Data Journalism

Estrazione dati da PDF

Angelica Lo Duca
angelica.loduca@iit.cnr.it

Estrazione manuale

- Si può utilizzare con tabelle singole e semplici
- Strumenti necessari
  - editor di testo
  - foglio di calcolo per la verifica
- Esempio
  - https://www.epicentro.iss.it/coronavirus/bollettino/Infografica_22marzo%20ITA.pdf

Sorveglianza Integrata COVID-19 in Italia

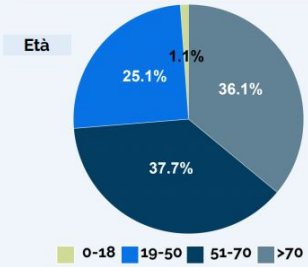
(Ordinanza n. 640 del 27/02/2020)

AGGIORNAMENTO 22 marzo 2020

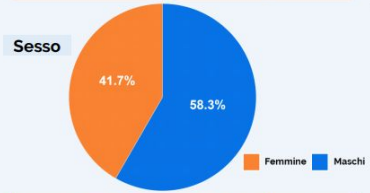
52.796 casi di COVID-19* di cui:

4.824 operatori sanitari $

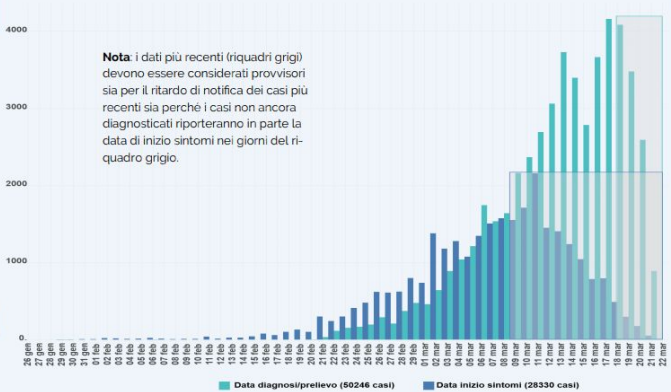
4.465 deceduti



Età mediana dei casi: **63 anni**

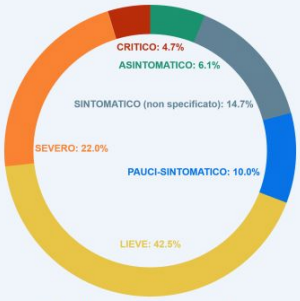


Fascia d'età (anni)	Deceduti [n (%)]	Letalità (%)
0-9	0 (0%)	0%
10-19	0 (0%)	0%
20-29	0 (0%)	0%
30-39	12 (0.3%)	0.3%
40-49	38 (0.9%)	0.6%
50-59	138 (3.1%)	1.3%
60-69	469 (10.5%)	5%
70-79	1585 (35.5%)	15.3%
80-89	1806 (40.4%)	23.3%
>=90	416 (9.3%)	24.1%
Non noto	1 (0%)	0.3%
Totale	4465 (100%)	8.5%



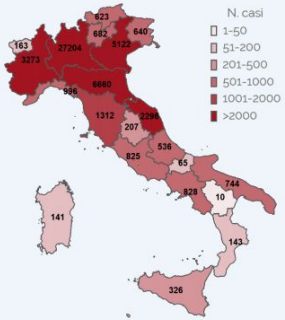
Nota: i dati più recenti (riquadri grigi) devono essere considerati provvisori sia per il ritardo di notifica dei casi più recenti sia perché i casi non ancora diagnosticati riporteranno in parte la data di inizio sintomi nei giorni del riquadro grigio.

Sono risultati positivi il **99%** dei campioni processati dal Laboratorio nazionale di riferimento presso l'Istituto Superiore di Sanità



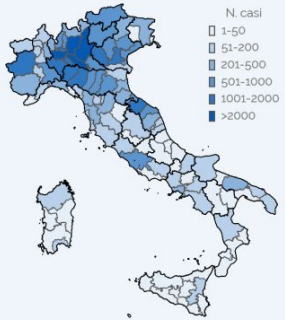
Dato disponibile per 16.335 casi

Numero totale di casi di COVID-19 diagnosticati dai laboratori regionali di riferimento



per Regione/PA di diagnosi

(dato disponibile per 52.796)



per Provincia di domicilio/residenza

(dato disponibile per 50.235)

*La definizione internazionale di caso prevede che venga considerata caso confermato una persona con una conferma di laboratorio del virus che causa COVID-19 a prescindere dai segni e sintomi clinici

https://www.ecdc.europa.eu/en/case-definition-and-european-surveillance-human-infection-novel-coronavirus-2019-ncov

*Il flusso ISS raccoglie dati individuali di casi con test positivo per SARS-COV-2 diagnosticati dalle Regioni/PPAA. I dati possono differire dai dati forniti dal Ministero della Salute e dalla Protezione Civile che raccolgono dati aggregati. $ Dato non riferito al luogo di esposizione ma alla professione.

A cura di: Task force COVID-19 del Dipartimento Malattie Infettive e Servizio di Informatica
Istituto Superiore di Sanità

Procedimento

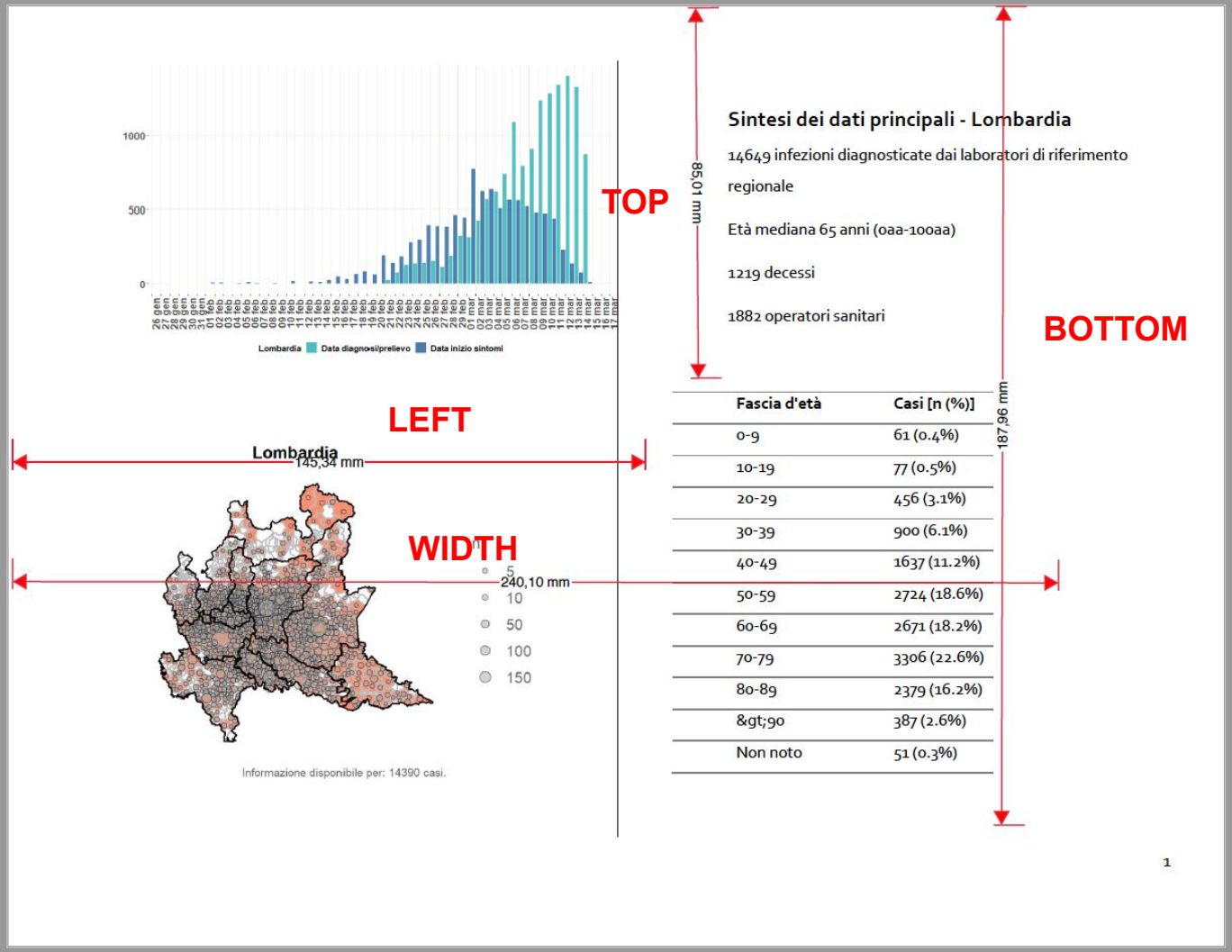
- Con il mouse selezionare tutte le celle della tabella
- Copiare la selezione
- Incollare in un editor di testo
- Sostituire il carattere di separazione (ad esempio lo spazio) con una virgola o un punto e virgola
- Correggere eventuali errori
- Salvare il file come csv
- Aprire il file con un programma per i fogli di calcolo (ad esempio Microsoft Excel)

Estrazione automatica

- Si usa per tabelle complesse o ripetute
- Strumenti richiesti
  - python
  - libreria tabula-py
    - pip install tabula-py
    - pip3 install tabula-py

Passi

Selezionare i margini della tabella attraverso un visualizzatore di PDF, come ad esempio Adobe Reader



Passi (cont.)

Selezionare le pagine da cui estrarre le tabelle

3,5,6,8,9,10,12,14,16,18,22,24,26,28,30,32,34,36,38,40

Utilizzare la libreria tabula-py

https://tabula-py.readthedocs.io/en/latest/