LA STATISTICA NELL'ERA DELL'INNOVAZIONE

di *Roberta Siciliano*[1], *Massimo Aria*[2], *Antonio D'Ambrosio*[1]

[1]Dipartimento di Ingegneria Industriale,
Università degli Studi di Napoli Federico II

[2]Dipartimento di Scienze Economiche e Statistiche,
Università degli Studi di Napoli Federico II

roberta.siciliano@unina.it; massimo.aria@unina.it; antonio.dambrosio@unina.it

1. L'evoluzione storica della statistica

La statistica nasce con l'uomo e con il suo fabbisogno di conoscenza della realtà che lo circonda: la percezione del mondo esterno, la registrazione di comportamenti, la classificazione di oggetti, la codifica di stili di vita, che accomunano un gruppo, una tribù, un popolo o una nazione, la ricerca di leggi empiriche, l'esigenza di sviluppare una conoscenza qualitativa e quantitativa, una cultura della misurazione per confronti temporali e/o spaziali.

La derivazione etimologica di "statistica" è "status" (i.e., paese, principato, regno), dal latino "statera" (i.e., bilancia), in tedesco "Stadt" (i.e., città), nel '700 il filosofo, storico, economista Gottfried Achenwall definisce la statistica come "*scienza deputata a raccogliere dati utili per governare meglio*", prima ancora lo storico Gerolamo Ghilini introduce la statistica come

> descrizione delle qualità che caratterizzano e degli elementi che compongono uno stato, ossia raccolta e presentazione del quadro numerico dei fatti economici e sociali importanti, ad esempio il movimento della popolazione, della ricchezza nazionale, etc., organizzato e gestito dallo Stato con l'obiettivo di offrire l'agio di prevedere e provvedere.

La statistica nasce con il conteggio dei cittadini e dei loro beni nei primi censimenti della storia, in Egitto, Israele, Antica Roma, si sviluppa con la rilevazione di certe ricorrenze nei rapporti demografici e la costruzione

delle tavole della mortalità o speranza di vita (che descrive, per ogni generazione di nati in un dato anno, l'andamento del numero dei sopravvissuti, dal momento della nascita fino alla morte dell'ultimo). La statistica spesso è associata alle statistiche economiche che descrivono uno stato-paese, ad esempio, il numero dei residenti, la ricchezza di uno Stato in termini di Prodotto Interno Lordo, la forza lavoro espressa come somma del numero di occupati e di persone in cerca di occupazione (disoccupati), il potere d'acquisto della moneta o tasso di inflazione, etc. Carlo Alberto Salustri, detto Trilussa, citava le statistiche in una famosa poesia:

> da li conti che se fanno seconno le statistiche d'adesso risulta che te tocca un pollo all'anno: e, se non entra nelle spese tue, t'entra ne la statistica lo stesso perch'è c'è un antro che ne magna due.

La statistica si afferma come applicazione della matematica ai giochi d'azzardo, il calcolo delle probabilità. Il là fu dato da Fra Luca Pacioli (*Summa de Arithmetica, Geometria, Proportioni et Proportionalità*, 1494) con il problema dei punti o delle parti:

> Due giocatori di pari abilità disputano una serie di partite; vince il gioco chi, per primo, raggiunge un totale di sei vincite. I giocatori, però, devono sospendere il gioco prima che questo abbia termine. Si domanda: se al momento della sospensione un giocatore ha vinto cinque partite e l'altro tre, e la posta in gioco è di 24 euro, come deve essere ripartita tale somma fra i due giocatori in modo tale che la ripartizione sia equa?

Quando si parla di statistica, non si può prescindere dall'empirismo, ossia l'osservazione "sensibile" di una realtà fenomenica per coglierne sinteticamente le caratteristiche qualitative e quantitative. In uno dei dialoghi di Platone, si ritrova il concetto di distribuzione statistica intorno ad una media:

> Credi forse che sia tanto facile trovare un uomo o un cane, o un altro essere qualunque, molto grande o molto piccolo o, che so io, uno molto veloce o molto lento, o molto brutto o molto bello, o tutto bianco o tutto nero? Non ti sei mai accorto che in tutte le cose, gli estremi sono rari mentre gli aspetti intermedi sono frequenti, anzi numerosi?...

Oltre alla media, si evincono anche i concetti di variabilità e mutabilità, ossia l'attitudine del fenomeno ad assumere espressioni numeriche o qualitative diverse nelle diverse unità del collettivo oggetto di studio.

L'immaginazione, o conoscenza sensibile, ha guidato illustri matematici, fisici, astronomi, nei secoli XVII e XVIII, nell'osservazione di alcune leggi

empiriche, desumendo le caratteristiche incognite e il modello teorico alla base del processo generatore dei dati osservati. Il matematico svizzero Jakob Bernoulli definì la distribuzione bernoulliana per descrivere la probabilità di successo in una prova con due esiti possibili e la sua generalizzazione, la distribuzione binomiale, per descrivere la probabilità di ottenere un numero k di successi in n prove bernoulliane. Inoltre, lo stesso Bernoulli definì la legge dei grandi numeri, nota anche come teorema di Bernoulli o legge empirica del caso, per la quale la media di una successione di n fenomeni aleatori, ripetuti in maniera indipendente e nelle medesime condizioni, tende alla media vera della distribuzione del fenomeno al tendere di n all'infinito.

Daniel Bernoulli, nipote di Jakob, oltre agli importanti studi condotti nella fluidodinamica, analizzò per primo dati sensibili sulla diffusione del vaiolo e i relativi dati di mortalità per dimostrare l'efficacia del vaccino, fu l'autore principale di una nuova teoria di misura del rischio, definendo i concetti di avversione al rischio e premio per il rischio (alla base delle teorie moderne di economia finanziaria), di utilità marginale in presenza di risorse scarse (alla base delle teorie neoclassiche in economia), attraverso la formalizzazione del paradosso di San Pietroburgo. Tale paradosso, la cui ideazione si deve in verità al cugino Nicolaus Bernoulli, tratta di un particolare gioco d'azzardo che suggerisce una linea di condotta apparentemente irragionevole ma con una speranza matematica di vincita tendente all'infinito:

> un giocatore punta su uno tra due esiti possibili e, se non vince, raddoppia la giocata fino a quando non vince; si dimostra che la speranza matematica tende all'infinito al crescere del numero di prove, e pertanto la somma da giocare potrebbe essere irragionevolmente un ammontare considerevole.

In statistica, la più nota legge empirica è la curva normale, descritta nel continuo e dalla forma simmetrica campanulare, nota come curva gaussiana, in onore a uno dei suoi scopritori, il matematico, fisico, astronomo tedesco Karl Friedrich Gauss, che nel 1821, a partire dall'osservazione del moto di corpi celesti nello spazio definì la legge degli errori di misurazione delle distanze tra corpi celesti, spiegando il comportamento e l'entità degli errori di misurazione. La curva degli errori si presta bene per descrivere molti fenomeni che presentino come unico elemento fondamentale il concetto di errore "casuale" di misurazione. La formalizzazione matematica della distribuzione normale è stata identificata anni addietro da due illustri studiosi. Il matematico francese Abraham De Moivre, nel 1733, utilizzando l'approssimazione dei termini fattoriali dovuta al matematico scozzese James Stirling, derivò la distribuzione normale come approssimazione della distribuzione binomiale al tendere del numero n di prove all'infinito. Qual-

che anno dopo, nel 1774, il marchese Pierre Simon Laplace, matematico, fisico, astronomo francese, formalizzò l'integrale di Eulero alla base della formulazione matematica ottenuta da De Moivre pur senza svilupparne proprietà e applicazioni. Laplace definì anche una nuova distribuzione di probabilità, per descrivere in un esperimento binomiale che l'esito successivo sia un successo, dati *s* successi osservati in precedenza e un totale di *n* prove. Egli definì la regola di successione rapportando il totale di successi osservati in precedenza più uno con il numero totale di prove effettuate più due, per indicare la probabilità che il prossimo esito sia un successo. Ciò gli ha consentito di dimostrare un'affermazione apparentemente ovvia, *"il sole è sorto e tramontato per miliardi di anni, il sole è tramontato anche stanotte, con un'elevata probabilità il sole domani sorgerà"*. La regola di Laplace è stata ritenuta la motivazione pionieristica del concetto bayesiano della probabilità quale livello di fiducia del verificarsi di un dato evento, avvicinandosi a uno o a zero all'aumentare dell'informazione disponibile, concetto ampiamente ripreso dall'italiano Bruno de Finetti e dall'americano Leonard Savage solo a metà del XX secolo.

La statistica si afferma con il fabbisogno di misurazione e quantificazione della realtà fenomenica, perseguendo l'approccio galileiano della ricerca sperimentale in ambito fisico e medico, con la raccolta di evidenze empiriche e misurabili attraverso l'osservazione e la sperimentazione, la formulazione di ipotesi e teorie più generali da sottoporre al vaglio dell'esperimento per testarne l'efficacia. L'utilizzo dei modelli di probabilità e delle leggi sulla convergenza ha permesso la declinazione del paradigma scientifico della statistica matematica fondato sull'inferenza statistica, ossia del metodo induttivo alla conoscenza, dal particolare al generale.

Precursore degli studi sull'induzione statistica è stato il riconosciuto "padre della genetica moderna", Gregor Johann Mendel, che a partire dal 1853, elaborando grandi quantitativi di dati sperimentali provò a generalizzare il processo sottostante i dati definendo le ben noti leggi dell'ereditarietà. Nel 1918, in pieno periodo neo-darwiniano, Ronald Fisher dimostrò matematicamente che i caratteri genetici seguono le leggi di Mendel e che si distribuiscono secondo la curva gaussiana. Del resto nella teoria sull'evoluzione della specie, già prima della sua pubblicazione, nel 1859 il famoso biologo inglese, Charles Darwin, si avvalse della collezione ed elaborazione di una moltitudine di dati raccolti in occasione del lungo viaggio intorno al mondo, per affermare una legge universale secondo la quale la specie, sia animale sia vegetale, nasce mediante un processo di "discendenza con variazione". Il cugino, Sir Francis Galton, eclettico studioso di geografia, antropologia, climatologia, genetica, etc., oltre all'introduzione della teoria dell'eugenetica

volta al miglioramento della specie umana, fu tra i promotori della cultura statistica nei diversi campi applicativi, fu considerato il padre della biometria (l'applicazione della statistica nelle scienze biologiche) e della psicometria (l'applicazione della statistica in psicologia). Egli utilizzò il concetto di variazione per introdurre sia il concetto di correlazione, affermando come non vi sia alcuna correlazione tra moralità di un individuo e la sua instabilità morale, sia il concetto di regressione, affermando come al crescere della statura dei padri, lo scarto dalla media della statura dei figli regredisce verso il centro, ossia ciascuna distribuzione della statura dei figli comprende valori più alti e più bassi della media, ma la media delle diverse distribuzioni cresce al crescere della statura dei padri.

Più tardi, il suo amico, il matematico e statistico inglese, Karl Pearson, sviluppò matematicamente tali concetti e generalizzò la problematica ai caratteri qualitativi, introducendo il concetto di indipendenza in distribuzione, si ha ad esempio se i maschi si ripartiscono tra fumatori e non fumatori nell'identica composizione delle femmine, in tal caso si concluderebbe che essere fumatore non dipende dal genere. In caso contrario, si analizza la connessione o dipendenza. In medicina, è rilevante stabilire meccanismi di causa ed effetto, ad esempio dosaggio di un medicinale e reazioni possibili, diagnosi e cura, etc.

Karl Pearson pose le basi della metodologia statistica nell'accezione moderna, assieme ai suoi allievi, Ronald Aylmer Fisher e Jerzy Neyman, allo stesso figlio, Egon Pearson, ai successori, uno su tutti l'ungherese Abraham Wald, definendo una prospettiva teorica riconosciuta dalla comunità degli scienziati e fondata sulle acquisizioni precedenti, un paradigma scientifico della statistica, l'inferenza statistica, indirizzando la ricerca nell'individuazione dei fatti rilevanti da studiare, nella formulazione delle ipotesi e nell'approntamento delle tecniche empiriche necessarie per conoscere la realtà e prendere decisioni in condizioni di incertezza. Fondamentale è il concetto di verosimiglianza o plausibilità di "ciò che appare" dall'osservazione parziale della realtà, il campione statistico, rispetto a "ciò che si ipotizza" sul fenomeno sottostante che ha generato i dati osservati.

Si è portati, infatti, a sostenere che la statistica sia una scienza relativamente giovane, considerando la fine del XIX secolo e la prima parte del XX secolo, il periodo in cui sono state poste le fondamenta della disciplina con elementi distintivi e caratterizzanti la disciplina stessa, rispetto al calcolo delle probabilità e alla matematica applicata. Abraham Wald nel 1939 così si espresse: "La statistica è un gioco a due persone in cui la prima è il matematico e la seconda è la natura".

Per dimostrare l'inconsistenza logica del metodo induttivo, il filosofo e

logico inglese Bertrand Russell raccontò la metafora del "tacchino induttivista":

> Fin dal primo giorno questo tacchino osservò che, nell'allevamento dove era stato portato, gli veniva dato il cibo alle 9 del mattino. E da buon induttivista non fu precipitoso nel trarre conclusioni dalle sue osservazioni e ne eseguì altre in una vasta gamma di circostanze: di mercoledì e di giovedì, nei giorni caldi e nei giorni freddi, sia che piovesse sia che splendesse il sole. Così arricchiva ogni giorno il suo elenco di una proposizione osservativa in condizioni più disparate. Finché la sua coscienza induttivista non fu soddisfatta ed elaborò un'inferenza induttiva come questa: «Mi danno il cibo alle 9 del mattino». Purtroppo, però, questa concezione si rivelò incontestabilmente falsa alla vigilia di Natale, quando, invece di venir nutrito, fu sgozzato.

Il passaggio fondamentale dalla formulazione del problema reale all'osservazione statistica si ebbe con Karl Raimund Popper che valorizzò il ruolo congetturale e percettivo della conoscenza umana nell'osservazione sensibile della realtà, fatta di pregiudizi, intuizioni, ragionamenti, teorie, laddove occorre sapere "cosa osservare" attraverso la formulazione del problema da affrontare. Il metodo scientifico si avvale così di una logica deduttiva basata sul criterio della falsificabilità, in contrapposizione alla verificabilità. In altre parole, da un esperimento non è possibile indurre che una teoria è vera perché è stata verificata empiricamente, bensì è possibile dedurre che una teoria è falsa perché almeno in un caso non si è verificata empiricamente. Invero, già nella teoria della relatività di Albert Einstein nel 1926: "nessuna quantità di esperimenti potrà dimostrare che ho ragione; un unico esperimento potrà dimostrare che ho sbagliato".

Nella seconda metà del XX secolo, grazie al potenziamento della *computer science* e delle tecnologie informatiche a supporto, è stato possibile un cambiamento di rotta nell'inferenza statistica e nella ricerca dei modelli empirici che potessero cogliere gli elementi distintivi e caratterizzanti i fenomeni investigati. Si è passati alle tecniche di visualizzazione per la formulazione di teorie. Il matematico, chimico, statistico americano John Wilder Tukey, già verso la metà del secolo scorso, promuoveva un approccio alternativo all'analisi confermativa dei dati propria del paradigma statistico inferenziale, introducendo l'analisi esplorativa dei dati con l'idea che siano i dati a guidare l'analista nella formulazione di particolari ipotesi riguardanti il fenomeno sottostante, la scelta del modello e dello strumento più appropriato per verificare le ipotesi, altresì per scoprire fatti e regolarità nella struttura dei dati, definire nuove tecniche statistiche commisurate al fabbisogno di risoluzione di problemi reali.

Nasce la *computational statistics*, da distinguersi rispetto allo *statistical computing*. Lo statistico Carlo Lauro dell'Università di Napoli Federico II, ai tempi della sua presidenza dello IASC (*International Association for Statistical Computing*) nel biennio 1993-95, precisò: con lo *statistical computing* "si applicano gli strumenti della computer science alla statistica", con la *computational statistics* "si progettano procedure e algoritmi per implementare metodi statistici *ad hoc* ad uso intensivo del computer, quali ad esempio il *bootstrap* o le simulazioni, per la risoluzione di problemi analiticamente ingestibili prima dell'era del computer". Sulla stessa onda, lo statistico americano della Stanford University, Jerome Friedman, ha stigmatizzato come oggigiorno la statistica sia statistica computazionale, ovvero statistica nell'era del computer.

Lo statistico britannico John Hand è stato tra i promotori principali di questa nuova frontiera della metodologia statistica, coniando l'etichetta di *Data Mining* (i.e., scavare nella miniera di dati): "il processo che attraverso l'impiego di modelli non banali ha l'obiettivo di individuare relazioni tra i dati non banali, nascoste, utili e fruibili dall'utilizzatore." Il *Data Mining* è una fase del più ampio processo di estrazione della conoscenza dalle basi di dati (*Knowledge Discovery from Databases*) con la finalità di estrarre il contenuto informativo utile che porti valore aggiunto, sfruttando appieno l'informazione derivante da quantità sempre crescenti di dati a disposizione nell'epopea digitale dei bit.

Il binomio tra i due mondi, la statistica nell'era del computer o statistica computazionale e il *machine learning* o *computer science,* si è concretizzato in un matrimonio di fatto quando illustri statistici della Stanford University, Trevor Hastie, Robert Tibshirani e Jerome Friedman, hanno pubblicato nel 2001 (in seconda edizione nel 2009) il volume dal titolo *The Elements of Statistical Learning. Data Mining, Inference and Prediction.* Si offre così una trattazione organica e sistematizzata dei principali metodi statistici nel nuovo paradigma contemporaneo che si sta prospettando nella comunità scientifica internazionale, contemplando le problematiche di gestione ed elaborazione informatica dei dati oggigiorno disponibili in grande quantità e in tempo reale.

Già il matematico russo Vladimir Vapnik sul finire del XX secolo sancisce la nascita della *Statistical Learning Theory*, ossia teoria dell'apprendimento statistico, ponendo fine al dibattito tra induzione (dal particolare al generale) e deduzione (dal generale al particolare), con l'introduzione del nuovo concetto di *transduction inference,* dal particolare al particolare.

Sono gli stessi dati a guidarci nella comprensione della realtà e spesso non ci sono leggi universali da scoprire, piuttosto una soluzione generale

costruita sulla base dell'osservazione corrente, caratterizzanti il singolo e non l'universo, da comprendere e analizzare per arricchire il patrimonio informativo e conoscitivo, acquisire esperienza, da utilizzare in un successivo momento di apprendimento per riformulare la soluzione quando l'osservazione si aggiorna con nuovi casi.

A ben vedere, l'introduzione del metodo scientifico risale a Leonardo da Vinci, in pieno Rinascimento italiano; egli sottolineò l'importanza dei due fattori, la sperimentazione empirica, "perché non basta ragionare e fare uso dei concetti se poi non li si mette alla prova", e la dimostrazione matematica, come garanzia di rigore logico: "Nissuna umana investigazione si può dimandare vera scienza, s'essa non passa per le matematiche dimostrazioni."

Così che a rafforzare la teoria di Vapnik, la matematica è servita per comprendere e giustificare il nuovo modo di fare inferenza e previsione.

Sulla scorta dell'evoluzione storica del paradigma scientifico stiamo assistendo a una nuova epopea della statistica metodologica con un nuovo indirizzo scientifico per il futuro, apprendere dai dati rapidamente e utilmente per conoscere e innovare, declinando il trinomio: "statistica", "tecnologia", "analisi dei dati".

2. *Statistica, Tecnologia, Analisi dei Dati*

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 lettura della realtà passa per la definizione dell'unità o entità elementare (individuo, oggetto, caso, etc.) del collettivo oggetto di studio, osservato in forma esaustiva (i.e., popolazione) o parziale (i.e., campione), e per la scelta di uno o più caratteri (ossia dei fenomeni di interesse da investigare), osservando o misurando sperimentalmente, in determinate condizioni, le modalità o espressioni (numeriche o attributi qualitativi) di ciascun carattere per ciascuna unità. Tale osservazione o misurazione definisce il dato statistico. Potranno ritenersi utili eventuali confronti temporali e/o spaziali.

Etimologicamente, Statistica sta per "ciò che è", Tecnologia per "ragionamento sull'arte e il fare", Analisi dei Dati per "soluzione a partire da quantità note".

A ben vedere, il ruolo dello statistico non è solo di produrre e analiz-

zare le statistiche economiche, bensì di osservare *"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. A tal uopo, l'osservazione della realtà non può prescindere dalla "tecnologia", senza confondersi con gli elementi tecnologici a supporto dell'acquisizione e trasmissione dati (chiave USB, centraline, satelliti, sistemi GPS, etc.), del calcolo (computer, sistemi di calcolo parallelo, etc.), altresì della comunicazione e della visualizzazione di immagini e testi (ipod, cinema 3D, TV digitale, etc.).

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 del processo di apprendimento e alla costruzione di un progetto, ossia di una strategia per passare dalla teoria alla pratica, adottando un approccio euristico, quando si procede sperimentalmente attraverso prove ed errori per cercare soluzioni per un problema inedito, altresì un approccio algoritmico, quando si applicano soluzioni note per un problema per gran parte simile ad altri affrontati in precedenza.

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.

"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).

Per dato grezzo si intende un numero o un attributo osservato su ciascuna entità di un collettivo oggetto di studio. L'analisi dei dati offre una soluzione a un problema reale posto, estraendo dai dati grezzi il contenuto informativo utile che porti un valore aggiunto alla comprensione dei fatti della realtà fenomenica sotto osservazione. Tutto ciò consente di formulare previsioni e di prendere decisioni. La comprensione dei fatti da parte dell'utilizzatore finale in grado di coniugare l'informazione (a posteriori)

con l'esperienza personale (a priori), è conoscenza per innovare, ossia produrre cambiamenti e progredire nel processo di apprendimento della realtà nell'ambito in cui si opera.

3. Il processo di apprendimento dai dati statistici

Il percorso scientifico della ricerca quantitativa estrae dai dati il contenuto informativo utile a produrre conoscenza, da utilizzare in momenti decisionali e per fare previsioni. I passaggi fondamentali sono due, dal dato all'informazione e poi dall'informazione alla conoscenza. Il processo si attua declinando le diverse fasi della piramide dell'apprendimento statistico per innovare, descritta in figura:

1. *Formulazione del problema reale*, da parte di chi ha fame di conoscenza, nel contesto applicativo in cui opera.
2. *Brainstorming*, che vede coinvolti tutti i soggetti che potenzialmente possono dare un contributo nell'ambito applicativo in cui si lamenta il fabbisogno conoscitivo, assieme a uno o più statistici in grado di riformulare il problema reale in uno o più quesiti metodologici di tipo statistico.
3. *Ricerca*, che impegna gli statistici nella formulazione della strategia metodologica, individuando i metodi e le tecniche, oltre ai dati necessari per soddisfare il fabbisogno conoscitivo e fornire una soluzione o risposta al problema delineato.
4. *Condivisione della strategia da seguire e del tipo di risultato atteso*, altresì occorre stabilire le condizioni operative temporali, logistiche, finanziarie, etc.
5. *Accessibilità dati*, che consiste nella individuazione dei dati grezzi e nella comprensione delle condizioni e ipotesi per il loro utilizzo.
6. *Filtraggio*, necessario per selezionare i dati grezzi e produrre la base informativa qualitativa e quantitativa utile per applicare il metodo statistico.
7. *Elaborazione*, che consiste nell'applicazione della ricerca in statistica ai dati disponibili fornendo un *reporting* dei risultati ottenuti, trasformando i dati in informazione.
8. *Analisi*, fondamentale per trasformare i risultati statistici in risposte utili per la soluzione del problema reale.

Una volta consolidato il supporto metodologico necessario per rispondere al fabbisogno di conoscenza, l'elaborazione e analisi dei dati approderà in un momento interpretativo di fondamentale importanza per le risposte da

offrire, eventualmente ciò comporta anche un approfondimento scientifico e la reiterazione delle fasi 3-8 del procedimento scientifico di ricerca, partendo da nuovi dati e da una nuova impostazione della ricerca.

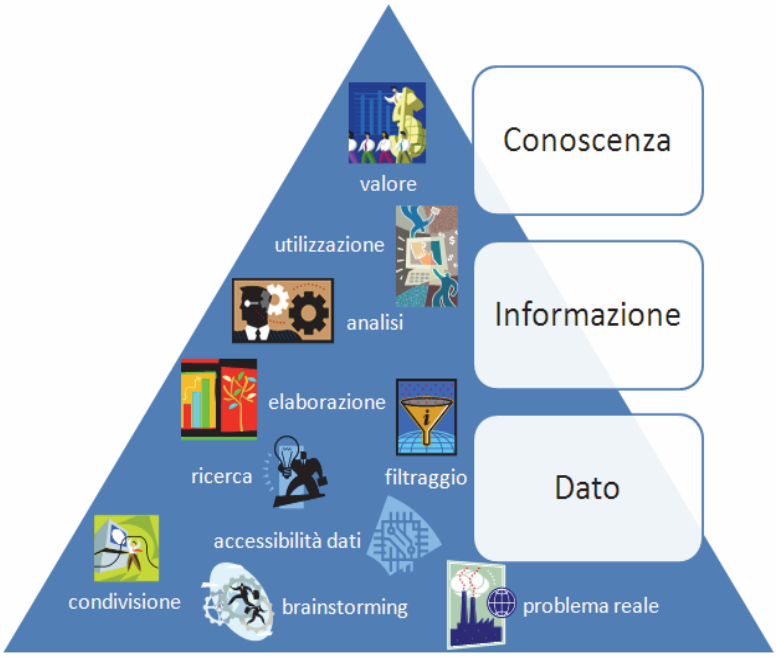

Figura 1: La piramide dell'apprendimento statistico per innovare

9. *Utilizzazione*, che consiste nella traduzione dei risultati statistici in risposte e azioni concrete a supporto del processo decisionale, altresì fornire previsioni, in funzione dell'ambito operativo. Il momento di diffusione dei risultati e il loro utilizzo nella realtà applicativa non possono prescindere dalle ricadute che la conoscenza acquisita ha in termini di "progresso conoscitivo" e innovazione.
10. *Valore*, che trasforma l'informazione acquisita e le azioni concrete in conoscenza per innovare, laddove il valore è espresso in termini monetari (redditività, utile, etc.), e/o di soddisfazione outilità e/o di *performance* o prestazione (efficienza, efficacia, affidabilità, qualità, etc.).

"*We need to tell people that Statisticians are the ones who make sense of the data deluge occurring in science, engineering, and medicine; that Statistics provides methods for data analysis in all fields, from art history to zoology; that it is exciting to be a Statistician in the 21st century because of the many challenges brought about*

by the data explosion in all of these fields." (Nancy Geller, American Statistical Association President, 2011).

Lo statistico è "*Data scientist, who uses both data and science to create something new*".

4. Le applicazioni

I progressi della tecnologia nell'acquisizione e l'elaborazione di dati in formato digitale hanno rappresentato, negli ultimi decenni, la spinta decisiva verso la nascita di un nuovo fabbisogno che potremmo sintetizzare come:

> il desiderio di estrarre conoscenza dall'enorme flusso di dati che oggi sono in grado di immagazzinare ed elaborare con gli strumenti tecnologici di cui dispongo.

La statistica moderna rappresenta la risposta naturale a questo nuovo fabbisogno informativo. Oggi le metodologie statistiche trovano applicazione in tutte le aree riguardanti la vita dell'uomo sia se esse investano le scienze economiche e sociali, sia le scienze della vita fino a considerare le scienze dell'ingegneria e della tecnologia.

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.).

Nell'ambito delle Scienze Economiche e Sociali, le applicazioni della statistica riguardano principalmente le tematiche dell'economia e della finanza, dell'economia aziendale, della sociologia e della psicologia. In tali contesti, la statistica si focalizza principalmente sull'analisi di dati osservazionali rilevati attraverso indagini campionarie che prevedono la somministrazione di questionari ad un campione statistico di unità estratte dalla popolazione di riferimento. Ne rappresentano un esempio, gli studi periodici svolti sulla popolazione per descriverne le evoluzioni delle abitudini di consumo e sul ruolo svolto dalle famiglie nell'accumulazione del risparmio. Ulteriori esempi sono gli studi territoriali per individuare gli elementi di criticità delle filiere produttive o le analisi di *customer satisfaction* effettuate dalle aziende sulla propria clientela.

In questo tipo di ricerche, 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 caratteristiche chiave 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 una 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 altamente rischiose quali l'amniocentesi per quei casi in cui il rischio si attesta su livelli ritenuti altamente tollerabili. Negli esempi forniti risulta 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 della Tecnologie le applicazioni della statistica riguardano numerosissime tematiche quali i trasporti, l'ambiente, la meccanica, l'energia, l'informatica, ecc. Anche in questi contesti applicativi, come nelle scienze della vita, il dato tipico oggetto di studio è di tipo sperimentale.

La statistica fornisce un supporto attraverso un ampio numero di metodologie mirate agli obiettivi più disparati. In ingegneria si impiegano 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.

Un ulteriore esempio del 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 contesto 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).

"*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?*".