Du discret au continu : Loi binomiale et loi normale

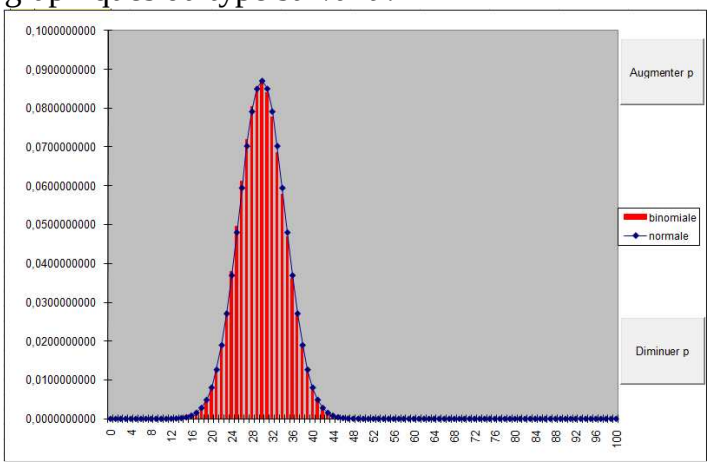
Ce que dit le programme :

«Pour introduire la loi normale $\mathcal{N}(0,1)$, on s'appuie sur l'observation des représentations graphiques de la loi de la variable aléatoire $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$ où X_n suit la loi binomiale $\mathcal{B}(n, p)$, et cela pour de grandes valeurs de n et une valeur de p fixée entre 0 et 1. Le théorème de Moivre Laplace assure que pour tous réels a et b, $P(Z_n \in [a,b])$ tend vers $\int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ lorsque n tend vers $+\infty$.»

Notre parti pris pour cet atelier

Nous avons choisi de nous placer du point de vue du professeur qui doit mettre en œuvre avec ses élèves l'illustration proposée par cet alinéa. L'outil tableur est couramment utilisé et la représentation d'une loi binomiale a été rencontrée en première.

Actuellement dans les classes de BTS on enseigne les approximations de la loi binomiale par la loi normale de même espérance et de même écart type. L'illustration de cette approximation se fait bien sur tableur, on trouve des graphiques du type suivant :



On représente ici la loi binomiale par un diagramme en bâtons et la courbe de la densité de la loi normale correspondante à partir des valeurs calculées sur chaque valeur prise par la loi binomiale.

Aussi il semble naturel de représenter avec le même outil les binomiales centrées réduites afin de les comparer à la loi normale centrée réduite.

La loi binomiale centrée réduite

Pour un réel p dans l'intervalle]0,1[et un entier naturel n, on note X_n la variable aléatoire qui suit la loi binomiale $\mathcal{B}(n, p)$.

X_n prend ses valeurs dans $\{0, 1, \ldots, n\}$ et $E(X_n) = np$, $V(X_n) = np(1-p)$, $\sigma(X_n) = \sqrt{np(1-p)} = \sigma$.

La variable centrée réduite associée à X_n est donc $Z_n = \frac{X_n - E(X_n)}{\sigma(X_n)} = \frac{X_n - np}{\sqrt{np(1-p)}}$.

La variable aléatoire Z_n est une variable discrète d'espérance 0 dont la variance et l'écart type sont égaux à 1.

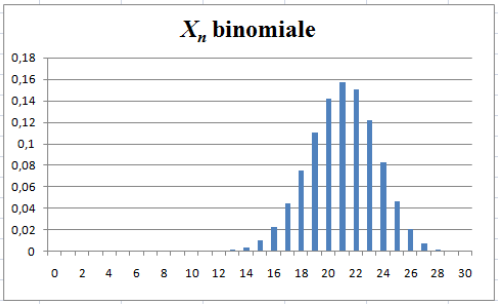
Les valeurs prises par Z_n sont les nombres réels $z_k = \frac{k - np}{\sqrt{np(1-p)}}$ avec $0 \le k \le n$.

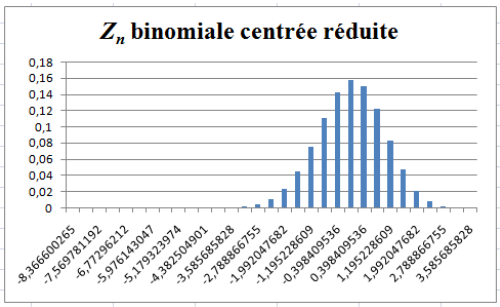
Ces valeurs ne sont pas entières mais réparties régulièrement sur l'intervalle $\left[\frac{-np}{\sqrt{np(1-p)}}, \frac{n(1-p)}{\sqrt{np(1-p)}}\right]$,

l'écart entre deux valeurs consécutives est $\frac{1}{\sqrt{np(1-p)}} = \frac{1}{\sigma}$.

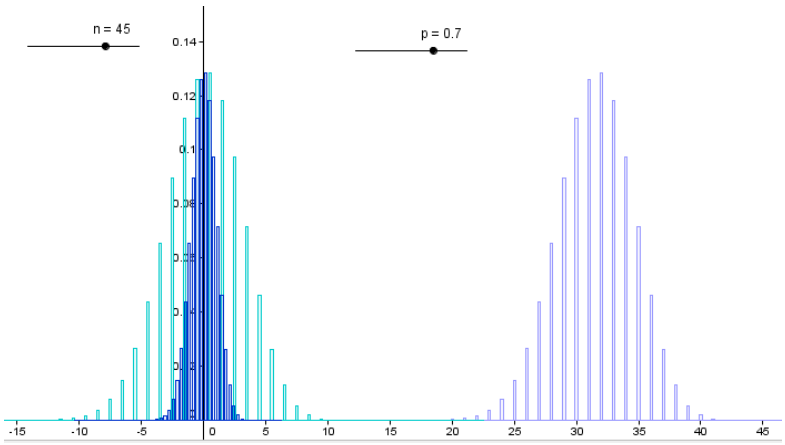
La représentation graphique de Z_n est donc un diagramme en bâtons.

Voici les représentations obtenues sur tableur avec $n = 30$ et $p = 0{,}7$





On observe un décalage horizontal des bâtons mais leur hauteur n'est pas modifiée, en effet, pour toute valeur de k on a $P(Z_n = z_k) = P(X_n = k)$.



Remarque : L'écart type de la série, lui, a été modifié donc les bâtons devraient être plus resserrés autour de l'espérance pour la variable Z_n que pour X_n . Mais le tableur ne permet pas de constater cela puisque les bâtons sont placés automatiquement et les données en abscisses ne sont que des étiquettes.
Il faut utiliser un autre logiciel, par exemple, Géogébra, pour contrôler les valeurs positionnées en abscisses et observer cette modification.
Les commandes utilisées sont les suivantes :
Diagramme en bâtons binomiale :

Barres[Séquence[k, k, 0, n], Séquence[Combinaison[n, k] p^k (1 − p)^(n − k), k, 0, n], 0.2]

Diagramme en bâtons binomiale centrée :

Barres[Séquence[k, k, − n p, n − n p], Séquence[Combinaison[n, k] p^k (1 − p)^(n − k), k, 0, n], 0.2]

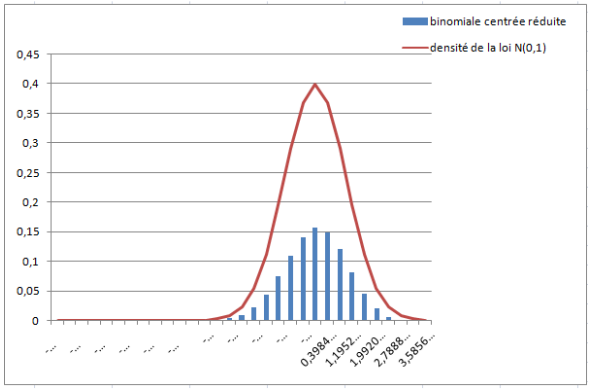
Diagramme en bâtons binomiale centrée réduite:

Barres[Séquence[(k − n p) / σ, k, 0, n], Séquence[Combinaison[n, k] p^k (1 − p)^(n − k), k, 0, n], 0.2]

Comparaison entre une loi binomiale centrée réduite et la loi normale centrée réduite

Il s'agit d'illustrer graphiquement que, pour n assez grand, $P(Z_n \in [a,b])$ est assez proche de l'aire sous la courbe de la fonction de densité f de la loi normale entre les valeurs a et b.

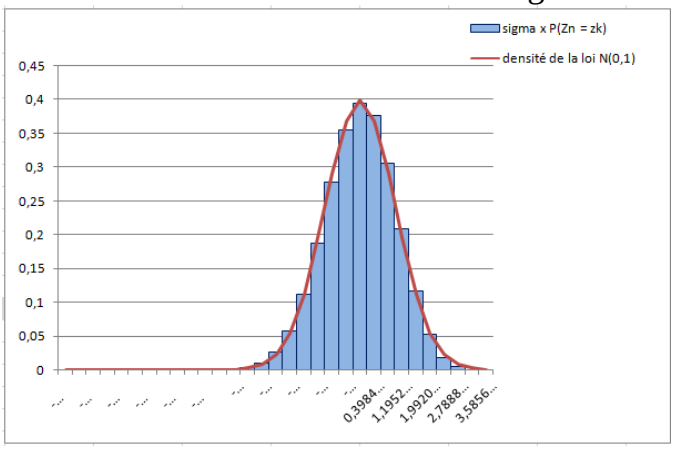
Voici ce que l'on obtient si on procède comme pour le graphique précédent, avec les valeurs de f calculées pour les z_k avec $0 \leq k \leq n$.



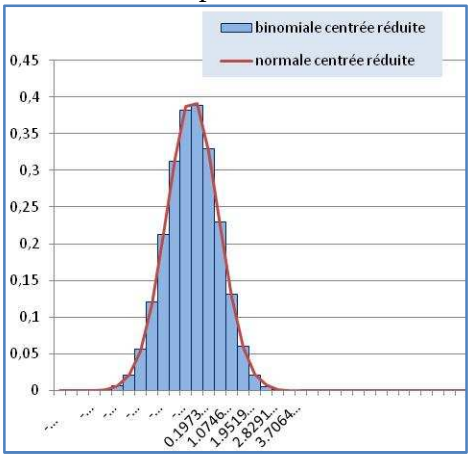
Cette façon de procéder ne convient pas, à l'évidence. En effet, contrairement à la situation d'approximation ci-dessus, ici, l'écart entre deux valeurs consécutives de Z_n n'est indépendant de n ni de p.

Pour que la comparaison des représentations soit possible, il faut se mettre en situation de comparer deux variables continues.
Considérerons alors la variable Z_n, qui pend un grand nombre de valeurs, avec un point de vue statistique. Il s'agit en quelque sorte de convertir son diagramme en bâton en histogramme d'une variable continue qui prendrait sur chaque intervalle $[z_k , z_{k+1}[$ une valeur constante égale à $P(Z_n = z_k)$ avec $0 \leq k \leq n - 1$.
On peut alors considérer Z_n comme une variable aléatoire continue dont la densité est une fonction en escalier qui sur chaque intervalle $[z_k , z_{k+1}[$ a pour valeur la hauteur du rectangle construit sur la "classe" $[z_k , z_{k+1}[$ de façon que l'aire de ce rectangle soit la probabilité $P(Z_n = z_k)$. L'amplitude de la "classe" $[z_k , z_{k+1}]$ est $\frac{1}{\sigma}$ donc : **La densité de Z_n sur $[z_k , z_{k+1}[$ est $\sigma \times P(Z_n = z_k)$ avec $0 \leq k \leq n - 1$.**
Sur le tableur, avec cette hauteur et une mise en forme des rectangles on obtient :

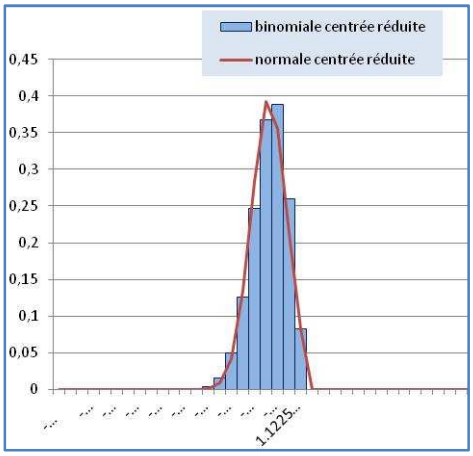
On peut constater que l'allure de l'histogramme se rapproche à chaque fois de la courbe de f . Cependant si p est proche de 0 ou de 1 il faut de grandes valeurs de n pour observer une coïncidence significative.



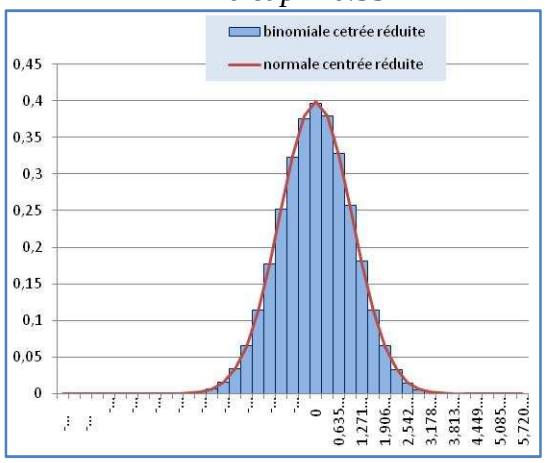
$n = 21$ et $p = 0.55$



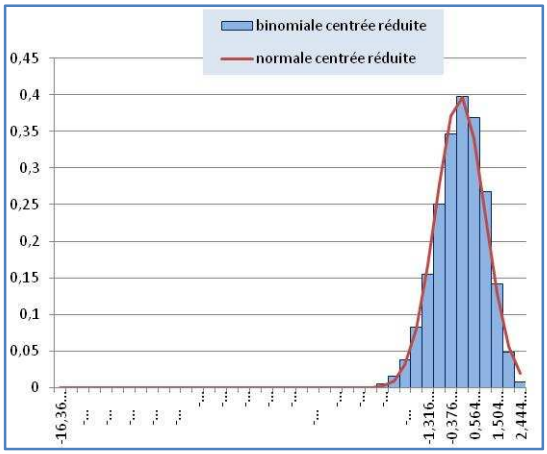
$n = 21$ et $p = 0.87$



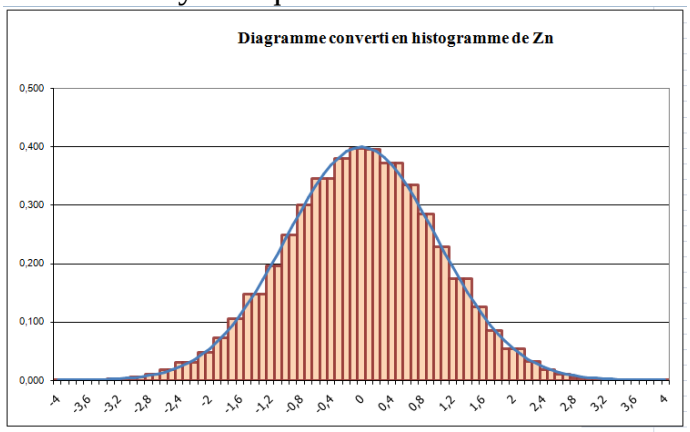
$n = 40$ et $p = 0.55$

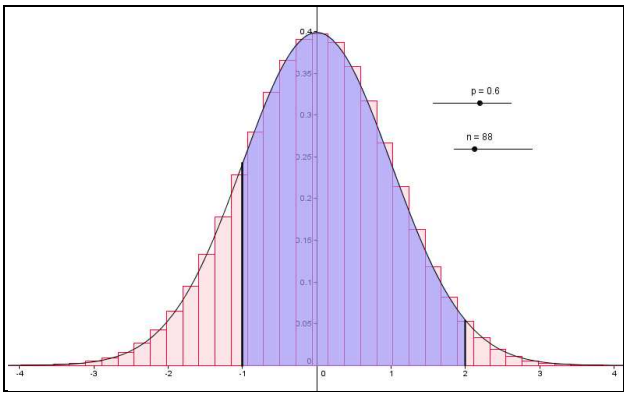


$n = 40$ et $p = 0.87$



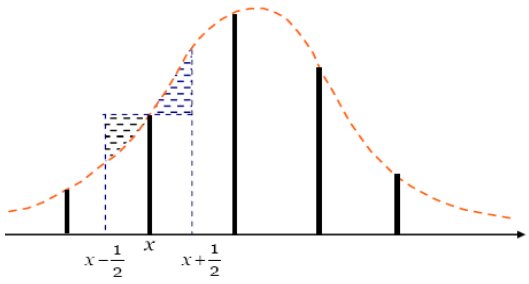
Pour aborder des illustrations correspondant à de plus grandes valeurs de n, on est amené à ne conserver que les valeurs de Z_n qui ont une probabilité significative. On choisit alors de travailler sur l'intervalle I = [–4 ; 4] et on détermine les valeurs entières k pour lesquelles Z_n prend ses valeurs dans I.
Voici alors les graphiques obtenus :
Illustration dynamique Excel et Géogébra





Ligne de commande Géogébra : (*s* désigne ici l'écart type de la loi binomiale)
Histogramme[Séquence[(k – n p) / s – 0.5 / s, k, 0, n + 1], Séquence[s × Combinaison[n, k] p^k (1 – p)^(n – k), k, 0, n]]
(on définit les classes de telle sorte que leurs centres soient les valeurs de Z_n)

Correction de continuité
Si X suit la loi B (n, p), X prend des valeurs entières entre 0 et n et la loi de X est proche de celle de la loi normale de même espérance et de même écart type.
Or pour une loi normale, la probabilité d'une valeur isolée est nulle. Il semble donc impossible de calculer P($X = k$) avec cette approximation.
Approcher la loi binomiale par la loi normale c'est remplacer une loi discrète (celle de X) par une loi continue (celle de X_c).
On remplace donc la probabilité de la valeur isolée x de la variable X par celle d'un intervalle d'amplitude 1 autour de x pour la variable X_c.
$P(X = x) \approx P(x - 1/2 < X_c < x + 1/2)$
Cette opération s'appelle la correction de continuité.



La variable discrète X étant approchée par la variable continue X_c , on utilise les règles suivantes d'approximation :

$P(X < n)$ s'obtient avec $P(X_c < n - 0,5)$ $P(X \leq n)$ s'obtient avec $P(X_c < n + 0,5)$
$P(X > n)$ s'obtient avec $P(X_c > n + 0,5)$ $P(X \geq n)$ s'obtient avec $P(X_c > n - 0,5)$

On calcule par exemple $P(a < X \leq b)$ avec $P(a + 0,5 < X_c < b + 0,5) = F(b + 0,5) - F(a + 0,5)$ où F désigne la fonction de répartition de la variable continue X_c .

Exemple
Soit X une variable aléatoire suivant la loi $\mathcal{B}(50, 1/2)$. Les conditions d'approximations de la loi de X par une loi normale sont remplies, et l'on peut considérer que X suit à peu près la loi $\mathcal{N}(25, 25/2)$.
Évaluons alors de deux façons P($24 \leq X \leq 26$)
En valeur exacte avec la loi binomiale : $P(X = 24) + P(X = 25) + P(X = 26) \approx 0,3282$
En valeur approchée avec la loi normale : $P(24 \leq X \leq 26) \approx 0,2222$
En valeur approchée avec la loi normale corrigée par continuité : $P(23,5 \leq X \leq 26,5) \approx 0,3286$.

Le résultat est bien meilleur en tenant compte de la correction par continuité.