Société de Calcul Mathématique SA
Outils d'aide à la décision
depuis 1995

La présentation des résultats d'une mesure :

Analyse critique de la conception des histogrammes et recommandations

par Bernard Beauzamy

juin 2019

I. Introduction

Lors du déroulement d'un process industriel, des mesures sont faites : résistance mécanique d'une pièce, teneur en un minerai, diamètre d'un objet, etc. On répartit ces informations en classes, généralement de même taille. On compte combien de mesures sont tombées dans chaque classe et on représente graphiquement le résultat. Cette pratique, très ancienne, est absolument courante ; nous l'appellerons ici "histogramme représentatif". Nous montrerons, et ceci est très important pour les démonstrations de sûreté, qu'il faut se garder d'utiliser un tel histogramme pour anticiper des résultats futurs ; il faut y substituer un autre type d'histogramme, que nous appellerons "histogramme prédictif".

Celui qui réalise un histogramme prête généralement peu d'attention au choix du nombre de classes. Il s'attend à ce que l'interprétation du résultat dépende de manière essentielle des mesures elles-mêmes. Cette vision est complètement erronée : quelles que soient les mesures, un choix inapproprié du nombre de classes conduit à une disparition complète de l'information. Il en va souvent ainsi dans les approches probabilistes : le choix de l'univers de référence conditionne la taille des zones dangereuses, que l'on peut ainsi faire passer pour négligeables. Dans une démonstration de sûreté, c'est inacceptable.

Il faut donc exiger, notamment lors des démonstrations de sûreté, que le choix du nombre de classes fasse l'objet d'un examen attentif et d'une justification, nécessairement fondée sur des éléments externes (en particulier sur la précision requise). Une banque, une compagnie d'assurances, vont la rencontrer les mêmes difficultés lorsqu'il s'agit d'évaluer les risques.

Nous montrons enfin que, si l'on prend la peine de se préoccuper de l'erreur de mesure (qui est généralement assez bien connue), toutes ces difficultés disparaissent : la décomposition en classes n'est plus utile.

II. Présentation du problème

Nous considérons un process, sur lequel une mesure est réalisée ; elle se traduit par un nombre réel. C'est le cas pour les process industriels (mesure de température, de pression, de composition chimique, etc.). C'est le cas pour les ventes (nombre de chaussures vendues par jour, selon la pointure), pour une compagnie d'assurances : nombre de sinistres par an, rangés par classes de gravité (remboursement associé, mesuré en Euros).

Cette énumération est intéressante, car elle illustre bien le problème :

- Pour les ventes de chaussures, la répartition en classes est immédiate, puisqu'il s'agit de la pointure. On a ici une répartition dite "discrète". On peut admettre qu'il n'y a aucune incertitude sur le résultat.

- Pour la compagnie d'assurances, on peut décider de compter en Euros, milliers d'Euros, millions d'Euros, etc. Le choix de l'unité va conditionner le nombre de classes et résultera en une incertitude sur le résultat : si nous comptons en milliers d'Euros, deux sinistres respectivement à 1000 et 1999 Euros seront considérés comme équivalents. Il n'y avait pas d'incertitude initiale (le coût du sinistre est en principe parfaitement connu), mais nous en avons introduit une, du fait de la répartition en classes.

- Pour la mesure d'une température ou du diamètre d'un objet, lors d'un process industriel, il y a une incertitude de mesure (généralement irréductible) et aucune décomposition en classe ne s'impose naturellement : elle est à la discrétion de l'ingénieur qui traite les données.

La présentation graphique de l'histogramme est évidemment destinée à montrer une compréhension globale du process : voilà ce que nous avons le plus vendu, voilà la variabilité du diamètre de nos pièces. Mais, et ceci est très important, elle est supposée avoir aussi une valeur prédictive : voici ce à quoi on peut s'attendre dans l'avenir, avec forte probabilité.

Dans la suite, nous notons N le nombre de mesures ; les valeurs mesurées sont notées $x_1,...,x_N$.

Nous pouvons toujours supposer que les mesures sont entre 0 et 1, en remplaçant x par $\frac{x-m}{M-m}$, où $[m,M]$ est un intervalle contenant toutes les mesures. Cela simplifie la présentation. Nous notons K le nombre de classes, qui seront donc :

$$B_1=[0,\frac{1}{K}[,\quad B_2=[\frac{1}{K},\frac{2}{K}[,...,B_K=[\frac{K-1}{K},\frac{K}{K}[$$

La question de savoir si les intervalles doivent être ouverts ou fermés à chaque extrémité est sans importance ici.

Une fois les classes construites, on compte combien de fois la mesure est tombée dans chacune : soit n_1 le nombre de fois où la mesure est dans l'intervalle B_1, n_2 pour B_2, etc. jusqu'à n_K pour B_K. Bien entendu, $n_1+\cdots+n_K=N$.

III. Deux types d'histogrammes

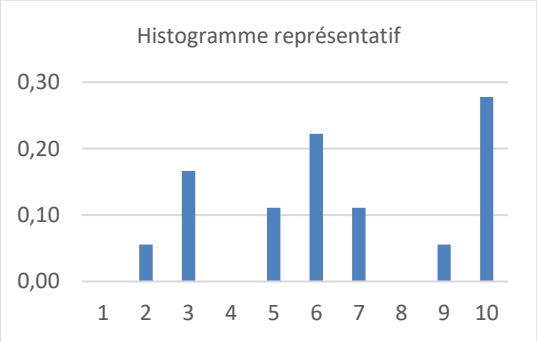
Nous avons alors deux manières de présenter les résultats sous forme d'histogramme, et il ne faut surtout pas les mélanger ; toute la difficulté vient de là.

A. Histogrammes représentatifs

Dans la classe B_k on met la quantité $\frac{n_k}{N}$: c'est l'approche la plus répandue. Elle permet de comparer les différentes classes. Si on n'a vendu aucune paire de chaussures de pointure 45, si aucune température ne s'est trouvée entre 35°C et 36°C, on met 0 dans la classe correspondante. C'est clair pour tout le monde. Nous dirons qu'il s'agit d'un "histogramme représentatif", parce qu'il permet de visualiser facilement les résultats de mesure.

Notons bien que, dans cette approche, les valeurs précises des mesures ont entièrement disparu : on ne conserve que les nombres d'occurrences $n_1,...,n_K$.

Voici un exemple. Les résultats de mesure sont les nombres 0,12, 0,22, 0,25, 0,28, 0,42, 0,45, 0,52, 0,55, 0,57, 0,58, 0,63, 0,66, 0,84, 0,92, 0,94, 0,95, 0,97, 0,98. On va les ranger dans des classes de largeur 1/10. Dans ces conditions, le nombre de résultats pour chaque classe est respectivement 0, 1, 3, 0, 2, 4, 2, 0, 1, 5 ; en divisant par le total (ici 18), on obtient l'histogramme représentatif suivant :



B. Histogramme prédictifs

Les choses sont fondamentalement différentes : on veut se servir des observations du passé pour établir une loi de probabilité en ce qui concerne le futur. Connaissant mes ventes de chaussure de l'an passé, quelle est la probabilité que le prochain client chausse du 45 ?

Et là, l'histogramme représentatif est complètement en défaut, puisqu'il met 0 là où aucune observation n'a été faite. Or, ce n'est pas parce qu'aucun client l'an passé n'a chaussé du 45 qu'il en sera de même dans l'avenir.

Pour construire correctement l'histogramme prédictif, il faut avoir recours à la notion de "taux de risque" (voir le livre [MPPR] pour les définitions détaillées). Le taux de risque est précisément la probabilité associée à chaque classe. Autrement dit, il répond à la question : connaissant mon historique, quelle est la probabilité que le prochain client chausse du 45 (ou toute autre pointure) ?

Rappelons que la loi conjointe des K taux de risque a pour densité :

$$f(\lambda_1,...,\lambda_K) = c\lambda_1^{n_1} \cdots \lambda_K^{n_K}$$

où c est une constante, calculée plus loin. On a $\lambda_1 + \cdots + \lambda_K = 1,$ parce que si une mesure est faite, il faut bien qu'elle tombe dans l'une quelconque des K classes. Pour calculer la constante $c,$ posons :

$$I(n_1,...,n_K) = \int_S \lambda_1^{n_1} \cdots \lambda_K^{n_K} d\lambda_1 ... d\lambda_K$$

où S est le simplexe (intersection de demi-espaces) :

$$S = \left\{ (\lambda_1,...,\lambda_K) ; \lambda_k \geq 0 \ \forall k, \sum_{k=1}^{K} \lambda_k = 1 \right\}$$

Alors on sait (voir [MPPR]) que :

$$I(n_1,...,n_K) = \frac{n_1! \cdots n_K!}{(N+K-1)!}$$

et donc la valeur de la constante c est :

$$c = \frac{(N+K-1)!}{n_1! \cdots n_K!}$$

On a donc la formule complètement explicite :

$$f(\lambda_1,...,\lambda_K) = \frac{(N+K-1)!}{n_1! \cdots n_K!} \lambda_1^{n_1} \cdots \lambda_K^{n_K}$$

L'espérance de la 1ère classe (taux de risque moyen pour cette classe) sera :

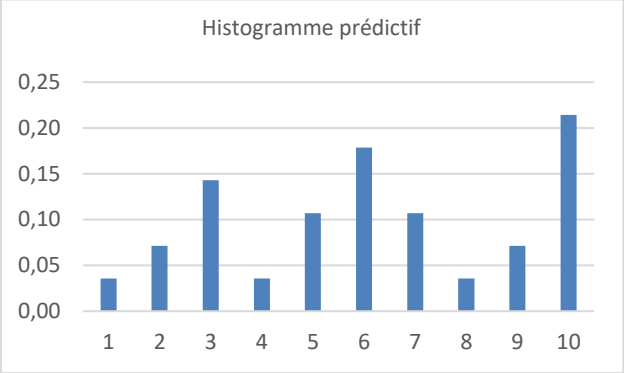
$$E_1 = E(\lambda_1) = \int_S \lambda_1 f(\lambda_1,...,\lambda_K) d\lambda_1 ... d\lambda_K = \frac{I(n_1+1, n_2,..., n_K)}{I(n_1, n_2,..., n_K)} = \frac{\frac{(n_1+1)! n_2! \cdots n_K!}{(N+K)!}}{\frac{n_1! \cdots n_K!}{(N+K-1)!}} = \frac{n_1+1}{N+K}$$

et de même pour toutes les autres classes :

$$E_k = \frac{n_k+1}{N+K}$$

On a bien $\sum_{k=1}^{K} E_k = 1$.

Autrement dit, la théorie générale des taux de risque conduit à attribuer à la $k^{ème}$ classe la valeur $\frac{n_k + 1}{N + K}$, au lieu de $\frac{n_k}{N}$ pour l'histogramme représentatif. C'est à peu près la même chose si n_k et N sont grands devant K, mais c'est fondamentalement différent si $n_k = 0$. Pour l'histogramme représentatif, à une classe vide on attribuera la probabilité 0, pour l'histogramme prédictif, on lui attribuera la probabilité $\frac{1}{N + K}$, qui n'est jamais nulle. Voici un exemple, avec les mêmes nombres que précédemment (0, 1, 3, 0, 2, 4, 2, 0, 1, 5) :



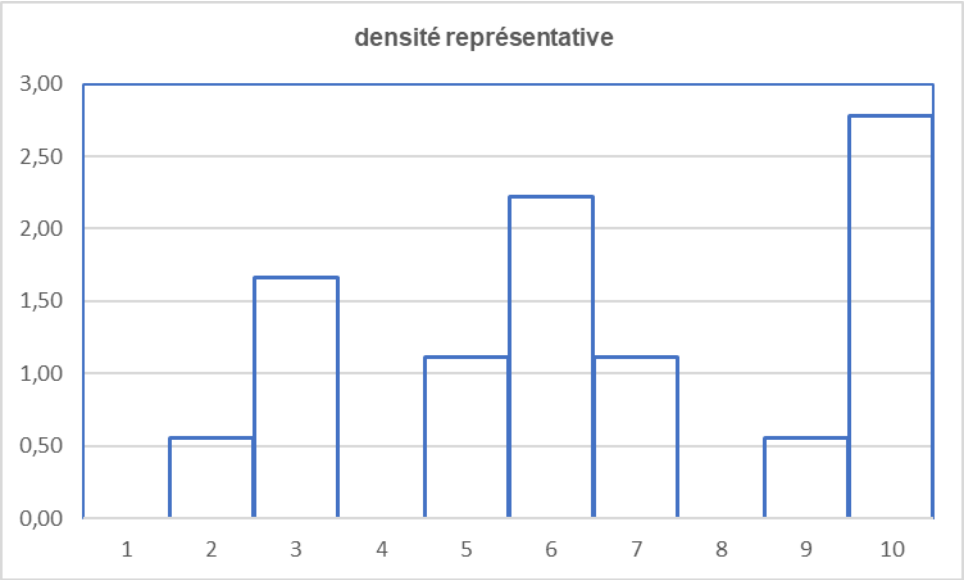
L'histogramme représentatif doit impérativement être évité dans le cadre d'une démonstration de sûreté : ce n'est pas parce qu'une situation ne s'est jamais produite dans le passé qu'on est assuré qu'elle ne se produira pas dans l'avenir. Et ceci est d'autant plus vrai que, en augmentant le nombre de classes, à nombre de mesures donné, on augmente mécaniquement le nombre de classes vides, donc le nombre de situations abusivement "sûres". Nous insisterons beaucoup sur ce point par la suite.

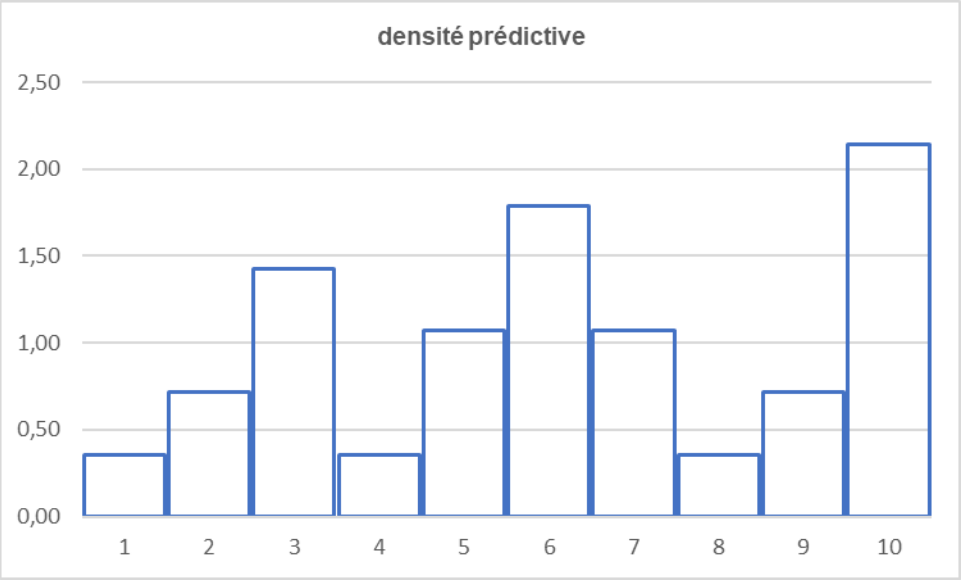
C. Représentation des histogrammes sous forme de densité de probabilité

Cherchons à représenter ces résultats au moyen d'une densité sur $[0,1]$. Par définition, l'aire sous la courbe doit valoir 1. Pour l'histogramme représentatif, c'est facile : on met la valeur $h_k = \frac{n_k K}{N}$ au-dessus du $k^{ème}$ intervalle, qui représente la $k^{ème}$ classe. Comme cet intervalle a largeur $\frac{1}{K}$, l'aire pour chaque intervalle sera $\frac{h_k}{K} = \frac{n_k}{N}$ et la somme des aires associées vaut bien 1.

Pour l'histogramme prédictif, il faut mettre au-dessus de la classe k une hauteur égale à :

$$h_k = \frac{(n_k + 1)K}{N + K}$$





IV. Augmenter le nombre de classes

Considérons maintenant la situation où N, nombre d'expériences, est fixé, et K, nombre de classes, tend vers l'infini : on veut passer du discret au continu. Ceci se fait, bien entendu, en diminuant la taille des classes ; pour les coûts, par exemple, on passe du million d'Euros au millier, puis à l'Euro.

Alors, pour chaque classe, lorsque K est assez grand, le nombre d'occurrences de cette classe vaudra en général 0 ou 1 (le plus souvent 0). Lorsque le nombre de classes dépasse le nombre d'expériences, il y a nécessairement $K-N$ classes qui ne reçoivent aucune valeur, c'est-à-dire pour lesquelles $n_k = 0$.

Voyons ce que l'on obtient, lorsque K augmente, pour chacun des deux histogrammes :

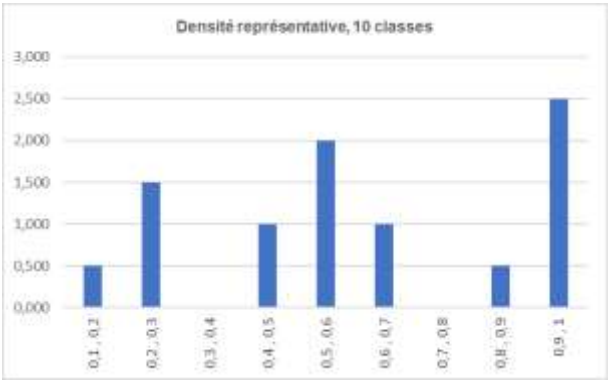
– Histogramme représentatif

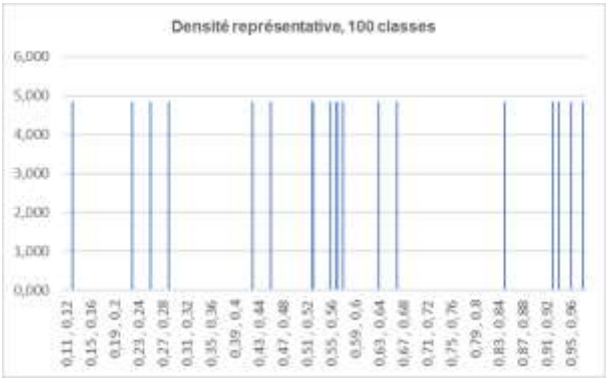
A chaque fois que $n_k = 0,$ c'est-à-dire la plupart du temps, la fonction limite est 0 ;

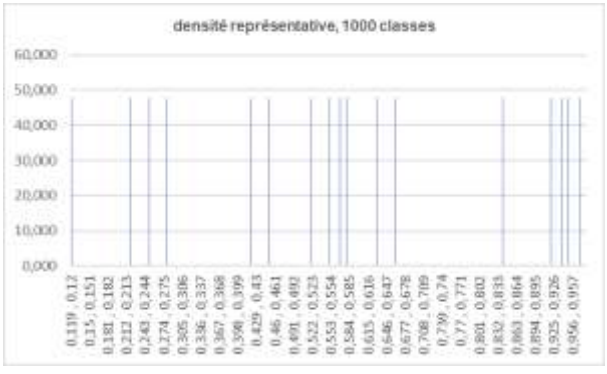
A chaque fois que $n_k = 1,$ la fonction limite est équivalente à $\frac{K}{N} \to +\infty.$

Autrement dit, la fonction limite sera faite d'un nombre fini (précisément N) de "bumps", de largeur $\frac{1}{K} \to 0$ et de hauteur $\frac{K}{N} \to +\infty.$

Considérons la situation où les résultats de la mesure sont, comme précédemment :
0,12, 0,22, 0,25, 0,28, 0,42, 0,45, 0,52, 0,55, 0,57, 0,58, 0,63, 0,66, 0,84, 0,92, 0,94, 0,95, 0,97, 0,98.
Voici les histogrammes construits, respectivement avec $K = 10, 100, 1000$ classes.





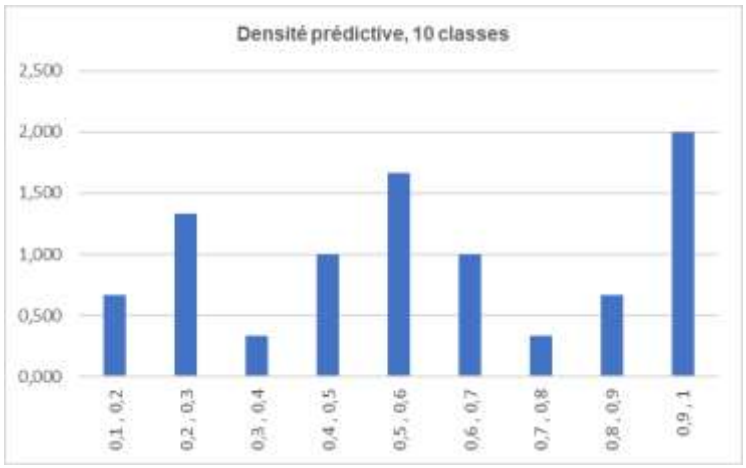


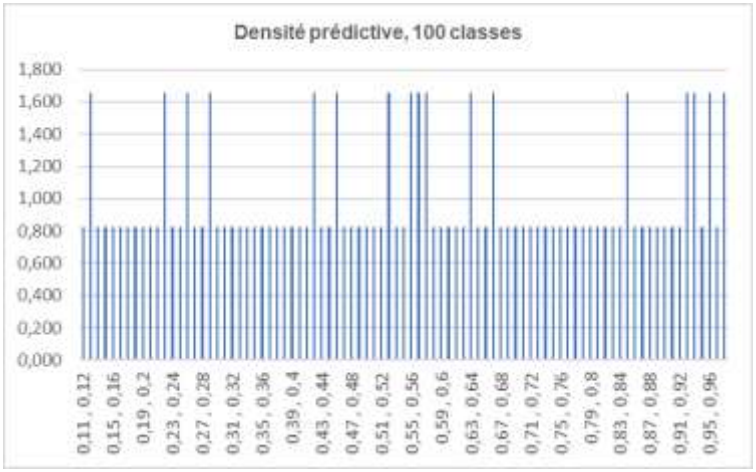
– Histogramme prédictif

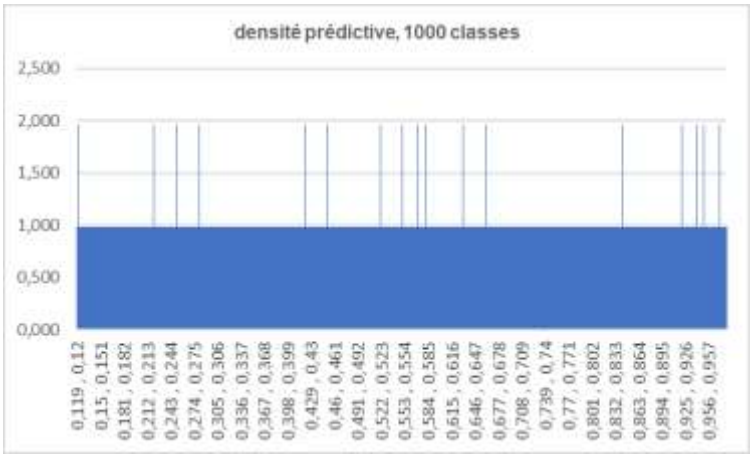
Les choses sont radicalement différentes : lorsque $K \to +\infty$, on a :

$$h_k \sim n_k + 1$$

et donc $h_k \to 1$ pour la plupart des classes (pour $K - N$) et $h_k \to 2$ pour K classes. Autrement dit, la fonction densité, à la limite, sera la loi uniforme sur $[0,1]$. Voici les densités prédictives, comme précédemment :







On constate donc ceci : lorsque le nombre de classes tend vers l'infini, l'information résultant de N expériences disparaît ; on se retrouve avec la loi uniforme, comme si aucune expérience n'avait été faite. On peut dire ainsi que, lorsque le nombre de classes augmente (c'est-à-dire lorsque le besoin en précision augmente), la valeur prédictive d'un nombre quelconque d'expériences devient de plus en plus faible.

V. Erreurs de mesure

Cet inconvénient, lié à la conception même des histogrammes, disparaît si on utilise les erreurs de mesure.

Disons que l'erreur de mesure est représentée par une densité de probabilité φ ; pour simplifier, on peut la supposer symétrique et même supposer que c'est une gaussienne. La probabilité que l'erreur soit supérieure à ε est $P\left(Erreur > \varepsilon\right) = \int_{\varepsilon}^{+\infty} \varphi(t)\,dt.$

En principe, cette représentation de l'erreur résulte de la calibration des instruments de mesure (voir le livre [MPPR]) ; elle est approximativement connue de l'industriel. On peut toujours faire une hypothèse préalable, quitte à la changer ensuite. La théorie s'accommode même des situations où φ n'est pas la même d'un bout à l'autre de la gamme de mesure (ce qu'on appelle un "facteur d'échelle" : l'instrument n'a pas la même précision partout). La fonction φ peut parfaitement n'être pas symétrique (les erreurs vers le bas sont, par exemple, plus nombreuses que les erreurs vers le haut).

Dans la suite, pour présenter la théorie, nous prendrons le cas où φ est une gaussienne centrée, de variance adaptée à chaque situation. Il faut bien sûr tronquer, pour respecter le fait que les valeurs sont entre 0 et 1, et renormaliser, pour que l'intégrale reste égale à 1, compte-tenu de la troncature. Si nous prenons une gaussienne :

$$\varphi(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

et si les essais ont donné les résultats numériques $x_1, ..., x_N$ (c'est le tableau Excel), alors la densité de probabilité du paramètre associé sera par définition, pour $0 \le t \le 1$:

$$f(t) = \frac{1}{N} \sum_{n=1}^{N} \frac{\varphi(t - x_n)}{\int_0^1 \varphi(t - x_n)\,dt}$$

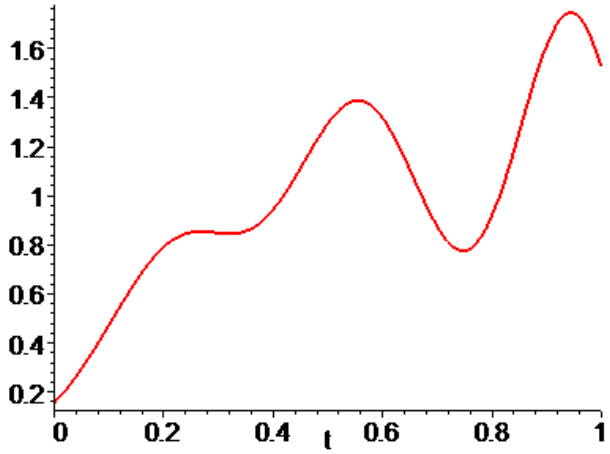
Pour établir cette formule, considérons d'abord le cas d'une seule mesure, x_1 ; on a :

$$f(t) = \varphi(t - x_1)$$

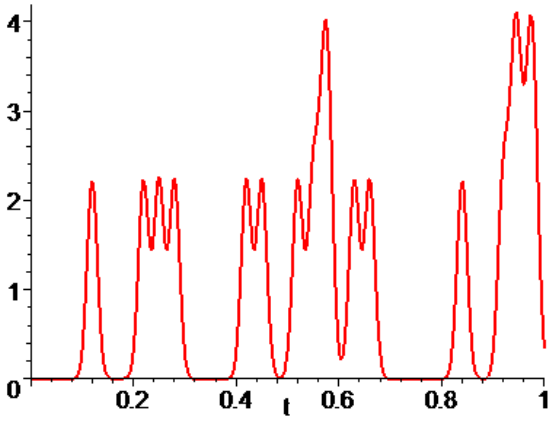
C'est bien une densité de probabilité ; elle est maximale en x_1 et décroît au fur et à mesure que l'on s'éloigne de x_1 ; il faut la tronquer pour rester entre 0 et 1 et la renormaliser. Dans le cas général, le choix de la formule moyenne $\frac{1}{N}\sum_{n=1}^{N}$ provient du fait que toutes les mesures x_n sont considérées comme d'importance équivalente : nous n'avons pas de raison de douter de certaines d'entre elles, ou d'en privilégier d'autres.

Ici, nous avons une densité continue et il n'y a plus aucun problème de définition.

Voici ce que l'on obtient dans le cas des 18 mesures ci-dessus, lorsque la densité de probabilité d'erreur est une gaussienne de variance 0.1 :



et voici le cas où la variance de la loi d'erreur est 0.01 :



Bien entendu, la fonction de répartition s'en déduit immédiatement, par intégration.

Comme loi de probabilité d'erreur, on peut utiliser une gaussienne (comme ci-dessus), une fonction triangle, ou même une loi uniforme sur un intervalle. Dans ce dernier cas, le résultat est alors discontinu, mais différent de ce que donne un histogramme, car la loi d'erreur ne fait pas référence à une classe particulière. Bien entendu, lorsqu'elle est disponible, on utilisera la loi d'erreur provenant d'un retour d'expérience.

VI. Détection des valeurs aberrantes

La méthode consistant à introduire les erreurs de mesure, pour satisfaisante qu'elle soit, n'est pas adaptée à la détection des valeurs aberrantes, parce que les "bosses" associées à de telles valeurs vont être imperceptibles, du fait du coefficient $\frac{1}{N}$ devant la somme. Les valeurs aberrantes doivent être traitées en se servant de l'histogramme représentatif.

VII. Bibliographie

[MPPR] Bernard Beauzamy : Méthodes Probabilistes pour l'étude des phénomènes réels. Ouvrage édité et commercialisé par la Société de Calcul Mathématique SA, ISBN 2-9521458-0-6, ISSN 1767-1175. Mars 2004. Seconde Edition, 2016.