

I. Introduction

Quel pourcentage des mots de la langue française connaissez-vous ?

Jouez 100 fois de suite à pile ou face avec une pièce de monnaie et notez la suite des résultats sur une feuille.

Imaginez que vous refaites l'expérience, mais notez au hasard sur le dos de la feuille 100 résultats que vous imaginez.

Une étude assez simple de la feuille permet de déterminer, avec un faible risque d'erreur, quel est le côté de la feuille contenant les résultats imaginés.

Il y a plusieurs dizaines d'années, un groupe de personnes a réussi à gagner méthodiquement à la roulette : pendant des semaines, elles ont observé et noté la liste de tous les numéros sortants.

Avec cette information elles ont joué et réussi à gagner en moyenne à la roulette.

Pouvez-vous imaginer comment elles ont fait ?

Voici 6 suites de 8 nombres. La moyenne de chaque suite est de 5.

Classez ces 6 suites par ordre croissant de dispersion de leurs nombres autour de la moyenne !

- | | | |
|----|-----------------|--|
| 1) | 5 5 5 5 5 5 5 5 | <i>dispersion = 0</i> |
| 2) | 1 1 1 1 9 9 9 9 | <i>dispersion = 4</i> |
| 3) | 1 2 3 5 5 7 8 9 | <i>dispersion = 2,692 (écart-type) ou = 2,250 (somme écarts absolus)</i> |
| 4) | 2 2 3 3 7 7 8 8 | <i>dispersion = 2,549 (écart-type) ou = 2,500 (somme écarts absolus)</i> |
| 5) | 4 4 4 4 4 4 8 8 | <i>dispersion = 1,732 (écart-type) ou = 1,375 (somme écarts absolus)</i> |
| 6) | 4 4 4 4 6 6 6 6 | <i>dispersion = 1,000 (écart-type) ou = 1,000 (somme écarts absolus)</i> |

II. Statistique descriptive, étude d'une série à valeurs discrètes

La statistique est une branche des mathématiques appliquées. **Elle est basée sur des observations d'événements réels à partir desquels on cherche à établir des hypothèses plausibles en vue de prévisions concernant des circonstances analogues.**

L'étude d'un problème statistique peut être décomposée en quatre étapes :

- le recueil des données.
- le classement et la réduction de ces données : c'est l'objet de la **statistique descriptive**.
- l'analyse des données.
- la déduction de prévisions.

La statistique descriptive se divise donc en deux étapes : le **classement**, puis la **réduction** des données.

Le classement des données consiste principalement à réorganiser les observations effectuées sous forme de tableaux et de graphiques. Pour la réduction des données, il s'agit de remplacer un grand nombre de données par quelques mesures significatives permettant de caractériser la série initiale.

Une variable statistique est dite **discrète** lorsqu'elle ne peut prendre que certaines valeurs isolées dans un intervalle de variation. Par exemple la note obtenue à une épreuve, le nombre de chevaux à l'arrivée d'une course hippique, le nombre de clients dans une salle de restaurant.

Voici un exemple qui va nous suivre tout au long de cette étude. Il va nous permettre d'explicitier les principales techniques statistiques que l'on peut appliquer aux séries discrètes :

Il est tiré du site de l'office cantonal de la statistique :

http://www.geneve.ch/statistique/statistiques/domaines/01/01_04/tableaux.asp#2 (T 01.04.2.03)

Nombre de personnes par ménage à Genève en l'an 2000, selon la taille du ménage.

taille du ménage :	1	2	3	4	5	6	7	8 ou +
nombre de personnes :	76'520	99'964	74'064	87'936	31'230	9'402	2'884	1'840

On négligera les ménages de plus de 8 personnes et considérerons qu'ils contiennent 8 personnes.

Passons à la première étape :

Organisation des données

<i>taille du ménage.</i>	<i>Nombre d'habitants</i>	<i>effectif</i>	<i>fréquence</i>	<i>effectif cumulé</i>	<i>fréquence cumulée</i>
x_i		n_i	f_i	Nc_i	Fc_i
1	76'520	76'520	0.4213	76'520	0.4213
2	99'964	49'982	0.2752	126'502	0.6965
3	74'064	24'688	0.1359	151'190	0.8324
4	87'936	21'984	0.1210	173'174	0.9534
5	31'230	6'246	0.0344	179'420	0.9878
6	9'402	1'567	0.0086	180'987	0.9965
7	2'884	412	0.0023	181'399	0.9987
8	1'840	230	0.0013	181'629	1,0000
total	383'840	181'629	1,0000		

Complétez les deux colonnes vides avec 4 chiffres après la virgule, où les notations utilisées sont :

effectif = n_i = le nombre de ménages constitués de x_i personnes.

fréquence = $f_i = \frac{n_i}{N}$, où $N = n_1 + n_2 + \dots + n_8 = \sum_{i=1}^8 n_i$.
= la proportion de ménages constitués de x_i personnes.

effectif cumulé = $Nc_i = n_1 + n_2 + \dots + n_i$.
= le nombre de ménages constitués d'au plus x_i personnes.

fréquence cumulée = $Fc_i = f_1 + f_2 + \dots + f_i = \frac{Nc_i}{N}$.

A partir de ce tableau, nous voyons qu'à Genève en l'an 2000 :

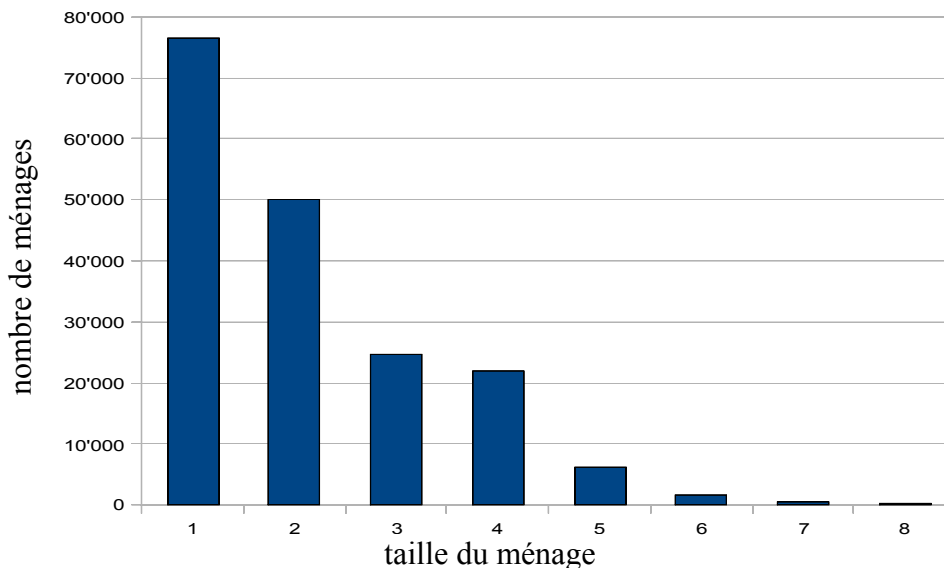
- Presque 70% des ménages sont constitués d'une ou deux personnes.
- Moins de 1,3% des ménages sont constitués de plus de 5 personnes.
- La majorité des ménages sont constitués de 1 personnes.

Dans cet exemple : $N =$ 181'629

Nombre d'effectifs : $k =$ 8.

Faisons une représentations graphiques déduites à partir des données initiales :

Histogramme du nombre de ménages en fonction de la taille du ménage.



Le tableau statistique, ainsi que les représentations graphiques, constituent la phase **d'organisation** et de **représentation** des données observées. Passons à la deuxième étape.

Réduction des données

Il s'agit de remplacer l'ensemble des données par quelques résultats représentatifs de la situation observée.

Il existe deux types de mesures que l'on peut dégager des données initiales :

- la **moyenne** de la série statistique, c'est-à-dire le nombre moyen de personne par ménage.
- la **dispersion** de la série, c'est-à-dire une quantité mesurant l'amplitude des variations autour de la moyenne.

Les mesures de la tendance centrale

Le nombre total d'habitant = $n_1 \cdot x_1 + n_2 \cdot x_2 + \dots + n_8 \cdot x_8 =$
 $76'520 \cdot 1 + 49'892 \cdot 2 + 24'688 \cdot 3 + 21'984 \cdot 4 + 6'246 \cdot 5 + 1'567 \cdot 6 + 412 \cdot 7 + 230 \cdot 8 =$ **383'840**.

Le nombre de ménages = $N =$ **181'629**.

Il y a donc en moyenne $\frac{383'840}{181'629} \approx 2,1133$ personnes par ménage. Elle se calcule aussi ainsi :

$$\frac{76'520 \cdot 1 + 49'892 \cdot 2 + 24'688 \cdot 3 + 21'984 \cdot 4 + 6'246 \cdot 5 + 1'567 \cdot 6 + 412 \cdot 7 + 230 \cdot 8}{181'629} =$$

$$= \frac{76'520}{181'629} \cdot 1 + \frac{49'892}{181'629} \cdot 2 + \dots + \frac{230}{181'629} \cdot 8 = \sum_{i=1}^8 f_i \cdot x_i \quad x_i = i = \text{taille du } i^{\text{ème}} \text{ ménage.}$$

définition :

La **moyenne** arithmétique d'une série statistique, notée \bar{x} , est définie par :

$$\bar{x} = \sum_{i=1}^k f_i \cdot x_i \quad \text{où } f_i = \frac{n_i}{N} \quad k = \text{le nombre de " } x_i \text{ ". } k = 8 \text{ dans notre exemple}$$

Dans l'exemple que nous regardons, nous avons $\bar{x} = \frac{383'840}{181'629} \approx 2,1133$. Chaque ménage est occupé, en moyenne, par 2,1133 personnes. A noter que ce résultat ne correspond à aucun ménage.

Le **mode** est la valeur qui apparaît le plus fréquemment dans la série. Pour les ménages, le mode vaut 1. Il y a 76'520 ménages avec 1 locataires.

La **médiane** partage en deux effectifs égaux les observations constituant la série préalablement rangée en ordre croissant. Dans notre exemple, si l'on range la liste dans l'ordre croissant :

$\underbrace{1,1,\dots,1}_{76'520 \text{ fois}}, \underbrace{2,2,\dots,2}_{49'982 \text{ fois}}, \underbrace{3,3,\dots,3}_{24'688 \text{ fois}}, \underbrace{4,4,\dots,4}_{21'984 \text{ fois}}, \underbrace{5,5,\dots,5}_{6'246 \text{ fois}}, \underbrace{6,\dots,6}_{1'567 \text{ fois}}, \underbrace{7,\dots,7}_{412 \text{ fois}}, \underbrace{8,\dots,8}_{230 \text{ fois}}$

Il y a 181'629 observations, la médiane est la $181'629 / 2 = 90'814^{\text{ème}}$ valeur, à savoir 2. Il y a 90'814 ménages occupés par 2 personnes ou moins, et 90'815 appartements occupés par 2 personnes ou plus.

Les mesures de dispersion

En plus de la **moyenne** et parfois de la **médiane**, une autre grandeur caractéristique importante d'une série statistique est "l'écart moyen à la moyenne".

Comment définir un tel "écart moyen à la moyenne" ?

Le tableau suivant indiquera des pistes.

Rappelons que $\bar{x} \approx 2,1133$.

x_i	n_i	f_i	$x_i - \bar{x}$	$f_i \cdot (x_i - \bar{x})$	$f_i \cdot x_i - \bar{x} $	$(x_i - \bar{x})^2$	$f_i \cdot (x_i - \bar{x})^2$
1	76'520	0,4213	-1,1133	-0,4690	0,4690	1,2394	0,5222
2	49'982	0,2752	-0,1133	-0,0312	0,0312	0,0128	0,0035
3	24'688	0,1359	0,8867	0,1205	0,1205	0,7862	0,1069
4	21'984	0,1210	1,8867	0,2284	0,2284	3,5596	0,4309
5	6'246	0,0344	2,8867	0,0993	0,0993	8,3330	0,2866
6	1'567	0,0086	3,8867	0,0335	0,0335	15,1064	0,1303
7	412	0,0023	4,8867	0,0111	0,0111	23,8798	0,0542
8	230	0,0013	5,8867	0,0075	0,0075	34,6532	0,0439
total	383'841	1,0000	19,0936	0,0000	1,0004	87,5707	1,5784

$\approx 1,2563^2$

La *calculatrice* permet de calculer facilement certains totaux comme montré au chapitre III.

Il est normal que la somme de la 5^{ème} colonne vaut toujours 0. Montrez-le !

$$\sum_{i=1}^k f_i \cdot (x_i - \bar{x}) = \sum_{i=1}^k (f_i \cdot x_i - f_i \cdot \bar{x}) = \sum_{i=1}^k f_i \cdot x_i - \bar{x} \cdot \sum_{i=1}^k f_i = \bar{x} - \bar{x} \cdot 1 = 0 \quad \text{CQFD}$$

La 6^{ème} colonne : $\sum_{i=1}^8 f_i \cdot |x_i - \bar{x}|$ est une mesure de "l'écart moyen à la moyenne", mais cette manière de faire n'est pas pratique, car la "valeur absolue" est désagréable à traiter mathématiquement.

Obtenir comme écart moyen à la moyenne la valeur nulle est tout à fait logique ! En effet, et c'est le propre de la moyenne, les écarts positifs à celle-ci compensent les écarts négatifs. Utiliser des valeurs absolues est désagréable mathématiquement.

C'est pour cette raison qu'on est amené à considérer les carrés des écarts à la moyenne.

La **variance**, notée ν , est la moyenne **des carrés des écarts à la moyenne** :

$$\nu = \sum_{i=1}^k f_i \cdot (x_i - \bar{x})^2 \quad \text{où} \quad f_i = \frac{n_i}{N}$$

Montrez que la variance peut également se calculer de la manière suivante, qui est souvent plus pratique.

$$\nu = \sum_{i=1}^k (f_i \cdot x_i^2) - \bar{x}^2 \quad \text{où} \quad f_i = \frac{n_i}{N}$$

$$\nu = \sum_{i=1}^k f_i \cdot (x_i - \bar{x})^2 = \sum_{i=1}^k f_i \cdot (x_i^2 - 2 \cdot x_i \cdot \bar{x} + \bar{x}^2) = \sum_{i=1}^k f_i \cdot x_i^2 - 2 \cdot f_i \cdot x_i \cdot \bar{x} + f_i \cdot \bar{x}^2$$

$$\nu = \sum_{i=1}^k f_i \cdot x_i^2 - 2 \cdot \bar{x} \cdot \sum_{i=1}^k f_i \cdot x_i + \bar{x}^2 \cdot \sum_{i=1}^k f_i = \sum_{i=1}^k f_i \cdot x_i^2 - 2 \cdot \bar{x} \cdot \bar{x} + \bar{x}^2 = \sum_{i=1}^k f_i \cdot x_i^2 - \bar{x}^2 \quad \text{CQFD}$$

Souvent les grandeurs x_i représentent une longueur ou une masse ou un temps, c'est-à-dire qu'elles ont une unité telle que le mètre, le kilogramme ou la seconde. **La variance aura le carré de cette unité** et ne pourra donc pas être comparée aux x_i . On ne peut pas comparer des mètres avec des mètres carrés.

C'est la raison pour laquelle on prend la racine carrée de la variance, appelée **écart type**, qui possède donc bien la même unité que x_i .

L'**écart type**, noté σ , est la racine carrée de la variance. $\sigma = \sqrt{\nu}$ par définition.

Dans l'exemple des ménages, nous avons : $\nu \approx 1,5784$ et $\sigma \approx \sqrt{1,5784} \approx 1,2563$.

Une autre grandeur que l'on peut rattacher à la série est le pourcentage des observations comprises dans l'intervalle $[\bar{x} - \sigma ; \bar{x} + \sigma]$.

C'est une façon de mesurer la concentration des observations autour de la moyenne.

Dans notre exemple, cela n'a pas beaucoup de sens, car il y a trop peu de différent type de ménages (seulement 8).

Remarque :

Il est courant que le pourcentage des x_i se trouvant dans l'intervalle $[\bar{x} - \sigma ; \bar{x} + \sigma]$ soit proche de 68 %.

III. Utilisation de la calculatrice pour effectuer des statistiques


Prenons un autre exemple. (La calculatrice du collège est dépassée par les données réelles !)


x_i	n_i
1	8
2	14
3	7
4	12
5	3
6	1
total	$n = 45$


Entrer dans le mode "Statistique" : **2nd** **STAT**


Choisir l'option : "1-VAR" : **=**

Pour débiter l'entrée des données : **DATA**


A l'affichage "X₁ =" taper la valeur de x_1 (dans cet exemple, $x_1 = 1$), puis 

A l'affichage "FRQ=1" taper la valeur de n_1 (dans cet exemple, $n_1 = 8$), puis 


A l'affichage "X₂ =" taper la valeur de x_2 (dans cet exemple, $x_2 = 2$), puis 

A l'affichage "FRQ=1" taper la valeur de n_2 (dans cet exemple, $n_2 = 14$), puis 

etc. jusqu'à la dernière valeur...

A l'affichage "X₆ =" taper la valeur de x_6 (dans cet exemple, $x_6 = 6$), puis 

A l'affichage "FRQ=1" taper la valeur de n_6 (dans cet exemple, $n_6 = 1$), puis **=** pour terminer.

Vous pouvez revenir en arrière pour corriger avec la touche 

Pour obtenir les résultats, taper sur **STATVAR**

n indique la somme des n_i : $n = \sum_{i=1}^k n_i$. Dans cet exemple, $k = 6$; $n = 45$. C'est un contrôle de saisie !

\bar{x} indique la moyenne : $\bar{x} = \frac{1}{N} \cdot \sum_{i=1}^k n_i \cdot x_i$. Dans cet exemple, $\bar{x} = 2,8$.

Sx indique un écart à la moyenne que nous n'utiliserons pas : $Sx = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^k n_i \cdot (x_i - \bar{x})^2}$

σx indique l'écart type : $\sigma x = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^k n_i \cdot (x_i - \bar{x})^2}$. Dans cet exemple, $\sigma x \approx 1,30979218$

Σx indique la somme : $\Sigma x = \sum_{i=1}^k n_i \cdot x_i$. Donc $\frac{\Sigma x}{n} = \bar{x}$. Dans cet exemple, $\Sigma x = 126$.

Σx^2 indique la somme : $\Sigma x^2 = \sum_{i=1}^k n_i \cdot x_i^2$. On a : $\sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} = \sigma x$. Dans cet exemple, $\Sigma x^2 = 430$.

Pour sortir du mode statistique : **2nd** **EXIT STAT** et confirmer avec **=**.

Référence : "T 01.04.2.03" du site Web : <http://www.geneve.ch/statistique/statistiques/domaines/>
"Ménages privés et population résidante vivant dans ces ménages, selon la taille du ménage, depuis 1960."

Office cantonal de la statistique - OCSTAT

Ménages privés et population résidante vivant dans ces ménages, selon la taille du ménage, depuis 1960.
Situation au début décembre.

		T01.04.2.03 Canton de Genève								
		Ménage de ... personnes ⁽¹⁾								Total
		1	2	3	4	5	6	7	8 ou plus	
1960	Nombre de ménages	18'418	29'367	19'315	13'198	5'933	2'424	1'049	920	90'624
	Répartition en pour mille	203	324	213	146	65	27	12	10	1'000
	Nombres de personnes	18'418	58'734	57'945	52'792	29'665	14'544	7'343	8'306	247'747
	Répartition en pour mille	74	237	234	213	120	59	30	34	1'000
1970	Nombre de ménages	35'200	41'187	24'204	19'036	6'674	1'915	665	355	129'236
	Répartition en pour mille	272	319	187	147	52	15	5	3	1'000
	Nombres de personnes	35'200	82'374	72'612	76'144	33'370	11'490	4'655	3'095	318'940
	Répartition en pour mille	110	258	228	239	105	36	15	10	1'000
1980	Nombre de ménages	60'897	44'920	23'088	21'536	5'114	1'113	200	77	156'945
	Répartition en pour mille	388	286	147	137	33	7	1	0	1'000
	Nombres de personnes	60'897	89'840	69'264	86'144	25'570	6'678	1'400	645	340'438
	Répartition en pour mille	179	264	203	253	75	20	4	2	1'000
1990	Nombre de ménages	66'484	49'565	26'179	21'759	5'010	1'043	215	106	170'361
	Répartition en pour mille	390	291	154	128	29	6	1	1	1'000
	Nombres de personnes	66'484	99'130	78'537	87'036	25'050	6'258	1'505	891	364'891
	Répartition en pour mille	182	272	215	239	69	17	4	2	1'000
2000	Nombre de ménages	76'520	49'982	24'688	21'984	6'246	1'567	412	212	181'611
	Répartition en pour mille	421	275	136	121	34	9	2	1	1'000
	Nombres de personnes	76'520	99'964	74'064	87'936	31'230	9'402	2'884	1'841	383'841
	Répartition en pour mille	199	260	193	229	81	24	8	5	1'000

⁽¹⁾ En 1960, 1970 et 2000, un ménage est constitué par l'ensemble des personnes vivant dans un même logement. En 1980 et 1990 par contre, les personnes sous-louant une chambre (sous-locataires) constituent des ménages distincts. Entre 1970 et 1980, l'augmentation du nombre de ménages d'une personne imputable à ces transferts est estimée à 7 300

Source : Office fédéral de la statistique - Recensements fédéraux de la population et des logements

Index

A venir peut-être ???