Report Statistica PCTO

Report creato da Di Vanni Tommaso, Rizzi Luca Ferruccio e Sansottera Sarah
In data 03/02/2025 a 07/02/2025

Introduzione

La statistica è la scienza che ha per oggetto lo studio dei fenomeni collettivi suscettibili di misurazione e di descrizione quantitativa, basandosi sulla raccolta di un grande numero di dati inerenti ai fenomeni in esame, e partendo da ipotesi più o meno direttamente suggerite dall'esperienza o da analogie con altri fenomeni già noti, mediante l'applicazione di metodi matematici fondati sul calcolo delle probabilità, si perviene alla formulazione di leggi di media che governano tali fenomeni, dette leggi statistiche. Spesso la raccolta dei dati viene limitata a un campione più ristretto, opportunamente predeterminato in modo da rappresentare il più fedelmente possibile le caratteristiche generali.

Le tecniche della statistica trovano applicazione nelle altre scienze sperimentali, in particolare nell'analisi statistica dei dati, sia in riferimento alla natura aleatoria (quindi casuale) dei risultati delle misure, in quanto affetti da errori, sia in presenza di fenomeni intrinsecamente aleatori per cui le proprietà dei fenomeni studiati devono essere dedotte, attraverso un procedimento di inferenza statistica, dalle proprietà di un campione statistico, costituito dagli eventi effettivamente osservati. Le applicazioni della statistica sono principalmente nella demografia, che studia la popolazione umana (ammontare e composizione), nella fisica, che studia la meccanica classica e quantistica, nell'economia, in particolare nell'econometria, e nella biostatica e medicina, che traduce i dati clinici e di laboratorio in espressioni quantitative.

I fenomeni statistici si dividono in qualitativi, che si identificano tramite attributi e che si misurano tramite le scale nominali (senza ordinamento) e le scale ordinali (in ordine naturale), e in quantitativi, che si identificano tramite numeri e che possono essere discreti (numeri naturali) o continui (numeri reali).
La dimensione temporale/spaziale può essere *cross-sectional* (dati di un preciso istante temporale), *longitudinal* (stesso soggetto in differenti momenti) o *panel* (dati di differenti stati temporali).
La copertura del dato può essere amministrativo (raccolto sulla totalità della popolazione) o da indagine o *survey* (raccolto su una parte della popolazione).

La Classificazione dei consumi individuali secondo lo scopo (COICOP) ha creato una struttura gerarchica dei beni che compongono la spesa totale di ogni famiglia: divisione; gruppo; classe; sottoclasse.
L'indagine dell'ISTAT, che si svolge ogni anno e che segue la prima voce di questa gerarchia con un questionario, è di tipo campionario e coinvolge circa 33 mila famiglie residenti in 540 comuni italiani.
In base al reddito si disegna il decile di reddito, che suddivide quindi le famiglie intervistate in 10 categorie. Da questo si comprende chi tra gli intervistati sia più ricco (9-10) e chi sia più povero (1-2).

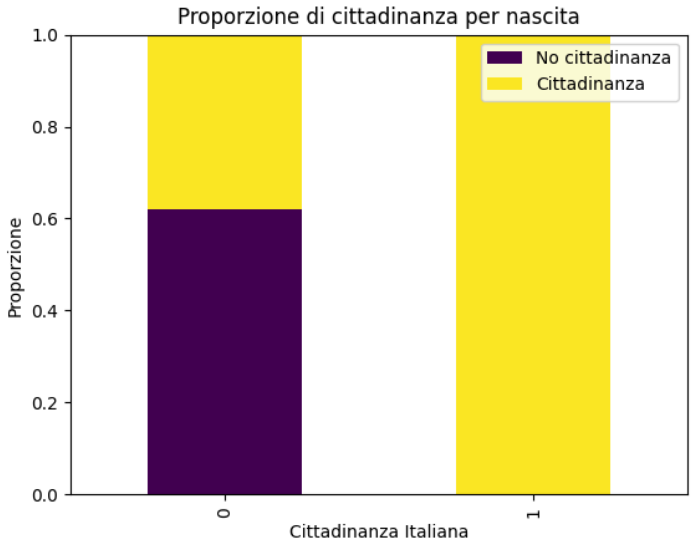
Altri dati rilevanti sono quelli riguardanti le spese di ogni divisione: la spesa 1 riguarda i prodotti alimentari e le bevande analcoliche; la spesa 2 le bevande alcoliche e i tabacchi; la spesa 3 abbigliamento e calzature; la spesa 4 abitazione, acqua, elettricità, gas e altri combustibili, con interventi di ristrutturazione; la spesa 5 mobili, articoli e servizi per la casa; la spesa 6 salute; la spesa 7 trasporti; la spesa 8 informazione e comunicazione; la spesa 9 ricreazione, sport e cultura; la spesa 10 istruzione; la spesa 11 servizi di ristorazione e di alloggio; la spesa 12 servizi assicurativi e finanziari, beni e servizi per la cura della persona, servizi di protezione sociale e altri beni e servizi.

Procedimento - Python e Dataset

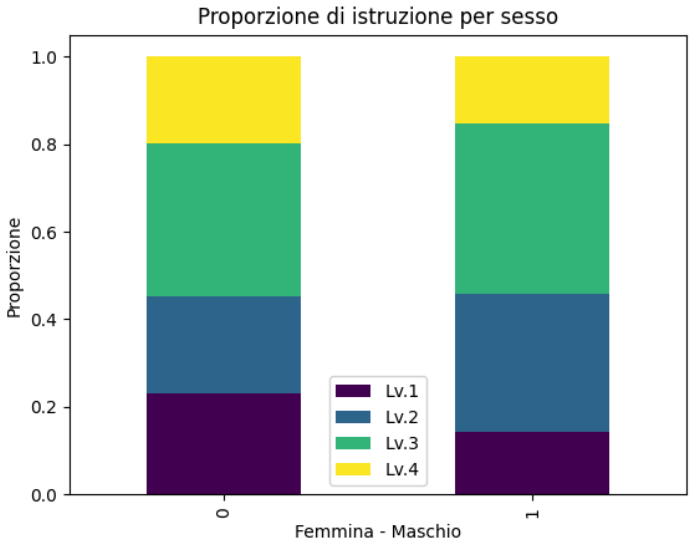
Siamo stati introdotti alla base della statistica dal Prof. Riganti e abbiamo fatto una ricerca sulle variazioni dei prezzi al consumo delle principali 12 categorie di spesa. Queste variazioni sono state studiate con i dati presi durante gli anni del 2019, 2020, 2021, e 2022.

La ricerca è stata svolta cercando i report dell'ISTAT sia manualmente, sia con l'aiuto dell'intelligenza artificiale, una volta raccolti tutti i report sono stati letti e studiati. Siamo stati successivamente introdotti a Python, un linguaggio di programmazione facile e molto vario con il quale si possono analizzare dati e programmare applicazioni. Abbiamo imparato poi cosa sono Numpy e Pandas caricando dei dataset: sono un open source contenente i metodi per analizzare i dati.
Come ultimo passaggio abbiamo visualizzato i dati di un dataset contenente informazioni di più di centomila famiglie. Sono raccolte in una tabella excel, dove le righe rappresentano le famiglie e le colonne le tipologie dei dati, in particolare: il reddito (valutato in decile), l'ampiezza del nucleo familiare, l'eventuale presenza di minori o anziani, le caratteristiche della casa in cui abitano, come ad esempio la metratura e il numero di stanze, e le spese divise nelle 12 categorie principali.
Dopo aver appreso come produrre le varie tipologie di grafici abbiamo svolto un'indagine sul notebook "colab" nel quale, dopo aver inserito i dati dalle nostre librerie, sono stati sviluppati numerosi grafici per un'analisi univariata e un'analisi multivariata. Queste analisi ci sono state utili per osservare delle correlazioni tra dei valori, seppur piccole.
Molti di questi grafici sono stati scartati per poi analizzare e commentare quelli con il risultato più interessante, confrontandoli con ciò che avevamo appreso dai report.

Grafici e Commenti

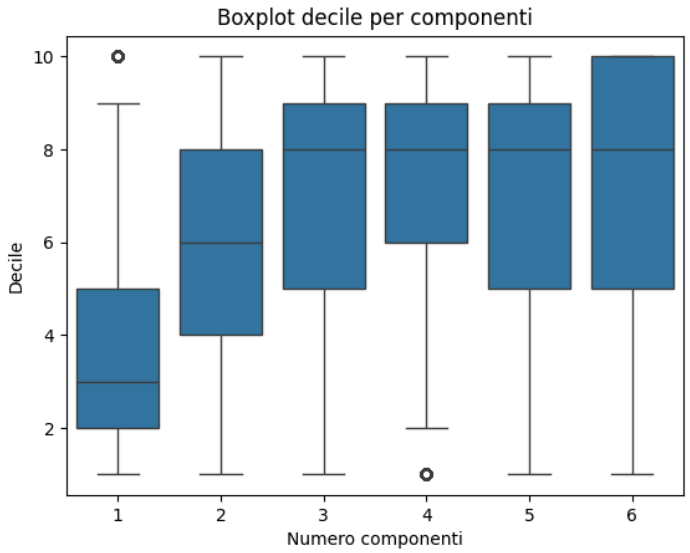


Questo grafico ha un'analisi multivariata, infatti le variabili in questione sono la cittadinanza italiana e il luogo di nascita, si può vedere nella colonna di sinistra che rappresenta i non nati in italia che circa il 60% di coloro che sono nati al di fuori dello stato italiano non hanno la cittadinanza mentre il 40% si. Nella colonna di destra si vede che coloro (partecipanti al sondaggio) che sono nati in italia posseggono tutti la cittadinanza, anche perche questo sondaggio è stato fatto verso le famiglie residenti in Italia.

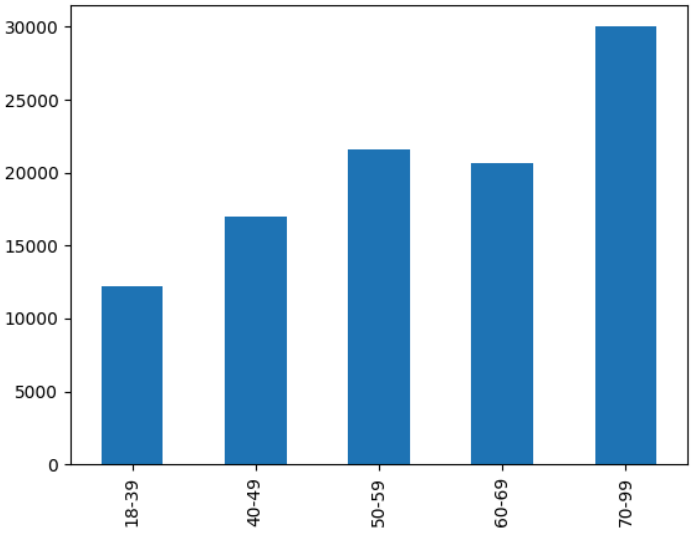


Questo grafico ha un'analisi multivariata: ha come dati il sesso e il livello di istruzione degli intervistati. La colonna a sinistra rappresenta la totalità della popolazione femminile intervistata mentre la colonna a destra quella maschile.
I soggetti femminili di livello di istruzione 1 (circa 20%) sono di più in proporzione dei soggetti maschili (circa 10%); i soggetti femminili di livello di istruzione 2 (circa 20%) sono meno in proporzione dei soggetti maschili (circa 30%); i soggetti femminili di livello di istruzione 3 (circa 35%) sono meno in proporzione dei soggetti maschili (circa 40%); i soggetti femminili di livello di istruzione 4 (circa 20%) sono di più in proporzione dei soggetti maschili (circa

15%). Le donne quindi sono la maggior parte del livello 3, poi del livello 1 e del 2/4. Gli uomini sono la maggior parte del livello 3, poi del livello 2 e del 4/1.

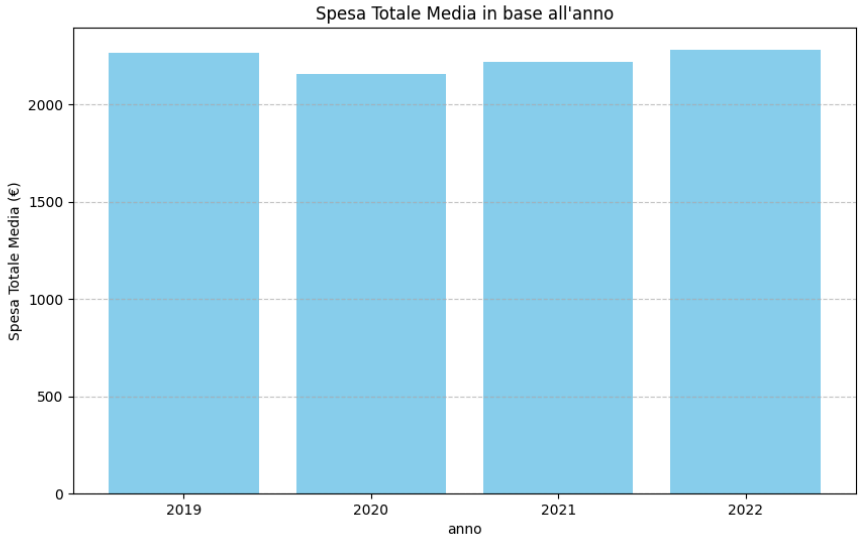


Questo grafico ha un analisi multivariata: ha come dati il numero di componenti e il decile di reddito degli intervistati. E' un grafico *box plot*, che è diviso in quattro quartili dove la linea della mediana divide il rettangolo. I cerchi sono gli outlier e rappresentano casi unici fuori dalla moda. La mediana del numero di componenti 1 è tra il secondo e il quarto decile; la mediana del numero di componenti 2 è al livello del sesto decile; le mediane del numero di componenti 3, 4, 5 e 6 sono al livello 8.
Quindi mediamente chi vive da solo ha un reddito più basso di chi ha una famiglia numerosa.

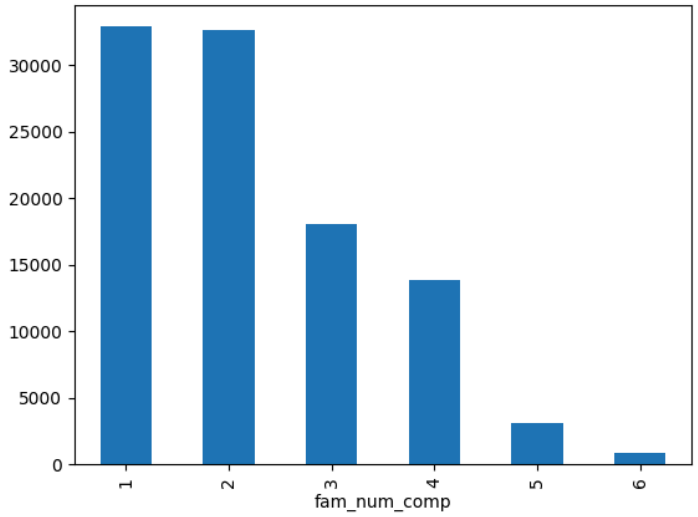


In questo grafico si può osservare la quantità di persone che hanno partecipato al sondaggio per età,la maggior parte della popolazione che ha partecipato ha un'età compresa tra i 40 e 69 anni di età che rappresenta l'età media italiana. Una grande quantità di persone è nell'ultima fascia d'età, perciò anche considerando i dati delle altre fasce d'età si conclude al

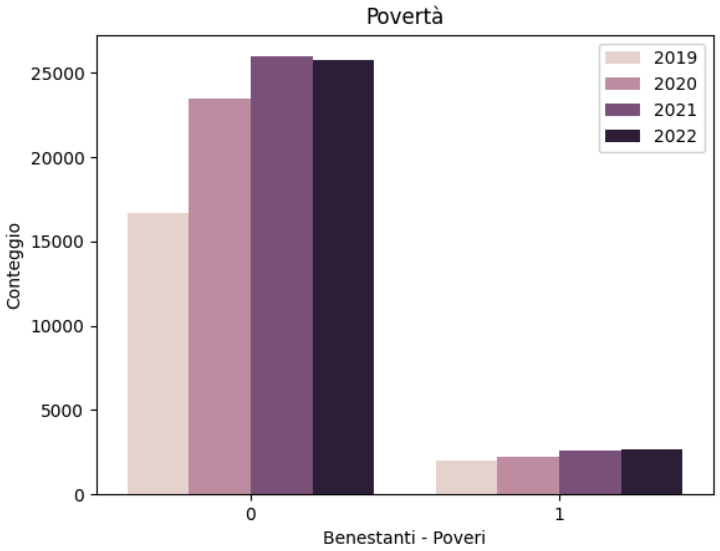
fatto che la maggior parte delle persone che hanno partecipato a questa indagine hanno un'età avanzata.



In questo grafico si può vedere la media della spesa totale nei 4 anni. Si può notare che la media della spesa totale ha tenuto lo stesso valore nel corso di questi anni. Ciò però era inaspettato, data l'inflazione e l'aumento delle bollette a causa della guerra in Ucraina. In conclusione possiamo dedurre che la gente ha rinunciato alla spesa di alcuni beni non primari, per mantenere la stessa spesa totale.



In questo grafico si può vedere il conteggio delle famiglie con un determinato numero di componenti.

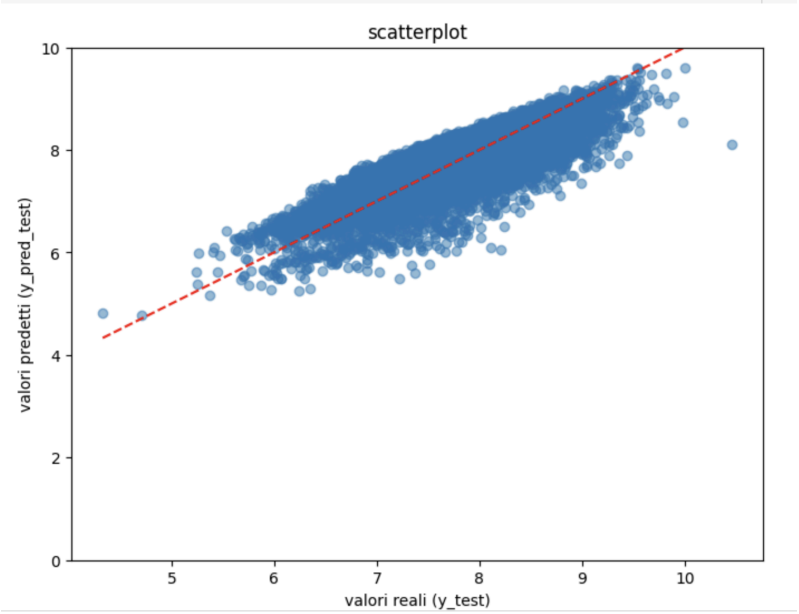


Questo grafico, indica la quantità di poveri e benestanti con il trascorrere degli anni, possiamo notare che le colonne riferite ai poveri crescendo anno per anno, forse perché la crisi ha indebolito evidentemente l'economia impedendone un recupero veloce, perciò molte persone, magari già in difficoltà economiche, vittime di questa economia disastrata si sono dirette verso la povertà. Appunto nell'anno precedente alla crisi si è verificato un aumento dei benestanti notevole (circa 40%), però questo trend si è rallentato fino a stopparsi durante gli anni della crisi; tra il 2020 e il 2021 si è verificato un lieve aumento dei benestanti da dopo il 2021 essi sono diminuiti. Questo è comprensibile anche con l'aumento dei soggetti in povertà.

Procedimento 2 - Grafici di Regressione

Il giorno seguente il Prof Riganti ci ha introdotto come prevedere una spesa in base a quali variabili, che possono essere continue e discrete, si consideravano. Ci ha spiegato come costruire un grafico, nel quale poi viene inserita una retta. Grazie a questa retta, bisettrice rappresentante la linea normale, tramite calcoli complessi basati sulla distanza punto retta possiamo determinare l'R2. l'R2 è un coefficiente il cui valore è compreso tra 0 e 1, che ci dice quanto il nostro grafico è preciso. Più l'R2 è vicino a 1, più il grafico sarà preciso. In seguito abbiamo appreso come disegnare questi grafici in Colab. Abbiamo fatto diverse prove, cambiando la variabile spesa nel target e le variabili predittive, fino a trovare l'R2 più vicino a 1 possibile.

Grafici e Commenti



Questo grafico rappresenta la previsione della spesa totale al variare di diverse componenti. Le variabili prese in considerazione sono: Numero di componenti, Decile di reddito di appartenenza, Spesa in prodotti Alimentare e Bevande Analcoliche, Crisi economica in corso e spesa nelle bollette. Il grafico presenta un R2 (coefficiente) del valore di 0,72. Ciò significa che il risultato della previsione è molto preciso.
Abbiamo deciso di usare la variabile del numero di componenti perché un numero di componenti maggiori, che può comprendere minori e anziani, porta a una spesa maggiore, e impatta sulle spese fatte. Poi abbiamo utilizzato il decile di reddito che influenza molto il valore totale, poiché limita la possibilità di spesa. Successivamente abbiamo scelto la spesa in prodotti alimentari e bollette, perché essendo spese primarie rappresentano le spese con percentuale maggiore all'interno della spesa totale.

Conclusioni

In conclusione abbiamo programmato una web app che date delle variabili esegue una predizione sulla base dell'ultimo grafico mostrato e commentato. Il programma è stato creato interamente con Python. All'interno della web app si trova pure un'introduzione alla statistica e il nostro progetto. Una volta pubblicato questa web app online chiunque avrà la possibilità di avere una previsione sulla propria spesa totale in base alle variabili da lui inserite.

Per quanto riguarda il dataset, i dati e i grafici ci dicono che nonostante l'inflazione la spesa totale media ha tenuto lo stesso valore nel corso dei 4 anni. Ciò significa che i cittadini italiani hanno deciso di rinunciare a dei beni secondari per la spesa di beni primari, come le bollette o prodotti alimentari.