File Standard

# Percorsi di studio e di lavoro dei diplomati Indagine 1998

*Manuale utente* (*)

**ISTITUTO NAZIONALE DI STATISTICA**

(*) A cura di Paola Ungaro

## PREMESSA

Il Decreto Legislativo n° 322 del 6/9/1989 regola la diffusione delle informazioni statistiche prodotte nell'ambito del Sistema Statistico Nazionale al fine di garantire la riservatezza dei rispondenti.
In particolare, per la diffusione di dati elementari, l'articolo 10, comma 2, dispone quanto segue: "Sono distribuite altresì, ove disponibili, su richiesta motivata e previa autorizzazione del Presidente dell'ISTAT, collezioni campionarie di dati elementari, resi anonimi e privi di ogni riferimento che ne permetta il collegamento con singole persone fisiche e giuridiche".
Nell'osservanza di tale legge l'ISTAT ha adottato misure e tecniche che rendono impossibile, o altamente improbabile, il collegamento dei dati rilasciati con l'unità statistica a cui si riferiscono. Per tale motivo sono state apportate alcune modifiche sui files originali delle indagini, nell'intento di garantire la massima protezione ai dati contenendo al minimo l'eventuale perdita di informazione.
Le metodologie applicate si concretizzano nell'accorpamento e/o riclassificazione di modalità di variabili e nell'oscuramento di variabili. In quest'ultimo caso, nei campi del tracciato record è riportata la dicitura "RISERVATO ISTAT".

Con l'occasione si ricorda al richiedente che si impegna a:
- utilizzare i dati soltanto per gli scopi dichiarati;
- non fornire a terzi i dati elementari, consentendone l'accesso, sotto la propria responsabilità, soltanto alle persone direttamente coinvolte nel lavoro per il quale essi sono stati richiesti;
- citare la fonte ISTAT nell'eventuale divulgazione di elaborazioni dei dati;
- inviare alla Biblioteca dell'ISTAT due copie delle pubblicazioni eventualmente prodotte con l'utilizzo dei dati ottenuti.

# ASPETTI GENERALI DELL'INDAGINE 1998 SUI PERCORSI DI STUDIO E DI LAVORO DEI DIPLOMATI[1]

## 1. I CONTENUTI

L'obiettivo dell'indagine è l'analisi delle condizioni di studio o di lavoro dei giovani a poco più di tre anni dal conseguimento del diploma di scuola secondaria superiore.

Fino ad ora, l'analisi del percorso formazione-lavoro ha potuto contare su una base informativa molto ampia per quanto riguarda i laureati, mentre sono mancate informazioni altrettanto ricche sul segmento dei maturi.
Sono però proprio questi ultimi a far registrare nel nostro paese i tassi di disoccupazione più elevati e a riversarsi "in massa" sull'università nel tentativo, spesso fallimentare, di migliorare la propria qualificazione e di sottrarsi così ad una lunga e scoraggiante ricerca di lavoro.

Proprio in considerazione delle particolarità che in Italia caratterizzano i percorsi all'uscita della secondaria superiore, l'indagine propone, per quei giovani che decidono di affrontare gli studi accademici, un adeguato approfondimento del percorso universitario; mentre offre un'ampia analisi degli esiti occupazionali, per quanti optano per un occupazione.
La scelta di condurre le interviste a distanza di tre anni dal conseguimento del diploma permette appunto di indagare, tanto sul primo inserimento dei giovani nel mondo del lavoro, quanto sull'impegno profuso da questi ultimi negli studi accademici (numero di esami sostenuti, frequenza delle lezioni, etc.) o, all'opposto, sull'eventuale interruzione della frequenza. Il fenomeno dell'abbandono degli studi, infatti, è particolarmente rilevante proprio nei primi anni di corso.

Nel tentativo di cogliere le determinanti dei diversi percorsi, l'analisi tiene in considerazione anche il contesto economico e sociale in cui si è formato lo studente, consentendo, per questa strada, di valutare sia l'interazione esistente tra estrazione sociale e selezione/espulsione operata dall'università, sia la capacità delle politiche del diritto allo studio di costituire un eventualmente correttivo.

## 2. LA METODOLOGIA

### 2.1 LA TECNICA DI RILEVAZIONE

L'indagine è di tipo campionario a due stadi di selezione: le unità di primo stadio sono le Scuole secondarie superiori, quelle di secondo stadio i diplomati dell'anno 1995.
L'indagine si è svolta in due fasi: la prima, rivolta alle Scuole, è stata di tipo postale; la seconda, rivolta ai diplomati, è avvenuta tramite telefono, col sistema C.A.T.I. (Computer Assisted Telephone Interview).

La prima fase ha avuto come obiettivo principale ottenere la lista e i recapiti telefonici dei diplomati su cui condurre l'indagine vera e propria. Nella seconda fase i maturi selezionati sono stati contattati telefonicamente da una ditta specializzata.

Le interviste ai diplomati si sono svolte dal 15 settembre al 16 dicembre 1998, circa 3 anni e 3 mesi dopo il conseguimento del diploma.

[1] Il presente paragrafo è tratto da: ISTAT, *Percorsi di studio e di lavoro dei diplomati. Indagine 1998*, collana Informazioni, in corso di stampa.

## 2.2 LA STRATEGIA DI CAMPIONAMENTO

Nelle pagine che seguono si illustrano gli obiettivi conoscitivi e gli aspetti più significativi della strategia di campionamento dell'indagine sugli sbocchi professionali dei maturi dell'anno 1995.
L'insieme delle unità statistiche è costituita dagli studenti che hanno conseguito il diploma di maturità nelle scuole secondarie superiori nell'anno 1995. Gli obiettivi conoscitivi più significativi riguardano le condizioni di lavoro o di studio ed in particolare siconsiderano, distintamente per sesso, tipo di scuola e ripartizione:

- i maturi che lavorano o non lavorano;
- i maturi che studiano o non studiano;
- i maturi che sono o non sono alla ricerca di un lavoro

I principali domini territoriali di riferimento delle stime sono:

1) l'intero territorio nazionale;
2) le cinque ripartizioni geografiche (Italia Nord Occidentale, Italia Nord Orientale, Italia Centrale, Italia Meridionale, Italia Insulare);

### 2.2.1 Descrizione generale del disegno di campionamento

Il disegno di campionamento è a due stadi di selezione con stratificazione delle unità primarie. Le unità primarie, costituite dalle scuole secondarie superiori, sono stratificate per regione geografica e tipo di scuola; l'incrocio delle regioni geografiche con i tipi di scuola ha portato alla formazione di circa 350 strati. Le unità secondarie sono gli alunni che hanno conseguito la maturità nell'anno 1995.
Le scuole appartenenti a ciascuno strato sono state selezionate senza reimmissione e con probabilità proporzionali alla loro dimensione in termini di maturi. Ciascuna scuola estratta al primo stadio ha proceduto alla selezione di un numereo approssimativamente costante di maturi campione, mediante scelta sistematica dalla lista riportante i maturi dell'anno 1995.
Il campione di maturi così identificato è stato intervistato mediante intervista telefonica. Tenendo conto che la rilevazione telefonica può dar luogo ad un'alta percentuale di mancate risposte si è deciso di estrarre un campione base ed un campione suppletivo di maturi da intervistare, nel caso in cui alcuni maturi del campione base fossero non rispondenti. Si è definito un campione suppletivo all'incirca della stessa dimensione di quello base.
La numerosità campionaria di primo e di secondo stadio è stata definita sia sulla base di criteri organizzativi e di costo, che in relazione agli errori di campionamento attesi delle principali stime di interesse. Nell'ambito di ciascuna scuola campione si è definito un numero approssimativamente costante, pari a 30, di maturi campione.
La preparazione delle liste base e suppletiva è stata effettuata dall'Istat suddividendo i 30 nominativi forniti da ciascuna scuola in 15 nominativi base e 15 nominativi suppletivi.
Allo scopo di illustrare la dimensione campionaria adottata nella indagine, viene riportata nel Prospetto 1 la distribuzione per regione delle scuole e dei maturi dell'anno 1995 nell'universo e nel campione dei rispondenti.

*Prospetto 1. Distribuzione regionale delle Scuole e dei Maturi nell'universo e nel campione. Anno 1995*

| Regione | Scuole Universo | Scuole Campione | Maturi Universo | Maturi Campione |
|---|---|---|---|---|
| Piemonte | 510 | 101 | 30220 | 1196 |
| Valle d'Aosta | 23 | 18 | 685 | 212 |
| Lombardia | 966 | 148 | 66116 | 1943 |
| Trentino A. A. | 119 | 41 | 6207 | 454 |
| Veneto | 497 | 106 | 35473 | 1325 |
| Friuli V. G. | 145 | 53 | 8920 | 598 |
| Liguria | 213 | 63 | 11458 | 661 |
| Emilia Romagna | 400 | 90 | 30921 | 1149 |
| Toscana | 400 | 86 | 28529 | 1099 |
| Umbria | 110 | 42 | 7377 | 493 |
| Marche | 187 | 57 | 12658 | 675 |
| Lazio | 697 | 124 | 51608 | 1569 |
| Abruzzo | 161 | 53 | 12232 | 670 |
| Molise | 54 | 31 | 3103 | 375 |
| Campania | 694 | 129 | 55918 | 1620 |
| Puglia | 491 | 106 | 39613 | 1197 |
| Basilicata | 119 | 45 | 6101 | 499 |
| Calabria | 325 | 76 | 21496 | 904 |
| Sicilia | 786 | 130 | 46088 | 1457 |
| Sardegna | 247 | 64 | 15625 | 747 |
| ITALIA | 7144 | 1563 | 490348 | 18843 |

### 2.2.2 Procedimento per il calcolo delle stime

Le stime prodotte dall'indagine sono principalmente stime di frequenze assolute riferite ai maturi dell'anno 1995, ad esempio, il numero totale dei diplomati che lavorano tre anni dopo la maturità.
Il principio su cui è basato ogni metodo di stima campionaria è che le unità appartenenti al campione rappresentino anche le unità della popolazione che non sono incluse nel campione.
Questo principio viene realizzato attribuendo ad ogni unità campionaria un peso che denota il numero di unità della popolazione rappresentate dalla unità medesima. Se, ad esempio, ad una unità campionaria viene attribuito un peso pari a 30, vuol dire che questa unità rappresenta se stessa ed altre 29 unità della popolazione che non sono state incluse nel campione.

Al fine di rendere più chiara la successiva esposizione, introduciamo la seguente simbologia:

d , indice di livello territoriale di riferimento delle stime;
i , indice della scuola;
j , indice del maturo;
h , indice dello strato di scuole;
$M_h$ , totale dei maturi dello strato h;
$M_{hs}$ , totale dei maturi dello strato h per sesso (s=1, maschi; s=2, femmine);
$M_{hi}$ , totale dei maturi della scuola i dello strato h;
$m_{hi}$ , campione dei maturi della scuola i dello strato h;
$m_{his}$ , campione dei maturi della scuola i dello strato h per sesso ;
$N_h$ , totale di scuole nello strato h;
$n_h$ , scuole campione nello strato h;
$H_d$ , numero totale di strati nel dominio d ;
$D_{hij} = \frac{1}{n_h}\frac{M_h}{m_{hi}}$, peso diretto da attribuire al j-mo maturo della scuola i dello strato h;
$W_{hij}$ , peso finale da attribuire al j-mo maturo della scuola i dello strato h;

x , generica variabile oggetto di indagine;
$X_{hij}$ , valore osservato della variabile x sul j-mo maturo della scuola i e strato h.

Ad esempio, se x rappresenta la condizione lavorativa (espressa dalle due modalità lavora, non lavora), si avrà $X_{hij}$ =1 se il maturo j lavora e $X_{hij}$ =0 altrimenti.

Ipotizziamo di voler stimare con riferimento ad un generico dominio d, il totale generico totale espresso dalla seguente relazione:

$$X_d = \sum_{h=1}^{H_d} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} X_{hij} \qquad (1)$$

La stima del totale (1), è ottenuta mediante uno stimatore del rapporto separato e post-statificato per sesso, espresso mediante la seguente formula:

$$\hat{\hat{X}}_d = \sum_{h=1}^{H_d} \sum_{s=1}^{2} \frac{\hat{X}_{hs}}{\hat{M}_{hs}} M_{hs} \qquad (2)$$

dove

$$\hat{M}_{hs} = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{his}} D_{hi} \quad \text{e} \quad \hat{X}_{hs} = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{his}} X_{hij} D_{hi} \qquad (3)$$

rappresentano rispettivamente, con riferimento allo strato h ed al sesso s, la stima del totale dei maturi, $M_{hs}$, e la stima del totale dei maturi, $X_{hs}$, che possiedono una determinata caratteristica.

Lo stimatore appena illustrato rientra nella classe degli stimatori di ponderazione vincolata, che è il metodo di stima standard per la maggior parte delle indagini ISTAT sulle imprese e sulle famiglie. Tale classe di stimatori viene utilizzata quando si dispone di informazioni espresse in forma di totali noti di variabili ausiliarie legate alle variabili di interesse. Una delle caratteristiche dei suddetti stimatori è quella di garantire l'uguaglianza tra i totali noti delle variabili ausiliarie e le corrispondenti stime ottenute con i pesi finali. Nell'indagine in oggetto i totali noti sono costituiti dal numero di maturi per sesso e strato (regione e tipo scuola).

Le formule di cui sopra sono state implementate in un software generalizzato che permette di ottenere i pesi finali da attribuire alle unità campionarie. Il software costruisce il peso da attribuire alle unità campionarie, seguendo le seguenti fasi :

1) si calcola dapprima, per ciascuna unità, il peso diretto $D_{hij}$ ottenuto come reciproco della probabilità di inclusione della unità;
2) si calcola il fattore correttivo $c_{hij}$ che consente di soddisfare, in ogni regione e tipo scuola, la condizione di uguaglianza tra i totali noti della popolazione e le corrispondenti stime campionarie;
3) il peso finale $W_{hij}$ è dato dal prodotto del peso base per il fattore correttivo sopra indicato.

E' possibile esprimere lo stimatore del rapporto in funzione del peso finale. Mediante semplici passaggi, a partire dalla formula (2), si ha:

$$\hat{\hat{X}}_d = \sum_{h=1}^{H_d} \sum_{s=1}^{2} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{his}} X_{hij} W_{hij} \qquad (4)$$

dove

$$W_{hij} = D_{hij} \frac{M_{hs}}{\sum_{i=1}^{n_h} \sum_{j=1}^{m_{his}} D_{hi}} = D_{hij}\, c_{hij} \qquad (5)$$

### 2.3 VALUTAZIONE DEL LIVELLO DI PRECISIONE DELLE STIME

### 2.3.1 Calcolo della varianza campionaria

Le principali statistiche di interesse per valutare la variabilità campionaria delle stime prodotte dall'indagine, sono l'errore di campionamento assoluto e l'errore di campionamento relativo.

La stima dell'errore di campionamento assoluto di $\hat{\hat{X}}_d$ è definita dalla seguente espressione:

$$\hat{\sigma}(\hat{\hat{X}}_d) = \sqrt{\hat{V}ar(\hat{\hat{X}}_d)} \qquad (6)$$

La stima dell'errore di campionamento relativo di $\hat{\hat{X}}_d$ è definita dall'espressione:

$$\hat{\varepsilon}(\hat{\hat{X}}_d) = \frac{\hat{\sigma}(\hat{\hat{X}}_d)}{\hat{\hat{X}}_d} \qquad (7)$$

La stima della varianza di $\hat{\hat{X}}_d$, indicata nella (6) come $\hat{V}ar(\hat{\hat{X}}_d)$, viene calcolata utilizzando il metodo di linearizzazione di Woodruff, che consente di ottenere un'espressione approssimata della varianza campionaria nel caso di stimatori, come quello espresso nella (2), che non sono funzione lineare dei dati campionari. In simboli si ha:

$$\hat{V}ar\left(\hat{\hat{X}}_d\right) \cong \sum_{h=1}^{H_d} \frac{n_h}{n_h - 1} (\hat{Z}_{hi} - \bar{\hat{Z}}_h)^2 \qquad (8)$$

in cui

$$\hat{Z}_{hi} = \sum_{j=1}^{m_{hij}} \sum_{s=1}^{2} \left( X_{hij} - \frac{\hat{X}_{hs}}{\hat{M}_{hs}} \right) \delta_{hijs} W_{hij} \qquad (9)$$

è l'espressione della variabile linearizzata relativa allo stimatore del rapporto utilizzato per l'indagine in parola e dove $\delta_{hijs}$ =1 se il j-mo maturo è del sesso s e $\delta_{hijs}$ =0 altrimenti.

Gli errori campionari espressi dalla (6) e dalla (7) consentono di valutare il grado di precisione delle stime; inoltre, l'errore assoluto permette di costruire un intervallo di confidenza, nel quale con una certa probabilità si trova il parametro oggetto di stima :

$$Pr\left\{\hat{\hat{X}}_d - k\hat{\sigma}(\hat{\hat{X}}_d) \le X_d \le \hat{\hat{X}}_d + k\hat{\sigma}(\hat{\hat{X}}_d)\right\} = P \qquad (10)$$

Nella (10) il valore di k dipende dal valore fissato per la probabilità P; ad es., per P=0,95 si ha k=2.

### 2.3.2 Presentazione degli errori campionari

Ad ogni stima $\hat{\hat{X}}_d$ corrisponde un errore campionario relativo $\hat{\varepsilon}(\hat{\hat{X}}_d)$; 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 motivi di tempi e costi eccessivi di elaborazione, sia perchè 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 questo si ricorre frequentemente ad una presentazione sintetica degli errori relativi, basata sul metodo dei modelli regressivi.
Questo metodo si basa sulla determinazione di una funzione matematica che mette in relazione ciascuna stima con il proprio errore relativo.
Il modello utilizzato, con riferimento alle stime di frequenze assolute, è del tipo seguente:

$$\log \hat{\varepsilon}^2(\hat{\hat{X}}_d) = a + b \log(\hat{\hat{X}}_d) \qquad (11)$$

dove i parametri a e b vengono stimati mediante il metodo dei minimi quadrati e d indica il generico dominio di stima. Nella presente indagine è stato stimato il modello (11) con riferimento ai seguenti domini di stima: il totale Italia e le cinque ripartizioni geografiche (Nord-Ovest, Nord est, Centro, Sud e Isole).
Per la presente indagine si è utilizzato un software generalizzato messo a punto dall'ISTAT che consente di calcolare gli errori di campionamento, gli intervalli di confidenza ed inoltre permette di costruire modelli regressivi per la presentazione sintetica degli errori di campionamento.

Il Prospetto 2 riporta i valori dei coefficienti a e b e dell'indice di determinazione $R^2$ del modello utilizzato per l'interpolazione degli errori campionari delle stime riferite ai maturi per il totale Italia e per le cinque ripartizioni geografiche. Sulla base delle informazioni contenute in tale prospetto è possibile calcolare l'errore relativo di una determinata stima di frequenza assoluta, $\hat{\hat{X}}_d^*$ , nel modo di seguito descritto. Dalla (11) mediante semplici passaggi si ha:

$$\hat{\varepsilon}(\hat{\hat{X}}_d^*) = \sqrt{\exp(a + b \log(\hat{\hat{X}}_d^*))} \qquad (12)$$

Se ad esempio la stima $\hat{\hat{X}}_d^*$ si riferisce alla ripartizione Nord-Ovest, è possibile introdurre nella (12) i valori dei parametri a e b (a=6,01102 , b=-1,27372) riportati nella prima riga del Prospetto 2.

*Prospetto 2. Valori dei coefficienti a e b e dell'indice di determinazione $R^2$ (%) del modello per l'interpolazione degli errori campionari delle stime riferite ai maturi per totale Italia e ripartizioni geografiche. Anno 1995*

| DOMINIO DI STIMA | a | b | $R^2$ |
|---|---|---|---|
| Nord-Ovest | 6.01102 | -1.27372 | 91.22 |
| Nord-Est | 4.63404 | -1.17267 | 91.98 |
| Centro | 4.99223 | -1.18480 | 92.56 |
| Sud | 5.05210 | -1.17886 | 93.42 |
| Isole | 4.34128 | -1.11386 | 94.61 |
| ITALIA | 5.96463 | -1.23966 | 94.15 |

Allo scopo di facilitare il calcolo degli errori campionari, nel Prospetto 3, per alcuni valori crescenti di stime di frequenze percentuali (colonna 1), sono riportati i corrispondenti valori assoluti delle stime (colonna 2 ) e i valori interpolati degli errori relativi percentuali (colonna 3 ), calcolati in base al modello (11).

Per ciascun dominio il suddetto prospetto ha la seguente struttura:

| **(1)** Valori percentuali delle stime | **(2)** Valori assoluti delle stime corrispondenti alle percentuali (1) | **(3)** Errori interpolati relativi percentuali |
|---|---|---|
| 1% | $\hat{\hat{X}}_d(1)$ | $\hat{\varepsilon}\,[\,\hat{\hat{X}}_d(1)\,]$ |
| 2% | $\hat{\hat{X}}_d(2)$ | $\hat{\varepsilon}\,[\,\hat{\hat{X}}_d(2)\,]$ |
| ..... | ..... | ..... |
| …. | ….. | …… |
| ..... | ..... | ..... |
| 50% | $\hat{\hat{X}}_d(50)$ | $\hat{\varepsilon}\,[\,\hat{\hat{X}}_d(50)\,]$ |

Le informazioni contenute nel Prospetto 3, permettono di calcolare l'errore relativo di un generica stima di frequenza assoluta o relativa mediante due procedimenti che risultano di facile applicazione, anche se conducono a risultati meno precisi di quelli ottenibili mediante l'espressione (12).
Il primo metodo consiste nel cercare nella colonna (2) del prospetto il livello di stima che più si avvicina alla stima di interesse $\hat{\hat{X}}_d^*$; l'errore relativo corrispondente, $\hat{\varepsilon}(\hat{\hat{X}}_d^*)$ , si trova sulla stessa riga della colonna (3).

Nel secondo metodo, l'errore campionario della stima $\hat{\hat{X}}_d^*$ , si ricava mediante la seguente espressione:

$$\hat{\varepsilon}(\hat{\hat{X}}_d^*) = \hat{\varepsilon}(\hat{\hat{X}}_d^{k-1}) + \frac{\hat{\varepsilon}(\hat{\hat{X}}_d^{k-1}) - \hat{\varepsilon}(\hat{\hat{X}}_d^{k})}{\hat{\hat{X}}_d^{k} - \hat{\hat{X}}_d^{k-1}} (\hat{\hat{X}}_d^{k} - \hat{\hat{X}}_d^{k-1}) \qquad (13)$$

dove $\hat{\hat{X}}_d^{k-1}$ e $\hat{\hat{X}}_d^{k}$ sono i valori più vicini delle stime, letti in colonna (2) del prospetto, entro i quali è compresa la stima $\hat{\hat{X}}_d^*$ e $\hat{\varepsilon}(\hat{\hat{X}}_d^{k-1})$ e $\hat{\varepsilon}(\hat{\hat{X}}_d^{k})$ sono i corrispondenti errori relativi letti sul prospetto.

*Prospetto 3. Valori interpolati degli errori campionari delle stime riferite ai maturi per totale Italia e ripartizioni geografiche. Anno 1995*

| ***Perc.*** | Ripartizioni | | | | | | | | | | Totale Italia | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Nord-Ovest | | Nord-Est | | Centro | | Sud | | Isole | | | |
| | Stima valore assol. | Errore relat. perc. | Stima valore assol. | Errore relat. perc. | Stima valore assol. | Errore relat. perc. | Stima valore assol. | Errore relat. perc. | Stima valore assol. | Errore relat. perc. | Stima valore assol. | Errore relat. perc. |
| ***1*** | 1085 | 23,6 | 815 | 19,9 | 1002 | 20,3 | 1385 | 17,6 | 617 | 24,5 | 4903 | 10,2 |
| ***2*** | 2170 | 15,2 | 1630 | 13,3 | 2003 | 13,4 | 2769 | 11,7 | 1234 | 16,6 | 9807 | 6,6 |
| ***3*** | 3254 | 11,7 | 2446 | 10,5 | 3005 | 10,6 | 4154 | 9,2 | 1851 | 13,3 | 14710 | 5,2 |
| ***4*** | 4339 | 9,7 | 3261 | 8,8 | 4007 | 8,9 | 5539 | 7,8 | 2469 | 11,3 | 19614 | 4,3 |
| ***5*** | 5424 | 8,5 | 4076 | 7,8 | 5009 | 7,8 | 6923 | 6,8 | 3086 | 10,0 | 24517 | 3,8 |
| ***6*** | 6509 | 7,5 | 4891 | 7,0 | 6010 | 7,0 | 8308 | 6,1 | 3703 | 9,0 | 29421 | 3,4 |
| ***7*** | 7594 | 6,8 | 5706 | 6,4 | 7012 | 6,4 | 9692 | 5,6 | 4320 | 8,3 | 34324 | 3,1 |
| ***8*** | 8678 | 6,3 | 6522 | 5,9 | 8014 | 5,9 | 11077 | 5,2 | 4937 | 7,7 | 39228 | 2,8 |
| ***9*** | 9763 | 5,8 | 7337 | 5,5 | 9015 | 5,5 | 12462 | 4,8 | 5554 | 7,2 | 44131 | 2,6 |
| ***10*** | 10848 | 5,4 | 8152 | 5,2 | 10017 | 5,2 | 13846 | 4,5 | 6171 | 6,8 | 49035 | 2,4 |
| ***15*** | 16272 | 4,2 | 12228 | 4,1 | 15026 | 4,1 | 20769 | 3,6 | 9257 | 5,4 | 73552 | 1,9 |
| ***20*** | 21696 | 3,5 | 16304 | 3,4 | 20034 | 3,4 | 27693 | 3,0 | 12343 | 4,6 | 98070 | 1,6 |
| ***25*** | 27120 | 3,0 | 20380 | 3,0 | 25043 | 3,0 | 34616 | 2,6 | 15428 | 4,1 | 122587 | 1,4 |
| ***30*** | 32544 | 2,7 | 24456 | 2,7 | 30052 | 2,7 | 41539 | 2,4 | 18514 | 3,7 | 147104 | 1,2 |
| ***35*** | 37968 | 2,5 | 28532 | 2,5 | 35060 | 2,5 | 48462 | 2,2 | 21600 | 3,4 | 171622 | 1,1 |
| ***40*** | 43392 | 2,3 | 32608 | 2,3 | 40069 | 2,3 | 55385 | 2,0 | 24685 | 3,1 | 196139 | 1,0 |
| ***45*** | 48816 | 2,1 | 36684 | 2,1 | 45077 | 2,1 | 62308 | 1,9 | 27771 | 2,9 | 220657 | 1,0 |
| ***50*** | 54240 | 2,0 | 40761 | 2,0 | 50086 | 2,0 | 9232 | 1,8 | 30857 | 2,8 | 245174 | 0,9 |

## METODOLOGIA PER LA STIMA DEL RISCHIO DI VIOLAZIONE[2]

L'Istat adotta il metodo presentato in Crescenzi (1992) e Coccia (1992) e in ulteriori studi approfonditi, Biggeri e Zannella (1991).
Secondo tale modello, sotto l'ipotesi di indipendenza tra le unità della collezione campionaria, la probabilità che si verifichi l'identificazione di una unità Y nella collezione campionaria è data dal prodotto di una serie di probabilità qui di seguito elencate.

1. Pr(U) = Pr(Y sia un caso unico nella popolazione)= $f_U$, ossia il rapporto tra il numero di casi unici U nella popolazione e la numerosità N della popolazione stessa. Il calcolo di questa quantità è possibile solo nel caso in cui siano disponibili tutti i dati della popolazione. Nel caso più classico in cui siano disponibili i soli dati relativi al campione sarà necessario determinare una stima del numero di casi unici nella popolazione sulla base degli stessi dati campionari. In letteratura sono stati proposti diversi modelli di stima, in particolare in Istat la stima del numero di casi unici $N_u$ nella popolazione, dato il numero di casi unici $n_u$ nel campione, viene effettuata utilizzando un modello proposto da Crescenzi (1992) che si basa sulla combinazione di una distribuzione binomiale negativa con una distribuzione Gamma:

$$N_u = n_u [ 1 + \gamma \log (N/n)]^{-(\alpha+1)}$$

   dove n è la numerosità del campione e N quella della popolazione di interesse e $\alpha$ e $\gamma$ sono i parametri del modello;

2. Pr(C | U) = Pr( Y $\in$ collezione campionaria | U ) che, nell'ipotesi di indipendenza sopra menzionata, è pari al tasso di campionamento, $f_C = \frac{n}{N}$, della collezione campionaria che viene rilasciata;

3. -Pr(A | C, U) = Pr(Y$\in$ archivio dell'utilizzatore | C, U) che è pari a $f_A = \frac{m}{N}$ se si suppone che l'utilizzatore dispone di un archivio con m unità;

4. Pr(Q | A, C, U) = Pr(codifica delle variabili chiave relative a Y nella collezione campionaria sia uguale a quella presente nell'archivio esterno | A, C, U) che è legata alla qualità delle variabili e viene indicata tramite $f_Q$ ; essa sarà pari a 1 solo nei casi in cui la combinazione delle variabili chiave evidenzi un caso unico di pubblico dominio o nel caso in cui l'utilizzatore abbia a disposizione un archivio pari all'intera popolazione.

5. Infine si deve tener conto che in realtà una identificazione si verifica solo se esiste una volontà di identificare un individuo nel file pubblicato.

La probabilità di identificazione può essere quindi vista come:

$$\Pr(\text{identificazione}) = \Pr(\text{tentativo}) * \Pr(\text{identificazione}|\text{tentativo})$$

ovvero

$$\Pr(\text{identificazione}) = f_D * \Pr(\text{identificazione}|\text{tentativo}) = f_D f_U f_C f_A f_Q,$$

dove $f_D$ è la probabilità che l'utilizzatore tenti l'identificazione.

[2] Di Alessandra Capobianchi

Tale probabilità di identificazione può costituire la base per una misura del rischio di violazione della riservatezza di una collezione campionaria da rilasciare. Se indichiamo con X la variabile casuale che rappresenta il numero di identificazioni, Biggeri e Zannella (1991) ipotizzano che la distribuzione di X può essere approssimata dalla distribuzione di Poisson con parametro $\lambda = nf_U f_A f_Q$. Il valore atteso del numero di unità identificabili

E(X)= $nf_U f_A f_Q f_D$=N $f_C f_U f_A f_Q f_D$,

rapportato alla numerosità totale *n* della collezione campionaria costituisce una possibile misura del rischio di violazione RV1= E(X)/n, solitamente espressa in termini di unità campionarie necessarie per avere una identificazione. Valori di RV1 inferiori ad 1 ogni 1.000.000 sono spesso ritenuti altamente protettivi.

Nel caso particolare del file per l'indagine sui percorsi post-diploma dei maturi, sono state analizzate diverse possibili protezioni. Quella adottata comporta un rischio di violazione al di sotto della soglia stabilita dall'Istat.

**BIBLIOGRAFIA**

Biggeri, L. e Zannella F. (1991) Release of microdata and statistical disclosure control in the new national system of Italy: main problems, some technical solutions, experiments, *Bullettin of the International Statistical Institute, Proceedings*, Tome LIV, Book 1, 1-25.

Coccia, G. (1992) Disclosure risk in Italian current population surveys. In *International Seminar on Statistical Confidentiality*, Dublin, 1992.

Crescenzi F. (1992) Estimating population uniques:methodological proposals and applications on Italian census data. In *International Seminar on Statistical Confidentiality*, Dublin, 1992.