# Cours 2 : Variables quantitatives

# Exemple

## Données de population des communes françaises (INSEE 2017)

- $n = 34995$ communes
- $x_i$ = population totale de la commune
- Données :

| $i$ | | $x_i$ |
|---|---|---|
| 1 | L'Abergement-Clémenciat | 794 |
| 2 | L'Abergement-de-Varey | 249 |
| 3 | Ambérieu-en-Bugey | 14428 |
| 4 | Ambérieux-en-Dombes | 1723 |
| 5 | Ambléon | 117 |
| 6 | Ambronay | 2841 |
| ⋮ | ⋮ | ⋮ |

(Lyon, Marseille et Paris apparaissent par arrondissements)

# Exemple

## Données de population des départements français : (INSEE 2017)

- $n = 100$ département (96 + 4)
- $x_i$ = population totale de la commune
- Données :

| $i$ | | $x_i$ |
|---|---|---|
| 1 | Ain | 659180 |
| 2 | Aisne | 546527 |
| 3 | Allier | 347035 |
| 4 | Alpes-de-Haute-Provence | 168381 |
| 5 | Hautes-Alpes | 145883 |
| 6 | Alpes-Maritimes | 1097496 |
| ⋮ | ⋮ | ⋮ |

# Critère de centralité

- Données :
  - Ensemble des observations $\{x_i\}_{1 \leq i \leq n}$
  - Vecteur des observations $\mathbf{x} = [x_1 \; x_2 \; \ldots \; x_i \; \ldots \; x_n]^T$
- Objectif :
  - Résumer l'ensemble des observations $\mathbf{x}$ par une valeur typique ou «centrale».
- Critère de centralité :
  - Une valeur $u$ est centrale si elle est «bien au centre» de l'ensemble des $x_i$. Il nous faut donc définir un *critère mesurant la centralité* de $u$.

## Exemples de critères de centralité

- L'*écart absolu* à la valeur $u$ est défini par

$$C_1(u) = \sum_{i=1}^{n} |x_i - u|.$$

- L'*écart quadratique* à la valeur $u$ est défini par

$$C_2(u) = \sum_{i=1}^{n} (x_i - u)^2.$$

Remarque : les fonctions $C_1$ et $C_2$ dépendent également du vecteur des observations $\mathbf{x}$ (on note $C_1(u)$ plutôt que $C_1(u; \mathbf{x})$ car $u$ inconnu et $\mathbf{x}$ donné).

On considère d'abord l'écart quadratique dont on verra qu'il est mathématiquement confortable.

**Proposition 1 (Minimisation de l'écart quadratique)**

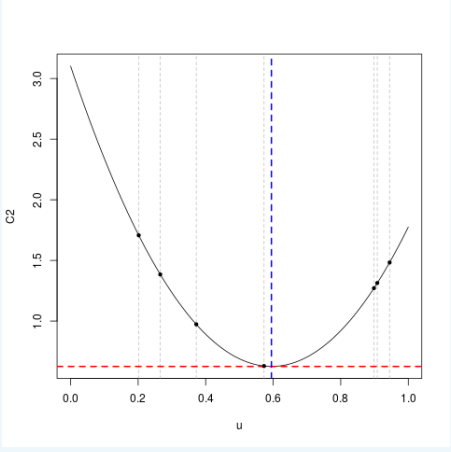
*La moyenne*

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

*minimise l'écart quadratique* $C_2(u)$.

## Exemple fictif : $n = 7$

$$\mathbf{x}^\mathsf{T} = [0.20\ 0.27\ 0.37\ 0.57\ 0.90\ 0.91\ 0.94]$$




$$\overline{x} = 0.595$$

## Écart quadratique : démonstration

On cherche le minimum en $u$ de la fonction

$$C_2(u) = \sum_{i=1}^{n}(x_i^2 - 2ux_i + u^2) = \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2ux_i + \sum_{i=1}^{n} u^2$$
$$= \left(\sum_{i=1}^{n} x_i^2\right) - 2u\left(\sum_{i=1}^{n} x_i\right) + nu^2 = Q - 2uS + nu^2$$

en notant

$$S = \sum_{i=1}^{n} x_i \qquad \text{et} \qquad Q = \sum_{i=1}^{n} x_i^2.$$

Pour cela on calcule sa dérivée

$$C_2'(u) = -2S + 2nu$$

qui est nulle si et seulement si

$$2S = 2nu \qquad \Leftrightarrow \qquad u = \frac{S}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}.$$

□

Utile pour une démonstration alternative du fait que $\overline{x}$ minimise $C_2(u)$.

**Proposition 2 (Formule de Huygens)**

*Pour tout réel $u$, on a*

$$\sum_{i=1}^{n}(x_i - u)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(u - \overline{x})^2.$$

# Formule de Huygens : démonstration

On a

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\overline{x}+\overline{x}-u)^2 &= \sum_{i=1}^{n}\left((x_i-\overline{x})^2+2(x_i-\overline{x})(\overline{x}-u)+(\overline{x}-u)^2\right)\\ &= \sum_{i=1}^{n}(x_i-\overline{x})^2+2(\overline{x}-u)\underbrace{\sum_{i=1}^{n}(x_i-\overline{x})}_{=0}+n(\overline{x}-u)^2\end{aligned}$$

car

$$\sum_{i=1}^{n}(x_i-\overline{x})=\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\overline{x}=\sum_{i=1}^{n}x_i-n\overline{x}=0$$

par définition de la moyenne $\overline{x}$. $\square$

## Démonstration alternative de la proposition 1

La formule de Huygens permet de montrer directement que $\overline{x}$ minimise $C_2(u)$. En effet, on a pour tout $u$

$$C_2(u) = \sum_{i=1}^{n}(x_i - u)^2 = C_2(\overline{x}) + n(\overline{x} - u)^2,$$

et donc, pour tout $u$,

$$C_2(u) \geq C_2(\overline{x}).$$

□

**Proposition 3 (Transformation linéaire)**

*Soit la variable $y$ telle que $y_i = a + bx_i$ (pour $1 \leq i \leq n$), on a*

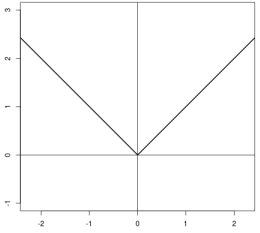
$$\overline{y} = a + b\overline{x}.$$

Démonstration :

$$\begin{aligned}
\overline{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} a + \frac{1}{n}b\sum_{i=1}^{n} x_i \\
&= \frac{1}{n}na + b\frac{1}{n}\sum_{i=1}^{n} x_i = a + b\overline{x}
\end{aligned}$$

□

- Difficulté : la fonction $f(x) = |x|$ n'est pas partout dérivable.




**Définition 1 (Statistique d'ordre $i$)**

Pour des données réelles $(x_i)_{1\leq i\leq n}$, on note $(x_{(i)})_{1\leq i\leq n}$ le même jeu de données ré-ordonné, c'est à dire tel que

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(i)} \leq \cdots \leq x_{(n)}.$$

$x_{(i)}$ est appelé la *statistique d'ordre $i$* de l'ensemble $\{x_i\}_{1\leq i\leq n}$.

**Proposition 4**

- *Si $n$ est pair ($n = 2k$), toute valeur comprise entre $x_{(k)}$ et $x_{(k+1)}$ minimise l'écart absolu $C_1(u)$.*
- *Si $n$ est impair ($n = 2k + 1$), $x_{(k+1)}$ minimise l'écart absolu $C_1(u)$.*

- Remarques :
  - Quand $n$ est impair ($n = 2k + 1$), la valeur qui minimise $C_1(u)$ est donc la médiane

$$m = x_{(k+1)}.$$

  - Par convention, quand $n$ est pair ($n = 2k$), on pose

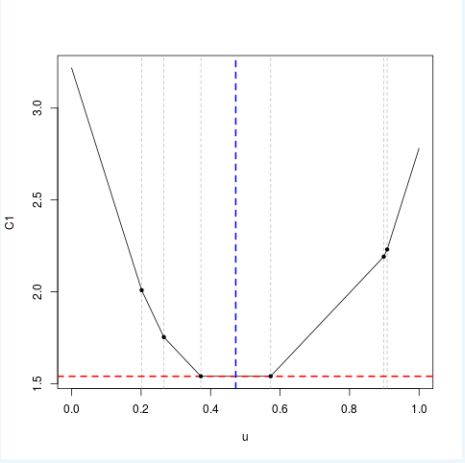
$$m = \frac{x_{(k)} + x_{(k+1)}}{2}.$$

# Écart absolu et médiane

## Exemples fictifs :

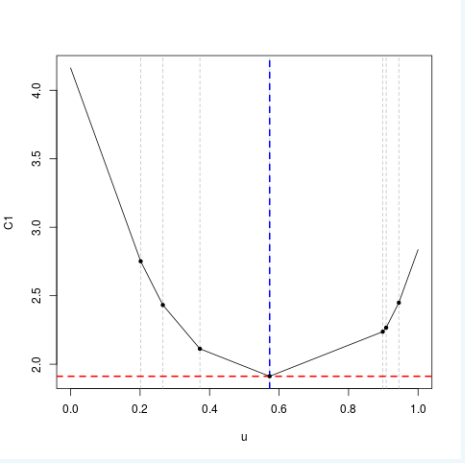
$n = 6$

$\mathbf{x}^\intercal = [0.20\ 0.27\ 0.37\ 0.57\ 0.90\ 0.91]$




$n = 7$

$\mathbf{x}^\intercal = [0.20\ 0.27\ 0.37\ 0.57\ 0.90\ 0.91\ 0.94]$



# Écart absolu : démonstration

Démonstration de la proposition 4

- La fonction valeur absolue n'est pas partout dérivable.
- Sa dérivée vaut $-1$ sur $\mathbb{R}^{-*}$, $+1$ sur $\mathbb{R}^{+*}$ mais n'est pas définie en 0.
- On va déterminer son optimum via son tableau de variation.
- On réécrit le critère $C_1(u)$ au moyen des statistique d'ordre :

$$C_1(u) = \sum_{i=1}^{n} |x_i - u| = \sum_{i=1}^{n} |x_{(i)} - u|.$$

# Écart absolu : démonstration

- Cas où $n$ est pair : $n = 2k$.
- La dérivée de la valeur absolue est égale au signe
- Pour tout $u \in \mathbb{R} \setminus \{x_i\}_{1 \leq i \leq n}$,

$$C_1'(u) = |\{i : x_{(i)} > u\}| - |\{i : x_{(i)} < u\}|.$$

- Comme la fonction $C_1(u)$ est continue sur tout $\mathbb{R}$, elle est :
  - strictement décroissante jusqu'en $x_{(k)}$
  - strictement croissante à partir de $x_{(k+1)}$
  - minimale et constante sur tout l'intervalle $[x_{(k)}, x_{(k+1)}]$ et la médiane
- $m = \frac{1}{2}\left(x_{(\lfloor (n+1)/2 \rfloor)} + x_{(\lfloor n/2 \rfloor + 1)}\right) = \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right)$ fait partie de cet intervalle.

# Écart absolu : démonstration

- Cas où $n$ est impair : $n = 2k + 1$
- On peut récrire $C_1(u)$ sous la forme

$$C_1(u) = \sum_{i=1}^{k} |x_{(i)} - u| + \sum_{i=k+2}^{n} |x_{(i)} - u| + |x_{(k+1)} - u|.$$

- La somme des deux premiers termes

$$\sum_{i=1}^{k} |x_{(i)} - u| + \sum_{i=k+2}^{n} |x_{(i)} - u|$$

  correspond à la fonction $C_1$ où on aurait retiré la données $x_{(k+1)}$.
- On se retrouve dans le cas $n$ pair
- La fonction est minimale dans l'intervalle $[x_{(k)}; x_{(k+2)}]$
- Reste la fonction $u \mapsto |x_{(k+1)} - u|$ :
  - partout positive
  - admet un unique minimum en $u = x_{(k+1)} \in [x_{(k)}; x_{(k+2)}]$.
- La fonction $C_1(u)$ admet donc un unique mimimum en $u = x_{(k+1)}$. $\square$

## Moyennes et médiane :

**Population par commune :**

| moyenne | médiane |
|---|---|
| 1936.3 | 468 |

**Population par département :**

| moyenne | médiane |
|---|---|
| 677610.9 | 546162.5 |

## Communes et départements

**Population par commune :** population de Toulouse = 484 809

| | moyenne | médiane |
|---|---|---|
| avec Toulouse | 1 936.3 | 468 |
| sans Toulouse | 1 922.5 | 468 |

**Population par département :** population du Nord = 2 635 255

| | moyenne | médiane |
|---|---|---|
| avec le Nord (59) | 677 610.9 | 546 162.5 |
| sans le Nord (59) | 657 836.7 | 545 798.0 |

La moyenne et la médiane dépendent toutes de l'ensemble des $\{x_i\}$, mais la moyenne varie plus quand on ajoute ou retranche une valeur «extrême».

1. Valeur centrale
   - Critère de centralité
   - Écart quadratique
   - Écart absolu

2. Dispersion
   - Déviation absolue (par rapport à la médiane)
   - Variance, écart type

3. Distribution, histogramme
   - Histogramme
   - Quantiles, intervalle inter-quartile.
   - Boîte à moustaches (*box-plot*)

# Dispersion

- Il est souvent utile de rendre également compte de la dispersion des valeurs autour de la valeur centrale.
- Des valeurs naturelles sont fournies par les critères (par exemple $C_1$ ou $C_2$) qu'on a choisit de minimiser pour déterminer les valeurs centrales.

**Définition 2 (Déviation absolue (par rapport à la médiane))**

$$\mathsf{mad}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} |x_i - m| = \frac{1}{n} C_1(m).$$

('MAD' = *median absolute deviation*).

**Définition 3 (Variance)**

$$\mathrm{var}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}C_2(\overline{x})$$

$\mathrm{var}(\mathbf{x})$ n'est pas homogène à la variable $x$ :

| | | |
|---|---|---|
| $x$ en mètres (m) | $\Rightarrow$ | $\mathrm{var}(\mathbf{x})$ en mètres carrés ($\mathrm{m}^2$) |
| $x$ en secondes (s) | $\Rightarrow$ | $\mathrm{var}(\mathbf{x})$ en secondes carrées ($\mathrm{s}^2$) |

**Définition 4 (Écart-type)**

L'écart-type de $\mathbf{x}$ est la racine carrée de sa variance : $\sqrt{\mathrm{var}(\mathbf{x})}$, qui est homogène à $x$.

**Proposition 5 (Variance d'une transformation linéaire)**

*Si on applique la transformation linéaire $y_i = a + bx_i$ pour $1 \leq i \leq n$, la variance des données transformées vaut*

$$\text{var}(\mathbf{y}) = b^2 \, \text{var}(\mathbf{x}).$$

# Variance d'une transformation linéaire : démonstration

Démonstration de la proposition 5
La proposition 3 assure que

$$\overline{y} = a + b\overline{x}.$$

On a donc

$$\begin{aligned}
\text{var}(\mathbf{y}) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}((a + bx_i) - (a + b\overline{x}))^2 = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i - a - b\overline{x})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(b(x_i - \overline{x}))^2 = \frac{b^2}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&= b^2\,\text{var}(\mathbf{x})
\end{aligned}$$

□

**Proposition 6 (Formule alternative de la variance)**

*La variance est la différence entre la moyenne des carrés et le carré de la moyenne :*

$$\text{var}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - (\overline{x})^2.$$

Démonstration :

$$\begin{aligned}
\text{var}(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\overline{x} + \overline{x}^2) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2 + \sum_{i=1}^{n} -2x_i\overline{x} + \sum_{i=1}^{n}\overline{x}^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2 - 2\overline{x}\sum_{i=1}^{n}x_i + n\overline{x}^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2 - 2\overline{x}n\overline{x} + n\overline{x}^2\right) = \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \overline{x}^2
\end{aligned}$$

**Définition 5 (Norme d'un vecteur)**

La norme d'un vecteur $\mathbf{x}$, noté $\|\mathbf{x}\|$ est la racine carré de son produit scalaire avec lui-même :

$$\|\mathbf{x}\|^2 = \mathbf{x}^\intercal \mathbf{x} = \sum_{i=1}^{n} (x_i)^2 \qquad \rightarrow \qquad \|\mathbf{x}\| = \sqrt{\mathbf{x}^\intercal \mathbf{x}} = \sqrt{\sum_{i=1}^{n} (x_i)^2}.$$

**Proposition 7 (Formule vectorielle de la variance)**

$$\text{var}(\mathbf{x}) = \frac{1}{n} \|\mathbf{x} - \overline{x}\mathbf{1}\|^2.$$

$$(\mathbf{x} - \overline{x}\mathbf{1})^\intercal = [x_1 - \overline{x}, \quad x_2 - \overline{x}, \quad \ldots, \quad x_n - \overline{x}]$$

# Exemple

## Communes et départements

**Communes :**

| moyenne | variance | écart-type | médiane | MAD |
|---|---|---|---|---|
| 1 936.3 | 75 624 488.9 | 8 696.2 | 468 | 487.8 |

**Départements :**

| moyenne | variance | écart-type | médiane | MAD |
|---|---|---|---|---|
| 677 610.9 | 261 045 643 813.1 | 510 926.3 | 546 162.5 | 401 168.6 |

Là encore, l'indicateur lié à $C_1$ (MAD) est plus robuste que celui lié à $C_2$ (écart-type)

## En écartant les valeurs «extrêmes»

| | écart-type | MAD |
|---|---|---|
| avec Toulouse | 8 696.2 | 487.8 |
| sans Toulouse | 8 304.4 | 487.8 |

| | écart-type | MAD |
|---|---|---|
| avec le Nord | 510 926.3 | 401 168.6 |
| sans le Nord | 473 899.4 | 394 024.7 |

1. Valeur centrale
   - Critère de centralité
   - Écart quadratique
   - Écart absolu
2. Dispersion
   - Déviation absolue (par rapport à la médiane)
   - Variance, écart type
3. Distribution, histogramme
   - Histogramme
   - Quantiles, intervalle inter-quartile.
   - Boîte à moustaches (*box-plot*)

**Objectif :** rendre compte de la répartition de l'ensemble des observations $\{x_i\}_{1 \leq i \leq n}$.

## Construction d'un histogramme régulier

1. On se donne un nombre d'intervalles $k$ ;
2. On se donne des bornes $x_{\min}$ et $x_{\max}$ ;
3. on définit la largeur d'un intervalle comme

$$\delta = \frac{x_{\max} - x_{\min}}{k};$$

4. on définit $k+1$ valeurs seuils $t_\ell$ $(0 \leq \ell \leq k)$, telles que

$$t_\ell = x_{\min} + \ell\delta;$$

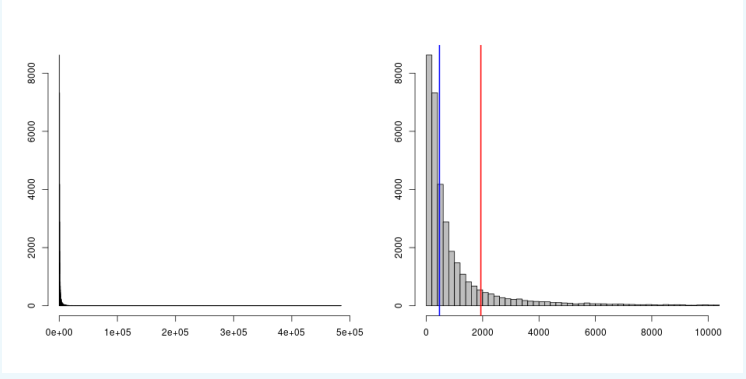
5. on associe à chaque intervalle le nombre d'observations qu'il contient :

$$n_\ell = |\{i : x_i \in [t_{\ell-1}; t_\ell]\}|.$$

## Population par commune

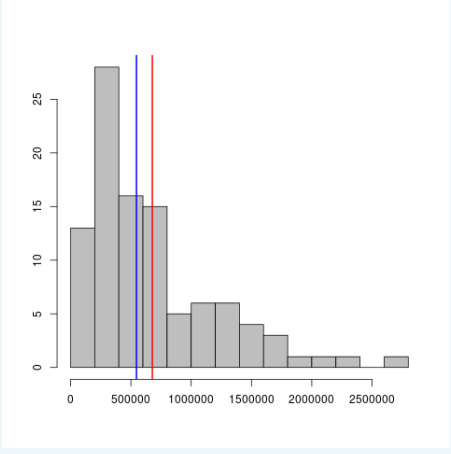
($\overline{x}$ : moyenne, $m$ : médiane)

Zoom sur les petites communes



## Population par département

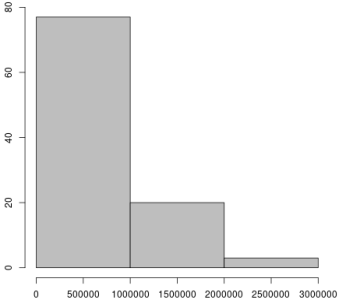
$k = 14, \quad x_{\min} = 0, \quad x_{\max} = 2\,800\,000, \quad \delta = 200\,000$
($\overline{x}$ : moyenne, $m$ : médiane)



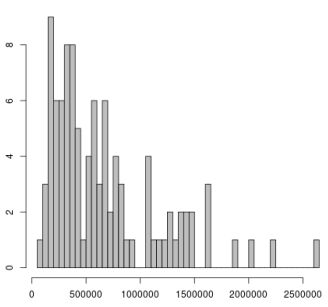
## Population par département

$(n = 100)$

$k = 3$ intervalles




$k = 50$ intervalles



**Objectif :** représenter la distribution de la variable $x$ à partir de valeurs qui sépare les observations selon des proportions prédéfinies.

**Quantile**

Pour tout $u \in [0, 1]$, le *quantile d'ordre u*, noté $q_u$ est tel que

$$\frac{|\{i : x_i \leq q_u\}|}{n} \leq u < \frac{|\{i : x_i \leq q_u\}| + 1}{n}$$

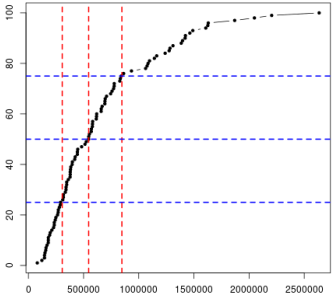
**Exemples**

- Le quantile d'ordre $1/2$ est la médiane : $m = q_{50\%}$.
- Le quantile d'ordre $0.1$ ($q_{10\%}$) laisse $10\%$ des données à sa gauche et $90\%$ à sa droite.

## Population par département

Quartiles = quantiles d'ordre 25%, 50%, 75%

| 25% | 50% | 75% |
|---|---|---|
| 306481 | 546162 | 847227 |



## Communes et départements

Quartiles = quantiles d'ordre 25%, 50%, 75%

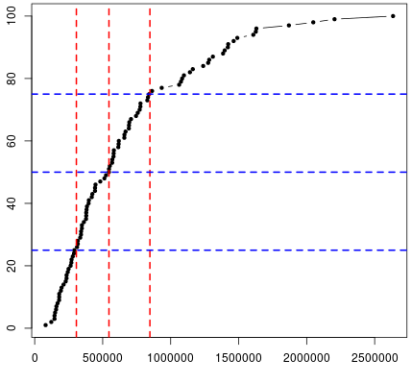
**Communes :**

| 25% | 50% | 75% |
|---|---|---|
| 203 | 468 | 1184 |

**Départements :**

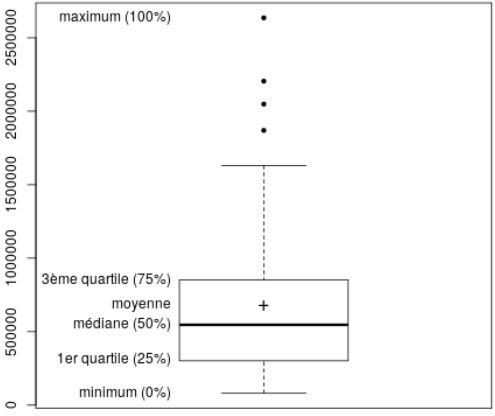
| 25% | 50% | 75% |
|---|---|---|
| 306 481 | 546 162 | 847 227 |

Départements



# Boîte à moustaches (*box-plot*)

Le graphique en *box-plot* reprend différents indicateurs vus jusqu'ici.



# *Box-plot* : distribution très déséquilibrée

## Population par commune