A propos du document de ressources pour la classe de première

Statistiques et probabilités

Juillet 2011 Claudine Schwartz

Voici quelques commentaires, questions et critiques sur le « document de ressources pédagogiques pour les statistiques et probabilités en classes de Première », publié le 2 mai 2011 sur le site de l'inspection générale de mathématiques[1] puis en juin sur le site de la DGESCO[2]. Ce document a été réalisé « par une équipe sous l'égide du Groupe Mathématiques de l'IGEN ».

Le point de vue exposé ci-dessous est celui d'une statisticienne et il est marqué par diverses expériences d'enseignement. Ce texte est écrit à l'attention des professeurs de mathématiques, dans le cadre de leur propre formation, et pas spécialement comme aide pédagogique directe pour leur enseignement.

Les commentaires du document de ressources (dans la suite doc.res.) sont regroupés en rubriques indépendantes :

1- Diagrammes en boîte et comparaisons de séries de données
2-Variables aléatoires
3-Loi géométrique tronquée
4-Loi binomiale

Les pages auxquelles ce texte se réfère sont celles de la version de juin publiée sur eduscol.

1- Diagrammes en boîte et comparaisons de séries de données

► Un diagramme en boîte est un outil d'aide à l'observation des données, utile pour l'étude d'une ou plusieurs série de données[3], comme on peut le voir dans l'exemple du tableau 1 et de la figure 1.

[1] http://igmaths.infos.st/spip/spip.php?article108

[2] http://eduscol.education.fr/cid56492/ressources-pour-les-nouveaux-programmes-premiere.html

[3] Le diagramme page 4 du doc.res. est victime de soucis d'édition qui le rendent ambigu : x_{min} et x_{max} sont des valeurs de la série et 25%,50%,75% des valeurs de la fonction de répartition empirique. Cela pose des problèmes, notamment du fait de certaines représentations proposées en sciences économiques et sociales avec des déciles. A propos de déciles, il peut être utile de signaler aux enseignants de maths que pour leur collègue de sciences économiques et sociales, un intervalle inter-décile est souvent encore appelé décile.

Une éruption du geyser Old Faithfull (Yellowstone National park, USA)

Statistiques descriptives
Eclaté par : mois

	Moy.	Dév. Std	Nombre	Minimum	Maximum	Médiane	Mode
durée, Total	90,274	7,741	5822	45,000	120,000	91,000	92,000
durée, 1	90,117	7,909	495	58,000	116,000	90,000	.
durée, 2	89,484	8,256	450	45,000	118,000	90,000	91,000
durée, 3	89,536	8,900	498	56,000	110,000	91,000	90,000
durée, 4	90,496	7,660	478	57,000	120,000	91,000	94,000
durée, 5	91,139	7,515	490	59,000	115,000	91,000	90,000
durée, 6	90,432	7,419	477	47,000	109,000	91,000	91,000
durée, 7	90,054	7,765	496	57,000	118,000	91,000	92,000
durée, 8	89,736	7,754	497	52,000	112,000	90,000	92,000
durée, 9	90,081	7,335	480	58,000	116,000	90,000	89,000
durée, 10	90,180	7,325	495	52,000	111,000	91,000	92,000
durée, 11	90,554	7,216	478	57,000	114,000	91,000	.
durée, 12	91,459	7,416	488	58,000	114,000	92,000	93,000

Résumés des durées inter-éruptions, en minutes, de 5822 éruptions de 2007 : le numéro des durées correspond au mois (durée 1=janvier)[4]
La deuxième colonne du tableau fournit les écarts-types : ceux-ci seraient utiles pour savoir si les moyennes des durées sont significativement différentes[5]. Mais en dehors d'une telle question et sans connaitre la distribution des données, ils sont peu utilisables.

Tableau 1

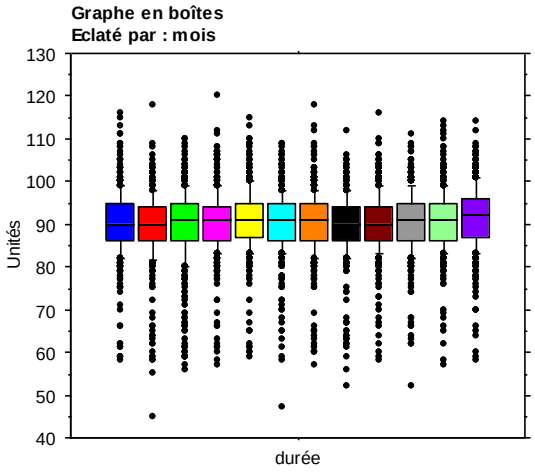


Figure 1 : En observant ces diagrammes, on peut se dire que les variations d'un mois à l'autre des durées d'éruption du geyser « Old Faithfull » aux USA, qu'elles soient significatives ou non, sont mineures.

►Un intérêt pédagogique de l'utilisation des diagrammes en boîtes, pour ce qui est de l'enseignement de la statistique par un professeur de mathématiques, est de faire prendre conscience aux élèves que « comparer des séries » :

[4] Mots clefs de recherche google : old faithfull datasets

[5] On pourrait alors, pour voir s'il y a un *effet saison* significatif, faire ce qu'on appelle une analyse de la variance (qui a pour but, comme son nom ne l'indique pas, de voir avec une seule analyse, si un ensemble de moyennes sont significativement différentes).

- ne saurait relever de la comparaison des seules moyennes. Même si ce sont justement ces moyennes qui nous intéressent, on ne peut *comparer* leurs écarts sans rapporter ceux-ci à l'ampleur de la variabilité des données et donc de la fluctuation d'échantillonnage.

- n'a de sens que dans un contexte donné. En S.V.T. et en physique, une situation fréquente est l'étude de l'influence d'un facteur donné. On se trouve alors devant deux séries de mesures correspondant à deux conditions expérimentales distinctes (présence ou absence d'un certain facteur) et on cherche à déduire de celles-ci une conclusion sur l'effet du facteur étudié. Compte tenu de l'effort fait en mathématiques sur la notion de fluctuation d'échantillonnage, il serait peu cohérent de se contenter de comparer les moyennes des deux séries, ou tout autre paramètre, sans dire qu'il convient de rapporter leur écart à celui qu'on peut attendre de la fluctuation d'échantillonnage. Il serait aussi incohérent que les différentes disciplines (S.V.T., maths, physique, sciences économiques et sociales) ne se mettent pas d'accord sur les outils statistiques à employer pour étudier l'effet d'un facteur. Etant donné qu'en mathématiques on peut introduire la notion de différence significative entre une fréquence et une probabilité de référence, on pourrait convenir d'une notion forte de différence significative définie par une intersection vide des intervalles de confiance des moyennes[6].

Le doc.res. parle de « comparaison *pertinente* de deux séries statistiques » (page 4). Etant donné les outils dont on dispose en première, en quoi consiste, à ce niveau d'études, une comparaison pertinente ? Que dit-on en classe de première à propos de l'influence d'un facteur, en mathématiques, SVT et physique ?

►Enfin, notons que les diagrammes en boîte permettent de visualiser une dissymétrie des données autour de la médiane : en tel cas, la moyenne est différente de la médiane et les deux paramètres sont utiles. Avec les moyens informatiques modernes, on calcule toujours ces deux quantités, tout en s'interrogeant sur leur sens éventuel tant au niveau de la réponse à la question qui a motivé le recueil des données qu'au niveau de la communication sur ces données[7].

2-Variables aléatoires

► Dans la plus grande partie du doc.res., on parle de variables aléatoires comme descripteurs du résultat d'une expérience aléatoire. Par exemple, pour dire qu'on choisit au hasard un nombre entier entre 1 et K, on parle d'une variable aléatoire X dont les valeurs possibles sont les entiers de 1 à K et telle que $P(X=i) = 1/K$ pour $i = 1...K$. La notation X et le terme *variable* aléatoire sont cohérents avec le fait de choisir une *variable* dans un ensemble selon une certaine loi P. Cette conception de la notion de variable aléatoire est historiquement fondée, efficace et simple à appréhender.

[6] S'il s'agit d'une expérience de Bernoulli, ces moyennes sont égales aux fréquences.

[7] Voir http://www.statistix.fr/spip.php?breve17

Dans le cadre de la classe de première, on ne considère souvent qu'une unique réalisation d'une unique expérience aléatoire. La loi P dont on parle est alors celle de X, l'ensemble sur lequel elle est définie est celui des valeurs que peut prendre X. Mais en statistique et en probabilité, le plus souvent, quand on écrit $P(X=i)$, la loi (ou mesure[8]) de probabilité P n'est pas définie sur l'ensemble des valeurs prises par X et n'est pas la loi de X. Pour fixer les idées, on peut avoir en tête le modèle d'échantillonnage. Si on fait n expériences identiques et indépendantes, dont les issues sont codées par les entiers de 1 à K, l'espace Ω associé à cette situation est $\{1,....,K\}^n$ muni du produit des lois (identiques) modélisant chaque expérience. La variable X_i qui donne le résultat de l'expérience i est la i-ème projection de Ω dans $\{1,...,K\}$ (c'est-à-dire la variable qui à un élément de Ω fait correspondre sa i-ème composante). On glisse ainsi de la notion intuitive de *variable* aléatoire, qui permet de parler du choix d'un élément variable d'un ensemble, à la notion formalisée d'*application* d'un ensemble Ω muni d'une loi de probabilité P dans un autre ensemble E sur lequel opérera la loi de probabilité P_X image de P par X. On a gardé dans la terminologie la trace de l'histoire de la notion de variable aléatoire : aujourd'hui, ce qu'on appelle une *variable* aléatoire est une *application* d'une espace probabilisé dans un ensemble.[9]

Dans le doc.res., on considère des espaces produits sans les nommer explicitement mais en les représentant par des arbres (un élément de Ω est alors un chemin de l'arbre) ; le poids de chaque arête permet de définir par produits la loi P sur l'ensemble Ω : c'est un choix pédagogique classique et qui facilite la compréhension.

On emploie souvent le mot univers dans l'enseignement secondaire : est-ce bien nécessaire ? Il n'est pas toujours clair de savoir si ce terme désigne $\{1,....,K\}^n$ ou $\{1,...,K\}$, s'il s'agit de l'ensemble des issues observables d'une expérience, de plusieurs expériences, ou d'un ensemble plus large comme nous allons le voir avec le lancer de deux dés.

► Il est aussi utile de prendre conscience qu'une fonction Z d'une ou plusieurs variables aléatoires est une nouvelle variable aléatoire. Ainsi, si (X,Y) est un couple de variables donnant le résultat de deux lancers de dés, $Z=X+Y$ est une nouvelle variable aléatoire qui en donne la somme. Dire que les dés sont équilibrés (c.a.d. que les lois de X et Y sont équiréparties), et que les lancers sont indépendants, (c.a.d. que la probabilité de $(X, Y)=(i,j)$ est 1/36), détermine entièrement la loi de (X,Y) et donc par calcul celle de Z. On n'a plus aucun choix de modélisation à faire pour la loi de Z, ce qui n'est pas évident pour les élèves ! Cela n'empêche pas de faire des simulations pour voir l'allure de cette loi et l'approximer.

► Revenons précisément sur le lancer de deux dés. Le choix de l'espace Ω est dicté par le fait qu'on parle de deux lancers identiques et indépendants : on prend $\Omega = \{1,....,6\}^2$. A ce niveau peu importe qu'on soit capable de distinguer tous les éléments de Ω : il n'a jamais été imposé que tous les résultats possibles de l'expérience soient effectivement observables avec

[8] Je n'ai pas vraiment repéré quelle était la terminologie choisie par les nouveaux programmes ni s'il y a changement par rapport aux anciens programmes.

[9] Pour unifier au plan de la théorie, l'aspect *variable* et l'aspect *application* dans le cas d'une unique expérience telle un lancer de dé, on peut dire que dans ce cas X est l'application *identité*.

les moyens techniques dont on dispose. Si on rajoute que les dés sont indiscernables on est amené à s'intéresser à une nouvelle variable Z qui vaut $\{i,j\}$ si $(X,Y)=(i,j)$ ou $(X,Y)=(j,i)$: ce n'est donc pas simple pour des élèves débutants en calcul des probabilités. La loi de Z n'est pas équirépartie. Les dés indiscernables, c'est un fleuron de l'enseignement français. Ils permettent l'étude de jolis problèmes historiques, mais l'importance qu'il convient de leur donner aujourd'hui n'est pas claire : de quelle compréhension et de quel apprentissage (que ce soit d'un savoir ou de procédures) sont-il les vecteurs et comment cela s'intègre dans l'ensemble du cursus ?

Enfin, on parle page 5 du doc.res. du lancer d'un dé « supposé équilibré ». Convenons de dire que le terme « dé équilibré », parle du modèle associé au lancer de dé et dit que ce modèle est la loi équirépartie. Cela peut être démotivant pour l'élève de « supposer » le dé équilibré (en maths, on ne peut *rien faire normalement*…), surtout si l'enseignant a du mal à répondre à la question : comment sait-on s'il est équilibré ? Le dé est un outil très ancien de simulation de la loi équirépartie : pourquoi mettre en doute d'entrée sa raison d'être ?

2- La loi géométrique tronquée

Dans le document ressource, l'approche de la loi géométrique à partir de la probabilité de désintégration d'un atome par unité de temps pose problème. On y définit une variable aléatoire X qui vaut k si l'atome s'est désintégré à la k-ème unité de temps avec $1 \leq k \leq 100$ et 0 si l'atome ne s'est pas désintégré en 100 unités de temps : cette variable aléatoire n'est donc pas la durée de vie de l'atome. Dès lors, on voit mal quel sens donner à son espérance et donc pourquoi on la calculerait. C'est comme si dans une pyramide des âges tronquée à 100 ans, on affectait à chaque individu ayant moins de 100 ans son âge et la valeur 0 aux plus que centenaires !

Pour ce qui concerne la radioactivité, le paramètre que les élèves doivent connaître est le « temps de demi-vie » qui au niveau d'un atome correspond à la médiane de durée de vie et non la moyenne. Instrumenter certains sujets au service de la pédagogie est inévitable, mais est-il indispensable de prendre un sujet aussi riche, de le dévier de son traitement standard (ici le calcul de la médiane) ? Heureusement, l'exercice sur la radioactivité page 71 a du sens car il s'agit d'une question qui se pose effectivement en pratique.

Dans le doc.res. on définit plus généralement une variable aléatoire de loi géométrique tronquée à n (page 14) comme une variable aléatoire X prenant ses valeurs dans $\{0,1,...,n\}$ et telle que $P(X=i)=(1-p)^{n-1}p$ pour $i \neq 0$ et $P(X=0)=(1-p)^n$. En fait, le cadre choisi dans le doc.res. est celui d'une variable aléatoire qui prend ses valeurs dans l'ensemble $\{1,..., n, a\}$, où a code la situation où on n'a pas eu de 1 en n expériences. Cette variable n'est pas nécessairement numérique et n'a donc pas nécessairement d'espérance. On peut ensuite définir une nouvelle variable Y fonction de X, telle que $Y=X$ si $X=1…n$, la valeur de Y pour $X=a$ dépendant du contexte (on peut imaginer prendre la valeur n si on s'intéresse au nombre de répétitions de l'expérience tronquée : l'espérance de Y est alors la moyenne théorique du nombre de répétitions faites). Passer de X à Y serait un formalisme bien inutile avec les élèves, aussi toutes les lois de telles variables Y pourraient être considérées comme des lois géométriques tronquées. Mais convenir de prendre systématiquement $a=0$ obscurcit la compréhension, même dans les cas où la valeur numérique de l'espérance dépend peu du choix de la valeur affectée à a.

Une note de bas de page du doc. res., page 14, donne une explication peu convaincante du choix systématique a=0. On y lit que si X et Y suivent des lois tronquées en n et n', avec ce choix de la valeur 0, les probabilités $P(X=k)$ et $P(Y=k)$ sont égales pour k=l ....min(n,n'). Mais il en est de même si on remplace 0 par toute autre valeur, que ce soit un nombre ou non. Cette note de bas de page reflète sans doute le fait qu'une définition plus classique d'une loi tronquée à un ensemble A consiste à prendre la loi conditionnée par l'événement A : dans ce cadre, la loi géométrique tronquée à n est la loi de X sachant que $X \leq n$. Dès lors, si X et Y suivent des lois tronquées par conditionnement en n et n', on n'a plus $P(X=k)=P(Y=k)$ pour k=l ....min(n,n'). Conditionner n'est pas au programme donc cette définition (classique) ne peut être envisagée. D'où l'idée d'une loi sur {1,..., n, a}.

4- Loi binomiale

► Il ne semble pas nécessaire d'impliquer la notion de succès ou d'échec dans la définition de la loi binomiale (encadré des pages 25 et 27 du doc.res.). Ainsi, si on considère une naissance, quel sexe sera considéré comme un succès ? La désintégration d'un atome, est-ce un succès ou un échec ? On peut ne pas parler de succès et échec et dire simplement que :

> Une expérience à deux issues est appelée épreuve de Bernoulli. Si on code les deux issues par 0 et 1, la loi de Bernoulli de paramètre p est celle d'une variable aléatoire prenant les valeurs 0 et 1 avec $P(X=1)=p$. Une telle variable aléatoire est aussi appelée variable de Bernoulli.

►Le terme de schéma de Bernoulli (page 25) est un grand classique au niveau de l'enseignement secondaire français, mais qu'est-ce que cela désigne ? Un protocole expérimental, un arbre de probabilité, l'ensemble $\{0,1\}^n$ muni des produits des lois de Bernoulli ? On pourrait parler d'échantillons de lois de Bernoulli, associés à la répétition de n épreuves de Bernoulli identiques (même valeur de p) et indépendantes. La loi binomiale est alors celle qui donne le nombre de 1 dans un échantillon de Bernoulli, c'est la loi de la somme de toutes les variables de Bernoulli en jeu. On calcule aisément l'espérance d'une loi de Bernoulli qui vaut p, et avec l'idée intuitive de linéarité de la moyenne, on peut conjecturer que l'espérance d'une variable X de loi binomiale $B(n,p)$ est np.

► Les probabilités fournissent de nombreuses occasions de mettre en œuvre des simulations, ne serait-ce que pour étudier la somme de 2 dés, 3 dés ou plus. Il y a un équilibre délicat à trouver a propos des simulations : si certaines doivent être programmées par les élèves eux-mêmes, il serait regrettable de ne pas apprendre à travailler avec des animations toutes faites, qui permettent de traiter des questions complexes au niveau des élèves [10] !

Des exercices de simulation sont proposés dans le doc. res. pour illustrer que la variance d'une variable aléatoire de loi binomiale $B(n,p)$ est une fonction linéaire de n, mais pas de p. Mais d'une part, dire que la variance est une fonction linéaire de n n'est pas forcément heureux : c'est l'écart-type et la racine de n qu'il convient d'avoir comme image mentale (celle-ci sera confortée dans des études ultérieures par le théorème central limite). D'autre part, même si le programme recommande de mettre en œuvre des simulations, avoir pour objectif de dire des choses sur un paramètre, ici la variance, dont les élèves ne voient pas bien

[10] On trouvera sur les sites suivants de nombreuses simulations utilisables en classe : http://jpq.pagesperso-orange.fr/proba/index.htm et http://www.statistix.fr/spip.php?article56

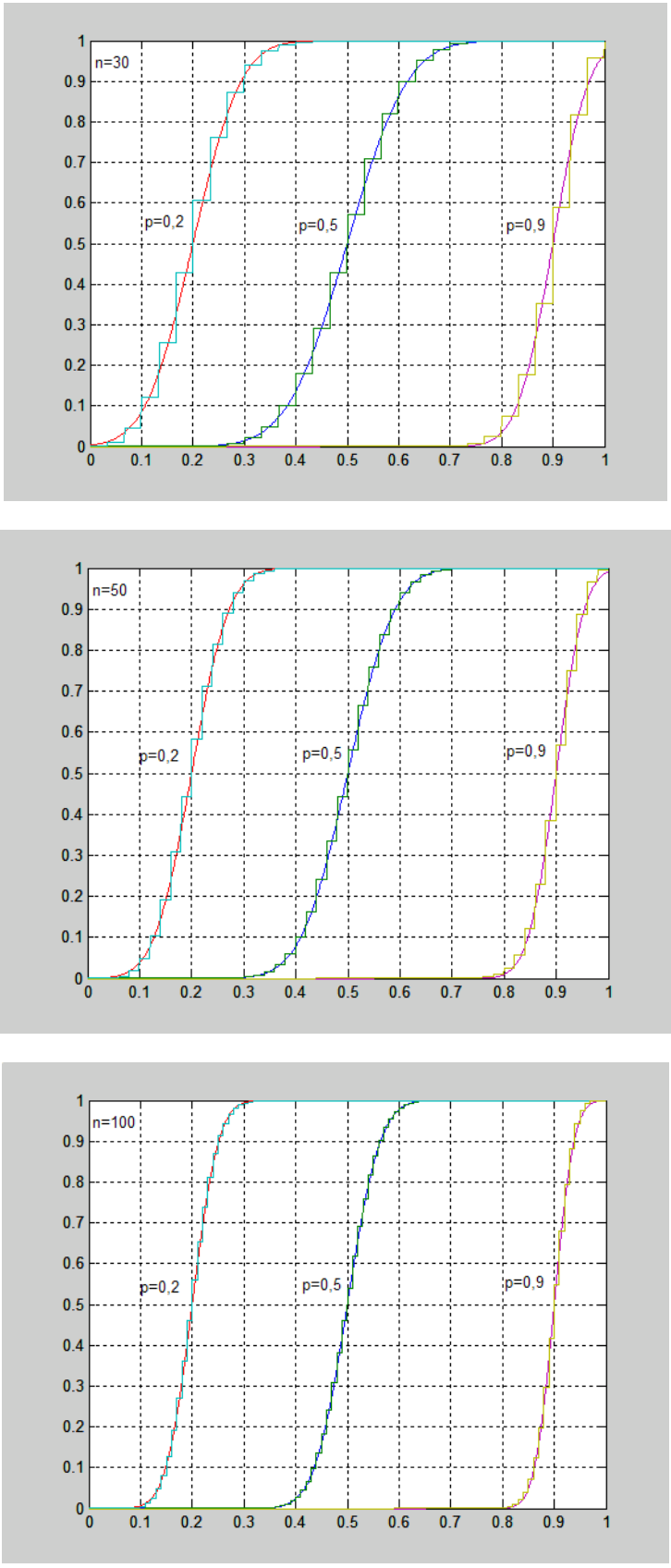
à quoi cela sert[11], est peut être inutilement chronophage. On pourrait calculer la variance d'une variable aléatoire de loi de Bernoulli (et montrer que cette variance est maximum pour p=1/2). Puis on pourrait directement admettre qu'avec des expériences indépendantes (ici expériences de Bernoulli), la variance de la somme des variables aléatoires associées est la somme de leurs variances . On pourrait ensuite faire un lien avec les intervalles de fluctuation étudiés en classe de seconde.

Pendant longtemps, les probabilités ont été au service de la combinatoire : soyons vigilants à ce que la statistique ne soit pas au service de l'apprentissage de l'algorithmique.

► Le lien avec les intervalles de confiance est abordé page 38 du doc.res. avec des notations déroutantes puisqu'on voit écrit page 38 un intervalle de fluctuation associé à un échantillon de taille n sous la forme [a/n, b/n] ; c'est une écriture qui choque la vue car elle ne laisse pas transparaitre que a dépend de n et plus précisément quasiment linéairement de $\sqrt{n}$. Evidemment, on y revient page 39 pour faire le lien avec les formules vues en seconde, mais c'est pour faire disparaitre la notation $1/\sqrt{n}$ au profit du $1/n$ dans les pages suivantes (pages 40,41, etc.).

► Enfin, en complément des annexes du doc.res., notons qu'on peut travailler sur les représentations graphiques des fonctions de répartition de lois binomiales et celles de lois normales approximantes. Elles sont particulièrement parlantes (cf. les figures 2 ci-dessous) et rendent moins ésotérique pour les professeurs l'approximation des intervalles de confiance ou de fluctuation. N'oublions pas que le théorème central limite se formule avec des convergences de fonctions de répartition.

[11] Y a-t-il un lien au niveau des programmes entre écart-type et présentation des mesures en physique et en SVT ?



Figures 2 : Les courbes en escalier sont les courbes représentatives des fonctions de répartition des fréquences empiriques X/n où X suit une loi $B(n,p)$ pour p=0,2, 0,5 ,0,9, n=30 ,50 et 100. Les courbes lisses sont les fonctions de répartition des approximations gaussiennes.

En guise de conclusion

►Le document de ressources est le document officiel. Il va être utilisé dans les académies au cours des séances de travail sur les nouveaux programmes. Il s'agira dans ce cadre d'aider les professeurs à préparer leur enseignement. Peut-on pour autant parler de formation ?

La possibilité qu'une partie des enseignants bénéficie d'une formation réelle, sans le souci constant et lancinant de savoir « que dire aux élèves » et avec un suivi sur un temps long (les moyens informatiques actuels le permettent) n'est, à ma connaissance, pas à l'ordre du jour.

Les institutions et associations en place peinent actuellement à s'inscrire dans une dynamique pouvant mener à des conceptions de formations continues des professeurs radicalement autres que celles mises en œuvre ces dernières décennies. Le manque de moyens n'est pas la seule raison qui bloque aujourd'hui la faculté d'anticiper. Les nouveaux modèles de formation continue conduiront à des changements culturels importants au niveau de l'éducation : c'est à ce niveau qu'opèrent les résistances majeures.

►Au delà des programmes, où pourrait aujourd'hui se développer une vraie réflexion sur la répartition de l'enseignement de la statistique entre différentes disciplines ? Les structures institutionnelles et associatives restent sur des actions ponctuelles, qui ne forment pas un tout cohérent. L'injonction faite aux professeurs de collaborer entre disciplines n'est pas nouvelle, mais hélas, aucune vision ne vient l'étayer.

Cependant, la porte est ouverte à de nouveaux chantiers…..