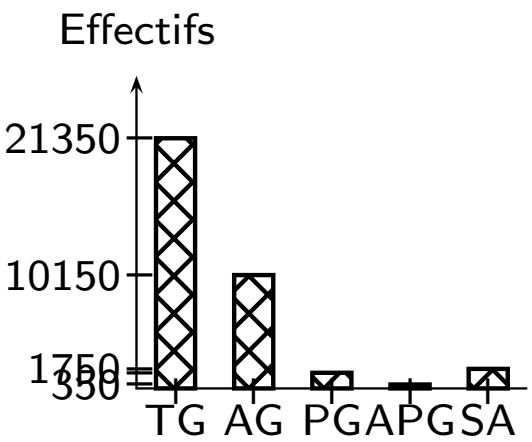
Statistique, analyse exploratoire des données et probabilités

Valérie Henry

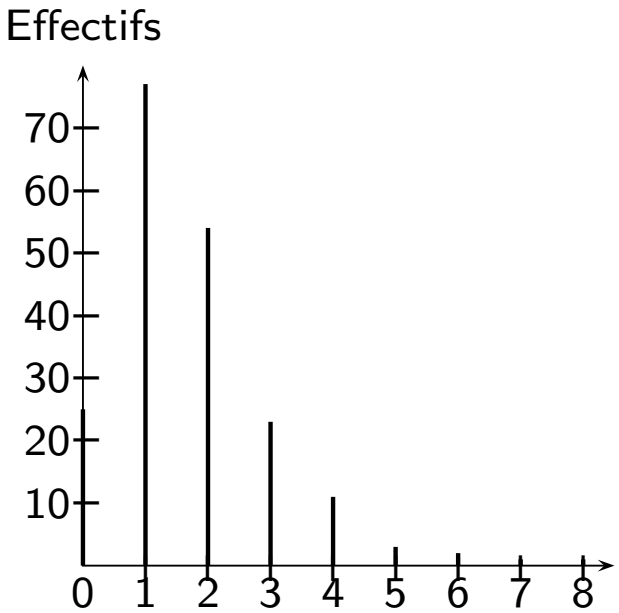
ULg, UNamur, CREM

24 avril 2015

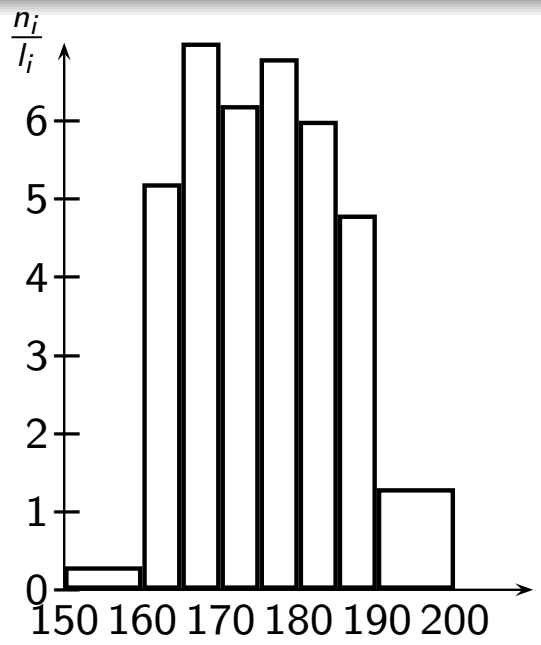
Interpréter et construire un graphique



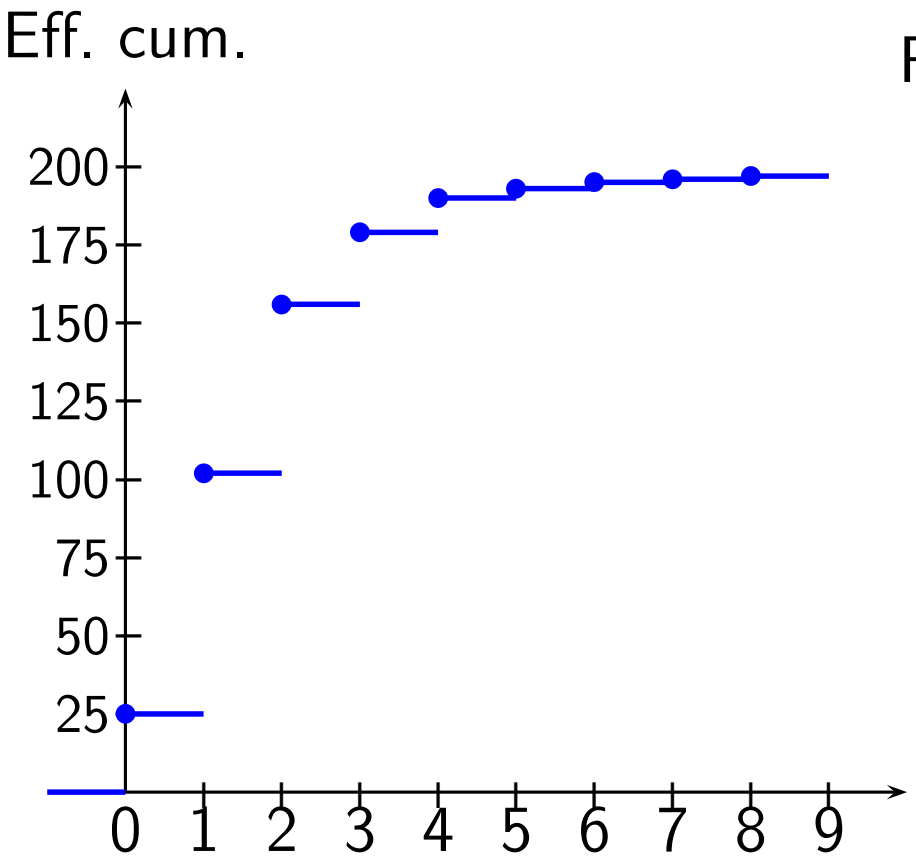
Réponse de 35 000 Wallons à la question suivante : "Estimez-vous que les faits qui sont reprochés à certains exdirigeants de La Carolorégienne sont
Très graves — Assez graves — Peu graves — Absolument pas graves — Sans avis"
Le Soir 10.10.05

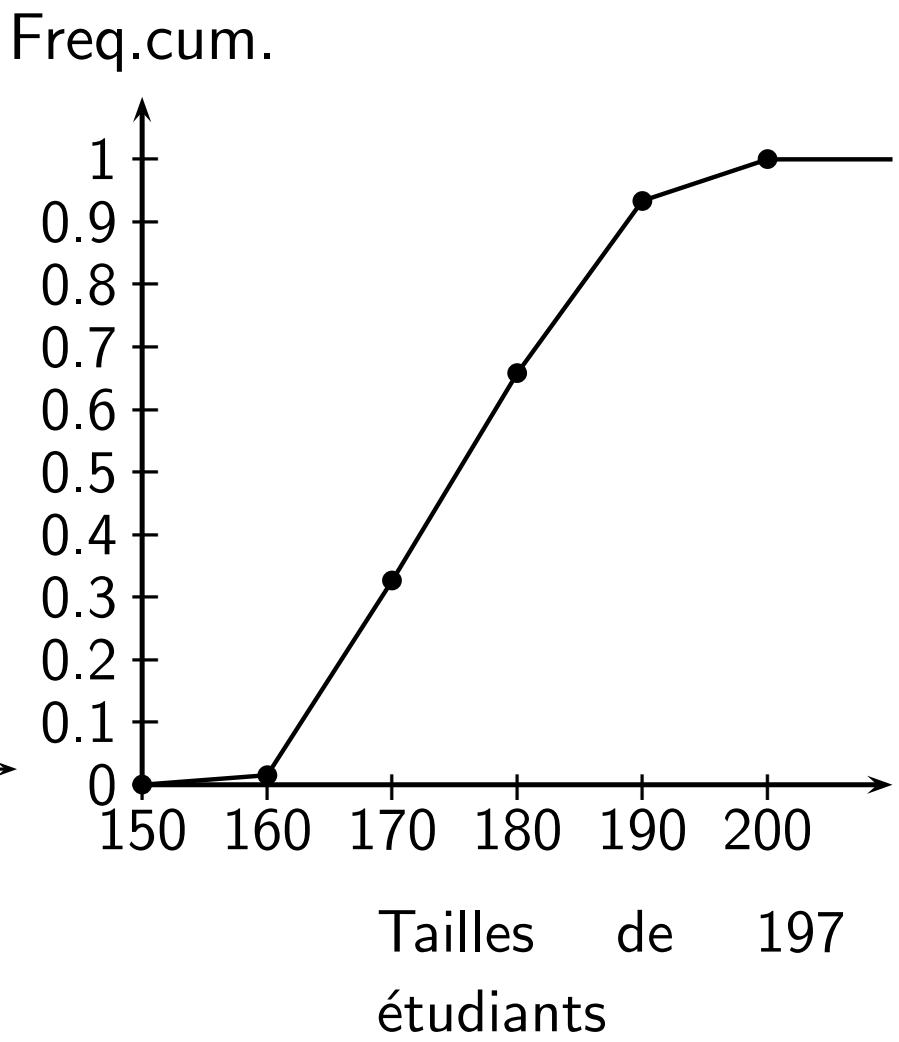


Nombre de frères et sœurs de 197 étudiants



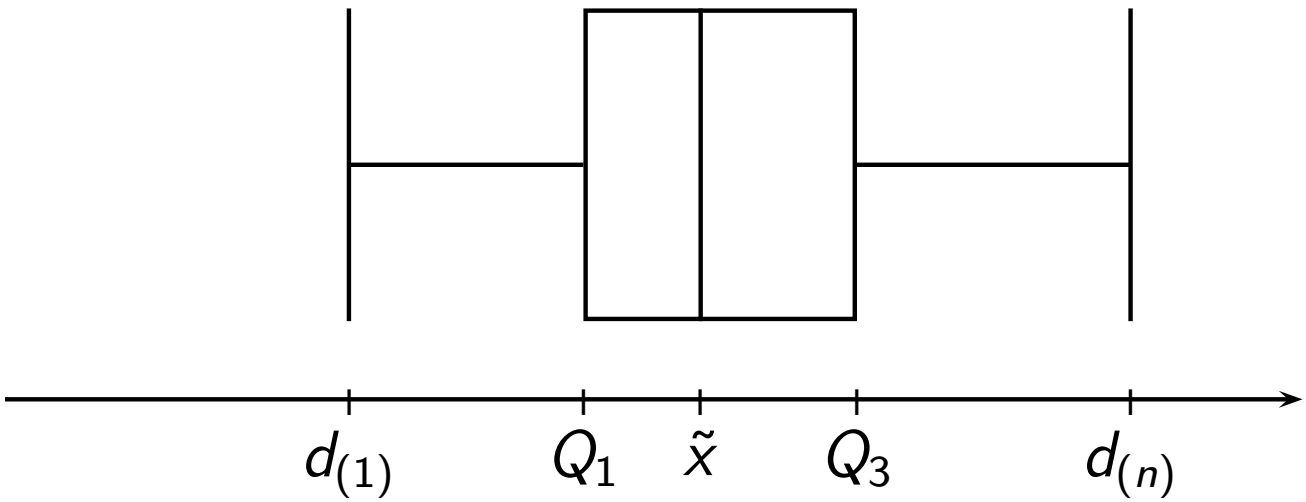
Tailles de 197 étudiants


Eff. cum.
200
175
150
125
100
75
50
25
0 1 2 3 4 5 6 7 8 9
Nombre de frères et sœurs de 197 étudiants

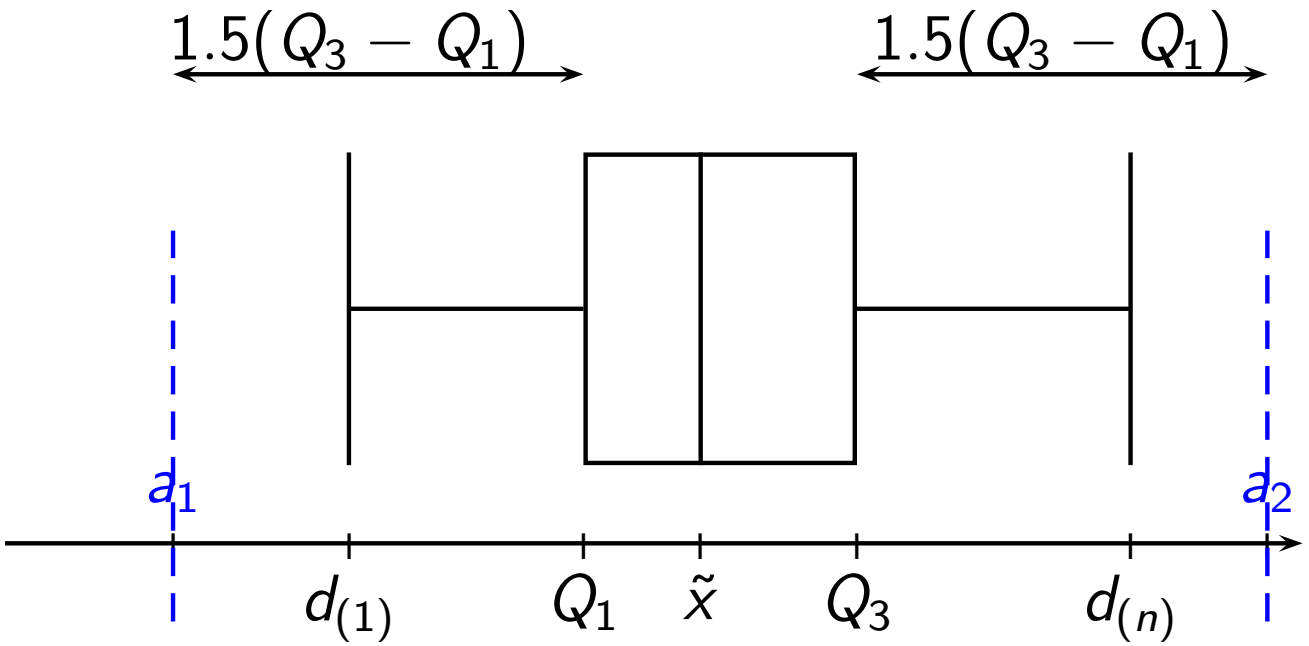

Freq.cum.
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
150 160 170 180 190 200
Tailles de 197 étudiants

	Qualitatif		Quantitatif	
	Nominal	Ordinal	Discret	Continu
Effectifs ou fréq.	barres ou secteurs		bâtons ou secteurs	Histo-gramme Polygone
Eff. cum ou fréq. cum.	////////		Courbe cumula-tive (en escalier)	Ogive (continue, affine par morceaux)

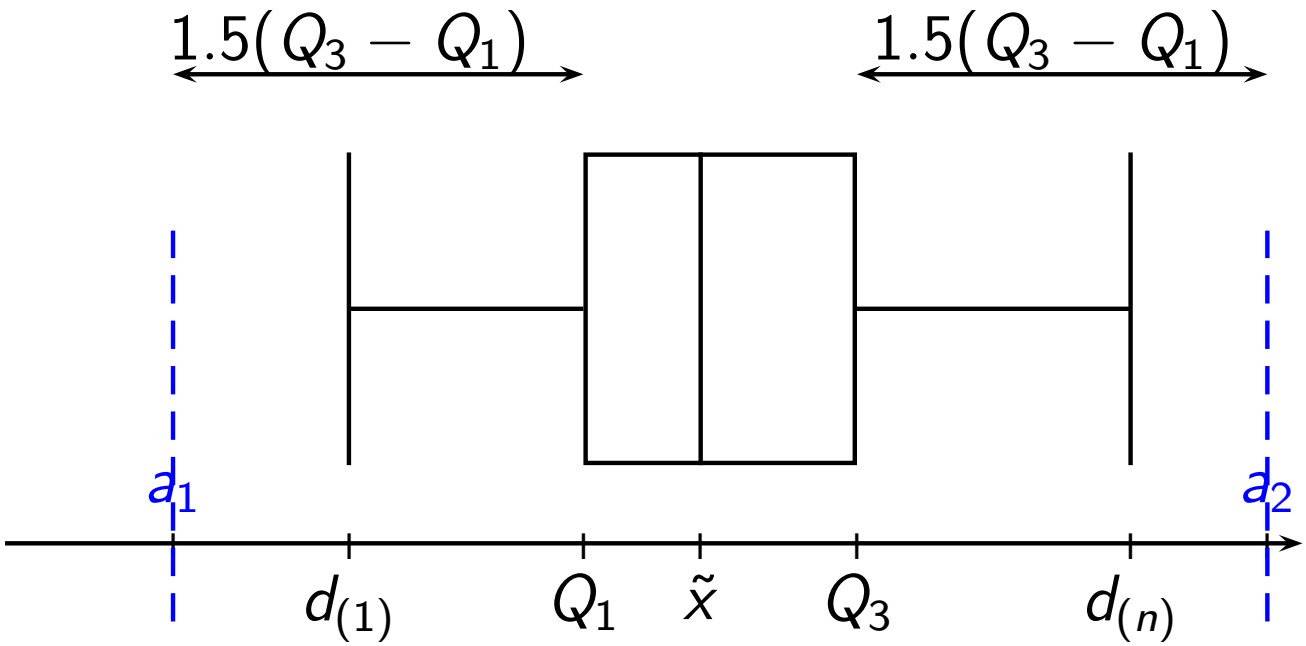
BAM concentrée



BAM concentrée



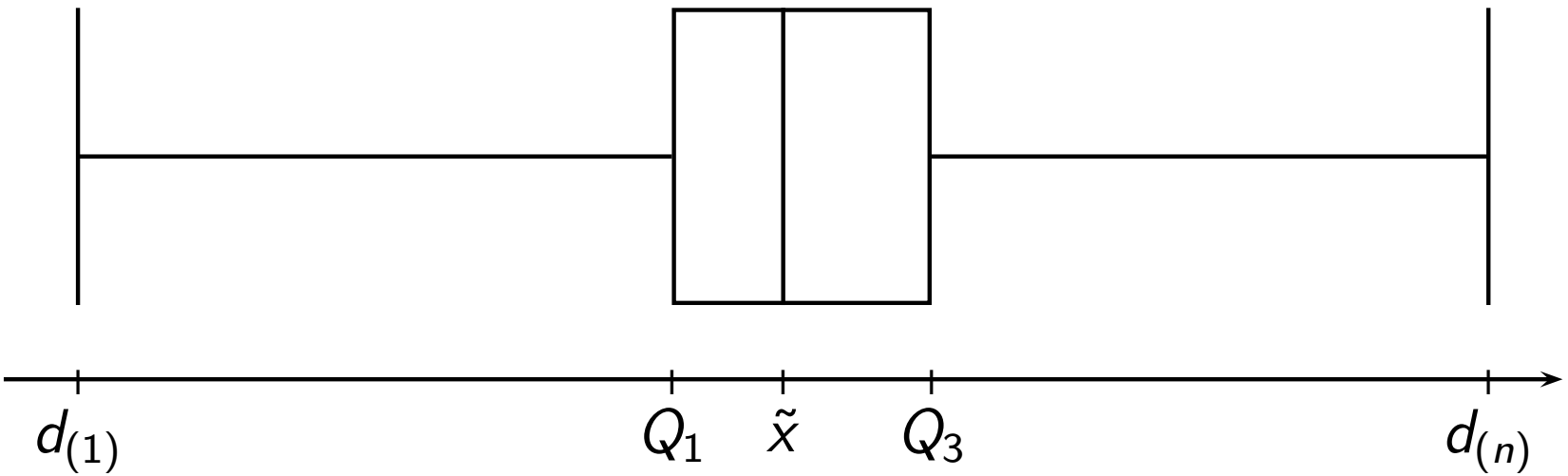
BAM concentrée



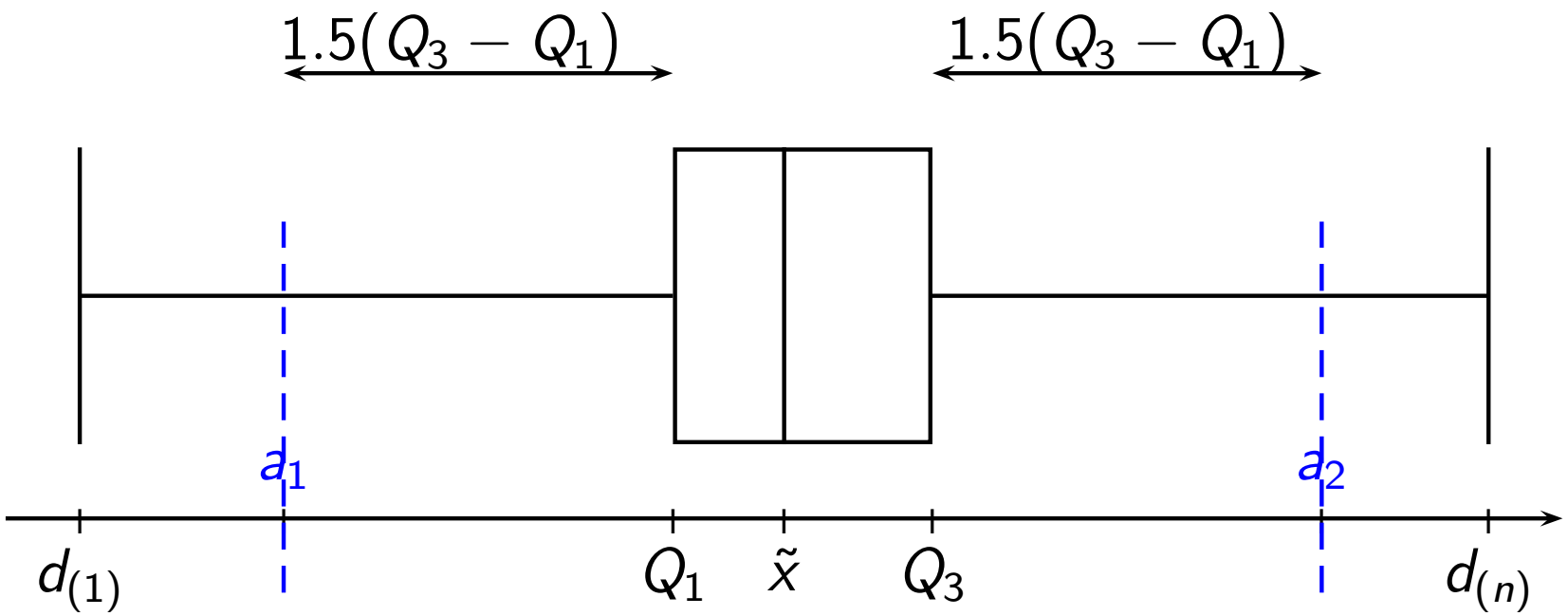
$$a_1 = Q_1 - 1.5(Q_3 - Q_1)$$

$$a_2 = Q_3 + 1.5(Q_3 - Q_1)$$

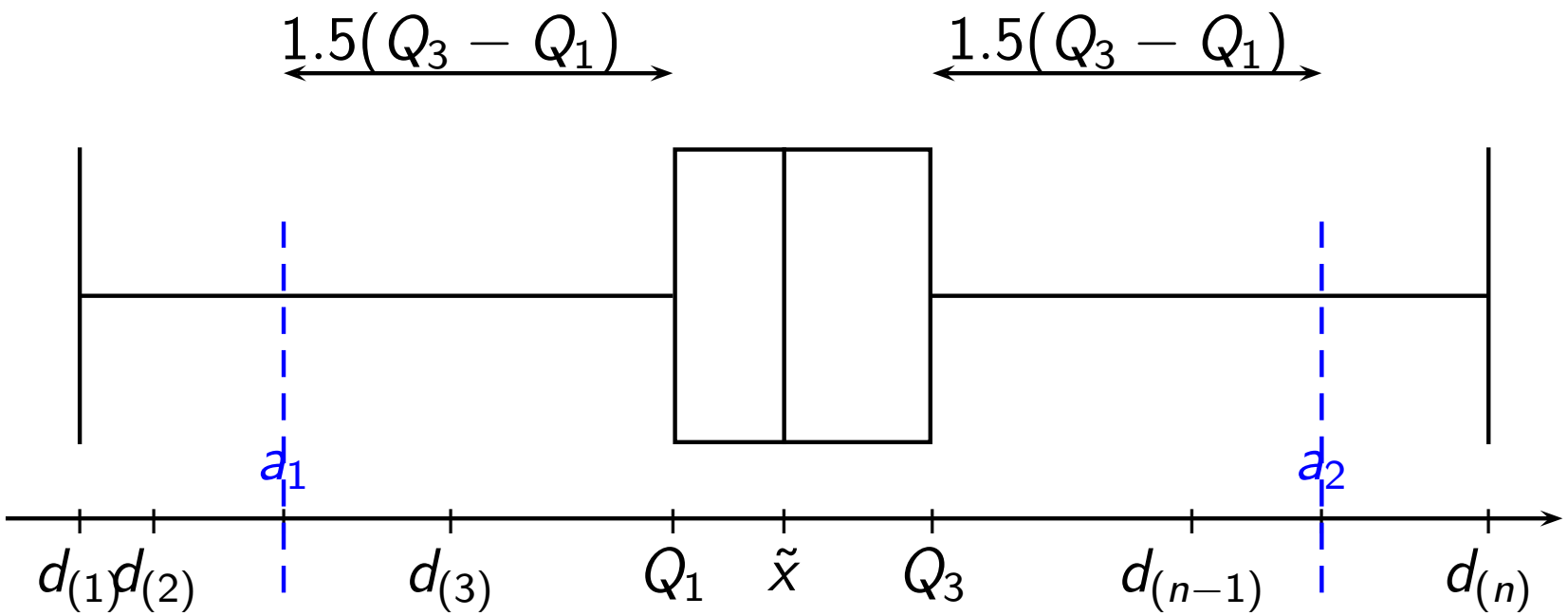
BAM dispersée



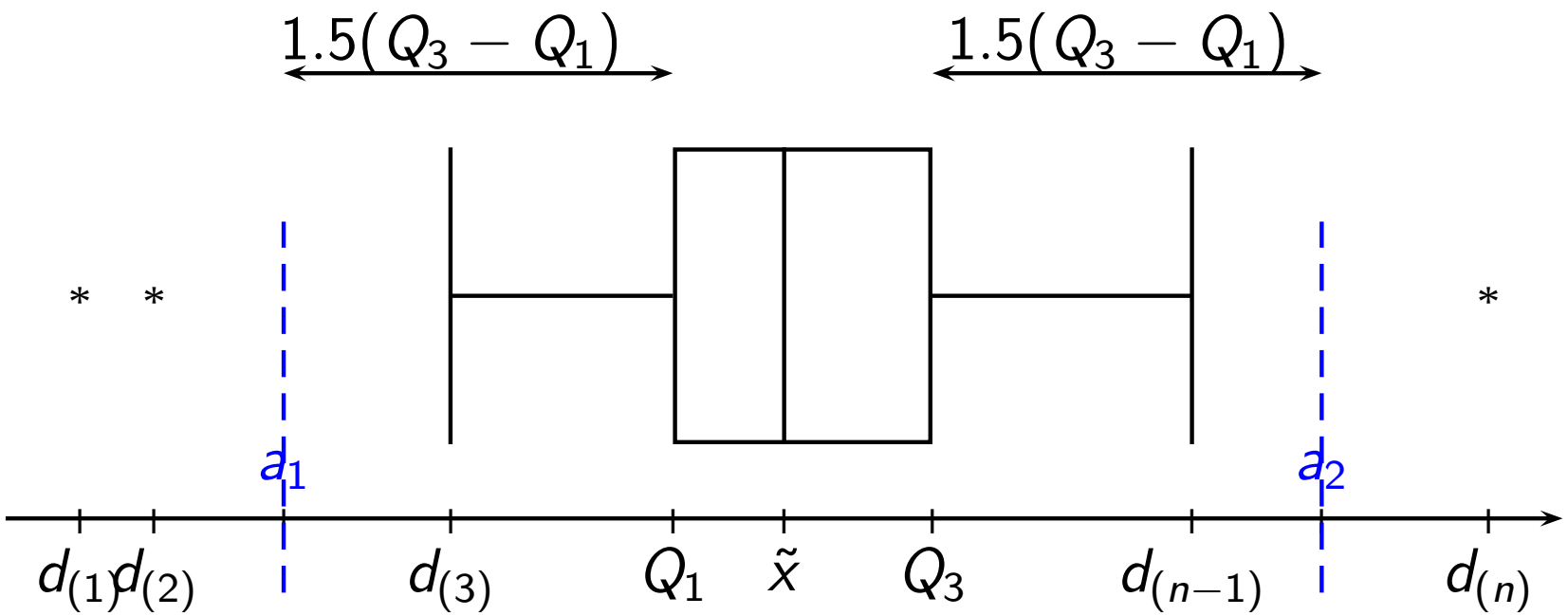
BAM dispersée



BAM dispersée



BAM dispersée



Taux de chômage (en %) observés dans 13 pays industrialisés pour les 25-64 ans, suivant le niveau d'études

	Taux de chômage selon le niveau d'études			
Pays	primaire	secondaire	sup. non univ.	univ.
Allemagne	13,3	7,9	5,2	4,7
Belgique	13,4	7,5	3,5	3,6
Canada	13	8,6	7,5	4,6
Danemark	14,6	8,3	5,3	4,3
Espagne	20,6	18,5	16,6	13,8
États-Unis	10	5	3,6	2,5
France	14	8,9	5,9	7
Grèce	6,3	9	10,1	7,1
Irlande	16,4	7,6	5	3,4
Norvège	6,5	4	3,4	1,7
Royaume Uni	12,2	7,4	4,1	3,5
Suède	10,1	8,7	4,8	4,2
Suisse	5,8	2,8	1,5	2,6

Supposons que la moyenne des notes obtenues à l'examen du cours de statistique en janvier soit de 70 sur 100 avec un écart-type de 5.
Combien d'étudiants ont obtenu une note entre 60 et 80 ?

Supposons que la moyenne des notes obtenues à l'examen du cours de statistique en janvier soit de 70 sur 100 avec un écart-type de 5.
Combien d'étudiants ont obtenu une note entre 60 et 80 ?

Inégalité de Tchebychev

Pour une série S, de moyenne $\bar{x}$ et de variance σ^2 et pour tout $t > 0$, la proportion d'observations qui appartiennent à l'intervalle $[\bar{x} - t\sigma, \bar{x} + t\sigma]$ est supérieure ou égale à $1 - \frac{1}{t^2}$.

Supposons que la moyenne des notes obtenues à l'examen du cours de statistique en janvier soit de 70 sur 100 avec un écart-type de 5.
Combien d'étudiants ont obtenu une note entre 60 et 80 ?

Inégalité de Tchebychev

Pour une série S, de moyenne $\bar{x}$ et de variance σ^2 et pour tout $t > 0$, la proportion d'observations qui appartiennent à l'intervalle $[\bar{x} - t\sigma, \bar{x} + t\sigma]$ est supérieure ou égale à $1 - \frac{1}{t^2}$.

Combien d'étudiants ont obtenu une note entre 48 et 92 ?

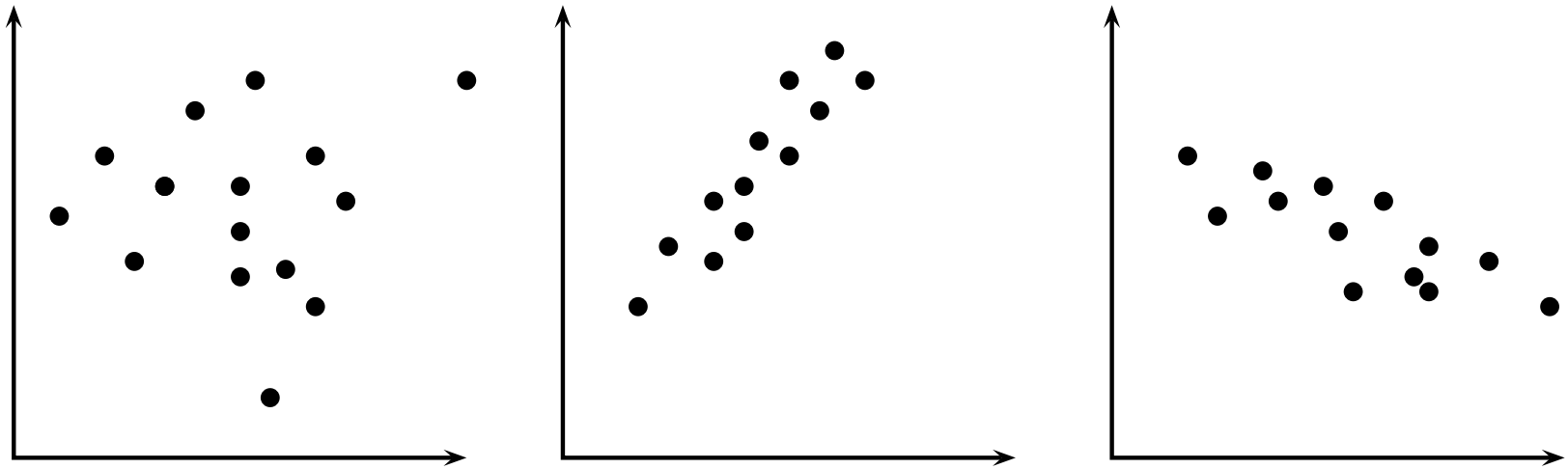
Démonstration

Soit $S = \{d_1, d_2, \ldots, d_n\}$, ordonnons les observations de telle sorte que $\begin{cases} |d_i - \bar{x}| \leq t\sigma & \forall\ 1 \leq i \leq k \\ |d_i - \bar{x}| > t\sigma & \forall\ k+1 \leq i \leq n \end{cases}$

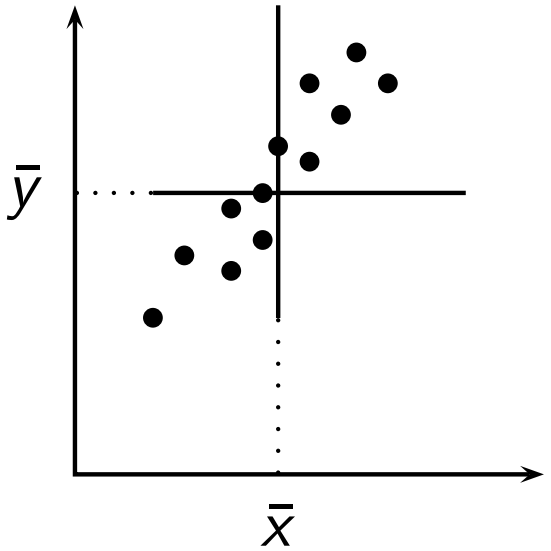
Par définition, on a $n\sigma^2 = \sum_{i=1}^{n}(d_i - \bar{x})^2$. Dès lors,

$$\begin{aligned} n\sigma^2 &> \sum_{i=k+1}^{n}(d_i - \bar{x})^2 \\ n\sigma^2 &> (n-k)t^2\sigma^2 \\ \frac{n}{t^2} &> n-k \\ k &> n - \frac{n}{t^2} \end{aligned}$$

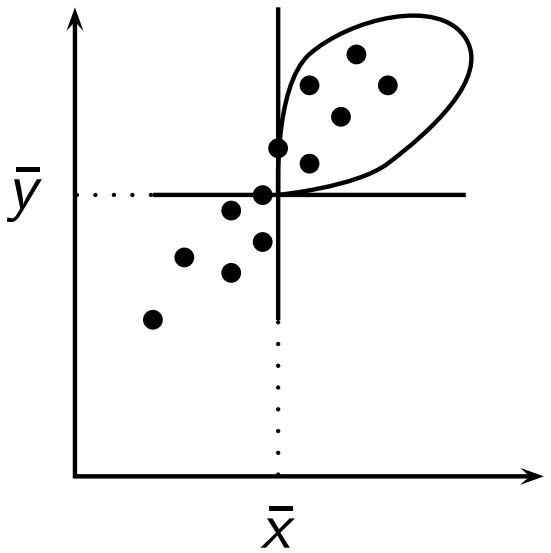
Nuage de points



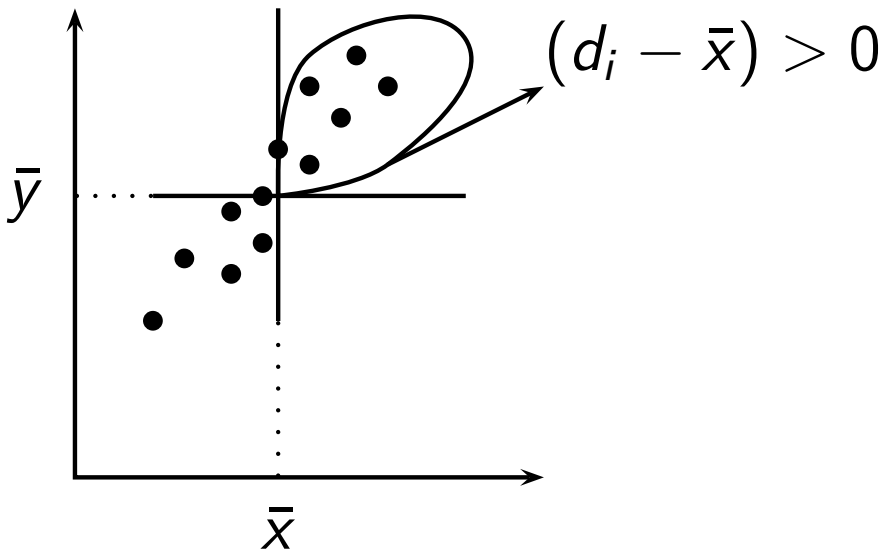
Nuage de points



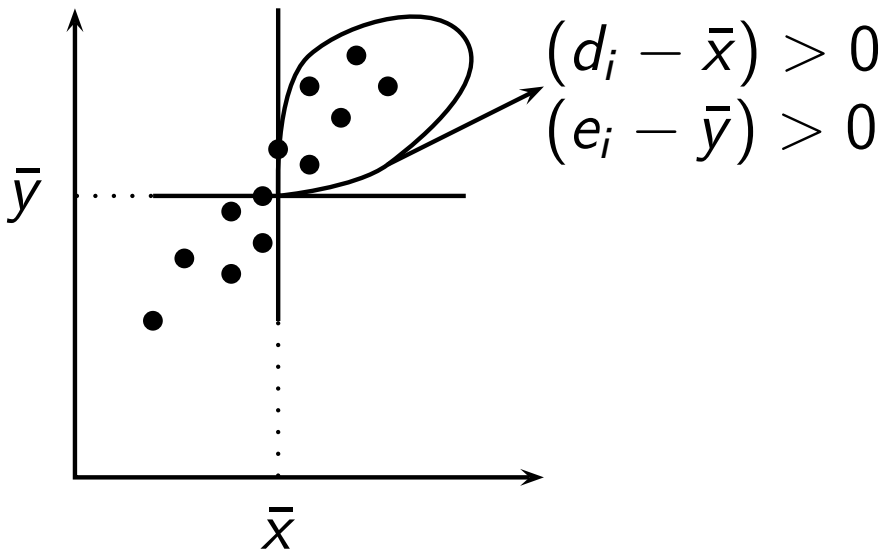
Nuage de points



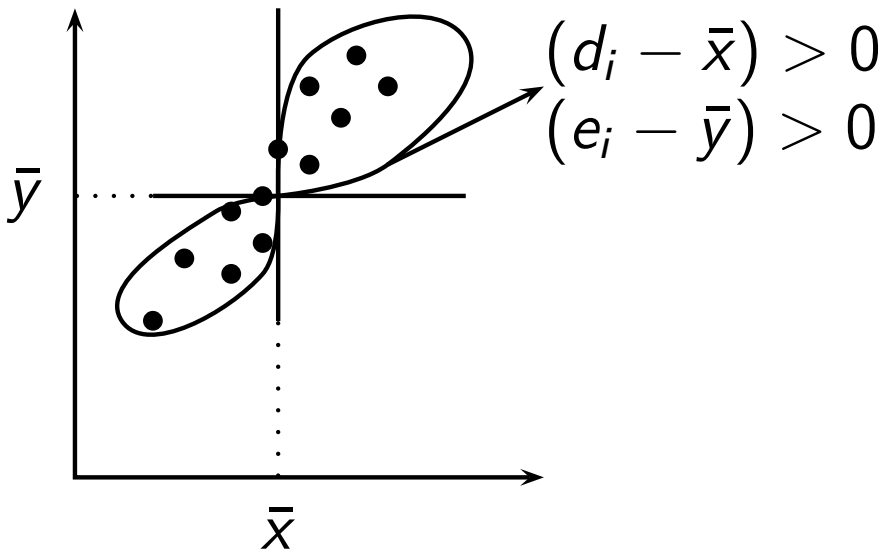
Nuage de points



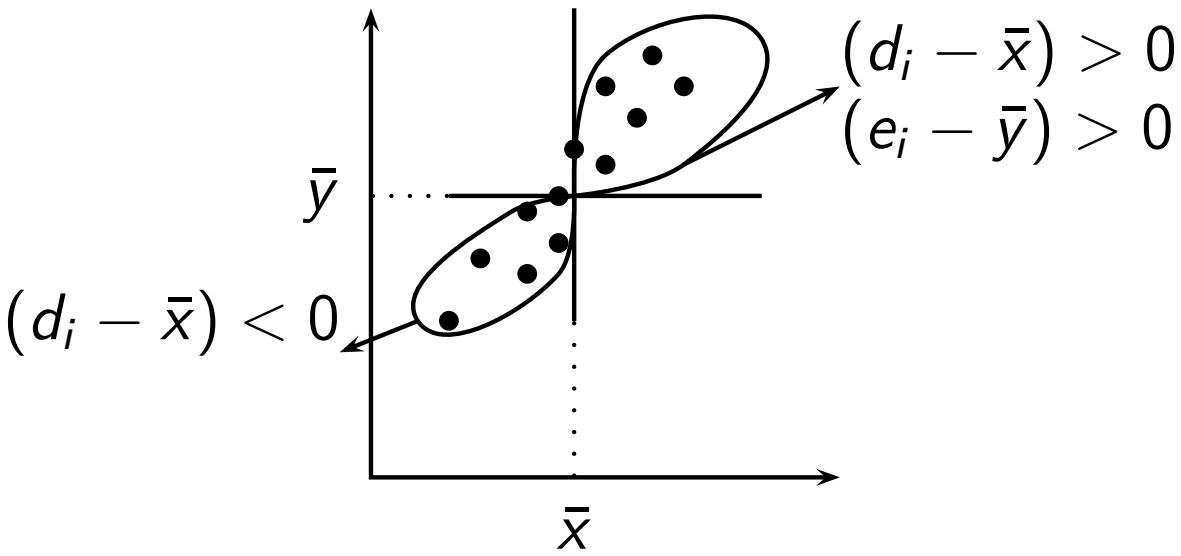
Nuage de points



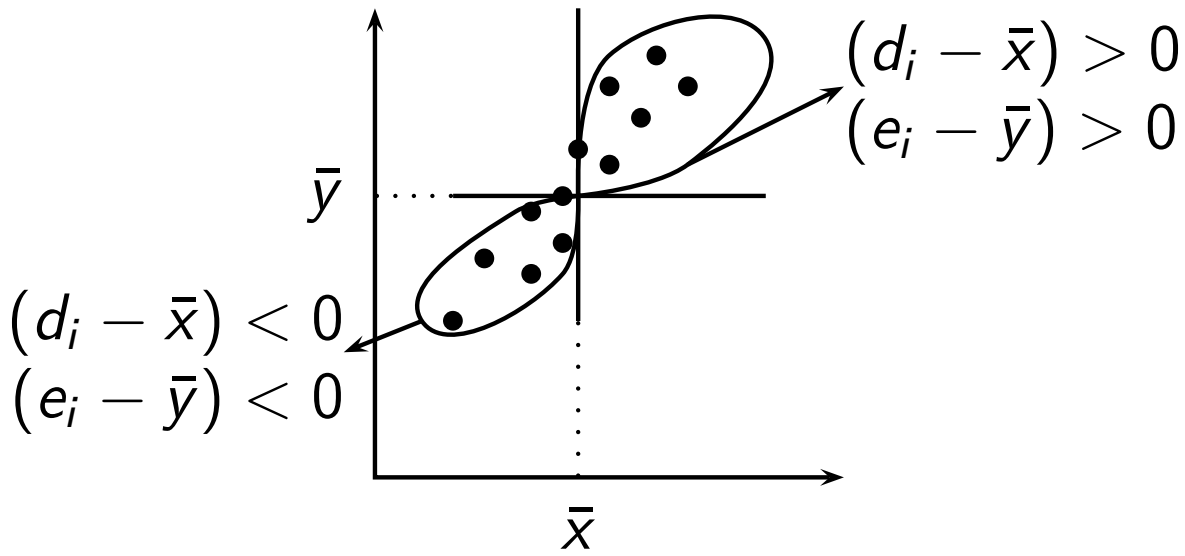
Nuage de points



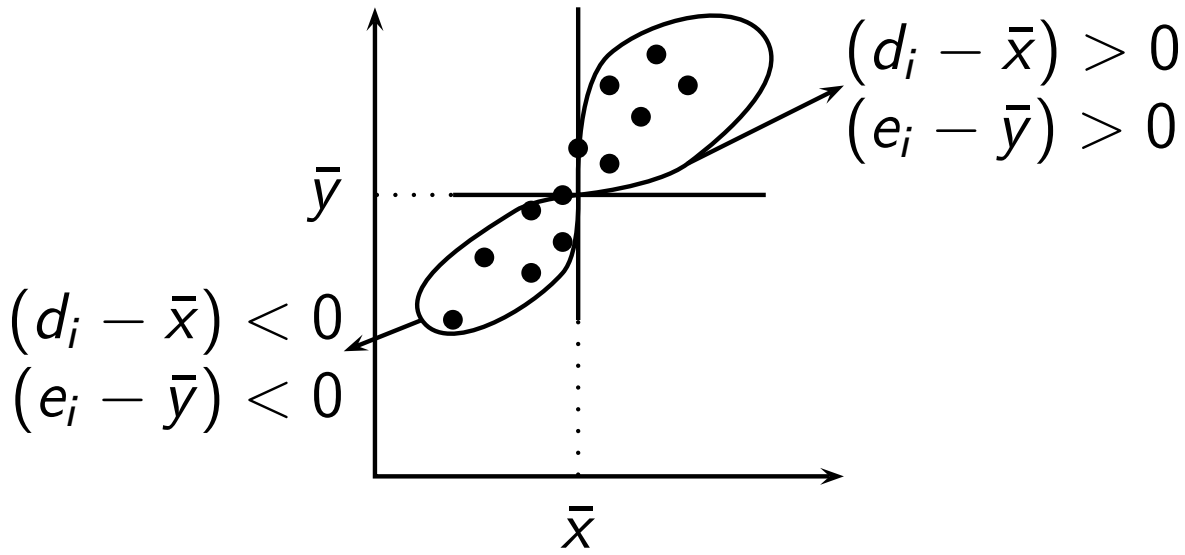
Nuage de points



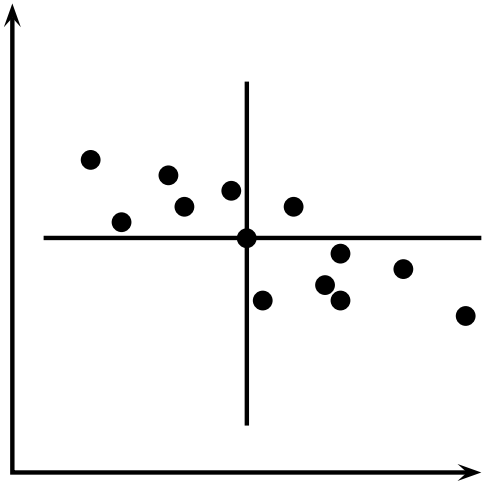
Nuage de points

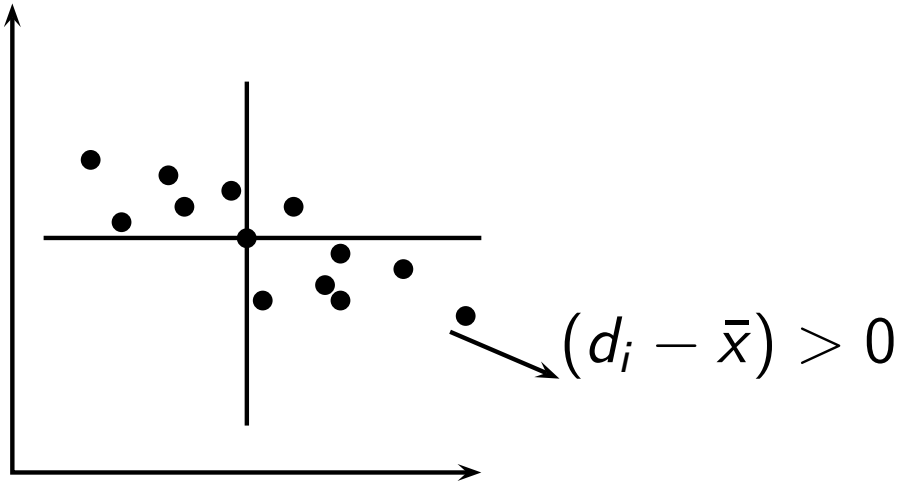


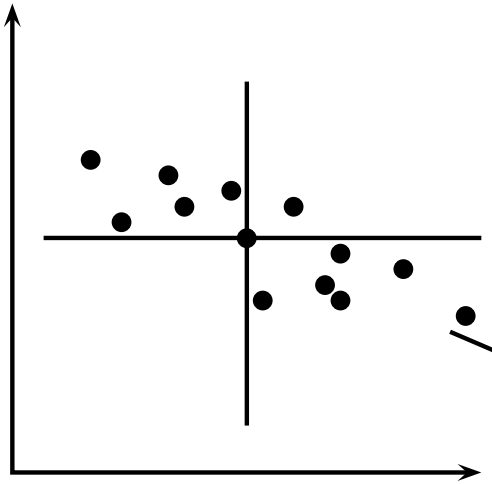
Nuage de points



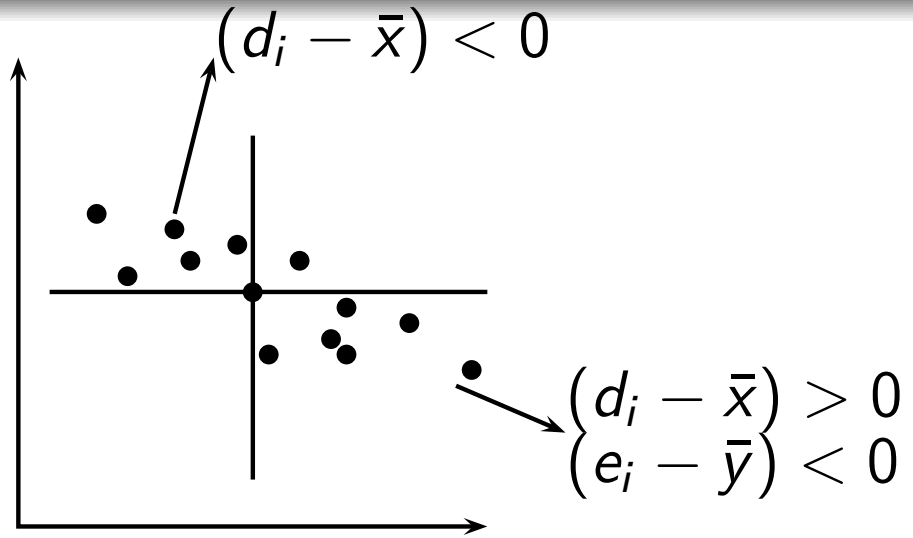
Globalement, $\sum_{i=1}^{n}(d_i - \bar{x})(e_i - \bar{y}) > 0$

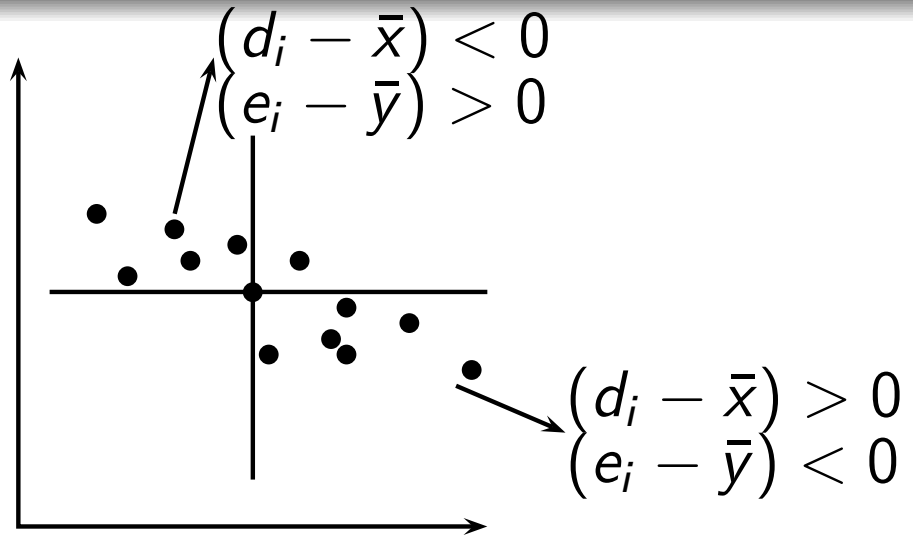


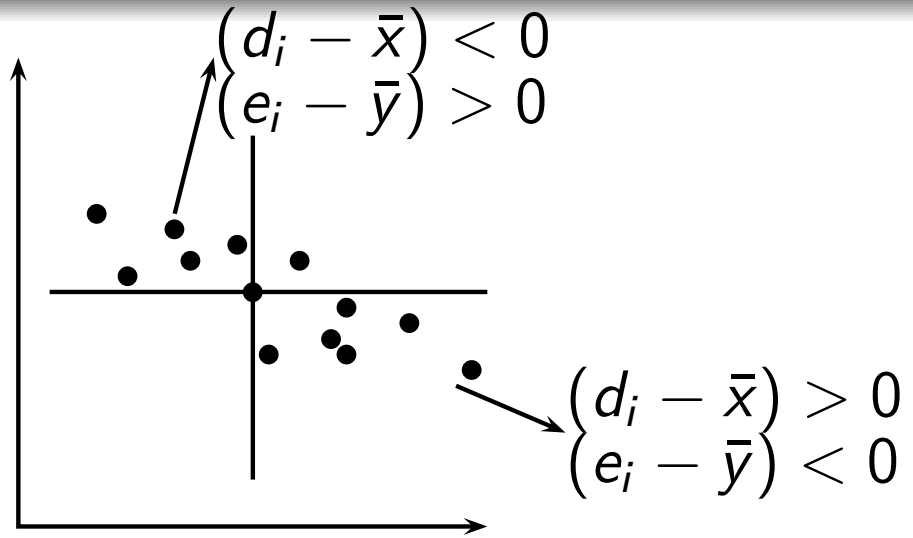

$(d_i - \bar{x}) > 0$



$$\begin{pmatrix} d_i - \bar{x} \end{pmatrix} > 0$$
$$\begin{pmatrix} e_i - \bar{y} \end{pmatrix} < 0$$

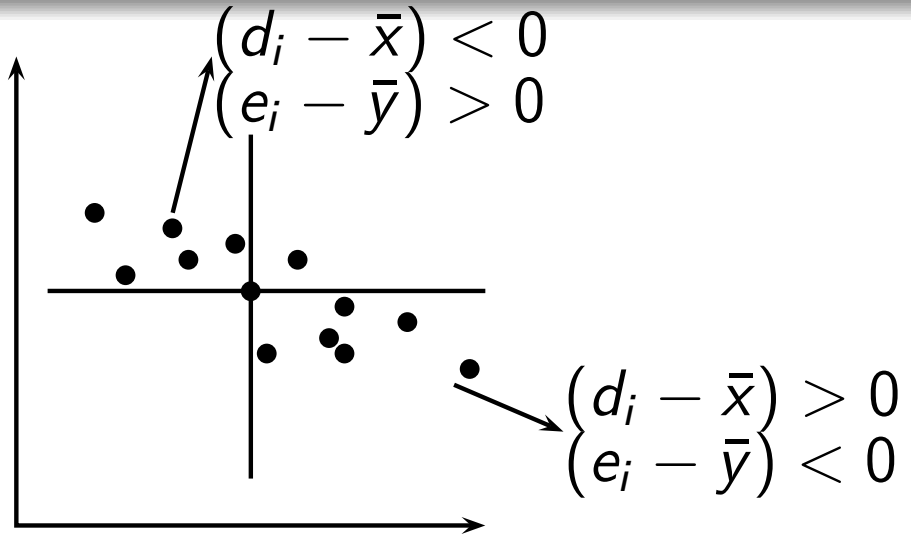

$(d_i - \bar{x}) < 0$
$(d_i - \bar{x}) > 0$
$(e_i - \bar{y}) < 0$


$(d_i - \bar{x}) < 0$
$(e_i - \bar{y}) > 0$
$(d_i - \bar{x}) > 0$
$(e_i - \bar{y}) < 0$



Globalement,

$$\sum_{i=1}^{n}(d_i - \bar{x})(e_i - \bar{y}) < 0$$



Globalement,

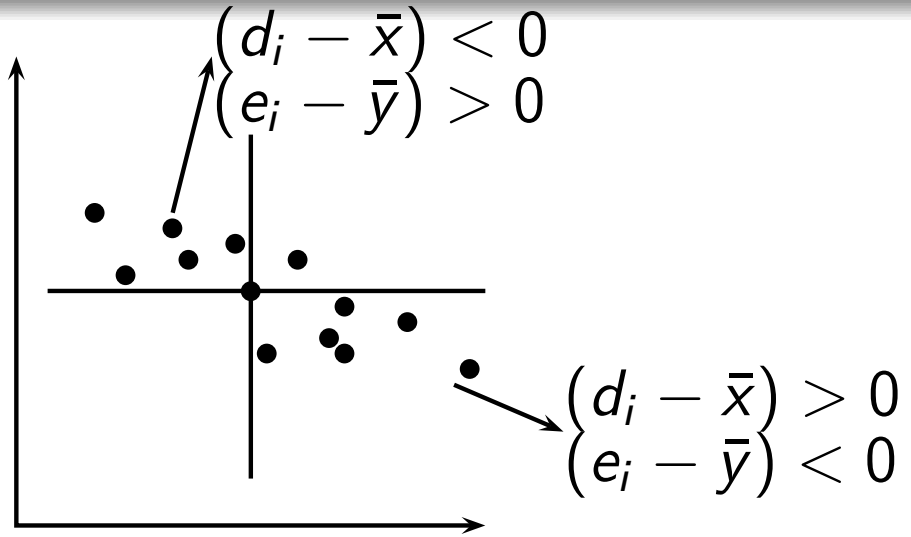
$$\sum_{i=1}^{n}(d_i - \bar{x})(e_i - \bar{y}) < 0$$

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(d_i - \bar{x})(e_i - \bar{y})$$

et on a

$$|\sigma_{xy}| \leq \sigma_x \sigma_y$$

Dès lors,



Globalement,

$$\sum_{i=1}^{n}(d_i - \bar{x})(e_i - \bar{y}) < 0$$

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(d_i - \bar{x})(e_i - \bar{y})$$

et on a

$$|\sigma_{xy}| \leq \sigma_x \sigma_y$$

Dès lors,

$$-1 \leq \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \leq 1$$

Tableau de contingence

Le tableau ci-dessous décrit la répartition d'une population constituée de 535 ménages selon les deux variables suivantes : X représente le nombre de pièces de l'habitation et Y correspond au nombre d'enfants du ménage.

Nombre de pièces	Nombre d'enfants				
	0	1	2	3	4
1	7	3	2	1	0
2	24	32	21	2	1
3	16	35	54	26	4
4	9	28	74	55	12
5	4	12	46	13	12
6	2	6	16	11	7

Nombre de pièces	Nombre d'enfants				
	0	1	2	3	4
1	7	3	2	1	0
2	24	32	21	2	1
3	16	35	54	26	4
4	9	28	74	55	12
5	4	12	46	13	12
6	2	6	16	11	7

$\sigma_{xy} =$

Nombre de pièces	Nombre d'enfants				
	0	1	2	3	4
1	7	3	2	1	0
2	24	32	21	2	1
3	16	35	54	26	4
4	9	28	74	55	12
5	4	12	46	13	12
6	2	6	16	11	7

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} n_{ij}(x_i - \bar{x})(y_j - \bar{y})$$

Probabilités

Exemple 1 : Jet de dés

- Jet d'un dé
  Résultat suit une loi uniforme discrète de paramètre $n = 6$,

Exemple 1 : Jet de dés

- Jet d'un dé
  Résultat suit une loi uniforme discrète de paramètre $n = 6$, soit X la variable « Résultat du dé », on écrit $X \sim U(6)$.

Exemple 1 : Jet de dés

- Jet d'un dé
  Résultat suit une loi uniforme discrète de paramètre $n = 6$, soit X la variable « Résultat du dé », on écrit $X \sim U(6)$.

k	1	2	3	4	5	6
p_k	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Exemple 1 : Jet de dés

- Jet d'un dé

  Résultat suit une loi uniforme discrète de paramètre $n = 6$, soit X la variable « Résultat du dé », on écrit $X \sim U(6)$.

k	1	2	3	4	5	6
p_k	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

- Jet de deux dés et somme obtenue

Exemple 1 : Jet de dés

- Jet d'un dé
  Résultat suit une loi uniforme discrète de paramètre $n = 6$, soit X la variable « Résultat du dé », on écrit $X \sim U(6)$.

k	1	2	3	4	5	6
p_k	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

- Jet de deux dés et somme obtenue
  X = somme des deux résultats = somme de deux variables de même loi (théorème central limite)

Exemple 1 : Jet de dés

- Jet d'un dé
  Résultat suit une loi uniforme discrète de paramètre $n = 6$, soit X la variable « Résultat du dé », on écrit $X \sim U(6)$.

k	1	2	3	4	5	6
p_k	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

- Jet de deux dés et somme obtenue
  X = somme des deux résultats = somme de deux variables de même loi (théorème central limite)

k	2	3	4	5	6	7	8	9	10	11	12
p_k	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Excel et l'aléatoire

Excel et l'aléatoire

- =ALEA() génère un nombre aléatoire entre 0 et 1 :

$$0 \leq \text{ ALEA() } < 1$$

Excel et l'aléatoire

- =ALEA() génère un nombre aléatoire entre 0 et 1 :

$$0 \leq \text{ ALEA() } < 1$$

- =ENT(*nombre*) renvoie la partie entière de *nombre*

Excel et l'aléatoire

- =ALEA() génère un nombre aléatoire entre 0 et 1 :

$$0 \leq \text{ ALEA() } < 1$$

- =ENT(*nombre*) renvoie la partie entière de *nombre*
- =ALEA.ENTRE.BORNES(m ;n) renvoie un nombre entier aléatoire entre m et n (compris tous les deux)

Les fonctions d'Excel pour le calcul des probabilités

Les fonctions d'Excel pour le calcul des probabilités

Soit $X \sim B(n, p)$

- LOI.BINOMIALE.N(x,p,n,0) renvoie $P(X = x)$
- LOI.BINOMIALE.N(x,p,n,1) renvoie $P(X \leq x)$
- LOI.BINOMIALE.INVERSE(n,p,α) renvoie x tel que $P(X \leq x) = \alpha$

Soit $X \sim N(\mu, \sigma)$

- LOI.NORMALE.N(x,μ,σ,1) renvoie $P(X \leq x)$
- LOI.NORMALE.INVERSE.N(α,μ,σ) renvoie x tel que $P(X \leq x) = \alpha$

Une étude publiée par des chercheurs de l'Université de Montréal en 2002[1] à propos de l'influence des pesticides sur le rapport garçons/filles à la naissance a été menée dans la ville d'Ufa (fédération de Russie) auprès de 198 personnes (150 hommes et 48 femmes) ayant été exposés, dans une usine agrochimique de 1961 à 1988, à des pesticides contenant de la dioxine.
Le rapport garçons/filles à la naissance pour cette ville est de 0,512.[2]
Sur la descendance des personnes étudiées, on observe 91 garçons et 136 filles, soit une fréquence observée de 0,4 garçons.

---

1. Sex Ratios of Children of Russian Pesticide Producers Exposed to Dioxin, *Environmental Health*, novembre 2002.
2. Ce rapport est reconnu comme valable dans le monde.

Simulation et théorie de l'échantillonnage

Simulation et théorie de l'échantillonnage

- Principe : simuler une population avec les mêmes caractéristiques que l'échantillon étudié et voir si la fréquence observée a des chances d'apparaître
- Caractéristiques : $n = 227$ et $p = 0.512$
- Code : 1 = garçon, 0 = fille