Correctif de la séance 1
Générateurs de nombres aléatoires

1. Algorithme de génération d'une variable aléatoire uniforme :

a) Si on génère suffisamment de nombres, on converge assez vite vers une moyenne de 0,5 et un écart type de 0,288 7. La moyenne d'une uniforme entre 0 et 1 est évidemment 0,5. Pour l'écart type, il faut se souvenir que l'écart type d'une loi uniforme entre a et b est donné par

$$\sigma = \frac{b-a}{\sqrt{12}}.$$

En prenant $b = 1$ et $a = 0$, on trouve bien un écart type de 0,288 675, ce qui est cohérent avec la valeur obtenue.

b) Pour des petits nombres d'événements générés (10 000 par exemple), la distribution semble bien aléatoire. Mais dès qu'on augmente ce nombre (avec 10 000 000 par exemple), on commence à remarquer des "patterns" (lignes obliques), qui indiquent une corrélation. Ceci est dû au fait que la fonction que l'on emploie est périodique : à partir d'un certain moment, on retombe sur la même série de valeurs. On atteint donc les limites de cette méthode de génération.

c) Pour un histogramme de 100 bins avec 100 000 événements générés (par exemple), on attend en moyenne 1 000 événements par bin. La distribution de ce nombre d'événements dans un bin donné n'est évidemment pas décrit par une loi uniforme, mais par une loi de Poisson (c'est un modèle d'"arrivées"). Pour une loi de Poisson, la moyenne est égale à la variance et l'écart type est donc la racine carrée du nombre d'événements attendu dans le bin. Dans notre cas, on attend donc un écart type de $\sqrt{1\,000} = 31{,}62$. C'est approximativement la taille des barres d'erreur données par `ROOT`. Cela n'a rien d'étonnant : par défaut, `ROOT` représente comme erreur la racine du nombre d'événements dans un bin.

2. Génération d'une variable aléatoire uniforme à l'aide des fonctions prédéfinies de `ROOT` :

a) Les distributions générées avec la classe `TRandom3` ont l'air uniformes même avec des grands nombres d'événements générés, ce qui prouve que ce générateur est beaucoup plus fiable. La documentation de `ROOT` nous indique que la période de l'algorithme est $2^{19937} - 1 \sim 10^{19}$.

b) Pour générer une distribution aléatoire uniforme entre $-1\,000$ et $1\,000$, il suffit de multiplier le nombre aléatoire par le "range" final désiré : $1\,000 - (-1\,000) = 2\,000$, et de décaler le résultat de $x_{\min} = -1\,000$. La moyenne donnée par `ROOT` est de 8,098, et l'écart type de 575,5. La question que l'on se pose est si la valeur de la moyenne est compatible avec 0 vu le nombre d'événements générés. Pour ce faire, on veut tester l'hypothèse que notre distribution de valeurs suit bien une loi uniforme de moyenne nulle. On va donc commencer par expliciter la distribution de la variable aléatoire "moyenne"
La loi de probabilité que suit l'estimateur de la moyenne peut s'écrire comme :

$$\begin{aligned}
\mu &= \frac{1}{N}\sum_{i=1}^{N} X_i \\
&= \frac{1}{N}\sum_{i=1}^{N} (-1\,000 + 2\,000\,U_i) \\
&= -1\,000 + \frac{2\,000}{N} \cdot \sum_{i=1}^{N} U_i \\
&= -1\,000 + \frac{2000}{N} \cdot X,
\end{aligned}$$

où X est une variable aléatoire qui suit une loi dite de *Irwin-Hall* de paramètre N (où N est le nombre de valeurs générées), qui est par définition la loi suivie par une somme de N nombres aléatoires distribués entre 0 et 1. La fonction de densité de probabilité d'une variable de Irwin-Hall est donnée par :

$$f_X(x;N) = \frac{1}{2(N-1)!}\sum_{k=0}^{N}(-1)^k\binom{N}{k}(x-k)^{N-1}\operatorname{sgn}(x-k).$$

Pour trouver la fonction de distribution de la moyenne, il faut changer de variables, en utilisant la conservation des probabilités :

$$f_\mu(\mu;N)\mathrm{d}\mu = f_X(x;N)\mathrm{d}X \quad \Rightarrow \quad f_\mu(\mu;N) = f_X(x;N)\frac{\mathrm{d}X}{\mathrm{d}\mu}.$$

Comme $\frac{\mathrm{d}X}{\mathrm{d}\mu} = \frac{N}{2000}$, il suffit donc de multiplier la fonction de densité de probabilité de X par $\frac{N}{2000}$ pour trouver la fonction de densité de probabilité que suit la moyenne :

$$f_\mu(\mu;N) = \frac{N}{2\,000}\cdot\frac{1}{2(N-1)!}\sum_{k=0}^{N}(-1)^k\binom{N}{k}\left(\frac{N}{2\,000}(1\,000+\mu)-k\right)^{N-1}\operatorname{sgn}\left(\frac{N}{2\,000}(1\,000+\mu)-k\right).$$

Pour vérifier si notre résultat est compatible avec l'hypothèse d'une moyenne nulle, il faudrait calculer la fonction cumulative, ce qui correspond à calculer l'intégrale de l'expression ci-dessus. Ce n'est pas très pratique... Il y a cependant moyen d'échapper à cette difficulté en remarquant qu'on a généré beaucoup d'événements (10 000) et qu'on est donc dans les conditions d'application du *théorème central limite* (TCL). On peut donc approximer notre loi par une loi normale de moyenne 0 (moyenne d'une loi uniforme sur $[-1\,000, 1\,000[$) et de variance donnée par la variance d'une loi uniforme sur $[-1\,000, 1\,000[$ divisée par N (cf TCL) :

$$\mu \sim \mathcal{N}\left(0, \left(\frac{2000}{\sqrt{12}}\right)^2/N\right).$$

Pour tester notre hypothèse, on veut trouver la probabilité (*p-value*) que l'on trouve une valeur pire que celle qu'on a obtenue (8,098), autrement dit $\mathcal{P}\,(\mu < -8{,}098 \text{ ou } \mu > 8{,}098)$ (on dit qu'on prend la p-value "2-tailed" car on accepte aussi les valeurs négatives). Pour obtenir cette valeur, il suffit de consulter une table donnant la cumulative d'une loi normale, en ayant préalablement converti notre variable en loi centrée réduite : la valeur de la variable aléatoire qui correspond pour une loi centrée réduite est :

$$Z = \frac{\mu-\mu_0}{\sigma_0} = \frac{8{,}098-0}{\sqrt{\left(\frac{2\,000}{\sqrt{12}}\right)^2/N}} = 1{,}403.$$

En consultant la table pour la loi normale centrée réduite, on trouve que :

$$\mathcal{P}(Z \geq 1{,}403) = 1 - 0{,}923\,6 = 0{,}076\,4.$$

De même (puisque la gaussienne est symétrique) :

$$\mathcal{P}(Z \leq -1{,}403) = 0{,}076\,4.$$

Donc, $\mathcal{P}\,(\mu < -8{,}098 \text{ ou } \mu > 8{,}098) = 0{,}076\,4 + 0{,}076\,4 = 0{,}152\,8$. On a donc une p-value de 15 %, ce qui est supérieur aux 5 % conventionnels ("2-sigma"), et on ne peut donc pas négliger l'hypothèse d'une moyenne nulle.

3. Génération d'une loi normale : Commençons par démontrer que la méthode de Box-Muller génère bien une loi normale centrée réduite.
On part d'une loi normale centrée réduite en 2 dimensions (qui est donc le produit de 2 lois normales) :

$$f(x,y) = \frac{1}{(\sqrt{2\pi})^2}e^{-\frac{x^2+y^2}{2}}.$$

En changeant de variables pour se placer en coordonnées polaires, on trouve :

$$f(r,\theta) = \frac{1}{2\pi} e^{-\frac{r^2}{2}} r \mathrm{d}r \mathrm{d}\theta.$$

En définissant $s \equiv r^2$, on peut encore une fois changer de variables et écrire f comme :

$$f(s,\theta) = \left(\frac{1}{2\pi}\mathrm{d}\theta\right)\left(\frac{1}{2}e^{-\frac{s}{2}}\mathrm{d}s\right).$$

On a donc exprimé f en termes de deux variables aléatoires indépendantes s et θ. De l'expression ci-dessus, on voit que θ suit une loi uniforme sur $[0, 2\pi[$ et que s suit une loi exponentielle de paramètre $1/2$.
Comme x et y suivent toutes les deux une loi gaussienne, il faut générer x et y à partir de θ et de s, de la façon suivante :

$$\begin{cases} x = r\cos\theta = \sqrt{s}\cos\theta \\ y = r\sin\theta = \sqrt{s}\sin\theta. \end{cases}$$

Cependant, on veut se ramener à des variables qui suivent une loi uniforme entre 0 et 1. Pour θ, c'est facile : θ suit une loi uniforme entre 0 et 2π, donc $\theta/2\pi$ suit une loi uniforme sur $[0, 1[$.
Pour s, qui suit une exponentielle, il faut employer la méthode de la transformée inverse. Cette méthode sera vue en détail lors de la séance 2, nous ne la redémontrons donc pas ici. Nous utiliserons uniquement le résultat final, qui est que l'on peut trouver une fonction d'une loi uniforme U sur $[0, 1[$ qui suit la même loi de distribution que s :

$$S = f(U) = -2\ln(U)$$

est distribuée comme une exponentielle.
Par conséquent, si on génère deux variables aléatoires u_1 et u_2 qui sont toutes les deux distribuées selon des lois uniformes *indépendantes* U_1 et U_2, alors les variables aléatoires

$$\begin{cases} X \equiv T_1 = \sqrt{-2\ln(U_1)}\cos(2\pi U_2) \\ Y \equiv T_2 = \sqrt{-2\ln(U_1)}\sin(2\pi U_2) \end{cases}$$

suivent toutes les deux une loi normale centrée réduite.

a) Pour générer une gaussienne $\mathcal{N}(\mu, \sigma^2)$, il suffit de décaler l'origine de μ (autrement dit, ajouter μ aux valeurs de T_1 et T_2 générées), puis de multiplier la valeur générée par σ. Remarquez que comme T_1 et T_2 sont indépendantes, une itération permet de générer deux valeurs indépendantes à la fois.

b) Les paramètres du fit (affichés à la ligne de commande) donnent également l'incertitude sur ces paramètres, ce qui nous permet de vérifier que les valeurs fittées sont compatibles avec les paramètres de la gaussienne que l'on a voulu générer.

c) On souhaite compter les événements en dessous de 0 ou au-dessus de 20 en générant une loi normale centrée en 10 et d'écart type 4. En ramenant à une normale centrée réduite, le nombre d'événements attendus au-dessus de 20 est donné par l'integrale d'une loi normale centrée réduite à partir de $2{,}5\,\sigma$ (car 20 se situe à 2,5 fois l'écart type à partir de 10, multipliée par le nombre d'événements tirés (idem pour le nombre d'événements en dessous de 0). En consultant la table de la loi normale centrée réduite, on trouve $\mathcal{P}(Z \geq 2{,}5) = 0{,}006\,2$. Puisqu'on compte générer 10 000 événements, on attend donc environ 62 événements au-dessus de 20.
Une fois l'histogramme dessiné dans `ROOT`, un clic droit dans la légende contenant les données statistiques permet de connaître entre autres le nombre d'événements dans l'underflow et l'overflow : on sélectionne `"SetOptStat"` et on remplit le champ avec `11111111`. On obtient 69 événements dans l'underflow et 52 événements dans l'overflow. Ces valeurs sont-elles compatibles avec notre hypothèse qu'on a généré une gaussienne ?
Le nombre d'événements arrivant dans une région donnée est décrit par une loi de Poisson ; ici, comme on attendait 62 événements, on a une loi de Poisson de paramètre $\lambda = 62$, ce qui correspond à un écart type de $\sqrt{62} = 7{,}87$. L'écart le plus grand que l'on observe par rapport

à la valeur centrale est dans l'underflow, où on a un écart de $62 - 52 = 10$ événements. Cela correspond à 1,27 fois l'écart type, ce qui donne une p-value (2-tailed) de 20,4 % (en consultant les tables de la loi normale). Comme c'est supérieur à 5 %, c'est une valeur acceptable et on ne peut pas rejeter l'hypothèse de départ.
On peut également regarder si la moyenne donnée par `ROOT` est compatible avec l'hypothèse qu'on a généré une gaussienne de moyenne 10. La moyenne donnée par `ROOT` est de 9,973. L'estimateur moyenne suit une loi normale $\mathcal{N}\left(\mu = 10, \sigma^2 = \frac{16}{10\,000} = 0{,}04^2\right)$. En "centrant-réduisant" la normale, on constate qu'on se situe à $\frac{9{,}973-10}{0{,}04} = -0{,}675$ fois l'écart type. La p-value correspondante (2-tailed) est de $2 \times (1 - 0{,}751\,7) = 49{,}7\,\%$, ce qui est compatible avec l'hypothèse qu'on a généré une loi normale.

d) On a accès aux paramètres de l'ajustement en faisant un clic droit dans le cadre de stat, en sélectionnant `"SetOptFit"` et en entrant `1111`. On peut alors avoir accès aux paramètres de l'ajustement, ainsi qu'aux incertitudes associées. En particulier, la valeur pour Prob est la p-value (2-tailed) de l'ajustement. Elle est de 23,8 %, ce qui est acceptable. Le χ^2 par nombre de degrés de liberté vaut 106,6/97, ce qui est proche de 1 et donc acceptable. On ne peut donc pas rejeter l'hypothèse de la loi normale, et heureusement puisqu'on a généré une loi normale...