# PROBABILITÉS ET STATISTIQUE
## 6. Variables aléatoires continues

Carmelo Vaccaro

Université de Paris-Saclay

M1 - Informatique Science des Données (ISD) - Apprentissage
2022/23: premier semestre

# Aperçu

Les distributions normales apparaissent souvent dans les applications réelles et jouent un rôle important dans les méthodes de statistique inférentielle.

Dans ce chapitre, nous présentons les concepts des distributions normales qui seront souvent utilisés dans les autres chapitres de ce cours.

# Contenu du chapitre

1. Variables aléatoires uniformes et normales
2. Théorème de la limite centrale
3. La loi normale comme approximation de la loi binomiale

## 1. Variables aléatoires uniformes et normales

# Remarque sur les variables aléatoires continues

Soit $X$ une variable aléatoire continue et soient $a$ et $b$ des nombres réels tels que $a < b$. Alors comme $P(X = d)$ est zéro pour tout réel $d$, donc on a que

$$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b).$$

# Centiles pour les variables aléatoires continues

Pour les variables aléatoires continues on définit les centiles de la manière suivante.

Soit $X$ une variable aléatoire continue et soit $\alpha \in ]0, 1[$. Le *centile* $\alpha$ de $X$ est la valeur $c_\alpha$ telle que

$$P(X \leq c_\alpha) = \alpha.$$

# Variable aléatoire uniforme

Soient $m_1$ et $m_2$ deux nombres réels tels que $m_1 < m_2$. Une variable aléatoire $X$ est dite *uniforme entre $m_1$ et $m_2$* si sa densité de probabilité est constante dans l'intervalle $[m_1, m_2]$ et zéro en dehors de cet intervalle.

Cela veut dire que

1. $P(X < m_1) = 0$ ;
2. $P(X > m_2) = 0$ ;
3. si $a$ et $b$ sont des nombres tels que $m_1 \leq a < b \leq m_2$ alors

$$P(a \leq X \leq b) = \frac{b-a}{m_2 - m_1}.$$

En particulier, la densité de probabilité de $X$ est égale à $\frac{1}{m_2 - m_1}$ dans l'intervalle $[m_1, m_2]$.

# Variable aléatoire uniforme : remarque

Soit $X$ une variable aléatoire uniforme entre $m_1$ et $m_2$. Alors si $a$ et $b$ sont des nombres tels que $a < m_1 \leq b \leq m_2$ alors

$$P(a \leq X \leq b) = P(m_1 \leq X \leq b) = \frac{b - m_1}{m_2 - m_1};$$

si $m_1 \leq a < m_2 < b$ alors

$$P(a \leq X \leq b) = P(a \leq X \leq m_2) = \frac{m_2 - a}{m_2 - m_1}.$$

# Variable aléatoire uniforme : calculer des probabilités

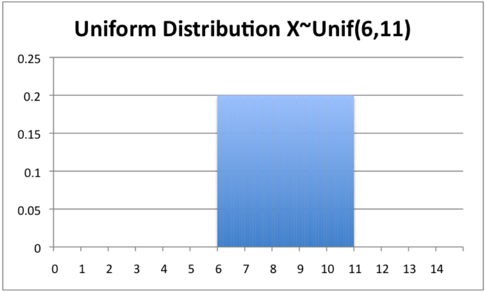
Soit $X$ une variable aléatoire uniforme entre $m_1$ et $m_2$. Soit $a$ et $b$ des nombres (finis ou infinis) tels que $a < b$. Calculons $P(a \leq X \leq b)$.

D'abord calculons l'intersection entre les intervalles $[m_1, m_2]$ et [a, b]. Si cette intersection est vide alors $P(a \leq X \leq b) = 0$. Sinon cette intersection sera un intervalle $[c, d]$ pour des opportuns réels $c$, $d$. Dans ce cas $P(a \leq X \leq b) = \dfrac{d - c}{m_2 - m_1}$.

# Exemple (1/3)

La figure ci-dessous montre le graphique de la distribution de probabilité d'une distribution uniforme comprise entre 6 et 11.

La densité de cette distribution est la fonction qui est égale à $1/(11 - 6) =$ 0,2 entre 6 et 11 et égale à 0 avant 6 et après 11. L'aire de cette fonction est 1 (comme pour toute fonction de densité).



# Exemple (2/3)

Pour la distribution de probabilité précédente (uniforme entre 6 et 11) déterminer la probabilité que la variable aléatoire soit

1. entre 7,25 et 8,75 ;
2. entre 9 et 13,15 ;
3. entre 2 et 4 ;
4. inférieure à 6,5.

# Exemple (3/3)

**RÉPONSE.**

1. comme l'intervalle $[7{,}25\ ,\ 8{,}75]$ est contenu dans l'intervalle $[6\ ,\ 11]$, alors la probabilité est égale à $0{,}2 \times (8{,}75 - 7{,}25) = 0{,}3$.
2. l'intersection entre l'intervalle $[9\ ,\ 13{,}15]$ et l'intervalle $[6\ ,\ 11]$ est $[9\ ,\ 11]$, donc la probabilité est égale à $0{,}2 \times (11 - 9) = 0{,}4$.
3. comme l'intervalle $[2\ ,\ 4]$ est extérieur à l'intervalle $[6\ ,\ 11]$, alors la probabilité est égale à 0.
4. ici on a l'intervalle $]-\infty, 6{,}5[$, dont l'intersection avec l'intervalle $[6\ ,\ 11]$ est $[6\ ,\ 6{,}5[$, donc la probabilité est égale à $0{,}2 \times (6{,}5 - 6) = 0{,}1$.

# Exercice

La compagnie d'électricité fournit de l'électricité avec des niveaux de tension qui sont uniformément répartis entre 123 volts et 125 volts. Si $X$ est la variable aléatoire des niveaux de tension, trouvez la probabilité d'avoir un niveau de tension

1. entre 123,2 et 123,9 volts ;
2. supérieur à 124,5 volts ;
3. entre 120 et 124 volts ;
4. entre 125 et 126 volts.

Soit $\mu$ et $\sigma$ deux nombres réels tels que $\sigma > 0$. Une *variable aléatoire normale de moyenne* $\mu$ *et écart type* $\sigma$ est une variable aléatoire continue dont la fonction de densité est égale à

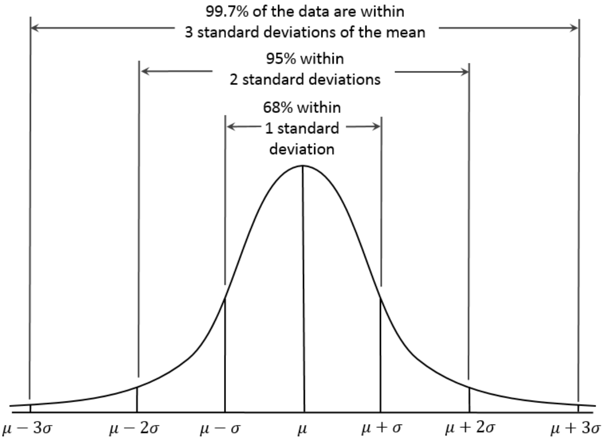
$$\frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}.$$

Cela veut dire que si $X$ est une variable aléatoire normale de moyenne $\mu$ et écart type $\sigma$ alors pour $a, b$ réels tels que $a < b$ on a

$$P(a \leq X \leq b) = \frac{1}{\sqrt{2\pi}\sigma} \int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

# Variable aléatoire normale

La distribution de probabilité d'une variable aléatoire normale de moyenne $\mu$ et écart type de $\sigma$ est montrée dans la figure suivante



# Variable aléatoire normale

La figure de la slide précédente nous montre que dans une variable aléatoire normale 68% des valeurs sont compris entre un écart type de la moyenne, 95% des valeurs sont compris entre deux écarts type de la moyenne et 99,7% des valeurs entre trois écarts type de la moyenne.

Donc seulement 5% des valeurs sont inhabituels.

# Variable aléatoire normale standard

Si $\mu = 0$ et $\sigma = 1$ on parle de *variable aléatoire normale standard*. Dans ce cas la densité de probabilité est égale à

$$\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}.$$

# Remarque sur la cote Z

Si on a un ensemble de données de moyenne 0 et écart type 1 (comme par exemple une variable aléatoire normale standard), alors la cote Z de chaque valeur est égale à la valeur elle même.

# Exemple 1 (1/2)

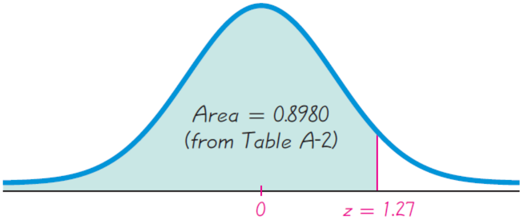
Dans un lot de thermomètres la mesure de la température à laquelle l'eau congèle est distribuée de manière normale avec une moyenne de $0^\circ$C et un écart type de $1^\circ$C.

Si un thermomètre est choisi au hasard dans le lot, trouver la probabilité que, au point de congélation de l'eau, la lecture de la température soit inférieure à $1{,}27^\circ$C.

# Exemple 1 (2/2)

Pour trouver cette probabilité il faut prendre le graphe de la distribution de probabilité de ces températures (normale standard), puis choisir sur l'axes des abscisses la valeur 1,27 et trouver l'aire correspondante à gauche car on veut trouver la probabilité que la lecture soit inférieure à cette valeur.

Cette aire est égale à 0,898, donc la probabilité recherché est 0,898.



# Exemple 2 (1/2)

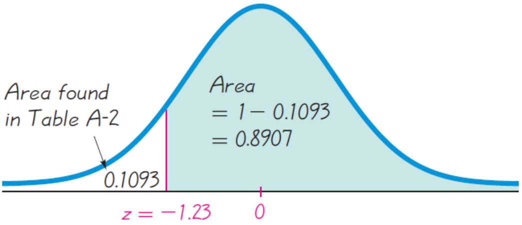
Toujours le même lot de thermomètres dont la mesure de la température de congélation de l'eau est distribué de manière normale avec une moyenne de 0°C et un écart type de 1°C.

Si un thermomètre est choisi au hasard, trouver la probabilité que, au point de congélation de l'eau, la lecture de la température soit supérieure à -1,23°C.

# Exemple 2 (2/2)

Pour trouver cette probabilité il faut choisir sur l'axes des abscisses la valeur -1,23 et trouver l'aire correspondante à droite car on veut trouver la probabilité que la lecture de la température soit supérieure à cette valeur.

L'aire de gauche est égale à 0,1093, donc celle de droite est la différence entre toute l'aire (égale à 1) et 0,1093, donc à 1 - 0,1093 = 0,8907.



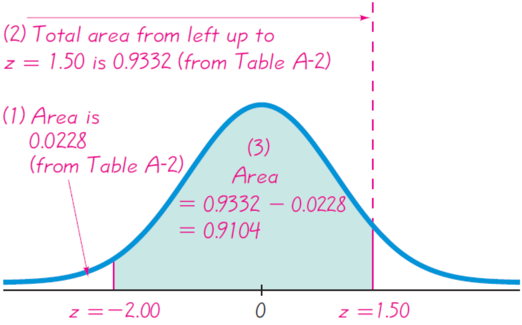
# Exemple 3 (1/2)

Toujours la même situation, trouver la probabilité que, au point de congélation de l'eau, la lecture de la température est comprise entre -2°C et 1,5°C.

# Exemple 3 (2/2)

Pour trouver cette probabilité il faut choisir sur l'axes des abscisses les valeurs -2 et 1,5 et faire la différence entre l'aire à gauche de 1,5 moins celle à gauche de -2.

L'aire à gauche de 1,5 est 0,9332, celle à gauche de -2 est 0,0228, donc l'aire recherchée est égale à 0,9332 - 0,0228 = 0,9104.



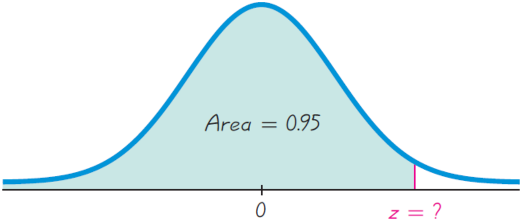
# Exemple 4 (1/2)

Dans les trois exemples précédentes nous sommes partis d'une valeur donnée et nous avons trouvé l'aire. Maintenant nous allons faire l'opposé, nous partons d'une aire et trouvons la valeur. En effet de cette façon nous trouvons les centiles.

Avec les mêmes thermomètres que dans les exemples précédentes, trouver la température correspondante au centile 0,95, c'est-à-dire la température qui sépare les 95% des valeurs les plus petites du 5% des valeurs les plus grandes.

# Exemple 4 (2/2)

La valeur $z$ indiquée dans la figure est telle que l'aire à gauche d'elle est égale à 0,95. Cette valeur correspond à 1,645.



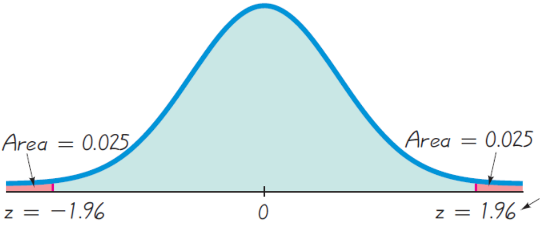
# Exemple 5 (1/2)

Avec les mêmes thermomètres que dans les exemples précédentes, trouver trouver les températures qui séparent les 2,5% les plus basses et les 2,5% les plus hautes.

# Exemple 5 (2/2)

On doit trouver les températures qui séparent les 2,5% les plus basses du 97,5% les plus hautes et les températures qui séparent les 2,5% les plus hautes du 97,5% les plus basses.

L'aire de 0,25 correspond à z = -1,96. L'aire de 0,975 correspond à z = 1,96.



# Exemple 6 (1/2)

Dans cet exemple on utilise une variable aléatoire non standard.

Assumons que les poids à la naissance des bébés soient normalement distribués avec une moyenne de 3420 g et un écart type de 495 g et qu'on désigne comme anormalement léger un poids inférieur à 2450 g et comme anormalement lourd un poids supérieur à 4390 g.

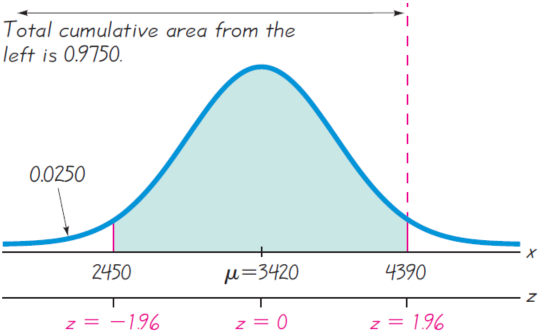
Quel est le pourcentage de bébés qui ont un poids jugé normale, entre 2450 g et 4390 g ?

# Exemple 6 (2/2)

Pour une normale standard les valeurs coïncident avec leurs Z cotes. Ici on a une normale non standard et donc on calcule les Z cotes

$$\frac{2450-3420}{495}=-1{,}96 \qquad \frac{4390-3420}{495}=1{,}96.$$

L'aire à gauche de z = -1,96 est 0,025, celle à gauche de z = 1,96 est 0,975, donc l'aire recherchée est égale à 0,975 - 0,025 = 0,95.



# Exercice

Pour une variable aléatoire $X$ normale de moyenne 4,5 et écart type 1,5 trouver :

1. la probabilité que $X$ soit inférieure à 6 ;
2. la probabilité que $X$ soit supérieure à 2 ;
3. la probabilité que $X$ soit comprise entre 1,75 et 5,75 ;
4. la valeur $\beta$ telle que $P(X \leq \beta) = 0{,}35$ ;
5. la valeur $\alpha$ telle que $P(X \geq \alpha) = 0{,}40$ ;
6. les valeurs $\gamma$ et $\delta$ telles que $[\gamma, \gamma]$ est un intervalle centré en 4,5 et $P(\gamma \leq X \leq \gamma) = 0{,}55$.

# 2. Théorème de la limite centrale

# Concept clé

Le théorème de la limite centrale dit que pour une population ayant une distribution quelconque, la distribution des moyennes des échantillons se rapproche d'une distribution normale à mesure que la taille de l'échantillon augmente.

La procédure décrite dans cette section constitue la base de l'estimation des paramètres de la population et des tests d'hypothèse.

# Théorème de la limite centrale : hypothèses

On a une variable aléatoire $X$ de distribution quelconque (normale ou non) de moyenne $\mu$ et d'écart type $\sigma$.

Des échantillons aléatoires simples, tous de taille $n$, sont sélectionnés dans la population.

Rappelons que pour avoir des échantillons aléatoires simples de taille $n$ tous les échantillons possibles de taille $n$ doivent avoir la même chance d'être sélectionnés.

# Théorème de la limite centrale : conclusions

La distribution des moyennes des échantillons de taille $n$ se rapproche d'une distribution normale pour $n$ qui va à l'infini.

La moyenne des moyennes des échantillons de taille $n$ est égale à la moyenne de la population, $\mu$, pour tout $n$.

L'écart type des moyennes des échantillons de taille $n$ est égal à $\sigma/\sqrt{n}$.

# Règles pratiques couramment utilisées

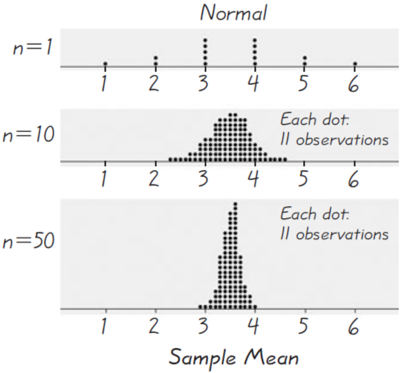
Pour les échantillons de taille $n$ supérieure à 30, la distribution des moyennes des échantillons peut être raisonnablement approximée par une distribution normale. L'approximation se rapproche d'une distribution normale à mesure que la taille de l'échantillon $n$ devient plus grande.

Si la population initiale est normalement distribuée, alors pour toute taille d'échantillon $n$, les moyennes des échantillons seront normalement distribuées (pas seulement les valeurs de $n$ supérieures à 30).

Si la population initiale n'est pas normalement distribuée et que la taille d'échantillon est inférieure à 30, alors on ne peut pas approximer la distribution des moyennes des échantillons par une normale.

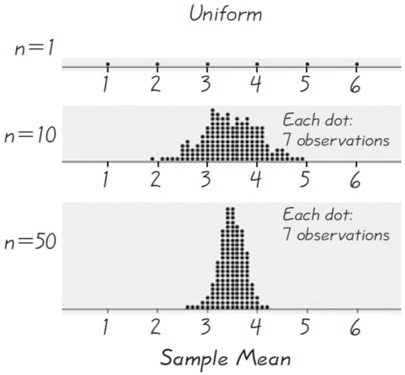
# Exemple - Distribution normale

Dans cet exemple on voit une distribution normale. Pour $n = 10$ la distribution des moyennes se rapproche à une distribution normale et pour $n = 50$ ce rapprochement est même plus évident.



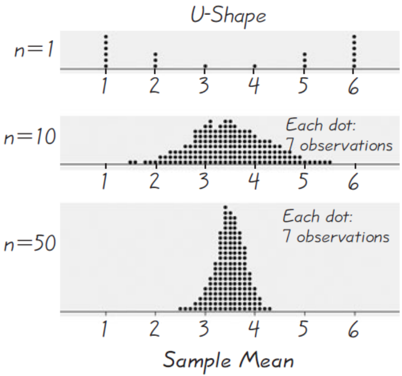
# Exemple - Distribution uniforme

Dans cet exemple on voit une distribution uniforme. Pour $n = 10$ la distribution des moyennes commence à se rapprocher à une distribution normale et pour $n = 50$ ce rapprochement est visible.



# Exemple - Distribution à U

Dans cet exemple on voit une distribution à U. Pour $n = 10$ la distribution des moyennes commence à se rapprocher à une distribution normale et pour $n = 50$ ce rapprochement est visible.



# Exemple (1/2)

Un bateau a coulé dans l'Inner Harbor de Baltimore. Parmi les 25 personnes à bord, 5 sont mortes et 16 ont été blessées.

La charge maximale de poids de passagers que le bateau pouvait transporter était de 1587 kg. En supposant un poids moyen des passagers de 63 kg, le bateau était autorisé à transporter 25 passagers.

La moyenne de 63 kg avait été déterminée il y a 44 ans, lorsque les gens n'étaient pas aussi lourds qu'aujourd'hui. (Le poids moyen des 25 passagers à bord du bateau qui a coulé s'est avéré être de 76 kg.)

# Exemple (2/2)

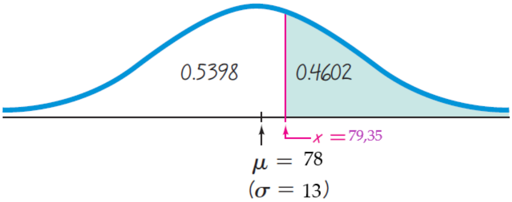
Les données plus récentes estiment que le poids des hommes a une moyenne de 78 kg et un écart type de 13 kg. Nous assumons aussi que les poids sont normalement distribuées.

a. Trouver la probabilité que si un homme est choisi au hasard, son poids soit supérieur à 79,35 kg.

b. Trouver la probabilité que 20 hommes choisis au hasard aient un poids moyen supérieur à 79,35 kg (de sorte que leur poids total dépasse la capacité de sécurité de 1587 kg).

# Solution de a

Dans le graphe représentant la distribution normale de moyenne 78 et écart type 13, on calcule l'aire à droite de la valeur 79,35. Cette aire est 0,4602.



# Solution de b (1/2)

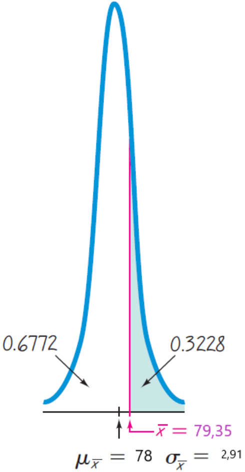
Nous utilisons le théorème de la limite centrale, car nous sommes intéressés à la moyenne d'un échantillon de 20 hommes, et non d'un homme individuel.

La taille de l'échantillon est inférieure à 30 mais la population originale a une distribution normale, donc nous pouvons utiliser le théorème de la limite centrale.

Comme la distribution originale est normale de moyenne 78 et écart type 13, alors la distribution des moyennes des échantillons de taille 20 peut être approximée par une distribution normale de moyenne 78 et écart type $13/\sqrt{20} = 2{,}91$.

# Solution de b (2/2)

Dans le graphe représentant la distribution normale de moyenne 78 et écart type 2,91, on calcule l'aire à droite de la valeur 79,35. Cette aire est 0,3228.



# Interprétation des résultats

Il y a une probabilité de 0,4602 qu'un homme pèse plus de 79,35 kg et une probabilité de 0,3228 que 20 hommes aient un poids moyen de plus de 79,35 kg, et donc un poids total supérieur à 1587 kg, le poids maximum des passagers permis.

Comme il y a une chance non négligeable que le bateau soit surchargé avec 20 hommes, la capacité maximale doit être réduite.

# Exercice

Le rendement des actions d'un portefeuille d'investissement est distribué normalement avec une moyenne de 20 et un écart type de 25.

a. Trouver la probabilité que si une action est choisie au hasard, son rendement soit inférieur à 0.

b. Trouver la probabilité que 10 actions choisies au hasard aient un rendement moyen inférieur à 0, en assumant que ces rendements soient indépendants.

# Correction pour une population finie

Lorsque l'échantillonnage est effectué sans remplacement et que la taille de l'échantillon $n$ est supérieure à 5% de la population finie de taille $N$ (c'est-à-dire $n > 0,05N$), il faut ajuster l'écart-type des moyennes de l'échantillon en le multipliant par le facteur de correction de la population finie $\sqrt{\frac{N-n}{N-1}}$.

Donc dans ce cas l'écart type des moyennes des échantillons de taille $n$ est

$$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \sigma\sqrt{\frac{N-n}{(N-1)n}}.$$

# 3. La loi normale comme approximation de la loi binomiale

# Concept clé

Cette section présente une méthode permettant d'utiliser une distribution normale comme approximation de la distribution de probabilité binomiale.

Si les conditions de $np \geq 5$ et $n(1-p) \geq 5$ sont toutes deux satisfaites, alors les probabilités d'une distribution de probabilité binomiale peuvent être bien approximées en utilisant une distribution normale avec moyenne $np$ et écart-type $\sqrt{np(1-p)}$.

# Variable aléatoire binomiale

Une **variable aléatoire binomiale** résulte d'une procédure qui répond à toutes les exigences suivantes :

1. La procédure a un nombre fixe d'essais.
2. Les essais doivent être indépendants. (Le résultat d'un essai individuel n'affecte pas les probabilités des autres essais).
3. Tous les résultats de chaque essai doivent être classés en deux catégories (communément appelées *succès* et *échec*).
4. La probabilité d'un succès reste la même dans tous les essais.

# Probabilités, moyenne, variance, écart type dans une distribution binomiale

Soient :
$n$ = nombre d'essais ;
$x$ = nombre de réussites parmi $n$ essais ;
$p$ = probabilité de succès dans un essai quelconque ;
$q$ = probabilité d'échec dans un essai quelconque ($q = 1 - p$).

Alors

$$P(x) = \frac{n!}{(n-x)!x!}\, p^x q^{n-x},$$

**moyenne :** $np$ ;
**variance :** $npq$ ;
**écart-type :** $\sqrt{npq}$.

# Approximation d'une distribution binomiale avec une distribution normale

Soit $Y$ une variable aléatoire binomiale de paramètres $n$ et $p$ avec $np \geq 5$ et $n(1-p) \geq 5$. Alors $Y$ peut être approximée avec une distribution normale de moyenne $np$ et écart-type $\sqrt{np(1-p)}$.

**Remarque.** Soit $X$ une variable aléatoire normale. Alors $X$ est continue et donc pour tout nombre $k$ on a que $P(X = k) = 0$. Donc si $Y$ est une variable aléatoire binomiale qui prend des valeurs entières et qui est approximée par $X$ alors

$$P(Y = a) \approx P(a - 0,5 < X < a + 0,5).$$

De la même manière,

$$P(Y < a) \approx P(X < a + 0,5)$$

et

$$P(Y > a) \approx P(X > a + 0,5).$$

# Exemple

Une compagnie de croisières maritimes envoie par courrier une enquête à la fin de laquelle on demande de communiquer son adresse électronique.

Supposons que l'enquête ait été envoyée à 40.000 personnes et que, pour ce type d'enquête, le pourcentage de réponses avec une adresse électronique est de 3%.

Si le véritable objectif de l'enquête était d'acquérir au moins 1150 adresses électroniques, trouvez la probabilité d'obtenir au moins 1150 réponses avec des adresses électroniques.

# Exemple

On a une distribution binomiale avec un nombre fixe d'essais ($n = 40.000$), qui sont indépendants. Il y a deux catégories pour chaque sondage : on obtient une réponse avec une adresse électronique ou on ne l'obtient pas. La probabilité de succès ($p = 0{,}03$) reste vraisemblablement constante d'un essai à l'autre.

Les calculs avec la formule de probabilité binomiale ne sont pas pratiques, car il faudrait l'appliquer 38851 fois, une fois pour chaque valeur de x de 1150 à 40.000 inclus,

$$P(\geq 1150) = \sum_{x=1150}^{40.000} \frac{40.000!}{(40.000-x)!x!} 0,03^x \, 0,97^{40.000-x}.$$

En approximant la distribution binomiale par une distribution normale on a des calculs beaucoup plus simples et moins longs.

# Exemple

Vérifions que les conditions pour approximer la distribution binomiale par la distribution normale sont vérifiées :

$$np = 40.000 \cdot 0{,}03 = 1200,$$

$$n(1-p) = 40.000 \cdot 0{,}97 = 38.800,$$

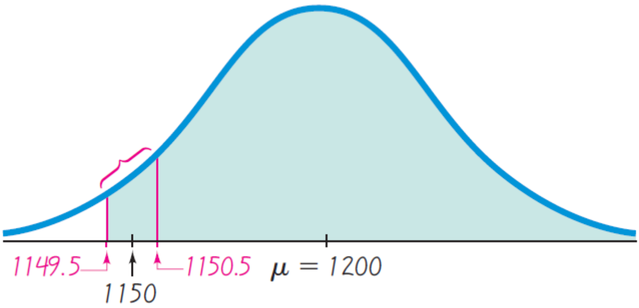
donc $np \geq 5$, $n(1-p) \geq 5$.

Donc on peut approximer la distribution binomiale ci-dessus par la distribution normale de moyenne $np = 1200$ et écart type

$$\sqrt{np(1-p)} = \sqrt{40.000 \cdot 0{,}03 \cdot 0{,}97} \approx 34{,}12.$$

# Exemple

Pour trouver la probabilité recherchée il faut choisir sur l'axes des abscisses la valeur 1149,5 et trouver l'aire correspondante à droite. Cette aire est égale à 0,9306, donc on a 93,06% de probabilité d'obtenir au moins 1150 réponses avec des adresses électroniques.



# Exercice

Supposons de lancer un dé à six faces 180 fois. Quelle est la probabilité d'obtenir le nombre 1 entre 25 et 40 fois ? (utiliser l'approximation d'une binomiale par une normale en justifiant cette approximation).

Fin.