Statistiques et probabilités
Cours n°4

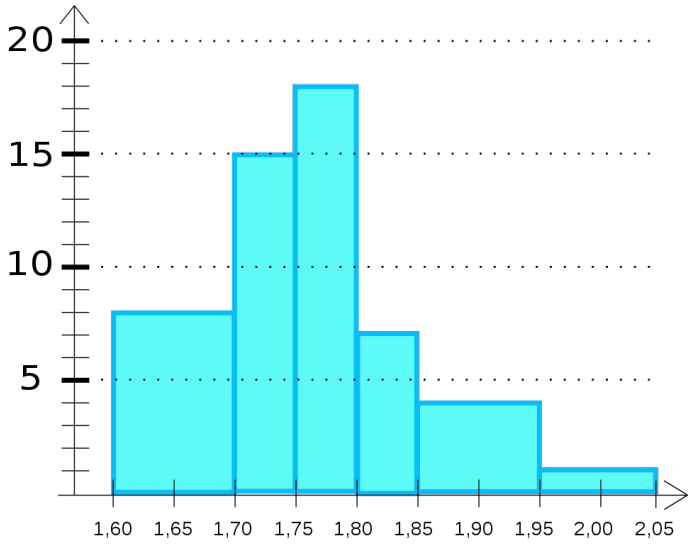
Guillaume Postic

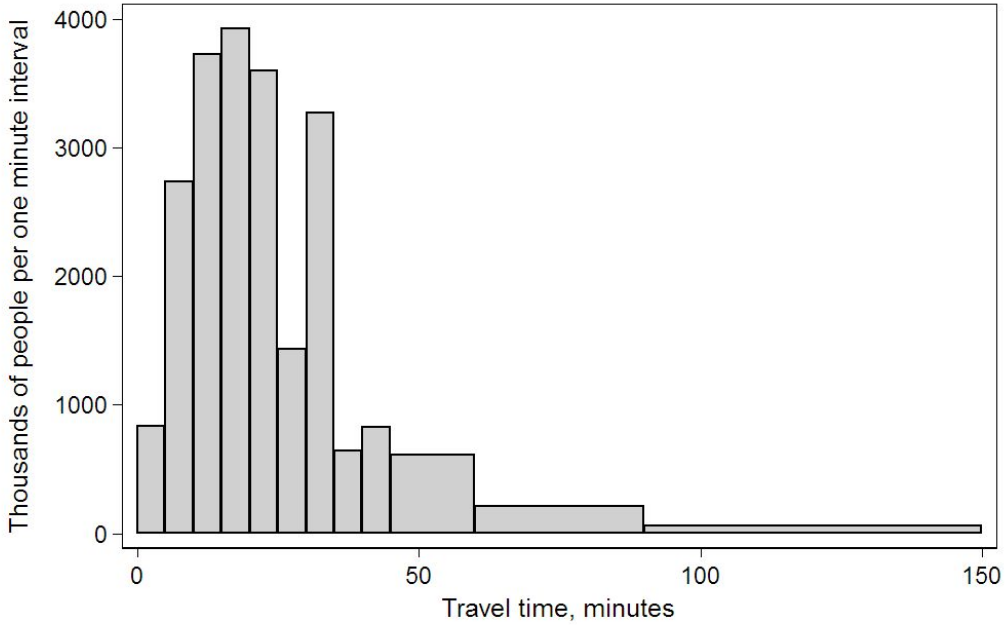
Université Paris-Saclay, Univ. Evry
Département informatique

Master 1 MIAGE - 2022/2023

Estimation de la densité : histogramme

Représentation graphique approximant la distribution d'une variable aléatoire par groupement des données en classes représentées par des colonnes.

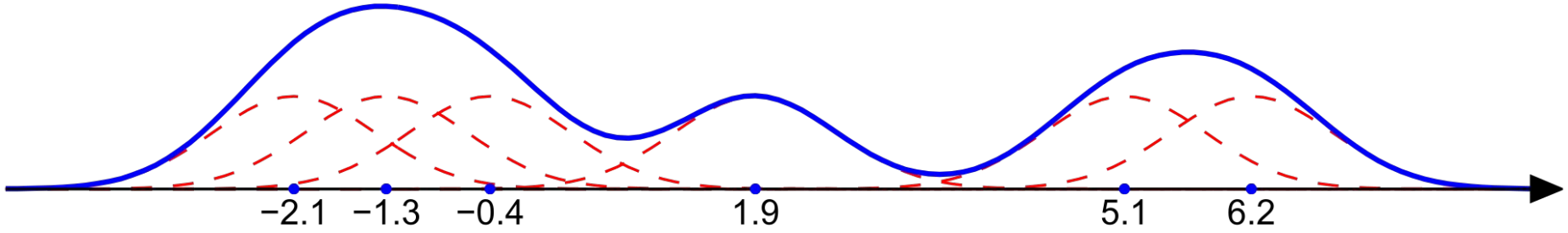


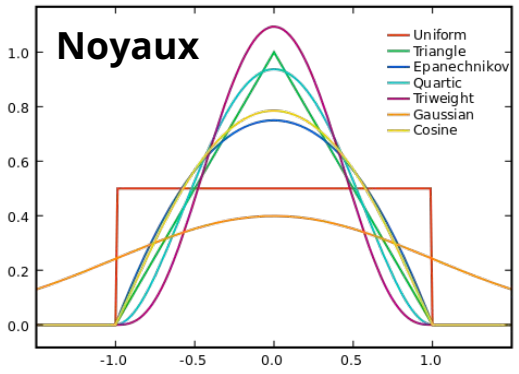


Estimation de la densité : par noyau

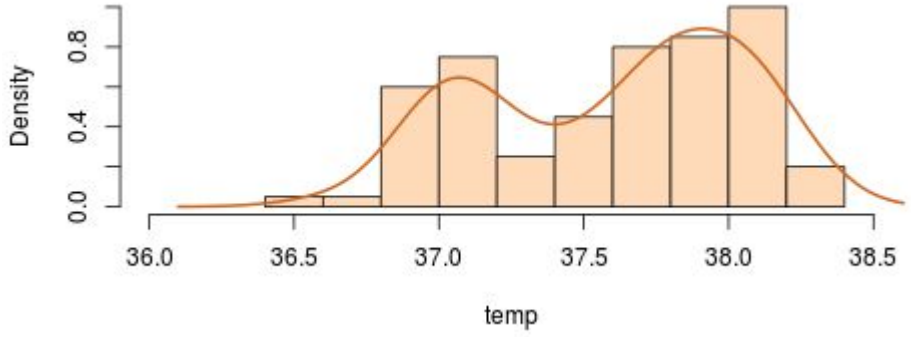
Généralisation de l'estimation par histogramme.

- Différents types de noyaux peuvent être utilisés (influe peu).
- Différentes méthodes pour choisir le paramètre de lissage (« largeur du noyau »).



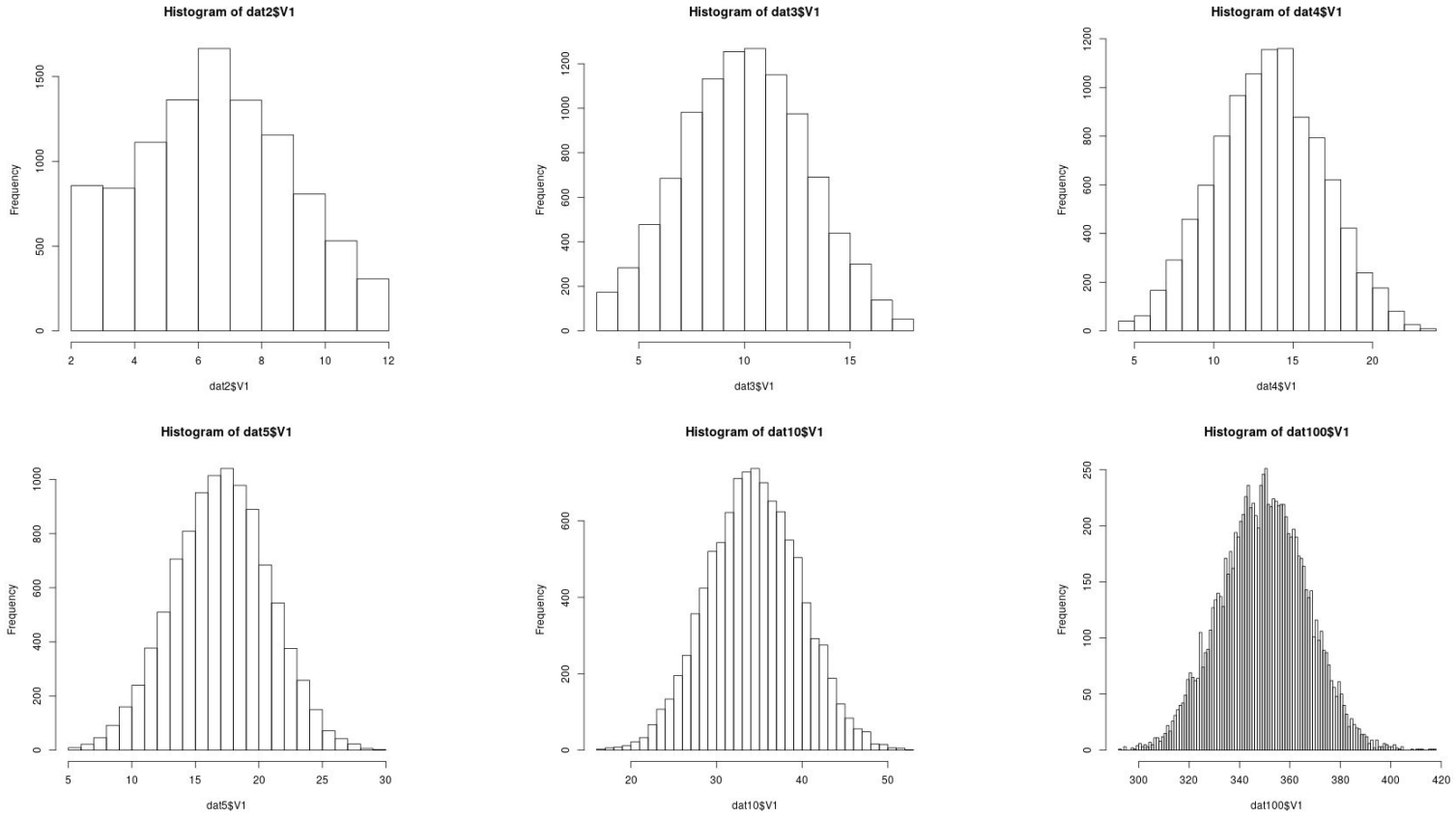


Exemple



Théorème de la limite centrale (1)

On lance 10 000 fois *n* dés équilibrés à 6 faces (loi uniforme) ; à chaque lancer, on calcule **la somme *S* des *n* dés**. Ci-dessous les représentations par histogrammes des distributions de *S*, pour des valeurs de *n* de 2, 3, 4, 5, 10 ou 100 :



Théorème de la limite centrale (2)

Pour n variables aléatoires, leur somme S est une variable aléatoire qui suit approximativement une loi normale de paramètres $\mu = E(S)$ et $\sigma^2 = V(S)$.

Conditions sur les variables aléatoires :

- **Indépendantes**
- Identiquement distribuées (c.-à-d. de même loi)
- n suffisamment grand, e.g. $n > 20$ ou 30, selon les auteurs
  - Loi des grands nombres

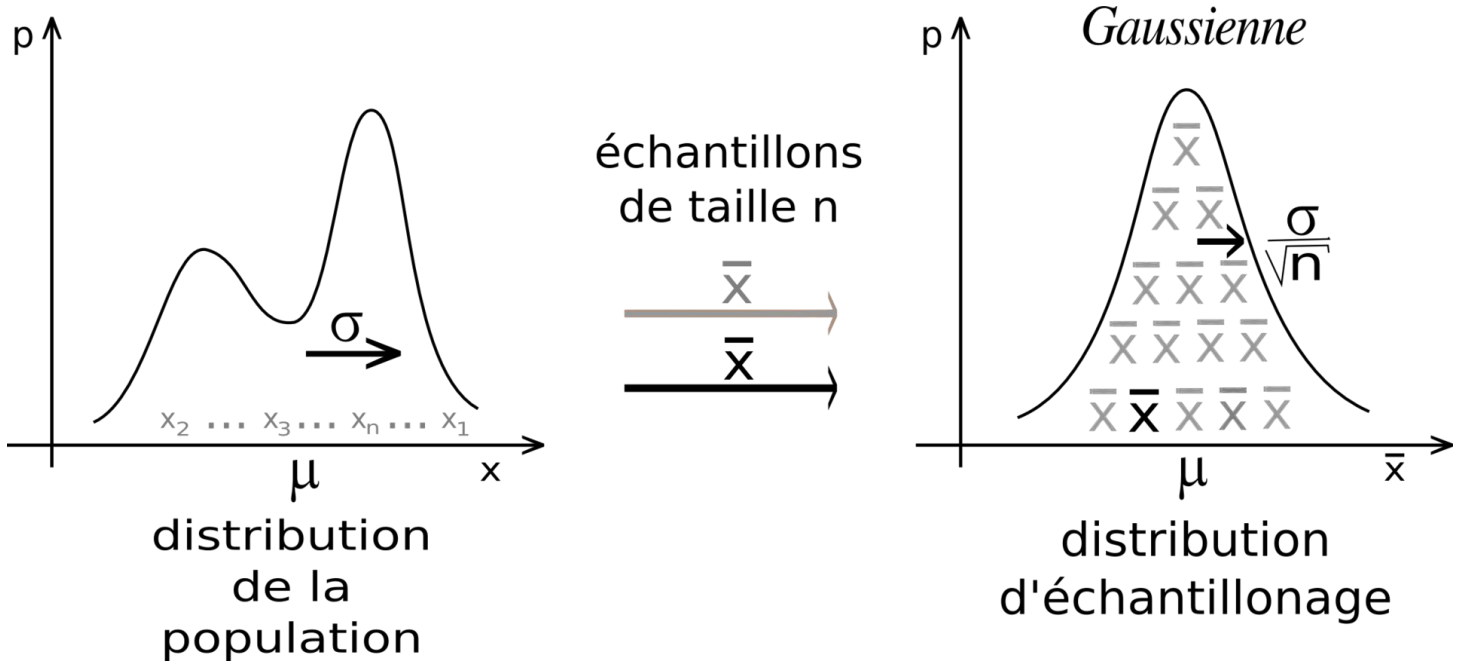
TLC avec la moyenne $M = S/n$ de n variables aléatoires :

M suit une loi normale de paramètres $E(M) = E(S)/n$ et $V(M) = V(S)/n^2$

Note : les conditions de Liapounov et Lindeberg permettent de supprimer l'hypothèse selon laquelle les variables aléatoires sont de même loi.

Théorème de la limite centrale (3)

Le théorème de la limite centrale **s'applique donc à la distribution des moyennes d'échantillons** de tailles égales (ces moyennes étant des sommes, toutes divisées par la même valeur).



Estimateur de l'écart-type : $S \sim N(n\mu, n\sigma^2)$

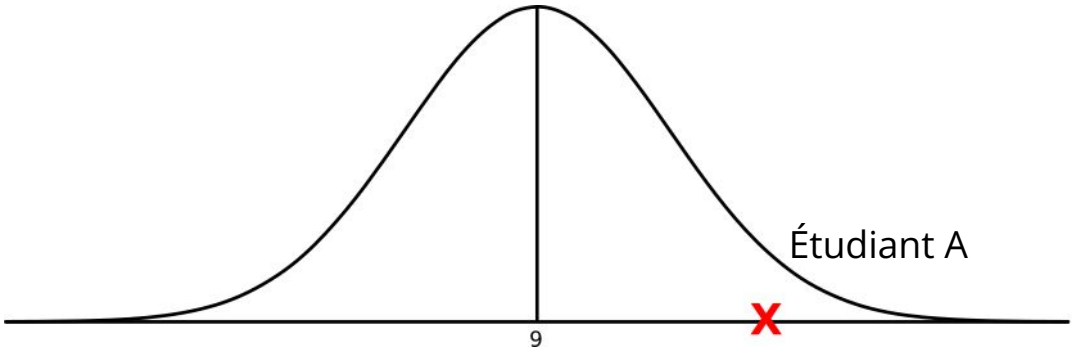
Variable centrée réduite (1)

Question) Deux étudiants veulent savoir qui est le meilleur, en comparant leurs notes obtenues à l'UE de statistiques. Ils appartiennent à deux groupes de TD différents, chacun noté par un enseignant différent.

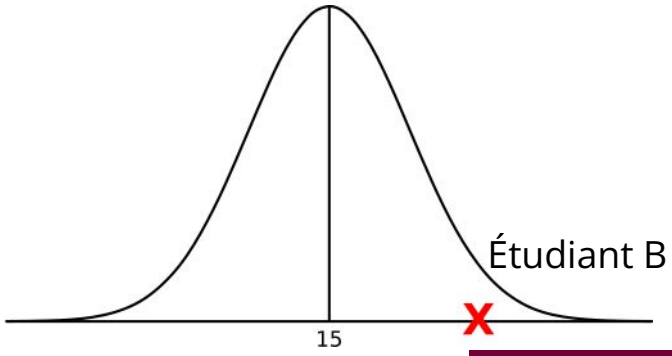
L'étudiant A a eu 15/20, dans le groupe 1, où la moyenne est de 9 et l'écart-type est de 5.

L'étudiant B a eu 19/20, dans le groupe 2, où la moyenne est de 15 et l'écart-type est de 3.

Illustration du problème avec des notes distribuées selon une loi normale et une représentation de la densité estimée :



Groupe 1



Groupe 2

Variable centrée réduite (2)

Pour comparer les notes des deux étudiants, il faut calculer leurs notes

- centrées, par soustraction avec la moyenne μ
- et réduites, en divisant par l'écart-type σ

Ainsi, pour toute variable centrée réduite : $\mu = 0$ et $\sigma = 1$

La variable centrée réduite est également appelée variable standardisée, **Z-score**, valeur Z, ou (improprement) variable normalisée.

$$z = \frac{x - \mu}{\sigma}$$

Ainsi, si $X \sim N(\mu, \sigma)$, alors $Z \sim N(0, 1)$

Variable centrée réduite (3)

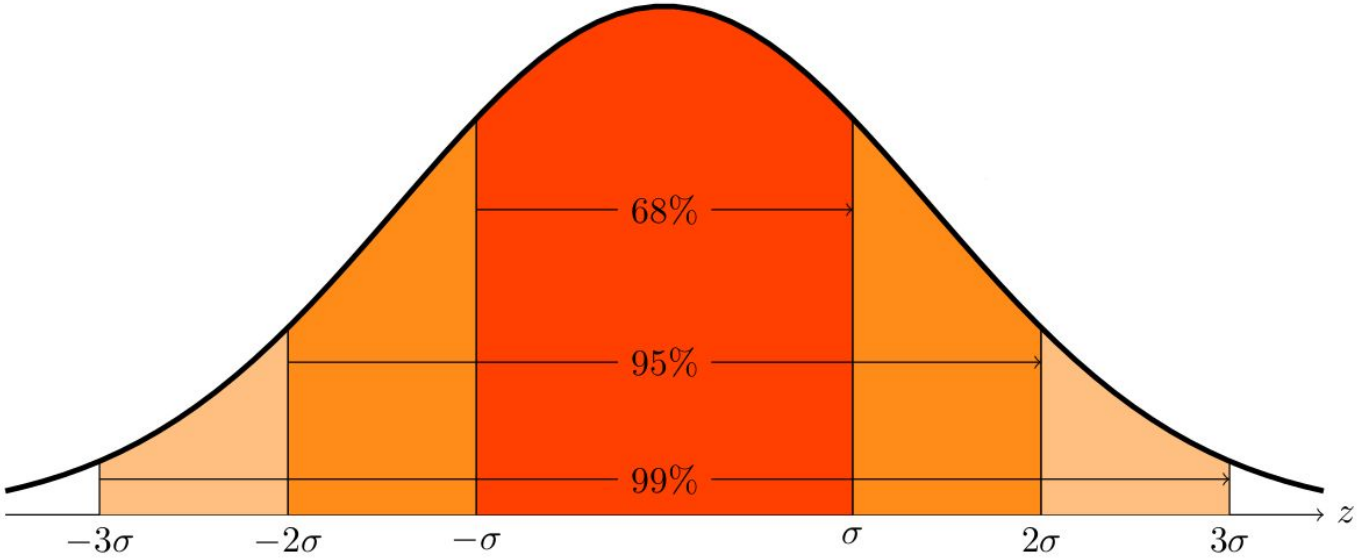
- Toute valeur Z suit une loi de probabilité de paramètres $\mu = 0$ et $\sigma = 1$
- Cette loi n'est pas nécessairement une loi normale
- Quelle que soit cette loi, **un Z-score égal à *α* signifie que la donnée *x* s'éloigne de la moyenne de *α* écarts-types** (donc $Z < 0$ si $x < \mu$)

Réponse : Les notes centrées réduites suivent donc toutes la même distribution et sont directement comparables.

L'étudiant A a une note standardisée de 6/5, inférieure à celle de l'étudiant B, 4/3.

Variable centrée réduite (4)

Lecture de la loi normale centrée réduite (ou standardisée) ou **fonction de distribution Z** :



1. $P(-1 < Z < 1)$ is

(a) 0.025 (b) 0.16 (c) 0.68 (d) 0.84 (e) 0.95

2. $P(Z > 2)$

(a) 0.025 (b) 0.16 (c) 0.68 (d) 0.84 (e) 0.95

Variable centrée réduite (5)

Soit $Z = (x-\mu)/\sigma > 2$

- si $X \sim N(\mu, \sigma)$, $p(X > x) = 2{,}5\%$ (environ)
- si $X \sim$ loi non-connue (de moyenne μ et variance σ^2)
  - **Inégalité de Bienaymé-Tchebychev** : donne la probabilité « dans le pire des cas »
    - $p(X > x) < 1/Z^2 = 25\%$
    - $p(X < x) > 1/Z^2 = 75\%$

Variable aléatoire qualitative

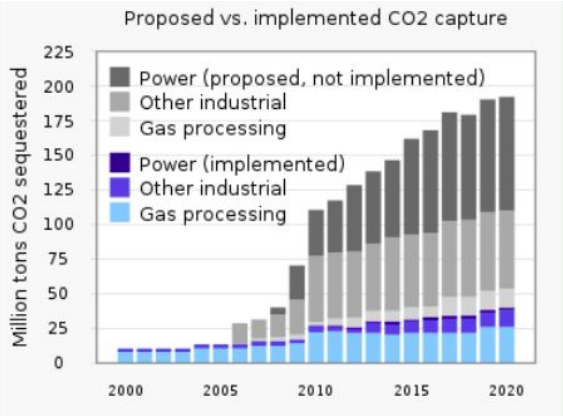
- Variable quantitative
  - Quantitative continue
    - Mesure de longueur, de temps…
  - Quantitative discrète
    - Résultat d'un dé, longueur discrétisée…
- **Variable qualitative**
  - **Qualitative catégorielle (ou nominale)**
    - **Couleurs (bleu, vert, rouge)…**
  - Qualitative ordinale
    - Lettres (A, B, C), mois de l'année…
    - Valeurs peuvent être traitées comme quantitatives discrètes, si on les positionne sur une échelle

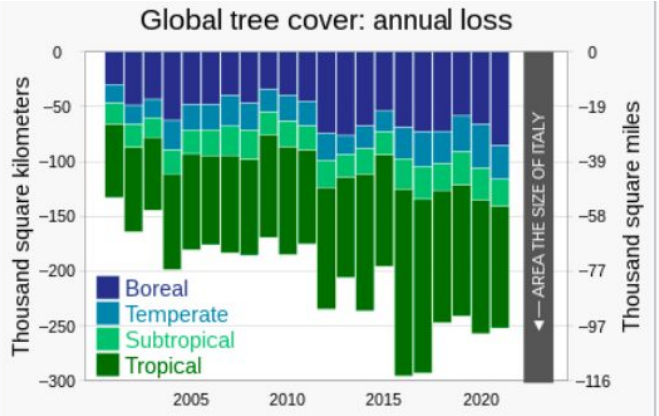
Diagramme en bâtons

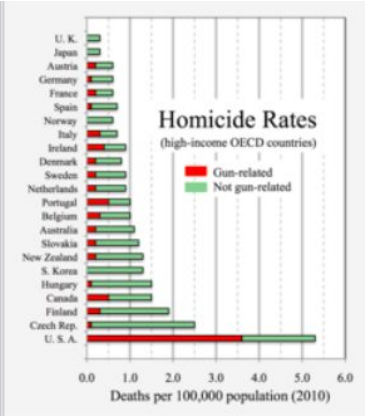
Bar chart, *bar graph*, *bar plot*…

- Graphique qui présente des **variables catégorielles** avec des barres rectangulaires avec des hauteurs ou des longueurs proportionnelles aux valeurs qu'elles représentent.
- Largeur des barres arbitraire et identique pour toutes

⚠ **Ne pas confondre avec histogramme**







Variable catégorielle nominale

- **Indicateurs de tendance centrale**
  - Les données n'étant pas ordonnées, la moyenne et la médiane ne sont pas calculables.
  - Néanmoins, le mode peut être calculé : catégorie la plus représentée.

- **Indicateurs de dispersion**
  - La variance et l'écart-type ne sont pas calculables.
  - Différentes mesures de la dispersion existent pour les variables nominales, comme l'entropie de Shannon.