INDAGINE DI SIEROPREVALENZASUL SARS-CoV-2

12 aprile 2021

GLOSSARIO

Asintomatici: persone che dal 1 febbraio 2020 alla data dell'intervista non hanno manifestato nessuno dei seguenti sintomi: dolori ossei/muscolari; senso di stanchezza; mal di testa; congiuntivite; diarrea; difficoltà a respirare; dolori addominali; perdita/alterazione del gusto; perdita/alterazione dell'olfatto; mal di gola; febbre; tosse; sindrome di tipo influenzale; nausea/vomito; confusione mentale.

Test sierologici: test che servono ad individuare tutte quelle persone che sono entrate in contatto con il virus. Attraverso i test sierologici è possibile individuare gli anticorpi prodotti dal sistema immunitario in risposta al virus.

Zona: La variabile zona è stata costruita attraverso una ricodifica delle regioni in tre macro aree geografiche: Zona rossa (Piemonte, Lombardia, Veneto, Emilia Romagna, Marche); Resto del Nord+Centro (Valle d'Aosta, P.A. di Trento, P.A. di Bolzano, Friuli Venezia Giulia, Liguria, Toscana, Umbria e Lazio); Mezzogiorno (Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria, Sicilia, Sardegna).

NOTA METODOLOGICA

Il disegno di campionamento

Il disegno di campionamento adottato per l'indagine di Siero-prevalenza è a due stadi di selezione con stratificazione sia delle Unità di Primo Stadio, sia delle Unità di Secondo Stadio (circa 150,000 individui distribuiti su tutto il territorio nazionale). Le UP sono i comuni stratificati all'interno di ciascuna provincia in base alla loro dimensione demografica (circa 2000, quasi il 25% dei comuni italiani), mentre le US sono gli individui stratificati sulla base di 6 classi di età (0-17; 18-34; 35-49; 50-59; 60-69 70 e più), sesso e 4 macro-aggregazioni dell'attività economica (non occupati, occupati non sospesi del comparto PA e istruzione, occupati non sospesi del comparto sanità, occupati non sospesi di altri comparti, occupati sospesi). Il disegno di campionamento effettuato dall'ISTAT permette la stima trasversale della siero-prevalenza della popolazione che vive in famiglia e la possibilità di ritorni longitudinali. Le variabili di stratificazione utilizzate per il campionamento casuale sono: regione di residenza; sesso; età, suddivisa in 6 differenti classi; settori di attività economica (aggregati in 4 macro aree secondo la classificazione ATECO e la eventuale sospensione su decisione del governo). Il campione obiettivo da osservare sull'intero territorio italiano era pari a 150.000 individui, residenti nei circa 2000 Comuni. unità. Le scelte di base del disegno di campionamento possono essere così riassunte:

1. Domini di stima territoriale costituiti dalle Regioni e Province autonome di Bolzano e Trento.
2. Domini strutturali, all'interno di ciascuna Regione geografica, costituiti da 6 classi di età, 0-17; 18-34; 35 - 49; 50-59; 60-69; 70 e più; sesso e maggiori e minori di 50 anni e da 4 macro-classi di ATECO (sotto-popolazione degli occupati);

Lo schema ha prodotto una buona dispersione spaziale del campione estratto sul territorio, in virtù del fatto che è stato selezionato un numero rilevante di comuni (circa 2.000).
E' stato inoltre assicurato che tutte le Aziende Sanitarie (USL) fossero rappresentate nel campione selezionato, e che quasi tutti i 610 Sistemi Locali del Lavoro (SLL) fossero inclusi nel campione stesso (ad eccezione di 82 SLL); i risultati dell'allocazione e successiva selezione hanno mostrato, in sintesi, una buona rappresentazione dei territori sub-regionali italiani, in rapporto alle prevalenze stimate e agli errori pianificati, e una soddisfacente copertura del campione a livello comunale, di Aziende Sanitarie e anche di SLL. La metodologia di allocazione adottata per il campione è di tipo multivariato e multidominio.

Test sierologici, centri prelievo, laboratori

Ai sensi dell'articolo 122 del decreto-legge 17 marzo 2020, n. 18, e del conseguente D.P.C.M.18 marzo 2020, il Commissario straordinario ha provveduto, mediante apposita procedura di gara, a selezionare test sierologici

che, oltre a rispondere a criteri di specificità, siano anche di facile realizzazione su larga scala e connotati da rapidità di ottenimento del risultato e caratterizzati dai seguenti elementi:

- kit CLIA e/o ELISA per la rilevazione di IgG specifiche/anticorpi neutralizzanti per SARS-CoV-2;
- validati, da laboratori qualificati o agenzie regolatorie presenti a livello nazionale o internazionale;
- con specificità non inferiore al 95%;
- con sensibilità non inferiore al 90%;
- capacità di applicazione su larga scala;
- rapidità di produrre il risultato dell'indagine.

Con decreto del Commissario Straordinario per l'attuazione e il coordinamento delle misure di contenimento e contrasto dell'emergenza epidemiologica COVID–19, del 25 aprile 2020, la fornitura è stata aggiudicata all' operatore economico Abbott.

I laboratori attivati sono stati 49 e i centri prelievo 2099.

Il lavoro sul campo e il disegno del sistema di monitoraggio

La metodologia di indagine ha previsto una prima fase, di contatto telefonico degli individui campione per l'acquisizione della disponibilità all'effettuazione delle analisi sierologiche, la somministrazione di un breve questionario, predisposto dall'ISTAT in accordo con il Comitato Tecnico Scientifico, e la definizione di un appuntamento per il prelievo del siero a domicilio o presso un centro prelievo. Il servizio di Contact Center è stato svolto da personale della CRI appositamente formato. Per la gestione informatica della rilevazione e il caricamento dei dati raccolti è stata utilizzata una piattaforma, appositamente progettata dal Ministero della Salute. Per monitorare il lavoro sul campo e l'andamento della rilevazione è stato progettato dall'Istat un articolato sistema di indicatori di qualità. Ogni giorno sono stati elaborati e resi accessibili gli indicatori aggiornati al giorno precedente, consentendo di monitorare al massimo livello di dettaglio tutti i possibili esiti del contatto telefonico, le specifiche ragioni del rifiuto a partecipare alla rilevazione, il mancato rispetto degli appuntamenti fissati per il prelievo, etc.

Rientrano tra gli indicatori calcolati giornalmente anche alcuni dei principali tassi tradizionalmente calcolati per monitorare non solo la propensione delle unità campionarie a collaborare, ma anche le performance della rete di rilevazione. A tal fine sono stati calcolati e quotidianamente aggiornati i tassi di completezza, di caduta, di rifiuto, di irreperibilità, di inattività, di pigrizia e di appuntamenti/contatti telefonici andati a buon fine. Tutti gli indicatori sono stati declinati rispetto a variabili di tipo territoriale e alle caratteristiche socioanagrafiche delle unità campionarie (sesso ed età). Alcuni indicatori sono stati prodotti anche a livello di singolo operatore in modo da avere il massimo dettaglio informativo sulle perfomance dell'intera rete di rilevazione.
Per l'assegnazione dei cittadini ai centri prelievo più vicini l'Istat ha messo a disposizione una procedura di calcolo delle distanze tra centri prelievo e domicilio di ciascun individuo estratto nel campione.

La correzione della mancata risposta

Per ridurre l'impatto della mancata risposta totale sono stati utilizzati dei fattori di correzione. I fattori di correzione sono stati calcolati a partire dalle probabilità di risposta predette tramite un modello logistico. Per evitare di aumentare eccessivamente la variabilità dei pesi campionari, con conseguenze negative sull'efficienza delle stime, sono state determinate delle classi di aggiustamento sulla base dei quintili delle distribuzioni regionali delle stesse probabilità di risposta predette dal modello. Per ciascuna classe di aggiustamento, il fattore di correzione è stato calcolato come l'inverso della media delle probabilità di risposta predette in ciascuna classe di aggiustamento.

Il modello logistico ha preso in considerazione un set di variabili disponibili per l'intero campione teorico (quindi sia rispondenti che non rispondenti). Le variabili considerate sono state: regione (Bolzano e Trento sono state trattate distintamente), tipologia comunale (città metropolitana; corona dell'area metropolitana; minore di 2000 abitanti; tra 2000 e 10000 abitanti; tra 10000 e 50000 abitanti; oltre 50000 abitanti), sesso, classe d'età (0-17; 18-34; 35-49; 50-59; 60-69; 70+), stato delle ATECO (occupati sospesi; occupati non sospesi, altro; occupati non sospesi, PA + Istruzione; occupati non sospesi, sanità; non occupati), prevalenze comunali (sulla base di previsioni fornite dall'Istituto Superiore di Sanità), differenza percentuale dei tassi di mortalità comunali rispetto allo stesso periodo dell'anno precedente.

Calcolo delle stime e modello per la presentazione sintetica degli errori campionari

Le stime prodotte per l'indagine sono principalmente stime di frequenze assolute o di frequenze relative riferite per diversi domini d (nazionale, regionale, provinciale, area Covid, stato ATECO, sesso e classi d'età ed alcuni incroci di questi).
Definendo la variabile Y come una variabile dicotomica che, dunque, sulla generica unità k $(k = 1, \dots, N)$ assume valore 1 se l'unità possiede la caratteristica Y e 0 altrimenti, la frequenza assoluta può essere scritta come il totale della variabile Y nel domino U_d:

$$t_{Y_d} = \sum_{k \in U_d} y_k.$$

La frequenza relativa, dunque, può essere vista come la media del carattere Y nel dominio U_d:ed è ottenuta dividendo t_{Y_d} per la numerosità della popolazione U:

$$\mu_{Y_d} = \frac{\sum_{k \in U_d} y_k}{N_d} = \frac{t_{Y_d}}{N_d}.$$

La stime di queste quantità sono ricavate attraverso lo stimatore calibrato (cfr. Deville, Särndal, 19921; Särndal, 20072; Tillé, 20193) che costituisce il principale metodo di stima correntemente utilizzato nella maggior parte delle indagini Istat.
Lo stimatore calibrato del totale è definito come:

$$\hat{t}_{Y_{CAL}} = \sum_{k \in R} y_k \, w_k$$

dove i pesi finali w_k sono determinati attraverso la risoluzione di un problema di minimo vincolato così definito:

$$\begin{cases} min\left\{\sum_{k \in \mathrm{R}} dist(d_k, w_k)\right\} \\ \sum_{k \in \mathrm{R}} \boldsymbol{x_k} \; w_j = \boldsymbol{t_X} \end{cases}$$

in cui d_k è il peso da disegno relativo all'unità k-esima che deriva dall'inverso della probabilità di inclusione dell'unità nel campione e dalla procedura di correzione per mancata risposta; $\boldsymbol{t_X}$ è il vettore dei delle variabili di calibrazione e $\boldsymbol{x_k}$ è il vettore delle variabili ausiliarie osservate sulla k-esima unità dei rispondenti.
I pesi w_k così ottenuti garantiscono la coerenza con i totali noti delle variabili ausiliarie considerate e, rispetto ad una opportuna funzione di distanza prescelta, sono il più vicino possibile ai pesi da disegno. In pratica, il peso indica il numero di unità della popolazione rappresentata dalla generica unità campionaria k. Per esempio, se a un'unità campionaria viene attribuito un peso pari a 30, ciò indica che questa unità rappresenta se stessa e altre 29 unità della popolazione che non sono state incluse nel campione.
La procedura di calibrazione è stata svolta con il pacchetto ReGenesees4 implementato in ambiente R.
Nella calibrazione si è tenuto conto dei seguenti totali di popolazione ricavati dal registro base degli individui (RBI):

- distribuzione regionale della popolazione per sesso e classe d'età (0-17; 18-34; 35-49; 50 59; 60-69; 70+);
- distribuzione regionale della popolazione per sesso e stato della ATECO (occupati sospesi; occupati non sospesi, altro; occupati non sospesi, PA + Istruzione; occupati non sospesi, sanità; non occupati);
- distribuzione provinciale della popolazione;
- distribuzione regionale della popolazione per cittadinanza;
- distribuzione della popolazione per ripartizione (Nord-Ovest, Nord-Est, Centro e Mezzogiorno), classe d'età (0-17; 18-34; 35-49; 50 59; 60-69; 70+) e titolo di studio (4 livelli).

Il numero complessivo di totali (vincoli) considerati è 11100. La funzione di distanza utilizzata è la funzione logaritmica troncata con estremi fissati a 0.74 e 6.50.

---

[1] Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, *87*(418), 376-382.
[2] Särndal, C. E. (2007). The calibration approach in survey theory and practice. *Survey methodology*, *33*(2), 99-119.
[3] Devaud, D., & Tillé, Y. (2019). Deville and Särndal's calibration: revisiting a 25-years-old successful optimization problem. *TEST*, *28*(4), 1033-1065.
[4] Zardetto D. (2015).ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys, (extended version). *Journal of Official Statistics*, 31(2):177-203.

Al fine di valutare l'accuratezza delle stime prodotte dall'indagine è necessario tenere conto dell'errore campionario che deriva dall'aver osservato la variabile di interesse solo su una parte (campione) della popolazione. Tale errore può essere espresso in termini di errore assoluto (standard error)

$$\hat{\sigma}(\hat{t}_{Y_d}) = \sqrt{\hat{Var}(\hat{t}_{Y_d})} \quad (1)$$

o di errore relativo (cioè l'errore assoluto diviso per la stima, che prende il nome di coefficiente di variazione, CV)

$$\hat{\varepsilon}(\hat{t}_{Y_d}) = \frac{\sqrt{\hat{Var}(\hat{t}_{Y_d})}}{\hat{t}_{Y_d}} \quad (2)$$

che spesso viene riportato in valore percentuale (CV%).
Gli errori campionari delle espressioni (1) e (2), consentono di valutare il grado di precisione delle stime; inoltre, l'errore assoluto permette di costruire l'intervallo di confidenza di livello $1-\alpha$, che, quindi, con probabilità $1-\alpha$ contiene il parametro d'interesse. Con riferimento alla generica stima $\hat{t}_{Y_d}$ tale l'intervallo di confidenza di livello $1-\alpha$ è:

$$\mathrm{IC}_{1-\alpha} = \left[\hat{t}_{Y_d} - k\,\hat{\sigma}(\hat{t}_{Y_d});\ \hat{t}_{Y_d} - k\,\hat{\sigma}(\hat{t}_{Y_d})\right],$$

dove k, nel caso di intervalli di confidenza al 95%, è 1.96 ovvero, pari al valore del $(1-\alpha/2)$% percentile della normale standard.

Di seguito è riportato il prospetto 1 che fornisce l'errore relativo associato a determinati valori della stima puntuale nei vari domini di studio.
Ad ogni stima $\hat{t}_{Y_d}$ corrisponde un errore campionario relativo $\hat{\varepsilon}(\hat{t}_{Y_d})$; ciò significa che per consentire un uso corretto delle stime sarebbe necessario pubblicare per ogni stima il corrispondente errore di campionamento relativo. Questo, tuttavia, non è possibile sia per limiti di tempo e di costi di elaborazione, sia perché le tavole di pubblicazione risulterebbero appesantite e di non facile consultazione per l'utente finale. Inoltre, non sarebbero comunque disponibili gli errori delle stime non pubblicate, che l'utente può ricavare in modo autonomo.
Per tali motivi si ricorre, in genere, ad una presentazione sintetica degli errori relativi basata sul metodo dei modelli regressivi (Wolter, 20075) fondata sulla determinazione di una funzione matematica che mette in relazione ciascuna stima con il proprio errore di campionamento. L'approccio utilizzato per la costruzione di questi modelli è diverso a seconda che si tratti di variabili qualitative o quantitative. Infatti, per quanto riguarda le stime di frequenze assolute (o relative) riferite alle modalità di variabili qualitative, è possibile utilizzare modelli che hanno un fondamento teorico, secondo cui gli errori relativi delle stime di frequenze assolute sono funzione decrescente dei valori delle stime stesse.
Il modello utilizzato per le stime di frequenze assolute, con riferimento al generico dominio d, è del tipo seguente:

$$\log\left(\hat{\varepsilon}^2(\hat{t}_{Y_d})\right) = a + b\,\log(\hat{t}_{Y_d})$$

dove i parametri a e b vengono stimati, utilizzando il metodo dei minimi quadrati, su un insieme di stime ottenute dall'indagine (con i rispettivi errori relativi) che coprono approssimativamente l'intervallo di variazione delle stime di frequenze che vengono pubblicate.

[5] Wolter, K. (2007). *Introduction to variance estimation*. Springer Science & Business Media.

Istat
Istituto Nazionale di Statistica

Ministero della Salute

PROSPETTO 1 - MODELLI SINTETICI DEGLI ERRORI

DOMINIO		a	b	r^2	10000	20000	50000	70000	100000	200000	500000	1000000	2000000
	ITALIA	11,99332	-1,36748	0,929	74,021	46,082	24,629	24,629	15,333	9,545	5,102	3,176	2,407
REGIONI													
1	Piemonte	11,23392	-1,40470	0,899	42,659	26,217	13,775	13,775	8,466	5,203	2,734	1,680	1,264
2	Valle d'Aosta	6,56678	-1,43177	0,886	3,651	2,223	1,154	1,154	0,702	0,428	0,222	0,135	0,101
3	Lombardia	11,07151	-1,37186	0,915	45,755	28,441	15,170	15,170	9,430	5,862	3,127	1,943	1,472
5	Veneto	10,92177	-1,40657	0,893	36,182	22,222	11,666	11,666	7,165	4,401	2,310	1,419	1,067
6	Friuli-Venezia Giulia	9,23261	-1,39454	0,867	16,435	10,136	5,350	5,350	3,300	2,035	1,074	0,663	0,499
7	Liguria	9,37704	-1,36231	0,905	20,492	12,780	6,847	6,847	4,270	2,663	1,427	0,890	0,675
8	Emilia-Romagna	10,98014	-1,38352	0,904	41,426	25,647	13,607	13,607	8,424	5,215	2,767	1,713	1,294
9	Toscana	10,70237	-1,35980	0,886	40,214	25,102	13,463	13,463	8,404	5,246	2,814	1,756	1,333
10	Umbria	8,07687	-1,30061	0,914	14,212	9,055	4,990	4,990	3,179	2,026	1,116	0,711	0,546
11	Marche	9,53435	-1,40351	0,876	18,338	11,274	5,927	5,927	3,644	2,240	1,178	0,724	0,545
12	Lazio	11,08717	-1,37309	0,907	45,852	28,490	15,187	15,187	9,437	5,863	3,126	1,942	1,470
13	Molise	8,86452	-1,34149	0,833	17,455	10,965	5,931	5,931	3,725	2,340	1,266	0,795	0,606
14	Abruzzo	6,26845	-1,20906	0,917	8,771	5,769	3,315	3,315	2,180	1,434	0,824	0,542	0,424
15	Campania	9,62466	-1,22958	0,905	42,737	27,909	15,889	15,889	10,376	6,776	3,857	2,519	1,963
16	Puglia	8,96387	-1,19634	0,920	35,794	23,645	13,668	13,668	9,029	5,964	3,448	2,278	1,787
17	Basilicata	6,72494	-1,24944	0,840	9,150	5,934	3,348	3,348	2,171	1,408	0,794	0,515	0,400
18	Calabria	8,39077	-1,20626	0,838	25,675	16,903	9,726	9,726	6,403	4,215	2,426	1,597	1,250
19	Sicilia	9,70389	-1,23824	0,857	42,726	27,818	15,774	15,774	10,270	6,687	3,792	2,469	1,921
20	Sardegna	7,84526	-1,18745	0,915	21,314	14,123	8,197	8,197	5,432	3,599	2,089	1,384	1,088
41	Bolzano	9,19680	-1,38728	0,850	16,691	10,320	5,466	5,466	3,379	2,090	1,107	0,684	0,516
42	Trento	8,88656	-1,39713	0,887	13,659	8,417	4,438	4,438	2,734	1,685	0,888	0,547	0,412
AREA COVID													
1	Rossa	11,47080	-1,38517	0,918	52,544	32,511	17,236	17,236	10,664	6,599	3,498	2,164	1,635
2	Resto del Nord + Centro	11,51270	-1,39316	0,911	51,716	31,911	16,855	16,855	10,400	6,417	3,390	2,092	1,577
3	Mezzogiorno	9,97216	-1,24815	0,909	46,681	30,288	17,097	17,097	11,093	7,198	4,063	2,636	2,047
RIPARTIZIONE a 3													
1	Nord	11,71879	-1,39914	0,918	55,772	34,342	18,090	18,090	11,139	6,859	3,613	2,225	1,675
2	Centro	11,18462	-1,37278	0,911	48,211	29,959	15,973	15,973	9,926	6,168	3,288	2,043	1,547
3	Mezzogiorno	9,97216	-1,24815	0,909	46,681	30,288	17,097	17,097	11,093	7,198	4,063	2,636	2,047

Per quanto riguarda la stima della varianza campionaria delle stime di frequenze assolute e relative, al fine di permettere il calcolo degli errori campionari delle stime pubblicate, mediante il metodo sopra descritto, nel prospetto 1 vengono riportati i valori di a e b e l'indice r^2 che fornisce una misura del grado di rappresentatività degli errori campionari stimati in base al modello.
Inoltre, allo scopo di facilitare il calcolo degli errori campionari, sempre nel prospetto 1 sono riportati, per i diversi domini territoriali di riferimento delle stime, i valori interpolati degli errori campionari relativi percentuali di alcuni valori tipici assunti dalle stime di frequenze assolute e di totali.
Nel prospetto 2 sono illustrate le modalità di calcolo per la costruzione dell'intervallo di confidenza al 95% delle stime puntuali riferite al numero di positivi in Italia, in Toscana, nell'Area Covid Rossa e nella ripartizione Nord.

PROSPETTO 2 – CALCOLO ESEMPLIFICATIVO DELL'INTERVALLO DI CONFIDENZA.

	NUMERO DI POSITIVI IN ITALIA	NUMERO DI POSITIVI IN TOSCANA
STIMA PUNTUALE:	1.482.146	38.041
ERRORE RELATIVO PERCENTUALE (CV%)	2,427	16,213
Errore relativo (CV)	2,427/100 =0,02427	16,213/100 =0,16213
STIMA INTERVALLARE:		
Semi ampiezza dell'intervallo	1,96*0,02427*1.482.146 = 70.499	1.96*0,16213*38.041 =12.089
Limite inferiore dell'intervallo di confidenza	1.482.146- 70.499 = 1.411.647	38.041-12.089 =25.953
Limite superiore dell'intervallo di confidenza	1.482.146+ 70.499 = 1.552.644	38.041+12.089 =50.130

	Numero di Positivi in Area covid rossa	Numero di Positivi NEL nord
STIMA PUNTUALE:	1.150.819	1.218.153
ERRORE RELATIVO PERCENTUALE (CV%)	1,963	1,938
Errore relativo (CV)	1,963/100 =0,01963	1,938/100 =0,01938
STIMA INTERVALLARE:		
Semi ampiezza dell'intervallo	1,96*0,01963*1.150.819 = 44.296	1,96*0,01938*1.218.153 = 46.267
Limite inferiore dell'intervallo di confidenza	1.150.819-44.296 =1.106.524	1.218.153-46.267 =1.171.886
Limite superiore dell'intervallo di confidenza	1.150.819+44.296 =1.195.115	1.218.153+46.267 =1.264.419

Per chiarimenti tecnici
Linda Laura Sabbadini, Istat
sabbadin@istat.it

Maria Clelia Romano, Istat
romano@istat.it

Per chiarimenti metodologici
Orietta Luzi
luzi@istat.it

Stefano Falorsi
stfalors@istat.it