Statistiques

Table des Matières

I. Introduction et définition 1

II. Critères statistiques 2
- II. A. Critère qualitatif . . . 2
- II. B. Critère quantitatif discret . . . 3
- II. C. Critère quantitatif continu . . . 4

III. Fréquences 5

IV. Indicateurs statistiques 7
- IV. A. Indicateurs de positions centrales . . . 7
  - IV. A. 1. Moyenne arithmétique . . . 7
  - IV. A. 2. Médiane . . . 8
- IV. B. Indicateurs de dispersion . . . 8
  - IV. B. 1. Quartiles et écart-interquartile . . . 8
  - IV. B. 2. Écart-type . . . 10

☙ Statistiques ❧

I. Introduction et définition

Dès que les premières structures sociales se sont mises en place, les hommes ont utilisé la science des statistiques pour d'abord compter le bétail, les hommes, les objets et garder une mémoire de ces comptages sur des écrits, les supports étaient des os ou du bois. Les plus anciennes traces de comptage datent des premières civilisation du Paléolithique (30 000 ans environ av. J.-C.).

L'utilisation des **recensements** a très vite pris sa place. Les premières traces de recensement apparaissent en Chine, XXIII ^{e} siècle av. J.-C.et en Égypte XVII ^{e} siècle av. J.-C..
À Rome Cicéron I^{er} siècle av. J.-C.insistait sur l'importance des recensements : *"Il est nécessaire au sénateur d'avoir une notion complète de l'État ; et cela s'étend loin : savoir l'effectif de l'armée, la puissance financière, les alliés, amis et tributaires que possède l'État ; [...] connaître les précédents traditionnels des décisions à prendre, l'exemple des ancêtres... Vous voyez enfin tout ce que cela comporte en général de savoir, d'application, de mémoire, et sur quoi un sénateur ne saurait en aucune manière se trouver pris au dépourvu."*
Le recensement romain permettait à la fois, de connaître les ressources en hommes mobilisables et en biens, et de classer les citoyens afin de répartir charges et avantages. Le recensement était également une démonstration de puissance, permettant de proclamer publiquement l'ampleur de la domination romaine.
Selon Tacite, l'empereur Auguste aurait été le premier à faire un bilan des richesses de l'empire romain (soldats, navires, ressources privées et publiques). Au III ^{e} siècle apparaissent à Rome des tables d'estimation des rentes viagères.
À partir du XIII ^{e} siècle, les données deviennent plus nombreuses. Les commerçants de Venise amassent des données sur le commerce extérieur, évaluent les risques maritimes. En Hollande, on étudie les rentes viagères. Au XV ^{e} siècle la tenue des registres des naissances est rendue obligatoire en France, par François I^{er}, puis, sous Henri III, ceux des mariages et naissances.
En Europe, ce système de recueil de données se poursuit jusqu'au XVII ^{e} siècle. En Europe, le rôle "statisticien" est souvent tenu par des guildes marchandes, puis par les intendants de l'État.
John Graunt dans Natural and Political Observations Made upon the Bills of Mortality a estimé la population de Londres en 1662 en s'aidant des registres paroissiaux. Il savait qu'il y avait environ 13 000 enterrements par an à Londres et que trois personnes pour onze familles mouraient par an. Il a estimé à partir des registres paroissiaux que la taille moyenne de la famille était de 8 et a calculé que la population de Londres était d'environ 384 000. Le mathématicien Laplace (1802), utilise une méthode similaire pour estimer la population française. À la suite des travaux fondateurs de Graunt (1620-1674) sur les bulletins de décès et les naissances (il découvre ainsi la proportion plus grande de naissances masculines : 107 pour 100 naissances féminines), l'économiste William Petty (1623-1687) systématise et théorise les études démographiques sur les naissances, décès, nombres de personnes par famille...
En 1696, l'astronome anglais Edmond Halley (1662 – 1742), en se basant sur cinq ans d'état civil de la ville de Breslau (Pologne), établit une table de mortalité, préfigurant les travaux d'actuariat.
En Hollande, le calcul des probabilités est appliqué à l'espérance de vie humaine (Christian et Louis Huygens en 1669) et à l'estimation du prix d'achat d'une rente, à l'aide de tables de mortalité (Jan De Witt en 1671).

Source : math93.fr

Actuellement, le mot **statistique** désigne à la fois un ensemble de données d'observation et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation.

Définition

Les statistiques consistent dans un premier temps à réunir des **données**, aujourd'hui on parle de "big data" (grosses données) de longues listes de données obtenues notamment par l'utilisation d'Internet.
Les premières analyses des données recueillies décrivent les données, on parle de **statistiques descriptives**.
Les indicateurs sont de positions et de dispersions sont les outils nécessaires à ces descriptions.
Enfin, les méthodes de traitement et d'interprétation des observations relient les données aux modèles théoriques, des modèles de probabilités. On parle d' **inférence statistique**.

II. Critères statistiques

II. A. Critère qualitatif

✒ Définition

Une série statistique à une variable est dite qualitative si la variable ne se compte pas, elle indique une qualité.

Représentation graphique :

Une série statistique à une variable qualitative est représentée par un diagramme circulaire, en tubes etc...

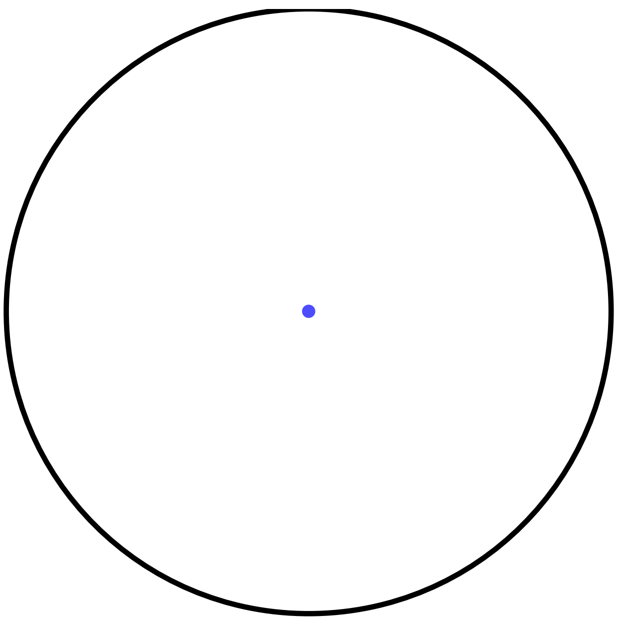
☞**Exercice 1**

Le tableau suivant présente la répartition des emplois suivant les catégories socioprofessionnelles en 2 015 (d'après l'INSEE) :

catégories socioprofessionnelles	Agriculteurs	Artisans commerçants	Cadres prof.Intellectuelle
nombre d'emplois	6038	8774	14 685
Angle			

catégories socioprofessionnelles	Professions intermédiaires	Employés Employés	Ouvriers Ouvriers
nombre d'emplois	30706	40988	36699
Angle			

1. Critère de la série :
2. Nommer la variable de la série :
3. Faire un diagramme circulaire de la série :



Florence Nightingale (1820-1910), une infirmière anglaise est la première à utiliser les diagrammes circulaires, développés par William Playfair en 1801. Après la Guerre de Crimée, elle se met à utiliser une version améliorée de ces diagrammes (équivalant aux diagrammes circulaires d'aujourd'hui), afin d'illustrer les causes saisonnières de mortalité des patients de l'hôpital militaire qu'elle gère. Par la suite, Nightingale réalise une étude statistique complète du système sanitaire dans les campagnes indiennes .

II. B. Critère quantitatif discret

Définition

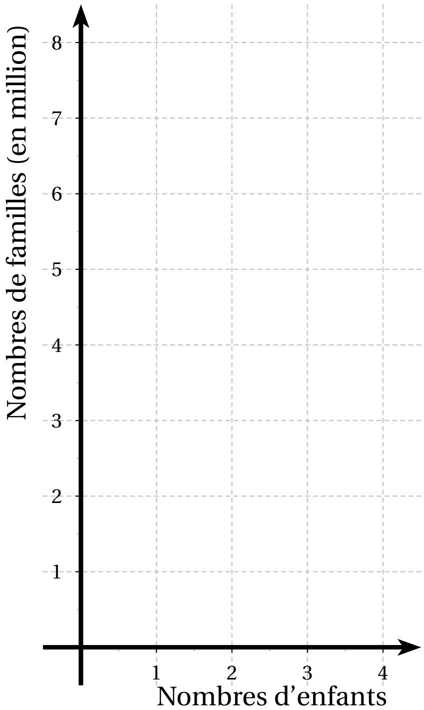
Une série statistique à une variable est dite quantitative discrète si la variable est dénombrable, généralement la variable est un ensemble fini d'entiers naturels.
Représentation graphique :
Une série statistique à une variable quantitative discrète est représentée par un diagramme en bâtons.

Exercice 2

Le tableau suivant présente le nombre d'enfants par famille française métropolitaine en 2015 (d'après l'INSEE) :

Nombre d'enfant(s)	0	1	2	3	4
Nombre de familles	7 492 332	3 615 859	3 255 259	1 267 979	465 353

1. Critère de la série :
2. Nommer la variable de la série :
3. Faire un diagramme un bâtons de la série :



II. C. Critère quantitatif continu

Définition

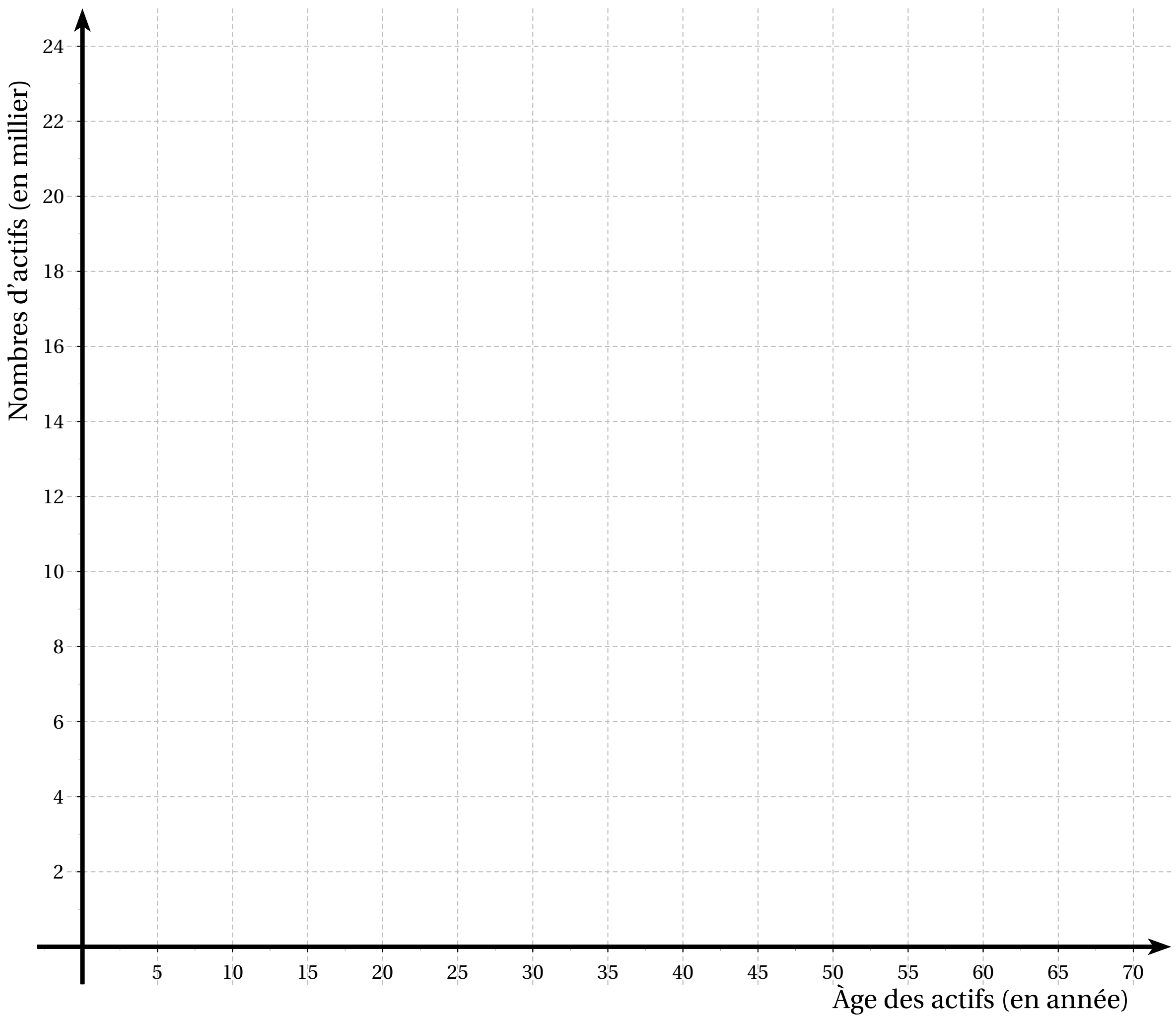
Une série statistique à une variable est dite quantitative continue si la variable est à valeurs dans un intervalle, généralement la variable est décrite par un nombre fini d'intervalles.
Représentation graphique :
Une série statistique à une variable quantitative continue est représentée par un histogramme.

Exercice 3

Le tableau suivant présente la répartition des actifs suivant l'âge, en France métropolitaine, en 2 015 (d'après l'INSEE) :

Âge	[15 ; 20[	[20 ; 25[	[25 ; 30[	[30 ; 35[	[35 ; 40[	[40 ; 45[	[45 ; 50[	[50 ; 55[	[55 ; 60[	[60 ; 65[
Effectifs	1 075	7 236	16 676	18 348	19 982	20 704	22 291	19 821	9 296	2 927

1. Critère de la série :
2. Nommer la variable de la série :
3. Faire un histogramme de la série :



III. Fréquences

Soit une série statistique quantitative rangée dans l'ordre croissant on donne les notations suivantes (* renvoie à la définition qui suit le tableau) :

Variable x_i	x_1	x_2	...	x_p	Total
Effectif n_i	n_1	n_2	...	n_p	n
Fréquence f_i *	f_1	f_2	...	f_p	1 (= 100%)
Fréquence cumulée croissante fcc_i*	$fcc_1 = f_1$	$fcc_2 = fcc_1 + f_2$	...	$fcc_p = 1$	
Fréquence cumulée décroissante fcd_i*	$fcd_1 = 1$	$fcd_2 = fcd_1 - f_1$	...	$fcd_p = f_p$	

Définition

- La fréquence (une proportion, une part) f_i du critère x_i d'une série statistique est le quotient de l'effectif n_i par l'effectif total n :

$$f_i = \frac{n_i}{n}$$

- La fréquence cumulée croissante fcc_i du critère x_i est la somme pour k variant de 1 à i des fréquences f_k :

$$fcc_i = f_1 + f_2 + ... + f_i = \sum_{k=1}^{i} f_k$$

- La fréquence cumulée décroissante fcd_i du critère x_i est la somme pour k variant de i à p des fréquences f_k :

$$fcc_i = f_i + f_{i+1} + ... + f_p = \sum_{k=i}^{p} f_k$$

Remarque

On ne calcule pas de fréquences cumulées pour les séries statistiques dont la variable est qualitative (il n'y a pas de relation d'ordre) ; par contre on peut calculer des fréquences pour ces séries qualitatives.

Exercice 4

Le tableau suivant présente la répartition des emplois suivant les catégories socioprofessionnelles en 2 015 (d'après l'INSEE) :

catégories socioprofessionnelles	Agriculteurs	Artisans commerçants	Cadres prof.Intellectuelle
nombre d'emplois	6038	8774	14 685
Fréquence			

catégories socioprofessionnelles	Professions intermédiaires	Employés	Ouvriers
nombre d'emplois	30706	40988	36699
Fréquence			

1. Compléter le tableau par les fréquences, données en pourcentage et arrondies à 1% soit à 0,01.
2. Interpréter la fréquence de la colonne Employés

☞**Exercice 5**

Le tableau suivant présente le nombre d'enfants par famille française métropolitaine en 2 015 (d'après l'INSEE) :

Nombre d'enfant(s)	0	1	2	3	4
Nombre de familles	7 492 332	3 615 859	3 255 259	1 267 979	465 353
Fréquences					
Fréquences cumulées croissantes					
Fréquences cumulées décroissantes					

1. Compléter le tableau, les fréquences seront arrondies à $0,1\%$ soit à $0,001$.
2. Interpréter la fréquence, puis la fréquence cumulée croissante, puis la fréquence cumulée décroissante pour 3 enfants.

☞**Exercice 6**

Le tableau suivant présente la répartition des actifs suivant l'âge, en France métropolitaine, en 2 015 (d'après l'INSEE) :

Âge	[15 ; 20[	[20 ; 25[	[25 ; 30[	[30 ; 35[	[35 ; 40[	[40 ; 45[	[45 ; 50[	[50 ; 55[	[55 ; 60[	[60 ; 65[
Effectifs	1 075	7 236	16 676	18 348	19 982	20 704	22 291	19 821	9 296	2 927
Fréquences										
Fréquences cumulées croissantes										
Fréquences cumulées décroissantes										

1. Compléter le tableau, les fréquences seront arrondies à 1% soit à $0,01$.
2. Interpréter la fréquences, puis la fréquence cumulée croissante, puis la fréquence cumulée décroissante pour la classe d'âge $[30\ ;\ 35]$.

IV. Indicateurs statistiques

IV. A. Indicateurs de positions centrales

IV. A. 1. Moyenne arithmétique

Avec les notations suivantes :

Variable x_i	x_1	x_2	...	x_p	Total
Effectif n_i	n_1	n_2	...	n_p	n
Fréquence f_i	f_1	f_2	...	f_p	1 (= 100%)

Définition

La **moyenne arithmétique** $\overline{x}$ d'une série statistique quantitative est le quotient de la somme des valeurs de la série par l'effectif total :

$$\overline{x} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_p n_p}{n} = \frac{\sum_{k=1}^{p} x_k n_k}{n} = \frac{1}{n} \sum_{k=1}^{p} x_k n_k$$

$$\overline{x} = x_1 \frac{n_1}{n} + x_2 \frac{n_2}{n} + \ldots + x_p \frac{n_p}{n} = x_1 f_1 + x_2 f_2 + \ldots + x_p f_p = \sum_{k=1}^{p} x_k f_k$$

Dans ce précédant cas on dit que la moyenne est **pondérée** par les fréquences.

Exercice 7

Le tableau suivant présente le nombre d'enfants par famille française métropolitaine en 2 015 (d'après l'INSEE) :

Nombre d'enfant(s)	0	1	2	3	4
Nombre de familles	7 492 332	3 615 859	3 255 259	1 267 979	465 353

1. Calculer la moyenne $\overline{x}$ de la série statistique.
2. Interpréter la moyenne $\overline{x}$ par une phrase.

Exercice 8

Le tableau suivant présente la répartition des actifs suivant l'âge, en France métropolitaine, en 2 015 (d'après l'INSEE) :

Âge	[15 ; 20[	[20 ; 25[	[25 ; 30[	[30 ; 35[	[35 ; 40[	[40 ; 45[	[45 ; 50[	[50 ; 55[	[55 ; 60[	[60 ; 65[
Centre classe										
Effectifs	1 075	7 236	16 676	18 348	19 982	20 704	22 291	19 821	9 296	2 927

1. Compléter le tableau par le centre de la classe.
2. Calculer la moyenne $\overline{x}$ en utilisant le centre de la classe pour représenter la classe.
3. Interpréter la moyenne $\overline{x}$ par une phrase.

Propriété

Soit une moyenne pondérée $\overline{x}$ et a et b deux nombres réels.

$$\overline{ax+b} = a\overline{x} + b$$

☞**Démonstration 1**

- $\overline{ax+b} = \dfrac{(ax_1+b)n_1 + (ax_2+b)n_2 + ... + (ax_p+b)n_p}{n}$
- $a\overline{x}+b = a\left(\dfrac{x_1n_1 + x_2n_2 + ... + x_pn_p}{n}\right) + b$

Partir de l'une ou l'autre des deux expressions pour aboutir, après calculs à l'autre expression.

☞**Exercice 9**

Le prix (en €) d'un carburant a été successivement de 1,55 ; 1,58; 1,60 ; 1,63 ; 1,62 ; 1,59.
Calculer mentalement le prix moyen $\overline{x}$ de ces six prix.
Expliquer la démarche.

IV. A. 2. Médiane

Edward Wright, dans le livre *Erreurs certaines en navigation* de 1599 considère que la médiane était la plus pertinente dans une série d'observation pour se localiser à partir d'une boussole.

Définition

Soit une série statistique quantitative discrète, telle que les valeurs de la variable sont rangées en ordre croissant.

- Dans le cas d'un effectif total n pair, la médiane M est la moyenne des deux valeurs centrales de la série.
- Dans le cas d'un effectif total n impair, la médiane M est la valeur centrale de la série.

IV. B. Indicateurs de dispersion

IV. B. 1. Quartiles et écart-interquartile

Définition

Soit une série statistique quantitative discrète, telle que les valeurs de la variable sont rangées en ordre croissant.

- Le premier quartile Q_1 est la plus petite valeur de la série telle qu'au moins 25% des valeurs de la série ait une valeur inférieure ou égale à Q_1.
- Le troisième quartile Q_3 est la plus petite valeur de la série telle qu'au moins 75% des valeurs de la série ait une valeur inférieure ou égale à Q_3.

On appelle écart-interquartile le nombre $Q_3 - Q_1$.

Remarque

On peut rédiger la définition ainsi :
Le i^{eme} quartile Q_i est la plus petite valeur de la série telle qu'au moins $\left(i \times \frac{1}{4} \times 100\right)$ % des valeurs de la série ait une valeur inférieure ou égale à Q_i, avec i dans $\{1;2;3\}$.
Ainsi le deuxième quartile peut être assimilé à la médiane de la série.
On peut définir n'importe quel partage fractionné de la série comme par exemple :
Les déciles :
Le i^{eme} décile D_i est la plus petite valeur de la série telle qu'au moins $\left(i \times \frac{1}{10} \times 100\right)$ % des valeurs de la série ait une valeur inférieure ou égale à D_i, avec i dans $\{1;2;3;4;5;6;7;8;9\}$.
Les centiles :
Le i^{eme} centile c_i est la plus petite valeur de la série telle qu'au moins $\left(i \times \frac{1}{100} \times 100\right)$ % des valeurs de la série ait une valeur inférieure ou égale à c_i, avec i dans $\{1;2;...;99;100\}$.

☞**Exercice 10**

Le tableau suivant donne le nombre de passagers (noté NbPas) de l'aéroport Limoges-Bellegarde suivant les années entre 1 997 et 2 018 :

Année	1 997	1 998	1 999	2 000	2 001	2 002	2 003	2 004	2 005	2 006	2 007
NbPas	120 208	130 042	139 668	133 165	132 504	158 566	187 491	223 841	283 849	376 558	391 220

Année	2 008	2 009	2 010	2 011	2 012	2 013	2 014	2 015	2 016	2 017	2 018
NbPas	382 398	356 353	339 047	334 816	306 837	299 654	290 792	292 607	291 564	309 641	301 493

1. Calculer la moyenne $\overline{x}$ de la série
2. Calculer l'étendue de la série.
3. Déterminer la médiane et les quartiles de la série.
4. Calculer l'écart-interquartile de la série. Interpréter par une phrase.

☞**Exercice 11**

Le tableau suivant donne le nombre de gares de France Métropole (hors Corse) par département en 2 018

Nombre de gares	5	7	12	14	15	16	17	18	21	22	23	24	25
Nombre de départements	1	2	4	1	1	1	1	1	4	1	3	5	2

Nombre de gares	26	27	29	30	31	34	35	36	39	40	41	42	44
Nombre de départements	2	4	3	1	4	2	3	1	2	2	1	1	3

Nombre de gares	48	49	50	52	53	54	55	59	61	62	64	70	74
Nombre de départements	2	4	2	2	2	3	1	1	2	1	1	1	1

Nombre de gares	75	77	79	80	82	87	91	93	102	127	183		Total
Nombre de départements	2	1	1	3	1	1	1	2	1	1	1		94

Source : Data.gouv.fr

1. Calculer la moyenne $\overline{x}$ de la série. Interpréter.
2. Calculer l'étendue de la série.
3. Déterminer la médiane M de la série. Interpréter.
4. Déterminer les quartiles Q_1 et Q_3 de la série. Calculer l'écart-interquartile et interpréter-le.

☞**Exercice 12**

Le tableau suivant donne le relevé des températures, exprimée en degré Celsius, à chaque heure, à Limoges, le 25 juillet 2 019.

Températures	24	25	26	27	29	32	33	34	35	36	37
Fréquences	4%	13%	16%	13%	8%	4%	4%	4%	13%	13%	8%
Fréquence cumulées croissantes											

Source : d'après infoclimat.fr

Lecture : 13% des relevés ont indiqué la température de 35° *C.*

1. Compléter le tableau par les fréquences cumulées croissantes.
2. Déterminer les quartiles Q_1, Q_2 (qu'on pourra assimiler à la médiane) et Q_3 de la série. Interpréter la médiane et le quartile Q_3.
3. Calculer l'écart-interquartile. Interpréter.

IV. B. 2. Écart-type

☞**Activité 1**

On s'intéresse au nombre d'essais par matches de deux équipes de Rugby.

nombre d'essais par match	0	1	2	3	4	5
nombre de matches pour l'équipe 1	4	6	2	1	2	1
nombre de matches pour l'équipe 2	7	2	1	3	2	1

1. Comparer les moyennes $\overline{x_1}$ et $\overline{x_2}$.
2. Comparer la moyenne X_1 et X_2 des écarts à la moyenne x_1 et x_2, par exemple pour l'équipe 1 on obtient le tableau suivant pour déterminer la moyenne X_1.

écart d'essai à la moyenne $\overline{x_1}$	$0-\overline{x_1}$	$1-\overline{x_1}$	$2-\overline{x_1}$	$3-\overline{x_1}$	$4-\overline{x_1}$	$5-\overline{x_1}$
nombre de matches pour l'équipe 1	4	6	2	1	2	1

   Que pouvez-vous dire ?

3. Comparer la moyenne V_1 et V_2 des carrés des écarts à la moyenne x_1 et x_2, par exemple pour l'équipe 1 on obtient le tableau suivant pur déterminer la moyenne X_1.

carré des écarts d'essais à la moyenne $\overline{x_1}$	$(0-\overline{x_1})^2$	$(1-\overline{x_1})^2$	$(2-\overline{x_1})^2$	$(3-\overline{x_1})^2$	$(4-\overline{x_1})^2$	$(5-\overline{x_1})^2$
nombre de matches pour l'équipe 1	4	6	2	1	2	1

   Que pouvez-vous dire ?

Avec les notations suivantes :

Variable x_i	x_1	x_2	...	x_p	Total
Effectif n_i	n_1	n_2	...	n_p	n
Fréquence f_i	f_1	f_2	...	f_p	1 (= 100%)

Définition

- La variance V_x d'une série statistique de moyenne $\overline{x}$ est la moyenne pondérée des carrés des écarts à la moyenne.

$$V_x = \frac{n_1(x_1-\overline{x})^2 + n_2(x_2-\overline{x})^2 + \ldots + n_p(x_p-\overline{x})^2}{n} = \frac{1}{n}\sum_{k=1}^{p}(x_k-\overline{x})n_k$$

- L'écart-type σ_x est le nombre $\sqrt{V_x}$.

Propriété

Avec les notations précédentes,

- $$V_x = \overline{x^2} - \overline{x}^2$$
- Soit un nombre réel a,

$$V_{ax} = a^2.V_x$$

☞**Démonstration 2**

✧✧✧ Laissée en exercice

☞**Exercice 13**

Le tableau suivant donne la taille des ménages français (hors Mayotte) en 1 982 et en 2 016.

Nombre d'occupants	Part des ménages 1 982	Part des ménages 2 016
1	24,4%	35,8%
2	28,4%	32,7%
3	18,7%	13,9%
4	16,2%	11,7%
5	7,4%	4,3%
6	4,9%	1,7%

Source : INSEE

Lecture : en 2 016, 13,9% des ménages français sont constitués de 3 personnes.

1. Calculer les moyennes $\overline{x_1}$ et $\overline{x_2}$ et les écarts-types σ_1 et σ_2 des deux séries statistiques.

2. Commenter par le critère de position, puis par le critère de dispersion.