FICHE T.D. 2 Introduction aux probabilités

Indications préliminaires

- *Objectif* : L'originalité de ce cours réside essentiellement dans l'axe qui a été choisi pour présenter les probabilités. Dans un cours classique, les développements mathématiques (de nature plutôt technique) sont proposés en priorité en laissant peu de place à l'interprétation des concepts théoriques véhiculés. Cette approche pour introduire les concepts de probabilités sera par la suite appelée **A.M.P.** pour désigner **A**pproche **M**athématique des **P**robabilités. La Statistique (Inférentielle ou Inductive, celle présentée dans ce cours) repose largement sur la théorie des Probabilités, mais de part sa vocation à être largement utilisée par les praticiens sous une forme plutôt méthodologique, il s'ensuit souvent une difficulté pour ces utilisateurs à appréhender les conditions d'applicabilité et les points-clés des outils statistiques qui bien souvent s'expriment en fonction des concepts probabilistes pas toujours faciles à assimiler (compte tenu de leurs aspects mathématiques). Afin de remédier à cet incovénient, nous avons choisi de proposer une approche complémentaire, appelée **A.E.P.** pour désigner **A**pproche **E**xpérimentale des **P**robabilités, qui nous semble plus intuitive car basée sur l'expérimentation et dont la difficulté technique se limite aux outils de Statistique Desciptive présentés en première année (faciles à appréhender par les praticiens motivés surtout lorsqu'ils en ont l'utilité). L'objectif de cette fiche T.D. est essentiellement de faire le lien entre les deux approches **A.M.P.** et **A.E.P.**. Notamment, il sera essentiel de comprendre comment les praticiens pourrons être éclairés via l'**A.E.P.** sur les résultats techniques obtenus grâce à l'**A.M.P.** par les mathématiciens.
- *L'**A.E.P.** en complément de l'**A.M.P.*** : Soit Y une variable aléatoire réelle dont on suppose disposer (via l'**A.E.P.**) d'un vecteur $\mathbf{y}_{[m]} := (y_{[\cdot]})_m := (y_{[1]}, y_{[2]}, \cdots, y_{[m]})$ de m (a priori très grand) réalisations indépendantes entre elles. En théorie, on pourra aussi imaginer disposer du vecteur $\mathbf{y}_{[+\infty]} := (y_{[\cdot]})_{+\infty}$ qui est l'analogue de $\mathbf{y}_{[m]}$ avec $m \to +\infty$. Supposons aussi que $m = 10000$ expériences aient été réalisées et les m composantes de $\mathbf{y}_{[m]}$ aient été stockées dans `R` sous le vecteur nommé `yy`.

Quantité	**A.M.P.**		**A.E.P.** $(+\infty)$		**A.E.P.**		**Traitement** `R`
Probabilité	$\mathbb{P}(Y = a)$	$=$	$\overline{(y_{[\cdot]} = a)}_{+\infty}$	$\simeq$	$\overline{(y_{[\cdot]} = a)}_{m}$	$\overset{R}{=}$	`mean(yy==a)`
Probabilité	$\mathbb{P}(Y \in]a,b])$	$=$	$\overline{(y_{[\cdot]} \in]a,b])}_{+\infty}$	$\simeq$	$\overline{(y_{[\cdot]} \in]a,b])}_{m}$	$\overset{R}{=}$	`mean(a<yy & yy<=b)`
Moyenne	$\mathbb{E}(Y)$	$=$	$\overline{(y_{[\cdot]})}_{+\infty}$	$\simeq$	$\overline{(y_{[\cdot]})}_{m}$	$\overset{R}{=}$	`mean(yy)`
Variance	$\mathbb{V}ar(Y)$	$=$	$\overline{(y_{[\cdot]})^2}_{+\infty}$	$\simeq$	$\overline{(y_{[\cdot]})^2}_{m}$	$\overset{R}{=}$	`var(yy)=sd(yy)^2`
Quantile	$q_Y(\alpha)$	$=$	$q_\alpha\left((y_{[\cdot]})_{+\infty}\right)$	$\simeq$	$q_\alpha\left((y_{[\cdot]})_{m}\right)$	$\overset{R}{=}$	`quantile(yy,alpha)`

Les formules d'obtention des quantités ci-dessus pour les colonnes **A.M.P.** et **A.E.P.** n'ont pas été fournies. Celles concernant l'**A.M.P.** requiert un niveau plutôt avancé en mathématiques et diffèrent selon la nature (discrète ou continue) de Y. Un point fort de l'**A.E.P.** est que les formules d'obtentions ne dépendent pas de la nature de Y et sont normalement déjà connues en 1ère année dans le cours de Statistique Descriptive (pour rappel, voir polycopié de notre cours).
IMPORTANT : L'objectif principal de la fiche T.D. est l'assimilation des concepts décrits dans le tableau ci-dessus.

- *Quelques résultats sur **A.M.P.*** : Soient Y, Y_1 et Y_2 trois variables aléatoires réelles (v.a.r.) et λ un réel.
  Fonction de répartition $F_Y(y) := \mathbb{P}(Y \leq y)$: Dans l'**A.M.P.**, elle permet de calculer, pour tout $a \leq b$:
  $$\mathbb{P}(a < Y \leq b) = \mathbb{P}(Y \leq b) - \mathbb{P}(Y \leq a) = F_Y(b) - F_Y(a).$$
  Moyenne (théorique) : $\mathbb{E}(\lambda \times Y) = \lambda \times \mathbb{E}(Y)$ et $\mathbb{E}(Y_1 + Y_2) = \mathbb{E}(Y_1) + \mathbb{E}(Y_2)$
  Variance : $\mathbb{V}ar(\lambda \times Y) = \lambda^2 \times \mathbb{V}ar(Y)$ et $\mathbb{V}ar(Y_1 + Y_2) = \mathbb{V}ar(Y_1) + \mathbb{V}ar(Y_2)$
  où Y_1 et Y_2 sont en plus supposées indépendantes.

Fin

Exercice 1 (Lancer d'un dé)

1. *Proposer le Schéma de Formalisation pour la variable aléatoire correspondant à un futur*

lancer de dé.

Réponse

- ***Expérience $\mathcal{E}$:*** *Lancer un dé*
- ***Variable d'intérêt :*** *Y la face supérieure du dé*
- ***Loi de proba :*** *$\mathbb{P}(Y=k)=1/6$ avec $k=1,\cdots,6$ (si le dé est équilibré).*

Fin

2. *Quelle expérimentation mettriez-vous en oeuvre pour vérifier qu'un dé est rigoureusement non pipé (i.e. parfaitement équilibré) ? Pensez-vous qu'il existe un tel type de dé ?*
3. ***Application :*** *Un expérimentateur propose l'expérience suivante avec un dé (en théorie vendu) équilibré et un autre dont il a volontairement légèrement déséquilibré une ou plusieurs de ses faces. Les résultats des deux dés sont fournis dans un ordre arbitraire dans les tableaux ci-dessous. Sauriez-vous reconnaître les deux dés et, en particulier, déterminer les probabilités d'apparition des faces (sachant que, pour chaque dé, il n'y a en théorie pas plus de 2 choix possibles pour celles-ci) ? A partir de combien de lancers (m) êtes-vous en mesure de faire votre choix ?*

m	$\overline{(y=1)}_m$	$\overline{(y=2)}_m$	$\overline{(y=3)}_m$	$\overline{(y=4)}_m$	$\overline{(y=5)}_m$	$\overline{(y=6)}_m$	$\overline{(y)}_m$
100	21%	14%	15%	22%	16%	12%	3.34
1000	15.5%	16.8%	17.3%	17.1%	15.9%	17.4%	3.533
10000	16.46%	16.43%	16.45%	17.23%	16.46%	16.97%	3.5171
100000	16.4%	16.52%	16.28%	17.05%	16.83%	16.92%	3.5214
1000000	16.47%	16.52%	16.49%	16.85%	16.77%	16.89%	3.5161

m	$\overline{(y=1)}_m$	$\overline{(y=2)}_m$	$\overline{(y=3)}_m$	$\overline{(y=4)}_m$	$\overline{(y=5)}_m$	$\overline{(y=6)}_m$	$\overline{(y)}_m$
100	13%	13%	16%	21%	23%	14%	3.7
1000	16.1%	18.1%	15.6%	17.3%	18.6%	14.3%	3.471
10000	16.92%	17%	16.47%	16.91%	17.13%	15.57%	3.4704
100000	16.73%	16.64%	16.53%	16.59%	16.88%	16.63%	3.5015
1000000	16.68%	16.66%	16.68%	16.67%	16.71%	16.61%	3.499

4. *Fournir les instructions `R` ayant permis de déterminer les résultats des tableaux précédents.*
5. *Ayant à présent identifié (du moins nous l'espérons !) le dé équilibré, sauriez vous compléter le tableau suivant correspondant à l'éventuelle dernière ligne du tableau précédent lui correspondant :*

m	$\overline{(y=1)}_m$	$\overline{(y=2)}_m$	$\overline{(y=3)}_m$	$\overline{(y=4)}_m$	$\overline{(y=5)}_m$	$\overline{(y=6)}_m$	$\overline{(y)}_m$
∞							

Comment noteriez-vous ces quantités via l'A.M.P. ?

6. *Considérons le dé (théoriquement) équilibré. Observons les expressions dans le tableau ci-dessous obtenues par le mathématicien (A.M.P.). Sauriez-vous les calculer (N.B. : c'est une question personnelle et il est donc possible de répondre NON) ? On rappelle (pour votre culture) les formules d'obtentions de la moyenne (ou espérance) de Y :*

$$\mathbb{E}(Y)=\sum_{k=1}^{6} k\times\mathbb{P}(Y=k)$$

ainsi que celle de la variance

$$\mathbb{V}ar(Y)=\sum_{k=1}^{6}(k-\mathbb{E}(Y))^2\times\mathbb{P}(Y=k)=\mathbb{E}\left(Y^2\right)-\mathbb{E}(Y)^2=\sum_{k=1}^{6}k^2\times\mathbb{P}(Y=k)-\mathbb{E}(Y)^2$$

$\mathbb{P}(Y\in[2,4[)$	$\mathbb{E}(Y)$	$\mathbb{V}ar(Y)$	$\sigma(Y)$	$q_{5\%}(Y)$	$q_{50\%}(Y)$	$q_{95\%}(Y)$
33.33%	3.5	2.9167	1.7078	1	3.5	6

Remarque (pour les amateurs) *: Puisque* $\mathbb{P}(Y=k)=\frac{1}{6}$*, les valeurs du tableau pour* $\mathbb{E}(Y)$*,* $\mathbb{V}ar(Y)$ *et* $q_p(Y)$ *(*$p=$*5%, 50% et 95%) ont simplement été obtenues en appliquant les formules de Statistique Descriptive pour la série de chiffres* $1,2,3,4,5,6$*.*

7. *Comprenons comment ces quantités peuvent être obtenues (ou intreprétées) par l'expérimentateur en les confrontant à ses résultats sur* $m=1000000$ *lancers (A.E.P.). Proposez aussi les instructions* `R` *ayant permis de les construire sachant que ces résultats ont été stockés dans le vecteur* **`yy`** *en* `R`*.*

$\overline{(y\in[2,4[)}_m$	$\overline{(y)}_m$	$\left(\overleftrightarrow{(y)}_m\right)^2$	$\overleftrightarrow{(y)}_m$	$q_{5\%}((y)_m)$	$q_{50\%}((y)_m)$	$q_{95\%}((y)_m)$
33.34%	3.499	2.9145	1.7072	1	3	6

8. *Quelle approche (A.M.P. ou A.E.P.) vous semble être la plus facile à appréhender ? Comprenez-vous les intérêts propres à chacune d'entre elles ?*

Exercice 2 (Somme de deux dés)

1. *Soient* Y_V *et* Y_R *deux variables aléatoires correspondant aux faces de 2 dés (Vert et Rouge) à lancer. Définissons* $S=Y_V+Y_R$ *correspondant à la somme de deux faces. Proposez le Schéma de Formalisation pour* S*.*

Réponse

- ***Expérience*** $\mathcal{E}$ ***:*** *Lancer de 2 dés*
- ***Variable d'intérêt :*** S *la somme des faces supérieures des 2 dés*
- ***Loi de proba :*** $\mathbb{P}(S=k)=$*??? avec* $k=2,\cdots,12$*.*

Fin

2. *Comparez* $\mathbb{P}(S=2)$*,* $\mathbb{P}(S=12)$ *et* $\mathbb{P}(S=7)$*). Sauriez-vous les évaluer ?*
3. *Que peut-on espérer en moyenne sur la valeur de* S *? (cette quantité rappelons-le est notée* $\mathbb{E}(S)$*).*
4. *Un joueur se propose de lancer* $m=5000$ *fois deux dés. A chaque lancer, il note la somme et stocke l'ensemble des informations dans un vecteur noté* `s` *en* `R`*. Voici quelques résultats d'instructions* `R` *:*

```
> s
   [1]  8  8  8  9  5  4  4  4  3  6  7  2  3 10  6  2  6  9  2  9 12  7 10 12
  [25]  3  5  9  6  6  7  7  6  7  8  9  8  7  3  4  9  8 10  5  8  7  6  8  8
...
[4969]  6 10  9  9  9 11  7  7 10  6  6 12  4  9  7  9 10  2  8  9  7  7  7  4
[4993]  8  7 12  8 10 11  6  9
> mean(s==2)
[1] 0.0314
> mean(s==12)
[1] 0.0278
> mean(s==7)
[1] 0.1698
> mean(s)
[1] 7.0062
> var(s)
[1] 5.872536
> sd(s)
[1] 2.423332
```

Pourriez-vous proposer les notations mathématiques (norme CQLS) correspondant aux résultats obtenus dans la sortie `R` *ci-dessus ?*

5. *Cette approche expérimentale confirme-t-elle le résultat du mathématicien affirmant que pour toute modalité* $k=2,\cdots,12$ *de* S*,*

$$\mathbb{P}(S=k)=\begin{cases}\frac{k-1}{36} & si\ k\leq 7\\ \frac{13-k}{36} & si\ k\geq 7\end{cases}$$

*Voici les résultats de l'****A.M.P.*** *présentés dans le tableau suivant (que vous pouvez vérifier si vous avez l'âme d'un mathématicien) :*

$\mathbb{P}(S=2)$	$\mathbb{P}(S=12)$	$\mathbb{P}(S=7)$	$\mathbb{E}(S)$	$\mathbb{V}ar(S)$
2.78%	2.78%	16.67%	7	5.8333

6. *Pourriez-vous aussi vérifier la validité des formules sur l'espérance et variance de la somme de variables aléatoires réelles fournies au début de cette fiche.*

Exercice 3 (Loi uniforme sur l'intervalle unité) *1. Soit Y_1 une variable aléatoire suivant une loi uniforme sur $[0,1]$ (en langage math., $Y_1 \rightsquigarrow \mathcal{U}([0,1])$), correspondant au choix "au hasard" d'un réel dans l'intervalle $[0,1]$. L'objectif est l'évaluation (exacte ou approximative) des probabilités suivantes $\mathbb{P}(Y_1 = 0.5)$ et $\mathbb{P}(0 < Y_1 < 0.5)$, le chiffre moyen $\mathbb{E}(Y_1)$ (espéré), l'écart-type $\sigma(Y_1)$ ainsi que la variance $\mathbb{V}ar(Y_1)$? Parmi ces quantités, lequelles sauriez-vous intuitivement (i.e. sans calcul) déterminer ?*

2. *Via **A.E.P.** : Un expérimentateur réalise cette expérience en choisissant 10000 réels au hasard (par exemple en tapant 10000 fois sur la touche RAND d'une calculatrice). Il stocke les informations dans son logiciel préféré (libre et gratuit) `R` dans un vecteur noté `y1`. Déterminez approximativement les quantités de la première question.*

```
> y1
    [1] 0.6739665526 0.7397576035 0.7916111494 0.6937727907 0.6256426109
    [6] 0.4411222513 0.8918520729 0.4331923584 0.4213763773 0.6879929998
...
 [9991] 0.3117644335 0.1422109089 0.4964213229 0.6349032705 0.3718051254
 [9996] 0.2839202243 0.7170524562 0.7066086838 0.9236146978 0.7250815830
> mean(y1)
[1] 0.4940455
> mean(y1==0.5)
[1] 0
> mean(0.25 <y1 & y1<0.5)
[1] 0.254
> var(y1)
[1] 0.08296901
> sd(y1)
[1] 0.2880434
> sd(y1)^2
[1] 0.08296901
```

3. *Via **A.M.P.** : Un mathématicien obtient par le calcul les résultats suivant pour une variable aléatoire Y représentant un chiffre au hasard dans l'intervalle $[a,b]$ (i.e. $Y \rightsquigarrow \mathcal{U}([a,b])$) :*
   (a) *pour tout $a \leq t_1 \leq t_2 \leq b$, $\mathbb{P}(t_1 \leq Y \leq t_2) = \frac{t_2 - t_1}{b-a}$.*
   (b) $\mathbb{E}(Y) = \frac{a+b}{2}$
   (c) $\mathbb{V}ar(Y) = \frac{(b-a)^2}{12}$

   Question optionnelle : lesquels de ces résultats sont intuitifs (i.e. déterminables sans calcul) ? Déterminez exactement les quantités de la première question.

```
> 1/12
[1] 0.08333333
> sqrt(1/12)
[1] 0.2886751
```

4. *L'**A.E.P.** confirme-t'elle les résultats théoriques de l'**A.M.P.** ?*

Exercice 4 (Somme de deux uniformes) *1. On se propose maintenant d'étudier la variable $S = Y_1 + Y_2$ où Y_1 et Y_2 sont deux variables aléatoires indépendantes suivant une loi uniforme sur $[0,1]$. Quel est l'ensemble des valeurs possibles (ou modalités) de S ? Pensez-vous que la variable S suive une loi uniforme ? Nous nous proposons d'évaluer (excatement ou approximativement) les probabilités $\mathbb{P}\left(0 < S \leq \frac{1}{2}\right)$, $\mathbb{P}\left(\frac{3}{4} < S \leq \frac{5}{4}\right)$, $\mathbb{P}\left(\frac{3}{2} < S \leq 2\right)$, la moyenne $\mathbb{E}(S)$, l'écart-type $\sigma(S)$ et la variance $\mathbb{V}ar(S)$. Lesquelles parmi ces quantités sont déterminables intuitivement ou via un simple calcul mental ? Etes-vous capable de comparer les trois probabilités précédentes ?*

2. *Via* ***A.E.P.*** *: Un expérimentateur réalise à nouveau l'expérience de choisir 1000 réels entre 0 et 1. Les informations sont stockées dans le vecteur* `y2`*. Déterminez approximativement les quantités de la premire question.*

```
> y2
    [1] 7.050965e-01 7.167117e-01 8.085787e-01 5.334738e-01 1.126156e-01
...
 [9996] 8.175774e-01 5.379471e-01 4.259207e-01 7.629429e-01 9.217997e-01
> s<-y1+y2
> mean(0<s & s <=1/2)
[1] 0.1361
> mean(3/4<s & s<=5/4)
[1] 0.4262
> mean(3/2<s & s<=2)
[1] 0.1244
> mean(s)
[1] 0.9907449
> var(s)
[1] 0.1709682
> sd(s)
[1] 0.413483
> 1/sqrt(6)
[1] 0.4082483
> 7/16
[1] 0.4375
```

3. *Via l'****A.M.P.*** *: Par des développements plutôt avancés, le mathématicien obtient pour tout réel t :*

$$\mathbb{P}(S \leq t) = \begin{cases} 0 & si\ t \leq 0 \\ \frac{t^2}{2} & si\ 0 \leq t \leq 1 \\ 2t - 1 - \frac{t^2}{2} & si\ 1 \leq t \leq 2 \\ 1 & si\ t \geq 2 \end{cases}.$$

Etes-vous en mesure de déterminer les valeurs exactes de la première question ?

4. *L'****A.E.P.*** *confirme-t'elle les résultats théoriques de l'****A.M.P.*** *?*

Exercice 5 (Loi d'une moyenne) *Cet exercice est à lire attentivement à la maison. Il permet d'appréhender via l'approche expérimentale le résultat suivant central en Statistique Inférentielle :*

> *Une moyenne d'un grand nombre de variables aléatoires i.i.d. (indépendantes et identiquement distribuées, i.e. ayant la même loi de probabilité) se comporte approximativement selon la loi Normale (qui tire son nom de ce comportement universel).*

Rappelons que les paramètres d'une loi Normale sont sa moyenne et son écart-type (les matheux préférant sa variance). Notons aussi que ce résultat s'applique dans un cadre assez général excluant tout de même le cas de moyenne de variables aléatoires n'ayant pas de variance finie (et oui, tout arrive !!!).

1. *A partir des exercices 2 et 4, pouvez-vous intuiter les comportements aléatoires des moyennes de 2 faces de dés et de 2 uniformes sur* $[0, 1]$.

   Réponse

   De manière expérimentale, il suffit de diviser par 2 les vecteurs `s` *en* `R` *pour obtenir les quantités d'intérêts désirées. Via l'A.M.P., on obtient très facilement la fonction de répartition de M_2 pour tout réel t :* $\mathbb{P}(M_2 \leq t) = \mathbb{P}(S/2 \leq t) = \mathbb{P}(S \leq 2 \times t)$.
   Les moyenne, variance et écart-type de M_2 se déduisent très facilement de ceux de S : $\mathbb{E}(M_2) = \mathbb{E}(S/2) = \mathbb{E}(S)/2$, $\mathbb{V}ar(M_2) = \mathbb{V}ar(S/2) = \mathbb{V}ar(S)/4$ *et* $\sigma(M_2) = \sigma(S)/2$.

   Fin

2. *On constate sur ces deux exemples que les modalités centrales (autour de la moyenne) sont plus probables pour la moyenne $M_2 := (Y_1 + Y_2)/2$ que sur l'une ou l'autre des variables*

aléatoires Y_1 et Y_2. Pensez-vous que ce phénomène reste vrai pour n'importe quelle paire de variables aléatoires i.i.d. selon Y ? (C'est votre avis qui est demandé !)

3. *Un expérimentateur, convaincu que ce principe est vrai, observe que la moyenne de 4 v.a. i.i.d. se décompose aussi comme une moyenne de 2 v.a. i.i.d. comme le montre la formule suivante :*

$$M_n := \frac{Y_1 + Y_2 + Y_3 + Y_4}{4} = \frac{\frac{Y_1+Y_2}{2} + \frac{Y_3+Y_4}{2}}{2}$$

Il en déduit alors que les valeurs centrales (autour de la moyenne des Y) de la moyenne de 4 v.a. i.i.d. selon Y sont plus probables que celles de la moyenne de 2 v.a. i.i.d. selon Y qui sont elles-mêmes plus probables que celles de Y. Itérant ce processus, il constate que les moyennes M_n de $n = 2^k$ (avec k un entier aussi grand qu'on le veut) v.a. i.i.d. s'écrit aussi comme une moyenne de 2 v.a. i.i.d. étant elles-mêmes des moyennes de 2^{k-1} v.a. i.i.d. elles-mêmes s'écrivant comme des moyennes de 2 v.a. i.i.d. .... En conclusion, il postule que les probabilités d'apparition des modalités centrales de Y augmentent pour la moyenne M_n de n v.a. i.i.d. selon Y lorsque n augmente. Qu'en pensez-vous au vu de son protocole expérimental suivant (les réalisations de M_n sont notées $\mu_{n,[k]}$ et correspondent aux moyennes des lancers de n dés) ?

n	$\overline{\left(\mu_{n,[\cdot]} \in [1,2[\right)}_m$	$\overline{\left(\mu_{n,[\cdot]} \in [2,3[\right)}_m$	$\overline{\left(\mu_{n,[\cdot]} \in [3,4]\right)}_m$	$\overline{\left(\mu_{n,[\cdot]} \in]4,5]\right)}_m$	$\overline{\left(\mu_{n,[\cdot]} \in]5,6]\right)}_m$
1	16.92%	17%	33.38%	17.13%	15.57%
2	8.46%	19.42%	44.63%	19.65%	7.84%
4	2.69%	20.96%	52.59%	21.05%	2.71%
8	0.4%	17.22%	64.83%	17.1%	0.45%
16	0.01%	10.76%	78.32%	10.91%	0%
32	0%	4.27%	91.45%	4.28%	0%
64	0%	0.7%	98.43%	0.87%	0%

4. *L'expérimentateur demande à son ami mathématicien s'il peut justifier sur un plan théorique (via A.M.P.) ces résultats. A sa grande surprise, le mathématicien lui annonce que ce résultat est central en statistique sous le nom de Théorème de la limite centrale (central limit theorem en anglais). Il s'énonce dans le cadre de la moyenne sous la forme suivante : pour toute v.a. Y et lorsque n est suffisamment grand (en général, $n \geq 30$)*

$$M_n := \frac{1}{n}\sum_{i=1}^{n} Y_i \overset{approx.}{\rightsquigarrow} \mathcal{N}\left(\mathbb{E}\left(M_n\right), \sqrt{\frac{\mathbb{V}ar\left(Y_1\right)}{n}}\right)$$

où $Y_1, \cdots, Y_n$ désignent n v.a. i.i.d. selon Y. La loi Normale tire son nom de ce résultat étonnant et combien important dans le sens où beaucoup de phénomènes réels peuvent être vus comme des moyennisations. Le premier paramètre d'une loi Normale correspond à l'espérance $\mathbb{E}\left(M_n\right)$ de M_n et le second à l'écart-type de M_n. Le fait marquant est que ce résultat est vrai indépendemment de la loi de Y. Afin de comparer ces résultats à ceux qu'il a déjà effectué sur la loi uniforme, il transforme toutes les réalisations des lois uniformes sur $[0,1]$ en les multipliant par 5 puis en les additionnant à 1 de sorte que toutes les nouvelles réalisations à moyenner soient celles d'une loi uniforme sur $[1,6]$. L'ensemble des modalités ainsi que celui du dés sont comprises entre 1 et 6. Ainsi, il lui semble possible de comparer les probabilités dans les deux exemples puisque les supports sont les mêmes ainsi que leurs espérances égales à 3.5.

n	$\overline{(\mu_{n,[\cdot]} \in [1,2[)}_m$	$\overline{(\mu_{n,[\cdot]} \in [2,3[)}_m$	$\overline{(\mu_{n,[\cdot]} \in [3,4])}_m$	$\overline{(\mu_{n,[\cdot]} \in]4,5])}_m$	$\overline{(\mu_{n,[\cdot]} \in]5,6])}_m$
1	16.92%	17%	33.38%	17.13%	15.57%
2	8.28%	23.81%	35.54%	23.99%	8.38%
4	1.75%	23.48%	49.27%	23.57%	1.93%
8	0.12%	16.08%	67.25%	16.33%	0.22%
16	0%	8.19%	83.46%	8.35%	0%
32	0%	2.3%	95.08%	2.62%	0%
64	0%	0.24%	99.46%	0.3%	0%

Qu'en pensez-vous ? Observez-vous à nouveau que le procédé de moyennisation concentre les probabilités vers les modalités centrales (en fait autour de l'espérance) ?

5. *Le mathématicien lui fait cependant remarquer qu'a priori les variances ne sont pas rigoureusement les mêmes (certainement assez proches) et qu'il n'est donc pas en mesure de comparer les résultats expérimentaux sur les 2 exemples. Pour comparer les résultats pour différentes v.a. Y, il faut au préalable les uniformiser (les contraindre à avoir les mêmes moyennes et variances). Une solution est de les centrer (soustraire l'espérance $\mathbb{E}(M_n)$) et les réduire (diviser ensuite par $\sqrt{\mathbb{V}ar(M_n)} = \sqrt{\frac{\mathbb{V}ar(Y_1)}{n}}$) de sorte à ce que les v.a. résultantes soient toutes d'espérances 0 et de variances 1 (et ainsi comparables). Cette transformation pourra plus tard (via une représentation graphique) être comparé au travail d'un photographe lors d'une photo de groupe qui demande d'abord à l'ensemble des photographiés de se recentrer (i.e. centrage) puis utilise son zoom (i.e. réduction ou plutôt changement d'échelle dans ce cas précis) pour bien les cadrer. Aidé par le mathémacien, il compare donc ses résultats en effectuant la dite transformation. Le mathématicien l'informe donc du nouveau résultat suivant :*

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

N.B. *: Ce résultat n'est valide que lorsque les notions d'espérance et de variance ont un sens ! Il existe en effet des v.a. (suivant une loi de Cauchy, par exemple) n'ayant pas d'espérance et variances finies !*
Voici les résultats expérimentaux pour $n = 64$ (i.e. la valeur de n la plus grande) et $m = 10000$ pour consécutivement les exemples du dé (i.e. $Y \rightsquigarrow \mathcal{U}(\{1,\cdots,6\})$), de la loi uniforme sur $[0,1]$ (i.e. $Y \rightsquigarrow \mathcal{U}([0,1])$) et sur sa loi transformée uniforme sur $[1,6]$ (i.e. $5Y+1 \rightsquigarrow \mathcal{U}([1,6])$). Les tableaux ci-dessous sont complétés par les résultats via l'A.M.P. correspondant (théoriquement) à $m = +\infty$.

loi de Y	$\overline{(\delta_{n,[\cdot]} < -3)}_m$	$\overline{(\delta_{n,[\cdot]} \in [-3,-1.5[)}_m$	$\overline{(\delta_{n,[\cdot]} \in [-1.5,-0.5])}_m$	$\overline{(\delta_{n,[\cdot]} \in [-0.5,0.5[)}_m$
$\mathcal{U}(\{1,\cdots,6\})$	0.11%	6.4%	25.16%	36.76%
$\mathcal{U}([0,1])$	0.11%	6.85%	24.12%	37.63%
$\mathcal{U}([1,6])$	0.11%	6.85%	24.12%	37.63%

loi de Δ_n	$\mathbb{P}(\Delta_n < -3)$	$\mathbb{P}(\Delta_n \in [-3,-1.5[)$	$\mathbb{P}(\Delta_n \in [-1.5,-0.5])$	$\mathbb{P}(\Delta_n \in [-0.5,0.5[)$
$\mathcal{N}(0,1)$	0.13%	6.55%	24.17%	38.29%

loi de Y	$\overline{(\delta_{n,[\cdot]} \in [0.5,1.5[)}_m$	$\overline{(\delta_{n,[\cdot]} \in [1.5,3[)}_m$	$\overline{(\delta_{n,[\cdot]} \geq 3)}_m$	$\overline{(\delta_{n,[\cdot]})}_m$	$\overline{\overline{(\delta_{n,[\cdot]})}_m}$
$\mathcal{U}(\{1,\cdots,6\})$	24.66%	6.74%	0.17%	$-8e-04$	0.9953
$\mathcal{U}([0,1])$	24.66%	6.53%	0.1%	0.0021	1.0036
$\mathcal{U}([1,6])$	24.66%	6.53%	0.1%	0.0021	1.0036

loi de Δ_n	$\mathbb{P}(\Delta_n \in [0.5,1.5[)$	$\mathbb{P}(\Delta_n \in [1.5,3[)$	$\mathbb{P}(\Delta_n \geq 3)$	$\mathbb{E}(\Delta_n)$	$\sigma(\Delta_n)$
$\mathcal{N}(0,1)$	24.17%	6.55%	0.13%	0	1

Commentez ces résultats et expliquez en particulier pourquoi les 2 lignes correspondant aux 2 exemples des lois uniformes (non transformée et transformée) sont identiques ?

6. *Fournir les instructions R permettant d'obtenir les probabilités des tableaux précédents pour $m = +\infty$.*

Réponse

Pour tout $a < b$,

$$\mathbb{P}(\Delta_n \in [a, b[) = F_{\mathcal{N}(0,1)}(b) - F_{\mathcal{N}(0,1)}(a) \overset{R}{=} \texttt{pnorm(b)} - \texttt{pnorm(a)}$$

puisque $F_{\mathcal{N}(0,1)}$ est obtenu en `R` *en utilisant la fonction* `pnorm`*.*

Fin

Quelques commentaires

- Un étudiant suivant ce cours n'est pas censé comprendre comment les résultats de l'**A.M.P.** ont été mathématiquement obtenus. Ils sont généralement proposés sans démonstration. Sa mission est en revanche de savoir comment les vérifier via l'**A.E.P.** en prenant soin de bien les interpréter. Autrement dit, l'**A.E.P.** permet à un praticien de mieux comprendre les tenants et les aboutissants des outils statistiques (qu'il utilise) développés dans le contexte de l'**A.M.P.**.
- Afin d'éviter de surcharger l'étude de l'**A.E.P.**, il a été décidé dans ce cours d'étaler son introduction en deux étapes. La première qui vous a été présentée dans cette fiche est naturellement complétée par une deuxième étape qui s'appuie sur la représentation graphique des répartitions de $\boldsymbol{y}_{[m]} := \left(y_{[\cdot]}\right)_m$ (avec m généralement très grand). Cette étape est présentée en Annexe. Un étudiant motivé pourra à sa guise choisir de compléter sa connaissance sur l'**A.E.P.** en lisant dès à présent la fiche Annexe **??** en Annexe consacrée à l'**A.E.P.** dans sa version "graphique". Il est toutefois important de rappeler que les 2 fiches T.D. **??** et **??** suivantes ne s'appuient que sur les outils présentées dans la fiche T.D. présentée ici.
- Dans la suite du cours (nous en avons déjà eu un aperçu dans la fiche introductive précédente), la plupart des variables aléatoires d'intérêt, appelées statistiques, seront de la forme $T := t(\boldsymbol{Y})$ où t est une fonction s'appliquant à $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ qui représente le "futur" échantillon, seule source d'aléatoire dans la variable aléatoire $t(\boldsymbol{Y})$. C'est en effet le cas pour l'estimation d'un paramètre inconnu θ qui s'écrit $\widehat{\theta}(\boldsymbol{y})$ lorsqu'il est évalué à partir de l'échantillon que l'on obtient le **Jour J** (i.e. jour d'obtention des données) et qui est la réalisation de $\widehat{\theta}(\boldsymbol{Y})$ représentant le procédé d'obtention de l'estimation à partir du "futur" échantillon $\boldsymbol{Y}$. L'étude **A.E.P.** consistera alors à construire m échantillons $\left(\boldsymbol{y}_{[\cdot]}\right)_m$ où $\boldsymbol{y}_{[k]} := (y_{1,[k]}, \cdots, y_{n,[k]})$ représente le $k^{ème}$ échantillon de taille n construit parmi les m. Le comportement aléatoire d'une statistique $T := t(\boldsymbol{Y})$ sera donc appréhendé via l'**A.E.P.** en proposant m réalisations indépendantes $\left(t_{[\cdot]}\right)_m := \left(t(\boldsymbol{y}_{[\cdot]})\right)_m$ avec $t_{[k]} := t(\boldsymbol{y}_{[k]})$ la $k^{ème}$ réalisation de T parmi les m.

Fin