

70	LE MODELE GAUSSIEN ET LE THEOREME CENTRAL	2
70.1	LA VARIABLE ALEATOIRE NORMALE	2
70.1.1	Utilité	2
70.1.2	Fonction de probabilité	2
70.1.3	Valeurs caractéristiques	3
70.1.4	Calcul de probabilité	4
70.2	VARIABLE NORMALE REDUITE	5
70.2.1	Utilité	5
70.2.2	Standardisation (ou réduction)	5
70.2.3	Valeurs caractéristiques	6
70.2.4	Propriétés	7
70.2.5	Tables et graphiques	7
70.3	THEOREME CENTRAL LIMITE	10
70.4	DISTRIBUTION D'ECHANTILLONNAGE DE M_X	12
70.4.1	Utilité	12
70.4.2	Augmentation de la précision	12
70.4.3	Erreur standard	15
70.5	DANS LE TABLEUR EXCEL	17

70 Le modèle Gaussien et le théorème central

70.1 La variable aléatoire normale

70.1.1 Utilité

La v.a. normale est un modèle qui permet de décrire la distribution de probabilité de nombreuses v.a. continues, ou de leur logarithme. Par exemple, la taille ou le poids des organismes, l'erreur de mesure, la concentration de nombreuses substances sanguines (taux de fibrinogène, cholestérol sanguin...), de substances dans l'eau (nitrites, nitrates)...

Ce modèle décrit aussi la distribution de la moyenne de tous les échantillons (parfois à partir d'une taille critique).

Le modèle normal est l'un des plus utilisés pour calculer des probabilités dans le domaine biomédical.

70.1.2 Fonction de probabilité

La fonction de densité de probabilité associée à cette variable est la fonction de Gauss -Laplace, qui représente la "courbe en cloche" bien connue.

Son équation est la suivante :

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Équation 70-1

Retenir cette équation n'est pas primordial, mais il est important de bien saisir la signification de ses paramètres.

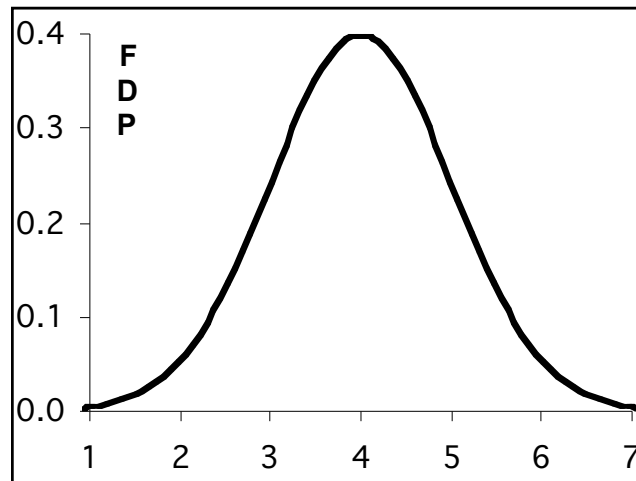


Figure 70 -1 Représentation graphique du modèle de la fonction de probabilité normale, décrite par l'équation de Gauss -Laplace.

70.1.3 Valeurs caractéristiques

Quelles sont les caractéristiques importantes de cette fonction?

Notons d'abord qu'elle est caractérisée par deux paramètres : μ et σ , la moyenne et l'écart -type (ou σ^2 la variance).

Elle est parfaitement symétrique (médiane = moyenne)

$f(X) \rightarrow 0$ lorsque $X \rightarrow \infty$.

Le maximum (dérivée première nulle) correspond au point $X = \mu$.

Il y a deux points d'inflexion (dérivée seconde nulle) qui correspondent aux points $\mu \pm \sigma$. La courbe sera donc d'autant plus étalée que σ est grand.

68% de la surface sous la courbe sont compris entre ces deux valeurs.

Au point $x = \mu$ correspond l'ordonnée $f(X)$

$$f(\mu) = \frac{1}{\sigma\sqrt{2\pi}} \quad \text{Équation 70-2}$$

Cette valeur est donc d'autant plus petite que σ est grand.

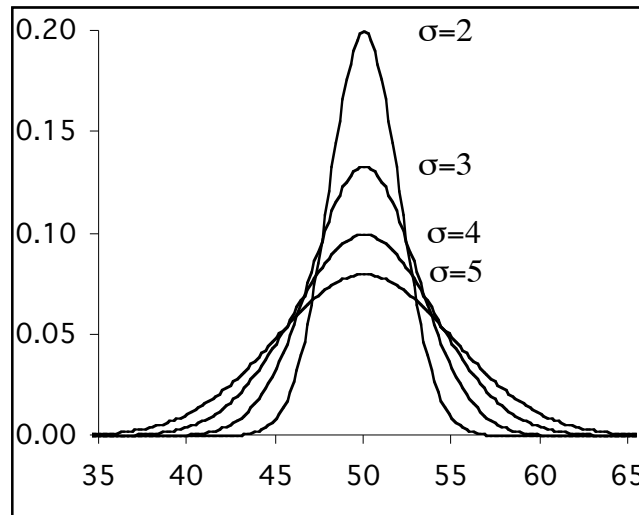


Figure 70 -2 Pour une valeur de $\mu = 50$, on trouve donc toute une famille de courbes, correspondant à différentes valeurs de σ . Ces courbes seront plates et étalées si σ est grand, pointues et étroites si σ est petit, la surface sous la courbe restant toujours égale à 1.

La variable aléatoire normale est notée conventionnellement

X : v.a. $N(\mu; \sigma^2)$

En appliquant une transformation linéaire on trouve que :

X : v.a. $N(\mu; \sigma^2)$,

$X' = a + bX$: v.a. $N(a + b\mu; b^2\sigma^2)$

70.1.4 Calcul de probabilité

La probabilité que X se trouve dans un intervalle donné correspond à la surface sous la courbe dans cet intervalle, et donc à l'intégrale de la fonction dans cet intervalle :

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} x \cdot f(X) dx$$

Cependant, il n'y a pas de solution analytique à cette intégrale : aucune formule ne donne $\Pi(X)$ en fonction des valeurs de x_1 , x_2 , μ et σ .

Si l'on a établi que le taux de fibrinogène est une v.a. ; $N(4; 1)$, comment résoudre la question de la firme pharmaceutique : déterminer la probabilité qu'un individu présente un taux de fibrinogène $\geq 4,5$?

Il faut recourir à une intégration numérique, c'est-à-dire estimer l'intégrale en sommant les surfaces d'un grand nombre de rectangles de base très petite.

Ce calcul étant long, les valeurs obtenues pour différents intervalles ont été tabulées. Etant donné qu'il y a une infinité de v.a. normales, chacune caractérisée par une moyenne et une variance propre, il a fallu choisir une distribution de référence, à partir de laquelle il est possible de recalculer rapidement les probabilités d'une v.a. normale particulière.

La variable de référence est la v.a. normale réduite; elle est notée Z , et son intégration numérique calculée une fois pour toutes est publiée dans une table de Z .

70.2 Variable normale réduite

70.2.1 Utilité

La v.a. normale réduite $Z \sim N(0;1)$ est une variable de référence dont la distribution de probabilité est tabulée. Elle permet de calculer des probabilités relatives à une v.a. normale pour n'importe quelle combinaison de paramètres, par exemple X , v.a. $N(5;2)$, X , v.a. $N(150;25)$ etc...

La transformation d'une variable en normale réduite, appelée standardisation ou réduction, permet également de supprimer les unités et de donner, dans une analyse impliquant plusieurs variables, un poids indépendant du système d'unité de référence. C'est ce que nous avons réalisé pour le calcul du coefficient de corrélation.

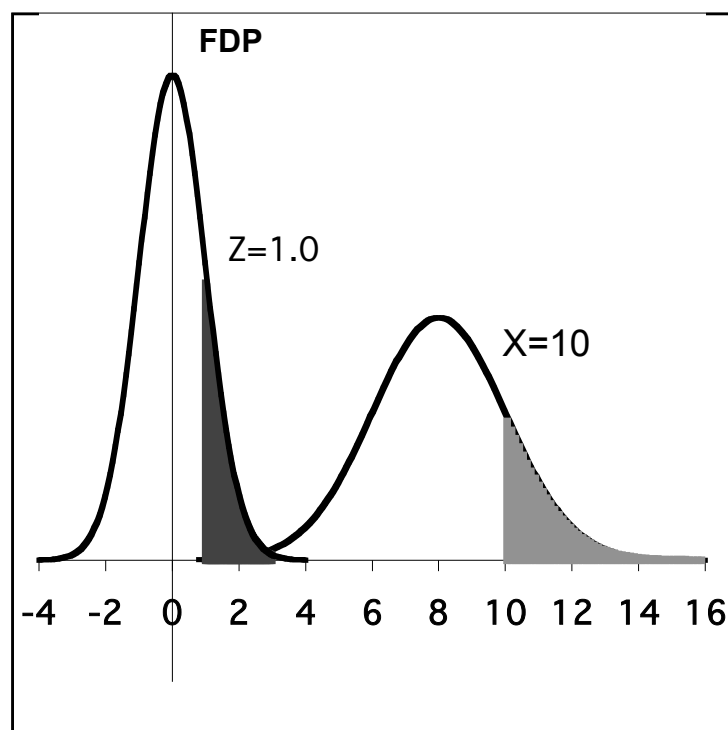


Figure 70 -3 Distribution de X v.a. $N(8 ; 4)$ et de Z v.a. $N(0 ; 1)$. $P(X \geq 10) = P(Z \geq 1)$.

70.2.2 Standardisation (ou réduction)

Disposant de la valeur $P(z_1 \leq Z \leq z_2)$, on doit donc retrouver x_1 et x_2 telles que

$$P(x_1 \leq X \leq x_2) = P(z_1 \leq Z \leq z_2)$$

Pour réduire une distribution quelconque $N(\mu, \sigma^2)$ en v.a. normale réduite $N(0;1)$, il faut effectuer les transformations suivantes :

1. centrer les valeurs , c'est -à -dire retirer la moyenne à chaque valeur

$$x' = x - \mu \text{ donc}$$

X' : v.a. $N(0, \sigma^2)$

2. Standardiser les valeurs centrées, c'est -à -dire les exprimer en unités écart - type :

$$z = (x - \mu)/\sigma \text{ ou } x = z\sigma + \mu \text{ avec } P(X \leq x) = P(Z \leq (x - \mu)/\sigma)$$

$$z_i = \frac{x_i - \mu}{\sigma} \quad \text{Équation 70-3}$$

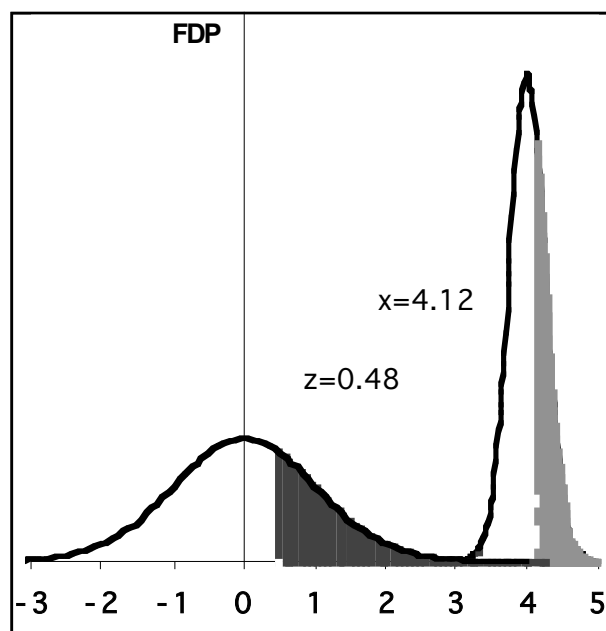


Figure 70 -4 Distribution de X v.a. $N(4 ; 0.0625)$ et de Z v.a. $N(0 ; 1)$. $\sigma = 0.25$; $P(X \geq 4,12) = P(Z \geq 0,48)$.

Cette transformation permet donc de passer d'une valeur particulière d'une variable normale quelconque à la valeur correspondante de la variable normale réduite, et donc de se référer à sa table de probabilité.

70.2.3 Valeurs caractéristiques

Suivant notre transformation, nous obtenons bien une v.a. $N(0;1)$. En effet,

$$E(Z) = E\left(\frac{x - \mu}{\sigma}\right) = E\left(\frac{x}{\sigma} - \frac{\mu}{\sigma}\right) = E\left(\frac{x}{\sigma}\right) - E\left(\frac{\mu}{\sigma}\right) =$$

$$\frac{1}{\sigma} E(x) - \left(\frac{\mu}{\sigma}\right) = \frac{1}{\sigma} \mu - \left(\frac{\mu}{\sigma}\right) = \left(\frac{\mu}{\sigma}\right) - \left(\frac{\mu}{\sigma}\right) = 0$$

$$VAR(Z) = VAR\left(\frac{x - \mu}{\sigma}\right) = VAR\left(\frac{x}{\sigma} - \frac{\mu}{\sigma}\right) = VAR\left(\frac{x}{\sigma}\right) - VAR\left(\frac{\mu}{\sigma}\right) =$$

$$VAR\left(\frac{x}{\sigma}\right) - 0 = \left(\frac{1}{\sigma^2}\right) VAR(X) = \left(\frac{1}{\sigma^2}\right) \sigma^2 = 1$$

70.2.4 Propriétés

Etant donné la symétrie des valeurs autour de 0, on trouve :

$$P(Z \leq -z) = P(Z \geq z)$$

La probabilité totale étant égale à 1, on trouve :

$$P(Z \geq z) = 1 - P(Z \leq z)$$

La probabilité que $Z = z$ étant nulle, on trouve :

$$P(z_1 \leq Z \leq z_2) = P(Z \leq z_2) - P(Z \leq z_1)$$

70.2.5 Tables et graphiques

Ces propriétés permettent de calculer la probabilité de n'importe quel intervalle sur Z en se référant à la fonction de répartition $p(Z \leq z)$ tabulée pour $Z \geq 0$

0,900	0,950	0,975	0,990	0,995
1,282	1,645	1,960	2,326	2,576

Tableau 70 -70-1 Extrait de la table Z. La première ligne donne une sélection des valeurs de probabilités $P(Z \leq z)$. La seconde ligne donne la valeur de Z correspondant à cette probabilité.

Pour des raisons de mise en page, cette table constitue la dernière ligne de la table t de Student (degrés de liberté ∞).

Une table plus complète est disponible dans les annexes de ce cours. Elle est un peu plus complexe à lire et est rarement nécessaire.

Pour des calculs spécifiques, nous vous renvoyons aux fonctions du tableur Excel.

Reprenons les paramètres du taux de fibrinogène, supposé être une v.a. $N(4 ; 1)$.

Quelle est la probabilité de trouver dans cette population un individu dont le taux de fibrinogène est supérieur à 4,5 ?

Calculer la valeur de z

A la valeur 4,5 dans cette distribution, correspond la valeur $(4,5 - 4)/1 = 0,5$ dans la distribution normale réduite.

Définir la probabilité recherchée

$$P(X \geq 4,5) = P(Z \geq 0,5)$$

Calcul de la probabilité

$$P(Z \geq 0,5) = 1 - P(Z \leq 0,5) = 1 - 0,691 = 0,309$$

On peut aussi répondre à la question : entre quelles limites 95% des individus de la population sont-ils situés?

Rechercher les valeurs de Z correspondant à cette probabilité

$$P(z_1 \leq Z \leq z_2) = 0,95$$

en utilisant la symétrie de la courbe :

$$P(Z \leq z_2) = 0,975$$

$$z_2 = 1,96 \text{ et donc } z_1 = -1,96$$

Calculer les valeurs x1 et x2 correspondantes

$$x_1 = -1,96 * 1 + 4 = 2,04$$

$$x_2 = 1,96 * 1 + 4 = 5,96$$

Ce calcul justifie notre approximation en statistique descriptive : 95% des valeurs environ sont comprises entre $M \pm 2S$, ce qui donne ici $\mu \pm 2\sigma = 4 \pm 2$.

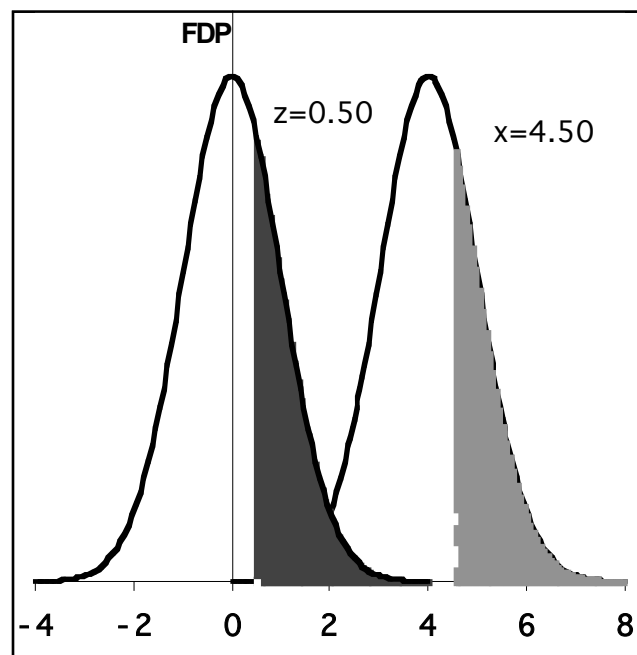


Figure 70 -5 Taux de fibrinogène $N(4 ; 1)$. $P(X \geq 4.5)$ et $P(Z \geq 0,5)$,

70.3 Théorème central limite

Imaginons les points obtenus lors d'un grand nombre de jets d'un seul dé bien équilibré. La distribution de fréquences obtenue montre une distribution rectangulaire :

$P(X = x) = \text{constante} = 1/6$ (soit 0,16).

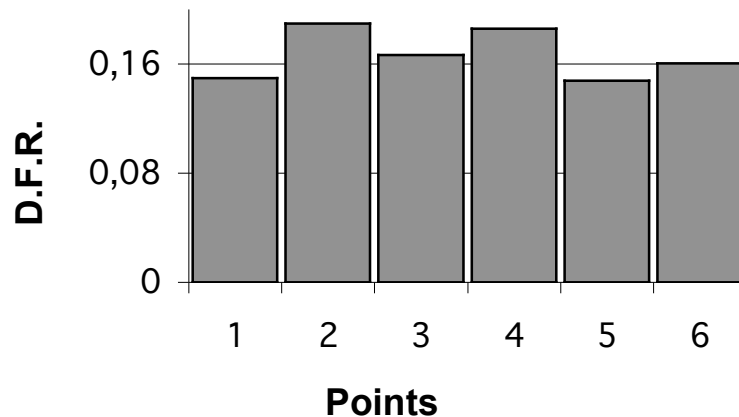


Figure 70 -6 Distribution des points obtenus pour le jet d'un seul dé réalisé un grand nombre de fois.

Si l'on effectue un grand nombre de fois le jet de 4 dés, la distribution d'échantillonnage de la moyenne des points obtenus, $Mx(4)$, est une pyramide symétrique.

Les valeurs proches de 1 et 6 sont rares, car pour obtenir une moyenne de 1 il faut obtenir 4 fois 1, avec une probabilité de réalisation de 0,00077. Par contre la probabilité d'obtenir une valeur proche de 3 est élevée car de nombreuses combinaisons de 3, de 4 et 2, de 5 et 1 forment une moyenne de 3.

Si l'on considère le jet de 16 dés, la distribution obtenue se superpose pratiquement à une variable aléatoire normale.

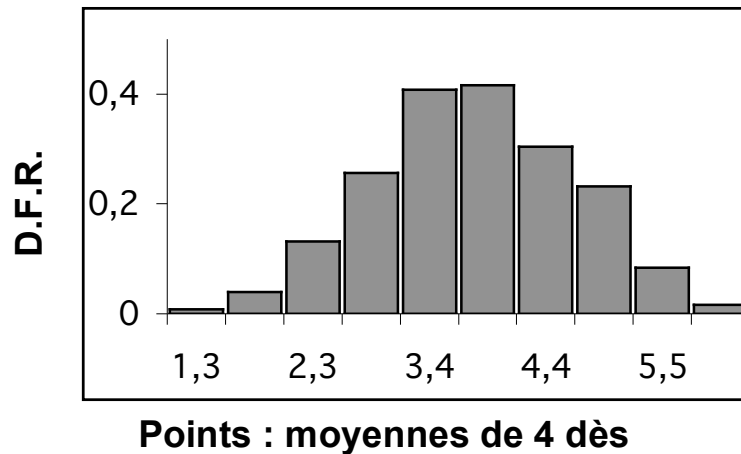


Figure 70 -7 Distribution d'échantillonnage de la moyenne des points obtenus pour le lancer de 4 dés, réalisé un grand nombre de fois.

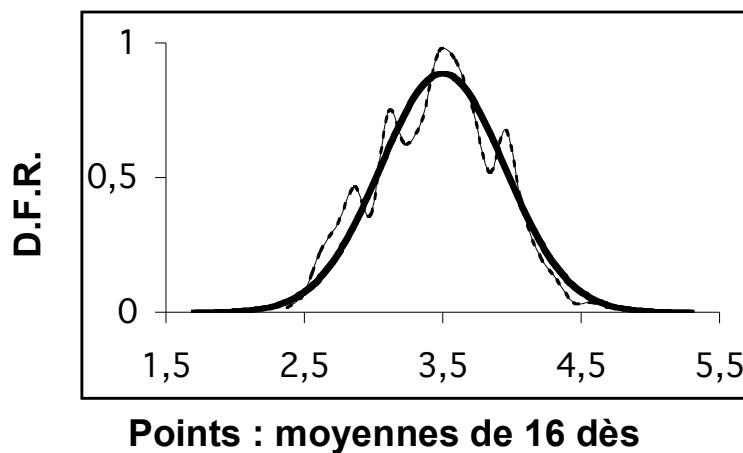


Figure 70 -8 Distribution d'échantillonnage de la moyenne des points obtenus pour le lancer de 16 dés, réalisé un grand nombre de fois (trait fin) et modèle de la v.a. Normale de variance σ^2/n (trait épais).

Le théorème central limite nous enseigne que la variance de la distribution normale obtenue est de σ^2/n .

Nous acceptons cela sans démonstration, en remarquant qu'il est logique que cette variance diminue lorsque n augmente, puisque la moyenne a de moins en moins de chances d'atteindre des valeurs extrêmes.

Si X est une v.a. de distribution quelconque, considérée dans une population de moyenne et variance finies μ et σ^2 , la distribution de M_x , moyenne d'un échantillon de taille n prélevé dans cette population, tend vers une v.a. Normale, de moyenne μ et de variance σ^2/n .

La distribution de la moyenne sera d'autant plus vite normale que la distribution de X se rapproche déjà de la normalité. On considère que la distribution de Mx est dans tous les cas v.a. N lorsque $n \geq 30$ observations, et pour n bien plus petit encore dans certaines conditions, concernant surtout la symétrie de la distribution de X .

Ceci nous montre l'importance fondamentale de la v.a. Normale en statistique :

- *dans le domaine biomédical, beaucoup de variables, ou le logarithme de ces variables, sont normales ;*
- *la Binomiale et la Poisson tendent vers la distribution normale lorsque $n\pi$ ou μ augmente ;*
- *la moyenne d'un échantillon d'une certaine taille est toujours une variable normale.*

70.4 Distribution d'échantillonnage de Mx

70.4.1 Utilité

Le théorème central limite nous a montré que les paramètres de la distribution de la moyenne d'un échantillon de taille n pouvaient être déduits des paramètres de la distribution de la variable, quelque soit n si la distribution est normale, et quelle que soit la distribution si n est suffisamment grand (disons $n > 30$ pour une distribution quelconque).

Dans de nombreux exemples, nous nous intéressons aux variables qui sont normales au départ :

$$X : \text{v.a. } N(\mu ; \sigma^2)$$

$Mx(n) : \text{v.a. } N(\mu ; \sigma^2/n) \{5.1\}$
--

La distribution d'échantillonnage de Mx , basée sur la v.a. normale, est utilisée comme modèle pour toutes les questions relatives à l'estimation de la moyenne d'une population. Ceci est l'objet d'un très grand nombre d'expériences biomédicales : étude de la croissance des poissons, comparaison d'un taux d'hormones entre mâles et femelles, évolution de la concentration en phosphates dans les nappes phréatiques, de la concentration en ozone dans l'atmosphère, etc...

70.4.2 Augmentation de la précision

Le théorème central limite explique pourquoi il est possible d'augmenter la précision d'une mesure en la répétant.

Lorsqu'on réalise un dosage de cholestérol dans le sang, l'erreur de mesure peut -être considérée comme une v.a. $N(0;0,16)$. La mesure est donc exacte ($\mu = 0$), mais imprécise. Cette imprécision étant liée au protocole de mesure, la seule façon d'obtenir une mesure plus précise est d'effectuer plusieurs mesures et d'en prendre la moyenne. Si je prends 4 mesures, par exemple, quelle est la probabilité que la moyenne calculée soit supérieure à la valeur réelle de plus de 0,1?

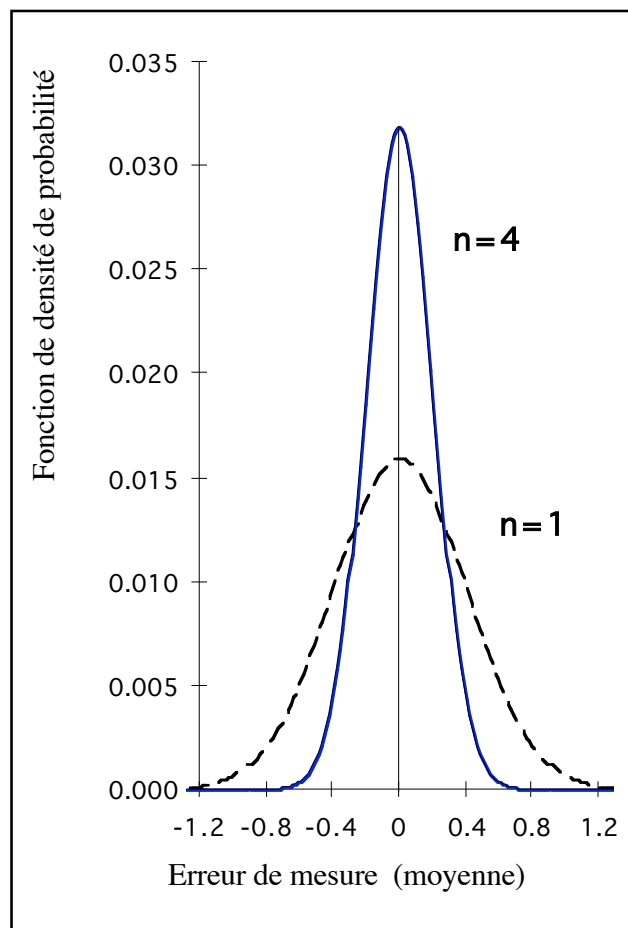


Figure 70 -9 Distribution de X et de $Mx(4)$ pour une variable $X \sim N(0;0,16)$.

Définir la variable :

X , erreur de mesure (écart entre valeur mesurée et valeur réelle) $Mx(4)$: moyenne de l'erreur de 4 mesures, v.a. $N(0; 0,04)$

Définir la probabilité recherchée :

$$P(Mx(4) \geq 0,1)$$

Standardiser la valeur $Mx(4)$:

$$z = \frac{(0.1 - 0)}{\sqrt{0.04}}$$

Rechercher la probabilité :

$$P(Z \geq 0,5) = 1 - P(Z \leq 0,5) = 0,31$$

De la même façon, nous pourrions calculer que pour X (une seule mesure) $p(x \geq 0,1) = 0,41$. Ainsi, le fait de prendre la moyenne de 4 mesures a permis de diminuer l'imprécision, mais elle reste encore grande.

Posons -nous donc la question dans un autre sens : combien de mesures faudrait -il prendre pour être presque sûr que la moyenne de ces mesures ne s'éloigne pas de la valeur réelle de plus de 0,2?

-Définir la variable :

X , erreur de mesure (écart entre valeur mesurée et valeur réelle)

$Mx(n)$: moyenne de l'erreur de n mesures, v.a. $N(0; 0,16/n)$,

-Définir la probabilité :

$$P(-0,2 \leq Mx(n) \leq 0,2) = 0,95$$

-Standardiser :

$$P(-z \leq Z \leq z) = 0,95 \rightarrow p(Z \leq z) = 0,975 \rightarrow z = 1,96$$

$$Mx(n) = \frac{z\sigma}{\sqrt{n}} + \mu$$

$$0.2 = \frac{1.96 \times 0.4}{\sqrt{n}} + 0$$

-Rechercher n :

$$\sqrt{n} > (1,96 \times 0,4) / 0,2 = 3,92$$

$$n > 15,36 \text{ et donc } n \geq 16$$

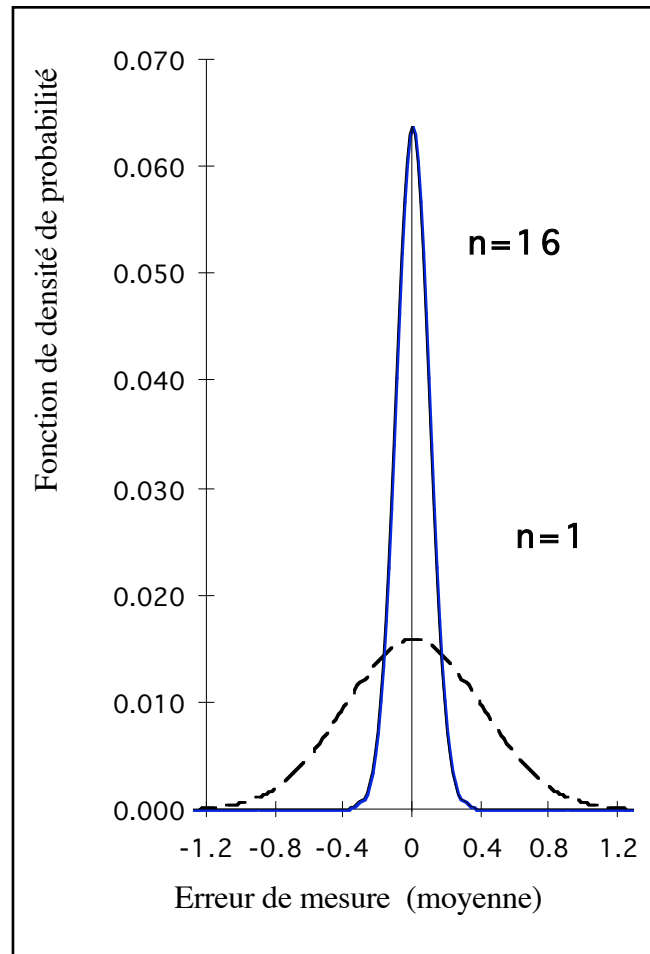


Figure 70 -10 Distribution de X et de $Mx(16)$ pour une variable $X \sim N(0; 0,16)$.

Il faudra donc 16 mesures au moins pour arriver avec quasi -certitude à la précision voulue.

70.4.3 Erreur standard

$$\sqrt{\frac{S^2}{n}} = \frac{S}{\sqrt{n}} \quad \text{Équation 70-4}$$

s'appelle erreur standard (S.E.) et se distingue fondamentalement de l'écart - type ou déviation standard (S.D.)

Dans des tableaux ou des graphiques, la variabilité peut s'exprimer sous forme de S.D. ou de S.E. En principe, ce choix doit être clairement exprimé dans la légende.

Supposons que la distribution du taux de cholestérol dans le sang (mg/100ml) pour la population humaine soit bien connue et caractérisée par une v.a.N(210;1600).

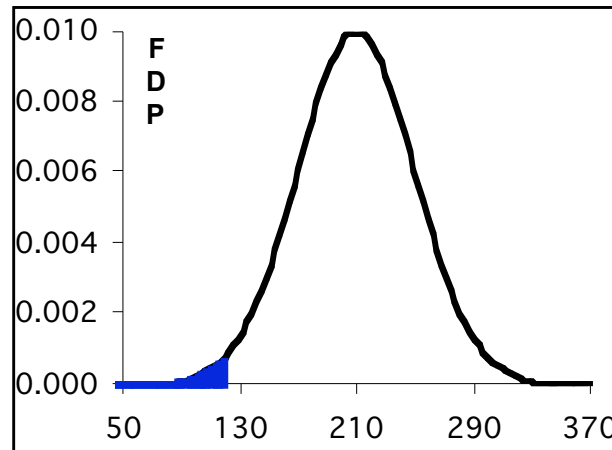


Figure 70 -11 Distribution du taux de cholestérol dans le sang v.a.N(210;1600) mg/100ml.

I

Quelle est la probabilité de trouver dans cette population un individu dont le taux est inférieur à 120 mg/100ml?

Calculer la valeur de z : à la valeur 120 mg/100ml, correspond la valeur $(120 - 210)/40 = -2,25$ dans la distribution normale réduite.

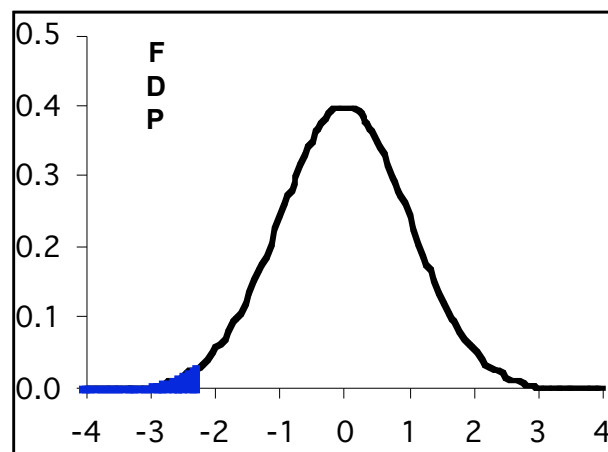


Figure 70 -12 Distribution de Z. La partie hachurée représente $P(Z \leq -2.25)$.

Définir la probabilité recherchée

$$P(X \leq 120) = P(Z \leq -2,25)$$

Calcul de la probabilité

En utilisant la symétrie de la v.a.normale, on trouve :

$$P(Z \leq -2,25) = P(Z \geq 2,25) = 1 - P(Z \leq 2,25) = 1 - 0,988 = 0,012$$

Entre quelles limites inférieure et supérieure 95% des individus de la population sont-ils situés?

Rechercher les valeurs de Z correspondant à cette probabilité

$$P(z_1 \leq Z \leq z_2) = 0,95$$

en utilisant la symétrie de la courbe :

$$P(Z \leq z_2) = 0,975$$

$$z_2 = 1,96 \text{ et donc } z_1 = -1,96$$

Calculer les valeurs x_1 et x_2 correspondantes

$$x_1 = -1,96 * 40 + 210 = 131,6$$

$$x_2 = 1,96 * 40 + 210 = 288,4$$

70.5 Dans le tableur Excel

la fonction $LOI.NORMALE(x;\mu;\sigma;VRAI)$ renvoie $P(X \leq x)$ pour X v.a. $N(\mu;\sigma)$

Dans notre exemple,

$LOI.NORMALE(120;210;40;VRAI)$ renvoie 0,012.

$LOI.NORMALE.INVERSE(\pi;\mu;\sigma)$ renvoie X v.a. $N(\mu;\sigma)$ pour $\pi = P(X \leq x)$

$LOI.NORMALE.INVERSE(0,975;0;1)$

renvoie 1,96

Prenez garde, les conventions d'Excel varient d'une fonction à l'autre.