Big Data

TD 3 : Gestion des données et traitements statistiques

BUT 3

Guillaume Metzler et Antoine Rolland
Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France
guillaume.metzler@univ-lyon2.fr; antoine.rolland@univ-lyon2.fr

Les exercices de cette fiche sont des exercices présentées lors du cours introductif et permettent de revenir sur des notions de calculs sur plusieurs serveurs.

Exercice 1 : Petit échauffement

Dans cet exercice, on suppose que l'on dispose de N serveurs sur lesquels sont stockées des données (plus ou moins sensibles) relatives à un ensemble de n individus.

Calcul de grandeurs statistiques et confidentialité

Nous faisons l'hypothèse que les informations relatives à un individu $i \in [\![1, n_s]\!]$ est stockée sur un seul et unique serveur et que chaque $s,\ s \in [\![1, N]\!]$ contient un nombre n_s d'informations.

1. On s'intéresse à une variable aléatoire Y qui est attribut sensible relative à des individus, cependant, nous souhaiterions connaître quelques grandeurs statistiques à son sujet. Pour cela, chaque serveur s ne peut retourner que deux informations : (i) la valeur moyenne de Y, notée $\bar{Y}_s$ sur le serveur, ainsi que le nombre n_s d'individus.

   (a) Est-il possible de déterminer la valeur moyenne de la variable Y sur l'ensemble des serveurs ? Si oui, expliquer comment.

   Oui, cela est possible, il suffit de repasser par la définition de moyenne et de la réécrire comme une moyenne des valeurs moyennes sur chaque serveur, pondérée par le nombre d'exemples enregistrés sur chaque serveur.

   Par définition, nous avons

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

En demandant aux différents serveurs de renvoyer la moyenne et le nombre d'exemples associés, on peut alors exprimer notre moyenne globale comme

$$\bar{Y}_n = \frac{1}{n}\sum_{s=1}^{N} n_s \bar{Y}_s.$$

(b) Est-il possible de déterminer la valeur variance de la variable Y sur l'ensemble des serveurs ? Si oui, expliquer comment. Si non, que nous faudrait-il comme information ?
Indice : penser à l'ANOVA et à la décomposition de la variance

On rappelle que la variance est mesure de dispersion des valeurs d'un échantillon autour de sa valeur moyenne. Ainsi, pour la calculer, nous devrions donc avoir besoin de connaître les différentes valeurs de notre échantillon.
Par le biais de cette définition, la réponse semble donc négative car nous ne serons donc pas en mesure de préserver le caractère confidentiel de nos données.

Nous pouvons aussi faire le lien avec l'ANOVA et regarder comment décomposer notre variance.
En effet, l'ANOVA repose sur l'étude de la variance d'échantillons issus de différents groupes (les serveurs dans notre cas).
Cette variance totale peut alors être décomposée comme la somme de deux termes :

(i) la variance **inter-classe**

$$V_{\text{inter}} = \frac{1}{n}\sum_{s=1}^{N} n_s (\bar{Y}_s - \bar{Y})^2$$

ce terme peut également se voir comme **la variance des moyennes**

(ii) la variance **intra-classe**

$$V_{\text{intra}} = \frac{1}{n}\sum_{s=1}^{N}\sum_{i=1}^{n_s} (Y_i^{(s)} - \bar{Y}_s)^2$$

ce terme peut également se voir comme **la moyenne des variances**

Nous avons ainsi

$$\underbrace{\frac{1}{n}\sum_{s=1}^{N}\sum_{i=1}^{n_s} (Y_i^{(s)} - \bar{Y})^2}_{=V_{\text{total}}} = \underbrace{\frac{1}{n}\sum_{s=1}^{N}\sum_{i=1}^{n_s} (Y_i^{(s)} - \bar{Y}_s)^2}_{=V_{\text{intra}}} + \underbrace{\frac{1}{n}\sum_{s=1}^{N} n_s (\bar{Y}_s - \bar{Y})^2}_{=V_{\text{inter}}},$$

où $Y_i^{(s)}$ désigne le i-ème individu sur le serveur s.

Ainsi, il suffirait d'avoir accès à la variance de chaque serveur pour estimer la variance globale.

2. Comment calculer une médiane si les données sont stockées sur N serveurs **ordonnés** (c'est-à-dire que les plus petites valeurs sont dans le serveur 1 et les suivantes dans le serveurs 2 et ainsi de suite jusqu'aux plus grandes dans le serveur N).

Oui. D'abord, nous demandons à chaque serveur combien il y a d'individus n_s et nous calculons n puis nous cherchons le serveur s tel que $\sum_{k=1}^{s} n_k \leq n/2 \leq \sum_{k=1}^{s+1} n_k$ et nous lui demandons la valeur correspondante à $n/2$ c'est-à-dire, pour lui la $n/2 - \sum_{k=1}^{s} n_k$.

Optimisation du stockage des données

On regarde maintenant un problème pour stocker des données de façon optimale pour des traitements pré-définis. Une commanditaire vient me voir car elle souhaite stocker des données temporelles représentant la température toutes les 30 secondes sur une année de la façon la plus pertinente possible pour avoir accès à la moyenne entre deux moments t_1 et t_2 dans l'année. Quelle(s) information(s) doit-elle stocker pour que le calcul soit rapide. Quel(s) problème(s) numérique(s) risque-t-elle de rencontrer ? Proposer une solution.

La première intuition serait de stocker les températures T_t à chaque instant et de calculer les moyennes en faisant la somme :

$$\frac{1}{t_2 - t_1} \sum_{t=t_1+1}^{t_2} Y_t.$$

Cependant, cela nécessiterait d'effectuer un grand nombre d'opérations (nombre de sommes) mais aussi de recherche des éléments. Ce qui n'en fait pas une procédure optimale sur le plan industriel.

Nous pouvons alors décider de faire différemment, en stockant par exemple la somme des valeurs jusqu'au temps t, que l'on notera S_t. Nous aurons donc

$$S_t = \sum_{i=1}^{t} Y_t$$

et ainsi, le calcul de chaque moyenne serait quasi instantané :

$$\frac{1}{t_2 - t_1} \sum_{t=t_1+1}^{t_2} Y_t = \frac{1}{t_2 - t_1}(S_{t_2} - S_{t_1}).$$

Cette solution présente cependant un problème : le fait de stocker des sommes va potentiellement faire croître les ressources nécessaires pour stocker les données (le nombre de bits nécessaires pour stocker la valeur de S_t augmente avec t).
Une technique est alors de stocker la somme des écarts à une valeur $\mu \in \mathbb{R}$; par exemple, la moyenne annuelle des températures. Ainsi, nous aurons :

$$S_t^{(\mu)} = \sum_{t'=1}^{t} (Y_{t'} - \mu)$$

et pour calculer la moyenne, nous remarquons que :

$$\begin{aligned}
\frac{S_{t_2}^{(\mu)} - S_{t_1}^{(\mu)}}{t_2 - t_1} &= \frac{1}{t_2 - t_1}\left[\sum_{t=1}^{t_2}(Y_t - \mu) - \sum_{t=1}^{t_1}(Y_t - \mu)\right] \\
&= \frac{1}{t_2 - t_1} \sum_{t=t_1+1}^{t_2} (Y_t - \mu) \\
&= \frac{1}{t_2 - t_1}\left[\sum_{t=t_1+1}^{t_2} Y_t - (t_2 - t_1)\,\mu\right] \\
&= \overline{T}_{t_1:t_2} - \mu.
\end{aligned}$$

Ainsi, en prenant $\frac{S_{t_2}^{(\mu)} - S_{t_1}^{(\mu)}}{t_2 - t_1} + \mu$, nous obtenons la moyenne souhaitée. De plus, le fait de prendre une estimation de la moyenne annuelle pour μ permettrait que les valeurs stockées oscillent autour de 0.

Exercice 2 : Données manquantes

Soit $\mathbf{Y}_n = (Y_1, \ldots, Y_n)$ un n échantillon avec une loi ayant un moment d'ordre 2 et d'espérance μ dont $n - m$ valeurs sont manquantes avec $m \in \{1, \ldots, n-1\}$. Comme nous avons des variables indépendantes et identiquement distribuées, l'ordre des valeurs est arbitraire et nous supposons que les m premières valeurs sont celles qui sont connues et les $n - m$ suivantes sont manquantes. Nous étudions le vecteur

$$\tilde{\mathbf{Y}}_n = \big(Y_1, \ldots, Y_m, \underbrace{\bar{Y}_m, \ldots, \bar{Y}_m}_{n-m \text{ fois}}\big) \text{ où } \bar{Y}_m = \frac{1}{m}\sum_{i=1}^{m} Y_i. \tag{1}$$

1. Explicitez la forme de la moyenne empirique $\tilde{\bar{Y}}_n$ basée sur le vecteur $\tilde{\mathbf{Y}}_n$ de l'équation (1) en fonction de la moyenne $\bar{Y}_m$.

   Nous avons directement, en reprenant la définition de $\tilde{\mathbf{Y}}_n$, l'expression suivante pour sa moyenne :

   $$\tilde{\bar{Y}}_n = \frac{1}{n}\left(\sum_{i=1}^{m} Y_i + (n-m)\bar{Y}_m\right).$$

   On peut ensuite réécrire cette expression en fonction de $\bar{Y}_m$ uniquement en utilisant le fait que $\sum_{i=1}^{m} Y_i = m\bar{Y}_m$, ce qui nous donne

   $$\tilde{\bar{Y}}_n = \frac{1}{n}\left(m\bar{Y}_m + (n-m)\bar{Y}_m\right) = \bar{Y}_m.$$

   Comme nos observations sont indépendantes, on rappelle que l'espérance de la moyenne empirique, $\mathbb{E}[\bar{\mathbf{Y}}_m]$ est égale à l'espérance de Y, $\mathbb{E}[Y]$.
2. Montrez que la moyenne empirique $\tilde{\bar{Y}}_n$ basée sur le vecteur $\tilde{\mathbf{Y}}_n$ de l'équation (1) est presque sûrement égale à la moyenne empirique $\bar{Y}_m$ basée uniquement sur les observations présentes dans le n échantillon $\mathbf{Y}_n$.

   Comme nos observations sont indépendantes, on rappelle que l'espérance de la moyenne empirique, $\mathbb{E}[\bar{Y}_m]$ est égale à l'espérance de Y, $\mathbb{E}[Y]$, nous avons directement le résultat.
   L'estimateur de la moyenne $\bar{Y}_m$ est donc non biaisée, il en est donc de même de l'estimateur $\tilde{\bar{Y}}_n$.

3. Donnez la forme de l'estimateur de la variance en prenant :

   (a) toutes les valeurs du vecteur $\tilde{\mathbf{Y}}_n$ de l'équation (1).

   L'estimateur de la variance du vecteur $\tilde{\mathbf{Y}}_n$ est donné par

   $$\hat{\sigma}_n^2 = \frac{1}{n}\left[\sum_{i=1}^{m}\left(Y_i - \tilde{\bar{Y}}_n\right)^2 + \sum_{i=m+1}^{n}\left(\bar{Y}_m - \tilde{\bar{Y}}_n\right)^2\right].$$

   (b) uniquement les valeurs connues du n échantillon $\mathbf{Y}_n$.

   On se concentre uniquement sur les valeurs connues de ce même échantillon, ce qui nous donne l'estimateur suivant

   $$\hat{\sigma}_m^2 = \frac{1}{m}\sum_{i=1}^{m}\left(Y_i - \bar{Y}_m\right)^2.$$

4. Démontrez que l'estimation (a) faite avec le vecteur $\tilde{\mathbf{Y}}_n$ de l'équation (1) est presque sûrement strictement plus petite que l'estimation (b) faite uniquement avec les valeurs connues du n l'échantillon $\mathbf{Y}_n$.

Si nous prenons l'estimateur empirique de la variance, nous avons :

$$\begin{aligned}
\tilde{\sigma}_n^2 &= \frac{1}{n}\left[\sum_{i=1}^{m}\left(Y_i - \tilde{\bar{Y}}_n\right)^2 + \sum_{i=m+1}^{n}\left(\bar{Y}_m - \tilde{\bar{Y}}_n\right)^2\right], \\
&= \frac{1}{n}\left[\sum_{i=1}^{m}\left(Y_i - \bar{Y}_m\right)^2 + \sum_{i=m+1}^{n}\underbrace{\left(\bar{Y}_m - \bar{Y}_m\right)^2}_{=0}\right] \text{ car } \tilde{\bar{Y}}_n = \bar{Y}_m \text{ p.s,} \\
&= \frac{m}{n}\frac{1}{m}\sum_{i=1}^{m}\left(Y_i - \bar{Y}_m\right)^2, \\
&= \underbrace{\frac{m}{n}}_{<1}\sigma_m^2.
\end{aligned}$$

Donc l'estimateur de la variance basé sur le vecteur $\tilde{\bar{Y}}_n$ de l'équation (1) est presque sûrement strictement plus petit que l'estimateur basé uniquement sur les valeurs connues du n échantillon $\mathbf{Y}_n$.