Université de Picardie Jules Verne **2010-2011**
UFR des Sciences - LAMFA CNRS UMR 6140

PAF Amiens - Formation Enseignement des Mathématiques - 28 janvier 2011

Mathématiques : statistiques et simulation

A propos de l'inégalité de Bienaymé-Tchebichev

Lors de la formation, un collègue a indiqué qu'il devait exister un résultat assurant que pour une série d'observations $x_1, \ldots, x_n$, on pouvait assurer qu'un pourcentage donné des observations se situait dans un intervalle centré sur la moyenne. Effectivement, le résultat existe bien : c'est l'inégalité de Bienaymé-Tchebychev "version statistique descriptive". Dans beaucoup d'ouvrages de Statistique et Probabilités, on rencontre plus souvent la version "variable aléatoire". Ce document présente les deux versions du résultat.

Cas d'une série d'observations.

Considérons une série d'observations numériques $x_1, \ldots, x_n$ non toutes égales. Certaines valeurs pouvant cependant être égales, on désigne par :

- $x_1, \ldots, x_k$ les valeurs différentes observées ;
- n_i l'effectif de chaque valeur x_i, c'est-à-dire le nombre de fois où l'on a observé la valeur (on a $\sum_{i=1}^{k} n_i = n$) ;
- $f_i = \frac{n_i}{n}$ la fréquence de chaque valeur x_i, c'est-à-dire la proportion de fois où l'on a observé la valeur (on a $\sum_{i=1}^{k} f_i$) ; les couples (x_i, f_i) constituent la distribution statistique étudiée ;
- $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$ la moyenne arithmétique des valeurs observées : on a $\overline{x} = \sum_{i=1}^{k} f_i x_i$;
- $var(x) = s_x^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \overline{x})^2 = \sum_{i=1}^{k} f_i (x_i - \overline{x})^2$ la variance des valeurs observées ;
- $s_x = \sqrt{s_x^2} = \sqrt{var(x)}$ l'écart-type des valeurs observées ;
- $\left[\, \overline{x} - s_x \,;\, \overline{x} + s_x \,\right]$ l'intervalle moyen ; on dit qu'en moyenne, les valeurs observées se trouvent dans l'intervalle moyen.

A noter que $s_x^2 \neq 0$; car sinon, toutes les observations x_i seraient égales (à la moyenne $\overline{x}$) , ce qui est contraire à l'hypothèse de départ.

Proposition (inégalité de Bienaymé-Tchebichev version statistique descriptive)

Pour tout réel $t > 0$, la fréquence des observations se trouvant dans l'intervalle $\left[\, \overline{x} - t \times s_x \,;\, \overline{x} + t \times s_x \,\right]$ est au moins égale à $1 - \frac{1}{t^2}$.

Autrement dit, au moins $\left(1 - \frac{1}{t^2}\right) \times 100$ % des observations se trouvent dans l'intervalle $\left[\, \overline{x} - t \times s_x \,;\, \overline{x} + t \times s_x \,\right]$. En pratique :

- soit on se donne t, et alors on obtient le pourcentage d'observations dans l'intervalle correspondant à t : par exemple, pour $t = 2$, on a $1 - \frac{1}{t^2} = \frac{3}{4} = 0.75$, et donc au moins 75 % des observations dans l'intervalle $\left[\, \overline{x} - 2s_x \,;\, \overline{x} + 2s_x \,\right]$;
- soit on se donne un pourcentage souhaité, et alors on obtient l'intervalle recherché : par exemple, pour avoir au moins 95 % des observations dans l'intervalle, on cherche t tel que $1 - \frac{1}{t^2} = 0.95$, ce qui donne $t = \sqrt{\frac{1}{0.05}} = \sqrt{20}$, et donc l'intervalle $\left[\, \overline{x} - \sqrt{20}\, s_x \,;\, \overline{x} + \sqrt{20}\, s_x \,\right]$.

Remarquons aussi que seul le cas $t > 1$ est utile ; en effet, pour $0 < t \leq 1$, on a $1 - \frac{1}{t^2} \leq 0$. En particulier, on n'obtient pas de résultat intéressant pour $t = 1$, c'est-à-dire pour l'intervalle moyen $\left[\, \overline{x} - s_x \,;\, \overline{x} + s_x \,\right]$.

Preuve.

Soit t un réel tel que $t > 1$. Considérons la partition de $I = \{1, \ldots, k\}$ en les deux ensembles suivants :

- $A_t = \left\{ i \in I \,/\, x_i \in \left[\overline{x} - t \times s_x \,;\, \overline{x} + t \times s_x \right] \right\} = \left\{ i \in I \,/\, |x_i - \overline{x}| \leq t \times s_x \right\}$;
- $B_t = \left\{ i \in I \,/\, x_i \notin \left[\overline{x} - t \times s_x \,;\, \overline{x} + t \times s_x \right] \right\} = \left\{ i \in I \,/\, |x_i - \overline{x}| > t \times s_x \right\}$.

Insistons sur le fait que l'ensemble A_t est l'ensemble des indices de toutes les valeurs x_i appartenant à l'intervalle $\left[\overline{x} - t \times s_x \,;\, \overline{x} + t \times s_x \right]$.

On a alors $s_x^2 = \sum_{i \in I} f_i (x_i - \overline{x})^2 = \sum_{i \in A_t} f_i (x_i - \overline{x})^2 + \sum_{i \in B_t} f_i (x_i - \overline{x})^2 \geq \sum_{i \in B_t} f_i (x_i - \overline{x})^2$, puisque $\sum_{i \in A_t} f_i (x_i - \overline{x})^2$ ne contient que des termes positifs. De plus, pour tout $i \in B_t$, on a $(x_i - \overline{x})^2 > (t \times s_x)^2$, d'où $\sum_{i \in B_t} f_i (x_i - \overline{x})^2 > \sum_{i \in B_t} f_i (t \times s_x)^2 = t^2 s_x^2 \sum_{i \in B_t} f_i$.

On a ainsi $s_x^2 > t^2 s_x^2 \sum_{i \in B_t} f_i$. Comme $s_x \neq 0$, on obtient $1 > t^2 \sum_{i \in B_t} f_i$ et donc $\sum_{i \in B_t} f_i < \frac{1}{t^2}$.

Remarquant que $1 = \sum_{i \in I} f_i = \sum_{i \in A_t} f_i + \sum_{i \in B_t} f_i$, on obtient $\sum_{i \in A_t} f_i = 1 - \sum_{i \in B_t} f_i > 1 - \frac{1}{t^2}$: la fréquence (cumulée) des valeurs x_i appartenant à l'intervalle $\left[\overline{x} - t \times s_x \,;\, \overline{x} + t \times s_x \right]$ est donc au moins égale à $1 - \frac{1}{t^2}$.

Cas d'une variable aléatoire discrète finie.

Soit X une variable aléatoire réelle discrète finie non constante définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$. Désignons par $x_1, \ldots, x_k$ les valeurs possibles de X et par p_i la probabilité $P(X = x_i)$, pour tout $i = 1, \ldots, k$. Les couples (x_i, p_i) définissent la loi de probabilité de X ; on a en particulier $\sum_{i=1}^{k} p_i = 1$.

Considérons l'espérance mathématique et la variance de X, à savoir $E(X) = \sum_{i=1}^{k} p_i x_i$ et $var(X) = \sum_{i=1}^{k} p_i (x_i - E(X))^2$; remarquer l'analogie avec les formules de statistique descriptive (vues plus haut) dans lesquelles p_i est remplacée par f_i. Désignons par $\sigma(X) = \sqrt{var(X)}$ l'écart-type de X.

A noter que $var(X) \neq 0$; car sinon, X serait une variable aléatoire constante vérifiant alors $P(X = E(X)) = 1$, ce qui est contraire à l'hypothèse de départ.

Proposition (inégalité de Bienaymé-Tchebichev version variable aléatoire)

Pour tout $t > 0$, la probabilité que X prenne une valeur dans l'intervalle $\left[E(X) - t \times \sigma(X) \,;\, E(X) + t \times \sigma(X) \right]$ est au moins égale à $1 - \frac{1}{t^2}$;

ce qui s'écrit $P(E(X) - t \times \sigma(X) \leq X \leq E(X) + t \times \sigma(X)) = P(|X - E(X)| < t\sigma(X)) > 1 - \frac{1}{t^2}$.

Preuve

Elle est analogue à la preuve précédente. On décompose la somme définissant la variance à l'aide des deux ensembles $A_t = \left\{ i \in I \,/\, |x_i - E(X)| \leq t \times \sigma(X) \right\}$ et $B_t = \left\{ i \in I \,/\, |x_i - E(X)| > t \times \sigma(X) \right\}$, ce qui donne $var(X) > t^2 var(X) \sum_{i \in B_t} p_i$, puis $\sum_{i \in B_t} p_i < \frac{1}{t^2}$. Remarquant que $1 = \sum_{i \in A_t} p_i + \sum_{i \in B_t} p_i$, on obtient $\sum_{i \in A_t} p_i > 1 - \frac{1}{t^2}$. Le résultat découle alors du fait que $P(|X - E(X)| < t\sigma(X)) = \sum_{i \in A_t} p_i$.

Remarques

1) On peut étendre cette proposition aux variables aléatoires discrètes infinies (démonstration analogue, les sommes devenant des séries) et aux variables aléatoires à densité, pourvu qu'elles admettent une variance : par exemple, variables aléatoires de loi de Poisson, de loi Normale, ...

2) A noter que dans la preuve, on minore grossièrement $\sum_{i \in A_t} p_i (x_i - E(X))^2$ par 0. Ainsi, si cette inégalité est utile d'un point de vue théorique (dans les problèmes de convergence de suites de variables aléatoires réelles par exemple), elle ne donne pas en pratique des approximations suffisantes, comme le montrent l'exemple suivant. D'après la proposition précédente, on a (pour $t = 2$ et $t = 3$) :

$$P(|X - E(X)| \leq 2\sigma(X)) \geq \frac{3}{4} = 0,75 \text{ et } P(|X - E(X)| \leq 3\sigma(X)) \geq \frac{8}{9} \simeq 0,8889.$$

Or si X suit une loi normale $\mathcal{N}(\mu; \sigma)$, on a $\mu = E(X)$ et $\sigma = \sigma(X)$ et on montre que

$$P(|X - E(X)| \leq 2\sigma(X)) = 0,9544 \text{ et } P(|X - E(X)| \leq 3\sigma(X)) = 0,9973.$$

3) Toujours pour X de loi normale $\mathcal{N}(\mu;\sigma)$, on a aussi $P(|X - E(X)| \leq \sigma(X)) = 0.6826$.

4) Dans la pratique, on voit parfois ces pourcentages 68,26 %, 95,44 % et 99,73 % (obtenus pour la loi Normale) utilisés en statistique descriptive : par exemple, on lit que 68 % des observations se trouvent dans l'intervalle $\left[\ \overline{x} - s_x\ ;\ \overline{x} + s_x\ \right]$, 95 % des observations se trouvent dans l'intervalle $\left[\ \overline{x} - 2s_x\ ;\ \overline{x} + 2s_x\ \right]$, ... Ce n'est bien sur qu'un pourcentage approximatif, à condition que les observations puissent être considérées comme provenant d'une variable X de loi Normale (le fait que le diagramme en bâton des fréquences ait une allure en cloche peut rassurer). Mais alors, les observations $x_1, \ldots, x_n$ ne sont que des observations d'un échantillon de taille n de X, $\overline{x}$ et s_x n'étant que des estimations de $E(X)$ et $\sigma(X)$: les pourcentages obtenus avec la loi Normale ne peuvent donc pas être utilisés automatiquement. Prudence ...