

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
<http://madesautels.rbind.io/>

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

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 utilisées 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]. En particulier, nous présentons au tableau 1 les estimations du nombre de chars d'assaut allemands en comparaison avec l'estimation par les services de renseignements et les archives allemandes, tel que présenté dans [6]. Comme vous pouvez le constater, les estimations statistiques étaient beaucoup plus précises que les estimations provenant des services de renseignements.

Mois	Estimation statistique	Estimation par les services de renseignements	Selon les archives allemandes
Juin 1940	169	1 000	122
Juin 1941	244	1 550	271
Août 1942	327	1 550	342

TABLEAU 1 – Comparaison des estimations produites par les estimateurs que nous présenterons, des estimations produites par les services de renseignements ainsi que de la production exacte selon les archives allemandes.

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 colon Dupuy indique que puisque chaque char 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 [8]. 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 [9].

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, ..., 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 aux annexes A et B quelques rappels sur les notions d'espérance, de variance et de covariance ainsi que la démonstration de $\mathbb{E}(X_{(A)})$. Les démonstrations de $\text{Var}(X_{(A)})$ et $\text{Cov}(X_{(A)}, X_{(B)})$ se font d'une manière similaire. Nous utiliserons donc, sans démonstration, les trois mesures du tableau 2.

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

À l'aide des trois mesures du tableau 2, 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 débuter en supposant que nous connaissons la valeur milieu m de la liste $1, 2, \dots, N$. Nous nous retrouvons dans la situation ci-dessous :

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

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 semble raisonnable de la remplacer par une estimation, par exemple la médiane \tilde{X} ou la moyenne \bar{X} . Nous pouvons maintenant obtenir nos deux premiers estimateurs.

À partir de maintenant, nous allons toujours supposer que nous arrondissons nos estimateurs à l'entier le plus près.

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.

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

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. En effectuant les calculs appropriés, nous obtenons les résultats du tableau 3.

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

TABLEAU 3 – L'espérance et la variance de l'estimateur \widehat{N}_1 .

2.3.3 La moyenne

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

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

En effectuant les calculs appropriés pour l'espérance et la variance, nous obtenons les résultats du tableau 4.

i	\widehat{N}_i	$\mathbb{E}(\widehat{N}_i)$	$\text{Var}(\widehat{N}_i)$
2	$2\bar{X} - 1$	N	$\frac{n+2}{3n} \cdot \frac{(N-n)(N+1)}{n+2}$

TABLEAU 4 – L'espérance et la variance de l'estimateur \widehat{N}_2 .

Malheureusement, nos deux estimés \widehat{N}_1 et \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 $n = 5$ suivant, tel que $X_1 = 2$, $X_2 = 10$, $X_3 = 3$, $X_4 = 1$ et $X_5 = 4$. Dans cette situation, la médiane de l'échantillon est 3 et la moyenne est 5. Nous obtenons donc :

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

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 devrons 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 :

$$\begin{aligned} N - X_{(n)} &= X_{(1)} - 1 \\ \widehat{N}_3 &= X_{(n)} + X_{(1)} - 1 \end{aligned}$$

En effectuant les calculs appropriés pour l'espérance et la variance, nous obtenons les résultats du tableau 5.

i	\widehat{N}_i	$\mathbb{E}(\widehat{N}_i)$	$\text{Var}(\widehat{N}_i)$
3	$X_{(n)} + X_{(1)} - 1$	N	$\frac{2}{n+1} \cdot \frac{(N-n)(N+1)}{n+2}$

TABLEAU 5 – L'espérance et la variance de l'estimateur \widehat{N}_3 .

Nous pouvons continuer le raisonnement précédent. En effet, 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 le nombre de numéros de série non-observés au-dessus de $X_{(n)}$, c'est-à-dire $N - X_{(n)}$ est égal à :

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

$$\widehat{N}_4 = \left(\frac{n+1}{n} \right) X_{(n)} - 1$$

En effectuant les calculs appropriés pour l'espérance et la variance, nous obtenons les résultats du tableau 6.

i	\widehat{N}_i	$\mathbb{E}(\widehat{N}_i)$	$\text{Var}(\widehat{N}_i)$
4	$\left(\frac{n+1}{n} \right) X_{(n)} - 1$	N	$\frac{1}{n} \cdot \frac{(N-n)(N+1)}{n+2}$

TABLEAU 6 – L'espérance et la variance de l'estimateur \widehat{N}_4 .

Bien que la démonstration des résultats présentés aux tableaux 3, 4, 5 et 6 ne soit pas présentée, nous pouvons remarquer les résultats suivants :

- Dans le cas où $n = 1$, la variance des quatre estimateurs est la même.
- Dans le cas où la taille de notre échantillon est la même que celle de notre population, c'est-à-dire $n = N$, les variances de nos quatre estimateurs sont nulles.
- Les variances de nos estimateurs diminuent lorsque i augmente, ce qui indique que \widehat{N}_4 est l'estimateur ayant la variance la plus petite.
- Il est possible d'utiliser le théorème de Lehmann-Scheffé pour démontrer que \widehat{N}_4 est l'estimateur non-biaisé ayant la plus petite variance.

La lectrice ou le lecteur intéressé pourra utiliser les idées présentées à l'annexe C pour démontrer les résultats des tableaux 3, 4, 5 et 6.

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.

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 lui aussi **inconnu**. Cette situation était celle des numéros de séries sur les boîtes de transmissions des chars d'assaut *Panther V*, comme présenté en [6].

Pour pouvoir continuer, nous allons utiliser les espérances des valeurs $X_{(A)}$ de notre échantillon, c'est-à-dire $\mathbb{E}(X_{(A)}) = \frac{A(N+1)}{n+1}$, que nous avons trouvé à l'annexe B.

Nous allons étudier l'espérance de la différence d entre le plus grand ($X_{(n)}$) et le plus petit ($X_{(1)}$) des numéros de série de notre échantillon. Ce résultat reste vrai même dans la situation où s est inconnu. Nous avons donc :

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

Il est donc possible d'estimer la taille de la population, même si s et N sont inconnus. En effectuant les calculs appropriés pour l'espérance et la variance, nous obtenons les résultats du tableau 7.

i	\widehat{N}_i	$\mathbb{E}(\widehat{N}_i)$	$\text{Var}(\widehat{N}_i)$
5	$\frac{(X_{(n)} - X_{(1)})(n+1)}{n-1} - 1$	N	$\frac{2}{n-1} \cdot \frac{(N-n)(N+1)}{n+2}$

TABLEAU 7 – L'espérance et la variance de l'estimateur \widehat{N}_5 .

Remarquons que les variances des tableaux 3, 4, 5, 6 et 7 ont toutes un même facteur commun, c'est-à-dire $\frac{(N-n)(N+1)}{n+2}$. Nous pouvons donc visualiser l'effet de la taille de notre échantillon sur nos variances en étudiant seulement le facteur multiplicatif devant le terme $\frac{(N-n)(N+1)}{n+2}$. Le tableau 8 indique les facteurs multiplicatifs de nos cinq estimateurs.

i	Facteur multiplicatif \widehat{N}_i
1	$\frac{1}{n}$ pour n impair $\frac{n+1}{n}$ pour n pair
2	$\frac{n+2}{3n}$
3	$\frac{2}{n+1}$
4	$\frac{1}{n}$
5	$\frac{n}{n-1}$

TABLEAU 8 – Les facteurs multiplicatifs des cinq estimateurs

La figure 1 représente le facteur multiplicatif pour les variances des cinq estimateurs, en fonction de la taille de l'échantillon.

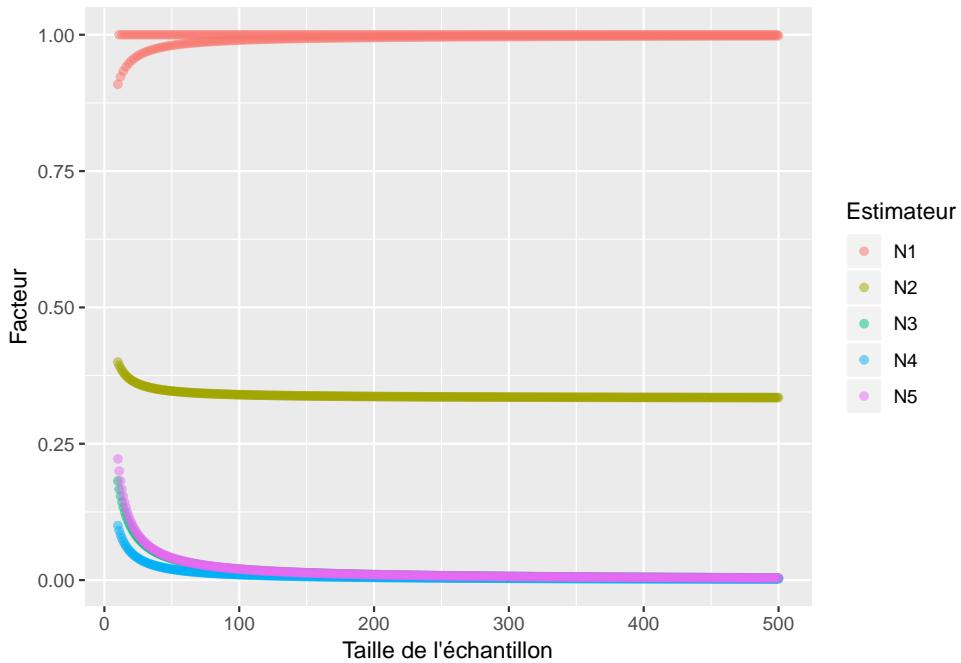


FIGURE 1 – Représentation du facteur multiplicatif pour les variances des cinq estimateurs, en fonction de la taille de l'échantillon.

Nous remarquons que les facteurs multiplicatifs des estimateurs \widehat{N}_3 , \widehat{N}_4 et \widehat{N}_5 s'approchent de 0 sensiblement à la même vitesse alors que celui pour \widehat{N}_1 s'approche de 1 et celui de \widehat{N}_2 s'approche de $1/3$. Nous pouvons donc supposer que les estimateurs \widehat{N}_3 , \widehat{N}_4 et \widehat{N}_5 sont meilleurs que les deux premiers.

3 Quelques simulations

Pour visualiser les différences entre les cinq estimateurs trouvés, nous allons effectuer des simulations avec le logiciel R. Les commandes R utilisées seront précédées du symbole `>`, qui représente l'invite de commandes (*prompt* en anglais). Ce symbole indique que l'interface de R est prête à recevoir une commande.

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 (d'où l'ajout de l'option `replace=FALSE`), 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 cinq estimateurs associés à l'échantillon précédent :

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

```

## [1] 418
> N5(ech)
## [1] 488

```

À partir d'ici, pour ne pas alourdir le texte, nous omettrons le code R utilisé pour produire les simulations qui suivent. Vous trouverez par contre à l'annexe E le code complet utilisé dans cet article.

Pour bien visualiser les différences entre nos cinq estimateurs, nous effectuerons quatre simulations distinctes. La figure 2 représente une simulation de 5 000 échantillons de tailles 5, 10, 25 et 50, pigés à partir d'une population de taille 500. Nous calculons les mesures de nos cinq estimateurs pour ces 5 000 échantillons.

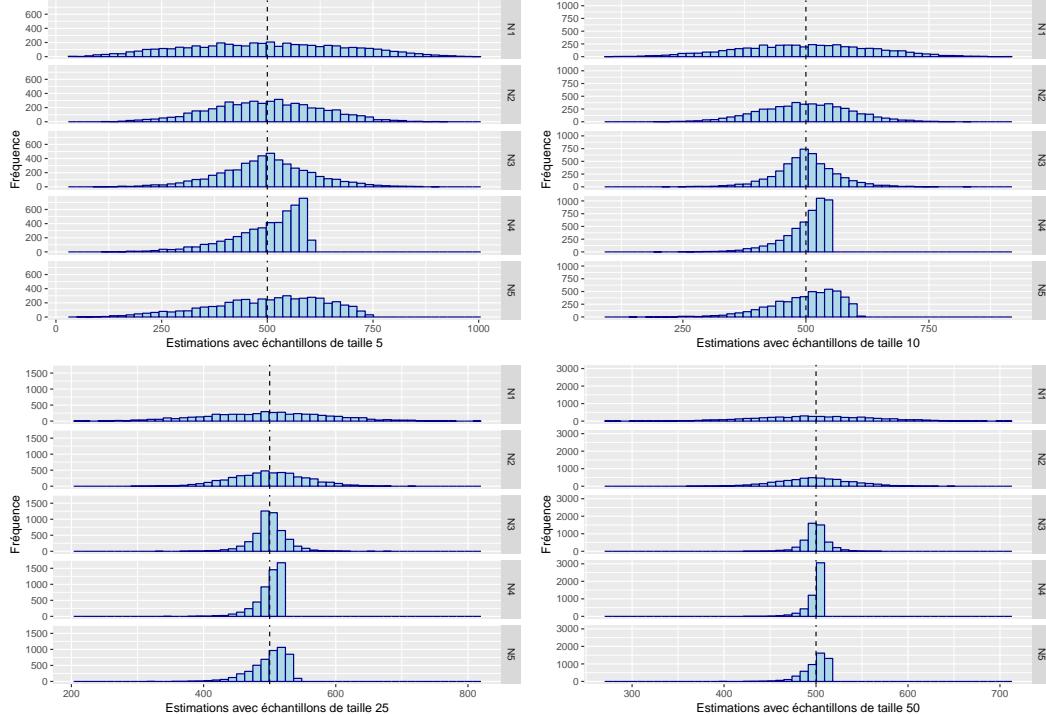


FIGURE 2 – Représentation sous forme d'histogrammes de 5 000 échantillons de taille 5 (en haut à gauche), de taille 10 (en haut à droite), de taille 25 (en bas à gauche) et de taille 50 (en bas à droite) pour les cinq estimateurs. La ligne pointillée correspond à la taille de la population, c'est-à-dire 500.

Pour représenter différemment les résultats de nos cinq estimateurs, nous ferons un autre type de simulation. Nous allons faire varier la taille de la population de 10 à 1000, par bonds de 5. Pour chacunes de ces populations, nous choisissons un échantillon de taille 5 et nous calculons la valeur des cinq estimateurs. Pour être en mesure de bien visualiser les résultats, nous allons ajouter une petite variation aléatoire à la position des points. De cette façon, nous verrons mieux la dispersion de nos données. La figure 3 représente nos résultats obtenus à partir de nos échantillons pour les cinq estimateurs. La droite en rouge représente la taille véritable de la population. Les idées utilisées pour obtenir les graphiques de la figure 3 sont basées sur [10].

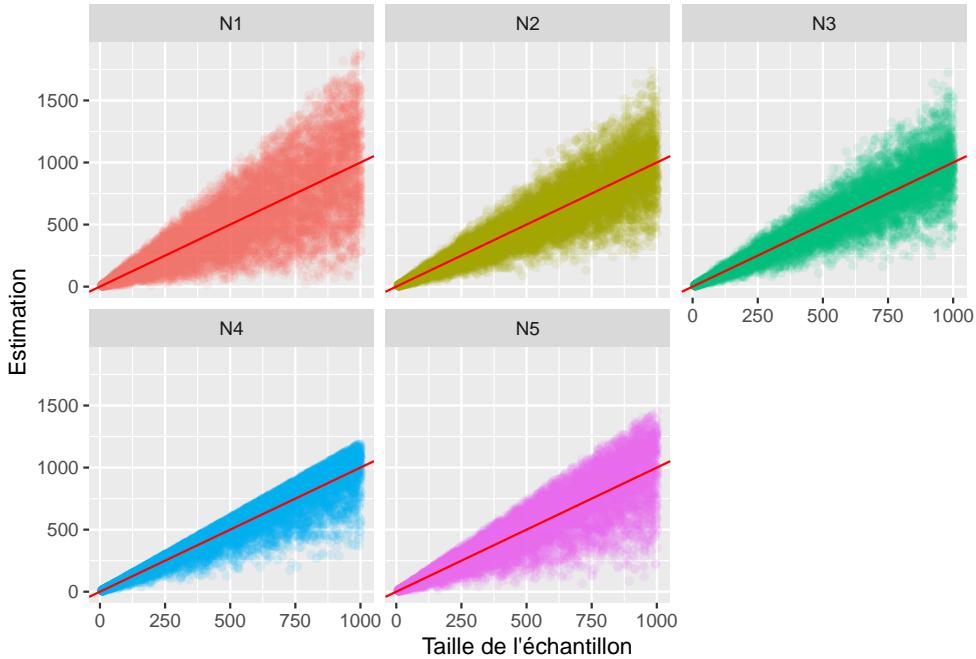


FIGURE 3 – Représentation de 50 échantillons de taille 5 pigés pour des populations de tailles 10 à 1 000, pour les cinq estimateurs. La droite en rouge représente la taille véritable de la population.

Nous allons maintenant partir d'une population de taille 500. Nous ferons varier la taille des échantillons de 1 à 500, par bonds de 1. Pour chacune de ces tailles, nous pigerons 50 échantillons sans remise. Nous représenterons ensuite l'estimation obtenu en fonction de la taille des échantillons. Pour être en mesure de bien visualiser les résultats, nous allons ajouter une petite variation aléatoire à la position des points. La figure 4 représente nos résultats obtenus.

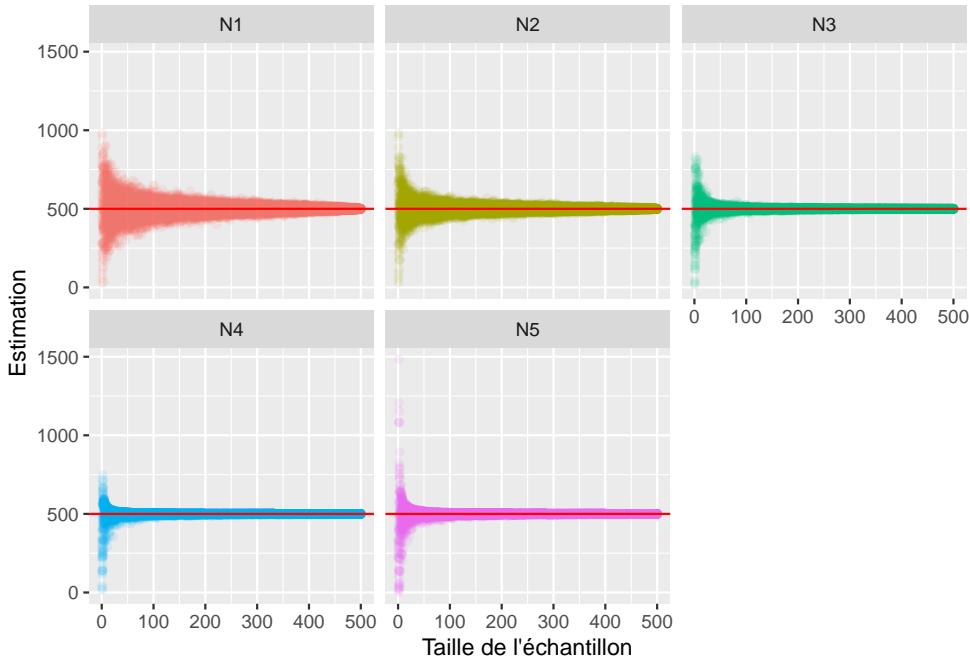


FIGURE 4 – Représentation de 50 échantillons de taille 1 à 500 pigés pour une population de taille 500, pour les cinq estimateurs. La droite en rouge représente la taille véritable de la population.

4 Une application

Nous trouvons en [3] une application concrète de l'utilisation de nos estimateurs. On nous indique que le département des sciences sociales de l'Université de Chicago utilisaient de l'équipement (bureaux, chaises, etc.) sur lesquels étaient placés des numéros de série.

Les numéros de série de 31 pièces d'équipement ont été observés. Les 31 numéros de série étaient :

```
## [1] 83 135 274 380 668 895 955 964 1113 1174 1210 1344 1387 1414
## [15] 1610 1668 1689 1756 1865 1874 1880 1936 2005 2006 2065 2157 2220 2224
## [29] 2396 2543 2787
```

Nous invitons les lectrices et les lecteurs intéressés à utiliser les cinq estimateurs présentés dans cet article pour estimer le nombre d'équipements achetés par l'Université de Chicago. Le

minimum de l'échantillon est 83 et le maximum de l'échantillon est 2787. Les réponses des cinq estimateurs ainsi que la véritable valeur du nombre d'équipements seront donnés à l'annexe D.

5 Conclusion

Dans cet article, nous avons étudié le problème de l'estimation de la taille d'une population à partir d'échantillons prélevés sans remise de cette population. Nous avons étudié le traitement statistique de ce problème d'un point de vue fréquentiste. Nous avons également introduit cinq estimateurs différents pouvant être utilisés dans un environnement pédagogique.

Pour trouver un autre exemple d'utilisation des notions présentées dans cet article, je vous invite à consulter le site suivant <http://datagenetics.com/blog/march22014/index.html>, qui discute du problème d'estimer le nombre de participants à une course en utilisant les numéros des coureurs rencontrés.

Il est également possible d'étudier ce problème d'un point de vue bayésien. Ce point de vue est partiellement utilisé par Roberts en [4]. Davantage d'informations est par contre disponible en [7] et en [11]. Il est intéressant de voir les différences entre les estimations trouvées dans cet article et les estimations trouvées dans un contexte bayésien.

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 :

$$\mathbb{E}(X) = x_1p_1 + x_2p_2 + \dots + x_kp_k = \sum_{i=1}^k x_i p_i$$

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 :

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

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 :

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

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 :

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

À l'aide de l'équation (1), nous pouvons démontrer les résultats du tableau 2. Pour montrer au lecteur la façon de faire, nous présenterons la façon de démontrer l'espérance de $X_{(A)}$. Nous avons :

$$\begin{aligned} \mathbb{E}(X_{(A)}) &= \sum_{i=A}^{N-n+A} i P(X_{(A)} = i) \\ &= \sum_{i=A}^{N-n+A} i \frac{\binom{i-1}{A-1} \binom{N-i}{n-A}}{\binom{N}{n}} \\ &= \frac{1}{\binom{N}{n}} \sum_{i=A}^{N-n+A} i \binom{i-1}{A-1} \binom{N-i}{n-A} \\ &= \frac{1}{\binom{N}{n}} \sum_{i=A}^{N-n+A} A \binom{i}{A} \binom{N-i}{n-A} \quad \left(\text{car } \binom{n}{k} = \frac{n}{k} \binom{n-1}{k-1} \right) \\ &= \frac{A}{\binom{N}{n}} \sum_{i=A}^{N-n+A} \binom{i}{A} \binom{N-i}{n-A} \\ &= \frac{A}{\binom{N}{n}} \binom{N+1}{n+1} \quad (\text{par l'équation 1 et changement d'indice}) \\ &= \frac{A(N+1)}{n+1} \end{aligned} \tag{2}$$

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 :

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

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 :

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

$$\begin{aligned}
&= (n^2 + n) \frac{(N+1)(N-n)}{(n+1)^2(n+2)} \\
&= \frac{n}{n+1} \frac{(N+1)(N-n)}{(n+2)}
\end{aligned}$$

D Réponses de l'exercice de la section 4

Les valeurs des cinq estimateurs de l'application de la section 4 sont :

$$\begin{aligned}
\widehat{N}_1 &= 3335 \\
\widehat{N}_2 &= 3010 \\
\widehat{N}_3 &= 2869 \\
\widehat{N}_4 &= 2876 \\
\widehat{N}_5 &= 2883
\end{aligned}$$

Les équipements ayant été achetés entre 1928 et 1934, l'auteur de [3] indique qu'il lui a été particulièrement difficile d'obtenir la véritable valeur du nombre d'équipements achetés. Après plusieurs jours et de nombreuses questions aux personnes concernées, les archives pertinentes ont été trouvées et le nombre N de pièces d'équipements était de 2885.

E Code R

Pour la lectrice ou le lecteur intéressé, vous trouverez ci-dessous le code R complet qui a servi à produire les diverses figures de cet article. Vous trouverez également le code complet de cet article à l'adresse https://github.com/desautm/serial_number_amq.

```

knitr:::opts_chunk$set(cache = TRUE)

## Pour obtenir les mêmes résultats malgré l'échantillonnage aléatoire
set.seed(39894095)

## La librairie utilisée
## Pour l'installer, décommenter la ligne suivante:
## install.packages("tidyverse")
library(tidyverse)

```

```

## Définition des fonctions pour le calcul des mesures sur les numéros de série.

## Estimateur N1
N1 <- function(ech){
  return(round(2*median(ech)-1))
}

## Estimateur N2
N2 <- function(ech){
  return(round(2*mean(ech)-1))
}

## Estimateur N3
N3 <- function(ech){
  if (length(ech) == 1) return(ech)
  else return(round(max(ech)+min(ech)-1))
}

## Estimateur N4
N4 <- function(ech){
  if (length(ech) == 1) return(ech)
  else{
    n <- length(unique(ech))
    return(round((n+1)/n*max(ech)-1))
  }
}

## Estimateur N5
N5 <- function(ech){
  if (length(ech) == 1) return(ech)
  else{
    n <- length(unique(ech))
    return(round((max(ech)-min(ech))*(n+1)/(n-1)-1))
  }
}

n <- 500

```

```

var_médiane <- function(x){
  if (x %% 2 == 0) {return(x/(x+1))}
  else {return(1)}
}

data <- tibble(
  id = 10:n,
  N1 = map_dbl(id, ~var_médiane(.x)),
  N2 = map_dbl(id, ~(.x+2)/(3*.x)),
  N3 = map_dbl(id, ~2/(.x+1)),
  N4 = map_dbl(id, ~1/.x),
  N5 = map_dbl(id, ~2/(.x-1))
)

data <- data %>%
  gather(key = Estimateur, value = Facteur, -id)

ggplot(data, aes(x=id, y=Facteur, color=Estimateur))+ 
  geom_point(alpha=0.5) +
  labs(
    x = "Taille de l'échantillon"
  )
pop <- c(1:500)
ech <- sample(pop, 5, replace = FALSE)
ech
N1(ech)
N2(ech)
N3(ech)
N4(ech)
N5(ech)
N <- 500
n <- 5
iter <- 5000
pop <- c(1:N)

ech5 <- tibble(id = map(1:iter, ~sample(pop, n, replace = FALSE)))

```

```

ech5 <- ech5 %>%
  mutate(N1 = map_dbl(id, N1),
         N2 = map_dbl(id, N2),
         N3 = map_dbl(id, N3),
         N4 = map_dbl(id, N4),
         N5 = map_dbl(id, N5)) %>%
gather(estimateur, valeur, N1, N2, N3, N4, N5)

n <- 10
ech10 <- tibble(id = map(1:iter, ~sample(pop, n, replace = FALSE)))
ech10 <- ech10 %>%
  mutate(N1 = map_dbl(id, N1),
         N2 = map_dbl(id, N2),
         N3 = map_dbl(id, N3),
         N4 = map_dbl(id, N4),
         N5 = map_dbl(id, N5)) %>%
gather(estimateur, valeur, N1, N2, N3, N4, N5)

n <- 25
ech25 <- tibble(id = map(1:iter, ~sample(pop, n, replace = FALSE)))
ech25 <- ech25 %>%
  mutate(N1 = map_dbl(id, N1),
         N2 = map_dbl(id, N2),
         N3 = map_dbl(id, N3),
         N4 = map_dbl(id, N4),
         N5 = map_dbl(id, N5)) %>%
gather(estimateur, valeur, N1, N2, N3, N4, N5)

n <- 50
ech50 <- tibble(id = map(1:iter, ~sample(pop, n, replace = FALSE)))
ech50 <- ech50 %>%
  mutate(N1 = map_dbl(id, N1),
         N2 = map_dbl(id, N2),
         N3 = map_dbl(id, N3),
         N4 = map_dbl(id, N4),
         N5 = map_dbl(id, N5)) %>%
gather(estimateur, valeur, N1, N2, N3, N4, N5)

```

```

graphe_ech5 <- ggplot(data = ech5, mapping = aes(x = valeur))+  

  geom_histogram(bins = 50, fill = "lightblue", color = "darkblue") +  

  facet_grid(estimateur ~ .) +  

  geom_vline(xintercept=N,linetype=2,color="black") +  

  labs(  

    x = "Estimations avec échantillons de taille 5",  

    y = "Fréquence"  

  )  
  

graphe_ech10 <- ggplot(data = ech10, mapping = aes(x = valeur)) +  

  geom_histogram(bins = 50, fill = "lightblue", color = "darkblue") +  

  facet_grid(estimateur ~ .) +  

  geom_vline(xintercept=N,linetype=2,color="black") +  

  labs(  

    x = "Estimations avec échantillons de taille 10",  

    y = "Fréquence"  

  )  
  

graphe_ech25 <- ggplot(data = ech25, mapping = aes(x = valeur)) +  

  geom_histogram(bins = 50, fill = "lightblue", color = "darkblue") +  

  facet_grid(estimateur ~ .) +  

  geom_vline(xintercept=N,linetype=2,color="black") +  

  labs(  

    x = "Estimations avec échantillons de taille 25",  

    y = "Fréquence"  

  )  
  

graphe_ech50 <- ggplot(data = ech50, mapping = aes(x = valeur)) +  

  geom_histogram(bins = 50, fill = "lightblue", color = "darkblue") +  

  facet_grid(estimateur ~ .) +  

  geom_vline(xintercept=N,linetype=2,color="black") +  

  labs(  

    x = "Estimations avec échantillons de taille 50",  

    y = "Fréquence"  

  )  

graphe_ech5  

graphe_ech10

```

```

graphe_ech25
graphe_ech50
ech1000 <- tibble(
  pop = rep(seq(10, 1000, 5), 50),
  id = map(pop, ~sample(.x, 5, replace = FALSE))
)
ech1000 <- ech1000 %>%
  mutate(N1 = map_dbl(id, N1),
         N2 = map_dbl(id, N2),
         N3 = map_dbl(id, N3),
         N4 = map_dbl(id, N4),
         N5 = map_dbl(id, N5)) %>%
  gather(estimateur, valeur, N1, N2, N3, N4, N5)

ggplot(data = ech1000, aes(x = pop, y = valeur, color=estimateur))+ 
  geom_jitter(alpha=0.1)+ 
  labs(
    x = "Taille de l'échantillon",
    y = "Estimation"
  )+ 
  geom_abline(intercept = 0, slope=1, color = "red") + 
  facet_wrap(estimateur ~ .) + 
  theme(legend.position="none")

pop100 <- tibble(
  ech = rep(seq(1, 500, 1), 25),
  id = map(ech, ~sample(1:500, .x, replace = FALSE))
)
pop100 <- pop100 %>%
  mutate(N1 = map_dbl(id, N1),
         N2 = map_dbl(id, N2),
         N3 = map_dbl(id, N3),
         N4 = map_dbl(id, N4),
         N5 = map_dbl(id, N5)) %>%
  gather(estimateur, valeur, N1, N2, N3, N4, N5)

ggplot(data = pop100, aes(x = ech, y = valeur, color=estimateur))+ 
  geom_jitter(alpha=0.1)+
```

```

  labs(
    x = "Taille de l'échantillon",
    y = "Estimation"
  ) +
  geom_abline(intercept = 500, slope=0, color = "red") +
  facet_wrap(estimateur ~ .) +
  theme(legend.position="none")
equip = c(83, 135, 274, 380, 668, 895, 955, 964, 1113,
        1174, 1210, 1344, 1387, 1414, 1610, 1668, 1689,
        1756, 1865, 1874, 1880, 1936, 2005, 2006, 2065,
        2157, 2220, 2224, 2396, 2543, 2787)
data <- tibble(
  id = 1:length(equip),
  equip = equip
)
equip

```

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 (1967). 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] Michael Höhle and Leonhard Held (2006). Bayesian estimation of the size of a population, *Technical report, SFB 386*, (no. 399).

- [8] 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>
- [9] 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>
- [10] Risto Hinno (14 février 2015). *German tank problem..*. Récupéré le 19 octobre 2018 : <http://bit.ly/2P7U9dE>
- [11] *German tank problem..*. Récupéré le 19 octobre 2018 : <https://en.wikipedia.org/wiki/GermanTankProblem>