# Analyse de données

## Données unidimensionnelles

Jamal Atif
jamal.atif@dauphine.fr

Université Paris-Dauphine, Licence MIDO


2014-2015

# Analyse de variables qualitatives

Dans ce tableau, on cherche à étudier la variable « Etude ».

| Salaire | Etude | Age |
|---|---|---|
| 2600 | Bac+4 | 31 |
| 2200 | IUT | 27 |
| 5000 | Ecoles | 53 |
| ⋮ | ⋮ | ⋮ |

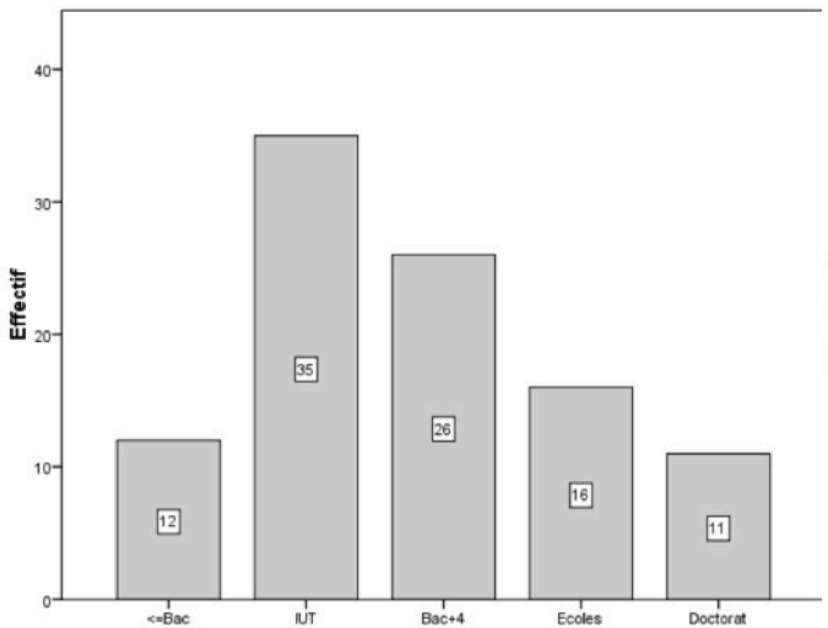
On effectue un « tri à plat »

| | Effectifs | Pourcentage | Pourcentage cumulé |
|---|---|---|---|
| Valide $\leq$ Bac | 12 | 12,0 | 12,0 |
| IUT | 35 | 35,0 | 47,0 |
| Bac+4 | 26 | 26,0 | 73,0 |
| Ecoles | 16 | 16,0 | 89,0 |
| Doctorat | 11 | 11,0 | 100,0 |
| Total | 100 | 100,0 | |

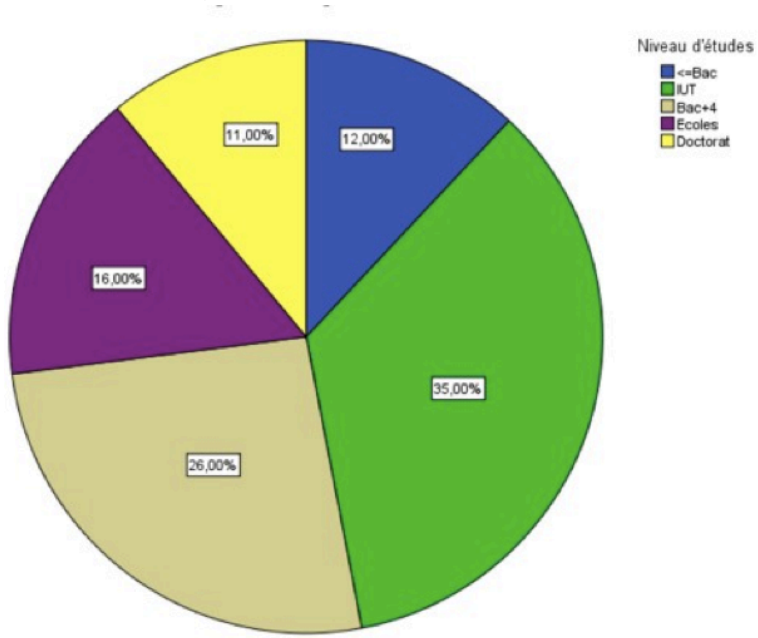
# Analyse de variables qualitatives

## Visualisation

### Diagramme en batons

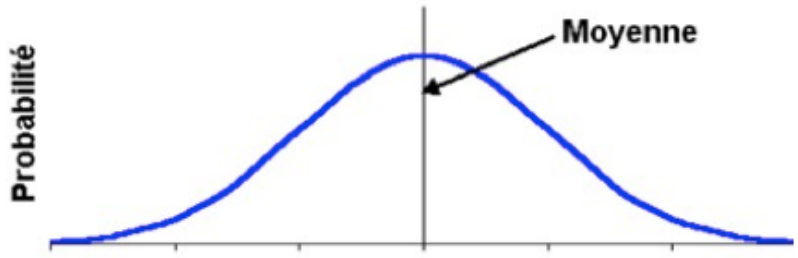



### Diagramme circulaire

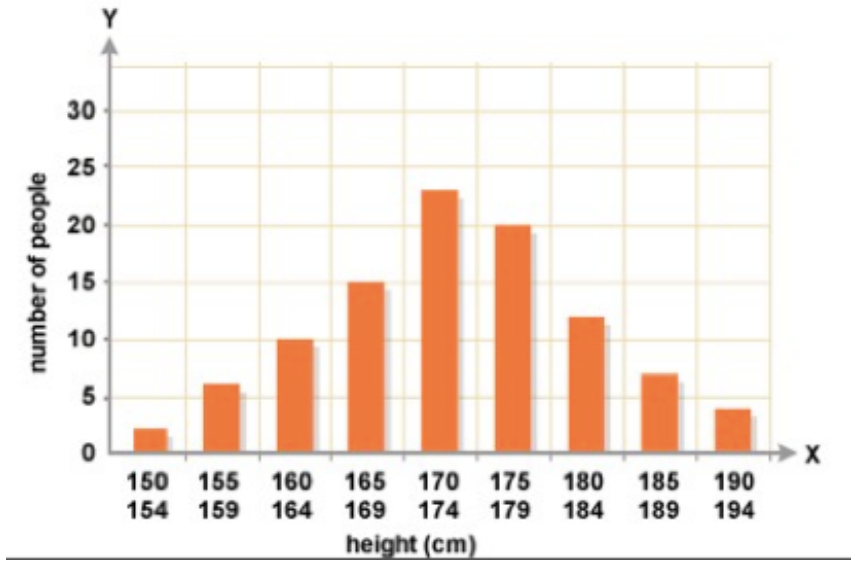


# Analyse de variables quantitatives

Idéalement, on souhaiterait avoir la distribution exacte d'une variable quantitative :




Pour représenter cette distribution à partir des données, on peut construire un *histogramme*



# Analyse de variables quantitatives

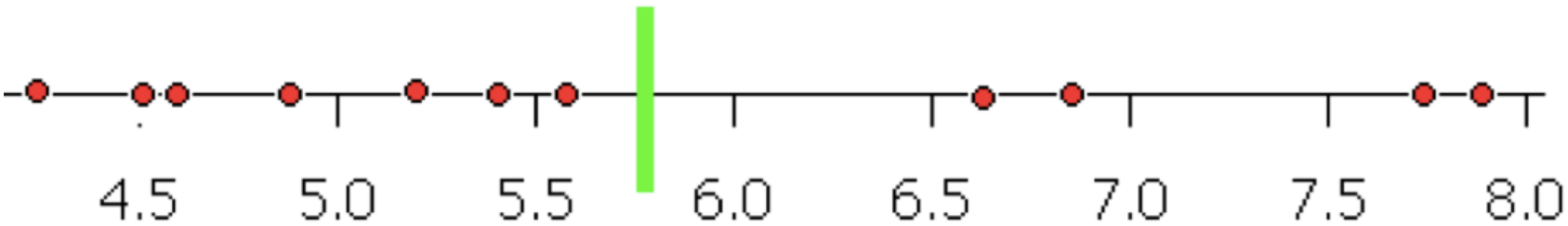
On peut aussi chercher les *caractéristiques* de cette distribution :

- Indicateurs de tendance centrale (moyenne, médiane,...)
- Indicateurs de dispersion (étendue, variance, écart type, coefficient de variation, interquartile...)
- Indicateurs de distribution (premier/dernier quartile, coefficients d'asymétrie,...)
- Outils de visualisation : fonction de répartition, boîte a moustache

# La moyenne... faut-il s'y fier ?

Centre de gravité, moyenne, $\cdots$

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




Moyenne de $\{4, 4.5, 4.6, 4.9, 5.2, 5.4, 5.6, 6.6, 6.8, 7.6, 7.8\} = 5.72$

# La moyenne... faut-il s'y fier ?

- 25 personnes gagnent 100 K€
- 20 personnes gagnent 80 K€
- 10 personnes gagnent 60 K€

$$\text{moyenne} = (25 \times 100 + 20 \times 80 + 10 \times 60)/55 = 85.45455\text{K€}$$

$\rightarrow$ Si on enlève 5 personnes du dernier groupe

$$\text{moyenne} = (25 \times 100 + 20 \times 80 + 5 \times 60)/45 = 97.77778\text{K€}$$

- Le salaire moyen a-t-il augmenté ?
- Un employé a-t-il vu son salaire augmenter ?

# La moyenne... faut-il s'y fier ?

Question :

On vous propose d'intégrer une entreprise A dont le salaire moyen des employés est de 3.545455€ de l'heure ou une entreprise B dont le salaire moyen est de 11.27273€.

Laquelle choisir ?

# La moyenne... faut-il s'y fier ?

Salaires de l'entreprise 1 : $\{0, 0, 0, 1, 1, 1, 3, 3, 5, 10, 15\}$
Moyenne=3.545455

Salaires de l'entreprise 2 : $\{0, 0, 0, 1, 1, 1, 3, 3, 5, 10, 100\}$
Moyenne=11.27273

## Questions :

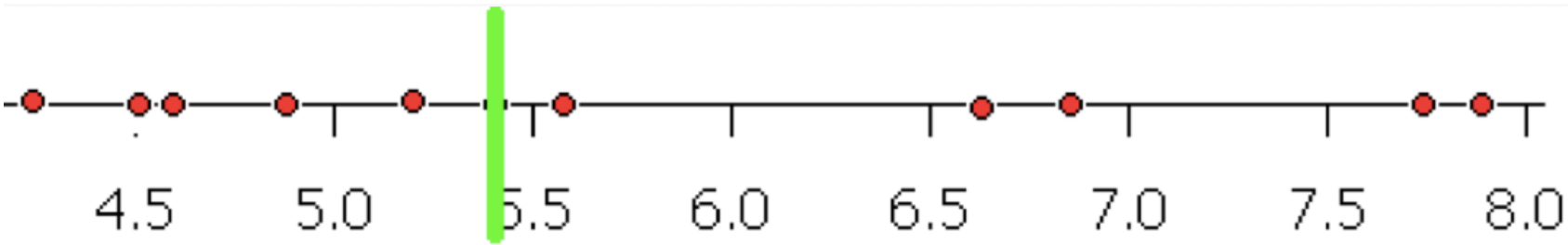
1. Est-ce vraiment plus intéressant de travailler pour l'entreprise 2 ?
2. La moyenne est-elle systématiquement un bon indicateur ?

# La médiane

## Définition

La médiane est une valeur $m$ permettant de diviser un échantillon en deux populations de taille égale.
Dans la première population tous les individus ont une valeur inférieur à $m$. Dans la deuxième population tous les individus ont une valeur supérieure à $m$.
Soit :

$$P(X \leq m) = P(X \geq m)$$



# La médiane

Médiane de $\{4, 4.5, 4.6, 4.9, 5.2, 5.4, 5.6, 6.6, 6.8, 7.6, 7.8\}$ = 5.4
Moyenne=5.72

50% des individus ont moins de 5.4 et 50% ont plus de 5.4

La médiane est parfois utilisée pour être « moins sensible » aux données aberrantes :
Médiane de
$\{4, 4.5, 4.6, 4.9, 5.2, 5.4, 5.6, 6.6, 6.8, 7.6, 100000000\} = 5.4$

# La médiane

- 25 personnes gagnent 100 K€
- 20 personnes gagnent 80 K€
- 10 personnes gagnent 60 K€

$$\text{moyenne} = 85.45455\text{K€}$$
$$\text{médiane} = 80\text{K€}$$

$\rightarrow$ Si on enlève 4 personnes du dernier groupe

$$\text{moyenne} = 87.45\text{K€}$$
$$\text{médiane} = 80\text{K€}$$

- Le salaire médian a-t-il augmenté ?
- Un employé a-t-il vu son salaire augmenter ?

Salaires de l'entreprise 1 : $\{0, 0, 0, 1, 1, 1, 3, 3, 5, 10, 15\}$
*Moyenne* = 3.545455
*Médiane* = ?

Salaires de l'entreprise 2 : $\{0, 0, 0, 1, 1, 1, 3, 3, 5, 10, 100\}$
*Moyenne* = 11.27273
*Médiane* = ?

## Questions :

1. Est-ce vraiment plus intéressant de travailler pour l'entreprise 2 ?
2. La moyenne est-elle systématiquement un bon indicateur ?

# Moyenne et médiane

- La médiane est plus robuste que la moyenne aux valeurs extrêmes
- Et donc$\cdots$ la médiane est un indicateur moins sensible que la moyenne. ex : une augmentation du salaire des 30% plus riche ne change pas la médiane
- Calcul de la moyenne plus robuste par trimage : éliminer k (%) des observations aux 2 extrémités de la distribution (k=1%, 2.5%, 5%)
- La différence entre moyenne et médiane peut donner une idée de la dissymétrie de la distribution :

# Indicateur de dispersion : l'écart-type

- L'évasement de la courbe correspond à l'écart-type.
- Selon ce que l'on mesure l'évasement peut être différent

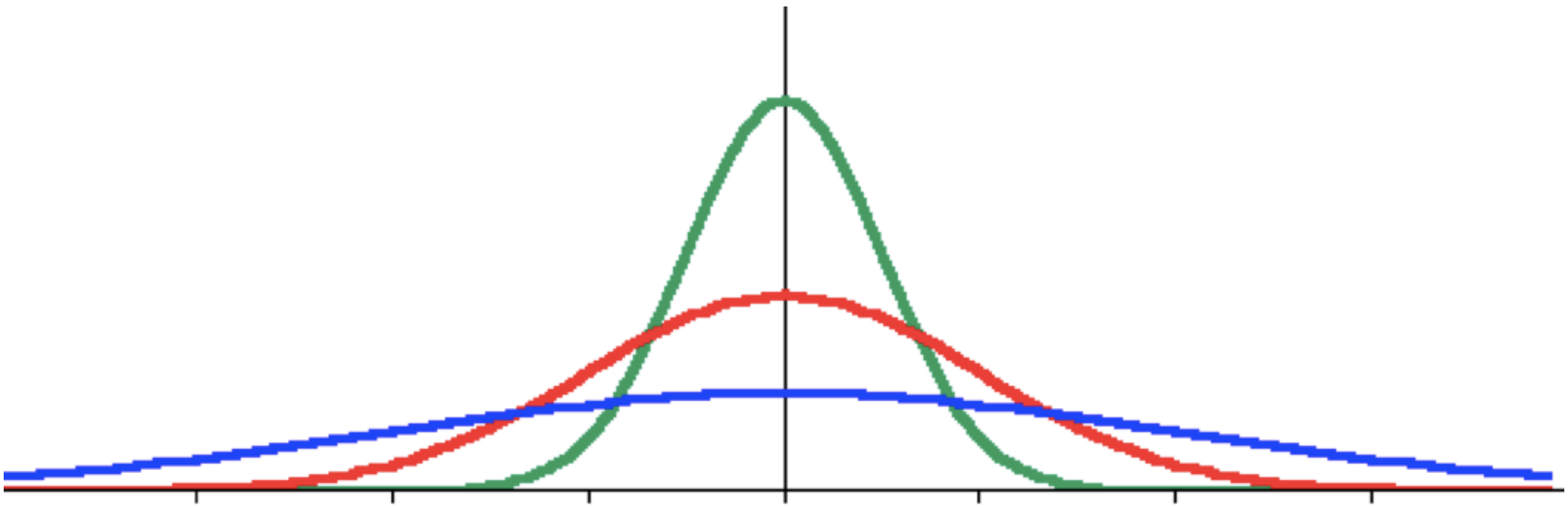


FIGURE : Différentes lois normales.

# L'écart-type

- Les mesures près de la moyenne sont plus probables que les mesures éloignées
- La distance parcourue par un véhicule avec un litre d'essence varie sur plusieurs essais (changement des conditions expérimentales, le hasard,$\cdots$)

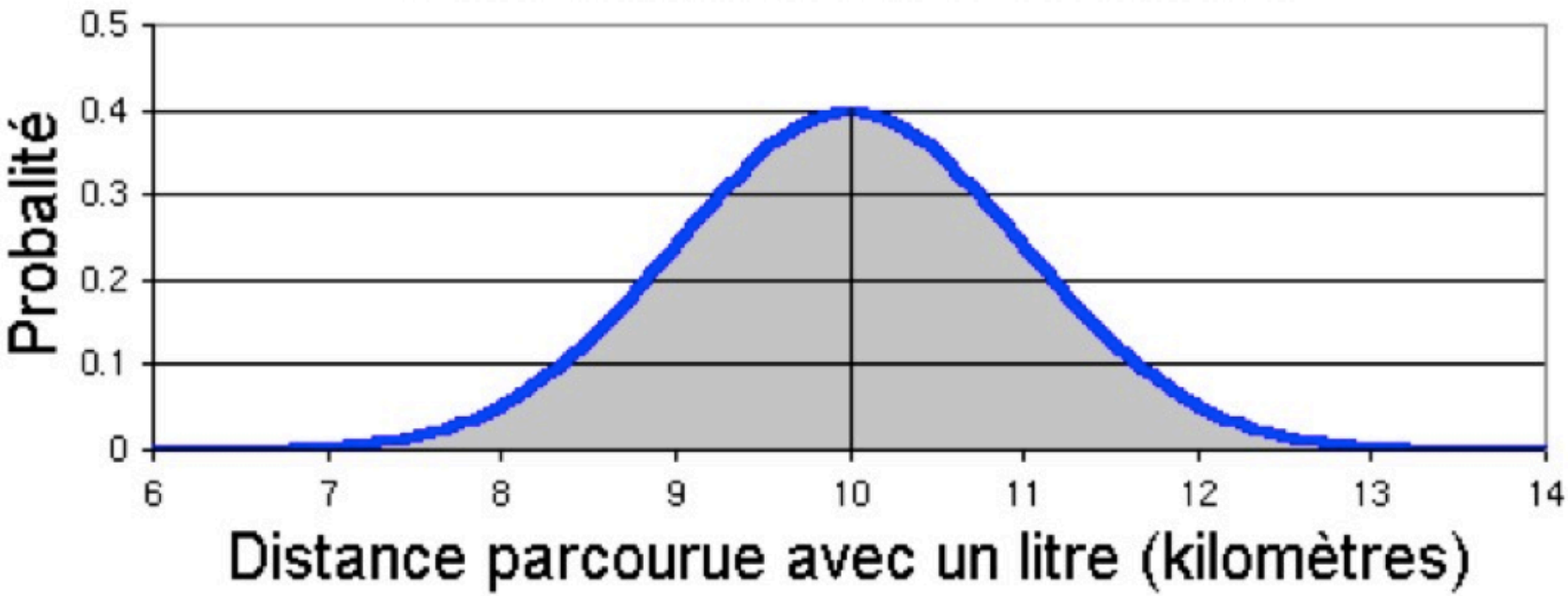



FIGURE : Consommation d'essence.

# L'écart-type

- L'écart-type exprime la dispersion des mesures
- Plus elle est élevée, plus les mesures sont dispersées et plus la courbe s'aplatit.
- La figure ci-dessous illustre trois courbes ayant la même moyenne et des écarts types différents

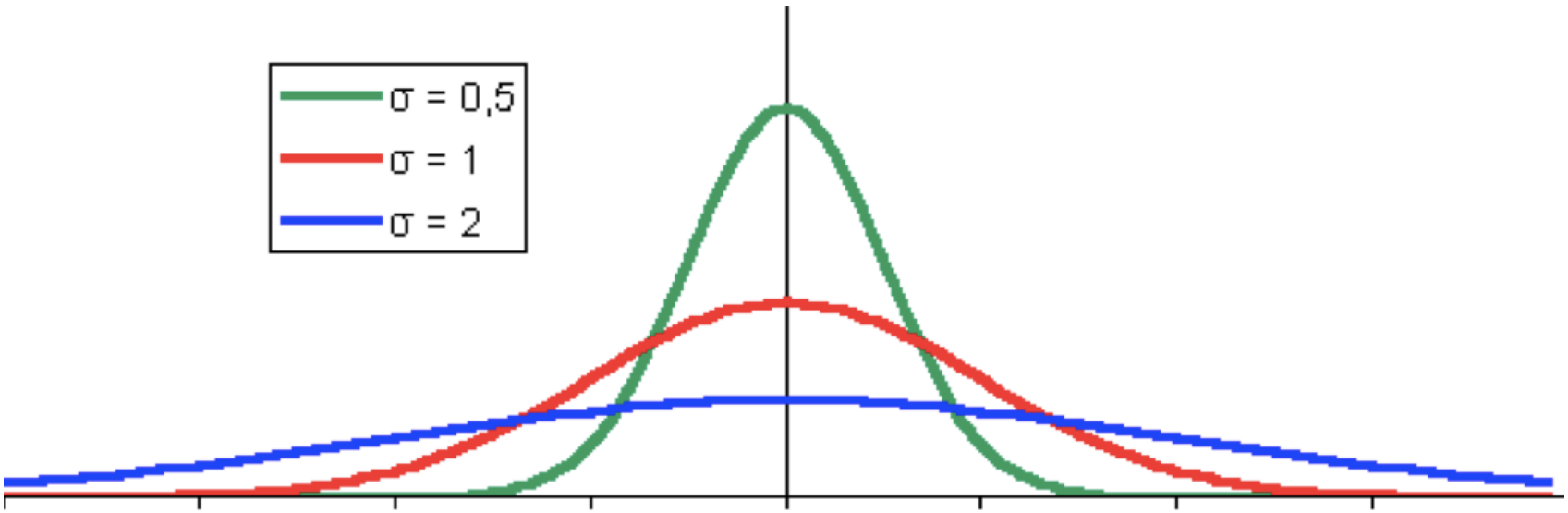



FIGURE : Loi normale selon différents écart-types $\sigma$.

# L'écart-type

Formule de calcul :

$$\sigma(x) = \sqrt{Var(x)} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

Rque : on divise souvent par $n - 1$ au lien de $n$

Salaires de l'entreprise 1 : $\{0, 0, 0, 1, 1, 1, 3, 3, 5, 10, 15\}$
Moyenne = 3.545455

$$\begin{aligned}\sigma &= \sqrt{\frac{(0 - 3.5)^2 + \cdots + (15 - 3.5)^2}{11}} \\ &= \sqrt{21.15702} = 4.599676\end{aligned}$$

# L'écart-type

Salaires de l'entreprise 2 : $\{0, 0, 0, 1, 1, 1, 3, 3, 5, 10, 100\}$
Moyenne = 11.27273

$$\sigma = \sqrt{\frac{(0 - 11.27273)^2 + \cdots + (100 - 11.27273)^2}{11}}$$
$$= \sqrt{855} = 29.24$$

## Analyse

Dans l'entreprise 2, les salaires sont beaucoup plus dispersés autour de la moyenne que dans l'entreprise 1. On suppose donc que dans l'entreprise 2 il y a des écarts plus importants entre les salaires que dans la première entreprise.

# Autres indicateurs

- *Quartiles, Déciles*
- Mesures de dispersion
  *Coefficient de variation, interquartile, min,max*
- Mesure d'assymétrie

$$skewness = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^3$$

- Visualisation :
  *Tracé de la fonction de répartition*
  *boîte à moustache*

# Exemple concret : analyse de salaires

Salaires net annuel par région(source insee)

| 3 | | Repartition des effectifs | Salaires nets annuels | | | | | |
|---|---|---|---|---|---|---|---|---|
| 4 | | 2006 | 2 001 | 2 006 | | | | |
| 5 | | % | Ensemble | Ensemble | Cadres * | Professions intermediaires | Employés | Ouvriers |
| 6 | Alsace | 3,1 | 19 748 | 21 996 | 41 669 | 23 458 | 16 209 | 17 780 |
| 7 | Aquitaine | 4,5 | 18 528 | 20 645 | 40 650 | 22 700 | 15 964 | 16 507 |
| 8 | Auvergne | 1,8 | 18 020 | 20 276 | 42 917 | 22 848 | 15 892 | 16 573 |
| | | | | | | | | |
| 26 | Provence-Alpe | 6,7 | 19 578 | 21 975 | 41 903 | 23 480 | 16 064 | 17 061 |
| 27 | Rhône-Alpes | 10,2 | 20 131 | 22 445 | 42 970 | 23 446 | 16 268 | 17 311 |
| 28 | **Métropole** | **98,1** | **20 873** | **23 306** | **23 446** | **23 566** | **16 325** | **17 083** |
| 29 | Guadeloupe | 0,5 | 18 859 | 21 257 | 46 576 | 24 511 | 16 626 | 16 063 |
| 30 | Guyane | 0,1 | 22 253 | 22 912 | 46 540 | 26 644 | 17 465 | 16 427 |
| 31 | Martinique | 0,5 | 19 377 | 21 198 | 47 903 | 24 351 | 16 244 | 16 000 |
| 32 | Réunion | 0,8 | 17 873 | 20 194 | 45 846 | 24 076 | 16 140 | 15 911 |
| 33 | **France** | **100,0** | **20 841** | **23 261** | **46 221** | **23 582** | **16 327** | **17 064** |

# Exemple concret : préparation des données

On vérifie :

- Qu'il ne manque pas d'information (cases vides) $\rightarrow ok$
- Les données sont elles adaptées aux outils d'analyse ? $\rightarrow \overline{ok}$ attention à ne pas faire des moyennes de pourcentages
- Lignes ou colonnes « polluantes » $\rightarrow \overline{ok}$ problème lignes 28 et 30

| 3 | | Repartition des effectifs | Salaires nets annuels | | | | | |
|---|---|---|---|---|---|---|---|---|
| 4 | | [illegible] | 2 001 | 2 006 | | | | |
| 5 | | % | Ensemble | Ensemble | Cadres * | Professions intermediaires | Employés | Ouvriers |
| 6 | Alsace | 3,1 | 19 748 | 21 996 | 41 669 | 23 458 | 16 209 | 17 780 |
| 26 | Provence-Alpes | 6,7 | 19 578 | 21 975 | 41 903 | 23 480 | 16 064 | 17 061 |
| 27 | Rhône-Alpes | 10,2 | 20 131 | 22 445 | 42 970 | 23 446 | 16 268 | 17 311 |
| [illegible] | **Métropole** | **98,1** | **20 873** | **23 306** | **23 446** | **23 566** | **16 325** | **17 083** |
| 29 | Guadeloupe | 0,5 | 18 859 | 21 257 | 46 576 | 24 511 | 16 626 | 16 063 |
| 30 | Guyane | 0,1 | 22 253 | 22 912 | 46 540 | 26 644 | 17 465 | 16 427 |
| 31 | Martinique | 0,5 | 19 377 | 21 198 | 47 903 | 24 351 | 16 244 | 16 000 |
| 32 | Réunion | 0,8 | 17 873 | 20 194 | 45 846 | 24 076 | 16 140 | 15 911 |
| 33 | [illegible] | 100,0 | 20 841 | 23 261 | 46 221 | 23 582 | 16 327 | 17 064 |

# Exemple concret : traitement des données

On calcule : moyenne, médiane et écart type pour chaque catégorie

| | somme2001 | somme2008 | Cadres | ProfInter | Employes | Ouvriers |
|---|---|---|---|---|---|---|
| Min. | 17577 | 19686 | 39742 | 22072 | 15589 | 15911 |
| Median | 18500 | 20618 | 41181 | 23040 | 15955 | 16540 |
| Mean | 18982 | 21160 | 42386 | 23289 | 16101 | 16748 |
| Max. | 26745 | 29940 | 51840 | 26644 | 17465 | 183 |
| ecart | 1867.9264 | 1975.7701 | 2956.3217 | 1041.3297 | 445.9654 | 636.6783 |

FIGURE : Traitement des variables

| | MoyMin08 | MoyMax08 | EcartT min | EcartT max | min evol | max evol |
|---|---|---|---|---|---|---|
| Région | Poit char | IDF | Bretagne | IDF | Guyane | IDF |

FIGURE : Traitement des individus

Et tout ce que vous voulez extraire...

# Exemple concret : interprétation des résultats

Voici quelques exemples d'une analyse (objective) des variables :

- En 2008, 50% des Français gagnent plus de 20618 k€, et 50% moins
- En 2008, le salaire moyen est de 21160 k€
- En moyenne, le salaire le plus élevé en 2008 est celui des cadres
- Entre 2001 et 2008 l'écart entre les plus faibles salaires et les plus gros a augmenté
- En 2008 ce sont les employés qui sont la catégorie avec le moins de disparité de salaire
- Les professions intermédiaires et cadres gagnent plus que la moyenne des français
- Le plus bas des salaires moyens (des régions) en 2008 est supérieur de 2109 k€ plus bas des salaires moyens (des régions) en 2009

Faites attention à ne pas faire dire aux chiffres plus de choses qu'ils ne le peuvent !

# Exemple concret : interprétation des résultats

Voici quelques exemples d'une analyse (objective) des individus :

- En moyenne, c'est en Poitou-Charentes que l'on gagne le moins bien sa vie en 2008 (resp en IDF pour le meilleur salaire)
- C'est en Bretagne que les salaires sont les moins dispersés autour de la moyenne
- C'est en Ile de France que l'écart à la moyenne entre les salaires est le plus élevé
- En moyenne, c'est en Guyane que les salaires ont le moins augmenté (resp IDF pour le plus)

Analyser des données est un travail différent de celui consistant à trouver des explications à cette analyse