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

# Construction d'une densité de probabilité continue

# à partir de relevés expérimentaux

Bernard Beauzamy

27/01/2019

## 1. Présentation du problème

Habituellement, dans les contrats que nous traitons, le résultat d'essais est donné par un tableau Excel où apparaissent des valeurs de mesure : on dispose, en colonne, de $N$ résultats numériques, représentant les différentes mesures d'un paramètre, par exemple une résistance mécanique, la teneur en un polluant, etc. On cherche à reconstituer la densité de probabilité de la variable d'intérêt.

La manière habituelle de procéder est de construire un histogramme, en découpant les valeurs possibles en tranches. Une fois l'histogramme construit, on divise par le nombre total d'occurrences pour obtenir une densité de probabilité pour le paramètre en question.

Mais cette façon de procéder, très courante, présente trois inconvénients :

- Le choix des tranches est arbitraire, aussi bien en ce qui concerne les bornes que la largeur des intervalles ; des choix différents conduiront à des conclusions différentes ;

- On perd de l'information en construisant l'histogramme, puisque toutes les valeurs à l'intérieur d'une même classe sont considérées comme identiques ;

- On se retrouve avec une densité discontinue (constante par paliers), et ces discontinuités sont arbitraires, puisqu'elles résultent de la discrétisation faite.

Ces inconvénients résultent du fait que les mesures (valeurs portées dans le tableau) sont considérées comme exactes. Ils disparaissent si on fait intervenir l'erreur sur la mesure. Voyons comment.

## 2. Prise en compte de l'erreur de mesure

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

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 $\varphi$ n'est pas la même d'un bout à l'autre de la gamme de mesure ; c'est ce qu'on appelle un "facteur d'échelle" : l'instrument n'a pas la même précision partout. La fonction $\varphi$ peut parfaitement n'être pas symétrique : les erreurs vers le bas sont, par exemple, plus importantes que les erreurs vers le haut, ou l'inverse.

Dans la suite, pour présenter la théorie, nous prendrons le cas où $\varphi$ est une gaussienne centrée, de variance $\sigma = 0.1$ :

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

Imaginons que les essais aient donné les résultats numériques $x_1, ..., x_N$ (c'est le tableau Excel). Alors la densité de probabilité du paramètre associé sera par définition :

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

Nous avons maintenant une densité continue et il n'y a plus aucun problème de définition.

Voici ce que l'on obtient dans le cas de 4 mesures : 1, 1.2, 2, 2.3 ; la densité d'erreur est une gaussienne de variance 0.1 :

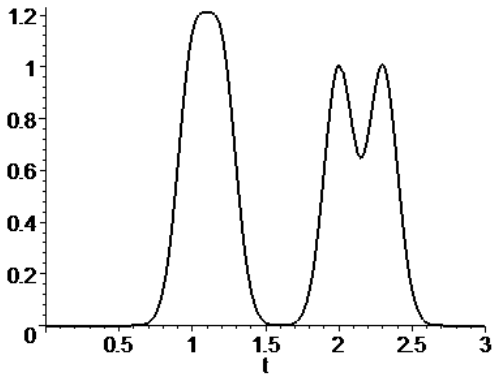


*Figure 1 : densité continue, 4 enregistements*

Les deux mesures proches (1 et 1.2) donnent naissance à une même grosse bosse ; les deux mesures plus différentes (2 et 2.3) donnent naissance à deux pics distincts.
Voici ce que l'on obtient si l'on rajoute la mesure 2.4 :

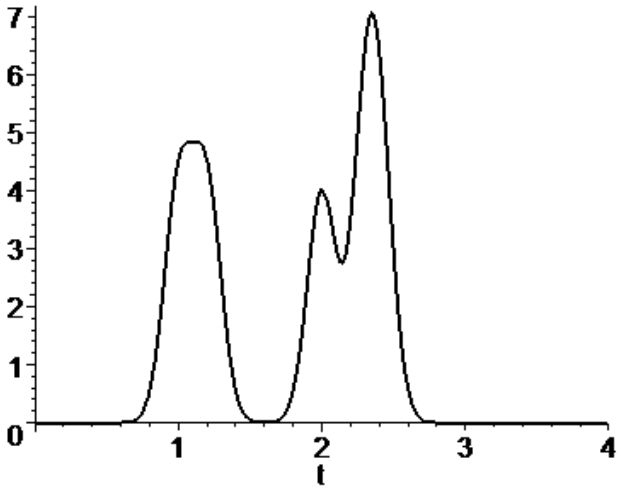


*Figure 2 : densité continue, 5 enregistements*

Le pic correspondant voit sa hauteur s'accroître.

Le choix dans (1) de la 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.

### 3. Lien avec l'histogramme

Nous allons voir le lien entre la construction de l'histogramme et la définition (1). Nous prenons l'exemple des 5 valeurs expérimentales données plus haut.

Pour un intervalle quelconque $A$, nous notons $1_A$ la fonction caractéristique de $A$ : elle est définie par $1_A(t)=1$ si $t \in A$, 0 sinon.

Prenons pour la définition de l'histogramme les intervalles $[0,1[,[1,2[,[2,3[,[3,4[$. Il y a deux points de mesure dans le second, trois dans le troisième ; nous aurons donc la fonction :

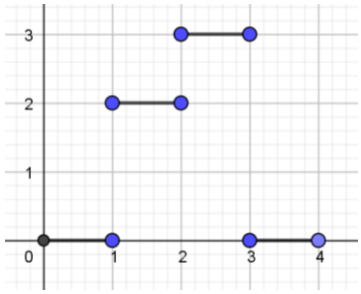


*Figure 3 : construction de l'histogramme*

Elle vaut 2 sur l'intervalle $[1,2[$ , 3 sur l'intervalle $[2,3[$ , 0 ailleurs.

Cette fonction, notée $f$, peut être représentée comme somme de fonctions caractéristiques d'intervalles. Notons :

$$A_1 = [0,1[,\ A_2 = [-0.2,0.8[,\ A_3 = [0,1[,\ A_4 = [-0.3,0.7[,\ A_5 = [-0.4,0.6[$$

On a :

$$f(t) = 1_{A_1}(t - x_1) + 1_{A_2}(t - x_2) + 1_{A_3}(t - x_3) + 1_{A_4}(t - x_4) + 1_{A_5}(t - x_5)$$

En effet, $1_{A_1}(t - x_1)$ vaut 1 si $0 \le t - x_1 < 1$, ou $1 \le t < 2$ et de même pour les autres termes.

L'histogramme peut donc être représenté comme somme de fonctions caractéristiques ; l'idée de la formule (1) est de remplacer ces fonctions caractéristiques, nécessairement discontinues, par des fonctions continues, représentant la probabilité d'erreur.

Certes, la fonction erreur, notée $\varphi$ plus haut, est assez arbitraire, mais elle peut être précisée par expérimentation, alors que, pour la construction de l'histogramme, les choix qui doivent être faits restent arbitraires quoi qu'on fasse.

### 4. Détection des valeurs aberrantes

On sait que la constitution de l'histogramme sert en particulier à la détection des valeurs aberrantes. Nous remarquons que la présente méthode n'est pas adaptée à cette question, parce que les "bosses" associées à de telles valeurs vont être imperceptibles, du fait du coefficient $\frac{1}{N}$ devant la somme. La recherche des valeurs aberrantes sera faite en se servant de l'histogramme.

### 5. Utilisation prédictive

La fonction de répartition se déduit immédiatement de (1) par intégration. La probabilité de trouver une valeur $> A$ lors d'une mesure ultérieure est :

$$p_A = P(mesure > A) = \int_A^{+\infty} f(t)\,dt$$

La probabilité d'avoir, dans l'avenir, exactement $n'$ mesures $> A$ alors que l'on a fait $N'$ essais est :

$$p(n', N'; n, N) = \binom{N'}{n'} p_A^{n'} (1 - p_A)^{N'-n'}.$$

### 6. Référence

[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.