

Chapitre 1 – La statistique descriptive

STT-1920 Méthodes statistiques

Département de mathématiques et de statistique



UNIVERSITÉ
LAVAL

1.1 Introduction

Pour commencer : 147 serpents nouveau-nés

Un chercheur étudie une espèce de serpents. Il a pesé 147 serpents à la naissance (en grammes). Voici le début de ses données :

33.90	34.87	34.49	34.16	35.14	34.10	34.41
33.10	35.59	35.20	34.35	33.87	35.18	34.54
35.71	34.74	36.42	34.68	34.05	34.76	35.49
35.44	36.03	34.19	35.49	35.30	34.26	35.36

⋮ (21 lignes en tout) ⋮

Difficile d'y voir clair ! Quel est le poids « typique » ? Les poids varient-ils beaucoup ? Y a-t-il des valeurs étranges ?

⇒ Il faut **résumer** les données : c'est le rôle de la statistique descriptive.

Deux types de méthodes statistiques

Statistique descriptive

L'art de **résumer** un ensemble de données (tableaux, graphiques, quelques nombres) pour mettre en évidence l'information qu'elles contiennent.

Statistique inférentielle

L'art de tirer des **conclusions** au sujet d'une population à partir d'un échantillon provenant de cette population.

Plan du cours : chapitre 1, statistique descriptive; chapitre 2, probabilités (le modèle du hasard); chapitres 3 à 7, statistique inférentielle (intervalles de confiance, tests, ANOVA, régression, khi-deux).

Décrire avant d'analyser

Une bonne analyse statistique est **toujours** précédée d'un résumé des principales caractéristiques des données.

- **Mesures de tendance centrale**

- moyenne, médiane

- **Mesures de position**

- quantiles, quartiles

- **Mesures de dispersion**

- écart-type, coefficient de variation
- écart interquartile, étendue

- **Représentations graphiques**

- histogramme
- diagramme en boîte
- diagramme quantile-quantile
- diagrammes en bâtons et en pointes de tarte
- diagramme de dispersion

Tout cela se fait en quelques lignes avec le logiciel **R**.

1.2 Population, échantillon et variable

Exemple 1 : les serpents

Le chercheur a pesé 147 serpents nouveau-nés de l'espèce étudiée.

Exemple 1 : les serpents

Le chercheur a pesé 147 serpents nouveau-nés de l'espèce étudiée.

- **Population** : ensemble de **tous** les individus qui nous intéressent.
⇒ tous les serpents de cette espèce.
- **Échantillon** : les individus effectivement observés.
⇒ les 147 serpents pesés par le chercheur ($n = 147$).
- **Variable statistique** : caractéristique mesurée sur chaque individu.
⇒ le poids à la naissance (en g). On pourrait aussi mesurer la longueur, le sexe, ...

Une **statistique** est une quantité calculée à partir de l'échantillon (par exemple, la moyenne des 147 poids).

Un jeu de données typique

Habituellement, on mesure **plusieurs variables** sur chaque individu. Une ligne = un individu ; une colonne = une variable.

Exemple : suivi de grenouilles léopards dans trois étangs.

Grenouille	Étang	Sexe	Masse (g)	Nombre de parasites	Stade
1	A	F	23.4	3	adulte
2	A	M	18.9	0	juvénile
3	B	F	31.2	7	adulte
4	C	M	12.5	1	juvénile
5	B	M	27.8	2	adulte
⋮	⋮	⋮	⋮	⋮	⋮

Ces variables ne sont pas toutes de la même nature...

Les types de variables

Variables quantitatives

- **Continue** : toute valeur numérique d'un intervalle.
Ex. : masse, taille d'un plant.
- **Discrète** : valeurs isolées (souvent entières).
Ex. : nombre de parasites.

Variables qualitatives

- **Ordinale** : catégories (non numériques) ordonnées.
Ex. : stade (juvénile, adulte).
- **Nominale** : catégories sans ordre.
Ex. : étang, sexe, espèce.

Attention! Un code numérique ne rend pas une variable quantitative : l'étang codé 1, 2, 3 ou le sexe codé 0/1 restent des variables **nominales**. Faire la moyenne de ces codes n'a aucun sens.

QUIZ!

Quel est le type (quantitative continue ou discrète, qualitative ordinale ou nominale) de chacune des variables suivantes ?

- a) Le groupe sanguin (A, B, AB, O)
d'un donneur.
- b) Le poids d'un loup à la naissance.
- c) Le stade d'un cancer (I, II, III, IV).
- d) Le nombre de tomates d'un plant.

QUIZ!

Quel est le type (quantitative continue ou discrète, qualitative ordinale ou nominale) de chacune des variables suivantes ?

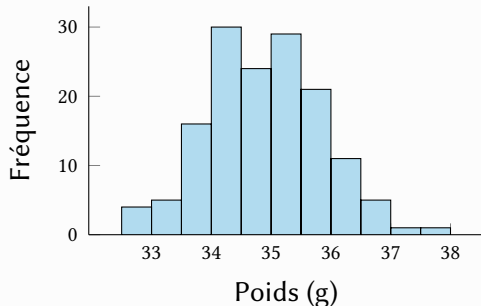
- a) Le groupe sanguin (A, B, AB, O) d'un donneur.
- b) Le poids d'un loup à la naissance.
- c) Le stade d'un cancer (I, II, III, IV).
- d) Le nombre de tomates d'un plant.
- e) La glycémie à jeun (en mmol/L).
- f) L'espèce d'un poisson capturé.
- g) L'intensité d'une douleur (aucune, légère, modérée, sévère).

1.3 L'histogramme

Exemple 1 (suite) : données groupées

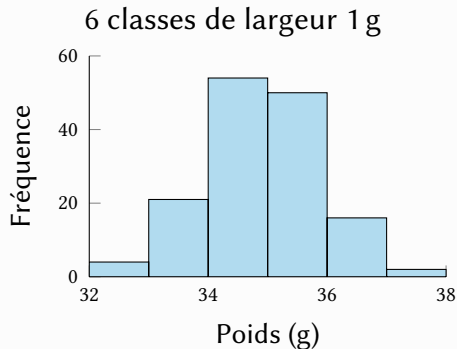
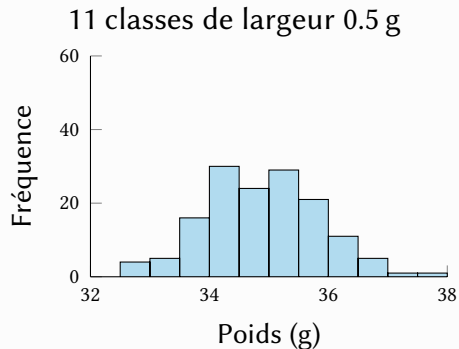
Pour voir la **forme** de la distribution d'une variable quantitative continue, on **groupe** d'abord les données en classes.

Poids (g)	Fréquence
(32.5, 33.0]	4
(33.0, 33.5]	5
(33.5, 34.0]	16
(34.0, 34.5]	30
(34.5, 35.0]	24
(35.0, 35.5]	29
(35.5, 36.0]	21
(36.0, 36.5]	11
(36.5, 37.0]	5
(37.0, 37.5]	1
(37.5, 38.0]	1
Total	147



Données **brutes** → données **groupées** (tableau de fréquences) → **histogramme**.

Le choix des classes est un peu arbitraire



- Trop de classes $\Rightarrow$ histogramme chaotique (classes presque vides).
- Trop peu de classes $\Rightarrow$ on ne voit plus la forme.
- En pratique : de 5 à 12 classes environ, selon n .

Procédure pour tracer un histogramme

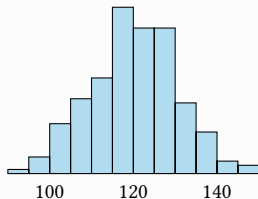
Soit un échantillon $x_1, \dots, x_n$.

1. On découpe l'étendue des données en ℓ classes $(c_0, c_1], (c_1, c_2], \dots, (c_{\ell-1}, c_\ell]$.
 2. On compte la **fréquence absolue** f_j (nombre d'observations dans la classe j) et on calcule la **fréquence relative** f_j/n .
 3. Sur chaque classe, on dessine un rectangle dont la **surface** est proportionnelle à la fréquence.
- **Classes de même largeur** : hauteur = fréquence absolue f_j (ou relative f_j/n).
 - **Classes de largeurs inégales** : histogramme de **densité**, de hauteur
$$h_j = \frac{f_j/n}{c_j - c_{j-1}},$$
 de sorte que la surface du rectangle j vaut exactement f_j/n .

Formes de distributions

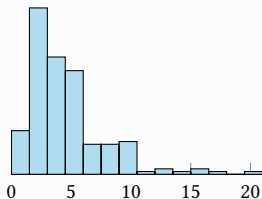
L'histogramme permet de reconnaître la **forme** de la distribution ($n = 200$ dans chaque cas, données simulées) :

Symétrique



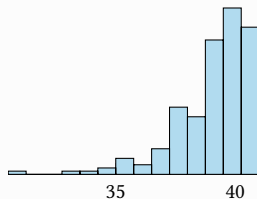
Pression systolique (mmHg)

Asymétrique à droite



Séjour à l'hôpital (jours)

Asymétrique à gauche



Durée de gestation (sem.)

On regarde aussi le nombre de « bosses » (**unimodale**, **bimodale**...) et la présence de valeurs isolées. Les poids des serpents : distribution unimodale et à peu près symétrique, « en cloche ».

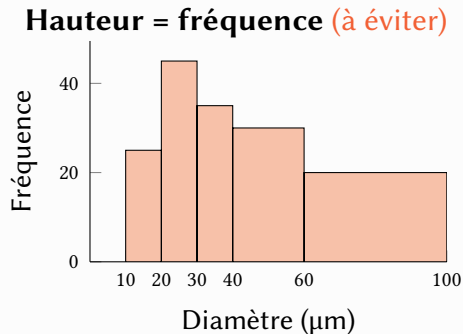
Exemple 2 : classes de largeurs inégales

Pour calibrer un pulvérisateur agricole servant à appliquer un fongicide, on a mesuré le diamètre (en microns) de 155 gouttelettes. Les données ne nous sont fournies que sous forme groupée :

Diamètre (μm)	Fréquence
(10, 20]	25
(20, 30]	45
(30, 40]	35
(40, 60]	30
(60, 100]	20
Total	155

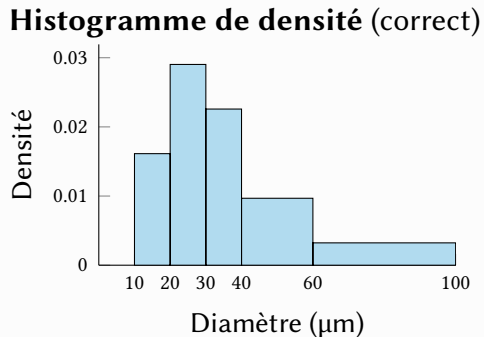
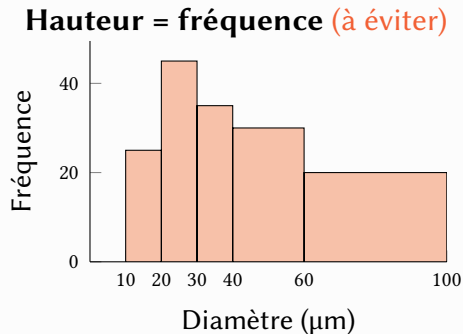
Les deux dernières classes sont **plus larges** que les autres. Peut-on simplement prendre les fréquences comme hauteurs ?

Exemple 2 (suite) : attention aux classes inégales!



À gauche, la classe $(60, 100]$ semble contenir presque autant de gouttelettes que $(10, 20]$... alors qu'elle est quatre fois plus large!

Exemple 2 (suite) : attention aux classes inégales !



À gauche, la classe $(60, 100]$ semble contenir presque autant de gouttelettes que $(10, 20]$... alors qu'elle est quatre fois plus large !

⇒ C'est la **surface** (et non la hauteur) qui doit représenter la fréquence. Avec des classes inégales, R trace automatiquement un histogramme de densité.

Exemple 2 (suite) : calcul des hauteurs

Classe	Fréquence f_j	Fréq. relative f_j/n	Largeur	Hauteur h_j
(10, 20]	25			
(20, 30]	45			
(30, 40]	35			
(40, 60]	30			
(60, 100]	20			
Total	155			

Exemple 2 (suite) : calcul des hauteurs

Classe	Fréquence f_j	Fréq. relative f_j/n	Largeur	Hauteur h_j
(10, 20]	25	0.1613	10	0.0161
(20, 30]	45	0.2903	10	0.0290
(30, 40]	35	0.2258	10	0.0226
(40, 60]	30	0.1935	20	0.0097
(60, 100]	20	0.1290	40	0.0032
Total	155	1		

Par exemple, pour la classe (40, 60] :

$$\frac{f_4}{n} = \frac{30}{155} = 0.1935 \quad \Rightarrow \quad h_4 = \frac{0.1935}{60 - 40} = 0.0097.$$

La surface totale de l'histogramme de densité vaut 1.

QUIZ!

Exemple 2 (gouttelettes du pulvérisateur).

- Quelle proportion des gouttelettes ont un diamètre entre 20 et 40 μm ?
Comment la lit-on sur l'histogramme de densité?

QUIZ!

Exemple 2 (gouttelettes du pulvérisateur).

- Quelle proportion des gouttelettes ont un diamètre entre 20 et 40 μm ?
Comment la lit-on sur l'histogramme de densité?
- Sur un histogramme de densité, que vaut la surface totale des rectangles?
La somme des hauteurs a-t-elle une interprétation?

1.4 La moyenne et l'écart-type

Moyenne et écart-type échantillonnaux

Moyenne échantillonnale (mesure de tendance centrale)

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

Écart-type échantillonnal (mesure de dispersion)

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (s^2 = \text{variance échantillonnale})$$

s est la **distance typique** entre les observations et leur moyenne, dans les **mêmes unités** que les données. (Pourquoi $n - 1$? Voir chapitre 3.)

Exemple 3 : calcul à la main

On a pesé 5 souris de laboratoire (en g) : 22 25 19 24 20.

Exemple 3 : calcul à la main

On a pesé 5 souris de laboratoire (en g) : 22 25 19 24 20.

$$\bar{x} = \frac{22 + 25 + 19 + 24 + 20}{5} = \frac{110}{5} = 22 \text{ g.}$$

x_i	22	25	19	24	20	Somme
$x_i - \bar{x}$	0	3	-3	2	-2	0
$(x_i - \bar{x})^2$	0	9	9	4	4	26

Exemple 3 : calcul à la main

On a pesé 5 souris de laboratoire (en g) : 22 25 19 24 20.

$$\bar{x} = \frac{22 + 25 + 19 + 24 + 20}{5} = \frac{110}{5} = 22 \text{ g.}$$

x_i	22	25	19	24	20	Somme
$x_i - \bar{x}$	0	3	-3	2	-2	0
$(x_i - \bar{x})^2$	0	9	9	4	4	26

$$s^2 = \frac{26}{5 - 1} = 6.5 \text{ g}^2 \quad \text{et} \quad s = \sqrt{6.5} = 2.55 \text{ g.}$$

Remarque : la somme des écarts $x_i - \bar{x}$ vaut **toujours** 0. C'est pourquoi on les met au carré avant de les additionner.

Exemple 1 (suite) : les serpents dans R

Pour les 147 serpents :

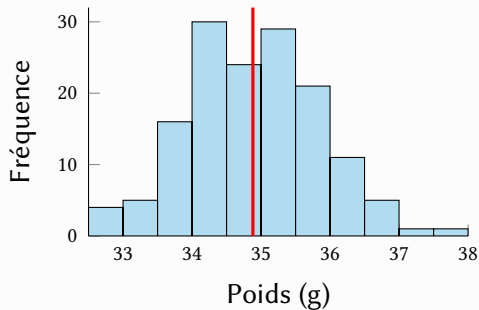
$$\bar{x} = \frac{33.90 + 34.87 + \dots + 35.56}{147} = 34.882 \text{ g},$$

$$s = \sqrt{\frac{(33.90 - 34.882)^2 + \dots + (35.56 - 34.882)^2}{146}} = 0.956 \text{ g}.$$

Heureusement, R fait le calcul pour nous :

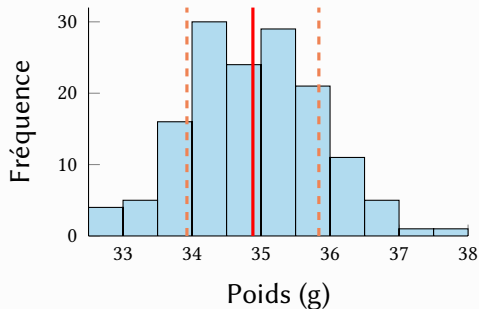
```
> serpents <- read.csv("serpents.csv")
> poids <- serpents$poids
> mean(poids)
[1] 34.88177
> sd(poids)
[1] 0.9560275
```

Interpréter l'écart-type



Trait plein : $\bar{x} = 34.882$ g.

Interpréter l'écart-type

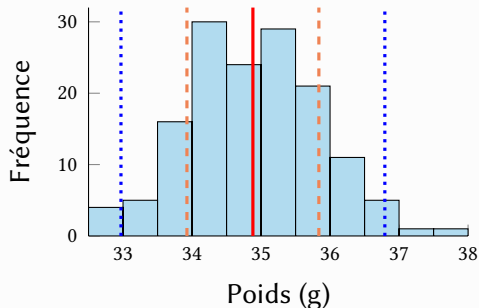


Trait plein : $\bar{x} = 34.882$ g.

Tirets : $\bar{x} \pm s = [33.93; 35.84]$

104 serpents sur 147, soit 70.7 %.

Interpréter l'écart-type



Trait plein : $\bar{x} = 34.882$ g.

Tirets : $\bar{x} \pm s = [33.93; 35.84]$
104 serpents sur 147, soit **70.7 %**.

Pointillés : $\bar{x} \pm 2s = [32.97; 36.79]$
141 serpents sur 147, soit **95.9 %**.

Règle empirique : pour une distribution **en cloche**, environ 68 % des observations sont à moins d'un écart-type de la moyenne et environ 95 % à moins de deux écarts-types (loi normale, chapitre 2).

Caractéristiques de la moyenne

- C'est le **point d'équilibre** de l'histogramme.
- Toutes les observations interviennent dans son calcul.
- Elle est **sensible aux valeurs extrêmes** : une seule valeur très grande (ou une erreur de saisie !) peut la déplacer beaucoup.

Autre mesure de dispersion : l'écart absolu moyen

$$d = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

Plus simple à interpréter, mais peu utilisé : l'écart-type a de bien meilleures propriétés mathématiques (chapitres 2 et 3). Pour les serpents, $d = 0.779$ g, proche de $s = 0.956$ g.

QUIZ!

La pression artérielle systolique moyenne de 30 patients vaut $\bar{x} = 112$ mmHg. On retire la plus grande et la plus petite observations.

Vrai ou faux? La moyenne des 28 observations restantes vaut toujours 112 mmHg.

Le coefficient de variation

Coefficient de variation (variable à valeurs positives)

$CV = s/\bar{x}$: mesure de dispersion **relative**, sans unité, souvent exprimée en %.

Exemple 4 : poids de rats et de lapins de laboratoire.

Population	n	$\bar{x}$ (g)	s (g)
Rats	75	110	11
Lapins	84	660	22

Les poids des lapins varient-ils plus que ceux des rats ?

Le coefficient de variation

Coefficient de variation (variable à valeurs positives)

$CV = s/\bar{x}$: mesure de dispersion **relative**, sans unité, souvent exprimée en %.

Exemple 4 : poids de rats et de lapins de laboratoire.

Population	n	$\bar{x}$ (g)	s (g)
Rats	75	110	11
Lapins	84	660	22

Les poids des lapins varient-ils plus que ceux des rats ?

$CV_{\text{rats}} = 11/110 = 10\%$ et $CV_{\text{lapins}} = 22/660 = 3.3\%$: l'écart-type est plus petit chez les rats, mais la variation **relative** est plus petite chez les lapins.

Moyenne et écart-type à partir de données groupées

Sans les données brutes, on fait comme si les f_j observations de la classe $(c_{j-1}, c_j]$ étaient toutes égales au **milieu** de la classe, $m_j = (c_{j-1} + c_j)/2$:

$$\bar{x} \approx \bar{x}^* = \frac{1}{n} \sum_{j=1}^{\ell} f_j m_j, \quad s \approx s^* = \sqrt{\frac{1}{n-1} \sum_{j=1}^{\ell} f_j (m_j - \bar{x}^*)^2}.$$

Moyenne et écart-type à partir de données groupées

Sans les données brutes, on fait comme si les f_j observations de la classe $(c_{j-1}, c_j]$ étaient toutes égales au **milieu** de la classe, $m_j = (c_{j-1} + c_j)/2$:

$$\bar{x} \approx \bar{x}^* = \frac{1}{n} \sum_{j=1}^{\ell} f_j m_j, \quad s \approx s^* = \sqrt{\frac{1}{n-1} \sum_{j=1}^{\ell} f_j (m_j - \bar{x}^*)^2}.$$

Exemple 2 (suite) : les milieux sont 15, 25, 35, 50 et 80 μm .

$$\bar{x}^* = \frac{25 \times 15 + 45 \times 25 + 35 \times 35 + 30 \times 50 + 20 \times 80}{155} = \frac{5825}{155} = 37.58 \mu\text{m},$$

$$s^* = \sqrt{\frac{25 \times (15 - 37.58)^2 + \dots + 20 \times (80 - 37.58)^2}{154}} = 19.86 \mu\text{m}.$$

QUIZ!

La balance du chercheur était mal tarée : chaque serpent a été pesé 2 g de trop.

- Que deviennent la moyenne $\bar{x} = 34.882$ g et l'écart-type $s = 0.956$ g une fois les poids corrigés ?

QUIZ!

La balance du chercheur était mal tarée : chaque serpent a été pesé 2 g de trop.

- Que deviennent la moyenne $\bar{x} = 34.882$ g et l'écart-type $s = 0.956$ g une fois les poids corrigés ?
- On exprime ensuite les poids corrigés en milligrammes. Que deviennent $\bar{x}$, s et le coefficient de variation ?

1.5 Les quantiles

Statistiques d'ordre

Exemple 5 : longueurs (en cm) de 7 truites mouchetées capturées dans un ruisseau :

22.3 17.9 20.4 24.6 19.5 26.2 18.7

Statistiques d'ordre

On note $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ les observations placées en **ordre croissant**. $x_{(1)}$ est le minimum et $x_{(n)}$ le maximum.

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$
17.9	18.7	19.5	20.4	22.3	24.6	26.2

Les quantiles

Grosso modo, le **quantile d'ordre γ** ($0 \leq \gamma \leq 1$) est la valeur q_γ telle que $100\gamma\%$ des observations lui sont inférieures (et $100(1 - \gamma)\%$ lui sont supérieures).

Quantile échantillonnal d'ordre γ

$$q_\gamma = x_{(1+(n-1)\gamma)}$$

Si $1 + (n - 1)\gamma$ n'est pas entier, on **interpole** entre les deux statistiques d'ordre voisines. Par exemple, $x_{(2.8)} = x_{(2)} + 0.8 \times (x_{(3)} - x_{(2)})$.

Il existe plusieurs définitions des quantiles échantillonnaux (très proches quand n est grand). Celle-ci est la définition par défaut de la fonction `quantile()` de R.

Exemple 5 (suite) : quantile d'ordre 0.3

Rappel : 17.9 18.7 19.5 20.4 22.3 24.6 26.2 ($n = 7$)

Exemple 5 (suite) : quantile d'ordre 0.3

Rappel : 17.9 18.7 19.5 20.4 22.3 24.6 26.2 ($n = 7$)

Position : $1 + (n - 1)\gamma = 1 + (7 - 1) \times 0.3 = 2.8$, donc $x_{(2.8)}$ est à 80 % du chemin entre $x_{(2)}$ et $x_{(3)}$:

$$q_{0.3} = x_{(2)} + 0.8 \times (x_{(3)} - x_{(2)}) = 18.7 + 0.8 \times (19.5 - 18.7) = 19.34 \text{ cm.}$$

Exemple 5 (suite) : quantile d'ordre 0.3

Rappel : 17.9 18.7 19.5 20.4 22.3 24.6 26.2 ($n = 7$)

Position : $1 + (n - 1)\gamma = 1 + (7 - 1) \times 0.3 = 2.8$, donc $x_{(2.8)}$ est à 80 % du chemin entre $x_{(2)}$ et $x_{(3)}$:

$$q_{0.3} = x_{(2)} + 0.8 \times (x_{(3)} - x_{(2)}) = 18.7 + 0.8 \times (19.5 - 18.7) = 19.34 \text{ cm.}$$

Serpents ($n = 147$) : $1 + 146 \times 0.4 = 59.4$, donc

$$q_{0.4} = x_{(59)} + 0.4 \times (x_{(60)} - x_{(59)}) = 34.56 + 0.4 \times (34.59 - 34.56) = 34.572 \text{ g.}$$

Environ 40 % des serpents pèsent moins de 34.572 g.

Quartiles et centiles

Quartiles

- Premier quartile : $Q_1 = q_{0.25}$
- Deuxième quartile, ou **médiane** : $Q_2 = m = q_{0.50}$
- Troisième quartile : $Q_3 = q_{0.75}$

Les trois quartiles séparent l'échantillon en quatre parties de même effectif.

Le quantile $q_{k/100}$ est aussi appelé k^{e} **centile** (ou percentile). Ainsi, Q_1 est le 25^e centile, m le 50^e et Q_3 le 75^e.

La médiane

Avec la formule $q_{0.5} = x_{(1+(n-1)/2)}$, on obtient :

Médiane

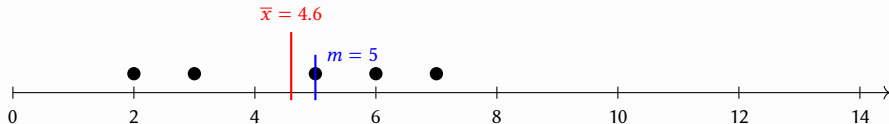
$$m = \begin{cases} x_{(\frac{n+1}{2})} & \text{si } n \text{ est impair (valeur du milieu),} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & \text{si } n \text{ est pair (moyenne des deux valeurs du milieu).} \end{cases}$$

Exemple 5 (suite) : $n = 7$, donc $m = x_{(4)} = 20.4$ cm.

Exercice : calculez Q_1 et Q_3 pour les 7 truites.

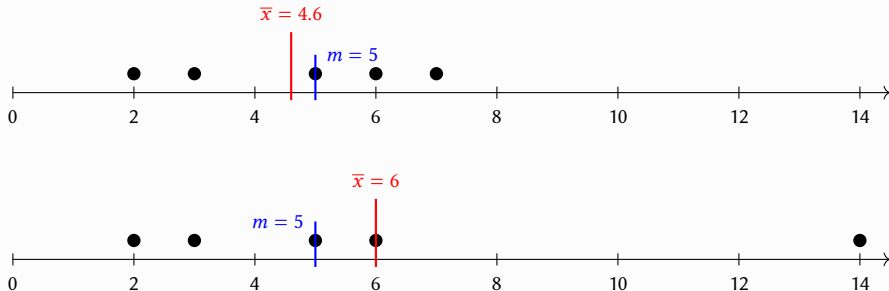
Moyenne ou médiane ?

La médiane est une autre mesure de **tendance centrale**. Elle ne fait intervenir que la ou les observations du milieu : elle est **peu sensible aux valeurs extrêmes** (on dit qu'elle est *robuste*).



Moyenne ou médiane ?

La médiane est une autre mesure de **tendance centrale**. Elle ne fait intervenir que la ou les observations du milieu : elle est **peu sensible aux valeurs extrêmes** (on dit qu'elle est *robuste*).



Remplacer 7 par 14 déplace la moyenne, mais pas la médiane.

QUIZ!

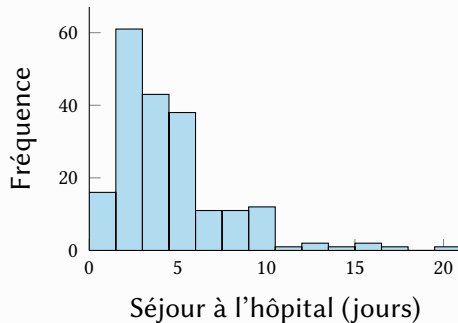
La pression artérielle systolique **médiane** de 30 patients vaut $m = 114$ mmHg. On retire la plus grande et la plus petite observations.

Vrai ou faux? La médiane des 28 observations restantes vaut toujours 114 mmHg.

QUIZ!

Voici la durée de séjour à l'hôpital de 200 patients.

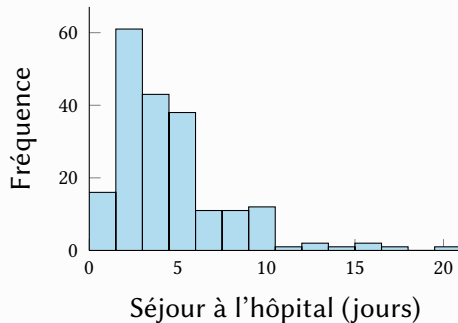
- La moyenne est-elle plus grande, plus petite ou à peu près égale à la médiane? Pourquoi?



QUIZ!

Voici la durée de séjour à l'hôpital de 200 patients.

- La moyenne est-elle plus grande, plus petite ou à peu près égale à la médiane? Pourquoi?
- Même question pour une distribution asymétrique à gauche, puis pour une distribution symétrique.



Résumé à cinq nombres et mesures de dispersion

Avec $\gamma = 0$ et $\gamma = 1$: $q_0 = x_{(1)} = \min$ et $q_1 = x_{(n)} = \max$.

Résumé à cinq nombres

$$(x_{(1)}, Q_1, Q_2, Q_3, x_{(n)}) = (q_0, q_{0.25}, q_{0.50}, q_{0.75}, q_1)$$

- **Étendue** : $x_{(n)} - x_{(1)}$ (ne dépend que des deux valeurs les plus extrêmes);
- **Écart interquartile** : $ElQ = Q_3 - Q_1$ (étendue de la moitié centrale des données).

Serpents : (32.62, 34.19, 34.86, 35.535, 37.61), donc étendue = 4.99 g et $ElQ = 35.535 - 34.19 = 1.345$ g.

Exemple 1 (suite) : les quantiles dans R

```
> quantile(poids, 0.4)
 40%
34.572
> quantile(poids, c(0, 0.25, 0.5, 0.75, 1))
 0%    25%    50%    75%   100%
32.620 34.190 34.860 35.535 37.610
> summary(poids)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 32.62  34.19   34.86   34.88   35.53   37.61
> IQR(poids)
[1] 1.345
```

La moyenne (34.88 g) et la médiane (34.86 g) sont presque égales : la distribution est à peu près **symétrique**.

`summary()` arrondit à 4 chiffres significatifs (35.535 devient 35.53).

Les rangs centiles

Ne pas confondre **centile** et **rang centile** ! Les 99 centiles $q_{0.01}, \dots, q_{0.99}$ découpent l'axe en 100 segments, numérotés de 1 à 100 à partir de la gauche.

Rang centile

Numéro du segment où se trouve l'observation : un rang centile de k signifie que l'observation est entre le $(k - 1)^{\text{e}}$ et le k^{e} centile.

Exemple 1 (suite) : le premier serpent pesait 33.90 g. Or $q_{0.13} = 33.8994$ et $q_{0.14} = 33.9232$. Le rang centile de ce serpent est donc **14** : environ 13 % des serpents sont plus légers que lui.

Les observations aberrantes

Une observation **aberrante** est très éloignée du centre de la distribution. Critère usuel (règle de 1.5 EIQ) :

Observation aberrante

$$x_i < Q_1 - 1.5 \text{ EIQ} \quad \text{ou} \quad x_i > Q_3 + 1.5 \text{ EIQ.}$$

Les observations aberrantes

Une observation **aberrante** est très éloignée du centre de la distribution. Critère usuel (règle de 1.5 EIQ) :

Observation aberrante

$$x_i < Q_1 - 1.5 \text{ EIQ} \quad \text{ou} \quad x_i > Q_3 + 1.5 \text{ EIQ}.$$

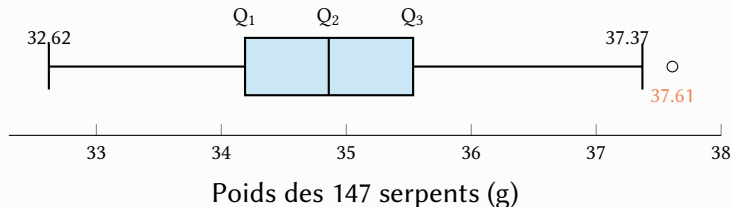
Serpents : $Q_1 - 1.5 \text{ EIQ} = 34.19 - 1.5 \times 1.345 = 32.1725$ et
 $Q_3 + 1.5 \text{ EIQ} = 35.535 + 1.5 \times 1.345 = 37.5525$. Le serpent le plus lourd (37.61 g) est donc une observation aberrante.

Attention ! « Aberrante » ne veut pas dire « à supprimer ». On vérifie d'abord s'il s'agit d'une erreur (de mesure, de saisie); sinon, on la garde : elle fait partie de la variabilité naturelle.

1.6 Le diagramme en boîte

Le diagramme en boîte

Le **diagramme en boîte** (*boxplot*) est, à quelques détails près, une représentation graphique du résumé à cinq nombres.



- La boîte va de Q_1 à Q_3 : elle contient la moitié centrale des données (longueur = IQ). Le trait indique la médiane.
- Les bras (« moustaches ») s'arrêtent aux observations non aberrantes les plus extrêmes ; les observations aberrantes sont des points.

Procédure de construction d'un diagramme en boîte

1. On trace un axe gradué (vertical ou horizontal).
2. On dessine une boîte allant de Q_1 à Q_3 .
3. On trace un trait à travers la boîte vis-à-vis de la médiane Q_2 .
4. On **calcule** (sans les tracer) les bornes $b_{\text{inf}} = Q_1 - 1.5 \text{ ElQ}$ et $b_{\text{sup}} = Q_3 + 1.5 \text{ ElQ}$.
5. **Bras inférieur** : de Q_1 jusqu'à la plus petite observation $\geq b_{\text{inf}}$.
6. **Bras supérieur** : de Q_3 jusqu'à la plus grande observation $\leq b_{\text{sup}}$.
7. Chaque observation aberrante (hors de $[b_{\text{inf}}, b_{\text{sup}}]$) est indiquée par un point ou un astérisque.

Exemple 6 : tiques sur des souris à pattes blanches

On a compté les tiques sur 11 souris à pattes blanches capturées en forêt (données déjà ordonnées) :

2 3 5 6 7 8 9 11 12 14 31

Construisons le diagramme en boîte.

Exemple 6 : tiques sur des souris à pattes blanches

On a compté les tiques sur 11 souris à pattes blanches capturées en forêt (données déjà ordonnées) :

2 3 5 6 7 8 9 11 12 14 31

Construisons le diagramme en boîte.

- $Q_1 = x_{(1+10 \times 0.25)} = x_{(3.5)} = 5 + 0.5 \times (6 - 5) = 5.5$, $Q_2 = x_{(6)} = 8$,
 $Q_3 = x_{(8.5)} = 11 + 0.5 \times (12 - 11) = 11.5$
- $ElQ = 11.5 - 5.5 = 6$, $b_{inf} = 5.5 - 1.5 \times 6 = -3.5$,
 $b_{sup} = 11.5 + 1.5 \times 6 = 20.5$

Exemple 6 : tiques sur des souris à pattes blanches

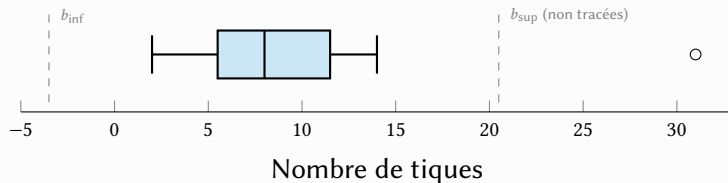
On a compté les tiques sur 11 souris à pattes blanches capturées en forêt (données déjà ordonnées) :

2 3 5 6 7 8 9 11 12 14 31

Construisons le diagramme en boîte.

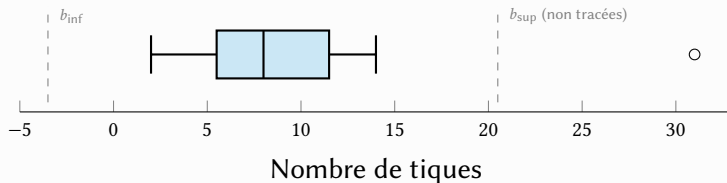
- $Q_1 = x_{(1+10 \times 0.25)} = x_{(3.5)} = 5 + 0.5 \times (6 - 5) = 5.5$, $Q_2 = x_{(6)} = 8$,
 $Q_3 = x_{(8.5)} = 11 + 0.5 \times (12 - 11) = 11.5$
- $ElQ = 11.5 - 5.5 = 6$, $b_{inf} = 5.5 - 1.5 \times 6 = -3.5$,
 $b_{sup} = 11.5 + 1.5 \times 6 = 20.5$
- Bras inférieur jusqu'à 2; bras supérieur jusqu'à 14; 31 est une observation aberrante.

Exemple 6 (suite) : le diagramme et la moyenne



```
> tiques <- c(2, 3, 5, 6, 7, 8, 9, 11, 12, 14, 31)
> summary(tiques)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 2.000   5.500   8.000   9.818  11.500  31.000
> boxplot(tiques, horizontal = TRUE)
```

Exemple 6 (suite) : le diagramme et la moyenne



```
> tiques <- c(2, 3, 5, 6, 7, 8, 9, 11, 12, 14, 31)
> summary(tiques)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 2.000  5.500   8.000   9.818 11.500  31.000
> boxplot(tiques, horizontal = TRUE)
```

La souris très infestée tire la moyenne vers le haut : $\bar{x} = 9.82 > m = 8$. Sans elle, $\bar{x} = 7.7$ et $m = 7.5$: la moyenne baisse de plus de 2 tiques, la médiane de 0.5 seulement.

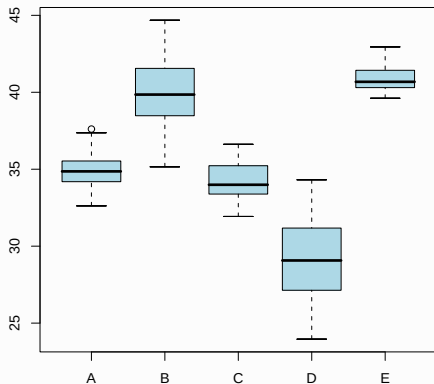
Que nous apprend un diagramme en boîte ?

En un coup d'œil, on voit :

- le **centre** de la distribution (la médiane);
- la **dispersion** (longueur de la boîte = EIQ, longueur des bras);
- la **symétrie** : la médiane est-elle au milieu de la boîte ? Les bras sont-ils de même longueur ?
- la présence d'observations **aberrantes**.

L'histogramme donne en général une meilleure idée de la forme de la distribution... mais le diagramme en boîte est **idéal pour comparer plusieurs échantillons** côte à côte.

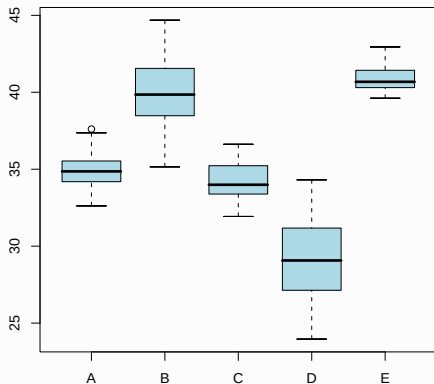
Diagrammes en boîte juxtaposés



On veut comparer le poids à la naissance dans 5 populations de serpents. Que remarquez-vous ?

Poids à la naissance (g) dans 5 populations de serpents
(données fictives)

Diagrammes en boîte juxtaposés



Poids à la naissance (g) dans 5 populations de serpents
(données fictives)

On veut comparer le poids à la naissance dans 5 populations de serpents. Que remarquez-vous ?

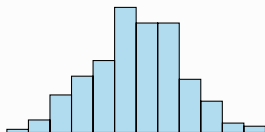
- Poids les plus élevés : populations **B** et **E**.
- Poids les plus faibles : population **D**.
- Plus grande variabilité : populations **B** et **D**.
- Population A : une observation aberrante.

Comparer les moyennes de plusieurs groupes : c'est l'objet de l'**ANOVA** (chapitre 5).

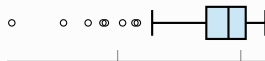
QUIZ!

Associez chaque histogramme (A, B, C) à son diagramme en boîte (1, 2, 3).

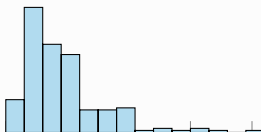
A



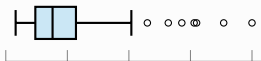
1



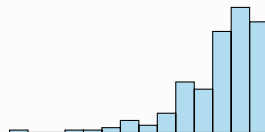
B



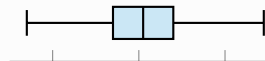
2



C



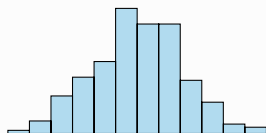
3



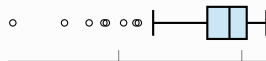
QUIZ!

Associez chaque histogramme (A, B, C) à son diagramme en boîte (1, 2, 3).

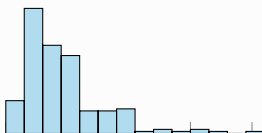
A



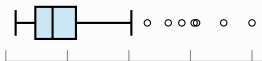
1



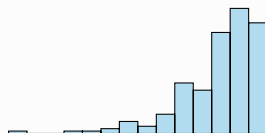
B



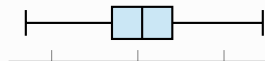
2



C



3



Pour chacune, la moyenne est-elle plus grande ou plus petite que la médiane?

Petit détail technique : la boîte de R

Lorsque n est **pair**, la fonction `boxplot()` de R ne place pas les côtés de la boîte exactement en Q_1 et Q_3 , mais à mi-chemin entre deux observations voisines (« charnières » de Tukey). Exemple avec 12 données :

```
> y <- c(1, 2, 3, 7, 7, 7, 8, 9, 10, 20, 21, 24)
> quantile(y, c(0.25, 0.75))    # Q1 et Q3 (définition du cours)
 25%   75%
 6.0 12.5
> boxplot.stats(y)$stats        # bras, charnières et médiane de boxplot()
[1]  1.0  5.0  7.5 15.0 24.0
```

Boîte de 6 à 12.5 selon la méthode du cours, de 5 à 15 dans R. Quand n est impair (serpents, tiques), les deux méthodes coïncident. En pratique, ces différences sont sans importance.

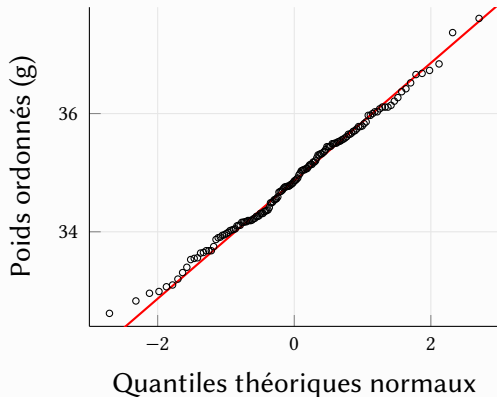
Complément : le diagramme quantile-quantile normal

Plusieurs méthodes des chapitres 3 à 6 supposent que les données proviennent d'une **loi normale** (distribution « en cloche », chapitre 2). Comment le vérifier ?

- **Principe** : on compare les quantiles de l'échantillon à ceux auxquels on s'attendrait sous une loi normale.
- **Technique** (*grosso modo*) : on associe à chaque statistique d'ordre $x_{(i)}$ le quantile de la loi normale d'ordre $\approx i/(n + 1)$, puis on trace les n points.
- **Lecture** : si le modèle normal est bon, les points sont à peu près **alignés** sur une droite.

Dans R : `qqnorm(poids)`, puis `qqline(poids)` pour ajouter la droite de référence.

Exemple 1 (suite) : les serpents

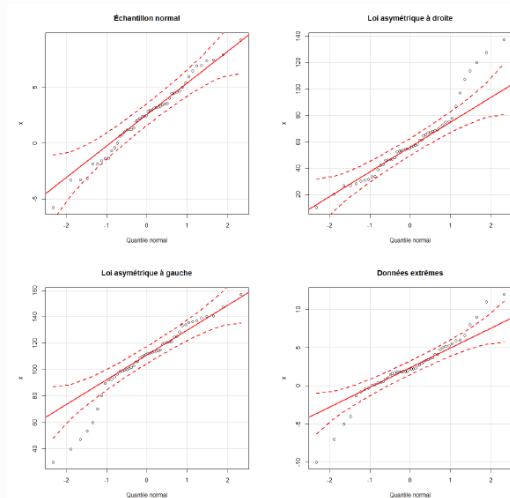


Les 147 points sont très bien alignés sur la droite (rouge).

⇒ La loi normale est un modèle raisonnable pour le poids à la naissance de ces serpents, comme le suggérait déjà l'histogramme « en cloche ».

Le serpent aberrant (37.61 g) est le dernier point en haut à droite : à lui seul, il ne remet pas en cause le modèle.

Comment lire un diagramme quantile-quantile



- **Échantillon normal** : points alignés (dans la bande pointillée).
- **Asymétrie à droite** : les points s'écartent vers le haut à droite (courbe convexe).
- **Asymétrie à gauche** : les points s'écartent vers le bas à gauche (courbe concave).
- **Valeurs extrêmes** (queues lourdes) : les deux extrémités s'écartent de la droite (forme de S).

1.7 Diagrammes en pointes de tarte et en bâtons

Exemple 7 : une variable qualitative

Pour une variable **qualitative**, on résume les données par un tableau de fréquences. On a demandé à 275 écoliers leur fruit préféré parmi les six fruits les plus consommés au Québec.

Fruit	Fréquence	Fréq. relative	Angle (degrés)
Banane	69		
Nectarine	35		
Orange	14		
Pêche	39		
Poire	41		
Pomme	77		
Total	275		

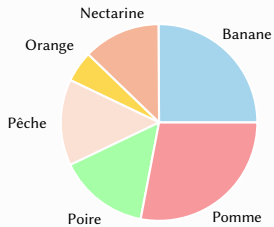
Exemple 7 : une variable qualitative

Pour une variable **qualitative**, on résume les données par un tableau de fréquences. On a demandé à 275 écoliers leur fruit préféré parmi les six fruits les plus consommés au Québec.

Fruit	Fréquence	Fréq. relative	Angle (degrés)
Banane	69	0.251	$\frac{69}{275} \times 360 = 90.33$
Nectarine	35	0.127	45.82
Orange	14	0.051	18.33
Pêche	39	0.142	51.05
Poire	41	0.149	53.67
Pomme	77	0.280	100.80
Total	275	1	360

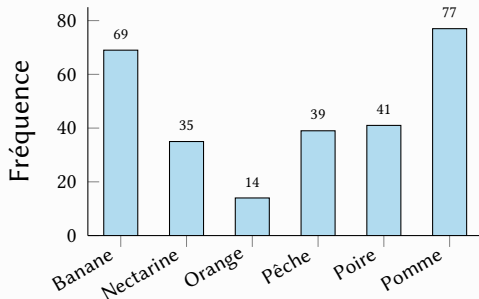
Exemple 7 (suite) : pointes de tarte et bâtons

Diagramme en pointes de tarte



Angle (et surface) de chaque pointe $\propto$ fréquence.

Diagramme en bâtons



L'œil compare mieux des **longueurs** que des **angles** : le diagramme en bâtons est généralement plus lisible (Poire ou Pêche?).

Diagramme en bâtons ou histogramme ?

- **Histogramme** : variable **quantitative** continue. L'axe horizontal est gradué ; les rectangles se touchent ; c'est la **surface** qui compte.
- **Diagramme en bâtons** : variable **qualitative** (ou quantitative discrète avec peu de valeurs). Les bâtons sont séparés ; l'ordre des catégories est arbitraire pour une variable nominale.

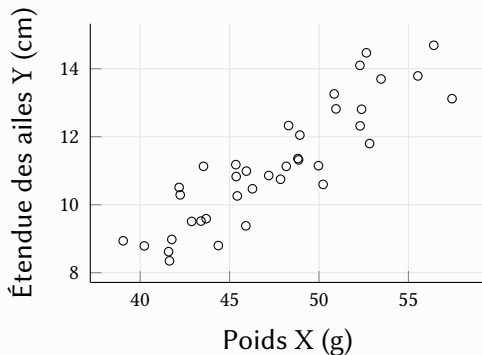
Dans R, pour une variable qualitative `f` : `table(f)` (tableau de fréquences), `barplot(table(f))` (bâtons), `pie(table(f))` (pointes de tarte).

Les tableaux de fréquences et leur analyse (test du khi-deux) seront approfondis au chapitre 7.

1.8 Les données bivariées

Exemple 8 : deux variables à la fois

On mesure deux variables quantitatives X et Y sur chacun des n individus : les données sont n couples $(x_1, y_1), \dots, (x_n, y_n)$.



Poids et étendue des ailes de 38 des 50 hirondelles de l'exemple 2 du chapitre 6.

Ce **diagramme de dispersion** (*scatterplot*) montre une association **positive** : les hirondelles plus lourdes ont tendance à avoir des ailes plus étendues.

Dans R : `plot(x, y)`.

Le coefficient de corrélation

Coefficient de corrélation échantillonnal

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- $-1 \leq r \leq 1$: il mesure le degré d'association **linéaire** entre X et Y.
- $r > 0$: association positive; $r < 0$: association négative; $r \approx 0$: pas d'association linéaire.

Le coefficient de corrélation

Coefficient de corrélation échantillonnal

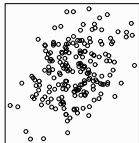
$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- $-1 \leq r \leq 1$: il mesure le degré d'association **linéaire** entre X et Y.
- $r > 0$: association positive; $r < 0$: association négative; $r \approx 0$: pas d'association linéaire.

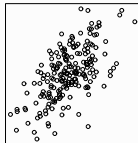
Exemple 8 (suite) : $\bar{x} = 47.53$ g, $s_x = 4.73$ g, $\bar{y} = 11.17$ cm, $s_y = 1.74$ cm, et $\text{cor}(x, y)$ donne $r = 0.890$: forte association linéaire positive.

Quelques valeurs de r

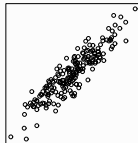
Corrélation = 0.3



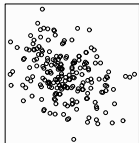
Corrélation = 0.6



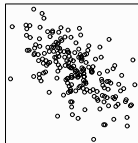
Corrélation = 0.9



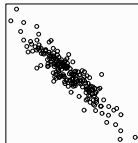
Corrélation = -0.3



Corrélation = -0.6



Corrélation = -0.9



Attention! Une forte corrélation ne prouve pas une relation de **cause à effet**. Et r ne mesure que l'association **linéaire** : une relation courbe peut donner $r \approx 0$. Toujours regarder le graphique! Nous y reviendrons au chapitre 6.

QUIZ!

Exemple 8 (hirondelles), $r = 0.890$.

- On exprime maintenant le poids en kilogrammes plutôt qu'en grammes. La valeur de r change-t-elle ?

QUIZ!

Exemple 8 (hirondelles), $r = 0.890$.

- On exprime maintenant le poids en kilogrammes plutôt qu'en grammes. La valeur de r change-t-elle?
- On calcule plutôt la corrélation entre l'étendue des ailes et le poids (on inverse les rôles de X et Y). Que vaut r ?

Que choisir ?

Type de données	Résumés numériques	Graphiques
Une variable quantitative	$\bar{x}$, m , quantiles; s , CV, EIQ	histogramme, boîte, quantile-quantile
Idem, plusieurs groupes	les mêmes, par groupe	boîtes juxtaposées
Une variable qualitative	fréquences (relatives)	bâtons, pointes de tarte
Deux variables quantitatives	coefficient de corrélation r	diagramme de dispersion

À retenir : la moyenne et l'écart-type sont sensibles aux valeurs extrêmes; la médiane et l'EIQ sont robustes. Toujours **regarder les données** (graphique) avant de calculer!

Prochain chapitre : les probabilités, pour passer de la description à l'inférence.

Aide-mémoire R

```
donnees <- read.csv("serpents.csv") # importer un fichier de données
x <- donnees$poids                  # extraire une variable
mean(x); median(x)                 # moyenne, médiane
sd(x); var(x); sd(x) / mean(x)     # écart-type, variance, CV
quantile(x, 0.4)                   # quantile d'ordre 0.4
summary(x); IQR(x)                 # résumé à cinq nombres + moyenne, EIQ
hist(x, breaks = seq(32.5, 38, by = 0.5)) # histogramme (classes choisies)
boxplot(x)                         # diagramme en boîte
boxplot(y ~ groupe, data = d)      # boîtes juxtaposées (une par groupe)
qqnorm(x); qqline(x)               # diagramme quantile-quantile normal
table(f); barplot(table(f))        # variable qualitative f
plot(x, y); cor(x, y)              # deux variables quantitatives
```

Réponses aux exemples

Quiz (types de variables) : a) qualitative nominale; b) quantitative continue; c) qualitative ordinale; d) quantitative discrète; e) quantitative continue; f) qualitative nominale; g) qualitative ordinale.

Quiz (exemple 2) : $(45 + 35)/155 = 0.516$, soit 51.6 %. C'est la surface des rectangles au-dessus de $(20, 40]$: $10 \times 0.0290 + 10 \times 0.0226 = 0.516$. La surface totale vaut 1; la somme des hauteurs n'a pas d'interprétation simple.

Exemple 3 : $\bar{x} = 22$ g, $s^2 = 6.5$ g², $s = 2.55$ g.

Quiz (moyenne, 30 patients) : Faux en général. La moyenne reste 112 seulement si les deux valeurs retirées ont une somme de $2 \times 112 = 224$.

Quiz (balance mal tarée) : $\bar{x}$ diminue de 2 g (32.882 g); s ne change pas (0.956 g). En mg : $\bar{x} = 32\,882$ mg et $s = 956$ mg (multipliés par 1000); le CV ne change pas en passant aux mg ($0.956/32.882 = 2.9\%$ dans les deux unités), mais il avait changé lors de la correction de 2 g ($2.74\% \rightarrow 2.91\%$).

Exemple 5 : $q_{0.3} = 19.34$ cm; $m = 20.4$ cm; $Q_1 = x_{(2.5)} = 18.7 + 0.5 \times 0.8 = 19.1$ cm; $Q_3 = x_{(5.5)} = 22.3 + 0.5 \times 2.3 = 23.45$ cm.

Réponses aux exemples (suite)

Quiz (médiane, 30 patients) : Vrai. Avec $n = 30$, la médiane est la moyenne de $x_{(15)}$ et $x_{(16)}$; après le retrait, ces deux valeurs sont toujours au milieu (ce sont les 14^e et 15^e des 28 valeurs restantes).

Quiz (séjour à l'hôpital) : asymétrie à droite, les grandes valeurs tirent la moyenne vers le haut : $\bar{x} = 4.61$ jours $> m = 3.8$ jours. Asymétrie à gauche : $\bar{x} < m$ (gestation : 39.1 contre 39.5 semaines). Symétrique : $\bar{x} \approx m$ (pression : 120.8 contre 121 mmHg).

Exemple 6 : résumé à cinq nombres (2, 5.5, 8, 11.5, 31); $ElQ = 6$; bornes -3.5 et 20.5 ; 31 est aberrante; $\bar{x} = 9.82$.

Quiz (histogrammes et boîtes) : A-3, B-2, C-1. A : $\bar{x} \approx m$; B : $\bar{x} > m$; C : $\bar{x} < m$.

Quiz (exemple 8) : r ne change pas quand on change d'unités (les données sont standardisées dans la formule); il est aussi symétrique en X et Y : $r = 0.890$ dans les deux cas.

Références

Ce document est rédigé avec Beamer ainsi qu'une classe .cls fournie par Jérôme Soucy.

Les diapositives sont adaptées des diapositives créées par Emmanuelle Reny-Nolin pour le cours STT-1920 ainsi que de celles des cours STT-1000 et STT-1900.

Référence : C. Bélisle, *STT-1920 Méthodes statistiques*, notes de cours, Université Laval, 2020, chapitre 1.