STATISTIQUES DESCRIPTIVES ET INFÉRENTIELLES

Résumé du cours (Lecture II) + 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

Indicateurs statistiques

Indicateurs de tendance centrale (Central Tendency)

La moyenne arithmétique (The Mean)

La moyenne géométrique

Le Mode (The Mode)

La médiane (The Median)

Indicateurs de dispersion (Variation)

L'étendue (The Range)

La variance et l'écart type

(The Variance and the Standard Deviation)

Le coefficient de variation

(The Coefficient of Variation)

Les quantiles & les centiles

L'écart interquartile

Indicateurs de forme (The Shape)

L'Asymétrie (The Skewness)

Aplatissement (The Kurtosi)

Indicateurs de tendance centrale

La moyenne arithmétique (The Mean)

La moyenne est alors donnée par la formule suivante :

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_i + \cdots + x_n}{n}$$

Exemple:

On veut effectuer une étude sur les heures supplémentaires pour un échantillon de 20 travailleurs qui a montré qu'ils ont effectué le nombre d'heures supplémentaires suivant pour le mois dernier:

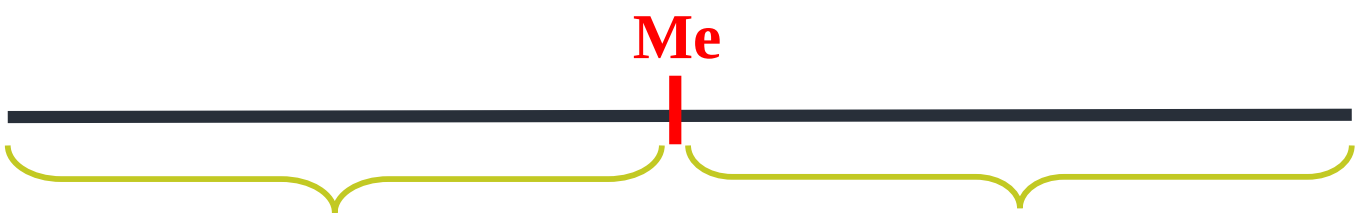
13	13	12	15	6	8	10	12	12	11
12	11	14	8	7	10	12	11	9	8

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{13+13+12+15+6+8+10+12+12+11+12+11+14+8+7+10+12+11+9+8}{20} = \frac{214}{20} = 10{,}7$$

Indicateurs de tendance centrale

La Médiane (The Median)

La médiane, notée **Me** est la valeur qui partage les éléments d'une série numérique en deux parties égales.



50 %
des données sont **inférieures ou égales** à la valeur de la médiane

50 %
des données sont **supérieures ou égales** à la valeur de la médiane

Indicateurs de tendance centrale

La Médiane (The Median)

Règle 1 : Si l'ensemble de données contient ***un nombre impair de valeurs (odd number of values)***, la médiane est la mesure associée à la valeur classée au milieu. C'est-à-dire la médiane est la valeur de l'observation numéro (n+1)/2.

Exemple pour une séries non groupées dont l'effectif est impair :

Soit les notes de **9** étudiants dans un examen de Math: 14 – 15 – 9 – 18 – 15 – 13 – 10 – 14 – 9

On trie la série dans l'ordre croissant: 9 – 9 – 10 – 13 – 14 – 14 – 15 – 15 – 18

Ici (n+1)/2 = (9+1)/2 = 10/2 = 5, donc la médiane est la valeur de l'observation n° 5 qui est égale à **Me = 14**.

Règle 2 : Si l'ensemble de données contient ***un nombre pair de valeurs (even number of values)***, la médiane est la moyenne des valeurs des deux observations n/2 , et (n/2)+1.

Exemple pour une séries non groupées dont l'effectif est pair :

Soit les notes de **10** étudiants dans un examen de Math: 14 – 15 – 9 – 18 – 15 – 13 – 13 – 10 – 14 – 9

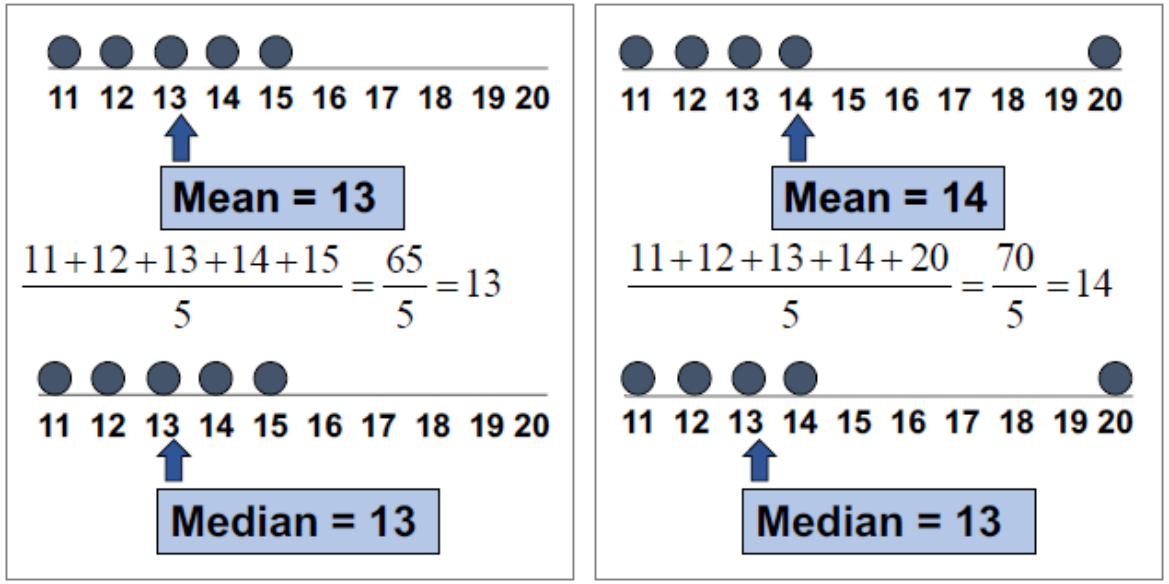
On trie la série dans l'ordre croissant: 9 – 9 – 10 – 13 – 13 – 14 – 14 – 15 – 15 – 18

Ici la valeur de l'observation n° (n/2 = 5) est 13, et la valeur de l'observation n° ((n/2 + 1) = 6) est 14. Donc **Me =** $\frac{13+14}{2}$ **= 13,5** .

Indicateurs de tendance centrale

Moyenne VS Médiane

On peut dire que ***la médiane est moins sensible que la moyenne aux valeurs extrêmes***.



En calculant la moyenne, si une ou plusieurs valeurs sont très grandes ou très petites par rapport aux autres valeurs de l'ensemble de données, la moyenne ***peut être grandement affectée***, car elle est sensible aux valeurs extrêmes. En revanche, la médiane est la valeur qui se trouve au milieu d'un ensemble de données trié par ordre croissant ou décroissant. Elle n'est pas affectée par les valeurs extrêmes car elle ne prend en compte que la valeur centrale de l'ensemble de données. Par conséquent, lorsque on analyse des données avec des valeurs extrêmes, ***la médiane est souvent préférée à la moyenne***, car elle donne une mesure ***plus stable et plus représentative*** de la valeur centrale de l'ensemble de données.

Indicateurs de tendance centrale

Le Mode (The Mode)

Le Mode, noté ***Mo***, est une mesure de tendance centrale qui **représente la valeur la plus fréquente** dans un ensemble de données. Autrement dit, c'est la valeur qui apparaît le plus souvent dans un ensemble de données.

Cas de variable quantitative discrète (séries groupées par valeurs):

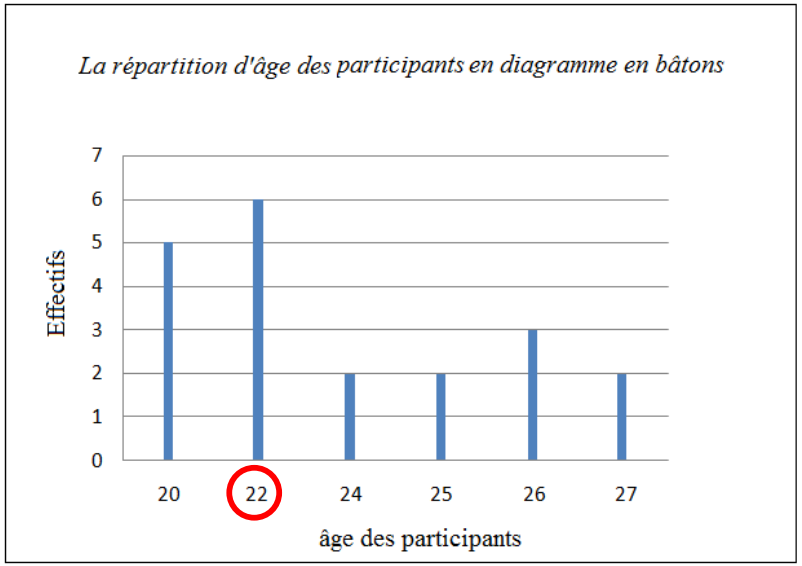
On considère la distribution d'âges des participants suivante:

La modalité 22 correspond à l'effectif le plus élevé. Il s'agit de la modalité la **plus fréquente**. Graphiquement, le Mode **est l'abscisse du plus haut bâton**.

Donc: Mode = 22

La majorité des participants sont âgés de 22 ans.

x_i	n_i
20	5
22	6
24	2
25	2
26	3
27	2
Total	20



Indicateurs de tendance centrale

Le Mode (The Mode)

Il est important de noter que dans certains ensembles de données, il peut y avoir plusieurs modes, c'est-à-dire deux ou plusieurs valeurs qui apparaissent avec la même fréquence maximale. Dans ce cas, on parle respectivement de distribution **bimodale**, et **multimodale**.

Exemple:

On suppose qu'il existe deux modalités ayant l'effectif le plus élevé ! Dans ce cas, nous avons une distribution statistique **bimodal**, c'est-à-dire qu'il y a deux modes.

x_i	n_i
20	**5**
22	**6**
24	6
25	**2**
26	3
27	2
Total	**20**

Mode = 22 et 24

Indicateurs de tendance centrale

La Moyenne Géométrique (The Geometric Mean)

La moyenne géométrique est une mesure utilisée pour calculer la moyenne des pourcentages, des ratios, des indices ou des taux de croissance. Elle est largement appliquée dans les domaines de la finance et de l'économie, où l'on cherche souvent à trouver les ***variations en pourcentage*** des ventes, des salaires, ou des indicateurs économiques tels que le produit intérieur brut. La formule de la moyenne géométrique pour un ensemble de *n* nombres positifs consiste à prendre la N ^{ième} racine du produit des *n* valeurs comme suit:

$$G = \left[\prod_{i=1}^{n} x_i\right]^{\frac{1}{n}} = (x_1 \times x_2 \times \ldots \times x_i \times \ldots \times x_n)^{\frac{1}{n}} = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_i \times \ldots \times x_n}$$

La moyenne géométrique sera être toujours **inférieure ou égale** **mais** **jamais supérieure à la moyenne arithmétique**. De plus, toutes les valeurs de données doivent être positives.

Indicateurs de tendance centrale

La Moyenne Géométrique (The Geometric Mean)

Exemple:

Supposons que vous ayez un terrain d'une superficie de 1 hectare que vous souhaitez diviser en deux parties pour y planter des cultures différentes. Vous voulez que la superficie des deux parties soit égale. Pour ce faire, vous pouvez utiliser la moyenne géométrique pour calculer la taille de chaque partie.

$$G = \sqrt[1]{1 \times 0{,}5} = 0{,}707$$

Cela signifie que chaque partie du terrain doit avoir une superficie de 0,707 hectares pour que les deux parties soient égales. La moyenne géométrique est utilisée dans ce cas pour obtenir une solution équitable en prenant en compte la relation exponentielle entre la superficie totale du terrain et la superficie de chaque partie.

Indicateurs de tendance centrale

La Moyenne Géométrique (The Geometric Mean)

Pourquoi on ne le divise pas la superficie par 2 tout simplement?

En effet, si on divise 1 hectare par 2 pour obtenir la superficie de chaque partie du terrain, ceci donnerait une superficie de 0,5 hectare pour chaque partie. Cependant, si le terrain n'est pas parfaitement carré ou rectangulaire, les deux parties ne seront pas nécessairement égales.

La moyenne géométrique est particulièrement utile lorsque ***les données sont liées de manière exponentielle***, comme c'est le cas ici. En effet, si nous voulions diviser le terrain en trois parties égales, il ne serait pas possible d'utiliser la simple division de la superficie totale par le nombre de parties, car cela ne tiendrait pas compte de la relation exponentielle entre la superficie totale et la superficie de chaque partie. Dans ce cas, nous devrions utiliser la moyenne géométrique pour obtenir une solution juste et précise. Cette méthode est donc plus précise et plus juste que la simple division de la superficie totale par le nombre de parties.

Indicateurs de tendance centrale

La Moyenne Géométrique (The Geometric Mean)

Dans ce cas, la relation exponentielle est liée à la racine carrée, car si l'on divise le terrain en deux parties égales, **la superficie de chaque partie** doit être égale à **la racine carrée de la superficie totale**. Dans ce cas, la racine carrée de 1 hectare est de 1, et la racine carrée de la superficie de chaque partie est de 0,707 hectares. En multipliant la superficie de chaque partie par elle-même, on obtient une superficie totale de 1 hectare.

Ainsi, en utilisant la moyenne géométrique, on obtient deux parties égales dont la superficie totale est égale à la superficie totale du terrain initial. C'est pour cette raison que la moyenne géométrique est utilisée dans ce cas pour obtenir une solution équitable et précise.

$$G = \sqrt[1]{1 \times 0{,}5} = 0{,}707$$

1 hectare

$$1 \text{ hectare} = (0{,}707)^2 + (0{,}707)^2$$

$$= 0{,}5 + 0{,}5$$

Indicateurs de tendance centrale

La Moyenne Géométrique (The Geometric Mean)

La moyenne géométrique peut également être utilisée pour calculer le ***taux de rendement moyen sur plusieurs périodes de temps***, en prenant en compte la capitalisation des intérêts. Si nous avons une série de rendements R_i sur *n* périodes de temps, alors le taux de rendement moyen géométrique est calculé de la manière suivante :

$$\bar{R}_G = [(1 + R_1) \times (1 + R_2) \times \ldots \times (1 + R_n)]^{\frac{1}{n}} - 1$$

Cette formule est utilisée pour calculer le rendement financier moyen sur une période donnée. Elle prend en compte les rendements individuels sur chaque période, en les combinant de manière géométrique pour déterminer le rendement moyen sur la période en question. Ceci permet de mesurer le rendement d'un investissement financier sur plusieurs périodes de temps. Notez que cette formule suppose que les rendements sont ***réinvestis***, c'est-à-dire que l'argent gagné à chaque période est ajouté à l'investissement initial et génère des rendements supplémentaires au cours des périodes suivantes. ***Si les rendements ne sont pas réinvestis, alors la moyenne arithmétique serait une mesure plus appropriée du rendement moyen sur la période***.

Indicateurs de tendance centrale

La Moyenne Géométrique (The Geometric Mean)

Exemple:

Supposons que vous avez investi 1000 € dans un fonds commun de placement qui a généré les rendements suivants au cours des **cinq dernières années** : 10%, 5%, 12%, -2% et 8%. Vous voulez calculer le taux de rendement moyen sur cette période en utilisant la formule du taux de rendement moyen géométrique.

$$\bar{R}_G = [(1+0{,}10) \times (1+0{,}05) \times (1+0{,}12) \times (1-0{,}02) \times (1+0{,}08)]^{\frac{1}{5}} - 1 = (1{,}353)^{\frac{1}{5}} - 1 = 0{,}065$$

Ainsi, le taux de rendement moyen géométrique pour cet investissement est de 6,5% sur la période de cinq ans. Cette mesure permet de prendre en compte la capitalisation des intérêts et donne une idée de la performance globale de l'investissement sur la période considérée.

Exercise 2:

Table below gives the total rate of return **percentage** for the Dow Jones Industrial Average (DJIA), the Standard Poor's 500 (SP 500) and the technology-heavy NASDAQ composite (NASDAQ) from 2013 through 2016.

Year	DJIA	S&P 500	NASDAQ
2013	26.5	29.6	28.3
2014	7.5	11.4	13.4
2015	-2.2	-0.7	5.7
2016	13.4	9.5	7.5

1. Compute the ***geometric mean rate of return per year*** for the DJIA, SP 500, and NASDAQ from 2013 through 2016.
2. What conclusions can you reach concerning the geometric mean rate of return of these three market indices ?

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

Year	DJIA	S&P 500	NASDAQ
2013	0,265	0,296	0,283
2014	0,075	0,114	0,134
2015	-0,022	-0,007	0,057
2016	0,134	0,095	0,075

On applique la formule suivante:

$$\bar{R}_G = [(1 + R_1) \times (1 + R_2) \times \ldots \times (1 + R_n)]^{\frac{1}{n}} - 1$$

1. Compute the geometric mean rate of return per year for the DJIA, SP 500, and NASDAQ from 2013 through 2016.

$$\bar{R}_G(DJIA) = [(1 + 0{,}265) \times (1 + 0{,}075) \times (1 + (-0{,}022)) \times (1 + 0{,}134)]^{\frac{1}{4}} - 1 = (1{,}508)^{\frac{1}{4}} - 1 = 0{,}108 = 10{,}8\%$$

$$\bar{R}_G(S\&P500) = [(1 + 0{,}296) \times (1 + 0{,}114) \times (1 + (-0{,}007)) \times (1 + 0{,}095)]^{\frac{1}{4}} - 1 = (1{,}5698)^{\frac{1}{4}} - 1 = 0{,}119 = 11{,}9\%$$

$$\bar{R}_G(NASDAQ) = [(1 + 0{,}283) \times (1 + 0{,}134) \times (1 + (0{,}057)) \times (1 + 0{,}075)]^{\frac{1}{4}} - 1 = (1{,}65319)^{\frac{1}{4}} - 1 = 0{,}133 = 13{,}3\%$$

2. What conclusions can you reach concerning the geometric mean rate of return of these three market indices ?

On remarque que le taux de rendement moyen géométrique pour l'indice DJIA qui est de 10,8% sur la période de 4 ans est le moins élevé par rapport aux autres. Alors que le taux le plus élevé est celui de l'indice S&P 500 avec 11,9%. Alors, on peut conclure que la performance globale de l'investissement pour l'indice NASDAQ est le plus important sur la période considérée.

Indicateurs de dispersion (Variation)

L'étendue (The Range)

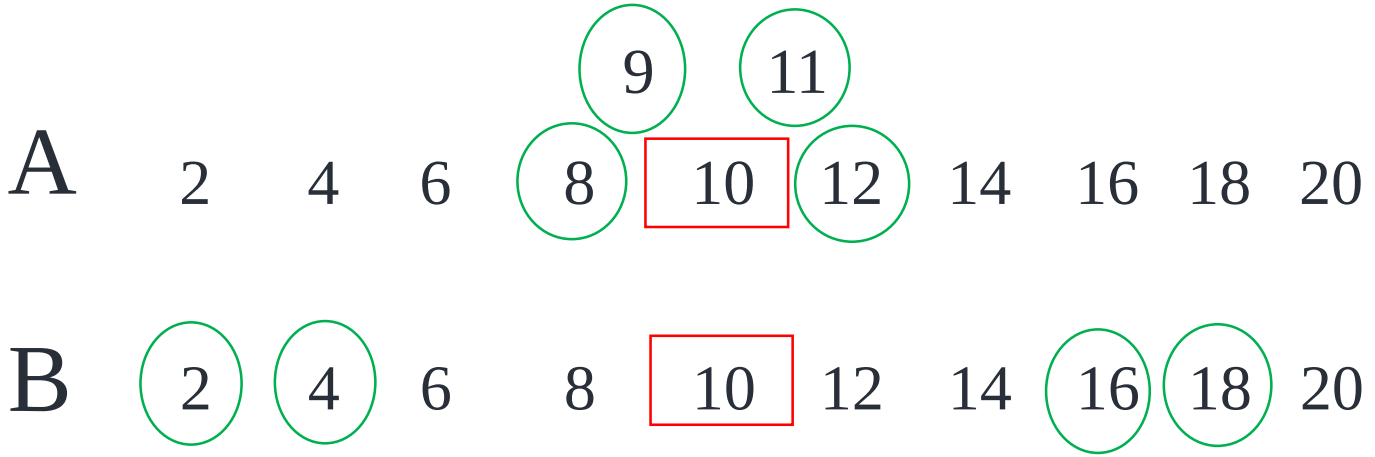
L'étendue est une mesure de dispersion qui indique la différence entre la plus grande et la plus petite valeur d'un ensemble de données.

Exemple:

Soit deux élèves dont les notes dans quatre matières ont été les suivantes : Élève A : $\{8, 9, 11, 12\}$; Élève B : $\{2, 4, 16, 18\}$

L'étendue des notes de A est $(12 - 8 = 4)$, tandis que l'étendue des notes de B est $(18 - 2 = 16)$.

On notera pourtant que la moyenne des deux élèves est de 10. Mais **B** a des notes beaucoup **plus dispersées** que **A**.



Ainsi, l'étendue est une mesure simple qui donne une idée de la ***dispersion des données***, mais elle peut être influencée par des valeurs extrêmes et ne prend pas en compte la répartition des données à l'intérieur de l'intervalle défini par l'étendue.

Indicateurs de dispersion (Variation)

La variance et l'écart type (Variance and Standard Deviation)

La ***variance*** est une mesure de la dispersion des valeurs d'une variable autour de sa moyenne. Elle est calculée en prenant la moyenne des carrés des écarts de chaque observation par rapport à la moyenne de l'échantillon:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

L'écart-type est une mesure de la dispersion des valeurs d'une variable autour de sa moyenne. Il est calculé comme la racine carrée de la variance:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Remarque: Comme la somme des carrés ne peut jamais être une valeur négative, la variance et l'écart-type seront toujours **des valeurs positives**. La variance et l'écart-type seront nuls, ***c'est-à-dire qu'il n'y a pas de variation***, uniquement dans le ***cas particulier*** lorsque *toutes les observations d'un échantillon ont la même valeur*.

Indicateurs de dispersion (Variation)

La variance et l'écart type (Variance and Standard Deviation)

Exemple:

Nous souhaitons calculer la variance et l'écart type de l'échantillon de 10 observations suivant:

39 – 29 – 43 – 52 – 39 – 44 – 40 – 31 – 44 – 35

1) Nous commençons le calcul de la moyenne de la série:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{39+29+43+52+39+44+40+31+44+35}{10} = 39{,}6$$

2) Nous préparons pour chaque valeur dans la série la valeur des $(x_i - \bar{X})$, et $(x_i - \bar{X})^2$:

3) D'après le tableau nous avons besoins maintenant de la somme des $(x_i - \bar{X})^2$ pour calculer la variance:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{(39-39{,}6)^2+(29-39{,}6)^2+(43-39{,}6)^2+\cdots+(35-39{,}6)^2}{10-1} = \frac{412{,}40}{10-1} = 45{,}82$$

4) Nous calculons maintenant l'écart type:

$$s = \sqrt{s^2} = \sqrt{45{,}82} = 6{,}77$$

x_i	$(x_i - \bar{X})$	$(x_i - \bar{X})^2$
39	-0,60	0,36
29	-10,60	112,36
43	3,40	11,56
52	12,40	153,76
39	-0,60	0,36
44	4,40	19,36
40	0,40	0,16
31	-8,60	73,96
44	4,40	19,36
35	-4,60	21,16
Total		**412,40**

Indicateurs de dispersion (Variation)

Le coefficient de variation (coefficient of variation)

Un ***coefficient de variation*** faible indique une faible dispersion relative et une plus grande homogénéité dans la distribution des données, tandis qu'un coefficient de variation élevé indique une plus grande dispersion relative et une plus grande hétérogénéité dans la distribution des données. Le CV peut varier de 0% à l'infini. En général, un CV élevé indique une dispersion élevée. Cependant, il est important de considérer le contexte de l'étude pour interpréter les résultats de manière appropriée.

$$CV = \frac{S}{\bar{X}} \times 100$$

Exemple:

Nous possédons quelques paramètres pour les salaires deux entreprises, et on souhaitons comparer la dispersion de leurs distributions. Calculez le coefficient de variation:

$$CV_1 = \frac{3{,}60}{15} = 0{,}24 = 24\%$$

$$CV_2 = \frac{3{,}16}{20} = 0{,}158 = 15{,}8\%$$

Données	
Entreprise 1	Entreprise 2
$\bar{X}$ = 15 S^2 = 13	$\bar{X}$ = 20 S^2 = 10
Calculs	
Ecart-type = 3,60 CV1 = 24 %	Ecart-type = 3,16 CV2 = 15,8 %

➔ CV1 > CV2

➔ L'entreprise 1 possède des salaires plus dispersés que la 2ème entreprise.

➔ L'entreprise 2 possède des salaires plus homogène que la 1ère entreprise.

Indicateurs de dispersion (Variation)

Les quartiles (Quartiles)

Les quartiles sont des mesures statistiques qui divisent une distribution en quatre parties égales. Ils sont utilisés pour décrire la dispersion des données d'une variable numérique. Voici les règles générales pour calculer les quartiles pour des variables numériques discrètes :

- **1er Quartile (Q1) :** Le premier quartile est la valeur qui divise la distribution en deux parties, la première partie contenant **25%** des données. Pour calculer Q1, on peut classer les données par ordre croissant et trouver la valeur qui est à **la position (n+1)/4** dans la série de données, où n est le nombre total de données.
- **2ème Quartile (Q2) :** Le deuxième quartile est la médiane de la distribution, qui divise la distribution en deux parties égales, chacune contenant **50%** des données. Pour calculer Q2, on peut classer les données par ordre croissant et trouver la valeur qui est à **la position (n+1)/2** dans la série de données (**Médiane**).
- **3ème Quartile (Q3) :** Le troisième quartile est la valeur qui divise la distribution en deux parties, la première partie contenant **75%** des données. Pour calculer Q3, on peut classer les données par ordre croissant et trouver la valeur qui est à **la position 3(n+1)/4** dans la série de données.

Indicateurs de dispersion (Variation)

Les quartiles (Quartiles)

- **Règle 1:** Si le rang de la donnée est un nombre entier, le quartile correspondant est égal à la mesure qui correspond à ce rang.
- **Règle 2:** Si le rang de la donnée est un nombre décimal demi (2,5; 4,5; etc.), le quartile correspondant est égal à la moyenne des mesures correspondant aux deux rangs impliqués.
- **Règle 3:** Si le rang de la donnée n'est ni un nombre entier ni un nombre décimal demi (exemple: 2,75; 3,25; etc.), nous arrondissons le résultat au nombre entier le plus proche, puis nous sélectionnons la mesure correspondant au rang.

Exemple :

Supposons que nous avons la distribution suivante pour le nombre d'enfants par famille : 1, 2, 2, 2, 3, 3, 4, 4, 4, 5.

Nous avons **10** données, donc :

Q1 = la valeur à la position (10+1)/4 = **2,5ème** position = la **moyenne des 2ème et 3ème valeurs** = (2+2)/2 = 2

Q2 = la valeur à la position (10+1)/2 = 5,5ème position = la moyenne des 5ème et 6ème valeurs = (3+3)/2 = 3

Q3 = la valeur à la position 3(10+1)/4 = **8,25ème** position = arrondissement à la **9ème** position = la 9ème valeur = 4

Indicateurs de dispersion (Variation)

Les quartiles (Quartiles)

Supposons que nous avons la distribution suivante pour le nombre d'enfants par famille : 2, 3, 3, 4, 4, 4, 5.

Nous avons **7** données, donc :

Q1 = la valeur à la position (7+1)/4 = **2ème position** = la **2ème valeur dans la série** = 3

Q2 = la valeur à la position (7+1)/2 = 4ème position = la 4ème valeur dans la série = 4

Q3 = la valeur à la position 3(7+1)/4 = 6ème position = la 6ème valeur dans la série = 4

Supposons que nous avons la distribution suivante pour le nombre d'enfants par famille : 2, 4, 6, 8, 10, 12, 14, 16

Nous avons **8** données, donc :

Q1 = la valeur à la position (8+1)/4 = 2,25ème position = arrondissement à la 3ème position = la 3ème valeur dans la série = 6

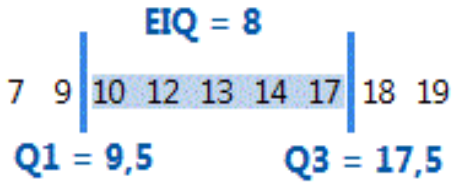
Q2 = la valeur à la position (8+1)/2 = 4,5ème position = la moyenne des 4ème et 5ème valeurs = (8+10)/2 = 9

Q3 = la valeur à la position 3(8+1)/4 = 6,75ème position = arrondissement à la 7ème position = la 7ème valeur dans la série = 14

Indicateurs de dispersion (Variation)

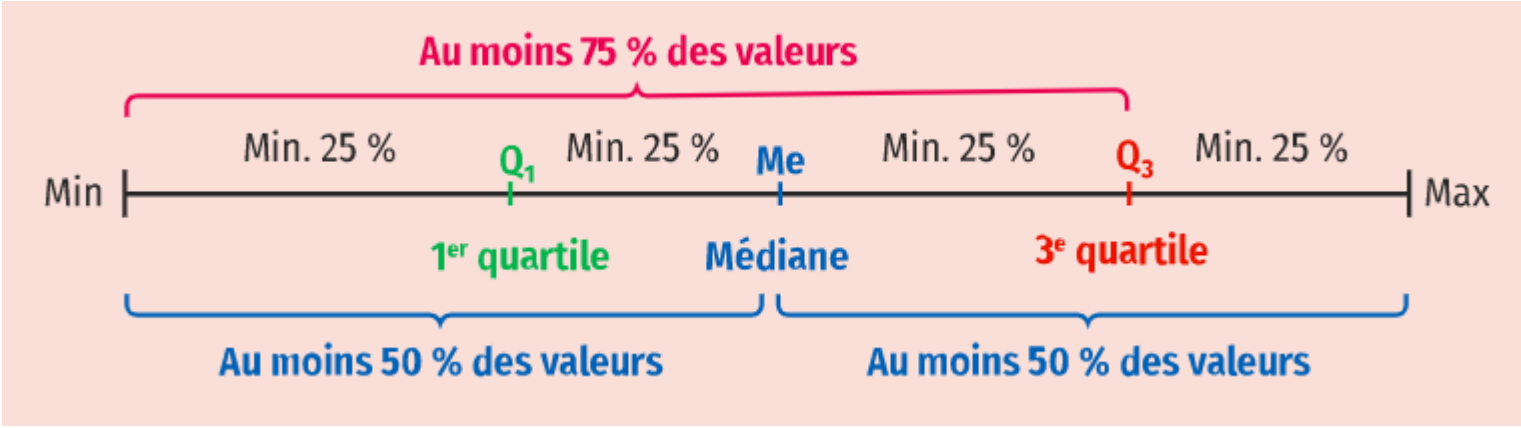
L'écart Interquartile (The Interquartile Range IQR)

L'écart interquartile (EI) appelé en anglais (the midspread) qui est la différence entre le troisième et le premier quartile (EI = Q3 - Q1) permettant de détecter des valeurs aberrantes ou des données extrêmes.



Pour ces données ordonnées, l'étendue interquartile est 8 (17,5 – 9,5 = 8). Donc, **50 % des données situées au milieu, se trouvent entre 9,5 et 17,5.**

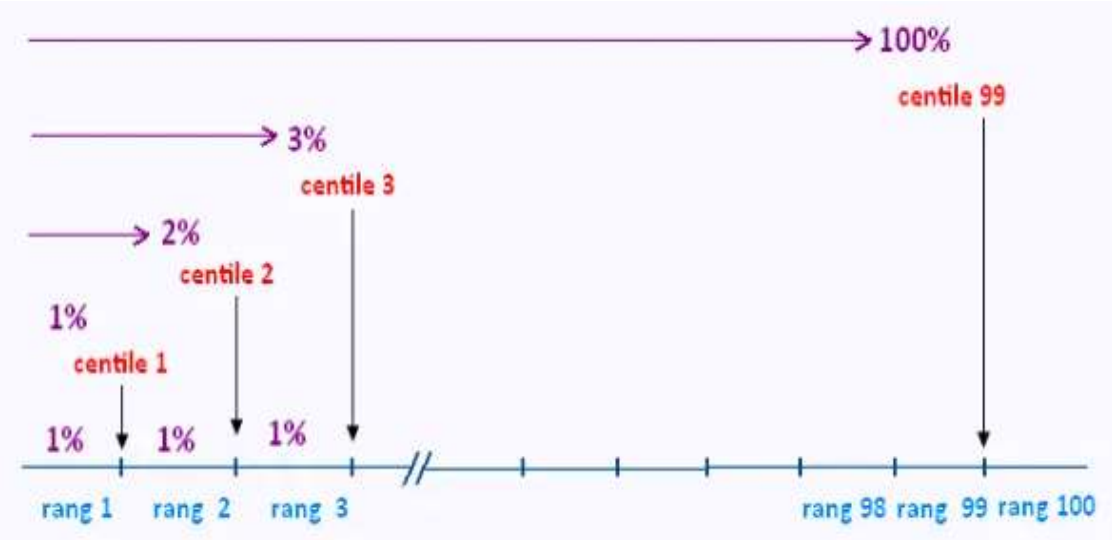
The Five-Number Summary



Indicateurs de dispersion (Variation)

Les centiles (The Percentiles)

Les centiles sont des mesures statistiques qui divisent une distribution de données en 100 parties égales. Dans une série statistique, le rang centile d'une donnée indique le pourcentage des données ayant une valeur inférieure ou égale à cette donnée. Par exemple, le percentile 90 d'une distribution de scores de test indique que 90% des scores se situent en dessous de cette valeur et seulement 10% des scores sont supérieurs.



Indicateurs de forme (The Shape)

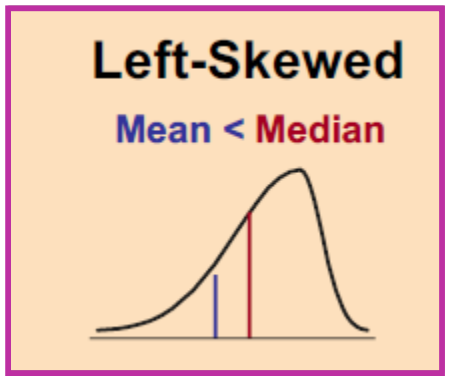
L'Asymétrie (The Skewness)

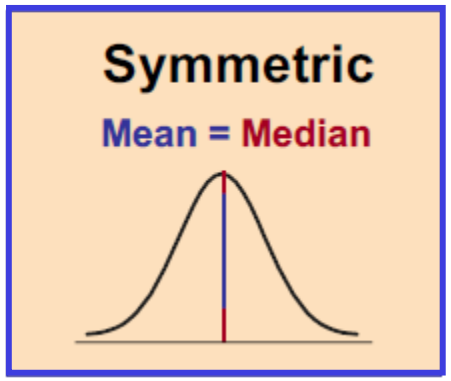
Le **skewness** est une mesure utile pour décrire la forme d'une distribution, et peut aider à identifier les distributions qui sont biaisées ou qui ne sont pas normales. C'est une mesure souvent utilisée dans l'analyse de données financières ou économiques, pour évaluer la forme des rendements ou des taux de croissance.

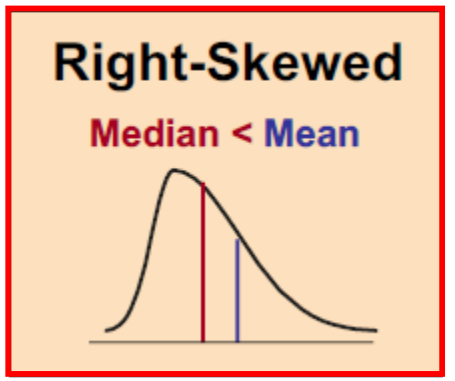
Mean < Median : Distribution asymétrique étalée à gauche = Oblique à droite (left-skewed distribution); Skewness < 0.

Mean = Median : Distribution symértique (Symmetrical distribution); skewness = 0.

Mean > Median : Distribution asymétrique étalée à droite = Oblique à gauche (right-skewed distribution); skewness > 0.





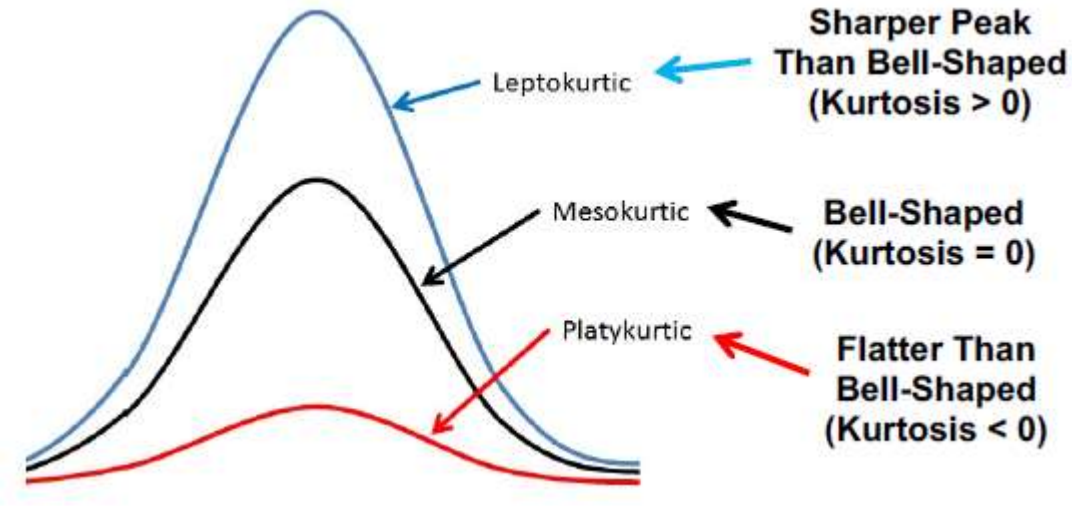


Indicateurs de forme (The Shape)

L'Aplatissement (The Kurtosis)

L'aplatissement est une mesure statistique qui décrit la forme d'une distribution de données. Cette mesure évalue l'ampleur des valeurs qui se trouvent dans les queues de la distribution par rapport à celles qui se trouvent autour de la moyenne, et indique si la distribution est plus ou moins étalée que la distribution normale. Plus précisément, cette mesure indique le degré ***d'aplatissement ou de pointu*** de la courbe de la distribution.

- Si la distribution ***est similaire à la distribution normale (bell-shaped)***, l'aplatissement sera proche de zéro ***(la courbe est mésokurtique)***.
- Si la distribution est ***plus pointue que la distribution normale***, l'aplatissement sera positif ***(la courbe est leptokurtique)***. Elle est plus pointue et possède des queues plus longues.
- Si la distribution est ***plus aplatie que la distribution normale***, l'aplatissement sera négatif ***(la courbe est platykurtique)***. Elle est plus arrondie et possède des queues plus courtes.



Exercise 3:

Excel sheets DataEx3 contains the overall download and upload speeds in mbps (megabits per second) for nine carriers (operators) in the United States. For the download speed and upload speed separately:

1. Compute the mean and the median
2. Compute the variance, standard deviation, range and coefficient of variation.
3. Are the data skewed? If so, how ?
4. Based on the results of (1) through (3), what conclusions can you reach concerning the download and upload speed of various carriers?

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

1. Compute the mean and the median

2. Compute the variance, standard deviation, range and coefficient of variation.

Les fonction utilisées sur Excel:

Mean = MOYENNE(C6:C14)

Median = MEDIANE(C6:C14)

Variance = VAR(C6:C14)

Standard D = ECARTYPE(C6:C14)

Range = MAX(C6:C14)-MIN(C6:C14)

Coefficient de variation = Standard D/ Mean = C18/C15

Skewness = COEFFICIENT.ASYMETRIE(C6:C14)

Carrier	Download Speed
Cricket	4,5
Straight Talk	7,1
Boost	10,3
Virgin Mobile	10,8
Sprint	11,2
Metro PCS	16,7
AT&T	20,8
T-Mobile	22,7
Verizon	24
Mean	14,23
Median	11,20
Variance	49,80
Standard deviation	7,06
Range	19,5
CV	50%
Skewness	0,19

Carrier	Upload Speed
Verizon	14,3
T-Mobile	13,2
AT&T	9,10
Metro PCS	11,10
Sprint	6,40
Virgin Mobile	6,20
Boost	6,00
Straight Talk	3,00
Cricket	3,80
Mean	8,12
Median	6,40
Variance	16,23
Standard deviation	4,03
Range	11,30
CV	50%
Skewness	0,39

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

3. Are the data skewed? If so, how ?

Pour les deux cas (**Download Speed & Upload Speed)**, nous remarquons que la valeur de la moyenne est supérieure à celle de la médiane ***(Mean > Median)***.
En plus le Coefficient de l'asymértie est supérieure à zéro ***(Skewness > 0)***.
Par conséquent, la distribution est asymétrique étalée à droite autrement dit oblique à gauche ***(right-skewed distribution)***.

Carrier	Download Speed
Mean	14,23
Median	11,20
Skewness	0,19

Carrier	Upload Speed
Mean	8,12
Median	6,40
Skewness	0,39

4. Based on the results of (1) through (3), what conclusions can you reach concerning the download and upload speed of various carriers?

Les deux cas (**Download Speed & Upload Speed)** présentent des coefficients de variation identiques de 50%, cela signifie que les deux séries ont la même dispersion relative par rapport à leur moyenne respective. Cependant, les moyennes sont différentes, 14,23 pour le **Download Speed** et 8,12 pour **Upload Speed**. Par conséquent, bien que les deux séries aient la même dispersion relative, la série **Download Speed** a des valeurs plus élevées en général, tandis que la série **Upload Speed** a des valeurs plus basses en général.

Exercise 4:

The following data give the average room price (in US$) paid by various nationalities while traveling abroad in 2016:

124 101 115 126 114 112 138 85 138 96 130 116 132.

1. Compute the mean, median, and mode.

2. Compute the range, variance, and standard deviation.

3. What conclusions can you reach concerning the room price paid by international travelers while traveling to various countries in 2016 ?

4. Suppose that the last value was 175 instead of 132. Repeat (1) through (3), using this value. Comment on the difference in the results.

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

1. Compute the mean, median, and mode.
2. Compute the range, variance, and standard deviation.

	room price
	85
	96
	101
	112
	114
	115
	116
	124
	126
	130
	132
	138
	138
Mean	117,46
Median	116,00
Variance	263,60
SD	16,24
Range	53,00
CV	14%
Skewness	-0,59

Les fonction utilisées sur Excel:

Mean = MOYENNE(C8:C20)

Median = MEDIANE(C8:C20)

Variance = VAR(C8:C20)

Standard D = ECARTYPE(C8:C20)

Range = MAX(C6:C14)-MIN(C8:C20)

Coefficient de variation = Standard D/ Mean = C24/C21

Skewness = COEFFICIENT.ASYMETRIE(C8:C20)

Remarque: comme dans ce cas, il est possible d'avoir une valeur de coefficient d'asymétrie négative et une moyenne supérieure à la médiane, car ces mesures statistiques sont basées sur des propriétés différentes de la distribution et peuvent refléter des aspects différents de la tendance centrale des données que nous ne somme pas sensé de les détailler à ce niveau dans ce cours.

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

3. What conclusions can you reach concerning the room price paid by international travelers while traveling to various countries in 2016 ?

En moyenne le prix des chambres payé par les voyageurs internationaux lors de leurs voyages dans différents pays est de 117$. On peut conclure que la moitié des prix payés par les voyageurs (50%) étaient inférieurs ou égaux à 116$ (médiane) et l'autre moitié était supérieure ou égale à 116$.

Un écart-type de 16 signifie que la plupart des prix de chambre payés se situent à environ 16 unités de la moyenne, soit entre 101 (117 – 16) et 133 (117 +16) $.

Le range est de 53$, cela signifie que la différence entre le prix le plus bas et le plus élevé est de 53$.
Un coefficient de variation de 14% indique que la variation relative des prix est relativement faible par rapport à la moyenne. Finalement, une valeur négative de l'asymétrie et (Mean > Median) indiquent que la distribution est asymétrique étalée à droite ***(Right-skewed distribution)***.

room price
85
96
101
112
114
115
116
124
126
130
132
138
138
Mean 117
Median 116
Variance 263
SD 16
Range 53,00
CV 14%

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

4. Suppose that the last value was 175 instead of 132. Repeat (1) through (3), using this value. Comment on the difference in the results.

<u>Le changement de cette valeur a influencé plusieurs paramètres:</u>

La moyenne des prix des chambres dans la série 2 (120,77) est devenu supérieure à celle de la série 1 (117,46), ce qui indique que les prix moyens dans la série 2 sont plus élevés que dans la série 1.Seulement, la médiane a conservé sa valeur parce qu'elle est moins sensible que les autres indicateurs par les valeurs extrêmes. Par conséquent, lorsque on analyse des données avec des valeurs extrêmes, la médiane est souvent préférée à la moyenne, car elle donne ***une mesure plus stable et plus représentative de la valeur centrale*** de l'ensemble de données.

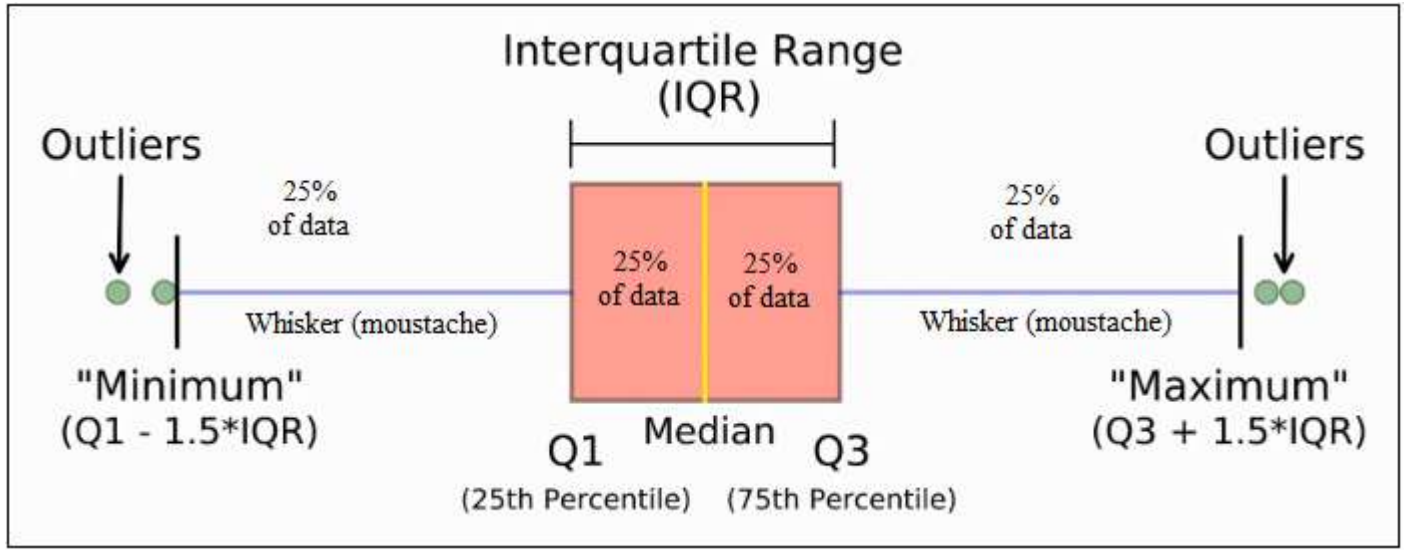
	room price
	85
	96
	101
	112
	114
	115
	116
	124
	126
	130
	132
	138
	138
Mean	117,46
Median	116,00
Variance	263,60
SD	16,24
Range	53,00
CV	14%

	room price
	85
	96
	101
	112
	114
	115
	116
	124
	126
	130
	138
	138
	175
Mean	120,77
Median	116,00
Variance	510,03
SD	22,58
Range	90,00
CV	19%

La nouvelle série est devenu plus dispersée que la première. Ceci est du à la valeur extrême (175) qui a pris la place de la valeur 132 et qui a influencé la moyenne ce qui a changé les résultats pour plusieurs indicateurs. Ce changement a également et largement changé le range vers 90$ entre le prix le plus bas et celui le plus élevé.

Box-Plot (Box-and-whisker plots)

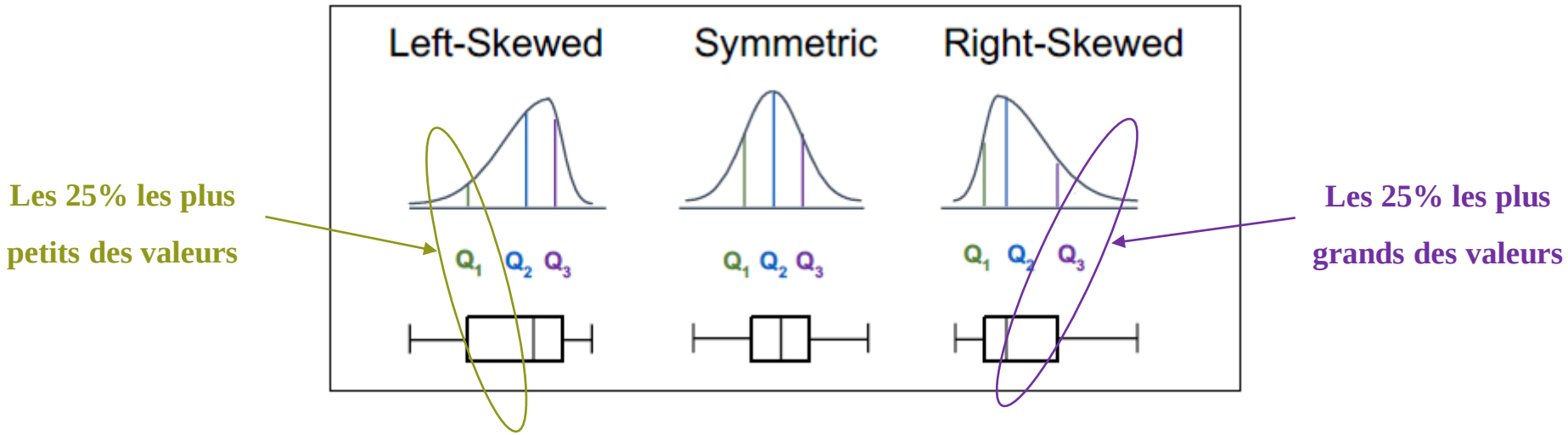
Le Box-Plot, également appelé diagramme en boîte à moustache, est un graphique qui représente visuellement une distribution de données numériques. Un Box-Plot est constitué d'une boîte qui s'étend de la médiane (valeur centrale) des données au premier et troisième quartiles (les valeurs qui divisent la distribution en quarts égaux) et de deux moustaches qui s'étendent jusqu'aux les observations les plus éloignées qui ne sont pas considérées comme des valeurs aberrantes.



Le Box-Plot permet de visualiser rapidement la dispersion et la forme d'une distribution de données, ainsi que les valeurs extrêmes qui pourraient indiquer la présence de valeurs aberrantes. Il peut également aider à comparer plusieurs distributions de données.

Box-Plot (Box-and-whisker plots)

Relation entre le Box-Plot et l'asymétrie d'une distribution:



- ❑ Dans la distribution symétrique, la moyenne et la médiane sont égales. La longueur du queue droite est égale à celui gauche.
- ❑ Dans la distribution asymétrique ***étalée à gauche***, il y a une longue queue gauche qui contient les ***25% les plus petits des valeurs***.
- ❑ Dans la distribution asymétrique ***étalée à droite***, il y a une longue queue droite qui contient les ***25% les plus grands des valeurs***.

Exercise 5:

The following data give the average room price (in US$) paid by various nationalities while traveling abroad in 2016

124 101 115 126 114 112 138 85 138 96 130 116.

1. Compute the first quartile Q1, and the third quartile Q3, and the interquartile range.
2. List the five-number summary.
3. Construct a Box-Plot and describe its shape.

On a $n = 12$

Les données triées à l'ordre croissant: 85 96 101 112 114 115 116 120 124 126 138 138

Q1 = la valeur à la position (12+1)/4 = 3,25ème position = arrondissement à la 4ème position = la 4ème valeur dans la série = 112

Q2 = la valeur à la position (12+1)/2 = 6,5ème position = la moyenne des 6ème et 7ème valeurs = (115+116)/2 = 115,5

Q3 = la valeur à la position 3(12+1)/4 = 9,75ème position = arrondissement à la 10ème position = la 10ème valeur dans la série = 126

IQR = Q3 – Q1 = 126 – 112 = 14

Les bornes des valeurs aberrantes: ***Min*** = Q1 – (1,5 × IQR) = 112 – 21 = ***91*** & ***Max*** = Q3 + (1,5 × IQR) = 126 + 21 = ***147***

La valeur 85 est une valeur aberrante (outlier) car elle est < à 91 qui représente le Min dans ce cas. On constate que la distribution est asymétrique ***étalée à droite***, dont le queue à droite contient les ***25% les plus grands des valeurs***. 50% des prix des chambres sont < à 115,5$ et 50% des prix sont > à 115,5$. Alors que 75% des prix sont < à 126$. 25% des prix sont compris entre 91 et 112.

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

Moyenne d'une population (population Mean)

La moyenne de la population est la principale mesure de tendance centrale dans une population. La lettre grecque minuscule mu, $\boldsymbol{\mu}$, représente ce paramètre qui est défini par l'équation suivante:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_i + \cdots + x_N}{N}$$

Variance et écart type d'une population (The Population variance and Standard Deviation)

La Variance de la population est représentée par la lettre grecque σ^2 défini par la formule suivante:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

L'écart type de la population est représentée par la lettre grecque σ défini par la formule suivante:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}$$

La règle empirique

Dans les ensembles de données symétriques, où la médiane et la moyenne sont identiques, produisant souvent une distribution normale en forme de cloche. La règle empirique énonce que pour des données de population provenant d'une distribution symétrique en forme de cloche telle que la distribution normale, les énoncés suivants sont vrais :

- Environ ***68%*** des valeurs se situent à $\pm$1 écart-type de la moyenne.
- Environ ***95%*** des valeurs se situent à $\pm$ 2 écart-types de la moyenne.
- Environ ***99,7%*** des valeurs se situent à $\pm$ 3 écart-types de la moyenne.

Exemple:

Supposons que vous collectiez des données sur le poids des pommes dans un verger. Si ces données suivent une distribution normale symétrique en forme de cloche, vous pouvez utiliser ***la règle empirique*** pour déterminer la proportion de pommes:

Supposons que la moyenne du poids des pommes soit de ***200 grammes*** et que ***l'écart-type soit de 20 grammes***. Selon la règle empirique, environ 68% des pommes pèseront entre 180 et 220 grammes, environ 95% des pommes pèseront entre 160 et 240 grammes, et environ 99,7% des pommes pèseront entre 140 et 260 grammes. Cela vous donne une idée de la plage de poids que vous pouvez vous attendre à trouver pour la plupart des pommes dans votre verger.

La Covariance (The Covariance)

La **covariance** est une méthode mathématique qui permet d'évaluer le sens de variation de deux variables et par la suite, de qualifier l'indépendance de ces derniers. La covariance d'un échantillon est défini par la formule suivante:

$$COV\ (X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

On utilise la covariance pour déterminer la direction d'une relation linéaire entre deux variables comme suit :

- Si le ***COV(X,Y) > 0*** , alors les deux variables tendent à augmenter ou à diminuer ensemble (they move in the same direction).
- Si le ***COV(X,Y) < 0*** , alors une variable tend à augmenter tandis que l'autre diminue (they move in the opposite direction).
- Si le ***COV(X,Y) = 0***, alors les deux variables sont indépendants, l'une n'influence l'autre (they are independent)

Remarque: Comme les données ne sont pas standardisées, on ne peut pas utiliser la covariance pour évaluer l'importance de la relation linéaire. Alors, pour évaluer **la force et la qualité d'une relation** entre deux variables à l'aide d'une échelle normalisée allant de −1 à +1, on utilise le coefficient de corrélation.

Le Coefficient de corrélation (The Coefficient of Correlation)

Le coefficient de corrélation est souvent utilisé en statistiques pour étudier les relations entre les variables et pour prédire les valeurs de l'une des variables en fonction des valeurs de l'autre variable. Cependant, il est important de noter que la corrélation ne signifie pas nécessairement une relation de cause à effet entre les variables.

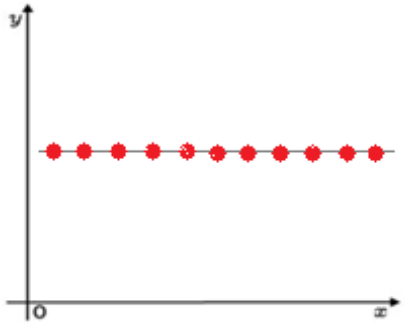
$$r = \frac{COV\ (X,Y)}{S_X\ S_Y} \qquad \text{Avec:} \qquad COV\ (X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

$$s_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \text{ , et } \quad s_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$

Plus *r* est proche de +1 ou de -1, plus les deux caractères sont dépendants. Plus il est proche de 0, plus les deux caractères sont indépendants:

- Si : **r = 0** : la corrélation linéaire observée est nulle : il n'y a pas de relation linéaire entre les variables.
- Si : **r = ±1** : : Les points sont alignés: La corrélation linéaire observée est parfaite.
- Si **r > 0** : la liaison est positive. Alors que, si **r < 0** : la liaison est négative.

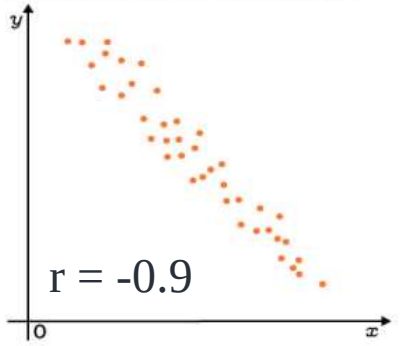
Le Coefficient de corrélation (The Coefficient of Correlation)



r = 0

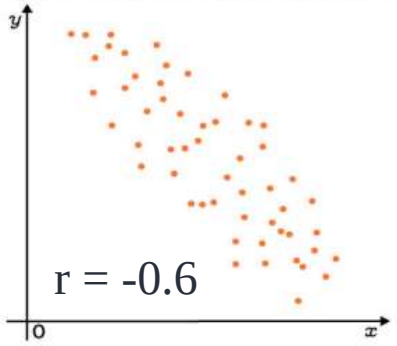
Corrélations linéaires négatives

Corrélation linéaire forte



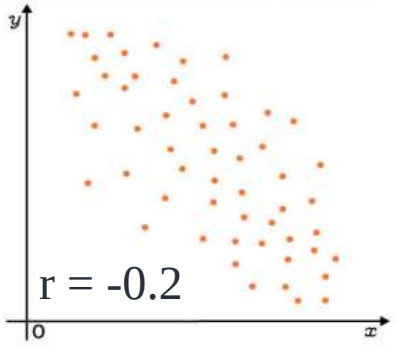
r = -0.9

Corrélation linéaire moyenne



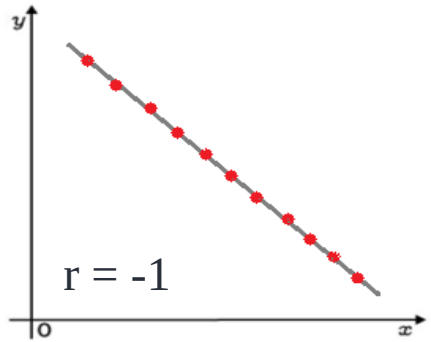
r = -0.6

Corrélation linéaire faible



r = -0.2

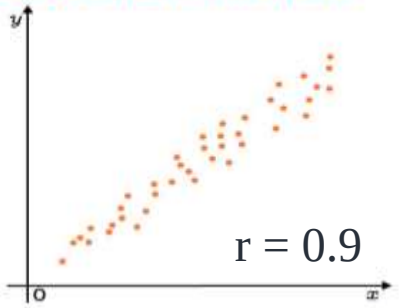
Corrélation linéaire négative parfaite



r = -1

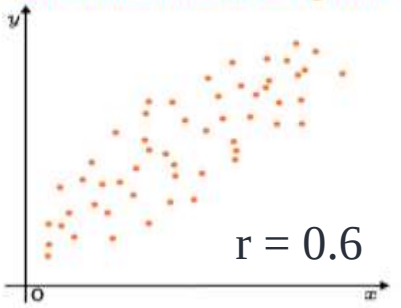
Corrélations linéaires positives

Corrélation linéaire forte



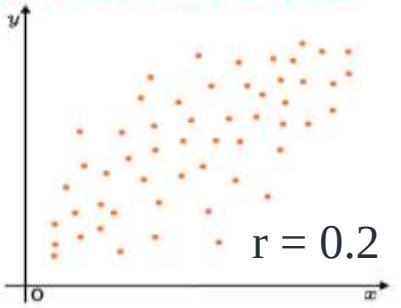
r = 0.9

Corrélation linéaire moyenne



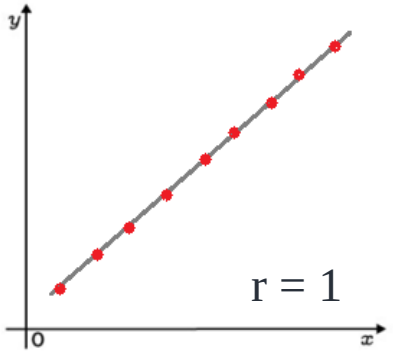
r = 0.6

Corrélation linéaire faible



r = 0.2

Corrélation linéaire positive parfaite



r = 1

Statistiques Descriptives Bidimensionnels: Distribution à 2 variables

Exercise 6:

The following is a set of data from a sample 11 items:

X	8	6	9	4	7	11	13	5	10	16	19
Y	22	15	25	10	19	31	37	13	27	45	54

1. Calculate the covariance
2. Calculate the coefficient of correlation
3. Describe the relationship between the two variables X and Y.

Statistiques Descriptives Bidimensionnels: Distribution à 2 variables

1. Calculate the covariance

$$COV\ (X,Y) = \frac{\Sigma_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

$$= \frac{641}{11-1} = \frac{641}{10} = 64{,}1$$

	X	Y	$x_i - \bar{X}$	$y_i - \bar{Y}$	$(x_i - \bar{X})(y_i - \bar{Y})$	$(x_i - \bar{X})^2$	$(y_i - \bar{Y})^2$
	8	22	-1,818	-5,091	9,256	3,306	25,917
	6	15	-3,818	-12,091	46,165	14,579	146,190
	9	25	-0,818	-2,091	1,711	0,669	4,372
	4	10	-5,818	-17,091	99,438	33,851	292,099
	7	19	-2,818	-8,091	22,802	7,942	65,463
	11	31	1,182	3,909	4,620	1,397	15,281
	13	37	3,182	9,909	31,529	10,124	98,190
	5	13	-4,818	-14,091	67,893	23,215	198,554
	10	27	0,182	-0,091	-0,017	0,033	0,008
	16	45	6,182	17,909	110,711	38,215	320,736
	19	54	9,182	26,909	247,074	84,306	724,099
Somme	108	298	0	0	641	218	1891
Moyenne	**9,82**	**27**					

2. Calculate the coefficient of correlation

$$r = \frac{COV\ (X,Y)}{S_X\ S_Y} = \frac{64{,}1}{4{,}66 \times 13{,}78} = \frac{64{,}1}{64{,}2148} = 0{,}9$$

$$avec: s_X = \sqrt{\frac{218}{10}} = 4{,}66 \text{ , et } s_Y = \sqrt{\frac{1891}{10}} = 13{,}78$$

3. Describe the relationship between the two variables X and Y.

Un coefficient de corrélation de 0,9 prouve une forte relation positive entre les deux variables. Cela signifie que lorsque la valeur d'une variable augmente, la valeur de l'autre variable a également tendance à augmenter dans une grande mesure. De même, si la valeur d'une variable diminue, la valeur de l'autre variable a également tendance à diminuer dans une grande mesure.