Indicateurs (analyse univariée)

Rappel : Une variable statistique (ou caractère statistique) est ce qui est observé ou mesuré sur les individus de la population statistique.

Notations : une variable statistique est en général notée par une lettre minuscule (x par exemple) et la valeur ou modalité observée chez le ième individu se notera par cette même lettre indicée par i (xi par exemple).

|  |
| --- |
| poids |
| 88 |
| 80 |
| 75 |
| 91 |
| 83 |
| … |

|  |
| --- |
| Variable x |
| x1 |
| x2 |
| x3 |
| x4 |
| x5 |
| … |

La suite est appelée **série statistique**

## Indicateurs de tendance centrale (ou de position)

Ces indicateurs correspondent à une valeur « centrale » autour de laquelle les valeurs de la série sont censées se concentrer.

### La moyenne (arithmétique)

Si tous les individus avaient la même valeur, la moyenne correspond à la valeur qui devrait être observée chez tous les individus, sans que la somme de toutes les valeurs changent.

Un inconvénient de la moyenne : son manque de « robustesse », c’est-à-dire sa sensibilité aux changements de valeurs dans la série et en particulier aux valeurs extrêmes.

### La médiane

La médiane est une valeur telle que la moitié des individus ont une valeur inférieure à la médiane. La médiane sépare donc la population en deux sous-populations de même effectif. *Nous avons vu en TP l’extension de cette définition à celle de quantile, décile d’ordre α% ou de quartile.*

Attention quand la taille de la population est petite ou quand la variable mesurée prend plusieurs fois la même valeur (valeurs entières par exemple), il n’est pas toujours possible d’obtenir deux sous-populations représentant 50% de la population.

Contrairement à la moyenne, la médiane est un indicateur « robuste », peu sensible aux variations de valeurs dans la série et aux valeurs extrêmes.

#### Exemple de situation où la médiane et la moyenne sont assez éloignées

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Distribution des salaires mensuels nets en 2013 et évolution entre 2012 et 2013 en euros constants** | | | | | | |
| **Déciles** | **Ensemble** | | **Hommes** | | **Femmes** | |
| **2013** | **Évolution (%)** | **2013** | **Évolution (%)** | **2013** | **Évolution (%)** |
| D1 | 1 200 | -0,6 | 1 254 | -0,9 | 1 154 | -0,3 |
| D2 | 1 342 | -0,5 | 1 415 | -0,6 | 1 268 | -0,3 |
| D3 | 1 471 | -0,4 | 1 559 | -0,5 | 1 374 | -0,2 |
| D4 | 1 609 | -0,3 | 1 709 | -0,3 | 1 485 | -0,1 |
| **D5 ou Médiane** | **1 772** | **-0,1** | **1 882** | **-0,2** | **1 619** | **0,1** |
| D6 | 1 974 | 0,0 | 2 100 | -0,1 | 1 794 | 0,3 |
| D7 | 2 244 | 0,0 | 2 405 | 0,0 | 2 029 | 0,3 |
| D8 | 2 682 | 0,0 | 2 921 | -0,1 | 2 368 | 0,3 |
| D9 | 3 544 | -0,2 | 3 892 | -0,1 | 3 036 | 0,0 |
| C95 | 4 526 | -0,2 | 5 030 | -0,4 | 3 756 | 0,2 |
| C99 | 8 061 | -0,5 | 9 253 | -0,6 | 6 053 | 0,1 |
| **Moyenne** | **2 202** | **-0,3** | **2 389** | **-0,4** | **1 934** | **0,0** |
| Champ : France, salariés en EQTP du privé et des entreprises publiques, y compris les bénéficiaires de contrats aidés. Sont exclus les apprentis, les stagiaires, les salariés agricoles et les salariés des particuliers employeurs. | | | | | | |
| *Source : Insee, DADS, fichier semi-définitif.* | | |  |  |  |  |

**Commenter les valeurs 1882, 2405 et 2389**

1882 :

2405 :

2389 :

### LE MODE

Le mode est la valeur la plus fréquente de la série statistique, on parle aussi de valeur modale. Attention, quand la variable est continue, on s’intéressera plutôt à la classe modale qu’au mode. A condition d’avoir mis en classe la variable continue.

Lorsque qu’une variable se répartit de façon symétrique, la médiane, la moyenne et le mode sont proches.

## Indicateurs de dispersion

Ils sont censés compléter l’information donnée par une caractéristique centrale (moyenne,…) en « mesurant » la tendance qu’ont les valeurs de la série à se concentrer ou se disperser autour de cet indicateur.

### Indicateurs donnés par des intervalles

**Intervalle Min-max** : il permet de calculer l’étendue Max-Min

**Intervalle Interquartile** : [Q1 ;Q3] représente 50% de la population

### Variance et ecart-type

La variance et l’écart-type résument la dispersion de valeurs de la série autour de sa moyenne.

La variance correspond à la moyenne des carrés des écarts de ces valeurs à la moyenne.

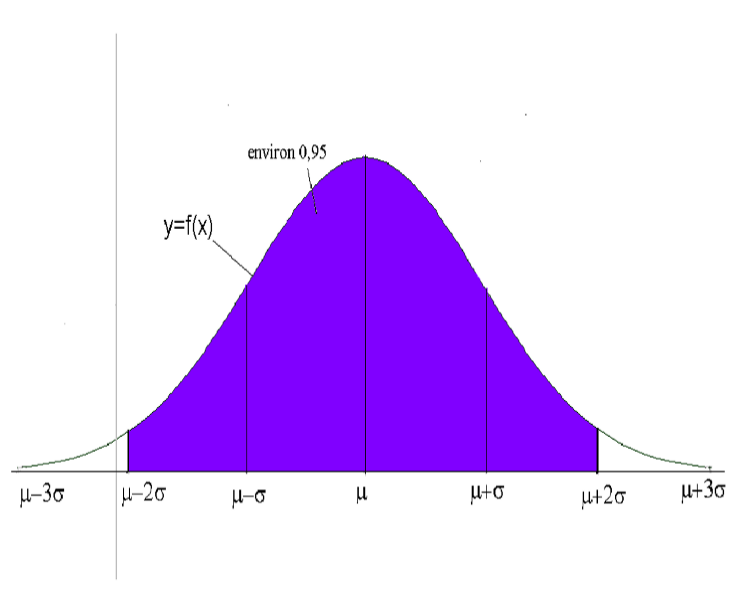
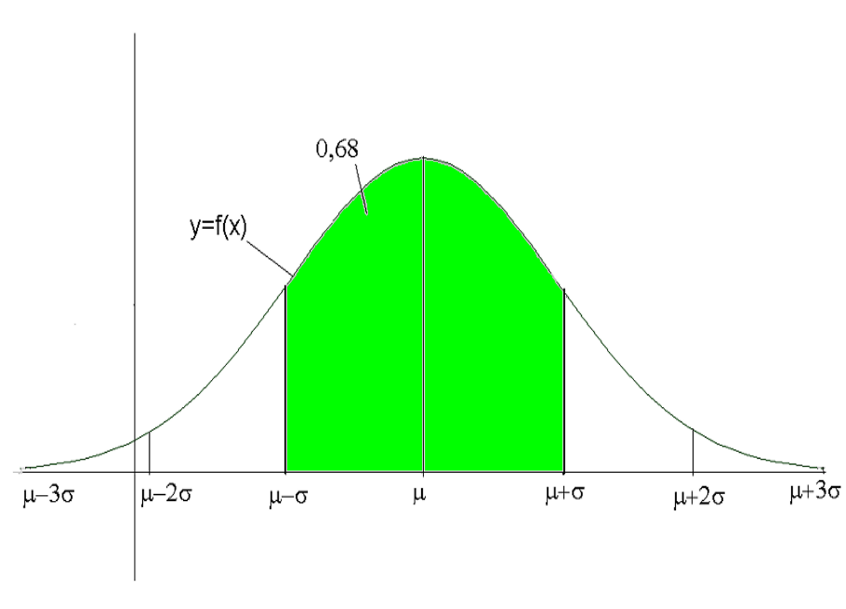
L’écart-type est la racine carrée de la variance. On le calcule pour avoir un indicateur de **même dimension** que la variable.

Il n’y a pas de « règle » d’interprétation de l’écart-type (ou la variance) :

* On peut donc l’utiliser pour **comparer** la dispersion de plusieurs séries :

*Par exemple dans un devoir, l’écart-type varie en général entre 2 et 4. La comparaison des écart-types permet de comparer les dispersions des notes.*

* Dans le cas particulier où la série observée serait la réalisation d’une **loi normale (distribution normale)**, l’interprétation est vraiment possible

**

## *Autres types d’indicateurs, les indicateurs de forme : asymétrie, dissymétrie, aplatissement,…*

# Calculs à partir de fréquences et des effectifs

Dans les formules précédentes chaque valeur xi est associée à un individu (le ième).

Il y a n valeurs xi, n étant le nombre d’individu.

Il arrive que les données de départ ne soient pas les données « brutes » (individuelles) mais un tableau de fréquence de la varia

|  |  |  |
| --- | --- | --- |
| Poids | Effectifs | Fréquences |
| … | … | … |
| 97 | 22 | 0,01 |
| 98 | 48 | 0,03 |
| 99 | 21 | 0,01 |
| 100 | 66 | 0,04 |
| 101 | 23 | 0,01 |
| 102 | 47 | 0,03 |
| 103 | 27 | 0,02 |
| 104 | 43 | 0,03 |
| … | … | … |

L’indice ***i*** de la variable correspond non plus à une numérotation des individus mais à une numérotation des valeurs prises par la variable.

Ainsi, dans la variable Poids des données rugby, s’il y a individus, le nombre de valeurs prises par le poids est ***p=72*** :

Pour les calculs de moyenne, on utilise alors les effectifs ou les fréquences comme pondération :