


## ■ STATISTIQUE DESCRIPTIVE

 Caractériser une population en déterminant un certain nombre de grandeurs qui la décrivent, de manière purement **descriptive**



On ne tire pas de conclusions à partir des données

La statistique descriptive peut aider à **formuler des hypothèses** :

- Telle variable semble suivre une distribution uniforme sur un intervalle
- Telle variable semble dépendre de telle autre
- Telle variable semble prendre une valeur plus élevée dans un segment de la population que dans un autre



# **STATISTIQUE DESCRIPTIVE UNIDIMENSIONNELLE**



Mettre en évidence les principales caractéristiques d'une unique variable statistique  $x$  observée sur  $n$  individus via la série statistique  $(x_1, x_2, \dots, x_n)$

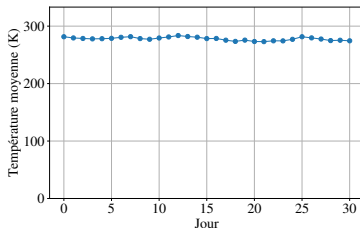
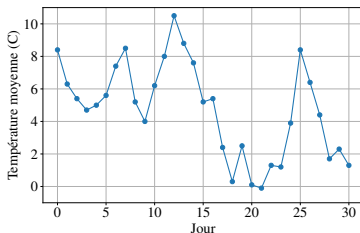
- Avec quelle **fréquence** nos données prennent-t-elles une valeur dans une plage donnée ?
- Observe-t-on une **tendance centrale** pour nos données ? Notions de **moyenne**, **médiane**, etc.
- A quel point les valeurs prises par nos données s'écartent-elles de cette tendance centrale ? Notions d'**écart-type**, de **variance**, de **quantile**.

**Rappel** une série statistique est un ensemble  $\{x_1, x_2, \dots, x_n\}$   
d'observations d'une variable aléatoire  $X$

Date	01/01	02/01	03/01	...	30/01	31/01
$T_{min}$ (°C)	7.6	5.6	4.1	...	0.5	-1

*Températures minimales journalières relevées à la station du Parc Montsouris (Paris)  
au cours du mois de janvier 2019*

# ■ Représentation graphique



*Températures moyennes journalières relevées à la station du Parc Montsouris (Paris) au cours du mois de Janvier 2019 en degrés Celsius (gauche) et en Kelvin (droite)*



Attention au choix des axes en ordonnée

# ■ Table de fréquences



Tranche d'âge (ans)	0 – 19	20 – 39	40 – 59	> 60
Effectif	14	36	62	86
Fréquence	7%	18%	31%	43%

*Appartenance de membres d'une population à une classe d'âge donnée.*

- variable **quantitative** : fréquence d'appartenance à un intervalle donné
- variable **qualitative** : fréquence d'apparition d'une valeur donnée

## ■ Diagramme en bâtons

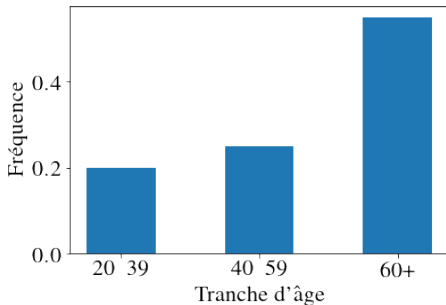


Figure – Diagramme en bâtons de la fréquence des tranches d'âges dans les données de remboursement.



## Règle de Sturges

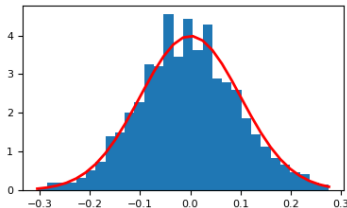
Découpage des valeurs observées en  $k = \lfloor 1 + \log_2(n) \rfloor$  intervalles de même taille  $\frac{\max(x_i) - \min(x_i)}{k}$ .



Présuppose que la variable observée suit une distribution gaussienne.

**Bonne pratique** si les valeurs s'étalent sur plusieurs ordres de grandeur, appliquer une transformation logarithmique.

- L'utilisation de fréquences permet la comparaison de populations de tailles distinctes
- La distribution des fréquences d'une série statistique de la v.a.  $X$  peut s'interpréter comme une **approximation** de la distribution de la loi de probabilité de  $X$

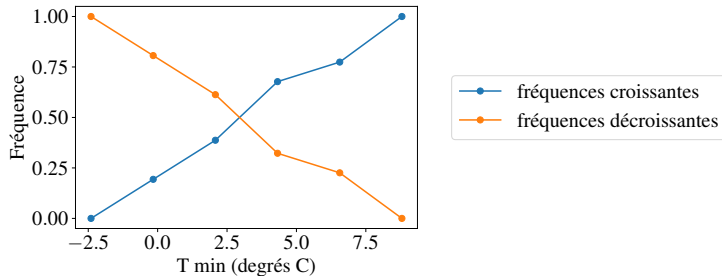


*Distribution des fréquences obtenues pour  $N = 1000$  réalisation d'une loi gaussienne de moyenne  $\mu = 0$  et d'écart-type  $\sigma = 0.1$ , et loi de distribution de cette même loi*

T min (°C)	< -0,16	< 2,08	< 4,32	< 6,56	< 8,80
Fréquence	0,19	0,38	0,67	0,77	1,0
T min (°C)	> -2,40	> -0,16	> 2,08	> 4,32	> 6,56
Fréquence	1,0	0,81	0,62	0,33	0,23

*Table des fréquences cumulées pour les températures minimales relevées à la station du Parc Montsouris (Paris) au cours du mois de janvier 2019*

# ■ Courbe des fréquences cumulées



*Courbes des fréquences cumulées pour les températures minimales relevées à la station du Parc Montsouris (Paris) au cours du mois de Janvier 2019.*

## ■ moyenne arithmétique

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



sensible à la présence de valeurs aberrantes

■ **médiane** valeur qui correspond à une fréquence cumulée de 50%,

■ **mode** valeur la plus fréquente dans la série statistique  
pour une variable continue, on utilise la **classe modale**

## ■ Variance de la série statistique



estimation **biaisée**

$$\text{var}(x_1, x_2, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

## ■ Variance d'échantillonnage

$$\text{var}^*(x_1, x_2, \dots, x_n) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

## ■ Ecart-type

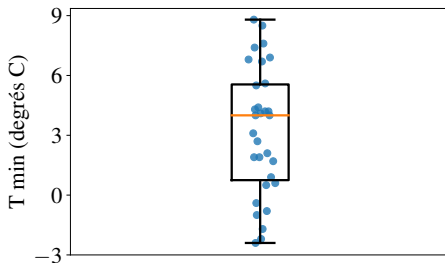
$$\sigma = \sqrt{\text{var}(x_1, x_2, \dots, x_n)}$$

- plus simple à interpréter que la variance

- Les  $q$ -quantiles divisent les valeurs prises par la variable en  $q$  intervalles de mêmes fréquences : le  $p$ -ème  $q$ -quantile de  $(x_1, x_2, \dots, x_n)$  est défini comme la valeur  $Q_p^q$  telle que

$$\text{Freq}(x \leq Q_p^q) = \frac{p}{q}.$$

- Lorsque  $q = 4$ , on parle de **quartiles**  
Lorsque  $q = 10$ , on parle de **déciles**



*Boîte à moustaches des températures minimales relevées à la station du Parc Montsouris (Paris)  
au cours du mois de Janvier 2019..*



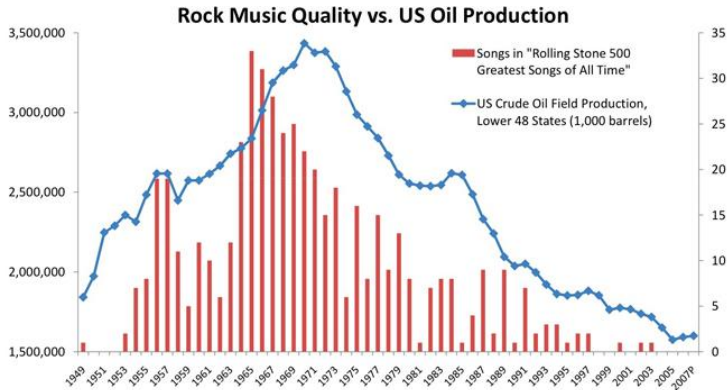


# **STATISTIQUE DESCRIPTIVE BIDIMENSIONNELLE**

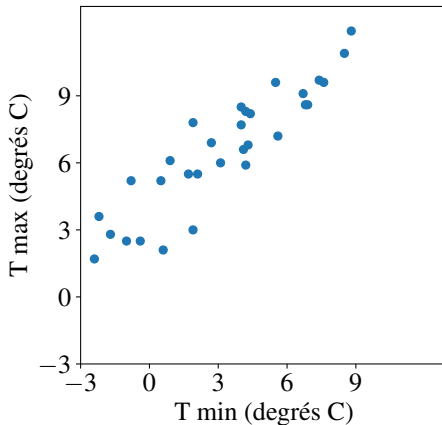
🎯 Mettre en évidence une éventuelle **liaison**, i.e une **variabilité simultanée**, entre deux variables statistiques  $x$  et  $y$  via les séries statistiques  $(x_1, x_2, \dots, x_n)$  et  $(y_1, y_2, \dots, y_n)$

- Une variable peut-elle dépendre d'une autre ?
- Une variable peut-elle permettre d'en prédire une autre ?
- Une variable peut-elle en suppléer une autre dans l'analyse d'un problème ?

# ■ Liaison $\neq$ causalité



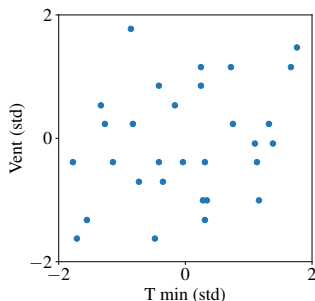
## ■ Nuage de points



*Températures maximales vs minimales. relevées à la station du Parc Montsouris (Paris) au cours du mois de Janvier 2019*

**Bonne pratique** lorsque  $x$  et  $y$  sont des grandeurs physiques distinctes, on préfère centrer et réduire les variables au préalable

$$x_i \leftarrow \frac{x_i - \bar{x}}{\sigma_x} \quad \text{avec } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{et} \quad \sigma_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$



*Vent vs températures minimales. Données mesurées à la station du Parc Montsouris (Paris) au cours du mois de Janvier 2019*

## ■ Covariance

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

## ■ Coefficient de corrélation de Pearson

$$r(x, y) = \frac{1}{n} \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$$

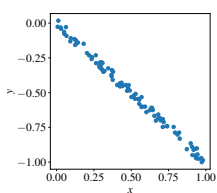


La covariance et le coefficient de corrélation de Pearson mesurent des liaisons **linéaires** entre deux variables

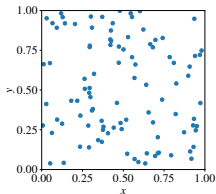
# ■ Coefficient de corrélation de Pearson



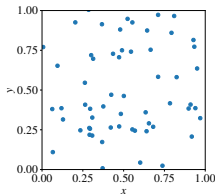
Une corrélation de Pearson proche de 1 ou de -1 indique une relation linéaire; une corrélation de Pearson proche de 0 indique une absence de corrélation.



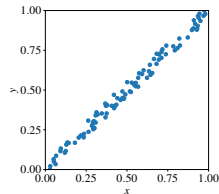
(a)  $r = -1$ .



(b)  $r = 0,03$ .



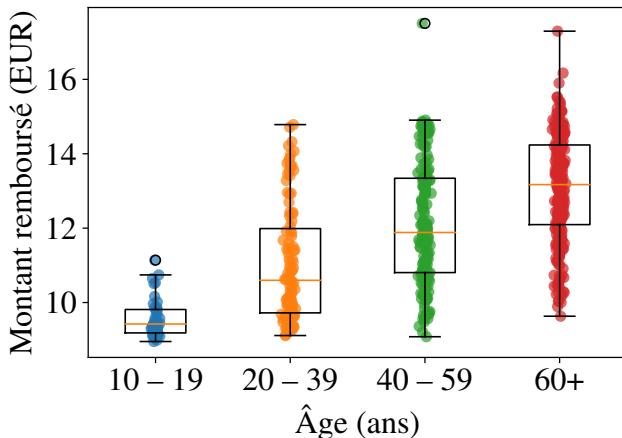
(c)  $r = 0,53$ .



(d)  $r = 1$ .

*Nuages de points entre deux variables simulées et leur corrélation de Pearson.*

# ■ Indicateurs de liaison quali-quant



*Montants remboursés par acte, par tranche d'âge, pour les données de remboursement.*



## ■ Variance expliquée



### ■ variance expliquée par $x$ de $y$

$$\sigma_E^2 = \frac{1}{n} \sum_{k=1}^K n_k (\bar{y}_k - \bar{y})^2,$$

où  $\bar{y}_k$  est la moyenne de  $y$  dans la sous-population  $k$  et  $\bar{y}$  la moyenne de  $y$  dans la population totale

### ■ variance résiduelle

$$\sigma_R^2 = \frac{1}{n} \sum_{k=1}^K n_k \sigma_k^2$$

où  $n_k$  est le nombre d'individus dans la sous-population  $k$  et  $\sigma_k^2$  est la variance de  $y$  dans cette sous-population

**Remarque** On peut montrer que  $\sigma_y^2 = \sigma_R^2 + \sigma_E^2$