Atelier AVa16 (Jean Marie Parnaudeau :Lycée Poitiers)
Quelques hypothèses sur les risques que l'on prend lorsque l'on souhaite enseigner les tests d'hypothèses !

« L'expérience peut-elle démontrer quelque chose ? »[1]

Ce texte relate la première partie de l'atelier éponyme.

Quelques remarques liminaires :

Dans ce que l'on appelle habituellement la vie courante, la plupart des individus, pour la plupart des problèmes, raisonnent par induction. Demandez à quelqu'un qui fait des « mots codés » comment il procède, demandez à un réparateur comment il procède pour trouver une panne[2].
Si mettre en œuvre un test statistique, c'est dérouler un algorithme ou appliquer une technique, comme lors de la résolution des équations du second degré par radicaux, pour ne prendre qu'un exemple, alors nul besoin de faire un effort didactique et dans ce cas il faut laisser l'enseignement des tests statistiques aux praticiens ou aux techniciens qui le font très bien.
Si, par contre, on souhaite enseigner le raisonnement plausible[3], au même titre que le raisonnement déductif ou le raisonnement par récurrence, alors cet enseignement est du ressort du professeur de mathématiques.
Sur le fond, tous les tests d'hypothèses relèvent de la même problématique, c'est pourquoi il est important que ce qui fait du « bruit » (au sens statistique du terme) soit élucidé dès le début. Le « bruit » ici étant, principalement, l'implicite, le vocabulaire et le formalisme. Une fois que l'on a compris ce que l'on fait, on peut mettre en place des pratiques (des routines) pour réduire le temps passé, comme le sont les formules pour la résolution des équations du second degré.
Ce ne sont pas le vocabulaire ou les formules qui apportent le sens.
Ce n'est pas le vocabulaire pour au moins deux raisons ; d'une part les définitions ne sont que des abréviations[4] et d'autre part si on se penche un peu sur la genèse des tests, les différents auteurs reconnaissent que le vocabulaire n'est peut être pas le plus approprié ; mais comme il est utilisé par tous, alors…
Ce ne sont pas les formules qui apportent le sens, car les résultats mathématiques utilisés en statistiques inférentielles sont en général difficiles à démontrer[5]. Certes, il est utile, pour un enseignant, de savoir comment tel ou tel résultat a été établi, mais il faut nous garder de nous réfugier dans les calculs. Ce n'est pas par les calculs (en statistiques, mais de façon plus générale en mathématiques) que l'on comprend ce que l'on fait, ni pourquoi on le fait.
Lorsque l'on consulte la littérature, le plus souvent les tests statistiques sont enseignés après les intervalles de confiance ; de plus l'approche est probabiliste et basée le plus souvent sur

[1] Extrait du sujet de philosophie du baccalauréat série S 2006.
[2] Ces deux situations sont de nature très différentes. Citons une différence entre les deux : dans l'une la base de connaissances est fixe, il s'agit de la distribution des fréquences des différentes lettres, et en réalité seules les fréquences de quelques lettres ; dans l'autre, toute nouvelle expérience (réparation) modifie la base de connaissances, ce que l'on appelle communément l'expérience.
[3] Pour reprendre le titre de l'ouvrage *Les mathématiques et le raisonnement plausible* de George POLYA édité chez Gauthier Villars 1958.
[4] « Car les géomètres et tous ceux qui agissent méthodiquement, n'imposent des noms aux choses que pour abréger le discours, et non pour diminuer ou changer l'idée des choses dont ils discourent » extrait de *De l'esprit géométrique* de Blaise PASCAL (1656).
[5] Un exemple, il n'y a pas, à ma connaissance, de démonstration élémentaire de l'énoncé « la somme de deux variables aléatoires normales indépendantes est une variable normale », énoncé qui est très fréquemment utilisé en BTS par exemple.

une loi continue, à savoir la loi normale. La proposition qui suit ne présuppose que des connaissances sur les fluctuations d'échantillonnage du programme de la classe de seconde. Hormis les notes en bas de page, tout le vocabulaire employé est à prendre au sens commun du terme. Le ton est celui d'une narration.

Entrons dans le vif du sujet :

L'activité ci-dessous a pour objectif d'introduire la problématique des tests d'hypothèse(s). La question, qui nous est posée, est :

Dans une usine, le responsable de la fabrication affirme que le nombre de produits fabriqués présentant un défaut est inférieur à 7%.
Sur la chaîne de fabrication, on a prélevé 38 produits, et on a relevé 6 produits présentant un défaut.
Qu'en pensez-vous ?

Avant d'aller plus loin, précisons un point technique, dans la suite, nous nommerons « défectueux » tout produit présentant un défaut. Il s'agira d'un défaut relatif à un seul attribut. Par exemple, un attribut des écrans plats (d'ordinateurs) est le nombre de pixels défectueux, leurs dispositions relatives en est un autre[6].

Comprendre ce qui ne sera pratiquement jamais écrit parce qu'allant de soi ou comment avoir une formulation claire des objectifs de la recherche

Dans la plupart des cas, pour qu'une analyse statistique soit menée, il faut « épurer le problème » ; c'est souvent une tâche difficile, mais elle est indispensable.
D'abord, il faut identifier le problème : quelqu'un a remis en cause l'affirmation du responsable de la fabrication (est-ce un client mécontent, un organisme de certification, une organisation de consommateurs, ou un membre de la société ?). Eh oui, car si on fait une étude statistique, c'est parce que l'on a de bonnes raisons ; le plus souvent, c'est parce qu'il y a un problème. Tant que vous ne l'aurez pas identifié (le problème), ce n'est pas la peine d'aller plus loin[7].
Qui dit problème dit le plus souvent, dans le domaine qui nous intéresse (ici un processus de fabrication), remise en cause d'un état de fait.[8]
Pour nous, **client mécontent**, il s'agit de la remise en cause de l'affirmation du responsable. Plus précisément, **il y a présomption que la proportion de défectueux soit supérieure à 7% (voire bien supérieure à 7%).**
On pourrait très bien dire, « comment a-t-il fait pour avoir ce pourcentage ? », « est-ce qu'il est sérieux ? ». Pour remettre en cause une affirmation, la polémique n'est pas une méthode acceptable. Il nous faut argumenter et pour cela, être informé et disposer de méthodes nous permettant d'avancer des arguments plus solides que de dire « le responsable de la fabrication est un menteur ».

[6] Consulter, par exemple sur internet, les normes en vigueur
[7] C'est le problème posé qui va induire la mise en place et déterminer la nature du test. D'autres problématiques donnent lieu à la mise en place de tests, mais nous n'aborderons pas cet aspect ici.
[8] Dans cette optique, les contrôles en cours de fabrication sont considérés comme des tests de « précaution ». L'entreprise met en place des procédures de contrôle en cours de fabrication afin de garantir un certain degré de qualité de la production, mais aussi pour se prémunir contre des plaintes éventuelles. Mais là aussi c'est un autre débat.

La règle générale, comme dans tout problème, est d'abord de prendre le maximum d'informations sur le sujet concerné. Par exemple, aller sur la chaîne de fabrication voir comment cela marche, qui fait quoi, comment, quels types de contrôles sont mis en place, suivant quel protocole, interroger les personnes qui effectuent les contrôles… Lire les compte-rendus sur les incidents ou problèmes de fabrication… Des questions comme celles qui nous intéressent ont-elles déjà été posées et comment ont-elles été résolues ?
Bien entendu, en classe, on ne peut faire effectivement cette démarche, mais dans la vie réelle, c'est ce qui se fait (ou devrait se faire). Il est important que les élèves soient sensibilisés avec ce type de questionnement.

Pourquoi prendre 38 produits ? Est-ce une norme professionnelle ou une habitude ou le fait du hasard (au sens usuel du terme) ? Comment ces 38 produits ont-ils été prélevés ? A quel moment ? Par exemple, si c'est en début de fabrication, on peut s'interroger sur le bien fondé de ce prélèvement. Quel contrôle a permis de décider si le produit était ou non défectueux ?
Admettons que tout cela soit fait.
Admettons que l'affirmation du responsable de la fabrication soit vraie. Il est légitime de se demander comment il a trouvé ce 7%. Est-ce qu'il sait que c'est 6,2% et qu'il ne veut pas le dire, est ce qu'il ne sait pas exactement ? Il est fort probable (au sens courant du terme) qu'il s'agisse d'une connaissance statistique, il doit y avoir des contrôles réguliers de la fabrication et il sait qu'il est très rare que le taux de défectueux dépasse 7% [et comme ça s'il y a un contrôle, il a peu de chances d'être pris en défaut ; les habitués du contrôle de qualité peuvent légitimement se poser la même question].
Faisons comme si le taux de défectueux dans la fabrication était de 7%[9] et admettons aussi que le prélèvement présente toutes les garanties que nous sommes en droit de réclamer (nous préciserons plus loin ces garanties).

Notre problème peut donc s'écrire :

Dans une usine, le responsable de la fabrication affirme que le nombre de produits fabriqués présentant un défaut est égal à 7%.
Sur la chaîne de fabrication, on a prélevé 38 produits, et on a relevé 6 produits présentant un défaut.
Qu'en pensez-vous ?

Il y a 6 produits présentant un défaut.

C'est simple, 6 divisé par 38 donne 15,7%.

15,7% est plus grand que 7% , donc le responsable de la fabrication est un menteur[10].

Est ce aussi simple ? D'ailleurs avec un nombre tel que 38, quel que soit le nombre de défectueux nous serions amenés à donner à chaque fois la même conclusion ; si k est le nombre de défectueux parmi les 38, k/38 sera toujours différent de 0,07.

Afin de donner une conclusion ou de prendre une décision, il faut aller au delà des résultats numériques

C'est-à-dire réfléchir au-delà de 7% et 6 défectueux sur 38.
Si l'on effectue un autre prélèvement de 38, aura-t-on encore 6 défectueux ?

[9] On décide de lui laisser le maximum de chances. Ceci ne pourra être justifié, aux élèves ou étudiants que plus tard. Ce n'est que lorsque nous aurons défini une règle de décision, en supposant que la proportion est de 7%, que cette phrase prendra tout son sens.
[10] C'est très souvent l'opinion des élèves. Il est temps de relire le chapitre I (la notion d'obstacle épistémologique) de *La formation de l'esprit scientifique* de Gaston BACHELARD 1938.

Si l'on considère beaucoup de prélèvements de 38 produits, quel est le nombre de défectueux que l'on est susceptible d'observer ? Dans le cas qui nous intéresse, ce nombre peut être n'importe quel entier entre 0 et 38. Sans trop faire de mathématiques, si le responsable de la fabrication dit vrai, c'est-à-dire si la proportion de défectueux est de 7%, il est clair que avoir 38 défectueux sur les 38 serait très surprenant, de même que 0 défectueux. Mais avoir 3 défectueux semble possible.
Il nous faut donc trouver un moyen de quantifier cet aspect.

Considérons tous les prélèvements possibles de 38 produits.
Que peut-on en dire de pertinent pour notre problème ?
Notons N le nombre de défectueux pour un prélèvement donné[11], N peut prendre toutes valeurs entre 0 et 38. Nous dirons que l'ensemble des valeurs possibles de N est {0, 1, 2, …., 38}.
Nous avons dit que intuitivement si ce que dit le responsable de la fabrication est vrai, alors certains cas seront plus rares ou plus fréquents que d'autres. Par exemple, on aura plus souvent 3 défectueux que 35. Dit autrement, cela signifie que si, au lieu de considérer tous les prélèvements de taille 38, on regarde seulement le nombre de défectueux de chaque prélèvement, alors on aura beaucoup plus de 3 que de 35. Ce que l'on peut aussi écrire : l'événement «avoir trois défectueux» (ce que l'on peut noter «N = 3 ») est *plus probable que* l'événement «N = 35»[12]. Pour quantifier ce «plus probable que», il faut examiner plus précisément la question.

Une modélisation est nécessaire

On peut d'abord se poser la question « comment fait-on pour savoir si le produit est défectueux ? ». Par exemple, si le contrôle est destructif, il n'est pas question d'envisager un prélèvement avec remise[13].
Pour aller plus loin, il est nécessaire de considérer le résultat de notre expérience (avoir 6 défectueux sur les 38 produits) comme un élément d'un ensemble. Cet ensemble serait l'ensemble des résultats de tous les prélèvements possibles de taille 38 de la fabrication.
Pour caractériser cet ensemble, nous devons savoir comment le prélèvement s'est effectué. Ce sont les garanties mentionnées précédemment.

- Admettons que le prélèvement a été fait de la façon suivante :
  on prélève un produit et on dit si il est défectueux ou non, on note le résultat, on prend un autre produit et ainsi de suite jusqu'au trente-huitième[14]. Et on donne le nombre de défectueux parmi les 38.
- Admettons aussi qu'il s'agit d'un produit fabriqué en très grande quantité. Ce qui signifie que le fait de prélever un (ou plusieurs) produit ne modifie pas la proportion de défectueux dans cette fabrication.
- Admettons aussi que le responsable de la fabrication dit vrai, c'est-à-dire que la proportion de produits défectueux dans la fabrication est de 7%.

---

[11] Pour être précis, il faudrait écrire que N est la fonction qui à tout échantillon associe le nombre de défectueux de cet échantillon, mais d'une part le mot échantillon ne figure pas dans le texte et d'autre part, de fait N est une variable aléatoire. Cette fonction sera nommée, beaucoup plus tard, statistique, variable aléatoire de décision ou fonction discriminante suivant les niveaux d'études, les « écoles » ou les secteurs d'activité.
Pour une première approche, comme indiqué au début, il me semble que tout le vocabulaire spécifique doit être laissé de coté.

[12] Sans le dire, on vient de définir un résumé du prélèvement. Ce qui sera nommé plus tard une statistique.

[13] Dans le cas d'un contrôle destructif ou assimilé, le prélèvement est un prélèvement sans remise par nature.

[14] C'est d'ailleurs ce qui se passe dans la réalité. On ne remet pas le produit contrôlé dans la fabrication, sauf si le contrôle est non destructif et que le produit est déclaré sans défaut. De plus la notion de « prélèvement au hasard », telle qu'elle est définie dans un cours de probabilité, n'est pas celle qui a cours sur une chaîne de fabrication. Mais c'est une autre histoire.

Il manque une condition très importante, c'est la place de l'aléatoire, c'est à dire à quel moment le hasard intervient. Dans notre cas, la condition à ajouter est que chaque produit est prélevé au hasard. On peut alors dire que pour chaque produit (prélevé au hasard dans la fabrication), la probabilité que ce produit soit défectueux est de 0,07 et la probabilité qu'il ne soit pas défectueux est de 0,93.[15]

Les conditions que nous venons d'imposer nous permettent de mettre en place un modèle et donc d'envisager de raisonner au-delà des valeurs numériques. Un modèle pourra être défini comme une expérience équivalente (mais que l'on pourra reproduire à l'identique plus facilement qu'un prélèvement sur la chaîne de fabrication). Dans le cas qui nous intéresse, ce sera un modèle d'urne[16].

Dans une urne on place 100 boules, 93 sont bleues et 7 sont jaunes. Les boules sont supposées indiscernables (sauf en ce qui concerne la couleur !).

On applique 38 fois le protocole suivant : on prélève, au hasard, une boule de l'urne, si la boule est jaune, on marque 1, si la boule est bleue, on marque 0, on la remet dans l'urne ;

A la fin, on fait la somme des nombres obtenus.

Simuler des prélèvements sur la chaîne de fabrication consiste à simuler le modèle.

Suivant le niveau de connaissances et suivant le but recherché, deux options pédagogiques sont possibles, soit procéder par simulation, soit utiliser la loi de probabilité associé au modèle retenu[17].

Pour avoir un instrument de mesure, une base de référence, nous allons simuler cette expérience (c'est à dire « faire tourner » le modèle) soit à la main, soit avec un tableur…[18]

Supposons que nous ayons fait 10 000 simulations[19]..

Nous obtenons alors un tableau qui peut avoir la forme suivante :

Nombre de défectueux	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Fréquences observées	0,06342	0,18145	0,25266	0,22821	0,15030	0,07693	0,03186	0,01066[20]	0,00319	0,00080	0,00048	0,00003	0,00001	0,00000

Les entiers supérieurs à 13 ne figurent pas car leurs fréquences sont inférieures à 10^{-5}.

[15] Tout ceci pour pouvoir assimiler la fabrication à une urne de Bernoulli à deux catégories dans laquelle on fait des tirages non exhaustifs.

[16] C'est un des modèles d'expérience aléatoire de référence, comme indiqué dans les programmes des classes de lycée (séries S et ES).

[17] Une des difficultés dans la mise en place d'un test est de définir la variable aléatoire de décision ou, ce qui revient au même, le modèle et l'instrument de mesure de l'écart. Si l'on prend l'option loi de probabilité, alors comme on suppose que le fait de prélever 38 produits ne modifie pas ou modifie très peu la proportion de défectueux (ce que l'on admet toujours dans le cas de fabrication en grande série), alors la loi de probabilité de N est la loi binomiale de paramètre 38 et 0,07 (on dit aussi de paramètres n = 38 et p = 0,07). Nous pouvons alors calculer la probabilité d'avoir 0, 1, 2, … 38 défectueux parmi les 38 produits prélevés. Le raisonnement se poursuit comme avec les fréquences. Dans cet exemple, la variable de décision « s'impose » d'elle même, mais ce n'est pas toujours le cas ; le choix de la meilleure variable de décision (en un sens à définir) est en général une question compliquée. Sans être polémique, les lois normales ont souvent bon dos.

[18] Est-ce qu'un tableur fait vraiment ce qu'il laisse croire qu'il fait ? Faire des simulations soi-même, à la main, est indispensable avant d'accepter de déléguer ce travail à une machine. Le phénomène « boîte noire » du tableur est un obstacle didactique souvent sous estimé. Comme le point de vue adopté est l'approche fréquentiste, on peut supposer que cet aspect a déjà été abordé (en classe de seconde ou après).

[19] Pourquoi prendre 10 000 et non 2 000 ou 100 000, cette question importante mérite d'être examinée de près, mais c'est un autre problème. Il ne s'agit pas de « botter » en touche » comme l'on dit, mais un vieux précepte dit « jamais deux difficultés pédagogiques en même temps ». De plus, pour les besoins de l'activité, seuls les fréquences nous intéressent.

[20] La théorie des tests statistiques est d'origine anglo-saxonne, on peut citer par exemple K Pearson, R A Fisher, E S Pearson et J Neyman (qui était d'origine polonaise), tout ceci pour dire que le 1066 est un clin d'œil.

A la lecture de ce tableau de fréquences, on peut remarquer que, si la proportion de défectueux dans la fabrication est de 7%, alors avoir 2 défectueux est un événement courant (une fois sur quatre en moyenne) et le fait d'en avoir 0 est peu courant (6 fois sur 100 en moyenne), le fait d'en avoir 7 très peu courant (1 fois sur 100 en moyenne).
Remarquons que, si dans notre prélèvement de 38 produits, nous avions obtenu 1 défectueux, nous ne serions pas allés nous plaindre, pour être plus précis, nous n'aurions même pas eu l'idée d'aller nous plaindre. Remarquons aussi que si nous avions obtenu 12 défectueux, nous serions extrêmement surpris, car cet événement est particulièrement rare (une fois sur 100 000 d'après le tableau de fréquences) et nous serions en droit d'avoir de forts soupçons quand à la validité de l'affirmation du responsable de la fabrication.
Il nous faut retenir un critère pour prendre une décision ; c'est à dire « rejeter l'affirmation du responsable » ou bien « ne pas la rejeter ». Le problème est : « pour nous, il s'agit de la remise en cause de l'affirmation du responsable. Plus précisément, il y a présomption que la proportion de défectueux soit supérieure à 7% ».
Nous allons choisir comme critère de décision la réalisation de l'événement « le nombre de produits défectueux est supérieur ou égal à 6 »[21]. Avec le résultat de notre simulation, la fréquence de cet événement est environ de 0,047.
Cela signifie que tout se passe comme si, dans une urne contenant 953 boules rouges et 47 boules vertes, on prélevait, au hasard, une boule et qu'elle soit verte. Cet événement est peu probable.
Notre conclusion sera que le fait d'avoir un nombre de défectueux supérieur ou égal à 6, si la proportion dans la fabrication est de 7%, est suffisamment rare pour que nous mettions fortement en doute l'affirmation du responsable de la fabrication.
Nous sommes en droit de rejeter son affirmation.

Nous avons rejeté l'affirmation du responsable de la fabrication, car nous avons décidé qu'un événement qui ne se produit que dans 47 cas sur mille est trop rare.[22].

Si ce que dit le responsable de la fabrication est vrai, alors avoir 6 défectueux est un événement rare, mais avoir 8 défectueux est un événement encore plus rare, c'est même quasi ment impossible cela signifie que si nous avions eu 8 défectueux au lieu de 6 nous serions encore plus enclins à rejeter son affirmation.
Supposons que dans notre prélèvement de 38 produits, au lieu d'avoir 6 défectueux, il y ait 3 défectueux. La fréquence de l'événement « avoir 3 défectueux ou plus » est d'environ 0,5 ; ce qui signifie que, si l'affirmation du responsable est vraie, alors une fois sur deux, en moyenne, on observera 3 défectueux ou plus parmi les 38 produits. Il ne serait pas raisonnable de rejeter l'affirmation du responsable, mais cela ne prouverait pas non plus que ce qu'il affirme est vrai. Nous ne pourrions énoncer qu'une conclusion de la forme « son affirmation est recevable[23] » ou bien « jusqu'à preuve du contraire, nous la considérerons comme vraie »[24].

[21] Retenir ce critère n'est pas spontané. L'acceptation de ce critère ne peut résulter que d'un débat en classe.
[22] On pourra lire dans *Chemins de l'aléatoire ; le hasard et le risque dans la société moderne* de Didier DACUNHA-CASTELLE édité chez Flammarion 1999 dans le chapitre 3 « le principe de la méthode est alors simple : observer la réalisation d'un événement rare dans une théorie conduit à douter de celle-ci » ou bien dans *Au hasard, la chance, la science et le monde* d'Ivar EKELAND, comme indiqué à la fin de l'introduction, vous avez lancé le dé, et vous avez un 6.
[23] Cette locution est utilisée, par exemple dans les ouvrages de génétique.
[24] Tout cela pour dire qu'un test statistique ne permet pas de démontrer qu'une affirmation (ou une hypothèse) est vraie.

Il ne faut pas perdre de vue que nous ne saurons jamais si cette décision est une bonne décision ou une mauvaise décision, mais en utilisant au mieux les informations dont nous disposons, nous avons pris une décision et nous sommes capable d'expliquer pourquoi nous l'avons fait.

Essayons de généraliser un peu ce qui vient d'être fait, toujours en nous plaçant dans le cas d'une personne ou d'un organisme qui remet en cause une affirmation (la proportion dans la fabrication est de 7%) au profit d'une autre (la proportion dans la fabrication est supérieure à 7%).
La méthode employée peut s'appliquer si la proportion déclarée par le responsable est 9%, 35% ou 1,3%. Elle peut aussi s'appliquer si le prélèvement est de taille 12, 57 ou 200. La modélisation se justifiera de la même façon et afin de prendre une décision, nous garderons la même méthode[25]. La seule part subjective est : à partir de quelle fréquence doit-on rejeter l'affirmation, c'est-à-dire à partir de quelle fréquence, un événement sera déclaré suffisamment rare pour que son observation nous permette de rejeter l'affirmation[26]. Cela ne peut résulter que d'une convention.
La convention la plus courante pour le seuil de rareté est 0,05 (un cas sur 20 en terme de fréquences). Ce qui se traduit dans le premier cas par le fait qu'un événement sera décrété suffisamment rare pour entraîner le rejet de l'affirmation du responsable si sa fréquence (ou sa probabilité) est inférieure ou égale à 0,05[27].(ce n'est pas un exposant, mais une note en bas de page !). D'autres seuils sont utilisés par exemple 0,01 ou 0,005. Sans entrer dans les détails, il faut bien comprendre qu'il s'agit, d'une convention, dans des applications pédagogiques, d'un consensus ou d'une norme dans les applications professionnelles.

Et le vocabulaire dans tout cela

La tradition didactique a consacré un vocabulaire et des procédures. Souvent coupé de son origine, le vocabulaire a perdu une partie de son sens.
Donnons juste un exemple.
Le mot échantillon n'a pas été écrit et pour cause, ce mot est d'une définition et d'un emploi difficile. Le choix du mot prélèvement, qui implique de préciser le mode opératoire de ce prélèvement me semble plus adapté pour une question telle que celle qui est posée. Dans la plupart des ouvrages scolaires de lycée, un échantillon désigne une partie d'un ensemble ; cette définition est inadaptée, de façon générale en statistique inférentielle[28].
L'affirmation du responsable de la fabrication porte souvent le nom d'hypothèse ou d'hypothèse nulle ou d'hypothèse principale.
Dans le cas traité ci dessus, l'hypothèse serait notée par exemple « $\pi = 7\%$ » (ou bien « $\pi = 0{,}07$ »), en français, la proportion de défectueux dans la fabrication est de 7%[29]. Le

[25] On pourra lire une version romancée, mais très sérieuse (il y est question de loi binomiale et de loi normale !), le chapitre 4 « le vieux loup de mer » dans *Elémentaire mon cher Watson* ! de Colin BRUCE édité chez Flammarion en 2002.
[26] On peut faire l'analogie, même si cela est discutable, avec : « c'est tellement rare que cela m'a mis la puce à l'oreille…»
[27] En fait, c'est un peu plus compliqué que cela, mais pour une première approche, une convention est suffisante pour travailler. En classes de terminale ES ou S, pour le test d'adéquation à une loi équirépartie, le seuil de rareté est de 0,1. Il est possible que ce choix, peu utilisé en dehors de mathématiques, soit un réinvestissement des diagrammes en boîtes vus en classe de première S.
[28] Dans notre activité, il faut admettre, au moins en théorie, d'avoir deux fois le même élément.
[29] Pour être plus précis, l'hypothèse « la proportion de défectueux dans la fabrication est de 7% » est une hypothèse simple, alors que l'hypothèse de départ, « la proportion dans la fabrication est inférieure au égale à 7% » est une hypothèse dite complexe. Seule une hypothèse simple permet une simulation ou un calcul de probabilité. Bien sûr, nous savons tous et toutes que $\pi = 3{,}14$ (comme disent les élèves), mais une convention

terme hypothèse principale sous-entend qu'il y a une autre hypothèse, dans notre exemple, l'autre hypothèse serait « $\pi > 7\%$ », cette deuxième hypothèse porte parfois le nom d'hypothèse alternative.
Remarquons tout de suite qu'il ne s'agit pas d'une hypothèse au sens où ce mot est utilisé en cours de mathématiques puisque dans ce contexte, pour un élève ou un étudiant, une hypothèse est une donnée d'un exercice ou une prémisse d'un théorème. Il ne s'agit pas non plus d'une conjecture, puisque face à une conjecture (peu importe comment on l'a obtenu) le travail mathématique consiste à l'infirmer ou à la démontrer, ce qui ne sera jamais le cas avec la méthode que nous avons utilisée. En simplifiant beaucoup, on peut dire que les hypothèses auxquelles nous nous intéressons sont des hypothèses susceptibles d'être mises à l'épreuve d'une expérience aléatoire. Ces hypothèses sont parfois qualifiées d'hypothèses statistiques ou probabilistes.
L'hypothèse principale est notée H_0. Dans l'exemple, l'hypothèse H_0 est « la proportion de défectueux dans la fabrication est inférieure ou égale à 7% ». Seule une analyse du problème et un élan magnanime ont fait que l'on a pu l'écrire sous la forme « la proportion de défectueux est de 7% ». L'hypothèse alternative est notée H_1.
Pourquoi a-t-on mis en place un test ? Uniquement parce que, pour des raisons non précisées dans le texte initial (premier encadré), il y avait un faisceau de présomptions contre H_0.
C'est donc H_1 qui motive toute cette histoire. On dit parfois que c'est H_1 qui détermine (ou induit) la nature du test[30].
Le fait de conclure un test statistique par « on rejette H_0 » ne signifie pas H_0 soit fausse ; le fait de conclure par « on ne rejette pas H_0 » ne signifie pas que H_0 soit vraie.

Généralisons un peu plus…

Allons un peu plus loin. Afin de faire face, dans l'avenir, à des réclamations de ce type, l'entreprise consulte un statisticien pour mettre en place une procédure qui permettra de régler la question.
Le responsable de la fabrication affirme que la proportion de produits défectueux est inférieure ou égale à 7% et une association de consommateurs affirme que cette proportion est supérieure à 7%.
Le statisticien écrit le problème sous la forme suivante :
L'hypothèse principale est $H_0 : \pi = 7\%$ et l'hypothèse alternative, notée $H_1 : \pi > 7\%$.
Le statisticien établit le protocole suivant :
Prélever, de façon aléatoire, 100 produits dans la fabrication. Compter le nombre de produits présentant un défaut.
Si le nombre de produits défectueux est inférieur ou égal à 10, on ne remet pas en cause l'affirmation du responsable, si le nombre de défectueux est supérieur ou égal à 11, on remet en cause son affirmation.
Le statisticien vient de nous donner une **règle de décision**, car il s'agit d'une procédure à appliquer et qu'elle aboutit à une décision.

Nombre de défectueux	0, 1, 2, 3, 4, 5, 6, 7 ,8, 9, 10	11, 12, 13, 14, …
Décision	On ne rejette pas l'affirmation	On rejette l'affirmation

Cette procédure a été établie en prenant un seuil de rareté de 0,05[31].

---

consiste à noter par des lettres grecques les « paramètres » (qu'ils soient connus ou non) d'une variable statistique, par exemple μ pour la moyenne ou π pour une proportion.
[30] Si H_1 était « la proportion est inférieure à 7% » ou bien « la proportion est différente de 7% », on voit bien que le raisonnement suivi serait encore valide, mais le critère de décision ne serait pas le même.
[31] On procède comme dans l'exemple précédent, mais avec 100 produits au lieu de 38.

Cette règle de décision permet de comprendre par exemple pourquoi un test statistique est souvent défini comme une règle permettant, à partir de n observations, de faire un choix entre deux hypothèses s'excluant mutuellement. De même l'ensemble {11, 12,…} porte le nom de **région critique**, c'est l'ensemble des valeurs qui entraîne le rejet de l'hypothèse principale. Attention « *décision n'est pas raison* ». Rien ne prouve que cette décision soit conforme à la réalité ; en effet, il se peut que l'on rejette l'affirmation du responsable, alors que dans la réalité, ce qu'il dit est vrai. Mais il se peut aussi que l'affirmation des consommateurs soit vraie et que pourtant on décide que c'est le responsable de la fabrication qui a raison.
Lorsque l'on prend des décisions, on peut faire des erreurs ! Pour la question qui nous intéresse, nous pouvons faire deux types d'erreurs.
La première erreur[32] est celle qui consiste à rejeter à tort l'hypothèse H_0 , c'est-à-dire à remettre en cause l'affirmation du responsable, alors qu'il a raison. Cette première erreur est nommée **erreur de première espèce**. « Il est possible que l'on réponde affirmativement par erreur, c'est-à-dire que l'on rejette à tort l'hypothèse H_0 : les fluctuations aléatoires peuvent être exceptionnellement fortes. C'est un risque que l'on prend, appelé **risque de première espèce**. »[33].
Dans le premier cas (prélèvement de 38 produits) le risque de première espèce était de 0,047, il a été calculé à partir des données du prélèvement (voir plus haut).
Dans le deuxième cas, le risque de première espèce a été fixé par le statisticien (0,05), mais en fait ce risque est de 0,0469[34]) et à partir de ce risque, il a trouvé une valeur, dite valeur « critique », puis il a donné une règle de décision :
Pour tout échantillon de taille 100, si le nombre de défectueux est inférieur ou égal à 10, on ne rejette pas H_0 et si le nombre de défectueux est supérieur ou égal à 11, on rejette H_0.[35]
Par habitude, le risque de première espèce est noté α. Remarquons, sans entrer davantage dans les détails que, dans le premier cas, le risque de première espèce est calculé à partir des données issues du prélèvement, alors que, dans le deuxième cas, il est fixé *a priori*.

La deuxième erreur possible est celle qui consiste à ne pas remettre en cause l'affirmation du responsable, alors qu'il a tort ; c'est à dire de ne pas rejeter H_0 alors que c'est H_1 qui est vraie. Cette erreur est nommée **erreur de deuxième espèce**, sa probabilité est **nommée risque de deuxième espèce** et est notée β. Pour la règle de décision que nous avons, nous ne pouvons pas calculer ce risque. En effet, pour calculer il faudrait avoir plus de renseignements sur l'hypothèse alternative H_1, la proportion est plus grande que 7%, ce n'est pas suffisant comme information. Dans notre exemple, pour pouvoir calculer le risque de deuxième espèce, il faut que H_1 soit une hypothèse simple.

Supposons maintenant que le problème posé au statisticien soit le suivant :

---

[32] Le mot *erreur* doit être pris dans un sens différent du sens classique. Dans notre exemple, nous ne saurons jamais si nous avons fait une erreur ou non, c'est à dire si nous avons pris une bonne décision ou non [il en est de même pour les fourchettes de sondages en classes de seconde (premier thème de statistique)]. Gaston BACHELARD dans *Essai sur la connaissance approchée* écrivait : « l'idée de loi statistique diffère philosophiquement de l'idée de loi approchée. (…). La loi statistique résulte au contraire d'un jeu qui court le risque d'une erreur complète, elle peut se trouver absolument en défaut. Le calcul d'erreur, au sens précis du terme, n'a pas d'emploi, il est remplacé par un calcul de chances ».

[33] Introduction de la norme AFNOR NF X 06-065.

[34] Si X est une variable aléatoire de loi binomiale B(100 ;0,07), alors la probabilité de l'événement $X \geq 11$ est égale à 0,0469 résultat obtenu au tableur ou à la calculatrice.

[35] Ces deux points de vue sont de natures très différentes. Par exemple, le statisticien met en place une procédure destinée à être utilisée souvent par un non spécialiste de la statistique. Ce n'était pas le cas du premier exemple. En particulier, dans le cas du statisticien, comme le risque de première espèce est fixé, mettre en place un test ne consiste plus, moyennant des calculs, qu'à déterminer des « zones de rejet » de H_0 et des « zones de non rejet » de H_0, c'est à dire trouver une (ou des) valeur « critique ».

Le responsable de la fabrication affirme que la proportion de défectueux est inférieure ou égale à 7% et une association de consommateurs affirme que la proportion de défectueux est au moins de 11%. Le statisticien va transformer le problème et écrire :
L'hypothèse principale est $H_0 : \pi = 7\%$ et l'hypothèse alternative est $H_1 : \pi = 11\%$.
Pour un prélèvement de taille 100, il trouvera la même règle de décision.
Mais comme H_1 s'écrit la proportion de défectueux dans la fabrication est de 11%, nous pouvons calculer le risque de deuxième espèce.
C'est la probabilité, si l'affirmation des consommateurs est vraie, de l'événement « le nombre de défectueux est inférieur ou égal à 10 ». Dans notre exemple, la probabilité de cet événement est de 0,453[36]. Si dans une fabrication la proportion de défectueux est de 11%, la probabilité d'observer un nombre de défectueux inférieur ou égal à 10 dans un prélèvement de 100 produits est de 0,453. Ce qui signifie que cet événement se produit environ une fois sur deux (si on faisait beaucoup de prélèvements, 453 fois sur 1000, en moyenne, le nombre de défectueux serait inférieur ou égal à 10).
Cela signifie aussi que l'on va avoir du mal à rejeter l'affirmation du responsable de la fabrication. Le « jeu » n'est pas loyal, car si on avait raison, on perdrait dans 453 cas sur 1000. Misère !
Il faudrait rééquilibrer la règle, c'est à dire pour notre cas, en admettant que α soit fixé, essayer de diminuer β. En restant dans de l'intuitif, on voit bien que si la taille du prélèvement augmente, le pouvoir discriminant augmente aussi, c'est à dire que pour H_0, H_1 et α fixé, si n augmente, alors β diminue. Mais, c'est une autre histoire.

Plutôt qu'une bibliographie qui souvent n'en finit pas, pour reprendre tout ce qui précède dans un autre contexte et avec un vocabulaire plus technique, on pourra travailler une norme AFNOR (par exemple la norme NF V 09-013 sur les essais triangulaires, cette norme « spécifie une méthode permettant de détecter s'il y a des différences perçues entre les échantillons de deux produits par comparaison triangulaire » l'outil mathématique est la loi binomiale) et lire la contribution de Pascal SCHLICH texte sur les statistiques de la commission de réflexion sur l'enseignement des mathématiques[37] (plus connue sous le nom de commission Kahane).

Qu'en est-il des mathématiques ? Elles interviennent comme aide à la décision.
Nous avons pris un problème de la vie courante (industrielle), moyennant quelques hypothèses raisonnables, nous avons modélisé la situation pour l'écrire sous la forme d'un problème de mathématiques (choix d'une variable aléatoire (on dit aussi statistique) sur laquelle sera basée le critère de décision, puis étude fréquentiste ou de probabilité pour cette variable aléatoire), puis nous avons émis une conclusion dans le contexte du problème initial.
La modélisation, c'est à dire la partie mathématique, comporte deux phases ; d'abord trouver (au moins) une variable aléatoire « rendant compte de la situation », ensuite parmi celles que l'on a trouvé choisir la meilleure (en un sens à déterminer !). On touche là à deux questions particulièrement difficiles, dites de statistique mathématique ou méthodologique pour reprendre l'expression de Paul Deheuvels[38]
Dans la deuxième partie de l'atelier, nous avons abordé ces deux questions à partir de deux exemples.

[36] Si la proportion de défectueux dans la fabrication est de 11%, alors N est de loi B(100 ;0,11) et il suffit de construire la table.
[37] Rapport du 15 mars 2003, disponible en librairie ou sur l'internet.
[38] Rapport sur la science et la technologie N°8 *« La statistique »* juillet 200 Edition TEC&DOC.