STATISTIQUES DESCRIPTIVES ET INFÉRENTIELLES

Résumé du cours (Lecture I) + correction des exercices

Mme Marwa Chalgham Abdennadher

Auditoires: 2ème année Licence en Science de l'Informatique: Analyse des Données et Big Data (D- LSI ADBD)

Introduction

Techniques qui consistent à résumer, simplifier, décrire et présenter les données concernant un phénomène particulier permettant de **fournir une présentation utile des données pour l'utilisateur.**

Statistiques

Techniques qui s'appuie sur l'utilisation de la théorie de probabilité. Il s'agit d'essayer de **tirer des conclusions concernant une population sur la base des résultats obtenus à partir d'un échantillon**.

Statistiques Descriptives

Statistiques Inférentielles

SD Unidimensionnels

Il s'agit de traiter des tableaux unidimensionnels présentent les données pour **une seule variable**. Leurs traitement et les représentations graphiques **varient de la nature des données**.

SD Bidimensionnels

Il s'agit de traiter des tableaux bidimensionnels qui présentent les données pour **deux variables.** Deux représentations possibles: tableaux **à 2 variables**, et tableaux de **contingence**.

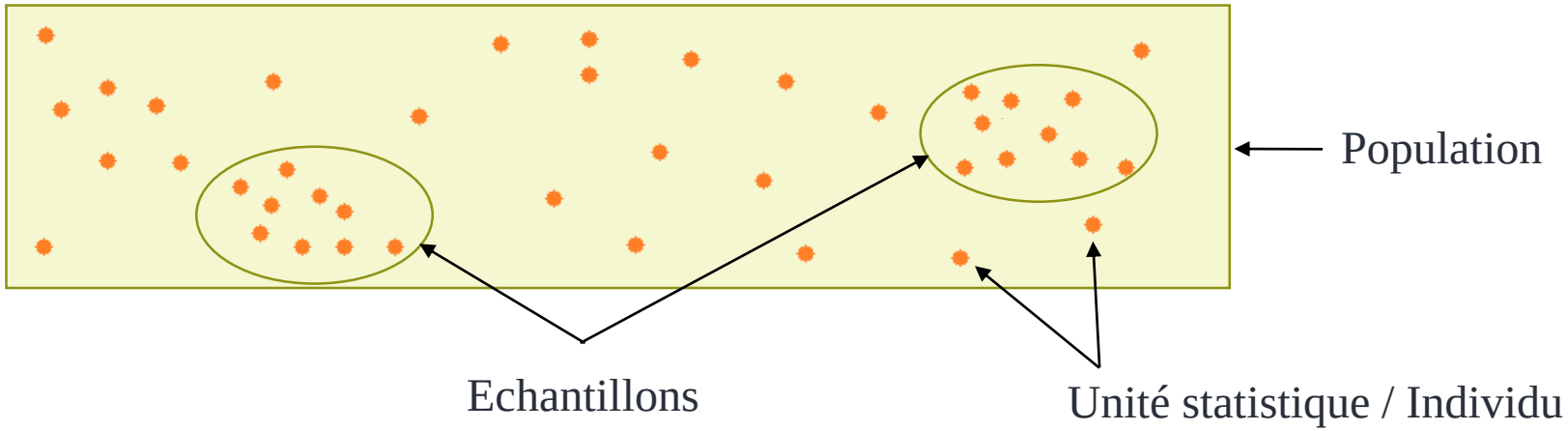
SD Multidimensionnels

Il s'agit de la discipline d'**Analyse des données**. Cette discipline traite les tableaux multidimensionnels qui présentent les données pour **plus de deux variables**.

Définitions et concepts fondamentaux

❖ Un échantillon:

Il s'agit d'un **sous-ensemble** qui représente **une partie de la population** sur laquelle porte l'étude. On fait recours au prélèvement d'un échantillon **lorsque la population est très importante**.



Exemple:

Dans une école primaire, il existe 450 élèves. Nous étudions l'âge de 50 élèves parmi eux.

Population : 450 élèves de l'école primaire.
Echantillon: 50 élèves de l'école primaire.
Individu : Un élève.

Définitions et concepts fondamentaux

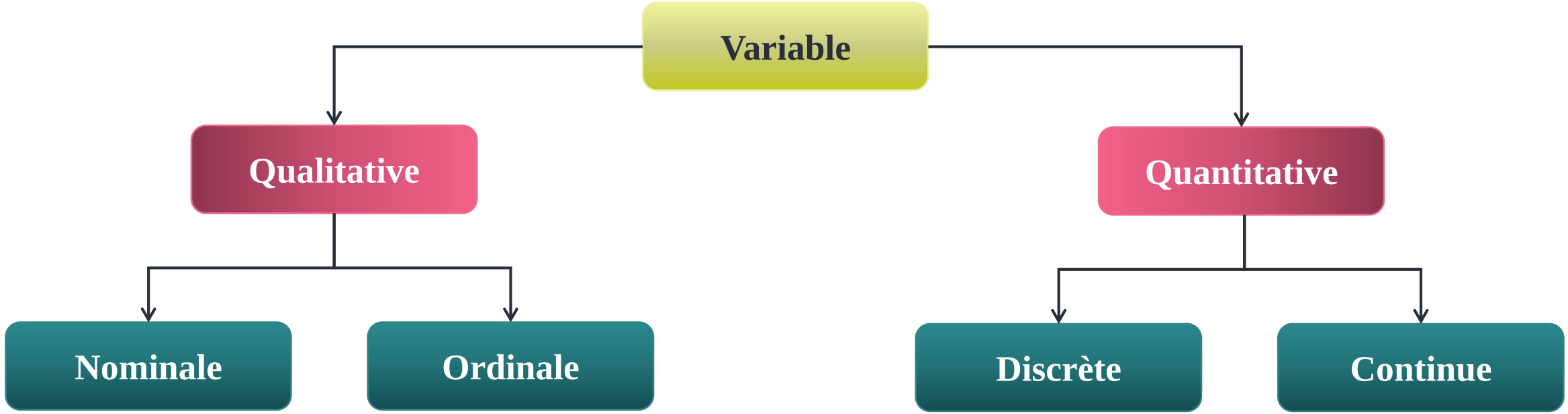
❖ Un caractère (variable):

Il s'agit de la **caractéristique mesurée** ou **observée** sur les individus de la population étudiée. Il peut être par exemple: l'âge, la taille, le salaire, la couleur des yeux, la nationalité, …. Le caractère (variable) peut être de nature **qualitative** ou **quantitative**.

- Une variable **qualitative** est une variable non mesurable tel que : la couleur des yeux, la nationalité, le sexe,…
  De même, il existe des variables qualitatives nominales (qui ne peuvent pas être ordonnées /classées), et des variables qualitatives ordinales (qui peuvent être ordonnées /classées sans ambigüité).

- Une variable **quantitative** est une variable numérique (représenté par des chiffres) tel que le salaire, le poids, l'âge,…
  De même, il existe des variables quantitatives discrètes (qui prend des valeurs entières et distinctes), et des variables quantitatives continues (qui ne sont pas précis c'est-à-dire qui s'inscrivent dans des intervalles).

Définitions et concepts fondamentaux



Exemples:

- Sexe
- Profession
- Nationalité
- Couleur des yeux

Exemples:

- « souvent » , « parfois »
- « faible » ,« bien » , « très bien »
- « satisfait » , « très satisfait »
- « grand » , « moyenne » , « petit »

Exemples:

- Nombre de joueurs
- Nombre de maisons
- Nombre d’enfants
- Nombre d’animaux

Exemples:

- Salaire mensuel
- Nombre de km
- Poids
- Taille d’une personne

Définitions et concepts fondamentaux

❖ **Une modalité:**

Il s'agit de l'ensemble des **valeurs possibles** qu'une variable peut la prendre.

Exemple:

Dans une école primaire, il existe 450 élèves. Nous étudions l'âge de 50 élèves parmi eux, sachant que les élèves sont regroupés selon la liste des âges suivante: « 6 ans », « 7 ans », « 8 ans », « 9 ans », « 10 ans », et « 11 ans ».

Population : 450 élèves de l'école primaire.

Echantillon: 50 élèves de l'école primaire.

Individu : Un élève.

Variable / caractère: âge.

Nature de la variable: quantitative discrète.

Modalité: « 6, 7, 8, 9, 10, 11 ».

Définitions et concepts fondamentaux

❖ **Série statistique:**

Il s'agit d'une **suite de valeurs** qu'une variable ***X*** a **pris sur les unités d'observation**. Le nombre d'unités d'observation est noté ***N***.

Exemple:

On s'intéresse à la variable « couleur de voiture préférée » notée *X* et à la **série statistique** des valeurs prises par *X* sur 20 personnes.

Soit la série statistique suivante:

$N = 20$

B	N	G	N	R
G	R	R	G	R
N	N	N	R	G
B	B	G	R	N

La codification de cette série est la suivante:

B :	Blanc
N :	Noir
R :	Rouge
G :	Grise

4 modalités

Définitions et concepts fondamentaux

❖ Tableau statistique:

Avec la série statistique des valeurs prises par *X* sur les 20 couleurs de voiture préférées de l'exemple précédant, on obtient **le tableau statistique** suivant. Dans la première colonne, les 4 modalités relatives aux couleurs préférées sont considérées. La deuxième colonne donne l'effectif de chaque modalité.

Série statistique :

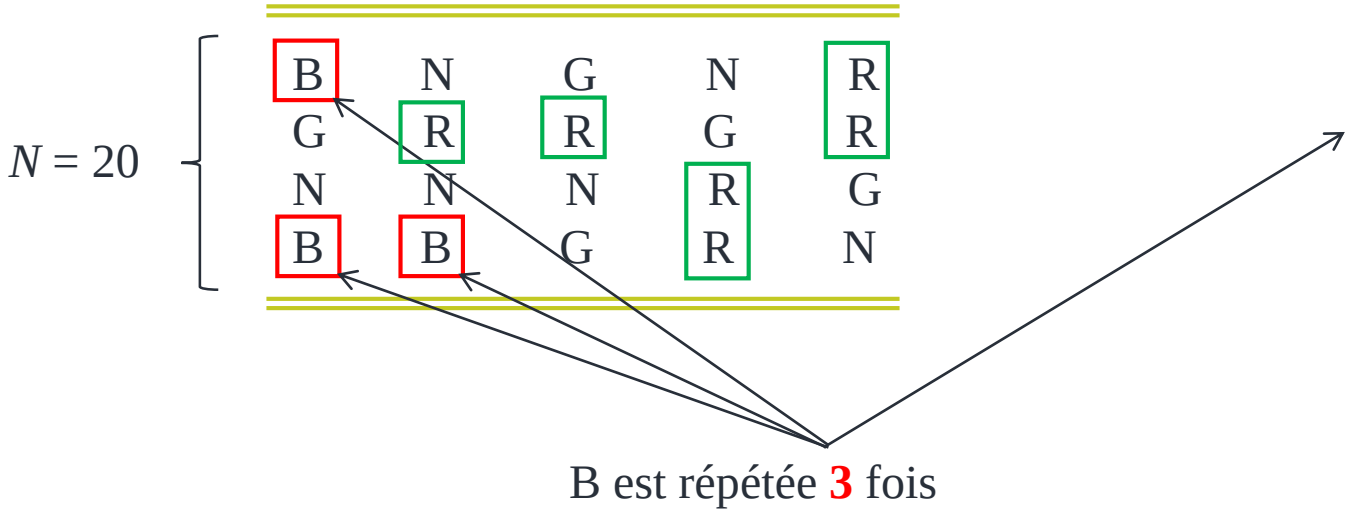


Tableau statistique :

Variables x_i	Effectifs n_i
B	**3**
N	6
R	**6**
G	5
Total	**N = 20**

Définitions et concepts fondamentaux

❖ Fréquences relatives :

Pour le tableau statistique précédant on calcule les fréquences relatives de chaque modalité:

6 personnes parmi 20 préfèrent avoir une voiture de couleur « Rouge ».

Variables	Effectifs	Fréquences relatives	Pourcentage	P. Cumulé
B	3	**3/20 = 0.15**	**15%**	**15%**
N	6	0.3	30%	45%
R	6	**0.3**	**30%**	**75%**
G	5	0.25	25%	**100%**
Total	20	**1**	**100%**	

15% de la population étudiée préfèrent avoir une voiture de couleur « Blanc ».

Les 6 personnes qui préfèrent la couleur « Rouge » représentent 30% de la population étudiée.

Statistiques Descriptives Unidimensionnels: Distribution à un seul caractère (variable)

Exercise 1:

A survey asked 40 college students majoring in business:

What is your major?

(A= Accounting; C= Computer Information Systems; M= Marketing).

Answers are: A C C M A C A A C C A A A M C M A A A C C C A A M M C A A A C C A A A A C C A C

Give the summary table of frequencies.

Variable	Effectif	Fréquence relative	Pourcentage
Accounting	20	20/40 = 0,5	50%
Computer Information Systems	15	15/40 = 0,375	37,5%
Marketing	5	5/40 = 0,125	12,5%
Total	40	1	100%

Définitions et concepts fondamentaux

❖ Tableau de contingence:

La 1ère ligne: Modalités de Y

X \ Y	Rarement absent	Moyennement absent	Fréquemment absent
Niveau faible	0	2	9
Niveau moyenne	2	2	4
Niveau élevé	9	6	1
Niveau excellent	12	3	0

La 1ère colonne: Modalités de X

il y en a 9 étudiants de niveau faible étant fréquemment absent

il y en a 2 étudiants de niveau moyenne étant moyennement absent

il y en a 9 étudiants de niveau élevé étant rarement absent

il y en a 3 étudiants de niveau excellent étant moyennement absent

Il n'y a pas d'étudiants de niveau excellent étant fréquemment absent

Définitions et concepts fondamentaux

❖ Tableau de contingence:

Contingency table

X \ Y	Rarement absent	Moyennement absent	Fréquemment absent	Total
Niveau faible	0	2	9	11
Niveau moyenne	2	2	4	8
Niveau élevé	9	6	1	16
Niveau excellent	12	3	0	15
Total	23	13	14	**50**

Contingency table based on total percentages

X \ Y	Rarement absent	Moyennement absent	Fréquemment absent	Total
Niveau faible	0%	4%	**18%**	22%
Niveau moyenne	4%	4%	8%	16%
Niveau élevé	18%	12%	2%	**32%**
Niveau excellent	24%	6%	0%	30%
Total	**46%**	**26%**	28%	100%

Quelques conclusions:

18% des étudiants ayant un niveau faible et s'absentent fréquemment

46% des étudiants sont rarement absents

26% des étudiants sont moyennement absents

32% des étudiants sont ont un niveau élevé

Définitions et concepts fondamentaux

❖ Tableau de contingence:

Contingency table based on row percentages

X \ Y	Rarement absent	Moyennement absent	Fréquemment absent	Total
Niveau faible	0%	2/11 = 18%	9/11 = 82%	100%
Niveau moyenne	2/8 = 25%	2/8 = 25%	4/8 = **50%**	100%
Niveau élevé	9/16 = 56%	6/16 = **37%**	1/16 = 7%	100%
Niveau excellent	12/15 = **80%**	3/15 = 20%	0%	100%

80% des étudiants ayant un niveau excellent, s'absentent rarement

50% des étudiants ayant un niveau moyenne, s'absentent fréquemment

37% des étudiants ayant un niveau élevé, s'absentent moyennement

Contingency table based on column percentages

X \ Y	Rarement absent	Moyennement absent	Fréquemment absent
Niveau faible	0%	2/13 = 15%	9/14 = 64%
Niveau moyenne	2/23 = 9%	2/13 = 16%	4/14 = **29%**
Niveau élevé	9/23 = 39%	6/13 = 46%	1/14 = 7%
Niveau excellent	12/23 = **52%**	3/13 = 23%	0%
Total	100%	100%	100%

29% des étudiants fréquemment absent, ont un niveau moyenne

52% des étudiants rarement absent, ont un niveau excellent

Statistiques Descriptives Bidimensionnels: Distribution à 2 variables

Exercise 2: The following data represent the responses to two questions asked in a survey of 40 college students majoring in business:

What is your gender? (M=Male; F=Female) and what is your major?

Gender	M	M	M	F	M	F	F	M	F	M	F	M	M	M	M	F	F	M	F	F
Major	A	C	C	M	A	C	A	A	C	C	A	A	A	M	C	M	A	A	A	C
Gender	M	M	M	M	F	M	F	F	M	M	F	M	M	M	M	F	M	F	M	M
Major	C	C	A	A	M	M	C	A	A	A	C	C	A	A	A	A	C	C	A	C

1. Tally the data into a contingency table where rows represent the gender categories and columns represent the academic major categories.
2. Construct a contingency table based on percentages of all 40 students' responses.
3. The contingency table based on row percentages of students' responses.
4. The contingency table based on column percentages of students' responses.
5. What conclusions can you reach from these analyses?

Statistiques Descriptives Bidimensionnels: Distribution à 2 variables

1. The contingency table:

X \ Y	Accounting	Computer Information Systems	Marketing	Total
Male	14	9	2	25
Female	6	6	3	15
Total	20	15	5	40

2. The contingency table based on percentages of **all** 40 students' responses:

X \ Y	Accounting	Computer Information Systems	Marketing	Total
Male	$\frac{14}{40} \times 100 = 35\%$	22,5%	5%	62,5%
Female	15%	15%	7,5%	37,5%
Total	50%	37,5%	12,5%	100%

Some conclusions :

35% of students are male following Accounting

7,5% of students are Female following Marketing

22,5% of students are male following Computer Information System

Statistiques Descriptives Bidimensionnels: Distribution à 2 variables

3. The contingency table based on **row** percentages of students' responses:

X \ Y	Accounting	Computer Information Systems	Marketing	Total
Male	$\frac{14}{25} \times 100 = 56\%$	$\frac{9}{25} \times 100 = 36\%$	$\frac{2}{25} \times 100 = 8\%$	100%
Female	$\frac{6}{15} \times 100 = 40\%$	$\frac{6}{15} \times 100 = 40\%$	$\frac{3}{15} \times 100 = 20\%$	100%

4. The contingency table based on **column** percentages of students' responses:

X \ Y	Accounting	Computer Information Systems	Marketing
Male	$\frac{14}{20} \times 100 = 70\%$	$\frac{9}{15} \times 100 = 60\%$	$\frac{2}{5} \times 100 = 62{,}5\%$
Female	$\frac{6}{20} \times 100 = 30\%$	$\frac{6}{15} \times 100 = 40\%$	$\frac{3}{5} \times 100 = 37{,}5\%$
Total	100%	100%	100%

5. What conclusions can you reach from these analyses?

8% of male'students are following Marketing

20% of female'students are following Marketing

60% of Computer Information Systems followers are male

30% of Accounting followers are Female

Représentations graphiques

Variable Qualitative

Le diagramme en Tuyau d'orgue (Bar Chart)

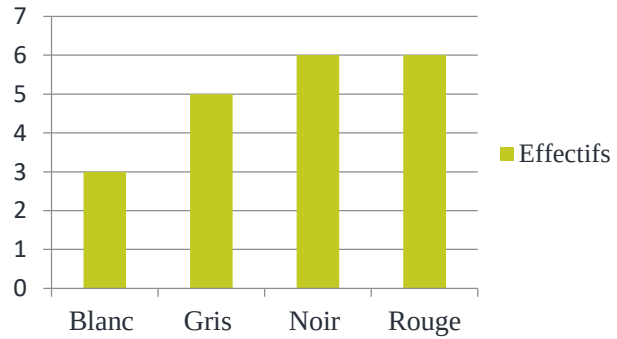
Des rectangles de hauteur = à l'effectif ou à la fréquence (%) qui se présente selon un ordre croissant ou décroissant de leurs effectifs correspondants. Les modalités se comparent directement les unes aux autres.

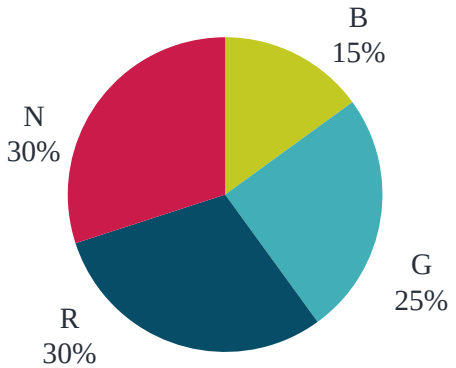
Le diagramme à secteurs (Pie Chart)

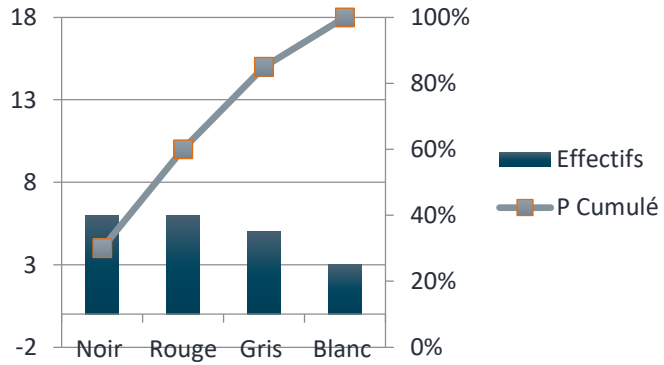
Représentation sous forme de cercles qui se compose de plusieurs parties. Il faut calculer la part que représente chaque modalité en multipliant sa fréquence par 360°. Il met en évidence une modalité par rapport aux autres.

Le diagramme de Pareto (Preto Chart)

Il visualise les causes d'un effet et identifie les causes les plus critiques. Il est basé sur la loi de Pareto, qui affirme que 80 % des effets peuvent être attribués à 20 % des causes. Le diagramme de Pareto comporte deux parties : une barre représentant la fréquence des différentes causes et une courbe représentant la somme cumulative des fréquences.







Répartition des couleurs préférées d'une voiture pour 20 personnes en utilisant 3 représentations graphiques différentes

Exercise 4:

Consider a bank study team that wants to enhance the user experience of automated teller machines (ATMs). The team identifies incomplete ATM transactions as a significant issue. It decides to collect data about the causes of such transactions using the bank's own processing systems. Data are organized in the table below.

Cause	Frequency
ATM malfunctions	32
ATM out of cash	28
Invalid amount requested	23
Lack of funds in account	19
card unreadable	234
Warped card jammed	365
Wrong keystroke	23
Total	**724**

Using Excel, construct a Pareto chart of causes of incomplete ATM transactions.

Statistiques Descriptives Unidimensionnels: Distribution à un seul caractère (variable)

Cause	Effectifs	Pourcentage
ATM malfunctions	32	4,42 %
ATM out of cash	28	3,87 %
Invalid amount requested	23	3,18 %
Lack of funds in account	19	2,62 %
Card unreadeble	234	32,32 %
Warped card jammed	365	50,41 %
Wrong keystroke	23	3,18 %
Total	**724**	**100%**

Tableau trié à l'ordre décroissant

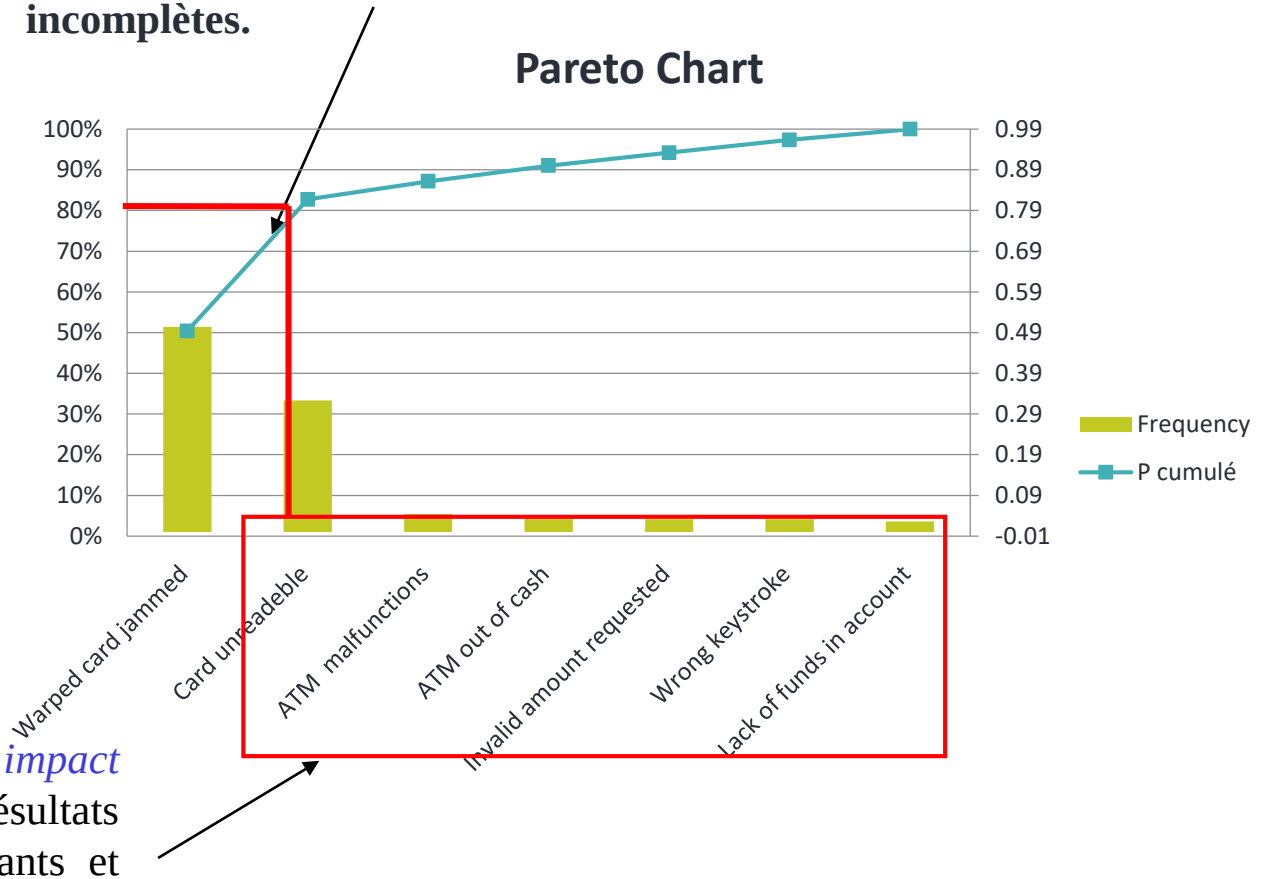
Cause	Effectifs	Pourcentage
Warped card jammed	365	50,41 %
Card unreadeble	234	32,32 %
ATM malfunctions	32	4,42 %
ATM out of cash	28	3,87 %
Invalid amount requested	23	3,18 %
Wrong keystroke	23	3,18 %
Lack of funds in account	19	2,62 %
Total	**724**	**100 %**

Statistiques Descriptives Unidimensionnels: Distribution à un seul caractère (variable)

Tableau trié à l'ordre décroissant

Cause	Effectifs	Pourcentage	% Cumulé
Warped card jammed	365	50%	50%
Card unreadeble	234	32%	83%
ATM malfunctions	32	4%	87%
ATM out of cash	28	4%	91%
Invalid amount requested	23	3%	94%
Wrong keystroke	23	3%	97%
Lack of funds in account	19	3%	**100%**
Total	**724**	**100%**	

Les « **vital few** » sont les **20%** des facteurs qui ont le *plus grand impact (80% d'impact)* et génèrent la plupart des résultats. Ces facteurs sont considérés comme les plus importants et sont au centre de l'attention pour l'amélioration. **Dans notre cas les cartes déformées coincées (Warped card jammed) sont les principales causes des transactions incomplètes.**



Les « **trivial many** » sont les **80%** des facteurs qui ont un *impact relativement mineur (20% d'impact)* et contribuent peu aux résultats globaux. Ces facteurs ne sont pas considérés comme importants et peuvent ne pas valoir l'effort nécessaire pour les aborder.

Exercise 5:

The Consumer Financial Protection Bureau reports complaints received from Louisiana consumers. The Table1 of Excel Sheet DataComplaints gives the number of complaints by category for 2016.

1. Construct a Pareto chart for the categories of complaints.
2. Discuss the “vital few” and “trivial many” reasons for the categories of complaints.

The Table2 of the same Excel Sheet gives the tally of the complaints received from Louisiana consumers by most-complained-about companies for 2016.

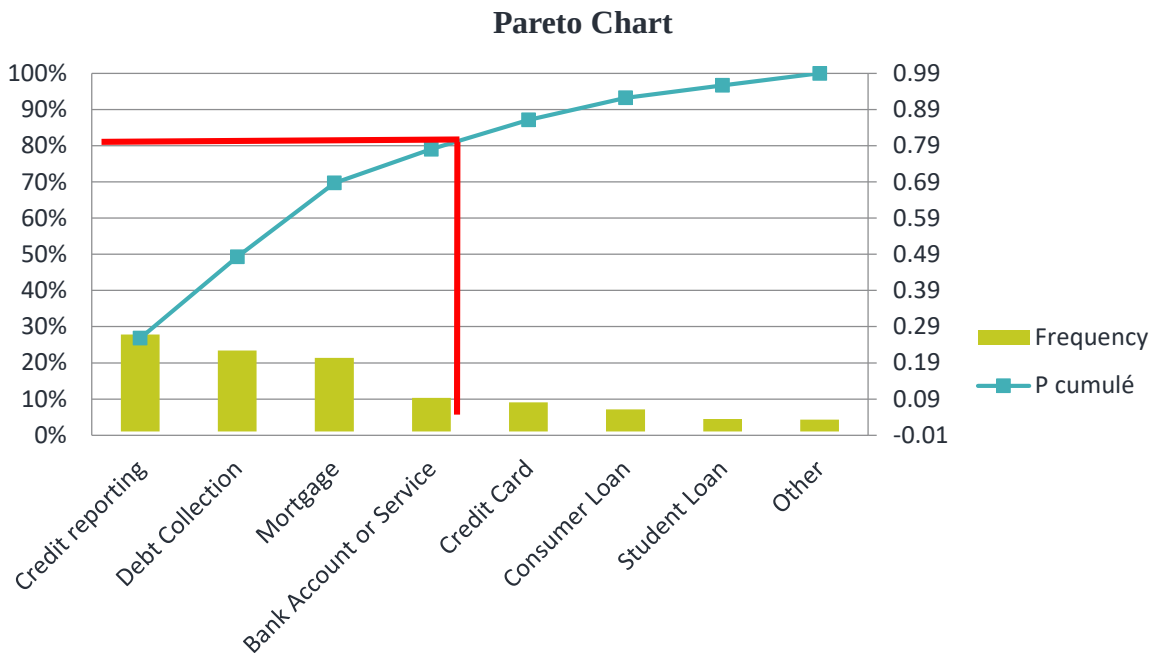
3. Construct a bar chart and a pie chart for the complaints by company.
4. What graphical method (Pareto, bar, or pie chart) do you think is best for portraying these data.

Statistiques Descriptives Unidimensionnels: Distribution à un seul caractère (variable)

1. Construct a Pareto chart for the categories of complaints:

Table 1 trié à l'ordre décroissant

Category	Number of Complaints	Frequency	Percentage	P cumulé
Credit reporting	581	0,27	27%	27%
Debt Collection	486	0,22	22%	49%
Mortgage	442	0,20	20%	70%
Bank Account or Service	202	0,09	9%	79%
Credit Card	175	0,08	8%	87%
Consumer Loan	132	0,06	6%	93%
Student Loan	75	0,03	3%	97%
Other	72	0,03	3%	100%
Total	2165	1	100%	



2. Discuss the "vital few" and "trivial many" reasons for the categories of complaints:

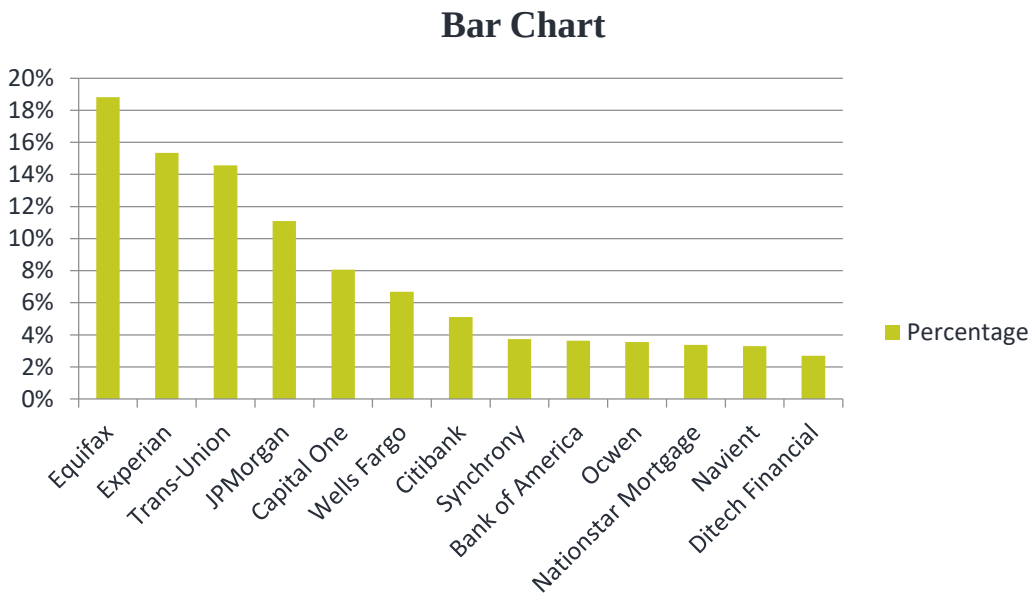
Credit reporting, debt collection, mortgage, et Bank account or service forment l'ensemble des catégories du « vital few » les plus importants et qui doivent être au centre de l'attention pour l'amélioration. Ces 4 catégories sont la source de 80% des réclamations. Le reste des catégories forment l'ensemble qui a un impact mineur (20% des réclamations) et qui ne nécessitent pas de l'effort pour les corriger.

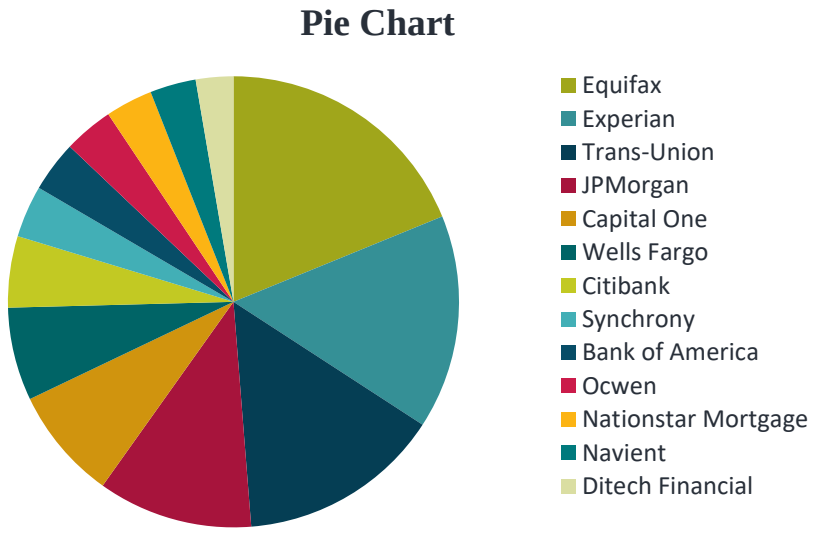
Statistiques Descriptives Unidimensionnels: Distribution à un seul caractère (variable)

3. Construct a bar chart and a pie chart for the complaints by company:

Table 2 trié à l'ordre décroissant

Company	Number of Complaints	Frequency	Percentage	P cumulé
Equifax	217	0,19	19%	19%
Experian	177	0,15	15%	34%
Trans-Union	168	0,15	15%	49%
JPMorgan	128	0,11	11%	60%
Capital One	93	0,08	8%	68%
Wells Fargo	77	0,07	7%	75%
Citibank	59	0,05	5%	80%
Synchrony	43	0,04	4%	83%
Bank of America	42	0,04	4%	87%
Ocwen	41	0,04	4%	91%
Nationstar Mortgage	39	0,03	3%	94%
Navient	38	0,03	3%	97%
Ditech Financial	31	0,03	3%	100%
Total	1153	1	100%	





Statistiques Descriptives Unidimensionnels: Distribution à un seul caractère (variable)

4. What graphical method (Pareto, bar, or pie chart) do you think is the best for portraying these data:

Dans cet exercice, les données présentent les entreprises sources des réclamations. Alors dans ce cas, l'objectif est de **visualiser** les entreprises auprès desquels nous avons reçu **le nombre le plus important de réclamation**. Ceci ne peut être fait qu'avec le diagramme de Pareto. En effet, ce diagramme est souvent utilisé pour identifier les problèmes les plus importants dans un système et pour optimiser par la suite les processus en concentrant les efforts sur les causes les plus critiques.

Le diagramme de Pareto est un outil puissant pour améliorer les processus en identifiant les causes les plus importantes d'un effet donné et en ciblant les efforts pour les corriger. Dans cet exemple, il **permet de répondre à la question: « d'où vient 80% des réclamations? »** .

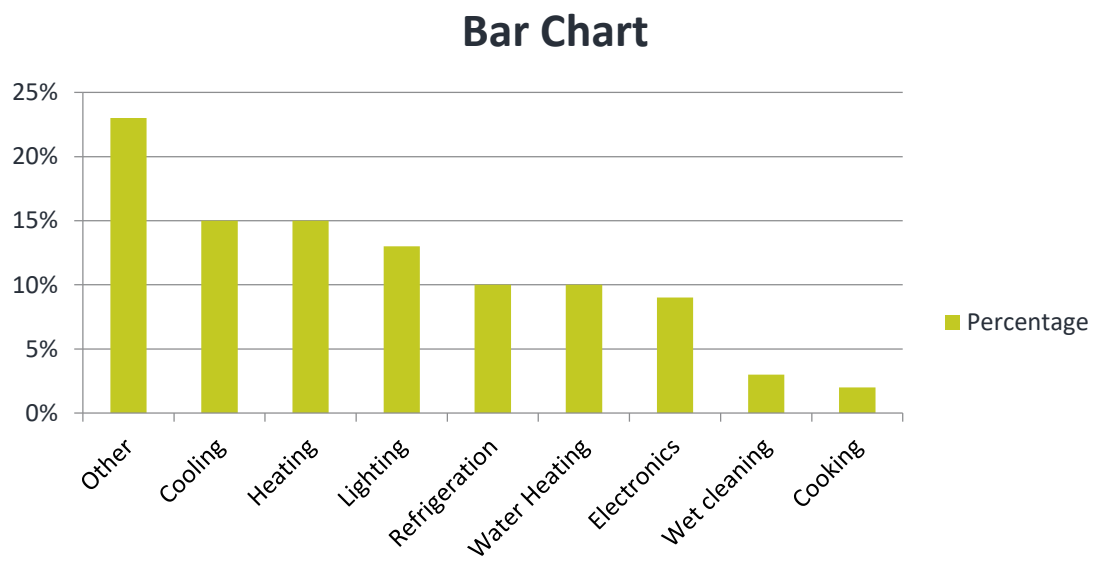
Exercise 6:

Table on Excel Sheet DataElectricity indicates the percentage of residential electricity consumption in the Unites States, in a recent year organized by type of use.

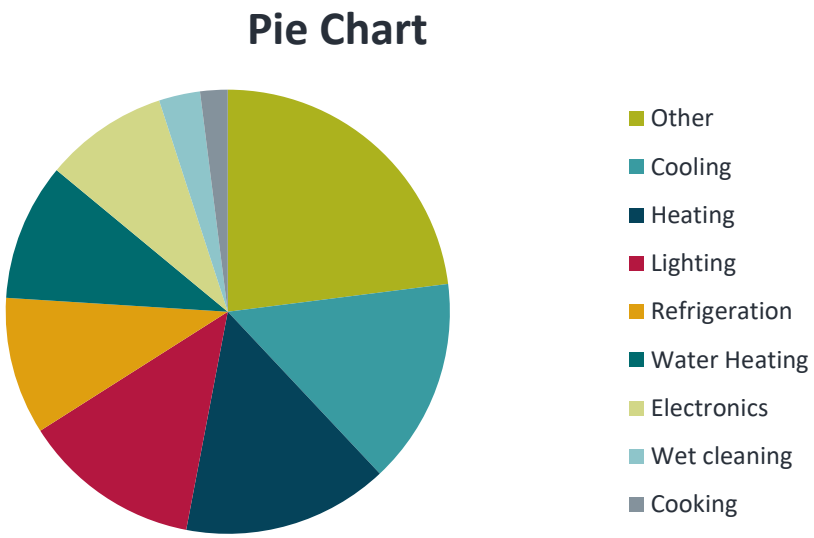
1. Construct a bar chart, a pie chart, and a Pareto chart.
2. Which graphical method do you think is best for portraying these data ?
3. What conclusions can you reach concerning residential electricity consumption in the United States?

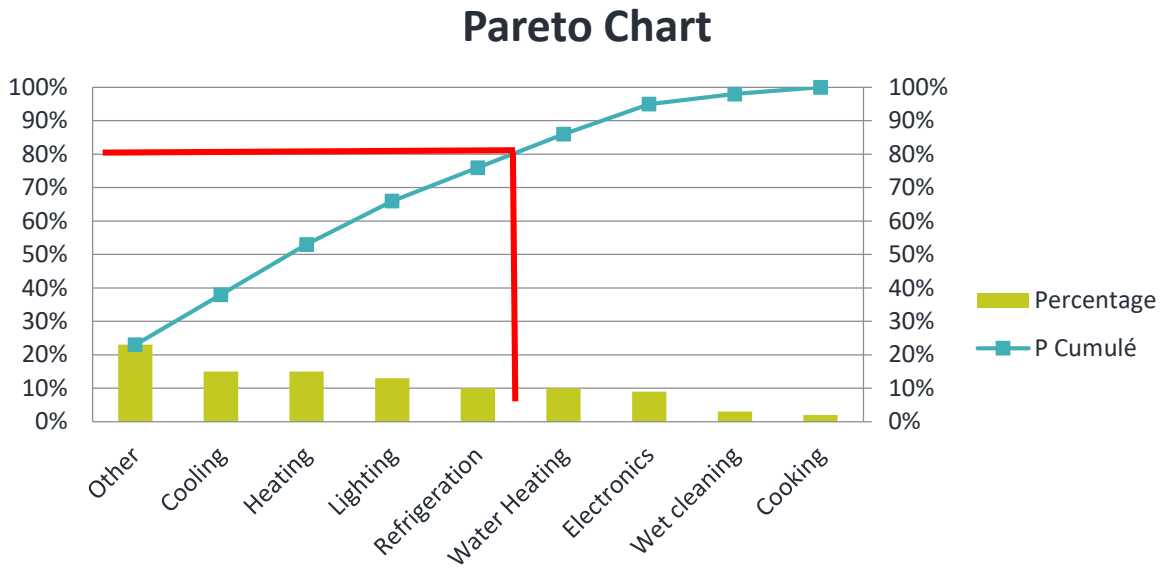
Table trié à l'ordre décroissant

Type of Use	Percentage	P Cumulé
Other	23%	23%
Cooling	15%	38%
Heating	15%	53%
Lighting	13%	66%
Refrigeration	10%	76%
Water Heating	10%	86%
Electronics	9%	95%
Wet cleaning	3%	98%
Cooking	2%	100%



Statistiques Descriptives Unidimensionnels: Distribution à un seul caractère (variable)





2. Which graphical method do you think is best for portraying these data ?

Les données dans cet exercice indiquent le pourcentage de la consommation d'électricité résidentielle aux États-Unis par type d'utilisation au cours d'une année. Alors, le diagramme Pie Chart pourrait être la meilleure présentation si on vise la mise en évidence de la source de consommation d'électricité la plus importante par rapport aux autres. Alors que, si on vise la comparaison entre les sources de de consommation d'électricité, le Bar chart convient plus à cet objectif. Cependant, le Pareto Chart, reste l'un des représentations les plus puissante pour mettre en évidence 80% de la source de consommation d'électricité.

Statistiques Descriptives Unidimensionnels: Distribution à un seul caractère (variable)

3. What conclusions can you reach concerning residential electricity consumption in the United States?

On peut conclure que la consommation d'électricité pour le refroidissement (cooling), et le chauffage (heating) sont des sources de consommation très importantes représentant 15% chacune. Cependant, plusieurs autres sources (Other) qui n'ont pas été détaillées dans l'exercice mais qui ont représenté le part le plus important parmi toutes les sources de consommation d'électricité (23%).

En plus, le Pareto Chart a montré que 80% des sources de consommation d'électricité sont à la base du refroidissement (cooling), chauffage (heating), Éclairage (Lighting), Réfrigération (Refrigeration), and autres sources (Other). Le reste des sources présentées dans le tableau représentent 20% seulement de toutes les sources de consommation d'électricité.

Définitions et concepts fondamentaux

❖ Amplitude (Interval width) :

Dans le cas de **variables quantitatives continues**, il est recommandé de choisir des amplitudes de classe qui ont une largeur suffisante pour capturer la variation des données et qui contient suffisamment de données pour permettre une analyse précise.

$$\textbf{Amplitude de classe} = \frac{\text{La plus grande valeur dans la série (Max)} - \text{La plus petite valeur dans la série (Min)}}{\text{Nombre de classes}}$$

Série statistique:

22	24	26	20	22
22	25	22	25	20
27	22	26	20	27
24	26	20	20	22

Amplitude = (27-20)/4 = 2

À partir de cette formule, on peut aussi trouver le nombre de classes en fixant l'amplitude:

$$\textbf{Nombre de classes} = \frac{\text{Max} - \text{Min}}{\text{Amplitude de classe}}$$

Définitions et concepts fondamentaux

❖ Définir les limites de chaque classe :

Les limites de classe doivent être choisies afin que chaque élément ou observation soit placé dans une seule classe. La limite de la classe inférieure identifie la plus petite valeur de données possible affectée à la classe. La limite de la classe supérieure identifie la plus grande valeur de données possible affectée à la classe.

Avec les données de notre exemple, des limites de classe sont nécessaires pour déterminer où chaque valeur de données appartient. Nous pouvons recalculer la valeur d'une des bornes en fixant l'autre:

Si on fixe le Min = 20 : $\frac{\boldsymbol{Max - 20}}{4} = 2$ *donc* $\boldsymbol{Max - 20 = 8}$ *alors* $\boldsymbol{Max = 28}$

Si on fixe le Max = 27 : $\frac{\boldsymbol{27 - Min}}{4} = 2$ *donc* $\boldsymbol{27 - Min = 8}$ *alors* $\boldsymbol{Min = 19}$

Dans ce cas, la valeur 27 ne sera pas prise dans le dernier intervalle donc il faut toujours fixer une valeur plus grande pour éviter ce problème!

Définitions et concepts fondamentaux

Nous avons choisir la valeur 20 comme borne inférieure pour le premier classe et la valeur 28 comme borne supérieur pour le dernier classe en passant d'un intervalle de 2 d'un classe au classe suivant:

22	24	26	20	22
22	25	22	25	20
27	22	26	20	27
24	26	20	20	22

Tenir uniquement les âges de 20 et 21

Tenir uniquement les âges de 22 et 23

Variables	Effectifs	Amplitudes a_i
[20, 22[	5	2
[22, 24[	6	2
[24, 26[	4	2
[26, 28[	5	2
Total	20	∅

amplitude d'une classe a_i $= Borne\ sup - Borne\ inf$

L'amplitude de la classe [20, 22 [:
22 – 20 = 2

Exercise 7:

The Excel Sheet Data5 contains the time in seconds to answer 50 incoming calls to a financial services call center.

1. Construct a frequency distribution and a percentage distribution.

2. Construct a cumulative percentage distribution.

3. The service target level is set at “80% of calls is answered within 20 seconds”. What do you conclude about call center performance ?

16	14	16	19	6	14	15	5	16	18	17	22	6	18	10	15	12
19	16	16	15	13	25	9	17	12	10	5	15	23	11	12	14	24
10	13	14	26	19	20	13	24	28	15	21	8	16	12	9	6	

Dans cette série, nous avons la valeur 5 est la plus petite (Min), et la valeur 28 la plus grande (Max). Alors, 28 – 5 = 23.

Si nous souhaitons avoir 5 classes alors l’amplitude sera calculer comme suivant:

$$\textbf{Amplitude de classe} = \frac{\text{Max} - \text{Min}}{\text{Nombre de classes}} = \frac{28 - 5}{5} = 4{,}6 \cong \textit{5}$$

On fixe maintenant les bornes limites du 1er et du 5ème classe:

Si on fixe le Min = 5 : $\frac{\boldsymbol{Max - 5}}{5} = 5$ *donc* $\boldsymbol{Max - 5 = 25}$ *alors* $\boldsymbol{Max = 25 + 5 = 30}$

Ça veut dire que les bornes limites du 1er classe sont : **[5, 10[** , et les bornes limites du 5ème classe sont : **[25, 30[**

Variables	frequency	Proportion (relative frequency)	Percentage	P Cumulés
[5, 10[	8	0,16	16%	16%
[10, 15[	15	0,3	30%	46%
[15, 20[	18	0,36	36%	82%
[20, 25[	6	0,12	12%	94%
[25, 30[	3	0,06	6%	100%
Total	50	1	100%	

L'objectif cible de ce service est fixé à "80 % des appels doivent être répondus dans les 20 secondes". Nous pouvons conclure que l'objectif est atteint puisque la courbe des pourcentages cumulative prouve que **82% des appels sont répondus dans un intervalle de 5 à 19 secondes.**

Exercise 8:

The file Data6 contains average age of the players (years, in 2018) of the 32 teams that qualified for the FIFA 2018 World Cup.

1. Organize these mean ages in an ordered array.
2. Construct a frequency distribution and a percentage distribution for these mean ages.
3. Around which class grouping, if any, are these mean ages concentrated? Explain.

Data

26.04	26.78	27.17	27.57
28.17	28.43	28.61	28.96
26.09	27.09	27.26	27.78
28.22	28.43	28.78	29.17
26.09	27.09	27.26	27.83
28.26	28.52	28.83	29.52
26.48	27.09	27.48	28.09
28.35	28.52	28.91	29.74

Ordered array

26.04	27.17	28.17	28.61
26.09	27.26	28.22	28.78
26.09	27.26	28.26	28.83
26.48	27.48	28.35	28.91
26.78	27.57	28.43	28.96
27.09	27.78	28.43	29.17
27.09	27.83	28.52	29.52
27.09	28.09	28.52	**29.74**

Dans cette série, nous avons la valeur 26.04 est la plus petite (Min), et la valeur 29.74 la plus grande (Max). Alors, si nous souhaitons avoir 4 classes alors l'amplitude sera calculer comme suivant:

$$\textbf{Amplitude de classe} = \frac{29.74 - 26.04}{4} = 0{,}925 \cong \textit{1}$$

On fixe maintenant les bornes limites du 1er et du 4ème classe:

Si on fixe le Min = 26 : $\frac{Max - 26}{4} = 1$ *donc* $Max - 26 = 4$ *alors* $Max = 4 + 26 = 30$

Ça veut dire que les bornes limites du 1er classe sont : **[26, 27[** , et les bornes limites du 5ème classe sont : **[39, 30[**

Frequency distribution and a percentage distribution

Variables	Frequency	Proportion (relative frequency)	Percentage
[26, 27[	5	0,16	16%
[27, 28[	10	0,31	31%
[28, 29[	**14**	0,44	**44%**
[29, 30[	3	0,09	9%
Total	32	1	100%

L'âge moyenne des joueurs est **concentré autour de 28 ans**. En effet, l'âge 28 ans présente l'effectif le plus élevé qui se traduit par 44% de la population étudiée.

Représentations graphiques

Variable Quantitatives

Le diagramme de « tige et feuilles » (Stem-and-Leaf)

Il peut être utilisé pour représenter des données quantitatives discrètes et continues. Il permet une interprétation simple et rapide pour les données présentées. En plus il aide à déterminer le mode, la médiane et la moyenne.

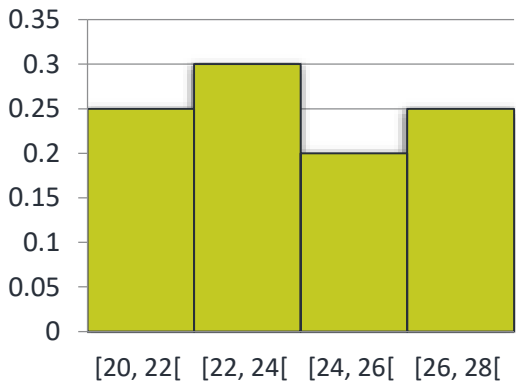
1	2 2 5
2	1 2 2
3	5 6 8
4	4 7 9

Tige (Stem) **Feuilles (Leaf)**

12	12	15	21	22	22
35	36	38	44	47	49

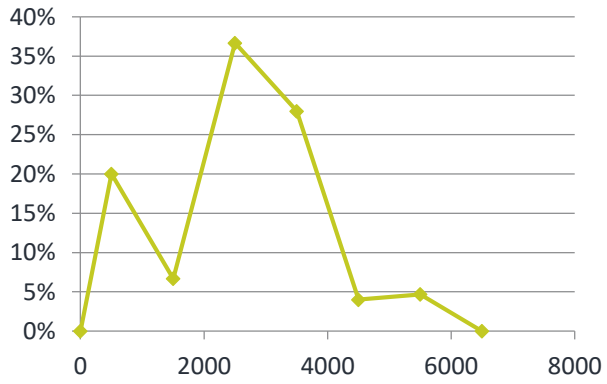
L'histogramme (Histogram)

Représentation appropriée pour les données continues car il permet de visualiser la distribution des données le long d'une échelle continue.



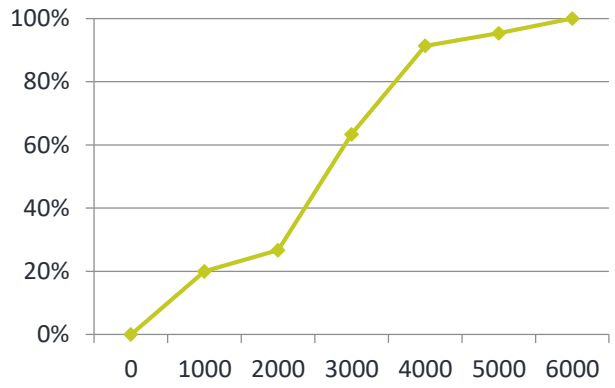
Le Polygone des fréquences (Percentage Polygon)

Représentation appropriée pour les données continues. Il se compose de segments de lignes reliant les **points milieux** des sommets de chaque classe. Il permet de comparer directement deux ou plusieurs distributions de fréquences.



Le polygone des % cumulées (Cumulative % polygon)

Ce graphique permet de visualiser rapidement la proportion de valeurs inférieures ou égales à une valeur donnée, ainsi que la forme générale de la distribution des données.

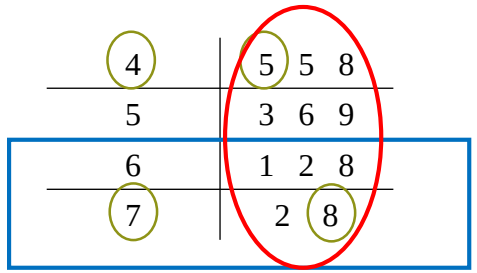


Exercise 9:

Le diagramme à tige et feuilles ci-dessous représente la répartition du temps en minutes obtenus à une course de 10 km.

1. À partir du diagramme à tige et feuilles, présenter la série de données initiale.
2. Combien de minutes séparent le coureur le plus rapide et le coureur le plus lent?
3. Combien de coureurs ont participer à cette course?
4. Combien de coureurs ont terminé la course dans moins qu'une heure?

Diagramme à tige et feuilles



La série de données initiale

45	45	48	53	56	59
61	62	68	72	78	

❖ Range = **78 – 45 = 33** minutes séparent le coureur le plus rapide et le coureur le plus lent.

❖ Le nombre de valeurs dans la partie feuilles (**11**) représente le nombre de coureurs qui ont participer à la course.

❖ Le nombre de valeurs dans la partie des feuilles en **carré bleu** représente les coureurs qui ont dépasser 60 minutes (1 heure). Donc le reste des valeurs se traduit par le nombre de coureurs qui ont terminé la course dans moins qu'une heure (**6 coureurs**).

Exercise 14:

Take again the Excel Sheet Data6 on the average age of players.

1. Construct a stem-and-leaf display.

2. Around which value, if any, are the mean ages of teams concentrated? Explain.

Data

26.04	26.78	27.17	27.57
28.17	28.43	28.61	28.96
26.09	27.09	27.26	27.78
28.22	28.43	28.78	29.17
26.09	27.09	27.26	27.83
28.26	28.52	28.83	29.52
26.48	27.09	27.48	28.09
28.35	28.52	28.91	29.74

Stem-and-leaf display

Stem	Leaf
26,	04 09 09 48 78
27,	09 09 09 17 26 26 48 57 78 83
28,	09 17 22 26 35 43 43 52 52 61 78 83 91 96
29,	17 52 74

L'âge moyenne des joueurs est **concentré autour de 28 ans**.

En effet, l'âge 28 ans présente l'effectif le plus élevé.

Exercise 10:

Les données ci-dessous représentent la répartition de150 clients selon le montant total de leurs achats.

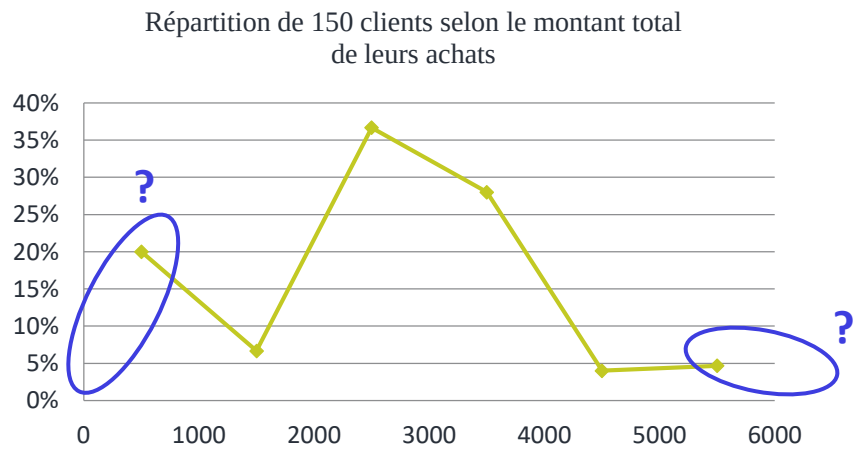
1. Tracer le Polygone des fréquences.

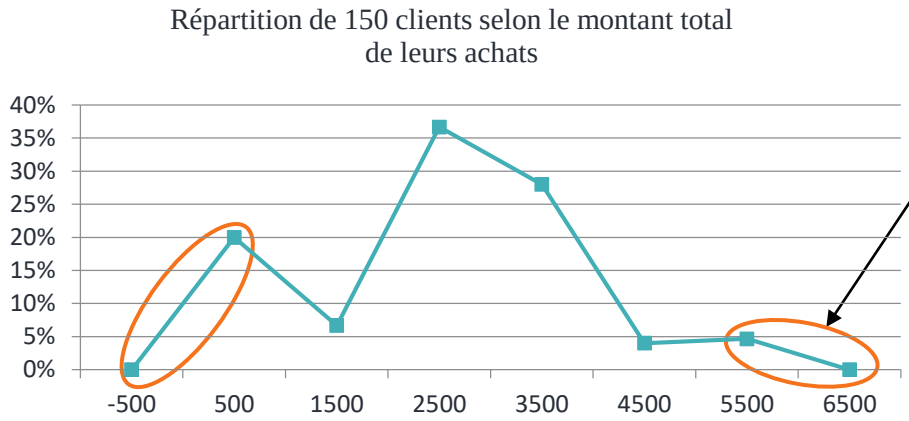
Répartition de 150 clients selon le montant total de leurs achats		
Montant total d'achats (DT)	Nombre de Clients	Pourcentage
[0, 1000[	30	20%
[1000, 2000[	10	7%
[2000, 3000[	55	37%
[3000, 4000[	42	28%
[4000, 5000[	6	4%
[5000, 6000[	7	5%
Total	150	100%

Le centre de classe = (B Inf + B Sup)/2

Présentation nécessaire pour la création d'un polygone			
Borne inférieure	Borne supérieure	Centre de classe	Pourcentage
-1000	0	-500	0%
0	1000	500	20%
1000	2000	1500	7%
2000	3000	2500	37%
3000	4000	3500	28%
4000	5000	4500	4%
5000	6000	5500	5%
6000	7000	6500	0%

Sur EXCEL, il faut ajouter 2 lignes pour une présentation correcte du polygone!

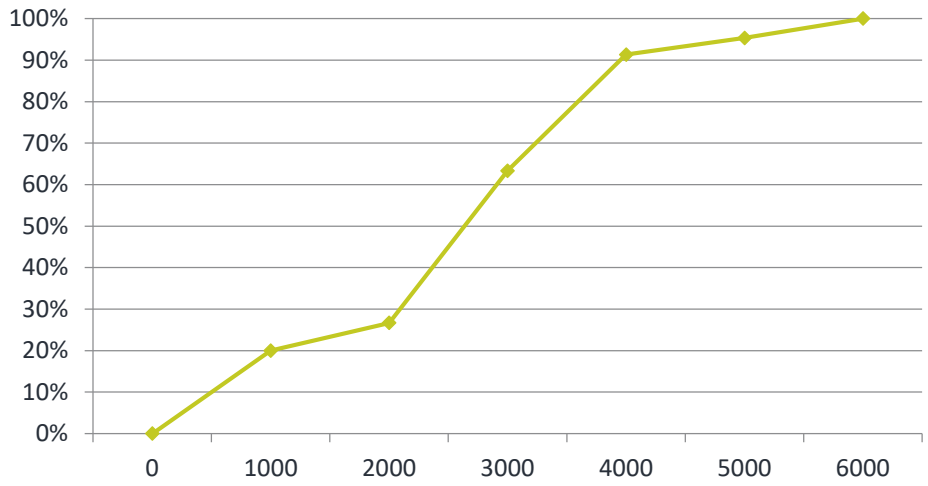




2. Tracer le Polygone des fréquences cumulées.

Répartition de 150 clients selon le montant total de leurs achats		
Montant total d'achats (DT)	Nombre de Clients	Pourcentage
[0, 1000[	30	20%
[1000, 2000[	10	7%
[2000, 3000[	55	37%
[3000, 4000[	42	28%
[4000, 5000[	6	4%
[5000, 6000[	7	5%
Total	150	100%

Présentation nécessaire pour la création d'une courbe cumulative	
Borne supérieure	% Cumulés
0	0%
1000	20%
2000	27%
3000	63%
4000	91%
5000	95%
6000	100%



Sur EXCEL, il faut ajouter une ligne qui représente la valeur 0 du début de la courbe pour une présentation correcte!

Exercise 12:

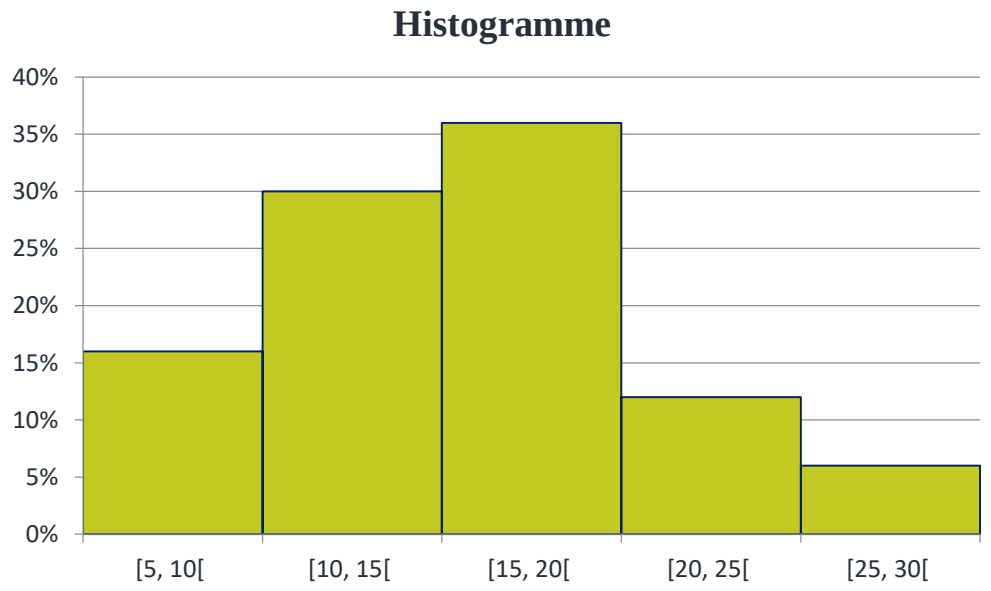
Take again the Excel Sheet Data5 containing time to answer 50 incoming calls.

1. Construct a percentage histogram and a percentage polygon.

2. Construct a cumulative percentage polygon.

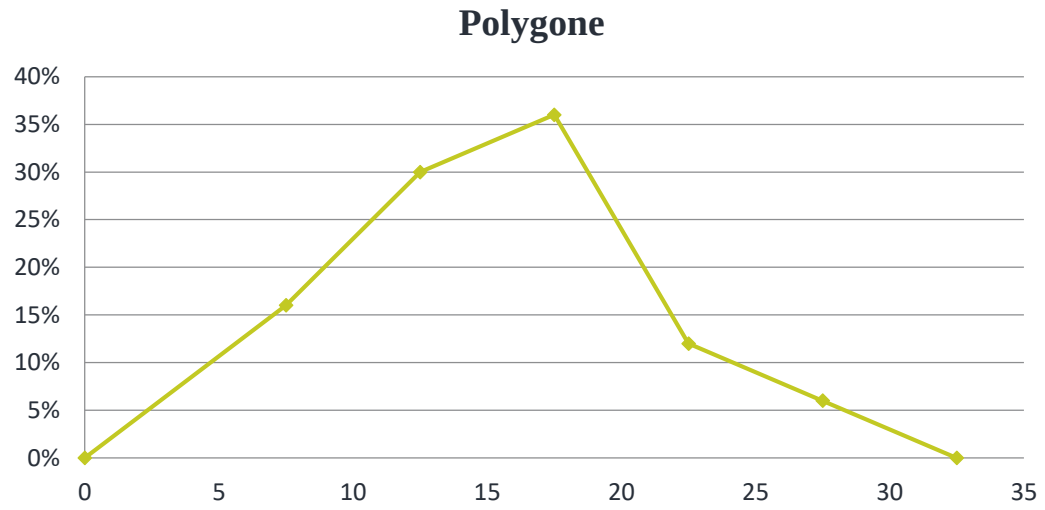
3. The service target level is set at “80% of calls is answered within 20 seconds”. What do you conclude about call center performance ?

Variables	Frequency	Proportion (relative frequency)	Percentage
[5, 10[	8	0,16	16%
[10, 15[	15	0,3	30%
[15, 20[	18	0,36	36%
[20, 25[	6	0,12	12%
[25, 30[	3	0,06	6%
Total	50	1	100%

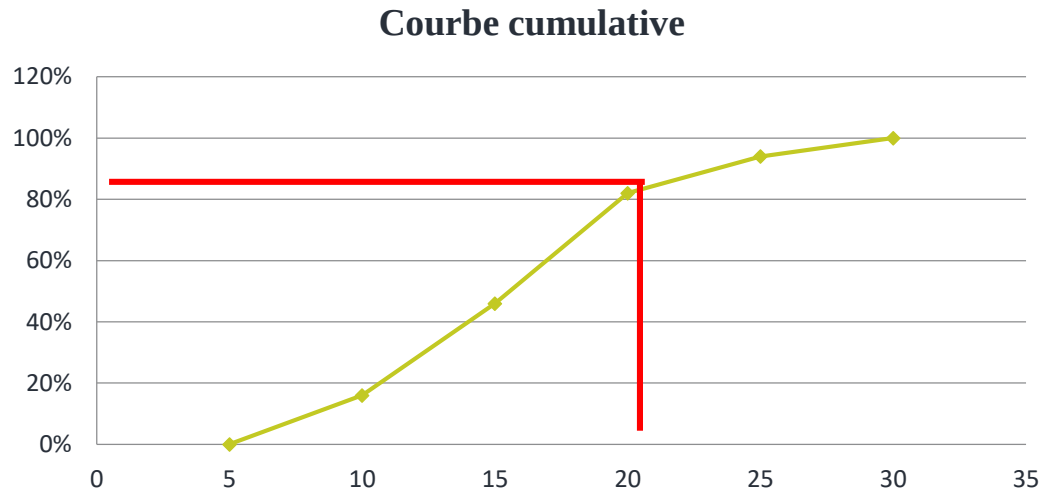


Statistiques Descriptives Unidimensionnels: Distribution à un seul caractère (variable)

Présentation nécessaire pour la création d'un polygone			
B.Inf	**B.Sup**	**Centre de Classe**	**Percentage**
		0	0%
5	10	7,5	16%
10	15	12,5	30%
15	20	17,5	36%
20	25	22,5	12%
25	30	27,5	6%
30	35	32,5	0%



Présentation nécessaire pour la création d'une courbe cumulative	
B.Sup	**P Cumulés**
5	0%
10	16%
15	46%
20	82%
25	94%
30	100%

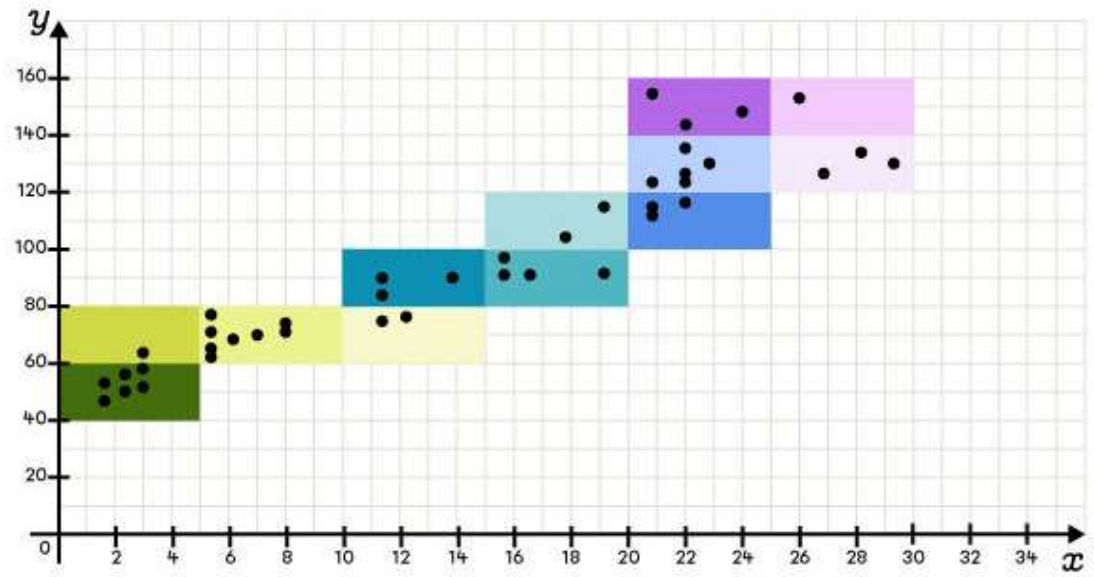


L'objectif cible de ce service est fixé à "80 % des appels doivent être répondus dans les 20 secondes". Nous pouvons conclure que l'objectif est atteint puisque la courbe des pourcentages cumulative prouve que **82% des appels sont répondus 20 secondes.**

Représentations graphiques

Les distributions à deux caractères pour un **tableau de contingence** peuvent être représentées graphiquement sous forme de **nuage de points** dans un plan suivant **un repère cartésien**. Chaque couple d'observation sera représenté par un point. La forme obtenu par ces nuages offre une information sur le type d'une éventuelle liaison.

x \ y	[40,60[	[60,80[	[80,100[	[100,120[	[120,140[	[140,160[
[0,5[	6	1	0	0	0	0
[5,10[	0	8	0	0	0	0
[10,15[	0	2	3	0	0	0
[15,20[	0	0	4	2	0	0
[20,25[	0	0	0	3	5	3
[25,30[	0	0	0	0	3	1



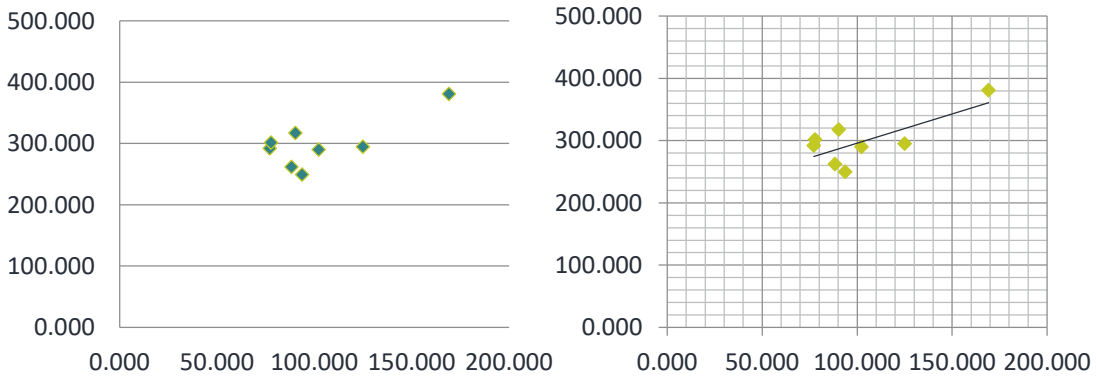
Représentation sous forme de nuage de points

Représentations graphiques

2 Variables Quantitatives

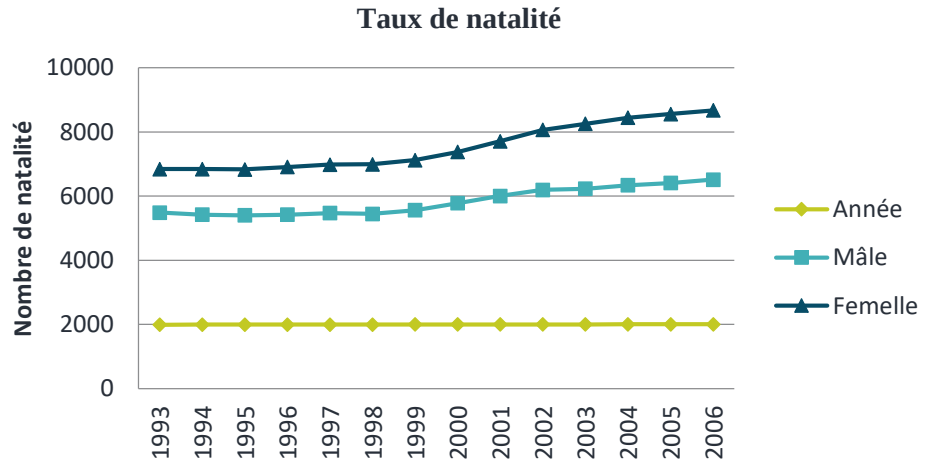
« Diagramme de dispersion » ou « nuage de points » (Scatter plot)

Il s'agit d'un graphique qui permet de représenter les relations entre deux variables quantitatives en utilisant des points placés sur un plan cartésien, où chaque axe correspond à une des variables étudiées. Le nuage de points permet de visualiser la répartition des données ainsi que la présence ou l'absence de corrélations entre les variables.



« Graphique temporel » ou « Graphique de série chronologique » (Time series plot)

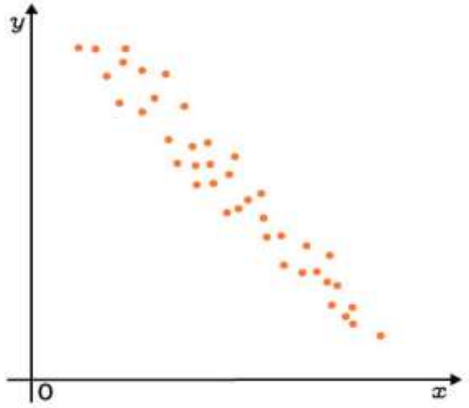
Il s'agit d'un graphique qui permet de représenter l'évolution d'une variable en fonction du temps, en plaçant les valeurs de la variable sur l'axe vertical et les dates ou périodes sur l'axe horizontal. Les graphiques de série chronologique permettent de visualiser les tendances, les variations saisonnières et les évolutions à long terme des données temporelles.



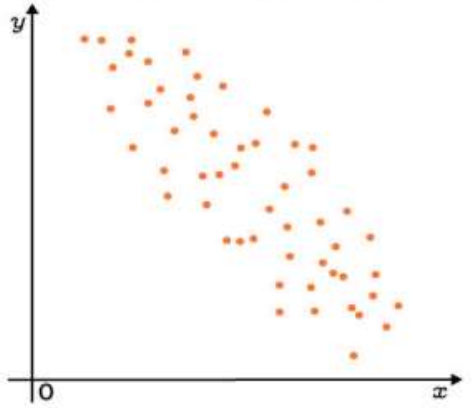
Représentations graphiques

Corrélations linéaires négatives

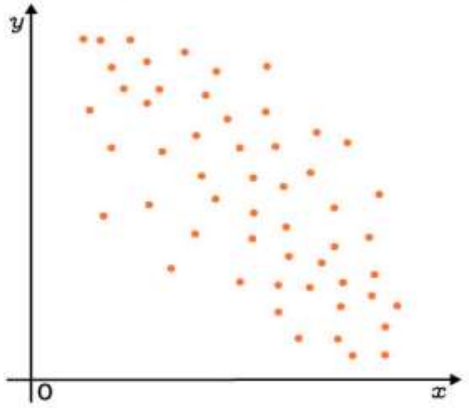
Corrélation linéaire forte



Corrélation linéaire moyenne

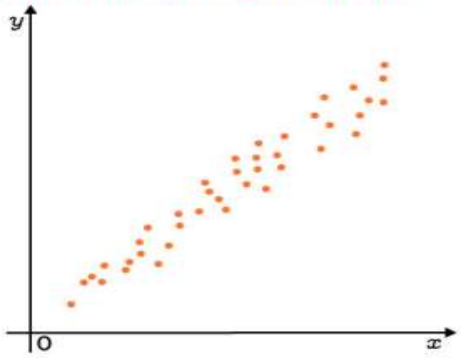


Corrélation linéaire faible

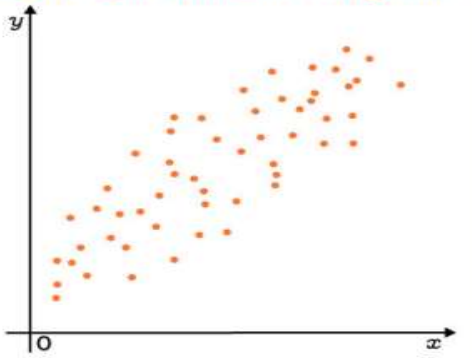


Corrélations linéaires positives

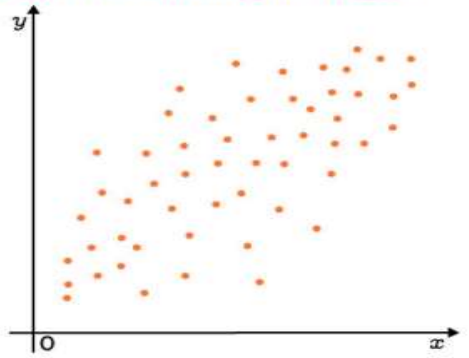
Corrélation linéaire forte



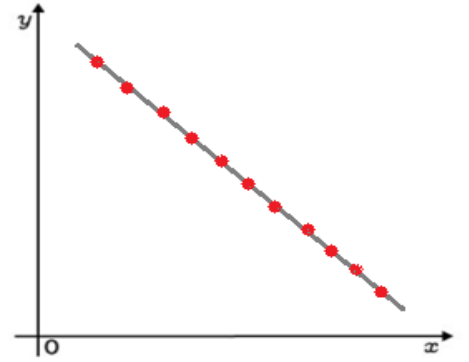
Corrélation linéaire moyenne



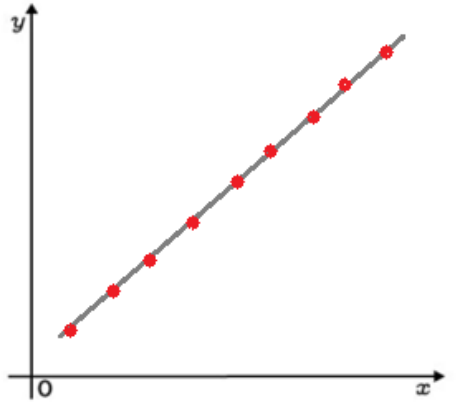
Corrélation linéaire faible



Corrélation linéaire négative parfaite



Corrélation linéaire positive parfaite



Statistiques Descriptives Bidimensionnels: Distribution à 2 variables

Exercise 16:

Movies companies need to predict the gross receipts of individual movies once a movie has debutes. Table in Data10 gives the first weekend gross, the US gross, and the worldwide gross (in $millions) of the eigth Harry Potter movies.

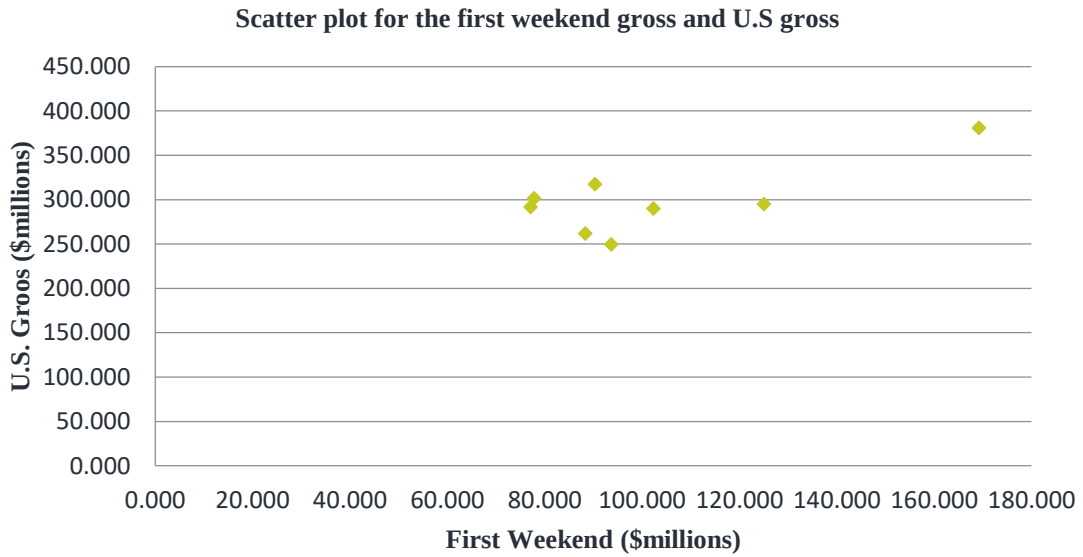
1. Construct a scatter plot with first weekend gross on the X-axis and U.S gross on the Y-axis.
2. Construct a scatter plot with first weekend gross on the X-axis and worldwide gross on the Y-axis.
3. What can you say about the relationship between first weekend gross and U.S gross, and between first weekend gross and worldwide gross.

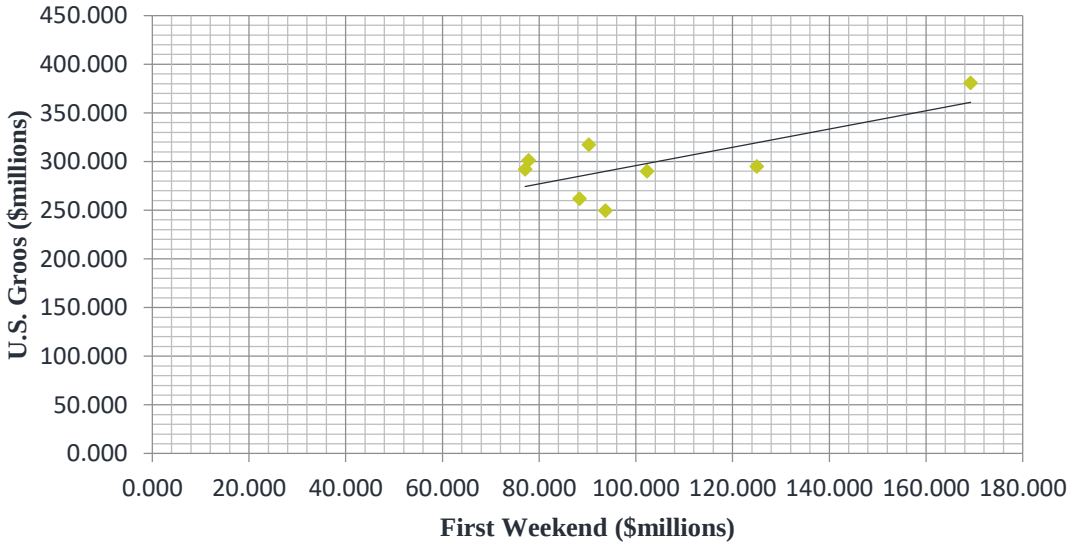
Title	First Weekend ($millions)	U.S. Groos ($millions)	World-wide Gross ($millions)
Sorcerer's Stone	90.295	317.558	976.458
Chamber of Secrets	88.357	261.988	878.988
Prisoner of Azkaban	93.687	249.539	795.539
Goblet of Fire	102.335	290.013	896.013
Order of the Phoenix	77.108	292.005	938.469
Half-Blood Prince	77.836	301.460	934.601
Deathly Hallows Part I	125.017	295.001	955.417
Deathly Hallows Part II	169.189	381.011	1,328.111

Statistiques Descriptives Bidimensionnels: Distribution à 2 variables

1. Construct a scatter plot with first weekend gross on the X-axis and U.S gross on the Y-axis.

Title	First Weekend ($millions)	U.S. Groos ($millions)
Goblet of Fire	102,335	290,013
Deathly Hallows Part I	125,017	295,001
Deathly Hallows Part II	169,189	381,011
Order of the Phoenix	77,108	292,005
Half-Blood Prince	77,836	301,460
Chamber of Secrets	88,357	261,988
Sorcerer's Stone	90,295	317,558
Prisoner of Azkaban	93,687	249,539

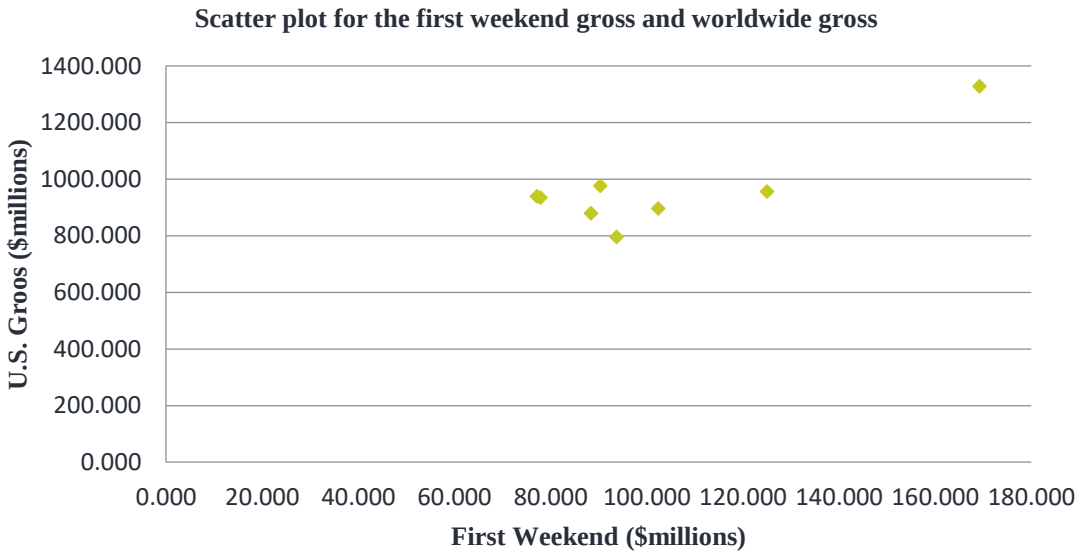


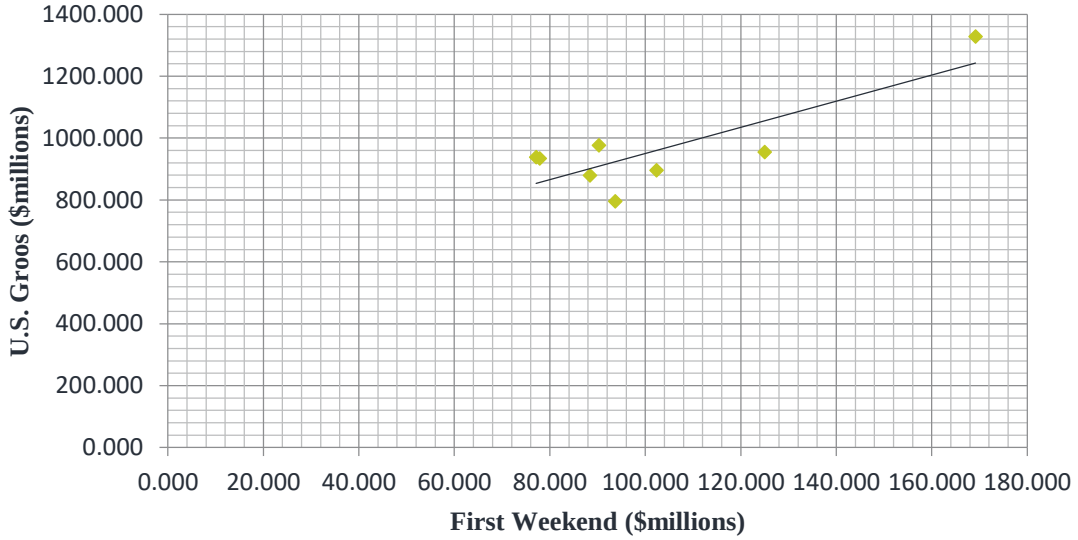


Statistiques Descriptives Bidimensionnels: Distribution à 2 variables

2. Construct a scatter plot with first weekend gross on the X-axis and worldwide gross on the Y-axis.

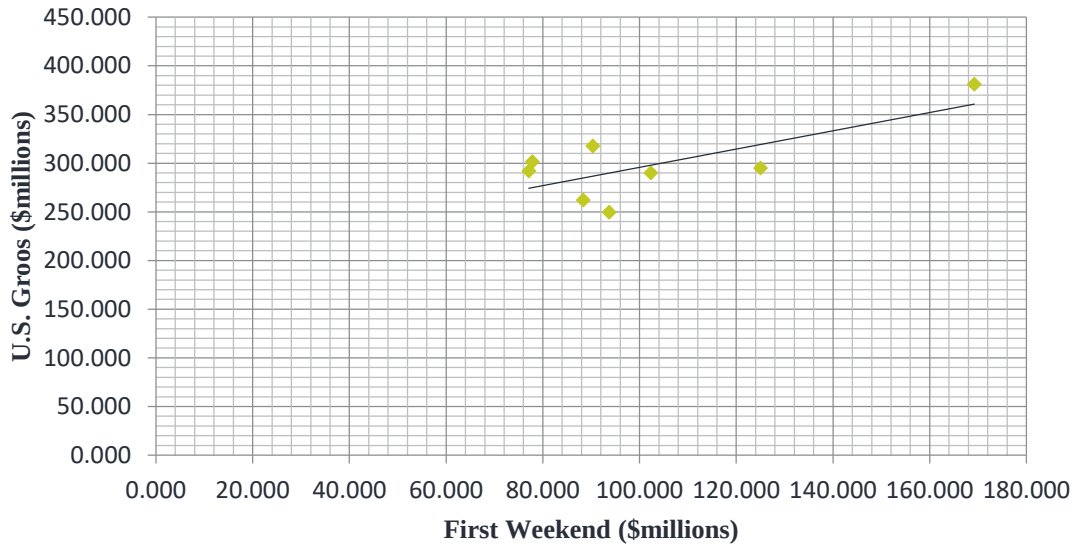
Title	First Weekend ($millions)	World-wide Gross ($millions)
Goblet of Fire	102,335	896,013
Deathly Hallows Part I	125,017	955,417
Deathly Hallows Part II	169,189	1328,111
Order of the Phoenix	77,108	938,469
Half-Blood Prince	77,836	934,601
Chamber of Secrets	88,357	878,988
Sorcerer's Stone	90,295	976,458
Prisoner of Azkaban	93,687	795,539

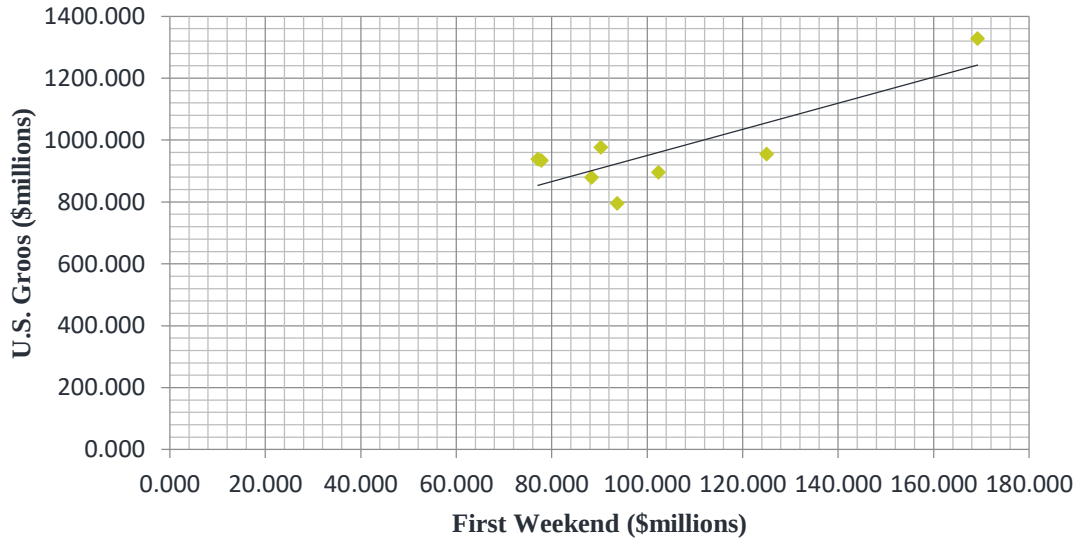




Statistiques Descriptives Bidimensionnels: Distribution à 2 variables

3. What can you say about the relationship between first weekend gross and U.S gross, and between first weekend gross and worldwide gross.





D'après les deux **Diagrammes de dispersions**, il est clair qu'il existe une liaison entre les recettes brutes des 8 films de Harry Potter pour les premier week-end et les recettes des États-Unis liées à ces films. Ceci se traduit par une **corrélation positive forte entre les 2 variables**. ***Autrement dit, la hausse des recettes brutes des 8 films de Harry Potter pour les premier week-end entraines l'augmentation des recettes des États-Unis liées à ces films.*** La même interprétation est valable pour les 2 variables : les recettes brutes des 8 films de Harry Potter pour les premier week-end et les recettes du monde entier liées à ces films.

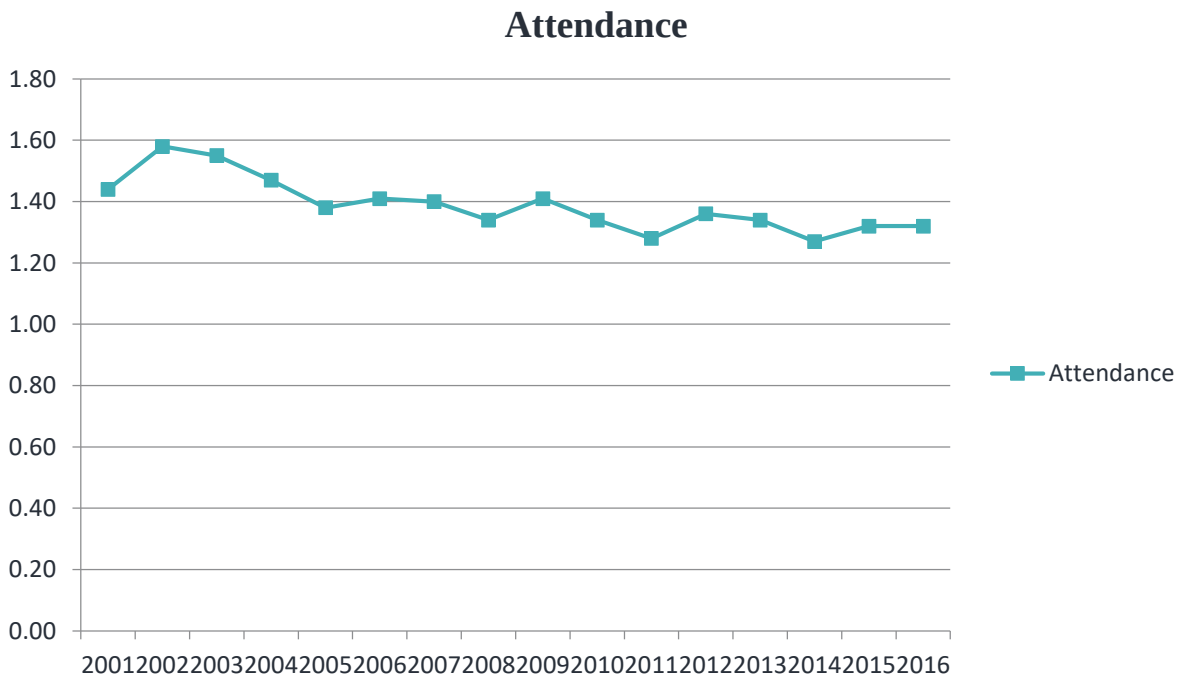
Statistiques Descriptives Bidimensionnels: Distribution à 2 variables

Exercise 18:

The table in Excel Sheet Data13 contains the yearly movie attendance (in billions) from 2001 to 2016.

1. Construct a time-series plot for the movie attendance (in billions).

2. What pattern, if any, is present in the data?



D'après la présentation graphique, on remarque que la fréquentation au film en question a vécue des hausses et des basses d'une années à une autre. Cependant, à partir de l'année 2002, la fréquentation au film en question est entrain de se dégrader. Ceci peut être dû à plusieurs raisons comme la sortie d'autres films qui offre une concurrence plus féroce. En plus, la qualité du film lui-même peut être un facteur important. Si le film est mal reçu par les critiques ou par le public, cela peut entraîner une baisse de fréquentation d'une année à l'autre….