Principes et Méthodes Statistiques
TP 2020
Chars d'assaut allemands et iPhones 3G

Le travail sera conduit par groupes de 2 ou 3 personnes, ces groupes étant constitués au hasard. Le délivrable de ce TP est une archive contenant deux fichiers. Le premier sera le compte-rendu du TP au format Rmd, qui comprendra le code `R` et vos réponses détaillées aux questions, selon les règles présentées ci-dessous. Le second sera le fichier pdf ou html résultant de la compilation (knit) du fichier Rmd. L'archive devra être déposée sur Teide avant le vendredi 17 avril 2020 à 22h. Tout retard sera pénalisé.

Le compte-rendu Rmd comprendra, suivant la nature des questions posées, des calculs mathématiques et/ou des sorties numériques et graphiques de `R`. Une grande importance sera accordée aux commentaires, visant à interpréter les résultats et mettre en valeur votre analyse du problème. Des conseils et des directives obligatoires pour la rédaction du compte-rendu sont disponibles sur Chamilo; les enseignants pourront y faire référence dans leur correction.

Le problème des chars d'assaut allemands

Voir `https://en.wikipedia.org/wiki/German_tank_problem`.

Le problème des chars d'assaut (ou tanks) allemands est un problème statistique consistant à estimer la valeur maximale d'une variable aléatoire de loi uniforme discrète. Son nom vient du fait qu'il a été utilisé par les alliés pendant la seconde guerre mondiale, dans le but d'estimer la production de tanks par la Wehrmacht. A l'issue de la guerre, les données de l'industrie d'armement allemande ont permis de montrer que les estimations obtenues par cette méthode étaient beaucoup plus proches de la réalité que celles déduites des informations fournies par les services de renseignement.

Chaque tank est identifié par un numéro de série. Une transformation permet de considérer que ces numéros de série sont des entiers allant de 1 jusqu'au nombre total de tanks produits. Ce nombre est un entier inconnu θ, que l'on cherche à estimer. Les alliés

ont relevé les numéros de série de tous les tanks capturés. Le principe de la méthode est de considérer que relever le numéro de série des tanks capturés revient à faire un tirage de variables aléatoires sur $\{1, \ldots, \theta\}$.

Comme un tank n'est capturé qu'une seule fois, il s'agit d'un tirage dit *sans remise*. Mais pour commencer, on va traiter le problème plus simple de tirage dit *avec remise*.

1 Tirage avec remise

Une autre façon de modéliser le problème est de considérer que l'on a un récipient contenant un nombre inconnu θ d'objets numérotés de 1 à θ. On effectue n tirages au hasard dans le récipient. Après chaque tirage, on note le numéro de l'objet obtenu et on remet celui-ci dans le récipient. Les numéros successifs des objets tirés sont notés $x_1, \ldots, x_n$. Les conditions de cette expérience font que ces données sont des réalisations de variables aléatoires $X_1, \ldots, X_n$ indépendantes et de même loi uniforme sur $\{1, \ldots, \theta\}$, $\mathcal{U}_{\{1,\ldots,\theta\}}$. L'objectif est d'estimer θ à partir de $x_1, \ldots, x_n$.

Une variable aléatoire X est de loi $\mathcal{U}_{\{1,\ldots,\theta\}}$ si elle est à valeurs dans $\{1, \ldots, \theta\}$ et que

$$P(X = k) = \frac{1}{\theta} \mathbb{1}_{\{1,\ldots,\theta\}}(k).$$

1. Calculer l'espérance et la variance de X.

2. Calculer l'estimateur des moments $\tilde{\theta}_n$ de θ. Montrer que cet estimateur est sans biais et calculer sa variance.

3. Calculer la fonction de répartition de X. Calculer la médiane de la loi de X et en déduire un estimateur $\tilde{\theta}'_n$ de θ basé sur la médiane empirique.

4. Soit X_n^* le maximum des observations. Calculer la fonction de répartition de X_n^* et les probabilités élémentaires $P(X_n^* = k)$, $\forall k \in \{1, \ldots, \theta\}$.

5. Montrer que l'estimateur de maximum de vraisemblance de θ est $\hat{\theta}_n = X_n^*$. Montrer qu'il est biaisé mais qu'on ne peut pas le débiaiser facilement.

6. Expliquer comment construire le graphe de probabilités pour la loi uniforme discrète. En déduire un estimateur graphique θ_g de θ.

   On peut en fait montrer que l'estimateur sans biais et de variance minimale de θ est :

$$\check{\theta}_n = \frac{X_n^{*n+1} - (X_n^* - 1)^{n+1}}{X_n^{*n} - (X_n^* - 1)^n}.$$

Dans la suite de cette première partie, on va comparer numériquement les 5 estimateurs $\tilde{\theta}_n, \tilde{\theta}'_n, \hat{\theta}_n$, θ_g et $\check{\theta}_n$ à l'aide de simulations en `R`.

7. En `R`, la simulation de la loi uniforme discrète se fait avec la commande `sample`. `sample(1:20,10,replace=T)` tire 10 nombres au hasard entre 1 et 20 avec remise, tandis que `sample(1:20,10)` tire 10 nombres au hasard entre 1 et 20 sans remise.

   Simuler un échantillon de taille $n = 20$ d'une loi $\mathcal{U}_{\{1,\dots,\theta\}}$, avec $\theta = 1000$. Tracer un histogramme et le graphe de probabilités pour la loi uniforme discrète. Calculez les 5 estimations de θ. Commentez les résultats.

8. Simuler m échantillons de taille n d'une loi $\mathcal{U}_{\{1,\dots,\theta\}}$, avec $\theta = 1000$. Pour chaque échantillon, calculer les valeurs des 5 estimations de θ. On obtient ainsi des échantillons de m valeurs de chacun des 5 estimateurs. Evaluer le biais et l'erreur quadratique moyenne de ces estimateurs. Faites varier m et n. Qu'en concluez-vous ?

9. Déterminer un intervalle de confiance asymptotique de seuil α pour θ, c'est-à-dire un intervalle aléatoire I_n tel que

$$\lim_{n\to\infty} P(\theta \in I_n) = 1 - \alpha.$$

10. Simuler m échantillons de taille n d'une loi $\mathcal{U}_{\{1,\dots,\theta\}}$. Calculer le pourcentage de fois où l'intervalle de confiance de seuil α pour θ contient la vraie valeur du paramètre θ. Faire varier n, m et α, et conclure.

2 Tirage sans remise

Dans le problème des chars allemands, un tank n'est capturé qu'une seule fois. Cela revient à considérer dans la modélisation précédente que l'objet tiré n'est pas remis dans le récipient. Par conséquent, les variables aléatoires $X_1, \dots, X_n$ représentant les numéros successifs des objets tirés ne sont ni indépendantes ni de même loi.

1. Déterminer la loi de X_1, puis celle de X_2 sachant $[X_1 = x_1]$, puis celle de X_3 sachant $[X_1 = x_1, X_2 = x_2]$, etc... Etant donné que la fonction de vraisemblance peut s'écrire

$$\begin{aligned}\mathcal{L}(\theta; x_1, \dots, x_n) &= P(X_1 = x_1, \dots X_n = x_n; \theta) \\ &= P(X_1 = x_1; \theta) \prod_{i=2}^{n} P(X_i = x_i | X_1 = x_1, \dots X_{i-1} = x_{i-1}; \theta),\end{aligned}$$

   montrer que l'estimateur de maximum de vraisemblance de θ est toujours $\hat{\theta}_n = X_n^*$.

2. Montrer que $\forall k \in \{n, \ldots, \theta\}$, $P(X_n^* = k) = \frac{\binom{k-1}{n-1}}{\binom{\theta}{n}}$. Calculer $E[X_n^*]$ et en déduire que $\hat{\theta}_n^{(1)} = \frac{n+1}{n} X_n^* - 1$ est un estimateur sans biais de θ.

3. Une façon intuitive de construire un autre estimateur est la suivante. Pour des raisons de symétrie, il est logique de s'attendre à ce que le nombre de numéros inférieurs au minimum des numéros tirés soit proche du nombre de numéros supérieurs au maximum des numéros tirés. Autrement dit, $x_1^* - 1 \approx \theta - x_n^*$. Cela amène à proposer un nouvel estimateur, $\hat{\theta}_n^{(2)} = X_n^* + X_1^* - 1$.

   A l'aide de `R`, faites des expérimentations numériques ayant pour objectif de comparer les estimateurs $\hat{\theta}_n$, $\hat{\theta}_n^{(1)}$ et $\hat{\theta}_n^{(2)}$, ainsi que l'estimateur $\tilde{\theta}_n$ calculé dans la question 1.2.

4. Pour estimer θ, peut-on se contenter de considérer que le tirage est avec remise ?

3 Estimation du nombre d'iPhones 3G produits

Le problème des chars d'assaut allemands a été réutilisé plus récemment dans un tout autre contexte. A l'occasion de la sortie de l'iPhone 3G en juillet 2008, des internautes ont voulu estimer par eux-mêmes le nombre d'unités produites. Pour cela, ils ont demandé aux possesseurs de ces mobiles de renseigner sur un fil consacré les deux numéros qui identifient un téléphone portable, le numéro IMEI et le code de production PC.

- Le numéro IMEI (*International Mobile Equipment Identity*) est délivré par une autorité indépendante. Il est constitué de 15 chiffres.

  1. Les 8 premiers constituent le TAC (*Type Allocation Code*).
     - Les deux premiers chiffres désignent le code du pays où le mobile a été immatriculé. Par exemple, 01 désigne les Etats-Unis.
     - Les 6 derniers chiffres fournissent un code permettant d'identifier un million de téléphones du même modèle. Par exemple, le code 161200 correspond au premier million de mobiles produits, le code 161300 correspond au deuxième million de mobiles produits, et ainsi de suite. Le code sera appelé *code TAC* et le numéro de million correspondant *numéro TAC*. La correspondance entre les deux est donnée dans la table 1.
  2. Les 6 chiffres suivants désignent le *numéro SNR* de fabrication du mobile.

code TAC	numéro TAC
161200	1
161300	2
161400	3
171200	4
171300	5
171400	6
174200	7
174300	8
174400	9
177100	10
177300	11
177400	12
177500	13
177600	14
180900	15

Table 1: IMEI : correspondance entre code TAC et numéro TAC

3. Le dernier chiffre est un chiffre de contrôle.

L'IMEI permet de reconstruire un numéro de série NS identifant un mobile :

$$NS = (\texttt{numéro TAC-1}) \times 10^6 + SNR$$

Par exemple, l'IMEI 011613006769038 = 01-161300-676903-8 donne comme numéro de série

$$NS = (\texttt{2-1}) \times 10^6 + 676903 = 1676903$$

- Le *code de production PC* est propre au constructeur, ici Apple. Il est constitué de 6 chiffres.

  1. Les deux premiers chiffres désignent l'usine de fabrication. Si on a *5K* à la place, il s'agit d'un produit reconditionné.
  2. Le troisième chiffre désigne l'année de production : 8 pour 2008, 9 pour 2009.
  3. Les quatrième et cinquième chiffres désignent la semaine de production.
  4. Le dernier chiffre est un chiffre de contrôle.

  Ainsi le code PC 878293=87-8-29-3 désigne un téléphone produit dans l'usine numéro 87, lors de la 29ème semaine de l'année 2008.

Par souci d'anonymat, seuls les 13 premiers chiffres de l'IMEI et les 5 premiers du PC ont été recueillis. Dans les exemples cités plus haut, le possesseur du mobile ayant pour IMEI 011613006769038 et pour PC 878293 a fourni les codes 0116130067690XX et 87829X. Il n'est donc pas possible de reconstituer le numéro de série exact 1676903. On l'approche en remplaçant le dernier chiffre inconnu par 0. On obtient donc comme numéro de série approché $NS = 1676900$.

Le fichier `iPhones.csv` contient un extrait de cette enquête contenant 139 réponses obtenues entre juillet 2008 et février 2009. Charger ce tableau de données dans `R` en utilisant les commandes :

```
> iPhones<-read.table("iPhones.csv", sep=";", header=T)
> names(iPhones)
> attach(iPhones)
```

1. Reconstituer les numéros de série NS de tous ces mobiles. Pour manipuler les chaînes de caractères en `R`, on pourra utiliser les commandes `as.character, as.numeric` et `substring`.

2. Estimer le nombre total d'iPhones produits durant la période concernée.

3. On veut suivre plus finement la progression de la production d'iPhones tout au long de cette période. Pour cela, on regroupe les données par paquets de 4 semaines : le premier groupe comporte tous les appareils produits entre les 25ème et 28ème semaine de 2008, etc... et le dernier tous les appareils produits entre les 1ère et 4ème semaine de 2009.

   Estimer le nombre d'iPhones produits sur chacune de ces sous-périodes.

4. Ces estimations reposent sur l'hypothèse d'uniformité des numéros de série. Que pensez-vous de la validité de cette hypothèse, sur l'ensemble de la période et sur chacune des sous-périodes définies dans la question précédente ?

5. Quelles conclusions tirez-vous de cette étude ?