

# Le problème du char d’assaut allemand

MARC-ANDRÉ DÉSAUTELS, DÉPARTEMENT DE MATHÉMATIQUES  
CÉGEP DE SAINT-JEAN-SUR-RICHELIEU  
[marc-andre.desautels@cstjean.qc.ca](mailto:marc-andre.desautels@cstjean.qc.ca)  
<https://www.cstjean.qc.ca/>

## Résumé

Durant la seconde guerre mondiale, les alliés avaient un besoin criant d’estimer avec précision la quantité de matériel militaire que l’Allemagne nazie produisait. Les estimations provenant des services de renseignements habituels étaient contradictoires et incertaines. Les gouvernements Britanniques et Américains se tournèrent donc vers des statisticiens pour savoir si leurs estimations pouvaient être améliorées. Nous présenterons une introduction aux notions mathématiques utilisées.

**Mots clés :** Statistiques, Estimation, Simulation

## 1 Introduction

TODO [https://rstudio-pubs-static.s3.amazonaws.com/59464\\_ce1bbf886d0841cd8ecc46c1852b352e.html](https://rstudio-pubs-static.s3.amazonaws.com/59464_ce1bbf886d0841cd8ecc46c1852b352e.html)

Au début de l’année 1943, la *Economic Warfare Division* de l’ambassade américaine à Londres commença à analyser divers marquages obtenus à partir d’équipements allemands capturés ou détruits sur le front. Plus particulièrement, les numéros de séries ont été utilisés pour estimer la force de production de la machine de guerre allemande. L’article *An Empirical Approach to Economic Intelligence in World War II* [6] explique en grand détail le développement des techniques utilisés pour faire cette estimation, les problèmes rencontrés et les solutions qui y ont remédiés. Nous invitons le lecteur intéressé par l’aspect pratico-pratique de ces techniques à lire cet article.

Le problème du char d’assaut allemand est nommé d’après son application par les alliés à l’estimation du nombre de chars d’assaut produits par l’Allemagne. Mais en fait, ce problème regroupe l’estimation du nombre de nombreux produits de guerre allemands, par exemple les camions, les fusils, les bombes et les fusées, voir [6]. Les russes ont également utilisé des techniques similaires pour estimer la production de char d’assaut allemands [5]. Durant les années 80, des militaires Américains ont eu accès aux lignes de production de chars d’assaut *Merkava* israéliens. Le nombre de chars produits était confidentiel, mais le colone Dupuy indique

que puisque chaque chars possédait un numéro de série, il aurait pu estimer la production, voir [1]

Les formules développées dans cet article ont aussi été utilisées dans des contextes non-militaires. Par exemple, elles ont été utilisées pour estimer le nombre de *Commodore 64* produits, voir [7]. Le premier modèle de iPhone, commercialisé en 2007 aux États-Unis, a été vendu au nombre de 9 190 680, suite à des estimations faites à partir des codes IMEI (*The International Mobile Equipment Identity*) de nombreux utilisateurs, voir [8].

Dans cet article, nous nous intéresserons à l'estimation du nombre d'items  $N$  à partir d'un échantillon aléatoire dans le cas où les items sont numérotés de façon séquentielle.

## 2 Les mathématiques

### 2.1 Préalables

Supposons que nous avons une population d'objets numérotés de la façon suivante : 1, 2, 3,  $\dots$ ,  $N$ , où  $N$  est **inconnu**. En d'autres mots, les objets de notre population doivent être numérotés de façon *séquentielle*. Nous pigeons, **sans remise**, un échantillon  $X_1, X_2, X_3, \dots, X_n$ , de taille  $n$  à partir de la population. Nous voulons estimer la valeur de  $N$  à partir de l'échantillon prélevé.

Pour calculer les diverses mesures statistiques dont nous aurons besoin, nous allons classer les unités statistiques de notre échantillon en ordre croissant. Nous avons :

$$X_{(1)} < X_{(2)} < X_{(3)} < \dots < X_{(n-1)} < X_{(n)}$$

où les valeurs  $X_{(1)}, X_{(2)}, X_{(3)}, \dots, X_{(n)}$  sont les valeurs ordonnées de l'échantillon  $X_1, X_2, X_3, \dots, X_n$ . En particulier,  $X_{(1)}$  est la plus petite valeur de l'échantillon et  $X_{(n)}$  est la plus grande.

À partir de nos définitions précédentes, il est possible de calculer l'espérance de la valeur  $X_{(A)}$  ( $\mathbb{E}(X_{(A)})$ ), la variance de la valeur  $X_{(A)}$  ( $\text{Var}(X_{(A)})$ ) et enfin la covariance des valeurs  $X_{(A)}$  et  $X_{(B)}$  ( $\text{Cov}(X_{(A)}, X_{(B)})$ ). Nous utiliserons ces mesures statistiques pour calculer l'espérance et la variance des estimateurs que nous construirons. Malheureusement, retrouver ces mesures statistiques nécessite des identités combinatoires et de fastidieux calculs. Pour ne pas alourdir le texte, nous donnerons ces mesures sans démonstration. Par contre, pour la lectrice ou le lecteur intéressé, vous pourrez trouver à l'annexe B une idée de la technique utilisée ainsi que la démonstration de  $\mathbb{E}(X_{(A)})$ .

Nous utiliserons donc, sans démonstration, les trois mesures du tableau 1. La lectrice ou le lecteur curieux pourront aller aux annexes A et B pour quelques rappels sur les notions d’espérance, de variance et de covariance ainsi que la démonstration de  $E(X_{(A)})$ . Les démonstrations de  $\text{Var}(X_{(A)})$  et  $\text{Cov}(X_{(A)}, X_{(B)})$  se font d’une manière similaire.

Mesure	Formule
$\mathbb{E}(X_{(A)})$	$\frac{A(N+1)}{n+1}$
$\text{Var}(X_{(A)})$	$\frac{A(n+1-A)(N+1)(N-n)}{(n+1)^2(n+2)}$
$\text{Cov}(X_{(A)}, X_{(B)})$	$\frac{A(n+1-B)(N+1)(N-n)}{(n+1)^2(n+2)}$

TABLEAU 1 – Les mesures de l’espérance, de la variance et de la covariance pour une valeur  $X_{(A)}$  de notre échantillon.

À l’aide des trois mesures du tableau 1, nous allons maintenant trouver quatre estimés de  $N$  en utilisant simplement notre “gros bon sens”. La structure des prochaines sections est calquée sur [1]. Nous utiliserons aussi les notions présentées en [2] et [3].

## 2.2 Les trois situations possibles

Supposons que nous avons une population d’objets numérotés de la façon suivante :  $s+1, s+2, s+3, \dots, s+N$ . Trois situations distinctes peuvent se produire :

1.  $s$  est **connu** et égal à 0 et  $N$  est **inconnu**.
2.  $s$  est **connu** mais différent de 0 et  $N$  est **inconnu**.
3.  $s$  est **inconnu** et  $N$  est **inconnu**.

Nous étudierons, dans l’ordre, les trois situations précédentes.

## 2.3 La situation où $s$ est connu et égal à 0 et $N$ est inconnu

Puisque  $s$  est connu et égal à 0, nous nous trouvons dans la situation où notre liste est numérotée de la façon suivante :  $1, 2, \dots, N$ . Cette situation est en fait un cas particulier d’une loi uniforme discrète de la forme  $U(1, N)$ .

### 2.3.1 Le milieu de la liste

Pour trouver nos deux premiers estimateurs, nous allons débiter en supposant que nous connaissons la valeur milieu  $m$  de la liste  $1, 2, \dots, N$ . Nous nous retrouvons dans la situation ci-dessous :

$$\underbrace{1, 2, 3, \dots, m-1}_{m-1 \text{ éléments}}, m, \underbrace{m+1, \dots, N-2, N-1, N}_{m-1 \text{ éléments}}$$

Il y aura donc  $m-1$  valeurs en-dessous de  $m$  et  $m-1$  valeurs au-dessus de  $m$ . Donc, si nous incluons la valeur milieu  $m$ , nous avons  $N = (m-1) + 1 + (m-1) = 2m-1$ . Puisque nous ne connaissons pas la valeur milieu  $m$ , il est raisonnable de le remplacer par une estimation, par exemple la médiane  $\tilde{X}$  ou la moyenne  $\bar{X}$ . Nous pouvons maintenant obtenir nos deux premiers estimateurs.

### 2.3.2 La médiane

Nous noterons notre premier estimateur  $\widehat{N}_1 = 2\tilde{X} - 1$ , où  $\tilde{X}$  représente la médiane de notre échantillon. Rappelons que pour  $k$  données discrètes, la médiane se calcule de deux façons différentes, dépendamment du fait que le nombre de données soit pair ou impair.

$$\tilde{X} = \begin{cases} \frac{1}{2} \left( X_{\left(\frac{k}{2}\right)} + X_{\left(\frac{k}{2}+1\right)} \right) & \text{si } k \text{ est pair} \\ X_{\left(\frac{k+1}{2}\right)} & \text{si } k \text{ est impair} \end{cases} \quad (1)$$

Nous voulons vérifier si l'estimateur  $\widehat{N}_1$  est non-biaisé, c'est-à-dire si l'espérance de  $\widehat{N}_1$  est égale à  $N$  ( $\mathbb{E}(\widehat{N}_1) = N$ ). Nous voulons également calculer la variance de  $\widehat{N}_1$ . Pour éviter d'alourdir le texte, nous avons mis un exemple de calcul d'espérance et de variance à l'annexe C. Nous obtenons les résultats du tableau

$i$	$\widehat{N}_i$	$\mathbb{E}(\widehat{N}_i)$	$\text{Var}(\widehat{N}_i)$
1	$2\tilde{X} - 1$	$N$	$\begin{aligned} & \frac{(N-n)(N+1)}{n+2} \text{ pour } n \text{ impair} \\ & \frac{(N-n)(N+1)}{n+2} \text{ pour } n \text{ pair} \end{aligned}$

TABLEAU 2 – Les mesures de l'espérance, de la variance et de la covariance pour une valeur  $X_{(A)}$  de notre échantillon.

En utilisant les mesures du tableau 1, nous obtenons :

$$\begin{aligned} E\left(\widehat{N}_1\right) &= E\left(2\tilde{X} - 1\right) \\ &= 2E\left(\tilde{X}\right) - 1 \quad \text{par propriétés des espérances} \end{aligned}$$

Pour continuer, nous devons distinguer le cas pair du cas impair et utiliser l'équation 1.

**Le cas pair** Dans le cas pair, nous avons :

$$\begin{aligned} E\left(\widehat{N}_1\right) &= 2E\left(\tilde{X}\right) - 1 \\ &= 2E\left(\frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2}\right) - 1 \\ &= E\left(X_{(\frac{n}{2})}\right) + E\left(X_{(\frac{n}{2}+1)}\right) - 1 \\ &= \frac{\left(\frac{n}{2}\right)(N+1)}{n+1} + \frac{\left(\frac{n}{2}+1\right)(N+1)}{n+1} - 1 \\ &= \frac{\cancel{(n+1)}(N+1)}{\cancel{n+1}} - 1 \\ &= N \end{aligned}$$

Nous avons donc bien que l'espérance de cet estimateur correspond à la valeur de  $N$  que nous désirons trouver.

**Le cas impair** Dans le cas impair, nous avons :

$$\begin{aligned} E\left(\widehat{N}_1\right) &= 2E\left(\tilde{X}\right) - 1 \\ &= 2E\left(X_{\left(\frac{n+1}{2}\right)}\right) - 1 \\ &= 2\frac{\left(\frac{n+1}{2}\right)(N+1)}{n+1} - 1 \\ &= \frac{\cancel{(n+1)}(N+1)}{\cancel{n+1}} - 1 \\ &= N \end{aligned}$$

Nous avons donc bien que l'espérance de cet estimateur correspond à la valeur de  $N$  que nous désirons trouver.

### 2.3.3 La moyenne

Notre second estimateur est  $\widehat{N}_2 = 2\overline{X} - 1$ , où  $\overline{X}$  représente la moyenne de notre échantillon. Rappelons que pour  $k$  données discrètes, la moyenne se calcule de la façon suivante :

$$\overline{X} = \frac{X_{(1)} + X_{(2)} + \dots + X_{(k-1)} + X_{(k)}}{k} \quad (2)$$

En utilisant les mesures du tableau 1 et l'équation 2, nous obtenons :

$$\begin{aligned} E(\widehat{N}_2) &= 2E(\overline{X}) - 1 \\ &= 2E\left(\frac{X_{(1)} + X_{(2)} + \dots + X_{(n-1)} + X_{(n)}}{n}\right) - 1 \\ &= \end{aligned}$$

Malheureusement, nos deux estimés  $E(\widehat{N}_1)$  et  $E(\widehat{N}_2)$  présentent un problème. Les valeurs de ces deux estimateurs peuvent être plus petites que le plus grand entier dans notre échantillon, c'est-à-dire  $X_{(n)}$ . Il est bien sûr impossible que la valeur  $N$  que nous cherchons soit plus petite que la plus grande valeur de notre échantillon.

Pour vous convaincre, étudions l'échantillon de taille  $n = 3$  suivant, tel que  $X_1 = 2$ ,  $X_2 = 10$  et  $X_3 = 3$ . Dans cette situation, la médiane de l'échantillon est 3 et la moyenne est 5. Nous obtenons donc :

$$\widehat{N}_1 = 2\widetilde{X} - 1 = 5 \quad \text{et} \quad \widehat{N}_2 = 2\overline{X} - 1 = 9$$

Malheureusement, nous savons que  $N$  est supérieur ou égal à 10, le maximum de notre échantillon. Ces deux estimateurs ne sont donc pas adéquats, nous devons en trouver d'autres.

### 2.3.4 Deux autres estimés

Nous voulons maintenant trouver d'autres estimés qui sont toujours supérieurs ou égaux au plus grand entier de notre échantillon. Par symétrie, nous pouvons supposer que le nombre de numéros de série non-observés au-dessus de  $X_{(n)}$  soit le même que le nombre de numéros de série non-observés en-dessous de  $X_{(1)}$ . Nous avons donc :

$$N - X_{(n)} = X_{(1)} - 1$$

$$\widehat{N}_3 = X_{(n)} + X_{(1)} - 1$$

Si nous continuons le raisonnement précédent, il apparaît raisonnable de poser le nombre de numéros de série non-observés au-dessus de  $X_{(n)}$  comme étant la moyenne du :

- nombre de numéros de série non-observés en-dessous de  $X_{(1)}$  ;
- nombre de numéros de série non-observés entre  $X_{(1)}$  et  $X_{(2)}$  ;
- nombre de numéros de série non-observés entre  $X_{(2)}$  et  $X_{(3)}$  ;
- ...
- nombre de numéros de série non-observés entre  $X_{(n-1)}$  et  $X_{(n)}$ .

Nous avons donc que  $N - X_{(n)}$  est égal à :

$$\begin{aligned} N - X_{(n)} &= \frac{1}{n} [(X_{(1)} - 1) + (X_{(2)} - X_{(1)} - 1) + (X_{(3)} - X_{(2)} - 1) + \dots + (X_{(n)} - X_{(n-1)} - 1)] \\ &= \frac{X_{(n)}}{n} - 1 \\ \widehat{N}_4 &= \left( \frac{n+1}{n} \right) X_{(n)} - 1 \end{aligned}$$

## 2.4 La situation où $s$ est connu mais différent de 0 et $N$ est inconnu

Supposons que nous avons une population d'objets numérotés de la façon suivante :  $s+1, s+2, s+3, \dots, s+N$ , où  $N$  est inconnu mais  $s$  est **connu**. Cette situation est en fait un cas particulier d'une loi uniforme discrète de la forme  $U(s+1, s+N)$ .

Nous pouvons résoudre ce problème en utilisant les estimés précédents et en soustrayant la valeur  $s$  aux numéros de série obtenus.

Par exemple, si nous avons une population où les numéros de séries débutent à 500 et nous obtenons l'échantillon suivant :

## [1] 848 523 651 527 605

Pour estimer la valeur de  $N$ , nous soustrayons 499 à notre échantillon et nous utilisons notre estimé  $\widehat{N}_4$ . Nous avons donc :

ech-499

```
## [1] 349 24 152 28 106
```

```
N4(ech-499)
```

```
## [1] 418
```

—

La population est en fait de taille 500.

## 2.5 La situation où $s$ est inconnu et $N$ est inconnu

Supposons que nous avons une population d'objets numérotés de la façon suivante :  $s + 1, s + 2, s + 3, \dots, s + N$ , où  $N$  est inconnu mais  $s$  est **inconnu**.

Nous allons étudier la différence  $d$  entre le plus grand ( $X_{(n)}$ ) et le plus petit ( $X_{(1)}$ ) des numéros de série, que nous notons  $d$ .

$$\begin{aligned} X_{(n)} - X_{(1)} &= E(X_{(n)} - X_{(1)}) \\ &= \frac{n \cdot (N + 1)}{n + 1} - \frac{1 \cdot (N + 1)}{n + 1} \\ &= \frac{(n - 1)(N + 1)}{n + 1} \\ N + 1 &= \frac{(X_{(n)} - X_{(1)})(n + 1)}{n - 1} \\ \widehat{N}_5 &= \frac{(X_{(n)} - X_{(1)})(n + 1)}{n - 1} - 1 \end{aligned}$$

## 3 Quelques simulations

Pour visualiser les différences entre les quatre estimateurs trouvés, nous allons effectuer des simulations avec le logiciel R. Pour nos simulations, nous utiliserons une population de taille  $N = 500$ . Nous pouvons créer cette population dans R de la façon suivante :

```
pop <- c(1:500)
```



Pour modéliser le problème qui nous intéresse, nous voulons piger, **sans remise**, un échantillon de notre population. Pour cette première simulation, nous pigerons un échantillon de taille  $n = 5$ .

```
ech <- sample(pop, 5, replace = FALSE)
ech
```

```
## [1] 349 23 152 28 105
```

Le minimum de notre échantillon est 23 et le maximum est 349. Nous pouvons calculer les quatres estimés associés à l'échantillon précédent :

```
N1(ech)
## [1] 209
N2(ech)
## [1] 262
N3(ech)
## [1] 371
N4(ech)
## [1] 418
```

Pour bien visualiser les différences entre nos quatre estimateurs, nous effectuerons trois simulations distinctes. La figure 1 représente une simulation de 5 000 échantillons de taille 5 à partir d'une population de taille 500.

La figure 2 représente une simulation de 5 000 échantillons de taille 25 à partir d'une population de taille 500.

La figure 3 représente une simulation de 5 000 échantillons de taille 50 à partir d'une population de taille 500.

Nous simulons des populations de tailles 10 à 1 000. Pour chacune d'entre elles, nous choisissons 50 échantillons de taille 5 et nous calculons les quatre estimations.

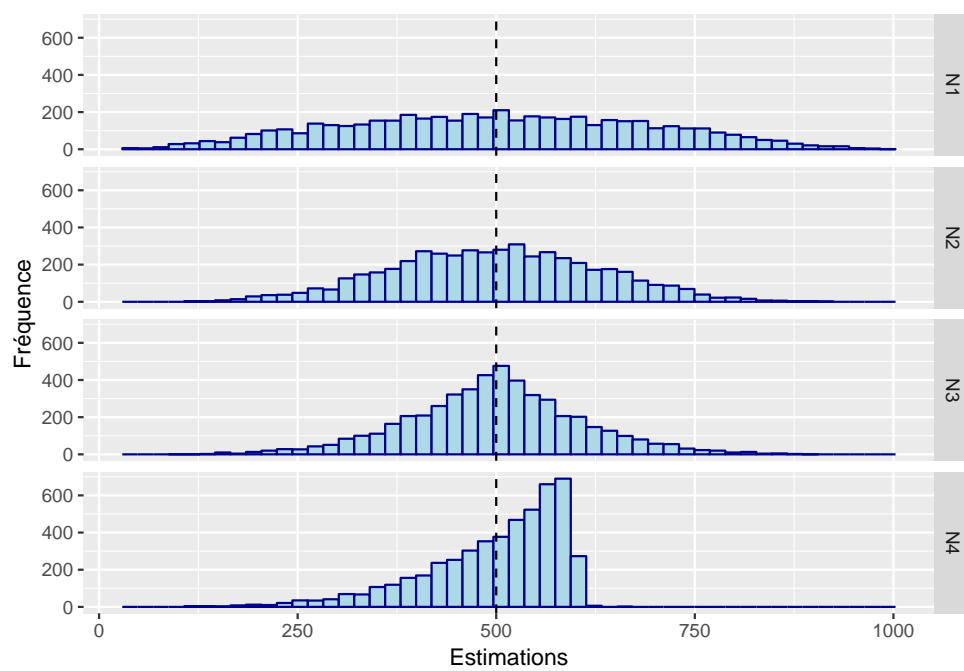


FIGURE 1 – Représentation sous forme d’histogrammes de 5 000 échantillons de taille 5, pour les quatres estimateurs

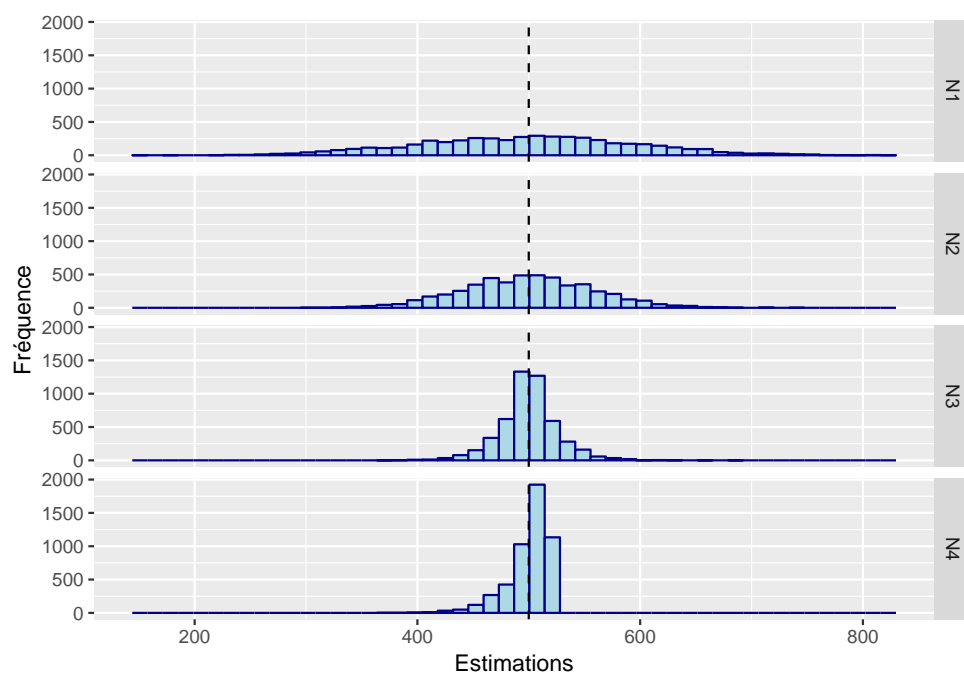


FIGURE 2 – Représentation sous forme d’histogrammes de 5 000 échantillons de taille 25, pour les quatre estimateurs

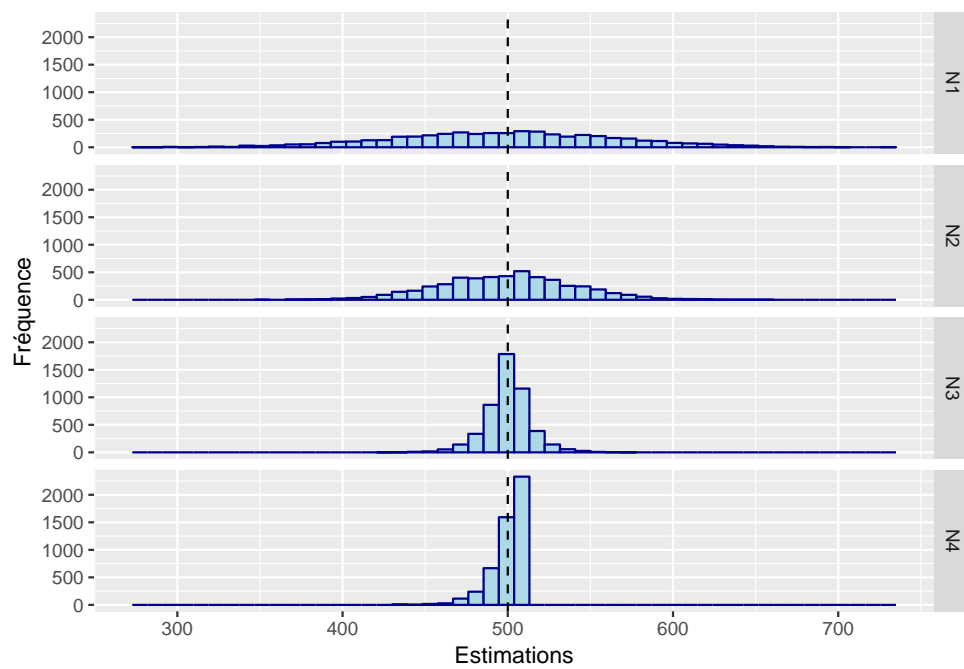


FIGURE 3 – Représentation sous forme d’histogrammes de 5 000 échantillons de taille 50, pour les quatres estimateurs

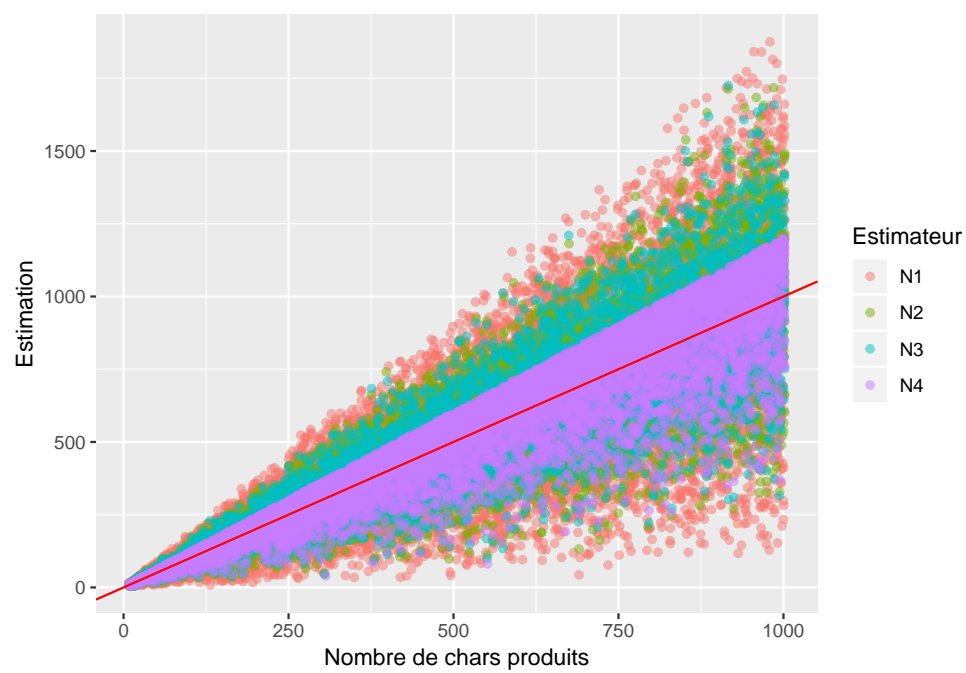


FIGURE 4 – Représentation de 50 échantillons de taille 5 pigés pour des populations de tailles 10 à 1 000, pour les quatre estimateurs

## A Rappels

### Définition 1 (Espérance d'une variable discrète prenant un nombre fini de valeurs)

Soit  $X$  une variable aléatoire discrète avec un nombre fini de valeurs  $x_1, x_2, \dots, x_k$  auxquelles sont associées les probabilités  $p_1, p_2, \dots, p_k$ . L'espérance de  $X$ , notée  $\mathbb{E}(X)$  est définie comme :

$$\mathbb{E}(X) = x_1 p_1 + x_2 p_2 + \dots + x_k p_k = \sum_{i=1}^k x_i p_i$$

Voici quelques propriétés élémentaires concernant l'espérance de variables aléatoires.

- Soit  $k \in \mathbb{R}$ , alors  $\mathbb{E}(k) = k$ .
- Soit  $X_i$  où  $i \in \{1, 2, \dots, n\}$ ,  $n$  variables aléatoires définies sur le même espace probabiliste et  $a_i \in \mathbb{R}$ , alors  $\mathbb{E}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \mathbb{E}(X_i)$ .

**Définition 2 (Variance)** Soit  $X$  une variable aléatoire discrète avec un nombre fini de valeurs  $x_1, x_2, \dots, x_k$  auxquelles sont associées les probabilités  $p_1, p_2, \dots, p_k$ . La variance de  $X$ , notée  $\text{Var}(X)$  est définie comme :

$$\text{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right]$$

**Définition 3 (Covariance)** Soit  $X$  et  $Y$  deux variables aléatoires définies sur le même espace probabiliste et ayant chacune une variance finie. La covariance de  $X$  et  $Y$ , notée  $\text{Cov}(X, Y)$  est définie comme :

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

Voici quelques propriétés élémentaires concernant la variance et la covariance de variables aléatoires.

- Le théorème de König-Huygens donne la formule de variance alternative suivante :  $\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$ .
- Soit  $X$  une variable aléatoire et  $a, b \in \mathbb{R}$ , alors  $\text{Var}(aX + b) = a^2 \text{Var}(X)$ .
- Soit  $X_i$  où  $i \in \{1, 2, \dots, n\}$ ,  $n$  variables aléatoires définies sur le même espace probabiliste et  $a_i \in \mathbb{R}$ , alors  $\text{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} a_i a_j \text{Cov}(X_i, X_j)$ .
- Une généralisation du théorème de König-Huygens pour la variance implique :  $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X) \mathbb{E}(Y)$ .
- $\text{Cov}(X, X) = \text{Var}(X)$
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$

## B Calcul de probabilités

Nous voulons calculer la probabilité reliée à l'événement  $X_{(A)} = i$ , c'est-à-dire l'événement où l'unité statistique  $X_{(A)} = i$ . Pour bien comprendre la situation, nous allons utiliser le schéma ci-dessous :

$$\underbrace{X_{(1)} < X_{(2)} < \dots < X_{(A-1)}}_{A-1 \text{ éléments}} < X_{(A)} < \underbrace{X_{(A+1)} < \dots < X_{(n-1)}}_{n-A \text{ éléments}} < X_{(n)}$$

Nous remarquons que  $A \in \{1, 2, 3, \dots, n\}$  et  $i \in \{A, A+1, A+2, \dots, N-n+A\}$ . En effet, la valeur de  $i$  ne peut pas être plus petite que  $A$  car il y a toujours  $A-1$  valeurs plus petites que  $A$ . De plus,  $A$  ne peut pas être plus grande que  $N-n+A$  car il y a toujours  $N-n+A-1$  valeurs plus grandes que  $A$ .

Pour calculer les probabilités, il faut se rappeler que nous devons choisir  $A-1$  éléments parmi  $i-1$ , ce qui correspond à  $\binom{i-1}{A-1}$ . De plus, nous devons choisir  $n-A$  éléments parmi  $N-i$ , ce qui correspond à  $\binom{N-i}{n-A}$ . Enfin, nous choisissons un échantillon de taille  $n$  parmi une population de taille  $N$ , ce qui correspond à  $\binom{N}{n}$ . Nous avons donc :

$$P(X_{(A)} = i) = \frac{\binom{i-1}{A-1} \binom{N-i}{n-A}}{\binom{N}{n}} \quad \text{où } i = A, A+1, \dots, N-n+A$$

Puisque les probabilités précédentes doivent sommer à un, nous avons :

$$\begin{aligned} \sum_{i=A}^{N-n+A} P(X_{(A)} = i) &= 1 \\ \sum_{i=A}^{N-n+A} \frac{\binom{i-1}{A-1} \binom{N-i}{n-A}}{\binom{N}{n}} &= 1 \\ \sum_{i=A}^{N-n+A} \binom{i-1}{A-1} \binom{N-i}{n-A} &= \binom{N}{n} \end{aligned} \tag{3}$$

À l'aide de l'équation (3), nous pouvons démontrer les résultats du tableau 1. Pour montrer au lecteur la façon de faire, nous présenterons la façon de démontrer l'espérance de  $X_{(A)}$ . Nous avons :

$$E(X_{(A)}) = \sum_{i=A}^{N-n+A} i P(X_{(A)} = i)$$

$$\begin{aligned}
&= \sum_{i=A}^{N-n+A} i \frac{\binom{i-1}{A-1} \binom{N-i}{n-A}}{\binom{N}{n}} \\
&= \frac{1}{\binom{N}{n}} \sum_{i=A}^{N-n+A} i \binom{i-1}{A-1} \binom{N-i}{n-A} \\
&= \frac{1}{\binom{N}{n}} \sum_{i=A}^{N-n+A} A \binom{i}{A} \binom{N-i}{n-A} \quad \left( \text{car } \binom{n}{k} = \frac{n}{k} \binom{n-1}{k-1} \right) \\
&= \frac{A}{\binom{N}{n}} \sum_{i=A}^{N-n+A} \binom{i}{A} \binom{N-i}{n-A} \\
&= \frac{A}{\binom{N}{n}} \binom{N+1}{n+1} \quad (\text{par l'équation 3 et changement d'indice}) \\
&= \frac{A(N+1)}{n+1}
\end{aligned} \tag{4}$$

Les démonstrations de  $\text{Var}(X_{(A)})$  et  $\text{Cov}(X_{(A)}, X_{(B)})$  se font d'une manière similaire.

## C Calculs d'espérance et de variance d'estimateurs

Nous démontrons ici le calcul de l'espérance et de la variance de l'estimateur  $\widehat{N}_1$ , pour donner une idée à la lectrice ou au lecteur des idées à utiliser lors de la démonstration des espérances et des variances des autres estimateurs. Nous étudierons le cas où  $n$  est pair.

### C.1 Calcul de l'espérance

Puisque  $n$  est pair, la médiane est donnée par  $\frac{1}{2} (X_{(n/2)} + X_{(n/2+1)})$  Ainsi :

$$\begin{aligned}
\mathbb{E}(\widehat{N}_1) &= \mathbb{E}(2\widetilde{X} - 1) \\
&= 2 \mathbb{E}(\widetilde{X}) - 1 \\
&= 2 \mathbb{E} \left( \frac{1}{2} (X_{(n/2)} + X_{(n/2+1)}) \right) - 1 \\
&= \mathbb{E}(X_{(n/2)}) + \mathbb{E}(X_{(n/2+1)}) - 1 \\
&= \frac{\binom{n/2}{n+1}}{n+1} + \frac{\binom{n/2+1}{n+1}}{n+1} - 1 \\
&= \frac{(n+1)(N+1)}{n+1} - 1
\end{aligned}$$



$$= N$$

## C.2 Calcul de la variance

Puisque  $n$  est pair, la médiane est donnée par  $\frac{1}{2} (X_{(n/2)} + X_{(n/2+1)})$  Ainsi :

$$\begin{aligned} \text{Var}(\widehat{N}_1) &= \text{Var}(2\tilde{X} - 1) \\ &= 4 \text{Var}(\tilde{X}) \\ &= 4 \text{Var}\left(\frac{1}{2} (X_{(n/2)} + X_{(n/2+1)})\right) \\ &= \text{Var}(X_{(n/2)} + X_{(n/2+1)}) \\ &= \text{Var}(X_{(n/2)}) + \text{Var}(X_{(n/2+1)}) + 2 \text{Cov}(X_{(n/2)}, X_{(n/2+1)}) \\ &= \frac{(n/2)(n+1-n/2)(N+1)(N-n)}{(n+1)^2(n+2)} + \frac{(n/2+1)(n+1-(n/2+1))(N+1)(N-n)}{(n+1)^2(n+2)} + \\ &\quad \dots + 2 \frac{(n/2)(n+1-(n/2+1))(N+1)(N-n)}{(n+1)^2(n+2)} \\ &= (n^2 + n) \frac{(N+1)(N-n)}{(n+1)^2(n+2)} \\ &= \frac{n}{n+1} \frac{(N+1)(N-n)}{(n+1)(n+2)} \end{aligned}$$

## Références

- [1] Roger W. Johnson (1994). Estimating the Size of a Population, *Teaching Statistics*, 16, (no. 2), pages 50-52.
- [2] Leo A. Goodman (1952). Serial Number Analysis, *Journal of the American Statistical Association*, 47, (no. 260), pages 622-634.
- [3] Leo A. Goodman (1954). Some Practical Techniques in Serial Number Analysis, *Journal of the American Statistical Association*, 49, (no. 265), pages 97-112.
- [4] Harry V. Roberts (1957). Informative Stopping Rules and Inferences about Population Size, *Journal of the American Statistical Association*, 62, (no. 319), pages 763-775.
- [5] Arthur G. Volz (2008). A Soviet Estimate of German Tank Production, *The Journal of Slavic Military Studies*, 21, (no. 3), pages 588-590.

- [6] Richard Ruggles and Henry Brodie (1947). An Empirical Approach to Economic Intelligence in World War II, *Journal of the American Statistical Association*, 42, (no. 237), pages 72-91.
- [7] Pagetable.com (4 février 2011). *How many Commodore 64 were really sold ?*. Récupéré le 17 octobre 2018 : <https://web.archive.org/web/20160306232450/http://www.pagetable.com/?p=547>
- [8] Charles Artur (8 octobre 2008). *Why iPhones are just like German tanks..*. Récupéré le 17 octobre 2018 : <https://www.theguardian.com/technology/blog/2008/oct/08/iphone.apple>