

# HLMA408: Traitement des données

Statistiques descriptives

Joseph Salmon

<http://josephsalmon.eu>

Université de Montpellier



# Sommaire

Exemple introductif: impact du tabac sur les nouveaux nés

Statistiques descriptives univariées et bivariées

Visualisation de distributions

# Sommaire

Exemple introductif: impact du tabac sur les nouveaux nés

Statistiques descriptives univariées et bivariées

Visualisation de distributions

# Objectifs

- ▶ Sur un jeu de données, comparer la masse à la naissance suivant le statut tabagique de la mère
- ▶ Revoir rapidement quelques outils de statistique descriptive

# Présentation des données babies23<sup>(1)</sup>

| Poids<br>Naissance | Statut<br>Tabagique | ... |
|--------------------|---------------------|-----|
| 120                | 0                   | ... |
| 113                | 0                   | ... |
| 128                | 1                   | ... |
| 123                | 0                   | ... |
| 108                | 1                   | ... |
| 136                | 0                   | ... |
| 138                | 0                   | ... |
| 132                | 0                   | ... |
| ⋮                  | ⋮                   | ... |

- ▶ Poids (onces):  $1 \text{ g} \approx 0.035 \text{ onces}$
- ▶ Statut :  
1, mère fumeuse  
0, mère non fumeuse
- ▶ Tableau entier :  
 $n = 1236$  observations
- ▶ Étude à l'œil nu impossible  
⇒ résumer les données par
  - quelques valeurs numériques
  - des graphiques parlants

---

<sup>(1)</sup> cf. <http://www.stat.berkeley.edu/users/statlabs/> pour une description, source <http://josephsalmon.eu/enseignement/datasets/babies23.data>

## Table de fréquences croisées

Étude sur données complètes<sup>(2)</sup> : le taux de mortalité infantile chez les enfants nés de mères fumeuses est plus faible<sup>(3)</sup> :

**Table:** Taux de mortalité infantile en fonction de la masse (g) à la naissance différencié selon le statut tabagique de la mère

| Masse du nourrisson (g) | Non fumeur | Fumeur |
|-------------------------|------------|--------|
| < 1500                  | 792 ‰      | 565 ‰  |
| 1500–2000               | 406 ‰      | 346 ‰  |
| 2000–2500               | 78 ‰       | 27 ‰   |
| 2500–3000               | 11.6 ‰     | 6.1 ‰  |
| 3000–3500               | 2.2 ‰      | 4.5 ‰  |
| ≥ 3500                  | 3.8 ‰      | 2.6 ‰  |

► Des critiques / commentaires sur le tableau?

---

<sup>(2)</sup>ici on n'a qu'une sous-partie de l'ensemble des données de l'étude initiale.

<sup>(3)</sup>D. Nolan and T. P. Speed. *Stat labs: mathematical statistics through applications*. Springer Science & Business Media, 2001.

## Corrigeons l'erreur ...

- ▶ Une autre étude préconise de travailler sur la masse à la naissance, après **standardisation**

masse **standardisée** de l'obs =  $\frac{\text{masse de l'obs.} - \text{masse moyenne}}{\text{écart-type de toutes les obs.}}$

- ▶ Cette standardisation est faite séparément pour les deux classes: pour les fumeurs et pour les non-fumeurs
- ▶ Intérêt: comparer ce qui est comparable ! Ici les bébés de mères fumeuses ont (en général) une masse plus faible

## Corrigeons l'erreur ...

- ▶ Une autre étude préconise de travailler sur la masse à la naissance, après **standardisation**

masse **standardisée** de l'obs = 
$$\frac{\text{masse de l'obs.} - \text{masse moyenne}}{\text{écart-type de toutes les obs.}}$$

- ▶ Cette standardisation est faite séparément pour les deux classes: pour les fumeurs et pour les non-fumeurs
- ▶ Intérêt: comparer ce qui est comparable ! Ici les bébés de mères fumeuses ont (en général) une masse plus faible
- ▶ Ainsi on comparera le taux de mortalité d'un bébé pesant 2680g (fumeur) à celui pesant 3000g (non-fumeur)



## Corrigeons l'erreur ...

- ▶ Une autre étude préconise de travailler sur la masse à la naissance, après **standardisation**

masse **standardisée** de l'obs = 
$$\frac{\text{masse de l'obs.} - \text{masse moyenne}}{\text{écart-type de toutes les obs.}}$$

- ▶ Cette standardisation est faite séparément pour les deux classes: pour les fumeurs et pour les non-fumeurs
- ▶ Intérêt: comparer ce qui est comparable ! Ici les bébés de mères fumeuses ont (en général) une masse plus faible
- ▶ Ainsi on comparera le taux de mortalité d'un bébé pesant 2680g (fumeur) à celui pesant 3000g (non-fumeur)

## Effets "cachés"<sup>(4)</sup>

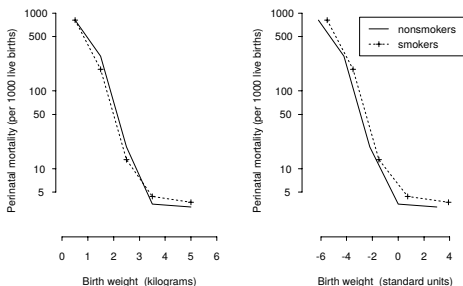


FIGURE 1.2. Mortality curves for smokers and nonsmokers by kilograms (left plot) and by standard units (right plot) of birth weight for the Missouri study (Wilcox [Wil93]).

- ▶ Il semblerait maintenant que les bébés de mères fumeuses aient un taux de mortalité plus élevé
- ▶ Faites attention aux effets cachés (**variables confondantes**)!
- ▶ Thème similaire: le paradoxe de Simpson

[https://www.youtube.com/watch?v=vs\\_Zzf\\_vL2I](https://www.youtube.com/watch?v=vs_Zzf_vL2I)

<sup>(4)</sup>D. Nolan and T. P. Speed. *Stat labs: mathematical statistics through applications*. Springer Science & Business Media, 2001.

# Sommaire

Exemple introductif: impact du tabac sur les nouveaux nés

## Statistiques descriptives univariées et bivariées

- Statistiques de tendance centrale/indicateurs de position

- Statistiques de dispersion

- Covariances et corrélations (de Pearson)

Visualisation de distributions

## Données étudiées

Pour les prochaines visualisations on utilise les tailles des pères (en cm), tirées de la base de données babies23.data obtenues par:

```
df_babies = pd.read_csv("babies23.data",  
                        skiprows=38)  
df_babies['dht'] # dht stands for ``dads height''
```

# Sommaire

Exemple introductif: impact du tabac sur les nouveaux nés

## Statistiques descriptives univariées et bivariées

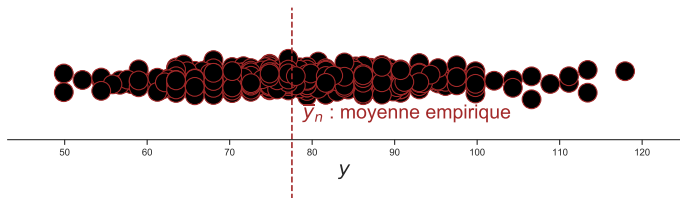
- Statistiques de tendance centrale/indicateurs de position

- Statistiques de dispersion

- Covariances et corrélations (de Pearson)

Visualisation de distributions

# Moyenne (arithmétique)

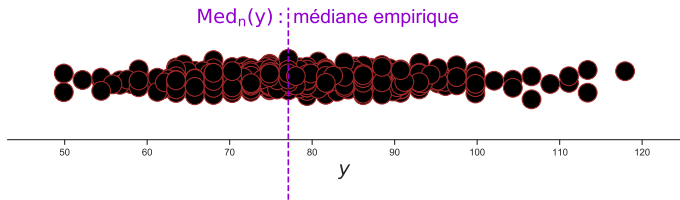


**Définition: Moyenne (arithmétique)**

$$\bar{y}_n = \frac{1}{n} \sum_{i=1}^n y_i$$

Notation:  $\mathbf{y} = (y_1, \dots, y_n)^\top$ , où le symbole  $\mathbf{y}^\top$  représente le transposé du vecteur  $\mathbf{y}$  (par convention on représente les vecteurs comme des colonnes :  $\mathbf{y} \in \mathbb{R}^n \iff \mathbf{y} \in \mathbb{R}^{n \times 1}$  )

# Médiane



On ordonne les  $y_i$  dans l'ordre croissant :  $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$

---

---

**Définition: Médiane**

---

---

$$\text{Med}_n(\mathbf{y}) = \begin{cases} \frac{y_{(\frac{n}{2})} + y_{(\frac{n}{2}+1)}}{2}, & \text{si } n \text{ est pair} \\ y_{(\frac{n+1}{2})}, & \text{si } n \text{ est impair} \end{cases}$$

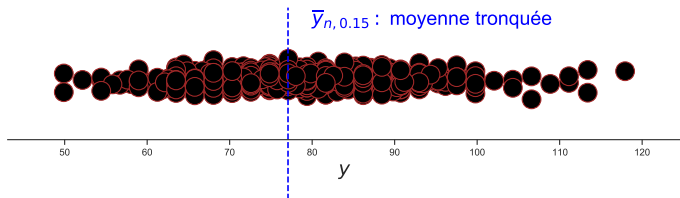
---

---

Rem. : utile pour décrire le niveau de richesse dans une population

Rem. : définition ambiguë : non unicité (idem pour les quantiles)

# Moyenne tronquée



Pour un paramètre  $\alpha$  (e.g.,  $\alpha = 15\%$ ), on calcule la moyenne en enlevant les  $\alpha\%$  plus grandes et plus petites valeurs

**Définition: Moyenne tronquée (à l'ordre  $\alpha$ )**

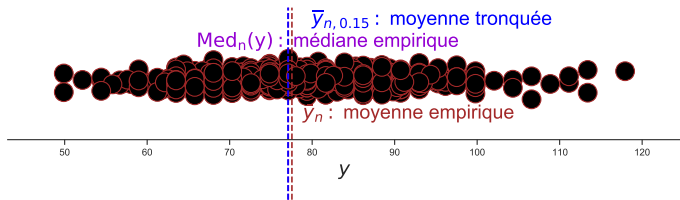
$$\bar{y}_{n,\alpha} = \bar{z}_n$$


où  $\mathbf{z} = (y_{(\lfloor \alpha n \rfloor)}, \dots, y_{(\lfloor (1-\alpha)n \rfloor)})$  est l'échantillon  $\alpha$ -tronqué

Rem. :  $\lfloor u \rfloor$  est le nombre entier tel que  $\lfloor u \rfloor \leq u < \lfloor u \rfloor + 1$

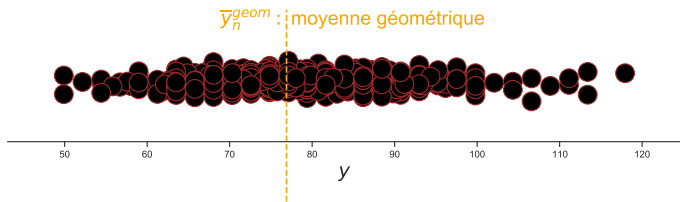


# Moyenne vs médiane



- ▶ Les trois statistiques ne coïncident pas
- ▶ Moyennes tronquées et médianes sont robustes aux points atypiques ( : *outliers*), la moyenne non!

# Moyenne géométrique (pour aller plus loin)



**Définition: Moyenne géométrique**

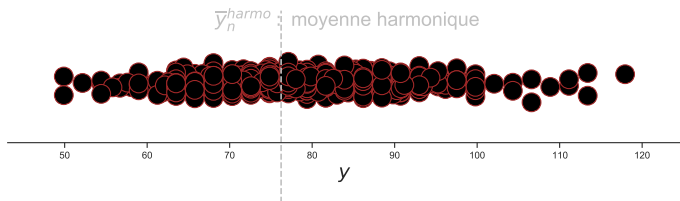
$$\bar{y}_n^{\text{geom}} = \left( \prod_{i=1}^n y_i \right)^{\frac{1}{n}} = \exp \left[ \frac{1}{n} \sum_{i=1}^n \ln y_i \right]$$

Rem. : définie uniquement pour des données positives

► usage: croissance exponentielle (virus, intérêts bancaires, etc.)

Voir aussi <https://www.youtube.com/watch?v=SmxKyTnfB2c>

# Moyenne harmonique (pour aller plus loin)



**Définition: Moyenne harmonique**

$$\bar{y}_n^{\text{harmonic}} = \frac{n}{\frac{1}{y_1} + \frac{1}{y_2} + \dots + \frac{1}{y_n}}$$

- Usage: physique (vitesse), apprentissage automatique ( $F_1$ ), etc.

Voir aussi

[https://fr.wikipedia.org/wiki/Moyenne\\_harmonique](https://fr.wikipedia.org/wiki/Moyenne_harmonique)

## Pour aller plus loin . . .

Inégalité reliant les diverses moyennes:

<https://twitter.com/TamasGorbe/status/1253987114104041472>

# Sommaire

Exemple introductif: impact du tabac sur les nouveaux nés

## Statistiques descriptives univariées et bivariées

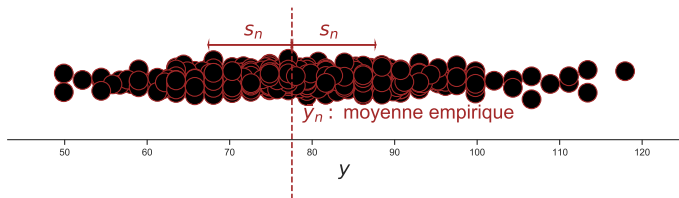
Statistiques de tendance centrale/indicateurs de position

Statistiques de dispersion

Covariances et corrélations (de Pearson)

Visualisation de distributions

# Dispersion: variance et écart-type



---

---

## Définitions

---

---

**Variance :** 
$$\text{var}_n(\mathbf{y}) = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_n)^2$$

**Écart-type :** 
$$s_n(\mathbf{y}) = \sqrt{\text{var}_n(\mathbf{y})}$$

---

---

Rem. : divers choix possibles pour le dénominateur ( $n$  ou  $n - 1$ ), soit un degré de liberté (🇬🇧 : *degree of freedom*) ddof=0 ou 1  
cf. <https://numpy.org/doc/stable/reference/generated/numpy.std.html>

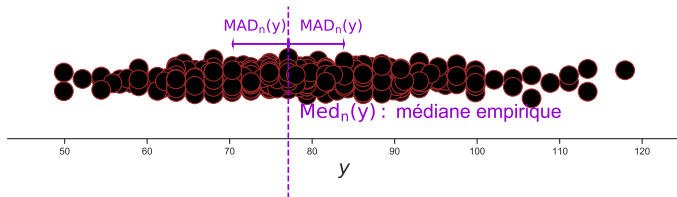
## Exercice (à faire seul)

---

**Exercice:** Décrire quels sont les vecteurs  $\mathbf{y} \in \mathbb{R}^n$  tels que  $\text{var}_n(\mathbf{y}) = 0$ .

---

# Dispersion: MAD



---

---

## Définition

---

---

**Déviations médiane absolue** (🇬🇧 : *Median Absolute Deviation*) :

$$\text{MAD}_n(\mathbf{y}) = \text{Med}_n (|\text{Med}_n(\mathbf{y}) - \mathbf{y}|)$$

où  $\text{Med}_n(\mathbf{y})$  est la médiane de l'échantillon  $\mathbf{y} = (y_1, \dots, y_n)^\top$

---

---



# Sommaire

Exemple introductif: impact du tabac sur les nouveaux nés

## Statistiques descriptives univariées et bivariées

- Statistiques de tendance centrale/indicateurs de position

- Statistiques de dispersion

- Covariances et corrélations (de Pearson)

Visualisation de distributions

# Covariances et corrélations empiriques

Soient deux échantillons  $\mathbf{x} = (x_1, \dots, x_n)^\top$  et  $\mathbf{y} = (y_1, \dots, y_n)^\top$

Notation:  $\mathbf{1}_n := (1, \dots, 1)^\top \in \mathbb{R}^n$  : vecteur constant

$$\langle \mathbf{x}, \mathbf{y} \rangle := \sum_{i=1}^n x_i y_i : \quad \text{produit scalaire de } \mathbf{x} \text{ et } \mathbf{y}$$

$$\|\mathbf{x}\| := \sqrt{\sum_{i=1}^n x_i^2} : \quad \text{norme}$$

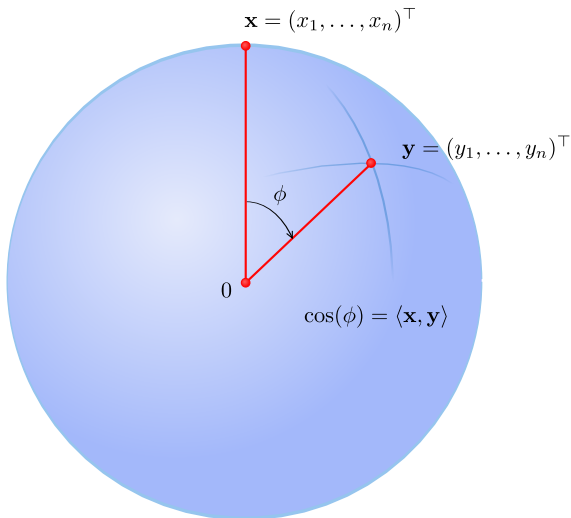
**Définition: covariance / corrélation empirique**

$$\text{cov}_n(\mathbf{x}, \mathbf{y}) := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n) = \frac{1}{n} \langle \mathbf{x} - \bar{x}_n \mathbf{1}_n, \mathbf{y} - \bar{y}_n \mathbf{1}_n \rangle$$

$$\text{corr}_n(\mathbf{x}, \mathbf{y}) := \frac{\text{cov}_n(\mathbf{x}, \mathbf{y})}{\sqrt{\text{var}_n(\mathbf{x})} \sqrt{\text{var}_n(\mathbf{y})}} = \frac{\langle \mathbf{x} - \bar{x}_n \mathbf{1}_n, \mathbf{y} - \bar{y}_n \mathbf{1}_n \rangle}{\|\mathbf{x} - \bar{x}_n \mathbf{1}_n\| \|\mathbf{y} - \bar{y}_n \mathbf{1}_n\|}$$

## Interprétation de la corrélation:

$$n = 3 \text{ et } \|\mathbf{x}\| = \|\mathbf{y}\| = 1$$



# Standardisation

Soit un échantillon  $\mathbf{x} = (x_1, \dots, x_n)^\top \in \mathbb{R}^n$

---

---

**Définition: échantillon standardisé**

---

---

On note  $\tilde{\mathbf{x}}$  l'**échantillon standardisé** de  $\mathbf{x}$  obtenu comme suit

$$\tilde{\mathbf{x}} = \frac{\mathbf{x} - \bar{x}_n \mathbf{1}_n}{s_n(\mathbf{x})} \iff \tilde{x}_i = \frac{x_i - \bar{x}_n}{s_n(\mathbf{x})}, \quad \forall i \in \llbracket 1, n \rrbracket$$

avec  $s_n(\mathbf{x}) = \sqrt{\text{var}_n(\mathbf{x})}$  son écart type et  $\bar{x}_n$  sa moyenne

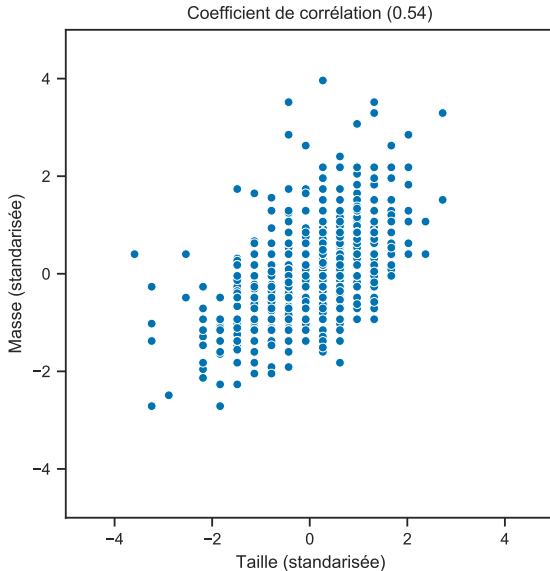
---

---

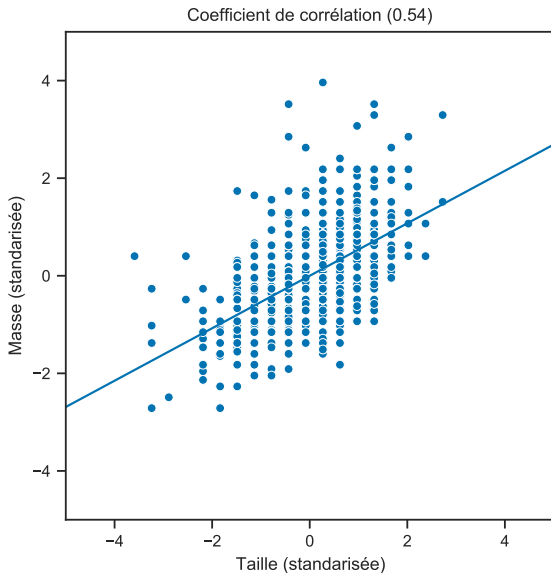
- ▶  $\tilde{\mathbf{x}}$  est
  - ▶ **centré** (moyenne nulle:  $\bar{\tilde{x}}_n = 0$ )
  - ▶ **réduit** (écart-type unitaire:  $s_n(\tilde{\mathbf{x}}) = 1$ )
- ▶  $\tilde{\mathbf{x}}$  est sans unité

Rem. : on passe de covariance à corrélation en standardisant ("réduire" suffirait) et la covariance des échantillons standardisés est la corrélation des échantillons originaux

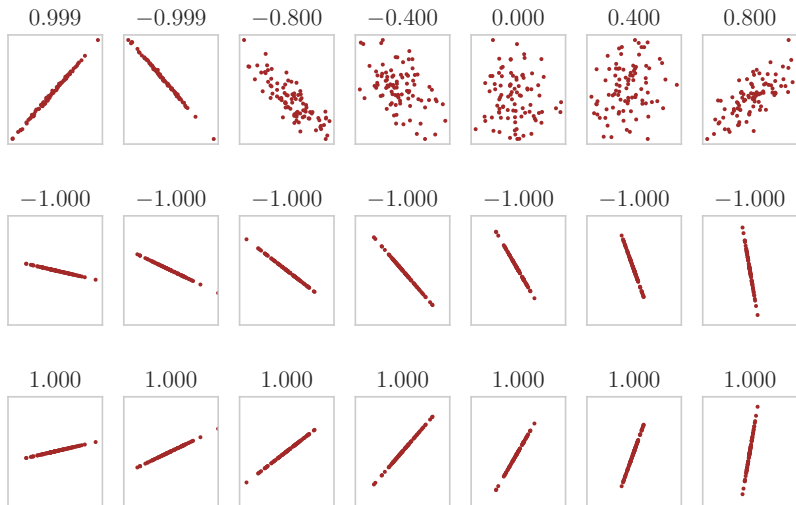
# Exemples de corrélations: taille du père / masse du père



# Exemples de corrélations: taille du père / masse du père

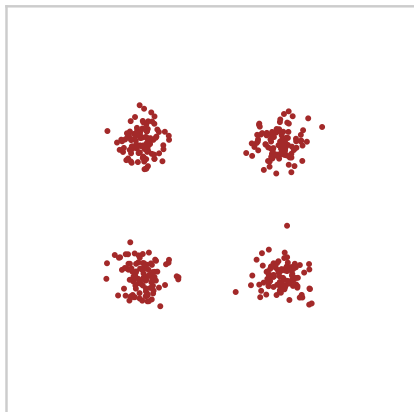


# Plus d'exemples



## Exemples de corrélations proches de zéro

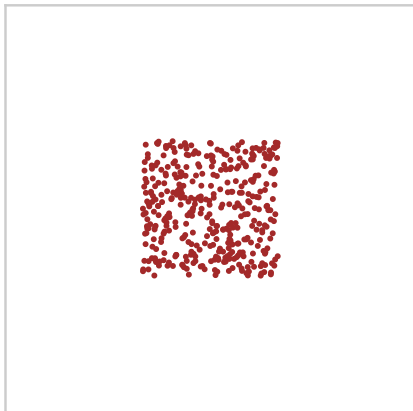
Corrélation =  $-0.021$





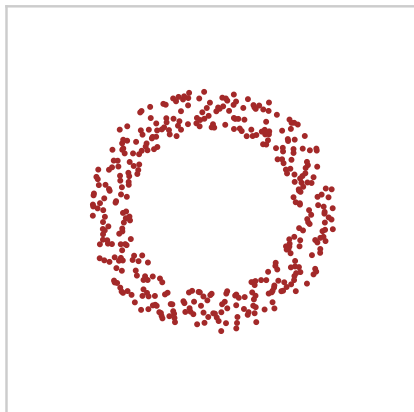
## Exemples de corrélations proches de zéro

Corrélation = 0.007



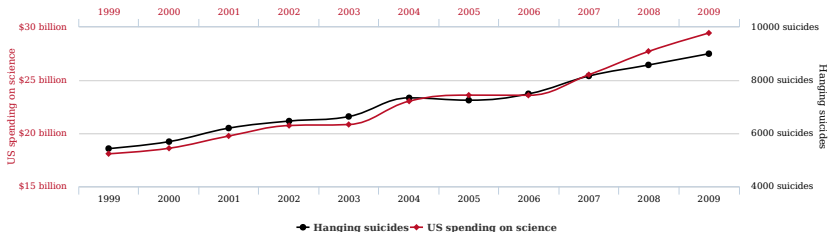
## Exemples de corrélations proches de zéro

Corrélation = 0.011



# Corrélation $\neq$ causalité

## US spending on science, space, and technology correlates with Suicides by hanging, strangulation and suffocation



tylervigen.com

Corrélation: 0.9979

cf. <http://www.tylervigen.com/spurious-correlations>

# Sommaire

Exemple introductif: impact du tabac sur les nouveaux nés

Statistiques descriptives univariées et bivariées

## Visualisation de distributions

- Rappels: quantiles et fonctions de répartition

- Histogrammes

- Boîtes à moustaches

- Méthode à noyau pour l'estimation de la densité

- Violons

# Sommaire

Exemple introductif: impact du tabac sur les nouveaux nés

Statistiques descriptives univariées et bivariées

## Visualisation de distributions

Rappels: quantiles et fonctions de répartition

Histogrammes

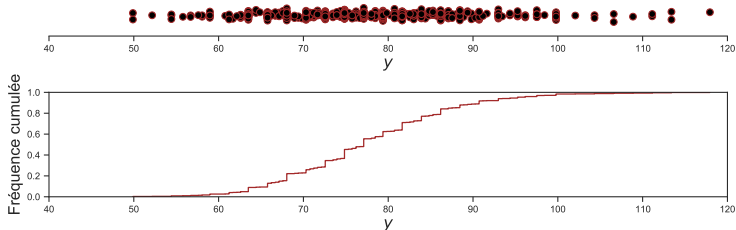
Boîtes à moustaches

Méthode à noyau pour l'estimation de la densité

Violons

# Fonction de répartition

Nombre d'échantillons:  $n = 695$



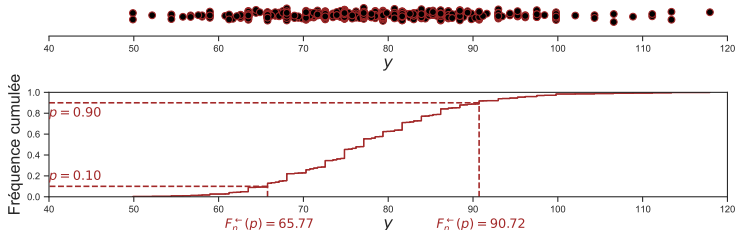
**Définition: fonction de répartition**

**Empirique :** 
$$F_n(u) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{y_i \leq u\}}, \text{ avec } \mathbb{1}_{\{y_i \leq u\}} = \begin{cases} 1, & \text{si } y_i \leq u, \\ 0, & \text{sinon.} \end{cases}$$

Interprétation: proportion d'observations sous un certain niveau

# Fonction quantile

Nombre d'échantillons:  $n = 695$



---

---

**Définition: Quantile**

---

---

Pour  $p \in ]0, 1]$ ,  $F_n^{\leftarrow}(p) = \inf\{u \in \mathbb{R} : F_n(u) \geq p\}$

---

---

Rem. : c'est l'inverse (généralisée) de la fonction de répartition; sa définition admet plusieurs conventions, cf. [percentile](#) dans Numpy

# Quantiles

En bref: “le quantile d'ordre  $p$  est le seuil tel que  $p \times 100\%$  des gens sont en dessous du seuil, et  $(1 - p) \times 100\%$  sont au-dessus”


- ▶ la médiane est le quantile d'ordre  $\frac{1}{2} = F_n^{\leftarrow}(\frac{1}{2})$
- ▶ le premier **quartile** ( $Q_1$ ) = quantile d'ordre  $\frac{1}{4} = F_n^{\leftarrow}(\frac{1}{4})$
- ▶ le troisième **quartile** ( $Q_3$ ) = quantile d'ordre  $\frac{3}{4} = F_n^{\leftarrow}(\frac{3}{4})$

Rem. : de manière similaire on parle de déciles et de centiles

---

## Définition

---

L'**Écart interquartile** ( : *Interquartile range*), noté IQR, est défini comme étant l'écart entre le 3<sup>e</sup> quartile et le 1<sup>er</sup> quartile:

$$IQR = F_n^{\leftarrow}(\frac{3}{4}) - F_n^{\leftarrow}(\frac{1}{4})$$

---




# Quantiles (seconde définition)

|                         |  |
|-------------------------|--|
| Échantillon :           | $y_1, \dots, y_n$  |
| Échantillon réordonné : | $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$   |
| Quantile d'ordre $p$ :  | valeur associée à l'indice $j \approx \lfloor pn \rfloor$<br>$= \begin{cases} \frac{y_{(j)} + y_{(j+1)}}{2}, & \text{si } j = pn \\ y_{(j+1)}, & \text{sinon} \end{cases}$ |

## Exemple :

- ▶  $n = 1000, p = \frac{1}{2} \implies j = 500 = \frac{1000}{2}$  et  $q_p(y) = \frac{y_{(500)} + y_{(501)}}{2}$
- ▶  $n = 1001, p = \frac{1}{2} \implies j = 500 \neq \frac{1001}{2}$  et  $q_p(y) = y_{(501)}$

 cette convention<sup>(5)</sup> ne coïncide pas avec la précédente, ici on choisit le milieu de l'intervalle au lieu de l'extrémité gauche

---

<sup>(5)</sup>voir <https://fr.wikipedia.org/wiki/Quantile> pour d'autres conventions possibles

# Sommaire

Exemple introductif: impact du tabac sur les nouveaux nés

Statistiques descriptives univariées et bivariées

## Visualisation de distributions

Rappels: quantiles et fonctions de répartition

**Histogrammes**

Boîtes à moustaches

Méthode à noyau pour l'estimation de la densité

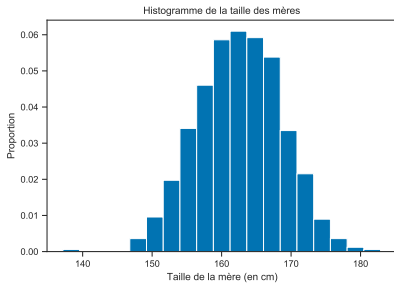
Violons

## Chargement (de nouveau)

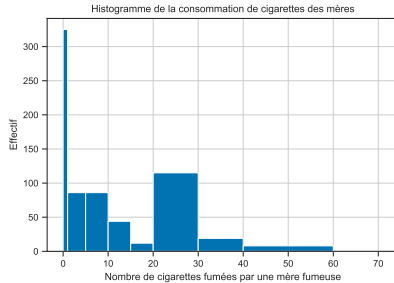
Pour les prochaines visualisations on utilise de nouveau la base de données babies23.data obtenues par:

```
df_babies = pd.read_csv("babies23.data",  
                        skiprows=38)
```

# Histogrammes




(a) Histogramme de la taille des mères: affichage d'une proportion densité (aire=1)



(b) Histogramme du nombre de cigarettes fumées par jour par les mères : effectif

# Qu'est-ce qu'un histogramme?

- ▶ Histogramme  $\neq$  diagramme en barre ( : *barplot*) !!!
- ▶ Décrit / estime la **distribution** : unimodalité, symétrie, étendue, ...
- ▶ Construction :
  - ▶ Axe horizontal : gradué (échelle des valeurs observées)
  - ▶ Axe vertical : **DENSITÉ!!!**  
de fréquence ou d'effectif

densité de fréquence de la classe  $k = \frac{\text{fréquence de la classe } k}{\text{longueur de la classe } k}$

densité d'effectif de la classe  $k = \frac{\text{effectif de la classe } k}{\text{longueur de la classe } k}$

- ▶ Attention à l'unité sur l'axe vertical (e.g., l'option `density=True/False` de `hist` en Matplotlib)

## Exemple de construction d'histogramme<sup>(6)</sup>

Nombre de cigarettes  
par jour pour les mères  
fumeuses:

| Nb de cig. | % de fumeurs |
|------------|--------------|
| 0          | 46.76        |
| 1-5        | 12.37        |
| 5-10       | 12.37        |
| 10-15      | 6.33         |
| 15-20      | 1.72         |
| 20-30      | 16.55        |
| 30-40      | 2.73         |
| 40-60      | 1.15         |
| 60-        | 0.00         |
| Total      | 100          |

- Problème de “bords”: le 5 appartient à quelle classe? choix = à la deuxième!

Rem. : toujours regarder l'aide pour savoir si `hist` est ouvert à droite ou à gauche (généralement :  $[a, b[$  )

- Hauteur du rectangle (cas densité):

$$h_0 = \frac{46.76}{1 \times 100} = 0.4676$$

$$h_1 = \frac{12.37}{4 \times 100} = 0.0309$$

$$\vdots = \vdots$$

$$h_{40} = \frac{1.15}{20 \times 100} = 0.00057$$

<sup>(6)</sup>[https://matplotlib.org/api/\\_as\\_gen/matplotlib.pyplot.hist.html](https://matplotlib.org/api/_as_gen/matplotlib.pyplot.hist.html)

## Exemple de construction d'histogramme<sup>(6)</sup>

Nombre de cigarettes  
par jour pour les mères  
fumeuses:

| Nb de cig. | % de fumeurs |
|------------|--------------|
| 0          | 46.76        |
| 1-5        | 12.37        |
| 5-10       | 12.37        |
| 10-15      | 6.33         |
| 15-20      | 1.72         |
| 20-30      | 16.55        |
| 30-40      | 2.73         |
| 40-60      | 1.15         |
| 60-        | 0.00         |
| Total      | 100          |

- Problème de “bords”: le 5 appartient à quelle classe? choix = à la deuxième!

Rem. : toujours regarder l'aide pour savoir si `hist` est ouvert à droite ou à gauche (généralement :  $[a, b[$  )

- Hauteur du rectangle (cas densité):

$$h_0 = \frac{46.76}{1 \times 100} = 0.4676$$

$$h_1 = \frac{12.37}{4 \times 100} = 0.0309$$

$$\vdots = \vdots$$

$$h_{40} = \frac{1.15}{20 \times 100} = 0.00057$$

<sup>(6)</sup>[https://matplotlib.org/api/\\_as\\_gen/matplotlib.pyplot.hist.html](https://matplotlib.org/api/_as_gen/matplotlib.pyplot.hist.html)

# Sommaire

Exemple introductif: impact du tabac sur les nouveaux nés

Statistiques descriptives univariées et bivariées

## Visualisation de distributions

Rappels: quantiles et fonctions de répartition

Histogrammes

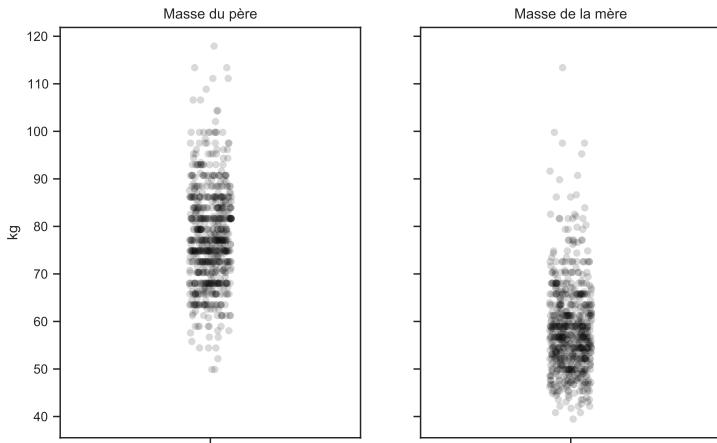
Boîtes à moustaches

Méthode à noyau pour l'estimation de la densité

Violons

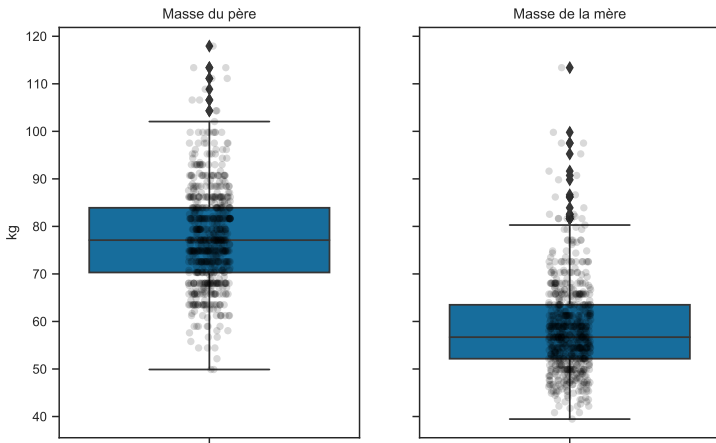


# Nuage de points (🇬🇧 : *scatterplot*)



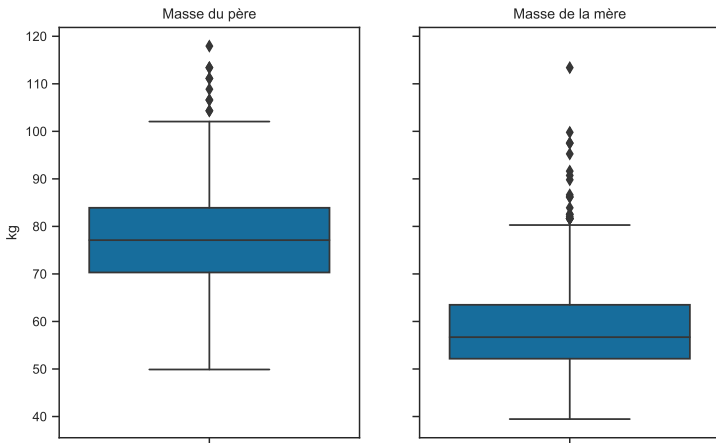
Masses des parents : nuages de points des

# Nuage de points (🇬🇧 : *scatterplot*) et boîte à moustaches (🇬🇧 : *boxplot*)



Masses des parents : nuages de points + boîtes à moustaches

# Boîte à moustaches (🇬🇧 : *boxplot*)



Masses des parents : boîtes à moustaches

# Qu'est-ce qu'une boîte à moustaches<sup>(7)</sup>?


- ▶ Représentation synthétique de la distribution d'une variable, similaire à l'histogramme, mais plus compacte
  
- ▶ Utilité :
  - ▶ Comparer des distributions
  - ▶ Permet de détecter les "valeurs aberrantes"
  - ▶ Visualiser un grand nombre de variables (compact), et représenter des variables quantitatives en fonction de variables qualitatives

---

<sup>(7)</sup>R. McGill, J. W. Tukey, and W. A. Larsen. "Variations of box plots". In: *The American Statistician* 32.1 (1978), pp. 12–16.

# Construction d'une boîte à moustaches

- ▶ la boîte est limitée par le 1<sup>er</sup> et le 3<sup>e</sup> quartiles
- ▶ elle est coupée en deux par la médiane
- ▶ les deux moustaches s'étendent de part et d'autre de la boîte sur une longueur (par défaut) de  $\frac{3}{2}$  fois l'écart inter-quartile

Rem. : il y a parfois des modifications à la marge pour les cas extrêmes, e.g., affichage de points aberrants ( : *outliers*)<sup>(8)</sup>

---

<sup>(8)</sup> cf. [https://matplotlib.org/api/\\_as\\_gen/matplotlib.pyplot.boxplot.html](https://matplotlib.org/api/_as_gen/matplotlib.pyplot.boxplot.html)

# Sommaire

Exemple introductif: impact du tabac sur les nouveaux nés

Statistiques descriptives univariées et bivariées

## Visualisation de distributions

Rappels: quantiles et fonctions de répartition

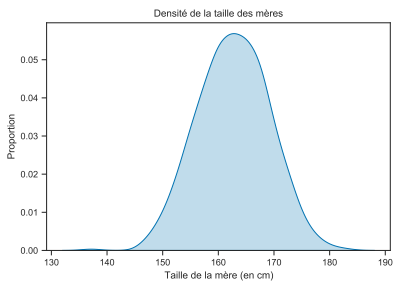
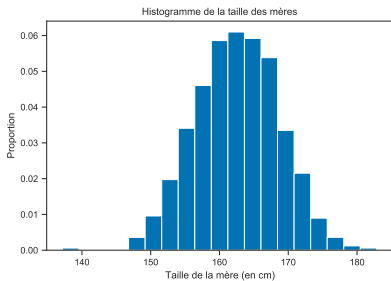
Histogrammes

Boîtes à moustaches

Méthode à noyau pour l'estimation de la densité

Violons

# Estimateur à noyau de la densité



(Gauche) histogramme — (droite) densité des tailles des mères

Rem. : on parle d'estimateur à noyau<sup>(9)</sup>,<sup>(10)</sup> de la densité

(🇬🇧 : *Kernel Density Estimator, KDE*)

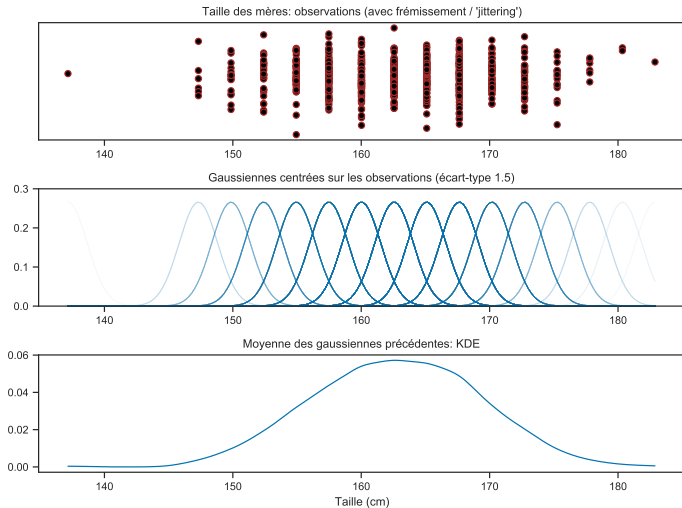
Rem. : les deux figures bleues ont une aire égale à 1

---

<sup>(9)</sup>M. Rosenblatt. "Remarks on some nonparametric estimates of a density function". In: *Ann. Math. Statist.* 27 (1956), pp. 832–837.

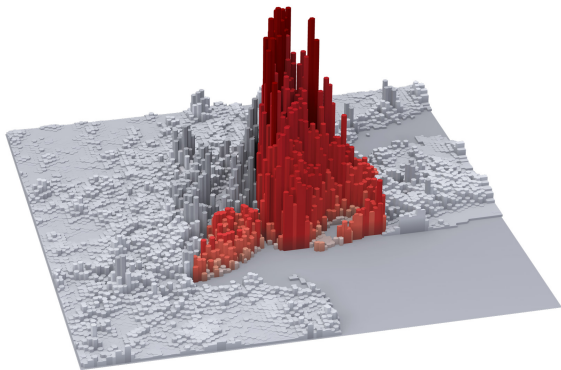
<sup>(10)</sup>E. Parzen. "On estimation of a probability density function and mode". In: *Ann. Math. Statist.* 33 (1962), pp. 1065–1076.

# Pour aller plus loin





# Histogramme pour données spatiales



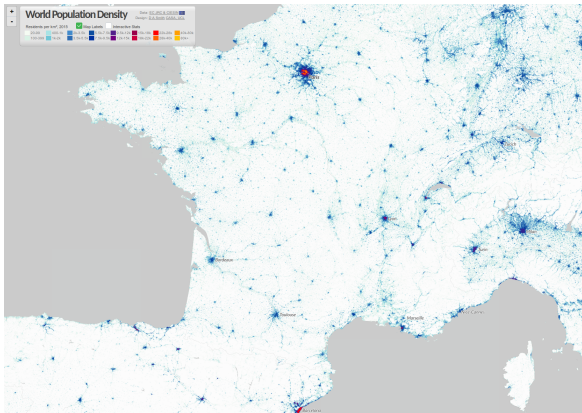
NE PAS UTILISER: effet masque !

Source:

<https://www.6sqft.com/see-how-nycs-urban-density-stacks-up-against-other-major-cities/>

# Densité spatiale

- Quantité numérique “intense” lorsqu’il y a beaucoup d’observations dans une région de l’espace et petite sinon

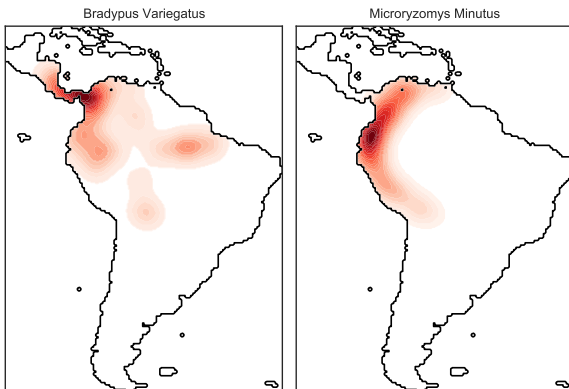


Densité de population humaine

Source: <https://luminocity3d.org/WorldPopDen/#7/45.491/2.115>

# Densité spatiale

- ▶ Quantité numérique “intense” lorsqu’il y a beaucoup d’observations dans une région de l’espace et petite sinon



Densité de population d'autres espèces...

Source: [https://scikit-learn.org/stable/auto\\_examples/neighbors/plot\\_species\\_kde.html](https://scikit-learn.org/stable/auto_examples/neighbors/plot_species_kde.html)

# Sommaire

Exemple introductif: impact du tabac sur les nouveaux nés

Statistiques descriptives univariées et bivariées

## Visualisation de distributions

Rappels: quantiles et fonctions de répartition

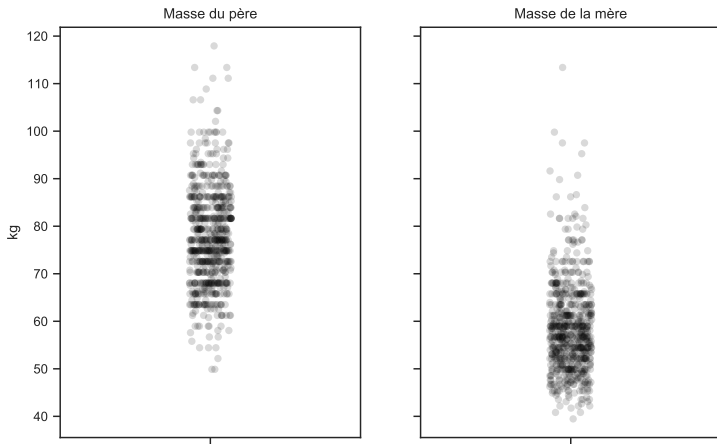
Histogrammes

Boîtes à moustaches

Méthode à noyau pour l'estimation de la densité

Violons

# Violons (🇬🇧 : *violins*)<sup>(11)</sup>

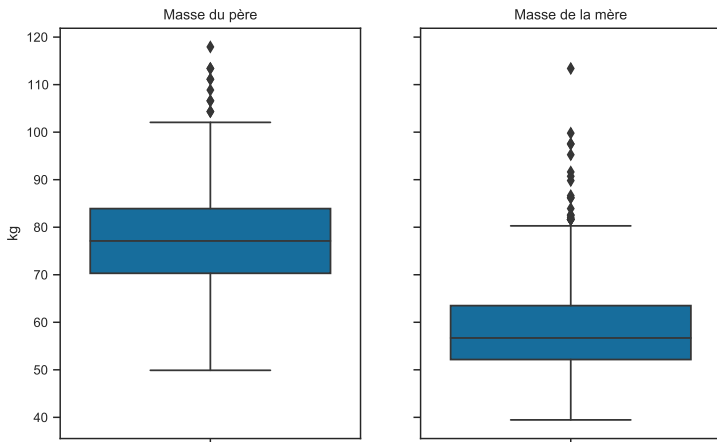


Nuages de points de la masse des parents

---

<sup>(11)</sup> J. L. Hintze and R. D. Nelson. "Violin plots: a box plot-density trace synergism". In: *The American Statistician* 52.2 (1998), pp. 181–184.

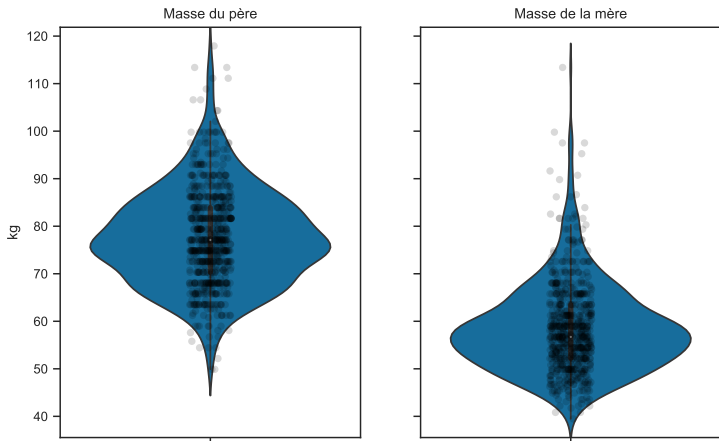
# Violons (🇬🇧 : *violins*)<sup>(11)</sup>



Boîte à moustaches, information condensée

<sup>(11)</sup> J. L. Hintze and R. D. Nelson. "Violin plots: a box plot-density trace synergism". In: *The American Statistician* 52.2 (1998), pp. 181–184.

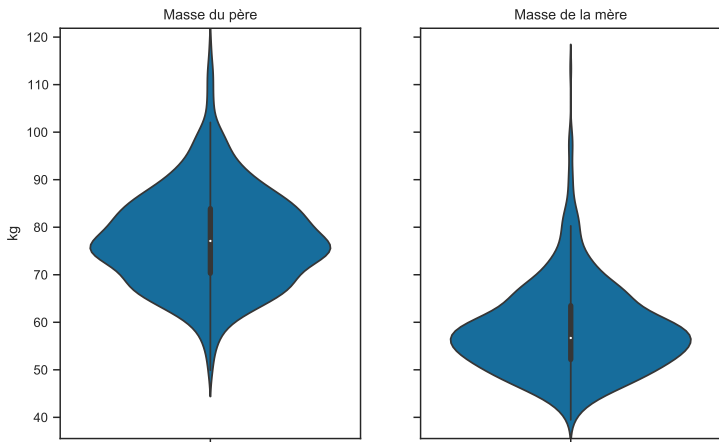
# Violons (🇬🇧 : *violins*)<sup>(11)</sup>



Violons de la masse des parents (estimateur de densité à noyau, pivoté et symétrisé) et nuage de points

<sup>(11)</sup> J. L. Hintze and R. D. Nelson. "Violin plots: a box plot-density trace synergism". In: *The American Statistician* 52.2 (1998), pp. 181–184.

# Violons (🇬🇧 : *violins*)<sup>(11)</sup>



Violons de la masse des parents (estimateur de densité à noyau, pivoté et symétrisé)

<sup>(11)</sup> J. L. Hintze and R. D. Nelson. "Violin plots: a box plot-density trace synergism". In: *The American Statistician* 52.2 (1998), pp. 181–184.



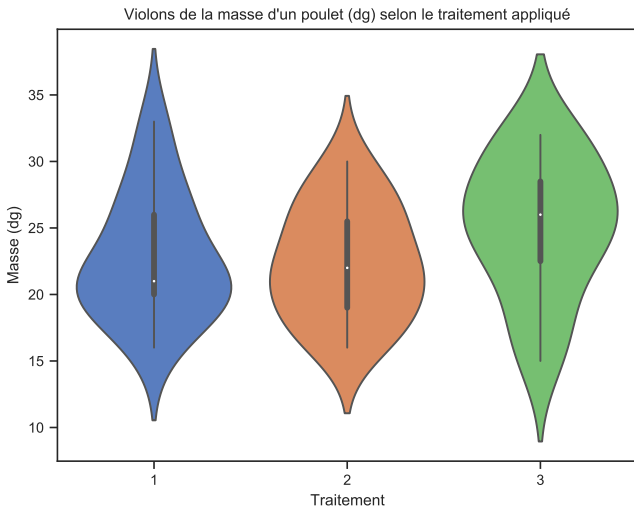
# Expérience: croissance des poussins

Scénario : des chercheurs veulent déterminer si parmi trois traitements possibles, il en existe un qui facilite la prise de masse des poussins.

Données : ils disposent des résultats des traitements (avec trois températures d'incubation différentes) sur la croissance de  $n = 45$  poussins.

- ▶ Les 45 œufs sont répartis aléatoirement entre les trois types de traitements (15 | 15 | 15)
- ▶ Au bout d'un nombre de jours fixé à l'avance, on note la croissance (masse, en dg) des poussins et leur sexe

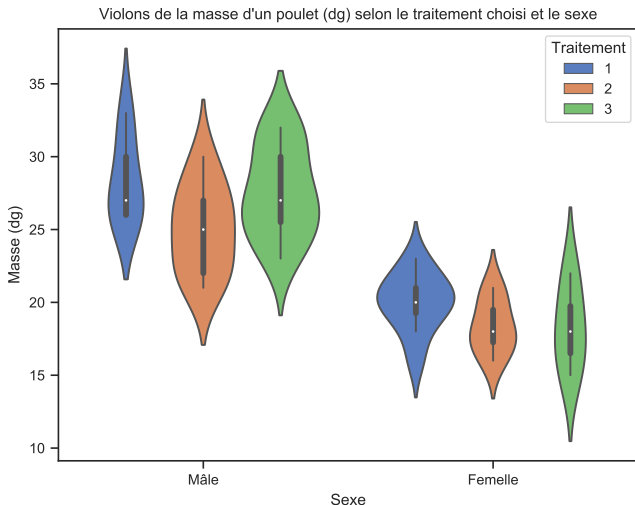
# Visualisation brute



Violons selon le type de traitement

Conclusion provisoire : le traitement 3 a le plus d'impact

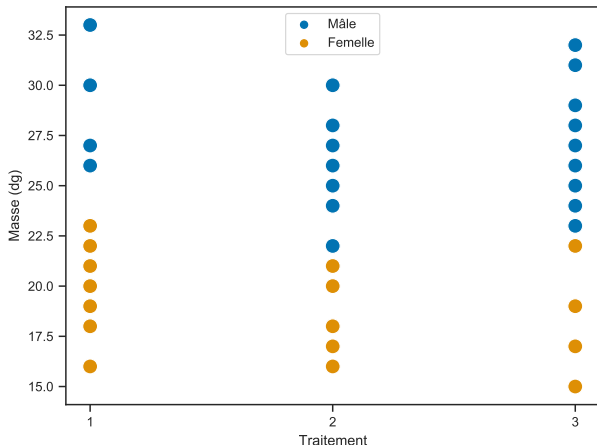
# Visualisation raffinée



Violons selon le type de traitement et le sexe

Conclusion : c'est le traitement 1 qui a le plus d'impact

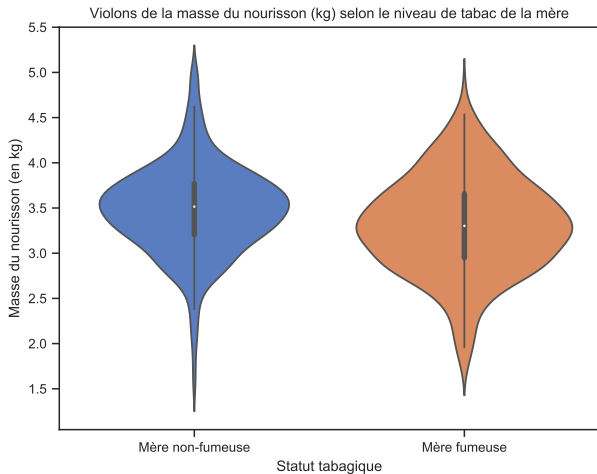
# Explication



Répartition des poussins par sexe et par traitement

Conclusion : il y avait trop de femelles dans le traitement 1, et l'effet sexe a caché l'impact du traitement (groupe inhomogène)

# Retour sur l'exemple du tabac



Masse du nourrisson selon le statut tabagique de la mère

Conclusion: les bébés de mères fumeuses ont une masse plus petite

## Biographie du jour : John Tukey<sup>(12)</sup>



- ▶ Mathématicien américain (1915-2000)
- ▶ Créateur de la Transformée de Fourier Rapide (FFT)
- ▶ Popularisa la **boîte à moustaches** en statistique
- ▶ Un des fondateurs de la statistique robuste (profondeur de Tukey)
- ▶ ...

---

<sup>(12)</sup>[https://fr.wikipedia.org/wiki/John\\_Tukey](https://fr.wikipedia.org/wiki/John_Tukey)

# Bibliographie I

- ▶ Hintze, J. L. and R. D. Nelson. “Violin plots: a box plot-density trace synergism”. In: *The American Statistician* 52.2 (1998), pp. 181–184.
- ▶ McGill, R., J. W. Tukey, and W. A. Larsen. “Variations of box plots”. In: *The American Statistician* 32.1 (1978), pp. 12–16.
- ▶ Nolan, D. and T. P. Speed. *Stat labs: mathematical statistics through applications*. Springer Science & Business Media, 2001.
- ▶ Parzen, E. “On estimation of a probability density function and mode”. In: *Ann. Math. Statist.* 33 (1962), pp. 1065–1076.
- ▶ Rosenblatt, M. “Remarks on some nonparametric estimates of a density function”. In: *Ann. Math. Statist.* 27 (1956), pp. 832–837.