Notion de variable aléatoire et distribution

Soit X une variable aléatoire et n réalisations $\{x_i\}$ de cette variable.

Imaginons des étudiants qui font des expériences de calorimétrie pour mesurer la capacité thermique de l'eau[1]. Les différents groupes mesurent les valeurs suivantes : {5100; 4230; 3750; 4560; 3980} J/K/kg. Que vaut alors la capacité ? Nous donnerons dans ce chapitre une réponse à cette question. Elle sera de nature probabiliste.

Centre de la distribution des $\{x_i\}$

1/ le mode → valeur la plus représentée
2/ Médiane → sépare la distribution en 2 parties égales
3/ La moyenne :

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_i + \ldots + x_n}{n} \quad \text{soit} \quad \boxed{\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}}$$

Pour la capacité thermique de l'eau nous obtenons :

$$\bar{c} = \frac{5100 + 4230 + 3750 + 4560 + 3980}{5} = 4324\ J/K/kg$$

la moyenne géométrique :

$$\bar{x} = \sqrt[n]{\prod x_i}$$

Par exemple, pour deux températures 20°C et 40°C, la moyenne géométrique est $\sqrt{20\,°C \cdot 40\,°C} \simeq 28{,}3\,°C$ alors que la moyenne arithmétique est 30°C. Dans la pratique on constate que la moyenne arithmétique est mieux adaptée.

Dispersion de la distribution des $\{x_i\}$

Etendue → différence entre min et max → sensible aux valeurs extrêmes → pas toujours représentative

Dans les faits, la grandeur la plus utilisée est l'***écart-type*** :

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

Pour l'écart-type de la capacité thermique de l'eau nous obtenons :

$$s_c = \sqrt{\frac{(5100-4324)^2 + (4230-4324)^2 + (3750-4324)^2 + (4560-4324)^2 + (3980-4324)^2}{4}}$$

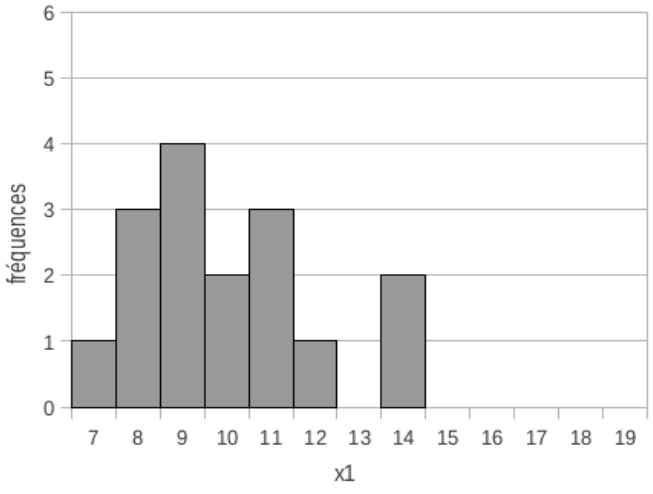
soit $s_c \simeq 530\,J/K/kg$

Si n au lieu de n-1 → écart quadratique moyen…. Si n grand → aucun impact

Exemples de distributions

Cas 1 :

	x_1
x_1^1	11
x_1^2	9
x_1^3	10
x_1^4	14
x_1^5	11
x_1^6	8
x_1^7	9
x_1^8	12
x_1^9	7
x_1^{10}	8
x_1^{11}	8
x_1^{12}	9
x_1^{13}	11
x_1^{14}	14
x_1^{15}	10

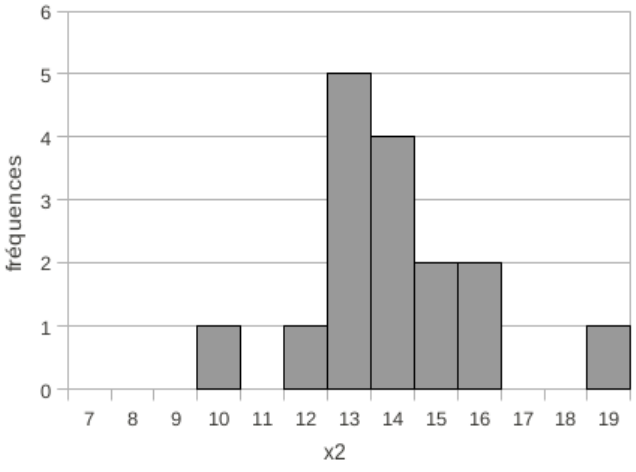


moyenne = 10 **écart-type= 2,07**

mode= 9 étendue= 7

Cas 2 :

	x_2
x_2^1	15
x_2^2	13
x_2^3	12
x_2^4	13
x_2^5	14
x_2^6	13
x_2^7	16
x_2^8	19
x_2^9	13
x_2^{10}	14
x_2^{11}	10
x_2^{12}	16
x_2^{13}	14
x_2^{14}	15
x_2^{15}	13
x_2^{16}	14



moyenne = 14 **écart-type= 2,00**

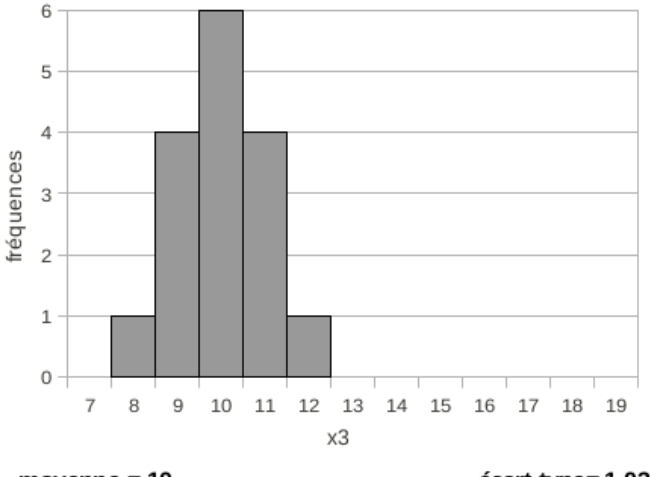
mode= 13 étendue= 9

médiane= 14 écart quadratique moyen= 1,94

La moyenne n'est pas toujours la plus représentée (cas 1 et 2)
Peut même ne pas être représentée

Cas 3 :

	x_3
x_3^1	10
x_3^2	10
x_3^3	12
x_3^4	11
x_3^5	9
x_3^6	8
x_3^7	10
x_3^8	9
x_3^9	9
x_3^{10}	11
x_3^{11}	9
x_3^{12}	11
x_3^{13}	10
x_3^{14}	10
x_3^{15}	11
x_3^{16}	10



moyenne = 10 **écart-type= 1,03**

mode= 10 étendue= 4

médiane= 10 écart quadratique moyen= 1,00

Courbe symétrique → égalité de la moyenne et de la médiane

Echantillonnage

1.1. Echantillons représentatifs et échantillons biaisés

Le but principal de la statistique est de déterminer les caractéristiques d'une population donnée à partir de l'étude d'une partie de cette population, appelée échantillon.

> La façon de sélectionner l'échantillon est aussi importante que la manière de l'analyser.

Il faut que l'échantillon soit *représentatif* de la population.

L'*échantillonnage aléatoire* est le meilleur moyen d'y parvenir.

Un *échantillon aléatoire* est un échantillon *tiré au hasard* dans lequel tous les individus ont la *même chance* de se retrouver.

Dans le cas contraire, l'échantillon est *biaisé*.

Un petit échantillon représentatif est, de loin, préférable à un grand échantillon biaisé.

Exemple :

Nous désirons déterminer la taille moyenne des étudiants de 2^{e} candi. commu. (97-98) qui étaient présents au 1er cours de statistique, à partir d'un échantillon de 10 individus.

(la réponse exacte, pour la population totale de 86 étudiants, est de 174,0 cm).

Mus par une bonne intention, sachant que les garçons sont, en général, plus grands que les filles, nous choisissons un échantillon contenant autant de filles que de garçons.

Soient 5 filles et 5 garçons choisis au hasard :

Taille des filles (cm)	Taille des garçons (cm)
171	193
165	187
173	180
174	185
166	178

A partir de cet échantillon de 10 individus, nous obtenons une taille moyenne de 177,2 cm, soit 3,2 cm de plus que la valeur exacte.

Avons-nous procédé correctement au choix de l'échantillon, sachant que la population contient 51 filles et 35 garçons ?

Non, car chaque garçon avait plus de chances d'être choisi que chaque fille.

En effet, les 5 garçons étant tirés au hasard dans une population de 35 individus, chacun d'eux avait 5 chances sur 35 d'être choisi, soit une probabilité de $5/35 \cong 0{,}143$.

Les 5 filles étant choisies dans une population de 51 individus, chacune d'entre elles avait 5 chances sur 51 d'être choisie, soit une probabilité de $5/51 \cong 0{,}098$, donc nettement plus faible que pour les garçons.

Nous avons biaisé l'échantillon en faveur des garçons. Il n'est donc pas surprenant que nous obtenions un résultat trop élevé.

La manière correcte de procéder est de choisir au hasard dans toute la population, sans considération du sexe.

Un tel tirage au hasard a donné les tailles suivantes (en cm) :

187, 165, 180, 168, 165, 160, 174, 183, 168, 176

La moyenne de l'échantillon est de 172,6 cm.

Elle est plus proche de la valeur exacte (erreur de – 1,4 cm).

[En fait, vu les petits échantillons utilisés, le hasard aurait pu donner un résultat inverse. Ce sera beaucoup moins probable pour de grands échantillons. Le raisonnement est néanmoins valable en toute généralité].

Précision de la moyenne

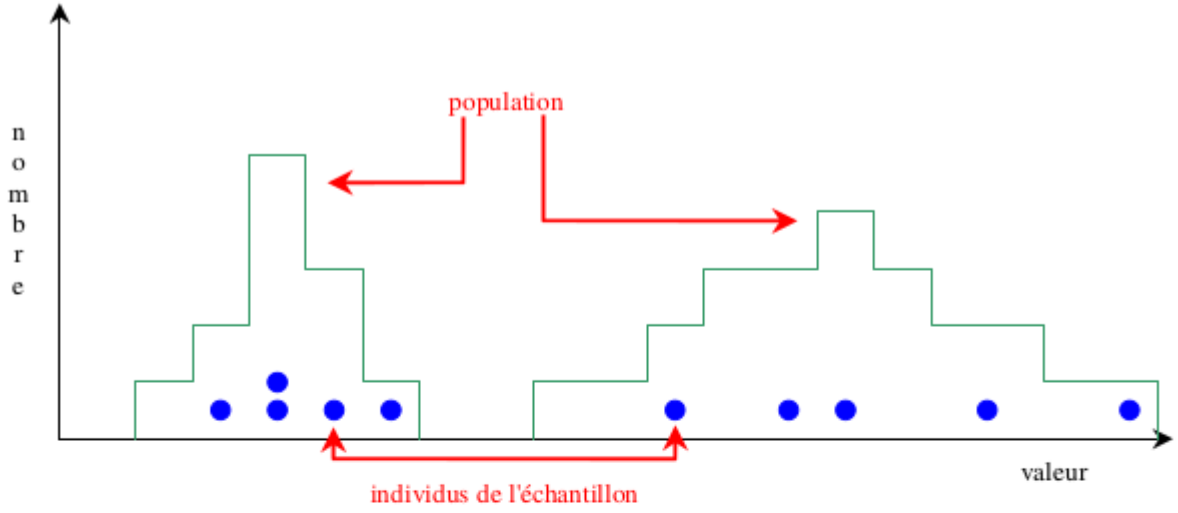
Nous supposons maintenant que notre échantillon est représentatif de la population.

La moyenne sur l'échantillon est donc une estimation de la moyenne sur la population.

Nous désirons savoir quelle est la précision de cette estimation, afin de connaître de quelle quantité la vraie valeur est susceptible de s'écarter de notre estimation.

En fait, la précision va dépendre :

- de la taille de l'échantillon
- de la dispersion de la population



Dans une population peu dispersée, toutes les valeurs de l'échantillon seront forcément proches de la moyenne.

Dans une population plus dispersée, les valeurs de l'échantillon seront généralement plus éloignées de la moyenne. La moyenne de l'échantillon pourra donc s'écarter plus fortement de celle de la population.

Soient:

- n le nombre d'individus dans l'échantillon,
- σ l'écart type de la population

Alors, la précision de la moyenne peut être mesurée par un écart type sur la moyenne :

$$\sigma\left(\overline{X}\right)=\frac{\sigma}{\sqrt{n}}$$

THÉORÈME CENTRAL LIMITE :

Nous prélevons au sein d'une population des échantillons aléatoires de taille n, la moyenne de l'échantillon $\bar{x}$ varie autour de la moyenne de la population μ avec un écart-type égal à σ/√n, où σ est l'écart-type de la population.

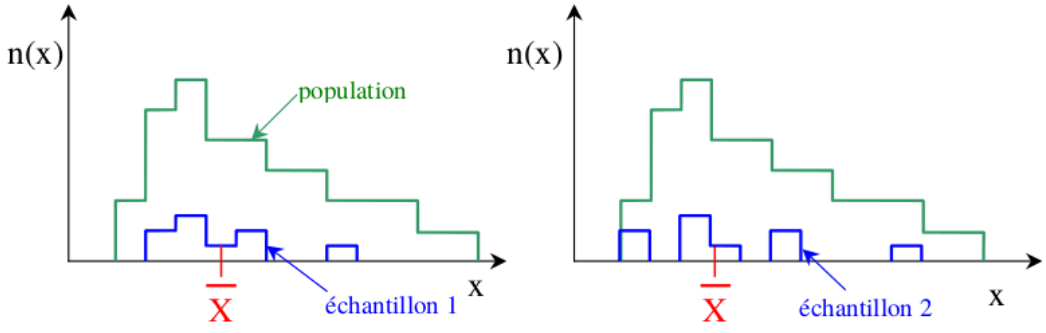
Quand n croît la distribution d'échantillonnage de $\bar{x}$ est de plus en plus concentrée autour de μ et devient de plus en plus proche d'une distribution de Gauss.

Supposons que nous analysions une population quelconque à partir d'un ensemble d'échantillons.

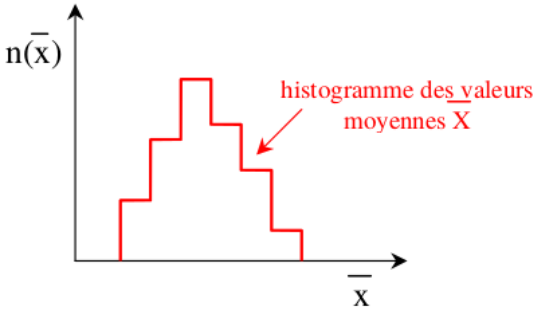
Pour chacun de ces échantillons, nous calculons une valeur moyenne $\overline{X}$ qui est une estimation de la moyenne de la population μ.

Bien entendu, les estimations $\overline{X}$ différeront généralement de la vraie moyenne μ.

Nous désirons savoir comment les différentes déterminations $\overline{X}$ vont se distribuer autour de la vraie moyenne μ.



Traçons l'histogramme des valeurs moyennes, c'est-à-dire le nombre d'échantillons pour lesquels la valeur moyenne $\overline{X}$ prend une certaine valeur (se situe dans une certaine classe).



$$\sigma\left(\overline{X}\right) = \frac{\sigma}{\sqrt{n}}$$

La précision sur la valeur moyenne sera donc d'autant meilleure que :

1. la population sera peu dispersée (σ petit)
2. l'échantillon sera grand (*n* grand)

La présence d'une racine carrée au dénominateur implique que :

- pour une précision 2 fois meilleure, il faut un échantillon 4 fois plus grand.
- pour une précision 10 fois meilleure, il faut un échantillon 100 fois plus grand.

→ la précision coûte cher !

1. Dans la population de 51 filles de 2e candi communication, la taille moyenne est de

$$\mu = 167{,}9 \text{ cm}$$

(nous noterons μ la valeur moyenne – généralement inconnue – pour la population et $\overline{X}$ la valeur moyenne pour l'échantillon)

L'écart type sur la taille est de:

$$\sigma = 5{,}3 \text{ cm}$$

Si on estime la taille moyenne à partir d'un échantillon de 4 personnes, on aura une précision (écart type) sur la moyenne de

$$\sigma\left(\overline{X}\right) = \frac{5{,}3}{\sqrt{4}} = \frac{5{,}3}{2} = 2{,}65 \text{ cm}$$

A partir d'un échantillon de 10 personnes, l'écart type serait de :

$$\sigma\left(\overline{X}\right) = \frac{5{,}3}{\sqrt{10}} \cong 1{,}7 \text{ cm}$$

2. Nous désirons déterminer la taille moyenne des hommes belges âgés d'une vingtaine d'années.

Nous disposons d'un échantillon de 35 étudiants de 2e candi communication.

Si cet échantillon est représentatif, sa taille moyenne est une estimation de celle de la population en question.

Elle est de 182,9 cm.

Pour estimer la précision de cette moyenne, il faudrait connaître l'écart type de la taille pour toute la population considérée, ce qui n'est pas le cas.

Si notre échantillon n'est pas trop petit (en principe, au moins 100 individus), nous pouvons remplacer l'écart type σ de la population par l'écart type s de l'échantillon.

Dans ce cas, il vaut $s = 6{,}7$ cm

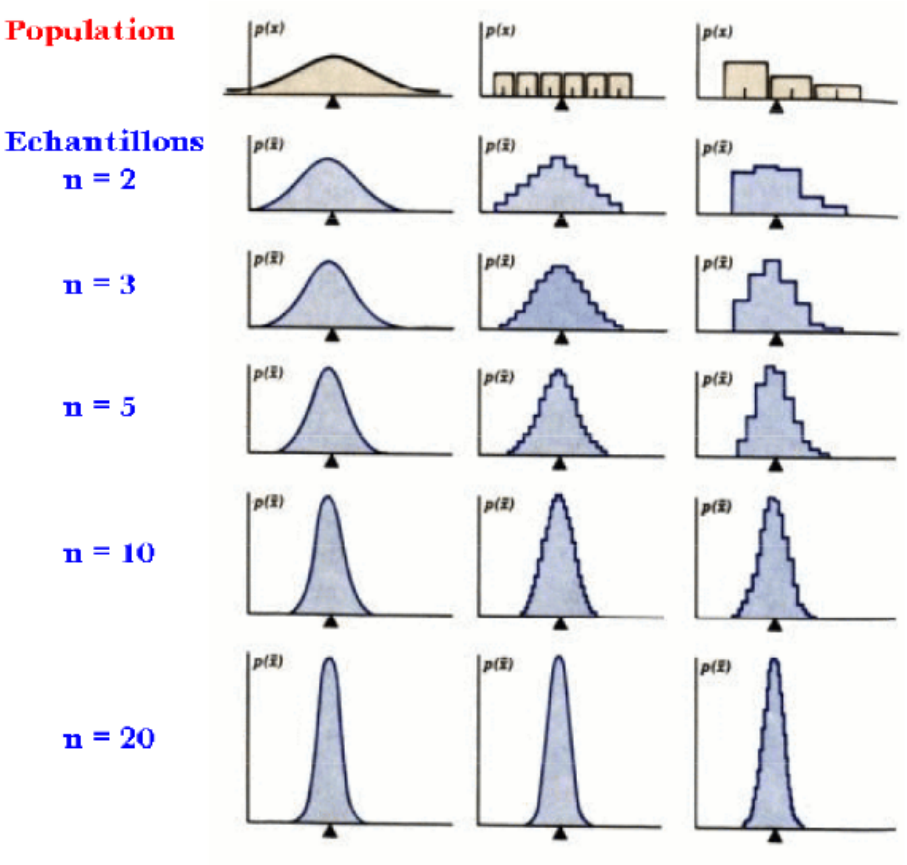
La précision sur la moyenne serait donc de :

$$\sigma\left(\overline{X}\right) = \frac{6{,}7}{\sqrt{35}} \cong 1{,}1 \text{ cm}$$

Si l'écart type de la grandeur analysée dans la population n'est pas connu, on peut le remplacer par l'écart type calculé dans l'échantillon, pour autant que cet échantillon soit suffisamment grand.

$$\sigma\left(\overline{X}\right) \cong \frac{s}{\sqrt{n}} \qquad (si \quad n \geq 100)$$

La figure suivante montre l'histogramme des valeurs moyennes $\overline{X}$ pour des échantillons de tailles croissantes tirés des populations indiquées sur la première ligne.



> Lorsque la taille de l'échantillon est suffisamment grande, ($n \geq 10$) la distribution de la moyenne a une forme approximativement normale.

L'écart type sur la moyenne est:

$$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

Quelle que soit la population sous-jacente, si on utilise des échantillons suffisamment grands (au moins 10 à 20 individus), la précision de la valeur moyenne peut être calculée à partir de la loi normale.

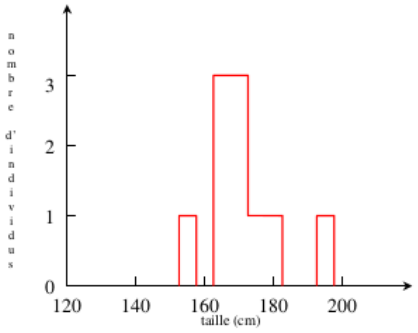
Loi normale ou de Gauss

Supposons que nous tirions des échantillons aléatoires d'une population dont la taille moyenne est de 170 cm, avec un écart type de 10 cm.

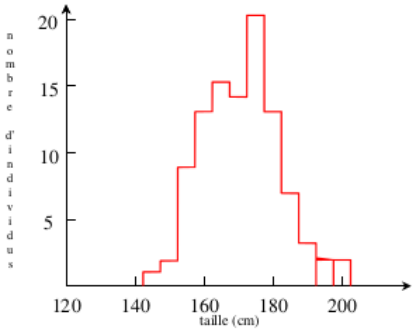
Traçons l'histogramme de la taille, avec des classes de 5cm de large.

Examinons l'aspect de ces histogrammes.

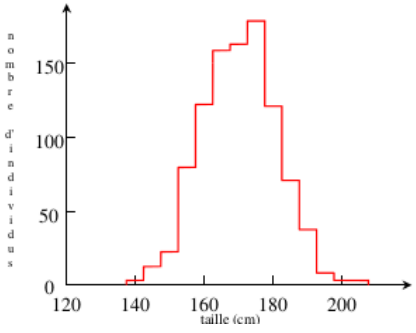
Echantillon de 10 individus



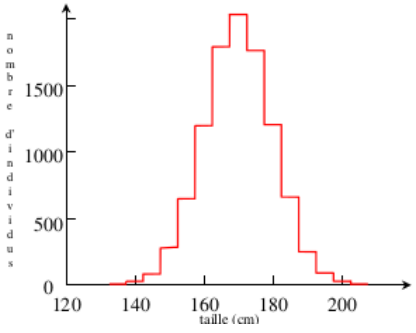
Echantillon de 100 individus



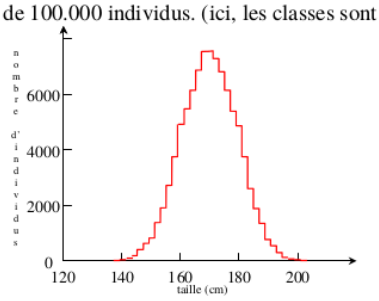
Echantillon de 1000 individus



Echantillon de 10.000 individus

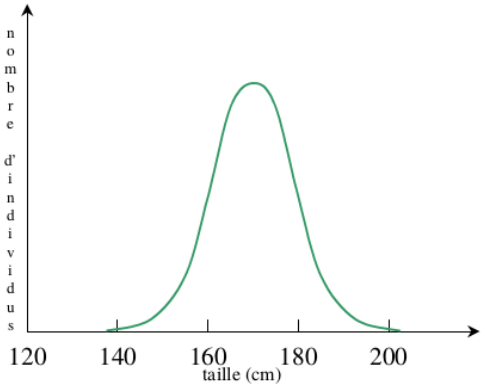


Echantillon de 100.000 individus. (ici, les classes sont de 2 cm)



Au fur et à mesure que la taille de l'échantillon augmente (et que la taille des classes diminue), l'histogramme devient de plus en plus régulier et se rapproche d'une courbe en cloche, appelée loi normale.

Loi normale



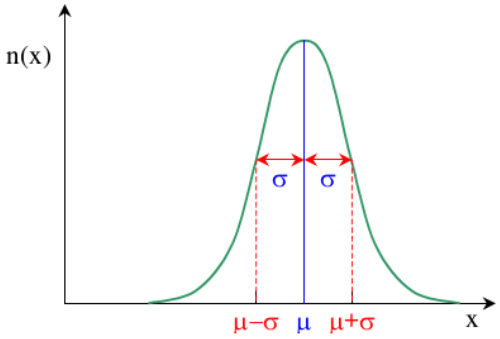
La loi normale est la loi statistique la plus répandue et la plus utile.

Elle représente beaucoup de phénomènes aléatoires.

De plus, de nombreuses autres lois statistiques peuvent être approchées par la loi normale, tout spécialement dans le cas des grands échantillons.

Son expression mathématique est la suivante:

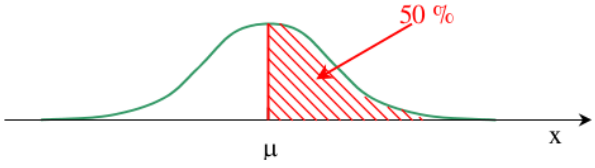
$$n(x) = \frac{n}{\sqrt{2\pi}\ \sigma}\ e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



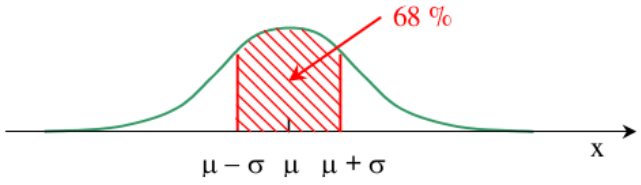
- μ est la moyenne
- σ l'écart type
- *n* le nombre total d'individus dans l'échantillon
- *n(x)* le nombre d'individus pour lesquels la grandeur analysée a la valeur *x*.

Lorsque la distribution des individus dans une population obéit à la loi normale, on trouve :

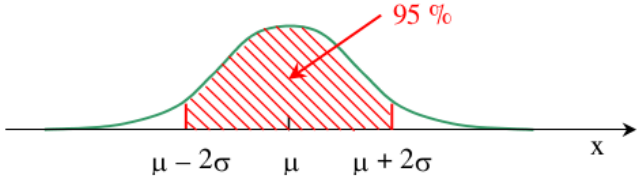
A. 50 % des individus en-dessous de la moyenne μ et 50 % au-dessus (la loi normale est symétrique)



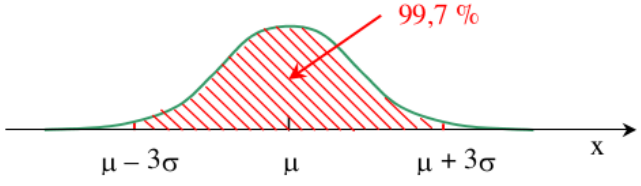
B. 68 % des individus entre μ–σ et μ+σ



C. 95 % des individus entre μ-1,96σ et μ+1,96σ, que nous arrondirons à l'intervalle [μ–2σ, μ+2σ]



D. 99,7 % des individus entre μ–3σ et μ+3σ (il y a donc très peu de chances qu'un individu s'écarte de la moyenne de plus de 3σ).

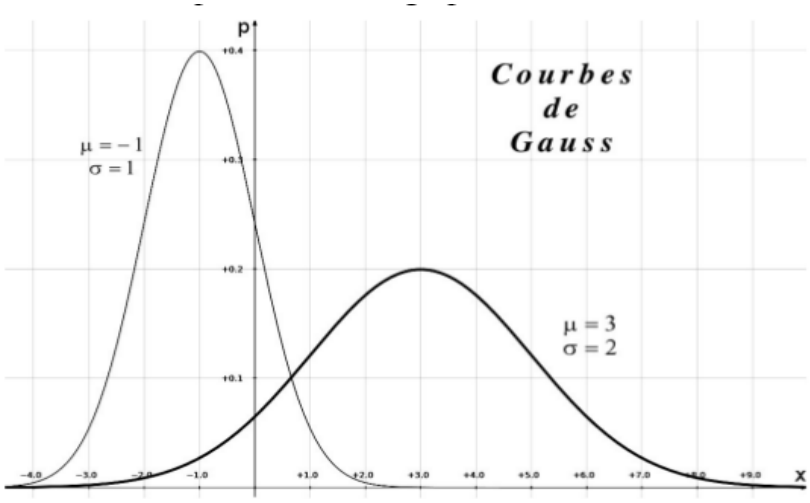


Calcul des probabilités

Pour calculer les probabilités associées à la loi normale, on utilise généralement la loi normale réduite : c'est une loi normale pour laquelle $\mu = 0$ et $\sigma = 1$.

La table suivante permet de déterminer la probabilité que la variable x s'écarte de la moyenne μ de plus de $z_0 \times \sigma$ vers le haut.

Pour obtenir z_0, on calcule l'écart par rapport à la moyenne : $\delta = x - \mu$, puis on divise par l'écart type : $z_0 = \frac{\delta}{\sigma}$

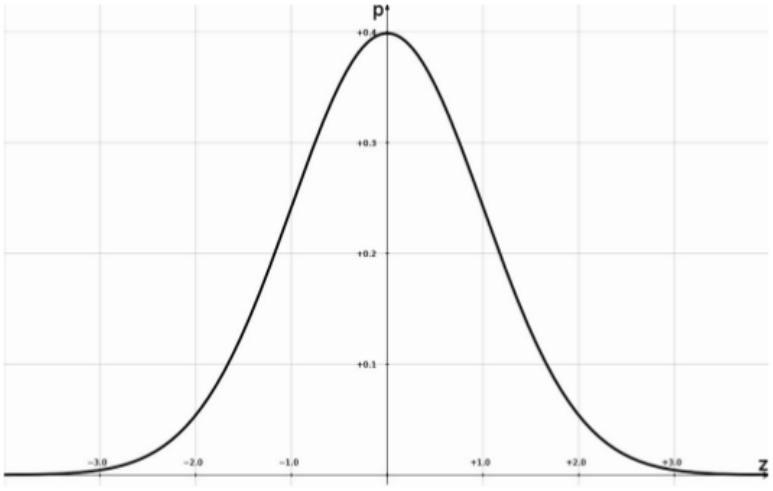


Loi normale standard

C'est la distribution normale de Gauss centrée et réduite. Pour le recentrage nous soustrayons la moyenne : $x' = x - \mu$. Pour la réduction nous divisons par l'écart-type :

$$z = \frac{x - \mu}{\sigma}$$ d'où : $$p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

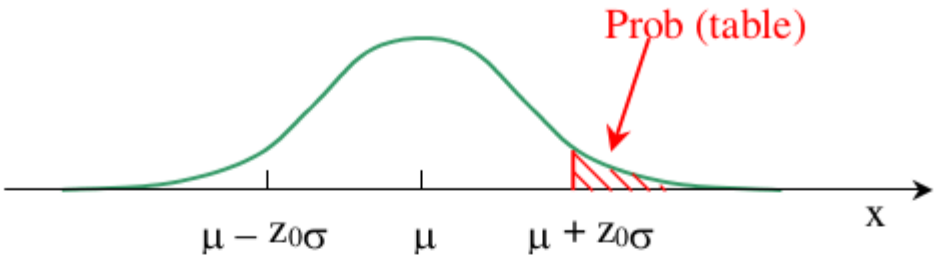
Nous avons alors une distribution normale de moyenne nulle et d'écart-type égale à un :



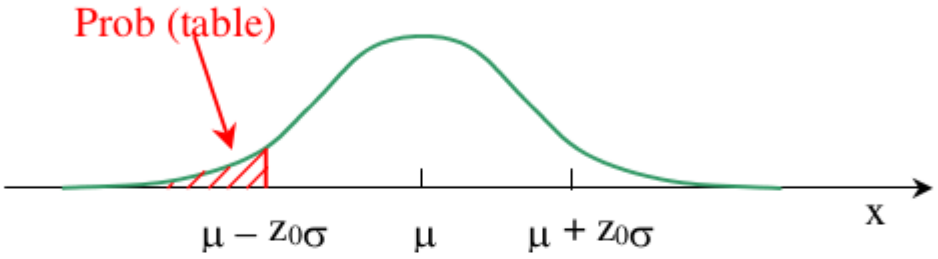
Z_o	2ème décimale de z_o									
	0	1	2	3	4	5	6	7	8	9
0.0	.500	.496	.492	.488	.484	.480	.476	.472	.468	.464
0.1	.460	.456	.452	.448	.444	.440	.436	.433	.429	.425
0.2	.421	.417	.413	.409	.405	.401	.397	.394	.390	.386
0.3	.382	.378	.374	.371	.367	.363	.359	.356	.352	.348
0.4	.345	.341	.337	.334	.330	.326	.323	.319	.316	.312
0.5	.309	.305	.302	.298	.295	.291	.288	.284	.281	.278
0.6	.274	.271	.268	.264	.261	.258	.255	.251	.248	.245
0.7	.242	.239	.236	.233	.230	.227	.224	.221	.218	.215
0.8	.212	.209	.206	.203	.200	.198	.195	.192	.189	.187
0.9	.184	.181	.179	.176	.174	.171	.169	.166	.164	.161
1.0	.159	.156	.154	.152	.149	.147	.145	.142	.140	.138
1.1	.136	.133	.131	.129	.127	.125	.123	.121	.119	.117
1.2	.115	.113	.111	.109	.107	.106	.104	.102	.100	.099
1.3	.097	.095	.093	.092	.090	.089	.087	.085	.084	.082
1.4	.081	.079	.078	.076	.075	.074	.072	.071	.069	.068
1.5	.067	.066	.064	.063	.062	.061	.059	.058	.057	.056
1.6	.055	.054	.053	.052	.051	.049	.048	.047	.046	.046
1.7	.045	.044	.043	.042	.041	.040	.039	.038	.038	.037
1.8	.036	.035	.034	.034	.033	.032	.031	.031	.030	.029
1.9	.029	.028	.027	.027	.026	.026	.025	.024	.024	.023
2.0	.023	.022	.022	.021	.021	.020	.020	.019	.019	.018
2.1	.018	.017	.017	.017	.016	.016	.015	.015	.015	.014
2.2	.014	.014	.013	.013	.013	.012	.012	.012	.011	.011
2.3	.011	.010	.010	.010	.010	.009	.009	.009	.009	.008
2.4	.008	.008	.008	.008	.007	.007	.007	.007	.007	.006
2.5	.006	.006	.006	.006	.006	.005	.005	.005	.005	.005
2.6	.005	.005	.004	.004	.004	.004	.004	.004	.004	.004
2.7	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003
2.8	.003	.002	.002	.002	.002	.002	.002	.002	.002	.002
2.9	.002	.002	.002	.002	.002	.002	.002	.001	.001	.001

Quelques cas concrets sont illustrés ci-dessous.

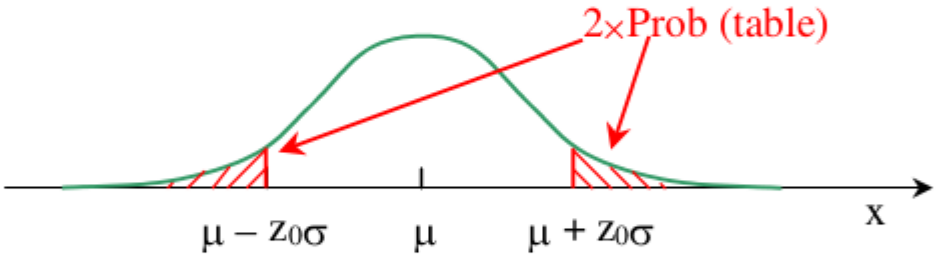
1) $x > \mu + z_0\sigma$



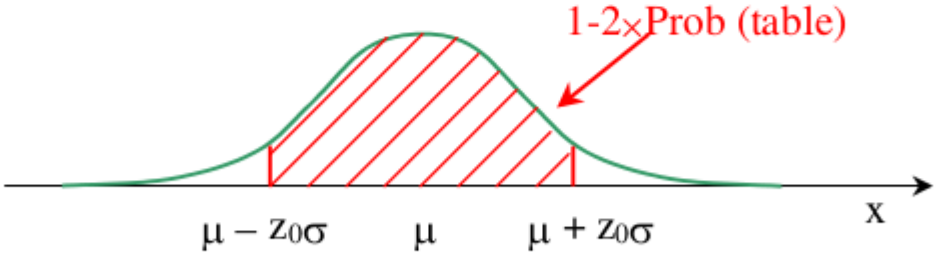
2) $x < \mu - z_0\sigma$



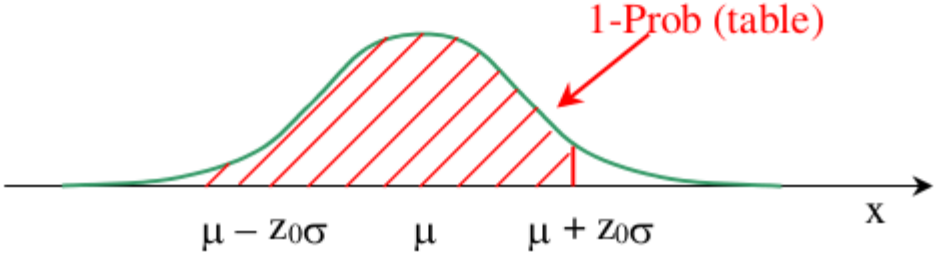
3) x plus éloigné de μ que $z_0\sigma$



4) x plus proche de μ que $z_0\sigma$



5) $x < \mu + z_0\sigma$



Exemples :

Le poids des tomates produites par un jardinier obéit à une loi normale de moyenne 200 gr et d'écart type 40 gr.

a. Calculez la probabilité que le poids d'une tomate excède 250 gr.

Solution:

$$\delta = 250 - 200 = 50\,\text{gr}$$

$$z_0 = \frac{\delta}{\sigma} = \frac{50}{40} = 1{,}25$$

$$\text{Prob} = 0{,}106 = 10{,}6\,\%$$

b. Calculez la probabilité que le poids d'une tomate soit inférieur à 100 gr.

Solution:

$$\delta = 100 - 200 = -100\,\text{gr}$$

la loi normale est symétrique → on ne s'occupe pas du signe

$$z_0 = \frac{\delta}{\sigma} = \frac{100}{40} = 2{,}5$$

moins de 100 gr: on s'écarte donc de la valeur moyenne μ = 200 gr de plus de $z_0 \times \sigma$

$$\text{Prob} = 0{,}006 = 0{,}6\,\%$$

c. Calculez la probabilité que le poids d'une tomate soit inférieur à 230 gr.

Solution:

$$\delta = 230 - 200 = 30\,\text{gr}$$

$$z_0 = \frac{\delta}{\sigma} = \frac{30}{40} = 0{,}75$$

L'intervalle (< 230 gr) considéré contient la valeur moyenne (200 gr) → on prend 1 – Prob(table):

$$\text{Prob} = 1 - 0{,}227 = 0{,}773 = 77{,}3\,\%$$

d. Calculez la probabilité que le poids d'une tomate ne s'écarte pas de la valeur moyenne de plus de 20 gr.

Solution: on calcule d'abord la probabilité que le poids s'écarte de plus de 20 gr, vers le haut ou vers le bas :

$$\delta = 20\,\text{gr} \qquad \sigma = 40$$

$$z_0 = \frac{\delta}{\sigma} = \frac{20}{40} = 0{,}5$$

$$\text{Prob} = 0{,}309 = 30{,}9\,\%$$

On doit multiplier par 2 car on considère les deux côtés → Prob = 2 × 0,309 = 0,618

On a donc une prob. de 0,618 que le poids s'écarte de μ de plus de 20 gr, et donc une prob. 1-0,618 que le poids ne s'écarte pas de plus de 20 gr.

Réponse: 0,382 = 38,2 %

Intervalles de confiance

Nous avons vu que la moyenne $\overline{X}$ d'un échantillon aléatoire permet d'estimer la vraie moyenne μ de la population.

Nous voudrions estimer également la précision de cette moyenne, c'est-à-dire donner une marge d'erreur ou un intervalle de confiance.

Nous pouvons utiliser les tables de la loi normale pour estimer ces intervalles de confiance.

En général nous adopterons l'intervalle de confiance à 95%, soit à $2\sigma(\overline{X})$.

Nous pourrons donc écrire, soit:

$$\mu = \overline{X} \pm 2\sigma(\overline{X})$$

soit, plus explicitement:

Il y a 95 chances sur 100 que μ se situe entre

$$\overline{X} - 2\sigma(\overline{X}) \quad \text{et} \quad \overline{X} + 2\sigma(\overline{X})$$

Si nous tirons une série d'échantillons aléatoires de la population, dans 19 cas sur 20 (en moyenne), μ se trouvera dans l'intervalle de confiance $\overline{X} \pm 2\sigma(\overline{X})$.

1. La taille moyenne d'un échantillon de 51 filles de 2ème candi. commu. est de 167,9 cm.
   L'écart type de cet échantillon est de 5,3 cm.

   Si nous supposons que cet échantillon est représentatif de la taille des filles belges âgées d'une vingtaine d'années, nous pouvons calculer la taille moyenne de cette population, avec sa marge d'erreur :

   $$n = 51 \qquad \overline{X} = 167{,}9 \qquad s = 5{,}3$$
   $$\sigma(\overline{X}) = \frac{5{,}3}{\sqrt{51}} = 0{,}74$$
   $$2\sigma(\overline{X}) = 1{,}48 \cong 1{,}5 \text{ cm}$$

   Avec 95 % de confiance, nous pouvons donc dire que la taille moyenne de la population vaut:

   $$\mu = 167{,}9 \pm 1{,}5 \text{ cm}$$

   ce qui revient à dire qu'il y a 95 chances sur 100 pour que la taille moyenne des filles belges de 20 ans se situe entre 166,4 et 169,4 cm.

2. La taille moyenne d'un échantillon de 35 garçons de 2ème candi. commu. est de 182,9 cm

   En supposant de même l'échantillon représentatif, nous pouvons donner un intervalle de confiance pour la taille des garçons belges de 20 ans.

   $$n = 35 \qquad \overline{X} = 182{,}9 \qquad s = 6{,}7$$
   $$\sigma(\overline{X}) = \frac{6{,}7}{\sqrt{35}} = 1{,}13$$
   $$2\sigma(\overline{X}) = 2{,}26 \cong 2{,}3 \text{ cm}$$

   Avec 95 % de confiance, on a donc:

   $$\mu = 182{,}9 \pm 2{,}3 \text{ cm}$$

Statistiques faibles ..

Le théorème central limite s'applique dans la limite des grands nombres, mais on sait l'adapter pour n petit grâce aux *coefficients de Student*. En effet on ne peut pas appliquer directement ce théorème car ne connaissant pas σ nous l'estimons avec s. Du fait d'une statistique faible il y a alors un élargissement connu :

$$\mu = \bar{x} \pm \boldsymbol{t} \cdot \frac{s}{\sqrt{n}}$$

Le coefficient de Student t dépend de n et inclut la *confiance* que nous voulons donner au résultat. Si la confiance est de 95%, nous avons 95 chances sur 100 que μ soit compris entre $\bar{x} - t \cdot s/\sqrt{n}$ et $\bar{x} + t \cdot s/\sqrt{n}$. Les valeurs de t sont lues dans la table page 166.

Nous reconnaissons ici la notion d'***incertitude*** Δx [4]:

$$x = \bar{x} \pm \Delta x \quad \text{avec} \quad \Delta x = t \cdot \frac{s}{\sqrt{n}}$$

Δx est aussi appelée l'*incertitude absolue* et $\Delta x / |\bar{x}|$ l'*incertitude relative*.

Reprenons l'expérience de calorimétrie décrite page 1, supposons que nous voulions maintenant connaître la capacité thermique de l'eau avec une confiance de 95%. Nous trouvons dans la table pour quatre *d*egrés *d*e *l*iberté (*ddl=n-1*) *t=2,78*.

D'où : $c = \bar{c} \pm t \cdot s_c/\sqrt{n} = 4320 \pm 660\ J/K/kg$ à 95%.

Ici suite à la dispersion des valeurs mesurées par les étudiants $\Delta c/\bar{c} \simeq 15\%$. Les mesures en calorimétrie ne sont pas très précises. La valeur attendue, ici connue, est bien dans l'intervalle : $3660 < 4180 < 4980$.

A. Coefficients de Student

Coefficient de Student t		Confiance (%)								
		50	**80**	**90**	**95**	**98**	**99**	**99,5**	**99,8**	**99,9**
Degrés de liberté (taille de l'échantillon moins le nombre de paramètres)	**1**	1,00	3,08	6,31	12,7	31,8	63,7	127	318	637
	2	0,82	1,89	2,92	4,30	6,96	9,92	14,1	22,3	31,6
	3	0,76	1,64	2,35	3,18	4,54	5,84	7,45	10,2	12,9
	4	0,74	1,53	2,13	2,78	3,75	4,60	5,60	7,17	8,61
	5	0,73	1,48	2,02	2,57	3,36	4,03	4,77	5,89	6,87
	6	0,72	1,44	1,94	2,45	3,14	3,71	4,32	5,21	5,96
	7	0,71	1,41	1,89	2,36	3,00	3,50	4,03	4,79	5,41
	8	0,71	1,40	1,86	2,31	2,90	3,36	3,83	4,50	5,04
	9	0,70	1,38	1,83	2,26	2,82	3,25	3,69	4,30	4,78
	10	0,70	1,37	1,81	2,23	2,76	3,17	3,58	4,14	4,59
	11	0,70	1,36	1,80	2,20	2,72	3,11	3,50	4,02	4,44
	12	0,70	1,36	1,78	2,18	2,68	3,05	3,43	3,93	4,32
	13	0,69	1,35	1,77	2,16	2,65	3,01	3,37	3,85	4,22
	14	0,69	1,35	1,76	2,14	2,62	2,98	3,33	3,79	4,14
	15	0,69	1,34	1,75	2,13	2,60	2,95	3,29	3,73	4,07
	16	0,69	1,34	1,75	2,12	2,58	2,92	3,25	3,69	4,01
	17	0,69	1,33	1,74	2,11	2,57	2,90	3,22	3,65	3,97
	18	0,69	1,33	1,73	2,10	2,55	2,88	3,20	3,61	3,92
	19	0,69	1,33	1,73	2,09	2,54	2,86	3,17	3,58	3,88
	20	0,69	1,33	1,72	2,09	2,53	2,85	3,15	3,55	3,85
	22	0,69	1,32	1,72	2,07	2,51	2,82	3,12	3,50	3,79
	24	0,68	1,32	1,71	2,06	2,49	2,80	3,09	3,47	3,75
	26	0,68	1,31	1,71	2,06	2,48	2,78	3,07	3,43	3,71
	28	0,68	1,31	1,70	2,05	2,47	2,76	3,05	3,41	3,67
	30	0,68	1,31	1,70	2,04	2,46	2,75	3,03	3,39	3,65
	40	0,68	1,30	1,68	2,02	2,42	2,70	2,97	3,31	3,55
	50	0,68	1,30	1,68	2,01	2,40	2,68	2,94	3,26	3,50
	60	0,68	1,30	1,67	2,00	2,39	2,66	2,91	3,23	3,46
	70	0,68	1,29	1,67	1,99	2,38	2,65	2,90	3,21	3,44
	80	0,68	1,29	1,66	1,99	2,37	2,64	2,89	3,20	3,42
	90	0,68	1,29	1,66	1,99	2,37	2,63	2,88	3,18	3,40
	100	0,68	1,29	1,66	1,98	2,36	2,63	2,87	3,17	3,39
	200	0,68	1,29	1,65	1,97	2,35	2,60	2,84	3,13	3,34
	300	0,68	1,28	1,65	1,97	2,34	2,59	2,83	3,12	3,32
	500	0,67	1,28	1,65	1,96	2,33	2,59	2,82	3,11	3,31
	1000	0,67	1,28	1,65	1,96	2,33	2,58	2,81	3,10	3,30
	∞	**0,67**	**1,28**	**1,64**	**1,96**	**2,33**	**2,58**	**2,81**	**3,09**	**3,29**

Pour info …. pas nécessaire de retenir…

Théorème central limite : quelle valeur de n est assez grande ?

- Si les lois des X_i sont proches d'une loi normale alors pour $n \geq 4$ l'approximation donnée par le théorème central limite est bonne.
- Si les lois des X_i sont moyennement proches d'une loi normale (p. ex. loi uniforme) alors pour $n \geq 12$ l'approximation donnée par le théorème central limite est bonne.
- Si les lois des X_i ne sont pas proches d'une loi normale alors pour $n \geq 100$ l'approximation donnée par le théorème central limite sera bonne (par exemple fonction de densité très asymétrique).

Distribution de Gauss

Notion de distribution continue :

Ex : taille d'une population → Spectre continu de tailles

densité de probabilité $p(x)$ avec $p(x)dx$ la probabilité d'un évènement d'être entre x et $x+dx$

$$\int_{x=-\infty}^{+\infty} p(x)\,dx = 1$$

Calcul de la moyenne et de l'écart-type d'une distribution continue :

$$\mu = \int_{-\infty}^{+\infty} x \cdot p(x)\,dx \qquad \sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2\, p(x)\,dx$$

Courbe de Gauss

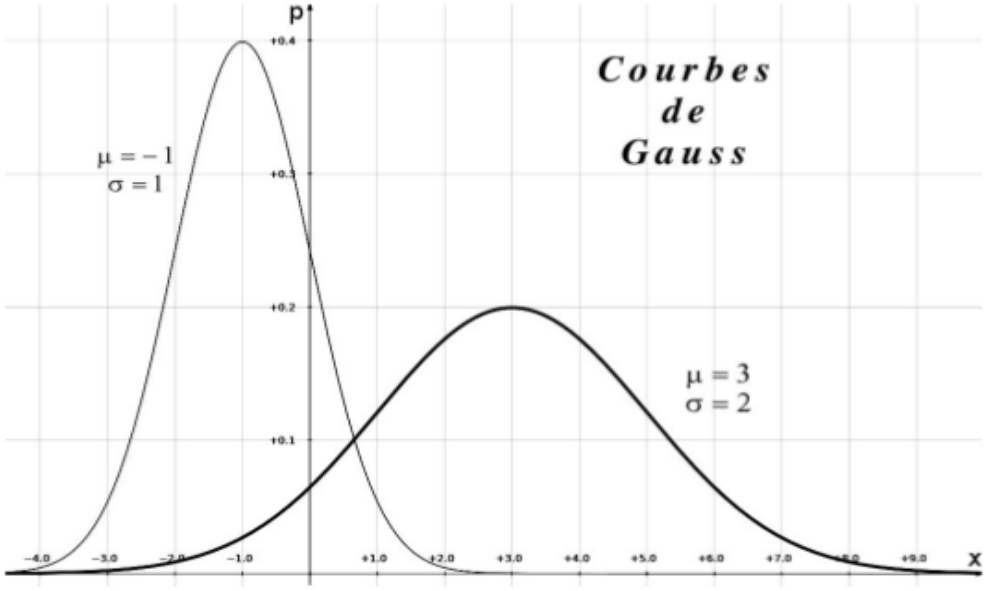
Le théorème central limite s'intéresse au cas où la taille de l'échantillon n est grande, et dans le cas limite où n tend vers l'infini nous considérons une distribution continue. Celle-ci est une gaussienne, l'expression mathématique est connue :

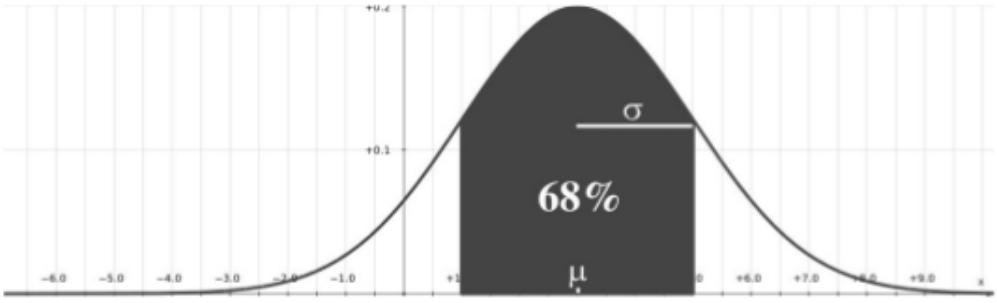
$$p(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Sachant que : $\int_{x=-\infty}^{+\infty} e^{-x^2}\,dx = \sqrt{\pi}$

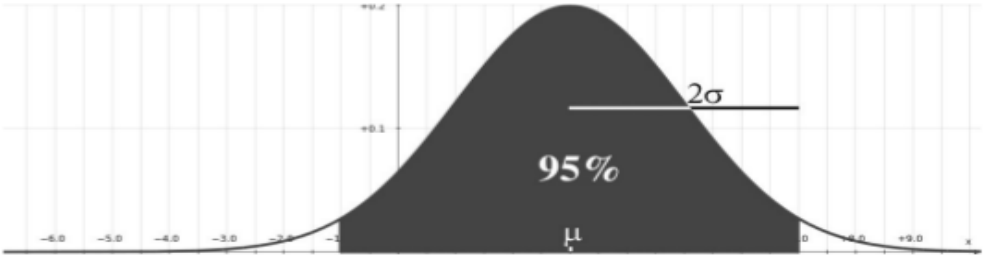
montrer que la distribution de probabilité Normale vérifie bien :

$$\int_{x=-\infty}^{+\infty} p(x)\,dx = 1$$

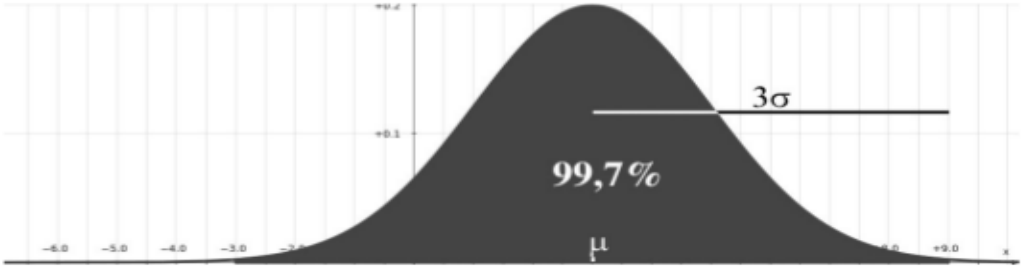




$$\int_{\mu-\sigma}^{\mu+\sigma} p(x)\,dx = 0{,}683\ldots \simeq 68\,\%$$



$$\int_{\mu-2\sigma}^{\mu+2\sigma} p(x)\,dx = 0{,}954\ldots \simeq 95\,\%$$



$$\int_{\mu-3\sigma}^{\mu+3\sigma} p(x)\,dx = 0{,}997\ldots > 99\,\%$$

3) Loi normale standard

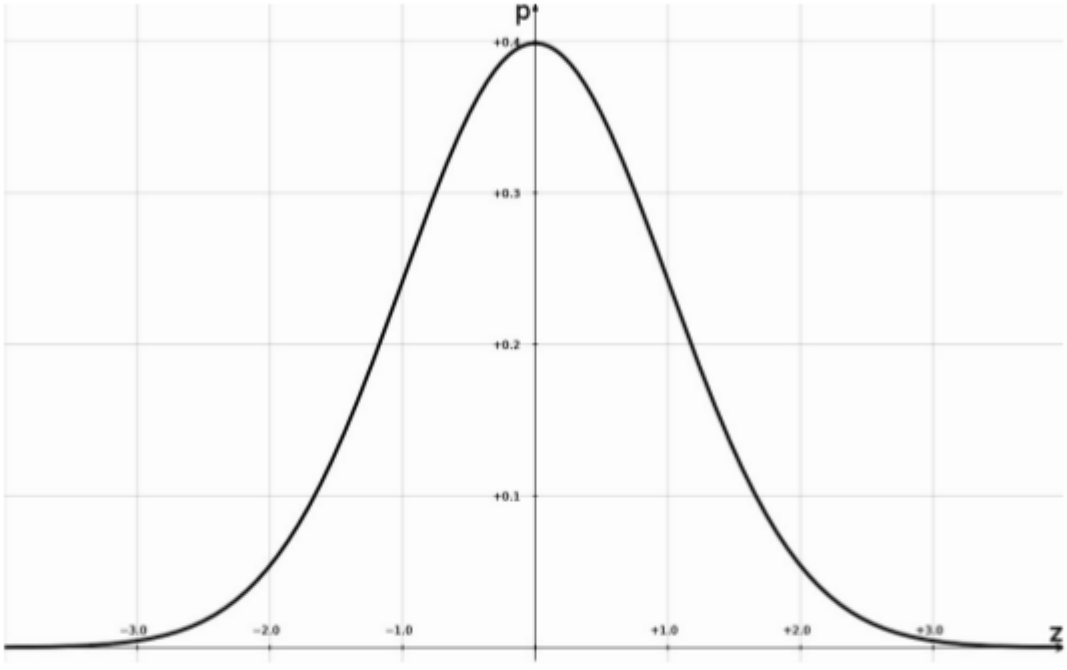
C'est la distribution normale de Gauss centrée et réduite. Pour le recentrage nous soustrayons la moyenne : $x' = x - \mu$. Pour la réduction nous divisons par l'écart-type :

$$z = \frac{x - \mu}{\sigma}$$

d'où :

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Nous avons alors une distribution normale de moyenne nulle et d'écart-type égale à un :

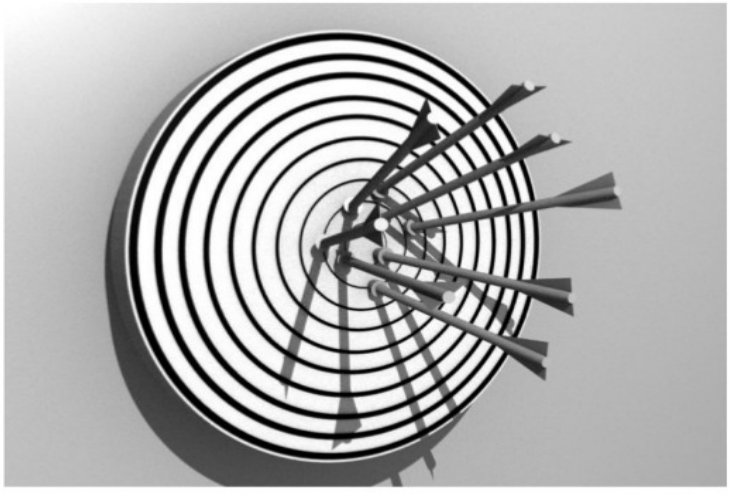


Incertitudes sur les mesures

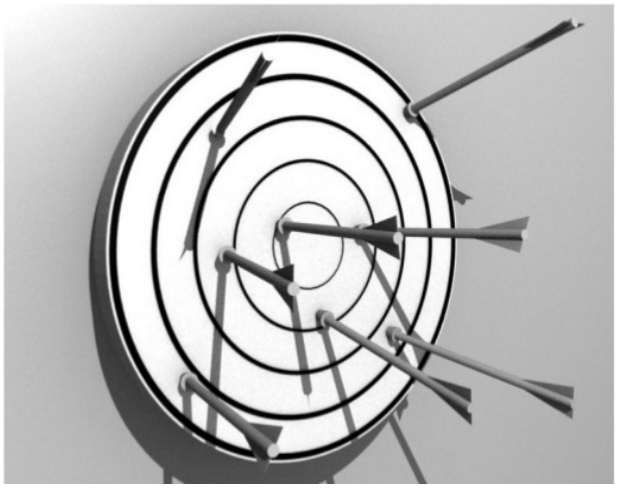
Qualité d'une mesure :

Fidélité → dispersion statistique faible
Justesse → pas d'erreur systématique (biais)
Bien résolue (système de mesure fin)

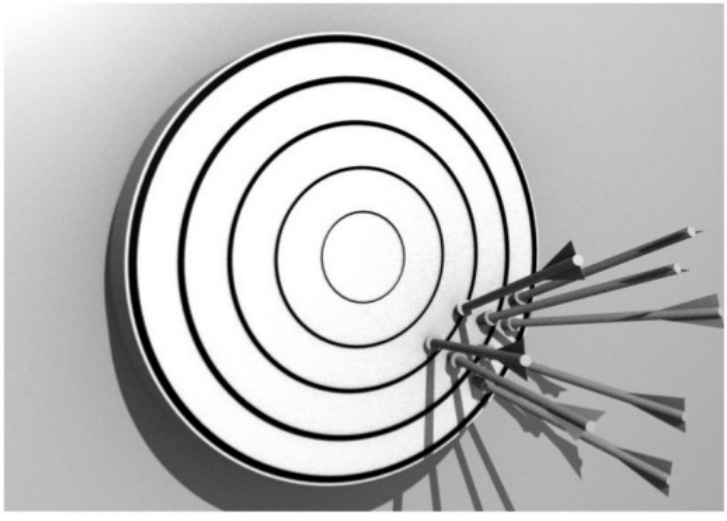
Mesure juste, fidèle et avec une bonne résolution :



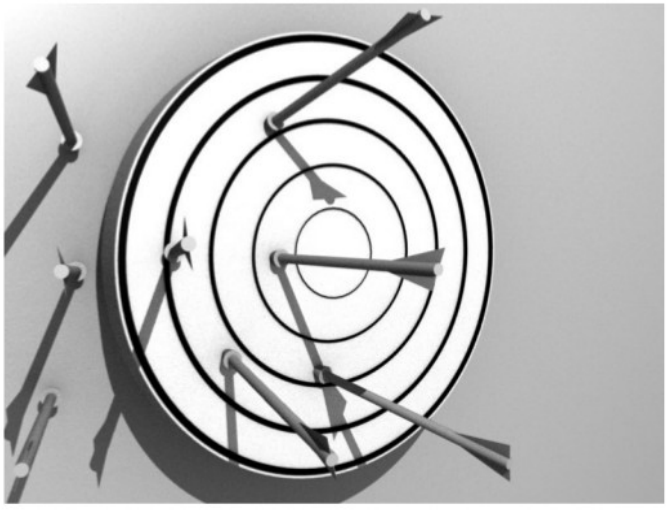
Mesure juste, mais peu fidèle et avec une faible résolution :



Mesure peu dispersée mais avec un biais, et mal résolue :



Mesure avec un biais, fortement dispersée et une faible résolution :



L'écart-type complet sera déterminé à partir des écarts de chaque source en ajoutant les carrés

$$\sigma = \sqrt{{\sigma_1}^2 + {\sigma_2}^2 + {\sigma_3}^2 + \ldots}$$

Propagation des Incertitudes et écarts types dans le cas de fonctions composées de plusieurs variables (aléatoires)

1) Formule de propagation des écart-types

Pour une approche générale du problème, soit une fonction f qui dépend de p variables indépendantes :

$$f(x_1, x_2, \ldots, x_j, \ldots, x_p)$$

A chacune de ces variables aléatoires est associée une valeur moyenne $\bar{x}_j$ et et un écart-type σ_j.

Que valent f et σ_f ?

La statistique donne la réponse et démontre la formule de propagation des écart-types :

$$\sigma_f^2 = \sum_{j=1}^{p} \left[\left(\frac{\partial f}{\partial x_j} \right)^2 \sigma_j^2 \right]$$

Plus utile, la propagation des incertitudes (éventuellement déterminées à partir d'écart types)

$$\Delta f^2 = \sum_{j=1}^{p} \left[\left(\frac{\partial f}{\partial x_j} \right)^2 \Delta x_j^2 \right]$$

Un exemple :

Soit un seau rempli d'un million de grains de sable. La masse d'un grain est de 10 mg à 1mg près. Qu'elle est la masse de sable contenue dans le seau ?

Pour notre seau : $M(m_1, m_2, \ldots, m_j, \ldots, m_p)$

avec $$M = \sum_{j=1}^{p} m_j$$

où nous appelons M la masse totale de sable dans le seau, m_j la masse de chaque grain et p le nombre de grains.

$$\Delta M^2 = \sum_{j=1}^{p} (\partial M / \partial m_j)^2 \Delta m_j^{\,2}$$

$$\partial M / \partial m_j = \partial m_1 / \partial m_j + \ldots + \partial m_j / \partial m_j + \ldots + \partial m_p / \partial m_j$$
$$\partial M / \partial m_j = 0 + \ldots + 1 + \ldots + 0 = 1$$

alors $$\Delta M^2 = \left(\sum_{j=1}^{p} 1^2 \right) \Delta m^2$$

d'où $\Delta M^2 = p \cdot \Delta m^2$ avec $\Delta m = \Delta m_j$ quelque soit *j*.

Finalement : $\Delta M = \sqrt{p} \cdot \Delta m = \sqrt{1000000} \cdot 0{,}001\, g$.

Le seau pèse donc dix kilos à un gramme près. La précision sur la masse du seau est donc de 0,01%. Naïvement, nous aurions pu penser que l'incertitude globale sur la masse du seau était la somme des incertitudes de chaque grain, nous aurions alors une incertitude absolue d'un kilo et une relative de 10%, ce qui est très différent de la réalité et ignorerait les compensations.

Ici la formule de propagation est très précise car nous avons un très grand nombre de grains. Elle est même exacte, dès les petits nombres, si la distribution de la masse des grains est gaussienne[10].

Corollaire des propriétés de la différenciation :

- Pour des sommes ou des différences les incertitudes absolues au carré s'ajoutent :

$$\Delta f^2 = \sum_{j=1}^{p} \Delta x_j^{\,2}$$

Par exemple si $d=x_2-x_1$ avec $\Delta x_2=\Delta x_1=1cm$ alors $\Delta d \simeq 1{,}4\,cm$.

- pour des produits ou des quotients les incertitudes relatives au carré s'ajoutent :

$$\left(\frac{\Delta f}{f}\right)^2 = \sum_{j=1}^{p} \left(\frac{\Delta x_j}{x_j}\right)^2$$

Par exemple si $R=U/I$ avec U et I à 1% alors R est connu à 1,4%.

Dans les cas plus complexes, il faut réaliser explicitement le calcul aux dérivées partielles.

Simplifications possibles selon les importances respectives des incertitudes

Par exemple comme les incertitudes s'ajoutent au carré, on peut considérer que l'incertitude la plus grande va l'emporter rapidement sur les autres. Dans l'exemple où $R=U/I$ si U est connu 1% et I à 0,1% alors R est connu à 1,005%≃1%, on peut négliger l'incertitude sur I.

Corrélations et dépendances

Pour illustrer, considérons un échantillon de quatre individus qui possèdent trois caractéristiques, la taille X_1, le poids X_2 et le mois de naissance X_3. A priori, nous nous attendons à une corrélation entre la taille et le poids : plus on est grand plus on a, en général, une masse importante (corrélation positive). Par contre, nous pouvons penser que le mois de naissance n'a aucune incidence sur le poids et la taille (X_3 non corrélée avec X_1 et X_2).

Coefficient de corrélation

Le coefficient de corrélation r permet d'identifier s'il y a une relation linéaire entre deux variables X_i et X_j :

$$r_{ij} = \frac{\sum_k [(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)]}{\sqrt{\sum_k [(x_{ik} - \bar{x}_i)^2]} \cdot \sqrt{\sum_k [(x_{jk} - \bar{x}_j)^2]}}$$

r varie entre *-1* et *+1*. Si $|r| = 1$ les variables sont parfaitement corrélées : $r = 1$ même sens de variation, $r = -1$ sens opposé. Si $r = 0$ il n'y a pas la moindre corrélation, les variables sont parfaitement indépendantes.

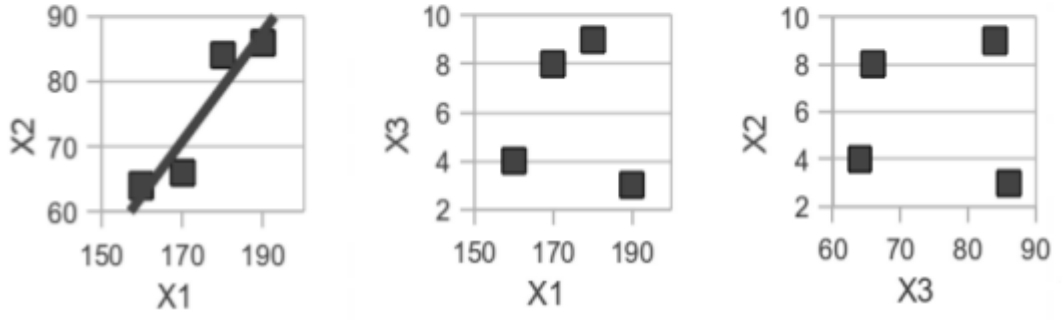
Calcul de r_{12} ,r_{13} et r_{23} :

	X_1 *(cm)*	X_2 *(kg)*	X_3	$x_1 - \bar{x}_1$	$x_2 - \bar{x}_2$	$x_3 - \bar{x}_3$	$(x_1 - \bar{x}_1)^2$
1	160	64	4	-15	-11	-2	225
2	170	66	8	-5	-9	2	25
3	180	84	9	5	9	3	25
4	190	86	3	15	11	-3	225
$\bar{x}$	175	75	6			Σ=	500

$(x_2-\bar{x}_2)^2$	$(x_3-\bar{x}_3)^2$	$(x_1-\bar{x}_1)$ $.(x_2-\bar{x}_2)$	$(x_1-\bar{x}_1)$ $.(x_3-\bar{x}_3)$	$(x_2-\bar{x}_2)$ $.(x_3-\bar{x}_3)$
121	4	165	30	22
81	4	45	-10	-18
81	9	45	15	27
121	9	165	-45	-33
404	26	420	-10	-2

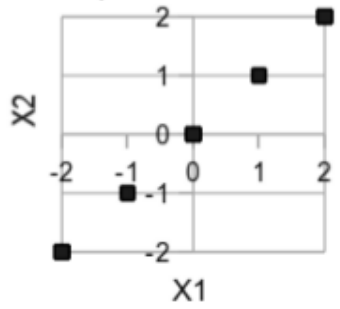
d'où : $r_{12}=\frac{420}{\sqrt{500}\sqrt{404}}\simeq 0{,}93$, $r_{13}\simeq -0{,}09$ et $r_{23}\simeq -0{,}02$.

r_{12} est proche de +1, nous avons donc une corrélation positive importante. r_{13} et r_{23} sont proches de zéro : X_3 est indépendante de X_1 et X_2 :

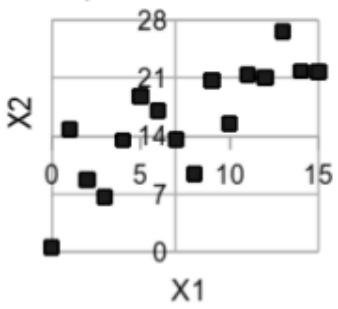


Exemples de nuages de points entre deux variables :

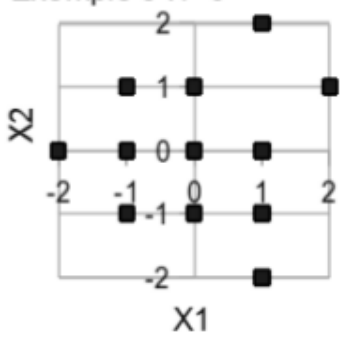
Exemple 1 : r=1



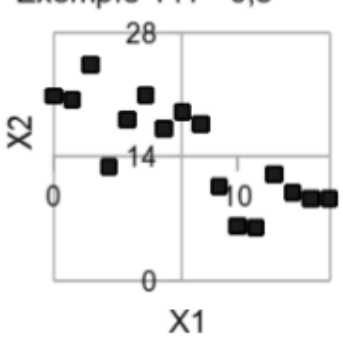
Exemple 2 : r=0,8



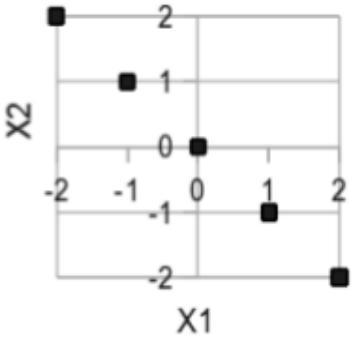
Exemple 3 : r=0



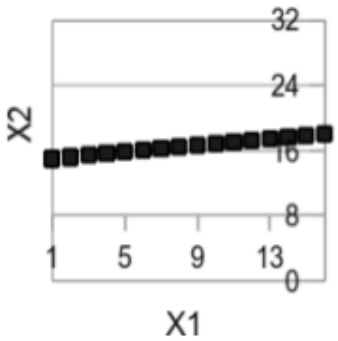
Exemple 4 : r=-0,8



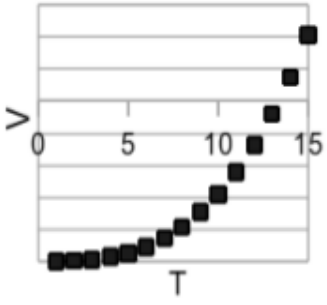
Exemple 5 : r=-1



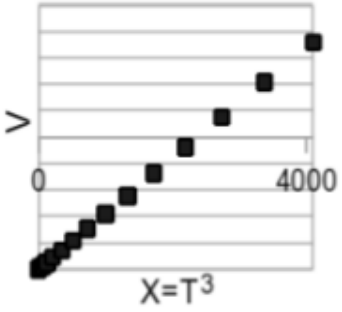
Exemple 6 : r=1



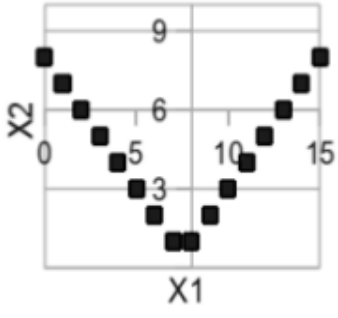
Exemple 7 : r=0,92



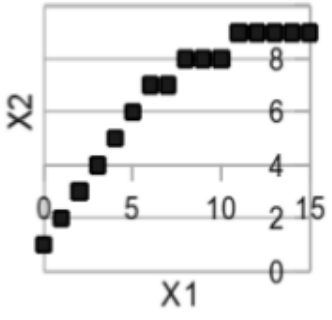
Exemple 8 : r=1



Exemple 9 : r=0



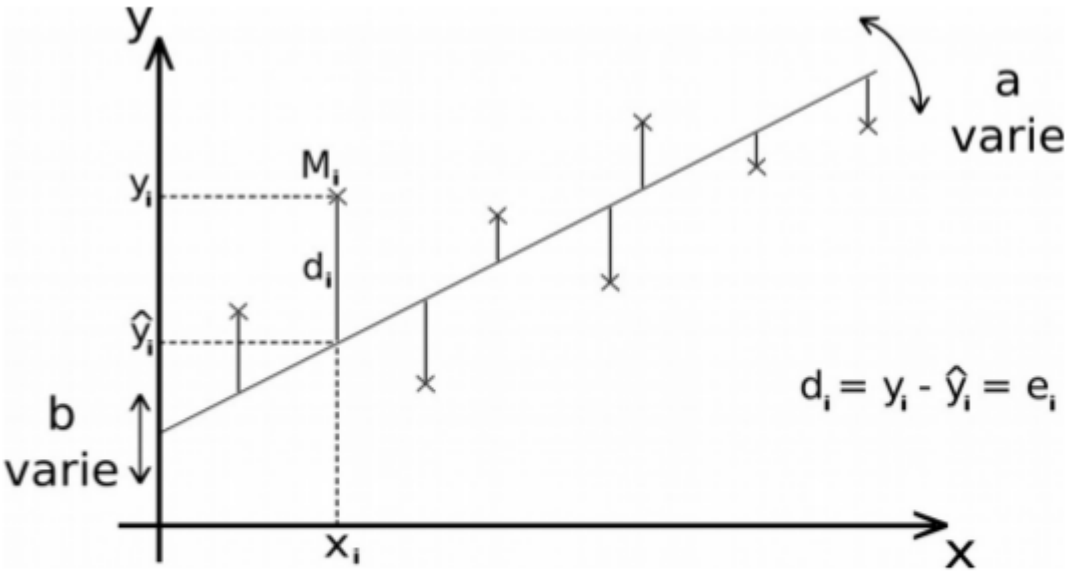
Exemple 10 : r=0,94



Régression linéaire

Si maintenant nous avons deux variables corrélées nous pourrions vouloir déterminer la relation la plus adaptée entre elles. Les variables aléatoires seront nommées X et Y et nous chercherons la droite qui passe au mieux par le nuage de points $y(x)$. Par exemple, quelle est la relation la mieux adaptée entre la taille X et le poids Y dans notre exemple initial : $y = ax + b$? Quelles sont les incertitudes Δa et Δb ?

1) Principe et formules



La méthode choisie est celle des moindres carrés : la droite considérée la meilleure est celle qui minimise la somme des carrés des distances à la droite, distances prises selon y (écarts).

L'ensemble des points se note $M_i(x_i, y_i)$. Pour x_i donné, l'ordonnée estimée sur la droite s'écrit $\hat{y}_i = a\,x_i + b$.

D'où la somme des distances au carré :

$$\sum d^2 = \sum_i (y_i - \hat{y}_i)^2$$

Les dérivées partielles de cette quantité selon a et b s'annulent pour la meilleure droite et nous obtenons les équations suivantes : $\sum_i (y_i - a x_i - b) x_i = 0$ et

$$\sum_i (y_i - a x_i - b) = 0 \quad .$$

Nous obtenons ainsi les résultats désirés :

$$\boxed{a = \frac{\overline{x\,y} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}} \quad \text{et} \quad \boxed{b = \bar{y} - a\,\bar{x}}$$

On nomme e_i le résidu tel que $y_i = \hat{y}_i + e_i$.

On trouve les différents écart-types suivants [11]:

- pour les résidus $s_r = \sqrt{\frac{\sum_i e_i^2}{n-2}}$
- pour la pente $s_a = \frac{s_r}{\sqrt{\sum_i (x_i - \bar{x})^2}}$
- pour l'ordonnée à l'origine $s_b = s_r \sqrt{\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}}$

Puis $\boxed{\Delta a = t_{n-2}\, s_a}$ et $\boxed{\Delta b = t_{n-2}\, s_b}$.

t_{n-2} : coefficients de Student pour $n-2$ degrés de liberté.

Application à l'exemple de la taille et du poids

	X_1 (cm)	X_2 (kg)	X_3
1	160	64	4
2	170	66	8
3	180	84	9
4	190	86	3
$\bar{X}$	175	75	6

$$\overline{xy} = (160\text{x}64 + 170\text{x}66 + 180\text{x}84 + 190\text{x}86)/4$$

$$\overline{x^2} = (160^2 + 170^2 + 180^2 + 190^2)/4$$

$$a = (13230 - 175\text{x}75)/(30750 - 175^2) = 0,84 \quad \text{et}$$

$$b = 75 - 0,84\text{x}175 = -72$$

$$s_r = \sqrt{[(64 - (0,84\text{x}160 - 72))^2 + (-4,8)^2 + 4,8^2 + (-1,6)^2]/2} \simeq 5,06$$

$$s_a \simeq 5,06/\sqrt{(160-175)^2 + (-5)^2 + 5^2 + 15^2} \simeq 0,226$$

$$s_b \simeq 5,06\sqrt{(160^2 + 170^2 + 180^2 + 190^2)/[4(15^2 + 5^2 + 25 + 225)]} \simeq 39,7$$

Exercice : donner la formulation du poids en fonction de la taille avec une confiance de 90 %

$$\Delta a \simeq 2,92\text{x}0,226 \simeq 0,66 \quad \text{avec une confiance de 90\%}$$

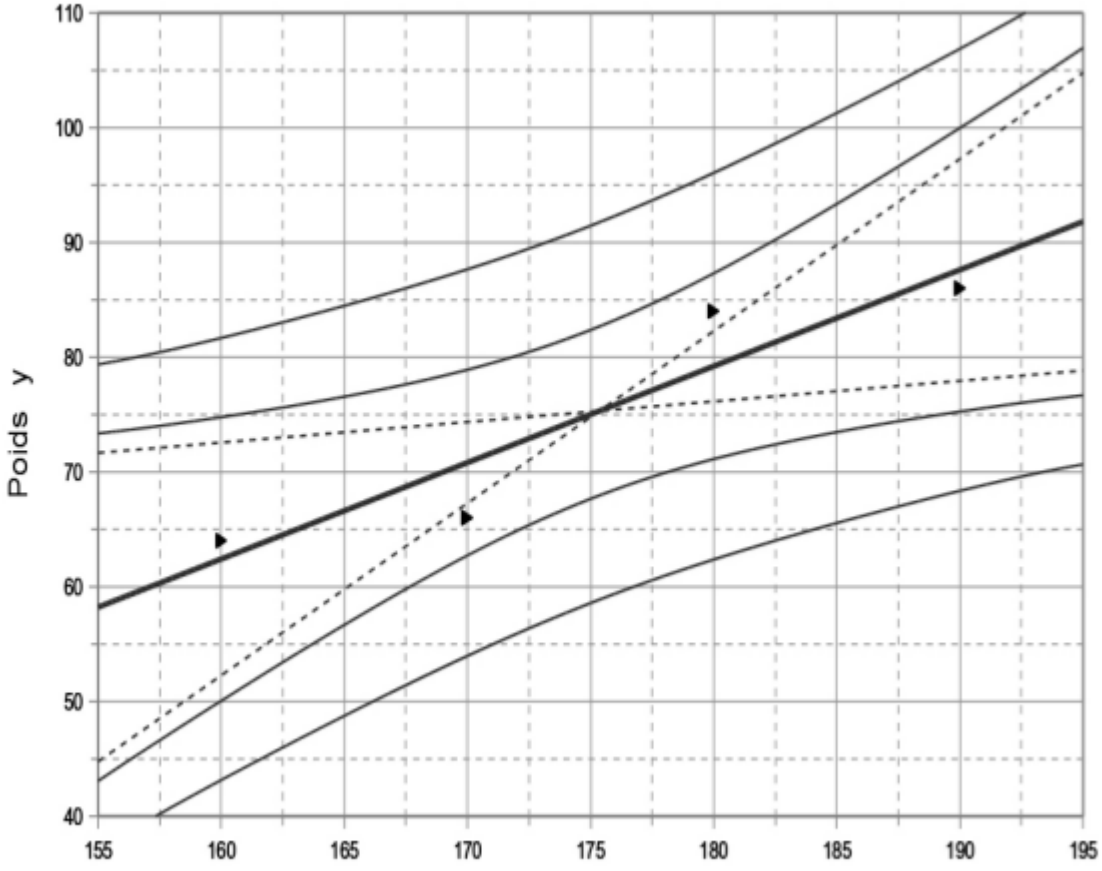
$$\Delta b \simeq 2,92\text{x}39,7 \simeq 116 \quad \text{à 90\%}$$

$$Poids = (0,84 \pm 0,66)Taille - (72 \pm 116) \quad \text{à 90\%.}$$

Formule ici très imprécise, ce qui n'est pas étonnant vu le peu de points et la dispersion des données.

Sur le graphique qui suit nous avons :

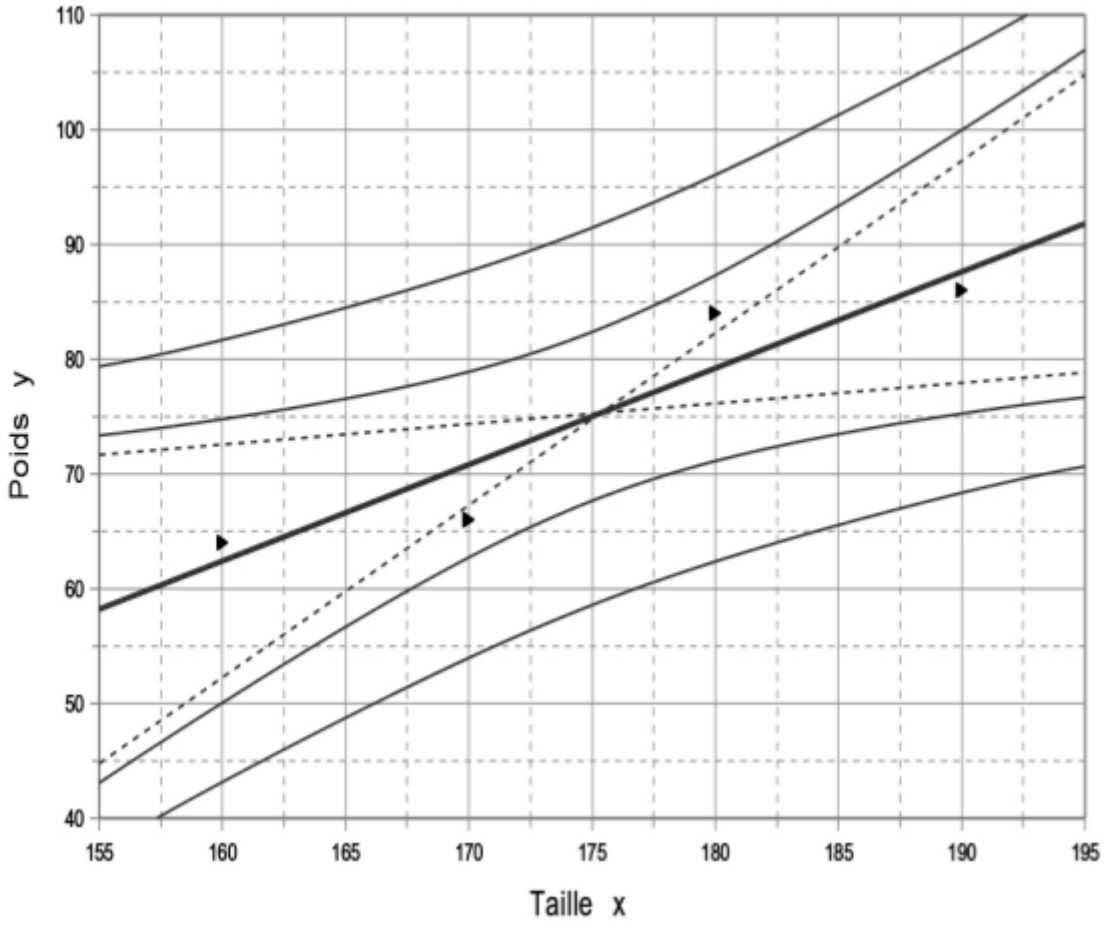
- au milieu la droite interpolée (le meilleur équilibre entre les points du dessus et ceux du dessous de cette droite).



- En pointillés sont représentés les deux droites extrêmes ($y=a_{min}x+b_{max}$ et $y=a_{max}x+b_{min}$).
- La première enveloppe correspond aux valeurs estimées de *y*. Intervalle de confiance de la moyenne de y_o pour une valeur x_o :

$$\Delta y_o = t_{n-2} s_r \sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Par exemple pour $x_o=175\ cm$ nous avons $y_o=75{,}0\pm7{,}4\ kg$. On peut même avoir une estimation en dehors de l'intervalle par exemple pour $x_o=195\ cm$ nous avons $y_o=92\pm15\ kg$.



- La deuxième enveloppe correspond à une prédiction si nous effectuions une nouvelle mesure. Intervalle de prédiction pour une observation y_o :

$$\Delta y_o = t_{n-2}\, s_r \sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (x_i - \bar{x})^2} + 1}$$

Par exemple, il y a 90% de chance pour une personne de 175 cm que sa masse soit entre 58 et 92 kg (en moyenne 90% des points sont dans cette seconde enveloppe et 10% en dehors).

Exercice

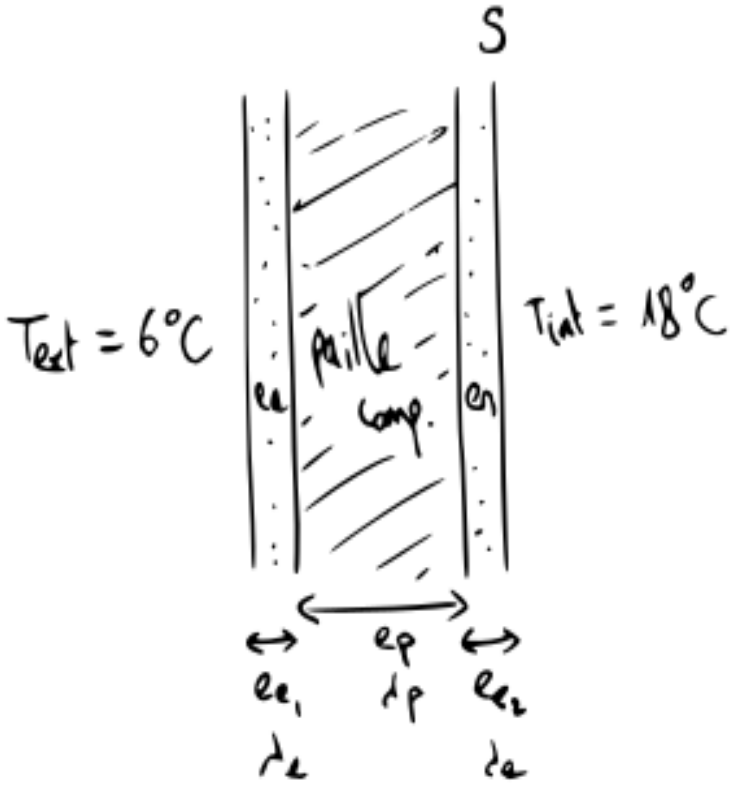
Nous avons un mur d'une surface S=72 m^2. La température extérieure est de 6°C et la température intérieure est maintenue à 18°C. Ce mur de 50 cm d'épaisseur est constitué de e_p=40 cm de paille compressée (conductivité thermique λ_p=45 mW/K/m) et de e_e=10 cm d'enduit (λ_e=200 mW/K/m). Les λ sont données à 10% près, les épaisseurs au cm et les température au demi degré.

1- Déterminez la résistance thermique, avec son incertitude,de la paille pour ce mur (R.λ.S=e)

2- Même chose pour l'enduit.

3- Sachant que les résistances thermiques s'associent comme les résistances électriques en série, déterminez la résistance thermique totale du mur avec son incertitude.

4- Quelle doit être la puissance minimale du chauffage de la maison rien que pour compenser les pertes par les murs ? ($\Delta T=R.\Phi$)



1-

$$R_p = \frac{e_p}{\lambda_p S} = \frac{0{,}4\,m\,K\,m}{0{,}045\,W\,.\,72\,m^2}$$

d'où $R_p = 0{,}123\,K/W$

$$\left(\frac{\Delta R_p}{R_p}\right)^2 = \left(\frac{\Delta e_p}{e_p}\right)^2 + \left(\frac{\Delta \lambda_p}{\lambda_p}\right)^2$$

et $\Delta R_p / R_p = 0{,}103$

$$\underline{R_p = 0{,}123 \pm 0{,}013\,K/W}$$

2- $\underline{R_e = 0{,}0069 \pm 0{,}0010\,K/W}$

3- Les résistances sont en série : $R = R_p + R_e$

et $\Delta R = \sqrt{(\Delta R_p)^2 + (\Delta R_e)^2}$ d'où $\underline{R = 0{,}130 \pm 0{,}015\,K/W}$

4- $\Phi = \frac{\Delta T}{R} = \frac{12}{0{,}130} = 92{,}3\,W$

$$(\Delta\Phi/\Phi)^2 = (\Delta(\Delta T)/\Delta T)^2 + (\Delta R/R)^2 \qquad \Delta T = T_{intérieur} - T_{extérieur}$$

d'où $\Delta(\Delta T) = \sqrt{2}.0{,}5\,°C$, $\Delta\Phi/\Phi$ = 12,9% et

$80\,W \leqslant \Phi \leqslant 104{,}3\,W$ d'où une puissance minimale de chauffage :

P_{min} = 105 W

Exercice 2 : Volumes

À l'aide d'une pipette jaugée nous remplissons quatre béchers avec 100 mL d'eau chacun. Pour tester la pipette et connaître précisément la quantité d'eau, nous effectuons quatre pesées au décigramme et nous obtenons, ramené en mL, les résultats suivants pour les différents béchers :

$$V_1=\{100,1 ; 100,0 ; 99,9 ; 100,0\}$$

1- Calculez la moyenne et l'écart-type de V_1. Estimez la précision de la pipette avec une confiance de 95%.

Nous remplissons maintenant deux béchers et rassemblons le contenu des deux dans un seul :

$$V = V_1 + V_2.$$

Correspondant à V_1, nous avons les mesures suivantes pour V_2 :

$$V_2=\{100,0 ; 100,1 ; 100,0 ; 99,9\}$$

Par exemple pour la troisième mesure nous avons $V_1=99,9$ mL et $V_2=100,0$ mL.

2- Montrez que V_1 et V_2 sont des grandeurs indépendantes.

3- Calculez la moyenne de V, son écart-type et l'incertitude ΔV à 95%.

4- Pourriez-vous retrouver ce résultat avec la formule de propagation des incertitudes?

(Pour affiner le test il faudrait prendre plus de mesures, mais le principe reste le même, et les résultats restent valides car nous avons élargi avec le Student, considéré des données décorrélées et des paquets globalement gaussiens. Nous devrions aussi tenir compte des incertitudes sur les mesures -résolution- en plus de leur dispersion.)

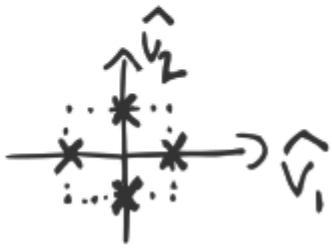
$\overline{V}_1$=(100,1+100,0+99,9+100,0)/4 d'où $\overline{\mathbf{V}}_1$ **= 100,0 mL**.

$$\sigma_1=\sqrt{\frac{0,1^2+0^2+(-0,1)^2+0^2}{4-1}}=\sqrt{\frac{2}{3}}\cdot 0,1\,mL \quad \text{d'où} \quad \sigma_1\simeq 0,082\,mL$$

D'après le théorème central limite : $\Delta V=t.\sigma/\sqrt{n}=3,18x0,0816/2$
$t_{ddl=3;\ 95\%}=3,18$ soit $\mathbf{\Delta\overline{V}_1\simeq 0,13\ mL}$ et $\Delta\overline{V}_1/\overline{V}1\simeq 0,13/100$
La pipette, avec une confiance de 95%, est à 0,13 mL près, soit pour 100 mL à 0,13% de précision.

$$\hat{V}^i=V^i-\overline{V} \quad \text{et}$$
$$\sum_i[(V_1{}^i-\overline{V_1})(V_2{}^i-\overline{V_2})]=\sum_i[\hat{V_1}{}^i\hat{V_2}{}^i]=0,1.0+0.0,1+(-0,1).0+0.0,1=0$$

or $$r_{12}=\frac{\sum_i[(V_1{}^i-\overline{V_1})(V_2{}^i-\overline{V_2})]}{\sqrt{\sum_i(V_1{}^i-\overline{V_1})^2\sum_i(V_2{}^i-\overline{V_2})^2}}$$



soit $\boldsymbol{r_{12}=0}$, les grandeurs sont totalement décorrélées et donc indépendantes.

V={200,1 ; 200,1 ; 199,9 ; 199,9}mL d'où $\overline{\mathbf{V}}$ **= 200 mL**.

$$\sigma_V=\sqrt{\frac{0,1^2+0,1^2+(-0,1)^2+(-0,1)^2}{4-1}}=\frac{2}{\sqrt{3}}\cdot 0,1\,mL$$

d'où $\sigma_V\simeq 0,115\,mL$ et $\mathbf{\Delta\overline{V}\simeq 0,183\ mL}$ et $\Delta\overline{V}/\overline{V}\simeq 0,09\%$

$V(V_1, V_2)$ d'où :

$$\Delta V^2=\left(\frac{\partial V}{\partial V_1}\right)^2_{V_2}\Delta V_1{}^2+\left(\frac{\partial V}{\partial V_2}\right)^2_{V_1}\Delta V_2{}^2=\Delta V_1{}^2+\Delta V_2{}^2$$

et $\Delta V=\sqrt{2}.\Delta V_1\simeq 0,18$ même résultat qu'au 3-.

Exercice 9 : Rendement

Des quantités d'engrais déterminées sont répandues sur des champs et nous obtenons les rendements suivants :

Engrais (kg/ha)	100	200	300	400	500	600	700
Rendement (Q/ha)	41	44	53	63	66	65	78

1- Déterminez la droite de régression qui passe par ce nuage de points. Pente, ordonnée à l'origine et incertitudes avec une confiance de 95%.

2- Pour 550 kg d'engrais par ha estimez le rendement.

3- Même chose en absence d'engrais.

4- Si un agriculteur répand 250 kg d'engrais par hectare, quelle est la probabilité qu'il obtienne 40 à 60 quintaux de céréales ?

$\bar{x} = \frac{100 + 200 + 300 + 400 + 500 + 600 + 700}{7} = 400$

$\bar{x^2} = \frac{(100)^2 + (200)^2 + (300)^2 + (400)^2 + (500)^2 + (600)^2 + (700)^2}{7} = 200\,000$

$\bar{xy} = \frac{100 \times 41 + 200 \times 44 + \ldots - 700 \times 78}{7} = 25500$

$\bar{y} = \frac{41 + 44 + 53 + 57 + 61 + 75 + 78}{7} = 58,57$

coefficients de la droite $a = \frac{\bar{xy} - \bar{x}\,\bar{y}}{\bar{x^2} - (\bar{x})^2}$

$b = \bar{y} - a\bar{x} = \quad = 0,0593$

$58,57 - 0,0593 \times 400 = 34,85$

Résidus $e_i = y_i - \hat{y}_i = y_i - (ax_i + b)$

$|e_1| = 0,12$
$|e_2| = 2,71$
$|e_3| = 0,31$
$|e_4| = 4,43$
$|e_5| = 1,5$
$|e_6| = 5,43$
$|e_7| = 1,64$

$\frac{\Sigma e_i^2}{n-2} = \frac{61,57}{5} = 12,31$

$s_r = \sqrt{\frac{\Sigma e_i^2}{n-2}} = \sqrt{12,31} = 3,509$

$s_a = \frac{s_r}{\sqrt{\Sigma (x_i - \bar{x})^2}} = 0,00853$

$s_b = s_r \sqrt{\frac{\Sigma x_i^2}{n \Sigma (x_i - \bar{x})^2}} = 2,965$

$\Delta a_{95\%} = 2,57\, s_a = 0,0173$

$\Delta b_{95\%} = 2,57\, s_b = 7,6205$

→ **Rendement = (0,0593 ± 0,017) Engrais + 34,85 ± 7,6**

2,57 : coef de Student pour $n-2 = 5$ degrés de liberté et 95 % de chance

Estimation du rendement pour 550 kg d'engrais

$0,0593 \times 550 + 34,85 = 67,465$

Incertitude $_{95\%} = t_{n-2/95\%} \cdot s_r \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\Sigma (x_i - \bar{x})^2}} = 4,26$

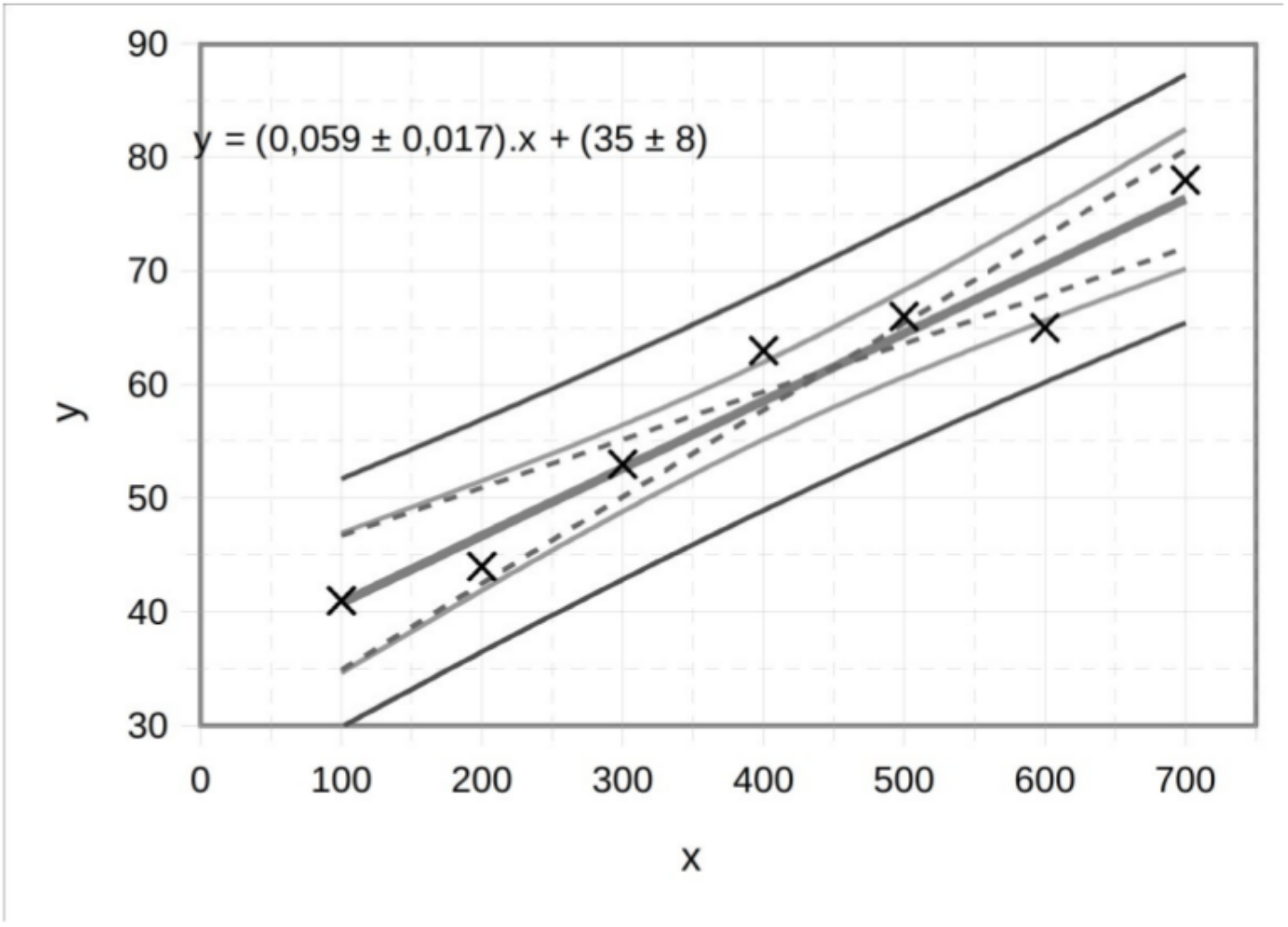
donc pour 550 kg d'engrais rendement = 67,5 ± 4,3 q/ha

Pour 0 engrais : $Rdt = 0,0593 \times 0 + 34,85$.

$\Delta_{95\%} = 2,57 \times 3,509 \times \sqrt{\frac{1}{7} + \frac{(400)^2}{250000}} = 7,02$

$Rdt_0 = 34,85 \pm 7,02$

Rendement en fonction de la quantité d'engrais :



4) Nouvelle mesure : il s'agit maintenant d'une prédiction

La formule pour l'incertitude devient

$$\underset{95\%}{\Delta y_0} = t_{n-2/95}\, S_r \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} + 1}$$

t Student pour $n-2 = 5$

→ Trouver cette fois ci la valeur du coefficient de Student telle que $\underset{x\%}{\Delta y_0}_{pred} = 10$

$$t_{n-2} = \frac{10}{S_r \sqrt{\frac{1}{n} + \frac{(250.400)^2}{280000}}}$$

$t_{n-2} = 2.57$ → Table de Student pour 5 degrés de liberté

→ intervalle de confiance à 95 %

A. Coefficients de Student

Coefficient de Student t		Confiance (%)								
		50	**80**	**90**	**95**	**98**	**99**	**99,5**	**99,8**	**99,9**
Degrés de liberté (taille de l'échantillon moins le nombre de paramètres)	**1**	1,00	3,08	6,31	12,7	31,8	63,7	127	318	637
	2	0,82	1,89	2,92	4,30	6,96	9,92	14,1	22,3	31,6
	3	0,76	1,64	2,35	3,18	4,54	5,84	7,45	10,2	12,9
	4	0,74	1,53	2,13	2,78	3,75	4,60	5,60	7,17	8,61
	5	0,73	1,48	2,02	2,57	3,36	4,03	4,77	5,89	6,87
	6	0,72	1,44	1,94	2,45	3,14	3,71	4,32	5,21	5,96
	7	0,71	1,41	1,89	2,36	3,00	3,50	4,03	4,79	5,41
	8	0,71	1,40	1,86	2,31	2,90	3,36	3,83	4,50	5,04
	9	0,70	1,38	1,83	2,26	2,82	3,25	3,69	4,30	4,78
	10	0,70	1,37	1,81	2,23	2,76	3,17	3,58	4,14	4,59
	11	0,70	1,36	1,80	2,20	2,72	3,11	3,50	4,02	4,44
	12	0,70	1,36	1,78	2,18	2,68	3,05	3,43	3,93	4,32
	13	0,69	1,35	1,77	2,16	2,65	3,01	3,37	3,85	4,22
	14	0,69	1,35	1,76	2,14	2,62	2,98	3,33	3,79	4,14
	15	0,69	1,34	1,75	2,13	2,60	2,95	3,29	3,73	4,07
	16	0,69	1,34	1,75	2,12	2,58	2,92	3,25	3,69	4,01
	17	0,69	1,33	1,74	2,11	2,57	2,90	3,22	3,65	3,97
	18	0,69	1,33	1,73	2,10	2,55	2,88	3,20	3,61	3,92
	19	0,69	1,33	1,73	2,09	2,54	2,86	3,17	3,58	3,88
	20	0,69	1,33	1,72	2,09	2,53	2,85	3,15	3,55	3,85
	22	0,69	1,32	1,72	2,07	2,51	2,82	3,12	3,50	3,79
	24	0,68	1,32	1,71	2,06	2,49	2,80	3,09	3,47	3,75
	26	0,68	1,31	1,71	2,06	2,48	2,78	3,07	3,43	3,71
	28	0,68	1,31	1,70	2,05	2,47	2,76	3,05	3,41	3,67
	30	0,68	1,31	1,70	2,04	2,46	2,75	3,03	3,39	3,65
	40	0,68	1,30	1,68	2,02	2,42	2,70	2,97	3,31	3,55
	50	0,68	1,30	1,68	2,01	2,40	2,68	2,94	3,26	3,50
	60	0,68	1,30	1,67	2,00	2,39	2,66	2,91	3,23	3,46
	70	0,68	1,29	1,67	1,99	2,38	2,65	2,90	3,21	3,44
	80	0,68	1,29	1,66	1,99	2,37	2,64	2,89	3,20	3,42
	90	0,68	1,29	1,66	1,99	2,37	2,63	2,88	3,18	3,40
	100	0,68	1,29	1,66	1,98	2,36	2,63	2,87	3,17	3,39
	200	0,68	1,29	1,65	1,97	2,35	2,60	2,84	3,13	3,34
	300	0,68	1,28	1,65	1,97	2,34	2,59	2,83	3,12	3,32
	500	0,67	1,28	1,65	1,96	2,33	2,59	2,82	3,11	3,31
	1000	0,67	1,28	1,65	1,96	2,33	2,58	2,81	3,10	3,30
	∞	**0,67**	**1,28**	**1,64**	**1,96**	**2,33**	**2,58**	**2,81**	**3,09**	**3,29**