

Mesures numériques (statistiques) de distributions

Nous nous limitons à l'étude de variables quantitatives. Les mesures (synthèses) numériques plus communes sont celles de:

1. *position* qui indiquent où se situe la distribution;
2. *dispersion* qui mesurent la variabilité (éparpillement).

Mesures de position

Moyenne arithmétique

Soient x_1, \dots, x_n les observations d'une variable X .

La *moyenne arithmétique* de la distribution de X (ou *moyenne* de X) est définie par:

$$m(X) = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

Souvent on utilise aussi la notation \bar{x} à la place de $m(X)$.

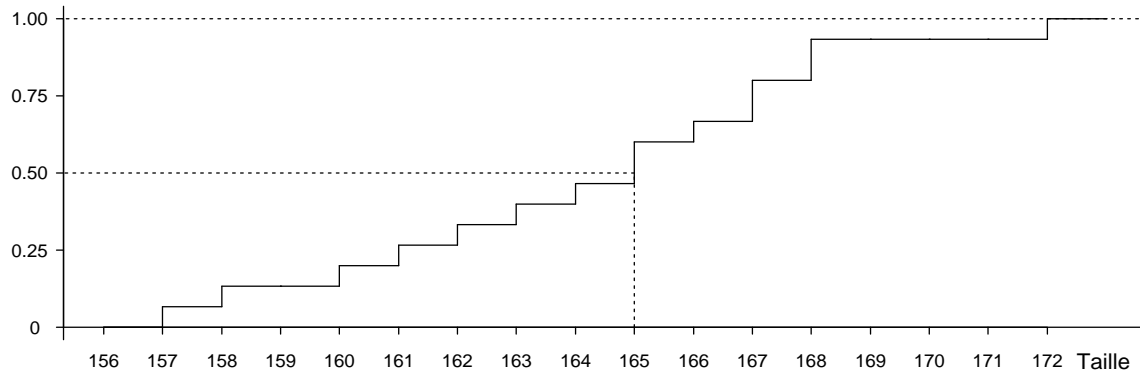
Exemple de calcul

Données: $x_1 = 15$, $x_2 = 25$, $x_3 = 31$, $x_4 = 10$, $x_5 = 75$;

$$m(X) = (15 + 25 + 31 + 10 + 75)/5 = 31.2.$$

Médiane

La *médiane* (de la distribution) de X est une valeur telle qu'une moitié des données se situe à sa droite et l'autre moitié à sa gauche. Pour la calculer on se sert de la fonction de distribution cumulative empirique $F_n(x)$.



$$F_n(x) \leq 0.5 \quad \text{si} \quad x < \text{med}(X), \quad \text{et} \quad F_n(x) \geq 0.5 \quad \text{si} \quad x > \text{med}(X).$$

Exemple. Tailles des 15 filles: $\text{med}(T) = 165$.

Pour le calcul à la main, soient x_1, \dots, x_n des observations de X et notons par $x_{[1]} \leq \dots \leq x_{[n]}$ les mêmes valeurs rangées en ordre croissant.

– si n est impair, $\text{med}(X) = x_{[(n+1)/2]}$,

– si n est pair, $\text{med}(X) = (x_{[n/2]} + x_{[n/2+1]})/2$.

Exemples

a) Données: 27, 29, 31, 31, 31, 34, 36, 39, 42.

$n = 9$, $(n + 1)/2 = 5$; la médiane est la cinquième valeur :

$$\text{med} = x_{[5]} = 31$$

b) Données: 27, 29, 31, 31, 31, 34, 36, 39, 42, 45.

$n = 10$, $n/2 = 5$, $n/2 + 1 = 6$ et

$$\text{med} = (x_{[5]} + x_{[6]})/2 = (31 + 34)/2 = 32.5$$

Moyenne ou médiane ?

- 1) Si la distribution de X est (approximativement) symétrique, son centre est bien défini. Dans ce cas, $\text{med}(X) \approx m(X)$.
- 2) $m(X)$ a des propriétés de calcul simples. Par exemple, soit X le coût hôtelier, Y le coût médical de séjours à l'hôpital. Alors $m(X+Y) = m(X) + m(Y)$ (en général: $m(aX + bY) = am(X) + bm(Y)$).
- 3) La moyenne se laisse influencer par les **outliers**. Par contre la médiane *résiste* lors de la modification (correction, élimination) de données éloignées. On dit que la médiane est plus *robuste* que la moyenne. Par exemple:

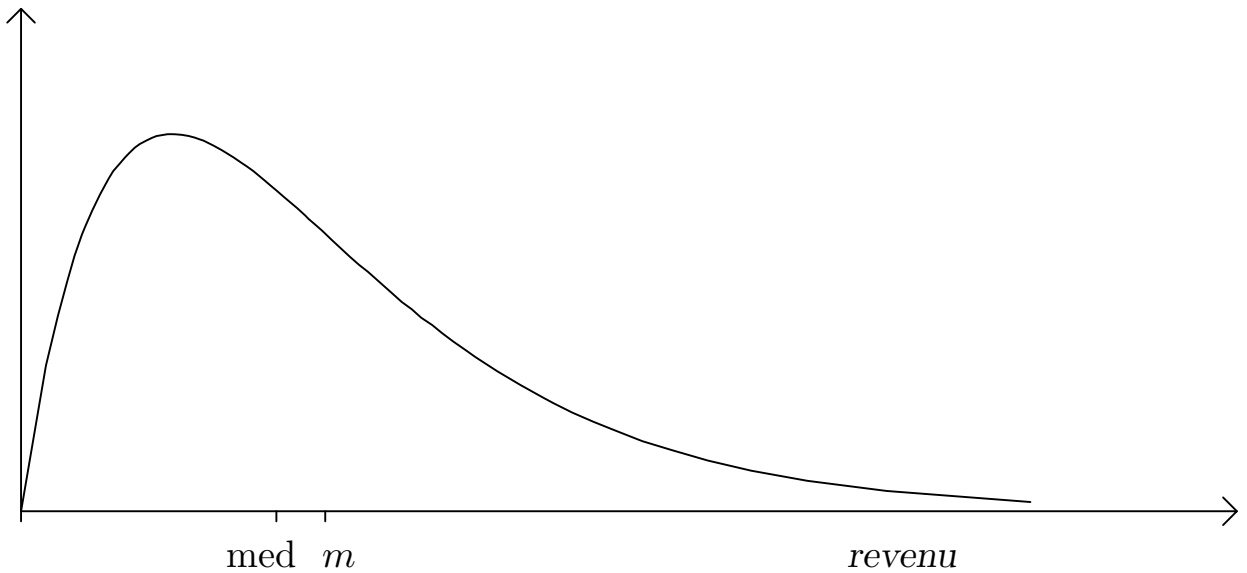
Données: 27, 29, 31, 31, 31, 34, 36, 39, 42, 45.

$n = 10$, $\text{med} = 32.5$, $m = 34.5$

Données: 27, 29, 31, 31, 31, 34, 36, 39, 42, 4500.

$n = 10$, $\text{med} = 32.5$, $m = 480$

- 4) Considérons la distribution des revenus dans le canton de Vaud.
La distribution des revenus est typiquement asymétrique.



En général, pour une telle distribution, $\text{med}(X) < m(X)$. Pour un habitant, il est intéressant de connaître la médiane dans le but de se situer dans la “moitié riche” ou dans la “moitié pauvre”. Pour le département des finances la moyenne est utile pour estimer le bénéfice des impôts (\approx revenu moyen \times coefficient moyen \times nombre d’habitants).

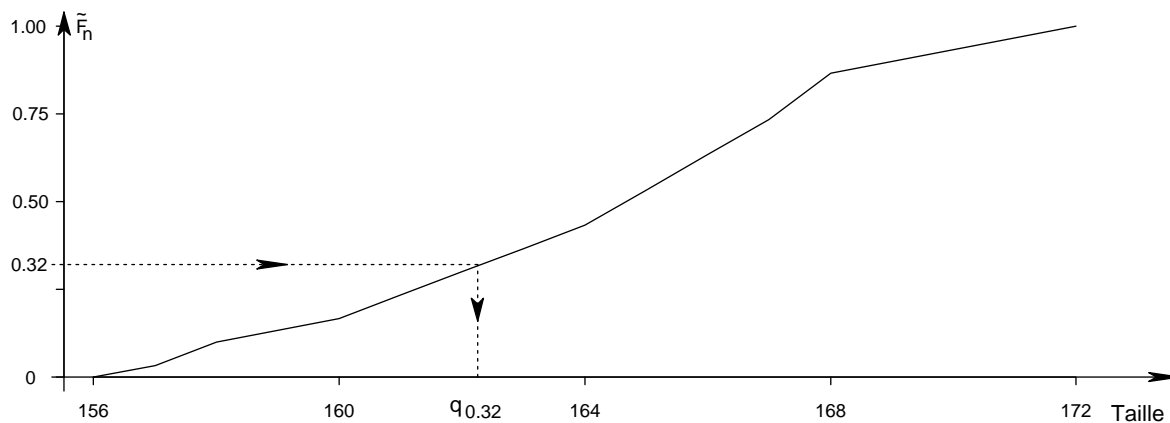
Quantiles, percentiles, déciles, quartiles

La médiane partage la distribution en deux: 50% des données sont plus petites que la médiane; 50% sont plus grandes. On peut partager la distribution en quatre, en dix, en cent, ou en un nombre quelconque de parties. Les valeurs ainsi obtenues sont appelées des *quartiles*, des *déciles*, des *percentiles* (ou *centiles*), ou des *quantiles*. Par exemple, le percentile 32%, est une valeur telle que 32% des données lui sont inférieures et 68% lui sont supérieures. La médiane est le percentile 50%. Le premier quartile est le percentile 25%; le troisième quartile est le percentile 75%.

Soit α un nombre compris entre 0 et 1. Pour définir le *quantile* α , que l'on note par q_α , on pourra se servir d'une version lissée et strictement croissante $\tilde{F}_n(x)$ de la fonction de distribution cumulative. Alors

$$q_\alpha = \tilde{F}_n^{-1}(\alpha),$$

où \tilde{F}_n^{-1} indique la fonction inverse de \tilde{F}_n .



En général, le calcul d'un quantile s'effectue à l'aide du graphique de $F_n(x)$ ou à l'aide d'un programme d'ordinateur.

Mesures de dispersion

Variance et écart-type

La *variance* de X est la moyenne des carrés des écarts entre les observations et leur moyenne.

Exemple de calcul

	x_i	$x_i - m(X)$	$(x_i - m(X))^2$
	3	-3	9
	5	-1	1
	12	6	36
	4	-2	4
Total	24	0	50
Total/4	6	0	12.5

Donc, $m(X) = 24/4 = 6$ et $s^2(X) = 12.5$. Si les données sont exprimées en centimètres (cm) on a $m(X) = 6\text{cm}$ et $s^2(X) = 12.5\text{cm}^2$. La variance n'a donc pas la même échelle que les données. Pour y remédier on utilise l'*écart-type*:

$$s(X) = \sqrt{s^2(X)}.$$

Exemple

Tailles des 15 filles : $s(T_f) = 3.95$ cm

Tailles de 30 garçons : $s(T_h) = 5.48$ cm

Note. Pour des raisons dépassant le cadre de ce cours, on utilise souvent une définition de la variance légèrement différente de celle donnée ci-dessus: à la place de diviser la somme des carrés des écarts entre les données et leur moyenne par le nombre de données n , on divise cette somme par $(n - 1)$.

Le MAD

Soient x_1, \dots, x_n les observations de X . Le *MAD* (*median absolute deviation*) de X est la médiane des écarts absolus entre les observations et leur médiane:

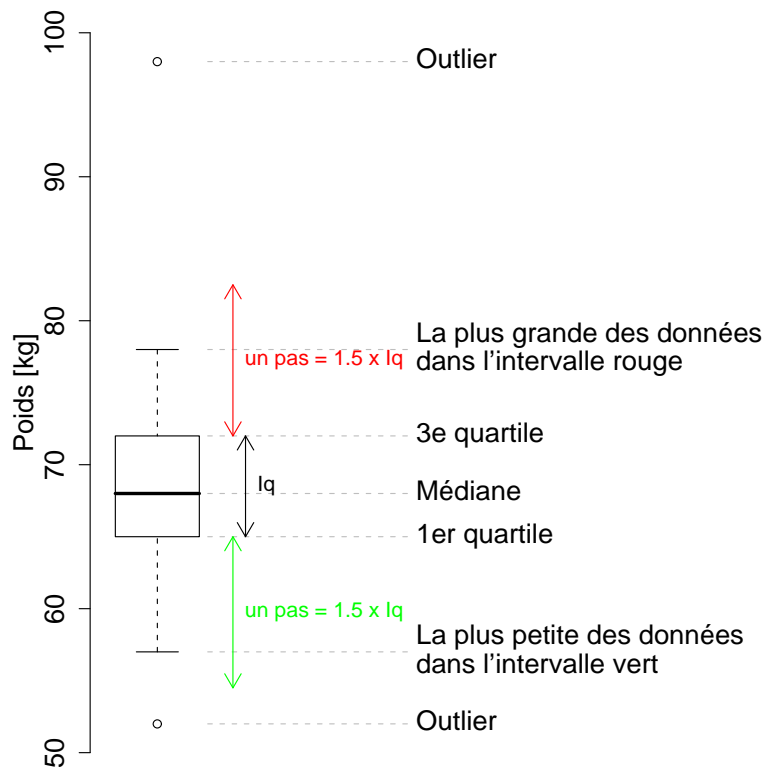
$$\text{MAD}(X) = \text{med}(|x_i - \text{med}(X)|).$$

Ecart interquartile

L'*écart interquartile* de X est la différence entre le troisième et le premier quartile de la distribution de X :

$$I_q = q_{0.75} - q_{0.25}.$$

Le Box-plot



Box-plot des poids des 30 étudiants garçons

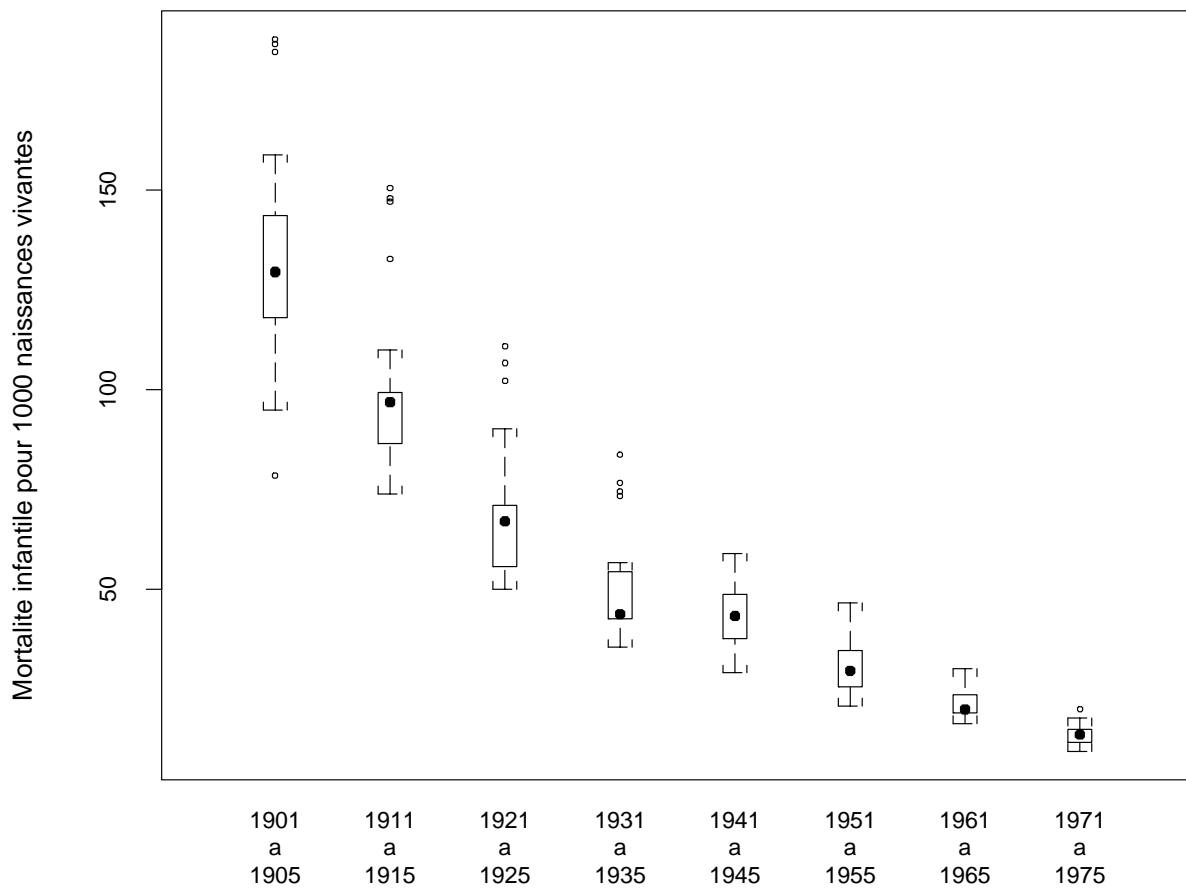
Le *box-and-whiskers plot* est une représentation graphique simple mais puissante d'un échantillon. L'échelle sur l'axe vertical est celle des données. On trace d'abord un rectangle (le box); le côté inférieur et le côté supérieur correspondent au premier et au troisième quartiles. Ainsi, le box contient idéalement la moitié (50%) centrale des données. Le box est partagé en deux par un trait horizontal au niveau de la médiane. On définit ensuite un *pas* comme le segment de longueur $1.5 \times I_q$, où I_q est l'écart interquartile, et on considère les données situées entre le côté supérieur et le côté supérieur plus un pas. Un trait vertical (*whisker*, *moustache en anglais*) s'étend du côté supérieur jusqu'à la plus élevée de ces données. De façon similaire, on définit l'autre moustache. Les données en dehors des moustaches sont marquées individuellement par le symbole "O" (*outlier*).

Exemple

Mortalité infantile pour 1000 naissances vivantes en Suisse

Source: Office fédéral de la statistique, 1982

	1901 à 1905	1911 à 1915	1921 à 1925	1931 à 1935	1941 à 1945	1951 à 1955	1961 à 1965	1971 à 1975
ZH.....	124.1	84.5	50.0	36.3	32.5	21.8	17.3	11.6
BE.....	119.8	85.3	54.8	42.6	36.0	26.7	17.0	11.6
LU.....	109.1	94.0	69.6	54.4	48.7	29.7	21.6	13.6
UR.....	125.3	109.9	87.8	56.7	51.1	46.6	26.1	15.2
SZ.....	138.8	98.6	71.0	45.6	46.1	34.6	24.3	13.7
OW.....	78.5	78.0	69.7	43.3	49.1	23.9	18.8	13.8
NW.....	94.9	73.9	69.1	44.9	43.5	29.5	29.1	16.3
GL.....	113.2	86.5	67.1	41.2	30.8	25.8	18.5	11.4
ZG.....	113.3	89.7	58.7	35.5	48.0	34.3	19.5	12.8
FR.....	186.6	150.5	90.2	74.5	56.4	42.6	30.1	14.5
SO.....	132.9	98.1	67.0	43.2	37.6	29.7	19.3	12.6
BS.....	133.3	80.4	51.9	43.8	34.7	23.9	19.0	11.2
BL.....	133.3	99.0	51.6	42.6	29.1	20.8	16.4	9.6
SH.....	129.5	94.4	54.3	48.6	41.3	23.3	20.6	13.9
AR.....	135.6	99.3	60.3	38.2	39.4	29.7	19.9	9.4
AI.....	184.6	148.0	110.9	83.7	54.4	38.2	19.5	20.0
SG.....	148.9	107.8	71.0	50.3	40.9	27.3	19.9	14.3
GR.....	118.0	98.9	71.9	54.6	43.5	34.6	24.7	16.0
AG.....	118.8	86.7	55.6	36.4	34.1	25.5	16.9	11.8
TG.....	123.2	96.9	106.7	73.4	54.2	38.8	23.6	17.2
TI.....	187.8	147.1	106.7	73.4	54.2	38.8	23.6	17.2
VD.....	143.8	97.2	55.7	48.9	39.6	30.5	21.8	12.8
VS.....	158.9	132.8	102.2	76.7	59.0	44.3	27.3	17.7
NE.....	143.7	96.1	57.7	43.2	43.3	27.8	19.4	14.9
GE.....	113.9	80.2	56.3	43.5	43.9	29.6	21.4	12.6



Box-plots de la mortalité infantile en Suisse selon la période.

Mesures d'association entre deux variables

Variables binaires qualitatives

Exemple. On s'intéresse à l'association entre le mode de vie, "seul" ou "en famille", et la présence ou l'absence d'une névrose. Dans un échantillon de 260 d'individus d'une certaine population on a trouvé les fréquences ci-dessous:

Mode de vie	Névrose		Total
	Présente	Absente	
En famille	40	60	100
Seul	100	60	160
Total	140	120	260

Cas général

	$B = 0$	$B = 1$	Total
$A = 0$	n_{11}	n_{12}	$n_{1.}$
$A = 1$	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

Pour mesurer l'association entre A et B on peut comparer

$n_{11}/n_{1.}$ = fréquence de $B = 0$ pour $A = 0$,

$n_{21}/n_{2.}$ = fréquence de $B = 0$ pour $A = 1$.

On calcule le *risque relatif* $(n_{11}/n_{1.})/(n_{21}/n_{2.})$

Exemple. Risque relatif “Névrose en famille” contre “Névrose seul”:

$$(40/100)/(100/160) = 0.64.$$

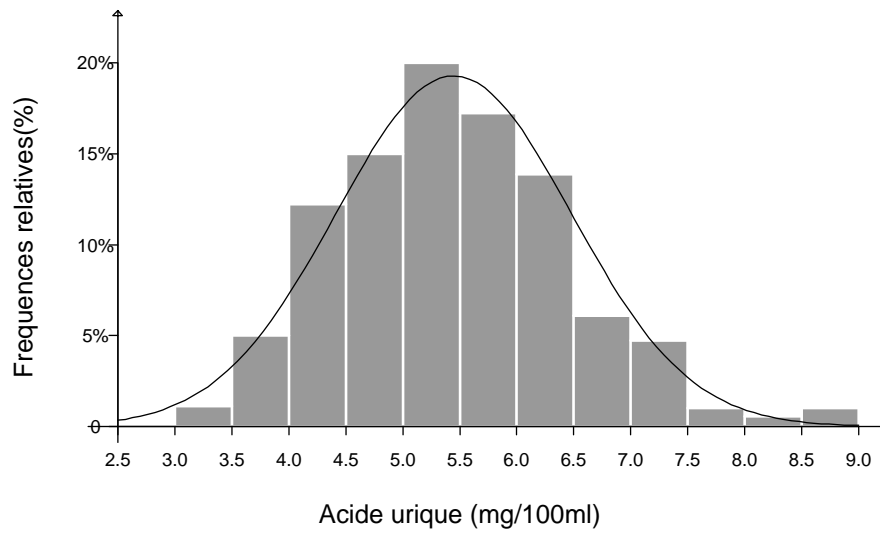
Une autre mesure d'association est le *odds ratio*

→ BOMS, Polycopié de biostatistique

Le modèle de Gauss ou modèle normal

Beaucoup de variables biologiques ont une distribution “en cloche”.

Exemple



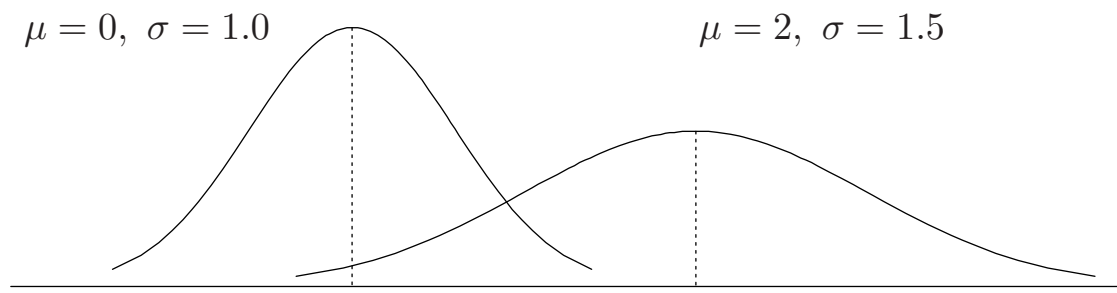
Histogramme de la concentration d'acide urique sérique
chez 267 donneurs de sang de sexe masculin

Dans ces cas, il est souvent utile de décrire la distribution à l'aide du **modèle mathématique de Gauss** (ou **normal**):

$$f_{\mu,\sigma}(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-(y - \mu)^2/(2\sigma^2)).$$

En changeant les valeurs de μ et de σ on obtient des cloches de différentes positions et largeurs.

μ est appelé la moyenne du modèle de Gauss correspondant, et σ son écart-type.



On peut adapter le modèle à un histogramme donné en posant

$$\mu = m(X) \quad \text{et} \quad \sigma = s(X).$$

Que signifie “valeur dans la norme”

La probabilité qu’une variable X Gaussienne s’éloigne de sa moyenne à plus de 2 écarts type est d’environ 2.5% :

$$P(X > \mu + 2 \cdot \sigma) \approx 2.5\%$$

et donc

$$P(X < \mu - 2 \cdot \sigma) \approx 2.5\%$$

$$P(\mu - 2 \cdot \sigma < X < \mu + 2 \cdot \sigma) \approx 95\%.$$

Souvent on dit qu’une certaine mesure “n’est pas dans la norme” si elle n’est pas comprise dans l’intervalle $[\mu - 2\sigma, \mu + 2\sigma]$ (voir exemple introductif no 1).

D’autres probabilités peuvent être calculées, par exemple

$$P(X > \mu + 1.960\sigma) = 2.5\%,$$

$$P(X > \mu + 1.645\sigma) = 5.0\%,$$

$$P(X > \mu + 1.282\sigma) = 10.0\%$$

En d’autres termes,

1.960 est le percentile 97.5%,

1.645 est le percentile 95%,

1.282 est le percentile 90%

de la distribution de Gauss standard (telle que $\mu = 0$ et $\sigma = 1$).