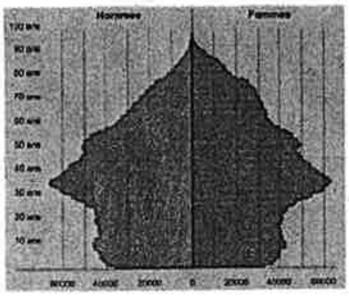
Thème 13: Quelques éléments de statistique descriptive I

13.1 Introduction, un peu d'histoire:

Introduction *Le mot statistique – de l'italien « statista », homme d'État – désignait à l'origine la collecte et l'évaluation des données concernant un État. Cette science de l'État était une représentation purement descriptive de faits géographiques et sociaux comme le climat, la population, les coutumes, les organisations économiques, etc..., à l'usage des hommes d'État ; à l'époque en France le roi et son conseil.*



Pyramide des âges Suisse 2000

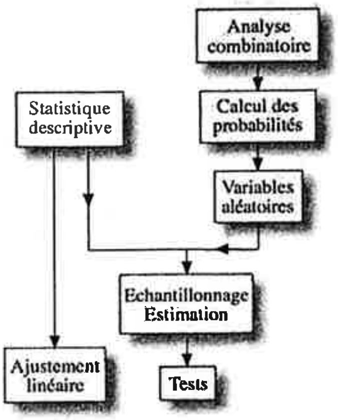
Dès la plus haute Antiquité, les dirigeants ont fait procéder à des enquêtes sur la population: l'Empereur Yao (vers 2200 av. J.-C.) pour connaître les productions agricoles, les pharaons égyptiens (dès 1700 av. J.-C.), l'Empereur Auguste à Rome pour le nombre de soldats, les revenus des citoyens.

Nous trouvons également de multiples exemples d'utilisation de statistiques dans les sciences :

Une page des données de l'astronome Tycho Brahé

- *Johannes Kepler (1571-1630) formula ses lois sur les mouvements des planètes en utilisant l'ensemble des données récoltées par l'astronome danois Tycho Brahé (1546-1601).*
- *Les premières études statistiques de Florence Nightingale, infirmière anglaise durant la guerre de Crimée de 1854 à 1856. permirent d'identifier les causes de mortalité des soldats et conduisirent à l'amélioration des conditions d'hygiène des hôpitaux militaires anglais.*

Aujourd'hui, cette partie des mathématiques a pris une grande place grâce aux nouvelles techniques et à la puissance des ordinateurs. Géographie, médecine, sciences humaines, sciences économiques, biologie, politique, aucun domaine n'est épargné.



On peut décomposer la méthode statistique en cinq étapes:

1. Identification précise de la population et du (des) caractère(s) à étudier
2. Récolte des données (recensement ou échantillonnage)
3. Regroupement, classification et présentation des données (statistiques descriptives)
4. Comparaison avec des modèles théoriques (calcul des probabilités et modèles probabilistes)
5. Interprétation, conclusion, prévision (inférence statistique)

13.2 Vocabulaire:

En statistique, le mot **population** représente un ensemble d'objets de même nature que l'on va étudier, analyser. Les éléments de la population, appelés **individus**, peuvent être des personnes, mais aussi des choses, des animaux, des objets, des faits, des notes de TE, etc... Le nombre d'individus est appelé l'**effectif**.

Souvent, il n'est pas possible de prendre en compte la totalité de la population. Dans ces cas, l'étude se limite à un ***échantillon****, pris au hasard, à partir duquel on peut tenter de déduire une tendance pour toute la population.*

Une population doit toujours être clairement définie afin que l'on puisse toujours déterminer si un élément quelconque fait ou non partie de la population étudiée. On pourra ainsi étudier une caractéristique que possède chacun des individus on appelle cela une **variable statistique (v.s)**.

Les différentes valeurs que peut prendre une variable statistique sont les **modalités** de cette variable.

Notation : On note une v.s par une lettre majuscule X, Y, ... et ses modalités par la même lettre minuscule affectée d'indices : x_1, x_2, ... pour la variable X ou y_1, y_2, ... pour la variable Y.

Modèle 1 : On fait une étude statistique auprès des élèves du gymnase de Morges. On aimerait connaître le sexe, l'âge au 1er janvier, la taille, la voie (ECGC ou EM) de chaque élève.

Population :

v.s	modalité des v.s
X : sexe	$x_1 =$ masculin $x_2 =$ féminin
Y : âge	$y_1 = 15$ ou $-$ $y_2 = 16$ $y_3 = 17$ $y_4 = 18$ $y_5 = 19$ ou $+$
Z : taille	$z_i \in [1.40 ; 2.10]$
U : voie	$u_1 =$ ECGC $u_2 =$ EM

Une v.s. est **quantitative** si les valeurs qu'elle peut prendre sont **numériques**. Une telle v.s est dite **quantitative discrète** si les valeurs qu'elle peut prendre sont isolées les unes des autres. Par contre, si celles-ci constituent des intervalles de nombres, la v.s est appelée **quantitative continue**. Si les valeurs d'une v.s sont descriptives ou nominatives, la v.s. est dite **qualitative**.

X est une v.s ..qualitative...., Y est une v.s ..quantitative..discrète

Z est une v.s ...quantitative continue..., U est une v.s ...qualitative....

Exercice 13.1: On a demandé aux employés d'une entreprise pour quel parti politique ils avaient voté lors des dernières élections. Voici les données brutes obtenues:

PS	PRD	PS	PDC	PS	UDC
PS	UDC	PRD	PS	verts	PDC
UDC	PRD	verts	UDC	UDC	UDC
PRD	PS	PRD	PDC	PRD	PDC
UDC	PDC	PS	UDC	UDC	UDC

a) Identifier la population ainsi que la variable statistique (v.s.).
b) Donner l'ensemble des modalités.
c) De quel type est cette variable statistique ?

Exercice 13.2: Un professeur de l'Uni a noté le nombre de points obtenus par 80 étudiants lors d'un test de statistiques.

2	3	5	5	4	6	6	5	4	3
7	7	7	6	2	7	7	9	8	10
5	6	6	8	6	6	3	7	3	5
9	7	6	4	7	5	9	9	6	9
6	3	9	8	8	7	5	6	10	6
9	7	7	7	4	7	10	8	7	10
3	5	8	5	8	7	4	8	10	7
4	6	6	8	7	7	7	8	8	9

a) Identifier la population ainsi que la variable statistique (v.s.).
b) Donner l'ensemble des modalités.
c) De quel type est cette variable statistique ?

Modèle 2 : En reprenant les données de l'exercice 13.1, on va sacrifier le caractère individuel de l'information afin d'obtenir un portrait d'ensemble. On calcule pour chaque modalité le nombre d'individus ayant cette modalité : l'**effectif** n_i de la modalité. Celui-ci ne permet pas de comparer deux populations inégales. Il sera alors naturel de calculer la proportion de la population qui a une telle modalité. On définit alors la **fréquence relative** f_i par le rapport entre l'effectif de chaque modalité et le nombre N d'individus de la population: $f_i = \frac{n_i}{N}$

$N = 30$

Pour dessiner le diagramme en secteur.

$f_i \cdot 360°$

Modalité x_i	Effectif n_i	Fréquence relative f_i	Angle
PS	7	7/30 ≅ 23,3 %	84°
PRD	6	6/30 = 20 %	72°
PDC	5	5/30 ≅ 16,7 %	60°
UDC	10	10/30 ≅ 33,3 %	120°
Verts	2	2/30 ≅ 6,7 %	24°
Total:	30	30/30 = 100 %	360°

Exercice 13.1: On a demandé aux employés d'une entreprise pour quel parti politique ils avaient voté lors des dernières élections. Voici les données brutes obtenues:

PS	PRD	PS	PDC	PS	UDC
PS	UDC	PRD	PS	verts	PDC
UDC	PRD	verts	UDC	UDC	UDC
PRD	PS	PRD	PDC	PRD	PDC
UDC	PDC	PS	UDC	UDC	UDC

a) Identifier la population ainsi que la variable statistique (v.s.).
b) Donner l'ensemble des modalités.
c) De quel type est cette variable statistique ?

Exercice 13.2: Un professeur de l'Uni a noté le nombre de points obtenus par 80 étudiants lors d'un test de statistiques.

2	3	5	5	4	6	6	5	4	3
7	7	7	6	2	7	7	9	8	10
5	6	6	8	6	6	3	7	3	5
9	7	6	4	7	5	9	9	6	9
6	3	9	8	8	7	5	6	10	6
9	7	7	7	4	7	10	8	7	10
3	5	8	5	8	7	4	8	10	7
4	6	6	8	7	7	7	8	8	9

a) Identifier la population ainsi que la variable statistique (v.s.).
b) Donner l'ensemble des modalités.
c) De quel type est cette variable statistique ?

Modèle 2 : En reprenant les données de l'exercice 13.1, on va sacrifier le caractère individuel de l'information afin d'obtenir un portrait d'ensemble. On calcule pour chaque modalité le nombre d'individus ayant cette modalité : l'**effectif** n_i de la modalité. Celui-ci ne permet pas de comparer deux populations inégales. Il sera alors naturel de calculer la proportion de la population qui a une telle modalité. On définit alors la **fréquence relative** f_i par le rapport entre l'effectif de chaque modalité et le nombre N d'individus de la population: $f_i = \frac{n_i}{N}$

$N = 30$

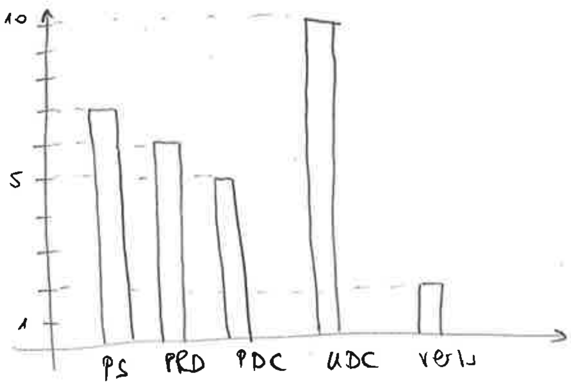
Pour dessiner le diagramme en secteur.

$f_i \cdot 360°$

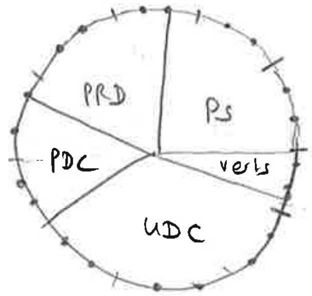
Modalité x_i	Effectif n_i	Fréquence relative f_i	Angle
PS	7	7/30 ≅ 23,3 %	84°
PRD	6	6/30 = 20 %	72°
PDC	5	5/30 ≅ 16,7 %	60°
UDC	10	10/30 ≅ 33,3 %	120°
Verts	2	2/30 ≅ 6,7 %	24°
Total:	30	30/30 = 100 %	360°

Le **tableau de distribution des effectifs et des fréquences** permet une bonne synthèse des informations, mais n'est pas très explicite. On l'accompagnera d'un graphique permettant de représenter ces données. On utilise fréquemment :

a) un diagramme en colonnes (histogramme)



b) un diagramme en secteurs (en "camembert")



Remarques :

- La somme des effectifs est toujours égale au nombre d'individus de la population:

$$n_1 + n_2 + \ldots + n_k = N$$

- La somme des fréquences est toujours égale à 1:

$$f_1 + f_2 + \ldots + f_k = 1$$

car:

$$f_1 + f_2 + \ldots + f_k = \frac{n_1}{N} + \frac{n_2}{N} + \ldots + \frac{n_k}{N} =$$

$$= \frac{n_1 + n_2 + \ldots + n_k}{N} = \frac{N}{N} = 1.$$

Lors de l'utilisation de la calculatrice, il peut arriver que la somme des fréquences ne soit pas exactement égale à 1 à cause des arrondis de calculs.

Exercice 13.3: Reprendre les données de l'exercice 13.2 afin d'en proposer :

a) le tableau de distribution des effectifs et des fréquences ;

b) un histogramme puis un diagramme en secteurs.

Exercice 13.4: On a demandé aux enfants de trois classes de $3^{ème}$ année primaire quel était leur sport d'hiver préféré. On a obtenu les données brutes suivantes:

Hockey	Glissade	Hockey	Hockey	Hockey
Hockey	Ski	Hockey	Ski	Raquette
Patinage	Ski	Ski	Hockey	Ski
Ski	Hockey	Ski	Raquette	Ski
Patinage	Ski	Hockey	Raquette	Raquette
Ski	Glissade	Hockey	Glissade	Glissade
Hockey	Glissade	Hockey	Hockey	Hockey
Ski de fond	Hockey	Patinage	Patinage	Hockey
Ski	Hockey	Ski	Raquette	Patinage
Hockey	Glissade	Ski	Ski	Ski de fond
Hockey	Patinage	Ski	Patinage	Hockey
Hockey	Patinage	Ski	Patinage	Raquette

a) Identifier la population.
b) Caractériser la variable statistique.
c) Donner l'ensemble des modalités.
d) Le tableau des distributions des effectifs et des fréquences.
e) Faire un diagramme en secteurs.

Exercice 13.5: On étudie l'état civil des 30 employés (numérotés de 1 à 30) d'une petite entreprise.

1	Marié	11	Marié	21	Célibataire
2	Mariée	12	Célibataire	22	Marié
3	Célibataire	13	Marié	23	Veuf
4	Divorcé	14	Veuve	24	Célibataire
5	Marié	15	Marié	25	Divorcée
6	Célibataire	16	Divorcé	26	Divorcé
7	Célibataire	17	Célibataire	27	Marié
8	Mariée	18	Mariée	28	Marié
9	Mariée	19	Marié	29	Marié
10	Divorcée	20	Marié	30	Marié

a) Identifier la population
b) Caractériser la variable statistique
c) Donner l'ensemble des modalités
d) Le tableau des distributions des effectifs et des fréquences
e) Proposer l'histogramme **des effectifs** de cette v.s.
f) Proposer l'histogramme **des fréquences** de cette v.s.
g) Comparer ces 2 représentations graphiques

13.3 Regroupement des données à l'intérieur de classes de valeurs

Souvent, lors d'une étude statistique portant sur une variable statistique quantitative discrète ou continue, les données recueillies diffèrent à peu près toutes les unes des autres et sont étalées sur un large intervalle de valeurs. L'objectif de la statistique descriptive étant de résumer de la façon la plus adéquate possible cet ensemble de données, les mesures seront regroupées dans des intervalles de valeurs que l'on appelle des classes du type $[b_{i-1} ; b_i[$.

b_{i-1} est **la borne inférieure** de la classe i;

b_i est **la borne supérieure** de la classe i;

$x_i = \dfrac{b_{i-1}+b_i}{2}$ est **le milieu** de la classe i;

$L_i = b_i - b_{i-1}$ est **la largeur** de la classe i.

Modèle 3 : Des chimistes viennent de composer une nouvelle fibre synthétique qui devrait se caractériser par sa résistance. Afin de vérifier sa capacité de tension, on prélève de la production, au hasard, un échantillon de 60 fibres qu'on soumet à des essais de résistance. Les résultats (en kg) sont les suivants :

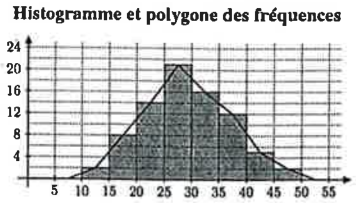
35	65	71	75	77	80	81	82	84	86	87	89	91	97	100
48	69	72	75	78	80	81	83	85	86	88	89	94	97	103
53	69	73	76	79	80	81	83	85	87	88	89	95	99	104
63	71	74	77	79	81	82	84	86	87	89	91	97	99	114

a) On regroupe les données en 6 classes d'amplitude 15 avec 30 comme valeur minimale. Compléter le tableau:

Classe	Centre	Effectif	Fréquence
[30 ; 45[	37.5	1	0,017
[45 ; 60[	52.5	2	0,033
[60 ; 75[	67.5	9	0,15
[75 ; 90[	82.5	35	0,583
[90 ; 105[	97.5	12	0,2
[105 ; 120[	112.5	1	0,017
Totaux		60	1

Représentations graphiques:

- L'**histogramme** est un diagramme en colonnes où les rectangles sont juxtaposés indiquant ainsi le caractère continu de la variable statistique.

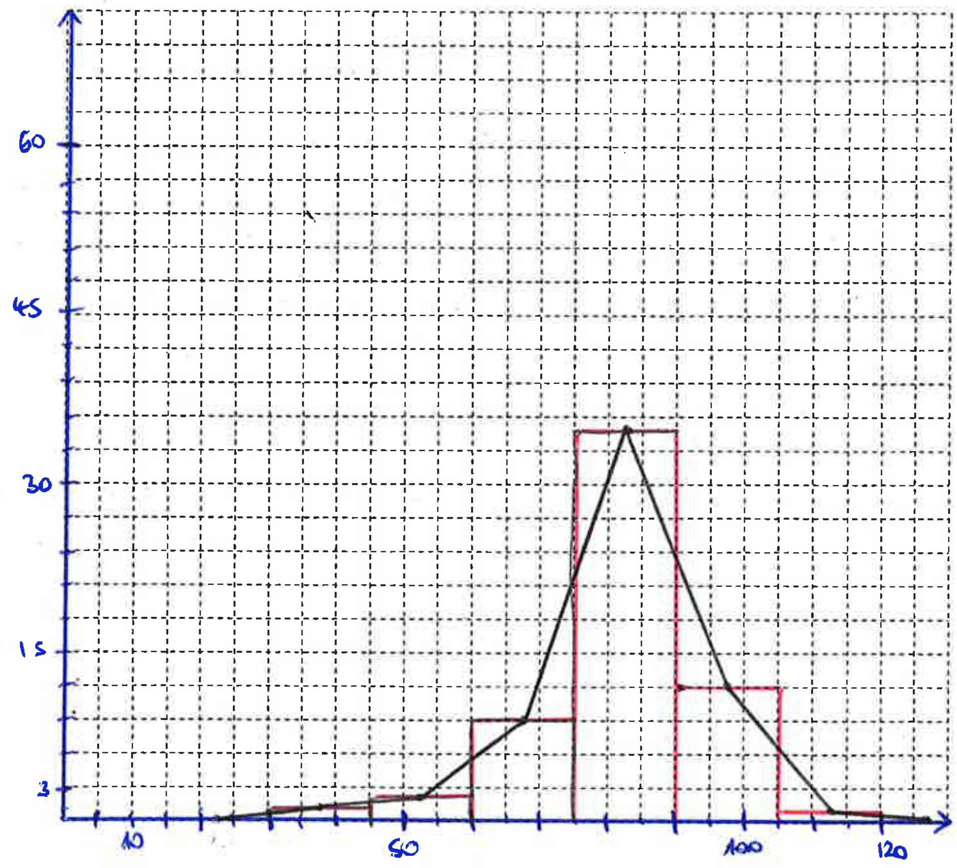


- Le **polygone des fréquences** est la ligne polygonale obtenue en joignant les points milieux consécutifs des sommets des rectangles de l'histogramme. On commence et on termine le polygone des fréquences en ajoutant une classe de fréquence nulle avant la première classe et une autre après la dernière classe.

Modèle 3 Suite : À propos de cette nouvelle fibre synthétique:

b) Effectuer l'histogramme.

c) Construire le polygone des fréquences.



Exercice 13.6: Une entreprise a enregistré le salaire de tous ses vendeurs pour l'année dernière. Voici les données rangées:

Classes (salaires)	x_i	n_i	f_i
[10000 ; 15000[	12500	2	2,50 %
[15000 ; 20000[	17500	8	10,00%
[20000 ; 25000[	22500	14	17,5 %
[25000 ; 30000[	27500	21	26,25%
[30000 ; 35000[	32500	16	20,00%
[35000 ; 40000[	37500	12	15,00%
[40000 ; 45000[	42500	5	6,25%
[45000 ; 50000[	47500	2	2,50%
	Totaux	80	100

a) Compléter le tableau des distributions des effectifs et des fréquences.
b) Faire un histogramme.
c) Construire le polygone des fréquences.

Exercice 13.7: En recevant les élèves qui désirent faire partie d'une équipe de foot du gymnase, l'entraîneur a pris note du poids de ces 60 joueurs:

72,6	81,9	84,7	88,1	89,4	91,6	93,7	95,8	99,1	103,2
75,8	82,6	85,4	88,1	90,2	92,4	93,9	96,6	99,4	103,9
77,5	82,9	86,2	88,3	90,9	92,5	94,4	97,1	99,8	104,0
78,3	83,0	86,9	88,7	91,1	92,8	94,7	97,2	100,4	105,2
79,6	83,5	87,3	89,0	91,2	93,0	94,8	97,5	101,7	106,1
81,5	84,1	87,8	89,1	91,3	93,3	95,2	98,3	102,1	118,7

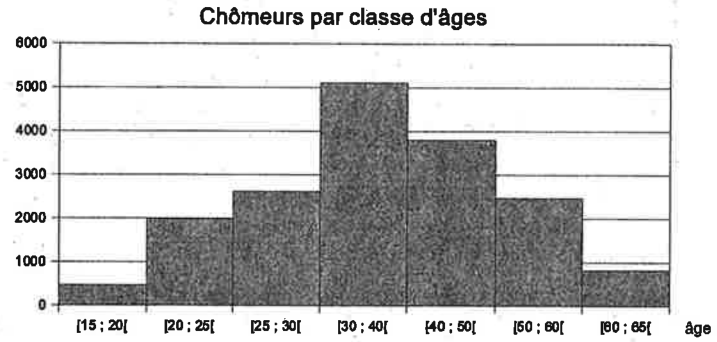
a) Identifier la population.
b) Identifier la variable statistique.
c) Cette variable statistique est-elle discrète ou continue ?
d) En utilisant des classes de largeur 5, construire le tableau des distributions des effectifs et des fréquences (valeur minimale: 70). Vous admettrez une classe plus large à l'extrémité (classe [105 ; 120[ ),
f) Construire le polygone des fréquences.

Exercice 13.8: Le tableau récapitulatif suivant donne la statistique trimestrielle par classe d'âges des chômeurs inscrits dans un office du travail dans le canton de Vaud en juin 2009:

Chômeurs par classe d'âges

Classe d'âges	Effectif	Fréquence [%]
[15 ; 20[	472	2,73
[20 ; 25[	1'990	11,51
[25 ; 30[	2'621	15,16
[30 ; 40[	5'110	29,56
[40 ; 50[	3'798	21,97
[50 ; 60[	2'476	14,32
[60 ; 65[	821	4,75
Totaux	17'288	100

a) En quoi l'histogramme suivant est-il trompeur ?



b) Proposer un nouvel histogramme corrigeant cet effet visuel trompeur.

13.4 Fréquences cumulées des v.s. quantitatives

Dans une étude statistique, si on souhaite connaître la proportion de chaque valeur que peut prendre la variable statistique étudiée, on regarde sa **fréquence** f_i.

Si par contre on souhaite connaître la proportion des individus qui présentent des valeurs inférieures ou égales à une valeur fixée, on regarde la **fréquence cumulée croissante** F_i.

Observons ceci sur le modèle suivant :

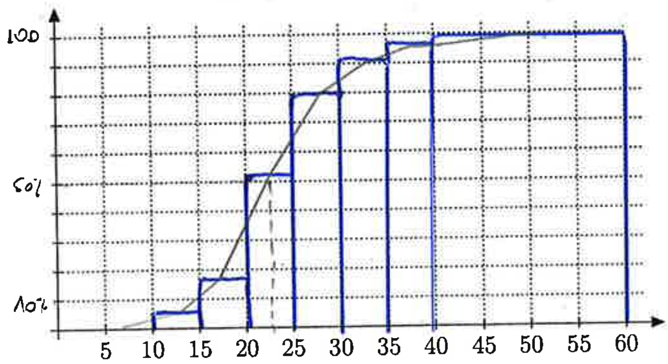
Modèle 4 : Lors d'un concours de pêche dans le lac de Bret, on a mesuré (en cm) toutes les prises et regroupées par classe dans le tableau qui suit :

a) Compléter le tableau que l'on a obtenu :

Classe $[b_{i-1} ; b_i[$	Centre x_i	Effectifs n_i	Fréquence f_i	Fréquence cumulée croiss F_i
[10 ; 15[	12,5	4	6,25 %	6,25 %
[15 ; 20[	17,5	8	12,5 %	18,75 %
[20 ; 25[	22,5	21	32,81 %	51,56 %
[25 ; 30[	27,5	18	28,13 %	79,69 %
[30 ; 35[	32,5	7	10,94 %	90,63 %
[35 ; 40[	37,5	5	7,81 %	98,44 %
[40 ; 60[	50	1	1,56 %	100 %
	Totaux :	64	100 %	

(= 6,25 + 12,5)

b) Représenter le diagramme des fréquences cumulées.



c) Déterminer la proportion des prises dont la longueur est plus petite que 30 cm.

C'est 79,69 %

d) Déterminer la proportion des prises dont la longueur est plus grande ou égale à 20 cm.

C'est 100 − 18,75 = 81,25 %

La médiane : L'abscisse du point correspondant à une fréquence cumulée croissante de 50% s'appelle **la médiane** de la v.s.

Dans le modèle précédent, la médiane vaut:

environ 22,5 cm.

Exercice 13.9: Compléter la solution de l'exercice 13.8 par:

c) Représenter les courbes des fréquences cumulées croissantes et en déduire la valeur de la médiane.
d) Déterminer la proportion de chômeurs de moins de 40 ans.
e) Déterminer la proportion de chômeurs de plus de 20 ans.
f) Déterminer la proportion des chômeurs entre 20 et 40 ans.

Exercice 13.10: Lors d'un cours de statistique, en 2008, 20 étudiants ont été invités à indiquer leur taille et leur poids.

N° d'ordre	taille en cm.	poids en kg
1	174	64
2	175	59
3	180	64
4	168	62
5	175	50
6	170	60
7	170	68
8	160	63
9	187	93
10	178	70

N° d'ordre	taille en cm.	poids en kg.
11	170	64
12	182	72
13	168	60
14	171	55
15	181	80
16	178	82
17	180	72
18	180	78
19	178	71
20	182	72

a) Regrouper les données de tailles et de poids en 6 classes de largeur égale.
b) Représenter l'histogramme et le polygone des fréquences des tailles des 20 étudiants.
c) le diagramme des fréquences cumulées croissantes des poids des 20 étudiants.