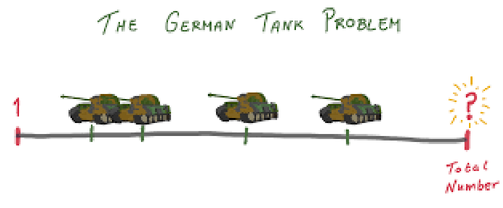
ECE 2 - Année 2021-2022
ENC Bessières
Mathématiques - F. Gaunard
http://frederic.gaunard.com



Devoir Maison n°7

German Tank Problem
À rendre le 14/02

Afin de motiver un peu le travail sur la notion d'estimateur, on propose ici un exemple historique.

Le contexte. Été 1943, les Alliés essaient de percer le bloc de l'Axe en créant un nouveau front via l'Italie. Ils rencontrent un nouveau type de char allemand, le bien nommé *Sonderkraftfahrzeug 171* plus connu des afficionados de machines de combat sous le nom de *Panther*.
Ce dernier est mieux équipé et plus performant que ceux rencontrés jusqu'alors. Il a été conçu en réponse à l'excellent *T-34* utilisé par les soviétiques sur le front de l'Est.
Sans rentrer dans les détails d'armement de cette subtile et sympathique machine, il peut percer les défenses et détruire la majorité des tank alliés.

Néanmoins, malgré sa puissance théorique, celui-ci ne peut avoir un réel impact sur l'issue de la guerre que si le nombre d'unités produites est suffisant. Il apparait crucial pour les Alliés de déterminer ou plutôt d'*estimer* combien de *Panther* étaient produits. La tâche fut confiée à [la] *Economic Warfare Division of the American Embassy in London*[1].

La modélisation. On suppose que l'ennemi produit une série de chars immatriculés par des entiers en commençant par 1. En plus de cela, quelle que soit la date de production du char, ses années de service, ou encore son numéro de série, la distribution des numéros d'immatriculation est considérée comme étant uniforme dès l'instant où on mène l'analyse.

Dans notre modélisation, les allemands disposent de N tanks numérotés de 1 à N. Les force alliées observent aléatoirement, uniformément et "avec remise" n numéros de séries $(X_1, ..., X_n)$ et cherchent à estimer le paramètre N.

[1] Comme le raconte l'article *An Empirical Approach to Economic Intelligence in World War II*, R. Ruggles & H. Brodie, Journal of the American Statistical Association **42**-237 (1947), 72-91

On considère dans tout le problème un $n-$échantillon $(X_1, ..., X_n)$ de la loi $\mathcal{U}([\![1, N]\!])$ et on commence par poser

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

(1) Déterminer $E(\overline{X}_n)$ et en déduire un estimateur T_n de N sans biais en fonction de $\overline{X}_n$.
(2) Calculer le risque quadratique de T_n et montrer que cet estimateur est convergent.

(3) **Application numérique &** `SciLab`.

(a) Écrire une fonction `SciLab` d'en-tête `function y=T(X)` qui, à partir d'un échantillon X renvoie l'estimation ponctuelle de N obtenue avec T_n.
(b) On observe

```
X=[8, 322, 15, 135, 69]
```

Que dire de l'estimation obtenue avec T_5? Commenter.
(c) On peut quand quand même imaginer que les renseignements à disposition sont un peu plus fournis que les numéros de série de 5 tanks.
Écrire une fonction `function y=T_moy(N,n)` qui renvoie le résultat moyen de T_n pour 1000 estimations ponctuelles obtenues sur un échantillon de taille n de la loi $\mathcal{U}([\![1; N]\!])$.
Que se passe-t-il pour $N = 1000$ avec $n = 20$? Avec $n = 100$?

On introduit alors le nouvel estimateur

$$M_n = \max(X_1, ..., X_n)$$

(4) Calculer, pour tout $k \in \mathbb{N}$, $P(M_n \leq k)$.
(5) Soit Y un v.a à valeurs dans $[\![1; N]\!]$. Montrer que

$$E(Y) = \sum_{k=0}^{N-1} P(Y > k).$$

(6) Montrer alors que

$$E(M_n) = N - \sum_{k=0}^{N-1} \left(\frac{k}{N}\right)^n.$$

(7) Vérifier que, pour tout $k \in [\![0; N-1]\!]$,

$$0 \leq \left(\frac{k}{N}\right)^n \leq N \int_{k/N}^{(k+1)/N} t^n \mathrm{d}t.$$

(8) En déduire que

$$N - \frac{N}{n+1} \leq E(M_n) \leq N$$

puis que l'estimateur M_n est asymptotiquement sans biais.

Si l'estimateur M_n parait naturel, il a clairement un défaut; il sous-estime nécessairement N (puisqu'il renverra toujours une valeurs inférieure (ou égale) à N). On va donc essayer d'y apporter une légère *correction*.

Commençons par introduire le numéro du plus petit tank observé $m_n = \min(X_1, ..., X_n)$.

Comme N est inconnu, on ne connait pas l'écart entre N et M_n, mais il parait raisonnable de penser qu'il y a (en moyenne) autant de tanks *non observés* entre M_n et N qu'entre 1 et m_n. Entre le plus

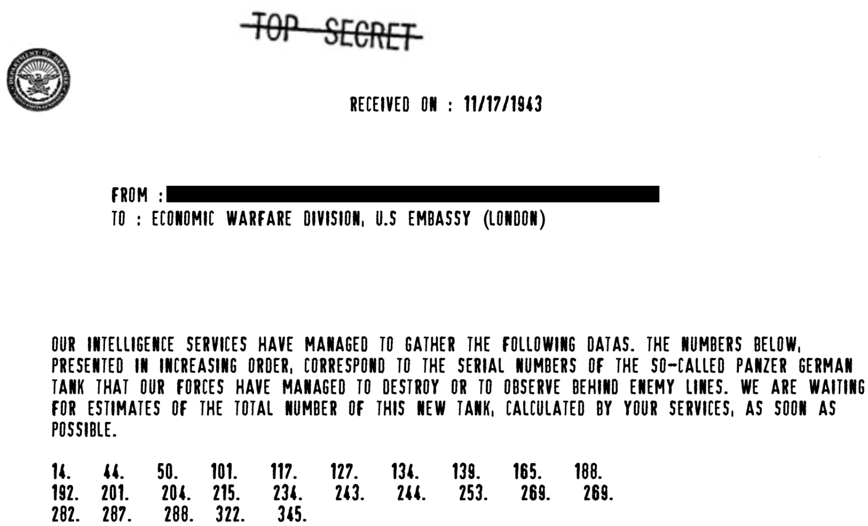
petit numéro observé et le tank avec le numéro de série 1, il y a $m_n - 1$ numéros de tank.
On pense alors à ajouter la correction

$$\tilde{M}_n = M_n + (m_n - 1).$$

(9) En s'inspirant des calculs précédents pour M_n, déterminer $E(m_n)$ sous forme d'une somme qu'on ne cherchera pas à simplifier.

(10) Montrer que $\tilde{M}_n$ maintenant un estimateur sans biais de N.

(11) Utiliser les données du faux document ci-dessous pour fournir une estimation de N à l'aide de $\tilde{M}_n$



~~TOP SECRET~~

RECEIVED ON : 11/17/1943

FROM : [redacted]

TO : ECONOMIC WARFARE DIVISION, U.S EMBASSY (LONDON)

OUR INTELLIGENCE SERVICES HAVE MANAGED TO GATHER THE FOLLOWING DATAS. THE NUMBERS BELOW, PRESENTED IN INCREASING ORDER, CORRESPOND TO THE SERIAL NUMBERS OF THE SO-CALLED PANZER GERMAN TANK THAT OUR FORCES HAVE MANAGED TO DESTROY OR TO OBSERVE BEHIND ENEMY LINES. WE ARE WAITING FOR ESTIMATES OF THE TOTAL NUMBER OF THIS NEW TANK, CALCULATED BY YOUR SERVICES, AS SOON AS POSSIBLE.

14.	44.	50.	101.	117.	127.	134.	139.	165.	188.
192.	201.	204.	215.	234.	243.	244.	253.	269.	269.
282.	287.	288.	322.	345.					

(12) **Comparaison**. On propose la fonction `SciLab` ci-dessous. Que fait-elle?

```
function [x,y,z]=mystere(N, n)
    T=zeros(1, 1000)
    M=zeros(1, 1000)
    Mtilde=zeros(1, 1000)
    for k=1:1000
        X=grand(1, n, "uin", 1, N)
        T(k)=2*mean(X)-1
        M(k)=max(X)
        Mtilde(k)=max(X)+(min(X)-1)
    end
    x=mean(T)
    y=mean(M)
    z=mean(Mtilde)
endfunction
```

On ajoute les instructions ci-dessous dont l'exécution permet d'obtenir la figure ci-contre. Interpréter. Quel estimateur semble le plus performant?

```
X=[]
Y=[]
Z=[]
N=1000
for n=10:10:200
    [x,y,z]=mystere(N,n)
    X=[X, x]
    Y=[Y, y]
    Z=[Z, z]
end

A=10:10:200
plot2d(A, X, -4)
plot2d(A, Y, -1)
plot2d(A, Z, -2)
plot2d([0,200], [N, N], 'red')
```