ☐ No ☐

Activity 3 Sources of bias in non-random sampling

When the individuals forming the sample are not chosen at random, there is a greater risk of bias.

a) The following situation represents a sampling technique that is not random: **accidental sampling** (sample where the individuals are chosen on the basis of being at a certain place at a certain time). A journalist asks pedestrians passing by about a current event.

1. Does the location of the interview influence the results? Justify your answer.
Yes. From one neighborhood to another, the social status of the individuals can influence the results.

2. Does the time of the interview influence the results?
Yes. During working hours, we automatically eliminate a segment of the population. After working hours, it is more likely to interview single people than people with children.

3. Does filming the interview contribute to more bias?
Yes. Certain individuals not wanting to be filmed will distance themselves to avoid being solicited.

b) The following situation represents a sampling technique that is not random: **voluntary sampling** (sample where the individuals are chosen on the basis of their willingness to be part of the sample). A radio show host asks his listeners to call in their opinion on a current event.

1. Does the time of the radio show influence the results? Justify your answer.
Yes. Depending on the time, the show will appeal to a certain type of individual. During working hours, we automatically eliminate a category of listeners.

2. Do people with more extreme opinions tend to express their opinions in larger proportion than the general population? *Yes*

Activity 4 Sources of bias in the processing of data

Is it possible in the processing of data

a) to make a coding error when transforming the answers into numerical codes? *Yes*

b) to make an error in data capture (in other words, while transferring the answers to a computerized system)? *Yes*

Each of the preceding errors is called a "**processing error**".

ACTIVITY 5 Sources of bias in the analysis of data

a) After a pre-election poll, while analyzing the results,

1. can we ignore the undecided? Justify your answer.

No, we must take the undecided into consideration. Certain undecided people refuse to answer but still have an opinion that will count on election day when they go to vote.

2. should we divide the undecided equally among each of the different categories of response?

No, we should divide them up according to the different weights of each category of answers or according to a former study on the behavior of the undecided for a similar statistical survey.

b) After conducting a poll, do we risk putting more emphasis on certain conclusions rather than others?

Yes, depending on their political allegiance, some newspapers will put more emphasis on certain conclusions and ignore others.

SOURCES OF BIAS

Bias is defined as any error associated with a statistical survey.

There are many sources of bias:

- The choice of sample (activity 1)
- The design of the questionnaire (activity 2)
- Non-random sampling (activity 3)
- Data processing (activity 4)
- Data analysis (activity 5)

1. Several different types of bias can occur at one stage or another of a census. These errors then influence the results and the accuracy of the census.

a) Is it possible for the person conducting a census to

1. forget to inventory the occupants of an apartment? ***Yes***
2. forget some of the people who are supposed to be part of the census? ***Yes***
3. to count people without actually surveying them? ***Yes***
4. to count the same person twice? ***Yes***

Each of the preceding errors is called an "**observation error**".

b) Is it possible to be unable to inventory a respondent due to a prolonged absence or the person moving? ***Yes***

This type of error is called a "**non-response error**".

c) 1. Is it possible for a respondent to misinterpret one of the surveyor's questions? ***Yes***

2. Is it possible for a surveyor to misinterpret one of the respondent's answers? ***Yes***

Each of the preceding errors is called a "**response error**".

2. In each of the following situations, specify the type of error.

a) The Landry family has gone to Florida and was unable to be reached for the census.

Non-response error

b) A study is being done on the students of Mountain View School who are in the Math 406 class, taken by students in secondary 4 and 5. Julie, who is conducting the survey, forgot to question the secondary 5 students. ***Observation error***

c) A survey is conducted on the students of Lakeview School concerning the school's hockey team. The question asked is: "Are you for or against the dissolution of the hockey team after their next game?" Anthony wants the team to win their next game. However, he does not want the dissolution of the hockey team no matter what. Frank wants the dissolution of the hockey team if they lose their next game. As for Nathalie, she wants the dissolution of the team no matter what.

Response error

d) A survey is conducted on the students of Riverside School. The question being asked is: "Are you for or against the election of a new student council?"

The answer "yes" is coded as a "1" and the answer "no" is coded by a "0". Franco makes the mistake of coding a "yes" by a 0. ***Processing error***

3. The capacity of a respondent to answer a question is influenced by the information they possess. Here are three questions:
- How old are you? (Question expressing an individual's state)
- Are you optimistic? (Question expressing behavior)
- Do you agree with the government? (Question expressing opinion)

Of the three types of questions (state, behavior or opinion), which one minimizes bias? Justify your answer.

The question expressing an individual's state. Everybody knows their age (their state). The respondent's level of conscience with regards to their behavior or their opinion is not as certain.

4. True or false?

a) The collaboration of the respondents is essential in a poll. ***True***

b) The respondent's comprehension of the question is essential for reducing bias. ***True***

c) The respondent must possess the necessary information to answer the question. ***True***

d) The sincerity of respondents is essential for reducing bias. ***True***

e) The formulation of the question must not suggest a particular answer. ***True***

5. In a local newspaper, we can read the following information:

In Quebec, high school students play badminton.

This is what is revealed by a poll conducted in Dollard-des-Ormeaux of 120 students from Royal School. The collected data is presented in the table on the right. Royal School is a 2nd cycle high school.

Do you play badminton?		
	Girls	Boys
yes	40	50
no	22	8

What are the sources of bias in this poll?

The students surveyed are 2nd cycle only and do not represent all high school students. Only students from one town in Quebec were surveyed.

9.5 Measures of central tendency

Activity 1 The mean

a) The number of passengers per car stopped at a red light was recorded.

2	3	1	2	1	1	3	1	4	1

1. Identify
   i) the population. ***All the cars stopping at a red light.***
   ii) the variable and its type. ***The number of passengers per car; quantitative variable.***
2. Calculate and interpret the mean (average) of the data.
   On average, there are 1.9 passengers per car.
3. When a discrete variable takes only integral values, can the mean be a value that is not an integer? ***Yes***

b) In a statistical survey, a quantitative variable X gives rise to the following series of data: $x_1, x_2, \ldots, x_n$.

1. How many data are there? ***n***
2. What does the quotient $\frac{x_1 + x_2 + \ldots + x_n}{n}$ represent? ***The mean***

Activity 2 The mode

a) The number of people sitting per table at a restaurant is recorded.

3	2	4	5	2	1	2	0	6	2

1. Identify
   i) the population. ***All the tables in the restaurant.***
   ii) the variable and its type. ***The number of people sitting per table; quantitative variable.***
2. The mode is the value of the variable that has the highest frequency in the data. Determine and interpret the mode.
   Mode = 2. There is most often 2 people per table.

b) The hair colour for a group of students is recorded..

blond, brown, brown, blond, black, blond, red, brown, brown, blond, black, brown, blond, brown, black, brown.

1. Identify
   i) the population. ***A group of students.***
   ii) the variable and its type. ***Hair colour; qualitative variable***
2. Determine and interpret the mode.
   Mode: brown. The most frequent hair colour is brown.

c) When a series of data comes from a qualitative variable, is it possible to determine

1. the mode of this data? ***Yes***
2. the mean of this data? ***No***

ACTIVITY 3 The median

The weekly salaries of a company's 8 employees and its manager are recorded. The recorded salaries are: $540, $500, $570, $580, $500, $540, $580, $540, $2400.

a) What is the mean of the data? ***$750***

b) What is the manager's salary? ***$2400***

c) 1. Put the data in increasing order.

500, 500, 540, 540, 540, 570, 580, 580, 2400.

2. The median is the value located in the middle of a distribution written in increasing order. What is the median of this distribution? ***$540***

d) Of the two measures of central tendency, mean and median, which is the most appropriate for representing this company's typical salary?

Justify your answer.

The median is the most appropriate. The mean in this situation is clearly bigger than the median, since the manager's salary inflates the mean.

e) Is it true to say that the mean of a distribution is influenced by the values located at the ends of the distribution? ***Yes***

MEASURES OF CENTRAL TENDENCY

- The measures of central tendency characterize the data at the centre of a distribution. We distinguish three measures: the **mean**, **mode** and **median**.
- The **mean** of a distribution is equal to the sum of the data divided by the number of data.

$$\text{mean} = \frac{\text{Sum of the data}}{\text{Number of data}}$$

If we designate $x_1, x_2, \ldots, x_n$ as the n data entries of the distribution, we designate the mean of the distribution of data $\bar{x}$.

We have: $$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Ex.: Given the distribution: 2, 4, 6, 9, 15. We have: $n = 5$ and $\bar{x} = 7.2$.

- The **mode** is the value of the variable with the highest frequency in a distribution. The symbol used for the mode is M_o.

Ex.: In the distribution: 2, 3, 1, 5, 3, 3, 1, 6, 3, 2, the value 3 appears most frequently (it appears 4 times). We have $M_o = 3$.

- The **median** corresponds to the value located at the centre of the distribution written in increasing or decreasing order. The symbol used for the median is M_d.
  We distinguish two cases:
  - **1st case: odd *n*.**
    Ex.: The distribution 2, 4, 5, 2, 6, 7, 6 when ordered becomes 2, 2, 4, ⑤, 6, 6, 7.
    The distribution has $n = 7$ datum. The median corresponds to the 4th datum $\left(\frac{n+1}{2} = 4\right)$.
    We have: $M_d = 5$.

    If *n* is odd the median is the datum with a rank of $\frac{n+1}{2}$.
  - **2nd case: even *n*.**
    Ex.: The distribution 3, 5, 2, 3, 4, 6, 7, 2 when ordered becomes 2, 2, 3, ③, ④, 5, 6, 7.
    The distribution has $n = 8$ datum. The median is equal to half the sum of the data ranked 4th $\left(\frac{n}{2}\right)$ and 5th $\left(\frac{n}{2} + 1\right)$.
    We have: $M_d = \frac{3+4}{2} = 3.5$.

    If *n* is even, the median is half the sum of the data ranked $\frac{n}{2}$ and $\frac{n}{2} + 1$.

  In a distribution of data, the number of data less than the median is equal to the number of data greater than the median.

1. The weekly salaries (in $) of a company's 25 employees are recorded. The salaries are ranked in increasing order.

450, 455, 455, 460, 460, 460, 465, 465, 465, 465, 465, 465, 475, 475, 480, 480, 485, 495, 500, 505, 505, 525, 545, 565, 565

a) Calculate and interpret

1. the mean. ***The employees' average weekly salary is $485.20.***
2. the mode. ***$465. The most frequent salary is $465.***
3. the median. ***$475. 50% of the employees make $475 or less.***

b) How can we explain the gap between the mean and the median?

The mean is greater than the median because certain employees (the last three) have salaries significantly higher than the others.

2. In one term, Jessica received the following marks in mathematics:

70, 55, 75, 85, 95

a) What is Jessica's term average in math? ***76***

b) What is her term average if

1. we exclude the lowest mark? ***81.25***
2. we exclude the highest mark? ***71.25***
3. we exclude the mark nearest the average? ***76.25***

c) Are the following statements true or false?

1. Excluding the lowest mark has the effect of increasing the average? ***True***
2. Excluding the highest mark has the effect of lowering the average? ***True***
3. The average is more affected by the data at the extremes (highest or lowest data) than the data nearest to the average. ***True***

d) 1. What is this distribution's median? ***75***

2. Is the median of a distribution affected by data at the extremes of the distribution? ***No***

3. Eric writes 5 tests and receives 5 different marks. He receives respectively for the first 4 tests the following marks: 74, 70, 86, 68.

a) What is the result of the 5th test if the average of the 5 tests is equal to 76? ***82***

b) What is the result of the 5th test if the median of the 5 tests is equal to 70 and the lowest mark remains 68? ***69***

4. A company has two branches, located in Quebec and Ontario. The table on the right indicates the number of employees at each branch and their average salary. What is the average salary of this company's employees?

$48 000

Province	Number of employees	Average Salary
Quebec	25	$47 600
Ontario	20	$48 500

ACTIVITY 4 Measures of central tendency for grouped data

The number of people living in each of a building's apartments is recorded. Here is the data given in no particular order.

2	3	4	3	5	4	3	6	5	4
4	5	4	3	2	4	5	4	2	4

a) Calculate and interpret

1. the mean. ***On average, there are 3.8 people living in each apartment.***
2. the mode. ***Most often, there are 4 people per apartment.***
3. the median. ***In at least 50% of the apartments, there are 4 people or less living there.***

b) Arrange the data in the table on the right.

c) 1. Verify that the sum of the frequencies is equal to the number n of data in the distribution.

Sum of the n_i = n = 20.

2. In this situation, what does the sum of the $n_i x_i$ represent

The total number of people living in this building.

3. Using the total of the 2nd and 3rd columns, calculate the mean of the data.

$\bar{x} = \frac{76}{20} = 3.8$

Number of people x_i	Frequency n_i	$n_i x_i$
2	***3***	***6***
3	***4***	***12***
4	***8***	***32***
5	***4***	***20***
6	***1***	***6***
Total	***20***	***76***

d) Explain how to determine the mode when the data is grouped in a table, and find the mode of this distribution.

We look for the value x_i with the highest frequency. Here, the highest frequency is 8. Thus, we have the Mode = 4.

ACTIVITY 5 Weighted mean

	HW 1 (10%)	Exam 1 (20%)	HW 2 (10%)	Exam 2 (20%)	Exam 3 (40%)	Final
Andy	75	70	85	75	80	[redacted]
Julie	70	75	90	70	[redacted]	81

In a college math course, there are two homework assignments each worth 10% of the final mark, two exams worth 20% each and a 3rd exam worth 40%

a) What is Andy's final mark in this course? ***77***

b) What must Julie's 3rd exam result be if her final mark in this math course is 81? ***90***

5. We record the number of spelling mistakes on a vocabulary test for a group of students. The data has been grouped in the table on the right

Number of mistakes x_i	Frequency n_i	$n_i x_i$
0	2	***0***
1	5	***5***
2	6	***12***
3	7	***21***
4	3	***12***
5	2	***10***
Total	***25***	***60***

a) 1. What is total number n of students who wrote this test? ***n = 25***

2. Calculate and interpret the mean. ***$\bar{x} = 2.4$***

On average, there are 2.4 mistakes on each test.

b) Determine and interpret

1. the mode. ***Mo = 3. Most often, there are 3 mistakes per copy.***
2. the median ***Md = 2. There are 2 mistakes or less on at least 50% of the copies.***

6. We record the number of defects per item from a sample size of 50 items taken from a machine's production line.

Calculate and interpret

Number of defects x_i	Frequency n_i	$n_i x_i$
0	15	***0***
1	10	***10***
2	8	***16***
3	7	***21***
4	6	***24***
5	4	***20***
Total	50	***91***

a) the mean. ***1.82***

On average, we found 1.82 defects per item.

b) the median. ***1.5***

50% of the items have less than 1.5 defects.

c) the mode. ***0***

Most often, there were no defects on the item.

7. A light fixtures company tests the lifespan (in hrs) of the electric light bulbs it produces. A random sample of 50 bulbs is taken.

When the data is not given and grouped in classes, we use the centre of each class to approximate the measures of central tendency.

Lifespan (h) classes	Centre c_i	Frequency n_i	$n_i c_i$
[80 – 120[	100	*4*	*400*
[120 – 160[	140	*8*	*1120*
[160 – 200[	180	*10*	*1800*
[200 – 240[	220	*14*	*3080*
[240 – 280[	260	*8*	*2080*
[280 – 320[	300	*6*	*1800*
Total		50	*10 280*

a) Give an approximation of the mean.
$\bar{x} \approx$ ***205.6 h***

b) 1. In which class is the median located?
The class [200 – 240[

2. We approximate the median using the centre of the medial class. Give an approximation of the median. ***Md ≈ 220 h***

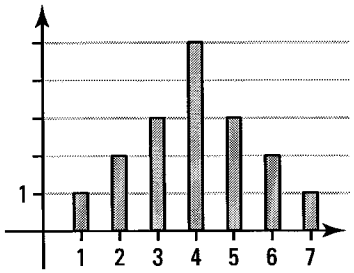
c) 1. The modal class is the class with the highest frequency.
What is the modal class? ***The class [200 – 240[***

2. The centre of the modal class is called the "raw mode". What is the raw mode? ***220 h***

8. **a)** The following distribution contains a single mode.

1 2 2 3 3 3 4 4 4 4 4 5 5 5 6 6 7

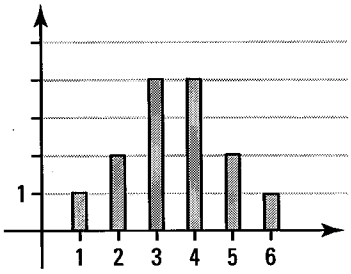
1. Is this distribution symmetrical? ***Yes***
2. Verify that mean = mode = median.
Indeed, mean = mode = median = 4.



b) The following distribution is bimodal, since it has the two modes 3 and 4.

1 2 2 3 3 3 4 4 4 5 5 6

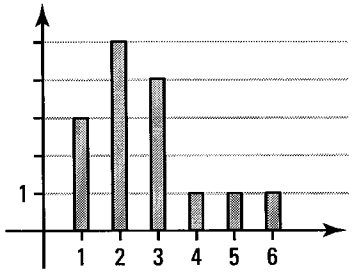
1. Is this distribution symmetrical? ***Yes***
2. Verify that $\bar{x} = M_d$.
Indeed, mean = median = 3.5



c) The following distribution is asymmetrical. The asymmetry is said to be positive because the distribution is skewed to the right.

1 1 1 2 2 2 2 2 3 3 3 3 4 5 6

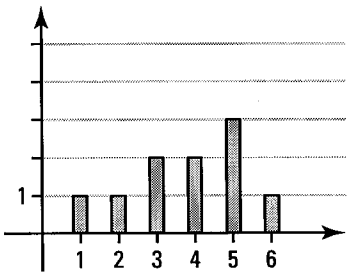
Verify that $\bar{x} > M_d$ when the asymmetry is positive.
Indeed, mean = 2.6̄ and median = 2



d) The following distribution is asymmetrical. The asymmetry is said to be negative because the distribution is skewed to the left.

1 2 3 3 4 4 5 5 5 6

Verify that $\bar{x} < M_d$ when the asymmetry is negative.
Indeed, mean = 3.8 and median = 2



9.6 Measures of position: Quartiles

Activity 1 Quartiles

The 25 marks of a mathematics test were placed in increasing order.

45	50	50	55	60
62	64	65	68	70
70	72	74	76	78
80	80	82	84	86
86	88	90	92	98

a) What is the median M_d of this distribution? ***$M_d = 74$***

b) Find the median of the data less than the median M_d of the distribution. ***63***

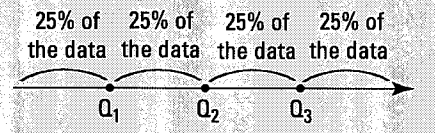
c) Find the median of the data greater than the median M_d of the distribution. ***85***

d) Give an interpretation of the measures of position found in a), b) and c).

1. ***Approximately 50% of the marks are less than or equal to 74.***
2. ***Approximately 25% of the marks are less than or equal to 63.***
3. ***Approximately 75% of the marks are less than or equal to 85.***

QUARTILES

- A **measure of position** enables you to establish the position of a given datum compared to the other data in the distribution.
- The **quartiles** Q_1, Q_2, Q_3 are measures of position dividing a distribution of n ordered data into four groups containing the same number of data.



Q_2 corresponds to the median M_d.

- To determine the 1st quartile Q_1, proceed as follows.
  - If $\frac{n}{4}$ is an integer, Q_1 is the average of the datum of rank $\frac{n}{4}$ and the next datum.
  - If $\frac{n}{4}$ is not an integer, round $\frac{n}{4}$ **upward** to get the rank of the datum corresponding to Q_1.
- To determine the 2nd and 3rd quartiles Q_2 and Q_3, use the same procedure considering the quotients $\frac{2n}{4}$ and $\frac{3n}{4}$.

Ex.: Consider the ordered distribution of $n = 8$ data.

2 3 5 7 8 8 9 12.

$Q_1 = 4$ $Q_2 = 7.5$ $Q_3 = 8.5$

- $\frac{n}{4} = 2$. Q_1 is therefore the average of the 2nd and 3rd data entries.
- $\frac{2n}{4} = 4$. Q_2 is therefore the average of the 4th and 5th data entries.
- $\frac{3n}{4} = 6$. Q_3 is therefore the average of the 6th and 7th data entries.

Ex.: Consider the ordered distribution of $n = 10$ data.

2 3 5 6 6 7 8 10 11 12.

$Q_1 = 5$ $Q_2 = 6.5$ $Q_3 = 10$

- $\frac{n}{4} = 2.5 \approx 3$. Q_1 is therefore the 3rd data entry. $Q_1 = 5$.
- $\frac{2n}{4} = 5$. Q_2 is therefore the average of the 5th and 6th data entries.
- $\frac{3n}{4} = 7.5 \approx 8$. Q_3 is therefore the 8th data entry. $Q_3 = 10$.

1. For each of the following distributions, complete the following table.

		Size n	1st quartile Q_1	2nd quartile Q_2	3rd quartile Q_3
a)	10, 20, 30, 40, 50, 60	***6***	***20***	***35***	***50***
b)	10, 20, 30, 40, 50, 60, 70	***7***	***20***	***40***	***60***
c)	10, 20, 30, 40, 50, 60, 70, 80	***8***	***25***	***45***	***65***
d)	10, 20, 30, 40, 50, 60, 70, 80, 90	***9***	***30***	***50***	***70***
e)	10, 20, 30, 40, 50, 60, 70, 80, 90, 100	***10***	***30***	***55***	***80***

2. Here is a distribution of sixteen data given in increasing order. The data corresponds to the marks received by sixteen students on a test.

45, 48, 58, 64
68, 70, 72, 72
76, 82, 84, 88
88, 90, 90, 95

a) Calculate and interpret.

1. $Q_1 =$ ***66.*** ***25% of the students got less than 66.***
2. $Q_2 =$ ***74.*** ***50% of the students got less than 74.***
3. $Q_3 =$ ***88.*** ***At least 75% of the students got less than 88.***

b) If the teacher decides to add 5 marks to each student's mark, what happens to the quartiles?

Each quartile increases by 5 marks.

3. We questioned 50 students from a school to determine the number of movies they have seen within the last month. Calculate and interpret

Number of movies	0	1	2	3	4	5
Frequency	10	8	16	8	6	2

a) the mean ***$\bar{x} = 1.96$. On average, the students have seen 1.96 movies that month.***

b) the mode. ***Mo = 2. Most often, a student has see two movies that month.***

c) the 1st quartile. ***$Q_1 = 1$. At least 25% of the students have seen one movie or less.***

d) the 2nd quartile. ***$Q_2 = 2$. At least 50% of the students have seen two movies or less.***

e) lthe 3rd quartile.. ***$Q_3 = 3$. At least 75% of the students have seen three movies or less.***

4. An English teacher tells the forty students in his class that the 1st quartile is equal to 64, the median is equal to 70 and the 3rd quartile is equal to 78. What is the maximum number of students that have a mark less than Francesca if she got

a) 62? ***10 students*** b) 69? ***20 students*** c) 76? ***30 students***

9.7 Measures of dispersion

Activity 1 Variation interval – Range

We recorded, at noon, the temperature during the first 10 days of March in the cities of Montreal and Toronto.

Montreal: −15°, −12°, −10°, −5°, 0°, 2°, 5°, 5°, 10°, 10°

Toronto: −5°, −5°, −4°, −3°, −2°, 0°, 1°, 1°, 2°, 5°

a) Verify that the average temperature was the same in Montreal and Toronto during the first 10 days of March. ***The temperature was −1° in both cities.***

b) Did we observe the same variation in temperatures in the two cities? In other words, are the data dispersed in the same way?

No, the temperature variation is greater in Montreal than in Toronto.

c) Let X_{min} designate the lowest datum and X_{max} the highest datum. Determine X_{min} and X_{max} in

1. Montreal. $X_{min} = -15°$; $X_{max} = 10°$
2. Toronto. $X_{min} = -5°$; $X_{max} = 5°$

d) The interval $[X_{min}, X_{max}]$ is called the "**variation interval**". What is the variation interval in

1. Montreal? *[−15°, 10°]*
2. Toronto? *[−5°, 5°]*

e) The length of the variation interval $(X_{max} - X_{min})$ is called the **range**. What is the range of the temperature distribution in

1. Montreal? *25°*
2. Toronto? *10°*

f) Is it true to state that the more dispersed the data is, the greater the range will be? ***Yes***

Activity 2 Interquartile interval – Interquartile range

Two groups of students wrote the same mathematics test. The quartiles in each group are indicated on the right.

Group	Q_1	Q_2	Q_3
A	70	75	80
B	65	75	85

a) Notice that both groups have the same median of 75. Can we say that the data are dispersed in the same way?

No, the data located at the centre of the distribution (50% of the data) are much more dispersed in group B than in group A.

b) The interval $[Q_1, Q_3]$ is called the "**interquartile interval**". What is the interquartile interval in

1. group A? *[70, 80]*
2. group B? *[65, 85]*

c) The length of the interquartile interval $(Q_3 - Q_1)$ is called the "**interquartile range**". What is the interquartile range in

1. group A? *10*
2. group B? *20*

d) Theoretically, what is the percentage of data located in the interquartile interval? ***50%***

e) 1. Is it true to say that the more homogeneous a group is, the lower the interquartile range will be? ***Yes***

2. Which one of the two groups is more homogeneous? ***Group A***

MEASURES OF DISPERSION

- A **measure of dispersion** indicates how the data is dispersed in a distribution.
- The **variation interval** is the interval with the lowest (X_{min}) and highest (X_{max}) data as extremities.

$$[X_{min}, X_{max}]$$

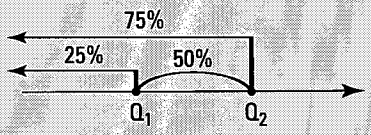
100% of the data are located in the variation interval.

- The **range** of the distribution, noted R, is equal to the length of the variation interval.

$$R = X_{max} - X_{min}$$

The range measures the dispersion of the data. The lower the range, the more **homogeneous** the data is.

- The **interquartile interval** is the interval with the 1st quartile Q_1 and the 3rd quartile Q_3 as extremities. It is noted: $[Q_1, Q_3]$.
  Approximately 50% of the data is located in the interquartile interval.



- The **interquartile range**, noted I, is equal to the length of the interquartile interval.

$$I = Q_3 - Q_1$$

The interquartile range measures the dispersion of the central data. The lower the interquartile range, the more homogeneous the data is.

Ex.: A statistical survey of the weekly salaries of a company's employees reveals the following results.
$X_{min} = \$425$; $Q_1 = \$575$; $Q_3 = \$750$; $X_{max} = \$875$
We have:
– The variation interval: [425, 875]; Range = \$450.
– The interquartile interval: [575, 750]; Interquartile range = \$175.
We notice that:
– 100% of the employees earn a salary between \$425 and \$875.
– 50% of the employees earn a salary between \$575 and \$750.

1. We asked some secondary 3 students how many hours of television they watched in the last week. Here are the answers in order (hours): 5, 6, 8, 8, 8, 10, 10, 12, 12, 16.

a) Determine

1. the mean. ***9.5 h*** 2. the mode. ***8 h*** 3. the median. ***9 h***
4. the 1st quartile. ***8 h*** 5. the 3rd quartile. ***12 h***

b) Determine

1. the variation interval. ***[5, 16]*** 2. the range. ***11 h***
3. the interquartile interval. ***[8, 12]*** 4. the interquartile range. ***4 h***

2. For the Providay company, we learn that 25% of the employees earn a weekly salary of less than \$375 and that 75% of the employees earn a salary of less than \$460.

a) Determine

1. the interquartile interval. *[375, 460]*
2. the interquartile range. *I = \$85*

b) For the Provinight company, we learn that 25% of the employees earn a weekly salary of less than \$400 and that 25% of the employees earn a salary of more than \$470. Which of the two companies has a more homogeneous salary distribution?

Provinight

3. A teacher experiments with two different pedagogical approaches to achieve one of the course's objectives, each on a group of ten students. Both groups are then given the same test. Here are the marks:

Group A: 40, 50, 60, 70, 70, 70, 70, 85, 90, 95.

Group B: 60, 60, 60, 65, 70, 70, 75, 75, 75, 90.

a) For each group, determine

1. the mean. *Group A: $\bar{x}$ = 70; group B: $\bar{x}$ = 70*
2. the median. *Group A: Md = 70; group B: Md = 70*
3. the 1st quartile. *Group A: Q_1 = 60; group B: Q_1 = 60*
4. the 3rd quartile. *Group A: Q_3 = 85; group B: Q_3 =75*

b) For each group, determine

1. the variation interval. *Group A: [40, 95]; group B: [60, 90]*
2. the range. *Group A: R = 55; group B: R = 30*
3. the interquartile interval. *Group A: [60, 85]; group B: [60, 75]*
4. the interquartile range. *Group A: I = 25; group B: I = 15.*

c) Comment on the results obtained in a) and b).

The measures of central tendency (mean, median) are the same in each group which is not the case for the measures of dispersion (range, interquartile range). The marks are more dispersed in group A than in group B. The distribution of marks is much more homogeneous in group B than in group A.

4. A distribution contains 7 numbers and has the following characteristics: $X_{min} = 3$; $Q_2 = 9$; $Q_3 = 15$; R = 15; I = 10 and $\bar{x} = 10$.

What are the numbers if $x_3 = 8$? *3, 5, 8, 9, 12, 15, 18*

5. **a)** For a distribution of three numbers, we observe: $\bar{x} = 4$, $x_{max} = 10$ and R = 15.

What is this distribution? *-5, 7, 10*

b) Determine the range of a distribution that is symmetrical and if $\bar{x} = 70$ and $x_{max} = 95$.

R = 50

c) For a symmetrical distribution of 5 numbers, we observe: $\bar{x} = 0$, $x_{max} = 20$ and $Q_3 = 10$.

What is this distribution? *-20, -10, 0, 10, 20*

9.8 Box-and-whisker plots

ACTIVITY 1 Illustrating quartiles

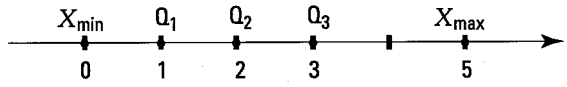
The number of goals scored on a team's goalie during the first 15 games of the season is recorded. The data is placed in increasing order.

0, 0, 0, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 4, 5

a) Determine

1. X_{min}. ***0*** 2. Q_1. ***1*** 3. Q_2. ***2*** 4. Q_3. ***3*** 5. X_{max}. ***5***

b) Locate on the scaled axis the results obtained in a).



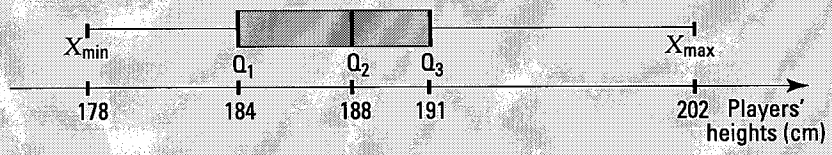
BOX-AND-WHISKER PLOTS

- The box-and-whisker plot illustrates:
  - the **minimum**, the **maximum** and therefore the range of a distribution;
  - the **first quartile**, the **third quartile** and therefore the **interquartile range**;
  - the median of the distribution.

Ex.: We recorded the heights X (in cm) of a college basketball team's players. We got:

$X_{min} = 178$ cm; $X_{max} = 202$ cm; $Q_1 = 184$ cm; $Q_2 = 188$ cm; $Q_3 = 191$ cm.

The box-and-whisker plot illustrating the distribution of the players' heights is:



In a box-and-whisker plot, we distinguish:

- a **rectangle** in place of the interquartile interval;
  The length of this rectangle is equal to the interquartile range I ($I = Q_3 - Q_1$).
- Three vertical lines called **hinges** in place of Q_1, Q_2 and Q_3.
- The left branch representing the distance between Q_1 and X_{min}.
- The right branch representing the distance between Q_3 and X_{max}.

- The box-and-whisker plot gives us information on the **dispersion** or **concentration** of the data. However, we can draw no conclusions with regards to the values, the mean or the mode.

1. Consider the box-and-whisker plot of the basketball players' heights (see shaded frame on the preceding page).

a) What are all the numerical values indicated on the diagram?

X_{min} = 178, Q_1 = 184 cm, Q_2 = 188 cm, Q_3 = 191 cm and X_{max} = 202 cm

b) Calculate and interpret the interquartile range.

I = 7 cm. 50% of this team's players have a height between 184 and 191 cm.

c) The quartiles divide the distribution into 4 quarters. In which quarter do we observe

1. the greatest concentration of players? *Between the median and the 3rd quartile*
2. the least concentration of players? *Between Q_3 and X_{max}*

d) By looking at the box-and-whisker plot, indicate if it is true or false to conclude the following propositions.

1. A quarter of the players have a height between 184 and 188 cm. *True*
2. Players with a height between 188 and 191 cm are less dispersed than those between 184 and 188 cm. *True*
3. The mean of the heights is equal to 188 cm. *False*
4. There are 25 players on the team. *False*
5. Half the players have a height greater than 188 cm. *True*
6. The players with a height greater than 191 cm are much more dispersed than those with a height of less than 184 cm. *True*

2. We have recorded the number of absences over the course of the year for the three secondary 4 groups at Mozart School.

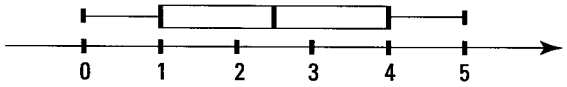
Class A

Number of absences	Frequency
0	3
1	4
2	5
3	5
4	4
5	3

a) Consider the 24 students in group A.

1. Is the data distribution symmetrical? *Yes*
2. Calculate

   Q_1 = *1* Q_2 = *2.5* Q_3 = *4*
3. Construct the box-and-whisker plot.



4. How is the distribution's symmetry reflected in the box-and-whisker plot?

   The hinge corresponding to the median divides the rectangle in two congruent rectangles. The left and right branches are equal.

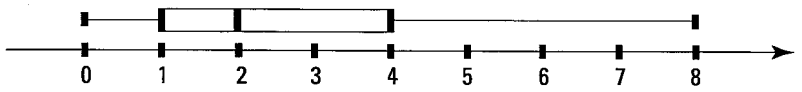
b) Consider the 25 students in group B.

Class B

Number of absences	Frequency
0	4
1	8
2	6
4	4
6	2
8	1

1. Calculate

   Q_1 = *1* Q_2 = *2* Q_3 = *4*
2. Construct the box-and-whisker plot.



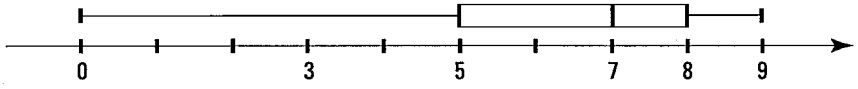
3. Do we observe any symmetry here? *No*

c) Consider the 31 students in group C.

1. Calculate

$Q_1 =$ *5* $Q_2 =$ *7* $Q_3 =$ *8*

2. Construct the box-and-whisker plot.

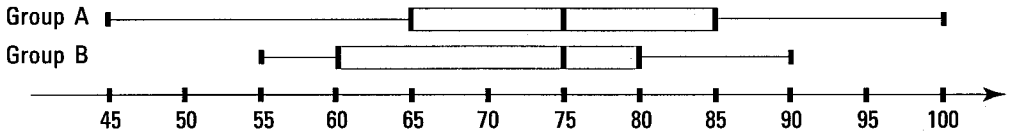


3. Do we observe any symmetry here? *No*

Class C

Number of absences	Frequency
0	2
3	2
5	6
7	7
8	9
9	5

3. The following box-and-whisker plots illustrate the results of two groups of students on a French test.



a) If the passing mark is 60, in which group do we observe the greatest failure rate? _____

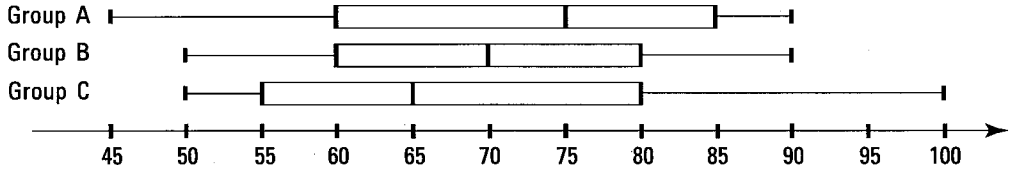
Group B

b) In which group is it more likely to have a mark

1. greater than 65? *Group A*
2. less than 90? *Group B*
3. between 75 and 80? *Group B*
4. between 65 and 75? *Group A*

c) Which group seems more homogeneous? *Group B*

4. The following box-and-whisker plots illustrate the results of three groups of students having written the same mathematics test.



a) If the passing mark is 60, what can we say about the failure rate in each group?

Group C has the highest failure rate (greater than 25%).

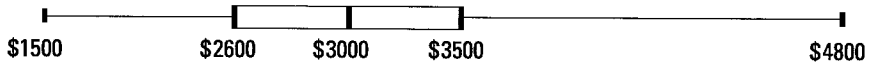
Groups A and B have the same failure rate (25%).

b) Which of the three groups has symmetry? *Group B*

c) Which of the three groups seems most homogeneous? *Group B*

d) In which group is it most likely to have a mark greater than 80? *Group A*

5. The monthly salaries of a company's 120 employees were recorded. The following box-and-whisker plot was constructed from this data.



True or false?

a) The average monthly salary of the employees is \$3000. ***False***

b) Of these employees, 35 earn a monthly salary of less than \$2600. ***False***

c) Approximately 60 employees earn a monthly salary greater than \$3000. ***True***

d) There are more employees whose monthly salary is greater than \$3500 than employees whose monthly salary is less than \$2600. ***False***

6. The students from two classes raise money for juvenile diabetes research. The following box-and-whisker plot was constructed from the amounts of money (in \$) raised by the 31 students in Mona's class.

The amount raised by Mona is the 8th highest amount in her class.

The following distribution represents, in increasing order, the amounts of money (in \$) raised by the students in Roger's class.

12	18	20	20	25	28	30	30	35	40
43	45	54	54	54	57	65	65	65	70
71	71	72	75	77	80	81	82	88	90

The amount raised by Roger is an even number between the median and the 3rd quartile

How much more did Roger raise than Mona? ***\$40***

7. Here are the results of a group's 26 students on an entrance exam.

49	54	58	58	60	61	61	62	64	66
70	71	78	79	79	83	85	86	87	88
90	90	93	94	97	98				

Karen and Alex's results on this exam reveal that
- Karen's mark is one of the quartiles.
- none of the other students in the group received the same mark as Karen.
- Alex's mark is between the median and the 3rd quartile.
- Alex's mark is an even number.

What are Karen and Alex's marks? **Karen: 88; Alex: 86**

9.9 Stem-and-leaf plots

STEM-AND-LEAF PLOTS

- Two groups of students were administered the same mathematics test. The results were written in increasing order.

Group A

35 45 48 50 55 56 58 60 60 62
64 65 65 68 68 70 70 72 75 75
77 78 80 80 82 85 88 90 95 98

Group B

40 45 48 50 50 55 58 60 60 55
65 68 68 70 70 72 72 75 75 78
78 80 80 82 85 88 90 95

- The diagram on the right, called a **stem-and-leaf plot**, enables us to better visualize the distribution of the data. The numbers located to the left of the vertical line are called stems and correspond to the tens digits of the data, whereas the number to the right of the vertical line are called leaves and correspond to the units digits of the data.

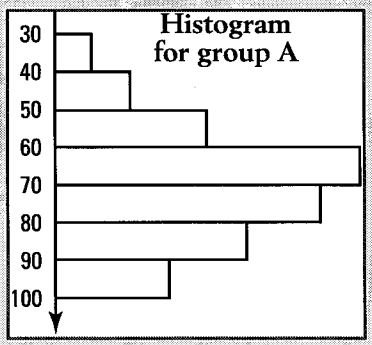
  Thus, the line 4 | 58 means that the data distribution contains the numbers 45 and 48. Note that there are as many leaves as the number of data in the distribution.

Group A results

3	5
4	5 8
5	0 5 6 8
6	0 0 2 4 5 5 8 8
7	0 0 2 5 5 7 8
8	0 0 2 5 8
9	0 5 8

- The resemblance between the stem-and-leaf plot and the histogram presented horizontally is striking.

  The stem-and-leaf plot not only gives us an idea of the dispersion of the data, but also gives us the actual values of the data.



- The juxtaposition of two stem-and-leaf plots enables us to easily compare two data distributions.

Group A results		Group B results
5	3	
8 5	4	0 5 8
8 6 5 0	5	0 0 5 8
8 8 5 5 4 2 0 0	6	0 0 5 5 8 8
8 7 5 5 2 0 0	7	0 0 2 2 5 5 8 8
8 5 2 0 0	8	0 0 2 5 8
8 5 0	9	0 5

1. The ages of a company's 25 employees are presented on the right in increasing order.

26 26 27 27 28 28 29 30 34 34
35 36 37 38 38 40 44 44 46 48
54 56 58 65 68

a) Construct a stem-and-leaf plot.

Age of employees

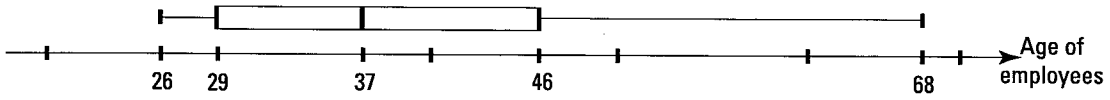
2	*6 6 7 7 8 8 9*
3	*0 4 4 5 6 7 8 8*
4	*0 4 4 6 8*
5	*4 6 8*
6	*5 8*

b) What is the range of the distribution? *42 years*

c) Calculate

1. the mean. ***39.4 yrs*** 2. the median. ***37 yrs*** 3. the 1st quartile. ***29 yrs***
4. the 3rd quartile. ***46 yrs*** 5. the interquartile range. ***17 yrs*** 6. the range. ***42 yrs***

d) Construct the box-and-whisker plot.



e) 1. Explain why the right branch is much longer than the left branch in the box-and-whisker plot.

Because there are a lot of data to the far right of the median.

2. Explain why the mean is greater than the median.

For the same reason.

2. Two groups of students were administered the same history test. Here are the results of each group.

Group A

89 98 88 40 50 60 68 70 80 85
90 92 88 38 54 62 72 74 82 86
48 58 66 76 76

Group B

69 70 58 72 58 58 65 78 69 79
50 80 56 88 62 62 76 68 68 68
60 96 60 58 74

a) Represent this situation with two juxtaposed stem-and-leaf plots.

Group A results		***Group B results***
8	***3***	
8 0	***4***	
8 4 0	***5***	***0 6 8 8 8 8***
8 6 2 0	***6***	***0 0 2 2 5 8 8 8 9 9***
6 6 4 2 0	***7***	***0 2 4 5 6 8***
9 8 8 6 5 2 0	***8***	***0 8***
8 2 0	***9***	***6***

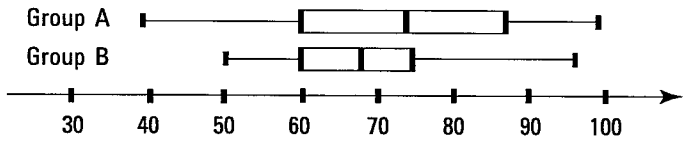
b) For each distribution, calculate

1. the range. ***Group A: 60; group B: 46***
2. the mean. ***Group A: 71.6; group B: 67.9***

c) For each distribution, calculate

1. the median. ***Group A: 74; group B: 68***
2. the 1st quartile ***Group A: 60; group B: 60***
3. the 3rd quartile. ***Group A: 86; group B: 74***
4. the interquartile range. ***Group A: 26; group B: 14***

d) For each group, construct a box-and-whisker plot juxtaposing them to compare the two distributions of marks.



e) Comment on the results.

There is a greater dispersion of marks in group A. Half of the students in group A have a mark between 60 and 86, whereas half of the students in group B have a mark between 60 and 74. The interquartile range of groups A is almost double that of group B.

EVALUATION 9

1. For each of the following situations, indicate

1. if it is a census, a poll or a study.
2. the population being surveyed.
3. the variable being studied.
4. the type of variable (qualitative, quantitative discrete, quantitative continuous)

a) The mother tongue of all a company's employees are recorded.

1. ***Census*** 2. ***All of the company's employees***
3. ***Mother tongue*** 4. ***Qualitative variable***

b) The first twenty students exiting a school are asked the number of times they have been to the cinema since the beginning of the year.

1. ***Poll*** 2. ***The school's entire student population***
3. ***Number of times to the cinema*** 4. ***Quantitative discrete***

c) We ask a doctor to determine if a new drug on the market increases the chances of curing hay fever or not.

1. ***Study*** 2. ***The set of patients***
3. ***The drug's effect*** 4. ***Qualitative variable***

d) The duration of the overtime period was recorded for all of the games in the last Stanley Cup playoffs.

1. ***Census*** 2. ***All of the playoff games***
3. ***Duration of the overtime period*** 4. ***Quantitative continuous***

2. The marks for a group of 25 students on a mathematics test are placed in increasing order.

45	52	58	60	64	66	68	70	70	74	75	78	78
78	78	80	82	84	84	85	88	90	95	95	98	

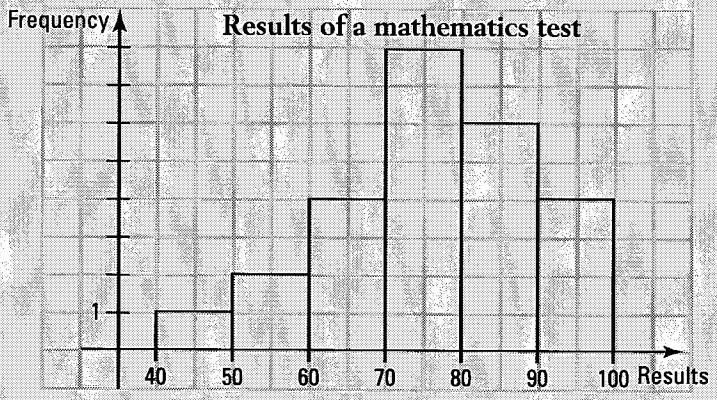
a) Determine and interpret

1. the mean. ***75.8. On average, the students scored 75.8 on the test.***
2. the median. ***78. Approximately half of the group's students scored 78 or less.***
3. the mode. ***78. The most frequently observed mark is 78.***

b) Group the data into 6 classes of amplitude 10.

	Class	Tally	Frequency	Rel. freq. (%)
1	*[40–50[*	|	***1***	***4***
2	*[50–60[*	||	***2***	***8***
3	*[60–70[*	||||	***4***	***16***
4	*[70–80[*	𝍸 |||	***8***	***32***
5	*[80–90[*	𝍸 |	***6***	***24***
6	*[90–100[*	||||	***4***	***16***
	Total		***25***	***100***

c) Construct a histogram representing the students' marks.



3. True of false?

a) In stratified sampling, we typically see little variation within the strata, but more variation from one stratum to another. ***True***

b) In cluster sampling, we typically see a lot of variation within the clusters, but little variation from one cluster to another. ***True***

c) In stratified sampling, the number of individuals chosen for each stratum is proportional to the weight of that stratum in the population. ***True***

d) In cluster sampling, we choose all the individuals belonging to the randomly chosen cluster. ***True***

4. We record the number of passengers in each vehicle entering a shopping centre's parking lot. We want a sample size of 50. Indicate if the sampling technique in the survey is random or systematic.

a) We randomly choose 50 numbers from a random number table. Each one of these numbers represents the car's number that we will record its number of passengers that day. (The car numbered 1 represents the first car of the day entering the parking lot.)

Random sampling

b) We record the number of passengers in the 3rd car, the 13th car, the 23rd car, ... until we get a sample of size 50.

Systematic sampling

5. To determine the favorite school activity of the secondary 3 students, we decide to ask 30 random students among the secondary 3 students. In this school, there are 5 groups of 30 secondary 3 students.

In each case, indicate the sampling technique.

a) We ask every student from one randomly chosen class.

Cluster sampling

b) In each group, we ask 6 students chosen at random.

Stratified sampling

c) After placing all the secondary 3 students in alphabetical order, we ask the 2nd student on the list, the 7th, the 12th ... until we get a sample of size 30.

Systematic sampling

d) After numbering the students 1 to 150, we randomly choose 30 numbers from 1 to 150 to determine the students who will form the sample.

Random sampling

6. In a class of 30 students, we recorded the total number of absences from mathematics class for each student throughout the school year.

1	0	2	1	3	2	0	1	0	1	2	1	0	1	5
0	0	0	1	1	4	0	2	2	3	0	0	1	0	2

Calculate and interpret

a) the mean. ***1.2. The average number of absences for each student is equal to 1.2.***

b) the mode. ***0. Most often, students did not miss any mathematics classes.***

c) the median. ***1. At least 50% of the students (actually 2/3) are absent once or less from their mathematics class.***

7. In a school, there are 5 groups in secondary 3. In the first four groups, we count 26, 30, 30, and 31 students. How many students are there in the 5th group if the average number of students per group is equal to 29?

There are 28 students in the 5th group

8. Sabrina is taking a university mathematics class. In this course, each student must do two homework assignments, a mid-term exam and a final exam. The weighting is 10% for each homework assignment, 30% for the mid-term exam and 50% for the final exam.

a) What is Sabrina's average in this course if she got 75 on the first homework, 85 on the second, 80 on the mid-term exam and 78 on the final exam? ***79***

b) Any student achieving an average of 85 or more is given a letter grade of A for the course. What is the minimum mark Sabrina would have needed on her final exam to get a letter grade of A? ***90***

9. We have recorded in order the weekly salaries of a company's 25 employees.

380, 380, 400, 400, 420, 430, 440, 450, 460, 460, 480, 480, 490, 500, 510, 520, 530, 540, 540, 580, 600, 650, 700, 850

a) Calculate and interpret

1. the median. ***$485. 50% of the employees earn a salary of less than $485.***
2. the 1st quartile. ***$435. 25% of the employees earn a salary of less than $435.***
3. the 3rd quartile. ***$540. 75% of the employees earn a salary of less than $540.***

b) Determine and interpret

1. the variation interval. *[380, 850]*

All the employees (100%) earn a salary located in the interval [380, 850].

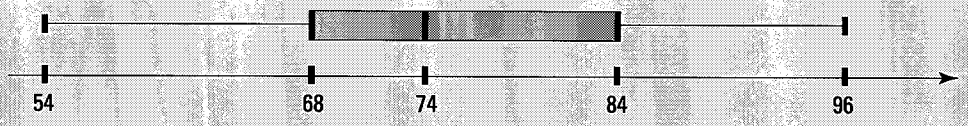
2. the interquartile interval. *[435, 540]*

Approximately half of the employees earn a salary in the interval [435, 540].

c) Determine

1. the range. *$470* 2. the interquartile range. *$105*

10.



The box-and-whisker plot above illustrates the students' results on a French test.

a) Determine and interpret

1. the median. *74. Approximately 50% of the students got less than 74.*
2. the 1st quartile. *68. Approximately 25% of the students got less than 68.*
3. the 3rd quartile. *84. Approximately 75% of the students got less than 84.*

b) True or false?

1. There are more students who received a mark between 74 and 84 than between 68 and 74. *False*
2. There is a larger concentration of students between 68 and 74 than between 74 and 84. *True*
3. The box-and-whisker plot enables you to calculate the mean. *False*
4. There are approximately as many students receiving a mark less than 68 as there are students receiving a mark greater than 84. *True*