GC – Probabilités et Statistique

http://moodle.epfl.ch/course/view.php?id=14271

Cours 1

- Notions de base
- Représentations graphiques
- Résumés numériques
- Le matériel d'aujourd'hui **ne sera pas examiné explicitement**, il est fourni à titre d'information (seulement)

Population

- La **population** est l'ensemble des éléments qui forment le champ d'analyse d'une étude particulière
  - Dans une étude sur les thérapies du cancer du sein, la population pourrait être l'ensemble des personnes souffrant du cancer du sein
  - Dans une étude sur l'effet de la lumière sur la plante *Arabidopsis thaliana*, la population pourrait être l'ensemble des plantes *Arabidopsis thaliana*
  - (Formuler vos propres exemples)
- Ne s'applique pas seulement aux êtres humains
- La population est constituée des **individus** ou **unités statistiques**

Variables (I)

- En statistique, *les caractéristiques qui varient* parmi les individus de la population sont appelées **variables**
- Les **modalités** d'une variable consistent en l'ensemble des *valeurs possibles*
- Variables de différentes sortes :
  - **Variables qualitatives :** les modalités sont des mots ou 'etiquettes' que l'on appelle des *catégories*
    Exemples : couleur des yeux ('bleu', 'brun', 'vert') ; le programme de télé préféré
  - **Variables quantitatives :** les modalités sont des valeurs numériques
    Exemples : âge, nombre de membres d'une famille, le poids

Variables (II)

- **Variables qualitatives** peuvent être classées comme :
  - Échelle *nominale* – les catégories ne sont pas naturellement ordonnées (p.ex. couleur des yeux, sexe)
    • *Même si* les modalités sont exprimés comme des codes numériques
    (p.ex. sexe = '0' pour 'mâle', = '1' pour 'femelle')
  - Échelle *ordinale* – les catégories peuvent être ordonnées (p.ex. 'toujours', 'de temps en temps', 'jamais')
- **Variables quantitatives** sont distinguées :
  - Variables *discrètes* – les valeurs possibles peuvent être énumérées sous la forme d'une *liste de chiffres* (le plus souvent : les entiers nonnegatifs 0, 1, 2, ...)
  - Variables *continues* – l'ensemble des valeurs possibles est constitué par *un (ou plusieurs) intervalle(s)*

Observations et données

- Les résultats observés d'une ou de plusieurs variables pour quelques individus d'une population constituent les **observations** ; p.ex. :
  - le sexe, le poids, la taille et le périmètre crânien des nouveau-nés dans un hôpital particulier
  - la survie, la classification histologique et le stade TNM pour des tumeurs du sein
- Un jeu de données générique :

	Variables					
Individus	X_1	X_2	$\ldots$	X_j	$\ldots$	X_p
i_1	x_{11}	x_{12}	$\ldots$	x_{1j}	$\ldots$	x_{1p}
i_2	x_{21}	x_{22}	$\ldots$	x_{2j}	$\ldots$	x_{2p}
$\ldots$						
i_i	x_{i1}	x_{i2}	$\ldots$	x_{ij}	$\ldots$	x_{ip}
$\ldots$						
i_n	x_{n1}	x_{n2}	$\ldots$	x_{nj}	$\ldots$	x_{np}

Analyse exploratoire des données

- Autrement dit 'la statistique descriptive', on *explore* les données avant l'analyse formelle
- Les données sont examinées pour évaluer leur qualité, ensuite 'netoyées' s'il le faut, ainsi que visualisées afin de fournir une impression de l'ensemble des résultats
- Nous allons examiner *deux types* de résumés :
  - résumés graphiques
  - résumés numériques
- Nécessaire à utiliser un logiciel statistique (tel que R)

Représentations graphiques des données : histogramme

- **L'histogramme** est un type (spécial) de *diagramme en barres*
- Il vous permet de visualiser *la répartition des valeurs* pour une variable numérique
- Lorsque dessiné avec une échelle de **densité** :
  - **la surface** (PAS la hauteur) de chaque barre est *la proportion* des observations dans l'intervalle
  - *la hauteur* représente *la densité*
- **La surface (l'aire sous la courbe) totale est 100% (ou 1)**

Quelques formes d'histgramme

étirée à gauche

étirée à droite

symétrique

Résumés numériques

- **Variables qualitatives :** tableau des fréquences
- **Variables quantitatives :**
  - mesures de *tendance centrale*
  - mesures de *dispersion*

Mesures de tendance centrale : la moyenne

- **La moyenne (arithmétique)** $\overline{x}$ est égale à la somme des valeurs observées, divisée par leur nombre total n :

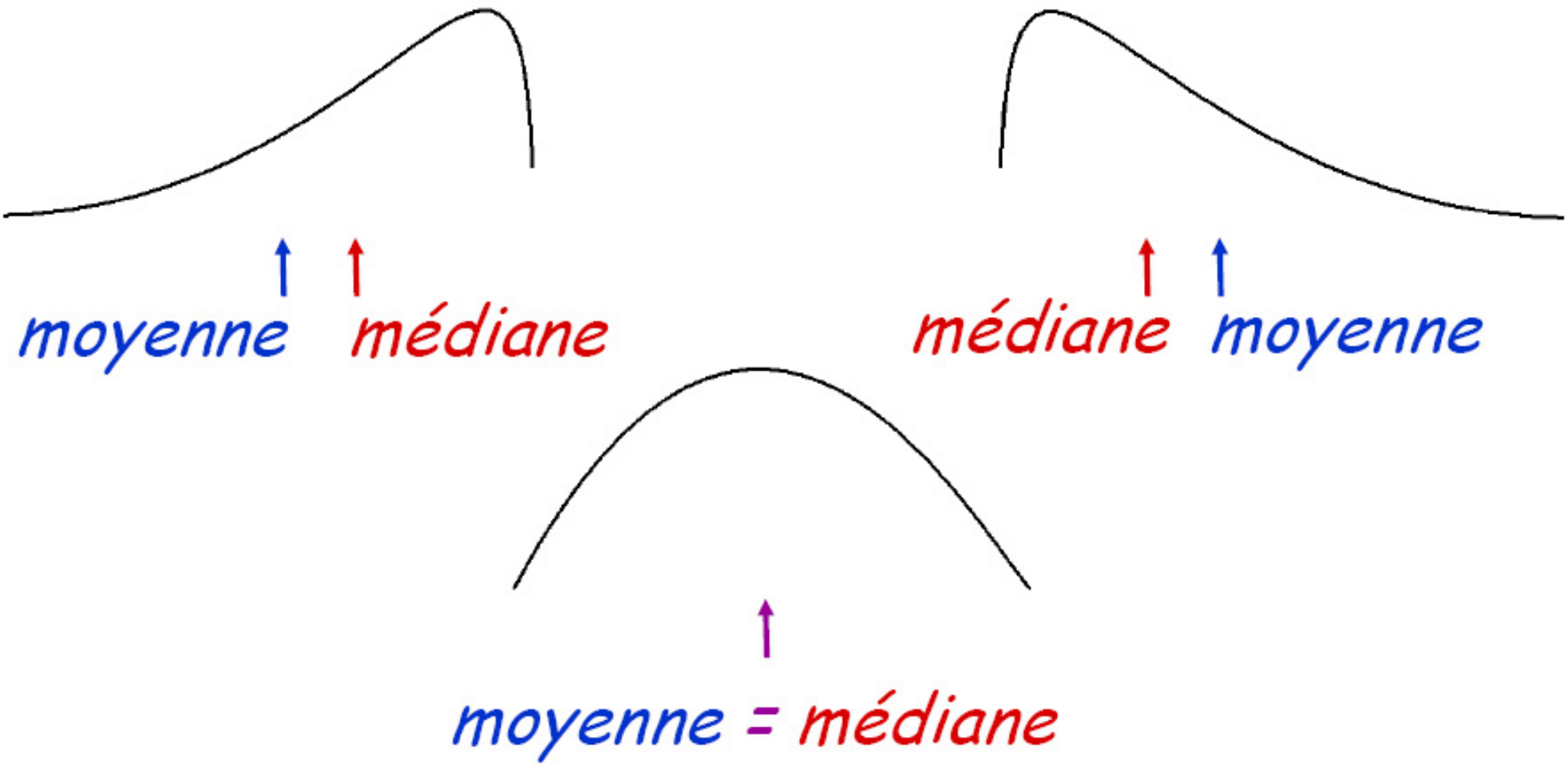
$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- La moyenne est une mesure appropriée pour une distribution (histogramme) qui est assez *symétrique*
- Puisque toutes les valeurs contribuent *également*, la moyenne est *sensible* à la présence d'observations abérrantes
- La moyenne est le point où l'histogramme s'équilibre

Mesures de tendance centrale : la médiane

- **La médiane** ($med(x)$) est le point qui partage la distribution d'une série (ordonnée) d'observations en *deux parties égales*
- La $((n+1)/2)^{\text{ème}}$ plus grande valeur parmi $x_1, \ldots, x_n$ définit la médiane
- Si le nombre d'observations n est *pair*, la médiane peut prendre n'importe quelle valeur située entre la $\left(\frac{n}{2}\right)^{\text{ème}}$ observation et la $\left(\frac{n+2}{2}\right)^{\text{ème}}$ observation – normalement on prend la valeur moyenne de ces deux comme valeur de la médiane
- La médiane *n'est pas sensible* à la présence de valeurs aberrantes, car elle 'ne tient pas compte' de la quasi-totalité des valeurs
- La médiane est donc généralement un résumé plus approprié pour les distributions *asymétriques*

Position relative de la moyenne et de la médiane



PAUSE

Mesures de dispersion : Variance et l'écart-type

- **La variance** s^2 d'une variable est la moyenne* des carrés des écarts de la moyenne :

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

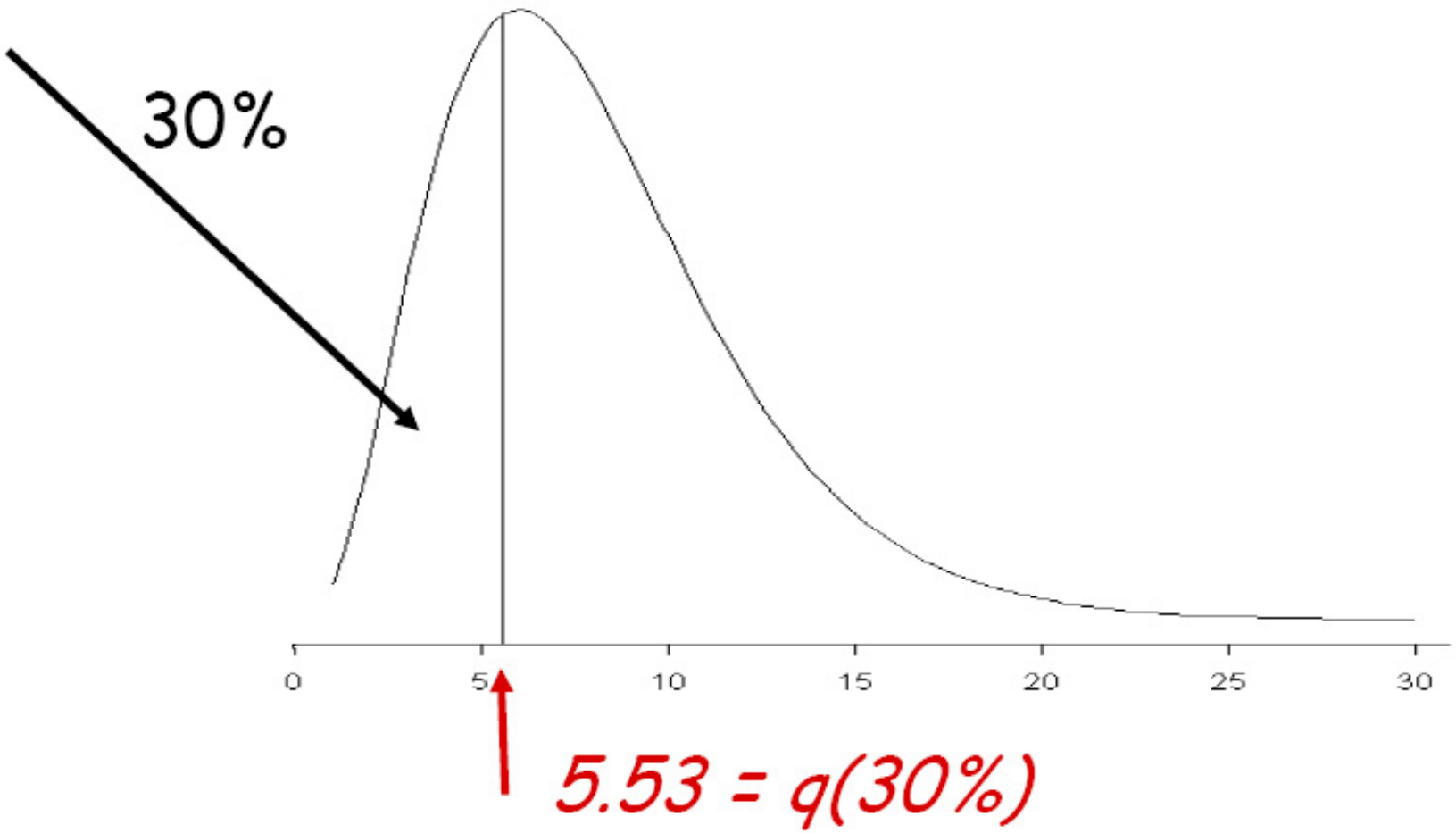
- **L'écart-type** s d'une variable est *la racine carrée de la variance* :

$$s = \sqrt{s^2}$$

- *Pour des raisons 'techniques', au lieu de diviser par le nombre de valeurs n, en générale on divise par $n-1$
- L'écart-type s est une mesure de dispersion appropriée lorsqu'on utilise *la moyenne* pour la tendence centrale

Quantiles

- Le **quantile (empirique)** $\hat{q}(p)$ est la valeur telle qu'*une proportion p des observations soit au plus $\hat{q}(p)$*



Mesures de dispersion : IQ

- Les quantiles $\hat{q}(25\%)$, la médiane, et $\hat{q}(75\%)$ divisent un ensemble d'observations en *quatre parties égales* (dont chacune contient 25% des observations)
- Ces quantiles spéciaux sont appelés les **quartiles**
- La distance entre les quartiles $\hat{q}(25\%) = Q_1$ et $\hat{q}(75\%) = Q_3$ est **l'intervalle interquartile (IQ)** :

$$IQ = Q_3 - Q_1$$

- L'IQ donne une mesure de dispersion lorsqu'on mesure le centre en utilisant la *médiane*

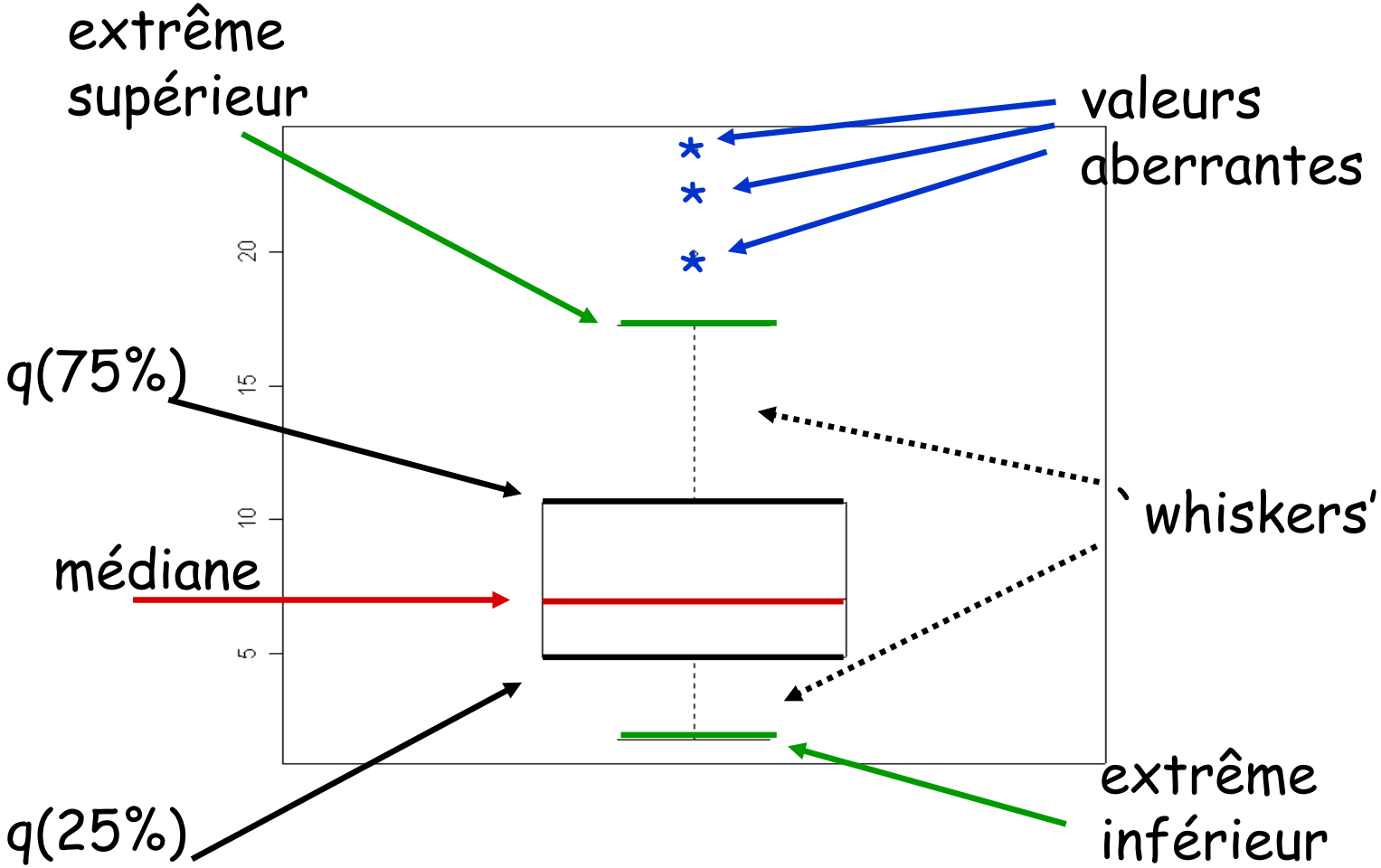
Mesures de dispersion : MAD

- La médiane des écarts de la médiane ('median absolute deviation', ou **MAD**) est obtenue comme le suit :
  1. trouver la médiane $med(x)$ des observations x_i, $i = 1, \ldots, n$
  2. calculer les écarts $|x_i - med(x)|$
  3. trouver la médiane des écarts calculés (l'étape 2 ci-dessus)
- Donc, analogue à l'écart-type
- Le *MAD* est bien *plus résistante* que l'écart-type s
- Le *MAD* donne une autre mesure de dispersion quand on utilise la médiane

Résumé à 5 valeurs et boxplot

- L'information essentielle dans une distribution est rapidment transmise au moyen du **résumé à 5 valeurs** :
  1. le minimum
  2. $\hat{q}(25\%)$ $(= Q_1)$
  3. la médiane
  4. $\hat{q}(75\%)$ $(= Q_3)$
  5. le maximum
- Le **boxplot** (la 'boîte à moustaches') représente graphiquement ces valeurs
- (**À noter** : le résumé à 5 valeurs dans PP est *différent* ; google '5-number summary')

Boxplot


extrême
supérieur
valeurs
aberrantes
q(75%)
médiane
`whiskers'
q(25%)
extrême
inférieur
5
10
15
20

Étapes pour créer un boxplot

1. Ordonner les valeurs
2. Calculer le résumé à 5 valeurs
3. Identifier des observations potentiellement aberrantes en calculant $d = 1.5 \times (Q_3 - Q_1)$ et en cherchant des valeurs

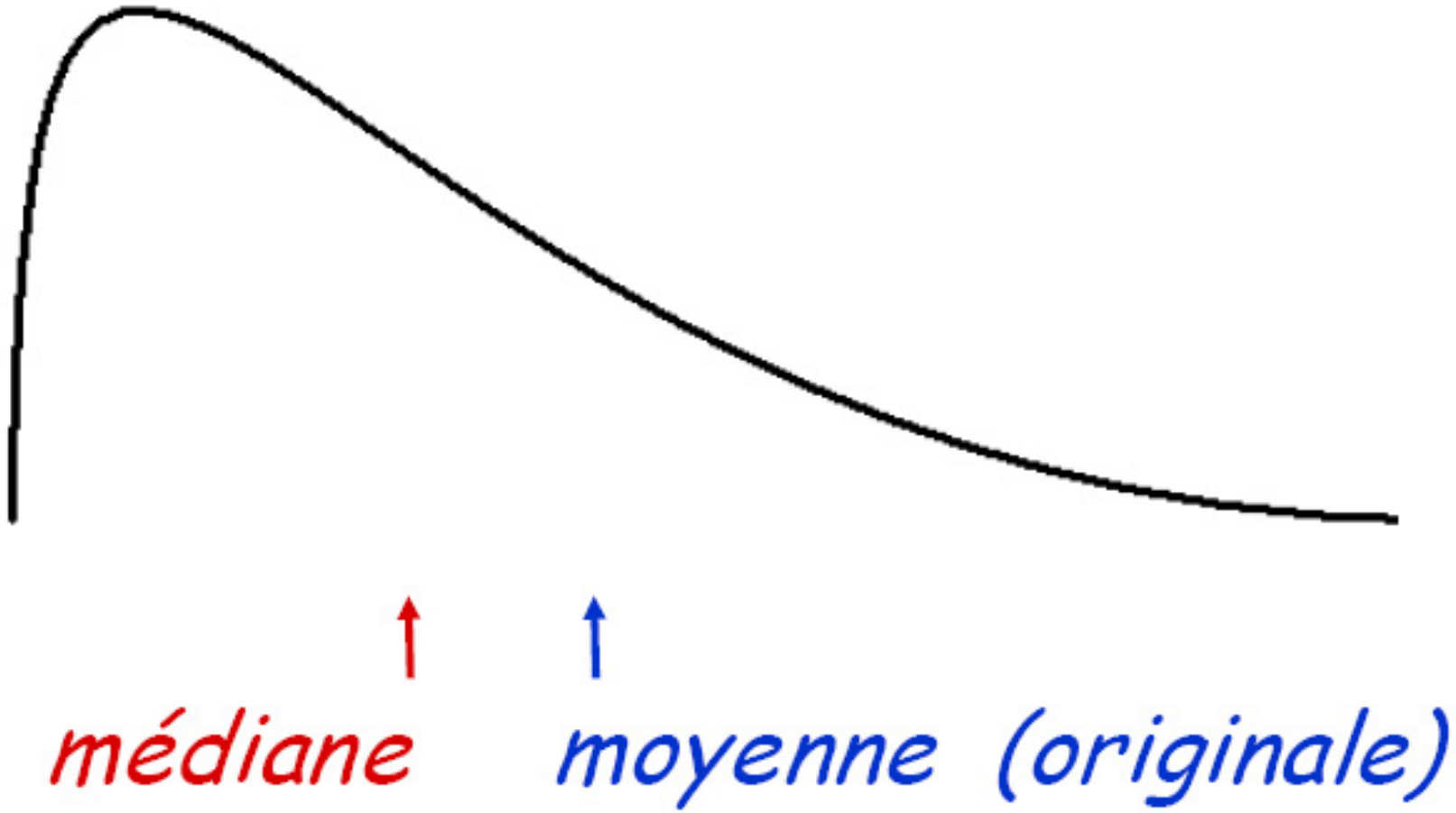
$$x_i < e_{inf} = Q_1 - d \quad \text{et} \quad x_i > e_{sup} = Q_3 + d$$

4. Esquisser le graphique :
   - faire le boîte (Q_1, Q_3)
   - mettre un trait dans la boîte au niveau de la médiane
   - ajouter des petits traits ('whiskers' ou 'moustaches') pour les extrêmes et les connecter à la boîte
   - s'il existe des observations aberrantes, les signifier en utilisant des étoiles

Résistance

- La **résistance** s'adresse à *l'insensibilité au 'mauvais comportement' des données*
- Une analyse ou un résumé est **résistant/e** si *un changement arbitraire dans n'importe quelle partie des données* **ne produit pas** *un grand changement* des résultats de l'analyse ou du résumé
- La résistance d'un résumé est *souhaitable* : on ne veut pas avoir des conclusions très fragiles
- *Exemple :* revenu 'typique' avec ou sans Mark Zuckerberg
- La médiane est bien résistante ; la moyenne n'est pas résistante

Résistance de la moyenne et de la médiane (1)



Résistance de la moyenne et de la médiane (2)