Thème 1 du colloque CORFEM 2021

Atelier 1-D

Focus sur le raisonnement inductif correspondant à l'approche empirique de l'intervalle de fluctuation

Intervention finale de Jannick Trunkenwald

Lycée International Alexandre Dumas (Alger)

Thème : Focus sur le raisonnement inductif correspondant à l'approche empirique de l'intervalle de fluctuation

- Depuis les années 2000, les évolutions institutionnelles en France accordent une grande importance à l'approche fréquentiste de la notion de probabilité.

➢ Ce principe est lié à la loi des grands nombres

➢ Abordé en classe dès le collège, et ceci en l'absence de tout référentiel théorique qui permettrait d'en donner une justification formelle…

➢ Cette approche empirique consiste à reproduire à l'identique une grande quantité d'expériences, puis à exploiter les données, ce qui met en relation le domaine des probabilités avec celui des statistiques.

➢ La nature répétitive de ce travail et son coût en manipulations, motive une automatisation des protocoles expérimentaux par des simulations informatiques (tableur ou éditeur d'algorithme).

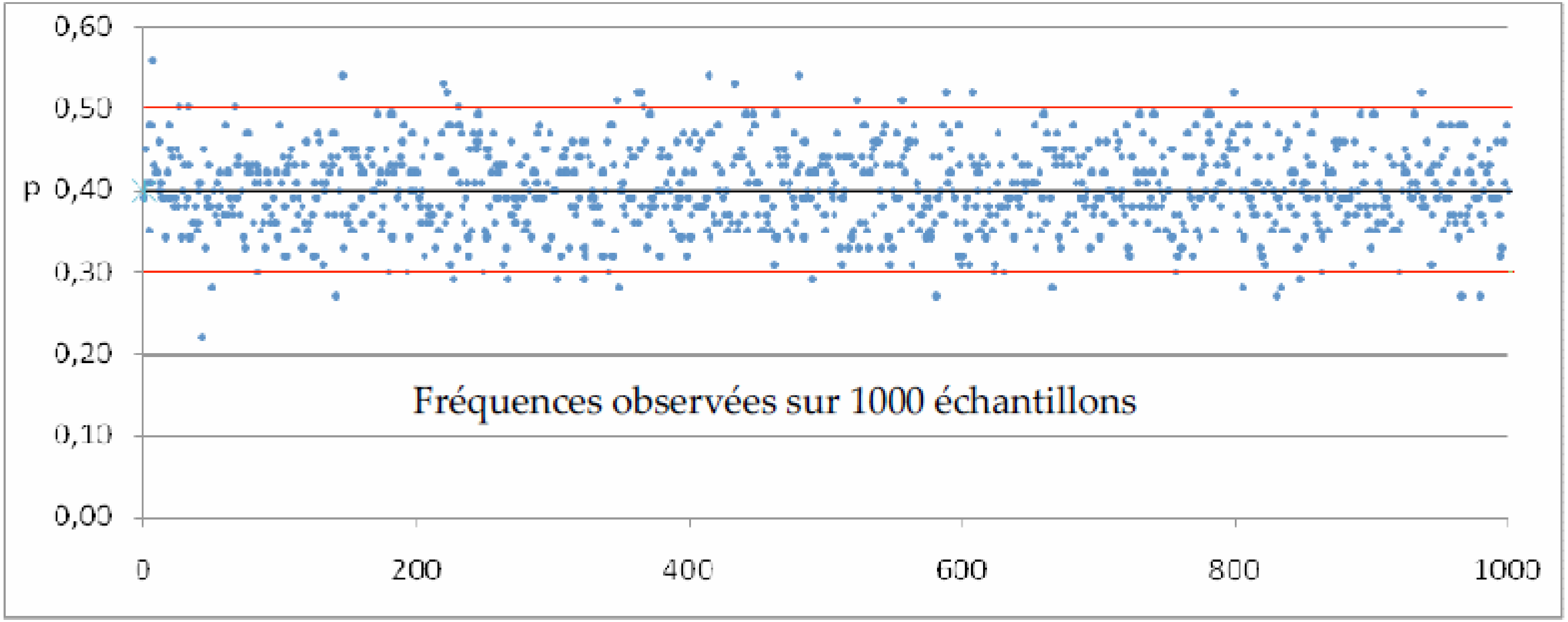
➢ En classe de 2nde, le travail de simulation permettant d'observer la fluctuation d'échantillonnage se trouve au cœur de cette problématique. Avec un enjeu lié à l'apprentissage ultérieur de la loi binomiale, puis de son approximation par la loi normale…

Textes officiels

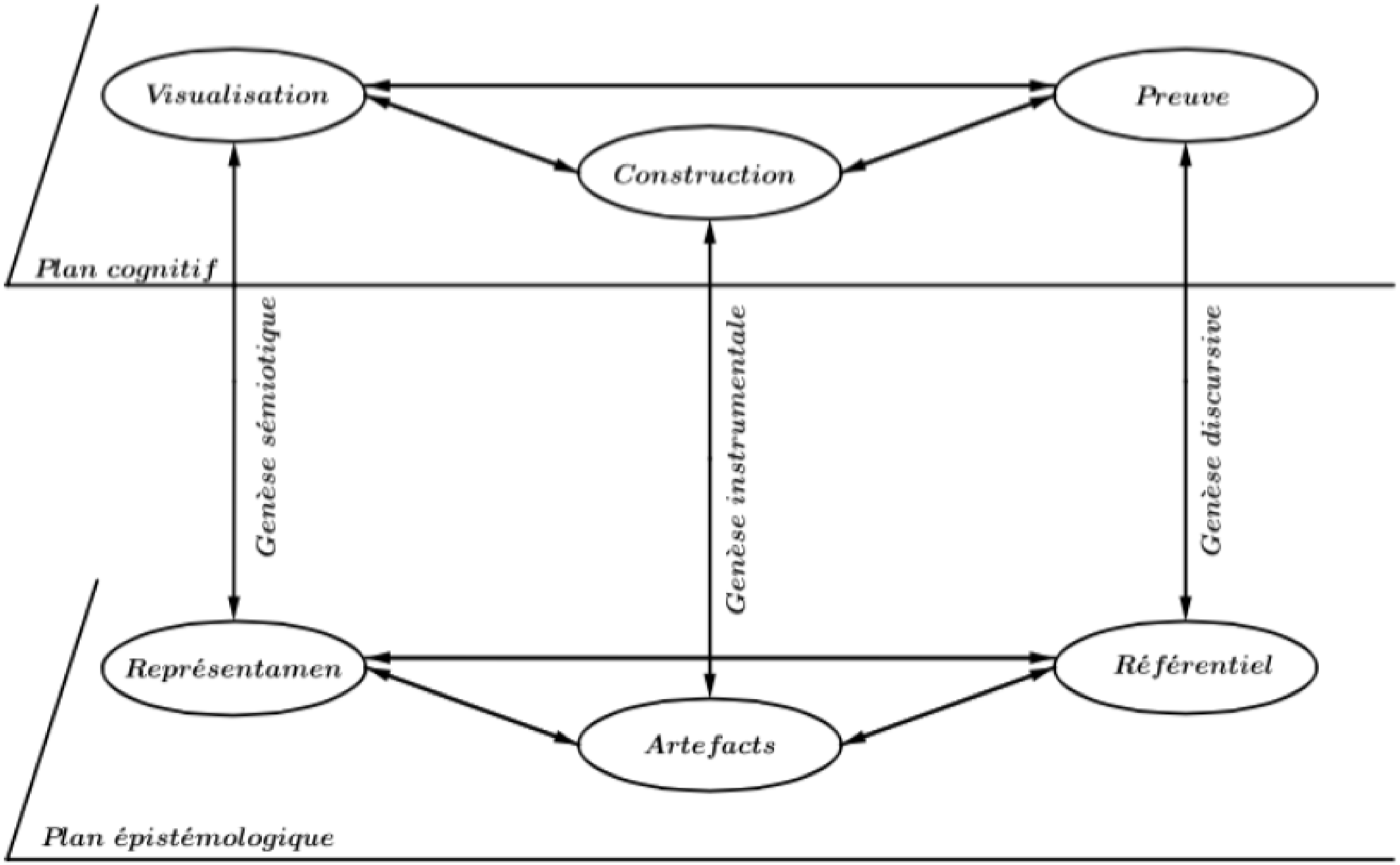
- « (…) en seconde, l'élève constatera expérimentalement qu'entre deux échantillons, de même taille ou non, les distributions de fréquences fluctuent (…) On observera aussi que l'ampleur des fluctuations des distributions de fréquences calculées sur des échantillons de taille n diminue lorsque n augmente. » Le choix pédagogique est ici d'aller de l'observation vers la conceptualisation et non d'introduire d'abord le langage probabiliste pour constater ensuite que tout se passe comme le prévoit cette théorie. (…) (**Mathématiques classe de seconde applicable à la rentrée 2000, CNDP, 2001**).
- « On peut, par expérimentation et simulation, faire observer aux élèves que les échantillons de taille n obtenus à partir d'un modèle de Bernoulli ont, pour environ 95% d'entre eux, des fréquences d'apparition du nombre 1 qui fluctuent dans un intervalle centré en p et d'amplitude $\frac{2}{\sqrt{n}}$. (…) Pour p donné, on peut faire calculer les bornes de cet intervalle pour quelques valeurs de n, et remarquer qu'il faut multiplier la taille de l'échantillon par k^2 pour diviser par k l'amplitude de l'intervalle. » (**Ressources pour la classe de seconde, Probabilités et statistiques, Juin 2009, Eduscol**)

Dernière évolution du programme en TS

- En liaison avec la partie « Algorithmique et programmation », on définit la notion d'échantillon. L'objectif est de faire percevoir, sous une forme expérimentale, la loi des grands nombres, la fluctuation d'échantillonnage et le principe de l'estimation d'une probabilité par une fréquence observée sur un échantillon.
- ***Connaissances***
- Échantillon aléatoire de taille n pour une expérience à deux issues.
- Version vulgarisée de la loi des grands nombres : « Lorsque n est grand, sauf exception, la fréquence observée est proche de la probabilité. »
- Principe de l'estimation d'une probabilité, ou d'une proportion dans une population, par une fréquence observée sur un échantillon.
- ***Expérimentations***
- Lire et comprendre une fonction Python renvoyant le nombre ou la fréquence de succès dans un échantillon de taille n pour une expérience aléatoire à deux issues.
- Observer la loi des grands nombres à l'aide d'une simulation sur Python ou tableur.
- Simuler N échantillons de taille n d'une expérience aléatoire à deux issues. Si p est la probabilité d'une issue et f sa fréquence observée dans un échantillon, calculer la proportion des cas où l'écart entre p et f est inférieur ou égal à $1/\sqrt{n}$.



1000 échantillons de taille 100 d'un modèle de Bernoulli avec p=0,4
(Probabilités et statistiques, 2009, Eduscol)



L’Espace de Travail Mathématique (Kuzniak, 2011)

« Aussi il faut bien comprendre que la proposition de structuration de l'ETM qui va suivre passe obligatoirement par une instanciation dans un domaine mathématique déterminé. Autrement dit, plus que la notion générale d'ETM, toute étude didactique suppose la description d'un Espace de Travail pour le domaine abordé. Le cadre des ETM se présente comme une coquille méthodologique sur laquelle il sera possible de s'appuyer pour développer de nouveaux Espaces de Travail spécifiques. »

Kuzniak 2011

Réflexion portant sur le simulation

- *« La simulation informatique peut jouer un rôle important dans l'acquisition de la notion de modèle probabiliste. En effet, la comparaison des procédures de simulation associées à diverses expériences aléatoires, ainsi que des tableaux qu'elles produisent, peut faire apparaître des similitudes conduisant à considérer comme légitime la substitution d'une expérience à une autre et à dégager l'idée d'un schéma d'expérience commun, sur lequel pourra s'élaborer la notion de modèle. » (Parzysz 2011 p.138)*
- *« A propos de simulation, le point essentiel est que ce qui est simulé est un modèle probabiliste, ce qui pose théoriquement problème au début de l'enseignement, avant la mise en place de cette notion. Cette difficulté peut néanmoins être contournée si l'on recourt à des simulations qui présentent une congruence sémantique avec l'expérience réelle et en explicitant les hypothèses probabilistes sous-jacentes. » (Parzysz 2011 p.137)*

Tutor :	Donc groupe B.
Student 1 :	Oui. Donc là en fait on a surtout envie de compter…
Student2 :	On a envie de comprendre.
Student1 :	… Le compteur moi je veux qu'il compte. Donc c'est la position…
Tutor :	D'accord. Donc comment on peut faire un compteur ? C'est… Donc il faut l'initialiser…
Student 1 :	D'accord. L'initialiser à combien ?
Tutor :	Au départ de 0, puisqu'il n'y a aucun… mais…
Student 1 :	Et quand est-ce que vous l'augmentez ?
Tutor :	C'est ça ! C'est la position où le mettre !... Est-ce
Student 1 :	que…
Tutor :	Il va compter quoi ?
Student 2 :	Il va compter le nombre de cas favorables…
Tutor :	Voilà, donc à côté du « oui » vous l'augmentez de 1.
Student1 :	… avant ou après ? …
Tutor :	Avant ou après. Au même niveau.
Student 2 :	On l'a mis ici… Ca n'a pas marché.
Tutor :	Vous mettez compteur devient compteur + 1.
Student 1 :	On l'a mis, mais ça n'a pas…
Tutor :	Où est votre compteur ?
Student 1 :	n
Tutor :	Ah non, n c'est pas le compteur, c'est le numéro de l'étape.
Student 1 :	Oui c'est ça !
Tutor :	Vous devez ajouter une autre variable qui, elle, va compter les succès. n c'est autre chose… n c'est le numéro de votre étape parmi les 100 étapes.
Student 1 :	Donc il nous faut deux… Donc on aura deux compteurs.
Tutor :	Il y a un compteur de l'étape en cours et un compteur du nombre de succès.
Student 1 :	Voilà ! Mais le compteur du nombre de succès c'est à la fin de la boucle ? Et après le sinon…
Tutor :	Il prend sa valeur à la fin de la boucle. Quand vous êtes sortis de la boucle.
Student 1 :	Voilà donc c'est après le « sinon ».
Tutor :	La fin de la boucle c'est pas dans le « sinon ». La fin de la boucle c'est après le « Fin pour ».
Student 1 :	Ah ! D'accord voilà !
Tutor :	C'est là que vous savez combien il y a eu de succès à la fin.

Extrait à partir d'un enregistrement de 8min14s en formation (LIAD, Alger, 2018). Implémentation du compteur de succès dans l'algorithme de simulation d'un calcul de fréquence.

SIMULATION ALGORITHMIQUE D’UNE FLUCTUATION D’ECHANTILLONNAGE

(classe de seconde dans le système français)

```
DEBUT_ALGORITHME
  POUR k ALLANT_DE 1 A 200
    DEBUT_POUR
    n PREND_LA_VALEUR 0
    POUR i ALLANT_DE 1 A 400
      DEBUT_POUR
      SI Alea(1,6)+Alea(1,6) ≥ 9 ALORS
        DEBUT_SI
        n PREND_LA_VALEUR n+1
        FIN_SI
      FIN_POUR
    f PREND_LA_VALEUR n/400
    TRACER_POINT (k,f)
    FIN_POUR
FIN_ALGORITHME
```

Le travail permettant de mettre en évidence la fluctuation d'échantillonnage peut s'appuyer sur la conception d'un algorithme ou l'utilisation du tableur et peut se décrire en deux temps :

- 1ère étape : Simuler le calcul d'une fréquence du nombre de succès à partir d'un certain nombre de répétition de l'expérience aléatoire.

	A	B	C	D
1	Premier point	Deuxième point	Distance	Supérieure à 1/2
2	0,208735038	0,078092793	0,130642245	0
3	0,73001426	0,059318148	0,670696112	1
4	0,014553273	0,973227772	0,9586745	1
5	0,895090907	0,509455168	0,385635739	0
6	0,757449513	0,921324623	0,163875111	0
7	0,582026286	0,373742788	0,208283498	0
8	0,772322473	0,608934617	0,163387856	0
9	0,356394902	0,887696564	0,531301662	1
10	0,401050985	0,759381244	0,358330259	0
11	0,716303478	0,97221226	0,255908783	0
12	0,942323555	0,566686222	0,375637333	0
13	0,12115469	0,318208569	0,197053879	0
14	0,723069047	0,157181089	0,565887958	1
15	0,113045272	0,902286994	0,789241722	1
16	0,685361589	0,27421229	0,411149299	0
17	0,784117151	0,379348643	0,404768508	0
18	0,711982266	0,049246903	0,662735363	1
19	0,485925993	0,902264194	0,416338201	0
20	0,377150075	0,821379848	0,444229773	0
21	0,48132643	0,312372995	0,168953435	0
22			Nombre de succès :	6
23			Fréquence des succès :	0,3

Saisie dans C2 : =Abs(A2-B2)
Saisie dans D2 : =SI(C2>0,5 ;1 ;0)
Manipulation de la souris reproduisant 20 lignes.
Saisie dans D22 : = NB.SI(D2:D21 ;1)
Saisie dans E22 : = D22/20

ALGORITHME SUR ALGOBOX :

```
n PREND_LA_VALEUR 0
POUR i ALLANT_DE 1 A 400
  DEBUT_POUR
  SI Alea(1,6)+Alea(1,6) ≥ 9 ALORS
    DEBUT_SI
    n PREND_LA_VALEUR n+1
    FIN_SI
  FIN_POUR
f PREND_LA_VALEUR n/400
```

Figure 1 : Modèle numérique simulant une fréquence au tableur et avec un algorithme (à gauche 20 tests sur Excel et à droite 400 tests sur Algobox)

- $2^{\text{ème}}$ étape : Construire un nuage de points représentant des fréquences ayant été obtenues en répétant plusieurs fois le processus de la $1^{\text{ère}}$ étape.

	A	B	C	D	E	F	G	H	I	J
1	Distance	>1/2	Distance	>1/2	Distance	>1/2	Distance	>1/2	Distance	>1/2
2	0,51211143	1	0,25709193	0	0,29944848	0	0,01048671	0	0,42530763	0
3	0,13405479	0	0,04771116	0	0,09079839	0	0,23554753	0	0,19504677	0
4	0,13886784	0	0,49284742	0	0,4911005	0	0,54050838	1	0,46093184	0
5	0,22690971	0	0,08530797	0	0,2531011	0	0,18433491	0	0,16975288	0
6	0,71988493	1	0,70536943	1	0,23533981	0	0,3164412	0	0,39111895	0
7	0,28451633	0	0,13285892	0	0,84301412	1	0,44460198	0	0,09206928	0
8	0,15390678	0	0,50058597	1	0,10695765	0	0,21756739	0	0,48420237	0
9	0,42546995	0	0,09006054	0	0,68635528	1	0,85519483	1	0,40148311	0
10	0,48057345	0	0,27474114	0	0,45169643	0	0,07937586	0	0,37353797	0
11	0,40768676	0	0,84492537	1	0,0691507	0	0,0832931	0	0,09009232	0
12	0,12933542	0	0,22956432	0	0,26411776	0	0,20874249	0	0,69646577	1
13	0,61282426	1	0,14614335	0	0,13941917	0	0,23871152	0	0,53092432	1
14	0,14139885	0	0,31869759	0	0,36389749	0	0,28126304	0	0,54423939	1
15	0,34501509	0	0,58903816	1	0,12642583	0	0,6584547	1	0,10516786	0
16	0,32060273	0	0,93502787	1	0,15011165	0	0,65189501	1	0,48701613	0
17	0,09747428	0	0,6815059	1	0,62686008	1	0,64266412	1	0,7635539	1
18	0,26836883	0	0,78167406	1	0,08609878	0	0,16860636	0	0,71526356	1
19	0,29118337	0	0,33720179	0	0,12888791	0	0,65099025	1	0,09103373	0
20	0,29820522	0	0,09542265	0	0,68186754	1	0,79259376	1	0,73805338	1
21	0,24577506	0	0,53873812	1	0,07612875	0	0,06291268	0	0,15602601	0
22	0,31981463	0	0,01754369	0	0,01852103	0	0,3464028	0	0,0777697	0
23	0,31183502	0	0,21831861	0	0,88945413	1	0,75406364	1	0,52781333	1
24	0,43850649	0	0,65691631	1	0,00688123	0	0,33653346	0	0,30633002	0
25	0,25448835	0	0,47551415	0	0,09922113	0	0,89889202	1	0,44485212	0
26	Nb succès :	3	Nb succès :	9	Nb succès :	5	Nb succès :	9	Nb succès :	7
27	Fréquence :	0,12	Fréquence :	0,36	Fréquence :	0,2	Fréquence :	0,36	Fréquence :	0,28

FEUILLE DE CALCUL SUR EXCEL :
Simple manipulation à la souris pour reproduire la feuille de calcul précédente vers la droite.

ALGORITHME SUR ALGOBOX :

```
POUR k ALLANT_DE 1 A 200
  DEBUT_POUR
  n PREND_LA_VALEUR 0
  POUR i ALLANT_DE 1 A 400
    DEBUT_POUR
    SI Alea(1,6)+Alea(1,6) ≥ 9 ALORS
      DEBUT_SI
      n PREND_LA_VALEUR n+1
      FIN_SI
    FIN_POUR
  f PREND_LA_VALEUR n/400
  TRACER_POINT (k,f)
  FIN_POUR
```

Figure 2 : Modèles numériques simulant la fluctuation des fréquences (à gauche sur Excel 5 échantillons de 25 tests et à droite sur Algob 200 échantillons de 400 tests).

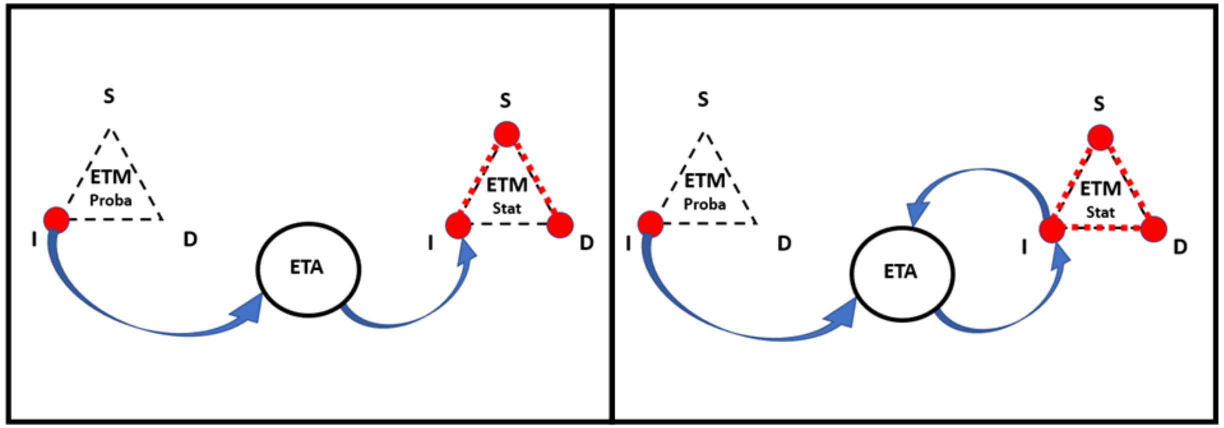
« In order to harmonize the notations, these specific workspaces, associated to the specific domains d, will be notated as ETM_d, i.e. $ETM_{algebra}$, $ETM_{analysis}$, $ETM_{arithmetic}$... The Space for Mathematical Work can be viewed as a network of diverse fibres constituting the ETM_d. Therefore, the issue is knowing how the links between the spaces or the browsing of the plans are organised. These interactions between the domains are essential for understanding the global fonctioning of the mathematical work (…). »

(Kuzniak, Richard, 2014)

Concernant la tâche choisie, nous présentons donc ci-après notre analyse a priori de ce travail mathématique à l'aide d'un tel schéma éclaté de notre ETM suivant ses projections dans les domaines en jeu.

Afin d'éviter une surcharge visuelle, nous représentons les trois projections de l'ETM par une vue du dessus. Sachant que le passage à une figure en deux dimensions nous permettra encore de présenter à l'aide de points et de traits mis en gras les différentes genèses Discursive (D), Instrumentale (I), et Semiotique (S) ainsi que les circulations dans les plans situés entre les axes de ces genèses.

<u>1ère conception/exploitation</u> : modèle numérique simulant l'expérience aléatoire.



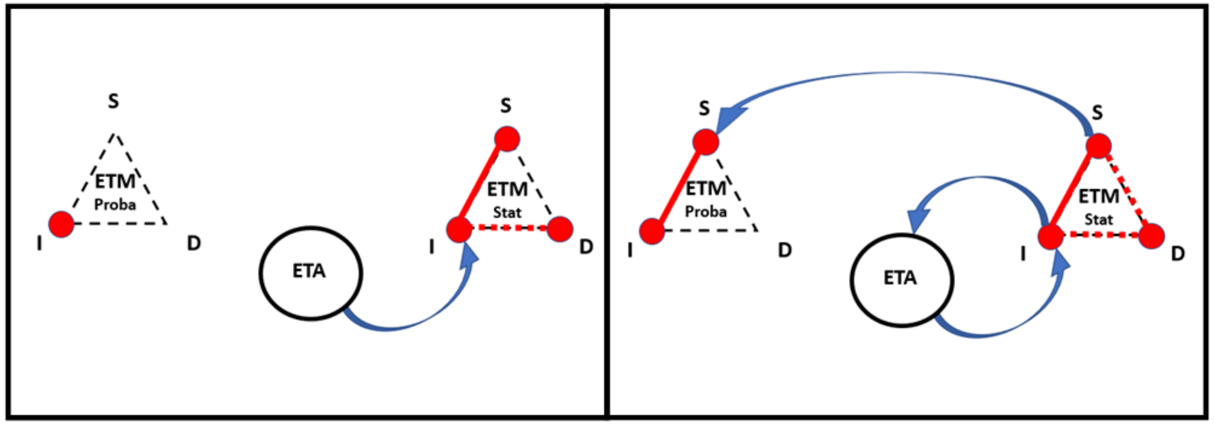
* Lancement d'une genèse instrumentale dans l'ETMProb avec l'idée de tester l'expérience aléatoire. L' ETMProb entre en interaction avec l'ETA pour simuler l'expérience aléatoire à l'aide du générateur pseudo-aléatoire Alea(). On instrumentalise dans l'ETA l'artefact pour construire l'expérience aléatoire simulée : Alea(1,6)+Alea(1,6).

* Répéter l'expérience aléatoire en relevant les résultats amène à exécuter plusieurs fois l'instruction (interaction de l'ETA vers l'ETMStat) qui sert d'artefact dans une genèse instrumentale de l' ETMStat pour construire une série de résultats.

* Dans l'ETMStat la dimension sémiotique est activée pour organiser visuellement les données (à partir de signes du type « dépouillement » et/ou « tableau »), puis la dimension discursive pour l'idée de mobiliser à partir du référentiel la notion de fréquence des succès obtenus à l'issue du protocole expérimental. Ces outils sont mobilisés par le plan cognitif pour alimenter la genèse instrumentale (pas de genèse dans les dimensions S et D de l'ETMStat).

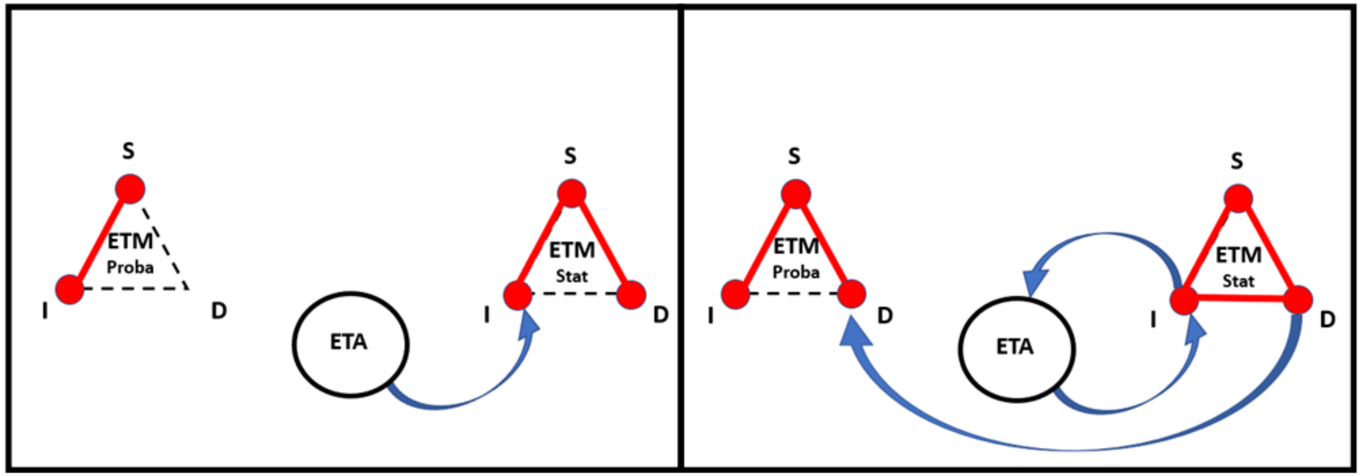
* L'idée d'automatiser l'ensemble du processus statistique réactive toutes les dimensions de l'ETMStat (expérience répétée, gérer les données, notion de fréquence) et engendre une interaction en retour de l'ETMStat vers l'ETA.

<u>2ème conception/exploitation :</u>
modèle numérique simulant une fréquence.

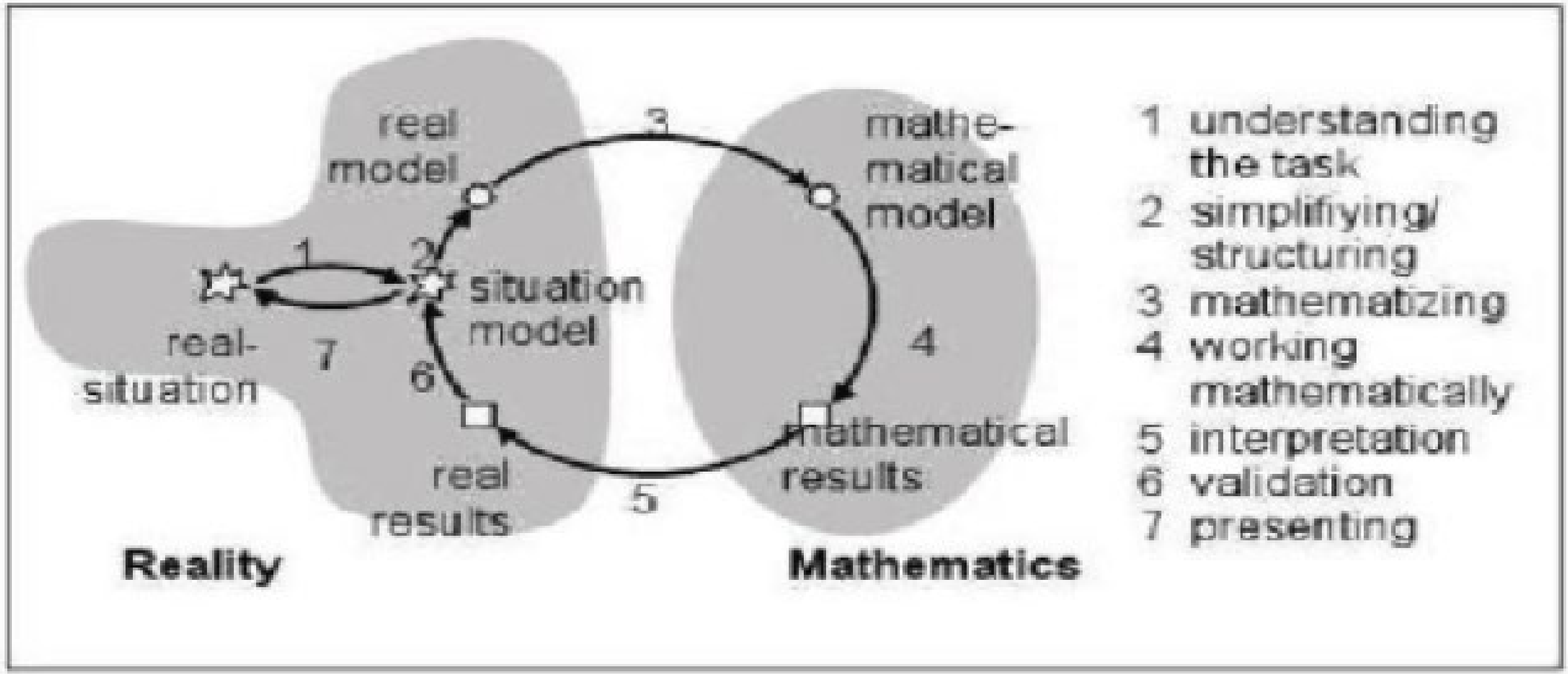


* L’objet « protocole statistique de simulation d’une fréquence » est repris dans l’ETA en tant qu’outil puis traduit sous la forme d’un objet du type « algorithme / feuille de calcul » confectionné pour simuler le calcul d’une fréquence.
* Plusieurs exécutions de l’objet « algorithme / feuille de calcul » mènent à collecter des valeurs de fréquences, et engendre une interaction de l’ETA vers l’ETMStat . L’objet « algorithme / feuille de calcul » devient un artefact qui engendre une genèse instrumentale dans l’ETMStat pour construire un nouveau protocole expérimental.
* Pour observer les résultats de fréquences on lance une représentation visuelle. La définition du repérage cartésien disponible dans le référentiel est mobilisée depuis du plan cognitif, activant la dimension discursive dans l’ETMStat.
* Le registre graphique est exploité dans une genèse sémiotique de l’ETMStat pour visualiser le nuage des points fréquences. Ce lien renforce la genèse instrumentale, d’où une circulation I-S dans l’ETMStat.
* Cette approche du type « papier » (ou par saisie dans un éditeur de graphique), donne un aperçu du nuage de points et de la fluctuation d’échantillonnage. La visualisation statistique du phénomène correspond à un phénomène lié au hasard. Ce qui engendre une interaction entre l’ETMStat et l’ETMProb (fibration S-S).
* Ce lien entre deux domaines entraîne une genèse sémiotique dans l’ETMProb. Un lien est établi avec la genèse instrumentale, en reliant graphiquement les idée d’expérience aléatoire et de fluctuation d’échantillonnage.
* L’idée d’automatiser l’ensemble du processus statistique réactive toutes les dimensions de l’ETMStat et engendre une interaction en retour de l’ETMS vers l’ETA.

3ème conception/exploitation : modèle numérique simulant la fluctuation.



* L’objet « protocole statistique de construction du nuage de points » est repris dans l’ETA comme un objet du type « algorithme / feuille de calcul » pour construire automatiquement un nuage de points fréquences.
* L’exécution de cet « algorithme/feuille de calcul » consolide les genèses et la circulation I-S dans l’ETMProba.
* Les exécutions multiples de l’outil « algorithme / feuille de calcul » permettent d’observer l’amplitude du nuage de points (interaction de l’ETA vers l’ETMStat). On construit un protocole (genèse instrumentale dans l’ETMStat).
* Ecarter les valeurs extrêmes (repère cartésien) entraîne une genèse discursive de l’ETMStat (tracés horizontaux pour élaguer). Ces signes graphiques sont exploités dans une genèse sémiotique de l’ETMStat (localiser le nuage).
* Des modifications successives dans l’objet-outil « algorithme/feuille de calcul » (nombres de tests par échantillon, et d’échantillons) permettent de conjecturer une formule pour l’intervalle de fluctuation. Ces interactions entre l’ETMStat et l’ETA renforcent les genèses, et la circulation dans l’ETMStat : I-S , S-D , et D-I .
* Conjecturer une formule pour les bornes de l’intervalle, nécessite un raisonnement inductif (genèse discursive dans l’ETMStat). Cette interaction entre l’ETMStat et l’ETMProb est du type fibration D-D.
* Cette genèse discursive dans l’ETMProb, prolonge la genèse sémiotique. Une circulation S-D se produit dans l’ETMProb. L’expérience aléatoire est reliée à la formule (modélisation formelle de l’intervalle de fluctuation).



Cycle de modélisation de Blum et Leiss (2005) (cité par Kuzniak, 2011

Le travail de conception, en classe, d'un programme de simulation d'échantillonnage mobilise des embryons de connaissance intuitive. La structuration progressive de ce travail montre un lien direct entre l'enchaînement des techniques, et les étapes de construction du programme de simulation par encapsulages successifs. Ces étapes algorithmiques entrent en résonnance avec les principales phases du protocole expérimental statistique, sémantiquement congruentes à l'organisation de type procédurale en blocs sous-programmes.

L'outil ETM met en évidence des interactions entre les trois domaines considérés : Le programme de simulation est exécuté dans l'ETM_{Stat} en tant qu'artefact, afin de construire dans une genèse instrumentale un nouveau protocole expérimental. L'ancien programme de simulation (outil théorique) va entrer en genèse discursive dans l'Espace de Travail Algorithmique (ETA) pour produire un objet algorithmique visant à automatiser ce nouveau protocole expérimental… Les projections statistique ou probabiliste de l'ETM et dans l'ETA, mettent en évidence une triple répétition de ce processus : une dialectique cyclique d'interactions évolutives, qui correspond aussi à l'enchaînement de trois cycles de modélisations…

La relation entre fréquence et probabilité est progressivement mise en évidence au travers des interactions (ou fibrations) du travail mathématique de l'ETM_{Stat} vers l'ETM_{Proba} d'abord d'un point de vue sémiotique lorsque le phénomène des fluctuations est mis en évidence visuellement, puis d'un point discursif lorsque l'intervalle de fluctuation est formalisé en tant que tel.

Nous obtenons sonc quelques réponses pour le sujet principal de cet atelier de réflexion portant sur la place du raisonnement dans le cadre d'une approche fréquentiste de l'intervalle de fluctuation. Si cette approche empirique des bornes de l'intervalle reste un raisonnement inductif du point de vue probabiliste, notre analyse montrer la mise en œuvre de nombreuses étapes de validation ayant trait au raisonnement déductif dans les autres domaines en jeu que sont les statistiques descriptives et l'algorithmique.

Bibliographie

Kuzniak, A. (2011a). L'espace de Travail Mathématique et ses genèses. In Annales de didactique et de sciences cognitives, volume 16, pages 9–24.

Kuzniak, A. et Richard, P. (2014). Spaces for mathematical work : viewpoints and perspectives. Relime, 17(4.1):17–26.

Laval, D. (2018). L'algorithmique au lycée entre développement de savoirs spécifiques et usage dans différents domaines mathématiques. *Thèse de doctorat de l'université Paris Diderot, Paris, France.*

Nechache, A. (2016). La validation dans l'enseignement des probabilités au niveau secondaire. *Thèse de doctorat de l'université Paris Diderot, Paris, France.*

Parzysz, B. (2011). Quelques questions didactiques de la statistique et des probabilités. In Annales de didactique et de sciences cognitives, volume 16, pages 127–147.

Trunkenwald, J. (2018). Entre probabilités et statistiques : un jeu algorithmique pour simuler la fluctuation d'échantillonnage. EMF 2018.

Trunkenwald, J. (2018). La simulation d'échantillonnage : Une synergie entre les domaines probabiliste et statistique. *Mathematical Working Space, Proceedings Sixth MWS Symposium (ETM 6)*. Valparaiso, Chili.

Trunkenwald, J. et Laval, D. (2019). Algorithms as a discovery process in frequentist approach to prediction interval. *Eleventh Congress of the European Society for Research in Mathematics Education (CERME11),* Utrecht University, Feb 2019, Utrecht, Netherlands. hal-02412851