CEA 2017 Première année
J. Malet

CEA 2017

Première année
J. Malet

Quantiles et simulation

1. Introduction

Les **quantiles** ont été introduits en statistique descriptive pour désigner des valeurs d'un caractère quantitatif repartissant la population étudiée en classes d'effectifs égaux. Les quartiles, premier quartile, médiane et troisième quartile, séparent la population en quatre classes d'effectifs égaux, les neuf déciles en dix, les quatre-vingt-dix-neuf centiles en cent.

Par exemple, le salaire médian sépare une population de salariés en deux classes d'effectifs égaux, ceux qui gagnent plus et ceux qui gagnent moins. Contrairement au salaire moyen, il n'est pas modifié par une augmentation des plus hauts salaires, dès lors que celle-ci concerne moins de la moitié de la population.

En probabilité, la notion de quantile s'applique de manière tout-à-fait générale aux variables aléatoires à valeurs réelles, et elle permet de définir une **fonction de répartition inverse** (ou fonction quantile). Comme on le verra plus loin, le procédé le plus courant de simulation d'une variable aléatoire de loi donnée s'appuie cette fonction.

2. Quantiles d'une variable aléatoire réelle

Pour une variable aléatoire X de fonction de répartition F_X, on appelle **quantile de niveau p** ou p-quantile de X (ou de sa loi de probabilité) une valeur x que la variable X a une probabilité p de ne pas dépasser :

$$\mathbf{x \text{ est un } p-\text{quantile de } X} \text{ si } \mathbf{P}[\mathbf{X} \leq \mathbf{x}] = \mathbf{p} \quad .$$

La terminologie de la statistique descriptive s'applique à ce contexte probabiliste, certains quantiles portant des noms spécifiques : on dit **médiane** plutôt que quantile de niveau $0,5$, **premier quartile** plutôt que quantile de niveau $0,25$ et **troisième quartile** plutôt que quantile de niveau $0,75$. Les quantiles dont les niveaux sont des multiples de un dixième (resp. un centième) sont appelés **déciles** (resp. **centiles**).
Par exemple un quantile de niveau $0,8$ est appelé huitième décile et un quantile de niveau $0,99$ est appelé quatre-vingt-dix-neuvième (et dernier) centile.

La notion actuarielle de **VaR**, utilisée pour mesurer le risque de marché d'un portefeuille d'instruments financiers est étroitement liée à celle de quantile. En effet, la VaR (de l'anglais Value at Risk) correspond au montant de pertes qui ne devrait être dépassé qu'avec une probabilité donnée, α, sur un horizon temporel donné ; il s'agit donc du quantile de niveau $1-\alpha$ de la variable aléatoire X représentant le montant de pertes à l'horizon donné.

2.a) Cas des lois continues usuelles

Dans le cas des lois usuelles, qui admettent une densité strictement positive sur un intervalle et nulle ailleurs (lois uniformes, lois exponentielles, lois normales, etc.), il existe un intervalle ouvert I tel que la fonction de répartition de la loi considérée réalise une bijection de I sur l'intervalle ouvert $]0,1[$. Autrement dit, pour tout p strictement compris entre 0 et 1, il existe un unique élément x de I pour lequel $F_X(x) = p$ (c'est-à-dire $P[X \leq x] = p$).

Ce nombre x est alors appelé **le** p-quantile de X ou de la loi de X. On peut le calculer par inversion de la fonction de répartition, mais les logiciels courants rendent ce calcul inutile, car la fonction réciproque de F_X y est fréquemment directement intégrée.

Exemple 1

Si X suit la loi uniforme $\mathcal{U}[a,b]$, F_X réalise une bijection de $I =]a,b[$ sur $]0,1[$ et, pour tout $p \in]0,1[$, le p-quantile de X est :

$$(F_X)^{-1}(p) = a + (b-a)\,p \quad .$$

Exemple 2

Si X suit la loi exponentielle $\mathcal{E}(\lambda)$ $(\lambda > 0)$, $F_X : x \longmapsto 1 - \exp(\lambda x)$ réalise une bijection de $I =]0,+\infty[$ sur $]0,1[$ et, pour tout $p \in]0,1[$, le p-quantile de X est :

$$(F_X)^{-1}(p) = \frac{-\ln(1-p)}{\lambda} \quad .$$

Exemple 3

Si X suit la loi normale $\mathcal{N}(\mu, \sigma^2)$ $(\sigma > 0)$, F_X réalise une bijection de $I =]-\infty,+\infty[$ sur $]0,1[$, et, pour tout $p \in]0,1[$, le p-quantile de X est :

$$(F_X)^{-1}(p) = \mu + \sigma\,\Phi^{-1}(p)$$

où Φ^{-1} est la bijection réciproque de la fonction de répartition Φ de la loi normale centrée réduite, définie par :

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\ \exp(-\frac{u^2}{2})\,\mathrm{d}u \quad .$$

Les quantiles des lois normales sont fournis directement par la fonction Excel "LOI.NORMALE.INVERSE" (voir coours sur les lois continues usuelles).

Exemple : si on cherche le segment $[a,b]$ le plus court contenant une variable aléatoire normale X d'espérance 1 et de variance 4 avec une probabilité égale à $0,95$, on cherche l'intervalle dont les bornes a et b sont les quantiles de niveaux $0,025$ et $0,975$.

En langage Excel :

"=LOI.NORMALE.INVERSE(0,025 ;1 ;2)" fournit $a \simeq -2,919922$
"=LOI.NORMALE.INVERSE(0,975 ;1 ;2)" fournit $b \simeq 4,919922$.

Sans surprise, l'intervalle est centré sur la valeur 1 :

$$\frac{a+b}{2} = 1 \quad .$$

2.b) Cas général

Dans le cas d'une variable aléatoire X à densité, la continuité de la fonction de répartition F_X assure, pour tout élément p de l'intervalle $]0,1[$, l'existence d'une valeur x pour laquelle $P[X \leq x] = p$, mais ce "quantile" n'est pas nécessairemment unique, lorsque la loi de X possède des zones de probabilité nulle.
Par exemple, si la loi de X est le mélange de deux lois uniformes continues sur des segments disjoints $\mathcal{U}[a,b]$ et $\mathcal{U}[c,d]$) (où $a < b < c < d$), pondérées par p et $1-p$, tous les éléments x du segment $[b,c]$ vérifient $P[X \leq x] = p$.

Dans le cas d'une variable aléatoire X discrète ou mixte, la fonction de répartition F_X n'est pas continue et l'existence d'une valeur x pour laquelle $P[X \leq x] = p$ n'est pas assurée pour toutes les valeurs de p.
Par exemple, si X est une variable aléatoire discrète dont les valeurs possibles sont $x_1, x_2, \ldots, x_n$ (classées par ordre croissant) et les probabilités ponctuelles correspondantes $p_1, p_2, \ldots, p_n$ (strictement positives et de somme égale à 1), les seules valeurs prises par la fonction de répartition F_X sont les $n+1$ probabilités cumulées $0, p_1, p_1+p_2, p_1+p_2+p_3, \ldots, 1$.

Ces deux difficultés amènent à assouplir la définition souhaitée des quantiles, en assurant leur existence, sans pour autant garantir leur unicité.

2.c) Définition générale des quantiles

Soit X une variable aléatoire réelle.
Pour tout nombre p strictement compris entre 0 et 1, on appelle **p-quantile de** X, ou de la loi de X, tout nombre x vérifiant les deux inégalités :

$$\boxed{\begin{cases} P[X < x\,] \leq p \\ P[X \leq x\,] \geq p \end{cases}} \quad .$$

L'ensemble des valeurs de x pour lesquelles $P[X < x\,] \leq p$ est un intervalle de la forme $]-\infty, M]$, contenant sa borne supérieure M. Ce nombre M est un quantile de niveau p de X parce que, pour tout $x > M$, on a $P[X \leq x\,] \geq p$, ce qui entraîne $P[X \leq M\,] \geq p$, par continuité à droite de la fonction de répartition F_X.

L'ensemble des valeurs de x pour lesquelles $P[X \leq x\,] \geq p$ est un intervalle de la forme $[m, +\infty[$, contenant sa borne inférieure m. Ce nombre est un quantile de niveau p de X parce que, pour tout $x < m$, on a $P[X < x\,] \geq p$, ce qui entraîne $P[X < m\,] \leq p$, par continuité à gauche de la fonction $x \longmapsto P[X < x\,]$.

Par conséquent, l'ensemble des quantiles de niveau p de la variable X est le segment $[m, M]$, qu'on notera désormais $[Q_X^-(p), Q_X^+(p)]$, où $Q_X^-(p)$ désigne donc le plus petit quantile de niveau p de X et $Q_X^+(p)$ le plus grand :

$$Q_X^-(p) = \mathrm{Inf}\{x \in \mathbb{R}; P[X \leq x\,] \geq p\}$$
$$Q_X^+(p) = \mathrm{Sup}\{x \in \mathbb{R}; P[X < x\,] \leq p\} \quad .$$

2.d) Exemples

Dans les deux exemples qui ont motivé plus haut l'assouplissement de la définition des quantiles, on peut maintenant indiquer ce que sont les quantiles, à n'importe quel niveau.

Exemple 4

Si la loi de X est le mélange de deux lois uniformes $\mathcal{U}[a,b]$ et $\mathcal{U}[c,d]$) (où $a < b < c < d$), pondérées par p et $1-p$, **les** médianes de X sont les éléments du segment $[b,c]$, car $Q_X^-(0,5) = b$ et $Q_X^+(0,5) = c$. Les autres quantiles sont uniques :

$$Q_X^-(p) = Q_X^+(p) = \begin{cases} a + 2p\,(b-a) & \text{si } 0 < p < \frac{1}{2} \\ c + (2p-1)(d-c) & \text{si } \frac{1}{2} < p < 1 \end{cases} \quad .$$

Exemple 5

On suppose ici que la variable aléatoire X est discrète et ne prend que trois valeurs 0, 1 et 2, avec des probabilités respectives $\frac{1}{4}$, $\frac{1}{4}$ et $\frac{1}{2}$. Les quantiles d'ordre $\frac{1}{4}$ de X sont les éléments du segment $[0,1]$ et les médianes de X les éléments du segment $[1,2]$.

Les autres quantiles sont uniques :

$$Q_X^-(p) = Q_X^+(p) = \begin{cases} 0 & \text{si } 0 < p < \frac{1}{4} \\ 1 & \text{si } \frac{1}{4} < p < \frac{1}{2} \\ 2 & \text{si } \frac{1}{2} < p < 1 \end{cases} \quad .$$

On observera pour l'anecdote que le troisième quartile est unique alors que le premier ne l'est pas. Ces choses-là arrivent.

3. Fonction de répartition inverse ("quantile function")

3.a) Quel choix ?

Pour toute variable aléatoire X on peut définir sur l'intervalle ouvert $]0,1[$ une "fonction quantile" qui associe à tout nombre p un quantile de niveau p de X. Il suffit pour cela de choisir pour chaque p un élément $Q_X(p)$ du segment $[Q_X^-(p), Q_X^+(p)]$ défini précédemment. N'importe quel choix systématique serait envisageable, $Q_X(p) = Q_X^-(p)$, $Q_X(p) = Q_X^+(p)]$ ou $Q_X(p) = \frac{1}{2}\big(Q_X^-(p) + Q_X^+(p)]\big)$ par exemple, et nous verrons que, dans l'objectif de la simulation, ce choix n'importe guère. Cependant, il semble que les statisticiens aient manifesté une certaine prédilection pour $Q_X^-(p)$ et nous allons nous conformer à ce choix pour définir **la** "fonction de répartition inverse".

3.b) **Définition** de Q_X

Pour toute variable aléatoire X, on appelle **fonction de répartition inverse** de X la fonction Q_X définie sur $]0,1[$ par :

$$\boxed{Q_X(p) = \mathrm{Inf}\{x \in \mathbb{R}; P[X \leq x)] \geq p\}} \quad .$$

3.c) Propriétés de Q_X

La fonction Q_X est croissante et continue à gauche sur $]0,1[$. Sa limite à droite en 0 est la borne inférieure du support de la loi de X et sa limite à gauche en 1 la borne supérieure de ce support.

La fonction Q_X est la réciproque de la fonction de répartition F_X lorsque celle-ci réalise une bijection continue d'un intervalle ouvert I sur $]0,1[$, ce qui est le cas pour les lois usuelles admettant une densité (voir plus haut). D'une manière générale, elle a des propriétés analogues à celles d'une application réciproque de F_X :

• Pour tout x, $Q_X\big(F_X(x)\big) \leq x$, avec égalité lorsqu'il n'existe aucun y strictement inférieur à x tel que $F_X(y) = F_X(x)$.

• Pour tout p strictement compris entre 0 et 1, $F_X\big(Q_X(p)\big) \geq p$, avec égalité lorsque p est une valeur prise par F_X.

• $Q_X \circ F_X \circ Q_X = Q_X$

• $F_X \circ Q_X \circ F_X = F_X$.

Pour ces raisons, la fonction de répartition inverse Q_X, "pseudo-inverse" de F_X, est souvent notée, abusivement, $(F_X)^{-1}$.

3.d) Exemples

Exemple 4 (suite)

Si la loi de X est le mélange de deux lois uniformes $\mathcal{U}[a,b]$ et $\mathcal{U}[c,d]$) (où $a < b < c < d$), pondérées par p et $1-p$:

$$Q_X(p) = \begin{cases} a + 2p\,(b-a) & \text{si } 0 < p \leq \frac{1}{2} \\ c + (2p-1)(d-c) & \text{si } \frac{1}{2} < p < 1 \end{cases} .$$

Exemple 5 (suite)

Si la variable aléatoire X est discrète et ne prend que trois valeurs 0, 1 et 2, avec des probabilités respectives $\frac{1}{4}$, $\frac{1}{4}$ et $\frac{1}{2}$:

$$Q_X(p) = \begin{cases} 0 & \text{si } 0 < p \leq \frac{1}{4} \\ 1 & \text{si } \frac{1}{4} < p \leq \frac{1}{2} \\ 2 & \text{si } \frac{1}{2} < p < 1 \end{cases} .$$

4. Simulation

4.a) Théorème fondamental de la simulation

Soit X une variable aléatoire, de fonction de répartition inverse Q_X. Si U suit la loi uniforme $\mathcal{U}\,[0,1]$, la variable aléatoire $Q_X \circ U$ suit la même loi que X.

Démonstration et remarques

Soit $x \in \mathbb{R}$.
D'après la définition de Q_X, les deux propriétés $Q_X(u) \leq x$ et $u \leq P[X \leq x]$ d'un élément u de l'intervalle $]0,1[$ sont équivalentes.
Il en résulte que les deux événements aléatoires $\{Q_X \circ U \leq x\}$ et $\{U \leq P[X \leq x]\}$ ont la même probabilité, donc que :

$$P[Q_X \circ U \leq x] = P\big[U \leq P[X \leq x]\big] = F_U\big(P[X \leq x]\big) = P[X \leq x] \quad .$$

Les deux variables aléatoires $Q_X \circ U$ et X ayant donc la même fonction de répartition, elles suivent la même loi, ce qui achève la démonstration.

On démontre de manière similaire que la variable aléatoire $Q_X^+ \circ U$ suit aussi la même loi que X, puisque, pour tout x :

$$P[Q_X^+ \circ U \geq x] = P\big[U \geq P[X < x]\big] = 1 - P[X < x] = P[X \geq x] \quad .$$

Il en résulte, comme il a été annoncé plus haut, que le choix adopté pour la définition de la "fonction quantile" Q_X est sans influence sur la loi de la variable aléatoire $Q_X \circ U$, qui coïncide toujours avec celle de X. Cela n'est pas étonnant, puisque les valeurs de p pour lesquelles un choix est effectivement possible pour le quantile correspondent aux plages constantes de la fonction de répartition. Ces valeurs constituent un ensemble vide, fini ou dénombrable, donc négligeable (référence) et par conséquent de probabilité nulle pour la variable à densité U. Changer la "fonction quantile" Q_X en ces points ne peut donc pas modifier la loi de $Q_X \circ U$.

4.b) Exemples : principe de simulation de lois usuelles

Exemple 1 (suite)

Si U suit la loi uniforme $\mathcal{U}\,[0,1]$, la variable aléatoire $a + (b-a)\,U$ suit la loi uniforme $\mathcal{U}\,[a,b]$.

Exemple 2 (suite)

Si U suit la loi uniforme $\mathcal{U}\,[0,1]$, la variable aléatoire $\dfrac{-\ln(1-U)}{\lambda}$ suit la loi exponentielle $\mathcal{E}(\lambda)$ (la variable aléatoire $\dfrac{-\ln(U)}{\lambda}$ suit aussi la loi $\mathcal{E}(\lambda)$ puisque U et $1-U$ suivent la même loi).

Exemple 3 (suite)

Si U suit la loi uniforme $\mathcal{U}\,[0,1]$, la variable aléatoire $\mu + \sigma\,\Phi^{-1}(U)$ suit la loi normale $\mathcal{N}(\mu, \sigma^2)$.