UNE BOITE DE PERLES
INTERVALLE DE FLUCTUATION – INTERVALLE DE CONFIANCE
THEOREME DE MOIVRE-LAPLACE

Michèle Gandit

Préambule

Le document suivant est destiné au(à la) professeur(e). L'objectif est de lui permettre de construire une séquence d'enseignement, favorisant l'investigation de la part des élèves à partir d'une question partant d'un dispositif expérimental simple, que chacun(e) peut construire facilement. Le niveau de classe visé est celui de Seconde, si l'on en reste à la première partie de ce document, à celui de Terminale. L'apprentissage de la démarche scientifique est annoncé clairement comme un objectif des programmes de mathématiques, du collège et du lycée (B.O. spéciaux n°6 du 28 août 2008, n°8 du 13 octobre 2011). C'est aussi ce que nous voulons mettre en avant dans ce document en proposant au départ un problème lui permettant d'appréhender le principe d'*un sondage aléatoire.* Le programme de terminale (B.O. n°8 du 13 octobre 2011) précise d'ailleurs que :

> L'enseignement des mathématiques au collège et au lycée a pour but de donner à chaque élève la culture mathématique indispensable pour sa vie de citoyen […].

Les contenus en jeu sont du niveau des classes de Seconde à Terminale :

- Echantillonnage (Seconde),
- la loi binomiale, la loi normale,
- le théorème central limite, ou, plus particulièrement le Théorème de Moivre Laplace,
- la notion d'intervalle de fluctuation, d'intervalle de confiance.

Nous proposons d'utiliser le logiciel Xcas pour visualiser des intervalles de fluctuation asymptotique au seuil de 95 %.

PREMIERE PARTIE

La situation de départ (pour le(la) professeur(e) et aussi pour les élèves)

A partir d'un dispositif simple, on cherche à faire saisir le principe d'un sondage aléatoire. Il est important de ne pas dévoiler à la classe cet objectif au départ. Le dispositif et la question posée ne relèvent pas *a priori* de l'aléatoire.

On considère une boîte, remplie de petites perles (environ 2 000) de deux couleurs : des rouges et des vertes. On la recouvre, sauf sur une petite lucarne, par de la bande adhésive

opaque de telle sorte qu'on ne puisse plus l'ouvrir. La question est de connaître la proportion p de perles vertes dans la boîte. Mais, évidemment, on ne peut pas ouvrir la boîte.

Cette proportion p est parfaitement déterminée, mais on ne peut l'obtenir directement. Il s'agit, dans un premier temps, d'amener les élèves à comprendre qu'on ne pourra pas se contenter de répondre par un nombre (un nombre compris entre 0 et 1 ou bien un pourcentage) dont on est certain qu'il est égal à p. L'idée doit faire son chemin qu'on va avoir recours à une procédure aléatoire pour obtenir une *estimation* de p (par un intervalle, avec un certain niveau de confiance).

Aussi est-il important de laisser les élèves réfléchir à une méthode permettant de répondre à la question. La lucarne ménagée dans la boîte permet de voir un échantillon de perles. En quoi la détermination de la proportion de perles vertes dans un tel échantillon peut-elle permettre de répondre à la question ? En effet, les résultats sur deux échantillons différents ont de fortes chances d'être différents. Que penser d'une réponse qui serait une proportion issue d'un échantillon ou une proportion calculée par moyenne (ou médiane) de séries de résultats issus de différents échantillons ? Il y a fort à parier que les élèves vont proposer de faire un tel calcul de moyenne. Mais peut-on être sûr que la valeur de p est égale à cette moyenne ? Evidemment non, puisque les résultats fluctuent d'un échantillon à l'autre, et aussi d'une série d'expériences à l'autre. Il s'agit alors d'amener les élèves à étudier *cette fluctuation des résultats* et de comprendre que, même si les résultats sont variables, on peut tout de même « cerner » cette *variabilité*.

Afin que chaque élève comprenne réellement la situation, il faut lui donner l'occasion de faire l'expérience suivante qui consiste à agiter la boîte et à compter les nombres de perles vertes et rouges qu'on voit dans la lucarne. Des questions vont alors se poser sur le protocole et sur le but recherché : comment voir correctement les perles, quel est l'intérêt de voir beaucoup de perles par rapport au fait d'en appréhender seulement un petit échantillon, compter seulement les perles vues en entier ou dénombrer les perles dont on voit juste une partie, pourquoi agiter la boîte, que va-t-on faire des résultats, les tirages d'échantillons sont-ils ou non indépendants… ?

L'objectif du travail qui suit n'est pas de répondre à la question de l'estimation, mais de faire comprendre (aux élèves) que, pour traiter la question de l'estimation, *il faut préalablement comprendre la fluctuation des différents résultats des sondages*. Ainsi il faut changer de point de vue et se poser la question suivante : connaissant la valeur de la proportion de perles vertes dans la boîte, comment se répartissent, autour de cette valeur, les proportions obtenues sur des échantillons ?

Pour le(la) professeur(e), un peu de théorie pour comprendre où l'on va

On fait une expérience : on agite la boîte et, dans la lucarne, on compte[1]… 12 perles rouges et 20 perles vertes, soit donc 32 perles au total, 20/32 (= 0,625). On peut estimer que la proportion de perles vertes est de 0,63 (ou 63 %), mais ce résultat ne présente guère d'intérêt si l'on n'a aucune idée de sa précision.

Que peut-on dire de la précision de ce résultat ?

Cet échantillon de 32 perles est suffisamment petit (par rapport au total des perles contenues dans la boîte) pour qu'on puisse considérer qu'il a été obtenu par un tirage avec remise : on fait comme si l'on avait répété 32 fois l'expérience qui consiste à tirer au hasard

[1] On a intérêt à avoir un échantillon d'une trentaine de perles en tout.

une perle dans la boîte, noter sa couleur et la remettre avant de tirer la suivante. *Autrement dit, on assimile cette expérience à un sondage aléatoire dans la population des perles.*

Si l'on désigne par X_n le nombre de perles vertes observées dans un échantillon de taille n, on sait que X_n suit la loi binomiale de paramètres n et p. Sa moyenne est égale à $n\,p$ et son écart-type à $\sqrt{np(1-p)}$.

La variable aléatoire F_n telle que $F_n = \dfrac{X_n}{n}$ donne la fréquence du nombre de perles vertes dans l'échantillon. Sa moyenne théorique est égale à p et son écart-type à $\sqrt{\dfrac{p(1-p)}{n}}$. Ainsi la loi de probabilité de la fréquence observée F_n est centrée autour de la proportion p. Plus n est grand, plus l'écart-type de F_n est petit, plus on a de chances d'observer une valeur f_n de F_n proche de p.

Revenons à la question concernant la précision de la proportion de 63 % trouvée à partir de l'expérience. Nous allons ici compléter cette fréquence observée par la détermination, autour de cette valeur estimée, d'un intervalle dont on a de bonnes raisons de croire qu'il contient la valeur de la proportion p cherchée.

En fait on a un estimateur de p, qui est F_n et on va déterminer, de part et d'autre de F_n les bornes d'un intervalle aléatoire (*intervalle de fluctuation asymptotique*, comme le nomme le programme de Terminale 2012), qui a une *forte* probabilité – désignée par $1 - \alpha$, où α désigne le *risque* (choisi *petit*) – de contenir p.

On choisit $\alpha = 0{,}05$ pour la suite (qui est un risque assez « standard »), mais on peut reprendre le raisonnement avec d'autres valeurs du risque. Le choix de cette valeur du risque présente aussi l'avantage de permettre une simplification de l'écriture de l'intervalle de confiance. En effet, le théorème central limite permet d'obtenir qu'avec une probabilité environ égale à 0,95, l'intervalle $\left[F_n - 1{,}96\sqrt{\dfrac{p(1-p)}{n}}\ ;\ F_n + 1{,}96\sqrt{\dfrac{p(1-p)}{n}}\right]$ contient p. Mais les bornes de cet intervalle dépendent elles-mêmes de p (qu'on ne connaît pas). En majorant[2] $\sqrt{p(1-p)}$ par 0,5, puis $1{,}96\sqrt{p(1-p)}$ par 1, on obtient qu'avec une probabilité environ égale à 0,95 (en fait un peu supérieure à 0,95), l'intervalle $\left[F_n - \dfrac{1}{\sqrt{n}}\ ;\ F_n + \dfrac{1}{\sqrt{n}}\right]$ contient p (*intervalle de fluctuation asymptotique, au seuil* 0,95).

Dans l'expérience évoquée au début de ce paragraphe, la taille n de l'échantillon étant égale à 32 (donc de l'ordre d'une trentaine), on obtient [0,45 ; 0,81] (intervalle tout de même large) comme réalisation particulière de cet intervalle, nommée aussi *fourchette de sondage au niveau* 0,95 ou *intervalle de confiance au niveau de confiance* 0,95 (d'après le programme de Terminale 2012). En fait, si l'on répète l'expérience « de la boîte » un grand nombre de fois, obtenant ainsi un grand nombre de fourchettes de sondages, environ 95 % d'entre elles contiennent la valeur de p.

C'est justement ce dernier point qu'il est important de faire comprendre aux élèves, avant de passer à l'estimation. Le(la) professeur(e) pourrait se contenter de dire aux élèves, dans un premier temps (c'est ce qui est recommandé dans le programme de Seconde) qu'un théorème permet d'établir que, pour une proportion p comprise entre 0,2 et 0,8, la fourchette centrée en

[2] majoration d'autant meilleure que p est proche de 0,5

p et de demi-largeur $1/\sqrt{n}$ contient « la plupart du temps » la fréquence obtenue sur tout échantillon de taille n (supérieure à 30). Il s'agit de faire comprendre le sens ici de cette expression « la plupart du temps », ainsi que l'importance de la taille de l'échantillon pour obtenir un intervalle « moins large ».

Avant de passer à l'étape de simulation, il est important que les élèves refassent tous une expérience avec la boîte de perles et qu'ils obtiennent une proportion de perles vertes sur un échantillon de taille supérieure à 30

Les résultats doivent être visibles de toute la classe : pas seulement la fréquence de perles vertes, mais aussi le nombre de perles observées (plus simplement, le nombre de perles vertes et de perles rouges observées). Chacun peut vérifier si le résultat qu'il a obtenu à partir de son échantillon appartient ou non à la fourchette centrée en p et de demi-largeur $1/\sqrt{n}$, pour différentes valeurs de p proposées par le(la) professeur(e).

Il s'agit alors de questionner les élèves sur une méthode pour obtenir une fourchette moins large : en rassemblant les résultats de chacun dans un seul « gros » échantillon (de taille supérieure à 30 × nombre d'élèves), ce qui permet de diminuer la largeur des fourchettes.

L'étape suivante va consister à simuler des fourchettes de la taille de ce « gros » échantillon pour comprendre comment elles se situent par rapport à diverses valeurs de p choisies. Cette simulation est à proposer par le(la) professeur(e), avec un vidéo-projecteur.

Simulation de fourchettes de sondages au niveau 0,95 ou encore d'intervalles de fluctuation asymptotique au seuil de 95%

Le logiciel utilisé est Xcas.

On propose un programme, qui est en fait, une fonction de trois variables :

- le nombre N de fourchettes choisi, obtenues à partir de N sondages de taille identique,
- la taille n de chacun des sondages,
- la proportion p de perles vertes dans la boîte.

Ce programme doit permettre d'obtenir la proportion, sur N sondages (de taille n), de ceux qui contiennent la valeur de p, et de visualiser graphiquement les fourchettes de sondage si l'on saisit l'instruction `DispG()` dans une ligne de commande.

```
1  Prog  Edit  Ajouter        14        nxt        OK (F9)
fluctuation(N,n,p):={
  local fp,j,k,vertes,l,nbfourch;
  // n taille echantillon, N nombre de sondages, p proportion de noires
  l:=[];
  nbfourch:=0;
  pour j de 1 jusque N faire
  vertes:=0;
    pour k de 1 jusque n faire
      si alea(0,1)<p alors vertes++; fsi;
    fpour;
    fp:=evalf(vertes/n);
    si abs(fp-p)<=1/sqrt(n) alors nbfourch++; fsi;
    l:=append(l,[evalf(fp-1/sqrt(n)),evalf(fp+1/sqrt(n))]);
    segment(fp-1/sqrt(n)+j*i,fp+1/sqrt(n)+j*i);
  fpour;
  afficher("proportion d'echantillons contenant p ="+evalf(nbfourch)/N);
  return l;
}

:;
```

Figure 1 : Simulation d'intervalles de fluctuation asymptotique au seuil de 95%

Si, pour reprendre l'exemple précédent, on simule 50 sondages de taille 900, pour une proportion $p = 0{,}4$, on obtient, par exemple (voir figure 2), 100 % de fourchettes de sondage (au niveau 0,95) qui contiennent la valeur 0,4. On visualise ces fourchettes à la figure 3.

```
2  fluctuation(900,50,0.4)
   proportion d'echantillons contenant p =1.0
   Evaluation time: 0.41
                                  0.376666666667,  0.443333333333
                                  0.352222222222,  0.418888888889
                                  0.362222222222,  0.428888888889
                                  0.357777777778,  0.424444444444
```

Figure 2 : Avec l'hypothèse que p = 0,4, on a simulé 50 sondages de taille 900, il se trouve ici que tous les intervalles de fluctuation asymptotique au seuil de 95 % contiennent p.

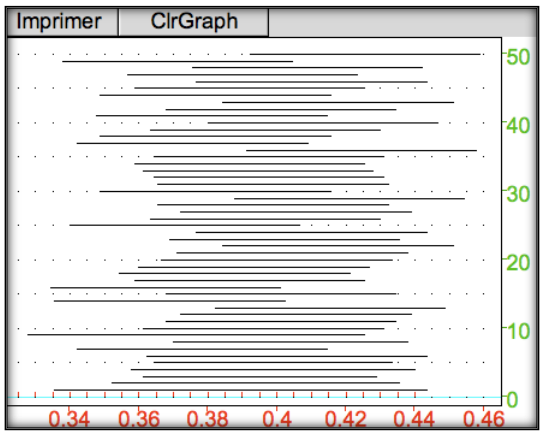

Figure 3 – Représentation des intervalles de fluctuation asymptotique au seuil de 95 % correspondant à l'expérimentation de la figure 2

Les élèves doivent pouvoir visualiser plusieurs graphiques, du type de celui de la figure 3, pour différentes valeurs de p choisies *a priori*, différentes valeurs du nombre de sondages.

Il s'agit ensuite d'amener les élèves à conclure relativement à la proportion de perles vertes dans la boîte.

Réponse à la question concernant la proportion de perles vertes dans la boîte

Ce qui précède permet de comprendre qu'avec un bon niveau de confiance (en réalité plus de 95 %), si on prend pour valeur la proportion f de perles vertes calculée à partir du « gros » échantillon, de taille n, l'intervalle de confiance $[\, f - 1/\sqrt{n}\, ;\ f + 1/\sqrt{n}\,]$ donne une bonne estimation de la proportion cherchée. L'inégalité $|f - p| \leq 1/\sqrt{n}$ peut en effet s'interpréter par un intervalle de demi-largeur $1/\sqrt{n}$ centré sur p ou par un intervalle de demi-largeur $1/\sqrt{n}$ centré sur f (on peut dire qu'il s'agit d'un changement de « point de vue »).

Cette conclusion doit rester à la charge des élèves et ne pas être précipitée.

DEUXIEME PARTIE

Le théorème central limite

Nous avons évoqué ci-dessus le théorème central limite :

> Ce théorème est au fondement de la statistique. Il explique une partie de l'incroyable efficacité de la théorie des probabilités, vue comme *théorie du hasard.*
> Il a une longue histoire, où circulent les noms de De Moivre (travaux de 1733), S.D. Poisson (1824), La place (1774 et 1810), Gauss (1809 et 1816), Lyapounov (1901), Paul Levy (1925).
> Ce théorème demande peu d'hypothèses, pour une conclusion qui ouvre la voie à de nombreuses applications. (C. Schwarz, 2006, p. 141)

Dans l'exemple pris au début de la partie précédente, nous avions un échantillon de taille $n = 32$. Nous avons considéré que nous réalisions, de manière identique et indépendante, le tirage d'une perle, dont nous notions la couleur. Ce faisant, on considère en fait un modèle où les résultats $(x_1,..,x_n)$ de ces expériences sont des réalisations de variables aléatoires $X_1,..,X_n$, indépendantes et de même loi P (dans cet exemple, P est la loi de Bernouilli de paramètre p). Mais nous considérons dans la suite une situation plus générale.

Dans ce cadre, la somme $s_n = x_1 + \ldots + x_n$ et la moyenne $\bar{x}_n = \dfrac{s_n}{n}$ des résultats sont des réalisations de variables aléatoires, notées $S_n = X_1 + \ldots + X_n$ et $\overline{X}_n = \dfrac{S_n}{n}$, respectivement.

On se place dans le cas où la loi P admet une espérance μ et un écart-type σ. L'espérance de la variable aléatoire $\overline{X}_n$ est alors aussi μ, et comme les variables $X_1,..,\ X_n$ sont indépendantes, l'écart-type de $\overline{X}_n$ est $\dfrac{\sigma}{\sqrt{n}}$.

La variable aléatoire $Z_n = \dfrac{\overline{X}_n - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ est donc centrée (soit d'espérance égale à 0) et réduite (soit d'écart-type égal à 1).

Le *théorème central limite* (à une dimension) dit que la suite (Z_n) converge en loi vers une variable de loi normale centrée réduite, $N(0,1)$, c'est-à-dire que, pour tout nombre réel t, la limite, quand n tend vers l'infini, de Prob $[Z_n \leq t]$ est égale à $\Phi(t)$, où Φ est la fonction de répartition de la loi $N(0,1)$, à savoir $\Phi(t) = \dfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-u^2/2}\, du$.

On peut faire une expérimentation informatique qui permette de visualiser cette convergence vers la loi normale.

Nous proposons divers petits programmes qui permettent au(à la) professeur(e) de faire voir à la classe de multiples résultats de simulation à partir desquels ils pourront comprendre le phénomène en jeu.

Un premier programme qui permet de simuler un sondage aléatoire de taille *n* dans une population contenant une proportion *p* d'un caractère

Ce premier programme (figure 4) permet de simuler un sondage de taille n, pour une proportion p de perles vertes dans la boîte, il renvoie une valeur approchée de la proportion de perles vertes dans l'échantillon.

Il s'agit en fait d'une fonction de deux variables :

- la taille n du sondage,
- la proportion p de perles vertes dans la boîte.

```
Prog  Edit  Ajouter
sondage(n,p):={
  local j,nbvertes;
  nbvertes:=0;
  pour j de 1 jusque n faire
    si alea(0,1)<p alors nbvertes++; fsi;
  fpour;
  return evalf(nbvertes/n);
}
:;
```

Figure 4 – Un programme qui donne le résultat d'un sondage de taille n dans le cas où la proportion de perles vertes dans la boîte est égale à p.

```
sondage(32,0.48)
                0.40625
```

Figure 5 – Résultat donné par le programme de la figure 4 pour n = 32 et p = 0,48.

Un second programme qui permet de répéter le sondage précédent de manière identique et indépendante

Ce second programme (en fait aussi, une fonction) permet d'obtenir une liste de 1000 résultats du sondage précédent, répété de manière identique et indépendante. On peut

évidemment le modifier suivant le nombre de résultats qu'on souhaite obtenir. Il s'agit d'une fonction des mêmes variables que le premier programme de la figure 4.

```
Prog  Edit  Ajouter                                    8
milleSondages(n,p):={
  local j,L;
  L:=seq(0,j,1,1000);
  pour j de 0 jusque 999 faire
    L[j]:=sondage(n,p);
  fpour;
  return L;
}
:;
```

Figure 6 – Cette fonction retourne une liste de 1000 résultats du sondage simulé à la figure 4.

Représentation des résultats obtenus grâce à ce programme

On ordonne ces résultats et on les représente dans un histogramme. Par exemple, pour $n = 32$ et $p = 0{,}48$, on obtient une liste de 1000 résultats de sondage, qu'on ordonne de manière croissante et qu'on représente à l'aide d'un histogramme, dont les classes sont, par exemple, d'amplitude 0,05. Il reste à faire varier les amplitudes des classes et observer les changements dans l'histogramme, comprendre comment sont construits les rectangle…

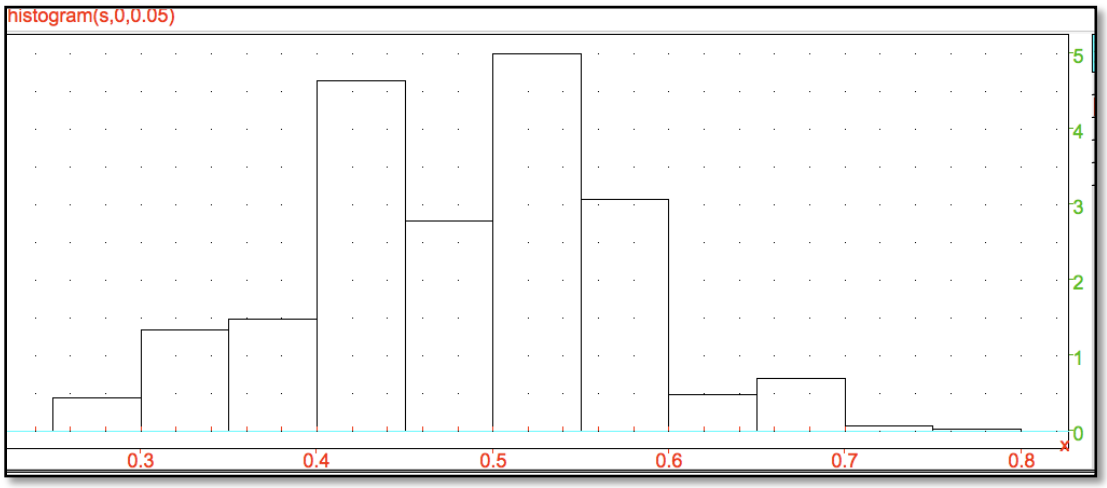

Figure 7 – Représentation par un histogramme, dont les classes ont pour amplitude 0,05, des 1000 valeurs obtenues par simulation, à l'aide du programme de la figure 6.

On peut ensuite représenter, sur le même graphique que l'histogramme des résultats sur les 1000 sondages, la densité de probabilité de la loi normale de moyenne p et d'écart-type $\sqrt{\dfrac{p(1-p)}{n}}$: dans le cas où $p = 0{,}48$ et $n = 32$, on obtient le graphique de la figure 8 suivante.

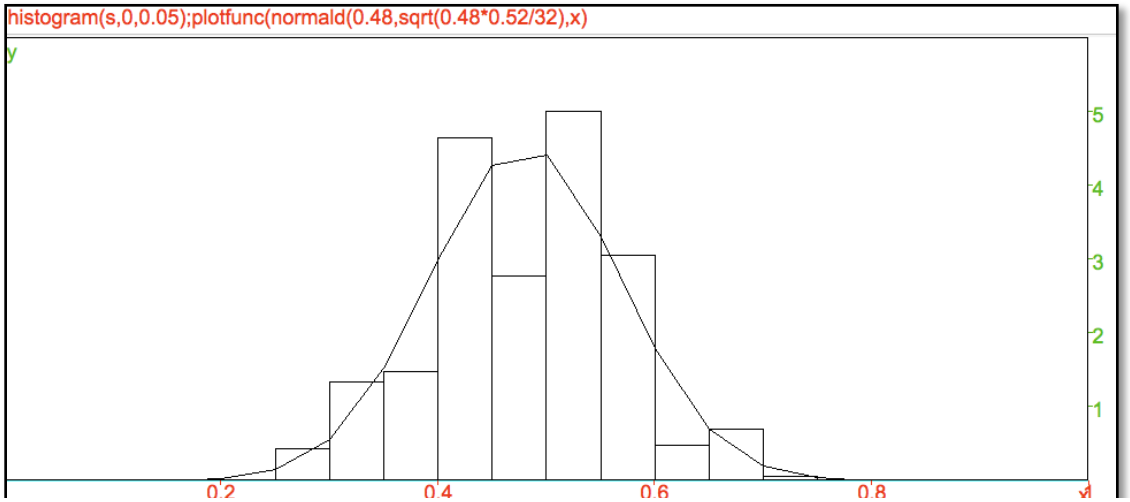

Figure 8 – On a ajouté à l'histogramme, construit à partir des 1000 résultats de sondages, la représentation graphique de la densité de probabilité de la loi normale de moyenne 0,48 et d'écart-type $\sqrt{\dfrac{0{,}48 \times 0{,}52}{32}}$.

On peut améliorer la précision du tracé de la courbe de Gauss, en précisant l'intervalle en abscisse et la discrétisation (voir figure 9).

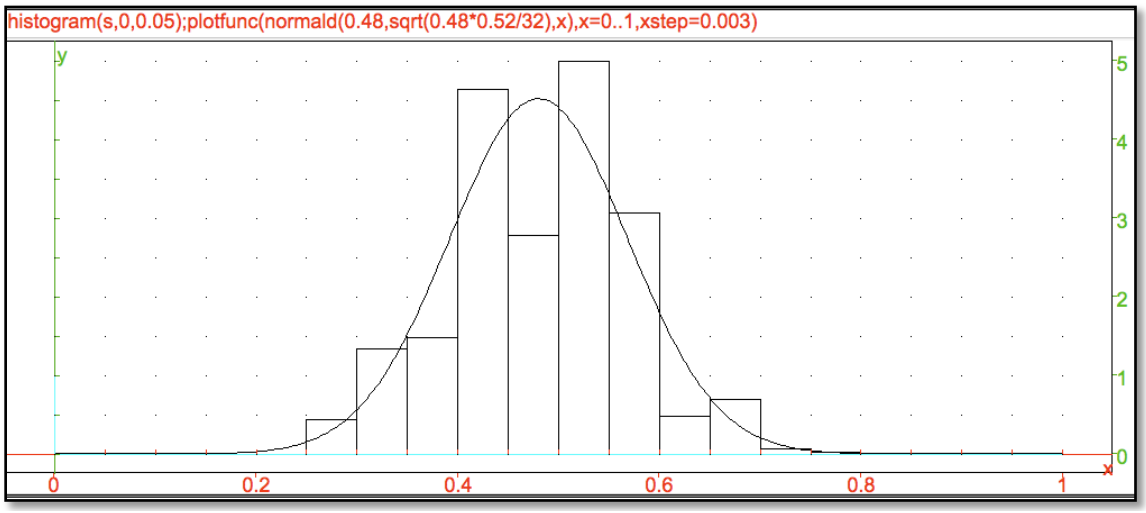

Figure 9 – Les arguments ajoutés concernant la représentation de la densité de la loi normale permettent de préciser le tracé

REFERENCES

Schwarz, C. (Dir.) (2006) *Pratiques de la statistiques. Expérimenter, modéliser et simuler.* Un livre de l'IREM de Grenoble. Paris : Vuibert.

Programmes officiels de mathématiques : B.O. spéciaux n°6 du 28 août 2008, n°8 du 13 octobre 2011.