UNIVERSITAT AUTÒNOMA DE BARCELONA

FACULTAT DE CIÈNCIES

DEPARTAMENT DE MATEMÀTIQUES

TREBALL FINAL DE GRAU:

EL PROBLEMA DELS TANCS ALEMANYS

Sílvia Prior Cayuela

Bellaterra, Juny 2015

Tutor:
Xavier Bardina

Abstract

El problema dels tancs alemanys és un problema d'estimació de la mida d'una població quan aquesta està numerada correlativament. Aquest problema va ser resolt durant la Segona Guerra Mundial per part dels Aliats i el seu resultat segueix vigent avui en dia.
En aquest projecte s'expliquen els càlculs fins arribar a l'estimador de la mida d'una població des d'una vessant freqüentista i també bayesiana. Se'n farà una simulació per comparar i verificar resultats i finalment es veurà un cas pràctic on es pugui aplicar aquest estimador en un problema actual.

Paraules clau:
Estimador de la mida d'una població, Semilongitud de l'Interval, Distribució Uniforme Discreta, Estadística bayesiana, Funció de Credibilitat, Interval de Credibilitat.

Agraïments

En primer lloc m'agradaria donar les gràcies al meu tutor, Xavier Bardina, per donar-me la oportunitat d'escriure el treball final de grau sota la seva supervisió, la seva paciència, dedicació i paraules d'alè en tot moment.

També m'agradaria agrair-li al meu cap, Jose Pallarés, per la flexibilitat que m'ha donat en els horaris per poder conciliar la feina amb la redacció d'aquest projecte i poder finalitzar els meus estudis.

Finalment, el record més especial per a la meva familia, per donar-me sempre el seu suport en tot el que em proposo i la confiança que sempre dipositen en mi.

Índex

Índex de figures

Índex de taules

Capítol 1

Introducció

La Segona Guerra Mundial desgraciadament forma part de la nostra història receent com a éssers humans. Tot i que la majoria de nosaltres no l'haguem patit directament, tots hem pogut llegir, documentar-nos i esgarrifar-nos amb aquest episodi de la història de la humanitat. Podem trobar molta informació de les barbaritats que es van dur a terme en aquella època la qual ens ajuda a no oblidar i no caure en el mateix error de nou, però no obstant això, qui havia de pensar que darrere de tots aquests fets l'estadística va jugar-hi un paper important?

La intel·ligència econòmica dels Aliats havia de proporcionar dades sobre la indústria i 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 pel problema dels tancs alemanys.

Per resoldre'l, el departament d'intel·ligència de les forces Aliades d'entrada va utilitzar les tècniques habituals d'espionatge com són descodificar missatges encriptats, interrogatoris, ... les quals 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 amb les que treballar i preparar les seves estratègies militars. En aquell mo-

ment 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 que contenien aquestes etiquetes eren 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 neumàtics, 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 a Nord Àfrica la qual 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 del tanc associat. Aleshores, amb uns quants tancs capturats als alemanys i amb l'ajuda dels estadístics es podria estimar la seva producció de carruatges de combat.

En 1.1 trobem reflectit en números la precisió de les estimacions dutes a terme pels estadístics i la gran diferència amb les que es van realitzar en un primer moment per l'inteligència dels Aliats.

Date	Estimated Monthly Production		Monthly Production Speer Ministry
	Serial Number Estimate	Munitions Record 10 Aug. 42	
June, 1940	169	1000	122
June, 1941	244	1550	271
August, 1942	327	1550	342

Figura 1.1: *Comparativa de les estimacions fetes pels estadístics, l'inteligència i el registre alemà real*

A continuació podem observar en un diagrama de punts l'ajust d'aquestes prediccions per cada tipus d'element que es va analitzar.

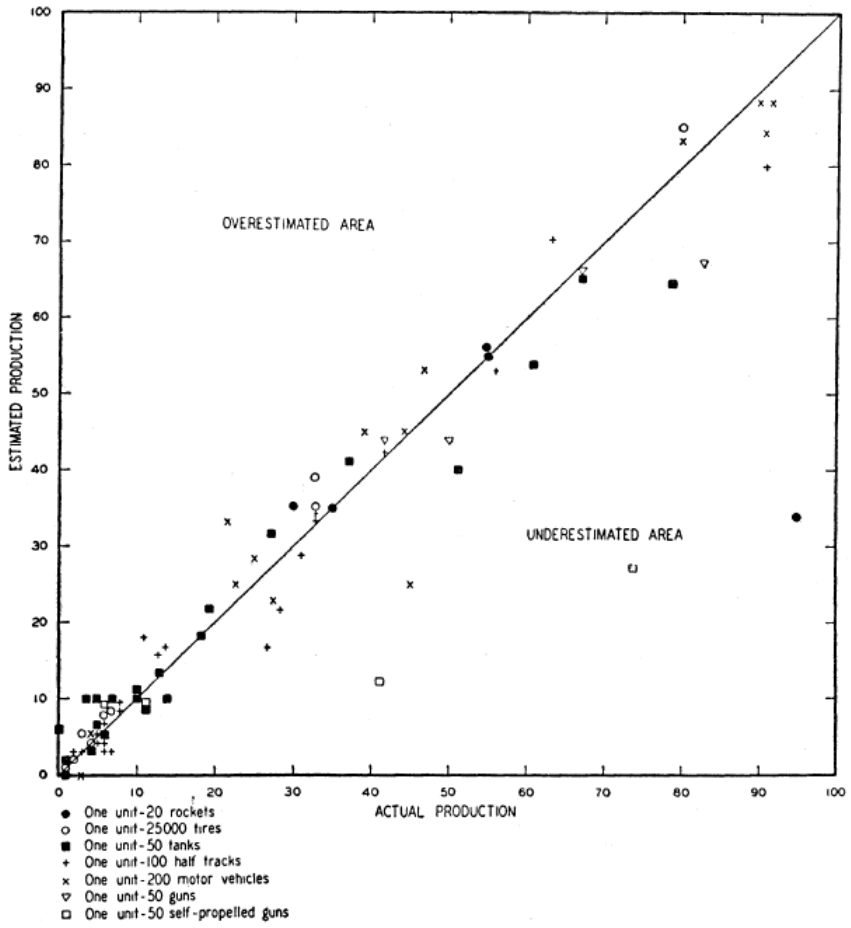

Figura 1.2: *Diagrama de punts de l'ajust de cada estimació individual diferenciat per tipus i model i per cada any pels set tipus d'equipament militar registrat.*

En aquest projecte 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, completitud,... i en farem una simulació amb R per comprovar i ratificar que l'estimador escollit és el millor.

La següent part del projecte consisteix en abordar el problema des de la vessant de l'estadística bayesiana. Això ens donarà un altre punt de vista diferent al vist utilitzant l'estadística més convencional i veurem si arribem a les mateixes conclusions o no.

Finalment veurem un cas pràctic en el que veurem que estimar la mida d'una població és un problema actual com per exemple, quants taxis hi ha a Barcelona? Quants corredors han participat en l'última marató de Nova York? Quants iphones 6 s'han venut des de la seva sortida al mercat?

Capítol 2

Estimant la mida d'una població

Suposem que tenim una població d'objectes numerada $1, 2, 3, \ldots, N$ amb N desconegut. A partir d'una mostra aleatòria simple sense reposició $X_1, X_2, \ldots, X_n$ de mida n volem estimar N.
Podem pensar que volem comptar, per exemple, el nombre de corredors d'una cursa, el nombre d'estands d'una fira o bé, els taxis que hi ha en una gran ciutat.

Aquest tipus d'estimacions les van utilitzar els Aliats durant la II Guerra Mundial per estimar la quantitat de tancs que tenia l'exèrcit alemany. Fins arribar a una bona estimació, el departament d'intel·ligència dels Aliats va cometre diversos errors. Els mètodes d'estimació classics donaven una producció militar exageradament gran la qual resultava inservible a l'hora de preparar l'estratègia militar. Però per sort se´n van adonar a temps. En aquell moment els estadístics van posar-se a treballar i van intentar trobar un estimador més coherent i aproximat a la realitat per la producció militar alemana.

2.1 Els estimadors

En aquesta secció deduïrem estimadors de la població només utilitzant el sentit comú (i una mica d'enginy).

Suposem que coneixem el valor mitjà m de la nostra població $1, 2, \ldots, N$ de la qual, recordem, és desconegut N. Aleshores hi ha $m-1$ valors per sota de m i $m-1$ per sobre, de manera que

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

Degut a que no coneixem qui és m, és natural substituir-lo per un estimador del valor mitjà com pot ser la mediana o la mitjana. D'aquesta manera obtenim els nostres primers dos estimadors.

Donada una mostra $X_1, X_2, \ldots, X_n$ denotem per $\tilde{X}$ la mediana de la mostra; $\tilde{X} = Mediana(X_1, X_2, \ldots, X_n)$ i denotem per $\bar{X}$ la mitjana de la mostra; $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$. Obtenim així els primers estimadors de la població que identifiquem per

$$\widehat{N_1} = 2\tilde{X} - 1 \tag{2.1}$$

i

$$\widehat{N_2} = 2\bar{X} - 1. \tag{2.2}$$

Observem que aquest últim estimador coincideix també al que trobem utilitzant el mètode dels moments. Ja que $X \sim Unif(1, \ldots, N)$. Per tant,

$$E(X) = \frac{N+1}{2}.$$

Pel mètode dels moments, substituïm l'esperança per la mitjana i tindrem que

$$\bar{X} = \frac{\widehat{N}+1}{2}$$

el qual reordenant obtenim que

$$\widehat{N} = 2\bar{X} - 1$$

coincidint així amb el resultat de $\widehat{N_2}$.

Però aquests estimadors tenen un gran inconvenient, poden donar estimadors clarmaent falsos. Pensem en un cas concret, per exemple si tenim una mostra de $n = 3$ tal que $X_1 = 2$, $X_2 = 10$ i $X_3 = 3$. Si estimem la mida de la mostra amb els estimadors obtinguts fins ara tenim que

$$\widehat{N_1} = 2 \cdot 3 - 1 = 6 - 1 = 5$$

$$\widehat{N_2} = 2 \cdot 5 - 1 = 10 - 1 = 9$$

quan clarament sabem que $N \geq 10$ ja que $X_2 = 10$.
Cal doncs buscar noves estimacions que no tinguin aquest problema.

Considerem utilitzar els estadístics d'ordre $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, la qual cosa vol dir que ordenem la nostra mostra $X_1, X_2, \ldots, X_n$ de menor a major:

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

Una primera consideració pot ser suposar que la distància que hi ha entre el número u i el primer valor després d'ordenar la mostra, $X_{(1)}$, serà la mateixa que entre l'N i el valor més gran observat,$X_{(n)}$. Obtenim així que podem igualar aquestes dues diferències de manera que en resulta el tercer estimador.

$$\widehat{N} - X_{(n)} = X_{(1)} - 1 \Longleftrightarrow \widehat{N_3} = X_{(n)} + X_{(1)} - 1. \tag{2.3}$$

Però observant bé aquest últim estimador podem encara refinar-lo més. Una derivació de $\widehat{N_3}$ seria estimar aquesta distància per la mitjana de les diferències entre observacions. És a dir:

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

Així doncs, el que s'obté és:

$$\widehat{N_4} = X_{(n)} + \frac{X_{(n)} - n}{n} \iff n \cdot \widehat{N_4} = n \cdot X_{(n)} + X_{(n)} - n \iff n \cdot \widehat{N_4} = (n+1) \cdot X_{(n)} - n \iff$$

$$\widehat{N_4} = \frac{(n+1)}{n} \cdot X_{(n)} - 1. \tag{2.4}$$

La primera observació que cal fer és que per construcció, $\widehat{N_3}$ i $\widehat{N_4}$ no tenen l'inconvenient de que poguessin ser menors que el valor observat més gran $X_{(n)}$.

D'altra banda, la segona observació que podem fer és que si busquem l'estimador de màxima versemblança de la mostra donada $X_1, X_2, \ldots, X_n$ tenim que

$$L(X_1, X_2, \ldots, X_n; N) = \left(\frac{1}{N}\right)^n \cdot \prod_{i=1}^{n} 1|_{\{1,\ldots,N\}}(X_i) = \left(\frac{1}{N}\right)^n \cdot 1|_{\{1,\ldots,N\}}(X_{(n)}).$$

Per tant $X_{(n)}$ és l'estimador de màxima versemblança. És a dir, $\widehat{N_4}$ és l'EMV reescalat i desplaçat.

De fet, $\widehat{N_4}$ és l'UMVUE (*uniformly minimum-variance unbaised estimator*) de N, és a dir, és un estimador no esbiaixat uniformement de mínima variància de N. Per demostrar-ho cal provar:

1. És un estadístic suficient

   Per demostrar aquesta condició enunciem el següent teorema:
   Teorema de Fisher-Neyman:
   La condició necessaria i suficient per la qual l'estadístic $T(X)$ sigui suficient és que la funció de versemblança $L(X;\theta)$ es pugui descomposar com a producte de dues funcions. Una d'aquestes funcions, $g(T(X);\theta)$, depenent del paràmetre i de la mostra, a través de l'estadístic $T(X)$, i l'altra, $h(X)$, independent del paràmetre θ. Això en altres paraules és que

   $$L(X;\theta) = g(T(X);\theta)h(X).$$

Si apliquem ara aquest teorema a la nostra funció de versemblança observem que:

$$L(X_1, X_2, \ldots, X_n; N) = \underbrace{\left(\frac{1}{N}\right)^n \cdot 1|_{\{1,\ldots,N\}}(X_{(n)})}_{g(X_{(n)};\theta)} \underbrace{1|_{\{1,X_{(n)}\}}(X_{(1)})}_{h(X)}$$

Fet que demostra que $X_{(n)}$ és un estadístic suficient.

2. És un estadístic complet

Recordem la definició d'estadístic complet:
Definició:
Sigui $X_1, X_2, \ldots, X_n$ una m.a. de $f_x(x|\theta)$ i sigui $T(X) = T(X_1, X_2, \ldots, X_n)$ un estadístic. Aleshores, direm que $T(X)$ és complet si i només si

$$E(g(T(X))) = 0 \Longrightarrow P(g(T(X)) = 0) = 1$$

per tot valor de θ i on $g(T(X))$ és una funció qualsevol de $T(X)$.

Comprovem si $X_{(n)}$ compleix la definició d'estadístic complet.
Si $X \sim Unif(1, \ldots, N)$, per la llei dels estadístics d'ordre que veurem en el punt 2.5 tenim que

$$P(X_{(n)} = k) = \frac{\binom{k-1}{n-1}}{\binom{N}{n}}$$

on k pren valors des de n fins a N.
Si,

$$E(g(X_{(n)})) = 0 \quad \forall N \qquad \Longrightarrow \qquad g(X) = 0$$

ja que,

$$E(g(X_{(n)})) = \sum_{k=n}^{N} g(k) \frac{\binom{k-1}{n-1}}{\binom{N}{n}}.$$

Això implica que en el cas en que $N = n$

$$g(n) \frac{1}{\binom{N}{n}} = 0 \Longrightarrow g(n) = 0.$$

Si considerem $N = n + 1$ el que s'obté és:

$$g(n)\frac{1}{\binom{N}{n}} + g(n+1)\frac{n}{\binom{N}{n}} = 0 \Longrightarrow g(n) = g(n+1) = 0.$$

Si fem doncs inducció sobre N, suposem que $g(k) = 0$ per $k = n, n+1, \ldots, m$, cal veure que per $N = m+1$, $g(m+1) = 0$. Llavors;

$$\sum_{k=n}^{m+1} g(k)\frac{\binom{k-1}{n-1}}{\binom{N}{n}} = g(m+1)\frac{\binom{m}{n-1}}{\binom{N}{n}} = 0 \Longrightarrow g(m+1) = 0$$

ja que els m valors anteriors la funció g és 0. Per tant, $P(g(X) = 0) = 1$.

Així doncs queda demostrat que $X_{(n)}$ compleix la definició per qualsevol valor de N i això ens assegura que $X_{(n)}$ és un estadístic complet.

3. Compleix el Teorema de Lehmann - Scheffé

Enunciem primerament el teorema:
Teorema de Lehmann-Scheffeé:
Sigui $X_1, X_2, \ldots, X_n$ una m.a. de $f_x(x|\theta)$ si:

(a) *$S(X)$ és un estadístic suficient de θ i complet.*

(b) *Sigui $T^*(X) = T^*(S(X))$ un altre estadístic que és funció de $S(X)$ tal que $E(T^*(X)) = \gamma(\theta)$ llavors $T^*(X)$ és un UMVUE per $\gamma(\theta)$.*

Ens falta doncs veure si $\widehat{N_4}$ compleix la condició (b) del teorema *Lehmann-Scheffeé* i haurem demostrat que $\widehat{N_4}$ és l'UMVUE per N.

Com hem dit anteriorment $\widehat{N_4}$ és l'estadístic $X_{(n)}$ reescalat i desplaçat. És a dir, $\widehat{N_4}$ és funció de $X_{(n)}$ i a més és no esbiaixat com veurem en l'anàlisi dels estimadors següent, $E(\widehat{N_4}) = N$. Per tant, es compleix també la segona condició del teorema i podem dir que $\widehat{N_4}$ és l'UMVUE per N.

En definitiva el que hem comprovat és que $\widehat{N_4}$ és un estimador no esbiaixat que té la variància més petita que cap altre estimador no esbiaixat per tots els valors possibles de N, és a dir, és l'UMVUE.

2.2 El biaix dels estimadors

El nostre objectiu és trobar quin dels quatre estimadors trobats és el millor. Per comparar-los començarem calculant el seu biaix i per fer-ho cal primerament calcular la *Llei dels estadístics d'ordre*.

2.2.1 Llei dels estadístics d'ordre

Sigui $X_1, X_2, \ldots, X_n$ una mostra aleatòria i sigui $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ la mostra ordenada de menor a major.
Observem que

$$P(X_{(j)} = k) = \frac{\binom{k-1}{j-1}\binom{N-k}{n-j}}{\binom{N}{n}} \tag{2.5}$$

on $k = j, j+1, \ldots, N-n+j$.
Amb aquest resultat podem doncs calcular ara l'esperança de $X_{(j)}$.

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

d'on s'extreu que

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

tal que $j = 1, \ldots, N$. Sabent això podem doncs calcular el biaix de cadascun dels estimadors trobats anteriorment.

2.2.2 Biaix de $\widehat{N_1} = 2\tilde{X} - 1$

Observem que l'estimador $\widehat{N_1}$ depèn de la mediana de la mostra X i aleshores podem calcular la $E(\widehat{N_1})$ en dues parts:

- **si** n **és senar;** $n = 2k+1$
  En aquest cas la mediana de X coincidiria amb $\tilde{X} = X_{(k+1)}$ i pel resultat anterior tindrem que

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

D'on, si substituïm en $E[\widehat{N_1}]$ tenim que

$$E[\widehat{N_1}] = E[2 \cdot \tilde{X} - 1] = N.$$

- **si n és parell;** $n = 2k$
  Llavors, la mediana és $\tilde{X} = \frac{X_{(k)}+X_{(k+1)}}{2}$. Calculem l'esperança de la mediana en aquest cas i tenim que

$$\begin{aligned} E[\tilde{X}] &= \frac{1}{2}\Big(E[X_{(k)}] + E[X_{(k+1)}]\Big) = \frac{1}{2}\Big(\frac{k}{2k+1}(N+1) + \frac{k+1}{2k+1}(N+1)\Big) \\ &= \frac{1}{2}\Big(\frac{2k+1}{2k+1}(N+1)\Big) = \frac{N+1}{2} \end{aligned}$$

  i per tant també tenim que

$$E[\widehat{N_1}] = E[2 \cdot \tilde{X} - 1] = N.$$

Amb aquests resultats el que comprovem és que $\widehat{N_1}$ és un estimador no esbiaixat d'N.

2.2.3 Biaix de $\widehat{N_2} = 2\bar{X} - 1$

L'esperança de l'estimador $\widehat{N_2}$ és:

$$E[\widehat{N_2}] = 2 \cdot E[\widehat{X}] - 1 = \cancel{2} \cdot \frac{N+1}{\cancel{2}} - 1 = N$$

ja que $E[X_i] = \frac{N+1}{2}$ i per tant $E[\widehat{X}] = \frac{N+1}{2}$.
Llavors, obtenim també que $\widehat{N_2}$ és un estimador no esbiaixat.

2.2.4 Biaix de $\widehat{N_3} = X_{(n)} + X_{(1)} - 1$

Cal calcular ara l'esperança de l'estimador $\widehat{N_3}$ la qual és:

$$E[\widehat{N_3}] = E[X_{(n)}] + E[X_{(1)}] - 1.$$

Com anteriorment hem trobat l'expressió de l'esperança dels estadístics d'ordre només cal substituir en el cas que $j = 1$ i $j = n$ i obtenim que:

$$E[X_{(n)}] = \frac{n}{n+1} \cdot (N+1) \tag{2.6}$$

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

i per tant obtenim que

$$E[\widehat{N_3}] = E[X_{(n)}] + E[X_{(1)}] - 1 = \frac{n}{n+1} \cdot (N+1) + \frac{1}{n+1} \cdot (N+1) - 1 = N.$$

Així doncs $\widehat{N_3}$ és també un estimador no esbiaixat de N.

2.2.5 Biaix de $\widehat{N_4} = \frac{(n+1)}{n} \cdot X_{(n)} - 1$

Finalment trobem l'esperança de l'últim estimador, $\widehat{N_4}$, la qual és molt fàcil de calcular amb tots els resultats que hem trobat anteriorment.

$$E[\widehat{N_4}] = E\Big[\frac{n+1}{n} X_{(n)} - 1\Big] = \frac{n+1}{n} E[X_{(n)}] - 1 = \frac{\cancel{n+1}}{\cancel{n}} \cdot \cancel{n} \cdot \frac{N+1}{\cancel{n+1}} - 1 = N.$$

Així doncs també $\widehat{N_4}$ és un estimador no esbiaixat.

Com que $\widehat{N_1}$, $\widehat{N_2}$, $\widehat{N_3}$ i $\widehat{N_4}$ són no esbiaixats, per saber quin dels quatre és més bon estimador calculem també la variància de cadascun d'ells.

2.3 La variància dels estimadors

Primerament recordem que per definició de variància sabem que

$$var(X) = E(X^2) - (E(X))^2 = E(X^2 + X) - E(X) - (E(X))^2 \qquad (2.7)$$

on X és una variable aleatòria.
I ara anem cas per cas a calcular la variància de cadascun dels estimadors trobats.

2.3.1 La variància de $\widehat{N_1} = 2\tilde{X} - 1$

Per calcular la variància d'aquest estimador cal diferenciar-ne dos casos, el cas parell i el cas senar. Per tant ho farem per parts.

- **Cas n senar;** $n = 2k + 1$

Com hem vist quan hem calculat la esperança de l'estimador $\widehat{N_1}$, si n és senar, $\tilde{X} = X_{(k+1)}$. Aleshores, en aquest cas, la variància de $\tilde{X}$ és la mateixa que per a $X_{(k+1)}$.

Utilitzant el resultat presentat en 2.7 per calcular la variància podríem doncs calcular $E[\tilde{X}^2]$ i $(E[\tilde{X}])^2$.
Per fer-ho observem que $E[\tilde{X}(N-\tilde{X}+1)] = (N+1)E[\tilde{X}] - E[\tilde{X}^2]$ i per tant calculant el primer terme d'aquesta igualtat obtindrem una expressió per a $E[\tilde{X}^2]$.

Per la llei dels estadístics d'ordre, 2.5, tenim que:

$$E[\tilde{X}(N-\tilde{X}+1)] = \frac{1}{\binom{N}{n}} \sum_{j=k+1}^{N-k} j(N-j+1)\binom{j-1}{k}\binom{N-j}{k}$$

$$= \frac{(k+1)^2}{\binom{N}{n}} \sum_{j=k+1}^{N-k} \binom{j}{k+1}\binom{N-j+1}{k+1}. \tag{2.8}$$

Fem ara les transformacions $j = j'+1$, $k+1 = k'$ i també $n' = n+2$ per substituir-ho en 2.8 i obtenir:

$$\frac{(k+1)^2}{\binom{N}{n}} \sum_{j=k+1}^{N-k} \binom{j}{k+1}\binom{N-j+1}{k+1} = \frac{(k+1)^2}{\binom{N}{n}} \sum_{j'=k+2}^{N-k+2} \binom{j'-1}{k+1}\binom{(N+2)-j'}{k+1}$$

$$= \frac{(k+1)^2}{\binom{N}{n}} \sum_{j'=k'+1}^{N-k'+2} \binom{j'-1}{k'}\binom{(N+2)-j'}{k'} = \frac{(k+1)^2}{\binom{N}{n}}\binom{N+2}{n+2}$$

$$= (k+1)^2 \frac{N+2}{n+2}\frac{N+1}{n+1} = \left(\frac{n+1}{2}\right)^2 \frac{N+2}{n+2}\frac{N+1}{n+1} = \frac{(n+1)}{4}\frac{(N+2)(N+1)}{n+2}. \tag{2.9}$$

Així doncs,

$$\begin{aligned}
& var(2\tilde{X}-1) = \\
= \;& 4var(\tilde{X}) = 4\Big(E[\tilde{X}^2] - (E[\tilde{X}])^2\Big) \\
= \;& 4\Big[(N+1)E[\tilde{(}X)] - \frac{(n+1)}{4}\frac{N+2}{n+2}(N+1) - \Big(\frac{N+1}{2}\Big)^2\Big] \\
= \;& 4\Big[\frac{(N+1)^2}{2} - \frac{(n+1)}{4}\frac{(N+2)(N+1)}{n+2} - \frac{(N+1)^2}{4}\Big] \\
= \;& (N+1)^2 - (n+1)\frac{(N+2)(N+1)}{n+2} \\
= \;& \frac{N+1}{n+2}\Big[(N+1)(n+2) - (n+1)(N+2)\Big] = \frac{(N+1)(N-n)}{n+2}.
\end{aligned}$$

- **Cas n parell;** $n = 2k$

En aquest cas $\tilde{X} = \frac{X_{(k)}+X_{(k+1)}}{2}$.
Com en el cas senar utilitzarem l'observació anterior adaptada a aquest cas la qual ens assegura que

$$E[X_{(k)}(N - X_{(k+1)} + 1)] = (N+1)E[X_{(k)}] - E[X_{(k)}X_{(k+1)}].$$

També utitlitzant la llei dels estadístics d'ordre, 2.5, i que

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

on $j = k, \cdots, N-k$ i $l = j+1, \cdots, N-k+1$ i per tant

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

Començarem calculant $E[X_{(k)}(N - X_{(k+1)} + 1)]$.

$$E[X_{(k)}(N - X_{(k+1)} + 1)] = \frac{k^2}{\binom{N}{n}}\sum_{j=k}^{N-k}\sum_{l=j+1}^{N-k+1}\binom{j}{k}\binom{N-l+1}{k} \qquad (2.10)$$

En aquest punt considerem les transformacions $k = k' - 1$, $j + 1 = j'$ i $l' = l + 1$ i ho apliquem a 2.10 per obtenir

$$\begin{aligned}
& \frac{k^2}{\binom{N}{n}} \sum_{j=k}^{N-k} \sum_{l=j+1}^{N-k+1} \binom{j}{k} \binom{N-l+1}{k} = \frac{k^2}{\binom{N}{n}} \sum_{j=k'-1}^{N-k'+1} \sum_{l=j+1}^{N-k'+2} \binom{j}{k'-1} \binom{N-l+1}{k'-1} \\
= & \frac{k^2}{\binom{N}{n}} \sum_{j'=k'}^{N-k'+2} \sum_{l=j'}^{N-k'+2} \binom{j'-1}{k'-1} \binom{N-l+1}{k'-1} \\
= & \frac{k^2}{\binom{N}{n}} \sum_{j'=k'}^{N-k'+2} \sum_{l=j'+1}^{N-k'+3} \binom{j'-1}{k'-1} \binom{N-l'+2}{k'-1} \\
= & \frac{k^2}{\binom{N}{n}} \binom{N+2}{2k+2} = k^2 \frac{(N+2)}{(n+2)} \frac{(N+1)}{(n+1)} = \frac{n^2}{4} \frac{(N+2)}{(n+2)} \frac{(N+1)}{(n+1)}. \qquad (2.11)
\end{aligned}$$

Per tant,

$$E[X_{(k)} X_{(k+1)}] = \frac{n}{2} \frac{(N+1)^2}{n+1} - \frac{n^2}{4} \frac{(N+2)(N+1)}{(n+2)(n+1)}.$$

Aprofitant els càlculs anteriors podem també trobar $E[X_{(k)}(N - X_{(k)} + 1)]$ que és:

$$\begin{aligned}
E[X_{(k)}(N - X_{(k)} + 1)] &= \frac{k(k+1)}{\binom{N}{n}} \sum_{j=k}^{N-k} \binom{j}{k} \binom{N-j+1}{k+1} \\
&= \frac{k(k+1)}{\binom{N}{n}} \sum_{j'=k}^{N-k'+2} \binom{j'-1}{k'-1} \binom{N-j'+2}{k'} \\
&= k(k+1) \frac{(N+2)}{(n+2)} \frac{(N+1)}{(n+1)} = \frac{n}{2} \frac{\cancel{(n+2)}}{2} \frac{(N+2)(N+1)}{\cancel{(n+2)}(n+1)} \\
&= \frac{n(N+2)(N+1)}{4(n+1)}.
\end{aligned}$$

Per tant,

$$E[X_{(k)}^2] = (N+1)E[X_{(k)}] - \frac{n(N+2)(N+1)}{4(n+1)} = \frac{n}{2} \frac{(N+1)^2}{n+1} - \frac{n(N+2)(N+1)}{4(n+1)}.$$

De la mateixa manera,

$$E[X_{(k+1)}(N - X_{(k+1)} + 1)] = \frac{n(N+2)(N+1)}{4(n+1)}$$

$$E[X^2_{(k+1)}] = \frac{(n+2)(N+1)^2}{2(n+1)} - \frac{n(N+2)(N+1)}{4(n+1)}$$

ja que com hem vist anteriorment $E[X_{(k+1)}] = (k+1)\frac{(N+1)}{n+1} = \frac{(n+2)(N+1)}{2(n+1)}$.

$$\begin{aligned}
& E[(X_{(k)} + X_{(k+1)})^2] = E[X^2_{(k)}] + E[X^2_{(k+1)}] + 2E[X_{(k)}X_{(k+1)}] \\
= \; & \frac{n}{2}\frac{(N+1)^2}{n+1} - \frac{n(N+2)(N+1)}{4(n+1)} + \frac{(n+2)(N+1)^2}{2(n+1)} - \frac{n(N+2)(N+1)}{4(n+1)} \\
& + n\frac{(N+1)^2}{n+1} - \frac{n}{2}\frac{(N+2)(N+1)}{(n+2)(n+1)} \\
= \; & (N+1)^2 - \frac{n(N+2)(N+1)}{2(n+1)} + n\frac{(N+1)^2}{n+1} - \frac{n^2}{2}\frac{(N+2)(N+1)}{(n+2)(n+1)} \\
= \; & (N+1)^2 + \frac{(N+1)}{(n+1)(n+2)}\Big[- \frac{n(N+2)}{2}(n+2) + n(N+1)(n+2) - \frac{n^2(N+2)}{2}\Big] \\
= \; & (N+1)^2 + \frac{(N+1)(N-n)n}{(n+1)(n+2)}.
\end{aligned}$$

Finalment amb tots els càlculs realitzats podem determinar el valor de $var(2\tilde{X} - 1)$ que és:

$$\begin{aligned}
var(2\tilde{X} - 1) &= var(2\tilde{X}) = var(X_{(k)} + X_{(k+1)}) \\
&= \cancel{(N+1)^2} + \frac{(N+1)(N-n)n}{(n+1)(n+2)} - \cancel{(N+1)^2} = \frac{(N+1)(N-n)n}{(n+1)(n+2)}
\end{aligned}$$

2.3.2 La variància de $\widehat{N_2} = 2\bar{X} - 1$

$$var(\widehat{N_2}) = var(2\bar{X} - 1) = 4 \cdot var(\bar{X}) = 4 \cdot \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$$

on σ^2 és la variància d'una $Unif(1, 2, \cdots, N)$. Per tant cal calcular qui és σ^2.

Si σ^2 denota la variància d'una distribució uniforme, sabem que la seva expressió és:

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}\Big(j - \frac{N+1}{2}\Big)^2 \tag{2.12}$$

I per tant, per calcular-la cal diferenciar dos casos:

- **Cas N senar; $N = 2k + 1$**

  Si $N = 2k+1$ fent uns simples càlculs aritmètics podem també expressar-ho com $\frac{N+1}{2} = \frac{2k+2}{2} = k + 1$.
  D'aquesta manera podem substituir en l'expressió anterior de σ^2,2.12 , i tenim que:

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\sum_{j=1}^{N}\left(j - \frac{N+1}{2}\right)^2 = \frac{1}{N}\sum_{j=1}^{N}\left(j - (k+1).\right)^2 = \frac{2}{N}\sum_{j=1}^{k}(j-(k+1))^2 \\ &= \frac{2}{N}\sum_{j=1}^{k}(k-j+1)^2 \qquad (2.13)\end{aligned}$$

  Si a l'equació anterior,2.13, denotem per $u = k - j + 1$ tindrem una sèrie coneguda, concretament:

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

  la qual el seu resultat podem utilitzar per simplificar els nostres càlculs. Així doncs continuant amb l'expressió de σ^2 de l'equació 2.13 tindrem que:

$$\begin{aligned}\frac{2}{N}\sum_{j=1}^{k}(k-j+1)^2 &= \frac{2}{N}\sum_{u=1}^{k}u^2 = \frac{2}{N}\frac{k(k+1)(2k+1)}{6} \\ &= \frac{\cancel{2}}{\cancel{N}}\frac{N-1}{\cancel{2}}\frac{N+1}{2}\frac{\cancel{N}}{6} = \frac{(N-1)(N+1)}{12}\end{aligned}$$

  ja que $k = \frac{N-1}{2}$.
  Així doncs veiem que per al cas senar, l'expressió de σ^2 és

$$\sigma^2 = \frac{(N-1)(N+1)}{12}.$$

- **Cas N parell; $N = 2k$**

  Com en el cas senar, fent una petita transformació aritmètica tenim que $N = 2k$ és equivalent a que $\frac{N+1}{2} = k + \frac{1}{2}$.

També com en el cas anterior partim de l'expressió 2.12 per trobar la variància en el cas parell i obtenim que:

$$\sigma^2 = \frac{2}{N}\sum_{j=1}^{k}(j-(k+\frac{1}{2}))^2 = \frac{2}{N}\sum_{j=1}^{k}(k-j+\frac{1}{2})^2 \qquad (2.14)$$

Considerant ara que $t = k - j + 1$ podem substituïr en 2.14 i tenim

$$\begin{aligned}
&\frac{2}{N}\sum_{t=1}^{k}(t-\frac{1}{2})^2 = \frac{2}{N}\sum_{t=1}^{k}(t^2+(\frac{1}{2})^2-t) = \frac{2}{N}\sum_{t=1}^{k}t^2+\frac{k}{4}-\sum_{t=1}^{k}t \\
= &\frac{2}{N}\Big(\frac{k(k+1)(2k+1)}{6}+\frac{k}{4}-\frac{k(k+1)}{2}\Big) = \frac{2}{N}\Big(\frac{2k(k+1)(2k+1)-3k(2k+1)}{12}\Big) \\
= &\frac{2}{N}\frac{(2k+1)(2k-1)k}{12} = \frac{(N+1)\cancel{N}(N-1)}{\cancel{N}12} = \frac{(N+1)(N-1)}{12}.
\end{aligned}$$

Per tant, tant si N és parell com si N és senar la vàriancia del segon estimador és la mateixa ja que

$$var(\widehat{N_2}) = 4\frac{N-n}{N-1}\frac{\sigma^2}{n} = 4\frac{N-n}{\cancel{N-1}}\frac{1}{n}\frac{(N+1)(\cancel{N-1})}{12} = \frac{(N-n)(N+1)}{3n},$$

és a dir,

$$var(\widehat{N_2}) = \frac{(N-n)(N+1)}{3n}$$

2.3.3 La variància de $\widehat{N_3} = X_{(n)} + X_{(1)} - 1$

Per calcular aquesta variància realitzarem uns càlculs previs per tal de facilitar la comprensió al lector.
Primerament escriurem l'expressió de les probabilitats condicionades aplicant la llei dels estadístics d'ordre, 2.5, en el cas que intervinguin els termes $X_{(n)}$ i $X_{(1)}$:

$$P(X_{(n)} = k, X_{(1)} = j) = \frac{\binom{k-j-1}{n-2}}{\binom{N}{n}}$$

on $1 \leq j \leq N-n+1$ i $j+n-1 \leq k \leq N$.
Per tant es compleix que

$$\binom{N}{n} = \sum_{j=1}^{N-n+1}\sum_{k=j+n-1}^{N}\binom{k-j-1}{n-2}.$$

Ara, com quan hem calculat la variància del primer estimador calcularem unes esperances que ens simplificaran els càlculs.

$$
\begin{aligned}
E[(X_{(n)} - X_{(1)} + 1)(X_{(n)} - X_{(1)})] &= \frac{1}{\binom{N}{n}} n(n-1) \sum_{j=1}^{N-n+1} \sum_{k=j+n-1}^{N} \binom{k-j-1}{n-2} \\
&= \frac{n(n+1)}{\binom{N}{n}} \sum_{j=1}^{N-n'+3} \sum_{k=j+n'-3}^{N} \binom{k-j-1}{n'+2} \\
&= \frac{n(n+1)}{\binom{N}{n}} \sum_{j=1}^{(N+2)-n'+1} \sum_{k'=j+n'-1}^{N+2} \binom{k'-j-1}{n'-2} \\
&= \frac{n(n+1)}{\binom{N}{n}} \binom{N+2}{n+2} = \frac{n(n-1)(N+2)(N+1)}{(n+2)(n+1)}. \qquad (2.15)
\end{aligned}
$$

on hem aplicat els canvis de variable $n = n' - 2$ i $k' = k + 2$.

Observem que l'esperança calculada també correspon a la següent:

$$
\begin{aligned}
&E[(X_{(n)} - X_{(1)} + 1)(X_{(n)} - X_{(1)})] = E[X_{(n)}^2 - 2X_{(1)}X_{(n)} + X_{(1)}^2 + X_{(n)} - X_{(1)}] \\
&= E[X_{(n)}^2] - 2E[X_{(1)}X_{(n)}] + E[X_{(1)}^2] + E[X_{(n)}] - E[X_{(1)}] \qquad (2.16)
\end{aligned}
$$

i aleshores, combinant els resultats obtinguts en 2.15 i en 2.16 tenim que:

$$
2E[X_{(1)}X_{(n)}] = -\frac{n(n-1)(N+2)(N+1)}{(n+2)(n+1)} + E[X_{(n)}^2] + E[X_{(1)}^2] + E[X_{(n)}] - E[X_{(1)}] \qquad (2.17)
$$

Finalment podem doncs calcular la variància de l'estimador $\widehat{N_3}$ el qual era el nostre objectiu inicial.

$$
\begin{aligned}
&var(\widehat{N_3}) = var(X_{(n)} + X_{(1)} - 1) = var(X_{(n)} + X_{(1)}) \\
&= E[(X_{(n)} + X_{(1)})^2] + (E[X_{(n)} + X_{(1)}])^2 \\
&= E[X_{(n)}^2] + 2E[X_{(1)}X_{(n)}] + E[X_{(1)}^2] - (N+1)^2
\end{aligned}
$$

En aquest pas podem substituïr $2E[X_{(1)}X_{(n)}]$ pel que hem trobat en 2.17 i

tindrem que:

$$
\begin{aligned}
& var(\widehat{N_3}) = \\
= \; & -\frac{n(n-1)(N+2)(N+1)}{(n+2)(n+1)} + 2E[X_{(n)}^2] + 2E[X_{(1)}^2] + E[X_{(n)}] - E[X_{(1)}] - (N+1)^2 \\
= \; & 2\Big[\frac{n(N+1)(N-n)}{(n+2)(n+1)^2} + \big(\frac{n(N+1)}{n+1}\big)^2\Big] \\
& +2\Big[\frac{n(n+1)(N+2)(N+1)}{(n+2)(n+1)} - (N+1)(N+2) + \frac{(N+1)+(N+2)}{n+1}(N+1)\Big] \\
& -\frac{n(n-1)(N+2)(N+1)}{(n+2)(n+1)} + \frac{n(N+1)}{n+1} - \frac{N+1}{n+1} - (N+1)^2 \\
= \; & -\frac{n(N+2)(N+1)}{n+2} + \frac{2n^2(N+1)^2}{(n+1)^2} - \frac{N(N+1)(n-1)}{n+1} + \frac{2n(N+1)(N-n)}{(n+2)(n+1)^2} \\
= \; & \frac{(N+1)}{(n+1)(n+2)}\Big[-n(N+2)(n+1) - N(n-1)(n+2) \\
& +\frac{2n^2(N+1)(n+2)}{n+1} + \frac{2n(N-n)}{n+1}\Big] \\
= \; & \frac{(N+1)}{(n+1)(n+2)}\Big[-2Nn(n+1) + 2N - 2n(n+1) + 2nN(n+1) + n\Big] \\
= \; & \frac{2(N-n)(N+1)}{(n+1)(n+2)}.
\end{aligned}
$$

2.3.4 La variància de $\widehat{N_4} = \frac{(n+1)}{n}X_{(n)} - 1$

En aquest apartat el que volem es calcular la variància concretament de $\widehat{N_4}$; és a dir, la variància de

$$var(\widehat{N_4}) = var\Big(\frac{(n+1)}{n}X_{(n)} - 1\Big) = \Big(\frac{(n+1)}{n}\Big)^2 var(X_{(n)})$$

la qual podem calcular utilitzant una de les equacions de la fórmula 2.7.

Comencem per calcular $E[(X_{(n)})^2 + X_{(n)}] = E[X_{(n)}(X_{(n)}+1)]$

$$\begin{aligned} E[X_{(n)}(X_{(n)}+1)] &= \frac{1}{\binom{N}{n}} \sum_{k=n}^{N} k(k+1)\binom{k-1}{n-1} = \frac{n(n+1)}{\binom{N}{n}} \sum_{k=n}^{N} \binom{k+1}{n+1} \\ &= \frac{n(n+1)}{\binom{N}{n}} \binom{N+2}{n+2} = \frac{\cancel{n(n+1)}}{\cancel{\binom{N}{n}}} \frac{N+2}{n+2} \frac{N+1}{\cancel{n+1}} \cancel{\binom{N}{n}} \\ &= \frac{n(N+2)(N+1)}{n+2} \end{aligned} \tag{2.18}$$

Ara, amb el resultat 2.18 i també el trobat en 2.6 podrem determinar qui és $var(X_{(n)})$.

$$\begin{aligned} var(X_{(n)}) &= E(X_{(n)}^2 + X_{(n)}) - E(X_{(n)}) - (E(X_{(n)}))^2 \\ &= \frac{n(N+2)(N+1)}{n+2} - \frac{n(N+1)}{n+1} - \frac{n^2(N+1)^2}{(n+1)^2} \\ &= \frac{n(N+1)}{(n+2)(n+1)^2}\Big((n+1)^2(N+2) - (n+1)(n+2) - n(n+2)(N+1)\Big) \\ &= \frac{n(N+1)(N-n)}{(n+2)(n+1)^2}. \end{aligned} \tag{2.19}$$

Finalment, ja tenim tots els càlculs preparats per determinar l'expressió de $var(\widehat{N_4})$ utilitzant sobretot el resultat de 2.19.

$$var(\widehat{N_4}) = \Big(\frac{(n+1)}{n}\Big)^2 var(X_{(n)}) = \frac{\cancel{(n+1)^2}}{\cancel{n^2}} \frac{\cancel{n}(N+1)(N-n)}{(n+2)\cancel{(n+1)^2}} = \frac{(N-n)(N+1)}{n(n+2)}.$$

Per acabar el apartat de les variàncies veurem quina d'elles és més petita. Observem que $1 \geq \frac{n}{n+1}$. A partir d'aquí anirem comparant les variàncies entre elles.

1. $\mathbf{var}(\widehat{N_1}) \geq \mathbf{var}(\widehat{N_2})$

- si n parell:

$$\begin{aligned} \frac{n}{n+1} \geq \frac{n+2}{3n} &\iff 3n^2 \geq n^2 + 3n + 2 \\ &\iff 2n^2 - 3n - 2 \geq 0 \\ &\iff (n-2)\Big(n + \frac{1}{2}\Big) \geq 0 \end{aligned}$$

  lo qual és cert per $n \geq 2$. En el cas de $n = 1$ i $n = 2$ els valors de la mitjana i la mediana coincideixen, $\tilde{X} = \bar{X}$.

- si n senar:

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

2. $\mathbf{var}(\widehat{N_2}) \geq \mathbf{var}(\widehat{N_3})$

$$\begin{aligned} \frac{n+2}{3n} \geq \frac{2}{n+1} &\iff n^2 + 3n + 2 \geq 6n \\ &\iff n^2 - 3n + 2 \geq 0 \\ &\iff (n-2)(n-1) \geq 0 \end{aligned}$$

   que és cert si $n \geq 2$.

3. $\mathbf{var}(\widehat{N_3}) \geq \mathbf{var}(\widehat{N_4})$

$$\begin{aligned} \frac{2}{n+1} \geq \frac{1}{n} &\iff 2n \geq n + 1 \\ &\iff n \geq 1 \end{aligned}$$

Per tant comprovem que cadascuna de les variàncies dels estimadors és més petita que l'anterior.

Per no perdre'ns entre tants càlculs i fòrmules resumim tots els resultats obtinguts en la taula presentada a continuació:

Estimadors de N	$E(\widehat{N_i})$	$Var(\widehat{N_i})$
$\widehat{N_1} = 2\tilde{X} - 1$	N	$\frac{(N-n)(N+1)}{(n+2)}$ (cas n senar)
		$\frac{n}{(n+1)}\frac{(N-n)(N+1)}{(n+2)}$ (cas n senar)
$\widehat{N_2} = 2\bar{X} - 1$	N	$\frac{(n+2)(N-n)(N+1)}{(n+2)}$
$\widehat{N_3} = X_{(n)} + X_{(1)} - 1$	N	$\frac{2(N-n)(N+1)}{(n+1)(n+2)}$
$\widehat{N_4} = \frac{(n+1)}{n}X_{(n)} - 1$	N	$\frac{(N-n)(N+1)}{n(n+2)}$

Taula 2.1: *Resum dels estimadors, esperances i variàncies*

2.4 Estimadors no esbiaixats de les variàncies

Continuem aquest capítol trobant estimadors no esbiaixats de les variàncies calculades en la secció anterior.
Cal remarcar que no ho podrem fer per a totes ja que no podem suposar la normalitat per totes elles. Així doncs només trobarem estimadors de les variàncies d'aquells N_i els quals podem suposar la normalitat.

2.4.1 Estimador de la variància de $\widehat{N_2}$

Per a $\widehat{N_2}$ podem suposar normalitat, si la mostra és gran, ja que es tracta d'una mitjana i és doncs clar. Dit això estimem la variància.
Observem que

$$E[(2\bar{X} - 1)^2] \;=\; var(2\bar{X} - 1) + (E[2\bar{X} - 1])^2 = \frac{(N-n)(N+1)}{3n} + N^2$$

i també cal notar que

$$(N-n)(N+1) = N^2 - (n-1)N - n.$$

Per tant

$$\begin{aligned} E[(2\bar{X}-1)^2-(n-1)(2\bar{X}-1)-n] &= \frac{(N-n)(N+1)}{3n}+(N-n)(N+1) \\ &= (N-n)(N+1)\Big(\frac{1}{3n}+1\Big) \\ &= (N-n)(N+1)\Big(\frac{3n+1}{3n}\Big). \end{aligned}$$

Aleshores, un estimador de la variància de $\widehat{N_2}$ sense biaix és:

$$\widehat{var(\widehat{N_2})} = \frac{1}{3n+1}\Big((2\bar{X}-1)^2-(n-1)(2\bar{X}-1)-n\Big).$$

2.4.2 Estimador de la variància de $\widehat{N_4}$

Com hem vist a l'inici d'aquest capítol 2, $X_{(n)}$ és l'estimador de màxima versemblança. D'altra banda el $\widehat{N_4}$ es construeix a partir d'aquest estadístic. Busquem doncs un estimador de la variància:

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

llavors

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

Per tant,

$$\widehat{var(\widehat{N_4})} = \frac{1}{(n+1)^2}\Big[\Big(\frac{n+1}{n}X_{(n)}-1\Big)^2-(n-1)\Big(\frac{(n+1)}{n}X_{(n)}-1\Big)-n\Big].$$

2.5 Intervals de confiança

Per acabar trobarem els intervals de confiança per a $\widehat{N_2}$ i $\widehat{N_4}$ i un interval alternatiu que anomenarem interval freqüentista.

2.5.1 Interval de confiança de $\widehat{N_2}$

Hem calculat un estimador no esbiaixat de la variància de $\widehat{N_2}$ en la secció anterior i sabem que aquest és

$$\widehat{var(\widehat{N_2})} = \frac{1}{3n+1}\Big((2\bar{X}-1)^2 - (n-1)(2\bar{X}-1) - n\Big).$$

D'aquesta manera tenim que l'interval de confiança del 95% és:

$$\Big(\widehat{N_2} \pm 1.96 \cdot \sqrt{\frac{1}{3n+1}\Big((2\bar{X}-1)^2 - (n-1)(2\bar{X}-1) - n\Big)}\Big).$$

2.5.2 Interval de confiança de $\widehat{N_4}$

Com en l'apartat anterior el cas de $\widehat{N_4}$ és anàleg sabent que un estimador no esbiaixat de la variància en aquest cas és

$$\widehat{var(\widehat{N_4})} = \frac{1}{(n+1)^2}\Big[\Big(\frac{n+1}{n}X_{(n)} - 1\Big)^2 - (n-1)\Big(\frac{(n+1)}{n}X_{(n)} - 1\Big) - n\Big].$$

Aleshores l'interval de confiança del 95% és:

$$\Big(\widehat{N_4} \pm 1.96 \cdot \sqrt{\frac{1}{(n+1)^2}\Big[\Big(\frac{n+1}{n}X_{(n)} - 1\Big)^2 - (n-1)\Big(\frac{(n+1)}{n}X_{(n)} - 1\Big) - n\Big]}\Big).$$

2.5.3 Interval de confiança freqüentista

Considerem una variable aleatoria Y que segueix una llei $Unif\{1,\dots,N\}$ i una mostra aleatoria simple sense reposició d'aquesta variable, $Y_1,\dots,Y_n$.

Considerem ara les variables $X_1 = \frac{Y_1}{N}, X_2 = \frac{Y_2}{N}, \dots, X_n = \frac{Y_n}{N}$.
Si N és gran podem suposar que X segueix una llei Uniforme en l'interval $(0,1)$, $X \sim Unif(0,1)$. Tenim així que $X_1, X_2, \dots, X_n$ que són v.a.i.i.d. amb llei $Unif(0,1)$.

D'aquesta manera tenim que

$$P(X_{(n)} \leq i) = P(X_1 \leq i, \dots, X_n \leq i) = \Big(P(X_1 \leq i)\Big)^n = i^n.$$

Observem ara que

$$0.05 = P(X_{(n)} < a) = a^n.$$

Si calculem doncs quins són els elements que ens donen una probabilitat del 0.95 tenim que:

$$\begin{aligned} 0.95 &= P(a \leq X_{(n)} \leq 1) \\ &= P\Big(a \leq \frac{Y_{(n)}}{N} \leq 1\Big) \\ &= P\Big(Y_{(n)} \leq N \leq \frac{Y_{(n)}}{a}\Big). \end{aligned}$$

Per tant, un interval de confiança del 95% per a N és:

$$\Big(Y_{(n)}, \frac{Y_{(n)}}{a}\Big) = \Big(Y_{(n)}, \frac{Y_{(n)}}{\sqrt[n]{0.05}}\Big). \tag{2.20}$$

Capítol 3

Anàlisi Bayesiana del problema

L'estadística Bayesiana és un subconjunt de l'estadística el qual es basa en l'evidència i en l'experiència de l'investigador sobre el que s'està intentant modelar.
La metodologia bayesiana consta de tres passos fonamentals:

1. Especificar un model de probabilitat que inclogui algun tipus de coneixement previ (a priori) sobre els paràmetres del model donat.

2. Actualitzar el coneixement sobre els paràmetres desconeguts condicionant a aquest model de probabilitat a les dades observades.

3. Evaluar l'ajust del model a les dades i considerar els possibles canvis sobre el model inicial proposat.

La diferència fonamental entre l'estadística freqüentista i la bayesiana és el concepte de probabilitat.
Per a l'estadística clàssica, la probabilitat, és un concepte *objectiu*, que es troba en les dades observades. Contrariament, per a l'estadística bayesiana la probabilitat la determina l'*observador*, i per tant és un concepte subjectiu. D'aquesta manera, en l'estadística clàssica només s'agafa com a font d'informació les mostres obtingudes a partir de l'observació. D'altra banda, en el cas bayesià, a més de les mostres, també juga un paper fonamental la

informació prèvia o externa que tenim en relació als fenòmens que volem modelitzar. És per aquest motiu que en estadística Bayesiana parlem de *graus de creença* o, més específicament, de probabilitats bayesianes.

Una de les diferències més significatives és que en comptes de parlar d'Intervals de confiança, parlem d'Intervals de Credibilitat degut a aquesta "creença"que cal tenir a la proposta inicial i el seu resultat ens dóna amb quina credibilitat s'assoleixen els estimadors proposats .

L'aproximació Bayesiana al problema dels tancs alemanys consisteix en considerar la credibilitat $(N = j|n = k, X_{(n)} = m)$ que el nombre de tancs enemics N és igual al nombre j, quan el nombre de tancs observats n és igual al nombre k i el nombre màxim de la mostra $X_{(n)}$ sigui el nombre m. Per simplificar considerarem que $(N = j|n = k, X_{(n)} = m)$ és igual que escriure $(j|k, m)$ i cal interpretar-ho com la credibilitat de que les variables N, n i $X_{(n)}$ assoleixin els nombres j, k i m respectivament.

Per la definició de les probabilitats condicionades, la *creença condicionada* ens dóna la següent igualtat:

$$P(j|m, k) = P(m|j, k)\frac{P(j|k)}{P(m|k)},$$

analitzem quin és el significat de cadascun dels termes que formen aquesta igualtat per separat.

L'expressió $P(m|j, k) = P(X_{(n)} = m|N = j, n = k)$ és la creença condicionada de que el nombre màxim observat sigui m, quan el nombre de tancs enemics es sap que és igual a j, i k és el nombre de tancs enemics observats. Això en altres paraules és

$$P(m|j, k) = \begin{cases} \frac{\binom{m-1}{k-1}}{\binom{j}{k}} & si \quad k \leq m \leq j \\ 0 & altrament \end{cases}$$

on el coeficient $\binom{j}{k}$ és el nombre de mostres k-dimensionals d'una població j-dimensional.

L'expressió $P(m|k) = P(X_{(n)} = m|n = k)$ és la creença de que el nombre màxim de la mostra (tancs observats) sigui m quan s'han observat k tancs, però, abans de que s'hagin observat els números explićits de la mostra. Es pot reescriure $P(m|k)$ en termes de les altres quantitats marginant sobre qualsevol nombre que prengui j.

$$\begin{aligned} P(m|k) &= P(m|k) \cdot 1 \\ &= P(m|k) \sum_{j=0}^{\infty} P(j|m,k) \\ &= P(m|k) \sum_{j=0}^{\infty} P(m|j,k) \frac{P(j|k)}{P(m|k)} \\ &= \sum_{j=0}^{\infty} P(m|j,k) P(j|k). \end{aligned}$$

D'altra banda, l'expressió $P(j|k) = P(N = j|n = k)$ és la credibilitat de que el nombre total de tancs és igual a j quan s'han observat k tancs, però, abans de que s'hagin observat els números de la mostra explícitament.
Suposem que es tracta d'alguna distribució uniforme discreta

$$P(j|k) = \begin{cases} \frac{1}{\Omega - k} & si \quad k \leq j < \Omega \\ 0 & altrament \end{cases}$$

on el límit superior Ω ha de ser finit ja que la funció

$$f(n) = \lim_{\Omega \to \infty} \begin{cases} \frac{1}{\Omega - k} & si \quad k \leq j < \Omega \\ 0 & altrament \end{cases}$$

és $f(n) = 0$ la qual no és una funció de distribució.

Llavors,

$$P(j|m,k) = \begin{cases} \frac{P(m|j,k)}{\sum_{j=m}^{\Omega-1} P(m|j,k)} & si \quad m \leq j < \Omega \\ 0 & altrament. \end{cases}$$

Si $\sum_{j=m}^{\infty} P(m|j,k) < \infty$ i prenent límits quan $\Omega \longrightarrow \infty$, la variable Ω desapareix de l'expressió i tenim que

$$P(j|m,k) = \begin{cases} \frac{P(m|j,k)}{\sum_{j=m}^{\infty} P(m|j,k)} & si \quad j \geq m \\ 0 & si \quad j < m. \end{cases} \tag{3.1}$$

3.1 La funció i l'Interval de Credibilitat

La probabilitat condicionada que el nombre més gran de les k observacions agafat a partir dels números de la sèrie $1, \ldots, j$, sigui igual a m, és

$$P(X_{(n)} = m|N = j, n = k \geq 2) = P(m|j,k) = 1|_{m \leq j} \frac{\binom{m-1}{k-1}}{\binom{j}{k}}$$

on $1|_{m \geq j}$ denota la condició que m ha de ser més petita o igual a j.

Considerem la següent identitat que més tard ens serà útil per simplificar resultats:

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

i que es dedueix a partir de [8].

La funció de màxima versemblança de j és doncs

$$L(j) = 1|_{j \leq m} \frac{\binom{m-1}{k-1}}{\binom{j}{k}}$$

i el total de la versemblança és finit per a $k \geq 2$ de manera que:

$$\begin{aligned} \sum_j L(j) &= \binom{m-1}{k-1} \sum_{j=m}^{\infty} \frac{1}{\binom{j}{k}} \\ &= \binom{m-1}{k-1} \cdot \frac{k}{k-1} \cdot \frac{1}{\binom{m-1}{k-1}} \\ &= \frac{k}{k-1}. \end{aligned}$$

La funció de credibilitat és doncs

$$
\begin{aligned}
&P(N = j | X_{(n)} = m, n = k \geq 2) = P(j|m,k) \\
&= \frac{L(j)}{\sum_j L(j)} \\
&= 1|_{j \geq m} \frac{k-1}{k} \frac{\binom{m-1}{k-1}}{\binom{j}{k}} \\
&= 1|_{j \geq m} \frac{m-1}{j} \frac{\binom{m-2}{k-2}}{\binom{j-1}{k-1}} \\
&= 1|_{j \geq m} \frac{m-1}{j} \frac{m-2}{j-1} \frac{k-1}{k-2} \frac{\binom{m-3}{k-3}}{\binom{j-2}{k-2}}.
\end{aligned} \tag{3.2}
$$

El resultat de la funció de credibilitat ens dóna doncs amb quina credibilitat, del 0 al 1, N pren el valor j quan $n = k$ i $X_{(n)} = m$.
Aquesta és una funció decreixent i pren el seu valor màxim en $X_{(n)}$ per tant per trobar l'interval de credibilitat partim de $X_{(n)}$ i anem sumant fins que la credibilitat sigui del 0.95 si el que estem buscant és un interval del 95% de credibilitat.
Aleshores, l'interval de credibilitat anirà des de $X_{(n)}$ fins al valor en que la suma de totes les "credibilitats" anteriors sigui 0.95.

3.2 L'estimador bayesià

L'estimador bayesià és troba arrel de la funció de credibilitat 3.2.
La idea és que l'estimador correspon a la mitjana dels valors possibles que pot prendre N, tot això tenint en compte els pesos que s'han donat per la mateixa funció trobada anteriorment.
Calculem doncs aquest estimador:

$$\begin{aligned}\widehat{N_b} \quad &= \mu = \sum_j j \cdot P(N = j | X_{(n)} = m, n = k) \\ &= \sum_j j1|_{j \leq m} \frac{m-1}{j} \frac{\binom{m-2}{k-2}}{\binom{j-1}{k-1}} \\ &= (m-1)\binom{m-2}{k-2} \frac{k-1}{(k-2) \cdot \binom{m-2}{k-2}} \\ &= \frac{(m-1)(k-1)}{(k-2)}. \end{aligned} \tag{3.3}$$

Amb aquest resultat concluïm el capítol d'anàlisi bayesià en el qual hem donat unes nocions de en que es basa aquest tipus d'anàlisi, hem fet els càlculs fins a trobar la funció de credibilitat, hem raonat com es calcula l'interval de credibilitat i finalment hem pogut donar un estimador bayesià del problema.

Capítol 4

Simulació

Una de les eines que tenim per comparar els estimadors entre ells és la simulació.
En aquest punt del projecte simularem una mostra de mida 100 d'una població amb valor mínim 1 i valor màxim 5000. La mostra serà considerada sense reposició ja que els estimadors trobats són vàlids quan es compleix aquesta condició. Un cop generada la mostra aleatòria calcularem els quatre estimadors trobats en el segon capítol i en calcularem els diferents intervals de confiança. A més a més en calcularem el estimador bayesià i l'interval de credibilitat els quals hem mostrat en el capítol anterior. Tota aquesta simulació l'hem fet amb el software lliure R i el codi utilitzat el podeu trobar en l'apèndix B.

Els resultats després de la simulació els hem recollit a la taula que mostrem a continuació.

Estimador	**Estimació puntual**	**Interval**	**Semilongitud de l'Interval**
$\widehat{N_1}$	4333.26	-	-
$\widehat{N_2}$	3807	(3323.1 , 4290.9)	483.91
$\widehat{N_3}$	4971	-	-
$\widehat{N_4}$	4972.24	(4876.7 , 5067.8) [I.Confiança]	95.52
		(4924 , 4998.3) [I. Freqüentista]	37.16
$\widehat{N_b}$	4973.24	(4924 , 4988)	32

Taula 4.1: *Comparativa dels resultats de la simulació fet amb una mostra de mida 100 amb nombres de l'1 al 5000 generats per una uniforme sense reposició*

En la taula 4.1 podem observar que els estimadors més acurats són $\widehat{N_4}$ i l'estimador bayesià el qual el supera per una unitat. D'altra banda, l'interval de confiança d'$\widehat{N_2}$ és més ampli que el de $\widehat{N_4}$ fet que és coherent amb que l'estimador $\widehat{N_2}$ sigui pitjor que $\widehat{N_4}$. En canvi, un fet sorprenent és que l'interval freqüentista i l´interval bayesià siguin tant bons i tant pròxims entre ells; només disten 10 unitats.

L'altra observació que cal fer és que els intervals de confiança per a $\widehat{N_2}$ i $\widehat{N_4}$ són més amplis que els altres dos. Això és coherent amb el fet que el seu límit inferior és menor que el número màxim que conté la mostra. Aquesta característica els altres dos intervals no la tenen i per aquest motiu són més estrets.

Capítol 5

Un cas pràctic

Un dels objectius d'aquest projecte és també veure una aplicació actual a l'estimador que fa tants anys va ser tant útil.
Per fer-ho hem utilitzat dades d'una entitat dedicada al finançament al consum. Aquesta empresa dóna finançament als clients a través dels diferents establiments que hi estan afiliats, no directament al client final. Per aquest motiu es té un registre de tots els establiments afilitas amb el seu CIF (Codi d'Identificació Fiscal) el qual es considera col·loquialment com el DNI de les empreses.

Seguidament desgloçarem l'estudi realitzat amb les dades aportades per aquesta empresa, quines són i com s'han utilitzat, quins han sigut els resultats i amb quina finalitat s'han emprat.

5.1 El Codi d'Identificació Fiscal (CIF)

Les dades de les que disposem són un registre des de l'1 de gener del 2013 fins al 31 de maig del 2015 de tots els establiments afiliats a aquesta finançera amb dades de la província on es troben i el seu codi CIF. La primera pregunta que hom es fa és perquè el codi CIF és útil en el nostre estudi si el que

necessitem és una població numerada correlativament per poder utilitzar l'estimador trobat.
El codi CIF consta de 9 posicions les quals ens aporten informació sobre l'establiment. A continuació detallem que ens indiquen els dígits que formen el codi CIF.

- **Posició 1:** Lletra que determina de quin tipus de societat es tracta.
  Per exemple, si el codi CIF en la primera posició hi té una A, això vol dir que es tracta d'una Societat Anònima (S.A.). En canvi, si hi té una B, és una Societat Limitada (S.L.) i així fins a 19 possibilitats diferents.

- **Posició 2 i 3:** Dígits numèrics que identifiquen en quina província es troba el negoci.
  Els codis de cadascuna de les províncies els podeu troba detallats en l'Annex C.

- **Posició 4 fins a la 8:** Cinc dígits numèrics correlatius que corresponen a la inscripció de la organització al registre provincial.
  Aquesta és la part del codi CIF imprescindible en el nostre estudi i la que compleix els requisits per poder utilitzar els estimadors trobats.

- **Posició 9:** Dígit de control que pot ser un número o una lletra.
  Es genera a partir dels dígits en les posicions anteriors.

Per com està definit el codi CIF observem que els nombres correlatius són únics sempre que determinem la lletra (primera posició) i la província (posicions 2 i 3). Així doncs podrem aplicar l'estimador poblacional sempre i quan fixem les tres primeres posicions.
Per aquest motiu, de les dades provinents de la base de dades inicial, només considerarem aquells negocis que comencin per la lletra B els quals corresponen a les Societats Limitades i donarem un estimador poblacional per cada província.

Amb això es preten estimar quantes Societats Limitades hi ha en cadascuna de les provincies de l'estat espanyol per després poder fer diferents anàlisis dins de l'empresa. Per exemple quina quota de mercat té aquesta entitat financera per tal de tenir un major control de la seva captació de clients.

Un dels fets que cal tenir en compte és que les grans ciutats com Barcelona o Madrid no tenen un únic codi de província, de fet en tenen 10 i 11

respectivament.
En aquest cas, per saber quin és el total d'establiments en aquestes ciutats podem sumar el resultat dels estimadors de cadascun dels codis ja que tots són independents entre ells i això ens donarà el nombre que estàvem buscant. En canvi, si el que volem és calcular la semilongitud de l'interval freqüentista (I.F.), com s'ha definit en 2.20, caldrà trobar-lo considerant que la mida de la mostra n és ara la suma de les mides de les mostres de cada codi de província.

Per fer-ho suposem que tenim una província la qual té k codis diferents, és a dir, tenim k mostres les quals tenen mida $n_1, n_2, \ldots, n_k$.
Volem trobar l'Interval de Confiança, pel mètode freqüentista, del 95%. Això és equivalent a dir que volem una probabilitat del 0.95 lo qual es pot expressar com:

$$0.95 = (0.95)^{\frac{1}{k}} \cdot \underbrace{\ldots}_{k} \cdot (0.95)^{\frac{1}{k}}.$$

Ara observem que

$$\begin{aligned} P\Big(X^1_{(n_1)} \leq N_1 \leq \frac{X^1_{(n_1)}}{a_1}\Big) &= (0.95)^{\frac{1}{k}} \\ P\Big(X^2_{(n_2)} \leq N_2 \leq \frac{X^2_{(n_2)}}{a_2}\Big) &= (0.95)^{\frac{1}{k}} \\ &\vdots \\ P\Big(X^k_{(n_k)} \leq N_k \leq \frac{X^k_{(n_k)}}{a_k}\Big) &= (0.95)^{\frac{1}{k}} \end{aligned}$$

on $X^i_{(n_i)}$ denota el valor màxim del i-èssim codi de la província que estem considerant. Amb això, si fem el producte tindrem

$$\begin{aligned} 0.95 &= P\Big(X^1_{(n_1)} \leq N_1 \leq \frac{X^1_{(n_1)}}{a_1}\Big) \cdot \underbrace{\ldots}_{k} \cdot P\Big(X^k_{(n_k)} \leq N_k \leq \frac{X^k_{(n_k)}}{a_k}\Big) = \\ &= P\Big(X^1_{(n_1)} \leq N_1 \leq \frac{X^1_{(n_1)}}{a_1} \underbrace{\ldots}_{k}, X^k_{(n_k)} \leq n_k \leq \frac{X^k_{(n_k)}}{a_k}\Big) \leq \\ &\leq P\Big(X^1_{(n_1)} + \cdots + X^k_{(n_k)} \leq N_1 + \cdots + N_k \leq \frac{X^1_{(n_1)}}{a_1} + \cdots + \frac{X^k_{(n_k)}}{a_k}\Big) \leq \\ &\leq P\Big(X^1_{(n_1)} + \cdots + X^k_{(n_k)} \leq N \leq \frac{X^1_{(n_1)}}{a_1} + \cdots + \frac{X^k_{(n_k)}}{a_k}\Big). \end{aligned}$$

D'aquesta manera el que s'obté és que l'Interval de Confiança per a N és:

$$\left(X^1_{(n_1)} + \cdots + X^k_{(n_k)}, \frac{X^1_{(n_1)}}{a_1} + \cdots + \frac{X^k_{(n_k)}}{a_k}\right)$$

on $a_i = 1 - (0.95)^{\frac{1}{k}} = \sqrt[n_i]{0.05}$.

Feta aquesta apreciació, en l'apèndix C s'ha recollit en una taula les estimacions que hem fet utilitzant l'estimador $\widehat{N_4}$ i la semilongitud de l'interval freqüentista classificat per codi de província i finalment agrupat pel total de la província. També es mostra l'estimació d'establiments en el total de la província.

Donats els resultats el que s'observa és que com major és la mida de la mostra més petita és la semilongitud de l'interval. També observem casos extrems com el de *Melilla* en el qual aquesta empresa només hi té un establiment afiliat i per tant no se'n pot extreure cap conclusió. El mateix passa en el cas de *Ceuta* tot i tenir-ne 2 establiments.
En el cas oposat es troba, per exemple, un dels codis d'*Alacant*, en concret el 54. En aquest cas es fa una estimació de gairebé 80.000 establiments i la semilongitud de l'interval és només de 861, la qual cosa ens dóna una bona estimació. El mateix passa amb alguns dels codis de província de grans ciutats com *Barcelona* i *Madrid* (sobretot el codi 81 i 83).

Els Intervals de Confinança són bastant ajustats, exceptuant els casos extrems en que tenim pocs establiments. En promig la longitud de l'interval és d'un 12% amb lo qual es pot considerar que les estimacions són prou bones, sempre amb un 95% de confiança.

Capítol 6

Conslusions

En aquest projecte volíem trobar un estimador d'una població numerada des d'un punt de vista de l'estadística clàssica i també des de la vessant bayesiana. Aquest fet l'hem assolit trobant els estimadors $\widehat{N_1}$, $\widehat{N_2}$, $\widehat{N_3}$ i $\widehat{N_4}$ amb l'estadística més convencional i $\widehat{N_b}$ amb l'estadística bayesiana. Entre els estimadors hem vist que es troba l'UMVUE, $\widehat{N_4}$.

A més hem trobat quatre intervals diferents, tres intervals de confiança i un interval de credibilitat. D'entre els intervals de confiança destaca el que hem anomenat Interval Freqüentista. La propietat que té aquest interval és que ens dóna un interval més estret degut a que és l'únic que el seu límit inferior és el màxim valor de la mostra. Els altres intervals consideren que pot haver-hi menys establiments que els que ens diu la propia mostra que tenim.
Aquesta propietat també la té l'interval de credibilitat fruit de l'anàlisi bayesiana.

Tot això ho hem pogut comprovar després de realitzar una simulació amb el software lliure R. La simulació l'hem fet generant una població de mida 5.000 de la qual n'hem extret una mostra de mida 100 i a la que li hem apli-

cat tots els estimadors i intervals trobats. Això ens ha permès comparar tots els resultats d'una manera objectiva i hem comprovat que com sospitàvem, l'interval freqüentista i l'interval de credibilitat són els més estrets, tot i que l'interval de credibilitat ho és una mica més que el freqüentista.
També hem provat que l'estimador $\widehat{N_4}$ és el més pròxim a la mida real de la població d'entre els estimadors que hem trobat mitjançant l'estadística clàssica i que ens dóna estimacions molt semblants a les que ens aporta $\widehat{N_b}$ (l'estimador bayesià).

Un altre dels objectius era aplicar els resultats a un cas pràctic. Ho hem fet utilitzant les dades d'una entitat financera per a estimar el nombre total d'establiments que hi ha en les diferents províncies de l'estat espanyol. Per fer-ho hem utilitzat l'estimador UMVUE i l'interval freqüentista que per la simulació hem vist que ens donaven les estimacions més ajustades.
Un cop estimat el nombre total d'establiments, els resultats es poden utilitzar per respondre preguntes tan comuns en el món de les finances com pot ser quina quota de mercat s'esta cobrint per tal de millorar o canviar les estratègies que s'estant utilitzant per fidelitzar clients en aquella entitat.

Això ens demostra que els resultats trobats en la Segona Guerra Mundial poden ser molt útilts avui en dia i que el problema dels tancs alemanys va ser resolt d'una manera molt enginyosa en uns moments en els que semblava més útil una arma que la tinta d'una ploma.

Bibliografia

[1] Ruggles, Richard and Brodie,Henry (1947) . **An Empirical Approach to Economic Intelligence in World War II.** *Journal of the American Statistical Association*, Vol. 42, No. 237, pp. 72-91. *American Statistical Association.*

[2] Touza Gil, Ramón (2012). **Los Panzer del mariscal Rommel y el Iphone de Paris Hilton.** *Revista general de marina*, Vol. 263, MES 4, 2012, pàgs. 687-693.

[3] Anònim. **German tank problem; Bayesian analysis.** *Wikipedia, the free encyclopedia.*
https://en.wikipedia.org/wiki/German_tank_problem

[4] Johnson, Roger W. (1994). **Estimating the Size of a Population.** *Teaching Statistics.*

[5] Anònim. **Análisis Bayesiano.** *Conselleria de Sanidade. Servizo Galego de Saúde. Xunta de Galicia.*
http://dxsp.sergas.es/ApliEdatos/Epidat/Ayuda/9-Ayuda%20An%E1lisis%20bayesiano.pdf

[6] Grima, Pere.(2010). **Quants taxis hi ha a Barcelona?.** *Departament d'Estadística i Investigació Operativa; Universitat Politècnica de Barcelona.*

[7] Hilbert A. David (1981). **Order Statistics.** *Wiley Series in Probability and Mathematical Statistics*, Second Edition, pàgs. 123-125.

[8] Anònim. **Binomial coefficient.** *Wikipedia, the free encyclopedia.* https://en.wikipedia.org/wiki/Binomial_coefficient

[9] Anònim. **Lehmann-Scheffe Theorem.** http://www.stat.ucdavis.edu/~yrsu/sta131b/handout_5.pdf

Apèndix A

Metadata

Títol: El problema dels tancs alemanys
Autora: Sílvia Prior Cayuela
Tutor: Xavier Bardina

Abstracts:

Català:
El problema dels tancs alemanys és un problema d'estimació de la mida d'una població quan aquesta està numerada correlativament. Aquest problema va ser resolt durant la Segona Guerra Mundial per part dels Aliats. En aquest projecte s'expliquen els càlculs fins arribar a l'estimador poblacional des d'una vessant freqüentista i també bayesiana. Se'n farà una simulació per comparar i verificar resultats i finalment es veurà un cas pràctic on es pugui aplicar aquest estimador en un problema actual.

Paraules clau:
Estimador de la mida d'una població, Semilongitud de l'Interval, Distribució Uniforme Discreta, Estadística bayesiana, Funció de Credibilitat, Interval de Credibilitat.

English:
The german tank problem is a problem of estimate the population size when the population is numbered consecutively. This problem was solved during the Second World War by the Allies. This project explains the calculations until the population estimatior from a frequentist and a bayesian point of view. In addition there is a simulation to compare and verify the results and finally we see a practical case where we can apply the estimator in a current problem.

Keywords:
Size population estimator, half length of interval, Discrete Uniform Distribution, Bayesian Statistics, Credibility Function, Credibility Interval.

Español:
El problema de los tanques alemanes es un problema de estimación del tamaño de una población cuando esta está numerada correlativamente. Este problema fue resulto durante la Segunda Guerra Mundial por parte de los Aliados. En este proyecto se explcan los calculos hasta llegar al estimador poblacional desde un punto de vista freqüentista y también bayesiana. Haremos una simulación para comparar y verificar resultados y finalmente se veurá un caso práctico donde se pueda aplicar este estimador en un problema actual.

Palabras clave:
Estimador del tamaño de la población, Semilongitud del Intervalo, Distribución Uniforme Discreta, Estadística Bayesiana, Función de Distribución de la Credibilidad, Intervalo de Credibilidad.

Apèndix B

Codi R per a la simulació

```
poblacio<-seq(1,5000,by=1)

head(poblacio)

length(poblacio)

n<-100

mostra<-sample(poblacio,n,replace=F)

head(mostra)


# ESTIMADORS

N1<-2*mean(mostra)-1

N2<-2*median(mostra)-1

N3<-max(mostra)+min(mostra)-1
```

```
N4<-((n+1)/n)*max(mostra)-1

Nb<-((max(mostra)-1)*(n-1))/(n-2)


# MARGE D'ERROR

var2<-(1/(3*n+1))*((2*mean(mostra)-1)^2-(n-1)*(2*mean(mostra)-1)-n)

marge2<-1.96*sqrt(var2)


var4<-(1/(n+1)^2)*((((n+1)/n)*max(mostra)-1)^2- (n-1)*(((n+1)/n)*max(mostra)-1)-n)

marge4<-1.96*sqrt(var4)


# INTERVALS DE CONFIANCA

## IC per a N2

minICN2 <-N2-marge2

maxICN2 <-N2+marge2

ICN2<-c(minICN2,maxICN2)

ICN2


## IC per a N4

minICN4 <- N4-marge4

maxICN4 <- N4+marge4
```

```
ICN4<-c(minICN4, maxICN4)

ICN4


## Interval Frequentista

minIF<-max(mostra)

maxIF<-(max(mostra)/(sqrt(0.05))^(1/n))

IF<-c(minIF,maxIF)

IF


semilongitudIF<-(maxIF-minIF)/2

semilongitudIF


# IINTERVAL DE CREDIBILITAT


m<-max(mostra)

suma<-0

while(suma<0.95)

{

k<-((m-1)/Nb)*((m-2)/(Nb-1))*((n-1)/(n-2))*((choose(m-3,n-3)/choose(Nb-2,n-2)))

suma<-suma+k
```

```
m<-m+1

}


ICredibilitat<-c(max(mostra),m)

ICredibilitat


semilongitudICred<-(m-max(mostra))/2

semilongitudICred
```

Apèndix C

Taula de resultats cas pràctic

Província	**Codi**	$X_{(n)}$	n	**Estimador Puntual** $(\widehat{N_4})$	**Semilongitud I.F.**	**Estimador Puntual Total** $(\widehat{N_4})$	**Semilongitud I.F. Total**
Àlaba	01	52.180	8	58.702	8.590	58.702	11.850
Albacete	02	56.713	23	59.178	2.712	59.178	3.945
Alacant	03	97.497	77	98.762	1.301	281.378	8.260
	53	99.447	31	102.654	3.441		
	54	79.121	94	79.962	861		
Almeria	04	79.456	28	82.293	3.068	82.293	4.486
Àvila	05	23.964	2	35.945	35.613	35.945	41.603
Badajoz	06	66.973	48	68.367	1.459	68.367	2.157
Illes Balears	07	98.992	26	102.798	4.141	196.199	10.102
	57	90.807	35	93.400	2.761		
Barcelona	08	90.792	22	94.918	4.559	955.456	41.893
	58	96.370	15	102.794	7.440		
	59	94.988	15	101.320	7.333		
	60	99.098	138	99.815	729		
	61	99.264	58	100.974	1.776		
	62	99.970	45	102.191	2.331		
	63	98.757	52	100.655	1.979		
	64	99.593	95	100.640	1.072		
	65	99.674	107	100.605	950		
	66	50.876	76	51.544	689		
Burgos	09	55.956	30	57.820	2.006	57.820	2.938
Càceres	10	45.859	28	47.496	1.771	47.496	2.589

Província	**Codi**	$X_{(n)}$	n	**Estimador Puntual** ($\widehat{N_4}$)	**Semilongitud I.F.**	**Estimador Puntual Total** ($\widehat{N_4}$)	**Semilongitud I.F. Total**
Cadis	11	96.572	57	98.265	1.759	124.832	4.054
	72	25.652	28	26.567	991		
Castelló	12	94.467	50	96.355	1.972	96.335	2.916
Ciutat Real	13	53.755	36	55.247	1.586	55.247	2.332
Còrdova	14	99.354	55	101.159	1.878	103.024	4.941
	56	1.244	2	1.865	1.849		
La Corunya	15	94.736	40	97.103	2.500	143.594	6.529
	70	44.703	25	46.490	1.952		
Conca	16	15.990	1	31.979	143.911	31.979	151.905
Girona	17	98.603	48	100.656	2.148	167.650	6.467
	55	64.901	31	66.994	2.246		
Granada	18	99.942	37	102.642	2.864	102.642	4.241
Guadalajara	19	57.020	14	61.092	4.767	61.092	6.802
Guipúscoa	20	98.902	19	104.106	5.843	113.135	10.951
	75	7.740	6	9.029	1.862		
Huelva	21	55.275	26	57.400	2.313	57.400	3.375
Osca	22	39.666	23	41.390	1.897	41.390	2.759
Jaén	23	74.435	40	76.295	1.964	76.295	2.894
Lleó	24	67.129	30	69.366	2.406	69.366	3.525
Lleida	25	79.358	47	81.045	1.768	81.045	2.611
La Rioja	26	50.605	8	56.930	8.330	56.930	11.493

Província	**Codi**	$X_{(n)}$	n	**Estimador Puntual** ($\widehat{N_4}$)	**Semilongitud I.F.**	**Estimador Puntual Total** ($\widehat{N_4}$)	**Semilongitud I.F. Total**
Lugo	27	79.936	21	83.741	4.226	83.741	6.128
Madrid	28	79.471	3	105.960	54.879	1.011.683	106.245
	78	97.468	14	104.429	8.147		
	79	83.160	14	89.099	6.951		
	80	98.914	58	100.618	1.769		
	81	91.183	304	91.482	302		
	82	96.028	36	98.694	2.833		
	83	99.276	136	100.005	741		
	84	95.043	48	97.022	2.071		
	85	97.893	88	99.004	1.139		
	86	98.991	93	100.054	1.089		
	87	24.164	21	25.314	1.278		
Màlaga	29	89.557	31	92.445	3.099	233.655	9.355
	92	98.568	40	101.031	2.601		
	93	39.542	62	40.179	660		
Murcia	30	86.808	44	88.780	2.072	174.889	5.263
	73	84.675	59	86.109	1.488		
Navarra	31	94.684	35	97.388	2.878	118.442	6.147
	71	19.885	17	21.054	1.332		
Ourense	32	43.516	6	50.768	10.464	50.768	14.089
Astúries	33	99.871	20	104.864	5.573	146.808	15.524
	74	37.751	9	41.945	5.358		
Palència	34	23.647	5	28.375	7.338	28.375	9.702

Província	**Codi**	$X_{(n)}$	n	**Estimador Puntual** ($\widehat{N_4}$)	**Semilongitud I.F.**	**Estimador Puntual Total** ($\widehat{N_4}$)	**Semilongitud I.F. Total**
Las Palmas	35	99.442	53	101.317	1.954	168.532	4.774
	76	65.994	54	67.215	1.272		
Pontevedra	36	99.728	59	101.417	1.753	120.144	91.555
	94	9.364	1	18.727	84.277		
Salamanca	37	49.894	13	53.731	4.547	53.731	6.465
Santa Cruz de Tenerife	38	99.781	57	101.531	1.817	101.531	2.692
Cantàbria	39	80.916	36	83.163	2.387	83.163	3.511
Segòvia	40	23.405	7	26.748	4.579	26.748	6.251
Sevilla	41	94.518	19	99.492	5.584	221.319	11.652
	90	19.534	27	20.256	785		
	91	99.985	63	101.571	1.642		
Sòria	42	18.476	1	36.951	166.285	36.951	175.522
Tarragona	43	95.192	33	98.076	3.081	98.076	4.523
Terol	44	24.342	10	26.775	3.035	26.775	4.251
Toledo	45	80.823	76	81.885	1.094	81.885	1.625
València	46	91.513	21	95.870	4.837	369.875	14.748
	96	99.423	46	101.583	2.265		
	97	98.185	89	99.287	1.129		
	98	71.394	41	73.134	1.836		
Valladolid	47	70.213	26	72.913	2.937	72.913	4.287

Província	**Codi**	$X_{(n)}$	n	**Estimador Puntual** ($\widehat{N_4}$)	**Semilongitud I.F.**	**Estimador Puntual Total** ($\widehat{N_4}$)	**Semilongitud I.F. Total**
Biscaia	48	99.540	17	105.394	6.664	184.828	13.370
	95	77.028	32	79.434	2.577		
Zamora	49	20.727	7	23.687	4.055	23.687	5.535
Saragossa	50	96.839	47	98.898	2.157	143.456	6.615
	99	42.437	20	44.558	2.368		
Ceuta	51	3.059	2	4.546	4.588	4.546	5.311
Melilla	52	50.037	1	100.073	450.334	100.073	475.352

Taula C.1: *Resultats de l'anàlisi cas pràctic*