

Chapitre 5:

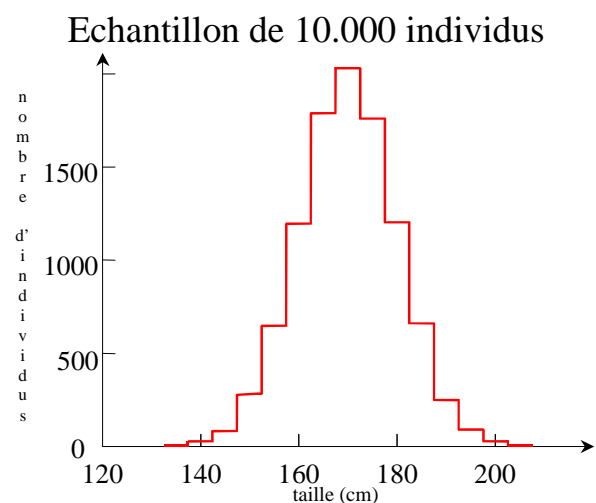
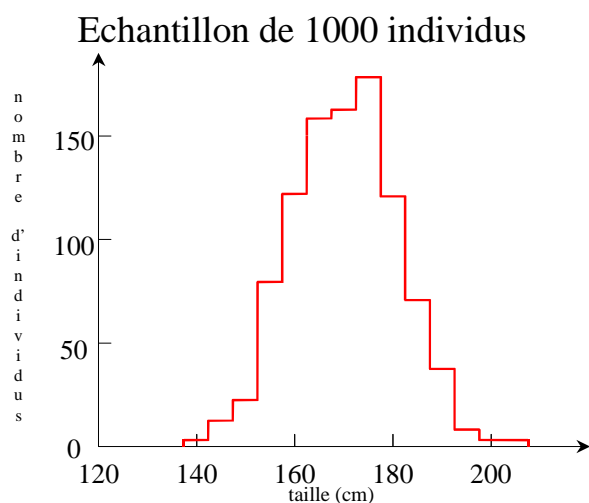
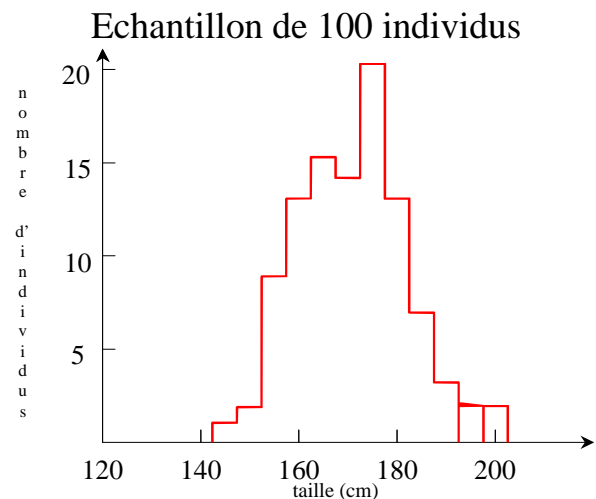
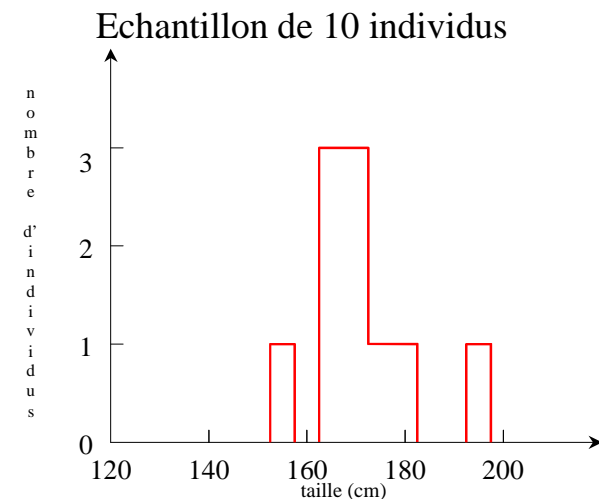
La loi normale

5.1. Loi normale ou de Gauss

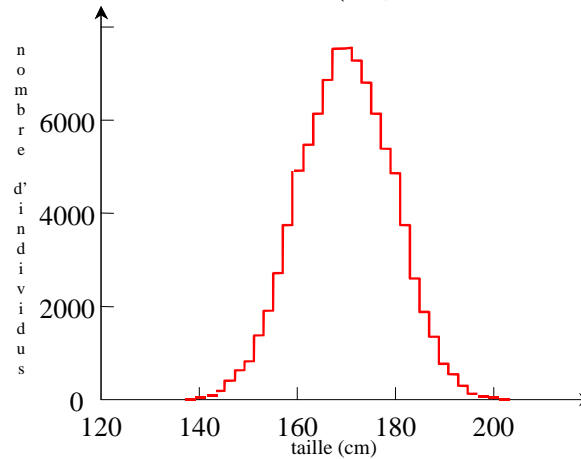
Supposons que nous tirions des échantillons aléatoires d'une population dont la taille moyenne est de 170 cm, avec un écart type de 10 cm.

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

Examinons l'aspect de ces histogrammes.

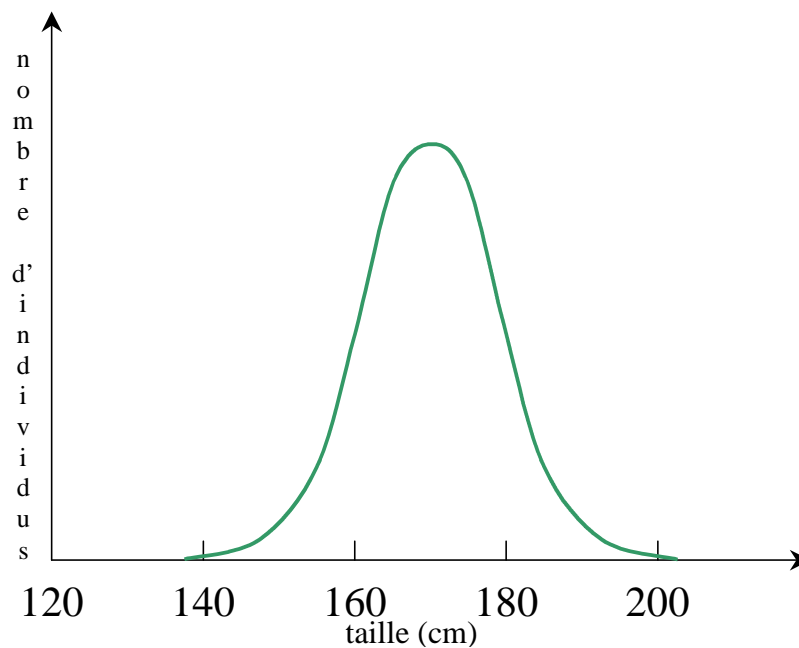


Echantillon de 100.000 individus. (ici, les classes sont de 2 cm)



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



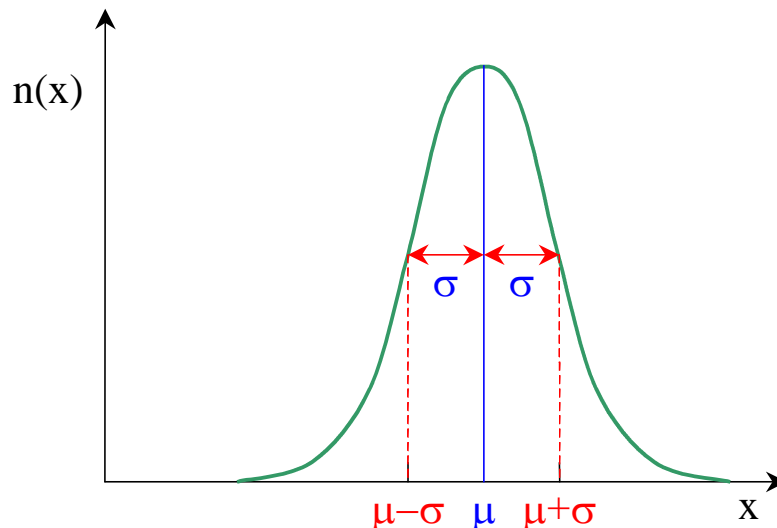
La loi normale est la loi statistique la plus répandue et la plus utile.

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:

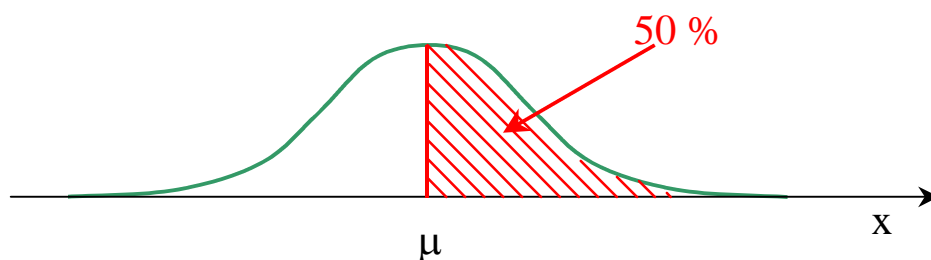
$$n(x) = \frac{n}{\sqrt{2\pi} \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



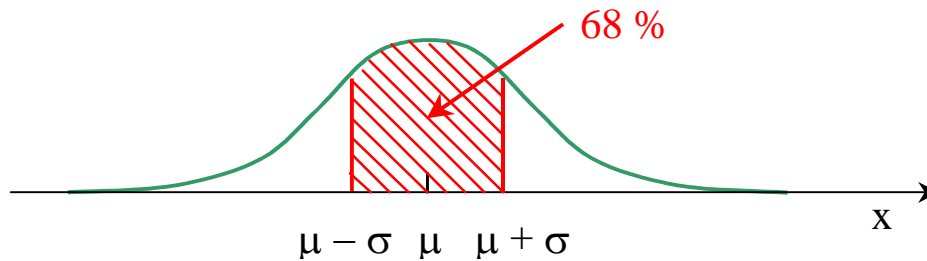
- μ est la moyenne
- σ l'écart type
- n le nombre total d'individus dans l'échantillon
- $n(x)$ le nombre d'individus pour lesquels la grandeur analysée a la valeur x .

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

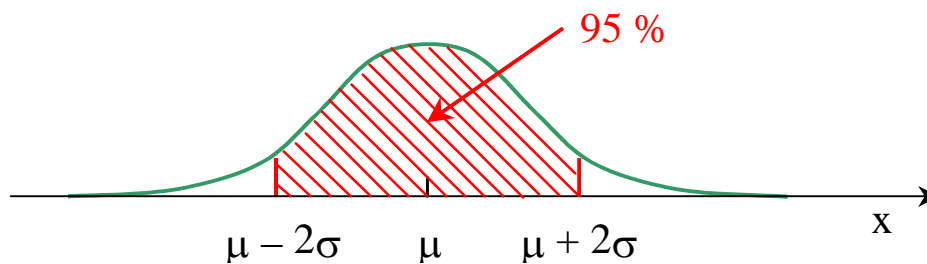
- A.** 50 % des individus en-dessous de la moyenne μ et 50 % au-dessus (la loi normale est symétrique)



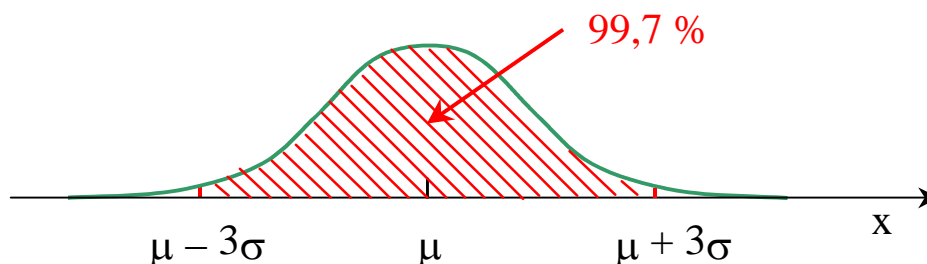
B. 68 % des individus entre $\mu - \sigma$ et $\mu + \sigma$



C. 95 % des individus entre $\mu - 1,96\sigma$ et $\mu + 1,96\sigma$, que nous arrondirons à l'intervalle $[\mu - 2\sigma, \mu + 2\sigma]$



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



5.2. 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.

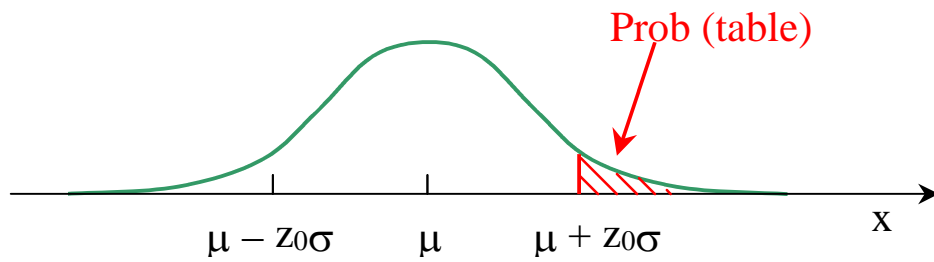
Pour obtenir z_0 , on calcule l'écart par rapport à la moyenne : $\delta = x - \mu$, puis on divise par l'écart type : $z_0 = \frac{\delta}{\sigma}$

z_0	2 ^{ème} décimale de z_0									
	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

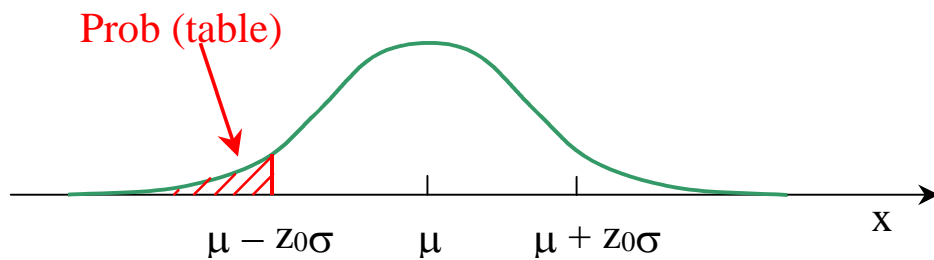
Lorsque l'on doit déterminer une probabilité à partir de la loi normale, on essaie de se ramener à une probabilité considérée dans la table.

Quelques cas concrets sont illustrés ci-dessous.

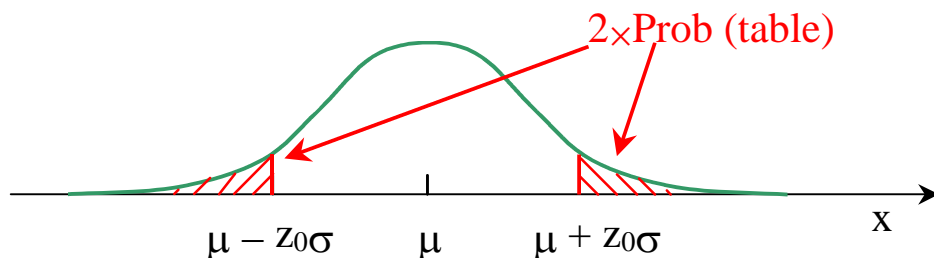
1) $x > \mu + z_0\sigma$



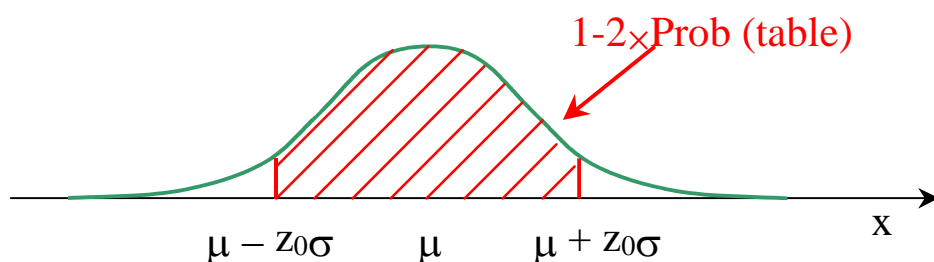
2) $x < \mu - z_0\sigma$



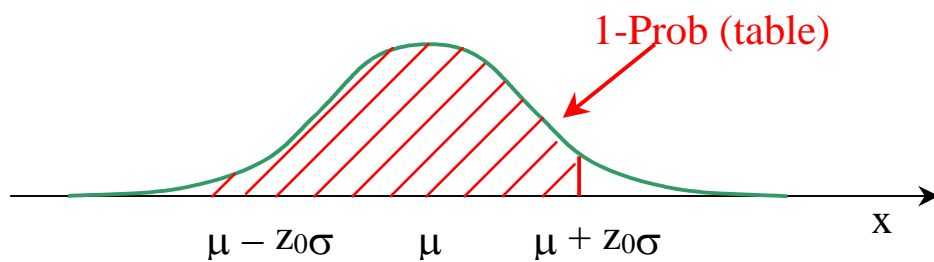
3) x plus éloigné de μ que $z_0\sigma$



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 \rightarrow 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 $\mu = 200 \text{ gr}$ 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 ($< 230 \text{ gr}$) considéré contient la valeur moyenne (200 gr) \rightarrow on prend $1 - \text{Prob}(\text{table})$:

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

- 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 :

$$\delta = 20 \text{ gr} \quad \sigma = 40$$

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

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

On doit multiplier par 2 car on considère les deux côtés $\rightarrow \text{Prob} = 2 \times 0,309 = 0,618$

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: $0,382 = 38,2 \%$

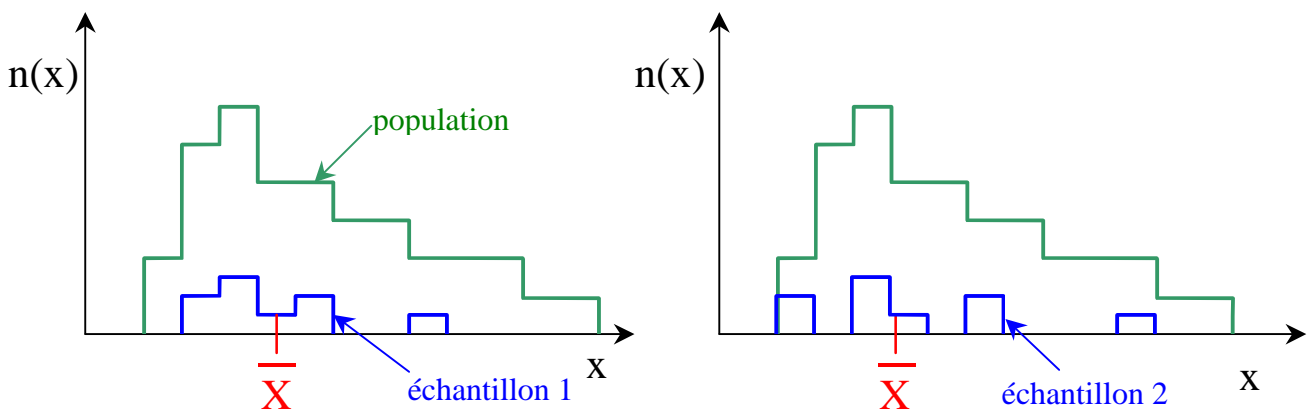
5.3. Forme de la distribution d'échantillonnage

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

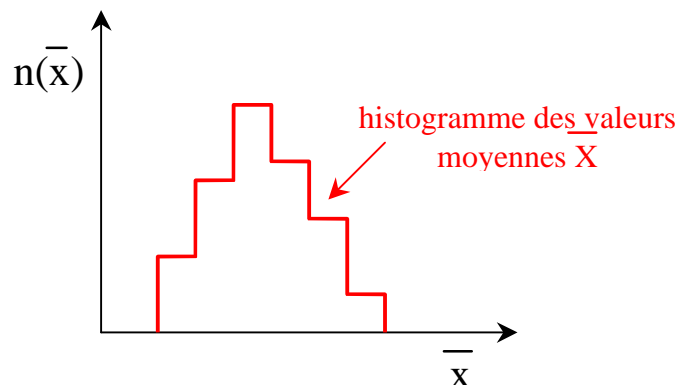
Pour chacun de ces échantillons, nous calculons une valeur moyenne \bar{X} qui est une estimation de la moyenne de la population μ .

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

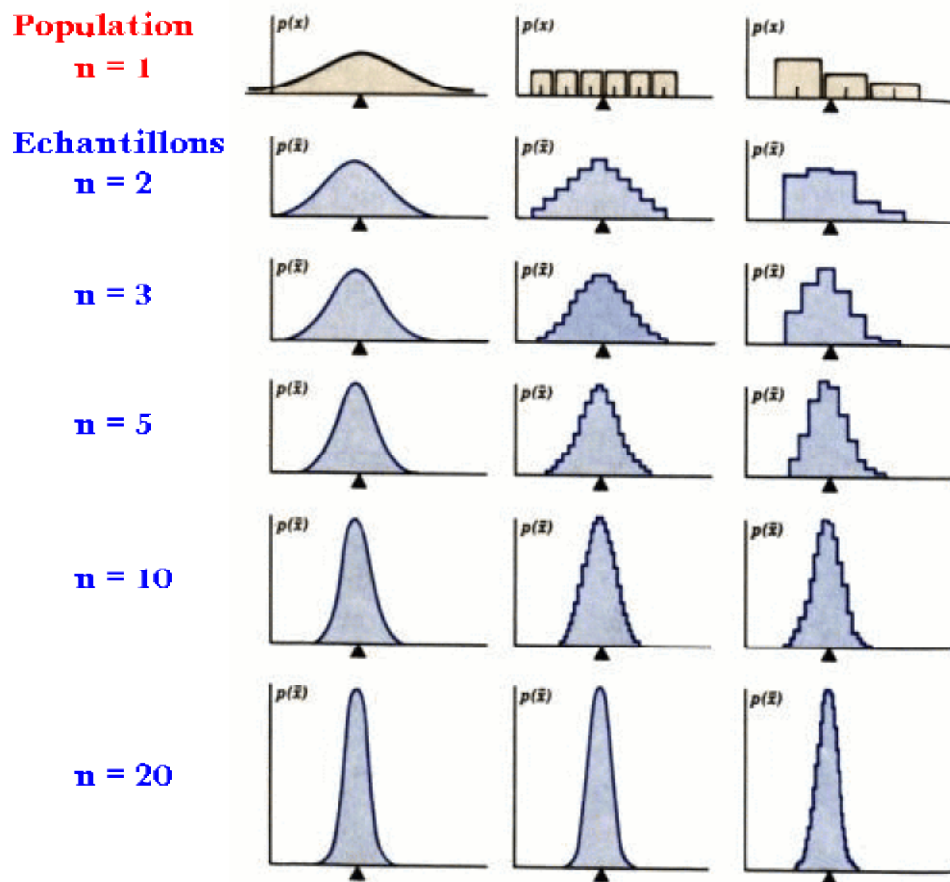
Nous désirons savoir comment les différentes déterminations \bar{X} vont se distribuer autour de la vraie moyenne μ .



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



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



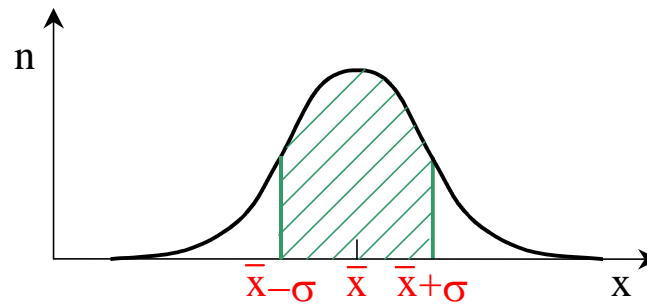
Lorsque la taille de l'échantillon est suffisamment grande, ($n \geq 10$) la distribution de la moyenne a une forme approximativement normale.

L'écart type sur la moyenne est:

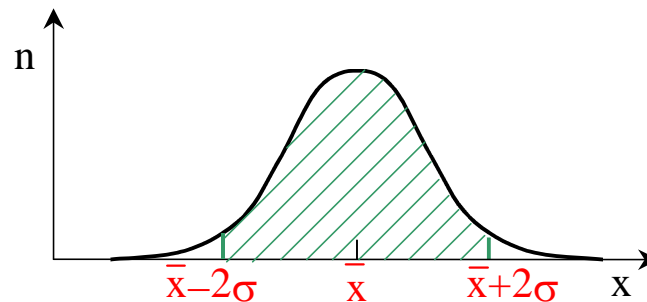
$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

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.

Il y a 68 % ($\cong 2/3$) de chances que la vraie moyenne μ soit dans l'intervalle compris entre $\bar{X}-\sigma$ et $\bar{X}+\sigma$.



Il y a 95 % de chances que la vraie moyenne μ soit dans l'intervalle compris entre $\bar{X}-2\sigma$ et $\bar{X}+2\sigma$.



Exemples

1. Une organisation de consommateurs désire savoir si le poids réel des pains de 800gr produits dans une boulangerie est bien conforme au poids annoncé.

Pour cela, elle prélève 100 pains au hasard. Elle obtient un poids moyen de 780 gr, avec un écart type de 80 gr.

Quelle est la probabilité que le boulanger triche en moyenne sur le poids de ses pains?

$$n = 100 \quad \bar{X} = 780 \text{ gr} \quad s = 80 \text{ gr}$$

Ecart type de \bar{X} :
$$\sigma(\bar{X}) = \frac{s}{\sqrt{n}} = \frac{80}{\sqrt{100}} = 8 \text{ gr}$$

Il n'y a pas tricherie si μ est de 800 gr au moins, c'est-à-dire 20 gr au-dessus de \bar{X} .

$$z_0 = \frac{20}{\sigma(\bar{X})} = \frac{20}{8} = 2,5$$

D'après la table, la probabilité que μ soit supérieur à \bar{X} d'au moins $2,5 \sigma(\bar{X})$ est de 0,006.

$$100 - 100 \times 0,006 = 100 - 0,6 = 99,4$$

Il y a 99,4 % de chance pour que le boulanger triche

→ on peut raisonnablement conclure à la tricherie !

2. Dans une autre boulangerie, l'échantillon de 100 pains conduit au résultat suivant :

$$n = 100 \quad \bar{X} = 795 \text{ gr} \quad s = 50 \text{ gr}$$

$$\sigma(\bar{X}) = \frac{s}{\sqrt{n}} = \frac{50}{\sqrt{100}} = 5 \text{ gr}$$

$$z_0 = \frac{800 - 795}{5} = \frac{5}{5} = 1$$

D'après la table, la probabilité que μ soit supérieur ou égal à 800 gr est de 0,159.

Il y a donc 84,1 % de chances que ce boulanger triche.

Même si la présomption est forte, ce n'est pas vraiment suffisant pour conclure à la tricherie !

3. Dans une troisième boulangerie, on obtient :

$$n = 100 \quad \bar{X} = 810 \text{ gr} \quad s = 50 \text{ gr}$$

$$\sigma(\bar{X}) = \frac{s}{\sqrt{n}} = \frac{50}{\sqrt{100}} = 5 \text{ gr}$$

$$Z_0 = \frac{800 - 810}{5} = -\frac{10}{5} = -2$$

D'après la table, la probabilité que μ soit inférieur ou égal à 800 gr est de 0,023.

Il n'y a que 2,3 % de chances que ce boulanger triche.

On peut conclure qu'il est honnête !

4. Le revenu moyen d'un échantillon représentatif de 16 ménages s'élève à 62 000 F net par mois, avec un écart type de 16 000 F.

- a. Quelle est la probabilité que le revenu moyen de l'ensemble de la population soit inférieur à 60 000 F ?

Solution: $n = 16 \quad \bar{X} = 62000 \quad s = 16000$

$$\sigma(\bar{X}) = \frac{s}{\sqrt{n}} = \frac{16000}{\sqrt{16}} = \frac{16000}{4} = 4000$$

$$\delta = 60000 - 62000 = -2000$$

$$z_0 = \frac{\delta}{\sigma(\bar{X})} = \frac{2000}{4000} = 0,5$$

< 60 000: s'écarte de \bar{X} de plus de $0,5 \times \sigma(\bar{X})$

→ on prend la valeur de la table

$$\text{Prob} = 0,309 = 30,6 \%$$

- b. Quelle est la probabilité que le revenu moyen de la population soit inférieur à 65 000 F ?

Solution: $\delta = 65000 - 62000 = 3000$

$$Z_0 = \frac{\delta}{\sigma(\bar{X})} = \frac{3000}{4000} = 0,75$$

< 65000: ne s'écarte pas de \bar{X} de plus de $0,75 \times \sigma(\bar{X})$ vers le haut

→ on prend la valeur p de la table et on calcule 1-p

$$p = 0,227$$

$$1-p = 0,773 = 77,3 \%$$

5.4. Intervalles de confiance

Nous avons vu que la moyenne \bar{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 $2\sigma(\bar{X})$.

Nous pourrions donc écrire, soit:

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

soit, plus explicitement:

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

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

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 $\bar{X} \pm 2\sigma(\bar{X})$.

Exemples :

1. La taille moyenne d'un échantillon de 51 filles de 2^{ème} 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 :

$$\begin{aligned} n &= 51 & \bar{X} &= 167,9 & s &= 5,3 \\ \sigma(\bar{X}) &= \frac{5,3}{\sqrt{51}} = 0,74 \\ 2\sigma(\bar{X}) &= 1,48 \cong 1,5 \text{ cm} \end{aligned}$$

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 2^{ème} 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.

$$\begin{aligned} n &= 35 & \bar{X} &= 182,9 & s &= 6,7 \\ \sigma(\bar{X}) &= \frac{6,7}{\sqrt{35}} = 1,13 \\ 2\sigma(\bar{X}) &= 2,26 \cong 2,3 \text{ cm} \end{aligned}$$

Avec 95 % de confiance, on a donc:

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

5.5. Comparaison de deux échantillons indépendants

Des échantillons sont indépendants lorsqu'une modification dans l'un d'eux n'a pas d'influence sur les autres.

Par exemple, un échantillon de filles et un échantillon de garçons sélectionnés pour déterminer le poids moyen des filles et celui des garçons sont indépendants.

Par contre, les personnes qui affirment voter pour un parti A ou pour un parti B dans un sondage politique ne forment pas deux échantillons indépendants car si une personne de plus déclare voter pour A, il y a un électeur potentiel en moins pour B (les résultats de A et B s'influencent).

La différence moyenne est simplement:

$$\bar{X}_D = \bar{X}_1 - \bar{X}_2$$

Son écart type est donné par:

$$\sigma_D = \sqrt{\sigma_1^2 + \sigma_2^2}$$

(formule approchée, mais suffisamment précise)

Pour la taille des filles et des garçons:

$$\bar{X}_D = 182,9 - 167,9 = 15 \text{ cm}$$

$$\sigma_D = \sqrt{0,74^2 + 1,13^2} = 1,35 \text{ cm}$$

L'intervalle de confiance à 95 % est de:

$$2 \times 1,35 = 2,7 \text{ cm}$$

$$\bar{X}_D = 15,0 \pm 2,7 \text{ cm}$$

Si nous désirons maintenant répondre à la question suivante: les garçons sont-ils plus grands, en moyenne, que les filles?

- différence moyenne de taille: $\bar{X}_D = 15,0 \text{ cm}$
- écart type de cette différence: $\sigma_D = 1,35 \text{ cm}$
- nombre d'écarts types au-dessus de 0 cm (0 cm \equiv pas de différence de taille)

$$z_0 = \frac{15,0}{1,35} = 11,1111$$

Dans une distribution normale, la probabilité d'avoir une valeur qui s'écarte de la moyenne de plus de 11 σ est inférieure à $2 \cdot 10^{-28}$ et donc complètement négligeable.

Si nos échantillons sont représentatifs, il n'y a donc aucune chance que la différence de taille soit due au hasard.

Sur base de nos échantillons, nous sommes donc pratiquement certains que les garçons sont, en moyenne, plus grands que les filles.

Exemple

On sélectionne un échantillon de 25 paysans syldaves. La superficie de leurs terres s'élève à 24 hectares en moyenne, avec un écart type de 5 hectares.

Pour un échantillon de 16 paysans bordures, la superficie moyenne des terres est de 26 hectares, avec un écart type de 8 hectares.

Quelle est la probabilité que les paysans syldaves aient, en moyenne, plus de terres que les bordures ?

Solution:

Ces échantillons sont indépendants.

- échantillon syldave: $n_1 = 25$ $\overline{X}_1 = 24$ $s_1 = 5$
écart type sur la moyenne: $\sigma(\overline{X}_1) = \frac{5}{\sqrt{25}} = \frac{5}{5} = 1$
- échantillon bordure: $n_2 = 16$ $\overline{X}_2 = 26$ $s_2 = 8$
écart type sur la moyenne: $\sigma(\overline{X}_2) = \frac{8}{\sqrt{16}} = \frac{8}{4} = 2$

différence moyenne: (syldave-bordure):

$$\overline{X_D} = \overline{X}_1 - \overline{X}_2 = 24 - 26 = -2$$

écart type sur la différence:

$$\sigma_D = \sqrt{\sigma(\overline{X}_1)^2 + \sigma(\overline{X}_2)^2} = \sqrt{1^2 + 2^2} = \sqrt{5} \cong 2,24$$

les syldaves ont plus de terre en moyenne que les bordures si la différence est $>0 \rightarrow$ s'écarte de la valeur moyenne calculée de plus de 2

$$z_0 = \frac{\delta}{\sigma_D} = \frac{2}{2,24} \cong 0,89$$

table \rightarrow Prob = 0,187 = 18,7 % de chances que la superficie moyenne soit supérieure en Syldavie.