Formules de stats descriptives pour les **échantillons** (notations : n, $\bar{x}$, s)

Présentation des données

Données isolées (ou énumérées)

Ex. : {2, 5, 5, 6, 8, 9, 12, 13, 15, 16, 16, 20, 21, 23, 24, 25}

$n = 16;\ x_1 = 2, \dots, x_{16} = 25$

Données groupées (valeurs et fréq.)

Ex. : {3, 3, 3, 3, 8, 8, 8, 8, 8, 8, 11, 11, 11, 11, 14, 14} (k=4 valeurs distinctes)

v_i	Fréq. f_i	Fréq.cum. F_i
v_1 =3	4	4
8	6	10
11	4	14
v_4 = 14	2	16 = n

Dans les notes de cours, v_i était noté aussi x_i ☺; $x_1 = 3, \dots, x_{16} = 16$

Données groupées par classes Ex. : (k=6 classes)	Pt milieu m_i	fréquence f_i	Fréq.cum. F_i
[50, 70[	60	2	2
[70, 90[	80	8	10
[90, 110[	100	12	22
[110, 130[	120	6	28
[130, 150[	140	3	31
[150, 170[	160	1	32=n

Mesures de tendance centrale

Moyenne

Données isolées :

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

$$\bar{x} = \frac{2+5+5+\cdots+24+25}{16} = \frac{220}{16} = 13{,}75$$

Données groupées (valeurs) :

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i v_i}{n}$$

$$\bar{x} = \frac{(4)(3)+(6)(8)+(4)(11)+(2)(14)}{16} = \frac{132}{16} = 8{,}25$$

Données groupées par classes :

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i m_i}{n}$$

$$\bar{x} = \frac{(2)(60)+(8)(80)+(12)(100)+(6)(120)+(3)(140)+(1)(160)}{32} = \frac{3\,260}{32} = 101{,}875$$

Médiane

$$\mathrm{Me} = x_{\left[\frac{n+1}{2}\right]}$$

où la notation [ ] en indice permet d'aller chercher dans les données **énumérées et triées** celle (ou une moyenne de celles) au centre.

Par exemple, $x_{\left[\frac{3+1}{2}\right]} = x_2$ tandis que $x_{\left[\frac{4+1}{2}\right]} = x_{[2.5]} = x_2 + 0.5(x_3 - x_2) = \frac{x_2+x_3}{2}$

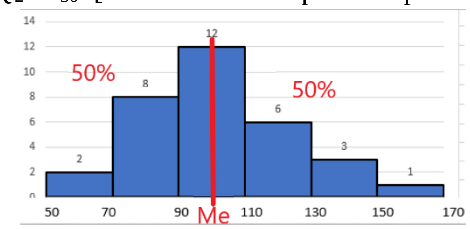
Données isolées — Ex. : $\mathrm{Me} = x_{\left[\frac{16}{2}+\frac{1}{2}\right]} = x_{[8.5]} = \frac{x_8+x_9}{2} = \frac{13+15}{2} = 14$

Données groupées (valeurs) — Ex. : $\mathrm{Me} = x_{\left[\frac{16}{2}+\frac{1}{2}\right]} = x_{[8.5]} = \frac{x_8+x_9}{2} = \frac{8+8}{2} = 8$

Données groupées par classes — Cas particulier de C_{50}, car $Me=Q_2=C_{50}$. [Voir les centiles plus bas pour description des paramètres.]

$$Me = L + \left(\frac{\frac{n}{2}-F}{f}\right)\cdot a$$

$$= 90 + \left(\frac{\frac{32}{2}-10}{12}\right)\cdot 20 = 100$$



Mode

Données isolées et groupées (valeurs) : Valeur(s) la plus fréquente

- Ex. : Mode = 5 ou 16 (bimodal)
- Ex. : Mode = 8

Données groupées par classes : Classe modale : classe avec la plus grande fréquence

- Ex. : Classe [90, 110[

Mesures de dispersion

Variance (pour échantillon)

Données isolées :

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

Ex. : $\sum_{i=1}^{n} x_i^2 = (2)^2 + (5)^2 + \cdots + (25)^2 = 3\,856$

$\left(\sum_{i=1}^{n} x_i\right)^2 = (2+5+\cdots+25)^2 = (220)^2$

d'où $s^2 = \frac{3\,856 - \frac{220^2}{16}}{15} = 55{,}4$

Données groupées (valeurs) :

$$s^2 = \frac{\sum_{i=1}^{k} f_i v_i^2 - \frac{\left(\sum_{i=1}^{k} f_i v_i\right)^2}{n}}{n-1}$$

Ex. : $\sum_{i=1}^{k} f_i v_i^2 = (4)(3)^2 + (6)(8)^2 + (4)(11)^2 + (2)(14)^2 = 1\,296$

$\left(\sum_{i=1}^{k} f_i v_i\right)^2 = \left((4)(3) + (6)(8) + (4)(11) + (2(14)\right)^2 = 132^2$

d'où $s^2 = \frac{1\,296 - \frac{132^2}{16}}{15} = 13{,}8$

Données groupées par classes :

$$s^2 = \frac{\sum_{i=1}^{k} f_i m_i^2 - \frac{\left(\sum_{i=1}^{k} f_i m_i\right)^2}{n}}{n-1}$$

$\sum_{i=1}^{k} f_i m_i^2 = (2)(60)^2 + (8)(80)^2 + \cdots + (1)(160)^2 = 349\,200$

$\left(\sum_{i=1}^{k} f_i m_i\right)^2 = \left((2)(60) + (8)(80) + \cdots + (1)(160)\right)^2 = 3\,260^2$

d'où $s^2 = \frac{349\,200 - \frac{3\,260^2}{32}}{31} = 551{,}2097$

Écart-type échantillon

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

- Ex. : $s = \sqrt{55{,}4} \approx 7{,}443$
- Ex. : $s = \sqrt{13{,}8} \approx 3{,}715$
- Ex. : $s = \sqrt{551{,}2097} \approx 23{,}48$

Mesures de position (une donnée x_0 vs groupe ou des balises globales sur le groupe)

Quartiles	On ramène le quartile au centile correspondant : $Q_1=C_{25}$, $Q_2=C_{50}$ et $Q_3=C_{75}$		
Centiles	$C_i = x_{\left[\frac{n\cdot i}{100}+\frac{1}{2}\right]}$	$C_i = x_{\left[\frac{n\cdot i}{100}+\frac{1}{2}\right]}$ mais il faut utiliser les fréquences cumulées pour connaître la valeur des $x_{[\ ]}$ (savoir à l'intérieur de quelles valeurs répétées).	$C_i = L + \left(\dfrac{\frac{n\cdot i}{100} - F}{f}\right)\cdot a$ L = limite inférieure de la classe contenant le quantile. n = taille de l'échantillon. f = fréquence absolue de la classe contenant le quantile. F = fréquence cumulée de la classe qui précède la classe contenant le quantile. a = amplitude de la classe contenant le quantile.
	Ex. : $Q_3 = C_{75} = x_{\left[\frac{16\cdot 75}{100}+\frac{1}{2}\right]} = x_{[12.5]}$ $= x_{12} + (0.5)(x_{13} - x_{12})$ $= 20 + (0.5)(21 - 20) = 21{,}5$	Ex. : $C_{20} = x_{\left[\frac{16\cdot 30}{100}+\frac{1}{2}\right]} = x_{[4.1]}$ $= x_4 + (0.1)(x_5 - x_4)$ $= 3 + (0.1)(8 - 3) = 3{,}5$	$Q_1 = C_{25} = L + \left(\dfrac{\frac{n\cdot 25}{100} - F}{f}\right)\cdot a = 70 + \left(\dfrac{\frac{32\cdot 25}{100} - 2}{8}\right)\cdot 20 = 85$ car d'après les fréq. cum., $\frac{32\cdot 25}{100} = 8^{\text{ième}}$ donnée; dans la classe [70,90[
Cote Z	Pour obtenir la cote Z d'une valeur d'une donnée particulière x_0 on soustrait la moyenne et divise par l'écart-type : $Z = \frac{x_0 - \bar{x}}{s}$		
	Ex. : la cote Z de la donnée 9 : $Z = \dfrac{x_0 - \bar{x}}{s} = \dfrac{(9 - 13{,}75)}{7{,}443} = -0{,}64$	Ex. : la cote Z de la donnée 11 : $Z = \dfrac{x_0 - \bar{x}}{s} = \dfrac{(11 - 8{,}25)}{3{,}715} = 0{,}61$	Ex. : la cote Z d'une donnée valant 142 : $Z = \dfrac{x_0 - \bar{x}}{s} = \dfrac{(142 - 101{,}875)}{23{,}48} = 1{,}71$

Formules de stats descriptives pour les **populations** (notations : N, μ, σ) [juste ce qui diffère des formules pour les échantillons plus haut]

Moyenne	$\mu = \dfrac{\sum_{i=1}^{n} x_i}{N}$	$\mu = \dfrac{\sum_{i=1}^{k} f_i v_i}{N}$	$\mu = \dfrac{\sum_{i=1}^{k} f_i m_i}{N}$
	Juste le symbole change (μ, lire «mu», au lieu de $\bar{x}$) pas les formules car N=n numériquement. Revoir les ex. ci-haut pour les échantillons.		
Variance (pour une population)	$\sigma^2 = \dfrac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{N}}{N}$	$\sigma^2 = \dfrac{\sum_{i=1}^{k} f_i v_i^2 - \frac{\left(\sum_{i=1}^{k} f_i v_i\right)^2}{N}}{N}$	$\sigma^2 = \dfrac{\sum_{i=1}^{k} f_i m_i^2 - \frac{\left(\sum_{i=1}^{k} f_i m_i\right)^2}{N}}{N}$
	Ex. : $\sum_{i=1}^{n} x_i^2 = idem = 3\,856$ $\left(\sum_{i=1}^{n} x_i\right)^2 = idem = (220)^2$ d'où $\sigma^2 = \dfrac{3\,856 - \frac{220^2}{16}}{16} = 51{,}9375$	Ex. : $\sum_{i=1}^{k} f_i v_i^2 = idem = 1\,296$ $\left(\sum_{i=1}^{k} f_i v_i\right)^2 = idem = 132^2$ d'où $\sigma^2 = \dfrac{1\,296 - \frac{132^2}{16}}{16} = 12{,}9375$	$\sum_{i=1}^{k} f_i m_i^2 = idem = 349\,200$ $\left(\sum_{i=1}^{k} f_i m_i\right)^2 = idem = 3\,260^2$ d'où $\sigma^2 = \dfrac{349\,200 - \frac{3\,260^2}{32}}{32} = 553{,}984375$ Idem : Mêmes sommes qu'avec les échantillons… …seule la 2e division par N au lieu de n-1change.
Écart-type population	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$
	Ex. : $\sigma = \sqrt{51{,}0375} \approx 7{,}207$	Ex. : $\sigma = \sqrt{12{,}9375} \approx 3{,}597$	Ex. : $\sigma = \sqrt{553{,}984375} \approx 23{,}11$
$Z = \dfrac{x_0 - \mu}{\sigma}$	Ex. : la cote Z de la donnée 9 : $Z = \dfrac{x_0 - \mu}{\sigma} = \dfrac{(9 - 13{,}75)}{7{,}207} = -0{,}66$	Ex. : la cote Z de la donnée 11 : $Z = \dfrac{x_0 - \mu}{\sigma} = \dfrac{(11 - 8{,}25)}{3{,}597} = 0{,}76$	Ex. : la cote Z d'une donnée valant 142 : $Z = \dfrac{x_0 - \mu}{\sigma} = \dfrac{(142 - 101{,}875)}{23{,}11} = 1{,}74$