HLMA408: Traitement des données

Statistiques descriptives

Joseph Salmon

http://josephsalmon.eu

Université de Montpellier

Sommaire

Sommaire

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`[1]

Poids Naissance	Statut Tabagique	...
120	0	...
113	0	...
128	1	...
123	0	...
108	1	...
136	0	...
138	0	...
132	0	...
⋮	⋮	...

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

[1] *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[2] : le taux de mortalité infantile chez les enfants nés de mères fumeuses est plus faible[3] :

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?

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

(3) 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**

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

- 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**

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

- 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**

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

- 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"[4]

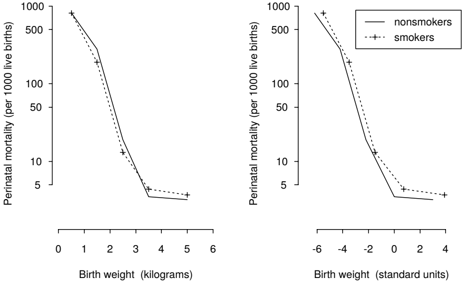


FIGURE 1.2. Mortality curves for smokers and nonsmokers by kilograms (left plot) and by standard units (right plot) of birth weight for the Missouri study (Wilcox [Wil93]).

- 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

[4] 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

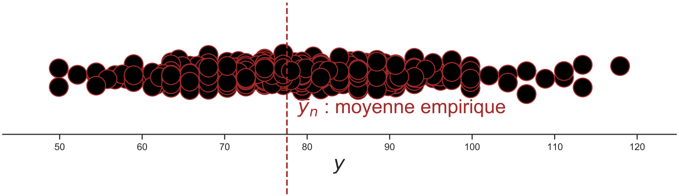
Données étudiées

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

Moyenne (arithmétique)

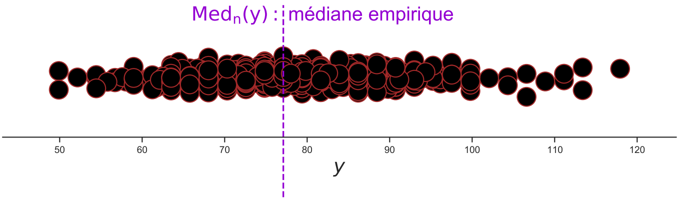


Définition: Moyenne (arithmétique)

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

Notation: $\mathbf{y} = (y_1, \ldots, y_n)^\top$, 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 : $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$

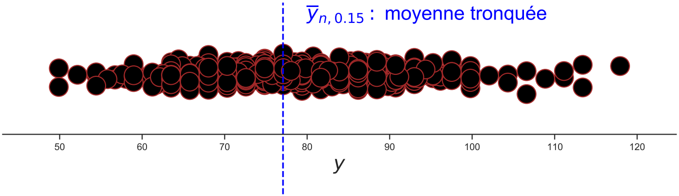
Définition: Médiane

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

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

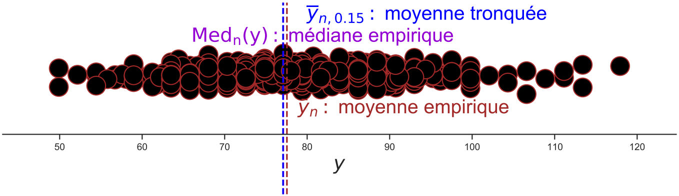
Définition: Moyenne tronquée (à l'ordre α)

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

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

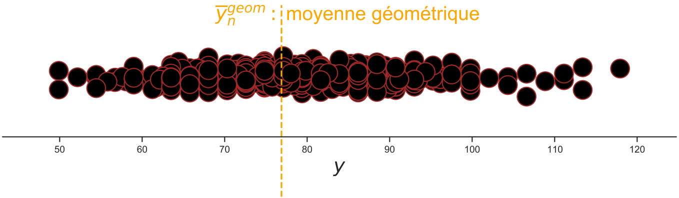
Rem. : $\lfloor u \rfloor$ est le nombre entier tel que $\lfloor u \rfloor \leq u < \lfloor u \rfloor + 1$

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

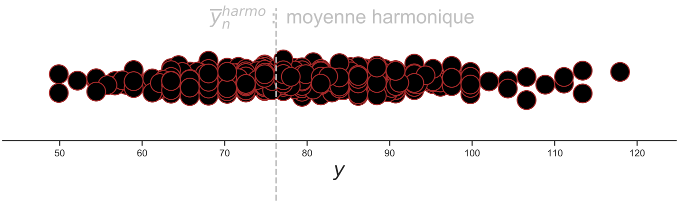
$$\overline{y}_n^{\text{geom}} = \left(\prod_{i=1}^{n} y_i\right)^{\frac{1}{n}} = \exp\left[\frac{1}{n}\sum_{i=1}^{n} \ln y_i\right]$$

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

$$\overline{y}_n^{\text{harmo}} = \frac{n}{\frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_n}}$$

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

Voir aussi
https://fr.wikipedia.org/wiki/Moyenne_harmonique

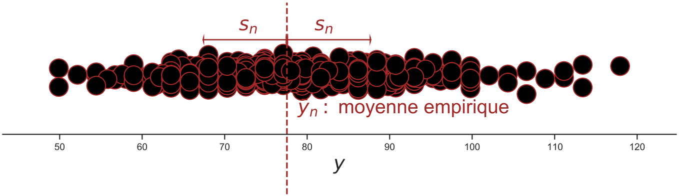
Pour aller plus loin . . .

Inégalité reliant les diverses moyennes:

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

Sommaire

Dispersion: variance et écart-type



Définitions

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

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

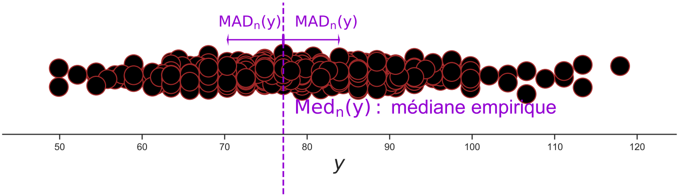
Rem. : divers choix possibles pour le dénominateur (n ou $n-1$), soit un degré de liberté (🇬🇧 : *degree of freedom*) ddof=0 ou 1

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 $\mathrm{var}_n(\mathbf{y}) = 0$.

Dispersion: MAD



Définition

Déviation médiane absolue (: *Median Absolute Deviation*) :

$$\mathrm{MAD}_n(\mathbf{y}) = \mathrm{Med}_n\left(|\mathrm{Med}_n(\mathbf{y}) - \mathbf{y}|\right)$$

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

Sommaire

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

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

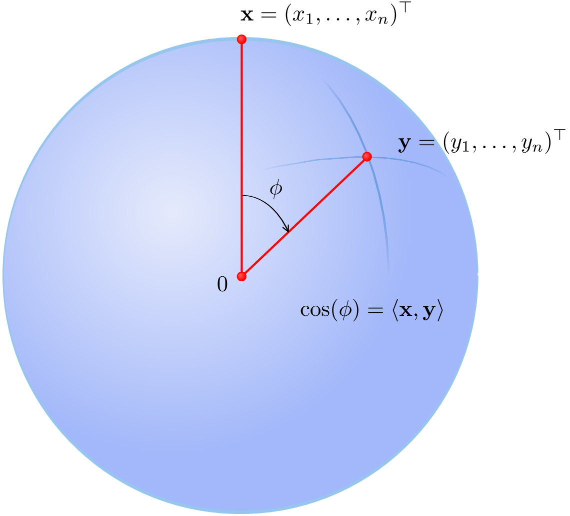
$$\|\mathbf{x}\| := \sqrt{\sum_{i=1}^n x_i^2} : \quad \textbf{norme}$$

Définition: covariance /corrélation empirique

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

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

Interprétation de la corrélation: $n = 3$ 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

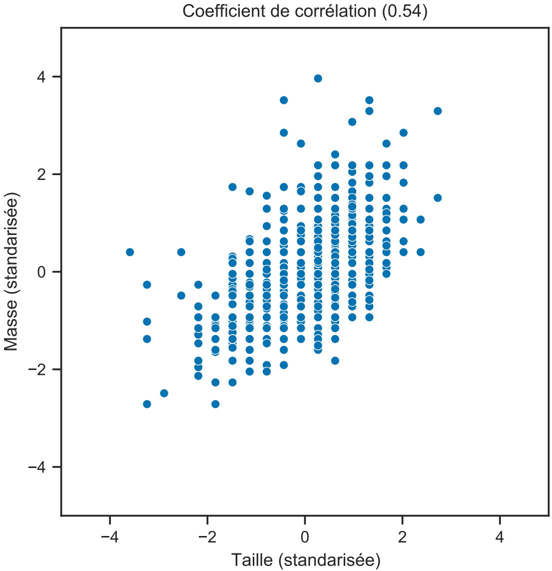
$$\tilde{\mathbf{x}} = \frac{\mathbf{x} - \bar{x}_n \mathbf{1}_n}{s_n(\mathbf{x})} \iff \tilde{x}_i = \frac{x_i - \bar{x}_n}{s_n(\mathbf{x})}, \quad \forall i \in [\![1, n]\!]$$

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

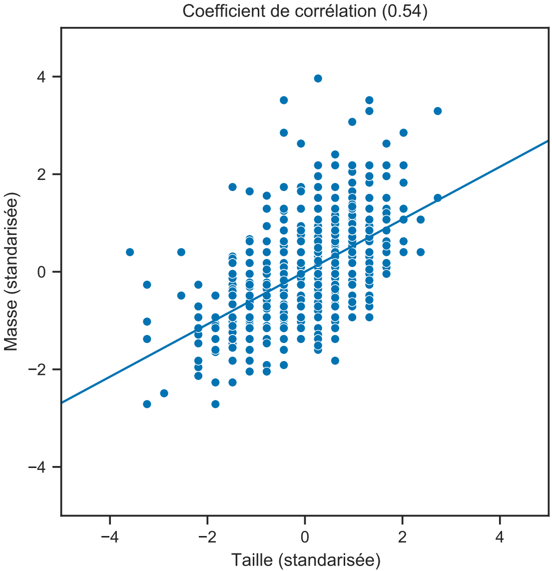
- $\tilde{\mathbf{x}}$ est
  - **centré** (moyenne nulle: $\bar{\tilde{x}}_n = 0$)
  - **réduit** (écart-type unitaire: $s_n(\tilde{\mathbf{x}}) = 1$)
- $\tilde{\mathbf{x}}$ est sans unité

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

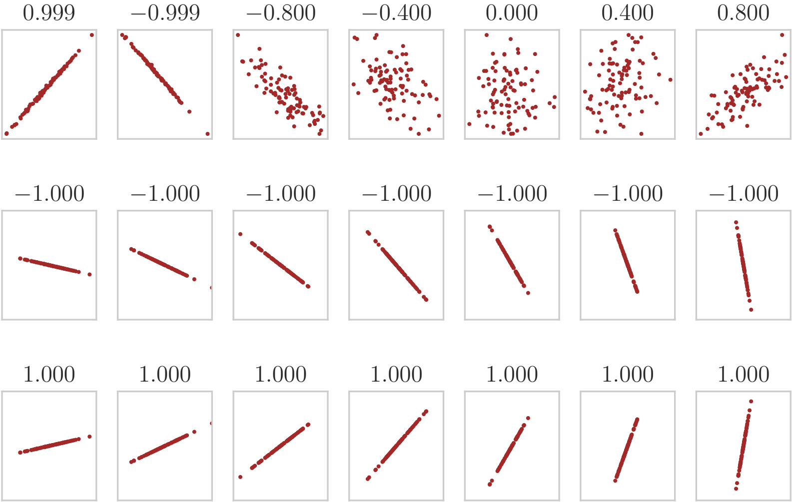
Exemples de corrélations: taille du père / masse du père



Exemples de corrélations: taille du père / masse du père

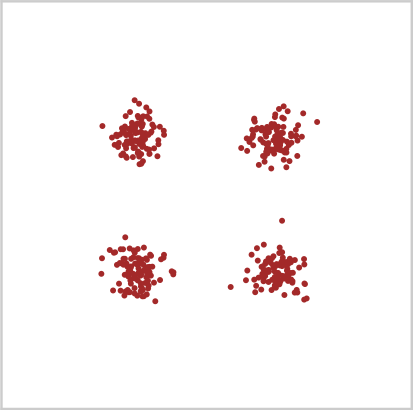


Plus d'exemples



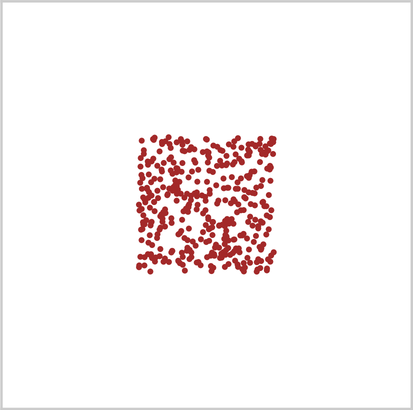
Exemples de corrélations proches de zéro

Corrélation $= -0.021$



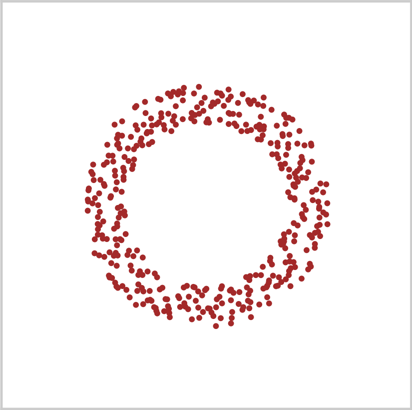
Exemples de corrélations proches de zéro

Corrélation $= 0.007$

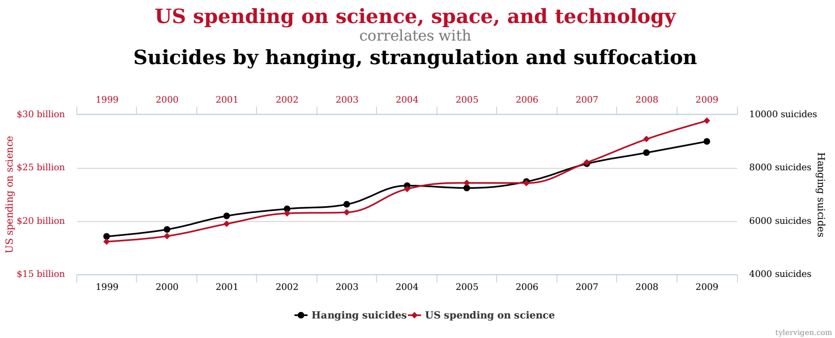


Exemples de corrélations proches de zéro

Corrélation $= 0.011$



Corrélation $\neq$ causalité



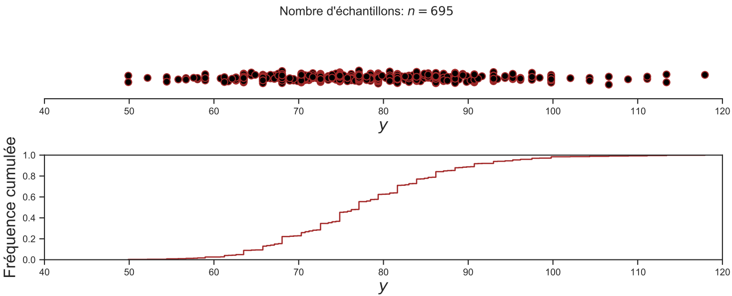
Corrélation: 0.9979

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

Sommaire

Sommaire

Fonction de répartition

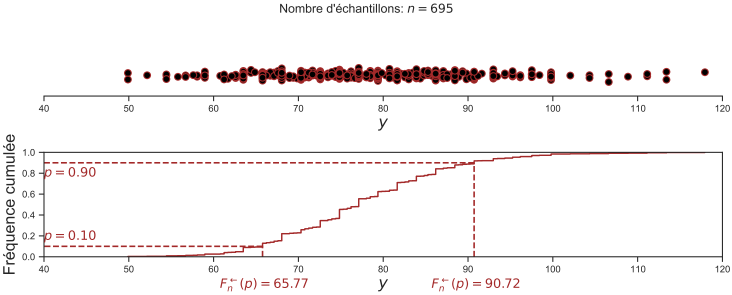


Définition: fonction de répartition

Empirique : $F_n(u) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{y_i \leq u\}}$, 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



Définition: Quantile

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

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 **quart**ile (Q_1) = quantile d'ordre $\frac{1}{4} = F_n^{\leftarrow}(\frac{1}{4})$
- le troisième **quart**ile (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 3e quartile et le 1er quartile:

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

Quantiles (seconde définition)

Échantillon : $y_1, \ldots, y_n$

Échantillon réordonné : $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$

Quantile d'ordre p : valeur associée à l'indice $j \approx \lfloor pn \rfloor$

$$= \begin{cases} \frac{1}{2}(y_{(j)} + y_{(j+1)}), & \text{si } j = pn \\ y_{(j+1)}, & \text{sinon} \end{cases}$$

Exemple :

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

⚠ cette convention[5] ne coïncide pas avec la précédente, ici on choisit le milieu de l'intervalle au lieu de l'extrémité gauche

[5] voir https://fr.wikipedia.org/wiki/Quantile pour d'autres conventions possibles

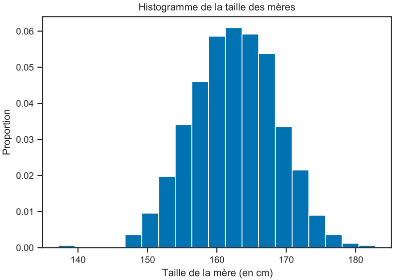
Sommaire

Chargement (de nouveau)

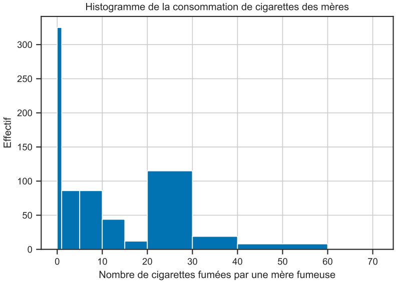
Pour les prochaines visualisations on utilise de nouveau la base de données `babies23.data` obtenues par:

```
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

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

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

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

Exemple de construction d'histogramme[6]

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):

$$
\begin{aligned}
h_0 &= \frac{46.76}{1 \times 100} = 0.4676 \\
h_1 &= \frac{12.37}{4 \times 100} = 0.0309 \\
\vdots &= \vdots \\
h_{40} &= \frac{1.15}{20 \times 100} = 0.00057
\end{aligned}
$$

(6) https://matplotlib.org/api/_as_gen/matplotlib.pyplot.hist.html

Exemple de construction d'histogramme[6]

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!

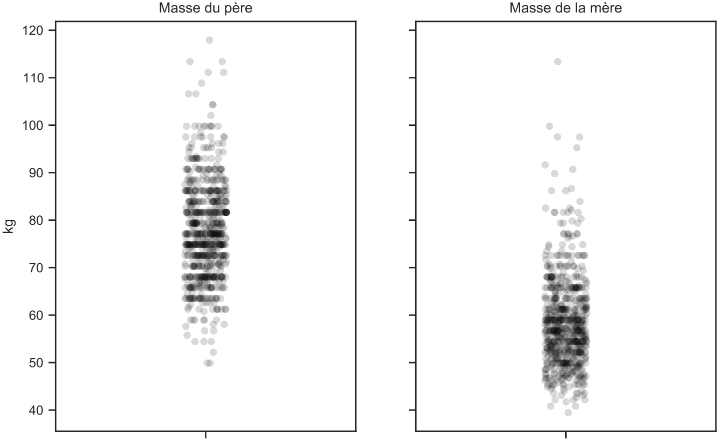
  Rem. : toujours regarder l'aide pour savoir si `hist` est ouvert à droite ou à gauche (généralement : $[a, b[$)
- Hauteur du rectangle (cas densité):

$$h_0 = \frac{46.76}{1 \times 100} = 0.4676$$
$$h_1 = \frac{12.37}{4 \times 100} = 0.0309$$
$$\vdots = \vdots$$
$$h_{40} = \frac{1.15}{20 \times 100} = 0.00057$$

[6] https://matplotlib.org/api/_as_gen/matplotlib.pyplot.hist.html

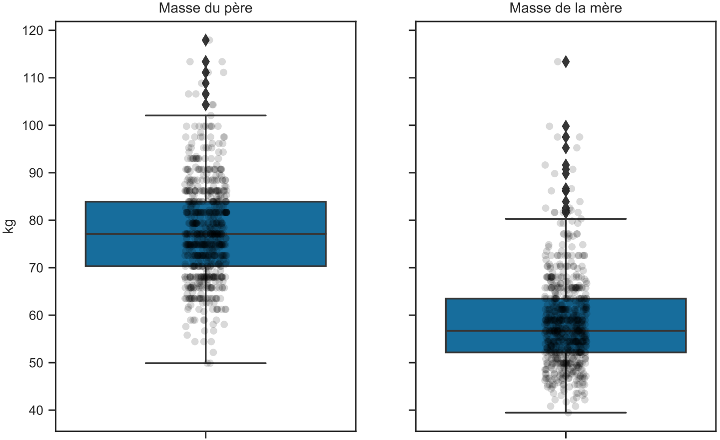
Sommaire

Nuage de points (: *scatterplot*)



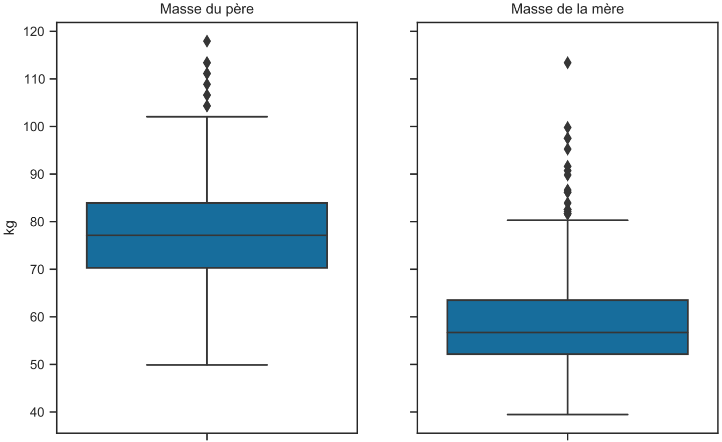
Nuages de points des masses des parents

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



Nuages de points et boîtes à moustache des masses des parents.

Boîte à moustache (: *boxplot*)



Boîtes à moustache des masses des parents.

Qu'est-ce qu'une boîte à moustache[7]?

- 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

[7] 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 moustache

- la boîte est limitée par le 1^{er} et le 3^{e} 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

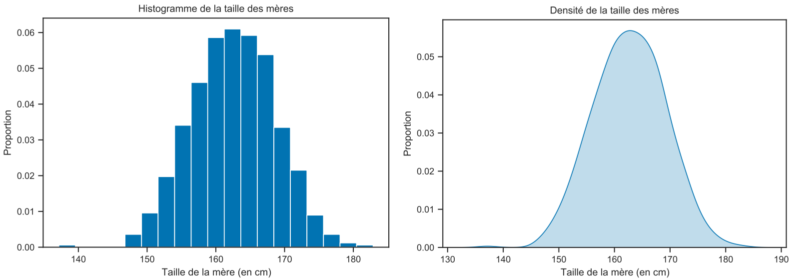
Rem. : il y a parfois des modifications à la marge pour les cas extrêmes, *e.g.*, affichage de points aberrants (: *outliers*)[(8)]

---

(8) *cf.* https://matplotlib.org/api/_as_gen/matplotlib.pyplot.boxplot.html

Sommaire

Estimateur à noyau de la densité



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

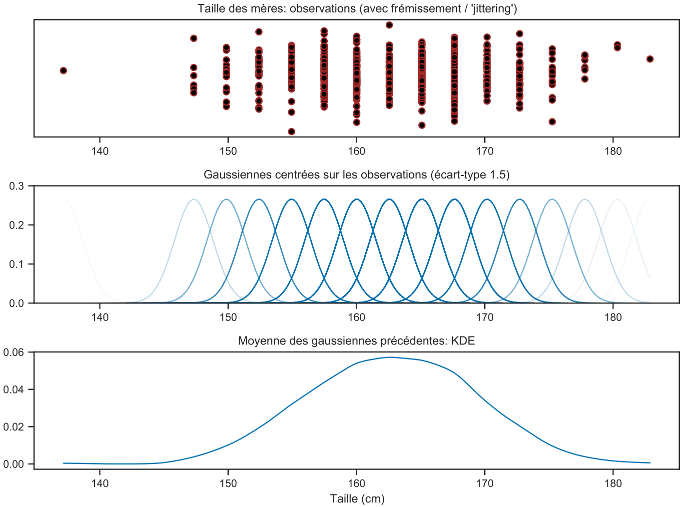
Rem. : on parle d'estimateur à noyau[(9), (10)] de la densité (🇬🇧 : *Kernel Density Estimator, KDE*)

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

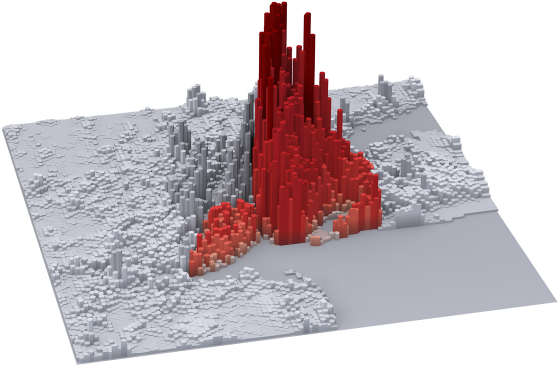
(9) M. Rosenblatt. "Remarks on some nonparametric estimates of a density function". In: *Ann. Math. Statist.* 27 (1956), pp. 832–837.

(10) 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



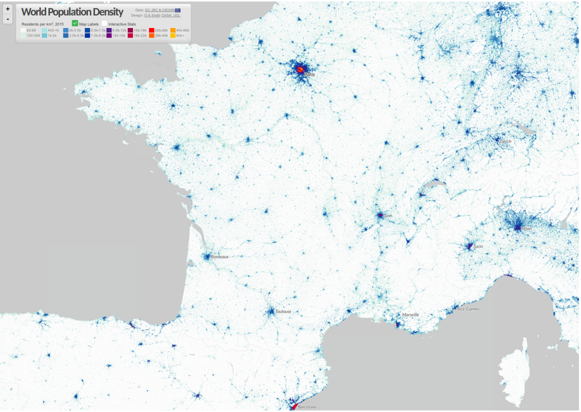
⚠ NE PAS UTILISER: effet masque !

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

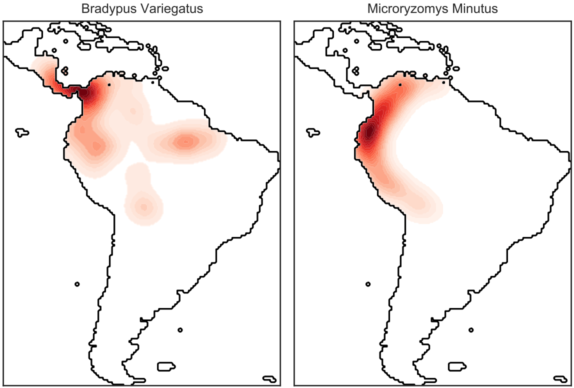


Densité de population humaine

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

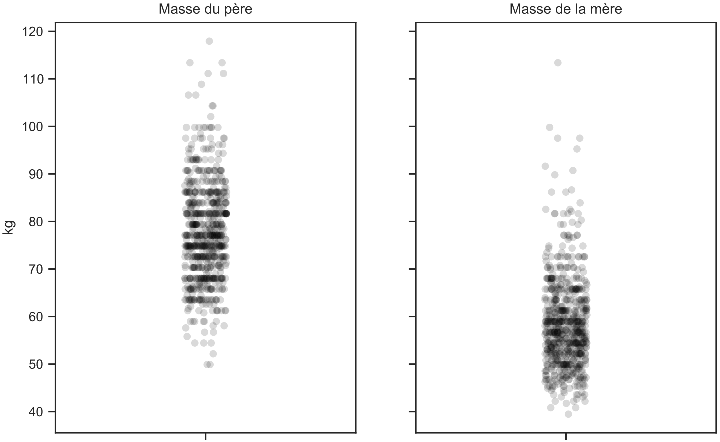


Densité de population d’autres espèces...

Source: https://scikit-learn.org/stable/auto_examples/neighbors/plot_species_kde.html

Sommaire

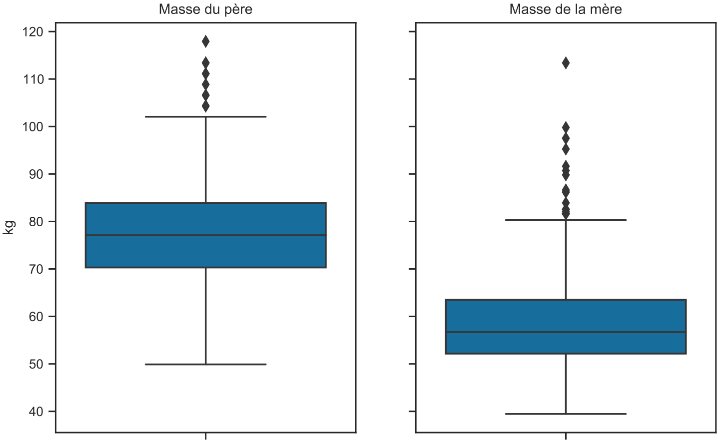
Violons (🇬🇧 : *violins*)[11]



Nuages de points de la masse des parents

[11] 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.

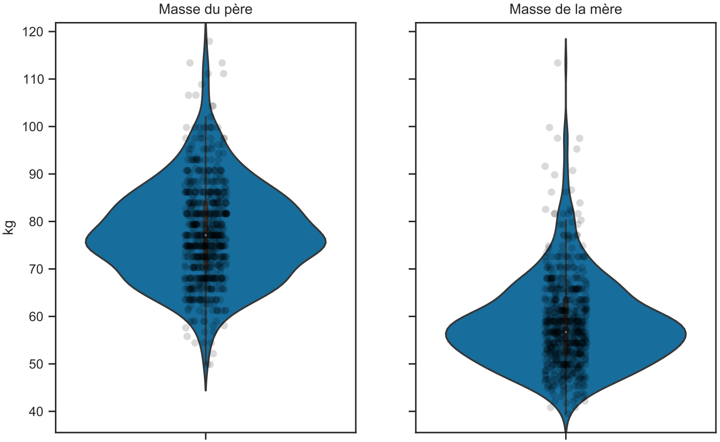
Violons (🇬🇧 : *violins*)[11]



Boîte à moustache, information condensée

[11] 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.

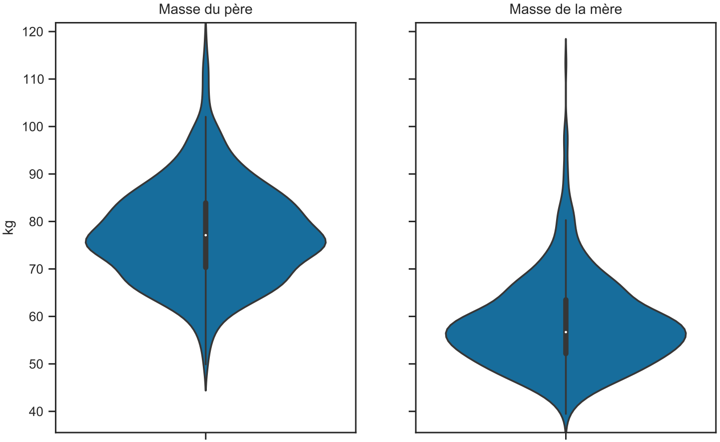
Violons (🇬🇧 : *violins*)[11]



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

[11] 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*)[11]



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

[11] 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.

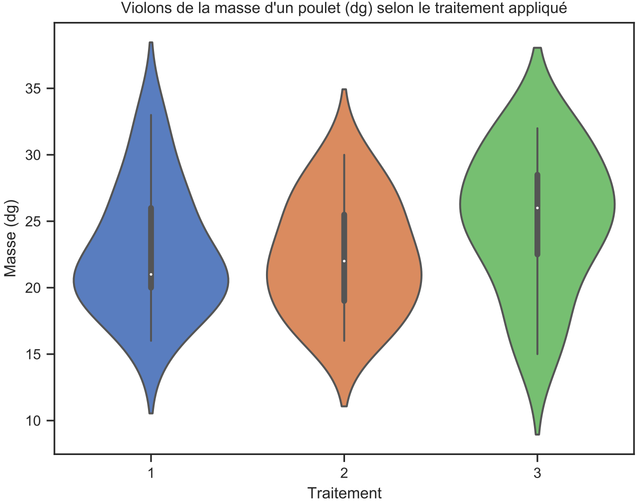
Expérience: croissance des poussins

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

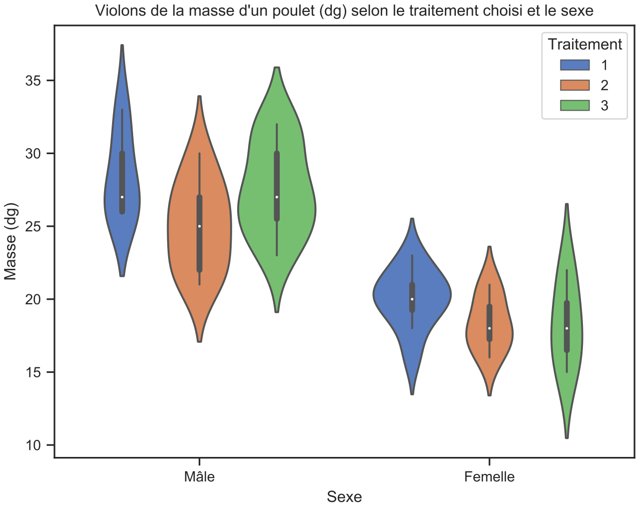
Visualisation brute



Violons selon le type de traitement

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

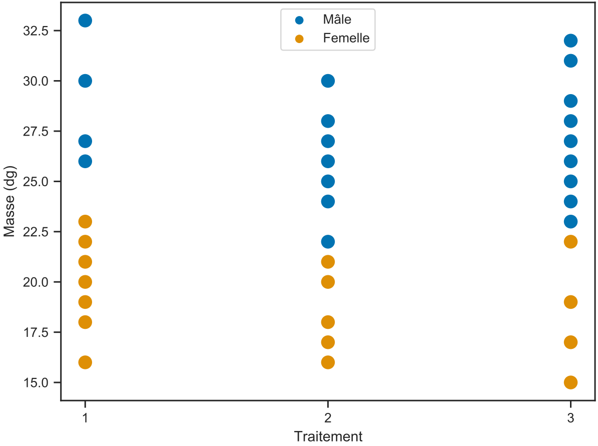
Visualisation raffinée



Violons selon le type de traitement et le sexe

<u>Conclusion</u> : c'est le traitement 1 qui a le plus d'impact

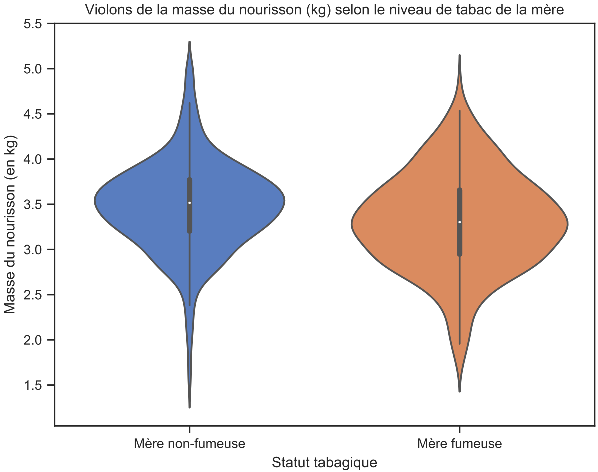
Explication



Répartition des poussins par sexe et par traitement

<u>Conclusion</u> : 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 nourisson selon le statut tabagique de la mère

<u>Conclusion</u>: les bébés de mères fumeuses ont une masse plus petite

Biographie du jour : John Tukey[12]

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

(12) 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.