Société de Calcul Mathématique SA

Outils d'aide à la décision

depuis 1995

Calcul de l'aire sous le graphe de la fonction de répartition

Eléments théoriques et règles de calcul pratiques

Bernard Beauzamy

18/01/2020, rev. 28/07/2020

La fonction de répartition, notée $F(y)$, d'une variable aléatoire Y est définie par la formule :

$$F(y) = P(Y \leq y).$$

C'est donc une fonction croissante de y, à valeurs entre 0 et 1. L'aire sous le graphe joue un rôle essentiel dans notre méthode de "hiérarchisation de paramètres" : on conditionne Y par diverses contraintes de type $X < m$ ou $X > m$ et on compare les aires dans chaque cas.

1. Approche théorique

En théorie, le domaine de définition de F peut s'étendre de $-\infty$ à $+\infty$, mais en pratique il s'agit toujours d'un intervalle borné $[a,b]$. On a $F(a)=0$, $F(b)=1$.

La fonction F est la primitive de la densité f, nulle en a.

L'aire sous le graphe de F vaut :

$$A = \int_a^b F(y)\,dy$$

Par intégration par parties :

$$A=\int_a^b F(y)\,dy=\left[yF(y)\right]_a^b-\int_a^b yF'(y)\,dy=b-\int_a^b yf(y)\,dy=b-E(Y)$$

où $E(Y)$ est l'espérance de Y.

Nous allons voir que cette formule très simple demeure correcte lorsque la loi de Y n'est pas connue : on dispose seulement de relevés expérimentaux.

2. Cas discret : relevés expérimentaux

On a fait N expériences et on a observé les valeurs y_1 avec multiplicité n_1,...., y_K avec multiplicité n_K, avec bien sûr $n_1+\cdots+n_K=N$; avec cette notation, K est le nombre de valeurs distinctes. On peut bien sûr supposer que les y_k sont rangés en ordre croissant :

$$a<y_1<\cdots<y_K<b$$

Pour chaque valeur de y, la "probabilité" $P(Y\le y)$ est définie comme étant le nombre de fois où, dans la liste, une valeur $\le y$ a été rencontrée, divisé par le nombre total d'expériences, noté ici N.

La plus petite valeur rencontrée dans la liste est y_1 ; par conséquent, si $y<y_1$, on ne rencontre jamais de valeur inférieure à y, donc $F(y)=0$ pour tout $y<y_1$.

En $y=y_1$, la fonction F prend la valeur $\frac{n_1}{N}$ et de même pour tout y, $y_1\le y<y_2$.

De même, $F(y)=\frac{n_1+n_2}{N}$ si $y_2\le y<y_3$ et, plus généralement, pour tout $k=1,\dots,K-1$:

$$F(y)=\frac{n_1+\cdots+n_k}{N} \text{ si } y_k\le y<y_{k+1}$$

$$F(y)=\frac{n_1+\cdots+n_{K-1}}{N} \text{ si } y_{K-1}\le y<y_K$$

et finalement :

$$F(y)=1 \text{ si } y\ge y_K.$$

La fonction F est donc constante sur une succession d'intervalles, fermés à gauche, ouverts à droite. Les valeurs prises vont en croissant. Le fait que les intervalles de définition soient fermés ou ouverts à chaque extrémité est sans importance du point de vue de l'aire.

Le $k^{ème}$ intervalle, noté I_k, va de y_k à y_{k+1} pour $k=1,...,K-1$; le dernier intervalle, I_K, va de y_K à b. Il n'est pas utile d'introduire un "$0^{ème}$" intervalle, entre a et y_1, puisque la fonction F est nulle dessus.

La taille du $k^{ème}$ intervalle est donc $y_{k+1}-y_k$ pour $k=1,...,K-1$ et la taille du dernier est $b-y_K$.

L'aire sous le graphe de F est donc :

$$A=\sum_{k=1}^{K-1} \frac{n_1+\cdots+n_k}{N}\left(y_{k+1}-y_k\right)+b-y_K$$

Ce qu'on écrit :

$$\begin{aligned}A&=\frac{n_1}{N}\left(y_2-y_1\right)+\frac{n_1+n_2}{N}\left(y_3-y_2\right)+\cdots+\frac{n_1+\cdots+n_{k-1}}{N}\left(y_k-y_{k-1}\right)+\\&\qquad+\frac{n_1+\cdots+n_k}{N}\left(y_{k+1}-y_k\right)+\cdots+\frac{n_1+\cdots+n_{K-1}}{N}\left(y_K-y_{K-1}\right)+b-y_K\\&=-\frac{n_1}{N}y_1-\frac{n_2}{N}y_2-\cdots-\frac{n_k}{N}y_k-\frac{n_{K-1}}{N}y_{K-1}-\frac{n_K}{N}y_K+b\end{aligned}$$

et donc :

$$A=b-\frac{1}{N}\sum_{k=1}^{K}n_k y_k$$

La quantité $\frac{1}{N}\sum_{k=1}^{K}n_k y_k$ peut être considérée comme l'espérance "expérimentale" de Y : c'est celle qui résulte de l'échantillon de mesure. La formule :

$$A=b-E\left(Y\right)$$

demeure donc valable.

Notons que la valeur $E\left(Y\right)$ est simplement la moyenne de tous les résultats expérimentaux de Y ; il n'est pas nécessaire de les trier par ordre croissant pour la calculer.

3. Conditionnement de Y

Pour l'application de la méthode de hiérarchisation de paramètres, on est amené à conditionner Y par des situations du type $X<m$ ou $X>m$, où X est un paramètre quelconque. Le domaine de variation $\left[a,b\right]$ pour Y est évidemment toujours le même.

L'application des règles précédentes est alors très simple. Par exemple, pour la situation $X < m$, on va extraire du tableau initial toutes les lignes pour lesquelles $X < m$; soit T_1 le tableau de données ainsi réduit. On va calculer la moyenne des valeurs de Y apparaissant dans ce tableau réduit ; notons-la E_1 ; l'aire au-dessous du graphe de la fonction de répartition dans la situation $X < m$ sera :

$$A_1 = b - E_1$$

On fait de même avec la situation $X > m$; on extrait le tableau T_2 et on calcule la moyenne E_2 des valeurs de Y dans ce tableau. L'aire au-dessous du graphe de la fonction de répartition dans la situation $X > m$ sera :

$$A_2 = b - E_2$$

Remarque. – Pour l'utilisation de la méthode de hiérarchisation, il est recommandé de tracer les deux courbes $F_1(y) = P(Y \le y \mid X < m)$ et $F_2(y) = P(Y \le y \mid X > m)$ de manière à vérifier que l'une est constamment au-dessous de l'autre.

4. Méthode pour déterminer le positionnement relatif des courbes, sans tracer d'histogramme.

On reprend le tableau de tous les résultats de Y. On le range par ordre de y croissants. On a deux sous-tableaux, l'un pour $X < m$ (lignes rouges), l'autre pour $X > m$ (lignes bleues) ; évidemment, une ligne ne peut être à la fois rouge et bleue. On élimine celles qui ne sont ni rouges, ni bleues. Soit N_1 le nombre de lignes rouges et N_2 le nombre de lignes bleues.

On dispose alors d'un tableau à deux colonnes : en première colonne, la valeur de Y (croissante au sens large) ; en deuxième colonne l'indication "rouge" ou "bleu".

On ajoute trois colonnes :

Colonne 3 : on parcourt le tableau en commençant par la première ligne. A chaque fois que l'on rencontre l'indication "rouge" sur cette ligne, on incrémente de $\frac{1}{N_1}$ (on met $1/N_1$ pour la première fois que l'on rencontre "rouge", $2/N_1$ la seconde, etc.).

Colonne 4 : la même chose, pour l'indication "bleu", et en mettant $1/N_2$ au lieu de $1/N_1$

Colonne 5 : la différence entre colonne 3 et colonne 4. Cette différence doit garder un signe constant pour que la méthode de hiérarchisation puisse être appliquée.