LINGE1113 - TP6 - Fonctions de variables aléatoires, distributions d'échantillonnage et théorème central limite - Solutions

Soit X une variable aléatoire dont on connait soit la fonction de répartition, soit la fonction de densité, soit la fonction génératice de moment (FGM). On va utiliser trois méthodes pour obtenir la distribution d'une fonction de cette variable aléatoire, notée $U = h(X)$: la méthode de fonction de répartition, la méthode de transformation et méthode de FGM. Ces méthodes sont résumées à la page 6 du formulaire. La première méthode peut être généralisée lorsque h est une fonction monotone admettant une fonction réciproque notée h^{-1} et la fonction de répartition $F_X(x) = P(X \leq x)$ de la variable X est connue. Dans ce cas, la fonction de répartition $F_U(u)$ de U est obtenue comme suit :

(i) Si h est une fonction **croissante**,

$$F_U(u) = P(U \leq u) = P(h(X) \leq u) = P(X \leq h^{-1}(u)) = F_X(h^{-1}(u)).$$

(i) Si h est une fonction **décroissante**,

$$F_U(u) = P(U \leq u) = P(h(X) \leq u) = P(X \geq h^{-1}(u)) = 1 - F_X(h^{-1}(u)).$$

Exercice 6.4

Soit Y la variable aléatoire représentant la quantité de farine utilisée quotidiennement dans une boulangerie. On sait que $Y \sim \text{Exp}(\beta = 4)$. La fonction de densité de Y est donc donnée par $f(y) = \frac{1}{4}e^{-y/4}$ si $y > 0$ et $f(y) = 0$ sinon.

On veut trouver la fonction de densité de $U = 3Y + 1$. Dans cet exercice, nous utilisons la méthode de la fonction de répartition. Toutefois, nous obtiendrons le même résultat en utilisant la méthode de transformation à l'exercice 6.24.
Dans un premier temps nous trouvons La fonction de répartition de U comme suit :

$$F_U(u) = P(U \leq u) = P(3Y + 1 \leq u) = P\left(Y \leq \frac{u-1}{3}\right)$$

Dès lors, si $\frac{u-1}{3} > 0$ càd $u > 1$, on a

$$\begin{aligned} F_U(u) = P(U \leq u) &= P\left(Y \leq \frac{u-1}{3}\right) \\ &= \int_0^{\frac{u-1}{3}} \frac{1}{4} e^{-y/4}\, dy = \Big[-e^{-y/4} \Big]_0^{\frac{u-1}{3}} = 1 - e^{-(u-1)/12} \end{aligned}$$

Pour $u \leq 1$ càd $\frac{u-1}{3} \leq 0$, on a $F_U(u) = P\left(Y \leq \frac{u-1}{3}\right) = 0$.
Finalement, il nous reste à dériver la fonction de répartition de U pour obtenir la fonction de densité de U :

$$f_U(u) = \frac{d}{du} F_U(u) = \begin{cases} \frac{1}{12} e^{-(u-1)/12} & \text{pour } u > 1 \\ 0 & \text{pour } u \leq 1 \end{cases}$$

Exercice 6.24

Le contexte de cet exercice est le même que celui de l'exercice 6.4. On a $Y \sim \text{Exp}(\beta = 4)$ et on cherche la fonction de densité de $U = 3Y + 1$. On va utiliser ici la méthode de transformation avec $h(y) = 3y + 1$ qui est bien une transformation monotone et qui justifie donc l'utilisation de cette méthode. Donc $h^{-1}(u) = \frac{u-1}{3}$ et la fonction de densité de U est donnée par

$$\begin{aligned} f_U(u) &= \boldsymbol{f_Y(h^{-1}(u)) \cdot \left|\frac{d}{du} h^{-1}(u)\right|} \\ &= f_Y\left(\frac{u-1}{3}\right) \cdot \left|\frac{d}{du}\left(\frac{u-1}{3}\right)\right| \\ &= \frac{1}{4} e^{-(u-1)/(3 \cdot 4)} \cdot \frac{1}{3} \quad \text{si } \frac{u-1}{3} > 0 \\ &= \frac{1}{12} e^{-(u-1)/12} \quad \text{si } u > 1 \end{aligned}$$

Et naturellement, on a $f_U(u) = 0$ si $u \leq 1$. On retrouve donc bien le même résultat final qu'à l'exercice 6.4

Exercice 6.34

(a) On va utiliser ici la méthode de transformation avec $h(y) = y^2$ qui est bien une transformation monotone (fonction croissante) car le domaine de définition de la fonction de densité se restreint aux valeurs positives de y. On a $h^{-1}(u) = \sqrt{u}$ et la fonction de densité de U est donnée par

$$\begin{aligned} f_U(u) &= \boldsymbol{f_Y(h^{-1}(u)) \cdot \left|\frac{d}{du} h^{-1}(u)\right|} \\ &= f_Y(\sqrt{u}) \cdot \left|\frac{d}{du}\sqrt{u}\right| \\ &= \frac{2\sqrt{u}}{\theta} e^{-(\sqrt{u})^2/\theta} \cdot \frac{1}{2\sqrt{u}} \quad \text{si } u > 0 \\ &= \frac{1}{\theta} e^{-u/\theta} \quad \text{si } u > 0 \end{aligned}$$

Il s'agit de la fonction de densité d'une loi exponentielle de paramétre θ. Donc $U \sim \text{Exp}(\theta)$.

(b) On veut calculer $E(Y)$ et $V(Y) = E(Y^2) - (E(Y))^2$.
En utilisant le point (a), puisque $U \sim Exp(\theta)$ on sait que $E(U) = \theta$, mais $E(U) = E(Y^2)$ donc $E(Y^2) = \theta$. Ensuite

$$E(Y) = \int_0^{+\infty} y f(y)\, dy = \int_0^{+\infty} \frac{2y^2}{\theta} e^{-y^2/\theta}\, dy$$

On réalise ensuite une intégration par parties avec $g_1(y) = y$, $g_2'(y) = \frac{2y}{\theta} e^{-y^2/\theta}$ et $g_1'(y) = 1$, $g_2(y) = -e^{-y^2/\theta}$. On obtient

$$E(Y) = \int_0^{+\infty} g_1(y)\, g_2'(y)\, dy = \Big[-ye^{-y^2/\theta} \Big]_0^{+\infty} + \int_0^{+\infty} e^{-y^2/\theta}\, dy = 0 + \int_0^{+\infty} e^{-y^2/\theta}\, dy$$

Or, on sait que l'intégrale de toute fonction de densité sur son domaine vaut 1. En particulier, si $X \sim N(0, \frac{\theta}{2})$, la fonction de densité de X vaut $f_X(x) = \frac{1}{\sqrt{\pi\theta}} e^{-x^2/\theta}$ pour $-\infty < x < +\infty$, et donc

$$\int_{-\infty}^{+\infty} f_X(x)\, dx = 1 \Longleftrightarrow \int_{-\infty}^{+\infty} \frac{1}{\sqrt{\pi\theta}} e^{-x^2/\theta}\, dx = 1$$

Or, la fonction de densité de la loi normale est symétrique par rapport à sa moyenne, càd par rapport à 0 ici, ce qui implique que $\int_0^{+\infty} \frac{1}{\sqrt{\pi\theta}} e^{-x^2/\theta}\, dx = \frac{1}{2}$. On conclut donc finalement que

$$E(Y) = \int_0^{+\infty} e^{-y^2/\theta}\, dy = \int_0^{+\infty} \frac{\sqrt{\pi\theta}}{\sqrt{\pi\theta}} e^{-y^2/\theta}\, dy = \sqrt{\pi\theta} \int_0^{+\infty} \frac{1}{\sqrt{\pi\theta}} e^{-y^2/\theta}\, dy = \frac{\sqrt{\pi\theta}}{2}$$

Aussi, $V(Y) = \theta - (\frac{\sqrt{\pi\theta}}{2})^2 = \theta(1 - \frac{\pi}{4})$.

Exercice 6.49

On a $Y_1 \sim \text{Bin}(n_1, p)$ et $Y_2 \sim \text{Bin}(n_2, p)$. Par l'exercice 3.145 ou encore le formulaire à la page 3, les fonctions génératrices des moments de Y_1 et Y_2 sont donc respectivement données par

$$m_{Y_1}(t) = (pe^t + 1 - p)^{n_1}$$

et

$$m_{Y_2}(t) = (pe^t + 1 - p)^{n_2}$$

Or les variables Y_1 et Y_2 sont indépendantes, la fonction génératrice des moments de $Y_1 + Y_2$ vaut donc le produit des fonctions génératrices des moments de Y_1 et Y_2 (voir la page 6 du formulaire), càd

$$m_{Y_1+Y_2}(t) = m_{Y_1}(t) \cdot m_{Y_2}(t) = (pe^t + 1 - p)^{n_1} \cdot (pe^t + 1 - p)^{n_2} = (pe^t + 1 - p)^{n_1+n_2}$$

On remarque que la fonction génératrice des moments de $Y_1 + Y_2$ est la fonction génératrice des moments d'une loi $\text{Bin}(n_1 + n_2, p)$. Comme la fonction génératrice des moments identifie la distribution, nécessairement $Y_1 + Y_2 \sim \text{Bin}(n_1 + n_2, p)$. Nous venons donc d'utiliser **la méthode de FGM** pour trouver la loi d'une fonction de variables aléatoires.

Démonstration Soient X, Y deux variables aléatoires indépendantes de FGM $m_X(t) = E(e^{tX})$ et $m_Y(t) = E(e^{tY})$. Alors la FGM de la variable aléatoire $U = X + Y$ est donnée par

$$m_U(t) = E(e^{tU}) = E(e^{tX+tY}) = E(e^{tX} e^{tY}).$$

La première égalité est donnée par définition de la FGM. La seconde égalité est obtenue en remplaçant U par $X + Y$. La dernière égalité est obtenue en utilisant la propriété mathématique $e^{a+b} = e^a e^b$. Maintenant on va utiliser l'indépendance de X et Y : comme X et Y sont indépendantes, alors les variables aléatoires $R = e^{tX}$ et $Q = e^{tY}$ sont indépendantes aussi et $E(RQ) = E(Q)E(R)$. Finalement, on obtient

$$E(e^{tX} e^{tY}) = E(e^{tX})E(e^{tY}) = m_X(t) m_Y(t).$$

Cette démonstration peut être généralisée pour plus de deux variables aléatoires indépendantes.

Exercice 7.11

Soit la variable aléatoire Y représentant la surface (en pouces carrés - pc) occupée par un pin. On sait que $Y \sim N(\mu, \sigma^2 = 4^2)$. Nous avons donc la valeur de la variance σ^2, mais la moyenne μ est inconnue. Notons par $Y_1, \ldots, Y_9$ les surfaces occupées par chacun des $n = 9$ pins que l'on considère dans notre échantillon. On suppose que $Y_1, \ldots, Y_9$ sont iid. Soit également $\overline{Y} = \frac{1}{9}\sum_{i=1}^{9} Y_i$ la moyenne des aires occupées par les 9 pins. Comme Y est normalement distribuée et les variables Y_i sont indépendantes , on aura aussi $\overline{Y} \sim N(\mu, \frac{\sigma^2}{n})$ - voir à la page 7 du formulaire pour ce résultat.

On veut calculer la probabilité que la moyenne de l'échantillon $\bar{Y}$ soit à l'intérieur d'un intervalle de demi-longueur 2pc autour de la vraie moyenne μ :

$$P(\mu - 2 \leq \bar{Y} \leq \mu + 2) = P(-2 \leq \overline{Y} - \mu \leq 2).$$

Pour calculer cette probabilité, il faut réduire $\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$ pour se ramener à une distribution normale centrée réduite $Z \sim N(0, 1)$. Nous allons donc diviser par l'écart type $\sqrt{\frac{\sigma^2}{n}}$ de $\bar{Y}$:

$$\begin{aligned} P\left(\frac{-2}{\sigma/\sqrt{n}} \leq \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \leq \frac{2}{\sigma/\sqrt{n}}\right) &= P(-1.5 \leq Z \leq 1.5) \\ &= P(Z \leq 1.5) - P(Z \leq -1.5). \end{aligned}$$

Notez que pour trouver ces probabilités en utilisant la table de la loi normale, il faut se ramener à des probabilités de la forme $P(Z \geq a)$ avec a positif. Par ailleurs, on sait que :

$$P(Z \leq a) = 1 - P(Z > a) \qquad\qquad P(Z \leq a) = P(Z \geq -a).$$

La deuxième égalité vient de la symmetrie de la loi normale. En utilisant ces relations on a :

$$\begin{aligned} P(Z \leq 1.5) - P(Z \leq -1.5) &= 1 - P(Z > 1.5) - P(Z \geq -(-1.5)) \\ &= 1 - 2P(Z \geq 1.5) = 0.8664 \end{aligned}$$

NB : Pour une variable aléatoire continue X les inégalités strictes $<, >$ ou non strictes $\geq, \leq$ n'ont pas d'importance pour le calcul de probabilités : $P(X > a) = P(X \geq a)$ ou $P(X < a) = P(X \leq a)$.

Exercice 7.43

Soit la variable aléatoire X représentant la taille d'un homme. Soit $E(X) = \mu$ et $V(X) = \sigma^2 = 2.5^2$. Notez qu'ici on ne connait pas la distribution de la variable X. Le fait que les deux paramètres $E(X) = \mu$ et $Var(X) = \sigma^2$ soient mentionnés ne signifie pas que la distribution de X est Normale.

Notons par $X_1, \ldots, X_{100}$ les tailles de chacun des $n = 100$ hommes que l'on considère dans notre échantillon. On suppose que $X_1, \ldots, X_{100}$ sont iid. Soit également $\overline{X} = \frac{1}{100}\sum_{i=1}^{100} X_i$ la moyenne des tailles de ces 100 hommes.

Comme à l'exercice précédent. On cherche la probabilité que l'écart entre la moyenne de l'échantillon $\bar{X}$ et la vraie moyenne μ soit d'au plus 0.5 pouces : $P(|\bar{X} - \mu| \leq 0.5)$.
Puisque la taille de l'échantillon est grande on applique le TCL. D'après le TCL, on sait que la

moyenne de l'échantillon standardisée suit approximativement une distribution normale centrée réduite : $\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \approx N(0,1)$. Ainsi,

$$\begin{aligned}
P(|\overline{X}-\mu| < 0.5) &= P(-0.5 < \overline{X}-\mu < 0.5) \\
&= P\left(\frac{-0.5\sqrt{n}}{\sigma} < \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} < \frac{0.5\sqrt{n}}{\sigma}\right) \\
&\approx P(-2 < Z < 2) \quad \text{par le TCL} \\
&\approx 1 - 2P(Z > 2) \\
&\approx 1 - 2 \cdot 0.0228 \\
&\approx 0.9544,
\end{aligned}$$

où $Z \sim N(0,1)$.
Notons qu'ici on a obtenu un résultat approximatif alors qu'à l'exercice 7.11, le résultat obtenu était exact. En effet, l'hypothèse de normalité faite à l'exercice 7.11 nous permettait de connaître exactement la distribution de la moyenne empirique (la moyenne de l'échantillon, $\bar{X}$) alors qu'ici aucune hypothèse de distribution n'a été faite sur notre population, nous obligeant à utiliser le théorème central limite.

Exercice 7.44

Le contexte de cet exercice est le même que celui de l'exercice 7.43. Avec $\sigma = 2.5$, on cherche maintenant n tel que $P(|\overline{X}-\mu| < 0.4) = 0.95$.
D'après le TCL, on sait que la moyenne de l'échantillon standardisée suit approximativement une distribution normale centrée réduite : $\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \approx N(0,1)$. D'où,

$$\begin{aligned}
P(|\overline{X}-\mu| < 0.4) = 0.95 &\iff P(-0.4 < \overline{X}-\mu < 0.4) = 0.95 \\
&\iff P\left(\frac{-0.4\sqrt{n}}{\sigma} < \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} < \frac{0.4\sqrt{n}}{\sigma}\right) = 0.95 \\
&\iff P\left(\frac{-0.4\sqrt{n}}{\sigma} < Z < \frac{0.4\sqrt{n}}{\sigma}\right) \approx 0.95 \text{ (par TCL)} \\
&\iff 1 - 2P\left(Z > \frac{0.4\sqrt{n}}{\sigma}\right) \approx 0.95 \\
&\iff P\left(Z > \frac{0.4\sqrt{n}}{\sigma}\right) \approx 0.025 \\
&\iff P\left(Z > \frac{0.4\sqrt{n}}{\sigma}\right) \approx 0.025
\end{aligned}$$

où $Z \sim N(0,1)$.
Pour obtenir le nombre $z = 0.4\sqrt{n}/\sigma$ tel que $P(Z > z) = 0.025$ on regarde dans la table de loi normale centrée réduite et on trouve $z = 1.96$. Donc

$$\frac{0.4\sqrt{n}}{\sigma} = 1.96 \iff \sqrt{n} = \frac{1.96 \cdot 2.5}{0.4}$$

On en déduit que $n = 150.0625$ et on prend donc $n = 151$.

Exercice 7.79

Soit Y le nombre de productions défectueuses. Y est une variable aléatoire qui compte le nombre de succés après un certain nombre de tentatives. En effet, le nombre de tentatives est le nombre de productions que l'on teste qui vaut 25. Le succès est le fait qu'une production soit défectueuse. Donc $Y \sim \text{Bin}(n = 25, p = 0.1)$ où p est la probabilité de succès d'une tentative, càd la probabilité qu'une production soit défectueuse.

On veut calculer $P(Y \geq 2)$. On va approximer cette probabilité en utilisant une loi normale au point (a) et on calculera la valeur exacte de cette probabilité en utilisant la distribution binomiale au point (b).

(a) Par le théorème central limite, $Y \sim \text{Bin}(n, p)$ a approximativement la même distribution que $X \sim N(np, np(1 - p))$ - voir à la page 7 du formulaire. En centrant et en réduisant on a : $Z = \frac{X-np}{\sqrt{np(1-p)}} \sim N(0, 1)$. Ainsi, sans faire une correction de continuité on obtient

$$\begin{aligned} P(Y \geq 2) &\approx P(X \geq 2) \\ &\approx P\left(\frac{X - np}{\sqrt{np(1-p)}} \geq \frac{2 - 25 \cdot 0.1}{\sqrt{25 \cdot 0.1 \cdot 0.9}}\right) \\ &\approx P(Z \geq -0.33) \quad \text{par le TCL} \\ &\approx 1 - P(Z > 0.33) \\ &\approx 0.6293. \end{aligned}$$

Avec correction de continuité (expliquée à la fin de cet exercice), on obtient

$$\begin{aligned} P(Y \geq 2) &= P(Y \geq 1.5) \\ &\approx P(X \geq 1.5) \\ &\approx P\left(\frac{X - np}{\sqrt{np(1-p)}} \geq \frac{1.5 - 25 \cdot 0.1}{\sqrt{25 \cdot 0.1 \cdot 0.9}}\right) \\ &\approx P(Z \geq -0.67) \quad \text{par le TCL} \\ &\approx 1 - P(Z > 0.67) \\ &\approx 0.7486 \end{aligned}$$

On voit donc que l'approximation de la probabilité exacte obtenue au point (b) est nettement meilleure en réalisant une correction de continuité.

(b) On sait que $P(Y = y) = C_y^n p^y (1 - p)^{n-y}$. Donc,

$$\begin{aligned} P(Y \geq 2) &= 1 - P(Y = 0) - P(Y = 1) \\ &= 1 - C_0^{25} 0.1^0 (1 - 0.1)^{25-0} - C_1^{25} 0.1^1 (1 - 0.1)^{25-1} \\ &= 1 - 0.0718 - 0.1994 \\ &= 0.7288 \end{aligned}$$

Correction de continuité La correction de continuité consiste à faire un ajout ou une soustraction de 0.5 pour calculer des probabilités lorsqu'on approxime une variable aléatoire discrète par une variable aléatoire continue. Dans la table ci-après, nous appliquons la correction de continuité pour toutes les formes de probabilité sur une Binomiale $Y \sim Bin(n, p)$ approximée par une Normale $X \sim N(np, np(1 - p))$.

Loi discrète	Loi continue
$P(Y = a)$	$P(a - 0.5 \leq X \leq a + 0.5)$
$P(Y \leq a)$	$P(X \leq a + 0.5)$
$P(Y < a) = P(Y \leq a - 1)$	$P(X \leq a - 1 + 0.5) = P(X \leq a - 0.5)$
$P(Y \geq a)$	$P(X \geq a - 0.5)$
$P(Y > a) = P(Y \geq a + 1)$	$P(X \geq a + 1 - 0.5) = P(X \geq a + 0.5)$

Table 1 – Correction de continuité appliquée aux probabilités sur une variable continue qui approxime une variable discrète.

Le raisonnement derrière la correction de continuité peut être compris à l'aide des figures ci-après, où on approxime $Y \sim Bin(20, 0.5)$ par $X \sim N(20 \times 0.5, 20 \times 0.5 \times 0.5) = N(10, 5)$.

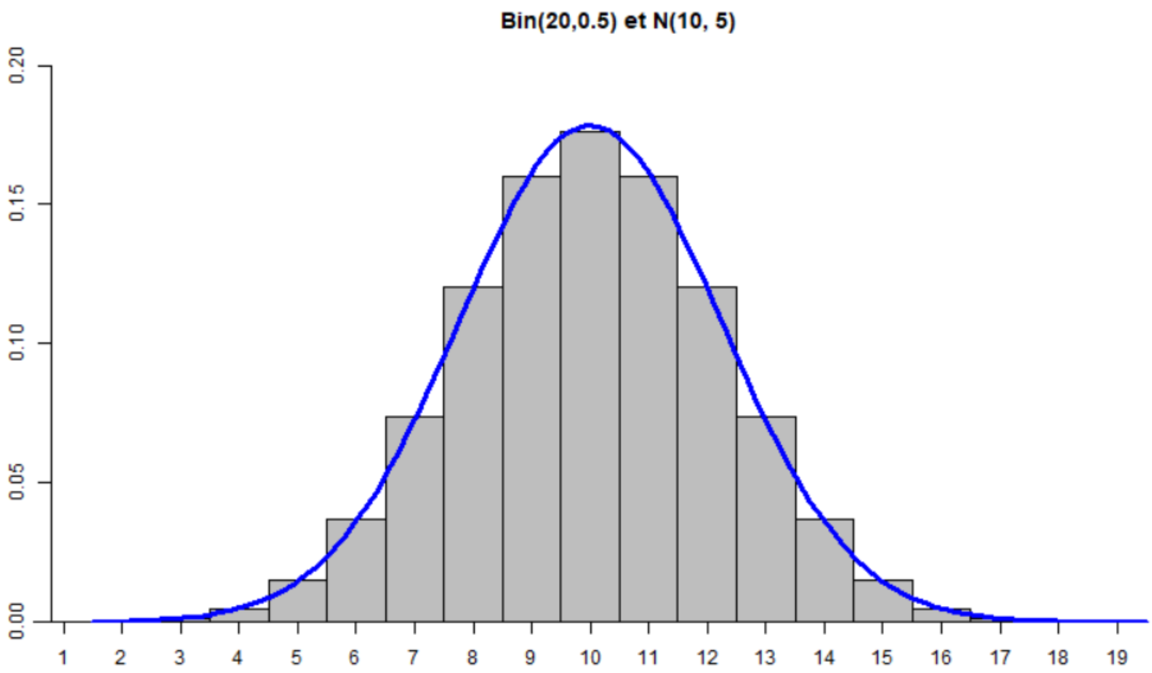

Figure 1 – Distribution Binomiale(20,0.5) et son approximation Normale(10,5). Notez que les barres sont centrées autour des nombres entiers.

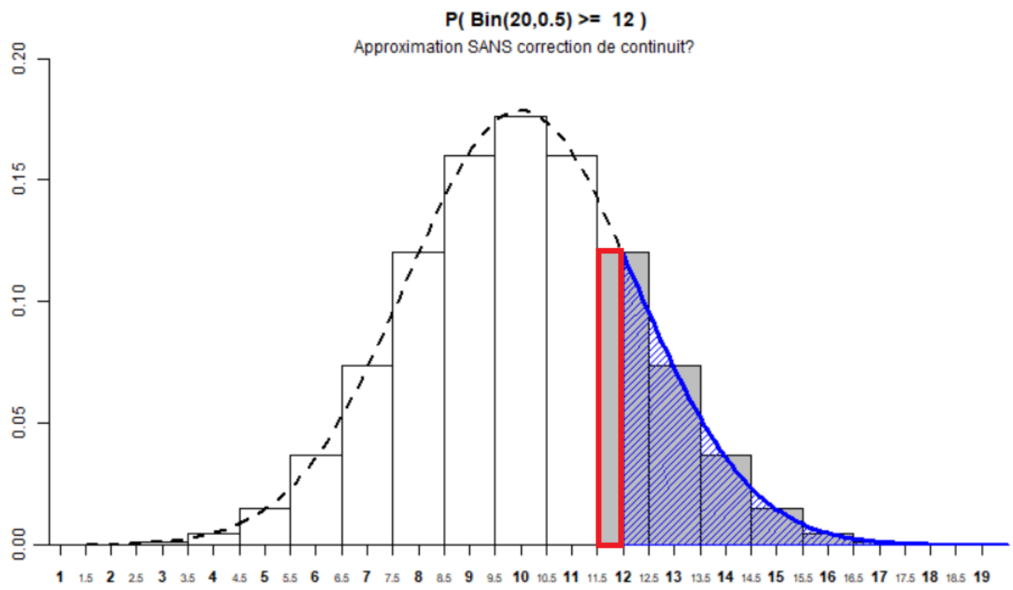

FIGURE 2 – La surface des barres en gris est la probabilité exacte $P(Y \geq 12)$. La surface en bleu est l'approximation de la probabilité $P(Y \geq 12)$ par $P(X \geq 12)$ sans correction de continuité. On remarque que la partie de la barre entourée en rouge est omise de l'approximation normale. Pour améliorer l'approximation de $P(Y \geq 12)$ il faut donc **soustraire** 0.5 de 12 comme suit : $P(X \geq 12 - 0.5) = P(X \geq 11.5)$.

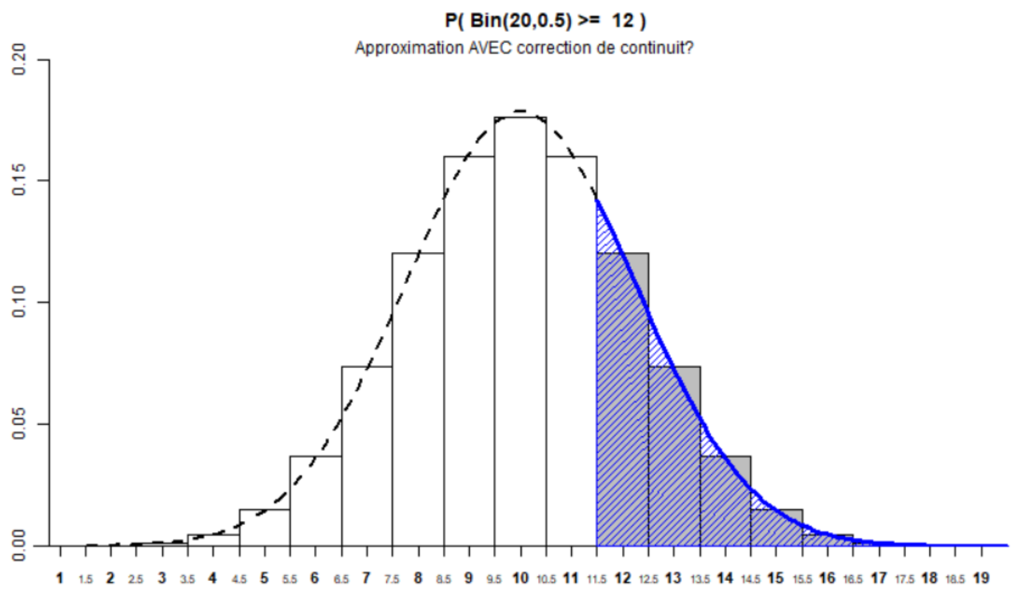

FIGURE 3 – La surface en bleu représente $P(X \geq 11.5)$. La surface en gris représente $P(Y \geq 12)$. On voit donc que l'approximation avec correction de continuité $P(X \geq 11.5)$ est meilleure que l'approximation sans la correction de continuité $P(X \geq 12)$ de la figure précédente.

En procédant de la même façon vous verrez pourquoi il faut ajouter 0.5 à 12 pour mieux approcher $P(Y \leq 12)$ lorsque la distribution de Y est approximée par celle d'une distribution normale.