MAT^2

MATerials MATemàtics
Volum 2017, treball no. 3, 31 pp. ISSN: 1887-1097
Publicació electrònica de divulgació del Departament de Matemàtiques
de la Universitat Autònoma de Barcelona
www.mat.uab.cat/matmat

El problema dels tancs alemanys

Xavier Bardina, Sílvia Prior, Carla Rodríguez i Ferran Rosado

La Segona Guerra Mundial forma part de la nostra història recent, però el que no és tan conegut és que darrere de tots aquells fets històrics l'estadística també va jugar-hi un paper important.

La intel·ligència econòmica dels aliats havia de proporcionar dades sobre la indústria i la producció del material de guerra de l'enemic, en aquest cas l'exèrcit de Hitler. La informació que aportava aquest departament es tenia molt en compte a l'hora de planejar l'estratègia dels aliats a Europa. En concret, la informació sobre la indústria i les plantes de producció alemanyes eren dades essencials per dissenyar el programa de bombardejos estratègics sobre el continent europeu. Aquest problema d'estimació de la producció militar alemanya en el món anglosaxó, i en el món de l'estadística en general, és conegut com el problema dels tancs alemanys.

Per resoldre'l, el departament d'intel·ligència de les forces Aliades va utilitzar d'entrada les tècniques habituals d'espionatge com són descodificar missatges encriptats, interrogatoris,... que donaven unes estimacions molt elevades i molt allunyades de la producció real. Així doncs van decidir que era el moment de buscar alternatives que donessin unes xifres més reals per poder treballar i preparar les seves estratègies militars. En aquell moment és quan va intervenir l'enginy dels estadístics amb l'ajuda inconscient dels alemanys.

Els alemanys eren molt meticulosos a l'hora d'etiquetar i marcar tots els components dels seus equips, cadascun d'ells portava les inscripcions gravades o estava etiquetat mitjançant plaques identificatives. La informació

d'aquestes etiquetes variava segons el component del tanc ja que ho etiquetaven absolutament tot: rodes, canó, xassís, palanca de canvis,... però generalment, la informació més rellevant que contenien aquestes etiquetes era el nom i la localització del component dins l'equip, la data de fabricació, el número de sèrie, el motlle que s'havia utilitzat per elaborar-lo, on s'havia produït, etc. A més a més, no només eren rigorosos amb el marcatge dels equips, també eren extremadament disciplinats amb els historials, manuals tècnics i tota la documentació de manteniment en general. Aquest rigor els permetia tenir un bon control de qualitat i de la gestió dels recanvis, però també eren una font d'informació molt valuosa pels aliats.

Així doncs, a principis de l'any 1943 la Divisió d'Economia de Guerra de l'ambaixada dels Estats Units a Londres va començar a analitzar els números de sèrie, les etiquetes i les marques de diferents components dels equips capturats als nazis. Inicialment els seus estudis es van centrar en els pneumàtics, dels quals van trobar molta informació, i posteriorment van ampliar el camp de treball a l'anàlisi de qualsevol component que trobessin marcat i etiquetat dels tancs de batalla, canons, camions i bombes V-1 i V-2.

Van fer un estudi exhaustiu de tots els components que capturaven a les batalles i també de tota la documentació trobada al nord d'Àfrica que incloïa llibres de registre que contenien els números de sèrie dels xassís dels tancs amb els corresponents codis de l'assemblador i la data de manufacturació. Amb tota aquesta informació analitzada es van adonar que cada una de les classes de carruatges de la Wehrmacht muntava un tipus de caixa de canvis que havia estat numerada de forma correlativa. Així doncs, existia una relació entre cada sèrie de caixa de canvis i el tipus de carruatge. Per tant, si es pogués determinar la producció d'una sèrie completa de caixes de canvi, s'hauria obtingut la producció dels tancs Panzer associats. Aleshores, amb uns quants tancs capturats als alemanys i amb l'ajuda dels estadístics es podria estimar la seva producció de carruatges de combat.

Aquest mètode va ser aplicat pels aliats en tots els equips i components que disposaven de número de sèrie. De fet, en les estimacions d'alguns models de Panzer van fer servir els números dels xassís dels motors enlloc dels de les caixes de canvi i, fins i tot, en el cas dels Panzer V es van fer servir els números de sèrie de les rodes per verificar les estimacions.

Després de la guerra, es van conèixer les dades reals i es va veure que les estimacions matemàtiques havien estat molt acurades mentre que els resultats de la intel·ligència distaven molt de la realitat. La taula que ve a continuació (que reprodueix la que apareix a [7]) reflecteix en números la precisió de les estimacions dutes a terme pels estadístics i permet comprovar la gran diferència amb les que es van realitzar en un primer moment per la intel·ligència dels aliats.

Data de producció	Estimació per número de sèrie	Estimació dels serveis d'intel·ligència	Dades reals dels serveis del Ministeri d'Armament
Juny de 1940	169	1000	122
Juny de 1941	244	1550	271
Agost de 1942	327	1550	342
Febrer de 1944	270	1400	276

Comparativa de les estimacions fetes pels estadístics, la intel·ligència i el registre alemany real (veure [7]).

En llenguatge estadístic ens trobem davant el problema d'estimar la mida d'una població numerada a partir de mostres sense reemplaçament. Això és:

> Suposem que tenim una població d'objectes numerada $1, 2, \ldots, N$, la mida de la qual, N, desconeixem. Escollim una mostra aleatòria simple sense reposició $X_1, X_2, \ldots, X_n$ d'aquesta població. És a dir, n'escollim n a l'atzar sense que es puguin repetir elements i l'objectiu serà estimar la mida de la població N a partir de la mostra.

Noteu que aquest és un tipus de problema que podem trobar en situacions aparentment molt diverses: comptar quants corredors hi ha en una cursa, el nombre de taxis d'una ciutat, el nombre d'iPhones fabricats el 2014, quantes parades hi ha en un mercat ambulant a partir dels números de llicència,... Qualsevol població numerada és susceptible d'aquest tipus d'estimació.

Com ja ha sortit al text, en la teoria estadística de les estimacions, el problema d'estimar el màxim d'una distribució uniforme discreta d'una mostra sense reemplaçament es coneix com el *Problema dels tancs alemanys*.

En aquest article intentarem donar resposta al problema que fa tants anys va portar de cap als aliats presentant primerament els possibles estimadors d'una població numerada, donant les característiques de cadascun d'ells i comparant-los entre si d'una manera objectiva. Mirarem les seves característiques com a estimadors: biaix, variància,... També abordarem el problema des de la vessant de l'estadística bayesiana. Això ens donarà un punt de vista diferent al que apareix utilitzant l'estadística freqüentista i es podrà veure si arribem a les mateixes conclusions o no.

La bibliografia del final permet aprofundir, des de punts de vista diferents, aquest tema. En l'article [6] trobem una explicació detallada del tipus de tancs i del tipus de dades que van fer servir els aliats. En [3] es presenten

els quatre estimadors freqüentistes que farem servir i les seves propietats. També recomanem el treball [7] de la Revista General de la Marina que fa una presentació divertida del tema i, a un nivell matemàtic més elevat, l'article [5]. Finalment també trobarem dos articles de divulgació científica sobre aquest problema, un de la revista Investigación y Ciencia ([4]) i l'altre del diari The Guardian ([2]).

1 Els estimadors

Passant a les matemàtiques, ens plantegem el problema següent: *volem estimar la mida d'una població numerada* $1, 2, \ldots, N$ *de la qual només coneixem uns quants elements amb la seva numeració* $X_1, X_2, \ldots, X_n$*, és a dir, tenim una mostra aleatòria simple sense reposició de mida* n. Comencem trobant estimadors de la mida de la mostra utilitzant només el sentit comú.

1.1 Mitjana i mediana

Suposem que coneixem el valor mitjà m de la llista $1, 2, \ldots, N$. Aleshores és clar que hi haurà $m-1$ valors per sota d'aquest valor i $m-1$ per sobre, és a dir que

$$N = (m-1) + 1 + (m-1) = 2\,m - 1\,.$$

Ara bé, d'entrada no coneixem el valor de m però és natural substituir-lo per un estimador del valor mitjà com poden ser la mediana o la mitjana.

Així obtindrem els nostres primers estimadors.

Donada una mostra $X_1, X_2, \ldots, X_n$, si denotem per

$$\widetilde{X} = \text{Mediana}(X_1, X_2, \ldots, X_n),$$

tenim el nostre primer estimador:

$$\boxed{\widehat{N}_1 = 2\,\widetilde{X} - 1\,.}$$

De la mateixa manera si utilitzem la mitjana de la mostra $\overline{X} = \frac{X_1 + \cdots + X_n}{n}$. Obtenim un altre estimador diferent:

$$\boxed{\widehat{N}_2 = 2\,\overline{X} - 1\,.}$$

Observem que aquest coincideix amb el que es troba a partir de l'anomenat *mètode dels moments*, que en aquest cas s'aplicaria de la forma següent:

sabem que X segueix una distribució Uniforme en $\{1,\ldots,N\}$. Per tant, $E(X)=\frac{N+1}{2}$. Llavors, igualant l'esperança a la mitjana, tenim que

$$\overline{X}=\frac{\widehat{N}+1}{2}\Rightarrow \widehat{N}=2\,\overline{X}-1.$$

I arribem, doncs, al mateix estimador que ja havíem obtingut.

Veiem però, que aquests dos estimadors presenten un inconvenient greu ja que és clar que un bon estimador de la mida de la població hauria de ser més gran o igual al màxim de la mostra, és a dir, $N\geq \max(X_1,X_2,\ldots,X_n)$. I això no es compleix sempre amb aquests dos estimadors, com es pot veure amb un exemple senzill on, en els dos casos, les estimacions seran clarament falses:

> *Suposem que tenim una mostra d'una població on* $X_1=4$, $X_2=10$ *i* $X_3=1$. *Si calculem les estimacions pels dos mètodes obtenim*
>
> $$\widehat{N}_1=2\cdot 4-1=7,$$
> $$\widehat{N}_2=2\cdot 5-1=9.$$
>
> *Quan és clar que* $N\geq X_2=10$ *!*

Per tant, hem de buscar nous estimadors que no presentin aquest problema.

1.2 Estadístics d'ordre

Passem a treballar amb els estadístics d'ordre $X_{(1)},X_{(2)},\ldots,X_{(n)}$, és a dir, ordenem la nostra mostra de menor a major:

$$1\leq X_{(1)}<X_{(2)}<\cdots<X_{(n)}\leq N.$$

Com que $X\sim \text{Unif}(1,\ldots,N)$, una primera consideració pot ser suposar que la distància que hi ha entre l'1 i el primer valor $X_{(1)}$ sigui la mateixa que entre N i el valor més gran observat $X_{(n)}$.

$$\underbrace{1\leq X_{(1)}}<X_{(2)}<\cdots<\underbrace{X_{(n)}\leq N}.$$

Obtenim així que:

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

a partir del qual obtenim un tercer estimador:

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

Una altra manera d'enfocar-ho és estimar aquesta diferència per la mitjana de les diferències entre observacions, és a dir, considerant la diferència entre l'1 i $X_{(1)}$, juntament amb la de diferència entre $X_{(1)}$ i $X_{(2)}$, $X_{(2)}$ i $X_{(3)}$, ..., $X_{(n-1)}$ i $X_{(n)}$ i fent la mitjana de totes elles.

$$\overbrace{1 \leq \underbrace{X_{(1)}}} < \overbrace{\underbrace{X_{(2)}} < X_{(3)}} < \cdots < \underbrace{X_{(n-1)} < X_{(n)}} \leq N.$$

Observem que

$$1 \leq X_{(1)} < X_{(1)}+1 \leq X_{(2)} < X_{(2)}+1 \leq \cdots \leq X_{(n-1)} < X_{(n-1)}+1 \leq X_{(n)} \leq N.$$

Així, igual que abans, tenim:

$$\widehat{N} - X_{(n)} = \frac{(X_{(1)}-1)+(X_{(2)}-X_{(1)}-1)+(X_{(3)}-X_{(2)}-1)+\cdots+(X_{(n)}-X_{(n-1)}-1)}{n}.$$

D'on obtenim

$$\begin{aligned}\widehat{N} &= X_{(n)} + \frac{(X_{(1)}-1)+(X_{(2)}-X_{(1)}-1)+(X_{(3)}-X_{(2)}-1)+\cdots+(X_{(n)}-X_{(n-1)}-1)}{n} \\ &= X_{(n)} + \frac{X_{(n)} - n}{n} = \frac{n+1}{n}\,X_{(n)} - 1\,.\end{aligned}$$

Obtenim així un quart estimador per a la mida de la població:

$$\boxed{\widehat{N}_4 = \frac{n+1}{n}\,X_{(n)} - 1\,.}$$

Aquests dos estimadors nous donaran sempre, per construcció, valors més grans o iguals que el valor més gran de les observacions.

Un exemple concret

Veiem tot seguit un exemple numèric. Suposem que s'han capturat 7 tancs a l'enemic i que els números de sèrie són: 131, 91, 19, 149, 100, 130 i 15. Llavors es té:

$$\begin{aligned}\widetilde{X} &= 100 \\ \overline{X} &= 90.71 \\ X_{(1)} &= 15 \\ X_{(7)} &= 149\end{aligned}$$

Per tant les estimacions del nombre de tancs N que s'obtindrien amb els nostres estimadors serien:

$$\widehat{N}_1 = 199$$
$$\widehat{N}_2 = 180.42$$
$$\widehat{N}_3 = 163$$
$$\widehat{N}_4 = 169.29$$

D'on, segons l'estimador utilitzat, el nombre de tancs prevists seria 199, 180, 163 o 169 .

2 El Biaix

En estadística, s'anomena biaix d'un estimador a la diferència entre la seva esperança matemàtica i el valor del paràmetre que estima. Per tant, no tenir biaix és una propietat desitjable dels estimadors. Un estimador amb un biaix nul es diu que és no esbiaixat

Per tal de comparar els quatre estimadors començarem calculant-ne el biaix.

2.1 Biaix $\widehat{N}_2$

Per veure si $\widehat{N}_2$ té biaix n'hem de calcular l'esperança. Recordem $\widehat{N}_2 = 2 \cdot \overline{X} - 1$, on $X_1, X_2, \ldots, X_n$ segueixen una distribució Unif$(1, \ldots, N)$. Per tant, $E(X_i) = \frac{N+1}{2}$ i també, com que l'esperança és lineal, $E(\overline{X}) = \frac{N+1}{2}$.

$$E[\widehat{N}_2] = 2\,E(\overline{X}) - 1 = 2\,\frac{N+1}{2} - 1 = N.$$

Per tant,

$\widehat{N}_2$ és un estimador no esbiaixat.

2.2 Eines per a calcular les esperances

Per calcular la mediana necessitem tenir la mostra ordenada i, per tant, treballar amb els estadístics d'ordre. Per calcular els estimadors $\widehat{N}_3$ i $\widehat{N}_4$ també necessitarem el valor més petit i el valor més gran obtinguts en la mostra.

Per això començarem calculant la llei i l'esperança dels estadístics d'ordre $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ en general.

Observació 2.1. *En una mostra de mida n obtinguda a partir d'una població de N elements numerats, la funció de probabilitat de l'estadístic d'ordre $X_{(j)}$, per a $j = 1, \ldots, n$, ve donada per*

$$P(X_{(j)} = k) = \frac{\binom{k-1}{j-1}\binom{N-k}{n-j}}{\binom{N}{n}} \qquad k = j, j+1, \ldots, N-n+j.$$

En efecte, hi ha $\binom{N}{n}$ possibles mostres diferents de les quals, si fixem que k ha de ser el $j-$èssim un cop ordenats, n'hem d'escollir $j-1$ entre els primers $k-1$ (ho podem fer de $\binom{k-1}{j-1}$ formes diferents) i $n-j$ entre els $N-k$ més grans (ho podem fer de $\binom{N-k}{n-j}$ formes diferents).

Com que es tracta d'una funció de probabilitats, la suma de les probabilitats de tots els valors possibles ha de ser igual a 1.

$$\sum_{k=j}^{N-n+j} \frac{\binom{k-1}{j-1}\binom{N-k}{n-j}}{\binom{N}{n}} = 1.$$

D'on obtenim el lema següent.

Lema 2.2.

$$\sum_{k=j}^{N-n+j} \binom{k-1}{j-1}\binom{N-k}{n-j} = \binom{N}{n}.$$

Aquesta fòrmula servirà per demostrar el resultat del lema següent, que és l'ingredient principal per poder fer els càlculs del biaix de la resta d'estimadors que hem proposat.

Lema 2.3. *En una mostra sense reposició de mida n obtinguda a partir d'una població de N elements numerats, l'esperança de l'estadístic d'ordre $X_{(j)}$, per a $j = 1, \ldots, n$, ve donada per*

$$E[X_{(j)}] = \frac{j}{n+1}(N+1).$$

Demostració. Utilitzant que

$$\binom{k}{j} = \frac{k!}{j!\,(k-j)!} = \frac{k \cdot (k-1)!}{j \cdot (j-1)!\,(k-j)!} = \frac{k}{j}\binom{k-1}{j-1},$$

tenim que

$$\begin{aligned} E[X_{(j)}] &= \sum_{k=j}^{N-n+j} k\, P(X_{(j)}=k) \\ &= \frac{1}{\binom{N}{n}} \sum_{k=j}^{N-n+j} k \binom{k-1}{j-1}\binom{N-k}{n-j} \\ &= \frac{j}{\binom{N}{n}} \sum_{k=j}^{N-n+j} \binom{k}{j}\binom{N-k}{n-j}. \end{aligned} \tag{1}$$

Ara, fent els canvis de variable $j = j' - 1$, $k = k' - 1$, $N = N' - 1$ i $n = n' - 1$ tenim que

$$\begin{aligned} \sum_{k=j}^{N-n+j} \binom{k}{j}\binom{N-k}{n-j} &= \sum_{k'=j'}^{N'-n'+j'} \binom{k'-1}{j'-1}\binom{N'-k'}{n'-j'} \\ &= \binom{N'}{n'}, \end{aligned}$$

on en el darrer pas hem utilitzat el Lema 2.2. Si ara desfem el canvi de variable tenim que

$$\binom{N'}{n'} = \binom{N+1}{n+1} = \frac{N+1}{n+1}\binom{N}{n}.$$

Llavors, tornant a la fórmula (1) tenim

$$E[X_{(j)}] = \frac{j}{\binom{N}{n}} \frac{N+1}{n+1}\binom{N}{n} = \frac{j}{n+1}(N+1),$$

tal com volíem demostrar. □

2.3 Biaix $\widehat{N}_1$

A l'hora de calcular el biaix de $\widehat{N}_1$ hem de distingir dos casos, n parell i n senar.

- Si n és senar, $n = 2k+1$, per a un cert $k \in \mathbb{N}$, i aleshores la mediana és $\widetilde{X} = X_{(k+1)}$ i pel Lema 2.3

$$E[\widetilde{X}] = E[X_{(k+1)}] = \frac{k+1}{2\,k+2}(N+1) = \frac{N+1}{2}.$$

- Si n és parell, $n = 2\,k$, per a un cert $k \in \mathbb{N}$, i aleshores la mediana és $\widetilde{X} = \frac{X_{(k)}+X_{(k+1)}}{2}$ i pel Lema 2.3

$$\begin{aligned} E[\widetilde{X}] &= \frac{1}{2}\,E[X_{(k)} + X_{(k+1)}] \\ &= \frac{1}{2}\left(\frac{k}{2\,k+1}\,(N+1) + \frac{k+1}{2\,k+1}\,(N+1)\right) \\ &= \frac{1}{2}\,\frac{2\,k+1}{2\,k+1}\,(N+1) = \frac{N+1}{2}. \end{aligned}$$

Per tant, en els dos casos,

$$E[\widehat{N}_1] = E[2\widetilde{X} - 1] = N.$$

Així doncs

$\widehat{N}_1$ és un estimador no esbiaixat.

2.4 Biaix $\widehat{N}_3$

Calculem el biaix de $\widehat{N}_3$. Recordem que $\widehat{N}_3 = X_{(n)} + X_{(1)} - 1$. Per tant, pel Lema 2.3

$$\begin{aligned} E[\widehat{N}_3] &= E[X_{(n)}] + E[X_{(1)}] - 1 \\ &= \frac{n}{n+1}\,(N+1) + \frac{N+1}{n+1} - 1 = N. \end{aligned}$$

Així doncs

$\widehat{N}_3$ és un estimador no esbiaixat.

2.5 Biaix $\widehat{N}_4$

Calculem el biaix de $\widehat{N}_4 = \frac{n+1}{n}\,X_{(n)} - 1$, novament pel Lema 2.3

$$\begin{aligned} E[\widehat{N}_4] &= E[\frac{n+1}{n}\,X_{(n)} - 1] = \frac{n+1}{n}\,E[X_{(n)}] - 1 \\ &= \frac{n+1}{n}\,\frac{n(N+1)}{n+1} - 1 \\ &= N. \end{aligned}$$

Per tant

$\widehat{N}_4$ és un estimador no esbiaixat.

Com que $\widehat{N}_1, \widehat{N}_2, \widehat{N}_3, \widehat{N}_4$ són estimadors no esbiaixats, per saber quin és millor necessitarem calcular les variàncies. L'estimador millor serà el que tingui menys variància, és a dir, el que tingui menys dispersió. Com que tots els estimadors són no esbiaixats prenen valors al voltant del valor central que és el paràmetre N que es vol estimar. Si hi ha menys dispersió tindrem més probabilitat d'estar a prop de N.

3 Variàncies dels estimadors

El resultat que s'enuncia al lema següent permetrà calcular les variàncies dels estimadors. La seva demostració es basa en aplicar repetidament els arguments que ja hem utilitzat per al càlcul de les esperances i la trobareu a l'apèndix.

Lema 3.1. *Considerem una mostra sense reposició de mida n obtinguda a partir d'una població de N elements numerats. Sigui $a, b \in \{1, ..., n\}$. Aleshores,*

a) $\mathrm{Var}(X_{(a)}) = a\,\dfrac{N+1}{n+1}\,\dfrac{(N-n)(n-a+1)}{(n+1)(n+2)}$,

b) $\mathrm{Var}(X_{(a)} + X_{(b)}) = \dfrac{(N+1)(N-n)}{(n+1)^2(n+2)}\,((3(a\wedge b)+(a\vee b))(n+1)-(a+b)^2)$,

c) $\mathrm{Var}(X_1 + ... + X_n) = \dfrac{(N+1)(N-n)\,n}{12}$,

on $\wedge$ i $\vee$ designen respectivament el més petit i el més gran dels dos nombres.

3.1 Variància de $\hat{N}_1$

Cal diferenciar entre el cas en que n és senar del que n és parell.

- Si n és senar aleshores, $n = 2\,k+1$ i $\widetilde{X} = X_{(k+1)}$.

$$\begin{aligned}\mathrm{Var}(\hat{N}_1) &= \mathrm{Var}(2\,X_{(k+1)} - 1) = 4\ \mathrm{Var}(X_{(k+1)}) \\ &= 4\,(k+1)\,\frac{N+1}{n+1}\,\frac{(N-n)(n-k-1+1)}{(n+1)(n+2)} \\ &= \frac{(N+1)(N-n)}{n+2}.\end{aligned}$$

- Si n és parell aleshores, $n = 2\,k$ i $\widetilde{X} = \frac{X_{(k)}+X_{(k+1)}}{2}$

$$\begin{aligned}\mathrm{Var}(\hat{N}_1) &= \mathrm{Var}(2\,\frac{X_{(k)}+X_{(k+1)}}{2} - 1) = \mathrm{Var}(X_{(k)}+X_{(k+1)})\\ &= \frac{(N+1)(N-n)}{(n+1)^2(n+2)}\,((4\,k+1)(n+1)-(2\,k+1)^2)\\ &= \frac{(N+1)(N-n)\,n}{(n+1)(n+2)}.\end{aligned}$$

3.2 Variància de $\widehat{N}_2$

Recordem que $\widehat{N}_2 = 2\,\overline{X} - 1$. El valor de la variància s'obtindrà amb els càlculs següents

$$\begin{aligned}\mathrm{Var}(\widehat{N}_2) &= \mathrm{Var}(2\,\overline{X}-1) = 4\ \mathrm{Var}(\overline{X}) = 4\ \mathrm{Var}(\frac{X_{(1)}+\cdots+X_{(n)}}{n})\\ &= \frac{4}{n^2}\,\frac{(N+1)(N-n)\,n}{12} = \frac{(N+1)(N-n)}{3\,n}.\end{aligned}$$

3.3 Variància de $\widehat{N}_3$

Si l'estimador és $\widehat{N}_3 = X_{(n)} + X_{(1)} - 1$ la variància s'obtindrà a partir de:

$$\begin{aligned}Var(\widehat{N}_3) &= \mathrm{Var}(X_{(n)}+X_{(1)}-1) = \mathrm{Var}(X_{(n)}+X_{(1)})\\ &= \frac{(N+1)(N-n)}{(n+1)^2(n+2)}((3\,\cdot 1+n)(n+1)-(n+1)^2)\\ &= 2\,\frac{(N+1)(N-n)}{(n+1)(n+2)}.\end{aligned}$$

3.4 Variància de $\widehat{N}_4$

Finalment, quan l'estimador és $\widehat{N}_4 = \dfrac{n+1}{n}\,X_{(n)} - 1$ els càlculs són:

$$\begin{aligned}\mathrm{Var}(\widehat{N}_4) &= \mathrm{Var}(\frac{n+1}{n}X_{(n)}-1) = \frac{(n+1)^2}{n^2}\,\mathrm{Var}(X_{(n)})\\ &= \frac{(n+1)^2}{n^2}\,n\,\frac{N+1}{n+1}\,\frac{(N-n)(n-n+1)}{(n+1)(n+2)}\\ &= \frac{(N+1)(N-n)}{n(n+2)}.\end{aligned}$$

3.5 Taula resum

Per poder visualitzar amb facilitat tots els resultats obtinguts els mostrem tots junts en una taula:

Taula resum

Estimadors de N	$E(\widehat{N}_i)$	$\mathrm{Var}(\widehat{N}_i)$
$\widehat{N}_1 = 2\widetilde{X} - 1$	N	$\frac{(N-n)(N+1)}{(n+2)}$ (n senar)
	N	$\frac{n}{(n+1)}\,\frac{(N-n)(N+1)}{(n+2)}$ (n parell)
$\widehat{N}_2 = 2\overline{X} - 1$	N	$\frac{(n+2)}{(3\,n)}\,\frac{(N-n)(N+1)}{(n+2)}$
$\widehat{N}_3 = X_{(n)} + X_{(1)} - 1$	N	$\frac{2}{(n+1)}\,\frac{(N-n)(N+1)}{(n+2)}$
$\widehat{N}_4 = \frac{n+1}{n}X_{(n)} - 1$	N	$\frac{1}{n}\,\frac{(N-n)(N+1)}{(n+2)}$

Observem que en l'expressió de les variàncies el terme $\frac{(N-n)(N+1)}{(n+2)}$ és comú en totes elles.

Per acabar de fer l'estudi d'aquests estimadors ens queda per veure quina variància és més petita.

3.6 Comparació de variàncies

Per tal d'establir quin és el millor estimador hem de veure quin és el que té la variància més petita i això és el que farem en aquest apartat.

A la taula hem pogut observar clarament que hi ha una part de la variància, $\frac{(N-n)(N+1)}{(n+2)}$, que és comú a tots els estimadors de N que hem estudiat. A l'hora de comparar-les només haurem de tenir en compte la resta de l'expressió.

Les desigualtats que es poden obtenir de forma ràpida i que permeten decidir l'estimador de variància mínima són:

- $\mathrm{Var}(\widehat{N}_1) \geq \mathrm{Var}(\widehat{N}_2)$. Com que l'estimador $\widehat{N}_1$ fa servir la mediana hem de distigir dos casos segons si n és parell o senar.

  – Si n és parell, hem de veure:

  $$\frac{n}{n+1} \geq \frac{n+2}{3\,n} \Longleftrightarrow (n-2)(n+\frac{1}{2}) \geq 0.$$

  I aquesta desigualtat és certa per a $n \geq 2$. En el cas $n = 2$, $\widetilde{X} = \overline{X}$, els estimadors coincideixen i no s'ha de comparar res.

– Si n és senar, la desigualtat que hem de comprovar és:

$$1 \geq \frac{n+2}{3\,n} \iff n \geq 1.$$

Per tant, la desigualtat és certa si $n \geq 1$. Com abans, en el cas límit $n = 1$ els estimadors coincideixen ($\widetilde{X} = \overline{X}$).

- Veiem ara si $\mathrm{Var}(\widehat{N}_2) \geq \mathrm{Var}(\widehat{N}_3)$. Caldrà fer la comparació:

$$\frac{n+2}{3\,n} \geq \frac{2}{n+1} \iff (n-2)(n-1) \geq 0.$$

Cert si $n \geq 2$. Observem també que per a $n = 2$ els dos estimadors coincideixen ja que en aquesta cas $\overline{X} = \frac{X_{(1)}+X_{(2)}}{2}$ i per tant $\widehat{N}_3 = X_{(1)} + X_{(2)} - 1 = 2\,\overline{X} - 1 = \widehat{N}_2$.

- Resta comprovar si $\mathrm{Var}(\widehat{N}_3) \geq \mathrm{Var}(\widehat{N}_4)$. Però això és:

$$\frac{2}{n+1} \geq \frac{1}{n} \iff n \geq 1.$$

Així doncs, $\widehat{N}_4$ és un estimador no esbiaixat, i és el millor estimador de tots els que havíem plantejat, ja que té la variància més petita.

4 L'estimador $\widehat{N}_4$ és l'UMVUE

A la secció anterior hem arribat a la conclusió que $\widehat{N}_4$ és el millor estimador de tots els que havíem plantejat, ja que és no esbiaixat i té la variància més petita. Però de fet és, a més, el millor possible de tots els estimadors que es puguin proposar. No trobarem un altre estimador amb variància més petita. Encara que, pot ser, n'hi hagi un altre amb la mateixa variància.

És fàcil veure que el nostre estimador $\widehat{N}_4$ és l'estimador de màxima versemblança d'una $\mathrm{Unif}(1, \ldots, N)$ reescalat i desplaçat.

A més, $\widehat{N}_4$ és l'UMVUE *(uniformly minimum-variance unbiased estimator)*, és a dir, és un estimador no esbiaixat que té varància més petita que cap altre estimador no esbiaixat que estimi el paràmetre N a partir de la mostra. No farem aquí la demostració però aquesta es pot obtenir seguint els passos següents:

1. És un estadístic suficient. Això vol dir a grans trets que el coneixement de tots els elements de la mostra $X_1, \ldots, X_n$ no afegeix informació addicional sobre el paràmetre N que volem estimar que no aporti el propi estadístic $\widehat{N}_4$. És fàcil demostrar-ho utilitzant el Teorema de Fisher-Neyman que permet demostrar la suficiència observant la forma de la funció de versemblança.

2. És un estadístic complet. En estadística la completitut és una propietat d'un estadístic en relació amb un model teòric, que depèn d'uns paràmetres, per a un conjunt de dades observades. En essència, és una condició que garanteix que els paràmetres de la distribució de probabilitat que representen el model es poden calcular en funció de l'estadístic: assegura que les distribucions corresponents a diferents valors dels paràmetres siguin diferents. Aquesta propietat també és fàcil de comprovar a partir de la definició matemàtica d'estadístic complet aplicada a aquest estadístic.

3. Un cop s'han demostrat els dos punts anteriors, s'aplica el Teorema *Lehmann-Scheffe* per obtenir que l'estimador és l'UMVUE.

5 Interval de confiança per a $\widehat{N}_4$

En la secció anterior hem demostrat que $\widehat{N}_4$ és el millor estimador (no en trobarem un altre amb variància més petita). Un cop hem demostrat això, és lògic preguntar-se quin nivell de confiança tindrem quan estimem N amb aquest estimador. Per fer-ho calcularem l'interval de confiança. En estadística, un interval de confiança és un tipus d'estimació d'un paràmetre desconegut a partir de les dades d'una mostra donant un interval de valors. El nivell de confiança γ representa la proporció d'intervals que contindrien el valor veritable del paràmetre si l'experiment es repetís infinites vegades. Per exemple, quan es diu que el nivell de confiança és del 95 % vol dir que amb el 95 % de les mostres possibles s'obté un interval que conté el valor real del paràmetre mentre que el 5 % complementari dona intervals que no el contenen.

Per tal de simplificar els càlculs en aquesta secció suposarem que hi ha reemplaçament. Això farà que la nostra estimació sigui més conservadora, és a dir, l'interval de confiança serà una mica més gran del que s'obtindria sense aquesta suposició.

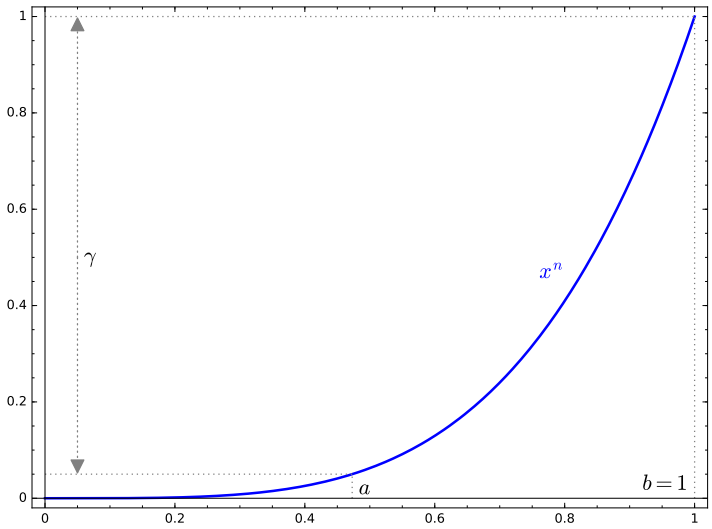
Recordem que X segueix una distribució $\text{Unif}(1, \ldots, N)$. Si considerem $Y := \frac{X}{N}$, aquest segueix una llei $\text{Unif}(\frac{1}{N}, \ldots, 1)$. Si N és prou gran podem aproximar aquesta distribució discreta per la distribució uniforme contínua

Unif$(0,1)$. (Això és un resultat prou clàssic que podeu consultar, per exemple, a [1, Exemple 1.14])

Considerem doncs que $Y \sim \text{Unif}(0,1)$. Aleshores per a $c \in (0,1]$

$$\begin{aligned} P(Y_{(n)} \leq c) &= P(Y_1 \leq c, Y_2 \leq c, \ldots, Y_n \leq c) \\ &= P(Y_1 \leq c)\, P(Y_2 \leq c) \cdots P(Y_n \leq c) \\ &= (P(Y \leq c))^n \\ &= (F_Y(c))^n = c^n. \end{aligned}$$

Per trobar l'interval de confiança començarem dibuixant aquesta funció de distribució.



Gràfica de la funció x^n fixant n i amb $x \in (0,1]$

Com que, fixat un *nivell de confiança* γ, busquem l'interval de longitud mínima tal que $\gamma = P(a \leq Y_{(n)} \leq b)$, és obvi que hem de prendre $b = 1$, ja que és on la funció té el pendent màxim.

Per tant, tenim que:

$$\begin{aligned} \gamma = P(a \leq Y_{(n)} \leq 1) &= P\left(a \leq \frac{X_{(n)}}{N} \leq 1\right) \\ &= P\left(X_{(n)} \leq N \leq \frac{X_{(n)}}{a}\right). \end{aligned}$$

De forma que

$$N \in \left[X_{(n)}, \frac{X_{(n)}}{a}\right].$$

Calculem el valor d'a en funció del nivell de confiança γ:

$$\gamma = P(a \leq Y_{(n)}) = 1 - P(Y_{(n)} < a) = 1 - a^n \quad \Rightarrow \quad a = \sqrt[n]{1-\gamma}.$$

Per tant l'interval de confiança és:

$$N \in \left[X_{(n)}, \frac{X_{(n)}}{\sqrt[n]{1-\gamma}} \right] \quad \text{amb un nivell de confiança de } \gamma.$$

Observem que:

- Per a mostres petites, l'interval de confiança és molt ample, reflectint la gran incertesa en l'estimació. (Per a n petit, el denominador $\sqrt[n]{1-\gamma}$ estarà a prop de 0, mentre que per a n gran es va acostant a 1).
- N no pot ser més petit que el màxim de la mostra, però pot ser arbitràriament superior a aquest.

6 Anàlisis Bayesià

Fins ara hem enfocat el problema des d'un punt de vista d'anàlisi freqüentista, tal i com van fer els aliats a la Segona Guerra Mundial per comptar els tancs de l'enemic. Ara però, volem donar una altra volta al problema i fer una aproximació des del punt de vista de l'estadística bayesiana.

L'anàlisi bayesià té la peculiaritat que incorpora en l'anàlisi les experiències que es tenen a priori sobre les dades, és a dir, és el mètode que, fent us de la fórmula de Bayes, permet corregir unes probabilitats a priori o de partida, generalment de caràcter subjectiu, en funció de la nova informació experimental o objectiva obtinguda mitjançant una mostra, i obtenir unes segones probabilitats revisades o "a posteriori".

Les premisses del problema són les mateixes: tenim una població numerada de mida desconeguda N, el nombre total de tancs, en el cas dels aliats, i volem estimar-ne la mida a partir d'una mostra aleatòria simple sense reposició $X_1, X_2, \ldots, X_n$, els números de sèrie dels n tancs capturats.

L'aproximació bayesiana al problema considera la credibilitat d'estimar bé N a partir de les dades de les que disposem. L'estadística bayesiana tracta els paràmetres, per exemple N, com a variables aleatòries mentre que en l'anàlisi freqüentista es tracten com a fixes. Per altra banda, tracta les dades com a variables fixes, mentre que els freqüentistes les tracten com a variables aleatòries.

Tot i que en estadística Bayesiana es parla de credibilitat i no de probabilitat, utilitzarem la notació de les probabilitats condicionades, és a dir,

$P(N = j|K = n, X_{(n)} = m)$, on K és el número d'elements observats i $X_{(n)}$ és la numeració màxima d'aquests.

A continuació farem alguns càlculs previs per poder calcular aquesta probabilitat.

Primer calcularem la probabilitat que el màxim de la mostra sigui m quan la mida de la població j és coneguda i s'han observat n elements (la mida de la mostra és n).

$$P(X_{(n)} = m|N = j, K = n) = \mathbb{1}_{[n\leq m\leq j]}\frac{\binom{m-1}{n-1}}{\binom{j}{n}},$$

on $\mathbb{1}_{[n\leq m\leq j]} = \begin{cases} 1 & \text{si } n \leq m \leq j \\ 0 & \text{en cas contrari} \end{cases}$

Assumim que la probabilitat a priori, abans de prendre la mostra, és alguna distribució uniforme discreta:

$$P(N = j|K = n) = \mathbb{1}_{[n\leq j\leq \beta]}\frac{1}{\beta - n},$$

el límit superior β ha de ser finit, perquè la funció

$$f(j) = \lim_{\beta\to\infty} \mathbb{1}_{[n\leq j\leq \beta]}\frac{1}{\beta - n},$$

és $f(j) = 0$, que no és una distribució de probabilitats.

Recordem que la fórmula de Bayes ens diu que si $A_1, \ldots, A_k$ formen una partició d'Ω aleshores:

$$P(A_i|B) = \frac{P(B|A_i)\,P(A_i)}{\sum_{\ell=1}^{k} P(B|A_\ell)\,P(A_\ell)}.$$

Aplicant-ho al nostre cas tenim,

$$\begin{aligned} &P(N = j|X_{(n)} = m, K = n) \\ &\quad = \mathbb{1}_{[m\leq j\leq \beta]}\frac{P(X_{(n)} = m|N = j, K = n)\cdot P(N = j|K = n)}{\sum_{\ell=m}^{\beta} P(X_{(n)} = m|N = \ell, K = n)\cdot P(N = \ell|K = n)} \\ &\quad = \mathbb{1}_{[m\leq j\leq \beta]}\frac{P(X_{(n)} = m|N = j, K = n)}{\sum_{\ell=m}^{\beta} P(X_{(n)} = m|N = \ell, K = n)}. \end{aligned}$$

Si $\sum_{\ell=m}^{\infty} P(X_{(n)} = m|N = \ell, K = n) \leq \infty$ fent el límit quan β tendeix a infinit, tenim que

$$P(N = j|X_{(n)} = m, K = n) = \mathbb{1}_{[m\leq j]}\frac{P(X_{(n)} = m|N = j, K = n)}{\sum_{\ell=m}^{\infty} P(X_{(n)} = m|N = \ell, K = n)},$$

on la $P(X_{(n)} = m|N = j, K = n)$ també s'anomena versemblança corresponent a j. Es tracta d'un número que quantifica com de versemblant seria que s'haguessin obtingut aquestes dades, si la veritable mida de la població fos $N = j$.

En el lema següent es calcula un dels termes que apareix posteriorment més d'un cop.

Lema 6.1.

$$\sum_{j=m}^{\infty} \frac{1}{\binom{j}{n}} = \frac{n}{n-1} \frac{1}{\binom{m-1}{n-1}}.$$

Demostració. De fet demostrarem que:

$$\frac{n-1}{n} \sum_{j=0}^{M} \frac{1}{\binom{j+x}{n}} = \frac{1}{\binom{x-1}{n-1}} - \frac{1}{\binom{M+x}{n-1}}.$$

L'equació que volem demostrar n'és un cas particular, quan $x = m$ i $M \to \infty$. Fem la demostració per inducció sobre M.

Si $M = 0$ observem que

$$\frac{n-1}{n} \frac{1}{\binom{x}{n}} = \frac{n-1}{n} \frac{n!\,(x-n)!}{x!} = \frac{(n-1)\,(n-1)!\,(x-n)!}{x!},$$

i també

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

Suposem ara que és cert per a $M = s$, és a dir

$$\frac{n-1}{n} \sum_{j=0}^{s} \frac{1}{\binom{j+x}{n}} = \frac{1}{\binom{x-1}{n-1}} - \frac{1}{\binom{s+x}{n-1}}.$$

Volem veure que aleshores també es compleix la igualtat per $M = s + 1$. És a dir, hem de demostrar que

$$\frac{n-1}{n} \left(\sum_{j=0}^{s} \frac{1}{\binom{j+x}{n}} + \frac{1}{\binom{s+1+x}{n}} \right) = \frac{1}{\binom{x-1}{n-1}} - \frac{1}{\binom{s+1+x}{n-1}}.$$

Però aplicant la hipòtesi d'inducció el terme de l'esquerra és igual a

$$\frac{1}{\binom{x-1}{n-1}} - \frac{1}{\binom{s+x}{n-1}} + \frac{n-1}{n}\frac{1}{\binom{s+1+x}{n}}.$$

Per tant és suficient provar que

$$-\frac{1}{\binom{s+x}{n-1}} + \frac{n-1}{n}\frac{1}{\binom{s+x+1}{n}} = -\frac{1}{\binom{s+x+1}{n-1}}.$$

I, efectivament, si desenvolupem en ambdós costats de la igualtat obtenim

$$-\frac{(s+x-n+2)!\,(n-1)!}{(s+x+1)!}.$$

Per tan hem demostrat el cas general i en particular hem demostrat que

$$\sum_{j=m}^{\infty} \frac{1}{\binom{j}{n}} = \frac{n}{n-1}\frac{1}{\binom{m-1}{n-1}}.$$

□

Ara ja tenim tot allò que és necessari per calcular $P(N = j|X_{(n)} = m, K = n)$. Per tal d'evitar allargar massa l'exposició estudiarem només la situació amb $n \geq 2$.

Ja havíem vist que:

$$P(X_{(n)} = m|N = j, K = n) = \mathbb{1}_{[m\leq j]}\frac{\binom{m-1}{n-1}}{\binom{j}{n}}.$$

Que, tal i com havíem dit, és la funció de versemblança de j i designarem com $L(j) := P(X_{(n)} = m|N = j, K = n)$.

D'altra banda, la funció de versemblança total és finita per $n \geq 2$, ja que pel Lema 6.1

$$\begin{aligned}\sum_{\ell=m}^{\infty} L(\ell) &= \sum_{\ell=m}^{\infty} P(X_{(n)} = m|N = \ell, K = n)\\ &= \binom{m-1}{n-1}\sum_{\ell=m}^{\infty}\frac{1}{\binom{\ell}{n}}\\ &= \binom{m-1}{n-1}\frac{n}{n-1}\frac{1}{\binom{m-1}{n-1}}\\ &= \frac{n}{n-1}.\end{aligned}$$

Aleshores, la funció de distribució de credibilitat (probabilitat a posteriori) és:

$$\begin{aligned} P(N=j|X_{(n)}=m,K=n) &= \frac{L(j)}{\sum_{l=m}^{\infty} L(\ell)} \\ &= \mathbb{1}_{[j\geq m]} \frac{n-1}{n} \frac{\binom{m-1}{n-1}}{\binom{j}{n}} \\ &= \mathbb{1}_{[j\geq m]} \frac{m-1}{j} \frac{\binom{m-2}{n-2}}{\binom{j-1}{n-1}}. \end{aligned}$$

Ara ja podem calcular el valor esperat de N. Utilitzant un altre cop el Lemma 6.1 tenim que

$$\begin{aligned} \widehat{N_b} &:= \sum_{j=m}^{\infty} jP(N=j|X_{(n)}=m,K=n) = \sum_{j=m}^{\infty} j \frac{m-1}{j} \frac{\binom{m-2}{n-2}}{\binom{j-1}{n-1}} \\ &= (m-1)\binom{m-2}{n-2} \sum_{j=m}^{\infty} \frac{1}{\binom{j-1}{n-1}} \\ &= (m-1)\binom{m-2}{n-2} \frac{n-1}{n-2} \frac{1}{\binom{m-2}{n-2}} \\ &= \frac{(m-1)(n-1)}{n-2}. \end{aligned}$$

Per tant, l'estimador de la mida de la població que es dedueix de l'anàlisi bayesià és:

$$\widehat{N_b} = \frac{(m-1)(n-1)}{n-2}.$$

Que en la notació original és igual a:

$$\boxed{\widehat{N_b} = \frac{(X_{(n)}-1)\,(n-1)}{n-2}.}$$

6.1 Interval de credibilitat

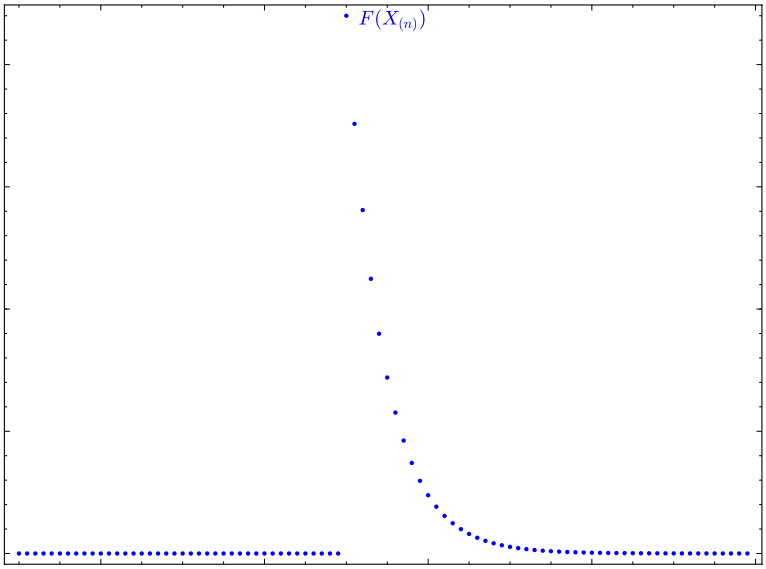
Recordem que en l'anàlisi bayesià, el que es calcula és el grau de credibilitat. Per tant calcular la regió de credibilitat és equivalent a dir que, donades les dades observades, hi ha una certa probabilitat que el valor de N pertanyi a la regió de credibilitat.

En l'anàlisi freqüentista en canvi, N és considerat un valor fix i les dades (i totes aquelles quantitats derivades de les dades, incloent els extrems de

l'interval de confiança) són variables aleatòries. Per tant donar l'interval de confiança freqüentista és equivalent a dir que hi ha un tant per cent de probabilitat que la mostra sigui bona en el sentit que el valor real de N pertanyi a l'interval obtingut amb aquella mostra.

Anem doncs a calcular l'interval de credibilitat. Igual que en l'interval de confiança, busquem el mínim interval $[I, S]$ d'entre aquells que tenen sota la corba de la funció de credibilitat una àrea de mida γ, és a dir, tals que $P(I \leq N \leq S) = \gamma$.

Representant gràficament la funció de distribució de credibilitat s'obté



Gràfic de la funció de distribució de credibilitat $F(j) = P(N = j|X_{(n)} = m, K = n) = \mathbb{1}_{[j \geq m]} \frac{m-1}{j} \frac{\binom{m-2}{n-2}}{\binom{j-1}{n-1}}$.

On veiem clarament que $I = X_{(n)}$ per tal que l'interval sigui de longitud mínima possible. Calcular S seria complicat, així que el que es fa a la pràctica és buscar el mínim S, tal que $\sum_{x=X_{(n)}}^{S} F(x) \geq \gamma$ essent $F(j) = P(N = j|X_{(n)} = m, K = n) = \mathbb{1}_{[j \geq m]} \frac{m-1}{j} \frac{\binom{m-2}{n-2}}{\binom{j-1}{n-1}}$. Obtenint així l'interval de credibilitat $[X_{(n)}, S]$.

7 Exercicis.

7.1 Simulació amb R

El codi en R següent simula un exemple d'aquest tipus de problemes. Genera una mostra aleatòria de mida $n = 100$ d'una $\text{Unif}(1, 2, \ldots, 5000)$ sense reposició. En aquest cas, doncs, $N = 5000$. Fa les estimacions de la mida de la població N a partir de la mostra, tant pel mètode freqüentista com pel mètode bayesià. I calcula els intervals de de confiança i credibilitat respectivament.

Per tal que els càlculs siguin reproduïbles s'imposa una *llavor* al generador de nombres aleatoris amb la instrucció `set.seed(2000)`.

```
N=5000
n=100
set.seed(2000)
u=sample(1:5000,100,replace=FALSE,prob=NULL)
u1=sort(u)
N1=2*median(u)-1
N1
N2=2*mean(u)-1
N2
N3=u1[1]+u1[n]-1
N3
N4=((n+1)/n)*u1[n]-1
N4
Nb=(u1[n]-1)*(n-1)/(n-2)
Nb
a=(1-0.95)^(1/n)
liminfN4=u1[n]
liminfN4
limsupN4=u1[n]/a
limsupN4
F=function(N)(u1[n]-1)*choose((u1[n]-2),(n-2))/(N*choose(N-1,n-1));
b=u1[n];
b
while(sum(F(u1[n]:b))<0.95)b=b+1;
b
```

Amb aquesta simulació s'obtenen els valors següents pels estimadors freqüentistes $\widehat{N_1} = 5725$, $\widehat{N_2} = 5131.52$, $\widehat{N_3} = 5062$ i $\widehat{N_4} = 4979.31$. Pel que fa a l'estimador Bayesià s'obté $\widehat{N_b} = 4980.31$.

Per tant clarament l'estimador $\widehat{N_4}$ és el que millor ha aproximat el valor real de N entre els estimadors freqüentistes, i gairebé ha pres el mateix valor que l'estimador Bayesià.

Si calculem els intervals de confiança i de credibilitat al 95%, ambdós coincideixen i donen $[4931, 5080]$.

Fixeu-vos que a partir de les simulacions no podem observar quasi diferència entre encarar el problema des d'un punt de vista freqüentista o bayesià, ja que els resultats dels dos mètodes són gairebé iguals. Només cal fer notar que en el cas de l'anàlisi Bayesià nosaltres hem posat com a distribució a priori la distribució uniforme. És a dir, suposem que no sabem res sobre el nombre N de tancs. Si enlloc d'aquesta distribució a priori hi posessim una distribució de N subjectiva basant-nos en les nostres suposicions els resultats millorarien en el cas que l'encertéssim i empitjorarien en cas que ens equivoquéssim.

7.2 Comptar corredors

Acabem amb un petit exercici: A partir de la fotografia següent, quants participants estimeu que hi va haver a la cursa?

Els autors hem comptat $n = 28$ dorsals, el més gran és $X_{(n)} = 639$ i els estimadors que hem obtingut són $\widehat{N}_4 = 660.8$ i $\widehat{N}_b = 662.5$. D'altra banda,

els intervals de confiança i de credibilitat que s'obtenen per a N són $[639, 711]$ i $[639, 712]$, respectivament.

8 Apèndix

8.1 Demostració del Lema 3.1

Veurem a continuació la demostració del Lema 3.1. Un pas previ és el resultat enunciat i demostrat tot seguit.

Lema 8.1. *En una mostra sense reposició de mida n obtinguda a partir d'una població de N elements numerats el moment d'ordre dos de l'estadístic d'ordre, $X_{(j)}$, per a $j = 1, \ldots, n$, ve donat per*

$$E[X_{(j)}^2] = \frac{j\,(N+1)\,((j+1)(N+2) - (n+2))}{(n+2)(n+1)}.$$

A més, si considerem $a < b$ amb $a, b \in \{1, \ldots, n\}$,

$$E[X_{(a)}X_{(b)}] = a\,\frac{N+1}{n+1}\,((N+1) - (n-b+1)\,\frac{N+2}{n+2}).$$

Demostració. Observem que

$$\begin{aligned} E[X_{(j)}^2] &= \sum_{k=j}^{N-n+j} k^2 P(X_{(j)} = k) \\ &= \frac{1}{\binom{N}{n}} \sum_{k=j}^{N-n+j} k^2 \binom{k-1}{j-1}\binom{N-k}{n-j} \\ &= \frac{j}{\binom{N}{n}} \sum_{k=j}^{N-n+j} k \binom{k}{j}\binom{N-k}{n-j}. \end{aligned}$$

Si en el darrer sumatori fem els canvis de variable $k = k' - 1$, $j = j' - 1$, $n + 1 = n'$ i $N + 1 = N'$ obtenim que

$$\begin{aligned}
&\sum_{k=j}^{N-n+j} k\binom{k}{j}\binom{N-k}{n-j} \\
&\quad = \sum_{k'=j'}^{N'-n'+j'} (k'-1)\binom{k'-1}{j'-1}\binom{N'-k'}{n'-j'} \\
&\quad = \sum_{k'=j'}^{N'-n'+j'} k'\binom{k'-1}{j'-1}\binom{N'-k'}{n'-j'} - \sum_{k'=j'}^{N'-n'+j'} \binom{k'-1}{j'-1}\binom{N'-k'}{n'-j'} \\
&\quad = j' \sum_{k'=j'}^{N'-n'+j'} \binom{k'}{j'}\binom{N'-k'}{n'-j'} - \binom{N'}{n'}.
\end{aligned}$$

Fent ara els canvis de variable $k' = k'' - 1$, $j' = j'' - 1$, $n' + 1 = n''$ i $N' + 1 = N''$ en el primer sumant obtenim que la darrera expressió és igual a

$$\begin{aligned}
j' \sum_{k''=j''}^{N''-n''+j''} \binom{k''-1}{j''-1}\binom{N''-k''}{n''-j''} - \binom{N'}{n'} &= j'\binom{N''}{n''} - \binom{N'}{n'} \\
&= (j+1)\binom{N+2}{n+2} - \binom{N+1}{n+1}.
\end{aligned}$$

Per tant,

$$\begin{aligned}
E[X_{(j)}^2] &= \frac{j}{\binom{N}{n}}\left((j+1)\binom{N+2}{n+2} - \binom{N+1}{n+1}\right) \\
&= j\left(\frac{(j+1)(N+2)(N+1)}{(n+2)(n+1)} - \frac{N+1}{n+1}\right) \\
&= \frac{j(N+1)((j+1)(N+2) - (n+2))}{(n+2)(n+1)}.
\end{aligned}$$

Això demostra la primera part del lema. Anem a demostrar ara que, si $a < b$ amb $a, b \in \{1, ..., n\}$, aleshores

$$E[X_{(a)}X_{(b)}] = a\,\frac{N+1}{n+1}\left((N+1) - (n-b+1)\frac{N+2}{n+2}\right).$$

Observem que

$$P(X_{(a)} = j, X_{(b)} = \ell) = \frac{\binom{j-1}{a-1}\binom{\ell-j-1}{b-a-1}\binom{N-\ell}{n-b}}{\binom{N}{n}}.$$

Com que es tracta d'una distribució de probabilitats, tenim

$$\sum_{j=a}^{N-n+a}\sum_{\ell=j+1}^{N-n+b}\frac{\binom{j-1}{a-1}\binom{\ell-j-1}{b-a-1}\binom{N-\ell}{n-b}}{\binom{N}{n}}=1.$$

Aleshores

$$\begin{aligned}E[X_{(a)}X_{(b)}] &= \sum_{j=a}^{N-n+a}\sum_{\ell=j+1}^{N-n+b} j\,\ell\,\frac{\binom{j-1}{a-1}\binom{\ell-j-1}{b-a-1}\binom{N-\ell}{n-b}}{\binom{N}{n}}\\ &= \frac{1}{\binom{N}{n}}\sum_{j=a}^{N-n+a}\sum_{\ell=j+1}^{N-n+b} j(N+1)\binom{j-1}{a-1}\binom{\ell-j-1}{b-a-1}\binom{N-\ell}{n-b}\\ &\quad -\frac{1}{\binom{N}{n}}\sum_{j=a}^{N-n+a}\sum_{\ell=j+1}^{N-n+b} j(N-\ell+1)\binom{j-1}{a-1}\binom{\ell-j-1}{b-a-1}\binom{N-\ell}{n-b}\\ &= \frac{a(N+1)}{\binom{N}{n}}\sum_{j=a}^{N-n+a}\sum_{\ell=j+1}^{N-n+b}\binom{j}{a}\binom{\ell-j-1}{b-a-1}\binom{N-\ell}{n-b}\\ &\quad -\frac{a(n-b+1)}{\binom{N}{n}}\sum_{j=a}^{N-n+a}\sum_{\ell=j+1}^{N-n+b}\binom{j}{a}\binom{\ell-j-1}{b-a-1}\binom{N-\ell+1}{n-b+1}.\end{aligned}$$

Observem que, si fem els canvis de variable $\ell=\ell'-1$, $b=b'-1$, $j=j'-1$, $a=a'-1$, $N'=N+1$ i $n'=n+1$ en el primer sumand, tenim

$$\begin{aligned}&\sum_{j=a}^{N-n+a}\sum_{\ell=j+1}^{N-n+b}\binom{j}{a}\binom{\ell-j-1}{b-a-1}\binom{N-\ell}{n-b}\\ &\qquad=\sum_{j'=a'}^{N'-n'+a'}\sum_{\ell'=j'+1}^{N'-n'+b'}\binom{j'-1}{a'-1}\binom{\ell'-j'-1}{b'-a'-1}\binom{N'-\ell'}{n'-b'}\\ &\qquad=\binom{N'}{n'}=\binom{N+1}{n+1}.\end{aligned}$$

Pel que fa al segon sumand, si fem els canvis de variable $\ell=\ell'-1$,

$b = b' - 1$, $j = j' - 1$, $a = a' - 1$, $N'' = N + 2$ i $n'' = n + 2$, obtenim

$$\begin{aligned}
&\sum_{j=a}^{N-n+a} \sum_{\ell=j+1}^{N-n+b} \binom{j}{a} \binom{\ell - j - 1}{b - a - 1} \binom{N - \ell + 1}{n - b + 1} \\
&\qquad = \sum_{j'=a'}^{N''-n''+a'} \sum_{\ell'=j'-1}^{N''-n''+b'} \binom{j' - 1}{a' - 1} \binom{\ell' - j' - 1}{b' - a' - 1} \binom{N'' - \ell'}{n'' - b'} \\
&\qquad = \binom{N''}{n''} = \binom{N + 2}{n + 2}.
\end{aligned}$$

Finalment substituint en el sumatori obtenim

$$\begin{aligned}
E[X_{(a)}X_{(b)}] &= a\,(N+1)\,\frac{\binom{N+1}{n+1}}{\binom{N}{n}} - a\,(n-b+1)\frac{\binom{N+2}{n+2}}{\binom{N}{n}} \\
&= a\,(N+1)\,\frac{N+1}{n+1} - a\,(n-b+1)\,\frac{(N+1)(N+2)}{(n+1)(n+2)} \\
&= a\,\frac{N+1}{n+1}\,(N+1-(n-b+1)\,\frac{N+2}{n+2}).
\end{aligned}$$

□

Demostració Lema 3.1. Estem ja en condicions de demostrar el Lema 3.1. Comencem demostrant la primera afirmació.

Demostració apartat a) Lema 3.1.

Utilitzant els Lemes 8.1 i 2.3 tenim que

$$\begin{aligned}
\mathrm{Var}(X_{(a)}) &= E[X_{(a)}^2] - (E[X_{(a)}])^2 \\
&= a\,\frac{N+1}{n+1}\,\frac{(a+1)(N+2) - (n+2)}{n+2} - a^2\,\frac{(N+1)^2}{(n+1)^2} \\
&= a\,\frac{N+1}{n+1}\,\frac{(N-n)(n-a+1)}{(n+1)(n+2)}.
\end{aligned}$$

Demostració apartat b) Lema 3.1.

Comencem calculant la covariància. Utilitzant els Lemes 8.1 i 2.3 tenim que

$$\begin{aligned}\mathrm{Cov}(X_{(a)}, X_{(b)}) &= E[X_{(a)}X_{(b)}] - E[X_{(a)}]E[X_{(b)}] \\ &= (a \wedge b)\,\frac{N+1}{n+1}\,(N+1-(n-(a\vee b)+1)\frac{N+2}{n+2}) \\ &\qquad - a\,b\,\frac{(N+1)^2}{(n+1)^2} \\ &= (a \wedge b)\,\frac{N+1}{n+1}\,\frac{(N-n)(n-(a\vee b)+1)}{(n+1)(n+2)}.\end{aligned}$$

Aleshores,

$$\begin{aligned}\mathrm{Var}(X_{(a)} + X_{(b)}) &= \mathrm{Var}(X_{(a)}) + \mathrm{Var}(X_{(b)}) + 2Cov(X_{(a)}, X_{(b)}) \\ &= a\,\frac{N+1}{n+1}\,\frac{(N-n)(n-a+1)}{(n+1)(n+2)} \\ &+ b\,\frac{N+1}{n+1}\,\frac{(N-n)(n-b+1)}{(n+1)(n+2)} \\ &+ 2\,(a \wedge b)\,\frac{N+1}{n+1}\,\frac{(N-n)(n-(a\vee b)+1)}{(n+1)(n+2)} \\ &= \frac{(N+1)(N-n)}{(n+1)^2(n+2)}\,((3\,(a \wedge b) + (a \vee b))(n+1) - (a+b)^2).\end{aligned}$$

Demostració apartat c) Lema 3.1.

Finalment, utilitzem aquests resultats per donar una forma alternativa de calcular la variància de la suma de totes les variables de la mostra.

En el càlcul necessitarem utilitzar els sumatoris següents

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2},\ \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6},$$

$$\sum_{i=1}^{n} i^3 = \frac{n^2(n+1)^2}{4} \quad \text{i} \quad \sum_{i=p}^{q} i = \frac{(q+p)(q-p+1)}{2}.$$

Observem que

$$\begin{aligned}
\mathrm{Var}(X_1 + \cdots + X_n) &= \mathrm{Var}(X_{(1)} + \ldots + X_{(n)}) \\
&= \sum_{i=1}^{n} \mathrm{Var}(X_{(i)}) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathrm{Cov}(X_{(i)}, X_{(j)}) \\
&= \sum_{i=1}^{n} i\,\frac{N+1}{n+1}\,\frac{(N-n)(n-i+1)}{(n+1)(n+2)} \\
&\qquad + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} i\,\frac{N+1}{n+1}\,\frac{(N-n)(n-j+1)}{(n+1)(n+2)} \\
&= \frac{(N+1)(N-n)}{(n+1)^2(n+2)}\Big[(n+1+2\,n\,(n+1)-n^2-n)\sum_{i=1}^{n} i \\
&\qquad\qquad - 2\,(n+1)\sum_{i=1}^{n} i^2 + \sum_{i=1}^{n} i^3\Big] \\
&= \frac{(N+1)(N-n)}{(n+1)^2(n+2)}\Big[(n+1)^2\,\frac{n\,(n+1)}{2} \\
&\qquad\qquad - 2\,(n+1)\,\frac{n\,(n+1)(2n+1)}{6} + \frac{n^2\,(n+1)^2}{4}\Big] \\
&= \frac{(N+1)(N-n)\,n}{12}.
\end{aligned}$$

□

Referències

[1] A. DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer Texts in Statistics. Springer, New York (2008).

[2] G. Davies. *How a statistical formula won the war*. Gavyn Davies does the maths, The Guardian, 20 de juliol de 2006. https://www.theguardian.com/world/2006/jul/20/secondworldwar.tvandradio

[3] R. W. Johnson. *Estimating the Size of a Population*. Teaching Statistics, 16, 50-52. (1994)

[4] B. Luque. *El problema de los tanques alemanes*. Investigación y Ciencia, 447, 90-91. (2013)

[5] J.S. Rao. *Problems for Rectangular Distributions (Or the Taxi Problem Revisited)* Metrika, Volume 28, 1981, page 257-262.

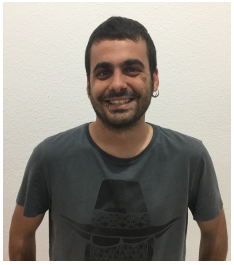
[6] R. Ruggles i H. Brodie. *An Empirical Approach to Economic Intelligence in World War II.* Journal of the American Statistical Association, Vol. 42, No.237, 1947, pp. 72-91.

[7] R. Touza Gil. *Los Panzer del Mariscal Rommel y el Iphone de Paris Hilton.* Revista general de marina, Vol. 263, MES 4, 2012, 687-693.

Xavier Bardina
Departament de Matemàtiques
Univ. Autònoma de Barcelona
bardina@mat.uab.cat

Sílvia Prior
Graduada en Matemàtiques i
Estadística Aplicada
Schibsted Media Group -Spain
silvia.prior@gmail.com

Carla Rodríguez
Graduada en Matemàtiques
Institut Bages Sud



Ferran Rosado
Graduat en Matemàtiques
ferranrosadoconejo@hotmail.com

Publicat el 28 de novembre de 2017