

Statistiques descriptives

Cours 2

Structure des variables

Structure des variables

- ◆ **Lorsqu'on fait une mesure en psychologique**
 - ✓ Il est nécessaire de savoir comment on va procéder à cette mesure
 - ✓ Attention: mesurer n'implique pas obligatoirement l'utilisation du système numérique

- ◆ **Variable statistique**
 - ✓ Une variable statistique est une quantité ou une qualité définie sur une population et qui est susceptible de varier d'un individu à l'autre
 - ✓ Les différentes valeurs ou aspects de la variable seront appelées modalités

- ◆ **Il existe 2 types de variables : qualitatives & quantitatives**

S. S. Stevens publie en 1946 son célèbre article « On the theory of scales of measurement* » contribution majeure du 20e siècle dans le domaine de la mesure en psychologie

On the Theory of Scales of Measurement

S. S. Stevens

Director, Psycho-Acoustic Laboratory, Harvard University

FOR SEVEN YEARS A COMMITTEE of the British Association for the Advancement of Science debated the problem of measurement. Appointed in 1932 to represent Section A (Mathematical and Physical Sciences) and Section J (Psychology), the committee was instructed to consider and report upon the possibility of "quantitative estimates of sensory events"—meaning simply: Is it possible to measure human sensation? Deliberation led only to disagreement, mainly about what is meant by the term measurement. An interim report in 1938 found one member complaining that his colleagues "came out by that same door as they went in," and in order to have another try at agreement, the committee begged to be continued for another year.

For its final report (1940) the committee chose a common bone for its contentions, directing its arguments at a concrete example of a sensory scale. This was the Sone scale of loudness (S. S. Stevens and H. Davis. *Hearing*. New York: Wiley, 1938), which purports to measure the subjective magnitude of an auditory sensation against a scale having the formal properties of other basic scales, such as those used to measure length and weight. Again the 19 members of the committee came out by the routes they entered, and their views ranged widely between two extremes. One member submitted "that any law purporting to express a quantitative relation between sensation intensity and stimulus intensity is not merely false but is in fact meaningless unless and until a meaning can be given to the concept of addition as applied to sensation" (Final Report, p. 245).

It is plain from this and from other statements by the committee that the real issue is the meaning of measurement. This, to be sure, is a semantic issue, but one susceptible of orderly discussion. Perhaps agreement can better be achieved if we recognize that measurement exists in a variety of forms and that scales of measurement fall into certain definite classes. These classes are determined both by the empirical operations invoked in the process of "measuring" and

by the formal (mathematical) properties of the scales. Furthermore—and this is of great concern to several of the sciences—the statistical manipulations that can legitimately be applied to empirical data depend upon the type of scale against which the data are ordered.

A CLASSIFICATION OF SCALES OF MEASUREMENT

Paraphrasing N. R. Campbell (Final Report, p. 340), we may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules. The fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurement. The problem then becomes that of making explicit (a) the various rules for the assignment of numerals, (b) the mathematical properties (or group structure) of the resulting scales, and (c) the statistical operations applicable to measurements made with each type of scale.

Scales are possible in the first place only because there is a certain isomorphism between what we can do with the aspects of objects and the properties of the numeral series. In dealing with the aspects of objects we invoke empirical operations for determining equality (classifying), for rank-ordering, and for determining when differences and when ratios between the aspects of objects are equal. The conventional series of numerals yields to analogous operations: We can identify the members of a numeral series and classify them. We know their order as given by convention. We can determine equal differences, as $8-6=4-2$, and equal ratios, as $8/4=6/3$. The isomorphism between these properties of the numeral series and certain empirical operations which we perform with objects permits the use of the series as a *model* to represent aspects of the empirical world.

The type of scale achieved depends upon the character of the basic empirical operations performed. These operations are limited ordinarily by the nature of the thing being scaled and by our choice of procedures, but, once selected, the operations determine

678
that there will be certain types of scales
issued in Table 1.

The decision to discard the scale names commonly encountered in writings on measurement is based on the ambiguity of such terms as "intensive" and "extensive." Both ordinal and interval scales have at

the case that stands at the median (mid-point) of a distribution maintains its position under all transformations which preserve order (isotonic group), but an item located at the mean remains at the mean only under transformations as restricted as those of the linear group. The ratio expressed by the coefficient

TABLE 1

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
NOMINAL	Determination of equality	Permutation group $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater or less	Isotonic group $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	General linear group $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	Similarity group $x' = ax$	Coefficient of variation

times been called intensive, and both interval and ratio scales have sometimes been labeled extensive.

It will be noted that the column listing the basic operations needed to create each type of scale is cumulative: to an operation listed opposite a particular scale must be added all those operations preceding it. Thus, an interval scale can be erected only provided we have an operation for determining equality of intervals, for determining greater or less, and for determining equality (not greater and not less). To these operations must be added a method for ascertaining equality of ratios if a ratio scale is to be achieved.

In the column which records the group structure of each scale are listed the mathematical transformations which leave the scale-form invariant. Thus, any numeral, x , on a scale can be replaced by another numeral, x' , where x' is the function of x listed in this column. Each mathematical group in the column is contained in the group immediately above it.

The last column presents examples of the type of statistical operations appropriate to each scale. This column is cumulative in that all statistics listed are admissible for data scaled against a ratio scale. The criterion for the appropriateness of a statistic is *invariance* under the transformations in Column 3.

* A classification essentially equivalent to that contained in this table was presented before the International Congress for the Unity of Science, September 1941. The writer is indebted to the late Prof. G. D. Birkhoff for a stimulating discussion which led to the completion of the table in essentially its present form.

of variation remains invariant only under the similarity transformation (multiplication by a constant). (The rank-order correlation coefficient is usually deemed appropriate to an ordinal scale, but actually this statistic assumes equal intervals between successive ranks and therefore calls for an interval scale.)

Let us now consider each scale in turn.

NOMINAL SCALE

The *nominal scale* represents the most unrestricted assignment of numerals. The numerals are used only as labels or type numbers, and words or letters would serve as well. Two types of nominal assignments are sometimes distinguished, as illustrated (a) by the 'numbering' of football players for the identification of the individuals, and (b) by the 'numbering' of types or classes, where each member of a class is assigned the same numeral. Actually, the first is a special case of the second, for when we label our football players we are dealing with unit classes of one member each. Since the purpose is just as well served when any two designating numerals are interchanged, this scale form remains invariant under the general substitution or permutation group (sometimes called the symmetric group of transformations). The only statistic relevant to nominal scales of Type A is the number of cases, e.g. the number of players assigned numerals. But once classes containing several individuals have

- ◆ Stevens est le fondateur d'une branche de la psychophysique contemporaine prônant la mesure directe des sensations. Il a enseigné la psychophysiologie et la psychologie à Harvard à partir de 1938 et y a dirigé le laboratoire de psychophysique



- ◆ (Wiki) Il contribua à la psychophysique en proposant une loi de puissance reliant la grandeur physique d'un stimulus et l'intensité perçue, dite loi de Stevens.

Structure des variables

◆ Les variables qualitatives

- ✓ Les modalités d'une variable qualitative ne sont pas objectivement comparables. Ce sont souvent des noms ou des étiquettes qui permettent de répartir la population en plusieurs groupes sans aucune hiérarchie entre ces groupes
- ✓ On les appelle aussi variables nominales

◆ On parlera de variable ordinales

- ✓ Lorsque l'on a faire à des variables qualitatives dont les modalités sont ordonnées selon certains critères

◆ Exemple

- ✓ Appréciation d'un étudiant à son cours de stat.:
très faible, faible, médiocre, moyen, assez bien, bien et très bien
- ✓ Anxiété d'un individu :
très peu anxieux, peu anxieux, anxieux, pas anxieux et pas anxieux du tout
- ✓ On remarque alors une hiérarchie entre les modalités qui n'est pas exprimée en terme de nombres réels

Structure des variables

◆ ... exemple

- ✓ L'échelle de Likert (du psychologue américain Rensis Likert) est un outil psychométrique* permettant de mesurer une attitude chez des individus.
 - Elle consiste en une ou plusieurs affirmations (énoncés ou items) pour lesquelles la personne interrogée exprime son degré d'accord ou de désaccord.
 - - Tout à fait d'accord, - D'accord, - Ni en désaccord ni d'accord, - Pas d'accord, -Pas du tout d'accord
 - On associe à chaque modalité de réponse un nombre entier de 1 à 5 ou de -2 à +2 dans le cas à 5 modalités

* La psychométrie est la science qui étudie l'ensemble des techniques de mesures pratiquées en psychologie.

Structure des variables

◆ Les variables quantitatives

- ✓ Lorsque la variable désigne une quantité que l'on mesure, de sorte que les modalités de la variable sont des nombres
- ✓ Ces variables peuvent être de deux types : variables continues et variables discrètes

a) Variables continues

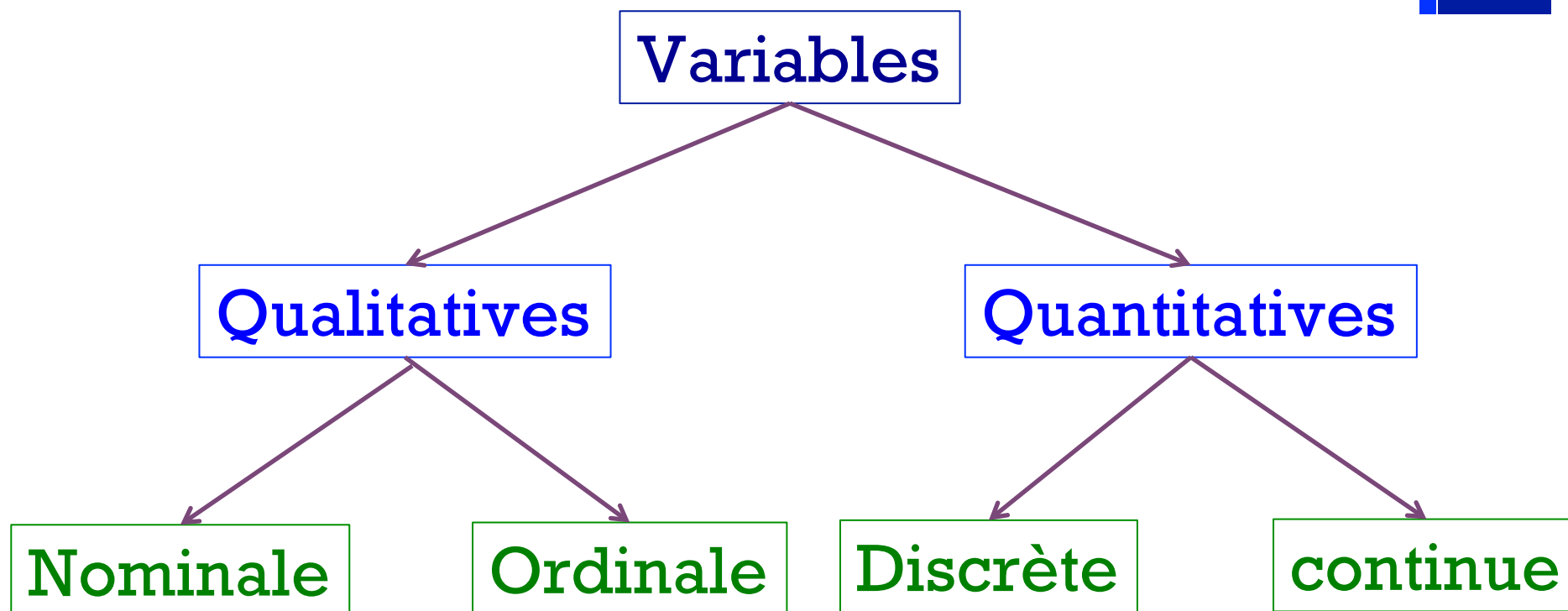
Les modalités de la variable peuvent prendre toutes les valeurs comprises dans un intervalle donné, notamment un nombre infini de valeurs

C'est notamment le cas pour des longueurs, des poids, etc.

b) Variables discrètes

La variable ne prend que des valeurs entières

En résumé



Présentation des variables statistiques: variables qualitatives

- ◆ À l'issue du dépouillement d'une étude, on choisit généralement de présenter les données en un tableau où apparaissent les différentes modalités (ou valeurs) du caractère ainsi que les effectifs de ces modalités.

Modalités	x_1	x_2	x_3	...	x_i	...	x_r
Effectifs	n_1	n_2	n_3	...	n_i	...	n_r

- ◆ n_i représente le nombre le nombre d'individus pour lesquels la variable x prend la modalité x_i
- ◆ La somme $n_1 + n_2 + \dots + n_r = n$ est l'effectif total c.à.d. la taille de la population étudié.
Que l'on écrit pour simplifier :

$$\sum_{i=1}^r n_i = n$$

Présentation des variables statistiques: variables qualitatives

- ◆ Soit x une variable statistique de modalité x_1, x_2, \dots, x_r et n l'effectif total, la fréquence relative de modalité x_i est par définition le rapport $f_i = n_i/n$ (c'est le quotient de cet effectif par celui de la population).

Remarque la somme

$$\sum_{i=1}^{i=r} f_i = 1 \quad (=100\%)$$

- ◆ L'intérêt de la notion de fréquence
 - ✓ Être un élément de comparaison pour un même caractère sur deux populations différentes (surtout si elles n'ont pas la même taille). L'expression en pourcentage (%) facilite les comparaisons
 - ✓ Les représentations graphiques sont en général associées aux fréquences et rarement aux effectifs

Présentation des variables statistiques: variables qualitatives

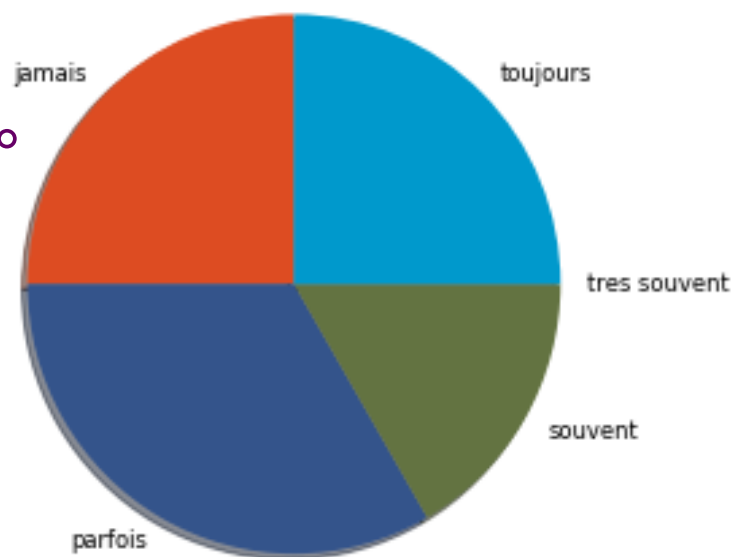
Les principales représentations graphiques pour des variables qualitatives sont les représentations en secteurs circulaires (ou camemberts) ou en barres (ou tuyaux d'orgue).

- ◆ Diagramme circulaire 120 candidats ont répondu à une même question

Réponses	Nombre de sujets	% de sujets
Jamais	30	25
Parfois	40	33,3
Souvent	20	16
Très souvent	0	0
Toujours	30	25
Total	120	100

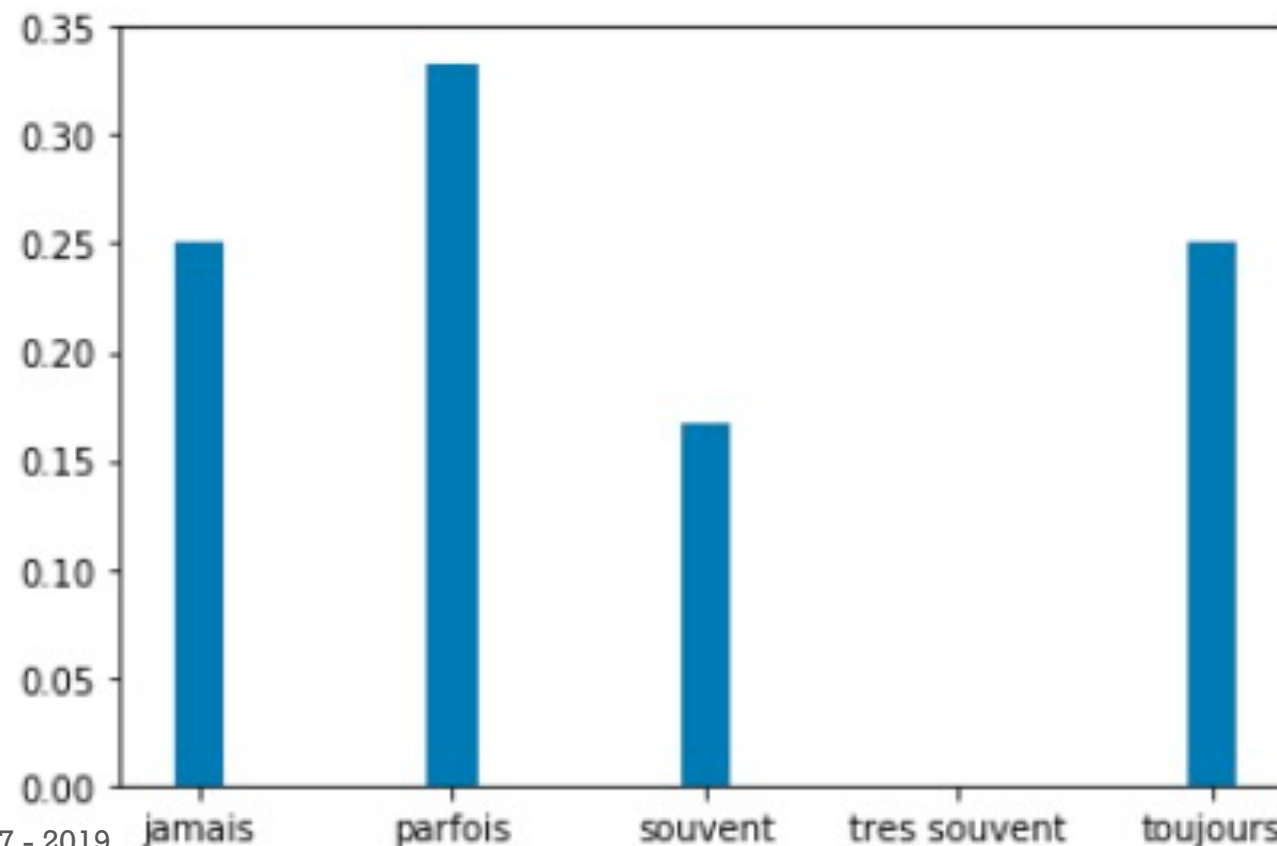
Présentation des variables statistiques: variables qualitatives

- ◆ Il permet de représenter un petit nombre de valeurs par des angles proportionnels à ces valeurs
 - ✓ Placer une étiquette de pourcentage peut faciliter la lecture du diagramme
 - ✓ À cause d'erreurs d'arrondis, la somme des pourcentages ne donne pas 100.
Tout comme la somme des angles au centre ne donne pas 360°



Présentation des variables statistiques: variables qualitatives

- ◆ Diagramme en barres
La hauteur de chaque barre correspond à la fréquence d'une modalité



Présentation des variables statistiques: variables quantitatives

- ◆ Quand on travaille avec de grands échantillons, les données brutes recueillies ne parlent pas d'elles-mêmes
- ◆ Il faut les organiser de manière lisible sur un tableau des effectifs. Deux cas:
 - ✓ Si la variable est de nature discrète et qu'il y a peu de modalités, il faut regrouper les effectifs des différentes valeurs, mais en prenant garde à ranger les modalités par ordre croissant.
 - ✓ S'il y a trop de modalités (par exemple pour des variables continues), on les organise en classes.

Présentation des variables statistiques: variables quantitatives

◆ Exemple

Voici les QI obtenus après un test réalisé auprès de 72 enfants âgés entre 10 à 13 ans

113	109	101	82	86	102	104	98	114	106
97	111	121	96	103	93	108	112	111	109
99	94	117	90	92	106	127	105	100	101
119	97	115	85	107	89	116	102	104	103
99	100	98	100	91	120	110	105	96	95
113	118	110	118	81	110	93	102	107	103
97	111	123	101	106	105	108	107	117	95
88	103								

Présentation des variables statistiques: variables quantitatives

◆ Exemple

Voici les QI obtenus après un test réalisé auprès de 72 enfants âgés entre 10 à 13 ans

QI	Effectifs (n_k)	Effectifs cumulés	Fréquences (%)	Fréquences cumulés (%)
82	2	2	2,78	2,78
87	4	6	5,55	8,33
92	6	12	8,33	16,66
97	11	23	15,27	31,91
102	15	38	20,83	52,74
107	13	51	18,06	70,80
112	10	61	13,89	84,69
117	7	68	9,73	94,42
122	3	71	4,17	98,59
127	1	72	1,39	99,98

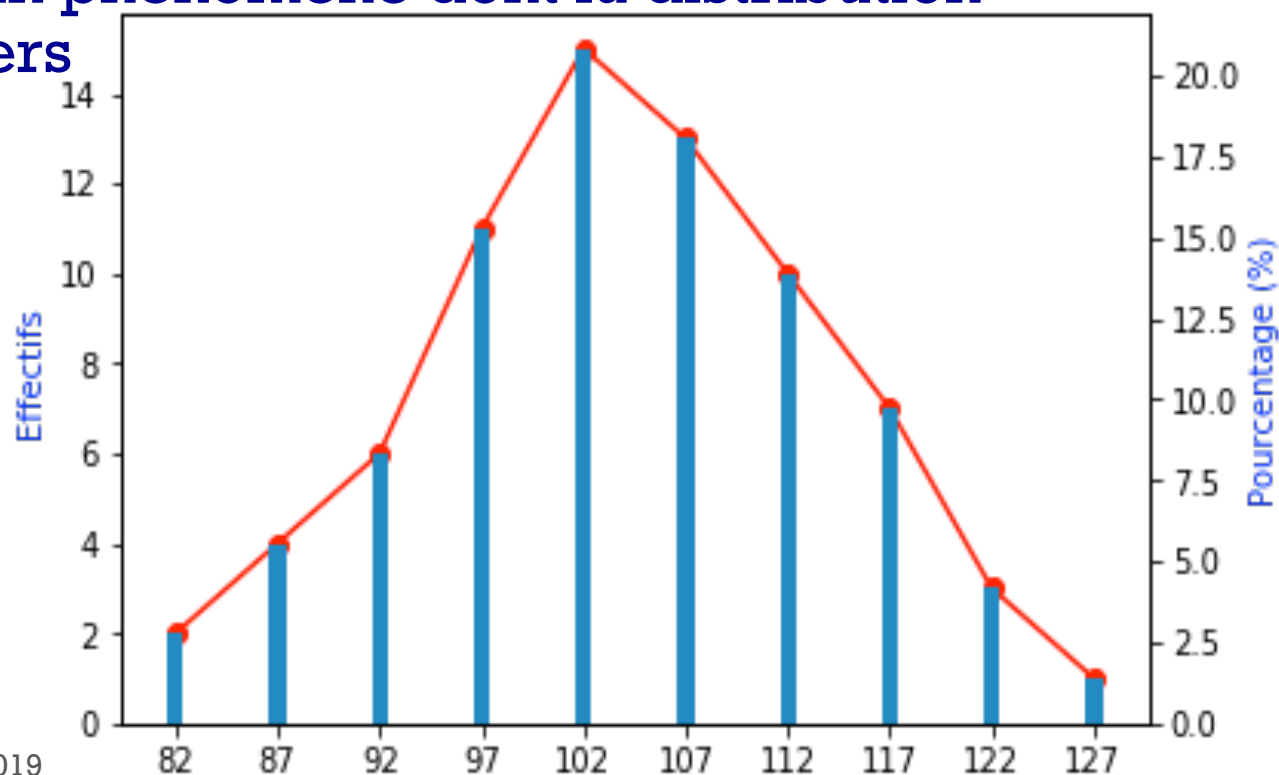
Présentation des variables statistiques: variables quantitatives

- ◆ Groupe par classe
 - ✓ Relevé et rangé en 10 classes
 - ✓ Effectifs tot. $N = \sum n_k = 72$
 - ✓ $\sum (\text{fréquence})_k \sim 100\%$

Présentation des variables statistiques: variables quantitatives

Représentation en barres des effectifs et des fréquences

- ◆ Représentation discutable elle traduit sous forme d'une ligne continue un phénomène dont la distribution est faite de paliers discontinus.

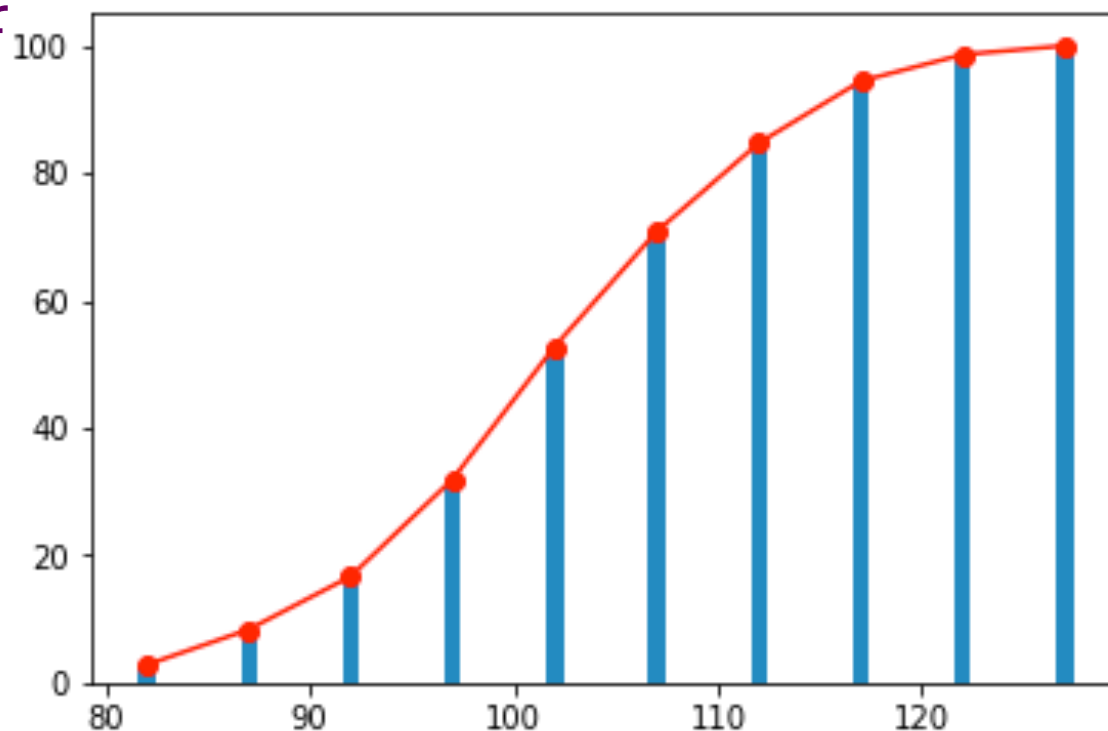


Présentation des variables statistiques: variables quantitatives

Représentation en barres des fréquences cumulés

◆ Remarque

- ✓ Environ 85% des sujets ont un QI < 112
- ✓ Permet d'avoir la population qui a une classe inférieure à une certaine valeur



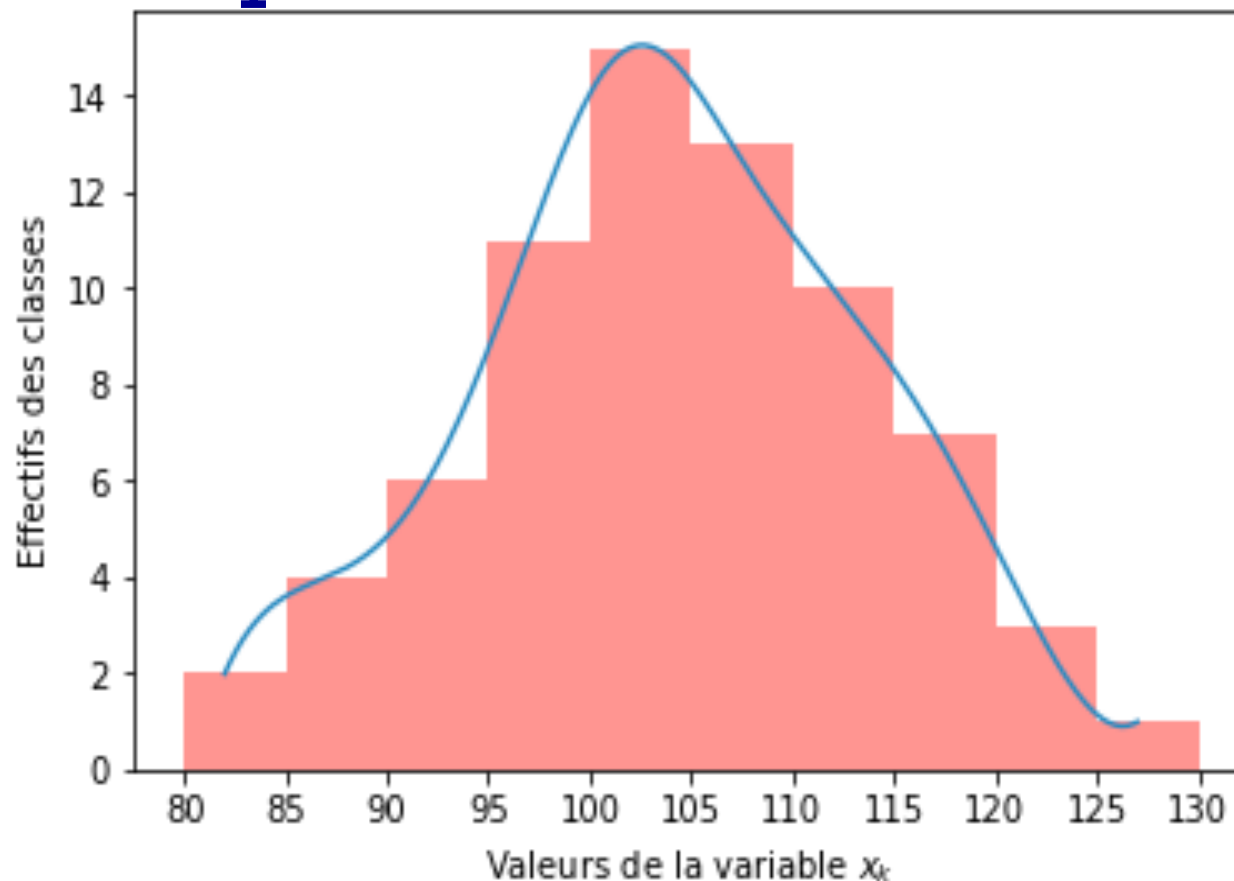
Présentation des variables statistiques: variables quantitatives

Représentation en histogramme

- ◆ Un histogramme est une représentation graphique d'une série statistique de variable quantitative
 - ✓ Il est constitué de rectangles contigus dont les aires sont proportionnelles aux effectifs des classes.
 - ✓ Sur l'axe des abscisses sont reportées les bornes des classes de la série;
- ◆ Si les amplitudes sont toutes égales, la hauteur du rectangle est proportionnelle à l'effectif (ou à la fréquence)
- ◆ L'amplitude est la différence entre la borne supérieure et la borne inférieure de la classe ($a_{i+1} - a_i$)
- ◆ Par ex. les QI par amplitudes de 5

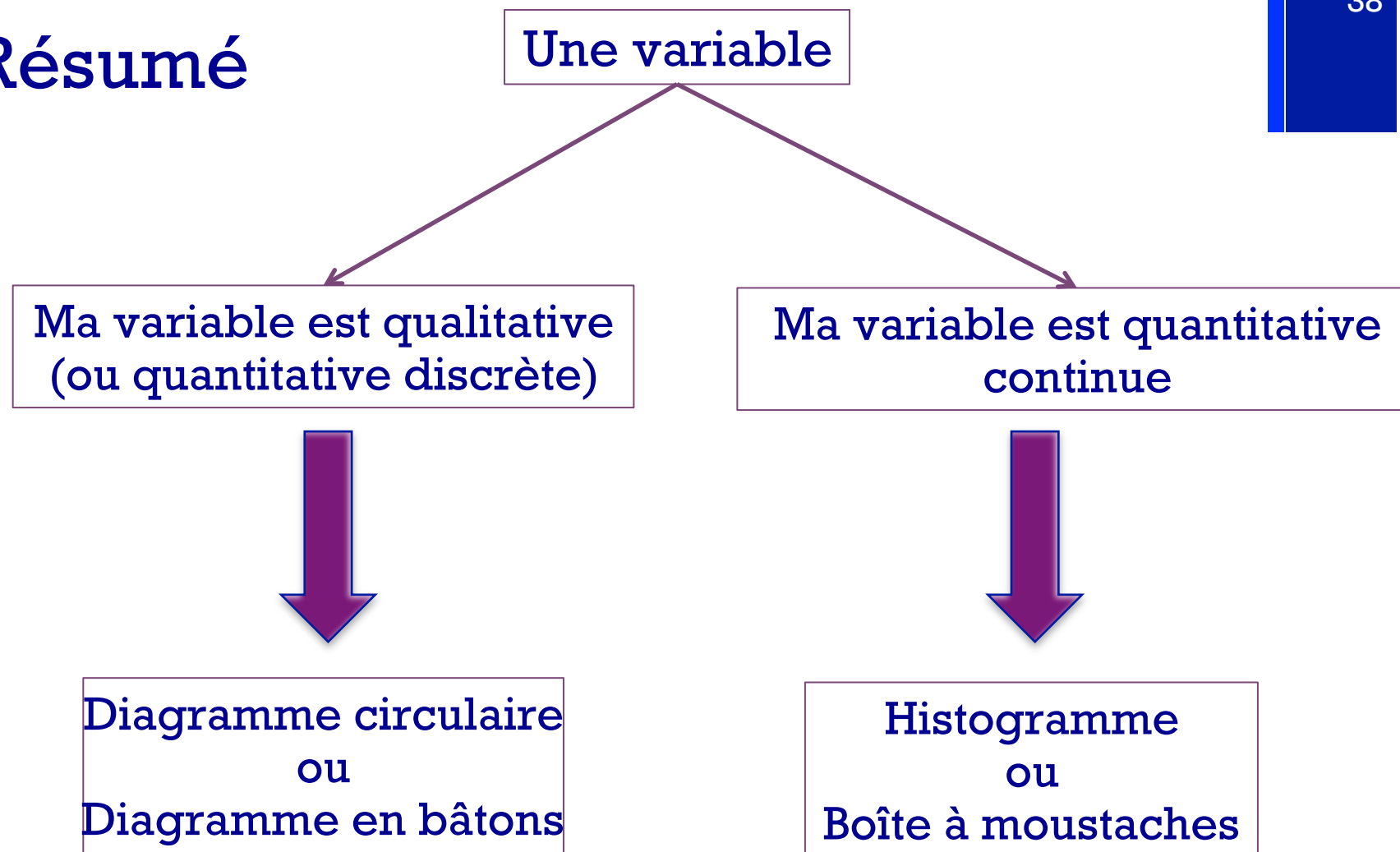
QI	Effectifs (n_k)
[80-85[2
[85-90[4
[90-95[6
[95-100[11
[100-105[15
[105-110[13
[110-115[10
[115-120[7
[120-125[3
[125-130[1

Présentation des variables statistiques: variables quantitatives



- ◆ Remarque: le centre de la classe correspondant à l'intervalle $[a_{i+1}, a_i[$ est le nombre $(a_{i+1} + a_i)/2$ (milieu de l'intervalle)

Résumé



Statistiques descriptives

- ◆ Un premier objectif des statistiques descriptives est de décrire au mieux les échantillons, c.à.d. de les synthétiser au travers d'indices descriptifs, appelés également résumés statistiques.
- ◆ Ces résumés visent avant tout à donner des renseignements sur la tendance centrale et sur la dispersion de la distribution
- ◆ On va parler dans ce qui suit de:
 - ✓ Mode
 - ✓ Moyenne
 - ✓ Médiane
 - ✓ Mesure de dispersion