Variables, répartition, paramètres de position et de dispersion

Guillaume Beraud

CHUO

15 Janvier 2024

Section 1

Statistiques descriptives

Subsection 1

Variables

Objectifs du cours

Vous devez

- Comprendre la notion de variable et leur classification
- Savoir comment décrire et représenter les données provenant d'un échantillon (paramètres de tendance centrale, de dispersion, tableaux et graphiques)

Les statistiques

- *Méthode de raisonnement permettant d'interpréter le genre de données très particulières, qu'on rencontre notamment dans les sciences de la vie, dont le caractère essentiel est la variabilité* D. Schartz.
- Variabilité : caractère essentiel des êtres vivants
  - de l'être humain, en particulier
- En outre, l'échantillon doit permettre de raisonner à propos de la population

Les variables

- Une variable est une caractéristique dont on peut observer des valeurs différentes au sein d'un groupe de sujets.
- Une variable peut être :
  - de nature catégorielle
  - de nature quantitative.

Variable catégorielle

- Une variable dite **catégorielle** ou **qualitative** est une caractéristique ayant un certain nombre de catégories ou modalités, exhaustives et mutuellement exclusives
- **Exhaustives** car toutes les modalités possibles sont citées
- **Mutuellement exclusives** car chaque individu peut être classé dans une catégorie et une seule.
- S'il n'y a que deux catégories, la variable catégorielle est dite **dichotomique** (ou binaire).
  - e.g. homme ou femme ; fumeurs ou non fumeurs ; atteints ou non d'allergie, douleur ou pas.

Variable catégorielle

- Lorsqu'une variable catégorielle a plus de deux catégories, elle peut être :
  - **nominales** : chaque classe désigne une catégorie de sujets (elle les nomme). Il n'existe pas d'ordre naturel entre les catégories.
    - C'est le cas du groupe sanguin : A / B / AB / O, ou encore de la situation familiale : marié / vivant en couple / célibataire / divorcé / séparé / veuf.
  - **ordinales** : il existe un ordre naturel entre les différentes catégories.
    - Au lieu de deux catégories (douleur / pas de douleur), on peut classer les individus selon les catégories suivantes : aucune / minime / modérée / sévère / insupportable.

Variable quantitative

- Les valeurs d'une variable **quantitative** sont obtenues par un instrument de mesure ou le résultat d'un dénombrement.
- Elles sont souvent accompagnées d'une unité de mesure.
- On peut toujours répondre à une question commençant par : " combien . . . ?"

Variable quantitative

- Une variable est **continue** si elle peut prendre, en théorie, un nombre infini de valeurs dans un intervalle donné,
- La précision avec laquelle on la mesure ne dépend que de l'exactitude de l'instrument de mesure.
  - âge,
  - pression artérielle systolique
  - quantité de sucre dans le sang
  - ...

Variable quantitative

- Lorsque l'on arrondit la valeur obtenue, on dit que l'on **discrétise** cette variable continue, car on lui impose de prendre certaines valeurs.
- On exprime couramment l'âge en années (22, 23, 24 ans, etc.) ou encore de dix ans en dix ans (20 à 29, 30 à 39 ans, etc.).
  1. l'**intervalle** entre chaque valeur a une amplitude d'une année,
  2. l'**amplitude d'intervalle** est de 10 ans. Dans ce dernier cas, il a été réalisé un **regroupement**, avec des intervalles de même amplitude.
- On parle aussi de variable **discrète** lorsque la variable ne peut prendre que certaines valeurs numériques.
  - Le nombre d'enfants d'une famille est une variable quantitative discrète qui peut prendre les valeurs : 0, 1, 2, 3, 4, 5, ... mais pas 1,4 enfants, ni 2,5 enfants.

Variable discrète vs. catégorielle ordinale

- On peut également choisir de ne pas utiliser les mêmes intervalles. Dans ce cas, la nouvelle variable créée après regroupement est une variable **catégorielle ordinale**.
  - Par exemple, l'âge des enfants pourrait être regroupé de cette manière : «$\leq$ 5 ans», «6-7ans», «8-9 ans», «$\geq$ 10 ans».
- *Bien distinguer*
  1. certaines variables catégorielles ordinales, comme le stade d'un cancer, qui pourrait être codé par exemple 1, 2, 3 ou 4
  2. les variables quantitatives discrètes, comme le nombre d'enfants.

Variable discrète vs. catégorielle ordinale : Petit test

- Pour une variable **catégorielle ordinale**, chaque différence entre les catégories ne signifie pas la même chose.
- Pour une variable **quantitative discrète**, chaque différence entre les catégories a toujours la même signification sur toute l'étendue des valeurs.
  1. pour la variable « stade de cancer », on ne peut pas dire que le stade 2 est deux fois plus grave que le stade 1 ; c'est donc une variable **catégorielle ordinale**.
  2. Pour la variable « nombre d'enfants » en revanche, on peut dire que deux enfants, c'est deux fois plus que un, et que trois enfants c'est trois fois plus que un ; c'est donc une variable **quantitative discrète**.

Paramètres décrivant une distribution

- Les caractéristiques permettant de décrire les êtres vivants n'ont pas une valeur unique.
- L'ensemble des valeurs observées sur un échantillon pour une caractéristique est sa distribution observée.
- Pour résumer les données, on va utiliser des paramètres, qui sont des fonctions des observations : ; ou
  1. les « paramètres de tendance centrale », comme la moyenne ou la médiane
  2. les « paramètres de dispersion », comme la variance, l'étendue et les percentiles.

Subsection 2

Paramètres de tendance centrale

Variable quantitative : La moyenne

- La **moyenne arithmétique**, appelée plus couramment **moyenne**.
- La moyenne est obtenue en faisant la somme des valeurs, puis en divisant cette somme par le nombre de valeurs, noté ici n.
- Exemple :
  - L'âge en années d'une population de cinq femmes qui viennent d'accoucher de leur premier enfant est : 24, 17, 35, 37, 32.
  - La somme est : 145 ans, et comme il y a 5 valeurs, la moyenne est : 145/5=29 ans.
  - L'âge moyen des femmes à l'accouchement de leur premier enfant est donc, pour la série de valeurs mesurées au sein de cet échantillon, de 29 ans.

La moyenne arithmétique

1. Chaque valeur est notée x_i
2. on a donc : $x_1 = 24$, $x_2 = 17$, $x_3 = 35$, $x_4 = 37$, $x_5 = 32$
3. La somme est notée $\sum_{i=1}^{n} x_i$
   - Somme de toutes les valeurs de la première à la dernière
4. $\mu = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{24+17+35+37+32}{5}$

Variable quantitative : La médiane

- La **médiane** est la valeur centrale de la distribution, qui divise l'échantillon en deux moitiés de taille égale.
- Pour trouver la médiane, il faut d'abord classer toutes les observations par ordre croissant.

Variable quantitative : La médiane

1. Si le nombre d'observations est **impair**, la médiane est la valeur correspondant à l'observation située au milieu, celle située au $\frac{n+1}{2}$ ème rang.
   - Pour notre série de 5 observations d'âge, après avoir ordonné les observations de façon croissante, la série s'écrit : 17, 24, 32, 35, 37
   - on voit facilement que la médiane est égale à 32 ans..
   - La médiane correspond bien à la valeur de la 3ème observation, car : (5+1)/2 = 3.
2. Si n est un nombre **pair**, la médiane est à mi-chemin entre les deux valeurs du milieu de la distribution (médiane = milieu).
   - Pour une série de 8 observations d'âge : 17, 24, 27, 27, 29, 32, 35, 37, la médiane se situe entre la 4ème observation (27 ans) et la 5ème observation (29 ans), car (8+1)/2 = 4,5.
   - La médiane vaut donc : (27 + 29)/2 = 28 ans.

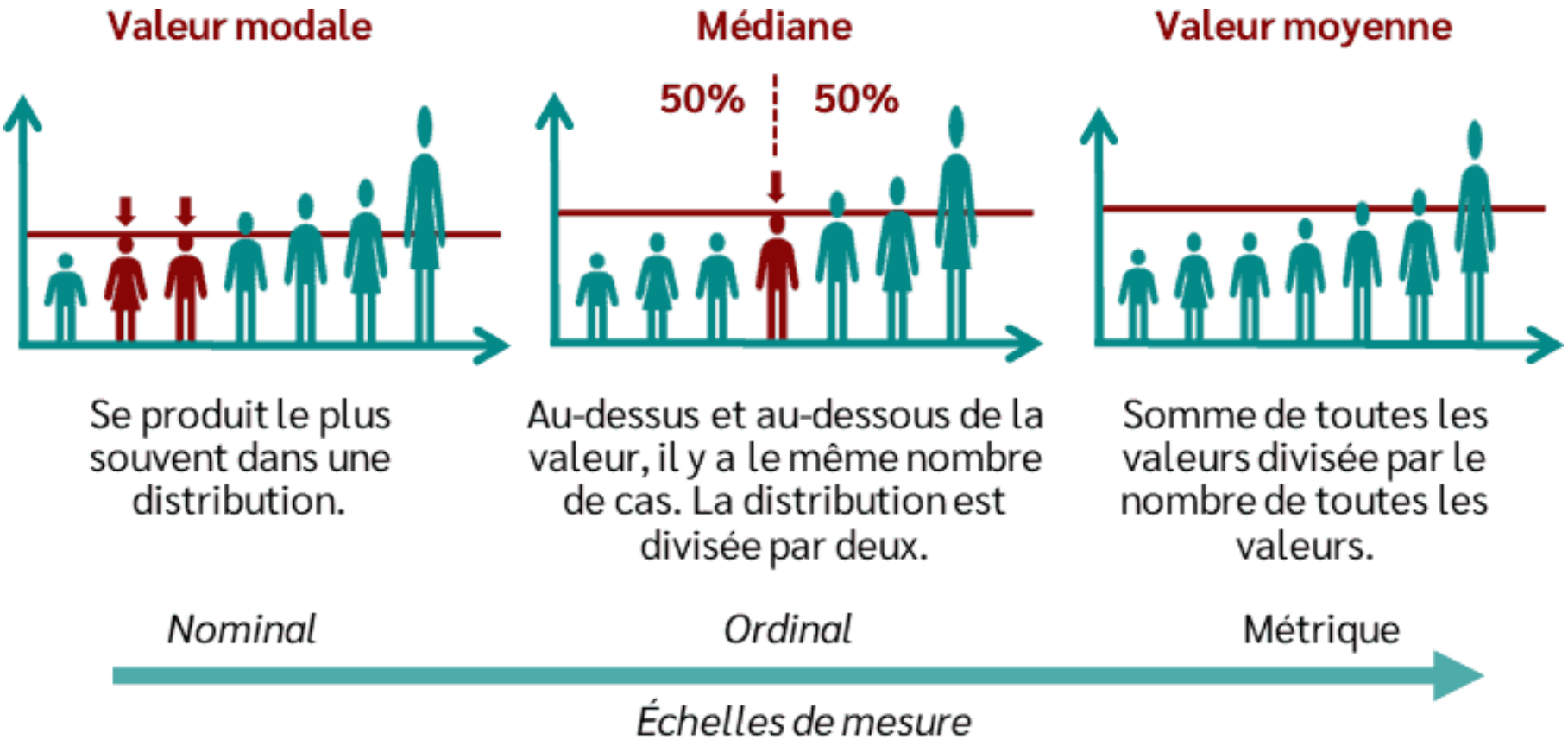
Les limites de la moyenne

Pause vidéo

Regardez **Pourquoi gagnez-vous moins que le salaire moyen**, c'est une vidéo de *La statistique expliquée à mon chat* sur les limites de la moyenne.

Variable quantitative : Le mode

- Le **mode** est la valeur la plus représentée de la série.
- Une série peut ne pas avoir de mode ou au contraire avoir plusieurs modes.
  - Pour la série des 8 observations d'âge (17, 24, 27, 27, 29, 32, 35, 37)
  - Le mode est 27 ans, car cette valeur apparaît deux fois, alors que les autres valeurs n'apparaissent qu'une seule fois
- La 1ère série des 5 valeurs d'âge (17, 24, 32, 35, 37) n'a pas de mode
- Dans la série : 17, 24, 27, 27, 29, 29, 32, 35, 37, il existe deux modes : 27 ans et 29 ans, valeurs qui apparaissent deux fois, alors que toutes les autres n'apparaissent qu'une seule fois.
  - Dans ce cas, on parle de **distribution bimodale**.


Valeur modale
Médiane
50%
50%
Valeur moyenne
Se produit le plus souvent dans une distribution.
Au-dessus et au-dessous de la valeur, il y a le même nombre de cas. La distribution est divisée par deux.
Somme de toutes les valeurs divisée par le nombre de toutes les valeurs.
Nominal
Ordinal
Métrique
Échelles de mesure

Variable catégorielle

- Pour une variable catégorielle (dichotomique, nominale ou ordinale), on présente la **proportion** des sujets dans les différentes catégories.
  - Parmi 100 malades atteints d'arthrite rhumatoïde, on observe 76 femmes et 24 hommes.
  - La proportion de femmes est le nombre de femmes rapporté au nombre total de sujets, soit 76 %.
  - La proportion d'hommes est de 24 %.
- Les proportions peuvent également être utiliséee pour une variable quantitative dont les valeurs ont été discrétisées ou bien regroupées.
  - Pour une série de femmes ayant consulté dans un centre de dépistage du cancer du sein, la proportion de femmes ayant plus de 60 ans est 70/120, soit 58,3 %.

Variable catégorielle

- On peut aussi utiliser la **fréquence cumulée** à la borne supérieure de classe.
- C'est la proportion des observations dans toutes les classes précédentes ajoutée à celle de la classe présente.
- Elle est utile pour une variable quantitative ayant fait l'objet d'une discrétisation
- Elle est également utile pour une variable catégorielle ordinale.

Titre : répartition selon l'âge des 402 sujets âgés de plus de 18 ans ayant participé à l'enquête « hypertension artérielle » du 19 au 21 avril 1999 à l'Ile Maurice

Tranche d'âge (ans)	effectif	Fréquence en pourcentage	Fréquences cumulées pourcentage
18-27	66	16,4	16,4
28-37	79	19,7	36,1
38-47	89	22,1	58,2
48-57	58	14,4	72,6
58-67	54	13,4	86,1
68-77	41	10,2	96,3
78-87	14	3,5	99,8
>88	1	0,2	100,0
Total	402	100,0	

Subsection 3

Paramètres de dispersion

Paramètres de dispersion

- Les paramètres de tendance centrale des observations ne donnent pas une idée de leur dispersion.
  - Durée d'incubation (délai entre exposition et symptômes), moyenne de 13 jours
  - Quelles mesures prendre pour les sujets exposés ?
  - Garder les patients 13 jours en isolement ?
  - Combien de personnes développent la maladie à J14, J15 etc.
- La plus petite (minimum) et la plus grande valeur (maximum) de la distribution : **l'étendue des observations**.
- L'étendue peut aussi être la différence entre valeurs minimum et maximum. Si l'on présente la différence, donner également la valeur minimum ou maximum.
  - Dans la série de valeurs d'âge (en années) : 17, 24, 27, 27, 29, 29, 32, 35, 37, le minimum est 17 ans et le maximum est 37 ans. L'étendue est 17-37 ans (*ou on peut aussi considérer que l'étendue est 20 ans*).

Paramètres de dispersion

- Mais l'étendue est insuffisante car les valeurs extrêmes sont assez particulières.
- Les **quantiles** sont les valeurs d'une distribution définies par la proportion de sujets qui se trouvent au-dessous et au-dessus de cette valeur.
- On parle de quartiles, déciles, percentiles.
- Les quartiles sont les trois valeurs qui partagent la distribution en quatre parties égales.
  - Le premier quartile correspond à la valeur de l'observation qui a 25 % de la distribution au-dessous et 75 % au-dessus,
  - le deuxième quartile est la médiane,
  - le troisième quartile correspond à la valeur de l'observation qui a 75 % de la distribution au-dessous et 25 % au-dessus.

Quartiles

Si on considère la durée de survie (en mois) de 42 patients atteints de cancer digestif :

Survie	Survie	Survie	Survie	Survie	Survie
1	15	30	40	52	60
3	16	32	41	54	64
3	17	33	41	54	64
5	23	34	42	58	74
8	24	36	44	58	74
10	28	36	44	58	74
12	28	38	49	60	90

Quartiles

- Il y a au total 42 observations et la médiane correspond à la valeur située entre les rangs 21 et 22 [(42+1)/2 = 21,5]. Comme la durée de survie est respectivement de 38 et 40 mois à ces deux rangs, la médiane vaut (38 + 40)/2 = 39 mois.
- Même méthode pour trouver la valeur de l'observation correspondant aux 1er et 3ème quartiles :
  - Le rang du 1er quartile est : (n+1)/4. Soit (42+1)/4 = 10,75 et donc une valeur le 10ème rang (17 mois) et le 11ème rang (23 mois). Avec le même calcul que pour la médiane, le 1er quartile vaut (17+23)/2 = 20 mois (ou de façon plus précise : 17 + 0,75x(23-17) = 21,5 mois).
  - Le rang du 3ème quartile est (n+1) x (3/4). Dans l'exemple, 43 x (3/4) vaut 32,25. Or, la valeur au 32ème rang et la valeur au 33ème rang valent toutes deux 58 mois. Le 3ème quartile de cette distribution est donc 58 mois.

Alternatives aux quartiles

- L'**étendue (ou intervalle) inter-quartiles** (25 % à 75 %)
- soit la partie centrale qui couvre 50 % de la distribution observée.
  - Dans l'exemple, l'étendue inter-quartiles est 20-58 mois.
- Même raisonnement avec les **quintiles**, les **déciles** ou les **centiles** (**percentiles**) qui partagent la distribution en 5, 10 ou 100 parties égales.
- Les valeurs correspondants au 5ème percentile et au 95ème percentile permettent d'obtenir l'étendue centrale couvrant 90 % de la distribution observée.

La variance

- Une autre façon de mesurer la variabilité consiste à calculer la variance σ^2
- C'est une mesure des distances de chaque individu à la moyenne
- $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$
- La variance a l'unité de la variable au carré.
  - Pour la variance de l'âge, il peut s'agir d'années au carré ($annes^2$) ou de jours au carré ($jours^2$)

Ecart-type

- Pour exprimer la variabilité dans la même unité que les valeurs observées, on en prend la racine carrée de la variance,
- Cela s'appelle l'**écart-type** (ou écart-type inter-individuel) (**standard deviation** en anglais) :
- $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$
- On utilise plus souvent l'écriture suivante, plus commode à utiliser pour les calculs manuels :
- $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i^2) - \frac{\sum_{i=1}^{n}(x_i)^2}{n}}{n}}$

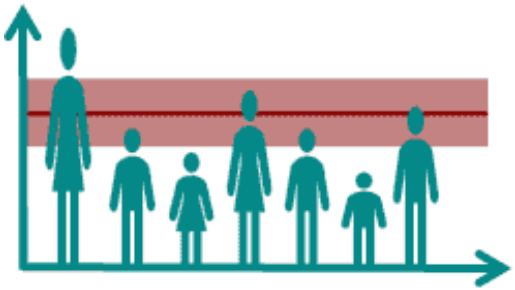
En pratique

Age en années, x_i	Age au carré (en années au carré), x_i^2
17	289
24	576
32	1024
35	1225
37	1369
$\sum_{i=1}^{n} x_i = 145$	$\sum_{i=1}^{n}(x_i^2) = 4483$

En pratique

- $\sqrt{\frac{4483-\frac{145^2}{5}}{5}} = 7,5$ ans
- Variance et écart-type sont très intéressants à titre descriptif
- Ils permettent d'apprécier à quel point la distribution est dispersée.
- Plus la variance et l'écart-type sont grands, plus la dispersion est grande (pour une même variable).

Écart-type



Distance moyenne de toutes les valeurs mesurées par rapport à la valeur moyenne

Étendue

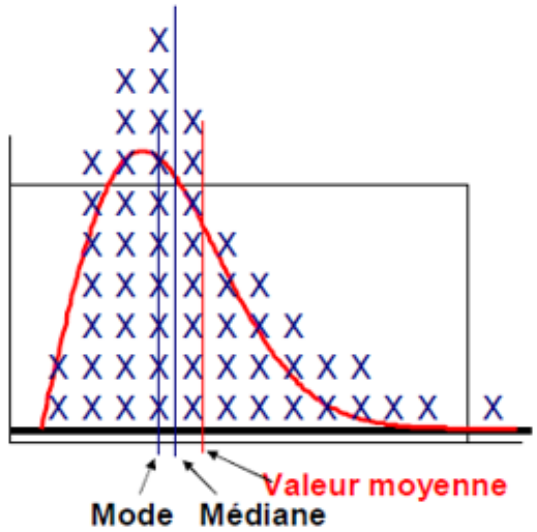
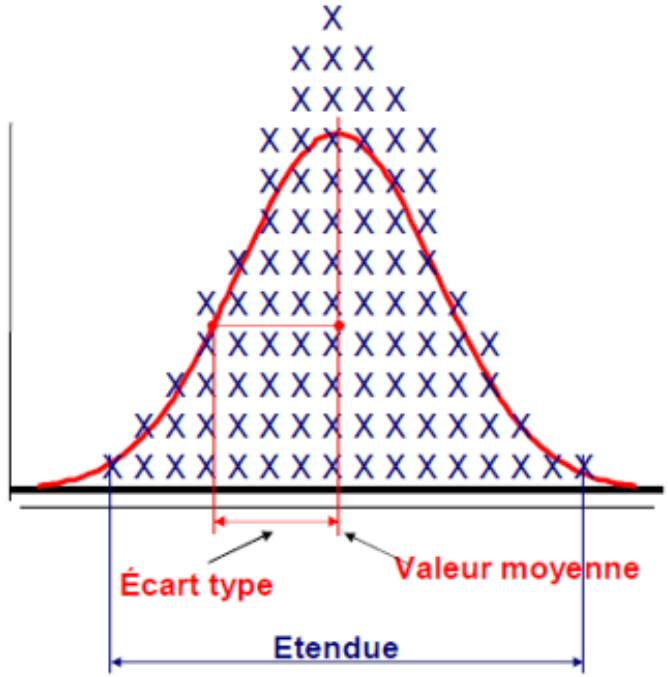
Distance entre la valeur la plus basse et la valeur la plus haute d'une distribution

Ecart interquartile

Spectre dans lequel se situent les 50% de valeurs moyennes. Différence entre le premier et le troisième quartile

Section 2

Conclusion

Tendance centrale	Dispersion
	

Merci de votre attention

Questions ?