Chapitre 2.
Statistiques à 1 variable

Yann Barsamian

École Européenne de Bruxelles 1

Année scolaire 2021–2022

Plan du chapitre

- Échantillonnage
- Rappels de S4 (moyenne, médiane, quartiles, boîte à moustaches)
- Écart-type

I/ Échantillonnage

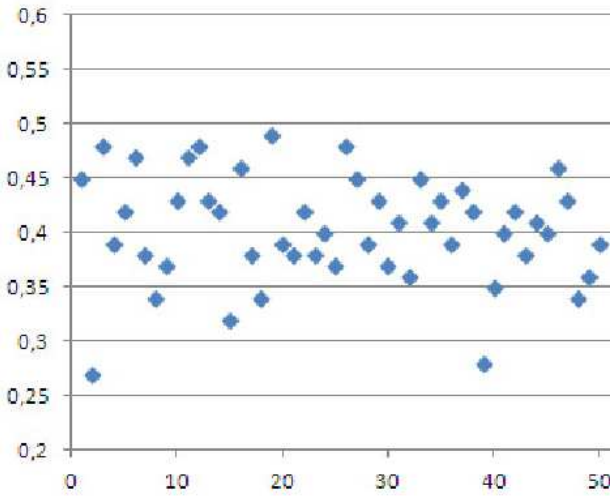
On s'intéresse à un caractère quantitatif dans une population (par ex. le pourcentage de votants pour le candidat A). Ne pouvant trouver la valeur dans toute la population, il faut prendre des échantillons, et faire un sondage pour ensuite extrapoler sur la population. C'est de l'inférence statistique.

Il faut faire attention à :

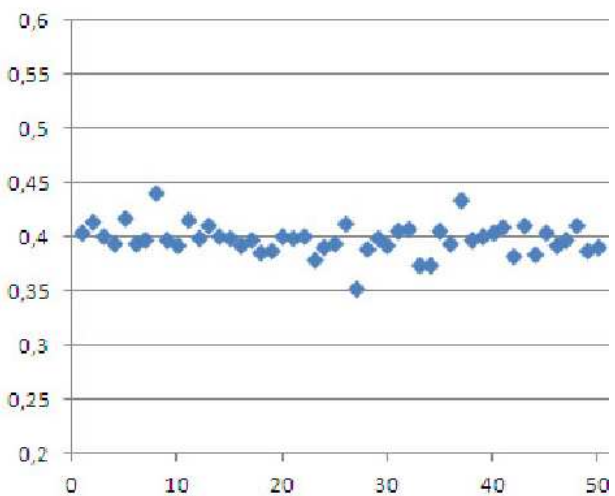
- avoir une taille d'échantillon correcte (en pratique de l'ordre de $\sqrt{N}$ environ, où N est la taille de la population)
- bien sélectionner son échantillon (aléatoirement !)
- dans un sondage, il faut traiter correctement les non-réponses
- dans un sondage, il ne faut pas introduire de biais lorsque l'on pose les questions

On a remarqué le phénomène de fluctuation d'échantillonnage : les échantillons donnent des valeurs différentes. Plus la taille de l'échantillon est grande, moins la fluctuation est grande.

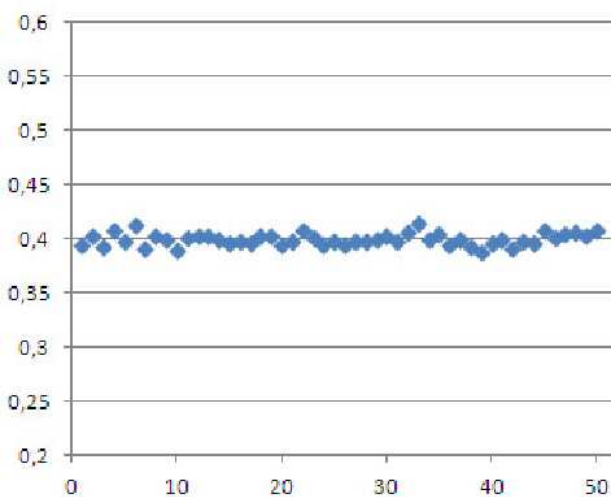
Exemple : 40% des -45 ans sont des fumeurs. . . quelques échantillons aléatoires :



Taille 100



Taille 1000



Taille 10000

Soit une série statistique prenant p différentes valeurs $x_1, x_2, \ldots, x_p$ (les x_i) et où chaque valeur x_i (pour $1 \leq i \leq p$) a pour effectif n_i. Par exemple, une étude sur 30 élèves concernant le temps de travail journalier (en minutes) à la maison donne les résultats suivants :

x_i	5	10	15	20	30	40	50	70
n_i	1	4	4	3	7	7	3	1

Dans ce tableau, on lit par exemple que 4 élèves travaillent 20 minutes par jour à la maison.

1) La moyenne[1] :

La moyenne est notée $\overline{x}$. C'est la somme des valeurs divisée par le nombre de valeurs. Lorsque l'on a des effectifs pour les valeurs, il faut pondérer la moyenne : on multiplie chaque valeur par l'effectif, et on divise par l'effectif total (qui se calcule par $n_1 + n_2 + \cdots + n_p$ si l'énoncé ne le donne pas).

$$\overline{x} = \frac{n_1 \cdot x_1 + n_2 \cdot x_2 + \cdots + n_p \cdot x_p}{n_1 + n_2 + \cdots + n_p}$$

x_i	5	10	15	20	30	40	50	70
n_i	1	4	4	3	7	7	3	1

$$\overline{x} = \frac{1 \cdot 5 + 4 \cdot 10 + 4 \cdot 15 + 3 \cdot 20 + 7 \cdot 30 + 7 \cdot 40 + 3 \cdot 50 + 1 \cdot 70}{30}$$

$= \frac{875 \div 5}{30 \div 5} = \frac{175}{6} \approx 29,17$. Ainsi, les élèves travaillent en moyenne $29,17$ minutes par jour.

1. https://www.youtube.com/watch?v=88_16UbkdZM

Remarque importante : quand on a une série donnée par classes d'intervalles, on prend pour valeur d'une classe la valeur centrale de la classe pour faire les calculs.

Exemple similaire à précédemment, où on a demandé aux élèves dans quelle tranche leur travail journalier se situait (toujours 30 élèves) :

Temps	$[0\,;\,15[$	$[15\,;\,30[$	$[30\,;\,45[$	$[45\,;\,60[$	$[60\,;\,75[$
Effectif	5	7	14	3	1

Pour la tranche $[0;15[$ on utiliserait la valeur $7,5$, pour la tranche $[15;30[$ on utiliserait la valeur $22,5$, etc.

$\overline{x} = \dfrac{5\cdot 7,5 + 7\cdot 22,5 + 14\cdot 37,5 + 3\cdot 52,5 + 1\cdot 67,5}{30} = 31,5$. Ainsi, les élèves travaillent en moyenne $31,5$ minutes par jour.

2) La médiane[2] :

La médiane sépare la série en deux parties de même effectif : au moins 50% des valeurs qui sont inférieures ou égales à la médiane, et au moins 50% des valeurs qui sont supérieures ou égales à la médiane.

Pour calculer la médiane d'une série de n nombres, on commence par ordonner les valeurs de manière croissante :

$$u_1 \leq u_2 \leq \cdots \leq u_n$$

- si n est impair, c'est la valeur centrale : celle de rang $\dfrac{n+1}{2}$
  Ex. : 2, 5, $\underline{7}$, 8, 9 : la valeur de rang $\frac{5+1}{2} = 3$: c'est 7
- si n est pair, c'est le nombre au milieu des deux valeurs centrales : la demi-somme des valeurs de rang $\dfrac{n}{2}$ et $\dfrac{n}{2}+1$
  Ex. : 2, 5, $\underline{7, 8}$, 9, 12 : la demi-somme des valeurs de rang $\frac{6}{2} = 3$ et $\frac{6}{2}+1 = 4$: c'est $\frac{7+8}{2} = 7,5$

2. https://www.youtube.com/watch?v=g1OCTw--VYQ

Médiane avec effectifs :

x_i	5	10	15	20	30	40	50	70
n_i	1	4	4	3	7	7	3	1
n_i cum.	1	5	9	12	19	26	29	30

Ici on a 30 valeurs, la médiane est donc la demi-somme des valeurs $\frac{30}{2} = 15$ et $\frac{30}{2} + 1 = 16$. Où sont ces valeurs ? Pour le savoir, on peut construire le tableau des effectifs cumulés.

On lit que les valeurs de rang 15 et 16 sont toutes les deux égales à 30, donc la médiane vaut $\frac{30+30}{2} = 30$. Il y a au moins 50% des élèves qui travaillent 30 minutes ou moins, et il y a au moins 50% des élèves qui travaillent 30 minutes ou plus.

3) Les quartiles [3] :

Le 1^{er} quartile Q1 ($3^{ème}$ quartile Q3) : la plus petite valeur de la série supérieure ou égale à au moins 25% (75%) des valeurs.

Le rang de Q1 (Q3) est le premier entier supérieur ou égal à $\frac{n}{4}$ $\left(\frac{3n}{4}\right)$, c'est-à-dire 25% (75%) de p.

x_i	5	10	15	20	30	40	50	70
n_i	1	4	4	3	7	7	3	1
n_i cum.	1	5	9	12	19	26	29	30

- $\frac{30}{4} = 7,5$ donc Q1 est la 8e valeur.
- $\frac{3 \times 30}{4} = 22,5$ donc Q3 est la 23e valeur.

3. https://www.youtube.com/watch?v=Yjh-9nMVmEw, https://www.youtube.com/watch?v=IjsDK0ODwlw

3) Les quartiles [3] :

Le 1^{er} quartile Q1 ($3^{ème}$ quartile Q3) : la plus petite valeur de la série supérieure ou égale à au moins 25% (75%) des valeurs.

Le rang de Q1 (Q3) est le premier entier supérieur ou égal à $\frac{n}{4}$ $\left(\frac{3n}{4}\right)$, c'est-à-dire 25% (75%) de p.

x_i	5	10	15	20	30	40	50	70
n_i	1	4	4	3	7	7	3	1
n_i cum.	1	5	9	12	19	26	29	30

- $\frac{30}{4} = 7,5$ donc Q1 est la 8e valeur. C'est 15.
- $\frac{3 \times 30}{4} = 22,5$ donc Q3 est la 23e valeur.

3. https://www.youtube.com/watch?v=Yjh-9nMVmEw, https://www.youtube.com/watch?v=IjsDK0ODwlw

3) Les quartiles [3] :

Le 1^{er} quartile Q1 ($3^{ème}$ quartile Q3) : la plus petite valeur de la série supérieure ou égale à au moins 25% (75%) des valeurs.

Le rang de Q1 (Q3) est le premier entier supérieur ou égal à $\frac{n}{4}$ $\left(\frac{3n}{4}\right)$, c'est-à-dire 25% (75%) de p.

x_i	5	10	15	20	30	40	50	70
n_i	1	4	4	3	7	7	3	1
n_i cum.	1	5	9	12	19	26	29	30

- $\frac{30}{4} = 7,5$ donc Q1 est la 8e valeur. C'est 15.
- $\frac{3 \times 30}{4} = 22,5$ donc Q3 est la 23e valeur. C'est 40.

3. https://www.youtube.com/watch?v=Yjh-9nMVmEw, https://www.youtube.com/watch?v=IjsDK0ODwlw

1) L'écart-type [4] :

L'écart-type $\sigma(x)$ représente une mesure de dispersion autour de la moyenne. Plus $\sigma(x)$ est grand, plus les x_i sont dispersées.

$$\sigma(x) = \sqrt{\frac{n_1 \cdot (x_1 - \overline{x})^2 + n_2 \cdot (x_2 - \overline{x})^2 + \cdots + n_p \cdot (x_p - \overline{x})^2}{n_1 + n_2 + \cdots + n_p}}$$

x_i	5	10	15	20	30	40	50	70
n_i	1	4	4	3	7	7	3	1

On avait calculé la moyenne $\overline{x} = \frac{175}{6}$. Le calcul donne $\sigma(x) = \sqrt{\frac{1 \cdot (5 - \frac{175}{6})^2 + 4 \cdot (10 - \frac{175}{6})^2 + \cdots + 1 \cdot (70 - \frac{175}{6})^2}{30}} \approx 15,17.$

Remarque : on définit également la variance $V(x)$, c'est le carré de l'écart-type ($V(x) = \sigma(x)^2$ ou $\sigma(x) = \sqrt{V(x)}$).

4. https://www.youtube.com/watch?v=CiFoBkipJQk.

2) Propriétés de la moyenne et de l'écart-type :

a) Quand on ajoute à toutes les valeurs d'une série x un nombre a, et qu'on nomme la nouvelle série y. C'est-à-dire, si la série initiale est $x_1, x_2, \ldots, x_p$, et que la nouvelle série est

$$y_1 = x_1 + a, y_2 = x_2 + a, \ldots, y_p = x_p + a$$

- $\overline{y} = \overline{x} + a$
- $\sigma(y) = \sigma(x)$

b) Quand on multiplie toutes les valeurs d'une série x par un nombre m, et qu'on nomme la nouvelle série z. C'est-à-dire, si la série initiale est $x_1, x_2, \ldots, x_p$, et que la nouvelle série est

$$z_1 = x_1 \times m, z_2 = x_2 \times m, \ldots, z_p = x_p \times m$$

- $\overline{z} = \overline{x} \times m$
- $\sigma(z) = \sigma(x) \times m$

Moyenne et écart-type

Pour une série statistique prenant p valeurs $x_1, x_2, \ldots, x_p$ (les x_i) et où chaque valeur x_i a pour effectif n_i, les formules pour la moyenne et l'écart-type sont les suivantes :

$$\overline{x} = \frac{n_1 \cdot x_1 + n_2 \cdot x_2 + \cdots + n_p \cdot x_p}{n_1 + n_2 + \cdots + n_p}$$

$$\sigma(x) = \sqrt{\frac{n_1 \cdot (x_1 - \overline{x})^2 + n_2 \cdot (x_2 - \overline{x})^2 + \cdots + n_p \cdot (x_p - \overline{x})^2}{n_1 + n_2 + \cdots + n_p}}$$

Résumé 2

Médiane et quartiles

Pour une série statistique comprenant n valeurs, on calcule la médiane et les quartiles de la manière suivante :

- Médiane :
  - n impair : valeur centrale ; rang $\frac{n+1}{2}$
  - n pair : au milieu des deux valeurs centrales ; rang $\frac{n}{2}$ et $\frac{n+1}{2}$ (dans ce cas, ce n'est pas forcément une valeur de la série)
- Q1 : son rang est le premier entier $\geq \frac{n}{4}$
- Q3 : son rang est le premier entier $\geq \frac{3n}{4}$