ANNEXE

A Représentations graphiques dans l'A.E.P.

Indications préliminaires

- *Objectif* : Comme nous l'avons vu à la fiche T.D. **??**, l'**A.E.P.** s'appuie sur l'étude descriptive de m (grand) réalisations indépendantes $\boldsymbol{y}_{[m]} := \left(y_{[\cdot]}\right)_m$ d'une variable aléatoire d'intérêt Y. Afin d'alléger l'introduction de l'**A.E.P.**, cette étude descriptive a été volontairement limitée à une étude quantitative n'utilisant aucune représentation graphique issue de la Statistique Descriptive. Les graphiques étant d'une grande aide pour représenter les répartitions de séries de données, ils vont donc nous aider à mieux appréhender le comportement aléatoire des variables aléatoires d'intérêt.
- *Représentations graphiques usuelles* : Les représentations graphiques des répartitions diffèrent selon la nature des variables. Ainsi, lorsque la variable est de nature discrète (i.e. les modalités ou valeurs possibles sont dénombrables), on utilise un diagramme en bâton, et lorsqu'elle est de nature continue (i.e. à valeurs dans un continuum qui est non dénombrable), un histogramme est utilisé. Ce choix pose problème lorsqu'une étude expérimentale nous amène à comparer sur un même graphique des répartitions de plusieurs variables n'ayant pas la même nature. Il n'est pas possible de représenter une variable continue par un diagramme en bâton mais il en est tout autrement pour une variable discrète qui peut se représenter via un histogramme discret que nous allons introduire très prochainement.
- *Histogramme (continu)* : Rappelons les règles générales pour construire un histogramme représentant la répartition de la série $z_1, z_2, \cdots, z_m$:
  1. L'ensemble des modalités est découpé en une partition d'intervalles pas forcément de même largeur.
  2. Chaque intervalle de la partition est représenté par un rectangle ayant pour base l'intervalle et de surface égale à la proportion des $z_1, z_2, \cdots, z_m$ appartenant au dit intervalle.
  3. La somme de tous les rectangles est donc égale à 100%=1.

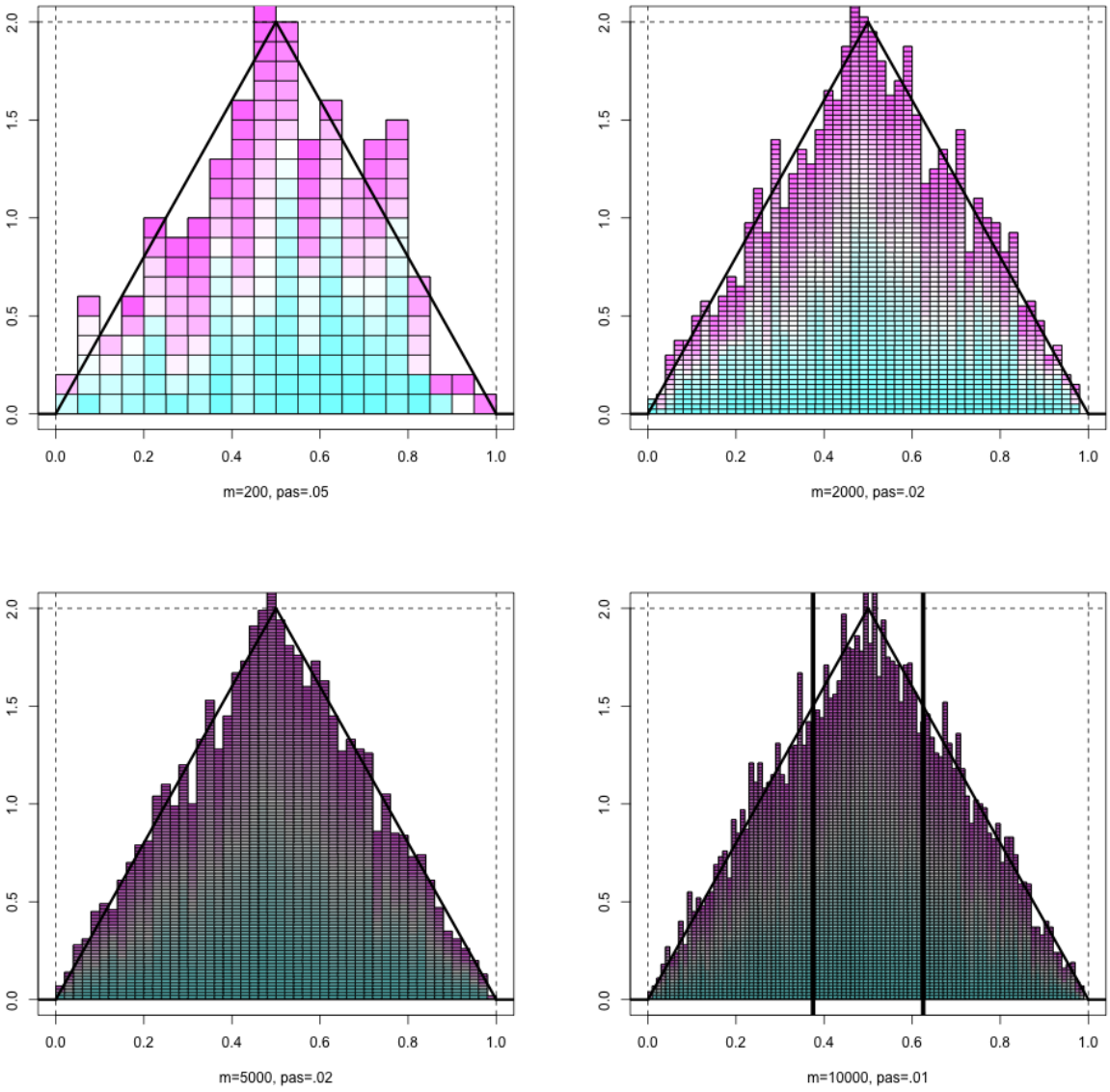
  Pour construire pratiquement un histogramme, il est conseillé au préalable de trier les $z_1, z_2, \cdots, z_m$ afin de les regrouper et ainsi de les affecter plus facilement à leurs intervalles d'appartenance. Au lieu de directement construire les rectangles associés aux intervalles de la partition, nous avons choisi d'associer à chaque donnée z_i un i$^{\text{e}}$ rectangle, que l'on appellera brique (l'histogramme étant un vu de manière imagée comme un "mur") :
  1. Toute brique associée à z_i a pour base l'intervalle d'appartenance de z_i et a une surface égale à $\frac{1}{m}$.
  2. La somme de toutes les briques associées aux $z_1, z_2, \cdots, z_m$ est donc égale à 1.
  3. Un rectangle associé à un intervalle est l'empilement de toutes les briques associés aux z_i de l'intervalle.

  Cette version de l'histogramme est équivalente à la version originale et propose en plus une représentation individualisée de toutes les données $z_1, z_2, \cdots, z_m$. Ce point particulier sera un atout pour comprendre la représentation graphique usuelle de loi de probabilités de variable aléatoire continue.
- *Histogramme discret* : Dans le même esprit, nous allons maintenant présenter la notion d'histogramme discret en l'adaptant à des données $z_1, z_2, \cdots, z_m$ à valeurs dans un espace dénombrable (donc pas dans un continuum) :
  1. Une brique associée à z_i a pour surface $\frac{1}{m}$ et sa base est centrée en z_i (à la différence d'une brique d'un histogramme continu où z_i doit appartenir à l'intervalle représentant la base de la brique).
  2. Les bases des briques sont fixées de sorte que le mur (i.e. l'histogramme) constitué de l'ensemble des briques ait le moins d'espace (trou) possible. Les briques voisines doivent se toucher dès que possible.
  3. La somme des briques est toujours égale à 100%=1 et l'empilement des briques associées aux données ayant même modalité forme un rectangle dont la surface est égale à la proportion des données égales à la modalité associée.

  IMPORTANT : Nous remarquons que dans la représentation d'histogramme (discret ou continu) les valeurs en ordonnée n'ont pas d'unité réellement interprétable. En revanche, lorsque les bases des rectangles sont toutes les mêmes, les hauteurs pourront donc être comparées entre elles pour nous éclairer sur la répartition des données puisque les aires des surfaces des rectangles (vues comme des empilements de briques) représentent naturellement des proportions de données.

Fin

Exercice 1 (Histogramme continu) *Cet exercice fait suite à l'exercice* **??** *mais dans l'esprit de l'exercice* **??** *puisqu'on s'intéresse à la moyenne plutôt que la somme. Voici 4 graphiques représentant les histogrammes continus des* $m = 200$*, 2000, 5000, 10000 premières réalisations de la variable* $M_2 := \frac{Y_1+Y_2}{2} = \frac{S}{2}$*. Nous rappelons que les* $m = 10000$ *réalisations de* S *avaient été stockées dans le vecteur* `s` *en* `R`*. Les* $m = 10000$ *réalisations* $\left(m_{2,[\cdot]}\right)_m$ *de* M_2 *sont donc accessibles en* `R` *via l'instruction* `s/2`*.*



*Dans le contexte de l'****A.E.P.****, les intervalles d'un histogramme continu peuvent sans restriction être de même largeur (car* m *est censé être suffisamment grand). Le "pas" d'un histogramme nomme en général la largeur du plus petit de ces intervalles.*

1. *Pour chaque graphique, indiquez quelle est la surface d'une brique.*
2. *Lequel de ces graphiques est le plus informatif ? A partir de ce dernier, êtes-vous en mesure de déterminer les valeurs de* M_2 *qui sont les plus probables ?*
3. *Identifiez les briques associées aux valeurs comprises entre* $\frac{3}{8}$ *et* $\frac{5}{8}$ *incluses. Quelle est la valeur de l'aire de la surface occupée par ces briques en vous rappelant que la proportion des* $\left(m_{2,[\cdot]}\right)_m$ *comprises entre* $\frac{3}{8}$ *et* $\frac{5}{8}$ *est fourni par :*

```
> mean(3/8<=s/2 &  s/2<=5/8)
[1] 0.4262
```

4. *Sauriez-vous alors évaluer approximativement la probabilité* $\mathbb{P}\left(M_2 \in [\frac{3}{8}, \frac{5}{8}]\right)$ *?*
5. *Est-il possible d'évaluer approximativement* $\mathbb{P}\left(M_2 = \frac{1}{2}\right)$ *qui via l'**A.E.P.** est approchée grâce à :*

```
> mean(s/2==1/2)
[1] 0
```

Que faudrait-il faire pour pouvoir y arriver ?

6. *En s'imaginant que le pas* $\to 0$ *au fur et à mesure que* $m \to +\infty$*, pouvez-vous décrire à quoi ressemblera une brique ? Même question pour le mur de briques ? Représentez-le sur le graphique en ne dessinant que le "dessus" (i.e. contour supérieur) du mur. Vu comme une fonction, comment interprèteriez-vous le contour supérieur du mur ?*
7. *Un mathématicien, sollicité pour nous assister dans l'étude de l'**A.M.P.**, nous apprend qu'il est classique de caractériser le comportement aléatoire de* M_2 *en fournissant la densité de probabilité (qui porte bien son nom !) s'exprimant ici mathématiquement par :*

$$f_{M_2}(t) = \begin{cases} 4t & si\ t \in [0, \frac{1}{2}] \\ 4 - 4t & si\ t \in [\frac{1}{2}, 1] \\ 0 & sinon \end{cases}$$

Représentez cette fonction sur le dernier graphique et comparez-la avec les histogrammes continus. Sont-ils très différents de la fonction ?

8. *Le mathématicien nous annonce que* $\mathbb{P}\left(M_2 \in [\frac{3}{8}, \frac{5}{8}]\right) = \int_{\frac{3}{8}}^{\frac{5}{8}} f_{M_2}(t)dt$ *qui est représentée graphiquement par la surface des points sous la courbe* $f_{M_2}(t)$ *et dont les abscisses sont compris entre* $\frac{3}{8}$ *et* $\frac{5}{8}$*. Représentez alors* $\mathbb{P}\left(M_2 \in [\frac{3}{8}, \frac{5}{8}]\right)$ *sur le graphique. Cela ne vous rappelle pas quelquechose ? Sachant qu'il n'est pas difficile de montrer que* $\mathbb{P}\left(M_2 \in [\frac{3}{8}, \frac{5}{8}]\right) = \mathbb{P}\left(S \in [\frac{3}{4}, \frac{5}{4}]\right) = \frac{7}{16} \simeq 43.75\%$ *(déjà évaluée à l'exercice **??**), évaluez l'aire de la surface représentant cette probabilité.*
9. *Quelle est la valeur exacte de* $\mathbb{P}\left(M_2 = \frac{1}{2}\right)$ *?*
10. *Sélectionnez la bonne réponse parmi les réponses (proposées en suivant entre parenthèses) : La modalité* $\frac{1}{2}$ *est le mode de la loi de* M_2 *car* $\frac{1}{2}$ *est la valeur qui maximise la fonction _____ (* $\mathbf{p_{M_2}}(\mathbf{x}) := \mathbb{P}(M_2 = x)$ *ou* $\mathbf{f_{M_2}}(\mathbf{x})$ *ou* $\mathbf{F_{M_2}}(\mathbf{x}) := \mathbb{P}(M_2 \leq x)$ *). Cela se traduit littéralement par :* $\frac{1}{2}$ *est la valeur de plus grande _____(**probabilité** ou **densité de probabilité** ou **fonction de répartition**).*

A retenir

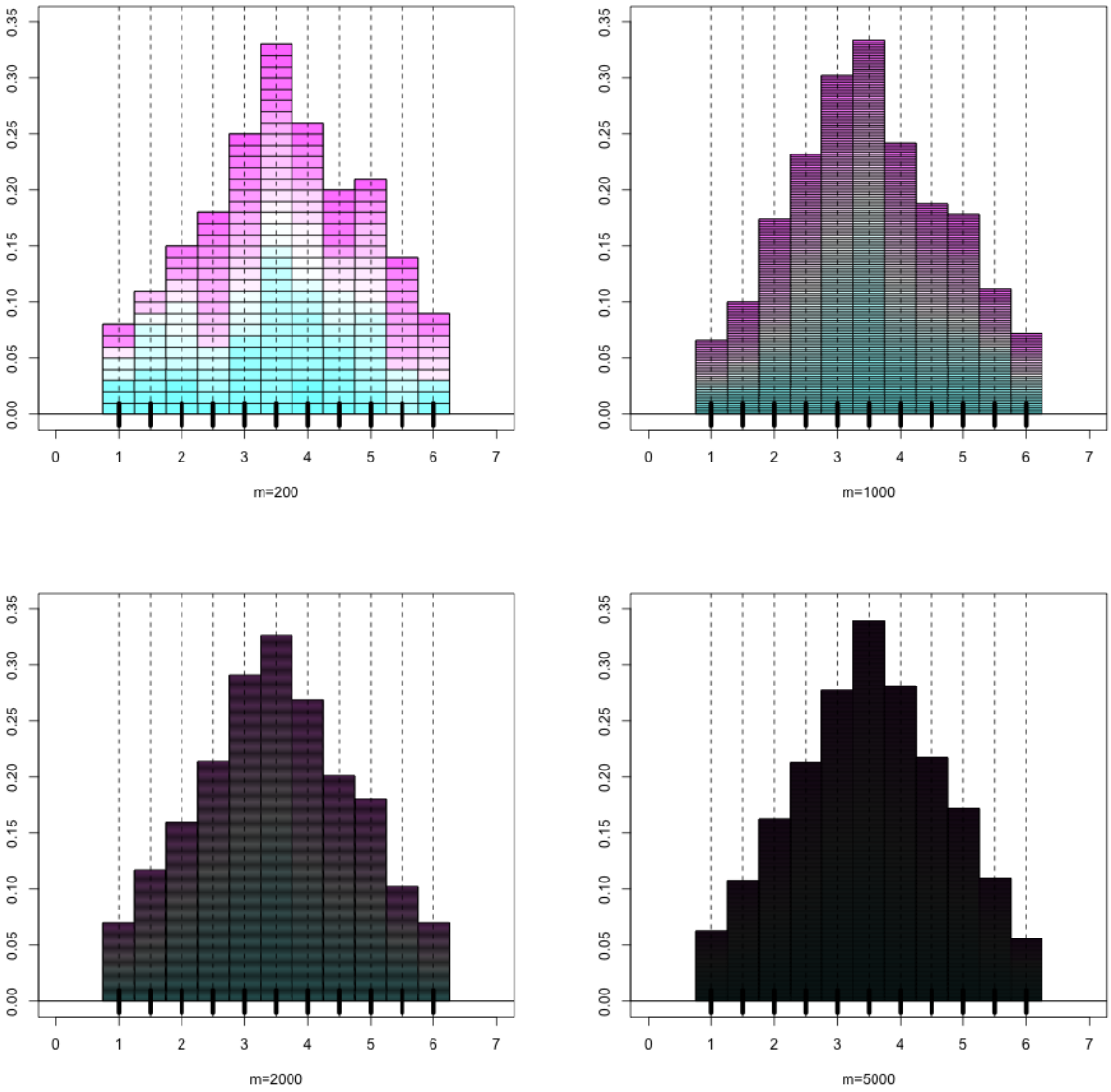
→ *La densité de probabilité caractérisant la loi de probabilité d'une v.a. continue* Y *est vue via l'**A.E.P.** comme le* `contour supérieur` *de **l'histogramme à "pas zéro" d'une infinité de ses réalisations** (i.e.* $\boldsymbol{y}_{[+\infty]} := \left(y_{[\cdot]}\right)_{+\infty}$*). De manière plus imagée, cet histogramme se décrit comme **un mur de briques devenues points** ou comme **un "tas de points"** (pour traduire la notion d'empilement) où chaque point est associé à une des réalisations. Autrement dit, tous ces objets permettent de décrire de manière très synthétique l'ensemble de "tous" les résultats possibles (représentés par les composantes de* $\boldsymbol{y}_{[+\infty]}$*) de la variable aléatoire* Y*.*

→ *La probabilité* $\mathbb{P}(Y \in [a, b]) = \int_a^b f_Y(t)dt$ *dans le contexte de l'**A.M.P.** correspond via l'**A.E.P.** à la proportion des composantes de* $\boldsymbol{y}_{[+\infty]}$ *appartenant à* $[a, b]$*. Du points de vue de l'**A.M.P.** ou de celui de l'**A.E.P.**, elle se représente par la surface occupée par les points sous la courbe* f_Y *et d'abscisses appartenant à* $[a, b]$*.*

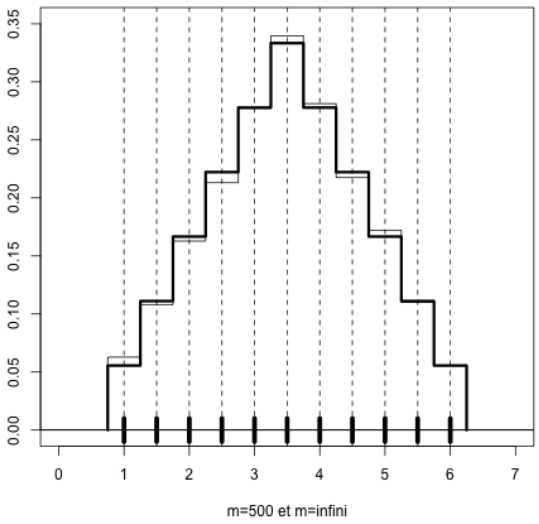
Exercice 2 (Histogramme discret) *On s'intéresse à la loi de probabilité de la moyenne* $M_2 := \frac{Y_1+Y_2}{2} = \frac{S}{2}$ *où* S *représente la somme de deux dés (introduite auparavant à l'exercice **??**). Le but est ici d'introduire la notion d'histogramme discret qui n'est pas la représentation graphique la plus usuelle pour représenter une variable aléatoire discrète. Voici 4 histogrammes*

discrets pour les $m = 200$, 1000, 2000 et 5000 premières somme des deux dés. Remarquons que dans le cadre de l'exercice, nous aurions pu nous limiter à la représentation graphique usuelle en diagramme en bâton puisque nous n'aurons aucune intention de comparer la loi de probabilité de M_2 avec celle d'une loi de probabilité associée à une variable aléatoire continue. Ce sera en revanche le cas dans l'exercice 3.

Afin de mettre en avant les caractéristiques les distinguant des histogrammes continus, nous avons complété les histogrammes en identifiant les modalités de M_2 par des petits traits en gras sur l'axe des abscisses prolongés par des lignes verticales en trait pointillé.



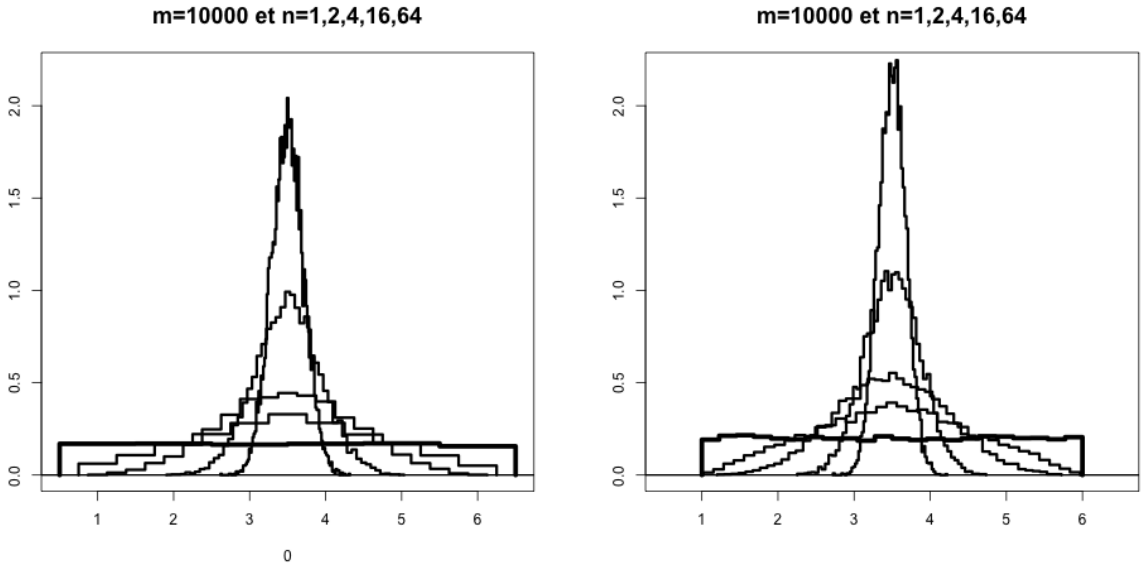
1. *Quelles sont les largeurs des briques utilisées dans ces histogrammes discrets ? Changent-elles lorsque m augmente (justifier votre réponse) ? Sur chacun des graphiques, indiquez l'aire de la surface de chaque brique.*
2. *A partir du premier graphique, donnez un ordre de grandeur de la probabilité $\mathbb{P}(M_2 = 1)$. Quelle est l'aire de la surface occupée par les briques associées aux $m_{2,[k]} = 1$? A-t'on vraiment besoin de voir les briques individuellement pour évaluer $\mathbb{P}(M_2 = 1)$? Si vous avez répondu non, appliquez cela sur le dernier graphique puisque les briques ne sont pas distinguables individuellement tellement elles sont plates.*
3. *Le graphique suivant fournit deux histogrammes discrets l'un correspondant à $m = 5000$ et l'autre à celui $m \to +\infty$. Deux types de trait (simple et en gras) ont été utilisés pour les représenter. Sauriez-vous les identifier ? Evaluez (le plus précisément possible) $\mathbb{P}(M_2 = 1)$ ainsi que $\mathbb{P}(M_2 \in \{1, 1.5, 6\})$.*



A retenir

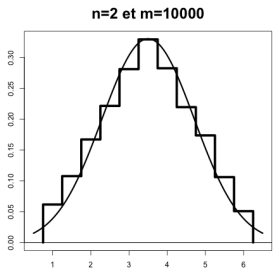
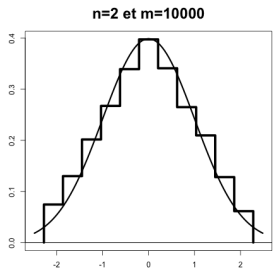
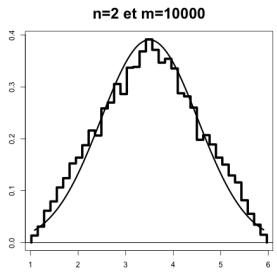
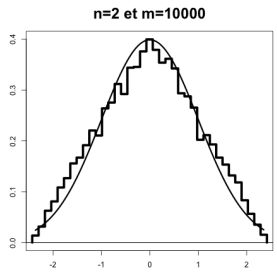
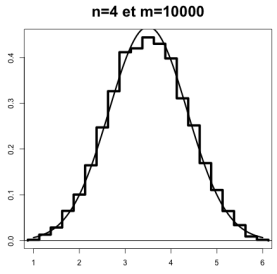
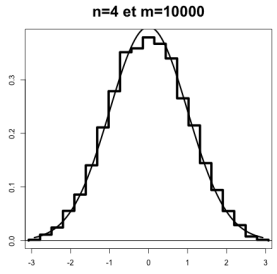
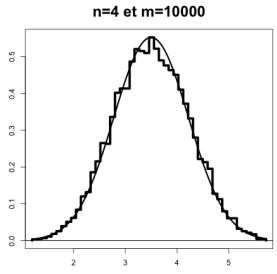
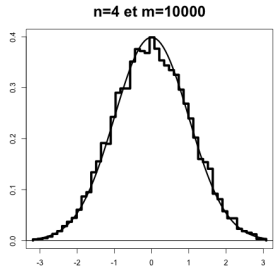
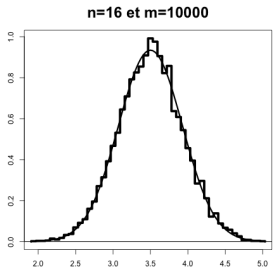
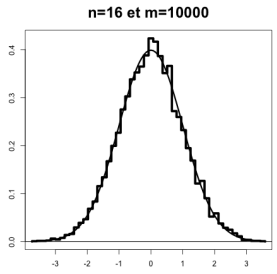
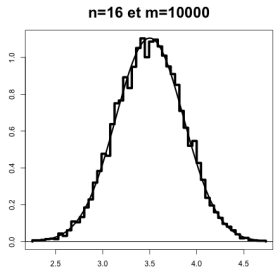
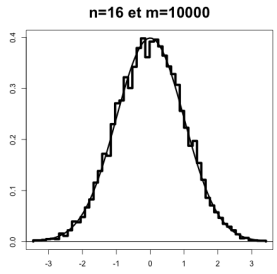
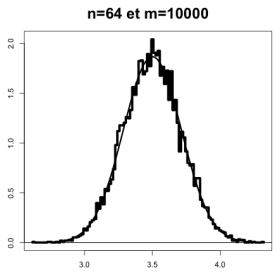
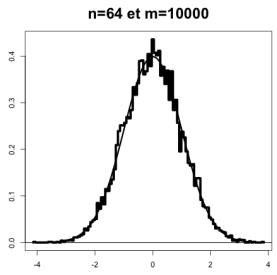
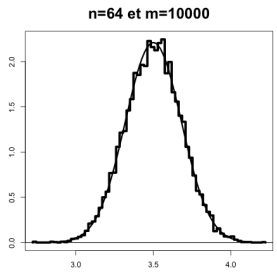
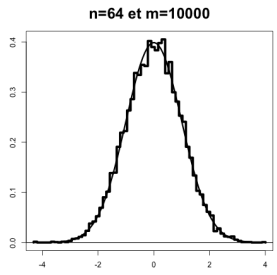
→ L'histogramme discret s'interprète de la même manière qu'un histogramme continu où les probabilités (ou proportions) sont mesurées via des aires de surfaces.
→ A la différence d'un histogramme continu, dans un histogramme discret :
– les briques sont de largeur fixe (même lorsque m varie)
– la base d'une brique n'a pas vraiment de sens
– seul le centre de la base d'une brique a un sens puisqu'il indique la valeur associée qui se lit en abscisse (surtout si la brique n'est pas la première à avoir été empilée).

Exercice 3 (Histogramme de moyenne) *L'étude menée est la suite de l'exercice* **??**. *Ayant introduit les notions d'histogrammes discret et continu, nous allons pouvoir apprécier de visualiser le théorème de la limite centrale notamment dans le cadre de l'exemple de la moyenne de dés. Voici pour commencer, 2 graphiques représentant les contours supérieurs des histogrammes (discrets pour l'exemple du dé à gauche et continu pour l'exemple de la loi uniforme sur* $[1, 6]$ *à droite) des lois de probabilité* M_n *pour* $n = 1, 2, 4, 16, 64$. *Les échelles sont identiques pour les 2 graphiques.*



1. *Pour chaque graphique, quelle est l'histogramme qui représente via l'**A.E.P.** la loi de probabilité approximative de* Y_1 *?*

2. *Ces représentations graphiques expriment-elles le résultat que nous avions décrit sur le procédé de moyennisation qui concentre les modalités ?*
3. *Comparez les 2 graphiques. Pour quelle étude (dé ou uniforme), la moyenne est de plus grande variance ?*
4. *Sauriez-vous anticiper les histogrammes pour le cas où $n \to +\infty$ avec $m \to +\infty$?*
5. *Comme il n'est pas possible d'observer la forme de l'histogramme dans le cas précédent, il est naturel de faire comme un photographe en rezoomant le graphique de sorte à pouvoir mieux cadrer l'histogramme sur le graphique. C'est aussi ce que fait automatiquement le logiciel `R` comme on peut le voir dans la série de graphiques suivants :*

Exemple du dé : $Y_i \rightsquigarrow \mathcal{U}(\{1,2,3,4,5,6\})$		*Exemple de l'uniforme : $Y_i \rightsquigarrow \mathcal{U}([1,6])$*	
loi de M_n	*loi de Δ_n*	*loi de M_n*	*loi de Δ_n*
			
			
			
			

6. *Pour les 2 exemples et pour chaque n, comparez la forme de l'histogramme (en trait le plus épais) des réalisations de M_n avec celle de l'histogramme (en trait le plus épais) des réalisations Δ_n ? Expliquez pourquoi il en est ainsi ?*
7. *Pourquoi ces histogrammes sont-ils de plus en plus irréguliers lorsque n augmente ? Qu'aurait dû faire l'expérimentateur pour qu'il n'en soit pas ainsi ? Pouvez-vous tout de même imaginer ce qui ce serait passé lorsque $m \to +\infty$?*
8. *D'après le Théorème de la limite centrale, on peut mathématiquement affirmer, lorsque n*

est suffisamment grand (convention simplifiée appliquée dans ce cours : $n \geq 30$) :

$$M_n \overset{approx.}{\rightsquigarrow} \mathcal{N}\left(\mathbb{E}\left(Y_1\right), \sqrt{\frac{\mathbb{V}ar\left(Y_1\right)}{n}}\right) \Leftrightarrow \Delta_n := \frac{M_n - \mathbb{E}\left(Y_1\right)}{\sqrt{\frac{\mathbb{V}ar(Y_1)}{n}}} \overset{approx.}{\rightsquigarrow} \mathcal{N}(0,1).$$

Aussi, on rappelle que, pour l'exemple du dé, $\mathbb{V}ar\left(Y_1\right) = 2.9167$ et que, pour l'exemple de la loi uniforme sur $[1,6]$, $\mathbb{V}ar\left(Y_1\right) = \frac{25}{12}$.
Pour chaque graphique, que représente la courbe en trait le plus fin ? Est-elle de plus en plus ressemblante à l'histogramme en trait le plus épais lorsque n augmente (Indication : éviter de tenir compte du caractère irrégulier de l'histogramme quand n augmente uniquement dû au fait que m aurait dû être augmenté en même temps que n) ?

9. *Dans le contexte de l'**A.E.P.**, comment décririez-vous ces courbes ? Dans l'exemple du dé, les deux histogrammes représentées sur chaque graphique sont-ils de la même nature ? Avez-vous une idée sur comment illustrer graphiquement le Théorème de la limite centrale sans l'utilisation de l'histogramme discret (Indication : Une réponse très courte est bien venue) ?*
10. *Le Théorème de la limite centrale s'appliquant pour tout Y_1 ayant n'importe quelle loi de probabilité admettant une moyenne et une variance finies, imaginez la même série de graphiques que précédemment mais pour d'autres exemples que ceux (uniformes) choisis dans cette étude.*