

# PROBABILITÉS ET STATISTIQUE

## 2. Résumé et représentation graphique des données

Carmelo Vaccaro

Université de Paris-Saclay

M1 - Informatique Science des Données (ISD) - Apprentissage  
2022/23: premier semestre

# Introduction au chapitre

Les échantillons de données sont souvent grands et pour les analyser, nous devons organiser, résumer et représenter les données sous une forme pratique et significative. Nous organisons et résumons souvent les données sous forme numérique dans des tableaux ou sous forme visuelle dans des graphiques, comme décrit dans ce chapitre.

La représentation que nous choisissons dépend du type de données que nous étudions.

Notre objectif n'est pas seulement d'obtenir un tableau ou un graphique, mais aussi d'analyser les données et de comprendre ce qu'elles nous disent.

# Caractéristiques des données

- ❶ **Centre** : Une valeur représentative ou moyenne qui indique où se trouve le milieu de l'ensemble des données.
- ❷ **Variation** : Une mesure du montant de la variation des valeurs des données.
- ❸ **Distribution** : La nature ou la forme de l'étalement des données sur la plage de valeurs (en forme de cloche, uniforme ou asymétrique).
- ❹ **Valeurs aberrantes** : Les valeurs de l'échantillon qui sont très éloignées de la grande majorité des autres valeurs de l'échantillon.
- ❺ **Temps** : évolution des caractéristiques des données dans le temps.

- 1 Distributions de fréquences
- 2 Histogrammes
- 3 Graphiques statistiques

# 1. Distributions de fréquences

Lorsque l'on travaille avec de grands ensembles de données, il est souvent utile d'organiser et de résumer les données en construisant un tableau appelé *distribution de fréquences*, qui nous aide à comprendre la nature de la distribution d'un ensemble de données.

## **Distribution de fréquences (ou tableau de fréquences) :**

montre comment un ensemble de données est réparti entre plusieurs catégories (ou classes) en énumérant toutes les catégories ainsi que le nombre de valeurs de données dans chacune des catégories.

# Tableau des fréquences

Supposons que j'ai la liste de nombres suivants et que je veux déterminer les fréquences :

8 9 8 9 7 8 9 9 8 9 7 10 9 8 6 7 8 8 9 6

Alors je compte combien de fois chaque nombre apparait et je reporte la fréquence de chaque élément.

Résultat	Fréquence
6	2
7	3
8	7
9	7
10	1

Ce tableau est appelé un **tableau des fréquences**.



# Autres données

Maintenant supposons que j'ai ces données, qui représentent des taux de pulsation des femmes et des hommes.

**Table 2-1 Pulse Rates (beats per minute) of Females and Males**

Females																			
76	72	88	60	72	68	80	64	68	68	80	76	68	72	96	72	68	72	64	80
64	80	76	76	76	80	104	88	60	76	72	72	88	80	60	72	88	88	124	64
Males																			
68	64	88	72	64	72	60	88	76	60	96	72	56	64	60	64	84	76	84	88
72	56	68	64	60	68	60	60	56	84	72	84	88	56	64	56	56	60	64	72

Calculer la fréquence de chaque valeur prendrait trop de temps et de plus cela ne donnerait pas des informations intéressantes comme : combien de valeurs y-a-t-il entre 80 et 89 ? (pour cela je devrais additionner les fréquences des nombres compris entre 80 et 89).

# Subdiviser les données en classes

Je subdivise les données en classes, par exemple en dizaines, et je calcule la fréquence de chaque classe et pas de chaque valeur.

La **fréquence** d'une classe particulière est le nombre de valeurs qui entrent dans cette classe.

**Table 2-2 Pulse Rates of Females**

Pulse Rate	Frequency
60-69	12
70-79	14
80-89	11
90-99	1
100-109	1
110-119	0
120-129	1

- Limite inférieure de classes.
- Limite supérieure de classes.
- Frontière de classe.
- Point médian de classe.
- Largeur de classe

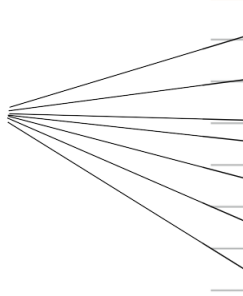
# Limite inférieure de classe

La limite inférieure d'une classe est le plus petit nombre qui appartient à la classe.

**Table 2-2 Pulse Rates of Females**

Pulse Rate	Frequency
60-69	12
70-79	14
80-89	11
90-99	1
100-109	1
110-119	0
120-129	1

**Lower Class Limits**



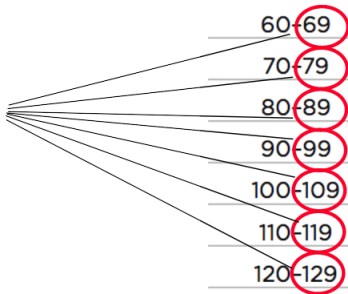
# Limite supérieure de classe

La limite supérieure d'une classe est le plus grand nombre qui appartient à la classe.

**Table 2-2 Pulse Rates of Females**

Pulse Rate	Frequency
60-69	12
70-79	14
80-89	11
90-99	1
100-109	1
110-119	0
120-129	1

**Upper Class Limits**



# Frontières de classe

Les frontières de classe sont les nombres utilisés pour séparer les classes. Normalement, il s'agit des points intermédiaires entre la limite supérieure d'une classe et la limite inférieure de la classe suivante.

**Class  
Boundaries**

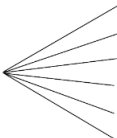


**Table 2-2 Pulse Rates of Females**

	Pulse Rate	Frequency
59.5	60-69	12
69.5	70-79	14
79.5	80-89	11
89.5	90-99	1
99.5	100-109	1
109.5	110-119	0
119.5	120-129	1
129.5		

# Point médian de classe

Le point médian d'une classe est la valeur située au milieu d'une classe et se calcule en ajoutant la limite inférieure de la classe à la limite supérieure de la classe et en divisant par deux.

**Class Midpoints** 

**Table 2-2 Pulse Rates of Females**

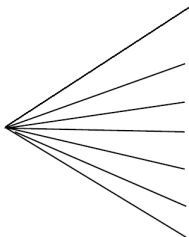
	Pulse Rate	Frequency
64.5	60-69	12
74.5	70-79	14
84.5	80-89	11
94.5	90-99	1
104.5	100-109	1
114.5	110-119	0
124.5	120-129	1

# Largeur de classe

La largeur d'une classe est la différence entre la limite inférieure de la classe suivante et celle de la classe même.

Pour la dernière classe la largeur peut être définie comme la différence entre la dernière et l'avant dernière frontière.

**Class  
Width**



10  
10  
10  
10  
10  
10  
10

**Table 2-2 Pulse Rates of Females**

Pulse Rate	Frequency
60-69	12
70-79	14
80-89	11
90-99	1
100-109	1
110-119	0
120-129	1



Pour la largeur de classe il faut éviter l'erreur facile qui consiste à faire de la largeur de classe la différence entre la limite inférieure de la classe et la limite supérieure de la classe.

# Exercice

Pour la table de fréquences ci-dessus, identifiez les limites inférieures des classes, les limites supérieures des classes, les frontières des classes, les points médians des classes et les largeurs des classes.

Nombre de citations	Fréquence
1-10	849
11-20	25
21-30	8
31-40	2
41-50	2
51-60	1
61-70	1
71-80	1
81-90	1

# Solution

Limites inférieures des classes : 1 11 21 31 41 51 61 71 81.

Limites supérieures des classes : 10 20 30 40 50 60 70 80 90.

Frontières des classes : 0,5 10,5 20,5 30,5 40,5 50,5 60,5 70,5 80,5

Points médians des classes : 5,5 15,5 25,5 35,5 45,5 55,5 65,5 75,5 85,5

Largeurs des classes : 10 pour toutes les classes.

Nombre de citations	Fréquence
1-10	849
11-20	25
21-30	8
31-40	2
41-50	2
51-60	1
61-70	1
71-80	1
81-90	1

# Raisons pour construire la distribution des fréquences

- ❶ Des grands ensembles de données peuvent être **résumés**.
- ❷ On peut **analyser** la nature des données.
- ❸ Nous disposons d'une base pour construire des **graphiques**.

# Tableau de fréquence pour des données qualitatives

Si les données sont qualitatives il faut déterminer un critère pour créer des classes.

Par exemple si on a une liste de noms d'étudiants on peut créer une classe contenant tous ceux dont le nom de famille est compris entre A et M et une classe pour le nom de famille entre N et Z.

Pour les données qualitatives les notions de limites inférieure et supérieure, frontière et point médian d'une classe n'ont pas de sens.

Si le nombre des valeurs différentes est petit on crée une classe par valeur.

# Exemple

Construire un tableau de fréquence pour les données qualitatives suivantes :

CBS Fox ABC Fox CBS CBS ABC Fox CBS ABC CBS CBS NBC CBS  
NBC NBC CBS CBS NBC NBC

Catégorie	Fréquence
ABC	3
CBS	9
Fox	3
NBC	5

# Construction d'un tableau de fréquences 1/2

Supposons d'avoir une liste de données numériques dont nous devons construire le tableau de fréquences. Supposons aussi d'avoir une valeur donnée comme point de départ et la largeur des classes. Alors on procède ainsi :

- 1 En utilisant le point de départ comme première limite inférieure de la première classe, on obtient les limites inférieures des autres classes en ajoutant la largeur de la classe à la limite inférieure de la classe précédente. Lister les limites inférieures de classe ainsi obtenus dans une colonne verticale.
- 2 La limite supérieure de la première classe est la valeur juste avant la limite inférieure de la deuxième classe. Les limites supérieures des autres classes sont obtenus en ajoutant la largeur de la classe à la limite supérieure de la classe précédente.
- 3 Comptez le nombre de valeurs de l'ensemble de données de chaque classe.

## Construction d'un tableau de fréquences 2/2

Supposons maintenant d'avoir toujours une liste de données numériques dont on doit construire le tableau de fréquences, mais que cette fois on nous donne le nombre de classes au lieu du point de départ et de la largeur des classes.

Alors on choisit comme point de départ soit la valeur minimale des données soit une valeur convenable en dessous de celle-ci. On choisit aussi comme point finale soit la valeur maximale des données soit une valeur convenable au dessus de celle-ci.

On définit la largeur des classes comme

$$\text{largeur des classes} = (\text{valeur maximale des données} - \text{valeur minimale des données}) / \text{nombre de classes}.$$

Maintenant qu'on a le point de départ et la largeur de classe on procède comme expliqué dans la slide précédente.



## Exemple : construction d'un tableau de fréquences

Construisons un tableau de fréquences pour les données numériques suivantes en 5 classes :

155 142 149 130 151 163 151 142 156 133

138 161 128 144 172 137 151 166 147 163

On peut prendre comme point de départ la valeur 125, comme point finale 175. On a  $(175 - 125)/5 = 10$ , donc les classes sont 125-134, 135-144, 145-154, 155-164, 165-174.

Construisez une distribution de fréquence pour l'ensemble de données suivant

48, 193, 3, 88, 24, 106, 79, 178, 163, 119, 148, 136, 175, 111, 35

Commencez avec une limite inférieure de classe de 0 et utilisez une largeur de classe de 50.

Valeurs	Fréquence
0-49	4
50-99	2
100-149	5
150-199	4

# Distribution des fréquences relatives

La distribution des fréquences relatives comprend les mêmes limites de classe qu'une distribution de fréquence, mais la fréquence d'une classe est remplacée par une fréquence relative (une proportion) ou une fréquence en pourcentage.

fréquences relative = fréquence de classe / somme de toutes les fréquences

La fréquence en pourcentage est la même que la fréquences relative mais exprimée en pourcentages.

La somme des fréquences relatives dans une distribution de fréquences relatives doit être égale à 1 (ou 100%) à moins d'approximations par arrondissement.

# Distribution des fréquences relatives

**Table 2-2** Pulse Rates of Females

Pulse Rate	Frequency
60-69	12
70-79	14
80-89	11
90-99	1
100-109	1
110-119	0
120-129	1

**Total Frequency = 40**

**Table 2-3** Relative Frequency Distribution of Pulse Rates of Females

Pulse Rate	Relative Frequency
60-69	30%
70-79	35%
80-89	27.5%
90-99	2.5%
100-109	2.5%
110-119	0
120-129	2.5%

**\*  $12/40 \times 100 = 30\%$**

Construisez la distribution de fréquences relatives correspondante à la distribution de fréquences suivante.

Valeurs	Fréquence
0-49	4
50-99	2
100-149	5
150-199	4

Construisez la distribution de fréquences relatives correspondante à la distribution de fréquences suivante.

Valeurs	Fréquence	Fréquence relative
0-49	4	27%
50-99	2	13%
100-149	5	33%
150-199	4	27%

# Distribution des fréquences cumulées

La **fréquence cumulée** d'une classe est la somme des fréquences de cette classe et de toutes les classes précédentes.

En utilisant les fréquences initiales de 12, 14, 11, 1, 1, 0 et 1, nous ajoutons  $12 + 14$  pour obtenir la deuxième fréquence cumulée de 26, puis nous ajoutons  $12 + 14 + 11$  pour obtenir la troisième, et ainsi de suite.

Notez que les limites de classe sont remplacées par des expressions “moins que” qui décrivent les nouvelles plages de valeurs.



# Distribution des fréquences cumulées

**Table 2-2** Pulse Rates of Females

Pulse Rate	Frequency
60-69	12
70-79	14
80-89	11
90-99	1
100-109	1
110-119	0
120-129	1

**Table 2-4** Cumulative Frequency Distribution of Pulse Rates of Females

Pulse Rate	Cumulative Frequency
Less than 70	12
Less than 80	26
Less than 90	37
Less than 100	38
Less than 110	39
Less than 120	39
Less than 130	40

Cumulative Frequencies

Construisez la distribution de fréquences cumulées correspondante à la distribution de fréquences suivante.

Valeurs	Fréquence
0-49	4
50-99	2
100-149	5
150-199	4

# Solution

Construisez la distribution de fréquences cumulées correspondante à la distribution de fréquences suivante.

Valeurs	Fréquence
0-49	4
50-99	2
100-149	5
150-199	4

Classe	Fréquences cumulées
Moins que 50	4
Moins que 100	6
Moins que 150	11
Moins que 200	15

**Table 2-2** Pulse Rates of Females

Pulse Rate	Frequency
60-69	12
70-79	14
80-89	11
90-99	1
100-109	1
110-119	0
120-129	1

**Table 2-3** Relative Frequency Distribution of Pulse Rates of Females

Pulse Rate	Relative Frequency
60-69	30%
70-79	35%
80-89	27.5%
90-99	2.5%
100-109	2.5%
110-119	0
120-129	2.5%

**Table 2-4** Cumulative Frequency Distribution of Pulse Rates of Females

Pulse Rate	Cumulative Frequency
Less than 70	12
Less than 80	26
Less than 90	37
Less than 100	38
Less than 110	39
Less than 120	39
Less than 130	40

Pour les données qualitatives on peut construire des tableaux de fréquences relatives mais pas de fréquences cumulées, à moins de donner un critère pour ordonner les différentes catégories.

## Exemple.

Catégorie	Fréquence	Fréquence relative
ABC	3	15%
CBS	9	45%
Fox	3	15%
NBC	5	25%

Dans cette section, nous avons abordé :

- Les caractéristiques importantes des données
- Les distributions de fréquences
- Procédures de construction des distributions de fréquences
- Distributions de fréquences relatives
- Distributions de fréquences cumulées

## 2. Histogrammes

Nous utilisons un outil visuel appelé **histogramme** pour analyser la forme de la distribution des données.



# Histogramme (1/3)

Un **histogramme** est un graphique constitué de barres de même largeur tracées les unes à côté des autres (sans espace).

L'échelle horizontale représente les classes de valeurs des données quantitatives et l'échelle verticale représente les fréquences.

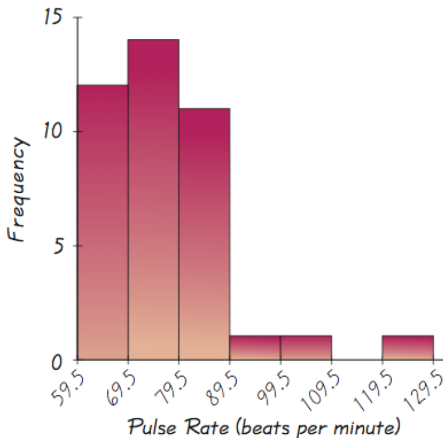
Les hauteurs des barres correspondent aux valeurs de fréquence.

# Histogramme (2/3)

Un histogramme est essentiellement une version graphique d'une distribution de fréquences.

**Table 2-2 Pulse Rates of Females**

Pulse Rate	Frequency
60-69	12
70-79	14
80-89	11
90-99	1
100-109	1
110-119	0
120-129	1



Les barres de l'échelle horizontale sont étiquetées par l'une des informations suivantes :

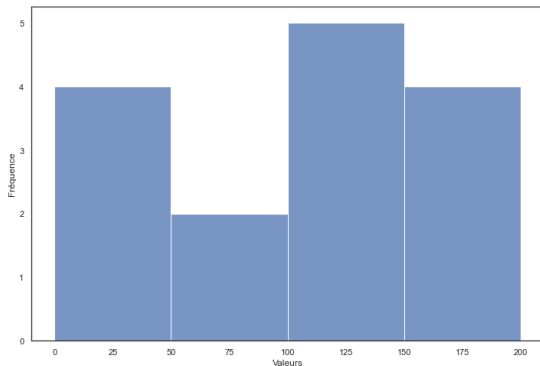
- 1 frontières des classes,
- 2 points médians des classes,
- 3 limites inférieures des classes.

Dessinez l'histogramme pour le tableau des fréquences suivant.

Valeurs	Fréquence
0-49	4
50-99	2
100-149	5
150-199	4

Dessinez l'histogramme pour le tableau des fréquences suivant.

Valeurs	Fréquence
0-49	4
50-99	2
100-149	5
150-199	4

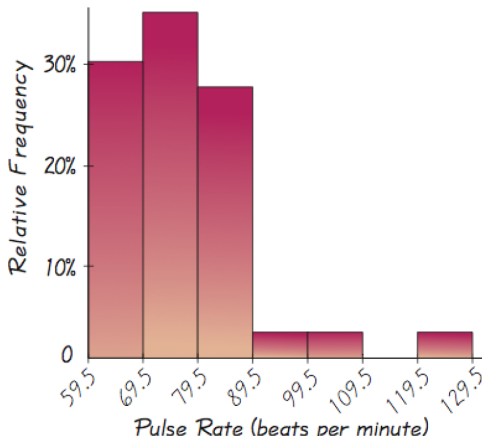


# Histogramme de fréquences relatives

Un **histogramme de fréquences relatives** a la même forme et la même échelle horizontale qu'un histogramme, mais l'échelle verticale est marquée par des fréquences relatives au lieu de fréquences réelles.

**Table 2-3** Relative Frequency Distribution of Pulse Rates of Females

Pulse Rate	Relative Frequency
60-69	30%
70-79	35%
80-89	27.5%
90-99	2.5%
100-109	2.5%
110-119	0
120-129	2.5%



### 3. Graphiques statistiques

Cette section aborde d'autres types de graphiques statistiques.

Notre objectif est d'**identifier** un graphique approprié pour **représenter** l'ensemble des données. Le graphique doit être efficace pour **révéler** les caractéristiques importantes des données.



# Types de diagrammes

Pour les données quantitatives :

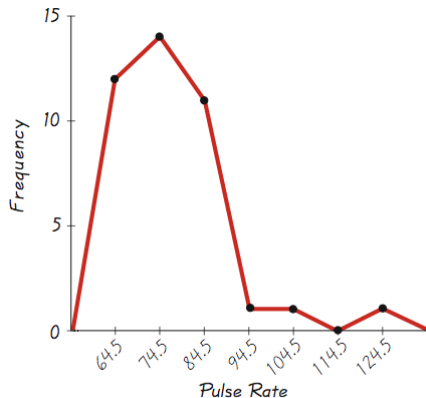
- histogrammes ;
- polygones de fréquences ;
- ogives ;
- nuages de points (pour les couples de données appariées) ;
- séries chronologiques (pour des données chronologiques).

Pour les données qualitatives :

- diagrammes de Pareto ;
- diagrammes circulaires.

# Polygone de fréquences

Un *polygone de fréquences* utilise des segments de ligne reliés à des points situés directement au-dessus des valeurs du point médian de la classe.



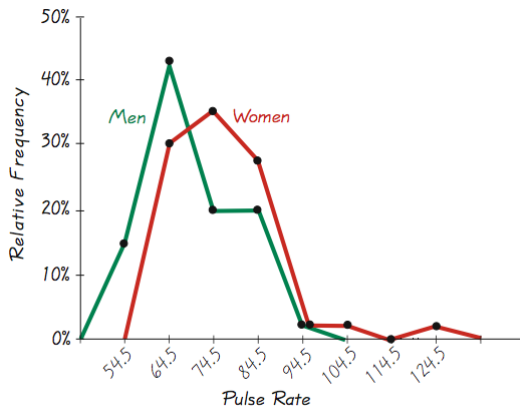
**Figure 2-5** Frequency Polygon: Pulse Rates of Women

# Polygone de fréquences relatives (1/2)

Un **polygone de fréquences relatives** utilise des fréquences relatives (proportions ou pourcentages) pour l'échelle verticale.

On peut superposer deux polygones de fréquences relatives pour comparer deux ensembles de données (ce qui avec les histogrammes pourrait donner un graphique confus et pas facilement lisible).

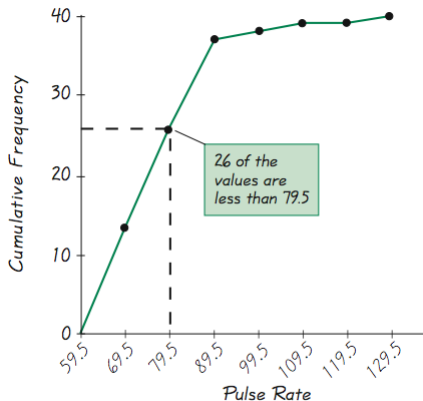
# Polygone de fréquences relatives (2/2)



**Figure 2-6** Relative Frequency Polygons: Pulse Rates of Women and Men

# Ogive

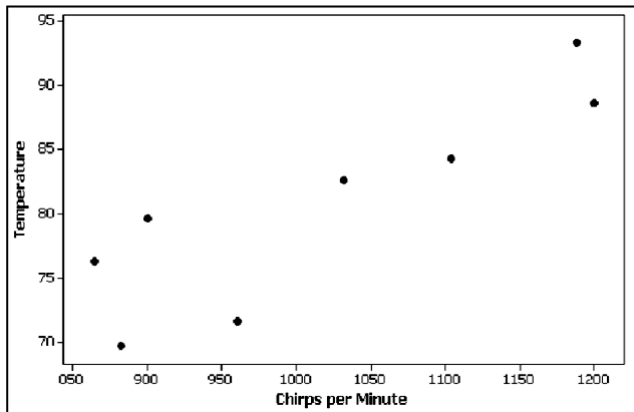
Une **ogive** est un graphique qui représente les fréquences cumulées par une ligne (équivalent du polygone de fréquences pour les fréquences cumulées).



**Figure 2-7 Ogive**

# Nuage de points

Un **nuage de points** est un diagramme de données appariées  $(x, y)$  avec un axe  $x$  horizontal et un axe  $y$  vertical. Utilisé pour déterminer s'il existe une relation entre les deux variables.



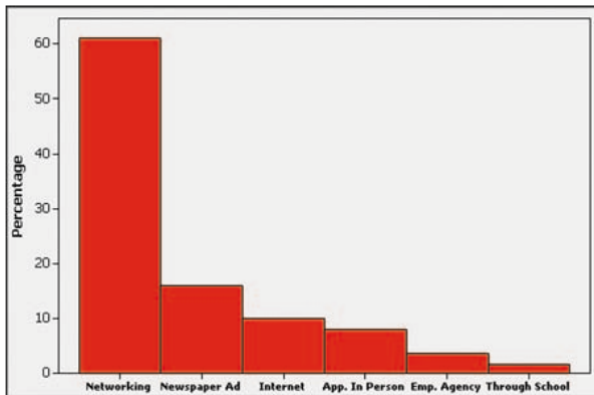
# Séries chronologiques

Un **graphique de série chronologique** est un graphique de données quantitatives recueillies à différents moments dans le temps.



# Diagramme de Pareto

Un **diagramme de Pareto** est un graphique à barres où les barres étant disposées en ordre décroissant selon les fréquences.





# Exercice

Dessinez le diagramme de Pareto pour la table de fréquences ci-dessous.

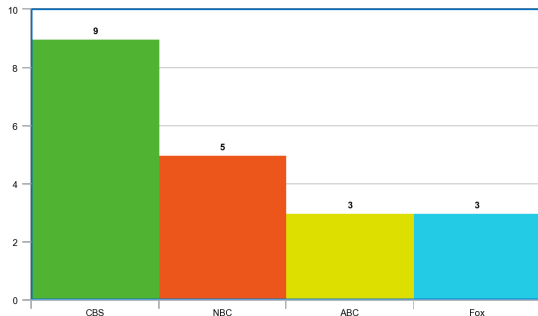
Catégorie	Fréquence
ABC	3
CBS	9
Fox	3
NBC	5

**Remarque.** Dans un diagramme de Pareto les barres sont disposées en ordre décroissant selon les fréquences.

# Solution

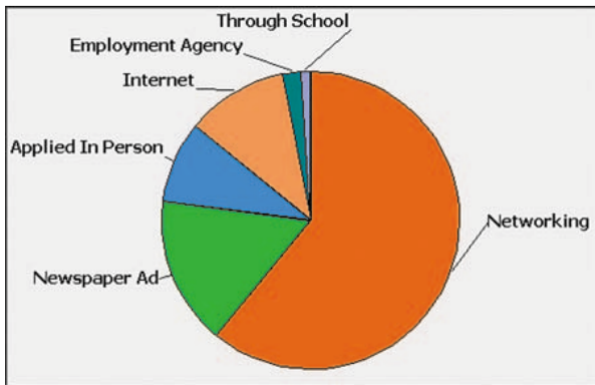
Dessinez le diagramme de Pareto pour la table de fréquences ci-dessous.

Catégorie	Fréquence
ABC	3
CBS	9
Fox	3
NBC	5



# Diagramme circulaire 1/2

Un **diagramme circulaire** est un graphique représentant des données qualitatives sous forme de tranches d'un cercle, la taille de la tranche étant proportionnelle au nombre de fréquences.



# Diagramme circulaire 2/2

Les diagrammes circulaires ne sont pas très recommandés à utiliser car ils peuvent être trompeurs.

En effet la quantité est représentée par des tranches et il n'est pas très aisé d'estimer la quantité à partir des angles.

Aussi les petits pourcentages (qui peuvent être importants) sont difficiles à représenter.

# Principes pour construire des bons graphiques statistiques

- Pour des petits ensembles de données de 20 valeurs ou moins, utilisez un tableau plutôt qu'un graphique.
- Un graphique de données doit concentrer l'attention sur la véritable nature des données, et non sur d'autres éléments, comme l'apparence esthétique du graphique.
- Ne déformez pas les données, construisez un graphique pour révéler la vraie nature des données.
- N'utilisez pas de surfaces ou de volumes pour des données qui sont en fait unidimensionnelles par nature.

Dans cette section, nous avons vu que les graphiques sont d'excellents outils pour décrire, explorer et comparer des données.

*Décrire des données* : Par exemple dans un histogramme on peut voir la distribution des données, le centre, la variation et les valeurs aberrantes.

*Exploration des données* : caractéristiques qui révèlent une caractéristique utile et/ou intéressante de l'ensemble des données.

*Comparer des données* : Construisez des graphiques similaires pour comparer des ensembles de données (polygones de fréquences).

Fin.