Statistiques univariées

1 Introduction

Le terme statistiques provient de l'allemand *Statistik*, emprunté de l'italien *statista* (Homme d'état) ou bien encore du latin *status* qui signifie état. Cette notion renvoie donc aux affaires de l'état, et par extension à la collecte, l'étude, l'analyse, puis l'interprétation, de données pouvant intéresser l'Etat dans son fonctionnement (le nombre d'habitants, répartition par sexe, par âge, par métiers etc.). Elle joue ainsi un rôle de prévisions dans des domaines comme l'économie ou la démographie.

On retrouve des traces de statistiques en Chine plus de 2000 ans avant Jésus-Christ, près de 1700 ans avant J.C en Égypte, et à peine plus tard dans les civilisations précolombiennes. Le premier bureau de statistique a été créé en France en 1800 par Napoléon. Cet organisme a pris en 1946 le nom d'Institut National de la Statistique et des Études Économiques (INSEE).

Les domaines d'applications des statistiques sont, de nos jours, très variés :

- Économie, assurance, finance : Etude de marchés, analyse de consommation des ménages, taxation des primes d'assurances etc.
- Biologie, médecine : Essais thérapeutiques, épidémiologie, dynamique de population etc.
- Sciences de la Terre : Prévisions météo, exploration pétrolière ...
- Sciences humaines : Enquête d'opinion, sondages, étude de population...
- Sciences de l'ingénieur, de l'information Contrôle qualité, sûreté de fonctionnement, traitement des images, des signaux etc.

Elles interviennent dès lors que les données collectées présentent des incertitudes, des variations. Cela peut provenir aussi bien du fait que les phénomènes observés ne sont pas prévisibles à l'avance, ou bien dès lors qu'on fait une mesure, il y a une marge d'erreur, ou encore du fait qu'on ne peut observer qu'un petit nombre d'individus de la population. Ces données sont ainsi issues de phénomènes aléatoires, et c'est la raison pour laquelle on fait intervenir les calculs de probabilités.

On distingue deux types de statistiques :

- **Les statistiques descriptives** : Elles consistent à résumer l'information contenu dans un ensemble de données, de façon rapide et claire. Cela passe par des représentations graphiques, des indicateurs de position, de dispersion, pour exprimer des tendances.
- **Les statistiques inférentielles** : Elles ont pour but de faire des prévisions dans le but de prendre des décisions sur des situations données. On fait des tests, des intervalles de confiances etc.

Dans ce chapitre nous allons seulement aborder des notions de statistique descriptive.

2 Vocabulaire

Définition 1 : Les statistiques s'intéressent aux propriétés des objets d'un ensemble, appelés **individus** (même s'ils ne s'agit pas de personnes). L'ensemble de ces individus est appelée la **population**. L'ensemble de la population est souvent noté par la lettre Ω, et un individu, par ω.

En général, la totalité de la population n'est pas directement accessible. Ce peut être simplement car elle est trop grande, ou pour d'autres raisons. L'étude statistique se fait alors sur un **échantillon**, c'est-à-dire sur un sous-ensemble fini de la population.

La **taille** d'un échantillon est le nombre N d'éléments qu'il contient.

■ **Exemple 1 :** Le ministère des solidarités et de la santé cherche à mieux comprendre sa population : l'ensemble des français de plus de 18 ans. Il ne peut bien évidemment pas étudier la totalité des plus de 51 millions d'individus de celle-ci, et demande à l'INSEE de mener une enquète.

Celle-ci va donc interroger un échantillon de 10000 français.

Définition 2 : Les observations faites sur chaque individu sont appelés **caractère**, **ou variables statistiques** et sont en général notées X.

Une variable statistique est ainsi une application $X : \Omega \mapsto E$, où E est appelé le support de la variable, et est égal $X(\Omega)$.

- Si ces observations sont des réels, la variable est dite numérique ou **quantitative**.
- Dans le cas contraire, elle est dite **qualitative**.

Dans le cas des variables numériques, on distingue de plus...

- si $X(\Omega)$ est fini ou dénombrable, auquel cas la variable est dite **discrète**,
- ou non, auquel cas elle est dite **continue**.

■ **Exemple 2 :** Parmi les caractères qui intéressent le ministère et qui seront analysés dans l'étude de l'INSEE, on peut compter la taille des individus (une variable statistique quantitative et continue), leur âge (une variable quantitative discrète), et leur nom (une variable qualitative)

Définition 3 : Les valeurs prises par une variable quantitative discrète sont appelées **modalités**, souvent ordonnées dans l'ordre croissant $x_1 < x_2 < \cdots < x_p$.

Dans le cas d'une variable continue, on regroupe plutôt les valeurs en classes, des intervalles $[x_1, x_2[$,$[x_2, x_3[$, ..., $[x_p, x_{p+1}[$. C'est aussi possible dans le cas d'une variable discrète, bien sûr.

■ **Exemple 3 :** Les modalités de l'âge de l'étude réalisée sont $18 < 19 < ... < 150$ (il n'est pas nécessaire de s'assurer que toutes les modalités sont remplies).

Pour l'étude de la taille, on peut considérer les classes suivantes (en centimètres) : $[0, 150[$, $[150, 160[$, $[160, 170[$, $[170, 180[$, $[180, 190[$, $[190, +\infty[$.

Définition 4 : Le nombre d'individus d'un échantillon dont le caractère est dans une modalité ($X(\omega) = x_i$) ou une classe particulière ($X(\omega) \in [x_i, x_{i+1}[$) est appelé **effectif** de celle-ci. Il est désigné par la lettre n_i.

L'effectif cumulé N_i en une modalité ou une classe est la somme des effectifs des modalités ou classes inférieures ou égales.

$$N_i = \sum_{k=1}^{i} n_k.$$

La fréquence f_i d'une modalité ou d'une classe est le rapport de son effectif par la taille de l'échantillon :

$$f_i = \frac{n_i}{N}.$$

La fréquence cumulée F_i en une modalité ou une classe est la somme des fréquences des modalités ou classes inférieures ou égales.

$$F_i = \sum_{k=1}^{i} f_k.$$

Propriété 1 : On a alors les relations suivantes

$$\sum_{i=1}^{p} n_i = N \qquad ; \qquad N_p = N \qquad ; \qquad \sum_{i=1}^{p} f_i = 1 \qquad ; \qquad F_p = 1$$

■ **Exemple 4 :** Le tableau ci-dessous représente les effectifs relatifs à la taille des 10000 personnes interrogées par l'INSEE(d'après les données ENNS, 2006-2007)

Classe	$[0,150[$	$[150,160[$	$[160,170[$	$[170,180[$	$[180,190[$	$[190;+\infty[$
Effectif	150	1700	3800	3280	955	1150
Fréquence	0,015	0,17	0,38	0,328	0,0955	0,0115
Effectif cumulé	150	1850	5650	8930	9885	10000
Fréquence cumulée	0,015	0,185	0,565	0,893	0,9885	1

Définition 5 : Une **série statistique** $(x_i, n_i)_{i \in [\![1,p]\!]}$ ou $([x_i, x_{i+1}[, n_i)_{i \in [\![1,p]\!]}$ est la donnée des différentes modalités ou classes d'un échantillon, accompagnées de leurs effectifs (ou effectif cumulés). Elle est dite **groupée** si les données sont regroupées par classes.

■ **Exemple 5 :** Un concessionnaire d'automobiles neuves a enregistré au cours de ses 40 premières semaines d'opération, le nombre X d'automobiles qu'il a vendu hebdomadairement. Il a obtenu les résultats suivants :

5, 7, 2, 6, 3, 4, 8, 5, 4, 3, 9, 6, 5, 7, 6, 8, 3, 4, 4, 0, 8, 6, 7, 1, 5, 5, 4, 6, 6, 10, 9, 8, 1, 5, 5, 6, 7, 8, 5, 5

La série statistique est la donnée du tableau suivant

x_i	0	1	2	3	4	5	6	7	8	9	10
n_i	1	2	1	3	5	9	7	4	5	2	1

3 Indicateurs statistiques

3.1 Indicateurs de position

Mode

Définition 6 : Un mode (ou une classe modale) d'une série statistique est une modalité (ou une classe) pour laquelle l'effectif est maximale.

■ **Exemple 6 :** La classe modale pour la taille dans l'exemple précédent est $[160, 170[$.

■ **Exemple 7 :** Un mode n'est pas forcément unique.

Si l'on considère la série statistique suivante,

x_i	1	2	3	8	9	10
n_i	3	1	4	4	2	1

Les modes sont 3 et 8.

Moyenne

Définition 7 : On appelle moyenne d'une série statistique discrète $X = (x_i, n_i)_{i \in [\![1,p]\!]}$ la quantité notée $\overline{X}$ qui vaut

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} f_i x_i.$$

Dans le cas d'une série groupée $X = ([x_i, x_{i+1}[, n_i)$, la moyenne pourra être calculée en utilisant les centres des classes : on pose $c_i = \frac{x_i + x_{i+1}}{2}$. On a alors

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{p} n_i c_i = \sum_{i=1}^{p} f_i c_i.$$

■ **Exemple 8 :** Dans l'exemple de la taille précédent, la moyenne groupée (en prenant $c_1 = 145$ et $c_6 = 195$), on obtient $\overline{X} = 168,5$cm.

C'est un peu plus que la moyenne non groupée fournie par le document, $\overline{X} = 168,05$cm.

Médiane

Définition 8 : On appelle médiane d'une série statistique tout réel m_e partageant la série statistique en deux séries d'effectifs égaux.

Propriété 2 : Dans le cas d'une série discrète d'effectif total N

- Si N est impair, la médiane est la $\dfrac{N+1}{2}$-ième valeur de la série.
- Si N est pair, toute valeur située entre la $\dfrac{N}{2}$-ième valeur et la $\left(\dfrac{N}{2}+1\right)$-ième valeur est une médiane. On choisit traditionnellement la moyenne de ces deux valeurs.

Dans le cas d'une série continue

- On dresse le tableau des fréquences cumulées ;
- On repère la classe $[x_m, x_{m+1}[$ pour laquelle les fréquences cumulées atteignent ou dépasse 0,5 ;
- Les médianes appartiennent alors à cette classe. On choisit en général

$$M_e = x_m + (x_{m+1} - x_m)\frac{0,5 - F_{m-1}}{F_m - F_{m-1}}.$$

■ **Exemple 9 :** On considère la série statistique suivante

x_i	1	2	3	8	9	10
n_i	3	1	4	4	2	1
N_i	3	4	8	12	14	15

Ici, l'effectif total est $N = 15$ qui est impair. La médiane correspond donc à la 8ème valeur de la série, c'est-à-dire 3.

■ **Exemple 10 :** On considère la série statistique suivante

x_i	1	2	3	8	9	10
n_i	3	2	4	4	4	3
N_i	3	5	9	13	17	20

Ici, l'effectif total est $N = 20$ qui est pair. On regarde les 10-ème et 11-ème valeurs. Elles sont toutes deux égales à 8. Ainsi, la médiane de cette série est 8.

■ **Exemple 11 :** Reprenons les données relatives à la taille dans l'échantillon de 10000 personnes interrogées par l'INSEE.

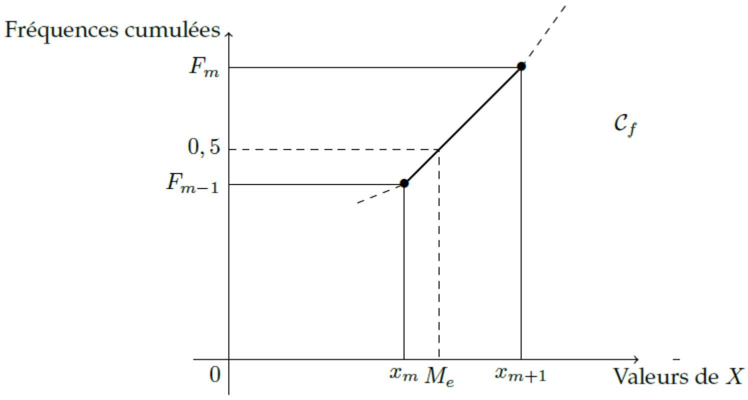
Classe	$[0, 150[$	$[150, 160[$	$[160, 170[$	$[170, 180[$	$[180, 190[$	$[190; +\infty[$
Effectif	150	1700	3800	3280	955	1150
Fréquence	0,015	0,17	0,38	0,328	0,0955	0,0115
Effectif cumulé	150	1850	5650	8930	9885	10000
Fréquence cumulée	0,015	0,185	0,565	0,893	0,9885	1

La classe qui correspond à une fréquence cumulée de 0,5 est la classe $[160, 170[$. La formule de calcul de la médiane donne alors

$$M_e = 160 + \frac{0,5 - 0,185}{0,565 - 0,185}(170 - 160) \simeq 168.$$

D'où provient cette formule ?

Il s'agit d'une méthode d'*interpolation linéaire* . On sait que la médiane appartient à la classe $[x_m; x_{m+1}[$ et on fait l'hypothèse que les valeurs dans la classe sont uniformément répartis.



Cela se traduit par le fait qu'entre les points (x_m, F_{m-1}) et (x_{m+1}, F_m), l'évolution est linéaire.

Le coefficient directeur de la droite reliant ces deux points est alors égal à $\dfrac{F_m - F_{m-1}}{x_{m+1} - x_m}$. L'équation de la droite est alors de la forme

$$y = \left(\frac{F_m - F_{m-1}}{x_{m+1} - x_m}\right) x + b.$$

Puisque cette droite passe par le point de coordonnées (x_m, F_{m-1}), on a donc

$$b = F_{m-1} - \left(\frac{F_m - F_{m-1}}{x_{m+1} - x_m}\right) x_m.$$

On trouve donc que l'équation de notre droite est

$$y = \left(\frac{F_m - F_{m-1}}{x_{m+1} - x_m}\right) x + F_{m-1} - \left(\frac{F_m - F_{m-1}}{x_{m+1} - x_m}\right) x_m = \left(\frac{F_m - F_{m-1}}{x_{m+1} - x_m}\right)(x - x_m) + F_{m-1}.$$

Pour trouver la médiane de la série statistique, on cherche alors la valeur de x pour laquelle on a $y = 0,5$. On a alors

$$\begin{aligned}
\left(\frac{F_m - F_{m-1}}{x_{m+1} - x_m}\right)(M_e - x_m) + F_{m-1} = 0,5 \quad &\Leftrightarrow \quad \left(\frac{F_m - F_{m-1}}{x_{m+1} - x_m}\right)(M_e - x_m) = 0,5 - F_{m-1} \\
&\Leftrightarrow \quad M_e - x_m = (0,5 - F_{m-1}) \times \frac{x_{m+1} - x_m}{F_m - F_{m-1}} \\
&\Leftrightarrow \quad M_e = x_m + (x_{m+1} - x_m) \times \frac{0,5 - F_{m-1}}{F_m - F_{m-1}}
\end{aligned}$$

■ **Exemple 12 :** Il est important de ne pas confondre la moyenne et la médiane !

En effet selon l'INSEE, le salaire mensuel net moyen des français est de 2733 euros alors que le salaire médian est de 2190 euros nets.

La moyenne est en effet sensible aux valeurs extrêmes : les très hauts salaires vont faire augmenter la moyenne, mais pas la médiane s'ils sont en nombre relativement faible.

Illustration simple : si un jour, Bill Gates vient au lycée, alors en moyenne, toutes les personnes dans le lycée seront millionaires.

Quartile et déciles

Définition 9 — Quartiles : On appelle premier quartile, et on le note q_1 ou Q_1, la valeur de la série telle que au moins 25% des valeurs de l'effectif total lui sont inférieures ou égales.

On appelle troisième quartile, et on le note q_3 ou Q_3, la valeur de la série telle que au moins 75% des valeurs de l'effectif total lui sont inférieures ou égales.

Définition 10 — Déciles : Soit $k \in [\![1,9]\!]$. On appelle k-ième décile, et on le note d_k ou D_k, la valeur de la série telle que au moins $k \times 10\%$ des valeurs de l'effectif total lui sont inférieures ou égales.

Le calcul des quartiles et des déciles se fait de manière analogue à celui de la médiane.

■ **Exemple 13 :** On considère la série statistique suivante

x_i	1	2	3	8	9	10
n_i	3	1	4	4	2	1
N_i	3	4	8	12	14	15

L'effectif total vaut $N = 15$. On a $\frac{N}{10} = 1,5$, $\frac{N}{4} = 3,25$ et $\frac{3N}{4} = 9,75$.

- Le premier décile est la 2ème valeur, c'est-à-dire $d_1 = 1$.
- Le premier quartile est la 4ème valeur, c'est-à-dire $Q_1 = 2$.
- Le troisième quartile est la 10-ième valeur, soit $Q_3 = 8$.

3.2 Indicateurs de dispersion

Considérons deux élèves A et B qui ont obtenue en mathématiques les notes suivantes :

- 7, 8, 11, 12, 13, 13, 13 pour A
- 4, 7, 9, 12, 13, 13, 19 pour B

Il est facile de vérifier que les séries de notes de A et B ont la même médiane (12), la même moyenne (11) et le même mode (13). Pourtant, ces deux séries statistiques sont bien différentes car un simple coup d'œil permet de voir que les notes de B sont plus dispersées.

Les paramètres de positions ne peuvent donc pas, à eux seuls, donner un aperçu de la répartition des valeurs. Pour être plus précis dans notre étude, on introduit plusieurs outils qui permettent de rendre compte de la dispersion, et ainsi décrire plus précisément la population.

Étendue

Définition 11 : On appelle **étendue** d'une série statistique l'écart entre la plus grande et la plus petite valeur.

■ **Exemple 14 :** Dans le cas des séries de notes, les étendues sont $e_A = 13 - 7 = 6$ et $e_B = 19 - 4 = 15$.

On en conclut que les notes de B sont plus étendues que celles de A.

Écart inter-quartile

Définition 12 : On appelle **écart inter-quartile** la différence $Q_3 - Q_1$ d'une série statistique.

L'intervalle [Q1;Q3] contient environ 50% des valeurs, dont la médiane. Plus l'écart interquartile est élevé, plus les valeurs peuvent dispersées.

Variance et écart-type

Définition 13 : Si X est discrète, la variance de la série $(x_i, n_i)_{i\in[\![1,p]\!]}$ est la quantité

$$V(x) = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \overline{X})^2 = \sum_{i=1}^{p} f_i(x_i - \overline{X})^2.$$

Si X est continue, la variance de la série $([x_i, x_{i+1}[, n_i)_{i\in[\![1,p]\!]}$ est la quantité

$$V(x) = \frac{1}{N}\sum_{i=1}^{p} n_i(c_i - \overline{X})^2 = \sum_{i=1}^{p} f_i(c_i - \overline{X})^2.$$

où $c_i = \frac{x_i + x_{i+1}}{2}$ est le centre de la classe $[x_i, x_{i+1}[$.

La variance représente donc la moyenne des carrés des écarts à la moyenne.

Propriété 3 — Formule de Koenig-Huygens : Pour une variable statistique, on a

$$V(X) = \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - \overline{X}^2 = \overline{X^2} - \overline{X}^2.$$

Démonstration 1 : Pour tout $i \in [\![1,p]\!]$, on a

$$(x_i - \overline{X})^2 = x_i^2 - 2x_i\overline{X} + \overline{X}^2.$$

Ainsi,

$$\sum_{i=1}^{p} n_i(x_i - \overline{X})^2 = \sum_{i=1}^{p} n_i(x_i^2 - 2x_i\overline{X} + \overline{X}^2) = \sum_{i=1}^{p} n_i x_i^2 - 2\sum_{i=1}^{p} n_i x_i \overline{X} + \sum_{i=1}^{p} n_i \overline{X}^2.$$

Puisque $\overline{X}$ ne dépend pas de l'indice i, on a alors

$$\sum_{i=1}^{p} n_i(x_i - \overline{X})^2 = \sum_{i=1}^{p} n_i x_i^2 - \overline{X} \times 2\sum_{i=1}^{p} n_i x_i + \overline{X}^2 \sum_{i=1}^{p} n_i.$$

Or, $\sum_{i=1}^{p} n_i = N$ et $\sum_{i=1}^{p} n_i x_i = N\overline{X}$. Ainsi,

$$\sum_{i=1}^{p} n_i(x_i - \overline{X})^2 = \sum_{i=1}^{p} n_i x_i^2 - 2N\overline{X}^2 + N\overline{X}^2 = \sum_{i=1}^{p} n_i x_i^2 - N\overline{X}^2.$$

On obtient la formule voulue en divisant cette égalité par N. □

Définition 14 : L'écart-type d'une série statistique est la racine carrée de sa variance.

$$\sigma(X) = \sqrt{V(X)}.$$

L'écart-type est la caractéristique de dispersion la plus souvent utilisée. On peut montrer que sous certaines conditions

- l'intervalle $[\overline{X} - \sigma(X), \overline{X} + \sigma(X)]$ contient environ 68% des valeurs de la série ;
- l'intervalle $[\overline{X} - 2\sigma(X), \overline{X} + 2\sigma(X)]$ contient environ 95% des valeurs de la série.

3.3 Transformation affine

Propriété 4 : Soit a un réel non nul, b un réel et X une série statistique. Alors

$$\overline{aX+b} = a\overline{X}+b \quad ; \quad V(aX+b) = a^2V(X) \quad ; \quad \sigma(aX+b) = |a|\sigma(X).$$

Par ailleurs, si M_e est une médiane de X, alors aM_e+b est une médiane de $aX+b$.

■ **Exemple 15 :** Si on augmente tous les salaires de tous les Français de 100 euros, alors le salaire moyen augmentera de 100 euros. La variance restera toutefois inchangé et les salaires seront tout aussi dispersé.

Si, en revanche, on augmente les salaires de tout le monde par 1,1 (soit une augmentation généralisée de 10%), alors le salaire moyen sera multiplié par 1,1, mais ce sera aussi le cas de l'écart-type des salaires. La variance sera quant à elle multipliée par $1,1^2$ soit $1,21$.

4 Représentations graphiques

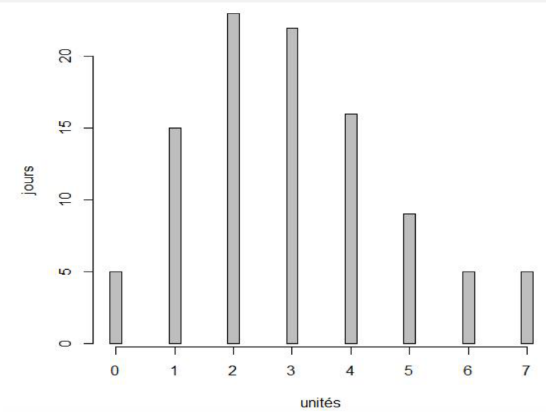
4.1 Diagramme en bâtons

C'est une représentation graphique particulièrement indiquée pour la représentation de variable discrètes.
La série $(x_i,n_i)_{i\in[\![1,p]\!]}$ se représente ainsi dans un repère orthogonal. On indique en abscisses les valeurs x_i et les n_i (ou les f_i) en ordonnées. Pour chacune des valeurs, on trace un bâton de hauteur n_i.

■ **Exemple 16 :** Le responsable des ventes d'un magasin note, au cours de 100 jours ouvrables, la demande journalière pour un de ses produit :

x_1	0	1	2	3	4	5	6	7
n_i	5	15	23	22	16	9	5	5

Le diagramme en bâtons correspondant est le suivant :



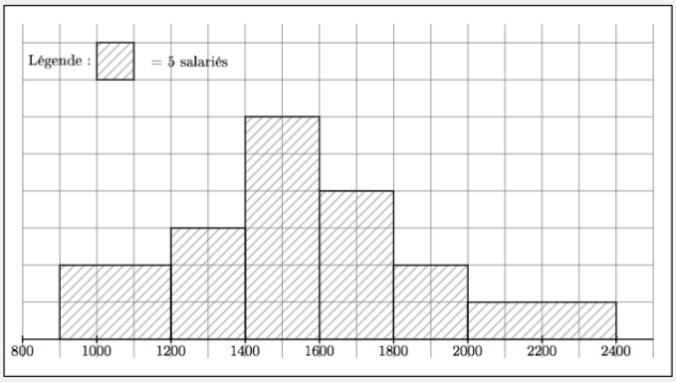
4.2 Histogramme

Lorsque le caractère étudié est quantitatif continu, et lorsque les données sont regroupées en classes, on peut représenter la série par un histogramme : l'aire de chaque rectangle est alors proportionnelle à l'effectif (ou a la fréquence) associée à chaque classe.

■ **Exemple 17 :** Une entreprise regroupe ses employés en différentes classes, en fonction de leur salaire.

Salaires	$[900;1200[$	$[1200;1400[$	$[1400;1600[$	$[1600;1800[$	$[1800;2000[$	$[2000;2400[$
Effectif	30	30	60	40	20	20

Voici un histogramme représentant la répartition des salaires dans cette entreprise.



4.3 Courbe des fréquences cumulées

Définition 15 : Soit X un série statistique quantitative.

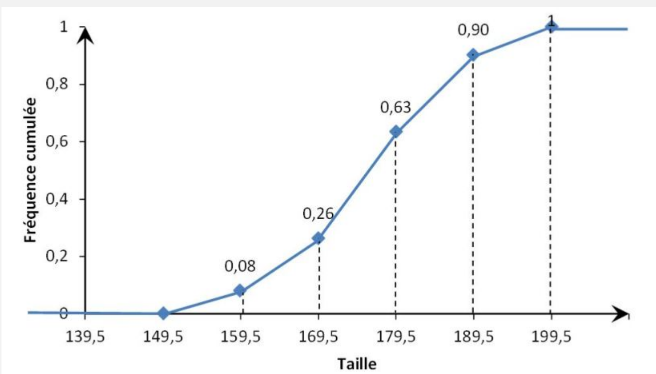
La fonction de répartition de X, notée F_X, est la fonction qui à tout réel t associe la proportion des observations inférieures ou égales à t.

La courbe des fréquences cumulées (ou diagramme des fréquences cumulées) est une approximation de la courbe représentative de cette fonction de répartition pour une série groupée. Elle s'obtient en plaçant dans un repère les points de coordonnées (x_i, F_{i-1}) pour i allant de 1 à p puis à les relier par des segments.

■ **Exemple 18 :** On considère un groupe de 175 étudiants, classés selon leurs tailles comme suit

Classes	$[149,5;159,5]$	$[159,5;169,5]$	$[169,5;179,5]$	$[179,5;189,5]$	$[189,5;199,5]$
Effectif	14	32	65	47	17
Effectif cumulé	14	46	111	158	175
Fréquence	0,08	0,18	0,37	0,27	0,10
Fréquence cumulée	0,08	0,26	0,63	0,90	1

La courbe des fréquences cumulées est représentée ci-dessous.



Remarque : l'antécédent de 0,5 donne une approximation de la médiane de la série.

4.4 Diagramme en boîtes

Le diagramme en boîte, (aussi appelé *boîte de Tukey* ou encore *boîte à moustaches*) est un diagramme qui regroupe les valeurs extrêmes, les quartiles Q_1 et Q_3 , ainsi que la médiane M_e.