
Selon les derniers sondages, 47% des statistiques sont fausses.
G&W

1 Définitions

1.1 Statistique

La **statistique** est l'ensemble des méthodes qui permettent de collecter, de traiter, d'interpréter et de présenter (numériquement et graphiquement) des données.

Le but de la statistique est de synthétiser l'information contenue dans les données (bases de données).

1.2 Population et individu

En statistique, l'ensemble que l'on observe et qui sera soumis à une analyse statistique est appelé **population**.

Exemples : les étudiants de l'Institut Saint-Laurent de Promotion Sociale, les étudiants de 12 à 18 ans, les coulées de fonte des 9 premiers mois de l'année, …

Chaque élément de cet ensemble est un **individu** (ou **unité statistique**).

1.3 Echantillon

Comme il est souvent difficile d'observer la totalité d'une population, on se contente généralement d'en observer un sous-ensemble.

Un **échantillon** est un sous-ensemble de la population observée.

Le nombre d'individus de l'échantillon est la taille ou l'effectif de l'échantillon.

Pour qu'un échantillon soit numériquement représentatif de la population observée, le nombre d'individus de cet échantillon doit être au moins de l'ordre de la racine carrée du nombre d'individus de la population.

Par exemple, la taille d'un échantillon numériquement représentatif de la population belge (11 467 923 individus, valeur provisoire, source : eurostat, mise à jour du 09/09/2019) doit être au moins de 3 387 individus.

1.4 Caractère ou variable statistique

Le **caractère** (ou **variable statistique**) est la propriété observée chez les individus d'une population ou d'un échantillon.

Exemples : l'âge, le sexe ou la couleur des yeux pour les étudiants, le pourcentage en carbone pour les coulées de fonte, …

Les variables statistiques se décomposent en deux familles : les variables qualitatives et les variables quantitatives.

1.4.1 Les variables qualitatives

Les **variables qualitatives** représentent une « qualité » et ne se traduisent pas par un nombre. Elles peuvent être ordonnées (**variables qualitatives ordinales**) ou non (**variables qualitatives nominales**).

Exemples de variables qualitatives nominales : le sexe, la couleur des yeux, …

Exemples de variables qualitatives ordinales : les stades d'une maladie, les niveaux d'études, …

1.4.2 Les variables quantitatives

Les **variables quantitatives** se traduisent par un nombre et se décomposent également en deux familles :

- les **variables quantitatives discontinues** ou **discrètes** qui ne peuvent prendre que des valeurs isolées les unes des autres ; les variables quantitatives discontinues résultent le plus souvent d'un comptage ;

  Exemples : le nombre d'enfants par famille, le nombre d'élèves par classe, …

- les **variables quantitatives continues** qui peuvent prendre toutes les valeurs d'un intervalle fini ou infini et qui résultent le plus souvent d'une mesure ;

  Exemples : la taille ou le poids d'une personne, le diamètre d'une pièce, …

1.5 Modalités

Les **modalités** sont les différentes situations où les individus peuvent se retrouver à l'égard du caractère observé. Les modalités d'un même caractère doivent être incompatibles et exhaustives.

Exemples :

- la variable qualitative nominale « sexe » possède deux modalités : homme ou femme ;

- les modalités de la variable quantitative discontinue « nombre d'enfants » pourraient être : 0, 1, 2, 3, 4, 5, supérieur à 5.

2 Séries statistiques

Collecte des données

2.1 Définition

Une **série statistique** est une énumération des observations.

2.2 Exemple dans le cas d'une variable qualitative nominale

Dans une classe d'un Institut de Promotion Sociale bien connu ☺, les étudiants ont été invités à préciser leur état civil. La liste ci-dessous reprend les choix proposés :

- Célibataire (C)
- Divorcé (D)
- Marié (M)
- Veuf (V)
- Autre (A)

Pour plus d'informations sur la modalité « Autre (A) », veuillez consulter le site internet be.STAT.

La série statistique ci-dessous reprend le choix de chaque étudiant de la classe :

C	C	C	C	C	V	C	C	C	C
C	C	C	M	A	C	C	A	C	C

Tableau 2.2.1 > Série statistique dans le cas d'une variable qualitative nominale

2.3 Exemple dans le cas d'une variable qualitative ordinale

Un site internet bien connu permet de « noter » un restaurant. La note de l'internaute est choisie dans la liste ci-dessous :

- Excellent (E)
- Très bien (TB)
- Moyen (M)
- Mauvais (MV)
- Epouvantable (EV)

La série statistique ci-dessous reprend, pour un restaurant « deux étoiles » belge, la note des internautes sur une année :

E	E	EV	TB	E	EV	TB	E	M	EV
M	MV	E	TB	MV	TB	M			

Tableau 2.3.1 > Série statistique dans le cas d'une variable qualitative ordinale

2.4 Exemple dans le cas d'une variable quantitative discontinue

La série statistique ci-dessous reprend, pour chaque étudiant d'une classe, la cote d'une question de l'examen de l'unité d'enseignement « Eléments de statistique » (la cote maximum est 5, valeurs entières uniquement) :

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

Tableau 2.4.1 > Série statistique dans le cas d'une variable quantitative discontinue

2.5 Exemple dans le cas d'une variable quantitative continue

Le site internet Meteo Belgique présente les relevés de stations météorologiques jour après jour.

La série statistique ci-dessous reprend, pour chaque jour du mois d'octobre 2011, la valeur de la température moyenne (°C) à Uccle :

19,2	18,4	19,3	16,8	16,4	13,4	10,2	9,0	12,0	16,4
15,2	14,8	11,3	9,3	8,8	9,1	11,6	9,9	8,0	6,0
6,1	6,1	9,6	9,7	11,3	10,7	12,2	14,3	14,5	14,0
12,6									

Tableau 2.5.1 > Série statistique dans le cas d'une variable quantitative continue

Remarque : le tableau se lit de gauche à droite et de haut en bas ; par exemple, la valeur de la température moyenne à Uccle le 20 octobre 2011 était de 6,0 °C.

3 Tableaux de fréquences (distributions)

Traitement et présentation numérique des données

3.1 Introduction

Toutes les données de la série statistique sont regroupées dans un **tableau de fréquences** indiquant la **distribution** des individus par rapport au caractère observé.

Le regroupement s'effectue par **classes**.

3.1.1 Classe dans le cas d'une variable qualitative ou d'une variable quantitative discontinue

Si le caractère est qualitatif ou quantitatif discontinu, une classe contient tous les individus ayant la même modalité.

Remarque importante

Dans le cas d'un caractère quantitatif discontinu présentant un nombre élevé de modalités, une classe est un intervalle de valeurs comme dans le cas d'une variable quantitative continue.

3.1.2 Classe dans le cas d'une variable quantitative continue

Si le caractère est quantitatif continu, une classe est un intervalle de valeurs.

Pour construire les intervalles de valeurs (classes), on respecte les règles suivantes :

- le nombre de classes est compris de préférence entre 5 et 20 et vaut environ la racine carrée du nombre d'observations si celui-ci est inférieur ou égal à 400 ;
- chaque fois que cela est possible, on s'arrange pour que toutes les classes aient la même largeur (ou étendue ou amplitude), la largeur d'une classe étant plus ou moins égale à l'**étendue** (ou **amplitude** c'est-à-dire la valeur maximum de la série statistique moins la valeur minimum de la série statistique) divisée par le nombre de classes ;
- chaque classe contient sa borne inférieure mais pas sa borne supérieure ou chaque classe ne contient pas sa borne inférieure mais bien sa borne supérieure.

Dans les calculs, une classe sera représentée par son **centre**, qui est le milieu de l'intervalle de valeurs.

Une fois la classe constituée, on considère les individus répartis uniformément entre la borne inférieure et la borne supérieure de la classe ce qui entraîne une perte d'informations par rapport aux données brutes de la série statistique.

3.1.3 Construction du tableau de fréquences

Le tableau de fréquences comprend les colonnes suivantes (*i* étant l'indice de la classe) :

- les différentes classes (x_i) ;
- si la classe est un intervalle de valeurs, le centre de la classe (c_i) ;
- la **fréquence absolue** ou **effectif** (n_i) c'est-à-dire le nombre d'individus de la classe ;
- la **fréquence relative** (f_i) c'est-à-dire le nombre d'individus de la classe (n_i) divisé par le nombre d'individus de l'échantillon ou de la population (N) ; la fréquence relative peut être exprimée en % ;

- la **fréquence absolue cumulée** ou **effectif cumulé** (N_i) c'est-à-dire le nombre d'individus de la classe augmenté du nombre d'individus des classes précédentes ; cette colonne n'a pas de sens dans le cas d'une variable qualitative nominale ;
- la **fréquence relative cumulée** (F_i) c'est-à-dire le nombre d'individus de la classe augmenté du nombre d'individus des classes précédentes (N_i) divisé par le nombre d'individus de l'échantillon ou de la population (N) ; la fréquence relative cumulée peut être exprimée en % ; cette colonne n'a pas de sens dans le cas d'une variable qualitative nominale.

3.2 Exemple dans le cas d'une variable qualitative nominale

Reprenons la série statistique du tableau 2.2.1 :

C	C	C	C	C	V	C	C	C	C
C	C	C	M	A	C	C	A	C	C

Tableau 2.2.1 > Série statistique dans le cas d'une variable qualitative nominale

Si le caractère est qualitatif, une classe contient tous les individus ayant la même modalité. Le nombre de classes est donc égal à 5 puisque les différentes valeurs possibles pour le choix des étudiants sont :

C (Célibataire), D (Divorcé), M (Marié), V (Veuf) et A (Autre)

Comme les classes ne présentent pas de gradation, l'ordre de présentation des données n'est pas imposé. Il est courant d'utiliser l'ordre croissant alphabétique des classes ou l'ordre décroissant des fréquences absolues.

Classe x_i	**Fréquence absolue** n_i	**Fréquence relative** f_i
C	16	80 %
A	2	10 %
M	1	5 %
V	1	5 %
D	0	0 %
Totaux	**20**	**100 %**

Tableau 3.2.1 > Tableau de fréquences dans le cas d'une variable qualitative nominale

3.3 Exemple dans le cas d'une variable qualitative ordinale

Reprenons la série statistique du tableau 2.3.1 :

E	E	EV	TB	E	EV	TB	E	M	EV
M	MV	E	TB	MV	TB	M			

Tableau 2.3.1 > Série statistique dans le cas d'une variable qualitative ordinale

Si le caractère est qualitatif, une classe contient tous les individus ayant la même modalité. Le nombre de classes est donc égal à 5 puisque les différentes valeurs possibles pour la note des internautes sont :

EV (Epouvantable), MV (Mauvais), M (Moyen), TB (Très bien) et E (Excellent)

Comme les classes présentent une gradation, l'ordre de présentation des données est imposé.

Classe x_i	Fréquence absolue n_i	Fréquence relative f_i	Fréquence absolue cumulée N_i	Fréquence relative cumulée F_i
EV	3	17,65 %	3	17,65 %
MV	2	11,76 %	5	29,41 %
M	3	17,65 %	8	47,06 %
TB	4	23,53 %	12	70,59 %
E	5	29,41 %	17	100,00 %
Totaux	**17**	**100 %**	**17**	**100 %**

Tableau 3.3.1 > Tableau de fréquences dans le cas d'une variable qualitative ordinale

3.4 Exemple dans le cas d'une variable quantitative discontinue

Reprenons la série statistique du tableau 2.4.1 :

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

Tableau 2.4.1 > Série statistique dans le cas d'une variable quantitative discontinue

Si le caractère est quantitatif discontinu ne présentant pas un nombre élevé de modalités, une classe contient tous les individus ayant la même modalité. Le nombre de classes est donc égal à 6 puisque les différentes valeurs possibles pour la cote de la question considérée de l'examen de l'unité d'enseignement « Eléments de statistique » (la cote maximum est 5, valeurs entières uniquement) sont : 0, 1, 2, 3, 4 et 5

Comme chaque classe correspond à un nombre, les données sont présentées dans l'ordre croissant ou décroissant des classes.

Classe x_i	Fréquence absolue n_i	Fréquence relative f_i	Fréquence absolue cumulée N_i	Fréquence relative cumulée F_i
0	1	5 %	1	5 %
1	2	10 %	3	15 %
2	3	15 %	6	30 %
3	2	10 %	8	40 %
4	5	25 %	13	65 %
5	7	35 %	20	100 %
Totaux	**20**	**100 %**	**20**	**100 %**

Tableau 3.4.1 > Tableau de fréquences dans le cas d'une variable quantitative discontinue

3.5 Exemple dans le cas d'une variable quantitative continue

Reprenons la série statistique du tableau 2.5.1 :

19,2	18,4	19,3	16,8	16,4	13,4	10,2	9,0	12,0	16,4
15,2	14,8	11,3	9,3	8,8	9,1	11,6	9,9	8,0	6,0
6,1	6,1	9,6	9,7	11,3	10,7	12,2	14,3	14,5	14,0
12,6									

Tableau 2.5.1 > Série statistique dans le cas d'une variable quantitative continue

Si le caractère est quantitatif continu, une classe est un intervalle de valeurs.

Le nombre de classes est environ égal à la racine carrée du nombre d'observations c'est-à-dire 5 ou 6 (racine carrée de 31 = 5,57).

L'étendue (ou amplitude) est égale à la valeur maximum de la série statistique moins la valeur minimum de la série statistique :

19,3 – 6,0 = 13,3

La largeur d'une classe est égale à l'étendue divisée par le nombre de classes :

13,3 / 5 = 2,66

Ce nombre peut être légèrement adapté pour faciliter la lecture : nous décidons de prendre, comme valeur pour la largeur d'une classe, non pas 2,66 mais 3,00.

La première classe est déterminée de manière à contenir la valeur minimum de la série statistique. Toutefois, il n'est pas conseillé de prendre cette valeur minimum comme valeur de la borne inférieure de la première classe : il faut au minimum soustraire la moitié de la précision de la mesure. Dans notre exemple, pour une précision de mesure de 0,1 °C, la valeur de la borne inférieure de la première classe doit être inférieure ou égale à :

6,00 – 0,05 = 5,95

Le plus souvent, on prend comme valeur de la borne inférieure de la première classe un nombre qui est multiple ou demi-multiple de la largeur de classe. Cependant, nous décidons de prendre, comme valeur de la borne inférieure de la première classe, non pas 5,95 mais 5,00.

La dernière classe doit contenir la valeur maximum de la série statistique.

Un point important est le caractère semi-ouvert des intervalles de valeurs : la notation [5,0 – 8,0[par exemple signifie que cette classe comprend toutes les observations supérieures ou égales à 5,0 °C et strictement inférieures à 8,0 °C. Une observation exactement égale à 8,0 °C appartiendra à la classe suivante [8,0 – 11,0[.

Classe x_i	**Centre** c_i	**Fréquence absolue** n_i	**Fréquence relative** f_i	**Fréquence absolue cumulée** N_i	**Fréquence relative cumulée** F_i
[5,0 – 8,0[	6,5	3	9,68 %	3	9,68 %
[8,0 – 11,0[	9,5	10	32,26 %	13	41,94 %
[11,0 – 14,0[	12,5	7	22,58 %	20	64,52 %
[14,0 – 17,0[	15,5	8	25,81 %	28	90,32 %
[17,0 – 20,0[	18,5	3	9,68 %	31	100,00 %
Totaux		**31**	**100 %**	**31**	**100 %**

Tableau 3.5.1 > Tableau de fréquences dans le cas d'une variable quantitative continue

4 Histogrammes

Présentation graphique des données

4.1 Introduction

Les tableaux de fréquences peuvent être représentés graphiquement sous la forme d'**histogrammes**. On distinguera d'une part, les histogrammes basés sur les fréquences absolues (n_i) ou les fréquences relatives (f_i) et, d'autre part, les histogrammes basés sur les fréquences absolues cumulées (N_i) ou les fréquences relatives cumulées (F_i).

L'axe horizontal de l'histogramme reprend l'ensemble des classes du tableau de fréquences. L'axe vertical de l'histogramme reprend les fréquences absolues ou relatives, non cumulées ou cumulées.

4.2 Exemple dans le cas d'une variable qualitative nominale

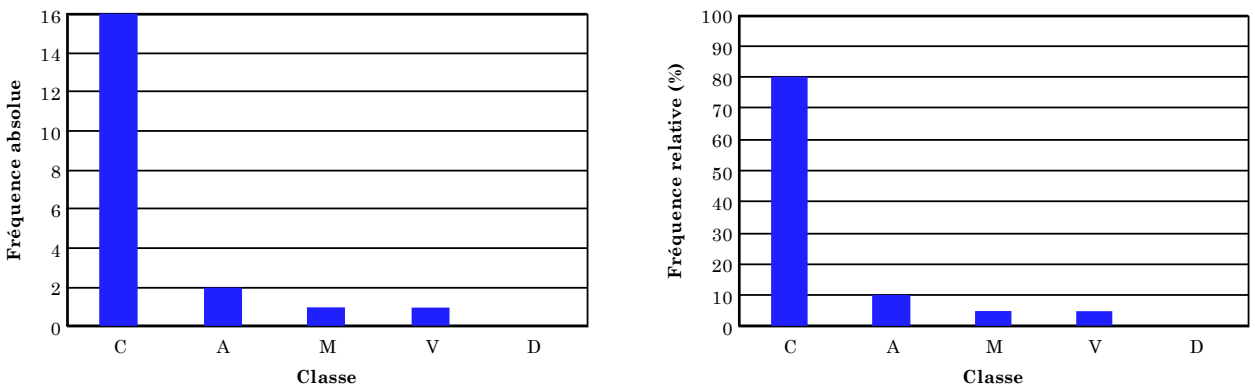
Reprenons le tableau de fréquences 3.2.1 :

Classe x_i	**Fréquence absolue** n_i	**Fréquence relative** f_i
C	16	80 %
A	2	10 %
M	1	5 %
V	1	5 %
D	0	0 %
Totaux	**20**	**100 %**

Tableau 3.2.1 > Tableau de fréquences dans le cas d'une variable qualitative nominale

Comme les classes ne présentent pas de gradation, l'ordre sur l'axe horizontal de l'histogramme n'est pas imposé. Il est courant d'utiliser l'ordre croissant alphabétique des classes ou l'ordre décroissant des fréquences absolues.

L'histogramme reprend les fréquences absolues non cumulées ou les fréquences relatives non cumulées ; les différentes colonnes de l'histogramme ne sont pas jointives afin de souligner le caractère discontinu de la variable :



Graphiques 4.2.1 / 4.2.2 > Histogrammes basés sur les fréquences absolues / fréquences relatives (%) dans le cas d'une variable qualitative nominale (graphiques générés par SAP Crystal Reports)

Comme les classes ne présentent pas de gradation, l'histogramme reprenant les fréquences absolues cumulées et l'histogramme reprenant les fréquences relatives cumulées n'ont pas de sens dans le cas d'une variable qualitative nominale.

4.3 Exemple dans le cas d'une variable qualitative ordinale

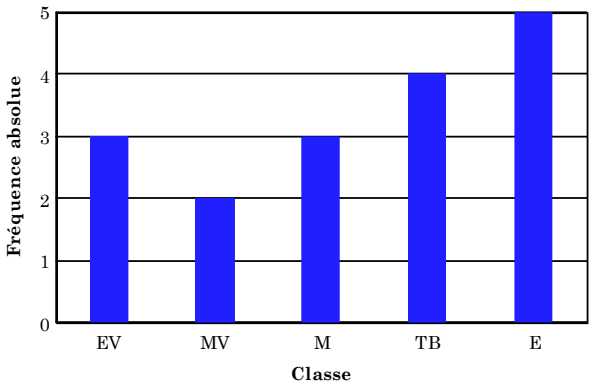
Reprenons le tableau de fréquences 3.3.1 :

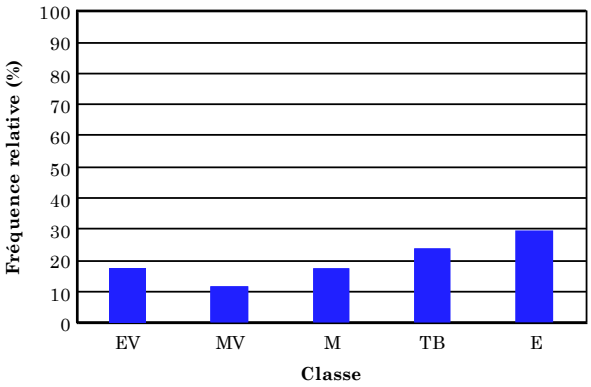
Classe x_i	**Fréquence absolue** n_i	**Fréquence relative** f_i	**Fréquence absolue cumulée** N_i	**Fréquence relative cumulée** F_i
EV	3	17,65 %	3	17,65 %
MV	2	11,76 %	5	29,41 %
M	3	17,65 %	8	47,06 %
TB	4	23,53 %	12	70,59 %
E	5	29,41 %	17	100,00 %
Totaux	**17**	**100 %**	**17**	**100 %**

Tableau 3.3.1 > Tableau de fréquences dans le cas d'une variable qualitative ordinale

Comme les classes présentent une gradation, l'ordre sur l'axe horizontal de l'histogramme est imposé.

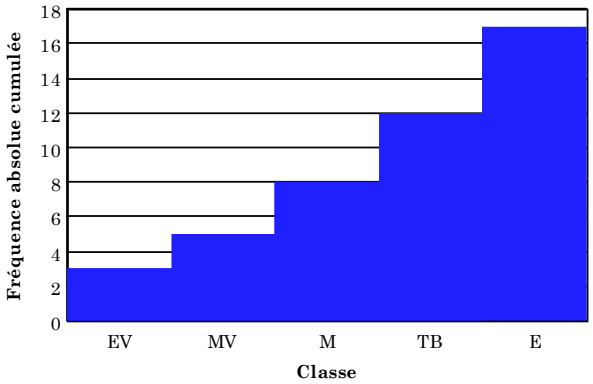
Si l'histogramme reprend les fréquences absolues non cumulées ou les fréquences relatives non cumulées, les différentes colonnes de l'histogramme ne sont pas jointives afin de souligner le caractère discontinu de la variable :

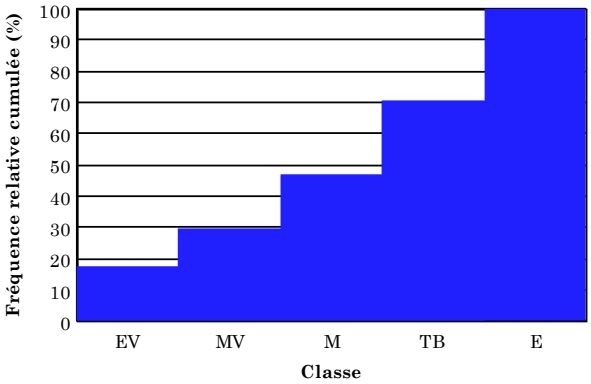




Graphiques 4.3.1 / 4.3.2 > Histogrammes basés sur les fréquences absolues / fréquences relatives (%) dans le cas d'une variable qualitative ordinale (graphiques générés par SAP Crystal Reports)

Si l'histogramme reprend les fréquences absolues cumulées ou les fréquences relatives cumulées, les différentes colonnes de l'histogramme sont jointives (l'histogramme cumulé représente l'intégrale de l'histogramme non cumulé) :





Graphiques 4.3.3 / 4.3.4 > Histogrammes basés sur les fréquences absolues cumulées / fréquences relatives cumulées (%) dans le cas d'une variable qualitative ordinale (graphiques générés par SAP Crystal Reports)

4.4 Exemple dans le cas d'une variable quantitative discontinue

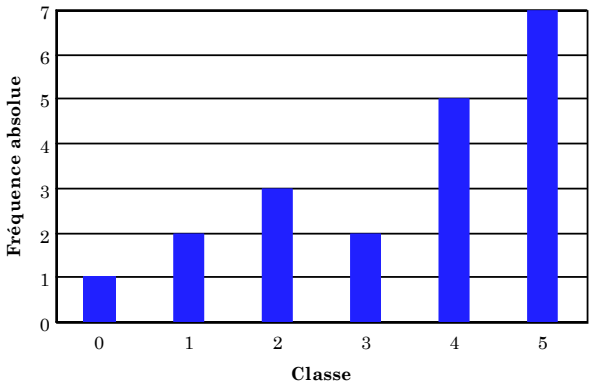
Reprenons le tableau de fréquences 3.4.1 :

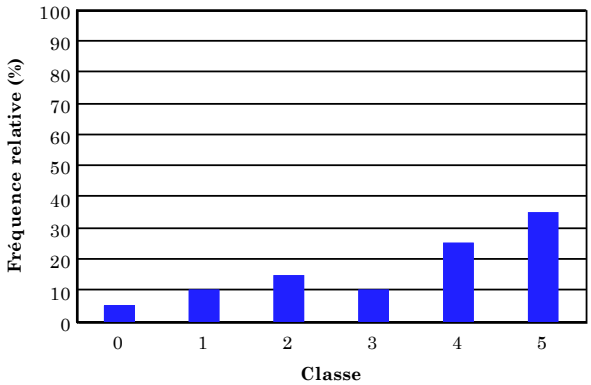
Classe x_i	Fréquence absolue n_i	Fréquence relative f_i	Fréquence absolue cumulée N_i	Fréquence relative cumulée F_i
0	1	5 %	1	5 %
1	2	10 %	3	15 %
2	3	15 %	6	30 %
3	2	10 %	8	40 %
4	5	25 %	13	65 %
5	7	35 %	20	100 %
Totaux	**20**	**100 %**	**20**	**100 %**

Tableau 3.4.1 > Tableau de fréquences dans le cas d'une variable quantitative discontinue

Comme chaque classe correspond à un nombre, l'ordre sur l'axe horizontal de l'histogramme est imposé (ordre croissant ou décroissant des classes).

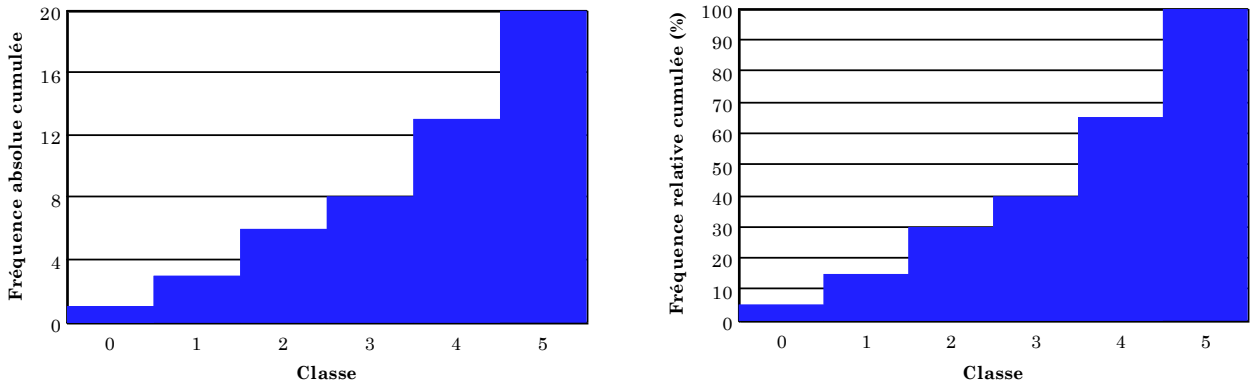
Si l'histogramme reprend les fréquences absolues non cumulées ou les fréquences relatives non cumulées, les différentes colonnes de l'histogramme ne sont pas jointives afin de souligner le caractère discontinu de la variable :





Graphiques 4.4.1 / 4.4.2 > Histogrammes basés sur les fréquences absolues / fréquences relatives (%) dans le cas d'une variable quantitative discontinue (graphiques générés par SAP Crystal Reports)

Si l'histogramme reprend les fréquences absolues cumulées ou les fréquences relatives cumulées, les différentes colonnes de l'histogramme sont jointives (l'histogramme cumulé représente l'intégrale de l'histogramme non cumulé) :



Graphiques 4.4.3 / 4.4.4 > Histogrammes basés sur les fréquences absolues cumulées / fréquences relatives cumulées (%) dans le cas d'une variable quantitative discontinue (graphiques générés par SAP Crystal Reports)

4.5 Exemple dans le cas d'une variable quantitative continue

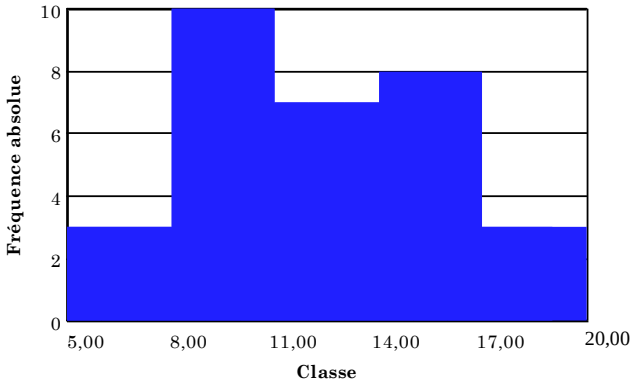
Reprenons le tableau de fréquences 3.5.1 :

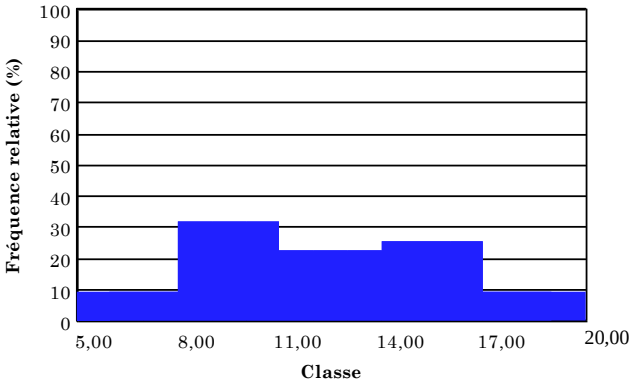
Classe x_i	**Centre** c_i	**Fréquence absolue** n_i	**Fréquence relative** f_i	**Fréquence absolue cumulée** N_i	**Fréquence relative cumulée** F_i
[5,0 – 8,0[	6,5	3	9,68 %	3	9,68 %
[8,0 – 11,0[	9,5	10	32,26 %	13	41,94 %
[11,0 – 14,0[	12,5	7	22,58 %	20	64,52 %
[14,0 – 17,0[	15,5	8	25,81 %	28	90,32 %
[17,0 – 20,0[	18,5	3	9,68 %	31	100,00 %
Totaux		**31**	**100 %**	**31**	**100 %**

Tableau 3.5.1 > Tableau de fréquences dans le cas d'une variable quantitative continue

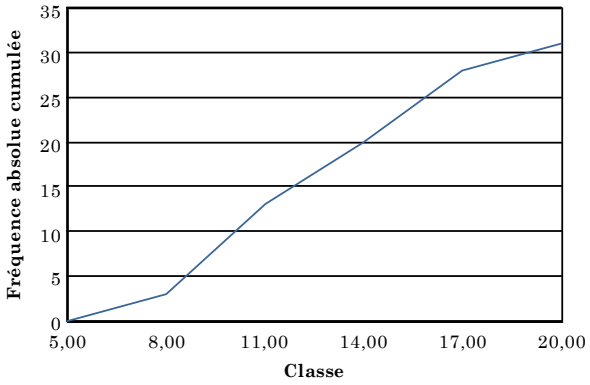
Si l'histogramme reprend les fréquences absolues non cumulées ou les fréquences relatives non cumulées, les différentes colonnes de l'histogramme sont jointives afin de souligner le caractère continu de la variable.

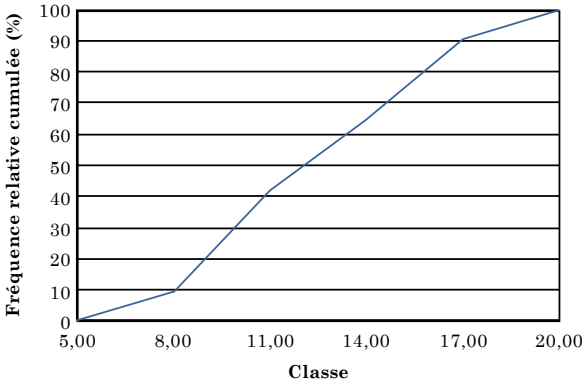
L'histogramme reprenant les fréquences absolues cumulées ou les fréquences relatives cumulées est un polygone, chaque classe étant traversée par un segment oblique. La borne inférieure de la première classe correspond à la valeur 0.





Graphiques 4.5.1 / 4.5.2 > Histogrammes basés sur les fréquences absolues / fréquences relatives (%) dans le cas d'une variable quantitative continue (graphiques générés par SAP Crystal Reports)





Graphiques 4.5.3 / 4.5.4 > Histogrammes basés sur les fréquences absolues cumulées / fréquences relatives cumulées (%) dans le cas d'une variable quantitative continue (graphiques générés par SAP Crystal Reports)