

Statistique descriptive à une variable

Vidéos de ce cours à <https://ophysis.net/course/section.php?id=110>

1 Longueurs des pétales de 50 fleurs iris setosa

1.1 Origine des données

Nous utilisons 50 mesures extraites d'une [expérience menée en 1935 par E. ANDERSON](#).

Ces données sont disponibles dans l'environnement R, [un logiciel libre pour effectuer des calculs statistiques et des graphiques](#).

```
[1] 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 1.5 1.6 1.4 1.1 1.2 1.5 1.3 1.4 1.7
[20] 1.5 1.7 1.5 1.0 1.7 1.9 1.6 1.6 1.5 1.4 1.6 1.6 1.5 1.5 1.4 1.5 1.2 1.3 1.4
[39] 1.3 1.5 1.3 1.3 1.3 1.6 1.9 1.4 1.6 1.4 1.5 1.4
```

Chaque nombre est la mesure, **en centimètres**, de la longueur des pétales d'un des 50 iris observés.

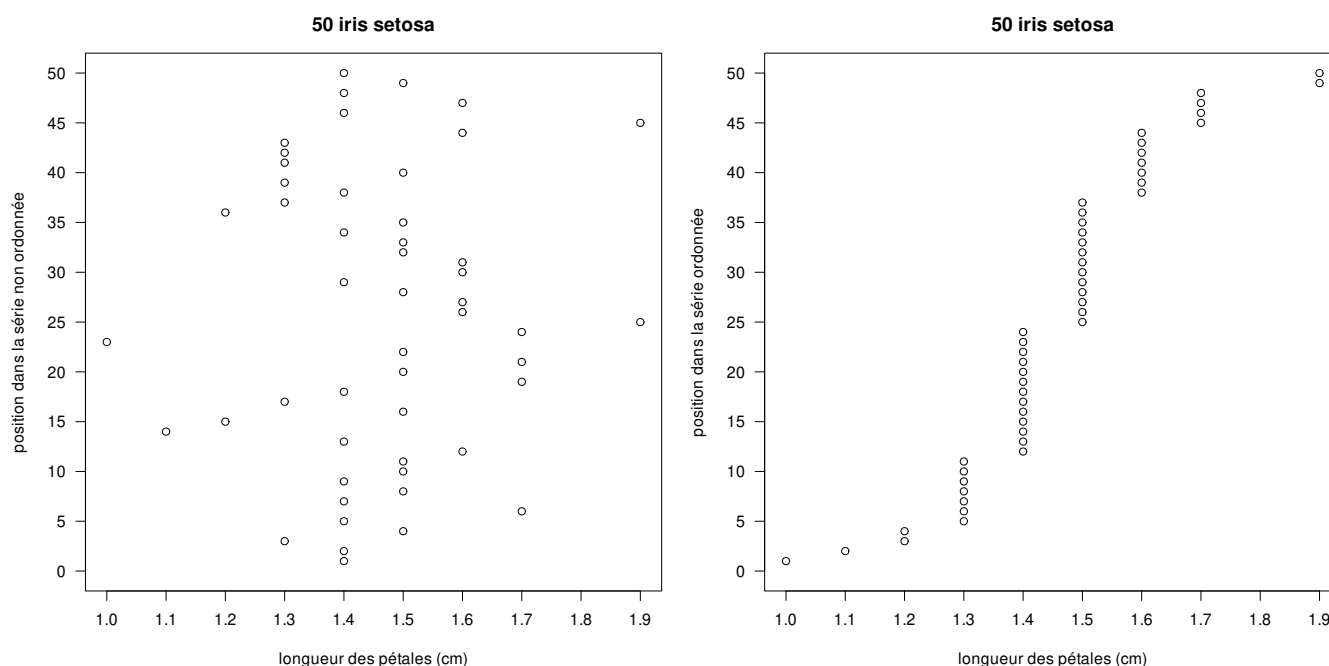
Le logiciel R, ([créé par des statisticiens, pour des statisticiens](#)), utilise comme la plupart des langages informatiques le point comme séparateur décimal.

1.2 Représentation graphique des mesures ordonnées

Un des principaux **objectifs de la statistique descriptive** est de représenter de façon efficace les données observées (mesurées), donc la suite précédente de nombres.

Nous ordonnons la série de données et la représentons graphiquement (tout cela avec R ...).

Chaque petit cercle correspond à un iris.



Nous représenterons toujours le **caractère étudié** (la longueur des pétales de 50 iris setosa) sur l'axe horizontal des abscisses x .

Les **valeurs** du caractère étudié sont aussi appelées **modalités**

Une fois rangées par ordre croissant, il est plus facile d'en déduire certaines informations.

[1] 1.0 1.1 1.2 1.2 1.3 1.3 1.3 1.3 1.3 1.3 1.3 1.4 1.4 1.4 1.4 1.4 1.4 1.4 1.4
 [20] 1.4 1.4 1.4 1.4 1.4 1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.5 1.6
 [39] 1.6 1.6 1.6 1.6 1.6 1.6 1.7 1.7 1.7 1.7 1.9 1.9

L'iris qui a la plus petite modalité (la plus petite longueur des pétales) a maintenant la position 1.

Les deux iris avec les pétales les plus longues ont les positions 49 et 50.

Les modalités observées sont : $x_1 = 1,0$ cm ; $x_2 = 1,1$ cm ; $x_3 = 1,2$ cm ; $x_4 = 1,3$ cm ; $x_5 = 1,4$ cm ;
 $x_6 = 1,5$ cm ; $x_7 = 1,6$ cm ; $x_8 = 1,7$ cm ; $x_9 = 1,9$ cm ;

Aucun des 50 iris observés a une longueur mesurée des pétales égale à 1,8 cm.

2 Tableau recensé, effectif et effectif cumulé d'une modalité

Définition 1 : Effectif d'une modalité

L'effectif $n(x)$ d'une modalité x (d'une valeur x du caractère étudié) est le **nombre d'individus ayant une modalité égale à x** .

i	x_i	n_i
1	1	1
2	1.1	1
3	1.2	2
4	1.3	7
5	1.4	13
6	1.5	13
7	1.6	7
8	1.7	4
9	1.9	2

Plutôt que d'écrire la suite des 50 modalités observées (**une observation par individu**), il est plus pratique de préciser pour chacune des 9 modalités observées x_i , l'effectif $n_i = n(x_i)$ qui lui correspond. Par exemple, $n_4 = n(1,3 \text{ cm}) = 7$ iris. Chaque valeur de l'indice i correspond à une ligne, et donc à une modalité x_i .

On obtient alors un tableau avec $t = 9$ lignes.

Ce tableau est appelé **tableau recensé**.

Il est important de remarquer que : $x_1 < x_2 < x_3 < x_4 \dots < x_8 < x_9$

Les 9 modalités sont rangées par ordre croissant !

Définition 2 : Effectif cumulé d'une modalité

L'effectif cumulé $N(x)$ d'une modalité x (d'une valeur x du caractère étudié) est le **nombre d'individus ayant une modalité inférieure ou égale à x** .

i	x_i	n_i	N_i
1	1	1	1
2	1.1	1	2
3	1.2	2	4
4	1.3	7	11
5	1.4	13	24
6	1.5	13	37
7	1.6	7	44
8	1.7	4	48
9	1.9	2	50

Le nombre $N(x)$ d'individus ayant une modalité inférieure ou égale à une certaine valeur x est la somme des effectifs des modalités x_i inférieures ou égale à x .

$$N(x) = \sum_{x_i \leq x} n_i$$

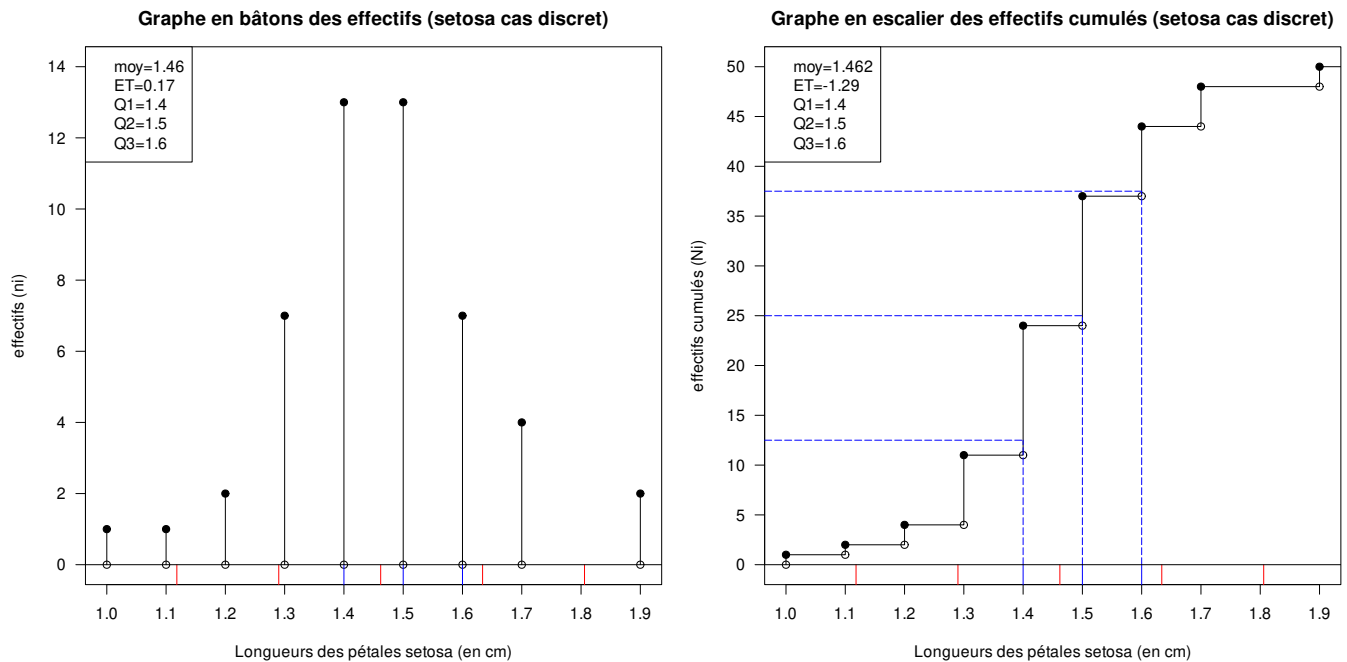
Ainsi, le nombre d'individus (d'iris) ayant une modalité inférieure ou égale à 1,2 cm est $N(1,2) = N(x_3) = N_3 = n_1 + n_2 + n_3 = 1 + 1 + 2 = 4$ iris.

Le nombre d'individus (d'iris) ayant une modalité inférieure ou égale à 1,39 cm est $1 + 1 + 2 + 7 = 11$ iris.

3 Graphes des effectifs et effectifs cumulés : cas discret

Dans le cas discret, les fonctions effectif $n(x)$ et effectif cumulé $N(x)$ associent à tout nombre réel x (toute modalité $x \in \mathbb{R}$), **un nombre entier**.

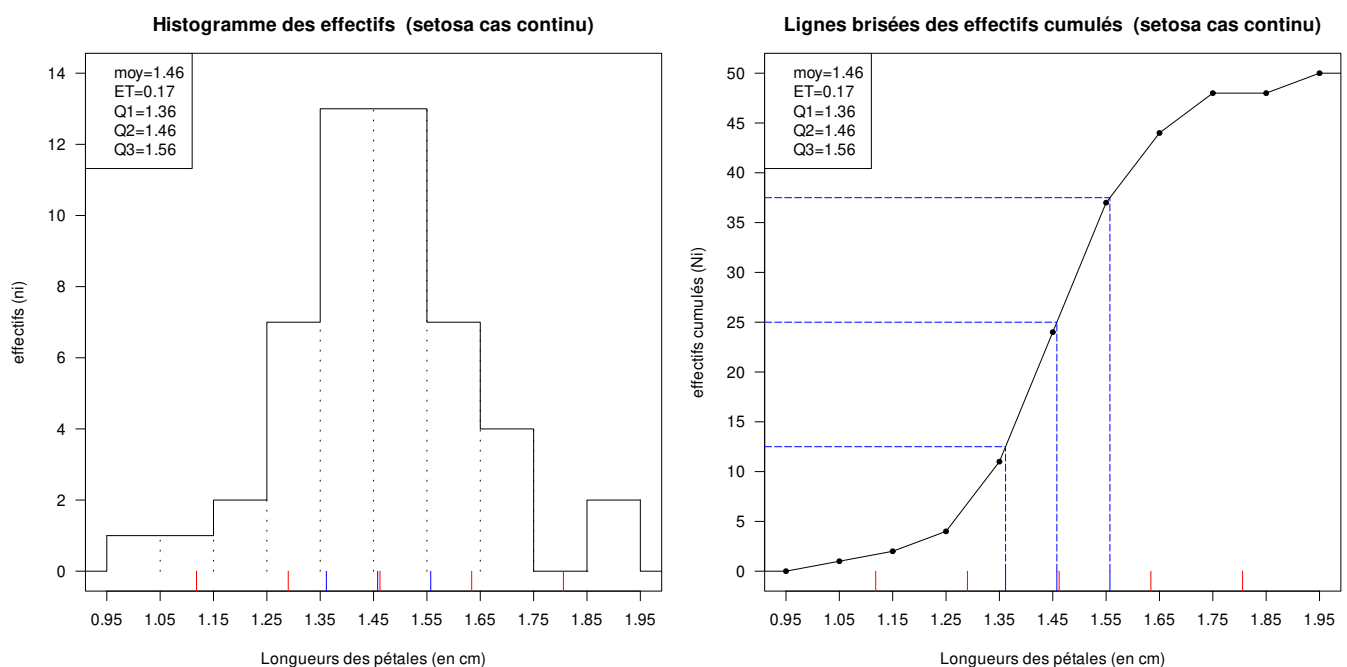
Voici, ci-dessous, leurs représentations graphiques construites « à la main » avec R.



Par exemple, $n(1,691 \text{ cm}) = 0$ iris et $N(1,691 \text{ cm}) = 44$ iris.

Aucun des 50 mesures a pour valeur 1,691 cm et 44 mesures sont plus petites ou égales à 1,691 cm.

4 Graphes des effectifs et effectifs cumulés : cas continu



Supposons par exemple que les mesures des longueurs des pétales des iris setosa aient une précision de $\pm 0,05$ cm. Quand on obtient une mesure de 1,5 cm, cela signifie que la longueur des pétales de l'iris est dans l'intervalle $[1,45; 1,55[$ cm, mais **certainement pas qu'elle est exactement égale à 1,5 cm**.

Dans le cas continu, chaque individu ayant pour valeur 1,5 cm peut être **comptabilisé uniformément** sur toutes les valeurs de l'intervalle $[1,45; 1,55[$, la classe $[1,45; 1,55[$ d'amplitude 0,1 cm **centrée sur cette mesure**.

On obtient l'histogramme ci-dessus.

Définition 3 : Effectif de la classe $[x_{i-1}; x_i[$

L'effectif n_i de la classe $[x_{i-1}; x_i[$ est le nombre d'individus dont les valeurs du caractère étudié sont dans l'intervalle $[x_{i-1}; x_i[$.

Cette classe est située sur la $i^{\text{ème}}$ ligne du tableau recensé **groupé**.

i	x_{i-1}	x_i	x_{ci}	n_i	N_i
1	0.95	1.05	1.0	1	1
2	1.05	1.15	1.1	1	2
3	1.15	1.25	1.2	2	4
4	1.25	1.35	1.3	7	11
5	1.35	1.45	1.4	13	24
6	1.45	1.55	1.5	13	37
7	1.55	1.65	1.6	7	44
8	1.65	1.75	1.7	4	48
9	1.75	1.85	1.8	0	48
10	1.85	1.95	1.9	2	50

L'effectif n_i de la classe $[x_{i-1}; x_i[$ correspond à l'aire du $i^{\text{ème}}$ rectangle dans l'histogramme.

La hauteur des rectangles d'un histogramme correspond à une **densité de présence** d_i définie par :

Densité de présence d_i dans la $i^{\text{ème}}$ classe : $d_i = \frac{n_i}{x_i - x_{i-1}}$.

Par exemple, $d_5 = \frac{13}{1,45 - 1,35} = 130$ iris/cm

d_5 est aussi la pente de la 5^{ème} ligne brisée du graphe des effectifs cumulés.

Ainsi, le nombre $n([1,38 \text{ cm}; 1,44 \text{ cm}])$ d'individus de modalités entre 1,38 cm et 1,44 cm est :

$$n([1,38 \text{ cm}; 1,44 \text{ cm}]) = 130 \cdot (1,44 - 1,38) = 7,8 \text{ iris}$$

Dans le cas continu, nous ne savons calculer des effectifs que sur des intervalles et les nombres obtenus ne sont pas forcément des nombres entiers !

De plus, le nombre d'iris ayant des pétales de longueurs inférieures ou égales à 1,4 cm est :

$$N(1,4 \text{ cm}) = N(1,35 \text{ cm}) + 130 \cdot (1,4 - 1,35) = 11 + 6,5 = 17,5 \text{ iris.}$$

Dans le cas discret, on avait $N(1,4 \text{ cm}) = 24$ iris, c'est à dire le résultat qui correspond effectivement aux 50 iris observés.

Dans le cas continu, nous n'utilisons plus les valeurs mesurées, simplement la connaissance de la classe à laquelle ces mesures appartiennent. Nous obtenons des résultats moins dépendants des 50 iris observés et nous l'espérons plus représentatifs de la population de **tous** les iris setosa.

Le modèle continu peut être aussi utilisé quand on a beaucoup de modalités différentes et que l'on souhaite une analyse plus globale. On **regroupe** alors **les modalités observées dans différentes classes bien choisies**, d'où le nom de tableau recensé **groupé** donné à un tableau de recensement contenant des classes.

Dans le **cas continu** (c'est à dire avec des classes), on retiendra que :

1. Les effectifs cumulés sont représentés par un graphe en **lignes brisées** qui est **continu** (pas de « saut » à une valeur précise du caractère étudié).
2. La pente de chaque ligne brisée i est la densité de présence d_i dans la classe $[x_{i-1}; x_i]$ correspondante située à la $i^{\text{ème}}$ ligne du **tableau recensé groupé**.

3. Chaque individu observé a été « **distribué uniformément** » sur toutes les valeurs de la classe (l'intervalle) à laquelle appartient sa modalité.
4. L'effectif **d'un intervalle** ou l'effectif cumulé d'une modalité peuvent être des nombres non entiers.

5 Calcul des indices de position et dispersion dans le cas discret

5.1 Indices de position

La **moyenne** est la somme des modalités divisée par leur nombre N (le nombre d'individus que comporte l'étude).

$$\bar{x} = \frac{1,0 + 1,1 + 1,2 + 1,2 + 1,3 + 1,3 + 1,3 + 1,3 + 1,3 + 1,3 + 1,3 + 1,4 + \dots + 1,9 + 1,9}{50} = \frac{73,1}{50} = 1,462 \text{ cm}$$

Quand le nombre N d'individus observés est très grand, on préfère faire le calcul à partir du tableau recensé beaucoup plus pratique en utilisant les effectifs n_i :

$$\bar{x} = \frac{1 \cdot 1,0 + 1 \cdot 1,1 + 2 \cdot 1,2 + 7 \cdot 1,3 + 13 \cdot 1,4 + \dots + 4 \cdot 1,7 + 2 \cdot 1,9}{50} = \frac{73,1}{50} = 1,462 \text{ cm}$$

On remarque aussi que :

$$\bar{x} = \frac{1}{50} \cdot 1,0 + \frac{1}{50} \cdot 1,1 + \frac{2}{50} \cdot 1,2 + \frac{7}{50} \cdot 1,3 + \frac{13}{50} \cdot 1,4 + \dots + \frac{4}{50} \cdot 1,7 + \frac{2}{50} \cdot 1,9 = \frac{73,1}{50} = 1,462 \text{ cm}$$

$\frac{1}{50}, \frac{2}{50}, \frac{7}{50}, \dots$ sont les fréquences f_i des modalités.

$$\bar{x} = \frac{\sum_{i=1}^t n_i \cdot x_i}{N} = \sum_{i=1}^t \frac{n_i}{N} \cdot x_i = \sum_{i=1}^t f_i \cdot x_i$$

Le(s) **mode(s)** est (sont) la (les) modalité(s) ayant le plus grand effectif.

On trouve le(s) mode(s) à partir du tableau recensé ou du graphe en bâtons des effectifs. La série statistique des 50 longueurs de pétales a deux modes 1,4 cm et 1,5 cm.

La **médiane** Q_2 est la plus petite modalité telle qu'au moins 50 % des individus ont une modalité inférieure ou égale à celle-ci.

On peut la trouver à **partir du tableau recensé**. On calcule le plus petit entier **supérieur** ou égal à $\frac{N}{2}$, ici $\frac{50}{2} = 25$ iris.

On cherche la plus petite valeur de l'effectif cumulé N_i telle que $N_i \geq 25$. Elle se situe sur la 6^{ème} ligne et $Q_2 = x_6 = 1,5 \text{ cm}$.

On peut aussi la trouver à **partir du graphe en escalier des effectifs cumulés**. On trace une horizontale correspondant à l'effectif cumulé 25 iris en partant de l'axe des ordonnées et on s'arrête dès qu'on touche le graphe (lignes verticales comprises). **L'abscisse** du point obtenu est la médiane Q_2 .

Le **premier quartile** Q_1 et le **troisième quartile** Q_3 .

On procède comme pour la médiane en remplaçant simplement 50 % par 25 % et 75 %.

5.2 Indices de dispersion

L'**intervalle interquartile** est égal à $Q_3 - Q_1 = 1,6 - 1,4 = 0,2$ cm.

La **variance** est la moyenne des carrés des écarts à la moyenne.

modalités $x_i = 1.0 \ 1.1 \ 1.2 \ 1.3 \ 1.4 \ 1.5 \ 1.6 \ 1.7 \ 1.9$

Pour chacune des 9 modalités (9 lignes du tableau recensé), on calcule l'écart à la moyenne $x_i - \bar{x}$ avec $\bar{x} = 1,462$ cm

écarts à la moyenne = -0.462 -0.362 -0.262 -0.162 -0.062 0.038 0.138 0.238 0.438

carrés des écarts = 0.213 0.131 0.069 0.026 0.004 0.001 0.019 0.057 0.192

Somme des carrés des écarts à la moyenne

(en n'oubliant pas de multiplier chaque carré par son effectif) :

1 fois 0,213 + 1 fois 0,131 + 7 fois 0,069 + ... + 2 fois 0,192 = 1,4778 cm²

Moyenne, on divise par 50 : 0,029556 centimètres carrés

Propriété 1 : Calcul simplifié de la variance d'une série statistique

La variance est aussi égale à la moyenne des carrés moins le carré de la moyenne :

$$\text{Var}(x) = \overline{x^2} - (\bar{x})^2$$

modalités au carré : 1.00 1.21 1.44 1.69 1.96 2.25 2.56 2.89 3.61

somme des carrés = 108,35 moyenne des carrés = 2,167

moyenne des carrés moins le carré de la moyenne :

2,167 - 1,462² = 0,029556 centimètres carrés

L'**écart type** est la racine carrée de la variance : $\sqrt{0,029556} = 0,1719186 \approx 0,17$ cm

L'écart type mesure l'écart des données par rapport à la moyenne

L'écart type est souvent appelé sigma : σ .

Théorème de Tchebychev : $\frac{n([\bar{x} - k \cdot \sigma; \bar{x} + k \cdot \sigma])}{N} \geq 1 - \frac{1}{k^2}$

$1 - \frac{1}{2^2} = \frac{3}{4} = 0,75$. Au moins 75 % des individus ont une modalité dans $[\bar{x} - 2 \cdot \sigma; \bar{x} + 2 \cdot \sigma]$.

6 Calcul des indices de position et dispersion dans le cas continu

On procède comme pour le cas discret en remplaçant les modalités x_i par les centres des classes xc_i .

Pour les quartiles Q_1 , Q_2 et Q_3 , on utilise le graphe **continu** des effectifs cumulés.

Avec le tableau recensé groupé (avec des classes), il faut effectuer une **interpolation linéaire**. On trouve dans quelle classe se situe la médiane (la 6^{ème}), avec $d_6 = d_5 = 130$ iris/cm :

$N(Q_2) = N(1,45 \text{ cm}) + 130 \cdot (Q_2 - 1,45)$ soit encore $25 = 24 + 130 \cdot (Q_2 - 1,45)$

Ainsi, $Q_2 - 1,45 = \frac{1}{130}$ et finalement $Q_2 = 1,45 + \frac{1}{130} = 1,4576... \approx 1,46$ cm