Paolo Bonanomi
Sukhraj Singh Ghuman
Romeo Silvestri
Andrea Succi

TEORIA E TECNICA DELL'INDAGINE STATISTICA E DEL CAMPIONAMENTO
PROVA LABORATORIALE NUMERO 17 (EXIT POLL) – 2018-19

INTRODUZIONE

In questo documento viene esposto il piano di campionamento relativo ad una rilevazione di tipo exit-poll in occasione delle elezioni europee che si svolgeranno in data 26 maggio 2019 in Italia.
Esso si basa su un campione composto da un determinato numero di persone votanti coerente con la precisione che si vuole ottenere sui risultati elettorali.

La rilevazione viene gestita da un insieme di intervistatori posizionati fuori dal seggio elettorale a debita distanza come stabilito da norma di legge per non interferire con le operazioni di voto.
In ciascun seggio l'individuazione dei votanti avviene con criterio sistematico.

Ad essi viene richiesta la compilazione anonima di un breve questionario composto da una sola pagina fronte e retro che verrà in seguito posto all'interno di un'urna svuotata regolarmente ad intervalli di 3 ore ai fini di comunicare i risultati all'istituto di sondaggi addetto all'analisi dei dati.

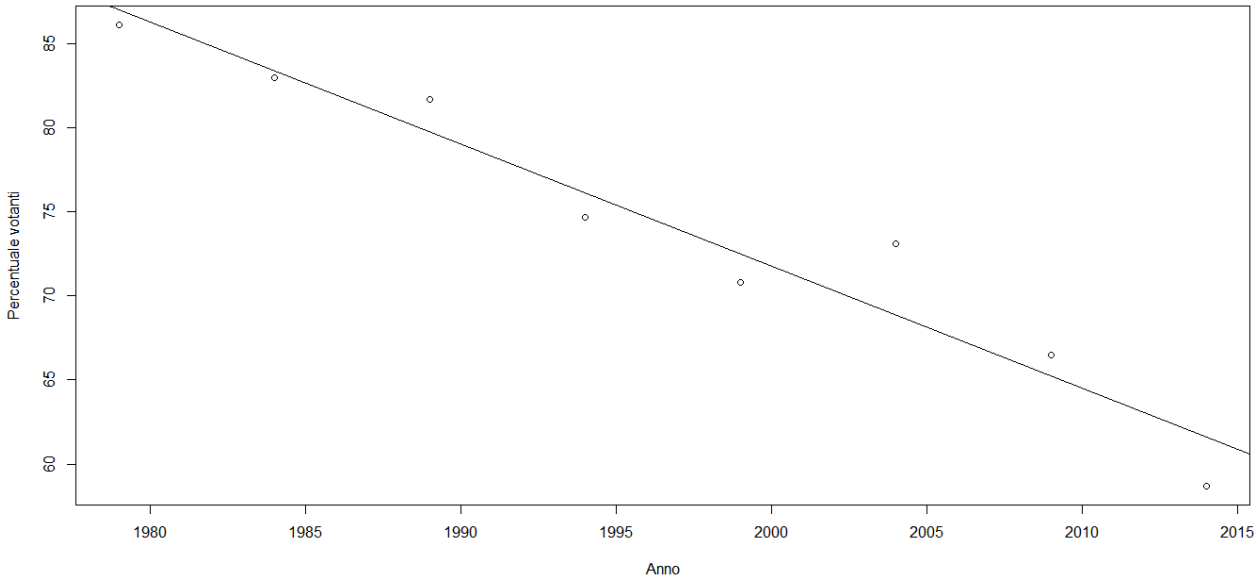
Il fronte del questionario somministrato risulta analogo alla scheda elettorale: sulla prima facciata infatti vengono riprodotti i simboli ufficiali dei partiti eleggibili per le elezioni europee con lo stesso ordine con cui si trovano nella scheda corrispondente al determinato seggio, sulla seconda invece sono presenti dei dati ascrittivi generici del votante (genere, età, titolo di studio e professione svolta) e le sue abitudini legate al voto (giornale e telegiornale più seguito, momento in cui ha deciso di votare e voto alle precedenti elezioni).
L'obiettivo principale è quello di ottenere un'indicazione anticipata su come si siano svolte le elezioni, in particolar modo riguardo alla previsione dei risultati elettorali espressi in percentuale.
In secondo luogo i dati ricavati potranno essere successivamente utilizzati per osservare i flussi elettorali e le caratteristiche dell'elettorato.

PIANO DI CAMPIONAMENTO

Non è possibile determinare la numerosità della popolazione di interesse, in quanto è impossibile conoscere in anticipo la percentuale di elettori che andranno a votare.

Si ritiene che utilizzare come numerosità della popolazione il numero di aventi diritto al voto sia scorretto perché si conosce a priori che una elevata percentuale di essi non andrà a votare.

Per questo motivo si è deciso di applicare un modello di regressione lineare utilizzando lo storico dell'affluenza alle elezioni europee per stimare la numerosità delle elezioni oggetto di indagine.



Dal modello lineare risulta che nel 2019 i votanti previsti saranno 28.603.123 (58,01714% dei votanti).

Si veda l'allegato corrispondente con i comandi in R

STRATIFICAZIONE

Si decide di stratificare per area geografica.

Questa metodologia consente di avere una maggiore precisione nel controllo dei dati e nella loro ponderazione, nonché nel gestire la trasmissione dei dati all'istituto di rilevazione.

In particolare si sceglie di formare 10 zone (agglomerati di province).

Considerando le affluenze delle singole province e la loro prossimità geografica, ogni zona conterrà all'incirca il 10% dei votanti rispetto al totale degli elettori alle scorse elezioni europee (lo strato con la percentuale più elevata avrà il 10.13% del totale mentre quello con la percentuale più bassa il 9.82%, portando la differenza percentuale massima tra due strati a 0.31%).

Questa metodologia consente di avere la massima autoponderazione raggiungibile.

Decidiamo di utilizzare le affluenze del 2014 perché è sensato aspettarsi che l'affluenza sia circa simile tra due elezioni successive.
Se l'istituto di rilevazione noterà distorsioni significative sull'affluenza rispetto a quella considerata, abbandonerà l'idea di autoponderazione e deciderà di utilizzare dei pesi consoni alle necessità.

CAMPIONAMENTO PER STADI:

Dovendo affrontare una struttura di campionamento definita su più liste delle unità della popolazione ad impostazione gerarchica, si decide di utilizzare un disegno di campionamento su tre stadi.

La probabilità di selezione nei primi due stadi è variabile, essendo fissata in funzione della dimensione delle unità al fine di favorire l'entrata nel campione di quelle aventi dimensione maggiore.
Nell'ultimo stadio si avrà una probabilità di selezione costante.

PRIMO STADIO:

La lista da cui attingeremo le informazioni per il campionamento sul primo stadio si trova sul sito del dipartimento degli affari interni e territoriali nella sezione elezioni/ archivio/Europee 2014.

Si sceglie di campionare per province e non per comuni al fine di evitare di poter poi selezionare al secondo stadio un solo seggio per comune (in quanto ci sono comuni con un unico seggio).

Questa tecnica permette di avere maggiore rappresentatività all'interno delle grandi città, potendo selezionare al secondo stadio due seggi per provincia.

Si procede quindi con un campionamento al primo stadio a probabilità variabile, pesando ogni provincia proporzionalmente all'affluenza alle ultime elezioni europee.

Si utilizza un campionamento a serpentina entro gli strati: ogni strato sarà ordinato in modo da consentire la contiguità geografica, così da considerare le diverse aree culturali dello strato.
Non c'è continuità tra le serpentine dei vari strati, in quanto si andrebbe a complicare il processo di campionamento aumentando il rischio di commettere errori.

Avendo un ammontare di province pari a 107, se ne campionano 4 per strato in modo da arrivare ad un totale di 40.

Nonostante vi sia un numero differente di province per strato, se ne campiona per ciascuno lo stesso numero, in quanto ogni strato ha un numero di votanti alle Europee 2014, circa uguale (vedere Stratificazione)

SECONDO STADIO:

Si campionano 10 seggi per provincia pesandoli sempre a seconda del numero di votanti per seggio alle scorse elezioni.

Sarà quindi necessario trovare il numero di votanti per ogni seggio delle province campionate.

Si continua ad utilizzare anche al secondo stadio un campionamento a serpentina per ogni provincia, in modo da mantenere la continuità geografica e culturale.

Ogni singolo seggio è un insieme di sezioni elettorali da cui possiamo recuperare i dati di cui necessitiamo consultando i siti web dei vari comuni.

Si campiona con queste modalità poiché non è possibile distinguere le sezioni all'interno di ogni singolo seggio, essendo il rilevatore obbligato a mantenersi a distanza per norma di legge dal luogo fisico in cui è collocata la sezione elettorale.
Si è deciso quindi di stimare i seggi presenti in Italia e in base al numero di sezioni selezionato si è trovato quanti seggi rilevare (vedere sezione numerosità campionaria).

TERZO STADIO:

Il rilevatore posizionato a debita distanza dal seggio elettorale, attrezzato con un'urna trasparente con delle schede bianche inserite (per evitare che l'intervistato

possa credere che il suo voto sia riconoscibile), fermerà coloro che hanno appena espresso il voto per sottoporre loro il questionario.

Verrà fermato un votante ogni 8.
Se l'elettore non collabora il rilevatore intervisterà la prima persona disponibile e riprenderà con la selezione sistematica a partire da questo intervistato.

AUTOPONDERAZIONE:

Non è possibile raggiungere la completa autoponderazione in quanto il nostro campione risulterà contaminato a seguito del problema delle mancate collaborazioni (che sarà trattata in seguito).

Ci si può però avvicinare all'autoponderazione degli strati ricorrendo ad una metodologia che formi degli strati con uguale numerosità (consultare la parte relativa alla stratificazione).

Per avere l'autoponderazione bisognerà stimare utilizzando le percentuali rilevate in ogni singolo seggio perché aggregando tutti i dati indistintamente si peserebbe 2 volte per il peso del seggio.

NUMEROSITA' CAMPIONARIA

Osservando lo storico degli studi degli exit-poll e considerando la variabilità richiesta, viene utile utilizzare la formula del campionamento casuale semplice per determinare l'n ottimale in quanto molto aderente alla realtà. Inoltre i coefficienti di correlazione intraclasse ci fanno intendere che la varianza di stima sarà circa uguale a quella di un campione casuale semplice.

Decidiamo di calcolare la numerosità in funzione delle sezioni e non delle persone votanti in quanto si vuole ottimizzare l'utilizzo dei rilevatori.

Considerando che la situazione con varianza maggiore sarà per un partito del circa 30%, e volendo trovare la numerosità necessaria per mantenere una variazione massima del +-2%, si è trovata una numerosità necessaria.

$\frac{S_y^2 k_{a/2}^2}{D^2}$=0.3*0.7*1.96^2/0.02^2.

Il risultato del calcolo è 2016 sezioni. Stimando un totale di circa 12000 seggi (sono presenti in Italia circa 61000 sezioni elettorali) andranno rilevati 400 seggi. Ovvero 10 seggi per provincia campionata al primo stadio.

Se in un determinato seggio il rilevatore notasse che la maggioranza dei votanti decide di non sottoporsi al questionario, l'istituto di rilevazione deciderà di pesarlo

di meno a favore di un seggio storicamente simile(ragionando anche per prossimità geografica e culturale).

MANCATE COLLABORAZIONI

Un grave problema delle rilevazioni exit poll risulta essere la mancata collaborazione di una elevata percentuale dei votanti intervistati.

Per ovviare a questa problematica si è deciso di utilizzare i dati ascrittivi e le abitudini legate al voto dei rispondenti.
Nello specifico si andrà a confrontare la media di questi dati con quella reale, per esempio nel caso dei telegiornali si misurerà la differenza tra share osservato e share reale.

Basandosi su queste differenze si andranno a rendere più importanti le unità statistiche del campione che hanno le caratteristiche più sottostimate. Per esempio se si noterà che a uno dei partiti verranno attribuite percentuali alle scorse elezioni inferiori rispetto a quanto effettivamente accaduto, si andrà a dare più peso alle unità presenti nel campione che hanno dichiarato di aver votato il suddetto partito alle scorse elezioni.

Questo tipo di ponderazione però va incrociato tra le varie variabili, considerando solo distorsioni considerevoli, in quanto sappiamo che al conto reale mancano le persone che non sono andate a votare.

La difficolta di questo metodo consiste nel riuscire a distinguere fra partiti effettivamente sottostimati e partiti i cui potenziali elettori non si sono recati alle urne.

Per cercare di risolvere questa complicazione si vuole creare un focus group con l'intenzione di capire le abitudini dei non votanti (telegiornali più seguiti, partiti più votati alle scorse elezioni etc..) in modo da non andare a correggere erroneamente le stime.

Una problematica costante che dovrà essere affrontata nella stessa maniera consisterà nel ripesare in base al sesso e all'età, in quanto le donne si fermeranno agli exit poll con frequenza nettamente minore rispetto agli uomini e gli anziani si fermeranno con minore frequenza rispetto ad altre fasce di età.

BIBLIOGRAFIA E SITOGRAFIA

Fabbris, Luigi, L'indagine campionaria. Metodi, disegni e tecniche di campionamento. Roma: NIS, 1989

Lohr, Sharon L., Sampling design and analysis. Boston: Brooks/Cole, 2010

Dipartimento per gli Affari Interni e Territoriali/Elezioni/Archivio storico delle elezioni: https://elezionistorico.interno.gov.it

https://www.linkiesta.it/it/article/2016/06/03/elezioni-ecco-perche-i-sondaggi-sbagliano-ma-spesso-ci-beccano/30620/

Strati:

Zona 1: Sassari, Nuoro, Oristano, Sud Sardegna, Cagliari, Trapani, Palermo, Agrigento, Caltanissetta, Enna, Messina, Catania, Siracusa, Ragusa, Reggio Calabria, Vibo Valentia, Catanzaro, Crotone

Totale Votanti Elezioni 2014 = 2.916.673(10.09% del totale dei votanti)

Zona 2: Cosenza, Taranto, Brindisi, Lecce, Potenza, Matera, Bari, Salerno, Barletta-Adria-Trani, Foggia

Totale Votanti Elezioni 2014 = 2.888.745 (9.99% del totale dei votanti)

Zona 3: Napoli, Avellino, Benevento, Caserta, Isernia, Campobasso, Frosinone, Latina, Chieti

Totale Votanti Elezioni 2014 =2.839.041(9.82% del totale dei votanti)

Zona 4: L'Aquila, Pescara, Teramo, Roma, Rieti, Viterbo, Terni, Ascoli Piceno

Totale Votanti Elezioni 2014 = 2.909.500(10.06% del totale dei votanti)

Zona 5: Perugia, Rimini, Forlì Cesena, Fermo, Macerata, Ancona, Pesaro-Urbino, Arezzo, Siena, Grosseto, Livorno, Pisa, Pistoia, Prato, Lucca, Massa-Carrara

Totale Votanti Elezioni 2014 = 2.904.581 (10.05% del totale dei votanti)

Zona 6: Firenze, Bologna, Ravenna, Modena, Reggio Emilia, Parma, Piacenza, Rovigo, La Spezia, Lodi

Totale Votanti Elezioni 2014 = 2.900.997(10.03% del totale dei votanti)

Zona 7: Genova, Savona, Imperia, Cuneo, Asti, Alessandria, Aosta, Torino, Biella, Verbano-Cusio-Ossola

Totale Votanti Elezioni 2014 = 2.865.626 (9.91% del totale dei votanti)

Zona 8: Milano, Novara, Vercelli, Pavia, Varese, Como

Totale Votanti Elezioni 2014 = 2.878.204 (9.96% del totale dei votanti)

Zona 9: Monza-Brianza, Lecco, Sondrio, Bergamo, Brescia, Trento, Bolzano, Vicenza

Totale Votanti Elezioni 2014 = 2.877.190 (9.95% del totale dei votanti)

Zona 10: Mantova, Cremona, Verona, Padova, Venezia, Treviso, Belluno, Pordenone, Udine, Gorizia, Trieste

Totale Votanti Elezioni 2014 = 2.927.447 (10.13% del totale dei votanti)