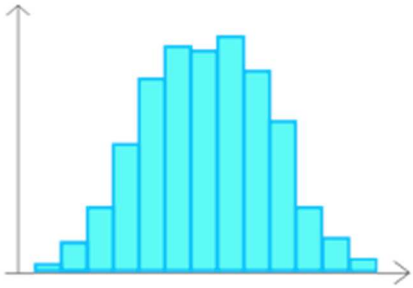
Rappel de quelques éléments simples de statistique

Imaginons que l'on s'intéresse à la taille des hommes de 40 ans en France. On aura alors une population de plusieurs centaines de milliers d'individus ayant cet âge. Si l'on trace un histogramme de la répartition de cette population par taille, on fera concrètement le schéma suivant : en ordonné, c'est-à-dire sur l'axe vertical, le nombre d'hommes mesurant une taille donnée par exemple 1,73 m (+/- 2 cm), ou, ce qui revient au même, le pourcentage des hommes de cette taille dans la population étudiée (ce qu'on peut appeler la fréquence de cette taille) et en abscisse, c'est-à-dire sur l'axe horizontal, les différentes tailles retenues, par tranches. On obtiendra un graphique du type :



Or, un résultat très intéressant de la statistique est résumé par la **loi des grands nombres** : quand on a affaire à une très grande population, la répartition de cette population suivant une variable donnée continue (comme ici la taille) donnera une répartition (en statistique, on parle de distribution) qui prendra la forme d'une courbe en cloche.
La figure 4.8 ci-dessous présente plusieurs courbes en cloche de ce type, également appelées Gaussienne ou **loi normale**.

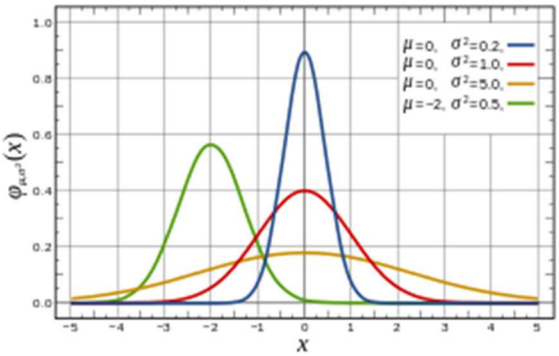


Fig.4.8 – Loi normale

Il est pour nous intéressant de se pencher sur cette loi normale pour deux raisons :
1- Un processus industriel qui se reproduit à l'identique, finit, dans la durée, par générer de grandes quantité, donc des mesures sur une grande population, et donc des quantités qui s'apparentent à des grands nombres… Les distributions que l'on peut construire à partir des mesures observées sur certains paramètres clés d'un processus peuvent ainsi être modélisées par la loi normale. C'est très utile du fait de ce qui suit.
2- Une loi normale peut être caractérisée assez simplement. C'est tout l'intérêt. D'une part, il y a la moyenne de la variable mesurée. (Par exemple 1,75m de taille moyenne pour les hommes adultes en France lors du recensement de 2007). Et puis il y a une mesure de l'aplatissement relatif de la courbe qui décrit la distribution. Sur la figure 4.8, la courbe jaune est très « aplatie », la courbe rouge un peu moins, la courbe verte encore moins et la courbe bleue est plutôt « pointue » : ceci signifie que la population représentée par la courbe jaune est très « dispersée », c'est-à-dire par exemple qu'il y aura peu de briques de lait avec la bonne quantité promise (1litre) et une part importante des briques avec soit trop, soit trop peu de lait. Cela revient à dire que le processus de remplissage des briques de lait, tel qu'il est organisé, ne permet pas d'obtenir ce qui est promis sur l'étiquette, à savoir un litre de lait, même si c'est implicitement « à peu de choses près ». A l'inverse, la population représentée par la courbe bleue, disons celle des sacs de ciment contrôlés en sortie de ligne, est pour l'essentiel proche de la cible de 35kg. Rares sont les « individus », c'est-à-dire les sacs contrôlés, qui sont « loin » de la cible. A ce titre, le processus est satisfaisant.
Pour caractériser l'aplatissement de la courbe vers le bas, ou au contraire, ce qui sera préférable, une forme pointue concentrée autour de la moyenne, on utilise en statistiques un concept appelé « l'écart type ». L'écart type mesure la dispersion des données. Il est traditionnellement symbolisé par la lettre grecque sigma : σ. La courbe jaune, dispersée, a un sigma élevé, la courbe bleue, pointue, a un sigma faible.
Nous ne rentrons pas tout de suite ici dans la définition mathématique formelle du σ. Nous ne nous préoccupons pas non plus du comment il se calcule, autrement que pour rappeler qu'il est calculé à partir des données mesurées. (Voir l'annexe ci-dessous). Mais arrive alors le point essentiel.
L'intérêt de la loi normale et du concept d'écart-type est que la forme connue de cette distribution permet de dire quelle est la part de la population qui est dans un intervalle de 2 ou 4 ou 6 ou n écarts-types autour de la moyenne m.
Ainsi, entre m – σ et m + σ, il y a 68% des individus de la population (c'est donc 68% des individus qui sont concentrés dans un intervalle large de 2σ autour de la moyenne) ; entre m – 2σ et m + 2σ, soit dans un intervalle large de 4σ, c'est 95%. Et entre m – 3σ et m + 3σ, soit dans un intervalle de 6σ, il y a 99,7% de la population, ce qui signifie, en creux, que 3 individus pour mille restent encore hors de cet intervalle. C'est dire que σ devient l'unité de compte de l'écart à la moyenne… D'où son nom, l'écart-type.

Pour mémoire, plutôt que de parler d'écart type, on peut aussi parler de variance (on sent bien l'idée de variation / dispersion véhiculée par le mot) : la variance se trouve être le carré de l'écart type : σ^2

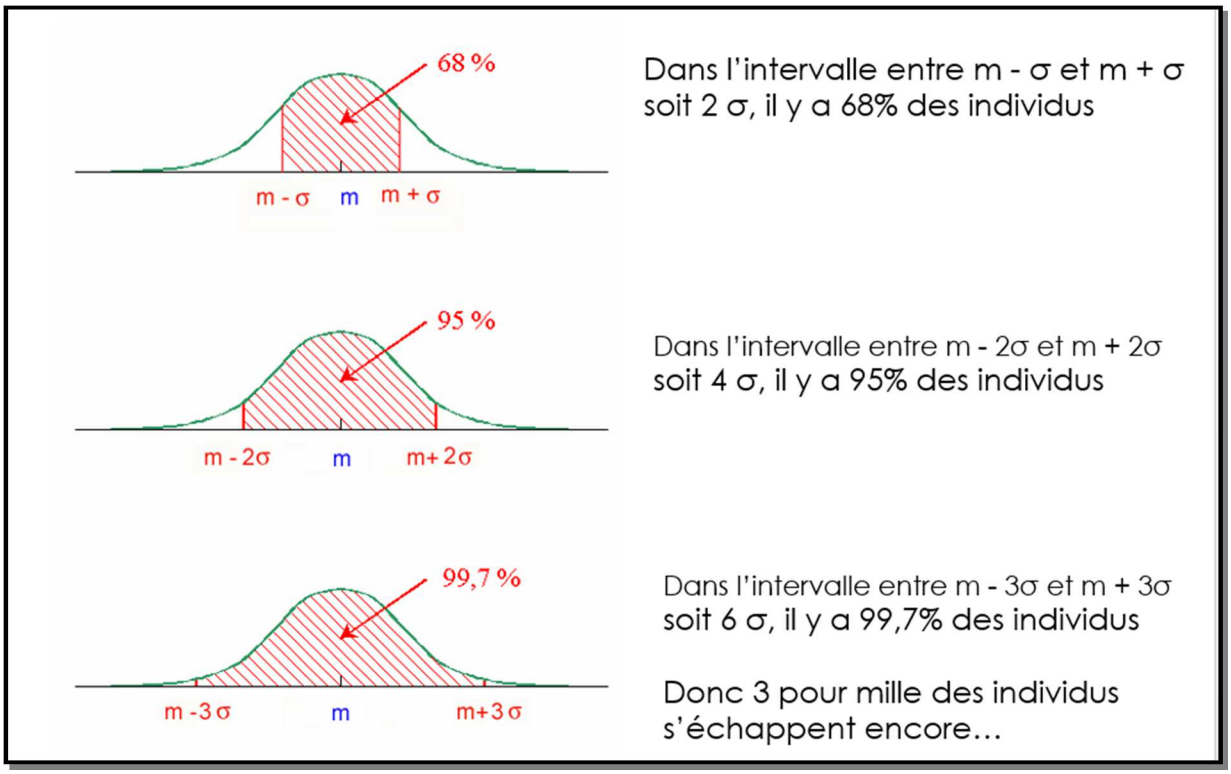


Fig.4.9 – Quelle fraction de la population dans quel intervalle ?

Et ainsi de suite. Mais d'où viennent ces chiffres ? De la table de la loi normale. Voir l'annexe ci-dessous.

Annexe :

Pour mémoire, mais ce n'est pas essentiel pour notre raisonnement, une courbe appelée la densité de la loi normale et représentée ci-dessous est donnée par :

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

où μ est la moyenne (notée m précédemment) et σ est l'écart type. Cette courbe dite de densité est, au sens mathématique, l'intégrale de la loi normale.

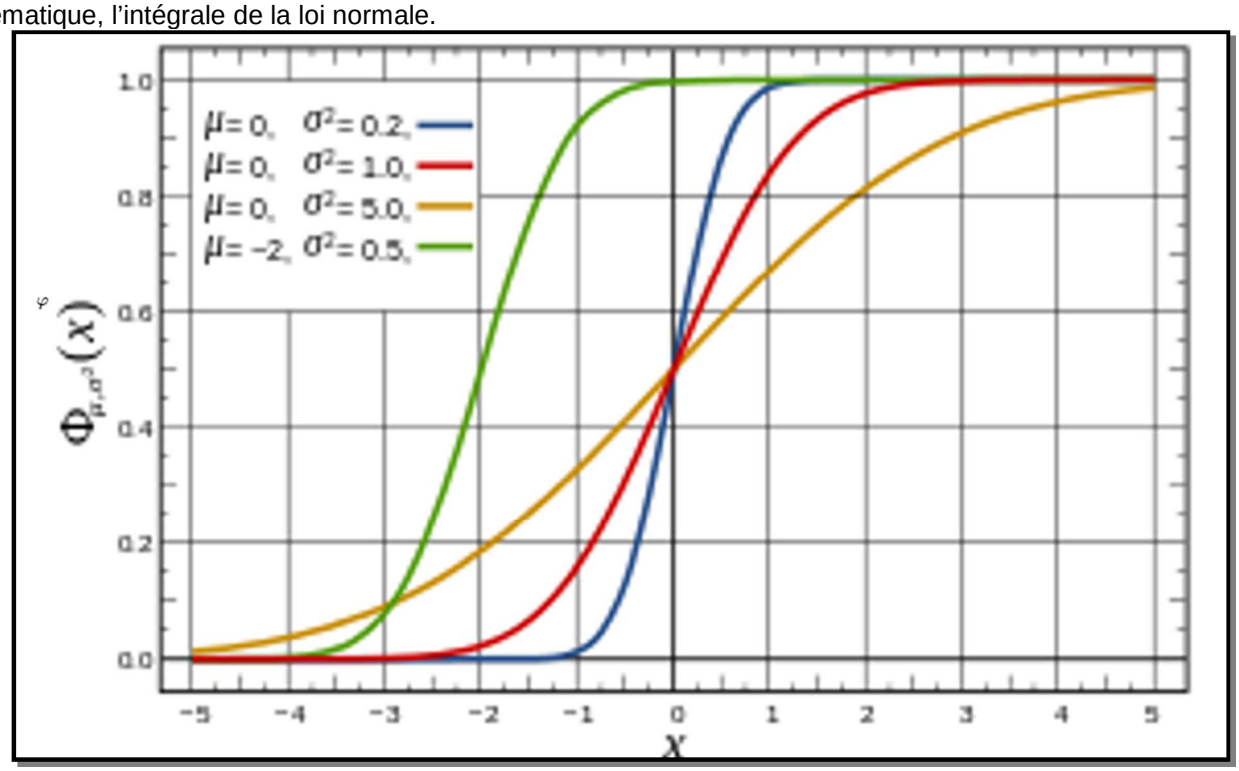
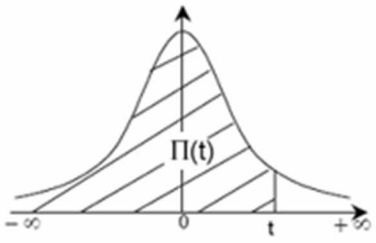


Fig.4.10 – Contenu de la table et Fonction de répartition

On parle de loi normale centrée réduite quand on a ramené la moyenne à 0 (en retirant la valeur moyenne à toutes les mesures, soit x-μ) et qu'on a réduit (normalisé) l'échelle des distances en la divisant par σ.
La table des valeurs de la loi normale est obtenue en calculant l'intégrale de la densité de répartition (le pourcentage de tous les individus de la zone grisée sur la courbe par rapport à la population totale).

Probabilité de trouver une valeur inférieure à t



$$\Pi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

t	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998

Table pour les grandes valeurs de t :

t	3	3,2	3,4	3,6	3,8	4	4,2	4,4	4,6	4,8
Π(t)	0,99865003	0,99931280	0,99966302	0,99984085	0,99992763	0,99996831	0,99998665	0,99999458	0,99999789	0,99999921

Fig.4.11 – Loi normale centrée réduite