Les statistiques descriptives

Mesures de tendance centrale,
Mesure de dispersion et
Mesure de position

Les mesures de tendance centrale

- résument l'ensemble des données à l'aide d'une seule valeur

- Ces mesures permettent de savoir :
- ce qui se passe au centre des données,
- quelles sont les valeurs centrales d'une distribution

- Nous verrons 3 mesures de tendance centrale :
  - le mode
  - la médiane
  - la moyenne

1) La moyenne $\bar{x}$ et μ

La moyenne représente le centre de gravité de l'ensemble des données, c.a.d. représente le centre d'équilibre des données.

- Comment la calculer ?

Voici les formules permettant de calculer la moyenne :

- Si les données sont non groupées :

Échantillon	Population
$\bar{X} = \frac{\sum x_i}{n}$	$\mu = \frac{\sum x_i}{N}$
Où x_i représentent chacune des données	

3

Exemple

- Soit le prix des maisons d'une rue d'un quartier. Les prix, en milliers de dollars sont les suivants :
- 167, 150, 155, 200, 210, 178, 170, 188, 189, 210
- Est-ce un échantillon ou une population ?
- Quelle est la moyenne de la variable?

4

- [illegible]

[illegible]	[illegible]
$\bar{X} = \frac{\sum x_i f_i}{n}$	$\mu = \frac{\sum x_i f_i}{N}$

[illegible]

- [illegible]

$$\bar{\boldsymbol{x}} = \boldsymbol{\mu} = \sum \boldsymbol{x_i f_i}$$

[illegible]

Prenons l’exemple du nombre de cours différents donnés en 2018-2019 par des profs de cégep

[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]

Exercices

Vous avez 5 évaluations dans la session : 2 examens qui valent 20% chacun, des devoirs et minitests qui valent 10% chacun et un final à 40%. Supposons que vous ayez eu les notes suivantes :

- Examen 1 : 40/100
- Examen 2 : 67/100
- Devoirs : 89/100
- Minitests : 70/100

Quelle note devez-vous avoir à l’examen final pour réussir le cours à 60%?

2) La médiane Md

- La médiane, que l'on note Md, est la valeur qui partage une série de données ordonnées en deux parties égales, chacune comprenant le même nombre de données.
- Pour trouver la médiane d'une série de données brutes, il faut les placer en ordre croissant et déterminer quelle valeur sépare le lot en 2 groupes identiques.
- Nombre impair de données :

On choisit la $\left(\frac{n+1}{2}\right)^e$ donnée

- Nombre pair de données :

On fait la moyenne entre la $\left(\frac{n}{2}\right)^e$ donnée et la $\left(\frac{n}{2}+1\right)^e$ données

7

Exemples

1. Quelle est la médiane de la série suivante?

4 5 5 7 8 9 10 11 11

Donc, il a la moitié des données qui se situent au-dessus de ________ et 50% des données se situent en-dessous de _________.

2. Quelle est la médiane de la distribution suivante?

RÉPARTITION DE 56 TIRAGES CONSÉCUTIFS DU LOTTO 6/49, SELON LE NOMBRE DE BILLETS GAGNANTS DU GROS LOT

Nombre de billets gagnants	Nombre de tirages	Cumul
0	16	16
1	12	28
2	11	39
3	8	47
4	7	54
5	1	55
6	1	56
Total	56	

8

Exercices

1. Quelle est la médiane de la série de données suivantes?

167, 150, 155, 200, 210, 178, 170, 188, 189, 210

Donc, il y a la moitié des données qui se situent au-dessus de ________ et 50% des données se situent en-dessous de _________.

Quelle est la médiane de la distribution suivante?

REPARTITION DE 75 PROPRIETAIRES D'ENTREPRISE, SELON LE NOMBRE DE VOYAGES D'AFFAIRES À L'ÉTRANGER AU COURS D'UN MOIS

Nombre de voyages	Nombre de propriétaires
0	25
1	35
2	10
3	3
4	2
Total	75

3) Le mode Mo

- Le mode est la valeur, la classe de valeurs ou la modalité qui a la plus grande fréquence. Il est significatif seulement si sa fréquence est nettement différente de celles des autres valeurs ou classes.
- Une distribution peut avoir plus d'un mode :
  - Unimodale (1 mode)
  - Bimodale (2 modes)
  - Multimodale (3 modes ou plus)
- Une distribution peut ne pas avoir de mode si aucune valeur ne se démarque.

RÉPARTITION EN POURCENTAGE DE CONSOMMATEURS, SELON LA MARQUE PRÉFÉRÉE

Marque	Pourcentage des consommateurs (%)
A	19
B	21
C	22
D	18
E	20
Total	100

Exemples

1. Quel est le mode de la série suivante?

Variable : Nombre d'heures par jour passées devant l'ordinateur

Échantillon : 15 étudiants du collège

2 – 3 – 3 – 1 - 3 – 3 – 2 – 2 – 1 - 3 – 3 – 2 – 4 – 1 – 2

Mo = ________________

2. Quel est le mode de la série suivante?

Variable : être fumeur ou non

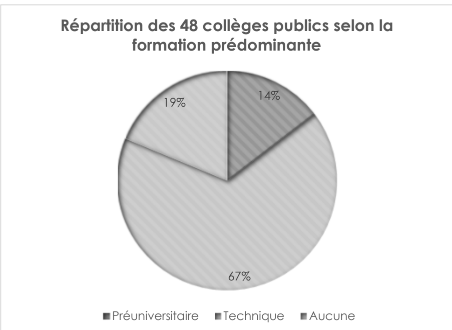
Échantillon: 10 résidents de Shawinigan

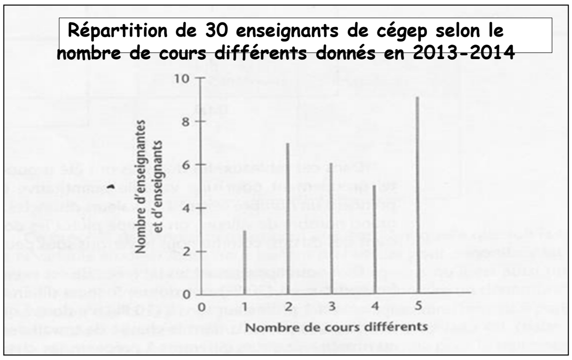
f, n-f, n-f, f, n-f, n-f, f, n-f, n-f, n-f

Mo = ________________

Exemple

Quel est le mode des distributions suivante?





Quelle mesure de tendance centrale utilisée?

[illegible]	[illegible]	[illegible]	**Variable**
[illegible]	- [illegible] - [illegible] - [illegible]	- [illegible] - [illegible]	- Quantitative seulement
[illegible]	- [illegible] - [illegible]	- [illegible]	- Quantitative seulement
[illegible]	- [illegible] - [illegible]	- [illegible]	- Qualitative et quantitative

13

Les mesures de dispersion

- [illegible]
- [illegible]
- [illegible]
- [illegible]
  - [illegible]
  - [illegible]
  - [illegible]
  - [illegible]

14

2…n~ «}n ø

Y£vjv·¡ ‡ai}na· «fibfln¡‡a¡‡}nfl¡£‡nflËn¡ Î ½e ·¡ n…a~ n¡ ln ~ a‡ufl«£·fi˝ tf£·«nflln ˙˝ o‡·lva¡‡fl

8 fi%	ˇ˝	˛ˇ	ˇ˚	—˙	—˘	ˇˇ	ˇ¸	—˙	ˇ—	ˇ˝	—˘	—˙
8 fi-	Æ˛	˝—	Æ¸	—Æ	˝˝	Æ˝	˝ˇ	—Æ	˝—	Æ˚	ÆÆ	—˝

. a}j· }£¡fl‡}nfl‡~ £%n¡¡nfl‡ln‡jnfl‡˙‡£fi£·«nfl‡ø

% ‡ø

- ‡ø

1) L'étendue E

- L'étendue est représentée par l'écart entre la valeur maximale et la valeur minimale de l'ensemble des données de la distribution.
- $E = V_{max} - V_{min}$
- L'étendue est une mesure de dispersion très facile à calculer, mais qui ne tient compte que de 2 valeurs, la plus grande et la plus petite. C'est pourquoi on l'utilise rarement seule, mais plutôt en combinaison avec d'autres mesures de dispersion.

- Exemple précédent :

Groupe A

Groupe B

2) L'écart type
s ou σ

- C'est la mesure de dispersion la plus utilisée, en combinaison avec la moyenne. Elle permet de préciser l'information.
- L'écart type permet de savoir comment, en moyenne, les données sont dispersées autour de la moyenne.
- Formules pour des données non groupées

3jua¡ ŧv}}£¡	M£«·}aŧv£¡
$s = \sqrt{\dfrac{\sum(x_i - \bar{X})^2}{n-1}}$	$\sigma = \sqrt{\dfrac{\sum(x_i - \mu)^2}{N}}$

E „ …v fın«fıbfın¡ŧnŧjua›·nŧl£¡¡on

¡ŧ£·ŧC ŧŧfın«fıbfın¡ŧnŧ}nŧ£~ ifınŧln ŧl£¡¡on

2…n~ «}n ø

Y£vjv·¡ ŧai}na· «fıbfın¡ŧa¡ŧ}nfl¡£ŧnflËn¡ Î ½e·¡ n…a~ n¡ ln ~ aŧufl«£·fi˝ tfı£·«nflln ˙˝ oŧ·lva¡ŧfın

8 fın%	ˇ˝	ˇ˛	ˇ˚	—¨	—˘	ˇˇ	ˇ¸	—¨	ˇ—	ˇ˝	—˘	—˙
8 fın-	Æ˛	˝—	Æ¸	—Æ	˝˝—	Æ˝	˝ˇ—	—Æ	˝—	Æ˚	ÆÆ	˝—

. a}j·}£¡flŧ}ŏjafiŧŧ‰«nŧln ŧjnflŧ˝ŧ£fı£·«nflŧø

%ŧø

-ŧø

Caractéristiques de l'écart type

- L'écart type tient compte de toutes les données comme c'est le cas pour la moyenne.
- Il possède les mêmes unités que la moyenne.
- Celui-ci peut être très influencé par les données extrêmes.
- Pratiquement toutes les données se situent entre + ou – 3 écarts types, si la distribution est normale.

3) La variance s^2 ou σ^2

- La variance se calcule en élevant au carré l'écart type.
- Elle donne les mêmes renseignements en ce qui a trait à la dispersion des données.
- Elle est surtout utilisée en inférence statistique pour faire l'analyse de variance ou pour tester les hypothèses.
- Exemple précédent :

Groupe A

Groupe B

4) Le coefficient de corrélation

CV

- Le coefficient de variation est une mesure relative de la dispersion.
- Le coefficient de variation permet de comparer n'importe quel ensemble de données même si leurs moyennes sont différentes ou que les unités sont différentes.
- Le coefficient de variation indique le degré d'homogénéité des données, c'est-à-dire si les données sont homogènes ou hétérogènes.
- Les formules sont :

3jua¡ŧv}}£¡	M£«·}aŧv£¡
$CV = \frac{s}{\overline{X}} \times 100$	$CV = \frac{\sigma}{\mu} \times 100$

- **Remarques :**

- Si CV ≤ 15 % ➔ la distribution est homogène, c.a.d. que les données sont peu ou modérément dispersées autour de la moyenne.
- Si CV > 15 % ➔ la distribution est hétérogène, c.a.d. que les données sont assez ou très dispersées autour de la moyenne.

Exemple

Un travailleur qui réside sur la Rive-Sud de Montréal peut utiliser différents trajets pour se rendre au travail. Voici les statistiques qu'il a recueillies après quelques essais de ces différents parcours.

DURÉE MOYENNE ($\overline{x}$) ET ÉCART TYPE (s) DE LA DURÉE, SELON L'ACCÈS UTILISÉ

	Hippolyte-Lafontaine	Jacques-Cartier	Champlain	Victoria
$\overline{x}$	46 min	40 min	35 min	45 min
s	8,5 min	8 min	12 min	10 min

Quel trajet est le plus homogène quant au temps de parcours?

Les mesures de position

- B nfl·fhfl «nfi~ n‡a¡‡ ln fl‡·nfi ·¡n l£¡¡on «afi fa««£fi‡ a·… a·‡fhfl l£¡¡onfl lð¡n flofin fl‡a‡fl‡ › ·nı
- 0 n·…~ nfl·fhfl ln «£fl‡€¡ flnf£¡‡o‡·lvonflø
  - }nfl › ·a¡‡v}nfl
  - }a j£‡n _

23

1) Les quantiles

- Les quantiles sont des valeurs qui partagent une distribution en un certain nombre de parties égales.
  - Les **quartiles** (Q_1, Q_2, Q_3) la partagent en 4 parties comprenant chacune 25% des données.
  - Les **quintiles** (V_1, V_2, V_3, V_4) la partagent en 5 parties comprenant chacune 20% des données.
  - Les **déciles** (D_1, D_2, …, D_9) la partagent en 10 parties comprenant chacune 10% des données.
  - Les **centiles** (C_1, C_2, …, C_{99}) la partagent en 100 parties comprenant chacune 1% des données.

24

Procédure

Pour les données non-groupées en classes, voici la façon dont on calcule un centile :

Pour calculer le centile C_x :

1. Ordonner les données.
2. Déterminer la position du centile cherché :
   ***a) Si $n \times x\%$ est un entier**, le centile C_x est la moyenne entre la donnée $n \times x\%$ et la suivante.*
   ***b) Si $n \times x\%$ n'est pas un entier**, le centile C_x est la donnée suivant $n \times x\%$*
3. Interprétation : Au moins $x\%$ des « *unités statistiques* » sont inférieures ou égales à « C_x ».

Remarque : Pour calculer un autre quantile, trouver le centile qui est équivalent.

Exemples

1. Soit les poids à la naissance de 10 nouveau-nés :

2350	3150	3252	3334	3552
3684	3843	3926	4125	4650

Trouver et interpréter le 2[e] décile.

Interprétation : _____ % des nouveau-nés pèsent _________ de _______ g.

2. Soit les poids à la naissance de 10 nouveau-nés :

2350	3150	3252	3334	3552
3684	3843	3926	4125	4650

Trouver et interpréter le 3[e] quartile.

Interprétation : Au moins _____ % des nouveau-nés pèsent ________ g ou moins.

2) La cote Z

La **cote Z** donne le nombre d'écarts types qui séparent une donnée de la moyenne.

Formules : $cote\ Z = \frac{X-\mu}{\sigma} = \frac{donnée - moyenne}{écart\ type}$

Caractéristiques de la cote Z :

- Si la cote Z est négative, alors la valeur x est sous la moyenne.
- Si la cote Z est positive, alors la valeur x est au-dessus de la moyenne.
- Les cotes Z varient généralement entre -3 et 3.
- La cote Z permet de situer un individu par rapport à son groupe.
- La cote Z permet comparer des données provenant de séries statistiques différentes.

Exemple

On veut trouver le meilleur vendeur du mois. Cet honneur sera accordé à la personne s'étant le plus distingué dans son domaine.

Voici la description de la performance, en un mois, de chacun des candidats :

- Lise a vendu 85 barres de chocolat, alors que la moyenne de vente a été de 52 barres par étudiant avec un écart type de 13 barres.
- Paul a vendu 25 polices d'assurance-vie, alors que la moyenne de vente est de 12 polices avec un écart type de 6.
- Lucie a vendu 75 abonnements au Journal de Québec, alors que la moyenne de vente est de 47 abonnements avec un écart type de 10.

Qui sera déclaré « vendeur du mois » ?

Exercices

Résultats d'un élève dans trois cours

Matière	Note de l'élève (x)	Moyenne (μ)	Écart à la moyenne ($x-\mu$)	Écart type (σ)	Cote z $[(x-\mu)/\sigma]$
Philosophie	85	87,5	−2,5	5	−0,5
Français	84	69	15	10	1,5
Méthodes quantitatives	80	70	10	5	2

Dans quel cours cet étudiant a-t-il le plus performé?