Société de Calcul Mathématique, S. A.
Algorithmes et Optimisation

Estimation d'intervalles de confiance : approche statistique et approche probabiliste

Bernard Beauzamy

Mai 2010

En résultat d'un calcul, on demande presque toujours un intervalle de confiance autour de la valeur numérique obtenue. C'est un élément essentiel de la prise de décision, en particulier pour les politiques publiques (environnement, santé).

Il y a deux approches possibles, l'une purement statistique, l'autre probabiliste. Nous allons en donner les grandes lignes, de manière à bien montrer les différences. Nous allons prendre deux exemples : qualité de l'eau (notre contrat AEE) et épidémiologie (IRSN, RTE) : les champs d'application sont différents, mais les principes sont exactement les mêmes. Comme on va le voir, l'approche statistique est très simple : à partir du relevé des mesures, on en déduit un intervalle de confiance, sans même se soucier de la qualité desdites mesures (!). L'approche probabiliste n'est pas plus difficile à mettre en œuvre, et elle permet d'incorporer l'incertitude sur la mesure, comme paramètre externe.

I. Approche statistique

1. Qualité de l'eau

On dispose de N mesures, notées $x_1,...,x_N$. On considère que chaque mesure est le résultat de l'observation d'une variable aléatoire, notée X_i $(i=1,...,N)$: l'observation de la $i-$ème variable a donné la valeur précise x_i. On forme la moyenne $Z=\frac{1}{N}\sum_{i=1}^{N} X_i$ (il s'agit aussi parfois de moyenne pondérée). Si les variables sont indépendantes et de même loi, pour N assez grand, la variable Z suit (asymptotiquement) une loi de Gauss. On se sert de la moyenne de l'échantillon, à savoir $m=\frac{1}{N}\sum_{i=1}^{N} x_i$ et de la variance de l'échantillon, à savoir $\sigma^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-m)^2$, pour estimer les paramètres de la loi de Z. Une fois ceci fait, la variable Z est entièrement spécifiée : c'est une variable gaussienne, de moyenne m et d'écart-type σ. On peut alors déterminer un intervalle de confiance, par exemple à 95 %, en se servant des tables de la loi de Gauss.

2. Epidémiologie

On dispose généralement d'un tableau de quatre nombres : nombre de décès et population totale, pour une zone "à risques" et une zone de comparaison. On note X_i la variable aléatoire qui dit, dans la zone 1, si la $i-$ème personne décède pour la raison qui nous intéresse, 0 sinon, et de même Y_j pour la zone 2. Les moyennes $\frac{1}{N_1}\sum_{i=1}^{N_1} X_i$ et $\frac{1}{N_2}\sum_{j=1}^{N_2} Y_j$ vont suivre des lois de Gauss. Pour leur quotient (appelé "risque relatif"), on passe aux logarithmes et on a une somme de deux lois log-normales, dont les paramètres sont estimés comme précédemment.

L'approche statistique se caractérise donc par ceci : à partir de la mesure, on fabrique une variable aléatoire dont on connaît la loi.

Il y a bien sûr des difficultés méthodologiques : les variables élémentaires X_i doivent être indépendantes et de même loi, ce qui n'est en rien assuré en pratique.

Mais la difficulté principale est celle-ci : l'erreur de mesure sur chacun des x_i n'est pas prise en compte. A partir de valeurs mesurées $x_1,...,x_N$, réputées précises, le calcul sort un encadrement pour la valeur moyenne de pollution (cas de la qualité des eaux) ou pour le risque relatif (cas de l'épidémiologie). En pratique, on dispose d'un logiciel : on rentre des valeurs de mesure, et il sort un encadrement !

Il s'agit donc d'un encadrement tout à fait factice, qui correspond à la description suivante : si la moyenne Z était effectivement une variable normale, avec les paramètres que nous lui attribuons, voici les valeurs entre lesquelles elle devrait se trouver. Mais l'incertitude sur les paramètres d'entrée n'est pas prise en compte !

II. La nécessité de tenir compte des erreurs de mesure

Dans les domaines de l'environnement et de l'épidémiologie, les erreurs de mesure ne peuvent être négligées, d'autant qu'il s'agit de sciences neuves. L'erreur ne porte pas tant sur la mesure elle-même que sur la nature de ce que l'on mesure. Par exemple :

- On parle de la mesure d'un polluant (en g/l), mais où et quand la mesure-t-on ? Suivant que la mesure est faite à un autre moment, ou à un autre endroit, elle peut donner des résultats différents.

- On cherche à déterminer la probabilité d'une maladie, par exemple Alzheimer au voisinage des lignes à haute tension. Mais diagnostique-t-on correctement cette maladie, et comment définit-on le "voisinage d'une ligne" ? Comment caractérise-t-on le temps qu'on y passe ?

III. L'approche probabiliste

Elle a au moins le mérite de chercher à tenir compte de l'erreur de mesure, de manière simplifiée. On considère que le résultat de la $i-$ème mesure est $x_i + E_i$: x_i valeur effectivement mesurée, E_i variable aléatoire représentant l'erreur de mesure (inconnue). On

fait les hypothèses suivantes sur les E_i : ce sont des variables indépendantes, de moyenne nulle et gaussiennes.

L'hypothèse "indépendantes" signifie que les erreurs commises à un endroit n'influent pas sur la mesure à un autre. Ce n'est pas toujours correct, surtout si ce sont les mêmes capteurs qui sont utilisés partout.

On suppose les E_i de moyenne nulle : les observations sont "sans biais".

On suppose les E_i gaussiennes. Grossièrement, cela signifie que les petites erreurs sont plus probables que les grandes.

Ces hypothèses sont réductrices, mais on peut toujours ensuite les remplacer par d'autres lorsque les lois d'erreur sont mieux connues (lorsqu'on a fait plus d'expériences). Elles sont acceptables pour commencer.

Ensuite, on décide de la précision : ceci est à dire d'expert. On décidera par exemple que la précision est à 10 % de la valeur de la mesure, et on prendra alors $\sigma_i = 0.1\, x_i$.

On voit ici que la précision de la mesure entre en ligne de compte et doit être renseignée par l'utilisateur : ceci est bien entendu indispensable. One ne peut concevoir de résultat où la précision de la mesure ne joue aucun rôle.

On peut alors déterminer l'intervalle de confiance pour la moyenne de manière très simple, puisque tous les paramètres sont connus. La moyenne vaut :

$$M = \frac{1}{N}\sum_{i=1}^{N}\left(x_i + E_i\right) = m + \frac{1}{N}\sum_{i=1}^{N} E_i$$

Les E_i étant des gaussiennes de moyenne nulle et de variance σ_i^2 (par exemple $\sigma_i^2 = 10^{-2}\, x_i^2$), la variable $\sum_{i=1}^{N} E_i$ est une gaussienne de moyenne nulle et de variance $\sum_{i=1}^{N}\sigma_i^2$. La variable $\frac{1}{N}\sum_{i=1}^{N} E_i$ sera donc une gaussienne de moyenne nulle et d'écart-type $\frac{1}{N}\left(\sum_{i=1}^{N}\sigma_i^2\right)^{1/2}$. On peut alors déterminer explicitement un intervalle de confiance à 95% pour cette variable, en se servant des tables de la loi de Gauss.