# A propos de test de qualité.

Syndrome du 95%

On lit souvent des expressions du genre "sûr à 95%", ou l'équivalent à connotation plus mathématique " à 2 $\sigma$ près". Par contre, il est rare de lire "avec une tolérance de …", ou "avec un écart-type de …".
Faut-il y voir une précaution prise *a priori*, en effet, si le résultat dépasse 2 $\sigma$, on se réserve la possibilité de répondre : "j'avais prévenu que c'était à 95% près !".
Autre explication possible, ce seuil de 95% est le seuil très utilisé avec le test du khi².

Soit une expérience constituée d'un ensemble de mesures indépendantes, réalisées suivant le même protocole. On sait que la moyenne arithmétique est le résultat le plus probable et que les écarts à la moyenne suivent la loi normale.
On peut donc calculer l'écart-type qui est l'écart moyen quadratique. On sait que 2/3 des écarts en valeur absolue sont inférieurs à 1 écart-type, que 95% des écarts en valeur absolue sont inférieurs à 2 écarts-type et que 99% des écarts en valeur absolue sont inférieurs à 3 écart-type. On a l'habitude de noter $\sigma$ la valeur de l'écart-type.

L'écart-type est une unité de mesure, rien de plus, sauf que son emploi sous-entend qu'on est dans un contexte de mesures aléatoires, c'est à dire dans le contexte de la loi normale.

Soit une liste quelconque. On peut calculer la moyenne arithmétique des valeurs élémentaires. Puis pour chaque valeur élémentaire, l'écart à la moyenne. Les tests de qualité consistent à chiffrer la façon dont se répartissent l'ensemble de ces écarts les uns par rapport aux autres.

Il est bien évident que la somme algébrique de ces écarts est nulle.
On peut calculer la moyenne arithmétique de ces écarts. Elle est égale à la somme des valeurs absolues des écarts divisée par le nombre d'écarts. On l'appelle ema.
On peut calculer la moyenne quadratique de ces écarts. Elle est égale à la racine carrée de la somme des carrés des écarts divisée par le nombre d'écarts. On l'appelle emq.
Le choix de la méthode est indifférent, puisqu'il y a une relation entre les deux :
$emq^2 / ema^2 = \pi/2$ soit, $emq/ema \sim 1.25$.

On utilise aussi fréquemment l'"erreur probable", notée ep. C'est la valeur qui partage l'aire sous la courbe de Gauss en deux parties équivalente. C'est à dire que la probabilité que les écarts soient inférieurs à 1 ep est égale à 50%.

Il ne faut jamais oublier qu'une expérience, quel que soit le contexte, réalisée dans de bonnes conditions, respecte la répartition des écarts à la moyenne suivant la loi normale. On trouve des méthodes pour vérifier cela, mais chaque fois que l'interprétation peut être douteuse, il vaut mieux changer de méthode de test. Le raisonnement qui consiste à calculer la probabilité que l'expérience est satisfaisante est, à mon avis, fallacieux.

Intervalle de confiance

En matière de statistique et de probabilité, il est important de définir la variable étudiée, c'est à dire le résultat que l'on veut obtenir. Deux exemples issus de lecture d'une encyclopédie libre bien connue.

1- Estimation de la taille des enfants de 10 ans. On va interroger un certain nombre d'écoles maternelles, dans différents contextes et de façon aléatoire. On va calculer la moyenne arithmétique des tailles, c'est la moyenne observée. On calculera ensuite l'écart-type par la formule bien connue. Si on reporte sur un graphique le résultat que l'on aura pris soin d'organier sous forme de classes, on obtiendra une courbe en forme de cloche, bien connue.
2- Estimation du nombre de voitures rouge. Dans ce cas aussi on va interroger un certain nombre de propriétaires, mais on ne peut pas déduire, encore moins calculer une moyenne ou un écart type, puisqu'on ne peut avoir qu'un seul chiffres, c'est à dire le pourcentage de voitures rouges.

Ces deux expériences qui utilisent la notion d'échantillon sont fondamentalement différentes.
Dans le premier cas, la taille des enfants de 10 ans, en supposant que les résultats sont cohérents, on connaît la moyenne et l'écart-type. On sait que la moitié des enfants ont une taille correspondant à +/- 2/3 de l'écart-type, 66% correspondant à +/- un écart-type etc. On dispose donc de toutes les informations nécessaires, moyenne et intervalle de confiance. Si on réalise un seconde étude satisfaisant aux mêmes conditions, on doit obtenir les mêmes résultats. Tant que le nombre de tailles observées est suffisant, la notion d'intervalle de confiance est ici sans grand intérêt.

A quoi peut donc servir cet "intervalle de confiance" ?

Passons sur les affirmation purement artificielles du genre : "95% des utilisateurs sont satisfaits !". Ceci est anecdotique, mais il faut reconnaître que 95% paraît suffisamment rassurant et que si on se trouve dans la tranche des 5%, la réponse est facile.

Imaginons un sondage concernant la taille des enfants de 10 ans. Le lecteur pourra transposer avec un autre exemple, à sa convenance. Pour un certain nombre de raisons techniques, on admettra que cette taille dépend de critères bien connus, la différence d'habitat, de localisation et je ne sais quoi d'autre. Lors du sondage, la fiche contenant la taille de l'enfant contiendra aussi les informations sur les différents critères. Les choix des enfants mesurés sera aléatoire et au dépouillement, on aura ainsi plusieurs listes, chacune fournissant une moyenne et un écart-type. Le but final étant naturellement de calculer une moyenne générale et éventuellement une fourchette d'incertitude.
Etant donné que l'on connaît la dimension de chaque groupe correspondant à chaque critère, on va pondérer chaque moyenne. Mais là le problème se pose de savoir si tous les groupes de l'échantillon sont comparables, au facteur de pondération près. Ce serait vraiment beaucoup de chance si la proportion des groupes de l'échantillon était la même que la proportion des groupes de l'ensemble des enfants de 10 ans. C'est là qu'on fait intervenir cette notion d'intervalle de confiance. Cette opération qui fait intervenir le terme 1/racine(N), N étant le nombre d'éléments du groupe concerné, permet de rendre comparable, au facteur de pondération près, les moyennes de chaque groupe et éventuellement la fourchette d'incertitude.

Autre exemple, question lue sur un forum de mathématiques, chapitre "Statistiques" :
"Je prends une base de données internet dans laquelle les internautes peuvent attribuer une note sur 20 à un ou plusieurs items. On obtient donc une moyenne sur 20 pour chaque item en fonction des notes. J'aimerais faire un classement de ces items selon leur moyenne. Cependant j'aimerais prendre en compte l'aspect *popularité* dans cette moyenne, c'est à dire si l'item A et l'item B obtiennent tous deux la moyenne 18.2, mais que le premier comptabilise 25 votes contre 120 pour le 2ème, je voudrais que ce dernier soit avantagé sur le 1er.
La formule que je cherche pourrait par exemple permettre à un item ayant une moyenne inférieure à l'autre de passer devant grâce à ce coefficient *popularité*.
Je cherche quelque chose de plus développé que « l'item ayant le plus de notes (ex : 1000) obtient le coefficient = 1 , un item ayant 250 notes aurait le coefficient 250/1000 = 0.25 ». Ce coefficient plomberait beaucoup trop la moyenne finale de l'item en question et ce n'est pas le but recherché."

La réponse n'est certainement pas "tu fais comme tu veux" comme on peut lire trop souvent, le coefficient à appliquer est racine(N), N étant la nombre de réponses. Donc dans le cas présent, le coefficient de pondération pour l'item A sera racine(25) = 5 et le coefficient de pondération pour l'item B sera racine(120) = 11. Les notes définitives seront calculées suivant la formule de pondération habituelle. Dans le cas présent on pourra par exemple prendre en compte la note par rapport à 10, cette valeur étant considérée comme "sans opinion", cela dépend de l'énoncé de la question.

La démonstration de ceci dépasse le cadre du présent papier.

Exemples

Exercice 1: *La proportion de pièces défectueuses dans un lot de pièces est 0.05. Le contrôle de fabrication des pièces est tel que si la pièce est bonne, elle est acceptée avec la probabilité de 0.96 et si la pièce est mauvaise, elle est ~~acceptée~~ refusée avec la probabilité 0.98. On choisi une pièce au hasard et on la contrôle. Quelle est la probabilité qu'il y ait une erreur de contrôle ?*
[Nota. Il est possible que cet énoncé soit une traduction. Auquel cas, il faudrait comprendre le terme "acceptée" comme "comptabilisée". J'ai préféré corriger.]

Cet énoncé peut dérouter certains probabilistes, pourtant les choses sont parfaitement claires.
La proportion de pièces défectueuses est connue. On ne sait pas comment on la connaît, et on veut pas le savoir. D'ailleurs, on ne sait pas si ces pièces "défectueuses" seront retirées ou non.
Par ailleurs, on dispose d'un contrôle de fabrication. Le résultat est du type binaire : la pièce est bonne ou pas.

- Si la pièce est bonne, elle est comptabilisée avec la probabilité de 0,96.
- Si la pièce est mauvaise, elle est comptabilisée avec la probabilité de 0,98.

Cela signifie que si la pièce est bonne, elle sera refusée avec la probabilité de 4%, et qu'au contraire, si la pièce est mauvaise, elle sera tout de même acceptée avec la probabilité de 2%.
C'est à dire qu'il y a 4% de pièces refusées à tort et 2% de pièces acceptées à tort.

Imaginons un lot de 1000 pièces. D'après l'hypothèse, 950 sont correctes et 50 sont défectueuses. Classons-les par ordre de qualité. A gauche, on aura les pièces "parfaites", et vers la droite les pièces de plus en plus défectueuses. On trace la limite 950-50.
Sur les 950 pièces conformes, la machine de contrôle va se tromper 4 fois sur 100, donc va déclarer 950 x 4% = 38 pièces défectueuses, alors qu'elles sont conformes.

Sur les 50 pièces défectueuses, la machine de contrôle va se tromper 2 fois sur 100, donc va déclarer conforme 50 x 2% = 1 pièce.

Dans la pratique, voyons ce que cela signifie.
Les pièces au voisinage de la frontière 950-50 sont "presque bonnes" ou "presque défectueuses". Puisqu'il s'agit de fabrication, les écarts suivent la distribution de la loi normale. C'est donc l'étalonnage de la machine de contrôle qu'il suffit de modifier.

Autre façon de voir le problème. Si cette frontière entre les pièces correctes et les autres était nette et sans doute possible, la machine de contrôle serait fiable à 100%, ce qui n'est pas le cas.

Cet exercice est intéressant dans le sens où on isole deux processus différents. D'une part le processus de fabrication qui met en œuvre une installation et est étudié de façon à satisfaire certains critères, en l'occurrence 95 % de pièces répondant à une certaine norme. D'autre part, un processus de contrôle, nécessairement ponctuel, la machine de contrôle ne peut pas être fiable à 100%. Les coefficients d'erreur de cette machine sont supposés avoir été fixés avec précision.

On pourra lire à ce propos le papier concernant le contrôle de boules de pétanque.

Exercices 2: *Répondre par Vrai ou Faux (en justifiant soigneusement la réponse).*
*Une entreprise fabrique en très grande quantité des gélules dont la masse est exprimée en milligrammes. Lors de la fabrication des gélules, une étude statistique a montré que 3% des gélules ont une masse non conforme.*

*Si l'entreprise conditionne les gélules par sachet de 10, il y aura au moins 96% des sachets qui comporteront 9 ou 10 gélules de masses conformes.*

Cet exercice a provoqué beaucoup d'émoi chez certains mathématiciens.
Les termes de l'énoncé ont été soigneusement choisis.

- "fabrique en grande quantité", cela implique la loi des grands nombres
- "unité milligramme" c'est une information inutile, juste pour détourner l'attention de l'élève,
- "étude statistique" cela sous-entend que l'on a pesé avec grande précision une cinquantaine de gélules prélevées aléatoirement.
- Enfin la question "Si le conditionnement est par sachet de10 ..." est très précise : le conditionnement a-t-il un impact que la probabilité du résultat ?

Si le conditionnement est par sachet de 1 gélule la probabilité de conformité est 97%, il en est de même avec 2 gélules, 10 gélules, 50 gélules. C'est à dire que le conditionnement n'a rien à voir avec la probabilité de proportion de gélules conformes. (Nota. Cette question a été posée dans le chapitre "logique" dans le cadre d'un concours).

Exercice 3: *Voilà, je fais une étude sur une cohorte et j'observe des événements à 2 moments différents.*
*Par exemple le pourcentage de fumeur à T0 et à T1.*
*J'ai donc 2 pourcentages différents, et j'aimerais savoir si la différence est statistiquement significative.*
*Il s'agit bien de la même population mais les observations ont lieu à 2 moments différents.*
*Par ailleurs, dans cette population, j'ai des perdus de vue. Par exemple à T0 j'ai 30% de fumeurs dans ma population de 100 personnes...*
*à T1 j'en ai 25% mais ma population ne fait plus que 70 personnes...*
*Est-ce que je peux savoir si ma différence est significative ou non, et si oui, quel test utiliser ?*

En fait, on ne sait pas vraiment s'il s'agit d'un exercice dont ce serait l'énoncé, ou une sorte de "traduction" d'un exercice pour éviter une recopie mot à mot, ou une question portant sur un cas plus ou moins réel. Quoi qu'il en soit, on peut faire les observations suivantes :

1- il n'est pas formellement précisé quels événements sont observés. On peut supposer que ces événement sont en relation avec le tabac, mais l'événement est-il du type "arrêter de fumer", ou "avoir contracté telle maladie" ou "ne pas avoir contracté telle maladie" ou "avoir changé d'activité professionnelle". Cette liste n'est pas limitative.
2- il est précisé que "il s'agit bien de la même population", mais on ne sait s'il s'agit réellement des mêmes individus. Si c'était le cas, sur les 70 personnes connues à l'instant T1, on pourrait savoir lesquels étaient fumeurs au temps T0, ce qui n'est pas le cas.

Concernant la première observation, il manque l'information importante qu'on peut exprimer par les termes "indépendance des éléments extérieurs" et "relation directe entre le fait connu (tabagisme) et l'élément observé (maladie [?]).
Concernant la seconde observation, calculons les cas possibles

1- Les 70 personnes non perdus de vue n'étaient pas fumeurs au temps T0 c'est à dire que les 30 personnes perdus de vue étaient fumeurs.
2- Les 30 personnes perdues de vue n'étaient pas fumeurs
3- La solution réelle est forcément entre ces deux extrêmes

Dans le cas 1 : 25% sont maintenant fumeurs, alors qu'ils ne l'étaient pas au temps T0. Cette situation n'est pas du tout invraisemblable, si le temps T0 concerne des élèves de 2$^{nd}$ et le temps T1, les mêmes cinq ans plus tard. En ce cas, la proportion de fumeurs au temps T1 est de l'ordre de 30% + 70x25% ~ 47.5%.

Dans le cas 2 : au temps T0, sur les 70 personnes que l'on n'a pas perdu de vue 30% étaient fumeurs, soit 21 personnes. Au temps T1, il n'y a plus que 25% de fumeurs, soit 17 personnes, autrement dit 4, personnes ont arrêté de fumer, ce qui n'est pas invraisemblable.

Le cas 3 : on a deux bornes 47% et 25%, soit pratiquement du simple au double.

Je laisse au lecteur le soin de conclure.

Utilisation de l'écart-type dans un contexte professionnel.

Soit une mesure à réaliser. Cette mesure n'est pas une mesure directe, c'est à dire qu'elle sera le résultat d'un calcul comprenant les résultats de plusieurs mesures, elles-mêmes pouvant être aussi le résultat d'un calcul. Il s'agit donc d'un cas parfaitement général qui comporte des mesures "élémentaires" qui sont des mesures directes et différentes opérations suivant des procédures précises pour parvenir au résultat final.
Quelque soit le contexte, on peut toujours se ramener à ce schéma.

Dans la pratique, pour effectuer la mesure dont on parle, on utilisera des appareils dont on connaît l'écart-type, celui-ci a été calculé par le fabricant qui a lui-même procédé suivant ce même schéma.
On se trouve à chaque instant à devoir calculer l'écart type d'une mesure, connaissant l'écart type de chacune des deux ou plusiieurs mesures du niveau inférieur.
Deux mesures peuvent se combiner de deux façons différentes, soit en série, soit en parallèle. La comparaison avec les montages de résistances électriques est une bonne image. On retrouve cette dualité dans de très nombreux contextes.

Les écarts-type se combinent quadratiquement. En série, le facteur est racine(N), en parallèle, le facteur est 1/racine(N), N étant le nombre de mesures de niveau inférieur. La démonstration de ceci dépasse le cadre de ce papier.
Exemples :
On mesure une grande distance avec une chaîne d'arpenteur. Si l'écart-type sur une portée est 1.5 cm, alors l'écart-type sur 25 portées sera 1.5 * 5 = 7.5 cm.
On mesure un angle avec un théodolite dont l'écart-type est 1.5cg. Si on fait 25 tours, c'est à dire que l'on mesure 25 fois le même angle, alors l'écart-type de la moyenne arithmétique des 25 lectures sera 1.5/5 = 3mg.

Remarquons qu'à aucun moment on n'a parlé d'intervalle de confiance. On aura, finalement, compte tenu de la procédure, des appareils de mesure et des calculs, un résultat numérique assorti d'une certaine précision, donnée par l'écart-type.

A chaque étape et à chaque niveau, l'écart-type est une valeur calculée. Il n'y a qu'à la première étape, c'est à dire la mesure directe de base que l'écart-type n'a pas été calculé mais fixé. Le détail de cette phase dépend de l'appareil concerné, il peut avoir été fait par mesures directes un grand nombre de fois, par expérience du fabricant en la matière ou de toute autre façon.

On observe que ce calcul d'intervalle de confiance est une source inépuisable d'exercices. On parle aussi d'intervalle de fluctuation. Ces deux notions sont sources de nombreuses discussions, généralement peu constructives. Les termes à préférer sont "précision" et "tolérance". Enfin, il y a lieu de proscrire le calcul d'un intervalle de confiance sur l'écart-type (à moins que l'on précise qu'il s'agit d'un exercice de calcul sur des notions abstraites).