

Simuler pour approcher la réalité

quelques exercices

Dans tous les secteurs d'applications des statistiques on utilise beaucoup les techniques de simulation. Nous en présentons ici des éléments qui pourront être utiles dans la pratique. Les données utilisées ici sont réelles même si elles ont parfois été simplifiées aux fins de votre apprentissage. Certains éléments du logiciel Statistica sont présentés dans la suite, mais tous les logiciels offrent les mêmes possibilités.

Toutes vos réponses doivent être expliquées et justifiées.

1 La technique (10 points)

On simule presque toute loi à partir de simulations de la loi uniforme $U \sim \mathcal{U}[0; 1]$. Nous n'expliquerons pas les techniques *non aléatoires* qui permettent d'obtenir des suites de nombres situés dans l'intervalle $[0; 1]$ qui ont toutes les apparences de réalisations d'une telle variable aléatoire. Sachons cependant que, quelle que soit la façon d'obtenir ces suites de nombres, il convient le plus possible de vérifier par observation et certains calculs simples (moyennes, *etc.*) que les suites qu'on utilise ont de bonnes propriétés [2, 3].

1. **(1 point)** Supposons qu'on sache obtenir des nombres au hasard dans l'intervalle $[0; 1]$ (c'est la fonction « `rnd(1)` » de Statistica). Comment obtenir des simulations d'une uniforme sur l'intervalle $[a; b]$?
Indice : Trouver la fonction $U \sim \mathcal{U}[0; 1] \longrightarrow U_1 \sim \mathcal{U}[a; b]$.
2. **(1 point)** Soit $X \sim \mathcal{U}[0; 1]$. Expliquer comment déterminer β_0 et β_1 de sorte que sa transformée $Y = \beta_0 + \beta_1 X$ ait une moyenne et une variance prescrites : $E(Y) = \mu, V(Y) = \sigma^2$.
3. **(2 points)** Soit X une variable continue de cumulative $F(x)$ strictement croissante.
 - Une telle fonction cumulative de probabilité possède-t-elle toujours une fonction inverse ?
 - La fonction inverse, notée F^{-1} , d'une fonction continue strictement croissante est-elle aussi une fonction croissante ?
4. **(1 point)** Plaçons-nous dans les hypothèses du numéro précédent. Si on se donne une suite de nombres u_i qui suivent la loi $U \sim \mathcal{U}[0; 1]$,

montrer alors que les $x_i = F^{-1}(u_i)$ suivent la loi de X , *i.e.* que $X = F^{-1}(U)$ est bien vérifiée.

Indice : On doit évidemment montrer $P(F^{-1}(U) \leq x) = P(X \leq x)$.

5. **(1 point)** Quel est l'avantage de simuler une loi X donnée en passant par des uniformes et la cumulative inverse de la loi de X tel que vu au numéro précédent, sur les simulations déjà programmées dans les logiciels ?

Indice : Penser à des lois de X ayant pour densités des formes malaisément interprétables.

6. **(2 points)** Dans le cas où la loi de X n'est pas continue, on peut remplacer F^{-1} dans les relations précédentes par

$$F^{-}(u) = \min\{x \mid F(x) \geq u\}.$$

Montrer comment utiliser ce fait pour simuler une loi discrète à 3 valeurs, de probabilité p_i de votre choix. (La généralisation à une loi discrète ayant n valeurs est immédiate.)

7. **(2 points)** Expliquer comment déterminer à l'aide d'un générateur de la loi $\mathcal{U}[0; 1]$ un sous-échantillon de proportion approximativement $p \in (0; 1)$ d'un échantillon donné.

Statistica. On comprend le rôle que jouent les cumulatives inverses dans les simulation de VA. Pour les lois les plus usuelles, **Statistica** préfixe leurs cumulatives par « V ». Ainsi on simulera une VA exponentielle (autant de lignes que dans le fichier de données) par la définition suivante : « = Vexpon(rnd(1); λ) ». Attention également, la loi gaussienne dans **Statistica** est paramétrée par la moyenne et l'écart *type* et non la variance comme la convention l'exige souvent.

2 Une mise en marché (20 points)

Une grande compagnie de distribution a décidé d'offrir une gamme de petits outils électriques pour bricoleurs. Vous êtes un des membres de l'équipe qui travaille sur le projet de mise en marché de ces produits. On trouvera dans le fichier *ventes.xls* (cliquer l'icone à droite) une partie des données pertinentes du premier six mois des opérations.

La compagnie possède des centres de distribution (CD) dans tous les états américains. Les États-Unis ont été répartis en 4 régions (1 à 4). Chacune d'entre elles comporte un certain nombre de CD qui sont répartis en 2 catégories de villes (11 et 12 pour la région 1, 21 et 22 pour la région 2, *etc.*), selon un classement socio-économique : plutôt industriel et primaire



Data

— avec 1 pour deuxième chiffre du code, ou plutôt tertiaire — avec 2 comme deuxième chiffre du code.

Une publicité massive a été utilisée dans chaque ville pour lancer le produit. Celle-ci a varié selon les villes et les désirs des gérants locaux qui ont décidé de répartir leur budget de publicité, de façon différente selon leur expérience, dans les médias (journaux et circulaires, télévision et radio, internet), mais aussi selon une stratégie de la compagnie. La variable *public* donne la somme totale consacrée à la publicité dans chacun des CD : c'est la somme des trois variables *Jour* pour la publicité sur papier, *Élec* pour celle dans les médias électroniques, enfin *Inter* pour la publicité sur Internet. Les valeurs sont en milliers de dollars.

Enfin, dernière variable du fichier : les ventes, *Ventes*, en milliers de dollars pour les premiers six mois après la mise en marché.

Après 6 mois de l'introduction des produits, la compagnie aimerait avoir un état de la situation dans les diverses régions et villes. L'ingénieur que vous êtes est en charge de faire (faire...) l'analyse de ces résultats semestriels, avec objectif d'optimiser les ventes. Vous êtes depuis peu au service de la compagnie, et vous désirez faire vos preuves...

1. **(10 points)** Faites une analyse complète (enfin, la plus complète possible dans la mesure de vos connaissances actuelles...) de ces données. Utilisez des moyens simples de descriptions statistiques pour « montrer » les influences s'il y en a de la ville, de la région, de la structure des budgets publicité, *etc.*, sur les ventes. Vous vous intéressez surtout aux différences des ventes par régions et par classes de villes, de même qu'à l'influence de la publicité et des types de publicité sur les ventes.
2. **(10 points)** Vous arrivez dans votre analyse à montrer que les structures de la publicité diffèrent selon le type de ville : c'est une stratégie de vente que la compagnie veut tester. Dans la première question, vous aurez étayé ces constatations à l'aide d'analyses de régression essentiellement pour valider des corrélations. Maintenant, vous aimeriez savoir si ces influences sont réelles et non des artefacts de l'échantillon.

Vous choisissez une des influences marquantes d'un type de publicité sur les ventes. Donc un modèle linéaire bien validé (du moins autant que vos connaissances le permettent). Une validation supplémentaire consiste à examiner l'effet de l'échantillonnage de la façon suivante. Vous considérez un grand nombre (on se contentera ici d'une vingtaine) de sous-échantillons de, disons, 50% de votre échantillon principal. Si le modèle linéaire n'est pas un artefact, vous devriez trouver sensiblement les mêmes paramètres des régressions que pour l'échantillon complet. Pensez-vous que votre effet est bien validé ? Utilisez des considérations de *rééchantillonnage*.

Note : Vous utilisez évidemment les techniques explorées brièvement à la section précédente pour déterminer les sous-échantillons. Vous décrirez la procédure dans votre rapport. Un peu de pratique de **Statistica** permet de sous-échantillonner très rapidement. Les paramètres des régressions peuvent être ajoutés aux fichier de données **Statistica** en passant les commandes *ctrl-c* et *ctrl-v*.

3 Transformer des VA : tolérances (20 points)

On a souvent besoin de transformer des VA, et il n'est pas toujours évident de déterminer les densités des VA transformées, ainsi que, bien évidemment, leurs moyennes, variances, les couvertures ou raretés de certains intervalles. Les techniques de simulation permettent bien souvent de se tirer d'affaire, au moins approximativement, du moins si on connaît les lois des VA avant transformation. On peut également utiliser quelques techniques mathématiques simples pour en déterminer approximativement les premiers moments dont les moyennes et variances.

Pour fixer les idées pensons ici à la résistance résultante d'un montage en parallèle de plusieurs résistances. Ce sera l'exemple de ce projet. Mais ces techniques s'appliquent à des fonctions complexes de VA, *i.e.* en fonction de paramètres dont les valeurs sont incertaines, l'incertitude étant modélisée par des lois de probabilité. Elles permettent notamment de prendre des décisions en présence d'incertitude (voir N. D. Cox [1]).

La formule de Taylor est à la base de la théorie de l'approximation¹. Considérons le cas d'une fonction $Y = f(X)$, une fonction d'une seule VA. On peut écrire pour f l'approximation de Taylor à l'ordre 2 centrée au point μ :

$$Y = f(X) \doteq f(\mu) + f'(\mu)(X - \mu) + \frac{f''(\mu)}{2} (X - \mu)^2. \quad (1)$$

Ainsi donc² :

$$E(f(X)) \doteq f(\mu) + \left[\frac{f''(\mu)}{2} \right] E((X - \mu)^2).$$

Si on néglige le dernier terme (on n'a qu'à tronquer l'approximation (1) à l'ordre 1), on a l'approximation suivante

$$E(f(X)) = f(E(X)).$$

¹Brook Taylor [1685-1731] —époque d'Isaac Newton [1642-1727]— dont la célèbre formule provient de son œuvre *Methodus incrementorum directa et inversa* (1715). Il s'est intéressé aussi aux solutions singulières des équations différentielles, à l'optique, et autres problèmes de la physique.

²On utilise ici les formules habituelles pour les espérances et les variances des VA.

1. **(2 points)** Écrire l'approximation pour $E(f(X))$ lorsqu'on utilise l'approximation (1) à l'ordre 2.
2. **(3 points)** Effectuez le produit (1) par lui-même, que vous réduirez en négligeant les termes d'ordre supérieur à 2. *i.e.* en éliminant les termes en $(X - \mu)^k$ avec $k > 2$, pour approcher $E(f(X)^2)$. Enfin, déduisez une approximation de $V(f(X))$.

Indice : On a le développement suivant pour la variance de toute VA :
 $V(X) = E(X^2) - E(X)^2$.

Nous allons étudier les propriétés des montages de résistances en parallèle.

Les chariots-élévateurs sont grandement utilisés dans les entrepôts de grande surface. Ils permettent la manutention et l'entreposage en hauteur des palettes. Ils pèsent jusqu'à environ 4 tonnes et peuvent lever des charges de 3 tonnes à 10 mètres de hauteur ! Afin d'assurer un bon fonctionnement mécanique, des dizaines de cartes de contrôle sont installées sur les chariots. Ces cartes possèdent souvent des composants munis de résistances en parallèle. On peut ainsi contrôler la présence d'anomalies, la vitesse du chariot lorsque les fourches sont en élévation, l'embrayage des vitesses, *etc.* Pour les fins de l'exercice, on considérera ici le cas de 3 résistances en parallèle.

Un bon nombre de résistances de nominal 50Ω et de *tolérance naturelle* 5% du nominal utilisées dans des montages en parallèle ont été testées pour leurs caractéristiques distributionnelles. On a conclu que le nominal était respecté de même que la tolérance, mais la répartition de la valeur des résistances était plutôt uniforme.

3. **(2 points)** L'intervalle dit de *tolérance naturelle* est celui situé à moins de 3σ de la moyenne. Déterminez à partir des considérations ci-haut le σ de la loi uniforme qui modélise les valeurs aléatoires des résistances. Donner la formule de la VA qui permet de simuler les valeurs des résistances.

Pour une fonction de deux variables aléatoires $Z = f(X, Y)$, on peut utiliser le développement limité à l'ordre 1 suivant, centré sur les moyennes (μ_X, μ_Y) :

$$Z = f(X, Y) \doteq f(\mu_X, \mu_Y) + \frac{\partial f}{\partial X}(X - \mu_X) + \frac{\partial f}{\partial Y}(Y - \mu_Y), \quad (2)$$

où les dérivées sont calculées au centre du développement.

4. **(3 points)** Admettez que les VA X et Y sont indépendantes. Obtenez alors des approximations pour la moyenne $\mu_Z = E(Z)$ et la variance $\sigma_Z^2 = E(Z^2) - E(Z)^2$. Appliquez au voltage maximal $V_{\max} = RI_{\max}$ où la résistance possède les caractéristiques plus haut et le courant maximal suit une loi $\mathcal{N}(10; 1)$.
5. **(2 points)** Utilisez la règle de Tchébycheff pour déterminer la couverture (probabilité) de l'intervalle de tolérance naturelle approximatif de la valeur de $V_{\max} = f(R, I_{\max})$. Calculez aussi cet intervalle. Pourrait-on utiliser ici la couverture selon l'inégalité de Camp-Meidel ?

Dès que les transformations des VA dépendent de plus de 2 VA, ou que l'expression en est assez complexe, les méthodes tayloriennes deviennent vite laborieuses. On préfère alors souvent utiliser les techniques de simulation. Cependant, il convient de les utiliser avec circonspection, comme le font voir les exercices suivants.

6. **(2 points)** Simulez 100 échantillons de valeurs d'une résistance résultante pour un montage de 3 résistances en parallèle. Pouvez-vous croire que la valeur de la résistance résultante suit une loi gaussienne ? Si oui, quelle est la couverture de l'intervalle de tolérance naturelle, et calculez une approximation de cet intervalle ; sinon, utilisez les approximations de Tchébycheff ou de Camp-Meidel pour obtenir une approximation de l'intervalle de même couverture.
7. **(2 points)** On comprendra que l'approximation des quantiles pour p petit ou grand demande plus que 100 échantillons... Recommencez la démarche du numéro précédent pour 1000 et 5000 échantillons. Que constatez-vous ? Commentez brièvement sur les utilités respectives des histogrammes et des diagrammes quantiles-quantiles pour jauger la normalité de données.
8. **(2 points)** Supposons que l'ampérage maximal que peut drainer un tel composant est de 10 ampères, et que la puissance maximale tolérable dissipée est de 2000W. En moyenne environ combien de tels composants à 3 résistances en parallèle feront-ils défaut sur 1000 achetés ?

Indice. On utilisera la formule $P = VI = RI^2$ et on supposera que $I_{\max} = 10$ n'est pas aléatoire.

9. **(2 points)** Supposons maintenant que le courant maximal est lui aussi aléatoire $I_{\max} \sim \mathcal{N}(10; 1)$. Déterminez la couverture approximative de l'intervalle de tolérance naturelle ; un intervalle de couverture 99%.

- Supposons que les valeurs des résistances soient gaussiennes plutôt qu'uniformes, En moyenne, environ combien de composants feront-ils défaut sur 1000 achetés ?

Références

- [1] Neil C. Cox, 1986 : *How to perform statistical tolerance analysis*. Milwaukee, Wisc. : ASQ Press (ASQ : The American Society for Quality ; <http://www.asq.org>). 55p.
- [2] William H. Press, 1992 : *Numerical recipes in Fortran (in C)*. Cambridge : Cambridge University Press.
- [3] Brian Ripley, 1987 : *Stochastic simulation*. New York : John Wiley.