Remarques sur le calcul des estimateurs

Position du problème

- Lors d'une étude de simulation, ou de l'analyse statistique d'un jeu de données, un problème récurrent consiste à calculer des moyennes arithmétiques ainsi que des variances empiriques.
- Plus spécifiquement, à partir d'une suite $X_1, X_2, \ldots, X_n$ de n réalisations (on parle aussi d'observations) i.i.d. d'une même variable aléatoire X, on est régulièrement amené à estimer l'espérance $E(X)$ ainsi que la variance $\mathrm{Var}(X)$ de la variable X.
- L'estimateur classique $\hat{X}$ de l'espérance $E(X)$ est simplement la moyenne arithmétique $\bar{X}$ des réalisations :

$$\hat{X} = \bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

- Cet estimateur est consistant, il converge en probabilité vers l'espérance lorsque n tend vers l'infini, et sans biais, son espérance est égale à $E(X)$.

Position du problème (2)

- Pour obtenir un estimateur de la variance $\mathrm{Var}(X) = E\left(\left(X - E(X)\right)^2\right)$ on utilise soit

$$s^2 = \frac{1}{n-1}\left(\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2\right) = \frac{1}{n-1}\sum_{k=1}^{n} X_k^2 - \frac{n}{n-1}\bar{X}^2$$

soit

$$\sigma^2 = \frac{1}{n}\left(\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2\right) = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \bar{X}^2$$

- Les deux estimateurs sont consistants mais seul le premier est non biaisé. En effet

$$E(\sigma^2) = \frac{n}{n-1}\mathrm{Var}(X) \quad \text{alors que} \quad E(s^2) = E(\tfrac{n}{n-1}\sigma^2) = \mathrm{Var}(X).$$

Pour n très grand, l'importance de ce biais est cependant plus théorique que pratique.

Difficultés pratiques

Lors du calcul de ces estimateurs dans un programme, typiquement à l'aide de variables de type `double`, deux difficultés peuvent potentiellement apparaître.

1) Dépassement de capacité : Un tel phénomène est peu probable lors du calcul de la moyenne, où seule la somme des X_k doit être maintenue, à moins que les valeurs à manipuler soient particulièrement importantes et le nombre gigantesque.

   Si on estime la variance en calculant

$$\sigma^2 = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \bar{X}^2 = \frac{1}{n}\left(\sum_{k=1}^{n} X_k^2\right) - \left(\frac{1}{n}\sum_{k=1}^{n} X_k\right)^2$$

   il faut aussi stocker la somme des carrées des X_k et le risque de dépassement est à peine plus important.

   Il n'y a pas vraiment de solution à ce problème si ce n'est le garder à l'esprit est s'adapter si nécessaire. L'utilisation d'estimateurs itératifs permet cependant de l'éviter.

Difficultés pratiques (2)

2) Perte de précision : Les types de variables disponibles dans un langage informatique pour stocker des réels, typiquement le type `double`, n'ont pas une précision infinie.

   Leur manipulation, même au travers d'opérations arithmétiques aussi simple que des soustractions, peut entraîner une perte de précision, autrement dit une diminution du nombre de décimales correctes.

   En temps normal, cette diminution reste limitée et n'a rien de préoccupant. Les estimateurs que l'on calcule n'ont qu'une précision limitée et fournir des résultats avec 15 chiffres, soit disant, significatifs n'a guère de sens.

   Mais dans certains cas, certes un peu pathologique, cette perte de précision peut être très importante et peut même aboutir à des résultats aberrants, comme des variances empiriques négatives !

Estimation de l'espérance

- Pour obtenir la moyenne $\bar{X}$ des réalisations X_k, afin d'estimer l'espérance de la variable X, on calcule simplement

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

  en stockant et mettant à jour la somme des X_k. Cette manière de faire ne pose pas de problèmes en général.

- Il est également possible de calculer cette moyenne itérativement en maintenant à jour les moyennes $\bar{X}(k)$ des k premières valeurs à l'aide de la relation de récurrence

$$\bar{X}(k) = \frac{(k-1)\bar{X}(k-1) + X_k}{k} = \bar{X}(k-1) + \frac{X_k - \bar{X}(k-1)}{k}, \qquad k = 1, \ldots, n$$

  avec l'initialisation $\bar{X}_0 = 0$.

- On a alors, pour tout k, $\bar{X}(k) = \frac{1}{k}\sum_{i=1}^{k} X_i$ et en particulier $\bar{X} = \bar{X}(n)$.

Estimation de la variance

- Pour estimer la variance à l'aide de σ^2 (les formules s'adaptent facilement au calcul de s^2) on peut

  1) calculer la moyenne des carrés des écarts à la moyenne :

$$\sigma^2 = \frac{1}{n}\left(\sum_{k=1}^{n} \left(X_k - \bar{X}\right)^2\right)$$

  2) utiliser la formule équivalente (en théorie) :

$$\sigma^2 = \frac{1}{n}\left(\sum_{k=1}^{n} X_k^2\right) - \bar{X}^2 = \frac{1}{n}\left(\sum_{k=1}^{n} X_k^2\right) - \left(\frac{1}{n}\sum_{k=1}^{n} X_k\right)^2$$

- Les deux approches donnent théoriquement la même valeur mais lorsqu'on utilise des nombres à virgule flottante, possédant une précision finie, les valeurs obtenues peuvent être très différentes !

Estimation de la variance (2)

- Le problème principal lors de l'évaluation des expressions précédentes dans un programme est la perte de précision qui se produit lorsque l'espérance de X est très grande par rapport à sa variance.

  Dans une telle situation chaque soustraction s'accompagne d'une perte de précision importante qui peut aboutir à des résultats avec très peu de décimales correctes voire complètement faux.

- Avant d'illustrer le phénomène (et de tenter d'y remédier), notons une différence importante entre les deux approches :

  1) La première formule aboutit à un algorithme à deux phases : dans la première on calcule la moyenne $\bar{X}$ et dans la seconde on calcule la moyenne des carrés des écarts.

     Une telle approche nécessite donc de stocker toutes les réalisations.

  2) La deuxième formule aboutit à un algorithme à une phase : il suffit de stocker et mettre à jour la somme de X_k et celle de leurs carrés pour estimer la variance.

     Il n'est donc pas nécessaire de stocker toutes les réalisations avec cette approche.

Exemple numérique

- On considère l'expérience suivante : Générer un échantillon de $N = 10^6$ réalisations indépendantes d'une variable X uniforme dans l'intervalle $[M, M+1]$ puis estimer son espérance par la moyenne $\bar{X}$ de l'échantillon ainsi que sa variance à l'aide des deux estimateurs précédents, notés respectivement σ_1^2 et σ_2^2.
- La valeur théorique de l'espérance de X est $E(X) = M + \frac{1}{2}$ alors que la valeur théorique de sa variance est $\mathrm{Var}(X) = \frac{1}{12} = 0.08333\overline{3}$ (indépendante de M).

M	$\bar{X}$	σ_1^2	σ_2^2
0	0.500098053223	0.083340367131	0.083340367131
10^3	1000.50009805	0.083340367131	0.083340349258
10^4	10000.5000981	0.083340367131	0.083349078894
10^5	100000.500098	0.083340367131	0.082859039307
10^6	1000000.5001	0.083340367131	0.100219726562
10^7	10000000.5001	0.083340367131	7.953125
10^8	100000000.5	0.083340367132	-708.0
10^9	1000000000.5	0.083340371436	-188032.0
10^{10}	10000000000.5	0.085779328889	867352576.0

Analyse de l'exemple

- Pour essayer de comprendre ces résultats rappelons d'abord la règle suivante : Si deux nombres à virgule flottante coïncident sur leurs p premiers chiffres significatifs alors il est possible de perdre jusqu'à p chiffres significatifs lors du calcul de leur différence.
- Rappelons également qu'une variable de type `double` (norme IEEE 754) offre 15 chiffres décimaux de précision (voire 16 ou 17 dans certains cas).
- Maintenant, si $M = 10^6$, lors du calcul de σ_1^2, X_k et $\bar{X}$ vont typiquement coïncider sur leurs 9 à 10 premiers chiffres significatifs. On peut donc conserver environ 5 à 6 chiffres significatifs lors des soustractions $(X_k - \bar{X})$.

  Lors du calcul de σ_2^2 la situation se dégrade : les valeurs

  $$\frac{1}{n}\left(\sum_{k=1}^{n} X_k^2\right) \qquad \text{et} \qquad \bar{X}^2 = \left(\frac{1}{n}\sum_{k=1}^{n} X_k\right)^2$$

  sont de l'ordre de 10^{18} alors que leur différence est de l'ordre de 10^{-2} (si les valeurs étaient calculées avec une précision infinie). Les deux grandeurs coïncident donc sur bien plus que les 15 chiffres significatifs à disposition !

Estimateur itératif de la variance

- L'estimation de la variance à l'aide de la méthode à deux phases est donc nettement plus robuste que par la seconde approche.
- Pour s'affranchir de l'obligation de stocker tout l'échantillon de valeurs, on peut utiliser une version itérative pour le calcul de la variance empirique dont la stabilité numérique en semblable (et même souvent meilleure) que celle de l'approche à deux phases.

  Comme précédemment, on note $\bar{X}(k)$ la moyenne des k premières valeurs et $\sigma^2(k)$ l'estimation de la variance de ces k valeurs. La mise à jour de ces grandeurs passe par les relations de récurrence :

  $$\bar{X}(k) = \frac{(k-1)\bar{X}(k-1) + X_k}{k} = \bar{X}(k-1) + \frac{X_k - \bar{X}(k-1)}{k}, \qquad k \geq 1$$

  et

  $$\sigma^2(k) = \frac{(k-1)\sigma^2(k-1) + \left(X_k - \bar{X}(k)\right)\left(X_k - \bar{X}(k-1)\right)}{k}, \qquad k \geq 1$$

  avec les valeurs initiales $\bar{X}(0) = 0$ et $\sigma^2(0) = 0$.

Méthode de Welford (1962)

- Plutôt que de mettre à jour $\sigma^2(k)$ il est plus intéressant de mettre à jour $M_2(k) = k\sigma^2(k)$ à l'aide de la récurrence

  $$M_2(k) = M_2(k-1) + \left(X_k - \bar{X}(k)\right)\left(X_k - \bar{X}(k-1)\right), \qquad k \geq 1$$

  en partant de $M_2(0) = 0$. On a alors $\sigma^2 = \sigma^2(n) = \dfrac{M_2(n)}{n}$.
- Le pseudo-code suivant, proposé par Knuth sur la base des travaux de Welford, propose une mise en œuvre numériquement stable pour l'estimation de la variance :

  1. Poser $M_1 = 0$ et $M_2 = 0$ // Initialisation : $M_1 =$ moyenne, $M_2 = k\sigma^2$
  2. Pour k de 1 à n faire
  3. $\quad \delta = X_k - M_1$
  4. $\quad M_1 = M_1 + \delta/k$
  5. $\quad M_2 = M_2 + \delta \cdot (X_k - M_1)$
  6. Retourner $\bar{X} = M_1$ et $\sigma^2 = M_2/n$ // ou $s^2 = M_2/(n-1)$ si on préfère

Exemple numérique

- On reprend la même expérience qu'avant. Cette fois σ_3^2 dénote l'estimation itérative de la variance obtenue avec la méthode de Welford et M_1 l'estimateur itératif de l'espérance.

M	$\bar{X}$	σ_1^2	M_1	σ_2^3
0	0.500098053223	0.083340367131	0.500098053223	0.083340367131
10^5	100000.500098	0.083340367131	100000.500098	0.083340367133
10^6	1000000.5001	0.083340367131	1000000.5001	0.083340367119
10^7	10000000.5001	0.083340367131	10000000.5001	0.083340367233
10^8	100000000.5	0.083340367132	100000000.5	0.083340362806
10^9	1000000000.5	0.083340371436	1000000000.5	0.083340367545
10^{10}	10000000000.5	0.085779328888	10000000000.5	0.083341251478
10^{11}	10^{11}	0.290386834018	10^{11}	0.083340984422
10^{12}	10^{12}	0.328820324109	10^{12}	0.083340335201
10^{13}	10^{13}	0.331488690334	10^{13}	0.083344633190
10^{14}	$10^{14}4$	0.318096739502	10^{14}	0.083622356934
10^{15}	$9.99999999996 \cdot 10^{14}$	19056500.1443	10^{15}	0.085948375000
10^{16}	$9.99999999994 \cdot 10^{15}$	3802695556.0	10^{16}	0.0

Estimation de la covariance

- Pour estimer la covariance $\mathrm{Cov}(X, Y)$ de deux variables aléatoires X et Y à partir d'une suite $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$ de n couples de réalisations on dispose d'options similaires à celles disponibles pour estimer la variance.

1) Estimer la covariance avec une approche à deux phases :

$$\sigma_{XY}^2 = \frac{1}{n}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)\left(Y_k - \bar{Y}\right) \quad \text{où} \quad \bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \quad \text{et} \quad \bar{Y} = \frac{1}{n}\sum_{k=1}^{n} Y_k$$

L'estimateur est plutôt stable numériquement mais il est nécessaire de stocker toutes les réalisations.

2) Estimer la covariance avec une approche à une phase :

$$\sigma_{XY}^2 = \frac{1}{n}\sum_{k=1}^{n} X_k Y_k - \bar{X}\bar{Y} = \frac{1}{n}\sum_{k=1}^{n} X_k Y_k - \frac{1}{n^2}\left(\sum_{k=1}^{n} X_k\right)\left(\sum_{k=1}^{n} Y_k\right)$$

Ici il n'est pas nécessaire de stocker toutes les réalisations mais la stabilité peut parfois être mise à mal.

Estimation de la covariance (2)

3) Estimer la covariance en une seule phase mais en utilisant les récurrences :

$$\bar{X}(k) = \frac{(k-1)\bar{X}(k-1) + X_k}{k} = \bar{X}(k-1) + \frac{X_k - \bar{X}(k-1)}{k}, \qquad k \geq 1$$

$$\bar{Y}(k) = \frac{(k-1)\bar{y}(k-1) + Y_k}{k} = \bar{Y}(k-1) + \frac{Y_k - \bar{Y}(k-1)}{k}, \qquad k \geq 1$$

et

$$\begin{aligned} C(k) &= C(k-1) + \left(X_k - \bar{X}(k)\right)\left(Y_k - \bar{Y}(k-1)\right), \qquad k \geq 1 \\ &= C(k-1) + \left(X_k - \bar{X}(k-1)\right)\left(Y_k - \bar{Y}(k)\right) \end{aligned}$$

L'estimateur de la covariance est alors

$$\sigma_{XY}^2 = C/n.$$

Comme dans le cas de la variance, cette approche offre une bonne stabilité numérique tout en ne nécessitant pas le stockage des réalisations.