How to get information from a data set ?

Elise Arnaud elise.arnaud@univ-grenoble-alpes.fr

How to get information from a data set ?

- to make statistics supposes that one studies a set of equivalent objects on which one observes characteristics called variables.

How to get information from a data set ?

- to make statistics supposes that one studies a set of equivalent objects on which one observes characteristics called variables.
- the group or set of equivalent objects is called the population.

How to get information from a data set ?

- to make statistics supposes that one studies a set of equivalent objects on which one observes characteristics called variables.
- the group or set of equivalent objects is called the population.
- objects are called individuals

How to get information from a data set ?

- to make statistics supposes that one studies a set of equivalent objects on which one observes characteristics called variables.
- the group or set of equivalent objects is called the population.
- objects are called individuals
- In general, the population is too large to be observed exhaustively. One then studies the variable on a subset of the population. A sample is studied.

Vocabulary

We wish to study a characteristic X on a population $\mathcal{P}$.

For example, the gender, the number of coffees consumed in a week, the weight or the height of a M2 student.

X takes its values in Ω.

Vocabulary

In general, we cannot observe this characteristic on all the individuals of a large population, but only on a sub-population of $\mathcal{P}$ of size n. We will note :

Vocabulary

In general, we cannot observe this characteristic on all the individuals of a large population, but only on a sub-population of $\mathcal{P}$ of size n. We will note :

the sub population : $\{i_1, ..., i_j, ..., i_n\}$ a set of n individuals chosen randomly in $\mathcal{P}$.

Vocabulary

In general, we cannot observe this characteristic on all the individuals of a large population, but only on a sub-population of $\mathcal{P}$ of size n. We will note :

the sub population : $\{i_1, ..., i_j, ..., i_n\}$ a set of n individuals chosen randomly in $\mathcal{P}$.

the data sample : $x_1, ..., x_j, ..., x_n$ the n observed values of the characteristic X on the individuals on the sub population

Vocabulary

In general, we cannot observe this characteristic on all the individuals of a large population, but only on a sub-population of $\mathcal{P}$ of size n. We will note :

the sub population : $\{i_1, ..., i_j, ..., i_n\}$ a set of n individuals chosen randomly in $\mathcal{P}$.

the data sample : $x_1, ..., x_j, ..., x_n$ the n observed values of the characteristic X on the individuals on the sub population

Two problems then arise:

1. What information about the character X can be drawn from the sample ?
2. What prediction could be made about an unobserved individual of $\mathcal{P}$ from the observed data $x_1, ..., x_j, ..., x_n$?

Vocabulary, types of variables

Each individual is described by a set of variables X. These variables can be classified according to their nature:

Vocabulary, types of variables

Each individual is described by a set of variables X. These variables can be classified according to their nature:

- a qualitative variable is a variable that isn't numerical. It describes data that fits into categories
  Ω={Woman, Man, Other} ; Ω={happy, not so happy, not happy}
  - qualitative nominal values, ex : color, gender
  - qualitative ordinal values, ex : ranking, opinion

Vocabulary, types of variables

Each individual is described by a set of variables X. These variables can be classified according to their nature:

- a qualitative variable is a variable that isn't numerical. It describes data that fits into categories
  Ω={Woman, Man, Other} ; Ω={happy, not so happy, not happy}
  - qualitative nominal values, ex : color, gender
  - qualitative ordinal values, ex : ranking, opinion

- a quantitative variable is expressed in numbers, for example the size of individuals or the results of a test.

Vocabulary, types of variables

Each individual is described by a set of variables X. These variables can be classified according to their nature:

- a qualitative variable is a variable that isn't numerical. It describes data that fits into categories
  Ω={Woman, Man, Other} ; Ω={happy, not so happy, not happy}
  - qualitative nominal values, ex : color, gender
  - qualitative ordinal values, ex : ranking, opinion

- a quantitative variable is expressed in numbers, for example the size of individuals or the results of a test.

- We distinguish discrete quantitative variables when Ω est une suite finie ou infinie d'éléments de $\mathbb{N}$ (ex : $\Omega = \{1, 2, 3\}$; $\Omega = \mathbb{N}$) from continuous quantitative variables if all values in an interval of $\mathbb{R}$ are acceptable.

Generalities on the statistical approach

The purpose of a statistical study is to answer a question in a particular field of application from a dataset. A statistical study takes place in several steps :

Generalities on the statistical approach

The purpose of a statistical study is to answer a question in a particular field of application from a dataset. A statistical study takes place in several steps :

1. Defining the protocol to be followed for data collection (experimental design, survey design, development of a questionnaire ...)

Generalities on the statistical approach

The purpose of a statistical study is to answer a question in a particular field of application from a dataset. A statistical study takes place in several steps :

1. Defining the protocol to be followed for data collection (experimental design, survey design, development of a questionnaire ...)
2. Data collection, coding, cleaning

Generalities on the statistical approach

The purpose of a statistical study is to answer a question in a particular field of application from a dataset. A statistical study takes place in several steps :

1. Defining the protocol to be followed for data collection (experimental design, survey design, development of a questionnaire ...)
2. Data collection, coding, cleaning
3. Data exploration (descriptive statistics, data analysis, data mining, ...), without trying to model them

Generalities on the statistical approach

The purpose of a statistical study is to answer a question in a particular field of application from a dataset. A statistical study takes place in several steps :

1. Defining the protocol to be followed for data collection (experimental design, survey design, development of a questionnaire ...)
2. Data collection, coding, cleaning
3. Data exploration (descriptive statistics, data analysis, data mining, ...), without trying to model them
4. Eventually, data pre-processing (recoding, aggregation, transformations, creation of new data)

Generalities on the statistical approach

The purpose of a statistical study is to answer a question in a particular field of application from a dataset. A statistical study takes place in several steps :

1. Defining the protocol to be followed for data collection (experimental design, survey design, development of a questionnaire ...)
2. Data collection, coding, cleaning
3. Data exploration (descriptive statistics, data analysis, data mining, ...), without trying to model them
4. Eventually, data pre-processing (recoding, aggregation, transformations, creation of new data)
5. If the data are from a sample, statistical modelling of the data: inferential statistics, use of a probabilistic model

Generalities on the statistical approach

The purpose of a statistical study is to answer a question in a particular field of application from a dataset. A statistical study takes place in several steps :

1. Defining the protocol to be followed for data collection (experimental design, survey design, development of a questionnaire ...)
2. Data collection, coding, cleaning
3. Data exploration (descriptive statistics, data analysis, data mining, ...), without trying to model them
4. Eventually, data pre-processing (recoding, aggregation, transformations, creation of new data)
5. If the data are from a sample, statistical modelling of the data: inferential statistics, use of a probabilistic model
6. Forecasting and/or decision making (answer to the initial question)

Data set

Initially, the data is often in the form of an array of individual data that contains:

- lines : statistical individuals
- columns : variables

A1 | sexe

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	sexe	situation	the	cafe	taille	poids	age	viande	poisson	fruit_crus	fruit_legume	chocol	matgras	
2	Femme	couple	2	0	165	69	77	1 fois/j	2	5	5	1 fois /sem	arachide	
3	Femme	seul	0	2	154	80	83	3 fois/sem	3	5	5	1 fois /sem	arachide	
4	Femme	seul	0	2	168	63	69	3 fois/sem	3	4	5	1 fois /sem	arachide	
5	Femme	couple	2	1	162	55	65	3 fois/sem	3	5	5	1 fois /sem	arachide	
6	Femme	couple	0	3	160	50	70	3 fois/sem	2	5	5	1 fois /sem	arachide	
7	Homme	couple	0	3	165	75	73	3 fois/sem	2	5	5	1 fois /sem	arachide	
8	Homme	seul	0	2	168	67	69	3 fois/sem	3	5	5	1 fois /sem	arachide	
9	Femme	seul	2	0	159	66	82	4 fois/sem	2	5	5	1 fois /sem	arachide	
10	Femme	couple	4	1	167	70	75	4 fois/sem	2	5	5	1 fois /sem	arachide	
11	Femme	couple	0	4	160	75	69	4 fois/sem	2	2	5	1 fois /sem	arachide	
12	Femme	couple	0	3	163	62	68	4 fois/sem	3	4	4	1 fois /sem	arachide	
13	Homme	couple	0	3	172	79	78	4 fois/sem	2	5	5	1 fois /sem	arachide	
14	Homme	couple	0	2	162	75	65	4 fois/sem	2	4	5	1 fois /sem	arachide	
15	Homme	couple	0	2	170	74	71	4 fois/sem	3	4	4	1 fois /sem	arachide	

Figure: Extract from the database of the file data_nutri.csv.

Descriptive statistics

Descriptive statistics refers to a set of techniques whose purpose is to

- explore, discover the information contained in the data
- represent them graphically
- detect the first trends

To each of these goals corresponds a technique:

explore the data summarize the information	statistical table statistical summaries
represent them graphically	graphs
detect the trends	link indicators

Exemple

Exemple: file data_nutri.csv

Survey on the diet of 226 elderly people in the Bordeaux region in 2000.

Source : "Le logiciel R" P. Lafaye de Micheaux, R. Drouilhet, B. Liquet.

- gender, family situation nominal qualitative variables
- caily consumption of tea, coffee (in number of cups) discrete qualitative variables
- height (in cm), weight (in kg), age (in years) on the day of the survey continuous quantitative variables
- weekly consumption of meat, fish, raw fruit, cooked fruit and vegetables, chocolatet (0 : never, 1 : $<$ 1 time, 2 : 1 time, 3 : 2 ou 3, 4 : 4 à 6 times, 5 : every day) ordinal qualitative variables
- fat preferentially used for cooking nominal qualitative variables

Qualitative variables

Statistical table

to summarise the information from the variables : statistical tables

- n sample size
- q number of modalities
- $m_i, i \in [1, q]$ modalities
- n_i (frequency) of m_i in the sample
- and f_i the corresponding relative frequency .

m_i	x_i	n_i	f_i	F_i

exemple for a nominal qualitative variable

sexe	Total	Fréquences
Femme	141	0,62
Homme	85	0,38
Total général	**226**	**1,00**

situation	Effectifs	Pourcentage
couple	119	52,65%
famille	9	3,98%
seul	98	43,36%
Total général	**226**	**100,00%**

matgras	Total	Pourcentages
arachide	48	21,24%
beurre	15	6,64%
canard	4	1,77%
colza	1	0,44%
isio4	23	10,18%
margarine	27	11,95%
olive	40	17,70%
tournesol	68	30,09%
Total général	**226**	**100,00%**

By default, in most programs, the modalities are sorted in alphabetical order.

Why such a table ?

- to check the quality of the data : one can easily see coding problems, or missing values
- to examine the distribution of the variable: is the variable spread over several modalities or on the contrary concentrated on a small number of modalities? What are the main modalities present?

barplot or Pareto chart

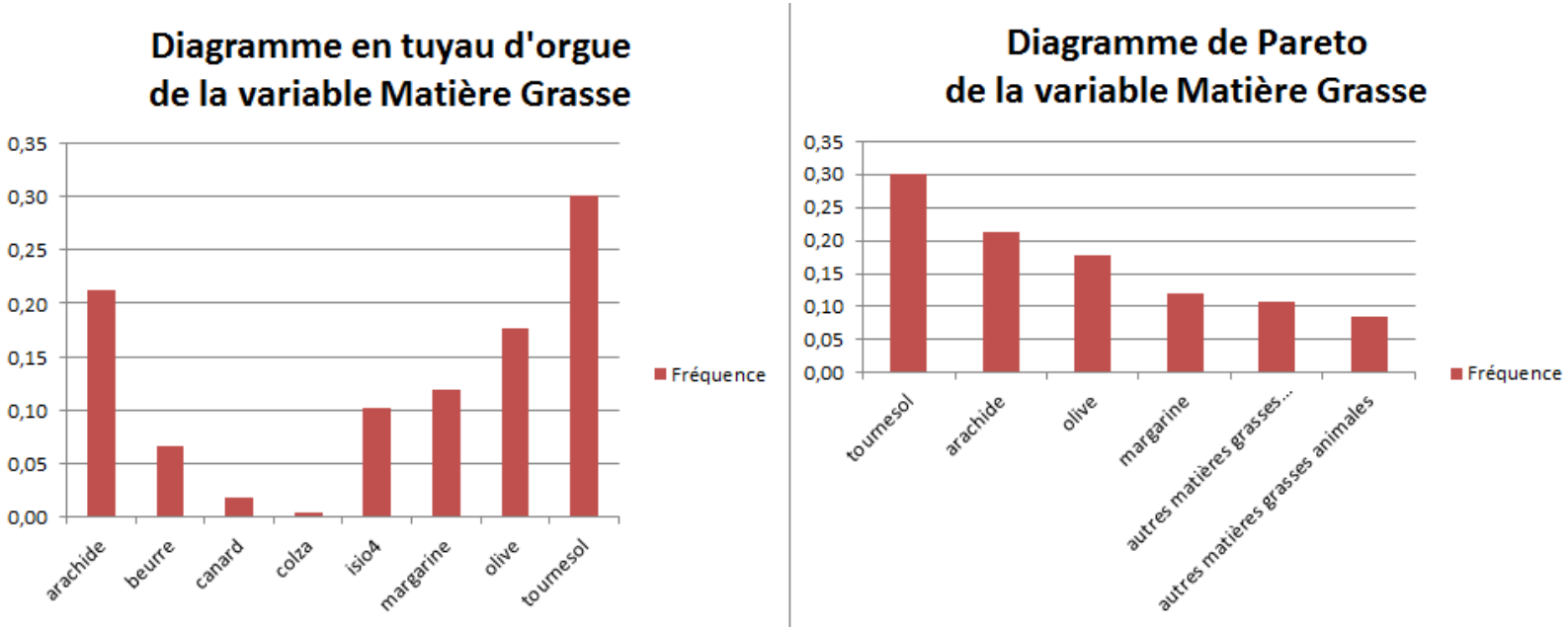
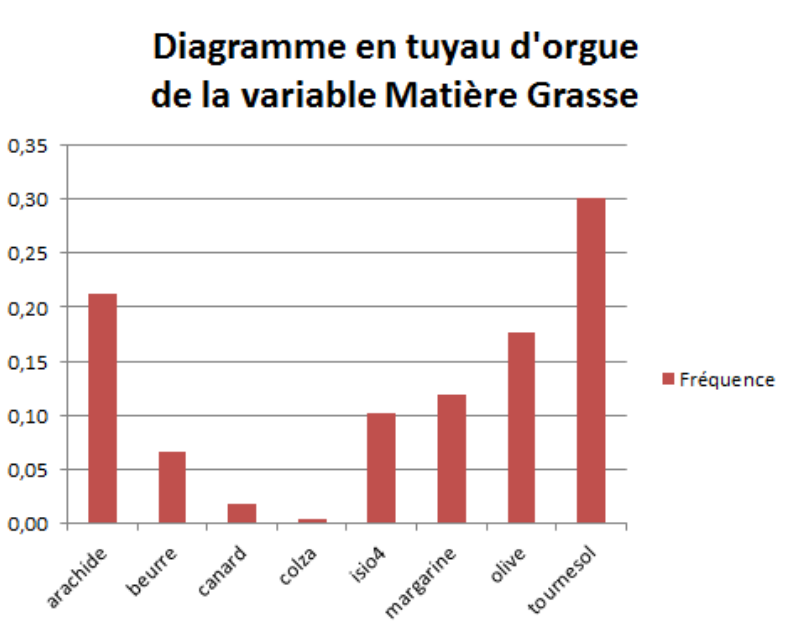


Figure: variable fat

- When there are many different modalities to represent, the Pareto chart is more readable.
- Avoid 3d diagrams

barplot or Pareto chart



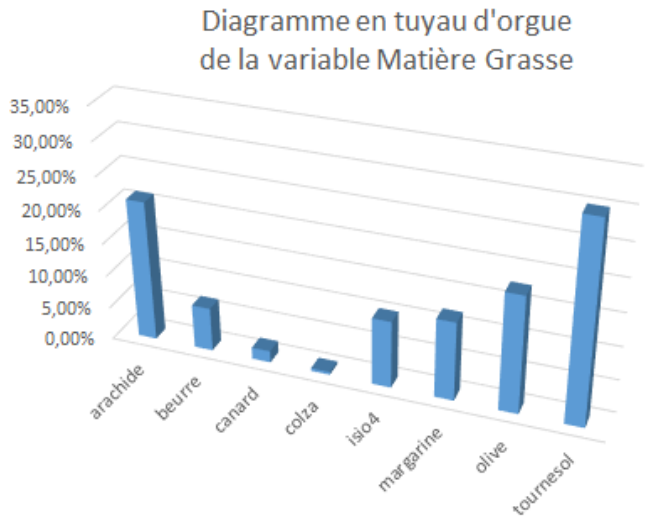


Figure: variable fat

- When there are many different modalities to represent, the Pareto chart is more readable.
- Avoid 3d diagrams

Representing nominal values

stacked chart

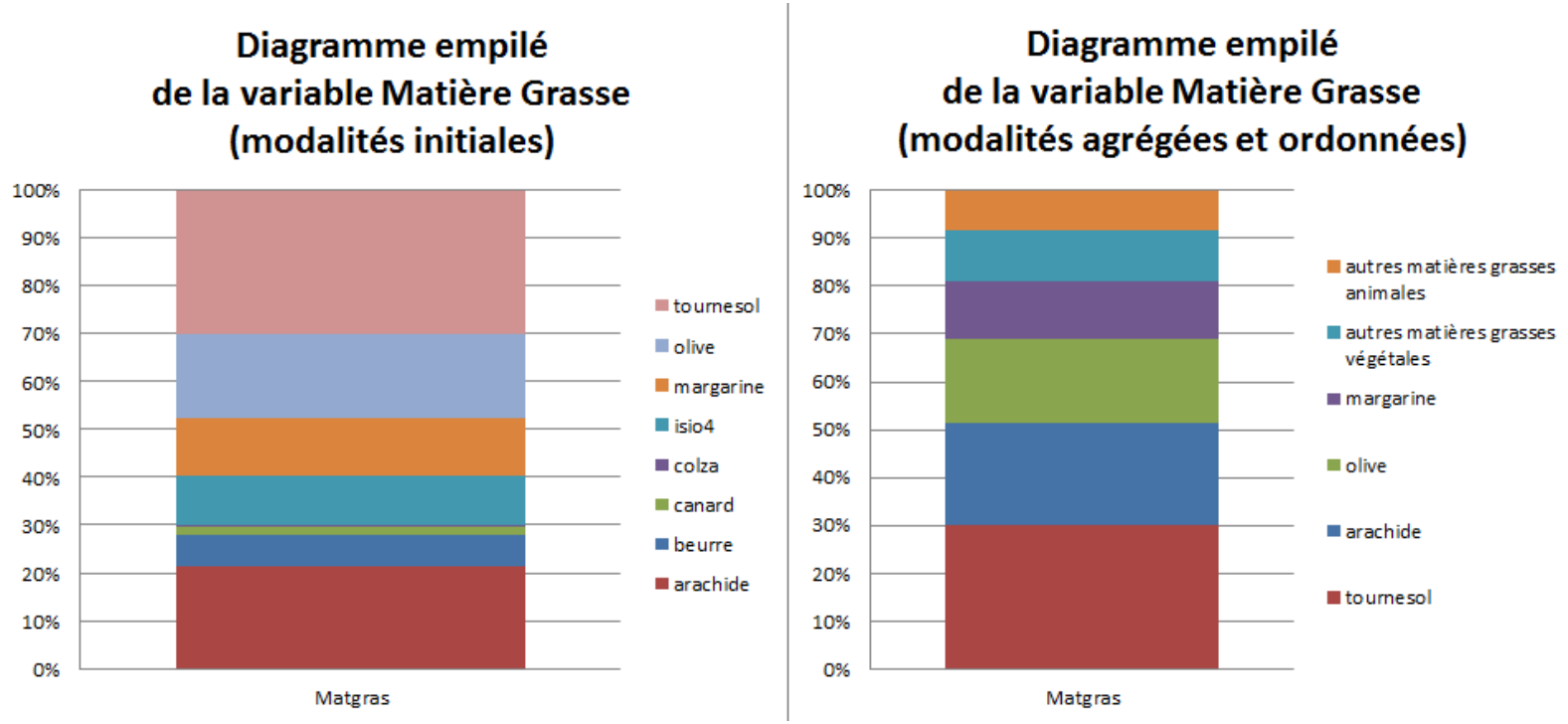


Figure: variable fat

- Also more readable if the modalities are ordered.
- Difficult to read if there are too many modalities at very low frequency.

Representing nominal values

Pie chart

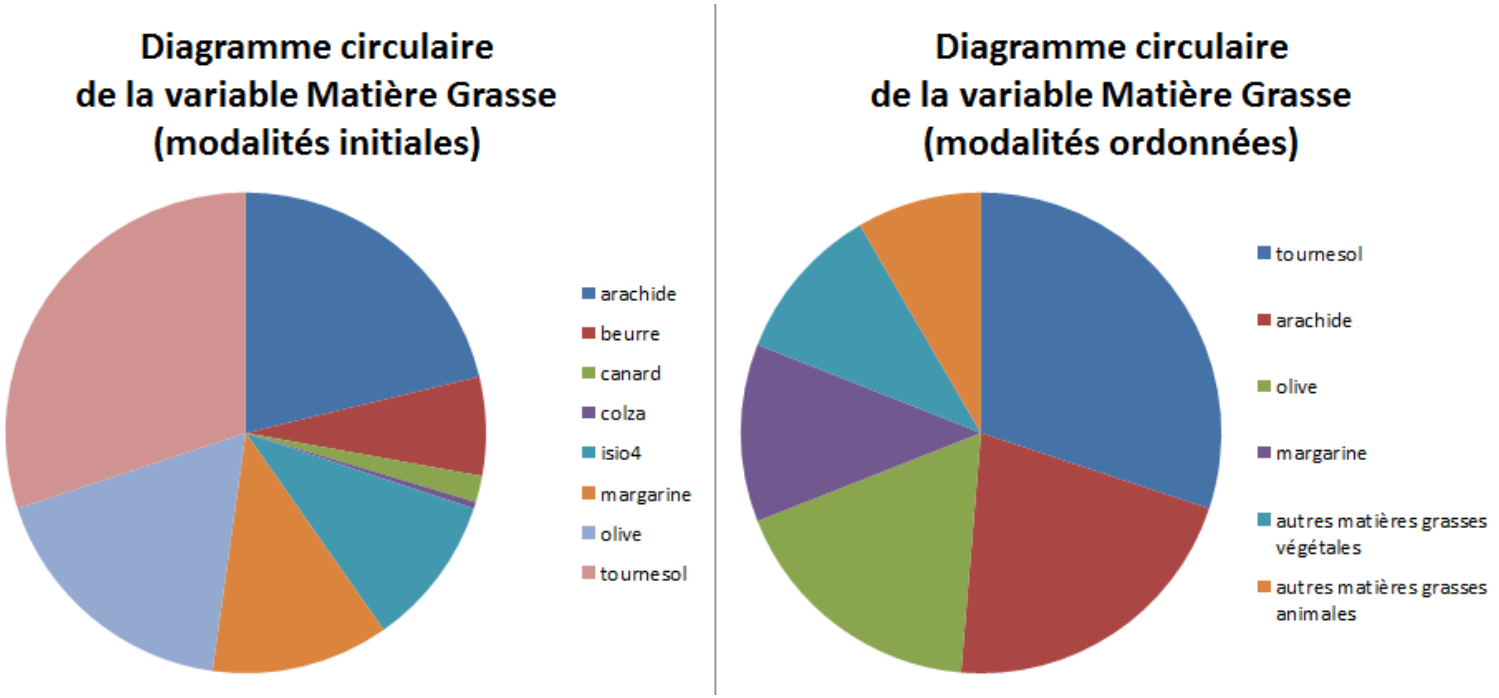


Figure: Variable fat.

- Very difficult to read if there are too many modalities
- Nice if the message is clear

Representing ordinal values

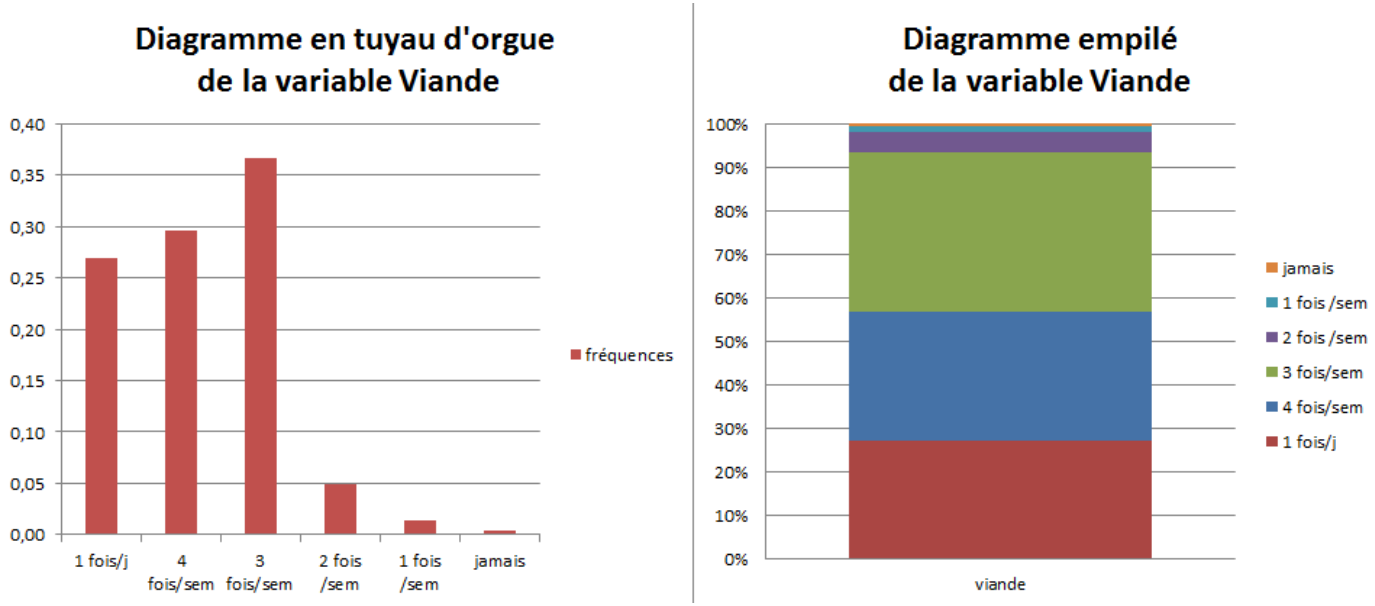


Figure: variable meet

- Avoid Pareto chart and pie chart

Quantitative variables

Quantitative variables

Where does the discrete end and the continuous begin?

Discret/continu

The discrete/continuous distinction is more about the number of repetitions of each modality than about the mathematical nature of the data.

Examples :

- number of children per family: many families will have 0, 1, 2 or 3 children $\Rightarrow$ discrete
- number of units put up for sale or sold: little chance of finding 2 quarters with exactly the same number of units $\Rightarrow$ continuous
- temperature: if recorded over time $\Rightarrow$ continuous ; if it is fixed by experience (for example, temperature of a room set at 18, 20, 22 degrees) $\Rightarrow$ discrete

statistical tables

Variable Thé	Effectifs	Fréquences	Fréquences cumulées
0	163	72,12%	72,12%
1	13	5,75%	77,88%
2	29	12,83%	90,71%
3	8	3,54%	94,25%
4	9	3,98%	98,23%
5	1	0,44%	98,67%
6	1	0,44%	99,12%
9	1	0,44%	99,56%
10	1	0,44%	100,00%
Total général	**226**	**100,00%**	

Variable Café	Effectifs	Fréquences	Fréquences cumulées
0	53	23,45%	23,45%
1	50	22,12%	45,58%
2	73	32,30%	77,88%
3	37	16,37%	94,25%
4	6	2,65%	96,90%
5	7	3,10%	100,00%
Total général	**226**	**100,00%**	

Figure: Tables of variables the and coffee.

statistical table for continuous variable

two many modalities :

$\Rightarrow$ The data are grouped into classes (sensitive choice !)

Tailles	Effectifs	Fréquences	Freq Cum
[140;150[	3	1,33%	1,33%
[150;160[	70	30,97%	32,30%
[160;170[	89	39,38%	71,68%
[170;180[	52	23,01%	94,69%
[180;190[	12	5,31%	100,00%
Total général	226	100,00%	

Poids	Effectifs	Fréquences	Freq Cum
[30;40[	1	0,44%	0,44%
[40;50[	13	5,75%	6,19%
[50;60[	52	23,01%	29,20%
[60;70[	68	30,09%	59,29%
[70;80[	57	25,22%	84,51%
[80;90[	23	10,18%	94,69%
[90;100[	12	5,31%	100,00%
Total général	226	100,00%	

Ages	Effectifs	Fréquences	Freq Cum
[65;70[	52	23,01%	23,01%
[70;75[	73	32,30%	55,31%
[75;80[	64	28,32%	83,63%
[80;85[	18	7,96%	91,59%
[85;90[	16	7,08%	98,67%
[90;95[	3	1,33%	100,00%
Total général	226	100,00%	

Figure: Grouping into classes of continuous quantitative variables Height, Weight and Age.

statistical summaries of qualitative variables

To summarize the information contained in quantitative variables we can also use statistical summaries.

statistical summaries

We may distinguish

- statistical summaries of position which give its order of magnitude ;
- statistical summaries of dispersion which express the variability of the values taken;
- statistical summaries of shape which express the general trend.

statistical summaries of position

- The mean
- The mode of the distribution : modality that appears with the highest frequency
- The median $Q_{0.5}$: central value that divides the population into two subpopulations of equal size
- The quartiles $Q_{0.25}, Q_{0.75}$

statistical summaries of dispersion

- The range : $\max - \min$
- The interquartile interval $[Q_{0.25}, Q_{0.75}]$
- The variance and the standard deviation

From n observations $x_1, \ldots, x_n$, the variance, denoted $var\,(x)$, is defined by :

$$var\,(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\,.$$

and the standard deviation $\hat{\sigma}_x$ is:

$$\hat{\sigma}_x = \sqrt{var\,(x)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Representing a discrete quantitative variable

Bar diagram

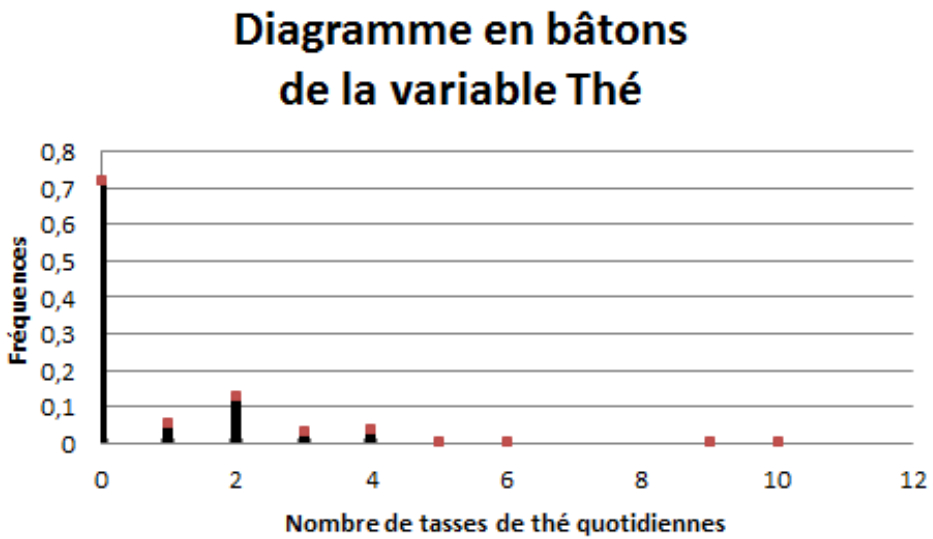


Figure: Bar diagram of variable the

Representing a continuous quantitative variable

Histogram

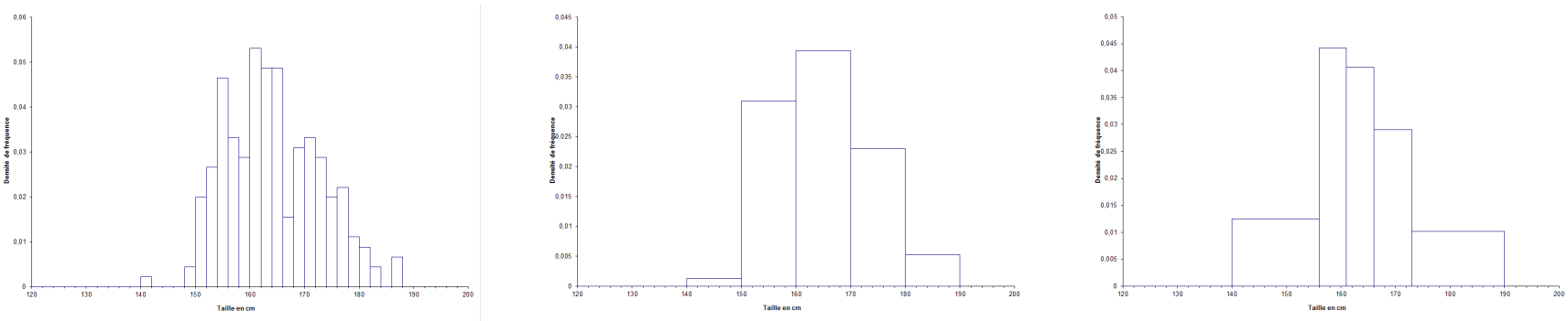


Figure: 3 histograms of variable height

- based on a grouping in classes, so the histogram inherits all the all the related problems: choice of classes, number of classes, etc.

Representing a quantitative variable

The empirical distribution function, based on the cumulative frequencies

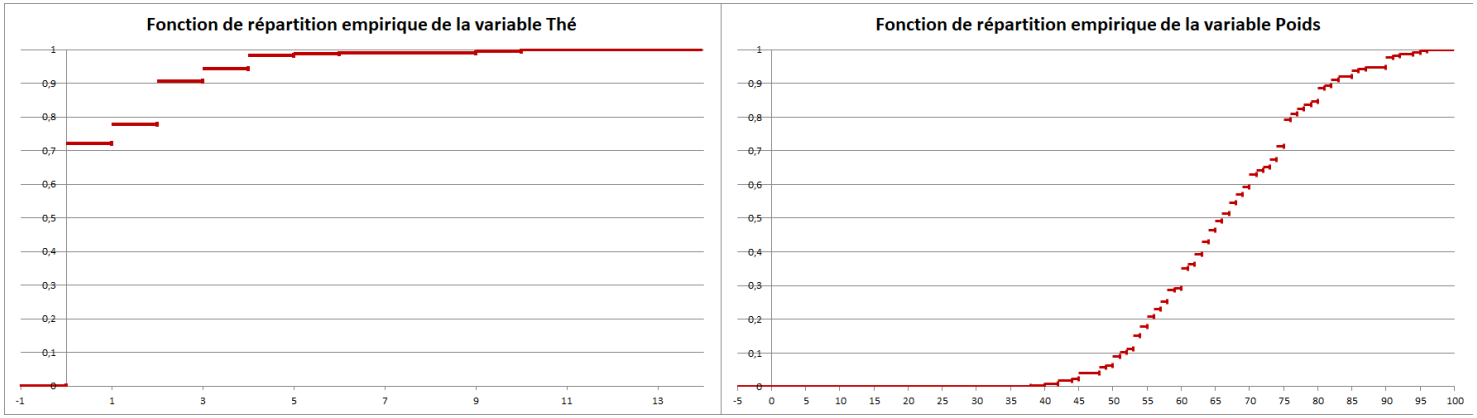


Figure: empirical distribution function, the and weight variables

- For a discrete variable, there are few jumps of significant size.
- For a continuous variable, there are many small jumps.

Representing a quantitative variable

Boxplot

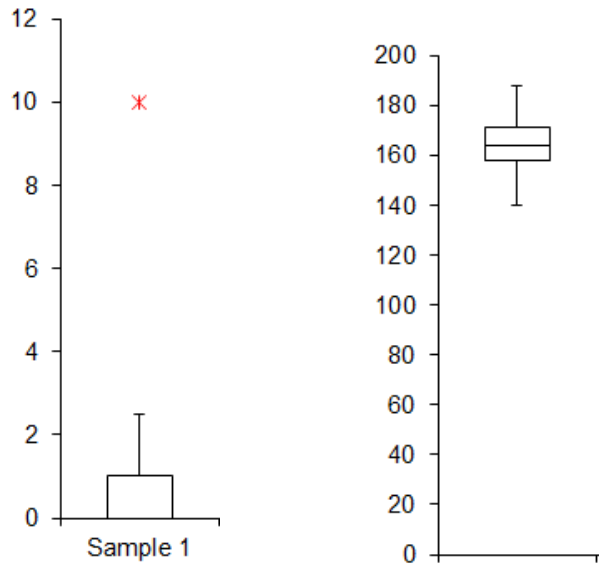


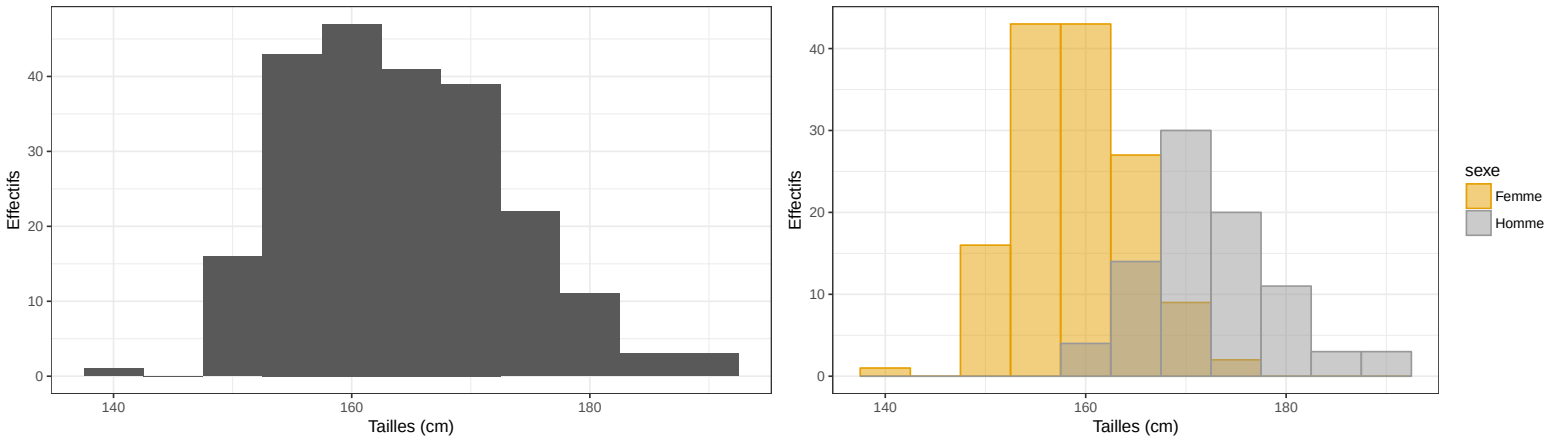
Figure: Boxplots of variables the and height

- for the variable The, we find the fact that the minimum, the first quartile and the median are equal.
- For the variable height, all statistical summaries are well distinguished.

Now ... multimodality ...

Multimodality

- try to take into account, measure, and analyze the links that may exist between two variables.



Representation by histograms of the variable height (in cm) by mixing the whole population (left) then dividing according to men and women (right).

Quali x Quali - stacked diagram

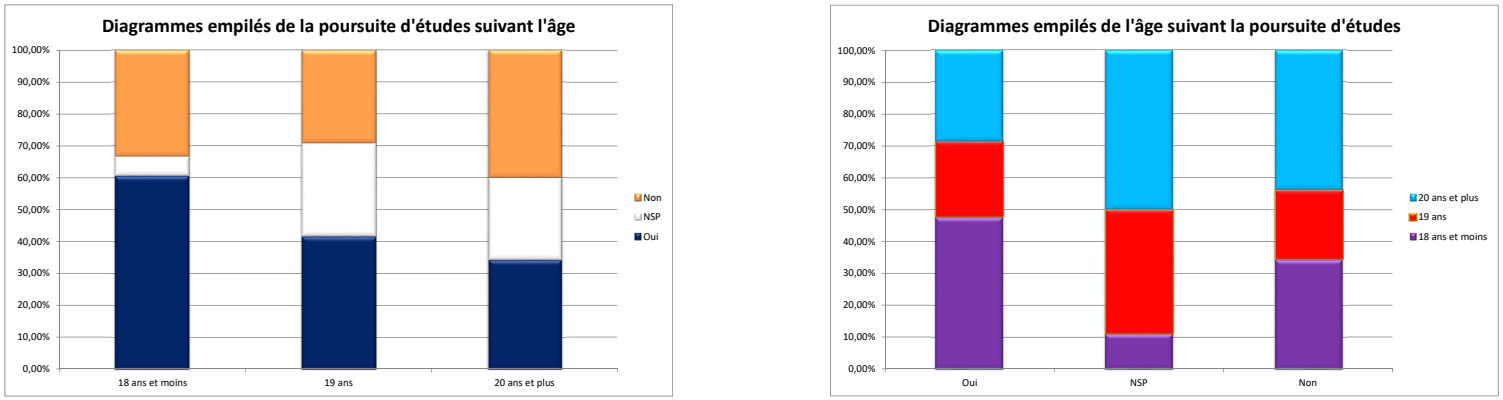


Figure: Representation of stacked diagrams of the different distributions: on the left, the distribution of further education as a function of age and, on the right, the distribution of age as a function of further education

Quanti x Quanti - scatterplot

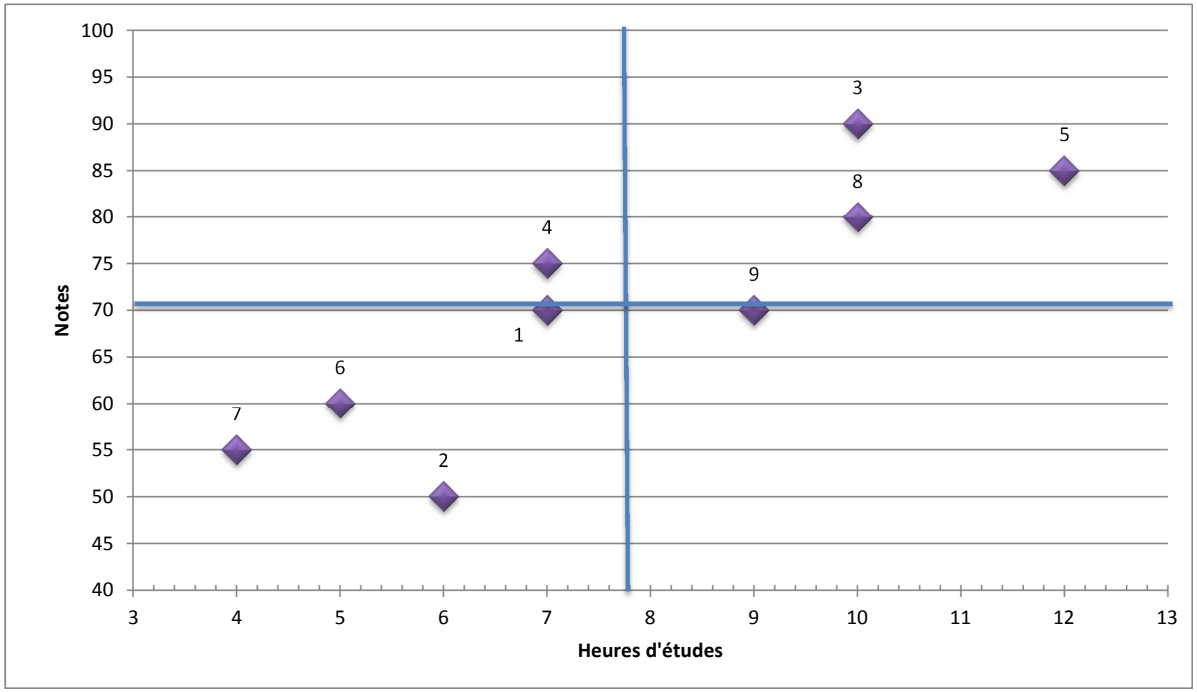


Figure: scatterplot

Quanti x Quanti - scatterplot

Warning! Correlation is not causality

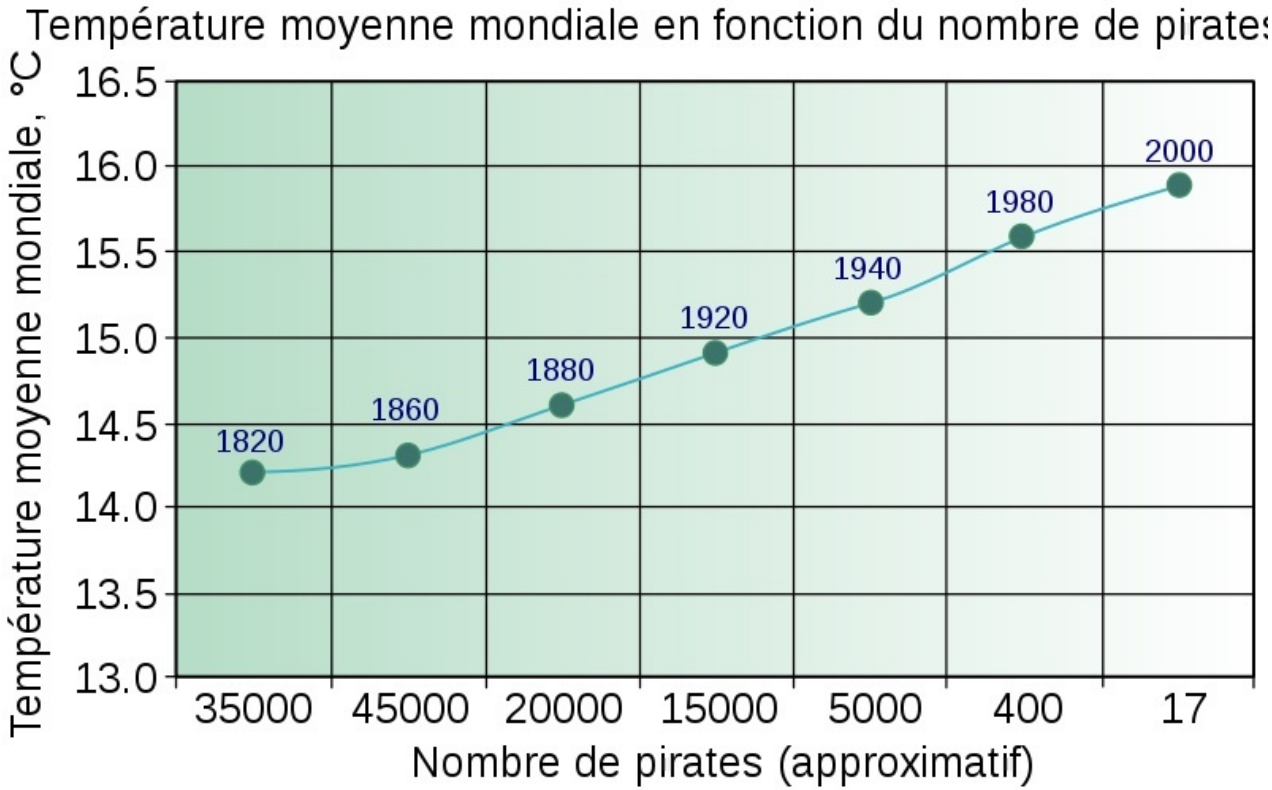


Figure: graph showing the relationship between global average temperature and the number of pirates

Quanti x Quanti - time series

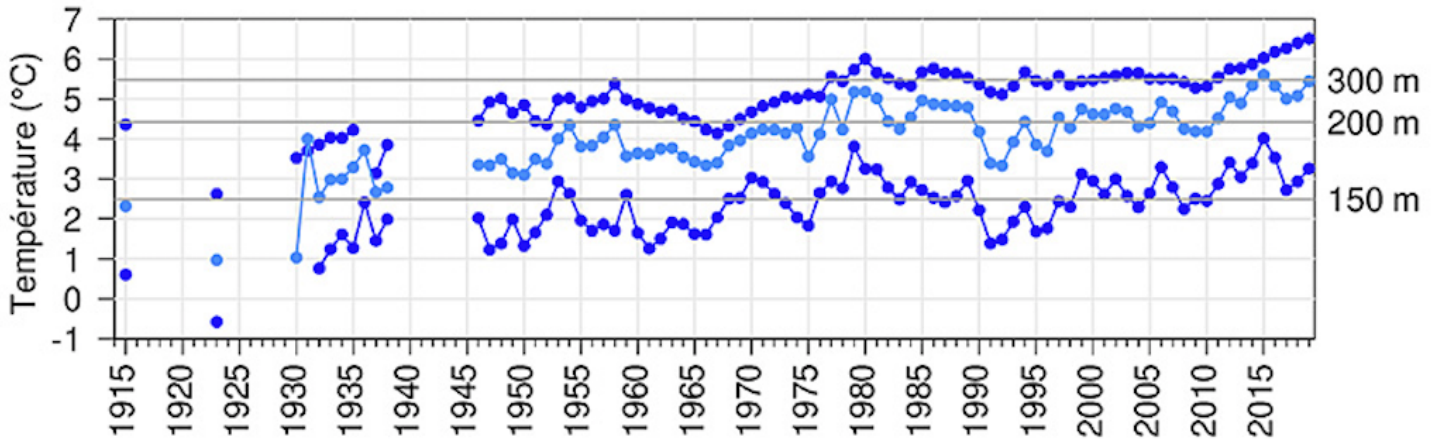


Figure: Time series of temperature averaged by depth layer for the Gulf of Saint Lurent. Annual averages at 150 m, 200 m and 300 m are shown and the horizontal lines are the 1981-2010 averages.