Ecole Supérieur de Technologie de Fkih Ben Salah

Statistique

I&II. Echantillon et Estimation

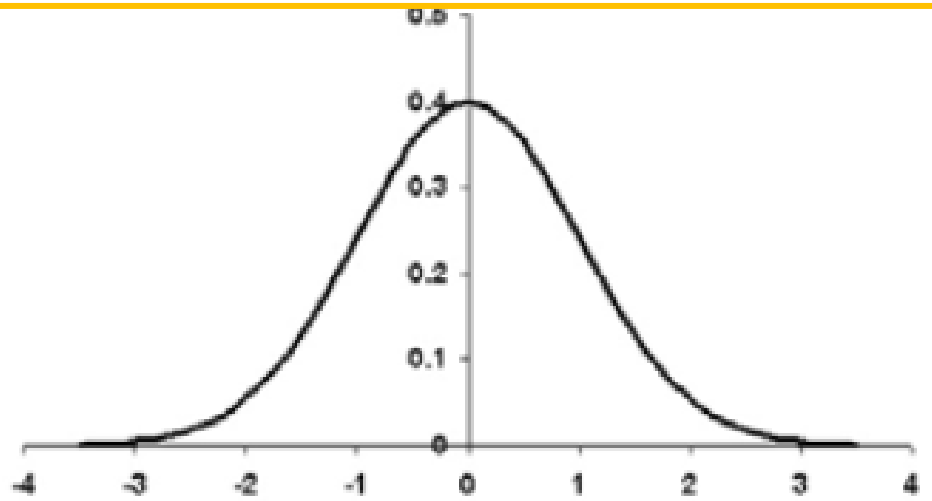
Génie Biologique & Industrie Agroalimentaire
2019-2020

M. MERZOUKI
m.merzouki@usms.ma

Loi Normale

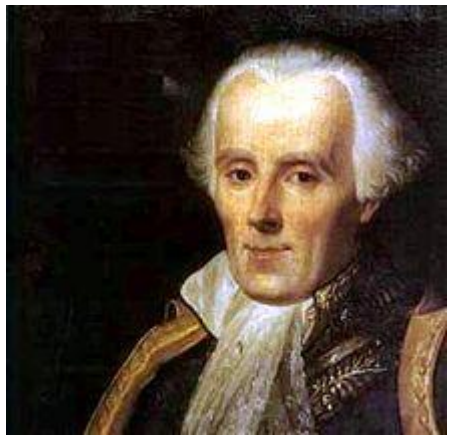
La **variable aléatoire** X est distribuée selon une loi normale si elle a une **fonction de densité** de la forme :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), \quad (\sigma > 0).$$



Loi normale, $\mu = 0$, $\sigma = 1$

Nous dirons que X suit une loi normale de moyenne μ et de variance σ^2. La loi normale est une **loi de probabilité continue**.



HISTORIQUE

La loi normale est souvent attribuée à **Pierre Simon de Laplace** et **Carl Friedrich Gauss** dont elle porte également le nom. Toutefois, son origine remonte aux travaux de **Jakob Bernoulli** qui dans son œuvre *Ars Conjectandi* (1713) a fourni les premiers éléments de base à la **loi des grands nombres**.

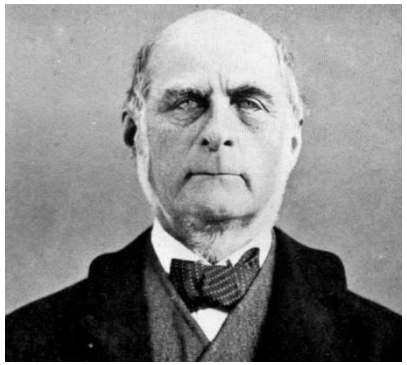
Abraham de Moivre fut le premier qui, en 1733, obtint la loi normale, comme approximation à la **loi binomiale**. Cet écrit était en latin ; il en publia une version anglaise en 1738. A. de Moivre parla de ce qu'il avait trouvé comme étant une « courbe » ;

il découvrit cette courbe alors qu'il devait calculer les probabilités de gain pour différents jeux de hasard.

P.S. de Laplace, à la suite d'A. de Moivre, étudia cette loi et obtint un résultat plus formel et général que l'approximation d'A. de Moivre. Il obtint en 1774 la distribution normale comme approximation de la **loi hypergéométrique**.

Remarquons que bien que la première approximation de cette loi de distribution soit due à A. de Moivre, Galilée avait déjà trouvé que les erreurs d'observations étaient distribuées de façon symétrique et tendaient à se grouper autour de leur vraie **valeur**.

La littérature propose de nombreuses dénominations de la loi normale. **Adolphe Quetelet** parlait de la « courbe des possibilités » ou « loi des possibilités ».



Notons également que **Francis Galton** parlait de « loi de **fréquence** des erreurs » ou de « loi de **déviation** d'après une **moyenne** ». S.M. Stigler (1980) présente une discussion plus détaillée sur les différents noms de cette courbe.

Echantillonnage

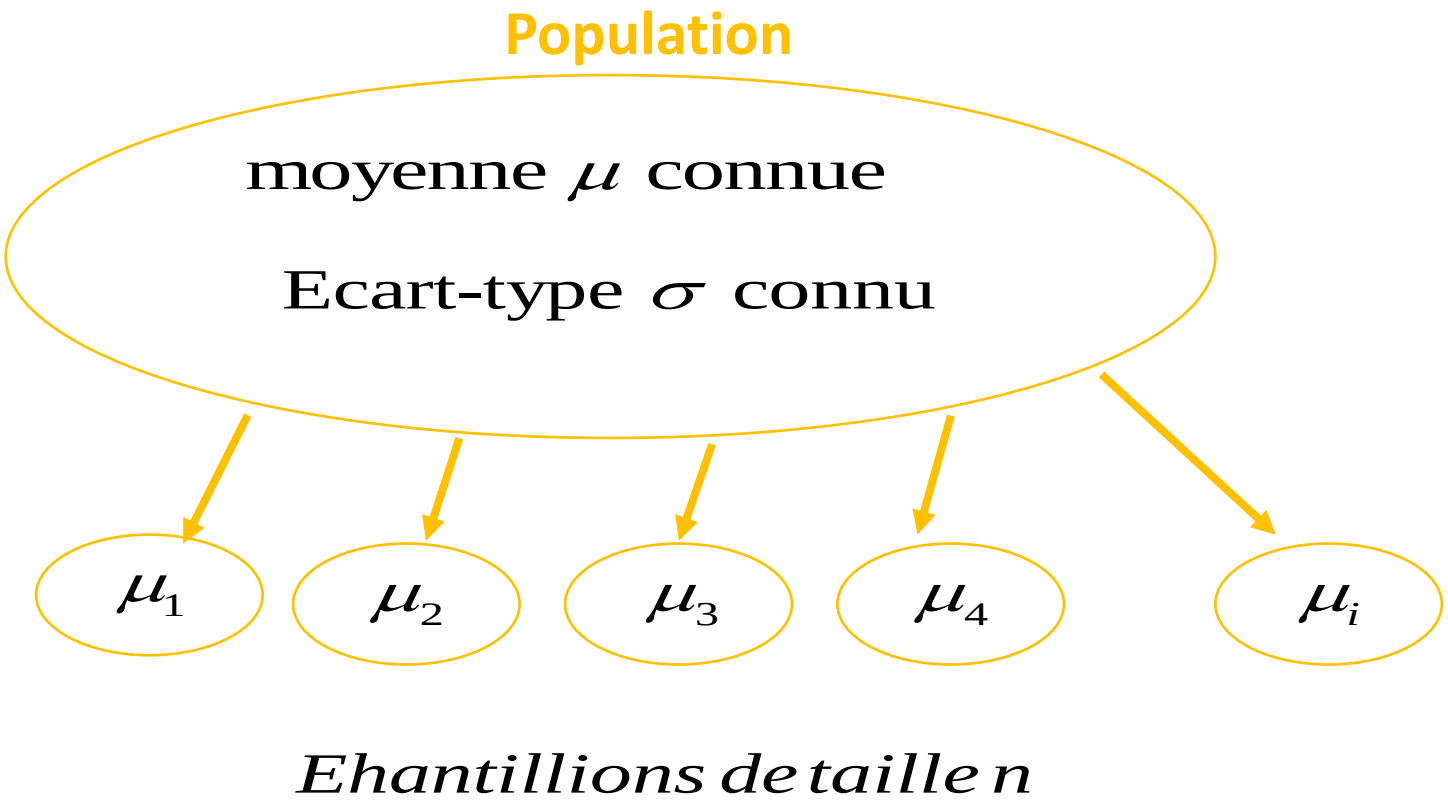
Étude de la moyenne et la proportion d'un échantillon

L'objectif de cette partie est de répondre à la problématique suivante : *comment, à partir d'informations (couple moyenne-écart-type ou proportion) connues sur une population, peut-on prévoir celles d'un échantillon ?*

Nous distinguerons deux cas : celui où l'on étudie une **moyenne** dans un échantillon et celui où l'on étudie une **proportion** dans un échantillon.

Étude de la moyenne d'un échantillon

On dispose d'une population sur laquelle est définie une variable aléatoire *X* dont on connaît l'espérance (ou la moyenne μ) m et l'écart-type σ.



On s'intéresse aux échantillons de taille *n*. Auront-ils tous la même moyenne ? Non, certains peuvent être constitués d'éléments atypiques et avoir une moyenne très différente de celle de la population (surtout si l'échantillon est de petite taille).

Notons $\overline{X}$ la variable aléatoire qui, à chaque échantillon de taille *n*, associe sa moyenne ($\overline{X}$ s'appelle encore la *distribution des moyennes des échantillons*).

Que peut-on dire de cette variable aléatoire *X* ?

Théorème Central Limite - Version 2 - (Version forte)

soit X une variable aléatoire qui suit une loi normale sur la population

avec E(X)= On prélève au hasard, un échantillon de taille *n* de moyenne $\bar{X}$.

Alors la variable aléatoire $\bar{X}$ suit également une loi normale :

$$\bar{X} \sim N(\mu; \frac{\sigma}{\sqrt{n}})$$

Théorème Central Limite - Version 2 - (Version forte)

soit X une variable aléatoire qui suit une loi normale sur la population

avec E(X)=μ et $\sigma(x)=\sigma$, On prélève au hasard, un échantillon de taille n, avec $n \geq 30$

de moyenne $\bar{X}$.

Alors la variable aléatoire $\bar{X}$ suit approximativement une loi normale :

$$\bar{X} \sim N(\mu; \frac{\sigma}{\sqrt{n}})$$

•X v.a continue, suit la loi normale de paramètres m et σ, notée N(m;σ), si sa densité de probabilité est:

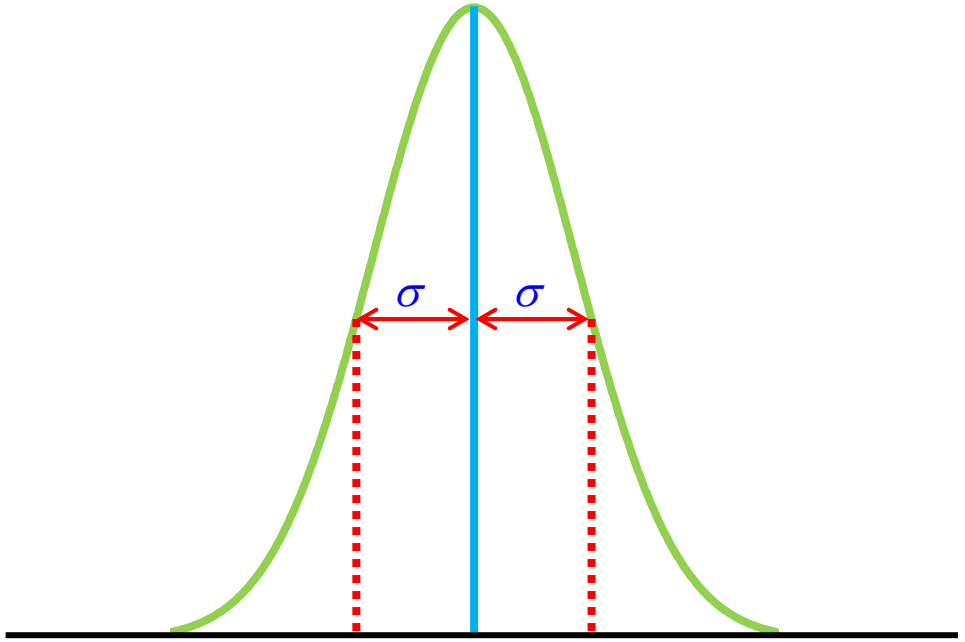
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

•Espérance $E(X)=\sigma$.

Ecart type $\sigma(X)=\sigma$

•Fonction de répartition $F(x) = \int_{-\infty}^{x} f(t)dt.$

•La représentation graphique de sa densité est une courbe en cloche (ou courbe de Gauss).



$\mu+\sigma$ μ $\mu+\sigma$

il ne faut pas confondre l'écart-type $\frac{\sigma}{\sqrt{n}}$ de la variable aléatoire $\bar{X}$ (qui est définie sur l'ensemble des échantillons possibles de taille n) avec l'écart-type d'un échantillon prélevé. L'écart-type de l'échantillon prélevé n'interviendra pas dans nos calculs dans cette partie. Pour éviter cette confusion, la quantité $\frac{\sigma}{\sqrt{n}}$ est appelée par fois "*erreur type*"

Exemple :

Les statistiques des notes obtenues en mathématiques au BAC Sc Math pour l'année 2006 sont :

Moyenne nationale: $\mu = 10,44$

Écart-type : $\sigma = 1,4$

Une classe filière de l'Economie comporte 200 étudiants dont 35 élèves en 2006/2007 issus d'un Bac Sc Math en 2006.

Calculer la probabilité que la moyenne de cette classe soit supérieure à 10.

Dans ce cas là on ne conanit pas la loi sur la population, mais l'effectif n de l'échantillon est supérieur à 30.

Nous allons donc pouvoir utiliser **le T.C.L. 2**. Notons X la variable aléatoire qui, à tout échantillon de taille $n = 35$, fait correspondre sa moyenne.

Dans ce cas là on a $\bar{X} \sim N(\mu; \frac{\sigma}{\sqrt{n}}) = N(10,44; \frac{1,46}{\sqrt{35}})$

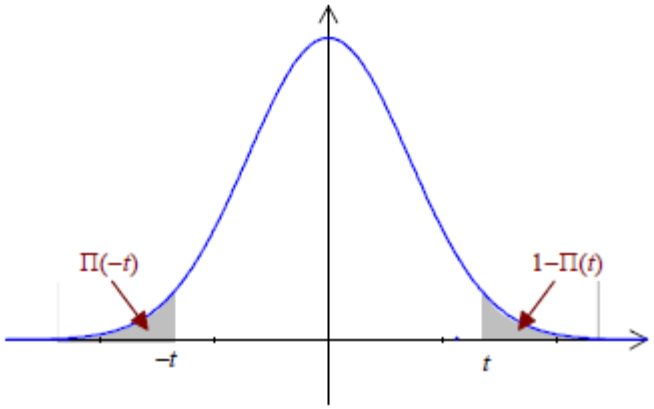
Poson $T = \frac{\bar{X} - 10,44}{\frac{1,46}{\sqrt{35}}}$, ainsi $T \sim N(0;1)$; (*loi normale centrée et réduite*)

$$P(\bar{X} \geq 10) = P\left(\frac{\bar{X} - 10,44}{\frac{1,46}{\sqrt{35}}} \geq \frac{10 - 10,44}{\frac{1,46}{\sqrt{35}}} \right)$$

$$= P(T \geq -1,78)$$

$$= P(T \leq 1,78)$$

$$= \prod(1,78)$$



Remarque : $P(T \geqslant t) = P(T \leqslant -t)$

En effet :

$P(T \geqslant t) = 1 - P(T \leqslant t) = 1 - \Pi(t) = \Pi(-t) = P(T \leqslant -t)$

TABLE III — AIRES LIMITÉES PAR LA COURBE NORMALE CENTRÉE RÉDUITE

La table fournit les valeurs de $\phi(z)$ pour z positif. Lorsque z est négatif il faut calculer le complément à l'unité de la valeur lue dans la table. La première colonne indique la première décimale de z et la première rangée fournit la deuxième décimale.

Exemples : pour $z = 1{,}21$, $\phi(z) = 0{,}8869$ et pour $z = -1{,}21$, $\phi(z) = 0{,}1131$

z	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

z	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
3	0,9987	0,9990	0,9993	0,9995	0,9997	0,9998	0,9998	0,9999	0,9999	1,0000
4	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

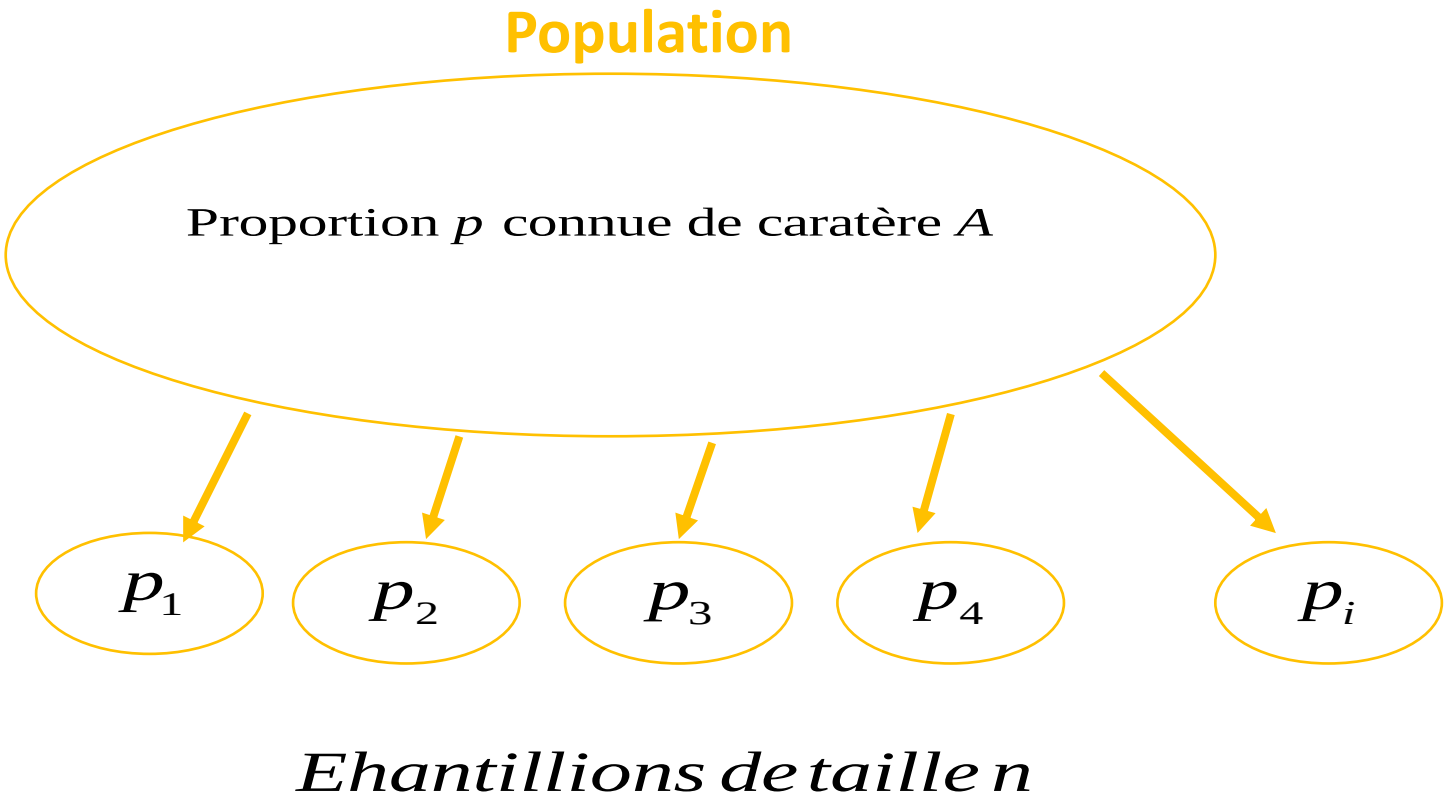
Et par lecture directe de la table de la loi normale centrée-réduite :

$$\Pi(1,78) = 0,9625$$

Conclusion : il y a environ 96% de chance que, dans cette classe, la moyenne des notes au baccalauréat de Mathématiques soit supérieure à 10.

Étude d'une proportion dans un échantillon

Cette fois-ci, on dispose d'une population sur laquelle on étudie un caractère (ou attribut) A dont on connaît la proportion p dans la population.



On s'intéresse aux échantillons de taille ***n***. **La proportion du caractère *A* dans les échantillons sera-t-elle toujours la même ?**

Evidemment non, cette proportion varie en fonction de l'échantillon choisi. Notons ***F*** la variable aléatoire qui, à chaque échantillon de taille n, associe **sa proportion du caractère *A*** (F s'appelle distribution des fréquence des échantillons).

Que peut-on dire de cette variable aléatoire *F* ?

Théorème

Theoreme

Soit une population sur laquelle on etudie un caractere A repandu avec une frequence p.

On preleve, au hasard, un echantillon de taille n avec $n \geq 30$.

On note F la frequence du caractere A dans l'echantillon.

Alors la variable aleatoire F suit approximativement une loi normale :

$$F \sim N\left(p; \sqrt{\frac{p(1-p)}{n}}\right)$$

Exemple :

Une élection **a eu lieu** et un candidat a eu 40 % des voix. On prélève un échantillon de 100 bulletins de vote.

Quelle est la probabilité que, dans l'échantillon, le candidat ait entre 35 % et 45 % des voix ?

On a $n = 100$ et $p = 0{,}4$. La variable aléatoire F correspondant à la fréquence des votes pour le candidat dans l'échantillon vérifie donc :

$$F \sim N\left(0,4;\sqrt{\frac{0,4\times 0,6}{100}}\right) = N\left(0,4;\sqrt{\frac{0,24}{10}}\right)$$

oit unePosons $T = \dfrac{F-0,4}{\dfrac{\sqrt{0,24}}{10}}$ ainsi $T \sim N(0;1)$. On obtenons alors par centrage et réduction:

$P(0,35 \le F \le 0,45) = P(-1,02 \le F \le 1,02) = 2\Pi(1,02) - 1.$

Par une lecture directe de la table de la loi normale centrée-réduite

on trouve que $\Pi(1,02) = 0.8461$. D'où $P(0,35 \le F \le 0,45) = 0,6922$

Il y a donc environ 69 % de chance que, dans un échantillon de taille *n* = 100, le candidat ait entre 35 % et 45 % des voix.

TABLE III — AIRES LIMITÉES PAR LA COURBE NORMALE CENTRÉE RÉDUITE

La table fournit les valeurs de $\phi(z)$ pour z positif. Lorsque z est négatif il faut calculer le complément à l'unité de la valeur lue dans la table. La première colonne indique la première décimale de z et la première rangée fournit la deuxième décimale.

Exemples : pour $z = 1{,}21$, $\phi(z) = 0{,}8869$ et pour $z = -1{,}21$, $\phi(z) = 0{,}1131$

z	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

z	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
3	0,9987	0,9990	0,9993	0,9995	0,9997	0,9998	0,9998	0,9999	0,9999	1,0000
4	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

On constate que l'on dispose des informations **sur la population** (ici, l'ensemble des votes) parce que **l'élection a déjà eu lieu**. On en **déduit des informations sur l'échantillon**. **Mais, dans la pratique, c'est souvent le phénomène réciproque que nous étudierons** : **les élections n'ont pas encore eu lieu et on voudrait retrouver les informations sur la population grâce un sondage réalisé sur un échantillon.**

D'où la deuxième partie de ce chapitre consacrée à **l'estimation**.

Estimation

L'objectif de cette partie est de savoir ***comment, à partir d'informations (couple moyenne/écart-type ou proportion) calculées sur un échantillon, retrouver ou plutôt estimer celles d'une population entière ?*** **L'estimation est le problème** inverse de l'échantillonnage.

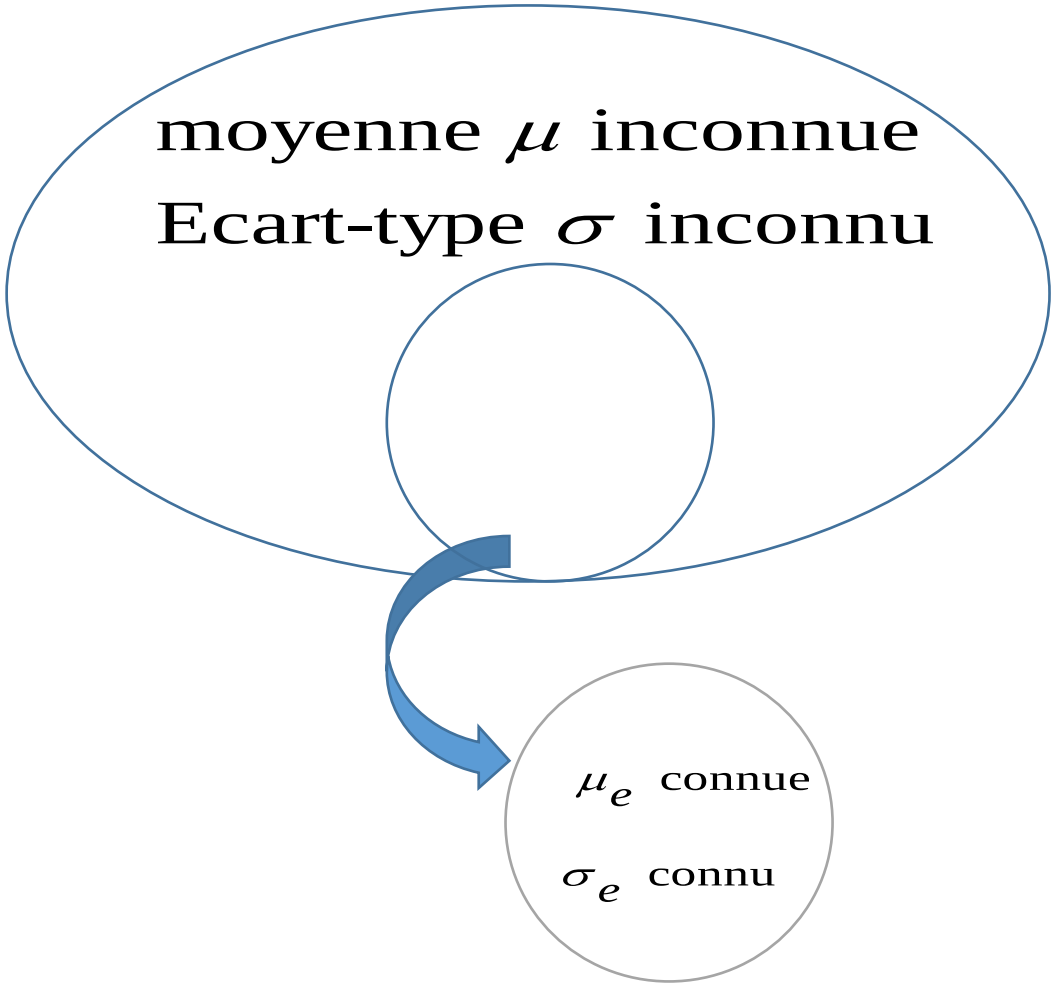
Pour passer à la phase estimative il faut avoir des résultats établis sur la théorie de l'échantillonnage

Il y a deux cas :

On cherche à estimer la moyenne *m* d'une variable aléatoire définie sur une population.

On cherche à estimer la proportion d'individus *p* ayant tel caractère dans la population.

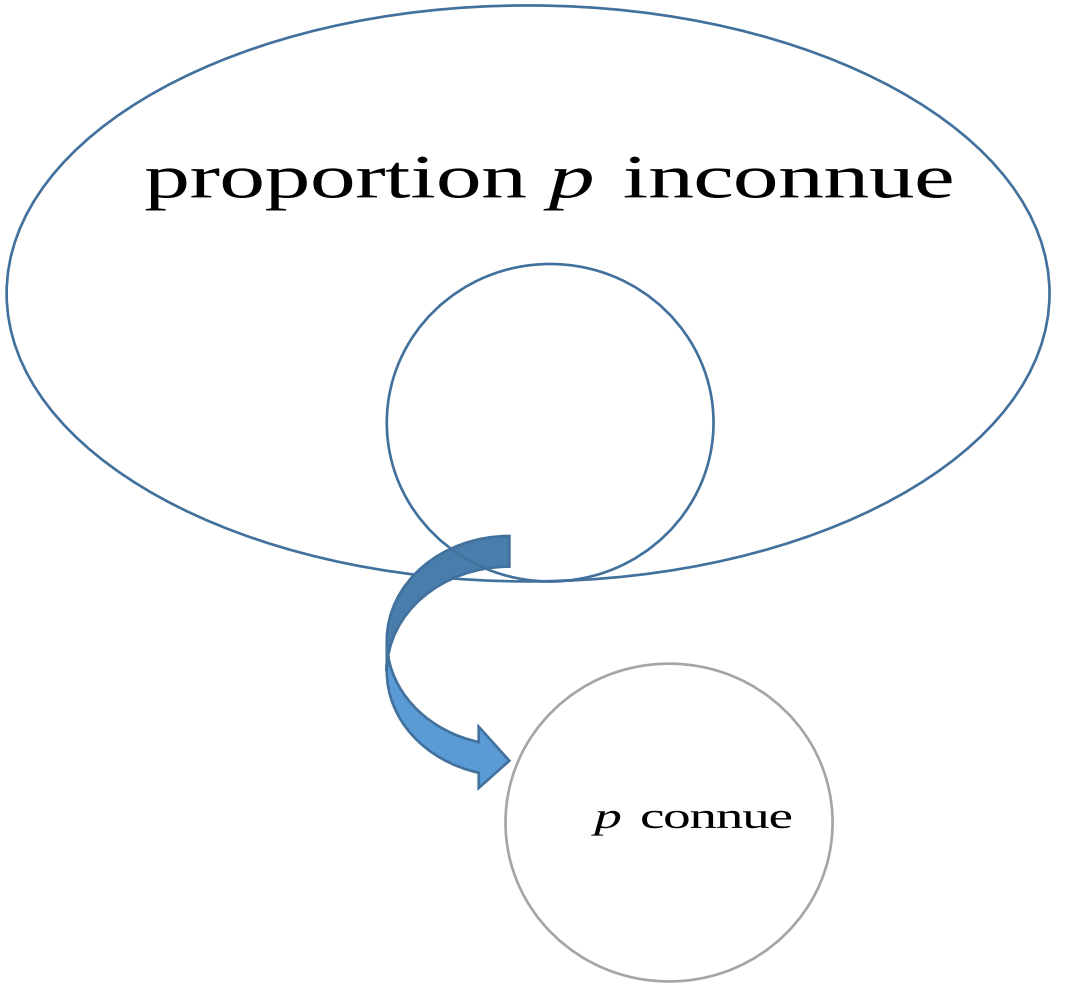
Population



Ehantillion de taille n

μ_2

Population



$Ehantillion\ de\ taille\ n$

μ_2

Estimation d'une moyenne

Estimation ponctuelle

Contexte : on considère une variable aléatoire *X* sur une population de moyenne (ou espérance) μ inconnue et d'écart-type σ inconnu (ou connu). On suppose que l'on a prélevé un échantillon de taille *n* (tirage avec remise ou assimilé) sur lequel on a calculé la moyenne μ_e et l'écart-type σ_e.

μ est l'éstimation ponctuelle de la moyenne μ.

$$\hat{\mu} = \mu_e.$$

$\hat{\sigma}$ est l'éstimation ponctuelle del'écart-type μ_e.

$$\hat{\sigma} = \sqrt{\frac{n}{n-1}} \times \sigma_e.$$

Le coefficient $\sqrt{\frac{n}{n-1}}$ s'appelle correction de biais. Lorsque la taille n de l'echantillon est assez grand (en pratique $n \geq 30$), ce coefficient est tres voisin de 1, si bien que, dans ce cas, on peut estimer $\hat{\sigma} = \sigma_e$.

Exemple :

Une université comporte 1500 étudiants. On mesure la taille de 20 d'entre eux.

La moyenne me et l'écart-type se calculés à partir de cet échantillon sont :

$$\mu_e = 176\,cm \; et \; \sigma_e = 6\,cm$$

On peut donc estimer les paramètres de la population :

$$\hat{\mu} = 176\,cm \; et \; \hat{\sigma} = \sqrt{\frac{20}{19}} \times 6{,}16\,cm$$

Remarque :

Nous n'avons fait qu'une estimation, il est bien sûr impossible de retrouver les vraies caractéristiques μ et σ de la population.

L'estimation ponctuelle permet surtout de disposer d'une **valeur de référence** pour poursuivre/affiner les calculs.

On souhaiterait notamment pouvoir faire une **estimation par intervalle**, en contrôlant le **risque** pris.

Estimation par intervalle de confiance

Le contexte est le meme que le precedent, sauf que nous allons raisonner en deux temps, une phase a priori (ou previsionnelle) dans lequelle on suppose que l'echantillon n'est pas encore prélevé et une phase a posteriori dans laquelle on suppose connue la moyenne μ_e et l'ecart-type σ_e de l'echantillon et donc la moyenne estimee $\hat{\mu}$ et l'ecart-type estime $\hat{\sigma}$ de la population.

PHASE *A PRIORI* - Mise en place du modèle prévisionnel

Nous avons vu, dans la théorie sur l'échantillonnage, que si $\bar{X}$ est la variable aléatoire correspondant à la moyenne d'un échantillon de taille n pris au hasard, alors le **Théorème Central Limite** permet d'affirmer que $\bar{X}$ suit approximativement une loi normale :

$$\bar{X} \sim N(\mu; \frac{\sigma}{\sqrt{n}})$$

Nous allons chercher un intervalle qui contient μ avec une **confiance arbitraire de 95%** (cela pourrait aussi être 99% ou un autre coefficient de confiance). Nous cherchons donc un rayon r tel que :

Probabilité que la moyenne μ de la population tombe dans un intervalle du type $[\bar{X} - r ; \bar{X} + r]$

$$P(\bar{X} - r \leq \mu \leq \bar{X} + r) = 0,95$$