# Principes et Méthodes Statistiques
# ENSIMAG 1ère année - TP 2016
# Chars d'assaut allemands et iPhones 3G

CARRE Ludovic EL IDRISSI BOUTAHER Mehdi
LEFOULON Vincent

Avril 2016

## Le problème des chars d'assaut allemands

## 1 Tirage avec remise

1. *Calculer l'espérance et la variance de $X$.*
   Comme $X$ est une variable aléatoire discrète, on a par définition :

$$E(X) = \sum_{k=0}^{+\infty} kP(X = k).$$

D'où :

$$\begin{aligned} E(X) &= \sum_{k=0}^{+\infty} k\frac{1}{\theta}1_{\{1,\dots,\theta\}}(k) \\ &= \sum_{k=1}^{\theta} k\frac{1}{\theta} \\ &= \frac{1}{\theta}\frac{\theta(\theta+1)}{2} \\ &= \frac{\theta+1}{2} \end{aligned}$$

De plus :

$$\begin{aligned} E(X^2) &= \sum_{k=1}^{\theta} k^2\frac{1}{\theta} \\ &= \frac{1}{\theta}\frac{\theta(\theta+1)(2\theta+1)}{6} \\ &= \frac{(\theta+1)(2\theta+1)}{6} \end{aligned}$$

Alors

$$Var(X) = E(X^2) - E(X)^2 = \frac{\theta^2 - 1}{12}.$$

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

   D'après la question précédente, $\theta = 2E(X) - 1$.

   Donc $\tilde{\theta}_n = 2\bar{X}_n - 1$, où $\bar{X}_n$ est une variable aléatoire désignant la moyenne empirique de $n$ observations.

   On a alors :

$$\begin{aligned} E(\tilde{\theta}_n) &= E(2\bar{X}_n - 1) \\ &= 2E(\bar{X}_n) - 1 \\ &= 2E(\frac{1}{n}\sum_{i=1}^{n} X_i) - 1 \\ &= 2\frac{1}{n}\sum_{i=1}^{n} E(X_i) - 1 \\ &= \theta \\ &= E(X) \end{aligned}$$

   Donc l'estimateur est sans biais.

   De plus :

$$\begin{aligned} Var(\tilde{\theta}_n) &= Var(2\bar{X}_n - 1) \\ &= 4Var(\bar{X}_n) \\ &= 4Var(\frac{1}{n}\sum_{i=1}^{n} X_i) \\ &= \frac{4}{n^2}\sum_{i=1}^{n} Var(X_i) \text{ par indépendance} \\ &= \frac{4}{n}Var(X) \\ &= \frac{4(\theta^2 - 1)}{12n} \end{aligned}$$

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 $\theta$ basé sur la médiane empirique.*

   $X$ est à valeurs dans $\{1, \ldots, \theta\}$, donc, avec $x$ réel :

$$\begin{cases} \forall x < 1, P(X \leq x) = 0 \\ \forall x > \theta, P(X \leq x) = 1 \end{cases}$$

   Soit $x \in [1, \theta]$. On a :

$$\begin{aligned} P(X \leq x) &= \sum_{i=1}^{\theta} P(X = i) 1_{i \leq x} \\ &= \sum_{i=1}^{\lfloor x \rfloor} \frac{1}{\theta} \\ &= \frac{\lfloor x \rfloor}{\theta} \end{aligned}$$

En somme :

$$\forall x \in \mathbb{R}, P(X \leq x) = \begin{cases} 0 \text{ si } x < 1 \\ \frac{\lfloor x \rfloor}{\theta} \text{ si } 1 \leq x \leq \theta \\ 1 \text{ si } x > \theta \end{cases}$$

Notons $m \in \mathbb{R}$ la médiane de $X$. Par définition, $m$ vérifie :

$$(S) \begin{cases} P(X \leq m) = \frac{1}{2} \\ P(X \geq m) = \frac{1}{2} \end{cases}$$

Comme $m \in \{1, \ldots, \theta\}$ :

$$\begin{aligned} (S) &\Leftrightarrow \begin{cases} \frac{\lfloor m \rfloor}{\theta} = \frac{1}{2} \\ P(X > m) + P(X = m) = \frac{1}{2} \end{cases} \\ &\Leftrightarrow \begin{cases} \frac{\lfloor m \rfloor}{\theta} = \frac{1}{2} \\ 1 - \frac{\lfloor m \rfloor}{\theta} + P(X = m) = \frac{1}{2} \end{cases} \\ &\Leftrightarrow m = \frac{\theta + 1}{2} \end{aligned}$$

Donc $\theta_n^{'} = 2\tilde{Med}_n - 1$.

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

$$\begin{aligned} P(X_n^* \leq x) &= P(X_1 \leq x \cap \ldots \cap X_n \leq x) \\ &= \prod_{i=1}^{n} P(X_i \leq x) \text{ par indépendance} \\ &= \prod_{i=1}^{n} \frac{\lfloor x \rfloor}{\theta} \\ &= (\frac{\lfloor x \rfloor}{\theta})^n \end{aligned}$$

En somme :

$$\forall x \in \mathbb{R}, P(X_n^* \leq x) = \begin{cases} 0 \text{ si } x < 1 \\ (\frac{\lfloor x \rfloor}{\theta})^n \text{ si } 1 \leq x \leq \theta \\ 1 \text{ si } x > \theta \end{cases}$$

Calculons maintenant les probabilités élémentaires. Pour $k \in \{1, \ldots, \theta\}$, on a :

$$\begin{aligned} P(X_n^* = k) &= P(X_n^* \leq k) - P(X_n^* \leq k-1) \\ &= \frac{k^n - (k-1)^n}{\theta^n} \end{aligned}$$

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

$$\begin{aligned} \mathcal{L}(\theta; x_1, \ldots, x_n) &= \prod_{i=1}^{n} P(X = x_i; \theta) \\ &= \begin{cases} 0 \text{ si } \theta < \max x_i \\ \dfrac{1}{\theta^n} \text{ sinon} \end{cases} \end{aligned}$$

   Comme $n \geq 0$ et $x_i \geq 1$, on en déduit $\max \mathcal{L}(\theta; x_1, \ldots, x_n) = \max x_i$.
   D'où $\hat{\theta}_n = X_n^*$.
   De plus :

$$\begin{aligned} E(\hat{\theta}_n) &= E(X_n^*) \\ &= \sum_{k=1}^{+\infty} P(X_n^* > k) \\ &= \sum_{k=1}^{+\infty} (1 - P(X_n^* \leq k)) \\ &= \sum_{k=1}^{\theta} (1 - \frac{k^n}{\theta^n}) \\ &= \theta - \sum_{k=1}^{\theta} \frac{k^n}{\theta^n} \\ &\neq \theta \end{aligned}$$

   Ainsi, l'estimateur est biaisé. On peut difficilement le débiaiser parce que le biais dépend de $\theta$, inconnu.

6. *Expliquer comment construire le graphe de probabilités pour la loi uniforme discrète. En déduire un estimateur graphique $\theta_g$ de $\theta$.*
   D'après la question 1.3, pour $1 \leq x \leq \theta$, $F(x) = \frac{\lfloor x \rfloor}{\theta}$.
   Donc $F(x)$ est directement de la forme $\alpha(\theta)g(x) + \beta(\theta)$, avec $g = \lfloor . \rfloor$, $\alpha(\theta) = \frac{1}{\theta}$ et $\beta = 0$.

Pour estimer graphiquement $\theta$ à partir des observations $x_i$, on regarde le nuage de points $(\lfloor x_i^* \rfloor, i/n)$. On devrait obtenir approximativement une droite, de pente $1/\theta_g$.

*On peut en fait montrer que l'estimateur sans biais et de variance minimale de $\theta$ 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 $\tilde{\theta}_n, \tilde{\theta}'_n, \hat{\theta}_n$, $\theta_g$ 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 $\theta$. 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 $\theta$. 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 $\alpha$ pour $\theta$, 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 $\alpha$ pour $\theta$ contient la vraie valeur du paramètre $\theta$. Faire varier $n$, $m$ et $\alpha$, 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 $\theta$ est toujours $\hat{\theta}_n = X_n^*$.*

On a :

$$P(X_1 = x_1) = \frac{1}{\theta}\mathbb{1}_{[0;\theta]}(x_1)$$
$$P(X_2 = x_2|X_1 = x_1) = \frac{1}{\theta - 1}\mathbb{1}_{[0;\theta]}(x_2)$$
$$P(X_i = x_i|\bigcap_{j=1}^{i-1}) = \frac{1}{\theta - i + 1}\mathbb{1}_{[0;\theta]}(x_i)$$

La fonction de vraisemblance s'écrit donc :

$$\begin{aligned}\mathcal{L}(\theta; (x_1, \ldots, x_n)) &= P(X_1 = x_1)\prod_{i=2}^{n} P(X_i = x_i|\bigcap_{j=1}^{i-1} X_j = x_j) \\ &= \frac{1}{\theta}\prod_{i=2}^{n}\frac{1}{\theta - i + 1}\mathbb{1}_{[0;\theta]}(x_i) \\ &= \prod_{i=0}^{n-1}\frac{1}{\theta - i}\mathbb{1}_{[1;\theta]}(\max_{0\leq j\leq\theta}(x_i)) \\ &= \frac{1}{(\theta - n)!}\mathbb{1}_{[1;\theta]}(\max_{0\leq j\leq\theta}(x_i))\end{aligned}$$

Donc, le $\hat{\theta}$ qui maximise $\mathcal{L}(\theta; (x_1, \ldots, x_n))$ est $\hat{\theta}_n = X_n^*$ puisque c'est le plus petit $\hat{\theta}$ qui permet à l'indicatrice de valoir 1.

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 estimateur sans biais de* $\theta$.

A)

$$P(X_n^* = k) = P(X_n^* \leq k)P(X_n^* \leq k - 1)$$

Or, par la formule des probabilités composées.

$$\begin{aligned}P(X_n^* \leq k) &= P(\bigcap_{i=1}^{n} X_i \leq k) \\ &= P(X_1 \leq k)\prod_{i=2}^{n} P(X_i \leq k|\bigcap_{j=1}^{i-1} X_j \leq k) \\ &= \frac{k}{\theta}\prod_{i=2}^{n}\frac{k - i + 1}{\theta - i + 1} = \prod_{i=0}^{n-1}\frac{k - i}{\theta - i}\end{aligned}$$

Nous avons ainsi :

$$\prod_{i=0}^{n-1} k-i = \frac{k!}{(k-n+1-1)!} = \frac{k!}{(k-n)!}$$
$$\prod_{i=0}^{n-1} \theta-i = \frac{\theta!}{(\theta-n+1-1)!} = \frac{\theta!}{(\theta-n)!}$$

Donc,

$$P(X_n^* \leq k) = \frac{\frac{k!}{(k-n)!}}{\frac{\theta!}{(\theta-n)!}}$$

$$= \frac{\frac{k!}{n!(k-n)!}}{\frac{\theta!}{n!(\theta-n)!}} = \frac{\binom{k}{n}}{\binom{\theta}{n}}$$

On a donc finalement,

$$P(X_n^* = k) = P(X_n^* \leq k)P(X_n^* \leq k-1)$$

$$= \frac{\binom{k}{n}}{\binom{\theta}{n}} - \frac{\binom{k-1}{n}}{\binom{\theta}{n}} = \frac{\binom{k-1}{n-1}}{\binom{\theta}{n}}$$

Car,

$$\begin{aligned}\binom{k}{n} - \binom{k-1}{n} &= \frac{k! - (k-1)!(k-n)}{n!(k-n)!} \\ &= \frac{k! - (k-1)!k + (k-1)!n}{n!(k-n)!} \\ &= \frac{(k-1)!}{(n-1)!(k-n)!} = \binom{k-1}{n-1}\end{aligned}$$

B)

$$\begin{aligned}E[X_n^*] &= \sum_{k=n}^{\theta} \frac{k(k-1)!n!(\theta-n)!}{(n-1)!(k-n)!\theta!} \\ &= \frac{n!(\theta-n)!}{\theta!} \sum_{k=n}^{\theta} \frac{nk!}{(k-n)!n!} \\ &= \frac{n!(\theta-n)!}{\theta!} n \sum_{k=n}^{\theta} \binom{k}{k-n}\end{aligned}$$

Or,

$$\sum_{k=n}^{\theta} \binom{k}{k-n} = \sum_{i=0}^{\theta-n} \binom{n+1}{i} = \binom{\theta+1}{\theta-1}$$

Donc,

$$\begin{aligned} E[X_n^*] &= \frac{n!(\theta-n)!}{\theta!} n \binom{\theta+1}{\theta-1} \\ &= \frac{n!(\theta-n)!n(\theta+1)!}{\theta!(\theta-n)!(n+1)!} \\ &= \frac{n!n(\theta+1)!}{\theta!(n+1)!} = n\frac{\theta+1}{n+1} \end{aligned}$$

Si $\hat{\theta}_n^{(1)}$ est un estimateur sans biais alors on a : $E[\hat{\theta}_n^{(1)}] = \theta$. D'où,

$$E[X_n^*] = n\frac{\theta+1}{n+1}$$
$$(n+1)\frac{E[X_n^*]}{n} - 1 = \theta$$

Et puisque :

$$(n+1)\frac{E[X_n^*]}{n} - 1 = E[\frac{n+1}{n}X_n^* - 1] = \theta$$

On a bien l'estimateur sans biais $\hat{\theta}_n^{(1)} = E[\frac{n+1}{n}X_n^* - 1]$.

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, $\hat{\theta}_n^{(2)} = X_n^* + X_1^* - 1$.*
   Je n'ai rien à répondre là-dessus si ?

4. 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.

5. Pour estimer $\theta$, 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 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.

| 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 |

TAB. 1 – IMEI : correspondance entre code TAC et numéro TAC

2. Les 6 chiffres suivants désignent le *numéro SNR* de fabrication du mobile.
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 ?