Laboratorio Stad

Statistica, Tecnologia, Analisi dei Dati

Prof. Roberta Siciliano

Di cosa si tratta?

Statistica, Tecnologia, Analisi dei Dati (STAD) sono parole chiave per la lettura della realtà, come questa si manifesta sia nelle espressioni qualitative (attributi, appartenenza a classi o gruppi, etichette, categorie, etc.) sia in quelle quantitative (numeri, misurazioni, valori, etc.), percepite e riconosciute dall'uomo nei diversi ambiti scientifici e applicativi, con la finalità di conoscere per innovare.

La derivazione etimologica di "**Statistica"** è "status" (i.e., paese, principato, regno), nel '700 il filosofo, storico, economista Gottfried Achenwall definisce la statistica come *"scienza deputata a raccogliere dati utili per governare meglio*". Il ruolo dello statistico non è solo di produrre e analizzare le statistiche economiche utili allo Stato, bensì di osservare la realtà - *"ciò che è"* - in economia e in tanti altri ambiti applicativi e/o fenomenici (psicologia, medicina, fisica, ingegneria, etc.), con l'ambizione di "scoprire" fatti seguendo un approccio esplorativo, altresì di "giustificare" teorie e leggi seguendo un approccio confermativo, in un processo di continuo apprendimento della realtà e delle sue manifestazioni. L'osservazione della realtà non può prescindere dalla tecnologia e dall'analisi dei dati. Il termine **"tecnologia"**, combinando le due parole greche *"technè"* (i.e., abilità concreta, il fare) e *"logìa"* (i.e., discorso o ragionamento sull'arte, il saper fare), si riferisce alla razionalizzazione di un processo di apprendimento e alla costruzione di un progetto, ossia di una strategia per passare dalla teoria alla pratica. **"Analisi dei dati"**, con le due parole "*anàlysis*" (i.e., soluzione) e "*dato*" (i.e., quantità nota), si riferisce al processo di apprendimento maturato con l'elaborazione e trasformazione dei dati grezzi (input) in informazione utile per uno scopo (output), restituendo con l'analisi un quadro sintetico e organizzato dei risultati ottenuti e dei collegamenti con l'esperienza pregressa, altresì fornendo una soluzione a un problema assegnato che implica una qualche azione o conseguenza (outcome).

Lo statistico, per la comprensione dei fatti e per la conferma delle teorie, si avvale di ragionamenti e di abilità concrete, per sfruttare al meglio i dati disponibili quale risultato dell'osservazione della realtà. In altri termini, combinando statistica e tecnologia, si può definire una metodologia statistica, formulando un metodo, le assunzioni, le proprietà, le condizioni di applicazione, etc., il fine ultimo è la sua applicazione a un problema reale attraverso l'analisi dei dati.

Oggi giorno le abilità concrete sono rappresentate dai progressi tecnologici nell'acquisizione ed elaborazione di dati in formato digitale promuovendo *"il desiderio di estrarre conoscenza dall'enorme flusso di dati digitali che oggi siamo in grado di immagazzinare ed elaborare grazie ai moderni strumenti tecnologici".* La statistica moderna rappresenta la risposta naturale a questo nuovo fabbisogno informativo.

A cosa serve?

Le metodologie statistiche trovano applicazione in tutte le aree riguardanti la vita dell'uomo. Queste applicazioni non coinvolgono, unicamente, contesti di ricerca specifica (come lo studio delle immagini dei corpi astrali, i database genetici, ecc.), ma in larga parte anche situazioni di vita quotidiana (si pensi ad esempio all'analisi dei dati collezionati dai supermercati e dalle compagnie che gestiscono le carte di credito, i dettagli delle bollette registrati dalle compagnie telefoniche, l'analisi delle statistiche sui referti

medici, ecc.).

Nelle applicazioni socio-economiche la statistica si focalizza principalmente sull'analisi di dati osservazionali rilevati attraverso indagini campionarie che prevedono la somministrazione di questionari ad un campione statistico rappresentativo della popolazione di riferimento. Ne rappresentano alcuni esempi, gli studi periodici svolti sulla popolazione per descriverne le evoluzioni delle abitudini di consumo e sul ruolo svolto dalle famiglie nell'accumulazione del risparmio, gli studi territoriali per individuare gli elementi di criticità delle filiere produttive, le analisi di *customer satisfaction* effettuate dalle aziende sulla propria clientela, le ricerche di mercato per individuare i segmenti di mercato da associare a tipologie di prodotti/servizi. In ambito aziendale, il management può sfruttare al meglio le informazioni derivanti dal sistema informativo aziendale che raccoglie dati contabili e finanziari, altresì dati sulla produzione e la logistica, la commercializzazione dei prodotti etc. Le banche e gli istituti finanziari svolgono le analisi di *credit-scoring* e di *rating finanziario* con i metodi statistici.

Nelle indagini campionarie, il problema chiave è rappresentato dalla corretta definizione delle diverse fasi che caratterizzano l'indagine campionaria, dalla raccolta dei dati, passando per la costruzione del questionario attraverso l'individuazione del set di domande che meglio consenta di rilevare le principali caratteristiche del fenomeno oggetto di studio, per giungere poi alla fase di elaborazione e interpretazione dei risultati ottenuti.

Nell'ambito delle scienze della vita le applicazioni della statistica riguardano principalmente le ricerche nel campo della medicina, della farmacia, della biologia e della veterinaria. In questi ambiti la statistica trova ampio utilizzo quale scienza a supporto della corretta formulazione e verifica e/o confutazione di ipotesi riguardanti l'efficacia di un trattamento, l'esistenza di legami causa-effetto tra due o più fenomeni, l'esistenza di gruppi caratteristici (cluster) nell'insieme oggetto di studio, ecc.. Si immagini, ad esempio, la sperimentazione di un nuovo farmaco in cui la metodologia statistica consente di valutare l'esistenza di un effetto migliorativo significativo del trattamento rispetto alla terapia classica, oppure lo studio mirato alla verifica di una relazione causale tra un fattore come il fumo e il rischio di contrarre una particolare malattia.

In queste scienze la statistica ha anche consentito di accrescere la rapidità e l'accuratezza dei sistemi di diagnostica. Oggi è possibile valutare la probabilità di malformazioni genetiche nel feto già alla 12esima settimana di gestazione grazie alla valutazione aggregata dei risultati di un'ecografia (la translucenza nucale) e di un prelievo ematico. La combinazione dei risultati è valutata attraverso un test statistico che genera quale output una misurazione della rischiosità di malformazioni. In questo modo è possibile evitare indagini invasive e molto rischiose quali l'amniocentesi per quei casi in cui il rischio si attesta su livelli ritenuti tollerabili. Negli esempi forniti è evidente come il dato tipico utilizzato dalle metodologie statistiche è rappresentato dal dato sperimentale, cioè da misure rilevate attraverso una sperimentazione che fa uso di strumenti di misurazione specifici.

Nell'ambito delle scienze dell'ingegneria e delle tecnologie le applicazioni della statistica riguardano numerosissimi temi quali i trasporti, l'ambiente, la meccanica, l'energia, l'informatica, ecc. Possono essere impiegati strumenti statistici sia di tipo classico, test di ipotesi e stima ad intervalli, sia tecniche di analisi multivariata basate sull'uso intensivo del calcolatore elettronico. Sono un esempio del primo tipo di analisi, gli studi di resistenza dei materiali al peso o alle sollecitazioni esterne (terremoti, agenti atmosferici, ecc.), mentre sono un esempio di analisi multivariata, l'utilizzo delle tecniche di classificazione ad albero per individuare i fattori caratterizzanti il rischio di incidenti stradali.

Altro esempio di supporto fornito dalla statistica, questa volta nell'informatica di uso quotidiano, è rappresentato dai motori di ricerca web (Google, Yahoo, Bing, ecc.). Essi utilizzano degli algoritmi per indicizzare le pagine web e creare uno score delle stesse rispetto ai termini di ricerca che impostiamo nel motore. Questi algoritmi sono basati su un insieme di metodi statistici, il *Text Mining*, che si

occupano di individuare i termini ricorrenti nei documenti tenendo conto del tema del discorso e associare in questo modo gli uni agli altri (che rappresentano le pagine visualizzate come output di una ricerca con una chiave di lettura dei diversi documenti in rete).

Che si fa?

Il laboratorio STAD si sviluppa in quattro incontri per introdurre gli aspetti metodologici fondamentali per l'elaborazione dei dati statistici di tipo osservazionale o sperimentale. L'obiettivo è di trasferire all'allievo la consapevolezza pratica della statistica, tecnologia e analisi dei dati per la comprensione di un problema reale.

Si mostrano le diverse fasi dell'indagine statistica alla luce di un caso studio. Si può far riferimento alle indagini condotte per l'analisi della *tourist satisfaction* nella provincia di Napoli dal gruppo di ricerca STAD dell'Università degli Studi di Napoli Federico II (www.stad.unina.it).

In alternativa, si può definire un caso studio in aula di concerto con gli allievi che partecipano al laboratorio. In passato sono state condotte delle vere e proprie indagini presso gli studenti delle scuole. In tal caso, agli allievi è richiesto lo svolgimento di attività collaterali alla partecipazione agli incontri in aula.

Nel 2013 fu condotta l'indagine sull'uso di Facebook, nel 2014 l'indagine Terra promessa: i ragazzi di oggi: dipendenze e ambizioni. I temi proposti sono stati affrontati dal punto di vista qualitativo in chiave sociologica per la definizione del questionario e dal punto di vista quantitativo svolgendo tutte le operazioni di raccolta e imputazione dei dati, per poi partecipare alla fase di elaborazione e analisi dei risultati conseguiti.

"I keep saying that the sexy job in the next 10 years will be statisticians", ha dichiarato Hal Varian, economista in Google, aggiungendo: *"And I'm not kidding. People think I'm joking, but who would have guessed that computer engineers would have been the sexy job of the 1990s?".*