

## Traitement de données – statistiques

Marie-Camille CAUMON  
Ingénieur de recherche  
GeoRessources - UMR 7359  
Entrée 3B - bureau A508  
+33 3 72 74 55 37

marie-camille.caumon@univ-lorraine.fr  
<http://georessources.univ-lorraine.fr/>



## S7-4 Traitement des données en Géosciences

### *Traitement de données – statistiques*

1 CM 3h

2 TP 4h en salle info

1 CM 3h

2 TP 4h en salle info

1 CC (TP 3)

1 contrôle terminal

# Objectifs et méthodes

- Utiliser de la manière la plus pratique possible un tableur (type EXCEL)
- Traitements statistiques de base
- Utilisation de fonctions spécifiques aux statistiques
- Analyses statistiques et factorielles sur études de cas



Traiter une population statistique de manière rigoureuse

Savoir interpréter les représentations graphiques issues du traitement statistique



Fonctions statistiques  
Tests statistiques  
Représentations

# Prérequis

- Bases de l'utilisation d'un tableur (type EXCEL)
  - Notions de variable, effectif, paramètres de position et dispersion
  - Représentations graphiques : histogrammes
  - Régression linéaire simple
- 
- Révisions rapides en CM
  - Exercices corrigés disponibles sur Arche
  - Utilisation des outils avancés d'Excel
  - Utilisation de R ( R, RStudio, packages Rmcd, FactoMineR, cluster, lattice)

# Plan du cours – partie I

1. Vocabulaire
2. Variables ou caractères
  1. Vocabulaire
  2. Notion de distribution
3. Grandeurs statistiques usuelles
  1. Paramètres de position
  2. Paramètres de dispersion
4. Représentations graphiques
5. Lois de distribution usuelles
6. Statistiques bivariées
  1. Représentation graphique
  2. Covariance
  3. Régression linéaire

# 1. Vocabulaire

Unité statistique



Population



Échantillon



Taille de l'échantillon =

Taille de la population =

Taux de sondage =





Unité statistique

Population

Échantillon

Taille de l'échantillon = 16

Taille de la population = 100

Taux de sondage = 16 %



## Vocabulaire :

- **Unité** statistique  
= individu  
= élément
- **Population**  
= ensemble statistique
- **Échantillon**
- **Taille** de la population ou de l'échantillon
- **Taux de sondage**

## Notations:

- |                            |   |                  |
|----------------------------|---|------------------|
| • individu ou observations | → | $i$              |
| • population               | → | $P$              |
| • échantillon              | → | $E$              |
| • taille de la population  | → | $N$              |
| • taille de l'échantillon  | → | $n$              |
| • taux de sondage          | → | $n / N$          |
| • variables                | → | $X, Y, Z, \dots$ |



# Plan du cours – partie I




1. Vocabulaire
2. Variables ou caractères
  1. Vocabulaire
  2. Notion de distribution
3. Grandeurs statistiques usuelles
  1. Paramètres de position
  2. Paramètres de dispersion
4. Représentations graphiques
5. Lois de distribution usuelles
6. Statistiques bivariées
  1. Représentation graphique
  2. Covariance
  3. Régression linéaire



Types de variable : qualitative nominale/ordinaire  
quantitative discrète/continue

1. Teneur en nitrate d'une eau minérale
2. Potabilité d'une eau
3. Nombre d'animaux dans un élevage
4. Coordonnées GPS d'une population (échantillons)
5. Porosité d'un réservoir
6. Occurrences d'un minéral dans une section polie
7. La saison à laquelle le prélèvement d'échantillons est effectué
8. Niveau de confort sonore d'une population à proximité d'une éolienne
9. Notes / 20 des étudiants d'une promo de M1 à un contrôle
10. La moyenne générale des étudiants d'une promo en fin de M1



Quantitative		Catégorielle	
continue	discrète	nominale	ordinaire
			



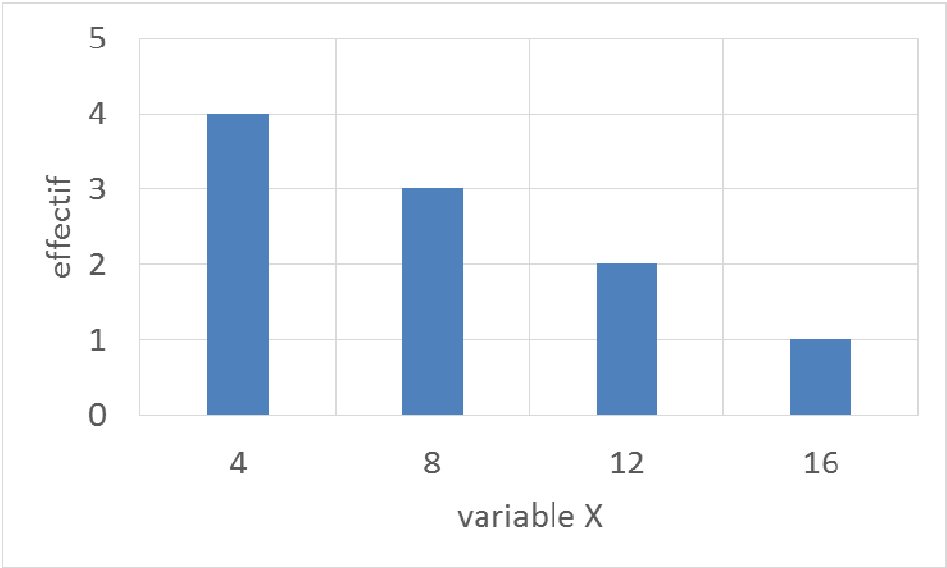
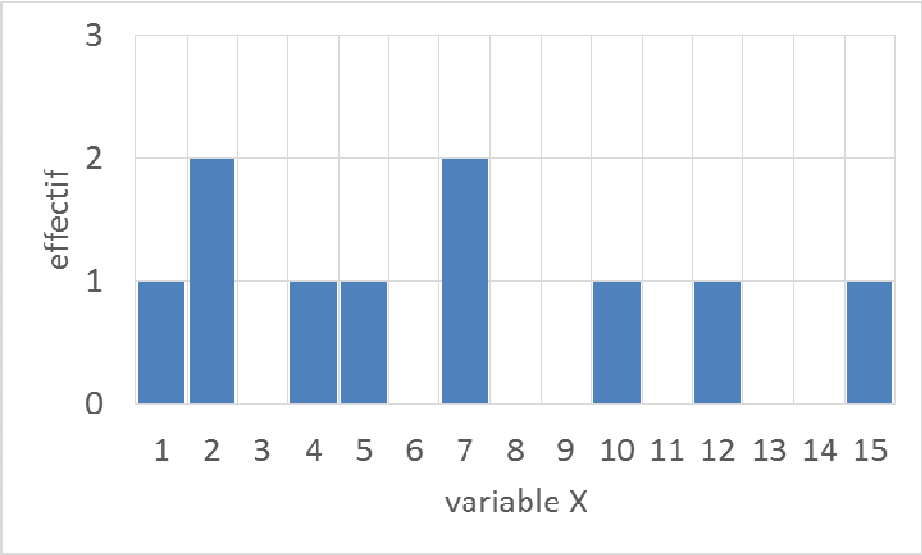
Types de variable : qualitative nominale/ordinaire  
quantitative discrète/continue

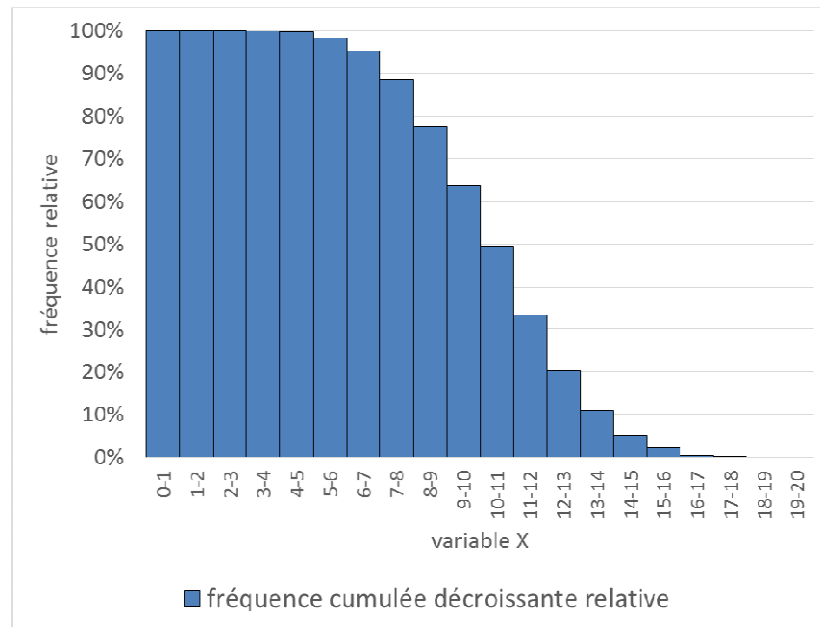
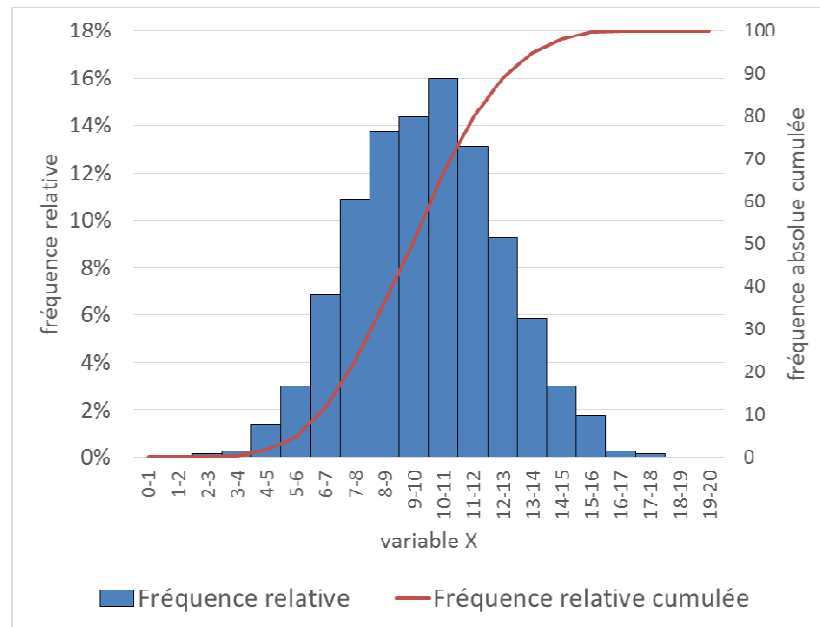
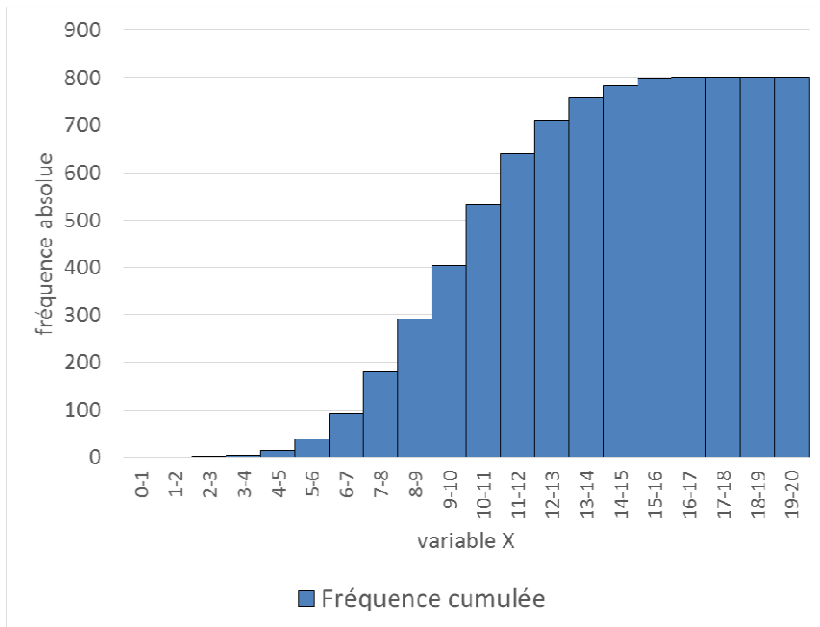
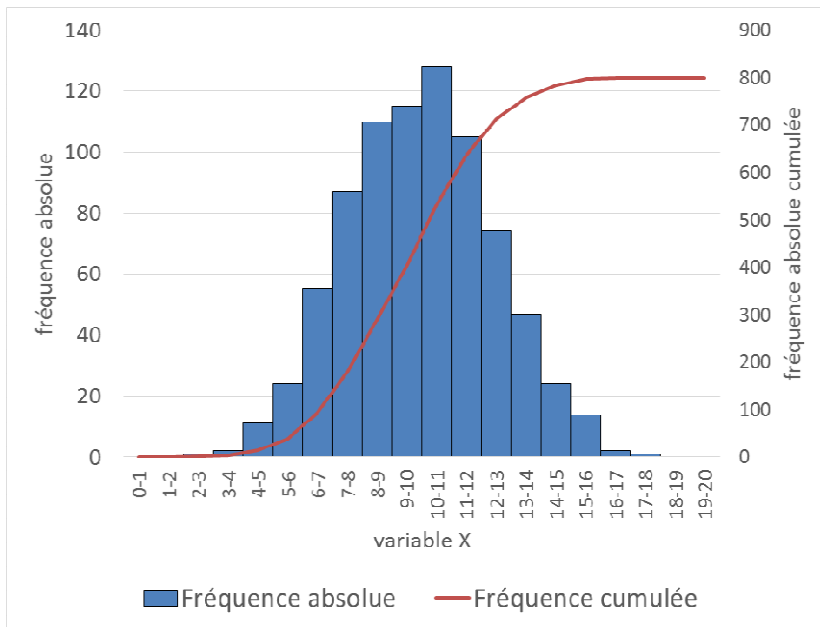
1. Teneur en nitrate d'une eau minérale
2. Potabilité d'une eau
3. Nombre d'animaux dans un élevage
4. Coordonnées GPS d'une population (échantillons)
5. Porosité d'un réservoir
6. Occurrences d'un minéral dans une section polie
7. La saison à laquelle le prélèvement d'échantillons est effectué
8. Niveau de confort sonore d'une population à proximité d'une éolienne
9. Notes / 20 des étudiants d'une promo de M1 à un contrôle
10. La moyenne générale des étudiants d'une promo en fin de M1

Quantitative		Catégorielle	
continue	discrète	nominale	ordinaire
1-5-10	3-6-9	2-7	4-8

## 2. Variables ou caractères : notion de distribution

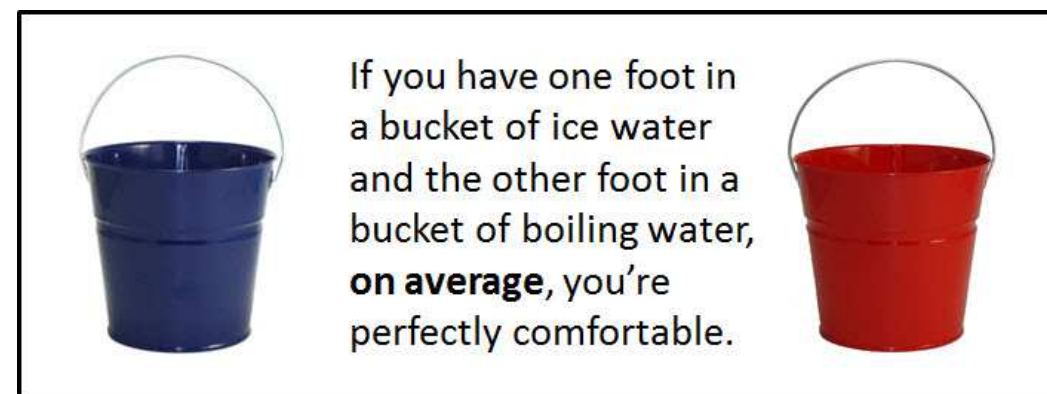
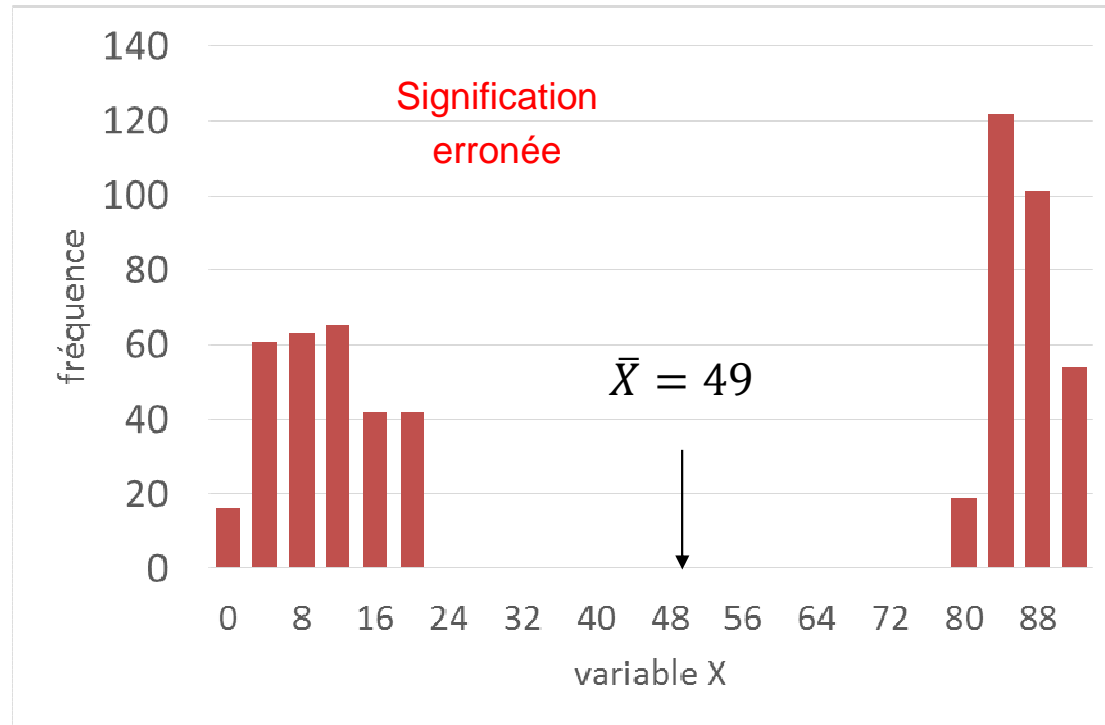
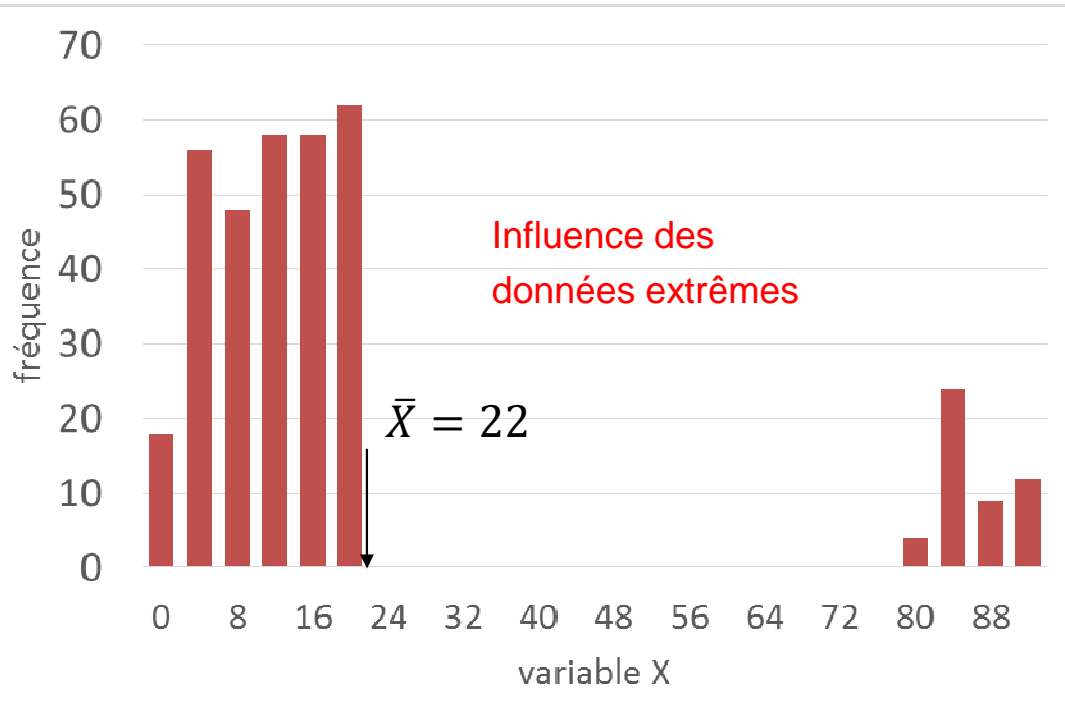
N° échantillon	Variable X
1	1
2	4
3	10
4	7
5	2
6	2
7	7
8	12
9	5
10	15



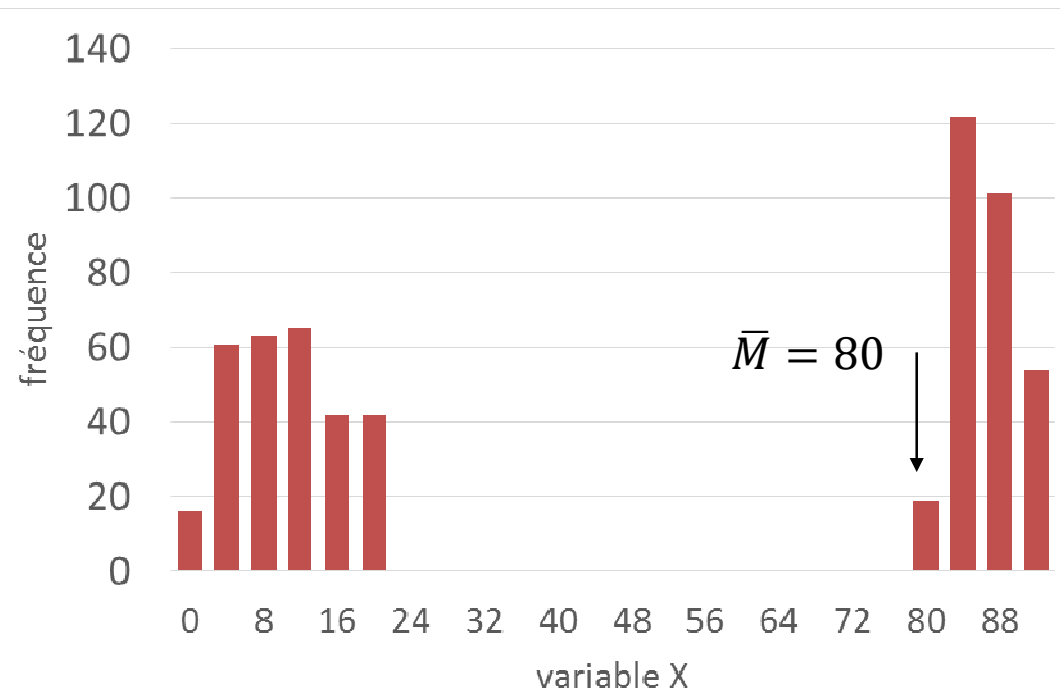
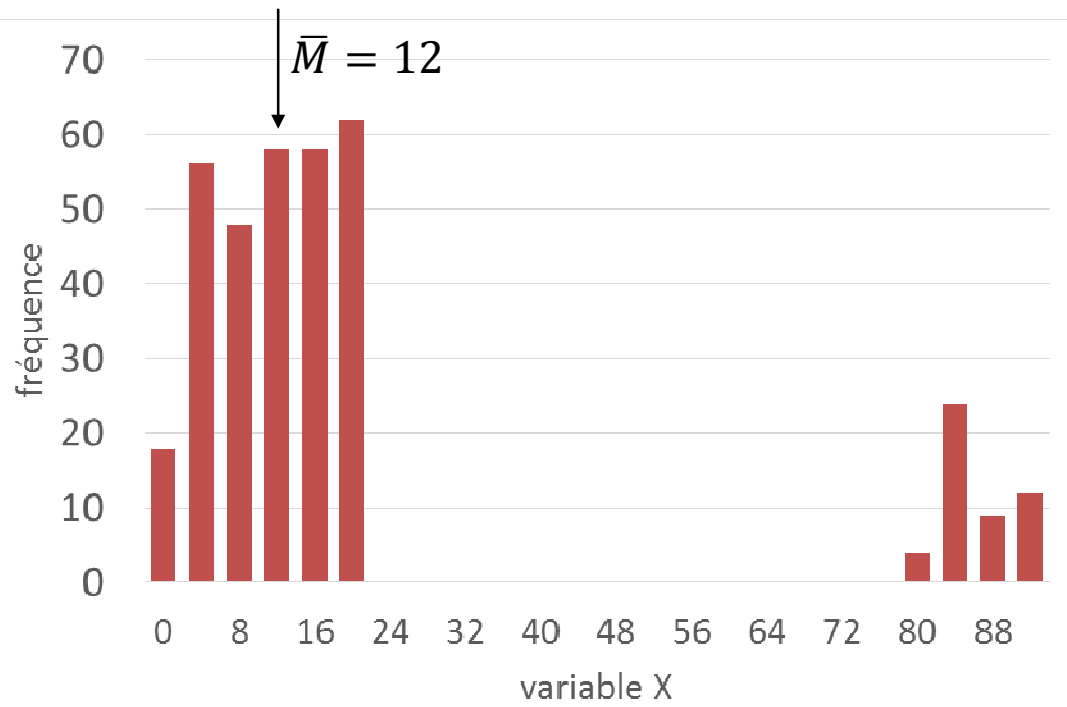


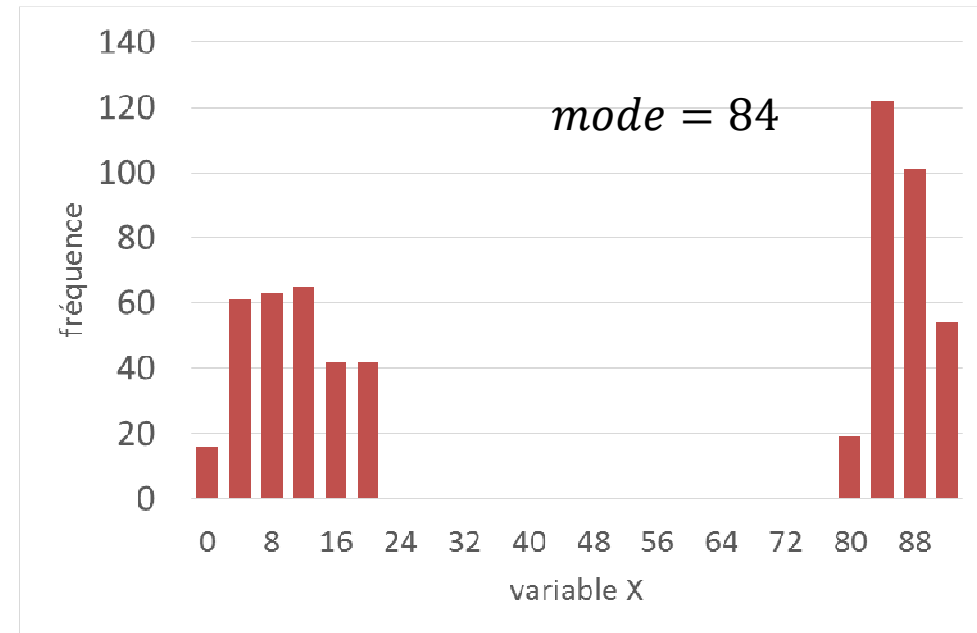
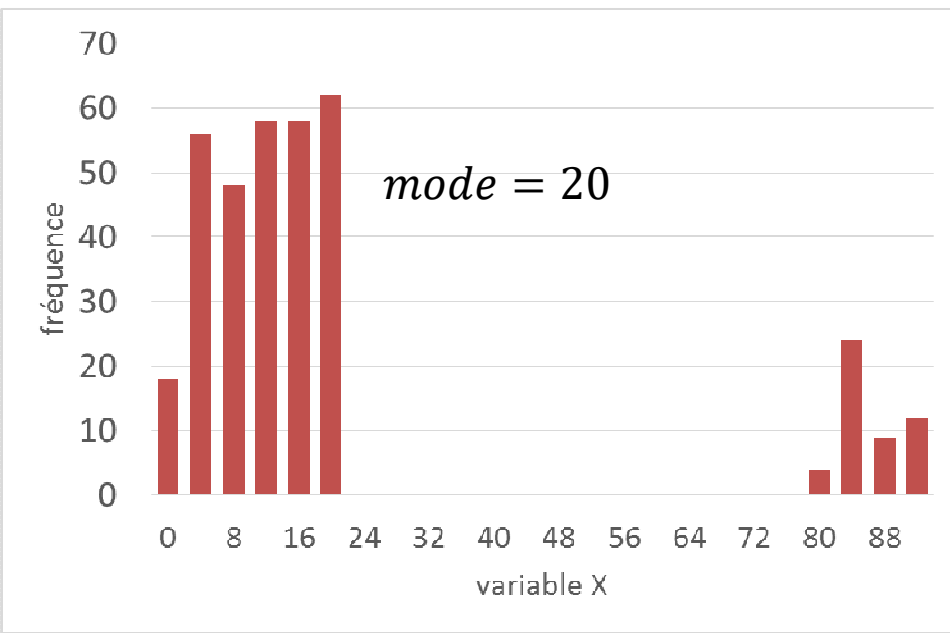
# Plan du cours – partie I

1. Vocabulaire
2. Variables ou caractères
  1. Vocabulaire
  2. Notion de distribution
3. Grandeurs statistiques usuelles
  1. Paramètres de position
  2. Paramètres de dispersion
4. Représentations graphiques
5. Lois de distribution usuelles
6. Statistiques bivariées
  1. Représentation graphique
  2. Covariance
  3. Régression linéaire

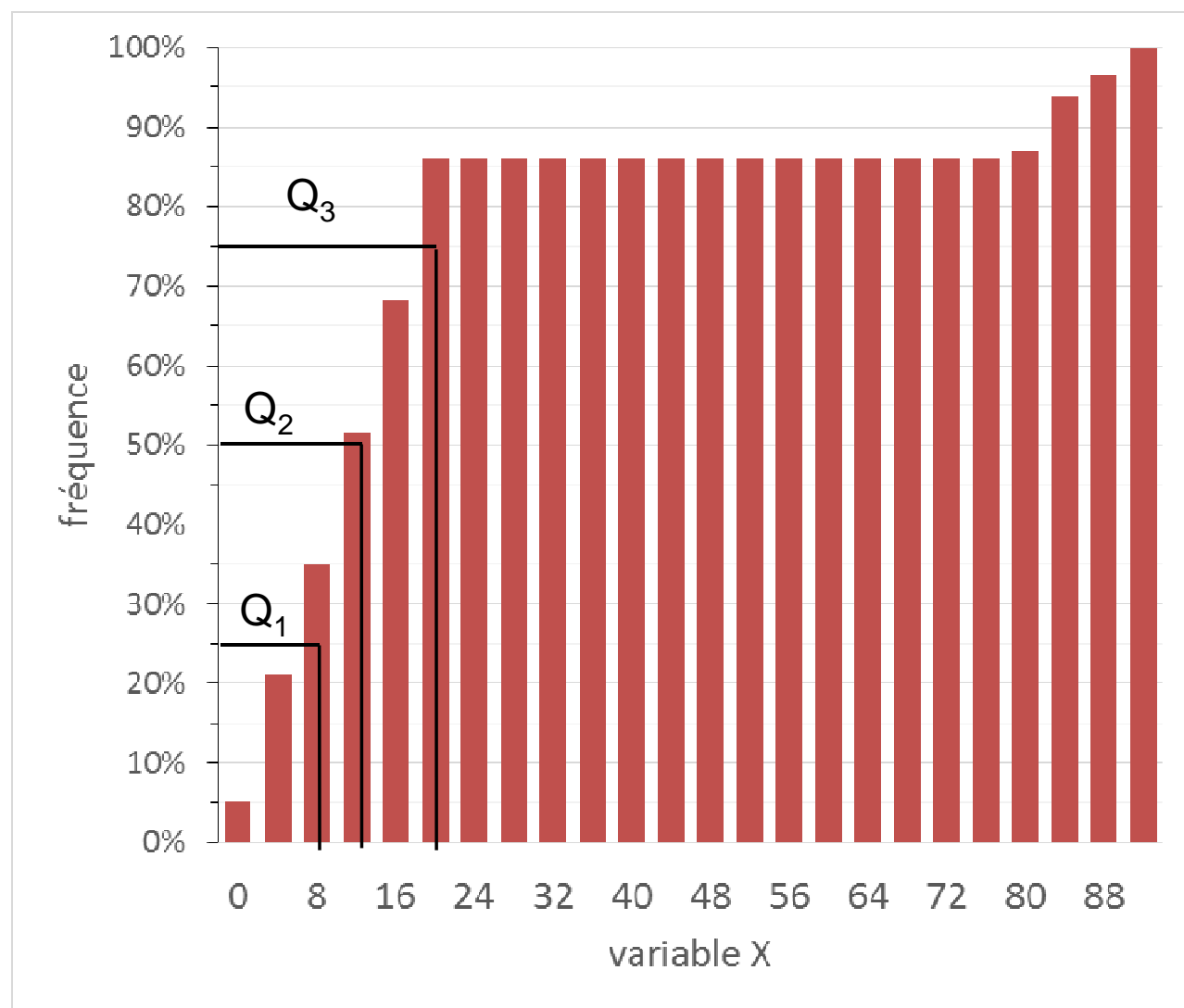
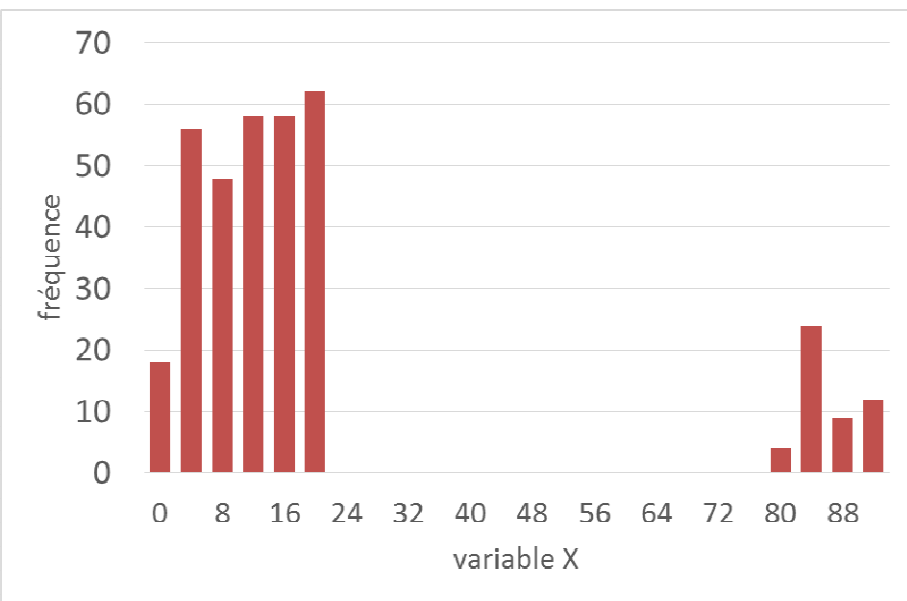








Le mode d'une distribution de variables **quantitatives continues** peut-il être déterminé ?



## Formules EXCEL pour les paramètres de position

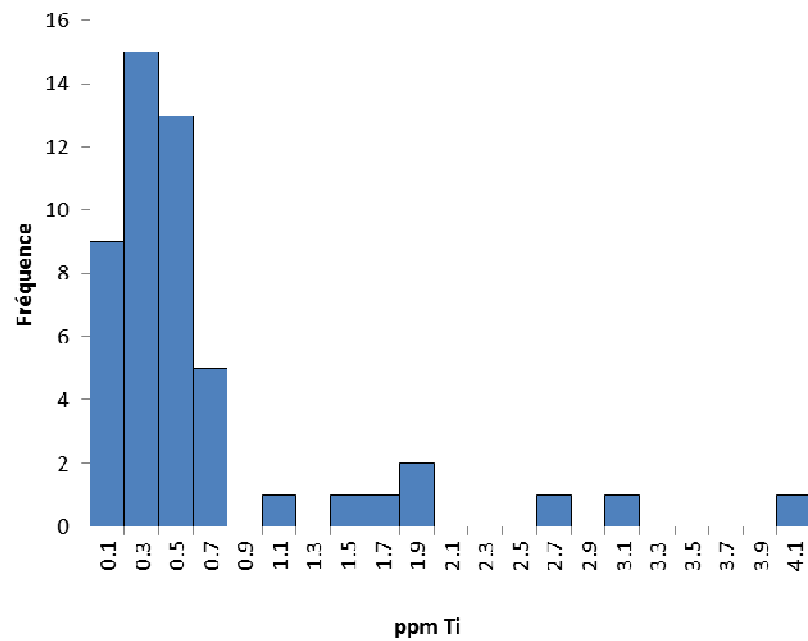
Paramètre	Formule	Arguments
Moyenne	= MOYENNE (matrice)	Tableau de données
Médiane	= MEDIANE (matrice)	Tableau de données
Mode	= MODE (matrice)	Tableau de données
Centile	= CENTILE.INCLURE (matrice;k)	Tableau de données;nb <u>quelconque</u> entre 0 et 1
Décile	= DECILE.INCLURE (matrice;k)	Tableau de données;nb <u>entier</u> entre 1 et 10
Quartile	= QUARTILE.INCLURE (matrice;k)	Tableau de données;nb <u>entier</u> entre 1 et 4

## Fonctions R pour les paramètres de position

Paramètre	Fonction	Arguments
Moyenne	= mean(x)	Tableau de données
Médiane	= median(x)	Tableau de données
Mode	= mode(x)	Tableau de données
Quantile	= quantile(x, probs)	Tableau de données, nb quelconque entre 0 et 1

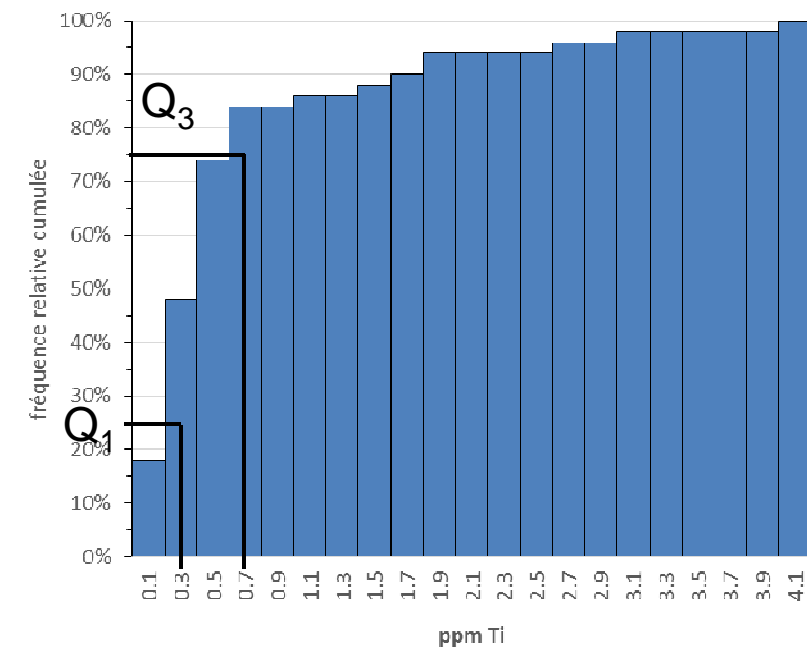


### teneur en Ti



Moyenne	0.67
Médiane	0.43
Mode	0.34
Écart-type	0.81
Variance de l'échantillon	0.65
Plage	4
Minimum	0.1
Maximum	4.1
Nombre d'échantillons	50
Q1	0.27
Q3	0.61
distance interquartile	0.34
coefficient de variation	1.20

### teneur en Ti



## Formules EXCEL pour les paramètres de dispersion

Paramètre	Formule	Arguments
Étendue	=MAX(matrice)-MIN(matrice)	Tableau de données
Variance	=VAR.S(matrice)	Tableau de données
Écart-type	=ECARTYPE.STANDARD(matrice)	Tableau de données



## Fonctions R pour les paramètres de dispersion

Paramètre	Fonction	Arguments
Étendue	<code>= max(x)-min(x)</code>	Tableau de données
Variance	<code>= var(x)</code>	Tableau de données
Écart-type	<code>= sd(x)</code>	Tableau de données

## Formules à mémoriser

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

*Moyenne arithmétique*

$$\text{Médiane} = Q_2 = C_{50}$$

$$I_Q = |Q_3 - Q_1|$$

*Intervalle inter-quartile*

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

*Variance*

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

$$CV = \frac{\sigma}{\bar{x}}$$

*Coefficient de variation*

# Plan du cours – partie I

1. Vocabulaire
2. Variables ou caractères
  1. Vocabulaire
  2. Notion de distribution
3. Grandeurs statistiques usuelles
  1. Paramètres de position
  2. Paramètres de dispersion
- 4. Représentations graphiques**
5. Lois de distribution usuelles
6. Statistiques bivariées
  1. Représentation graphique
  2. Covariance
  3. Régression linéaire

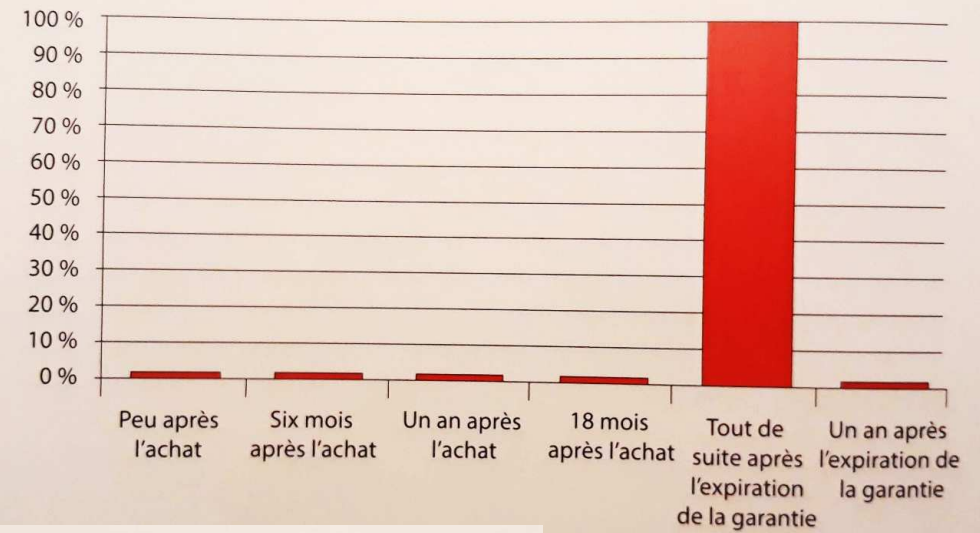


Ce que je fais en attendant que les toasts soient grillés

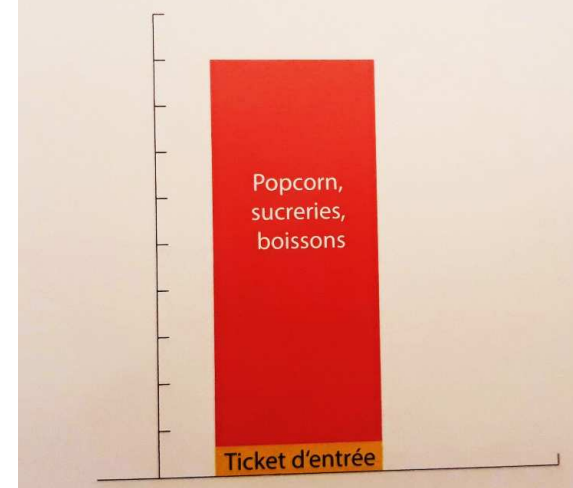


Diagramme en secteur ou circulaire

Quand les appareils tombent-ils en panne ?



Coût d'une sortie au cinéma



Histogrammes

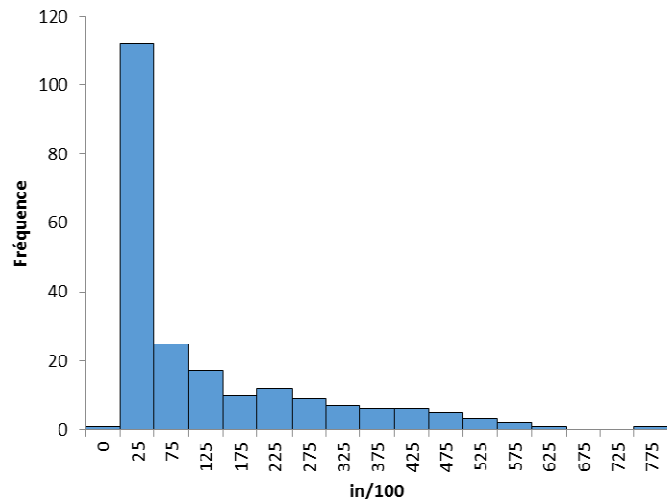
Ti / ppm				
1.15	0.73	0.44	0.14	1.78
1.87	0.54	0.36	0.14	0.3
1.84	0.52	0.26	0.34	1.56
0.47	0.45	0.23	0.36	0.26
0.34	0.46	0.23	0.3	0.61
2.78	0.4	0.13	0.1	0.45
3.04	0.36	0.12	0.12	0.7
0.16	0.34	0.12	0.61	0.59
0.42	0.32	0.11	0.6	0.4
0.45	0.5	0.63	0.46	4.1

Borne sup. des classes	Fréquence
0.5	33
1	9
1.5	1
2	4
2.5	0
3	1
3.5	1
4	0
4.5	1

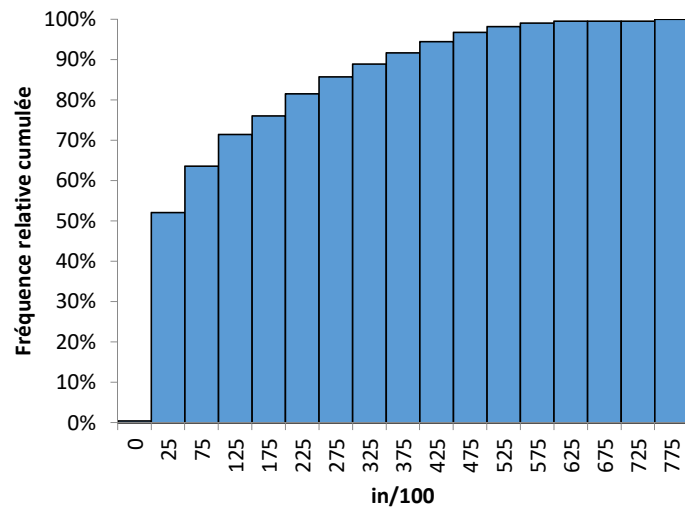
Borne sup. des classes	Fréquence
1	42
2	5
3	1
4	1
ou plus...	1

Borne sup. des classes	Fréquence
0.25	11
0.5	22
0.75	9
1	0
1.25	1
1.5	0
1.75	1
2	3
2.25	0
2.5	0
2.75	0
3	1
3.25	1
3.5	0
3.75	0
4	0
4.25	1

Histogramme

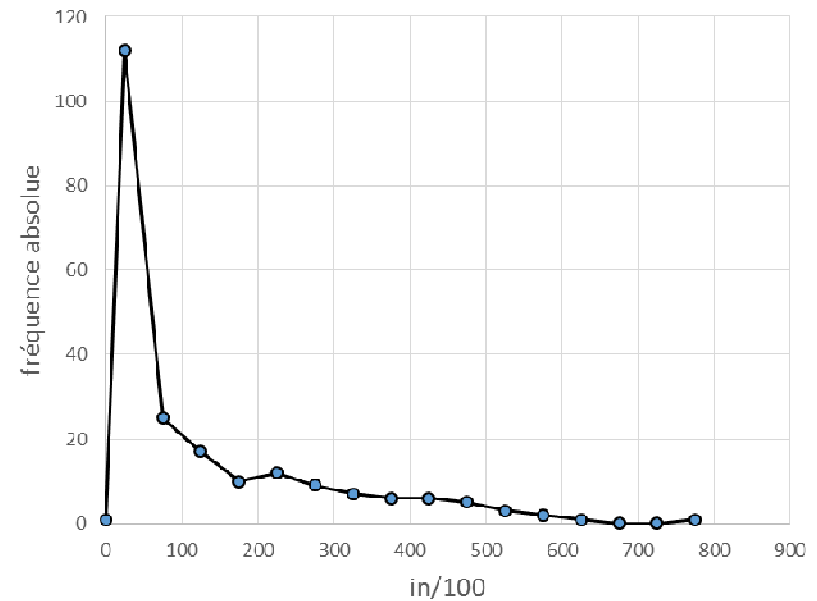


Histogramme

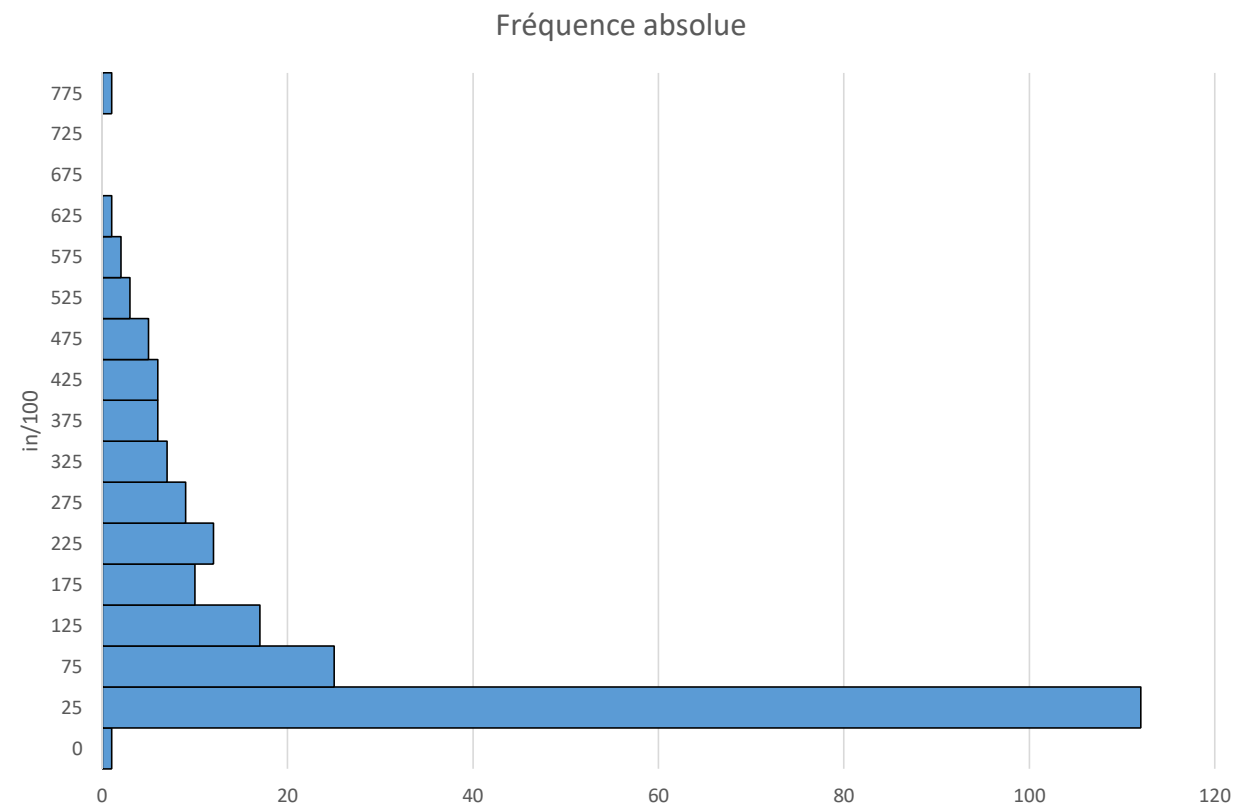


Histogramme

Polygone des fréquences

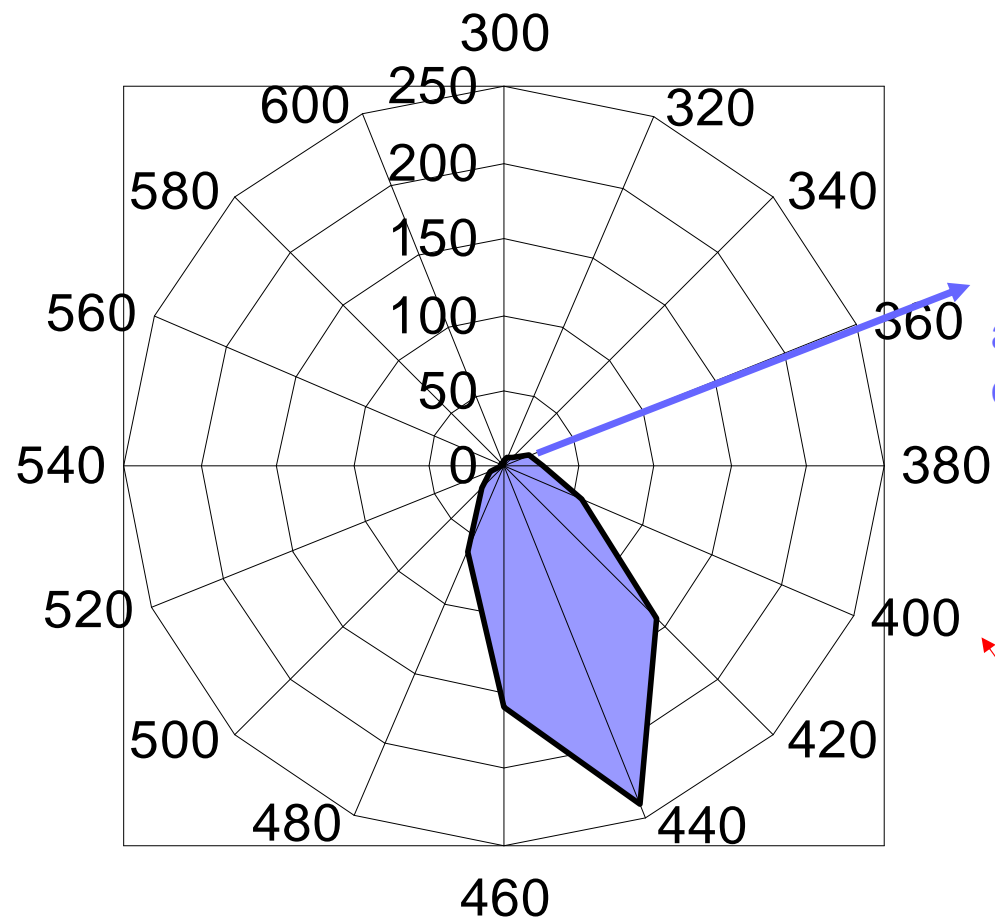


## diagrammes en barres



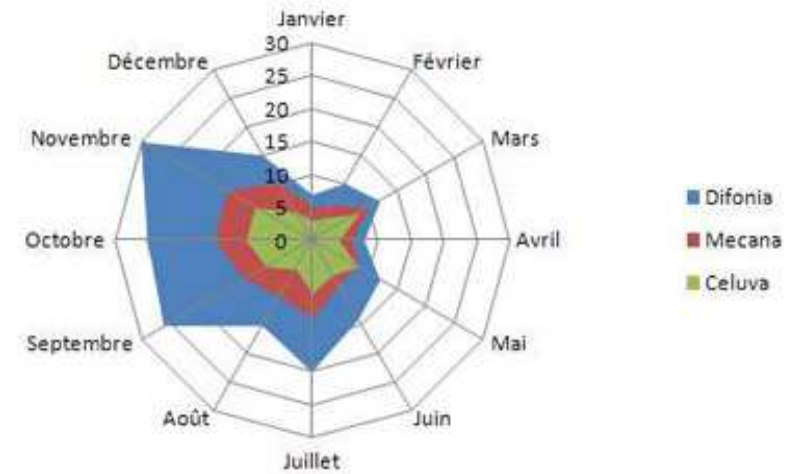


diagrammes circulaires → radars

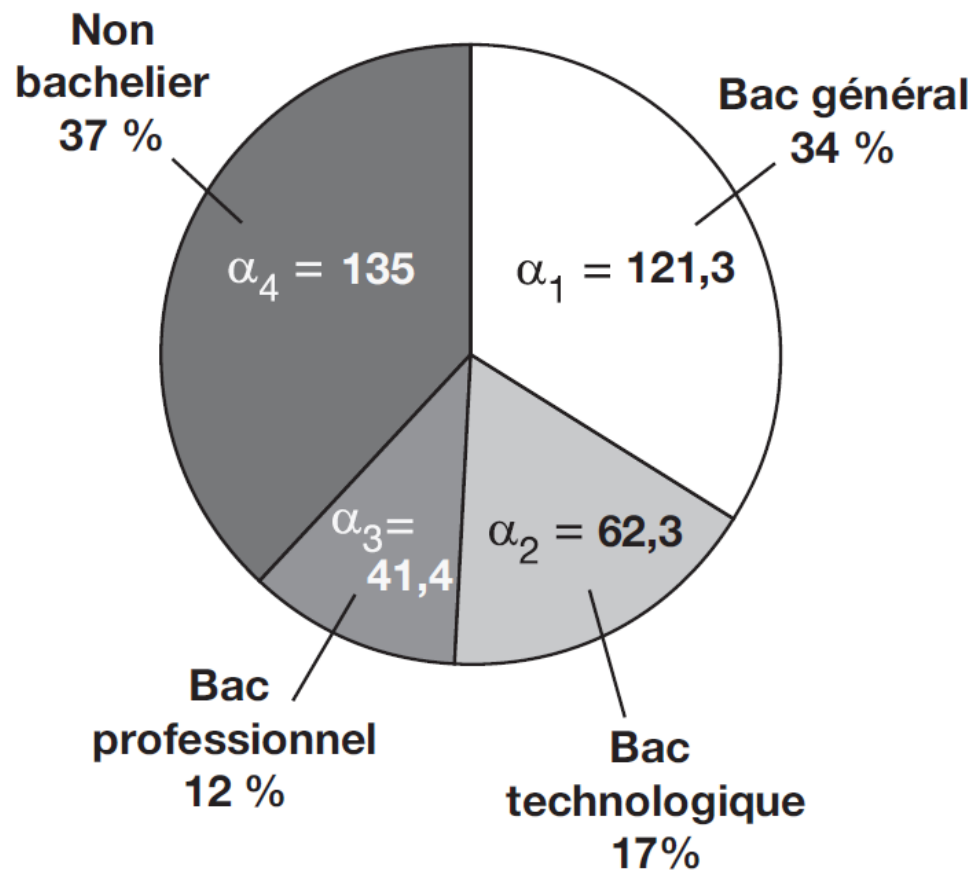


axe des fréquences -  
ordonnées

classes -  
abscisses



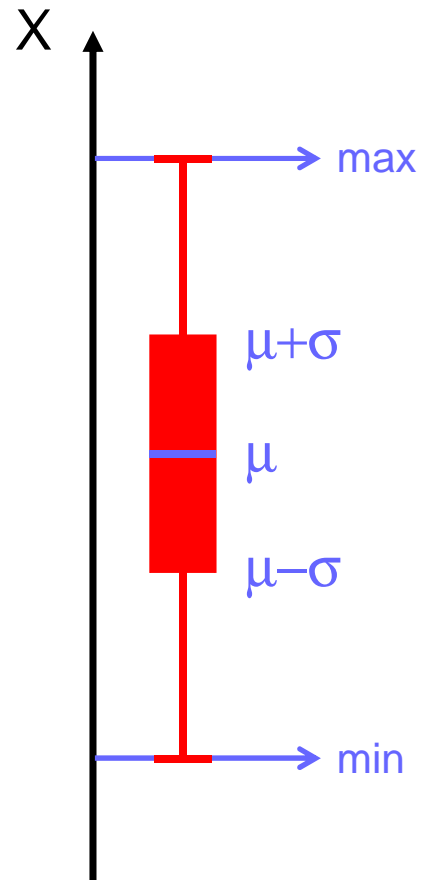
diagrammes circulaires → secteurs



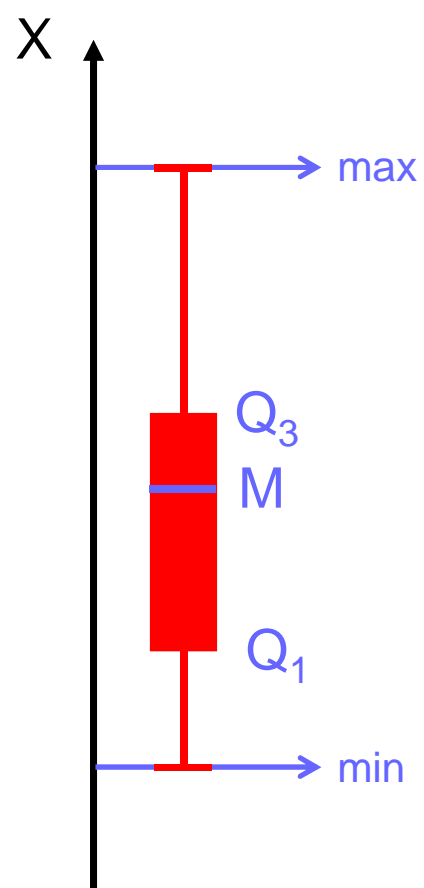
Angle du secteur :  
 $\alpha_i = f_i^{\text{relative}} \times 360$

# Boîtes à moustaches « boxplot »

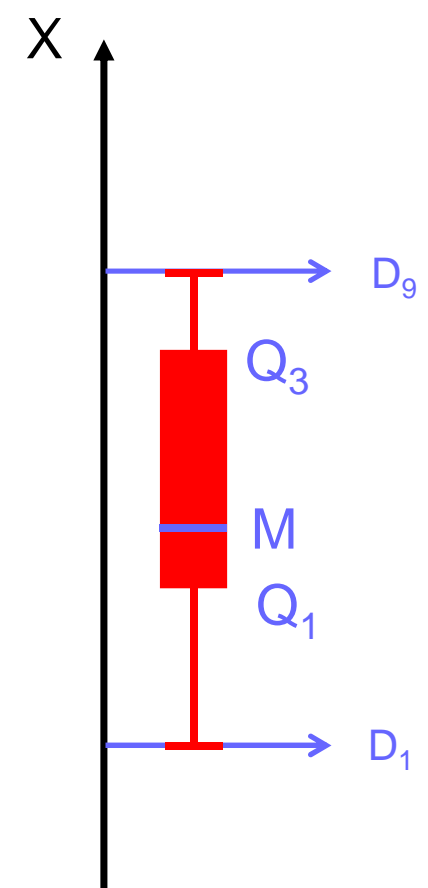
1. boîte des *écarts-types*

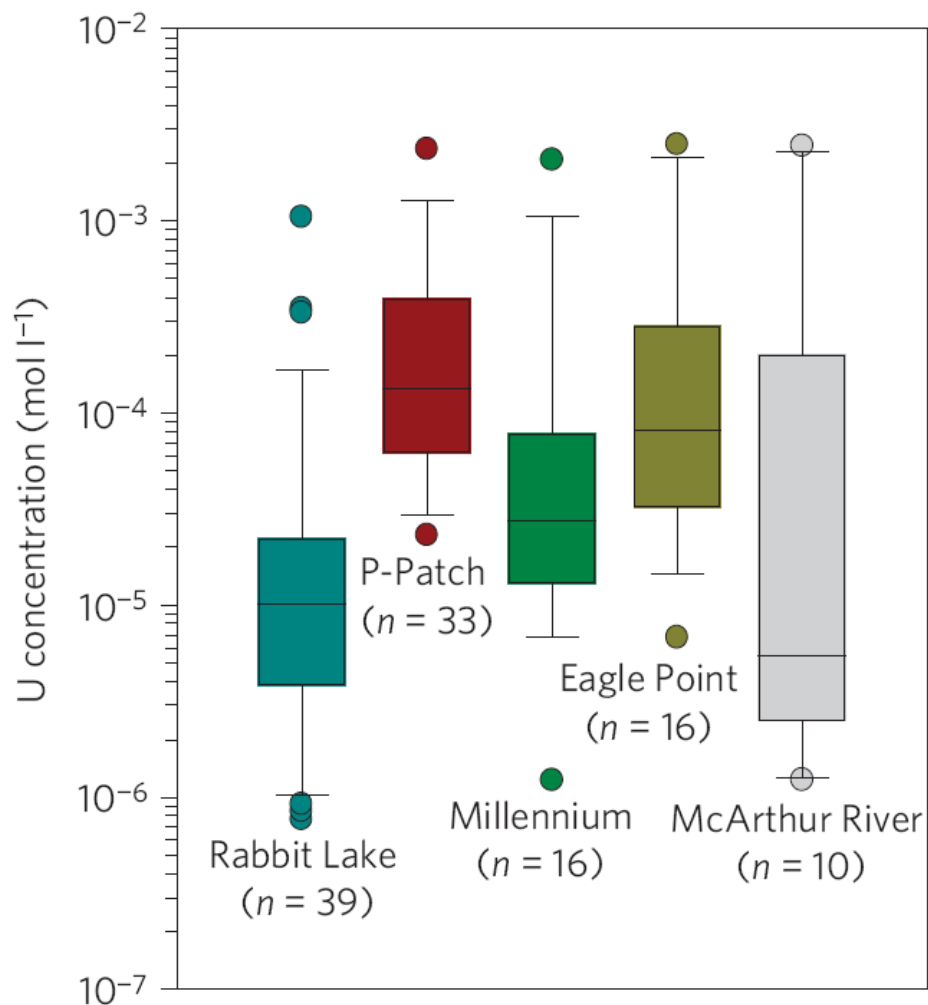


2.1 boîte de *dispersion*



2.2 boîte de *dispersion*





(Richard et al., 2012)

**Figure 2 | LA-ICP-MS determination of U concentration in fluid inclusions.**

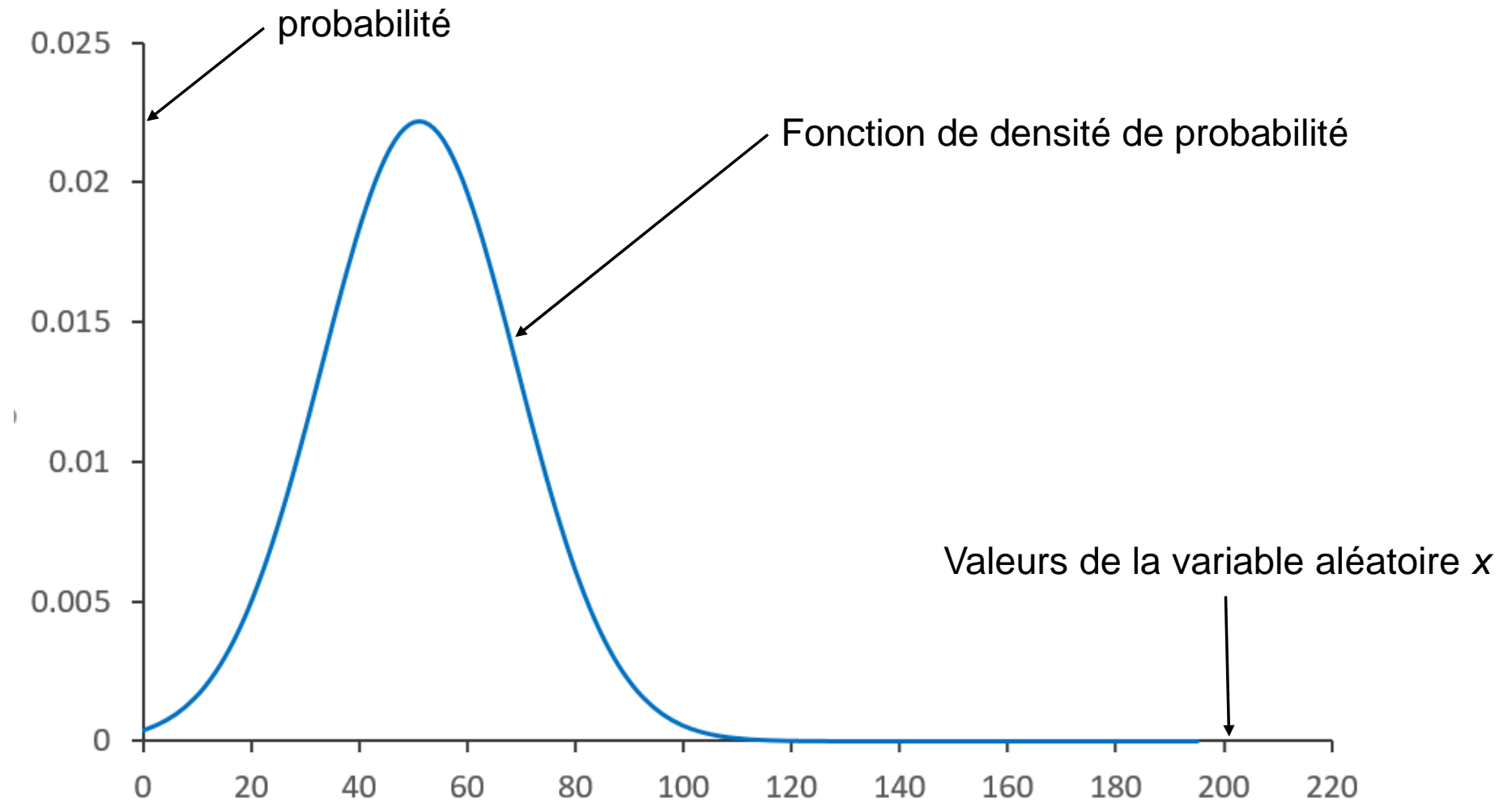
**a**, LA-ICP-MS signal for selected elements in a  $1.0 \times 10^{-3} \text{ mol l}^{-1}$  U fluid inclusion (sample RBL1Qz, Rabbit Lake deposit). U is absent from quartz (Qz) and is entirely fluid-inclusion hosted as no U signal is observed during quartz ablation before opening of the fluid inclusion (FI). a.u., arbitrary units. **b**, Box-and-whisker plots showing the distribution of U concentration in fluid inclusions among the studied deposits. Lower whiskers, bottoms of boxes, central lines, tops of boxes and upper whiskers represent 10th, 25th, 50th, 75th and 90th percentiles respectively; symbols represent outliers. McArthur River data have been published previously<sup>7</sup>. *n*, number of fluid inclusions analysed.

# Plan du cours – partie I

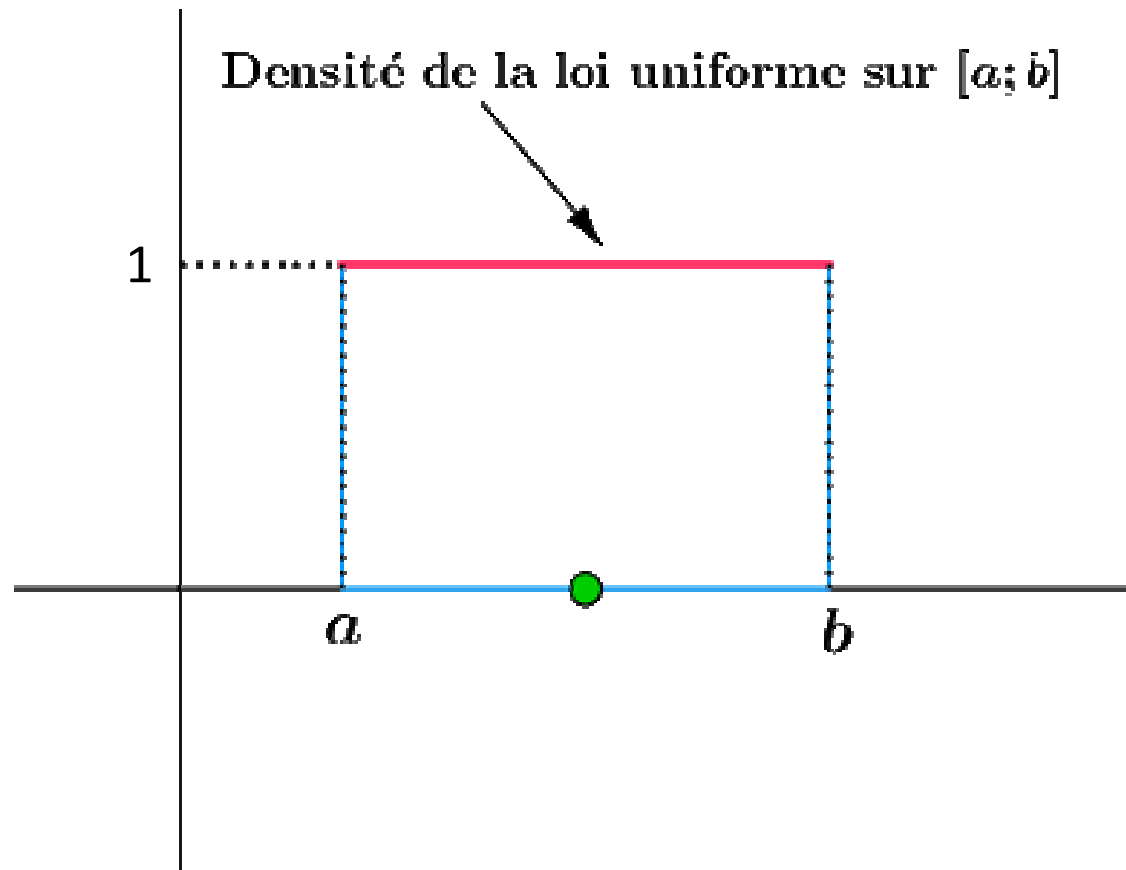
1. Vocabulaire
2. Variables ou caractères
  1. Vocabulaire
  2. Notion de distribution
3. Grandeurs statistiques usuelles
  1. Paramètres de position
  2. Paramètres de dispersion
4. Représentations graphiques
- 5. Lois de distribution usuelles**
6. Statistiques bivariées
  1. Représentation graphique
  2. Covariance
  3. Régression linéaire



## Loi de distribution de probabilité



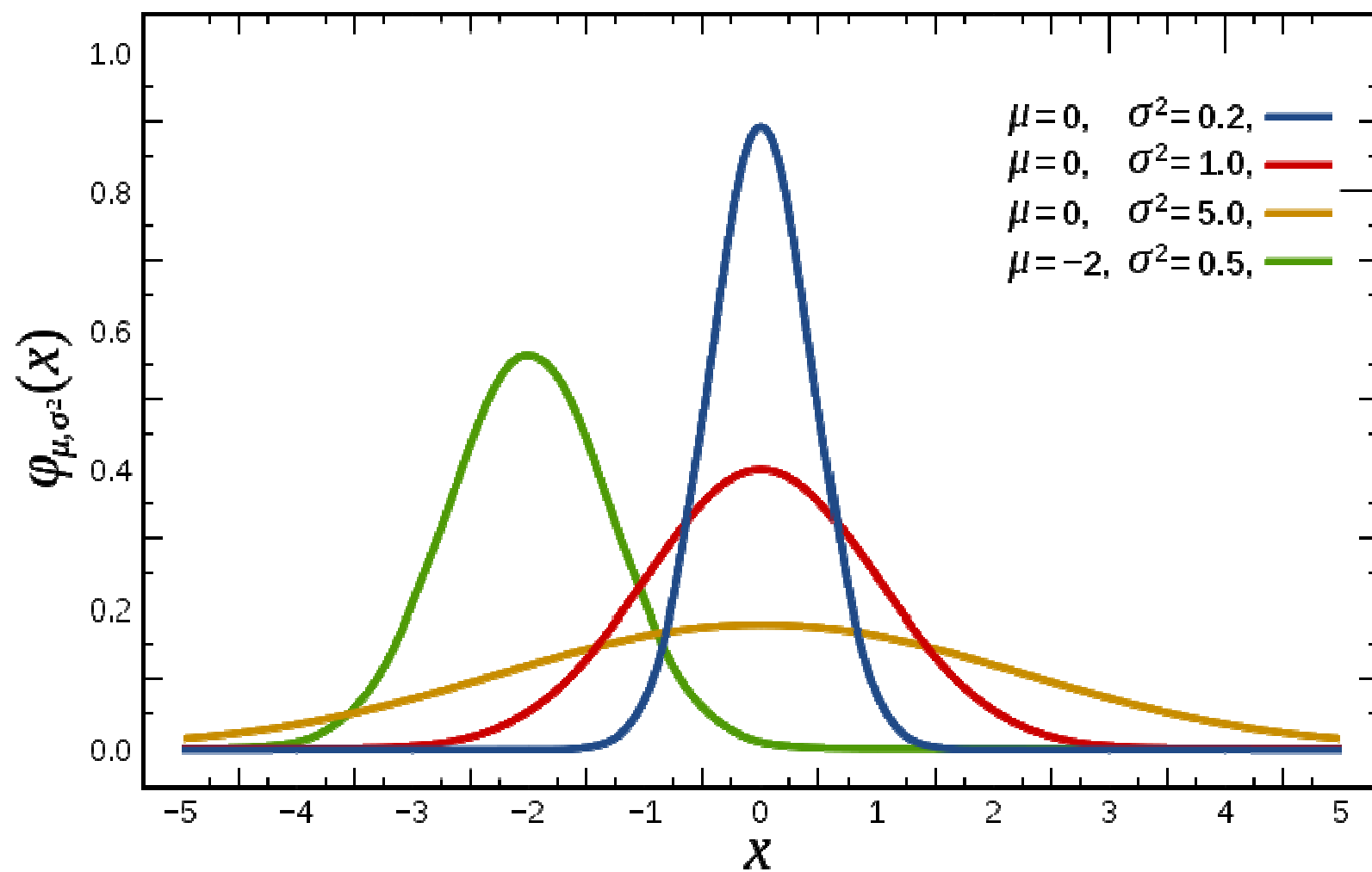
## Loi uniforme

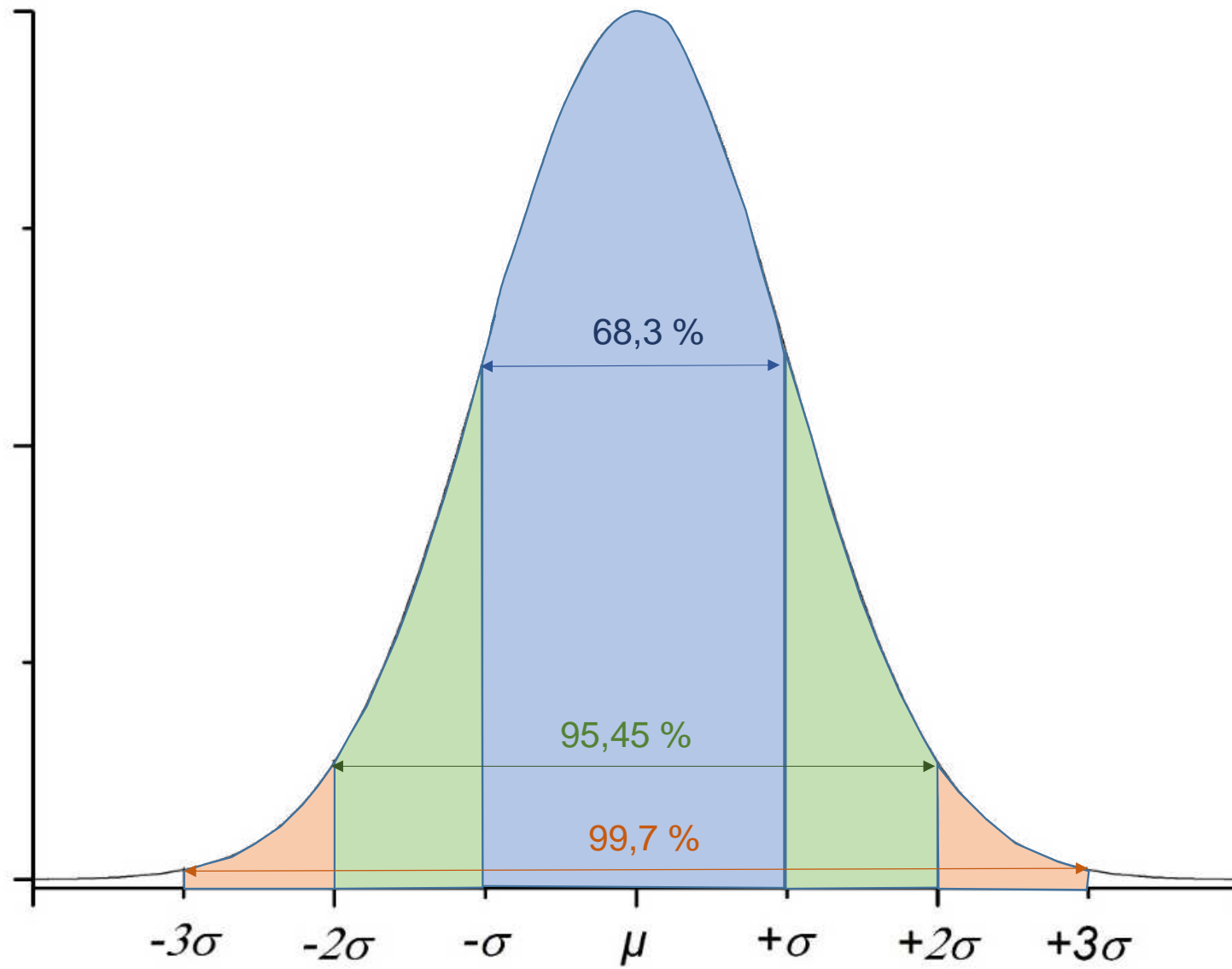






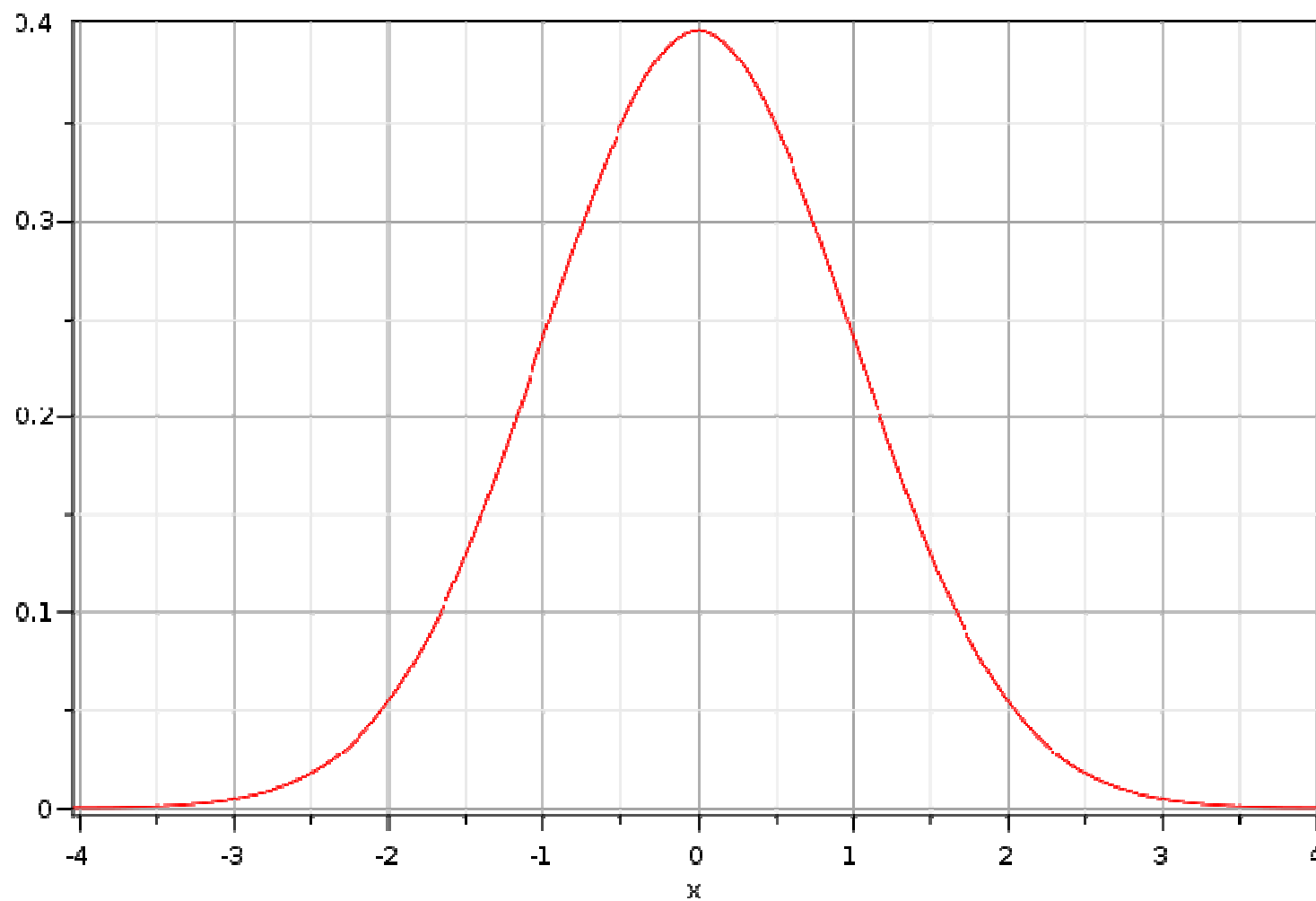
## Loi normale ou gaussienne



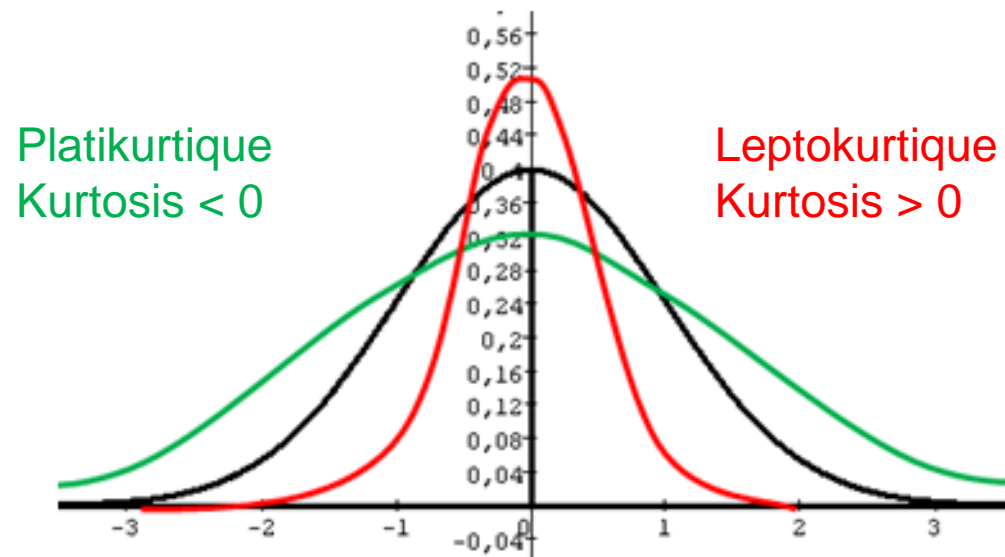




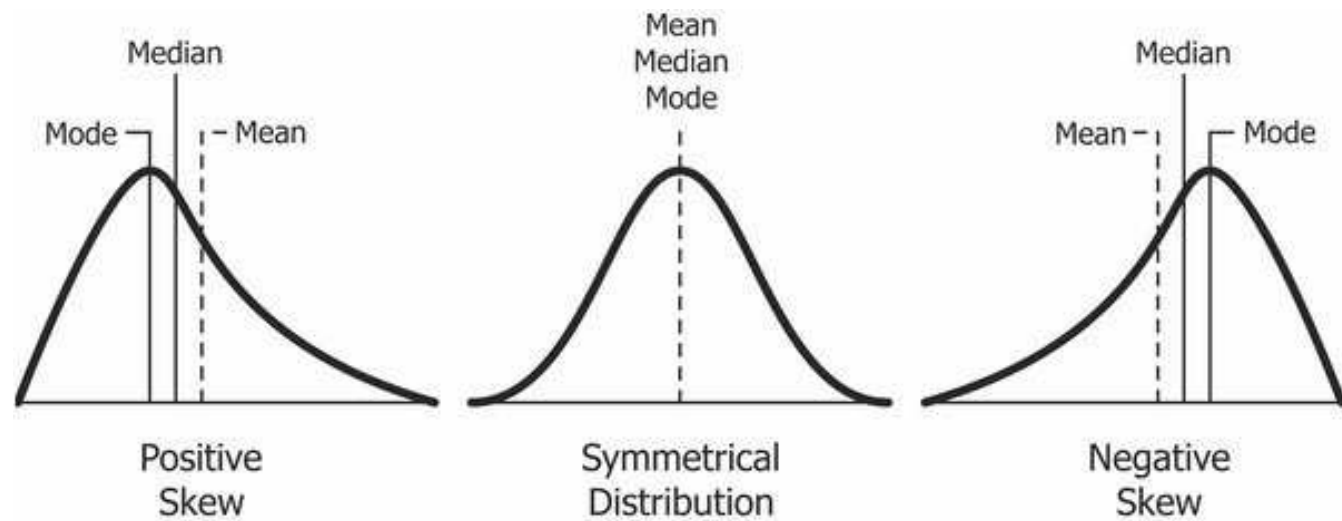
## Loi normale centrée réduite



## Kurtosis

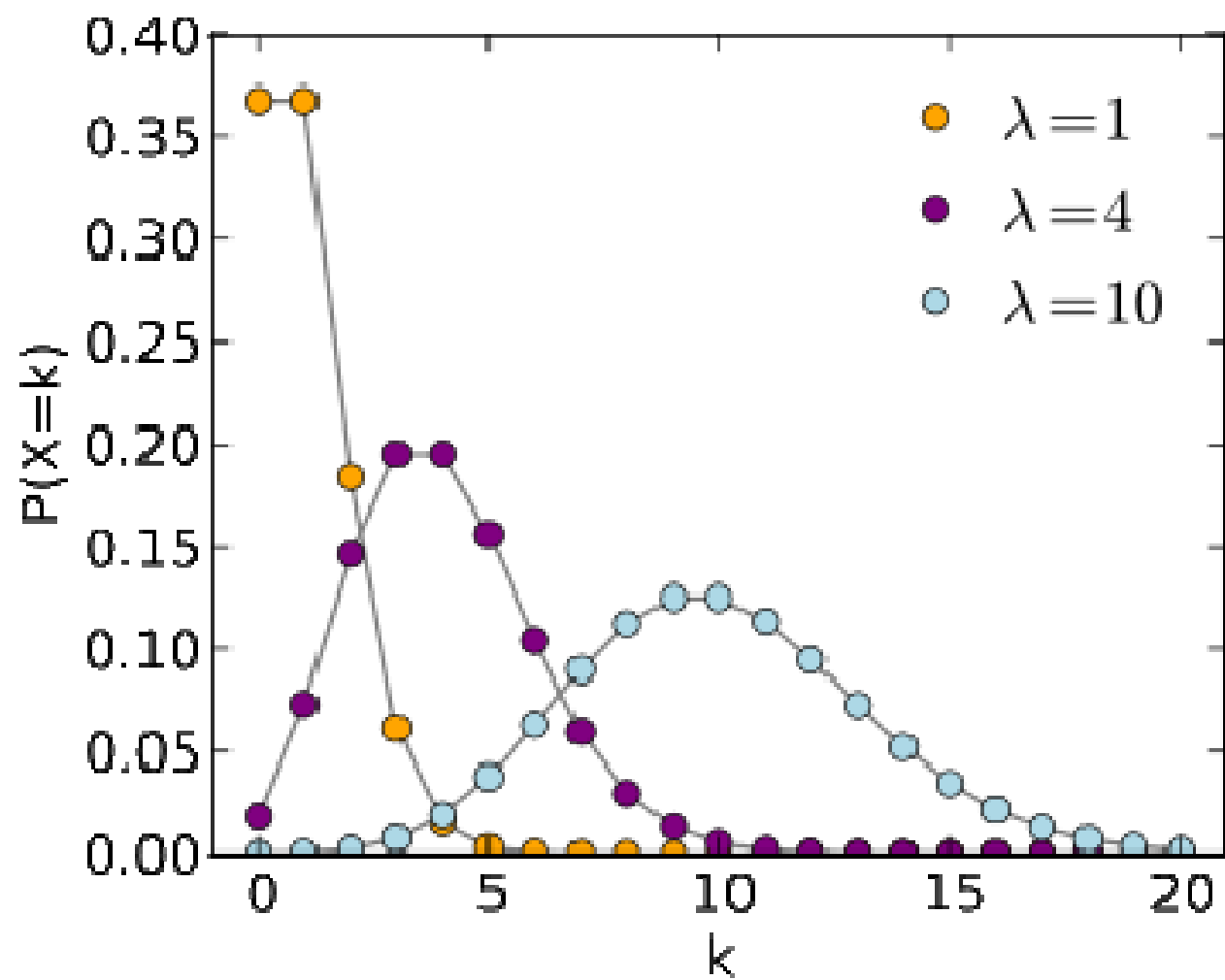


## Skewness





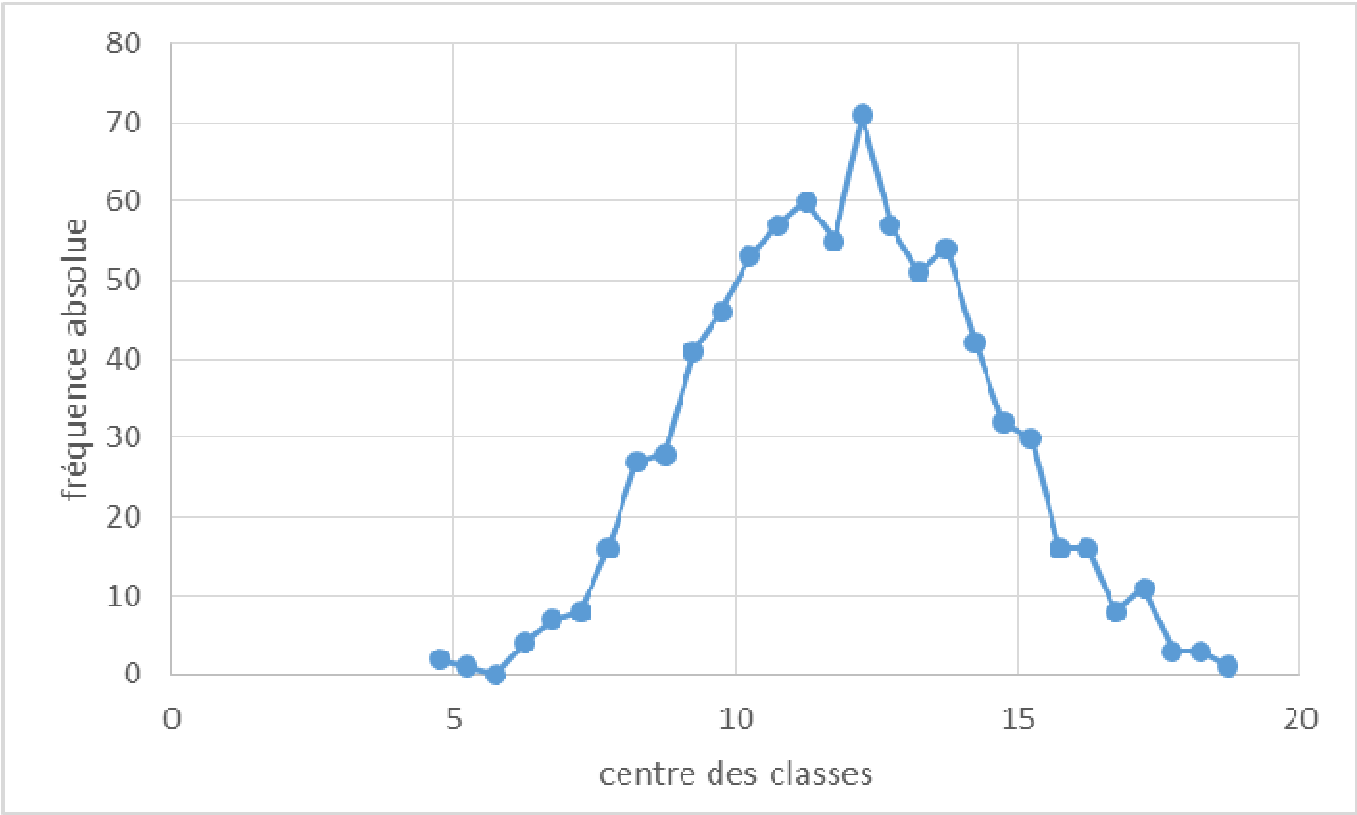
## Loi de Poisson de paramètre $\lambda$





# Représentation graphique (1) et première interprétation visuelle (2)

classes	centre des classes	Fréquence
5	4.75	2
5.5	5.25	1
6	5.75	0
6.5	6.25	4
7	6.75	7
7.5	7.25	8
8	7.75	16
8.5	8.25	27
9	8.75	28
9.5	9.25	41
10	9.75	46
10.5	10.25	53
11	10.75	57
11.5	11.25	60
12	11.75	55
12.5	12.25	71
13	12.75	57
13.5	13.25	51
14	13.75	54
14.5	14.25	42
15	14.75	32
15.5	15.25	30
16	15.75	16
16.5	16.25	16
17	16.75	8
17.5	17.25	11
18	17.75	3
18.5	18.25	3
ou plus...	18.75	1



$\mu = 11,93$   
Mediane = 11,94  
 $\sigma = 2,45$

Max pour x = 12,5

## Calcul des effectifs théoriques (3)

$$Z_c = \frac{X_c - \mu}{\sigma}$$

Pour  $X_c = 4,75$  :

$$Z_c = \frac{4,75 - 11,93}{2,45} = -2,93$$

classes	centre des classes	Fréquence	Zc
5	4.75	2	-2.93
5.5	5.25	1	-2.73
6	5.75	0	-2.52
6.5	6.25	4	-2.32
7	6.75	7	-2.11
7.5	7.25	8	-1.91
8	7.75	16	-1.71
8.5	8.25	27	-1.50
9	8.75	28	-1.30
9.5	9.25	41	-1.09
10	9.75	46	-0.89
10.5	10.25	53	-0.68
11	10.75	57	-0.48
11.5	11.25	60	-0.28
12	11.75	55	-0.07
12.5	12.25	71	0.13
13	12.75	57	0.34
13.5	13.25	51	0.54
14	13.75	54	0.74
14.5	14.25	42	0.95
15	14.75	32	1.15
15.5	15.25	30	1.36
16	15.75	16	1.56
16.5	16.25	16	1.77
17	16.75	8	1.97
17.5	17.25	11	2.17
18	17.75	3	2.38
18.5	18.25	3	2.58
ou plus...	18.75	1	2.79

classes	centre des classes	Fréquence	Zc	loi normale	effectif théorique
5	4.75	2	-2.93	0.005430363	1
5.5	5.25	1	-2.73	0.009678327	2
6	5.75	0	-2.52	0.016544541	3
6.5	6.25	4	-2.32	0.027126399	4
7	6.75	7	-2.11	0.042659187	7
7.5	7.25	8	-1.91	0.06434518	11
8	7.75	16	-1.71	0.093089881	15
8.5	8.25	27	-1.50	0.129173058	21
9	8.75	28	-1.30	0.171919213	28
9.5	9.25	41	-1.09	0.219462285	36
10	9.75	46	-0.89	0.268706634	44
10.5	10.25	53	-0.68	0.315558505	52
11	10.75	57	-0.48	0.355438445	58
11.5	11.25	60	-0.28	0.384000615	63
12	11.75	55	-0.07	0.397907796	65
12.5	12.25	71	0.13	0.395472223	65
13	12.75	57	0.34	0.376992345	62
13.5	13.25	51	0.54	0.344692702	56
14	13.75	54	0.74	0.302283647	50
14.5	14.25	42	0.95	0.254261279	42
15	14.75	32	1.15	0.205129826	34
15.5	15.25	30	1.36	0.158730507	26
16	15.75	16	1.56	0.11780806	19
16.5	16.25	16	1.77	0.083863429	14
17	16.75	8	1.97	0.05726025	9
17.5	17.25	11	2.17	0.037498756	6
18	17.75	3	2.38	0.023553937	4
18.5	18.25	3	2.58	0.014190351	2
ou plus...	18.75	1	2.79	0.008199848	1
			total	4.880977591	800

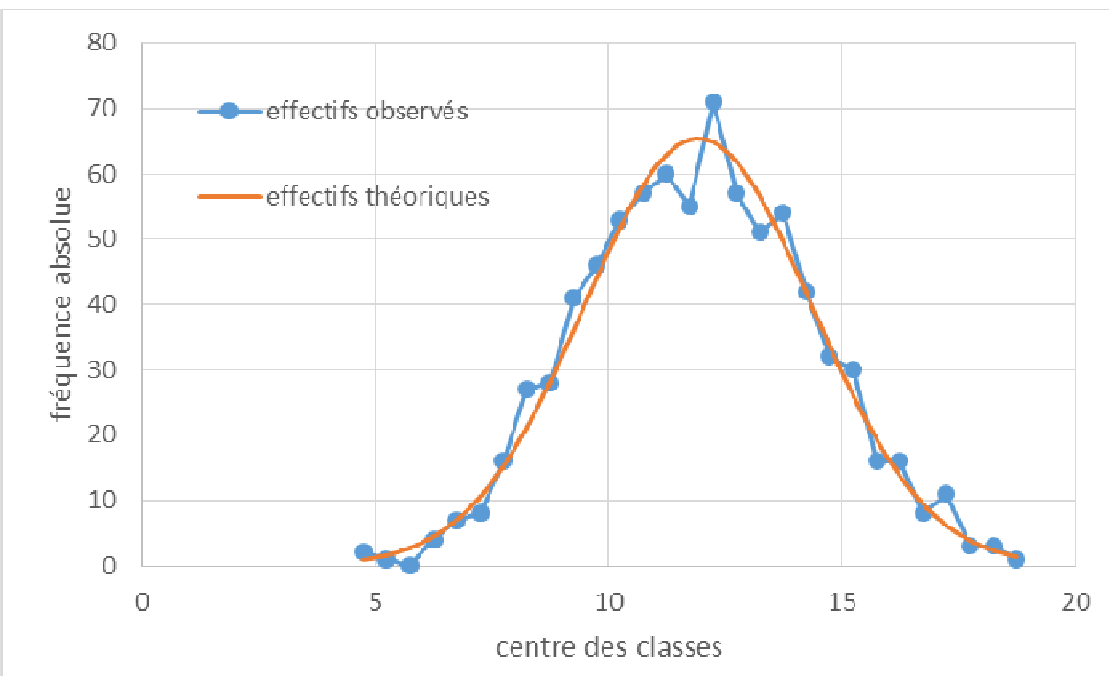
## Calcul des effectifs théoriques (3)

Loi théorique  
= LOI.NORMALE.STANDARD.N( $Z_C$ ;FAUX)

- $Z_C$  calculé avec  $X_C$  centre des classes
- Non cumulative
- Normalisée à  $\sigma$ /taille intervalle classe



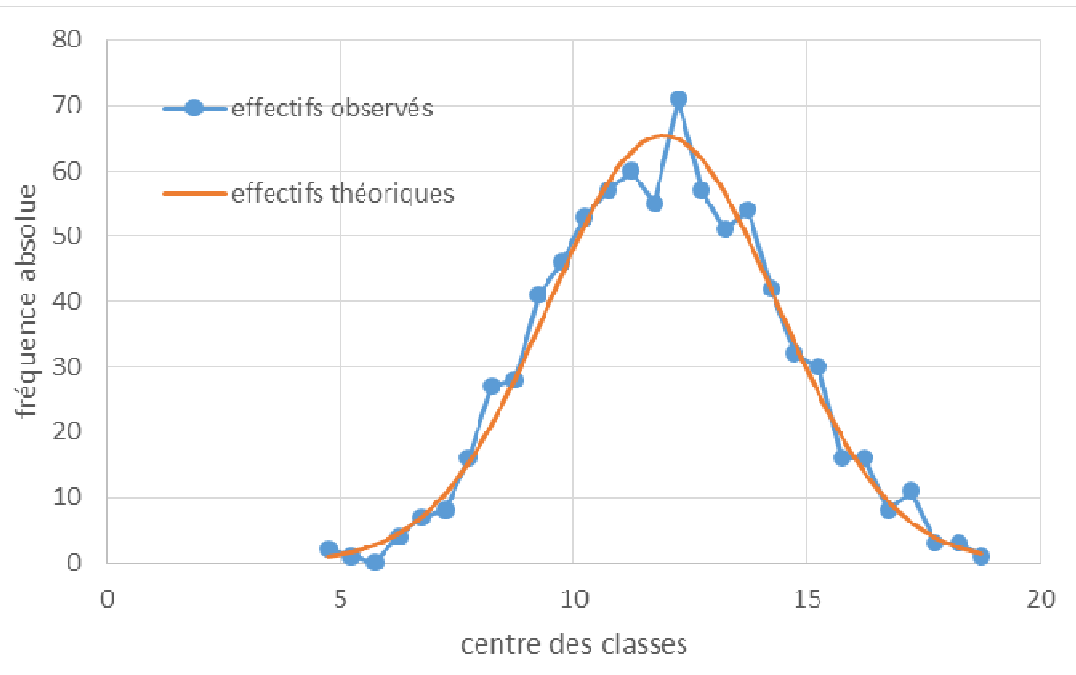
## Comparer les distributions observée et théorique (4)



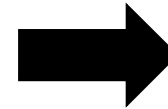
Conclusion ?



## Comparer les distributions observée et théorique (4)



Conclusion ?



Test du  $\chi^2$

On teste l'hypothèse : « la variable suit une loi de distribution normale de moyenne  $\mu$  et d'écart-type  $\sigma$  ».

**! Conditions de validité des tests statistiques !**

classes	centre des classes	Fréquence	Zc	loi normale	effectif théorique
5	4.75	2	-2.93	0.005430363	1
5.5	5.25	1	-2.73	0.009678327	2
6	5.75	0	-2.52	0.016544541	3
6.5	6.25	4	-2.32	0.027126399	4
7	6.75	7	-2.11	0.042659187	7
7.5	7.25	8	-1.91	0.06434518	11
8	7.75	16	-1.71	0.093089881	15
8.5	8.25	27	-1.50	0.129173058	21
9	8.75	28	-1.30	0.171919213	28
9.5	9.25	41	-1.09	0.219462285	36
10	9.75	46	-0.89	0.268706634	44
10.5	10.25	53	-0.68	0.315558505	52
11	10.75	57	-0.48	0.355438445	58
11.5	11.25	60	-0.28	0.384000615	63
12	11.75	55	-0.07	0.397907796	65
12.5	12.25	71	0.13	0.395472223	65
13	12.75	57	0.34	0.376992345	62
13.5	13.25	51	0.54	0.344692702	56
14	13.75	54	0.74	0.302283647	50
14.5	14.25	42	0.95	0.254261279	42
15	14.75	32	1.15	0.205129826	34
15.5	15.25	30	1.36	0.158730507	26
16	15.75	16	1.56	0.11780806	19
16.5	16.25	16	1.77	0.083863429	14
17	16.75	8	1.97	0.05726025	9
17.5	17.25	11	2.17	0.037498756	6
18	17.75	3	2.38	0.023553937	4
18.5	18.25	3	2.58	0.014190351	2
ou plus...	18.75	1	2.79	0.008199848	1
			total	4.880977591	800

**Effectifs < 5 !**

Regroupement de classes :

- Changement des bornes
- Classes plus larges

Min = 4,675  
 Max = 18,793  
 Étendue = 14,118



9 classes  
 Intervalle : 1,6  
 Min : 4,5  
 Max : 18,8

classe	centre des classes	effectif observé	Zc	Loi normale	effectifs théoriques
6.3	5.5	5	-2.625	0.013	6.67
7.9	7.1	30	-1.972	0.057	29.95
9.5	8.7	99	-1.318	0.167	87.76
11.1	10.3	167	-0.664	0.320	167.75
12.7	11.9	200	-0.011	0.399	209.16
14.3	13.5	161	0.643	0.324	170.13
15.9	15.1	94	1.296	0.172	90.27
17.5	16.7	37	1.950	0.060	31.25
19.1	18.3	7	2.604	0.013	7.06

**Effectifs > 5 !**



**Somme = 800**



**Somme = 1,53  
 =  $\sigma$ /intervalle**



**Somme = 800**

classe	centre des classes	effectif observé	Zc	Loi normale	effectifs théoriques	D <sup>2</sup>
6.3	5.5	5	-2.625	0.013	6.67	0.42
7.9	7.1	30	-1.972	0.057	29.95	0.00
9.5	8.7	99	-1.318	0.167	87.76	1.44
11.1	10.3	167	-0.664	0.320	167.75	0.00
12.7	11.9	200	-0.011	0.399	209.16	0.40
14.3	13.5	161	0.643	0.324	170.13	0.49
15.9	15.1	94	1.296	0.172	90.27	0.15
17.5	16.7	37	1.950	0.060	31.25	1.06
19.1	18.3	7	2.604	0.013	7.06	0.00

$$D^2 = \sum_{i=1}^k \frac{(O - T)^2}{T}$$

$$D^2 = 3,96$$

$$D^2 = 3,96$$

$$\chi^2_{\alpha} = ?$$

- calcul du degré de liberté

$$v = \text{nombre de classes} - 1 - \text{nb paramètres estimés}$$

- choix du risque (de se tromper)

$$\alpha = 1\%, \underline{5\%}, 10\%\dots$$

$$\chi^2_{\alpha} = \text{LOI.KHIDEUX.INVERSE.DROITE}(0,05;6) = 12,59 \quad (\text{Ou lecture dans une table})$$

$D^2 < \chi^2_{\alpha}$  : on ne peut pas rejeter l'hypothèse d'une distribution normale au risque 5% de se tromper.

Exemple :

Paramètres estimés = 2 (moyenne, écart-type)

Nombre de classes : 9

$$v = 9 - 1 - 2 = 6$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$p(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

$$\mu = \sigma^2 = \lambda$$

$$\mu = \frac{a+b}{2}$$

$$\sigma^2 = \frac{(b-a)^2}{12} ; \sigma = \frac{b-a}{2\sqrt{3}}$$

Formules à mémoriser :

$$Z_c = \frac{X_c - \mu}{\sigma}$$

$$D^2 = \sum_{i=1}^k \frac{(O - T)^2}{T}$$

= LOI.NORMALE.STANDARD.N(Z<sub>C</sub>;FAUX)

## Calcul des effectifs dans Excel

1. Définir les bornes supérieures des classes (col.1, k lignes)
2. Sélectionner la colonne voisine **+1 lignes (k+1)** (col.2)
3. =FREQUENCE (matrice X;col.1)
4. CTRL+SHIFT+ENTREE : {FREQUENCE (matrice X;col.1)}

## Représentation graphique dans Excel

1. Calculer le centre des classes (col.3)
2. Sélectionner la col.2 (fréquences)
3. Tracer un histogramme
4. Clic droit : sélectionner des données
5. Modifier les étiquettes de l'axe horizontal : centre des classes (col.3)

Ou : utilitaire d'analyse