

Universitatea Națională de Știință și Tehnologie
POLITEHNICA București
Școala Doctorală a Facultății de Științe Aplicate
Domeniul: Matematică

CONTRIBUȚII LA INFERENȚĂ
ÎN TEORIA SONDAJULUI

TEZĂ DE ABILITARE
REZUMAT

Conf. univ. dr. MARIUS ȘTEFAN
Facultatea de Științe Aplicate
Universitatea Națională de Știință și Tehnologie
POLITEHNICA București

Universitatea Națională de Știință și Tehnologie
POLITEHNICA București
Școala Doctorală a Facultății de Științe Aplicate
Domeniul: Matematică

CONTRIBUȚII LA INFERENȚĂ
ÎN TEORIA SONDAJULUI

TEZĂ DE ABILITARE
REZUMAT

Conf. univ. dr. MARIUS ȘTEFAN
Facultatea de Științe Aplicate
Universitatea Națională de Știință și Tehnologie
POLITEHNICA București

Rezumat

Nevoia de informație statistică este mare în societatea modernă. Date statistice sunt culese cu regularitate din mulțimi de elemente denumite *populații finite*. Unul din modurile de colectare este sondajul statistic care reprezintă o cercetare parțială, prin eșantion s , a unei populații finite U , unde $s \subset U$. Dacă un recensământ presupune cercetarea tuturor unităților care compun U , la sondajul statistic informația este culeasă doar pentru unitățile din eșantionul s . Din acest punct de vedere, sondajul statistic bazat pe eșantion prezintă avantajul de a fi mai eficient din punct de vedere cost și timp în comparație cu un recensământ. Pe de altă parte, estimațiile obținute pe baza lui s prezintă erori de eșantionaj, în sensul că nu sunt egale cu parametrii pe care îi estimează. În multe țări, Institutele Naționale de Statistică (INS) sunt însărcinate prin lege să furnizeze informație statistică privind starea națiunii, iar sondajul statistic reprezintă o parte importantă a activității acestora.

În cazul unui sondaj statistic, există una sau mai multe variabile de studiu (sau interes). În această teză, considerăm cazul când avem o singură variabilă de studiu, notată y . Pentru variabila y , fiecare unitate k din U are atașată o valoare, notată y_k . Un parametru de interes θ asociat populației U este o funcție care depinde de valorile $y_k, k \in U$. Exemple de astfel de parametri sunt: venitul mediu (sau total) al populației, randamentul mediu (sau total), proporția (sau numărul) de șomeri, etc... Cunoașterea cu exactitate a lui θ presupune realizarea unui recensământ în urma căruia se obțin toate valorile $y_k, k \in U$. Când un recensământ nu este posibil sau este considerat prea scump, se recurge la sondajul statistic. În acest caz, θ este necunoscut iar scopul sondajului este estimarea lui θ cu un estimator $\hat{\theta}$ care este funcție de datele din eșantionul s . În capitolele tezei facem presupunerea că toate elementele din s răspund.

Parametrul θ poate fi asociat cu U (parametru național), sau poate fi asociat unei subpopulații D (domeniu), unde $D \subset U$. În capitolele 1, 2, 3 și 4 ale tezei, ne ocupăm de estimarea mediei unei populații. În capitolul 5 ne ocupăm de estimarea mediei unui domeniu. Capitolul 6 conține detalii privind cercetarea prezentă și viitoare.

În capitolele 1, 2, 3 și 4, inferența este bazată pe *planul de sondaj*, în timp ce inferența din capitolul 5 este bazată pe un model statistic. Când inferența este bazată pe planul de

sondaj, s este un *eșantion probabilist*. La eșantionarea probabilistă, fiecare s are o probabilitate $p(s)$ de a fi selecționat. Pentru selectarea lui s se folosește planul de sondaj, un mecanism aleator conform caruia s este ales cu probabilitatea $p(s)$. Funcția $p(s)$ definește o distribuție de probabilitate pe mulțimea tuturor eșantioanelor posibile numită *distribuție de eșantionaj*, iar proprietățile unui estimator sunt evaluate în raport cu distribuția de eșantionaj. Pe de altă parte, când inferența este bazată pe un model statistic precum în capitolul 5, estimatorul $\hat{\theta}$ este obținut cu modelul statistic, iar proprietățile sale sunt evaluate în raport cu distribuția postulată de model. Astfel, validitatea inferenței depinde de cât de potrivit este modelul la datele de lucru.

Sub inferența bazată pe planul de sondaj, probabilitățile $p(s)$ definesc, pentru fiecare unitate k din U , probabilitatea de incluziune de ordin unu a lui k în eșantion. Notată π_k , ea este dată de $\pi_k = \sum_{s \ni k} p(s)$, unde $s \ni k$ indică că suma se face după toate eșantioanele s care îl conțin pe k . Necesare sunt și probabilitățile de incluziune de ordin doi care, pentru o pereche de unități (k, l) , reprezintă probabilitatea ca acestea să se găsească simultan în s . Notată π_{kl} , probabilitatea de incluziune a lui (k, l) este dată de $\pi_{kl} = \sum_{s \ni k, l} p(s)$, unde $s \ni k, l$ indică că suma se face după toate eșantioanele s care conțin simultan k și l . Un estimator $\hat{\theta}$ al lui θ se construiește cu probabilitățile de incluziune $\pi_k, k \in s$, iar π_{kl} se utilizează pentru a estima varianța lui $\hat{\theta}$. Spre exemplu, dacă θ este totalul populației, adică $\theta = \sum_{k \in U} y_k$, θ poate fi estimat nedeplasat sub planul de sondaj cu $\hat{\theta}^{HT}$, estimatorul său Horvitz-Thompson (HT), unde $\hat{\theta}^{HT} = \sum_{k \in s} \pi_k^{-1} y_k$. Se poate observa că $\hat{\theta}^{HT}$ este liniar, egal cu suma ponderată a valorilor observate y_k , unde ponderea de sondaj $d_k = \pi_k^{-1}$ este inversul probabilității de incluziune, pentru $k \in s$. Un estimator nedeplasat pentru $V(\hat{\theta}^{HT})$ este dat de $\hat{V}(\hat{\theta}^{HT}) = \sum_{k \in s} \sum_{l \in s} \pi_{kl}^{-1} \Delta_{kl} \pi_k^{-1} y_k \pi_l^{-1} y_l$, unde $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$, pentru $k, l \in s$. Observăm că estimatorul varianței $\hat{V}(\hat{\theta}^{HT})$ este exprimat ca sumă dublă pe s .

Variabilele auxiliare sunt utile deoarece pot ameliora precizia unui estimator. În general, o variabilă auxiliară x este orice variabilă cu valori cunoscute înainte de eșantionare pentru fiecare unitate k din U . Cel mai adesea, se dispune de un vector $\mathbf{x}$ compus din mai multe

variabile auxiliare. Pentru a crește precizia unui estimator, vectorul $\mathbf{x}$ poate fi folosit în două moduri: i. în estimare, pentru a construi formula estimatorului; ii. în eșantionare, pentru a construi planul de sondaj. Capitolele 1, 2, 3 și 5 se ocupă cu utilizarea lui $\mathbf{x}$ în estimare. Pe de altă parte, în capitolul 4 este studiată utilizarea lui $\mathbf{x}$ în eșantionare.

În estimare (punctul i. de mai sus), vectorul $\mathbf{x}$ este introdus în mod explicit în formula estimatorului $\hat{\theta}$. De regulă, când $\mathbf{x}$ este corelat cu variabila de interes y , un estimator care folosește $\mathbf{x}$ în formulă are o varianță mai mică, și deci o precizie mai bună, în comparație cu un estimator care nu folosește $\mathbf{x}$, așa cum este $\hat{\theta}^{HT}$. Un estimator, folosit de multe Institute Naționale de Statistica, și care exploatează eficient corelația dintre $\mathbf{x}$ și y , este estimatorul de regresie generalizat (GREG – *generalized regression estimator*).

Estimatorul GREG este construit cu ajutorul unui model de regresie liniară postulat între variabila y și vectorul $\mathbf{x}$. Ca și estimatorul HT, GREG poate fi exprimat sub forma unei sume ponderate, $\hat{\theta}^G = \sum_{k \in S} w_k y_k$, unde ponderile de sondaj w_k sunt calibrate pe vectorul de variabile auxiliare $\mathbf{x}$, adică $\sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k$. Inferența relativă la GREG este asistată de model, acest lucru însemnând că modelul de regresie servește doar la construcția estimatorului, iar proprietățile estimatorului, precum media sau varianța, sunt calculate în raport cu planul de sondaj. Astfel, valabilitatea concluziilor nu depinde de cât de bine este adaptat modelul la datele din eșantion. Totuși, calitatea modelului influențează eficiența lui GREG: cu cât modelul este mai bine adaptat datelor, cu atât GREG este mai eficient.

GREG este un estimator consistent și asimptotic nedeplasat. În consecință, când efectivul eșantionului este mare, estimările furnizate de GREG sunt precise. Dar când numărul de observații este mic, GREG poate avea o deplasare mare, mai ales dacă datele nu verifică modelul liniar cu care GREG este construit. Când estimatorul este deplasat, o practică comună constă în a corecta deplasarea cu ajutorul unui estimator al deplasării. Spre exemplu, acest estimator al deplasării se poate obține cu ajutorul *jackknife*-ului, o tehnică de reeșantionare binecunoscută. Obținerea unui estimator al deplasării lui GREG cu metoda jackknife se face folosind variabilitatea estimatorului GREG recalculat în subeșantioane ale eșantionului inițial obținute prin eliminarea pe rând a câte unei unități a eșantionului inițial. Estimatorul cu deplasare corectată (redușă), notat GREG-JK, denumit și versiunea jackknife a lui GREG, se obține scăzând estimarea deplasării din estimatorul inițial.

Capitolul 1 se ocupă cu estimarea deplasării lui GREG folosind metoda jackknife. GREG este funcție de estimatori HT. De aceea, obținerea lui GREG-JK este imediată folosind HT-JK, versiunea jackknife a estimatorului HT. Jackkniful din teoria sondajului este modificat față de versiunea sa originală, deoarece, pentru a avea sens, HT-JK trebuie să coincidă cu HT dat fiind că HT este nedeplasat. Această proprietate a lui HT-JK reprezintă justificarea pentru utilizarea lui $\hat{\theta} - JK$, versiunea corectată prin jackknife a lui $\hat{\theta}$, în cazul oricarui estimator $\hat{\theta}$ care este funcție de estimatori HT. În plus față de această justificare, capitolul 1 furnizează pentru GREG-JK și o justificare teoretică. În capitolul 1 demonstrăm că ordinul deplasării asimptotice a lui GREG-JK este mai mic decât cel al deplasării asimptotice a lui GREG. Aceasta înseamnă că deplasarea lui GREG-JK tinde mai repede la zero decât cea a lui GREG. În plus, simulațiile arată că deplasarea lui GREG-JK este neglijabilă când efectivul eșantionului este mic.

Eficiența unui estimator se măsoară cu varianța sa dacă estimatorul este aproximativ nedeplasat. Varianța lui GREG se poate obține pe baza tehnicii de dezvoltare în serie Taylor. Estimatorii varianței obținuți cu tehnica Taylor sunt asimptotic nedeplasați, însă pe eșantioane mici ei pot avea o deplasare mare. Mai precis, este bine cunoscut că deplasarea lor este negativă când numărul de observații este mic. Originalitatea capitolului 2 este că, în loc să aplicăm jackkniful pe GREG precum în capitolul 1, propunem să aplicăm jackkniful pe un estimator Taylor al varianței pentru a estima și reduce deplasarea estimatorului varianței în eșantioane mici.

Cum se poate vedea mai sus în cazul lui $\hat{V}(\hat{\theta}^{HT})$, un estimator al varianței este de regulă exprimat ca sumă dublă pe s . Pe de altă parte, jackkniful este introdus și definit pentru estimatori care se exprimă ca sumă simplă pe s . De aceea, a-priori nu este foarte clar cum se poate aplica jackkniful pe un estimator al varianței. În capitolul 2, arătăm că jackkniful poate fi aplicat sub forma sa cunoscută folosind o populație și un eșantion sintetice. Populația sintetică este formată din toate perechile (k, l) cu $k, l \in U$, iar eșantionul sintetic este format din toate perechile (k, l) cu $k, l \in s$. Această tehnică permite scrierea unui estimator al varianței sub forma unei sume simple pe eșantionul sintetic. În plus, un plan de sondaj sintetic poate fi definit conform căruia perechea (k, l) este selecționată cu probabilitatea π_{kl} . Artificiul bazat pe populația și eșantionul sintetic a fost deja folosit în literatură pentru a obține

estimatori calibrați pentru parametrii cuadratici precum varianța/covarianța unei populații, sau varianța unui estimator liniar. În capitolul 2, folosim această tehnică pentru a obține versiunea jackknife a unui estimator al varianței lui GREG.

În capitolul 3, sub un plan de sondaj oarecare, propunem o procedură de estimare a varianței lui GREG bazată pe *bootstrap*, o altă tehnică bine cunoscută de reeșantionare care poate fi folosită pentru estimarea deplasării și estimarea varianței. Ca și în cazul jackknife-ului, forma originală a bootstrapului propusă în afara teoriei sondajului nu poate fi aplicată identic în teoria sondajului. Pentru a da rezultate bune, practicienii teoriei sondajului au modificat forma inițială a bootstrapului ceea ce a condus la mai multe tipuri de bootstrap.

Tipul de bootstrap pe care îl folosim în capitolul 3 este cel bazat pe ideea de pseudo-populație U^* . După construcția lui U^* , un mare număr de eșantioane, s^* , se selecționează din U^* folosind planul de sondaj inițial $p(\cdot)$. Pe s^* se calculează un estimator $\hat{\theta}^*$ care are aceeași formulă ca și $\hat{\theta}$. Varianța lui $\hat{\theta}$ se estimează folosind variabilitatea valorilor $\hat{\theta}^*$, dat fiind că un mare număr de $\hat{\theta}^*$ sunt disponibile. Posibilitatea de a capta în mod natural variabilitatea unui estimator în raport cu planul $p(\cdot)$, prin reeșantionări repetate din U^* folosind $p(\cdot)$, a convins mulți cercetători să studieze bootstrapul bazat pe pseudo-populație. În literatura, metodele existente pentru a construi U^* prezintă dezavantaje, iar acestea au complicat algoritmi de construcție ai lui U^* .

Metoda propusă în capitolul 3 pentru a construi U^* este originală. Ea evită aceste probleme prin folosirea modelului de regresie liniară folosit pentru a obține estimatorul GREG. Sub condiții de regularitate frecvent folosite în teoria sondajului demonstrăm că estimatorul propus pentru varianța lui GREG este consistent și asimptotic nedeplasat. Aceste rezultate sunt stabilite sub distribuția comună dată de planul de sondaj și procesul aleator de construcție a lui U^* . De aceea, estimatorul bootstrap al varianței propus în capitolul 3 este asistat de model, ca și estimatorul GREG. În plus, simulațiile Monte Carlo, realizate atât cu date artificiale cât și cu date reale, arată că estimatorul varianței bazat pe bootstrap are o deplasare neglijabilă în eșantioane cu puține observații.

Am specificat mai sus (punctul ii.) că o variabilă auxiliară, x , poate fi folosită și în eșantionare, pentru a crea un plan de sondaj care să crească precizia estimatorului HT. Spre exemplu, acest lucru se poate realiza prin *eșantionare echilibrată*. Modul de lucru este

urmatorul. Se selectează un eşantion s folosind un plan de sondaj inițial denumit *plan de sondaj bazic*. Eşantionul ales s este păstrat numai dacă verifică *ecuațiile de echilibrare*, $\sum_{k \in s} \pi_k^{-1} \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k$, adică dacă estimatorii HT asociați variabilelor auxiliare din $\mathbf{x}$ sunt egali cu totalurile (cunoscute) ale acestor variabile. Pentru că este foarte probabil să nu existe nici un s care să verifice exact ecuațiile de echilibrare, în realitate, un eşantion echilibrat va verifica aproximativ aceste ecuații. Logica eşantionării echilibrate este următoarea: dacă există o relație strânsă între $\mathbf{x}$ și y , estimatorul HT asociat variabilei de studiu și calculat pe un eşantion echilibrat va furniza o valoare apropiată de totalul lui y , iar aceasta înseamnă o estimare precisă.

Capitolul 4 al tezei se ocupă cu *planul de sondaj rejectiv*, un plan propus în literatură pentru a selecționa eşantioane echilibrate. Conform acestei proceduri, distanța dintre estimatorii HT și totalurile variabilelor auxiliare din $\mathbf{x}$ este controlată de o distanță, Q , și de o constantă γ^2 , aleasă convenabil. Conform procedurii, se calculează valoarea lui Q asociată eşantionului selecționat cu planul de sondaj bazic. Dacă Q este mai mare decât γ^2 , eşantionul s este respins și un alt eşantion este selecționat. Procedura se oprește și eşantionul este reținut când Q este mai mic decât γ^2 . Prin valoarea lui γ^2 se controlează în ce măsură ecuațiile de echilibrare sunt verificate.

În capitolul 4 arătăm că un estimator $\hat{\theta}$, care este nedeplasat sub planul de sondaj bazic, este nedeplasat și sub planul de sondaj rejectiv dacă vectorul compus din $\hat{\theta}$ și estimatorii HT ai variabilelor $\mathbf{x}$ are o distribuție normală multivariată sub planul bazic. Sub această condiție de normalitate, obținem o formulă exactă pentru $V_r(\hat{\theta})$, varianța rejectivă a lui $\hat{\theta}$. Această formulă depinde de constanta γ^2 , dimensiunea p a vectorului $\mathbf{x}$, și de elemente ale planului bazic inițial. Pe baza formulei lui $V_r(\hat{\theta})$ construim un estimator $\hat{V}_r(\hat{\theta})$ al lui $V_r(\hat{\theta})$. Ipoteza de normalitate este de regula verificată pe esantioane mari. Așadar, rezultatele teoretice din capitolul 4 arată că $\hat{\theta}$ și $\hat{V}_r(\hat{\theta})$ sunt asymptotic nedeplasați sub planul de sondaj rejectiv.

Am menționat mai sus că sondajul statistic este folosit pentru a estima nu doar parametrii la nivel național ci și parametrii relativi la domenii. Domeniile pot fi arii geografice sau grupuri socio-demografice. Teoria clasică a sondajului bazată pe planul de sondaj poate fi aplicată pentru a obține estimări pentru domenii dacă se consideră o variabilă artificială de

studiu, y_d , asociată domeniului de interes. Variabila y_d se definește astfel: y_{dk} este egal cu y_k pentru o unitate k din domeniu, și este egal cu zero pentru k din afara domeniului. Estimatorii domeniului care rezultă când teoria sondajului bazată pe planul de sondaj este aplicată lui y_d se numesc *estimatori direcți*. Un estimator direct este mai puțin precis deoarece el folosește doar observațiile din eșantionul național care provin din domeniu. Un domeniu este considerat domeniu mare dacă numărul acestor observații este astfel încât estimatorii direcți au un nivel suficient de precizie. În caz contrar, domeniul este considerat *domeniu mic* sau *arie mică*.

În cazul ariilor mici, pentru a obține estimatori cu precizie suficient de bună, este necesară utilizarea de *estimatori indirecti*. Estimatorii indirecti se obțin pe baza unui model statistic. Pentru că modelul acționează ca o legătură între toate ariile mici ale populației, un estimator indirect folosește toate observațiile eșantionului național, nu doar pe cele din aria mică de interes. În consecință, un estimator indirect va avea o precizie mai bună decât un estimator direct. Însă, inferența legată de estimatorii indirecti nu este asistată de model, ca în cazul lui GREG, ci este bazată pe model. Aceasta înseamnă că proprietățile estimatorilor indirecti se calculează în raport cu distribuția postulată de model, iar valabilitatea rezultatelor depinde de cât de bine este modelul adaptat datelor din eșantion.

Capitolul 5 se ocupă de estimarea mediei unei arii mici cu ajutorul modelului de arie mică la nivel de unitate statistică. Pe baza acestui model, se obțin estimatori indirecti pentru toate ariile mici care compun o populație națională. Când acești estimatori sunt agregați, se obține un estimator la nivel național. Pe de altă parte, la nivel național, cu teoria sondajului bazată pe planul de sondaj se poate obține un estimator direct de bună precizie. Astfel, avem doi estimatori la nivel național care provin din două surse diferite. Este de preferat ca aceștia să coincidă, mai ales dacă al doilea este considerat standardul de aur. Egalitatea dintre cei doi este denumită *ecuație benchmark*. Pentru ca ecuația benchmark să fie satisfăcută este necesar ca estimatorii indirecti inițiali să fie ajustați. Estimatorii indirecti ajustați se numesc *estimatori benchmarkuiți* deoarece ei verifică ecuația benchmark.

Pentru a obține estimatori benchmarkuiți, ponderile de sondaj cu care este construit estimatorul direct al parametrului național trebuie introduse în formula estimatorilor indirecti. Tehnicile standard de încorporare a ponderilor de sondaj nu conduc în mod necesar la estimatori benchmarkuiți. Capitolul 5 al tezei investighează cum aceste tehnici pot fi folosite

astfel încât estimatorii care rezultă să verifice ecuația benchmark. Evaluăm estimatorii benchmarkuiți în simulații care folosesc date artificiale și date reale. Rezultatele acestor studii arată că benchmarkuirea estimatorilor oferă protecție contra riscului de a lucra cu un model incorect.